The Effect of Fluorination on Protein-Protein Interactions - A Molecular Dynamics Study

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
LEON WEHRHAN
Berlin 2024

The research of this thesis was carried out under the supervision of Prof. Dr. Bettina Keller at the Institute of Chemistry and Biochemistry between December 2019 and June 2024.

1st reviewer: Prof. Dr. Bettina Keller

2nd reviewer: Prof. Dr. Beate Paulus

Date of defense: 26.07.2024

Acknowledgements

First and foremost, I would like to thank my first reviewer and supervisor, Prof. Dr. Bettina Keller, for the opportunity to join her research group and for mentoring me throughout my doctoral studies. I am deeply grateful for all the helpful guidance and ideas, her approachable mentoring style, her encouragement and motivation, her trust in my abilities and the freedom to pursue my ideas, as well as her commitment to her doctoral students and skillful organization of the research group.

I would like to thank Prof. Dr. Beate Paulus for being the second reviewer of this thesis and for her helpful advice, her approachability and the pleasant atmosphere established in-between the research groups of theoretical chemistry.

Next, I want to thank Prof. Dr. Beate Koksch and Jakob Leppkes for the friendly and productive collaboration on our research project.

I would also like to thank Prof. Dr. Jörg Rademann, Markus Tiemann and Matteo Accorsi as well as Prof. Dr. Hajo Kries and Maximilian Müll for the fruitful collaboration and the opportunity to work on these exciting research projects.

Furthermore, I thank my colleagues of the theoretical chemistry at FU Berlin for thoughtful discussions and for making my doctoral studies a truly enjoyable time.

Finally, I would like to thank my friends, my family and my girlfriend for always supporting me through this time.

Authorship Declaration

I, Leon Wehrhan, hereby declare that I alone am responsible for the content of my doctoral dissertation and that I have only used the sources or references cited in the dissertation.

Berlin, 17.06.2024

Contents

List of Publications

1. "Fluorinated Protein-Ligand Complexes - A Computational Perspective"
   Leon Wehrhan, Bettina G. Keller
   J. Phys. Chem. B **2024**
   Accepted manuscript

2. "Water Network in the Binding Pocket of Fluorinated BPTI–Trypsin Complexes–Insights from Simulation and Experiment"
   Leon Wehrhan, Jakob Leppkes, Nicole Dimos, Bernhard Loll, Beate Koksch, Bettina G. Keller
   J. Phys. Chem. B **2022**
   DOI: 10.1021/acs.jpcb.2c05496

3. "Pre-bound State Discovered in the Unbinding Pathway of Fluorinated Variants of the Trypsin-BPTI Complex Using Random Acceleration Molecular Dynamics Simulations"
   Leon Wehrhan, Bettina G. Keller
   J. Chem. Inf. Model. **2024**
   DOI: 10.1021/acs.jcim.4c00338

4. "Pentafluorophosphato-Phenylalanines: Amphiphilic Phosphotyrosine Mimetics Displaying Fluorine-Specific Protein Interactions"
   Matteo Accorsi, Markus Tiemann, Leon Wehrhan, Lauren M. Finn, Ruben Cruz, Max Rautenberg, Franziska Emmerling, Joachim Heberle, Bettina G. Keller, Jörg Rademann
   Angew. Chem. Int. Ed. **2022**
   DOI: 10.1002/anie.202203579

5. "Biosynthetic Incorporation of Fluorinated Amino Acids into the Nonribosomal Peptide Gramicidin S"
   Maximilian Müll, Farzaneh Pourmasoumi, Leon Wehrhan, Olena Nosovska, Philipp Stephan, Hannah Zeihe, Ivan Vilotijevic, Bettina G. Keller, Hajo Kries
   RSC Chem. Biol. **2023**
   DOI: 10.1039/D3CB00061C

Abstract

Fluorine's outstanding capability to influence physico-chemical properties of organic molecules make fluorine substitution an attractive strategy to modulate the binding affinities in protein-ligand and protein-protein systems. The direction and magnitude with which fluorine influences the binding affinity is often difficult to predict, as multiple effects may play a role. Computational methods and Molecular Dynamics (MD) simulations in particular are an excellent tool to study such effects, as they allow for atomic precision and in-depth insights into the multiple components of the binding affinity. I have compiled selected aspects of using computational methods to study fluorination in protein systems in a perspective article, which is part of this thesis. Furthermore, I use MD simulations to study how fluorine impacts the binding properties of selected protein-protein or protein-amino acid systems. The first system is the complex of trypsin with fluorinated variants of Abu-BPTI. In Abu-BPTI, the crucial amino acid Lys15 is replaced by the shorter aliphatic amino acid Abu, which comes with a loss of inhibitor strength. Some of the inhibitor strength can be restored by fluorination. An MD based study of the water molecules in the binding pocket of these protein complexes revealed a highly dynamic water network, with strongly interconnected water molecules. The fluorinated moiety of Abu does not interact with the water molecules, making it unlikely that the inhibitor activity is restored via water mediated bonds between trypsin and fluorine. The analysis of the unbinding paths using Random Acceleration MD (RAMD) revealed a novel metastable state for all of the studied variants, which we call the pre-bound state. This state is clearly distinct from the fully bound state in position and orientation and is stable in long MD simulations. Moreover, the states differ in their interaction pattern, with fluorine possibly having a stabilizing effect on the formation of the pre-bound state. The second system studied here is the protein PTP1B in complex with a phosphotyrosine mimetic with a highly fluorinated headgroup. I characterized the binding pose and fluorine specific interactions between the mimetic and the protein binding pocket residues. The third system is the GrsA A domain in complex with fluorinated phenylalanine variants. Here, I characterized the interruption of an aromatic interaction by fluorination. With respect to the impact of fluorination on the interaction of proteins, the results of this thesis show no evidence for direct water mediated interactions between proteins and fluorine, but indicate that fluorine might influence pre-bound intermediates. For future research it might be interesting to investigate the effects on other components of binding affinities, such as binding entropy. Furthermore, not only the

impact of fluorination on fully bound states should be investigated, but also the impact on possible pre-bound states.

Kurzzusammenfassung

Die herausragende Fähigkeit von Fluor, die physikalisch-chemischen Eigenschaften von organischen Molekülen zu beeinflussen, machen Fluor-Substitution zu einer attraktiven Strategie um Bindungsaffinitäten von Protein-Ligand und Protein-Protein Systemen zu modulieren. Die Richtung und der Stärke, mit der Fluor die Bindunsaffinität beeinflusst, ist oft schwer vorherzusagen, da mehrere Effekte eine Rolle spielen könnten. Computergestützte Methoden, und besonders Moleküldynamische (MD) Simulationen, sind ein hevorragendes Mittel um solche Effekte zu untersuchen, denn sie erlauben atomare Präzision und tiefgehende Einblicke in die verschiedenen Komponenten der Bindungsaffinität. In einem Perspective-Artikel habe ich ausgewählte Aspekte der Benutzung von computergestützten Methoden zur Untersuchung von Fluorierung in Proteinsystemen zusammengestellt, welcher Teil von dieser Thesis ist. Außerdem nutze ich MD Simulationen um zu untersuchen wie sich Fluor auf die Bindungseigenschaften von ausgewählten Protein-Protein und Protein-Aminosäure Systemen auswirkt. Das erste System ist der Komplex aus Trypsin und fluorierten Varianten von Abu-BPTI. In Abu-BPTI wurde die wichtige Aminosäure Lys15 ersetzt durch die kürzere und aliphatische Aminosäure Abu, was mit einem Verlust von Inhibitorstärke einher geht. Ein Teil der Inhibitorstärke kann durch Fluorierung wiederhergestellt werden. Eine MD basierte Studie der Wassermoleküle in der Bindungstasche dieser Proteinkomplexe zeigte ein hoch-dynamisches Wassernetzwerk, mit stark vernetzten Wassermolekülen. Der fluorierte Bestandteil von Abu interagiert nicht mit den Wassermolekülen, was es unwahrscheinlich macht, dass die Inhibitoraktivität durch Wasser-vermittelte Bindungen zwischen Trypsin und Fluor wiederhergestellt wird. Die Analyse des Losbindungspfads mit RAMD dekte für alle untersuchten Varianten einen neuen metastabilen Zustand auf, den wir als vorgebundenen Zustand bezeichnen. Dieser Zustand unterscheidet sich in Position und Orientierung deutlich vom vollständig gebundenen Zustand und ist in langen MD Simulationen stabil. Außerdem unterscheiden sich die Zustände in ihrem Interaktionsmuster, wobei Fluor möglicherweise eine stabilisierende Wirkung auf die Bildung des vorgebundenen Zustands hat. Das zweite hier untersuchte System ist das Protein PTP1B im Komplex mit einem Phosphotyrosin-Mimetikum mit einer stark fluorierten Kopfgruppe. Ich habe die Bindungsposition und die fluorspezifischen Wechselwirkungen zwischen dem Mimetikum und den Aminosäuren der Proteinbindungstasche charakterisiert. Das dritte System ist die GrsA A-Domäne im Komplex mit fluorierten Phenylalaninvarianten. Hier habe ich die Störung einer aro-

matischen Wechselwirkung durch Fluorierung charakterisiert. Was die Auswirkungen der Fluorierung auf die Interaktion von Proteinen betrifft, so zeigen die Ergebnisse dieser Arbeit keine Hinweise auf direkte, durch Wasser vermittelte Wechselwirkungen zwischen Proteinen und Fluor. Sie deuten aber darauf hin, dass Fluor möglicherweise vorgebundene Zustände beeinflusst. Für die zukünftige Forschung könnte es interessant sein, die Auswirkungen auf andere Komponenten von Bindungsaffinitäten, wie die Bindungsentropie zu untersuchen. Außerdem sollten nicht nur die Auswirkungen von Fluorierung auf vollständig gebundene Zustände untersucht werden, sondern auch die Auswirkungen auf mögliche vorgebundene Zustände.

1 Introduction

Fluorine is a unique element in general chemistry and has an exceptional standing within the main group elements due to its very high reactivity in its elemental form and its exceptionally high electronegativity.[1] This high electronegativity, together with its small atomic size and low polarizability make fluorine a powerful modulator of physicochemical properties of organic molecules.[2–5] In these organic molecules, the substitution of a C-H bond with a C-F bond is highly interesting, as the C-F bond is very strong, only slightly longer than the C-H bond and reverses the dipole moment of the bond.[4] These properties of the C-F bond can be used to modulate key properties of organic molecules, like for example the pKa, lipophilicity, conformation or biologic effects such as cell penetration and metabolism of the molecules.[2,4,5] These properties are especially interesting in a medicinal chemistry context and the importance of fluorine in medicinal chemistry is demonstrated by the fact that more than 20% of globally registered drugs contain at least one fluorine.[6,7] Naturally, a highly interesting application of modulating molecular properties through fluorine substitution for medicinal chemistry is tuning protein binding properties. The rational design of these properties can be challenging, as the impact that fluorination has is not always straightforward. To demonstrate this point, I will describe examples of fluorine impacting the binding properties of protein binding molecules in a variety of ways, in the next sections of this chapter.

Computational methods are a possible option to get a better understanding of the effect fluorine can have on protein binding properties. One such method that has proven itself to be particular useful to study protein structure and dynamics, binding strength of protein-ligand and protein-protein systems, kinetic and thermodynamic properties is Molecular Dynamics (MD) simulations.[8–11] MD simulations make use of a classical approximation of a molecular system, using force field parameters and an integrator, to simulate the systems of interest and obtain a time trajectory of the positions.[12,13] These trajectories have atomistic precision, which allows for a particularly detailed insight into the molecular properties. The first MD simulation of a protein system was that of Bovine Pancreatic Trypsin Inhibitor (BPTI) and was only 9.2 ps long.[14] Nowadays, MD simulations of protein-protein complexes on the microsecond scale can be considered routine tasks on modern GPU systems and using specialized supercomputers, such as ANTON3, simulations on the millisecond scale of larger biological systems is possible.[15] However, even with modern supercomputers, MD simulations are inherently limited in their accessible simulation timescales as the fundamental integrator step needs to be small to ensure stability. This makes it difficult to study events like protein-ligand binding or protein folding, as these events usually happen on larger timescales than what is accessible with MD simulations. Enhanced Sampling methods help overcome this issue by accelerating MD simulations with respect to the processes of interest and providing tools to calculate useful information, like binding free energy, from the results of the simulations.[16]

In the following sections of this chapter, I will put a special emphasis on MD simulation methods in studying how fluorine impacts the properties of protein binding molecules.

1.1 Fluorine as a Hydrogen Bond Acceptor

Hydrogen bonds are one of the major driving forces of the binding of protein-ligand and protein-protein complexes.[17] Given the high electronegativity of fluorine it seems natural to assume that fluorine is a strong acceptor of hydrogen bonds. However, fluorine can be described as a weak hydrogen bond acceptor in the context of organic molecules and protein binding.[18,19] An attempt to explain this weak hydrogen bond acceptor ability is that the high electronegativity of fluorine only causes an initial attraction for external electronic charge, but the low charge capacity and low polarizability limit the amount of charge that fluorine can really absorb.[19] In consequence, this limits the magnitude of the most negative potentials $V_{\mathrm{S,min}}$ which organic fluorines can achieve and which can be considered as effective means to analyze and interpret hydrogen bonding.[19,20] For example, $V_{\mathrm{S,min}}$ for the oxygen of ethanol is -154 kJ/mol and for fluorine in fluoroethane it is -111 kJ/mol.[19] Considering the shortcomings of fluorine as a hydrogen bond acceptor, Dalvit et al. suggest that hydrogen bonds with fluorine as acceptor would only contribute to the overall binding affinity of a protein-ligand complex if they happen in an environment shielded from water and with no other acceptors available.[18]

Nonetheless, hydrogen bonds with fluorine as acceptor are not uncommon protein-ligand complexes, as showcased by many recent examples where these interactions have been observed for protein targets like FXa,[21] μ-Opioid-Receptor,[22] S1P receptor,[23] Akt1,[24] HIV protease,[25] Tyrosinase,[26] Bruton's tyrosine kinase,[27] Janus kinases[28] or the SarS-Cov2 main protease.[29–31] To assess whether these kind of hydrogen bonds make a contribution to binding strength, Pietrus et al. conducted a thorough survey of Protein DataBank (PDB) structures of aliphatic fluorines of small molecules bound to proteins.[32] They calculated interaction energies using Quantum Mechanics methods considering isolated pairs of donor and acceptors. They found interaction strengths between 0 and -5.02 kJ/mol.[32] Fluorine was observed to be an acceptor for hydrogen bonds from OH, NH and CH donors, with CH being the most frequent. Naturally, the magnitude of the interaction strength of fluorine hydrogen bonds depended on the donor-acceptor distance, but it was not dependent on the hydrogen bond angle.[32] When comparing the calculated energy minima of the isolated donor-acceptor pairs with the geometries actually observed in the PDB, the authors found that the geometries in actual protein-ligand systems did not match the calculated energetic minima.[32] This lead them to the conclusion that hydrogen bonds with fluorine as acceptor are likely not the driving force of binding affinity, but that these bonds form in addition to stronger interactions, that stabilize the protein-ligand complex.[32]

1.2 Disrupting Water Networks with Fluorine

Water is essential for the binding properties of protein systems,[33] meaning that any tool that impacts the hydration properties of a protein-protein or protein-ligand system can be considered a modulator of protein binding properties as well. Fluorine substitution could be such a tool, as fluorinated molecules, more specifically fluorinated amino acids, have been shown computationally to influence hydration energies through a variety of effects, such as hydrogen bond like interactions and hydrophobicity.[34–36] There is also experimental evidence for fluorinated molecules impacting the dynamics of water molecules, as ultrafast fluorescence spectroscopy experiments show that fluorinated amino acids can slow down the motion of water.[37]

Robalo et al. were researching how a series of fluorinated aliphatic amino acids would establish their hydrophobicity and compare it to the unfluorinated aliphatic amino acids.[35] Among these amino acids is α-aminobutyric acid (Abu), its trifluorinated variant trifluoro-ethylglycine (TfeGly) and, in a later study[34] the partially fluorinated amino acids difluoro-ethylglycine (DfeGly) and monofluoro-ethylglycine (MfeGly). They find that changes to the hydration free enrgy caused by the stepwise fluorination of Abu is composed of a variety of contributing and sometimes opposing effects like the change in surface area, changes in hydrogen bonds between the amino acid backbone carbonyls and amine groups, changes in the electrostatic potential and direct hydrogen bond like interactions between fluorine and water.[34] Moreover, fluorine cannot only affect its solvation shell by establishing hydrogen bond like interactions, but also by creating so-called "dangling" waters, as was found by studying fluorinated alcohols.[36] These dangling waters cannot find a binding partner in their vicinity and are entropically stabilized.[36]

The unique ways in which fluorine can interact with water and thereby possibly disrupt water networks and hydration shells becomes apparent in several examples of protein-ligand systems, described in the following.

One example is the μ-opioid receptor in complex with fentanyl and a mono-fluorinated variant of fentanyl.[22] The μ-opioid receptor hosts two clusters of complex water mediated hydrogen bond networks, whose specific structure depends on pH and also on the type of ligand that is bound.[22] The fluorination of fentanyl at just one single site causes multiple significant changes in the water mediated hydrogen bond network and also changes the conformation of the bound ligand.[22]

In other examples, fluorine impacts the binding energetics of protein-ligand systems by not only influencing the enthalpic contributions of the protein hydration shell, but also the entropic contributions. In the case of human carbonic anhydrase, bound to a series of similiar ligands, which only differ by fluorine substitution, the enthalpic and entropic contributions to the binding affinity vary widely across the series of ligands.[38] This is despite the fact that the overall binding affinity is similar.[38] These changes in energy contributions can be linked to the water network in the hydration shell of the protein, which is disrupted by fluorine in such a way that it causes less restricted water motion.[38]

Disruptions in water hydration shells can not only lead to enthalpy-entropy compensation, but also entropy-entropy compensation, as becomes apparent in the example of Galectin-3 bound to fluorinated inhibitors.[39] In this example, the comparison of three mono-fluorinated inhibitors, which only differ by the position of the fluorine substitution, shows that the entropic contribution to the binding affinity is realized in very different ways.[39] Especially one ligand is almost entirely stabilized by changes to the water entropy.[39]

Having highlighted the unique and sometimes unexpected ways in which fluorine can interact with water networks in protein-ligand systems, I will describe an example where such an interaction is discussed and that is particularly relevant for this thesis: The protein-protein complex of trypsin, bound to variants of its natural inhibitor BPTI.

In the first publication presented in this thesis, I review computational approaches to model fluorine in protein environments, including fluorine as hydrogen bond acceptor and its ability to disrupt water networks, in more detail.

1.3 Trypsin and BPTI Variants

Trypsin is an enzyme, that belongs to the subclass of serine proteases. Serine proteases are enzymes, that catalyze the cleavage of peptide bonds and are of utmost im-

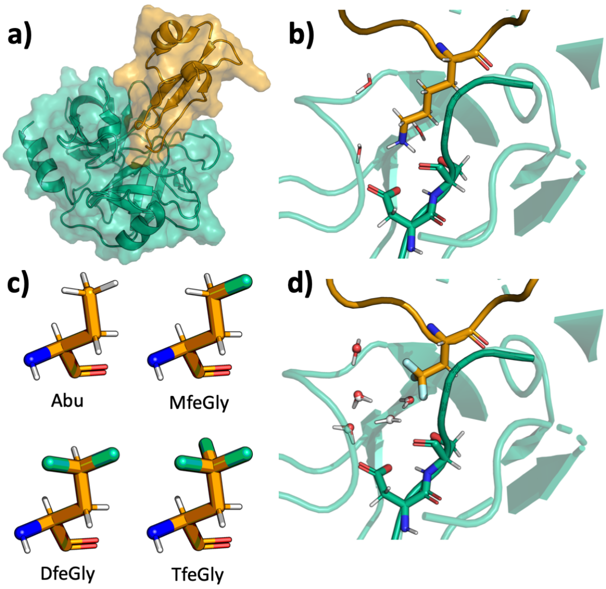


Figure 1.1: (a) The structure of the complex between trypsin (green) and BPTI (orange). (b) The S1 binding pocket of the complex between trypsin and wildtype BPTI. (c) The four Abu variants: Abu (unfluorinated), MfeGly (mono-fluorinated), DfeGly (di-fluorinated) and TfeGly (tri-fluorinated). (d) The S1 binding pocket of the complex between trypsin and TfeGly-BPTI.

portance in various vital processes, like protein degradation, signalling pathways and in pathologies such as cancer or cardiovascular diseases.[40,41] Trypsin is a particularly interesting example of a serine protease, as the trypsin-fold is the most prevalent fold for proteases in higher organisms.[42,43] The active site of trypsin, which is the catalytic triad that catalyzes the proteolytic cleavage, is located at the rim of a deep binding pocket.[44] Binding to the protein substrates is then realized by binding of long positively charged amino acid side chains, such as Lys or Arg, to the negatively charged bottom of this binding pocket.[44] The amino acid residue that reaches into this pocket is called the P1 residue and the deep binding pocket is called the S1 pocket.[45] Some proteins, like BPTI, that bind to trypsin are not cleaved, but instead act as an inhibitor.[46,47] The reason why BPTI is not cleaved by trypsin is discussed to be fast re-ligation of the peptide bond[46] or a "clogged gutter" mechanism, where the cleaved parts of the inhibitor form very stable complexes to trypsin, so that product release is hindered.[47]

BPTI binds to trypsin, among other interactions in the protein-protein binding interface, by reaching with the side chain of its P1 residue Lys15 into the S1 pocket of trypsin and forming water mediated bonds to the negatively charged amino acid residue Asp189 and Ser190 at the bottom of the S1 pocket (see fig. 1.1a-b).[48] The importance of these specific interactions of the Lys side chain is demonstrated by studying variants of the BPTI-trypsin complex, where Lys15 is mutated to Ala (K15A).[49] Here, the binding affinity of BPTI towards trypsin is dramatically decreased.[49] Trypsin and BPTI do not only form a very stable complex in their fully bound conformation, but it is likely that the two proteins already form more loosely bound encounter states upon recognition.[50]

Given the importance of the interactions of Lys15 and the residues at the bottom of the S1 binding pocket, it is not surprising that the inhibitor activity of BPTI is strongly reduced when Lys15 in BPTI is replaced with the shorter and aliphatic amino acid Abu. Interestingly, when Abu is replaced with the mono- di- and tri-fluorinated variants MfeGly, DfeGly and TfeGly (see fig. 1.1c), the inhibitor activity is partially restored in a stepwise manner, as evidenced in experimental inhibition assays.[51] Ye et al. propose a hypothesis for the reason of this inhibitor strength restoration, that is based on interactions of fluorine with the water molecules in the S1 pocket.[51] They label these water molecules with the first alphabetic letters A-E.[51] In the wildtype-BPTI-trypsin complex, two of these water molecules (A and E) are present at the bottom of the S1 pocket and are involved in the interactions of Lys15[48] and another water molecule (D) can be found towards the outside of the pocket.[51] In complexes of Abu-BPTI and its fluorinated variants with trypsin, there are two additional water molecules (B and C) in the S1 pocket, which are in direct proximity of the (fluorinated) Abu side chain (compare fig. 1.1b and fig. 1.1d).[51] The hypothesis, which is based on the measurement of B-factors in X-ray crystallography structures of the complexes, states that fluorine binds to the water molecules in the S1 pocket in a hydrogen bond like way and this bond is then extended through the water network to binding partners at the bottom of the S1 pocket.[51]

2 Research Question

This thesis is concerned with understanding how the introduction of fluorine into protein-protein interaction sites influences the properties of these interactions. As described in the last chapter, fluorine has been shown to influence the interactions of proteins with their binding partners in a great variety of ways, including direct hydrogen bond like interactions, inducing conformational changes, entropic effects and possibly through interactions with water molecules. Even if such an effect of fluorination is discovered, e.g. the enhancement of a protein inhibitor by fluorine substitution, and it is thoroughly studied, e.g. by structural experiments, it may be difficult to rationalize how the introduction of fluorine impacts the experimental observables. To tackle this problem, I use computational methods, more precisely methods based on MD simulations. These simulations are a well-recognized tool in the study of protein-ligand and protein-protein interactions and allow atomistic insight of the structure and dynamics of the studied system. MD simulations can provide understanding of structural, dynamic and also energetic properties of protein-protein systems, which makes them a particularly well suited tool to study the influence of fluorination on protein-protein interactions.

In the main project of this thesis, I focus on the protein-protein system of the enzyme trypsin bound to variants of its inhibitor BPTI. The variants are the same as described in the introduction, where the crucial amino acid Lys15 of wildtype-BPTI is replaced by Abu and its fluorinated variants. Inhibition assays show that while the substitution of Lys15 with Abu causes a substantial decrease in inhibitor strength, this effect can be partially reverted by fluorination of Abu. Moreover, X-ray crystallography shows additional water molecules in the main binding pocket of the system and B-factors indicate altered water dynamics after fluorination. The main project of this thesis is concerned with rationalizing these findings through computational insights. I will at first focus on the water molecules in the main binding pocket of the protein-protein complex and then analyze the unbinding process of the BPTI variants from trypsin.

In the rest of the thesis, I focus on two systems of a protein bound to single fluorinated amino acids. The first system is the protein tyrosine phosphatase PTP1B in complex with fluorinated phosphotyrosine mimetics. The second system is the A domain of the nonribosomal peptide synthetase GrsA in complex with fluorinated phenylalanine residues. While both complexes are not exactly a protein-protein complex, they feature an amino acid as their second binding partner, which means they are still relevant for studying protein-protein interactions. Just like for the system of the main project, there is experimental evidence for fluorine having a significant impact on the binding between the protein and the substrate/inhibitor. In the case of PTP1B, a highly fluorinated headgroup has been designed and shown to increase binding affinity of the phosphotyrosine mimetic, possibly by leveraging fluorine specific interactions. In the case of GrsA, it can be shown in activity assays that the A domain rejects 4-fluorinated phenylalanine as a substrate. For both of these cases, my research in the context of this thesis attempts to rationalize the influence of fluorine on the experimentally observed effects using Molecular Dynamics and structural modeling techniques.

In essence, this thesis aims at a better understanding of the question: How does fluorine influence the interactions of proteins?

3 Theory

3.1 Molecular Dynamics

The term Molecular Dynamics (MD) describes simulation methods, that propagate a molecular system $\mathbf{x}$, comprised of its N particles $\mathbf{x}_i$ in time to form a trajectory

$$\mathbf{x}(t) = (\mathbf{x}(0), \mathbf{x}(\Delta t), ..., \mathbf{x}(N_t \cdot \Delta t)) \tag{3.1}$$

with the timestep Δt and the number of steps N_t.

On a fundamental theoretical level, molecular systems consist of atomic nuclei and electrons and their evolution in time is described by the time-dependent Schrödinger equation.[12] In MD, the assumption is made that atoms can be described as point-particles and evolve in time according to classical mechanics. Following Newton's second law of motion, the time evolution of the point-particles, with mass m can be described as:

$$m_i \cdot \ddot{\mathbf{x}}_i = m_i \cdot \frac{\partial^2 \mathbf{x}_i}{\partial t^2} = \mathbf{F}_i = -\nabla_{\mathbf{x}_i} V(\mathbf{x}). \tag{3.2}$$

The potential energy function $V(\mathbf{x})$ governs how the particles interact with each other. The functional form and the set of parameters from which the potential energy is constructed is also called the force field. The selection of a force field that is appropriate for correctly capturing the physical properties of the simulated system is crucial for any MD simulation (see section 3.3). The propagation of the molecular system in time to obtain a trajectory as in eq. 3.1 works by integrating eq.3.2 numerically. The foundation of the numerical algorithm is the Taylor series of the position of the molecular system $\mathbf{x}$ around a point in time t.[12] If we consider the Taylor expansion of the classical trajectory of the atomic position $\mathbf{x}_i$ at time t forward and backward in time:[12]

$$\mathbf{x}_i(t + \Delta t) = \mathbf{x}_i(t) + \dot{\mathbf{x}}_i(t)\Delta t + \frac{1}{2}\ddot{\mathbf{x}}_i(t)\Delta t^2 + \frac{1}{6}\dddot{\mathbf{x}}_i(t)\Delta t^3 + \mathcal{O}(\Delta t^4) \tag{3.3}$$

and

$$\mathbf{x}_i(t - \Delta t) = \mathbf{x}_i(t) - \dot{\mathbf{x}}_i(t)\Delta t + \frac{1}{2}\ddot{\mathbf{x}}_i(t)\Delta t^2 - \frac{1}{6}\dddot{\mathbf{x}}_i(t)\Delta t^3 + \mathcal{O}(\Delta t^4) \tag{3.4}$$

we can add the two equations to obtain:[12]

$$\mathbf{x}_i(t + \Delta t) = 2\mathbf{x}_i(t) - \mathbf{x}_i(t - \Delta t) + \ddot{\mathbf{x}}_i(t)\Delta t^2 + \mathcal{O}(\Delta t^4). \tag{3.5}$$

By neglecting higher order terms and using eq.3.2 to substitute for $\ddot{\mathrm{x}}_i$, we get to the update rule for the numerical integration:[12]

$$\mathbf{x}_i(t + \Delta t) = 2\mathbf{x}_i(t) - \mathbf{x}_i(t - \Delta t) + \frac{\mathbf{F}_i(t)}{m_i}\Delta t^2. \tag{3.6}$$

This algorithm is called the Verlet integrator.[52] It is not necessary to calculate the particle velocities for the trajectory, however they are needed to calculate the instantaneous temperature and can be calculated at half-steps:[12]

$$\mathbf{v}_i\left(t + \frac{\Delta t}{2}\right) = \frac{\mathbf{x}_i(t + \Delta t) - \mathbf{x}_i(t)}{\Delta t}. \tag{3.7}$$

Another integrator algorithm that can be derived from the Verlet integrator and is numerically more precise, is the Leap-Frog integrator.[53] The idea of this algorithm is to compute the velocities at half-integer timesteps and use these velocities to calculate the new positions at full integer timesteps.[13] The name comes from the calculations of positions and velocities "leaping" over each other to be computed at full- and half-integer timesteps, respectively. By rearranging eq.3.7, we get the update rule for positions of the Leap-Frog integrator:[13]

$$\mathbf{x}_i(t + \Delta t) = \mathbf{x}_i(t) + \Delta t \cdot \mathbf{v}_i\left(t + \frac{\Delta t}{2}\right). \tag{3.8}$$

The following update rule for velocities can be obtained from the Verlet algorithm:[13]

$$\mathbf{v}_i\left(t + \frac{\Delta t}{2}\right) = \mathbf{v}_i\left(t - \frac{\Delta t}{2}\right) + \Delta t \cdot \frac{\mathbf{F}_i}{m_i}. \tag{3.9}$$

The instantaneous temperature can be calculated from the particle velocities:[12]

$$T(t) = \frac{2}{3Nk_B}\sum_{i=1}^{N}\frac{1}{2}m_i(\mathbf{v}_i)^2 \tag{3.10}$$

with the Boltzmann constant k_B and the number of particles N.

The integrator algorithms need a timestep Δt that is chosen to be sufficiently small to obtain stable simulations. For simulations of flexible molecules, a timestep that is ten times smaller than the fastest bond vibration is recommended.[54] In biomolecular systems, this fastest vibration is that of the C-H bond at a timescale of approximately 10 fs.[54] That would mean a time step of 1 fs has to be used. However, when running simulations of biomolecular systems, one is generally not interested in the C-H bond vibrations, so the C-H bonds are often constrained in their length throughout the simulation. This way the timestep can be increased to 2 fs. Bond constraints are realized by correcting the bond length after an unconstrained simulation step, which leads to a problem of undetermined Lagrange multipliers, which is solved by the LINCS algorithm.[55]

The simulation methods above study the natural time propagation of a system of particles according to Newton's laws of motion. This means the total energy of the system E is conserved. As the number of particles N and the volume V are also constant, the simulation samples states of the molecular system in the microcanonical (NVE) ensemble. As most experiments are carried out at constant temperature instead of at constant energy, the simulator is usually motivated to run their simulations in the NVT or NPT ensemble. To achieve this, an additional algorithm called the thermostat (for constant temperature) and/or barostat (for constant pressure) is needed.

While it is technically possible to scale the particle velocities, so the instantaneous temperature (eq. 3.10) is always at the target temperature, this is generally not recommended as this does not produce the correct temperature fluctuations and therefore does not simulate the true NVT ensemble.[13] Another option, which used to be widely used despite not being associated with a well-defined ensemble,[56] would be the Berendsen thermostat.[57] This thermostat algorithm models weakly coupling the system to a heat bath. The particle velocities are rescaled by introducing a scaling factor $\gamma(t)$, that is dependent on the deviation between the instantaneous temperature $T(t)$ and the target temperature T_0. The scaling factor is chosen so that $T(t)$ decays towards T_0 with a first-order rate:[57]

$$\frac{dT(t)}{dt} = \frac{T(t) - T_0}{\tau} \tag{3.11}$$

with the time-scale parameter τ.

The Berendsen thermostat allows for some temperature fluctuation, but it still violates the equipartition theorem and therefore does also not produce the correct NVT ensemble.

An improved version of the Berendsen thermostat is the velocity rescaling thermostat with a stochastic term by Bussi et al.,[56] where the change of the kinetic energy dK follows the equation:[56]

$$dK = (\bar{K} - K)\frac{dt}{\tau} + 2\sqrt{\frac{K\bar{K}}{N_f}\frac{dW}{\sqrt{\tau}}} \tag{3.12}$$

with the target kinetic energy $\bar{K}$.

For simulating at constant pressure, the volume of the simulation box has to be modified. This can be achieved with the Parrinello-Rahman barostat,[58] which couples the system to a fictitious pressure bath and allows for anisotropic scaling of shape and size of the simulation box.

Ensemble averages of an observable A of the simulated system can be calculated as time averages $\bar{A}$ of sufficiently long MD simulations of N_t snapshots t_k:[12]

$$\bar{A} = \frac{1}{N_t} \cdot \sum_{k=1}^{N_t} A(t_k) \tag{3.13}$$

Under the assumption of the ergodic hypothesis, a fundamental axiom of statistical mechanics, the time average (in the limit of infinite simulation time) of a system with N particles is equal to the ensemble average $\langle A \rangle$:[54]

$$\bar{A} = \langle A \rangle = \int \int A(\mathbf{p}^N \mathbf{x}^N) \rho(\mathbf{p}^N \mathbf{x}^N) d\mathbf{p}^N d\mathbf{x}^N \tag{3.14}$$

with the particle momenta $\mathbf{p}$, particle positions $\mathbf{x}$ and the probability density ρ to find the system at the position in phase space where the particles have the positions $\mathbf{x}$ and momenta $\mathbf{p}$.

This relation makes MD simulations suitable for calculating equilibrium properties of the simulated system. These properties include e.g. structural properties of proteins, such as amino acid side chain dihedrals or backbone conformation. For example, an MD simulation can be used to calculate the Root Mean Square Fluctuation (RMSF) of a simulated particle i with the time-average $\mathbf{x}_i$:[59]

$$RMSF = \sqrt{\frac{1}{T} \sum_{t=0}^{T} (\mathbf{x}_t - \langle \mathbf{x}_i \rangle)^2} \tag{3.15}$$

3.2 Enhanced Sampling

One of the main limitations of MD is that the integration timestep needs to be small enough for the integration to be stable and accurate. This limits the accessible total simulation time significantly, meaning that reachable timescales can be shorter than what is needed to study slow processes of interest.[16] Generally, it is not possible to sample processes with free energy barriers much higher than the thermal energy $k_B T$, because in this case there is a negligible amount of sampling points observed in the barrier region during equilibrium MD simulations.[13] This is particularly relevant for biomolecular simulations, as biological molecules often have rough conformational free energy landscapes with local minima separated by high free energy barriers.[60,61] Binding and unbinding of drug-like ligands to proteins are also such slow processes, that are hard to sample with unbiased MD, as demonstrated by k_{off} reaction rates of standard protein-ligand systems like trypsin/benzamidine with $k_{\text{off}} = 600 \pm 300\ s^{-1}$, or

μ-opiod-receptor/morphine with $k_{\text{off}} = 2.3 \pm 0.2\ s^{-1}$.[8] The dissociation of protein-protein complexes is also often a slow process, as demonstrated by the complex of trypsin and BPTI with $k_{\text{off}} = 5 \cdot 10^{-8}\ s^{-1}$.[49]

The term "enhanced sampling" collects a great variety of methods designed to overcome this sampling problem of MD described above.[8,16,60,62,63] Many enhanced sampling methods are based on reducing the high-dimensional simulation data to a set of a few dimensions, which can be calculated from the simulation data, called collective variables (CVs).[16,63] CVs are functions of the atomic coordinates x of the simulation.[63] A good set of CVs should be able to distinguish between initial and final state of the process of interest, as well as all intermediate metastable states.[63] Moreover, it should include all slow degrees of freedom that cannot be sampled by unbiased MD.[63] A CV that can distinguish different states by its value is also called an order parameter.[13] If the CV consists of a continuous sequence of order parameter values that represents the progress of transformation from one state (the reactants or initial state) to another state (the products or final state) it is also called the reaction coordinate.[13] The choice of good CVs is crucial for the success of CV based enhanced sampling methods.[16] The CVs are generally chosen using chemical and physical intuition,[16] in a process using any theoretical and experimental information that is available about the system and process of interest.[63] If the final state is not known or there is only little information available about the system, preliminary CVs might have to be created from unbiased simulation or non-CV based enhanced sampling, which are then refined.[63] CVs can also be estimated automatically, using Machine Learning techniques.[16,64] A typical main CV for the process of binding or unbinding of a ligand to and from a protein is the center-of-mass distance between protein and ligand.

3.2.1 Umbrella Sampling

Umbrella sampling is a method introduced by Torrie and Valleau[65] and is the earliest example of systematic biased sampling.[13] The basic idea is to sample a pre-defined reaction path along the reaction coordinate ξ by introducing an artificial biasing potential $\omega(\xi)$. Thus, the unbiased potential energy function V^u is amended by adding $\omega(\xi)$, which is dependent on the reaction coordinate ξ. Typically, $\omega(\xi)$ is a harmonic potential, which is chosen to confine the system in a small interval around the equilibrium position ξ_0.[66] This yields the biased potential energy of the umbrella sampling simulation to be:[67]

$$V^b = V^u + k(\xi - \xi_0)^2 \tag{3.16}$$

with the force constant k and the equilibrium position ξ_0.

Using only one simulation with a biased potential as described above, only a small region of ξ would be sampled thoroughly. For sufficient sampling along the whole region of interest of ξ it is necessary to run a number of simulations, called windows, at different regions of ξ.[66] In practice, the first step of an umbrella sampling run is to separate the phase space of the system along the reaction coordinate ξ into multiple windows i.[67] This can be done, e.g. by using Steered MD,[68] where time dependent force is used to propagate the system along ξ. Within each window, the additional potential energy term $\omega_i(\xi)$ is added to the unbiased potential energy to yield the biased potential energy V_i^b.[66,67] Then, MD simulations using the potentials V_i^b are run separately in each window.[67] The last step is to estimate the free energy along the reaction coordinate from the simulation data. The free energy along the reaction coordinate is sometimes also called the potential of mean force (PMF).[16] Historically, the PMF is a concept of statistical mechanics of liquids and complex molecular systems introduced by Kirkwood in 1935.[66,69] In this specific context, the PMF $\mathcal{W}(r)$ between two particles was calculated from the radial distribution function $g(r)$, which itself depends on the interparticle distance r.[16] The PMF here describes in a literal sense a potential that arises from an average force that would act on a particle at the specific location of r.[16] The historic PMF $\mathcal{W}(r)$ is related to the Helmholtz free energy along the interparticle distance $A(r)$, but is a distinct property:[16]

$$\mathcal{W}(r) = A(r) + 2k_BT \cdot \ln(r) + C \tag{3.17}$$

where C is an arbitrary constant. In modern literature, the colloquial meaning of PMF is synonymous to the free energy along the reaction coordinate (or more general the free energy surface in CV space).[16]

As the simulations were run with a biased potential, it is necessary to unbias the results of the window simulations and recombine the simulation data to eventually estimate the free energy along the reaction coordinate.[66] The Helmholtz free energy $A(\xi)$ is based on the canonical partition function, arising from simulations in the NVT ensemble.[67] As all simulations of umbrella sampling in this thesis are run in the NPT ensemble, I will express the free energy along the reaction coordinate as the Gibbs free energy $G(\xi)$ in the following. In simulations of condensed phase systems, differences in Helmholtz- and Gibbs free energy are numerically very similar, due to the incompressibility of the system.[67] The Gibbs free energy in the NPT ensemble can be expressed as:[67]

$$G^u(\xi) = -\frac{1}{\beta} ln(P^u(\xi)) \tag{3.18}$$

with $\beta = 1/k_BT$ and the unbiased probability distribution $P^u(\xi)$, which can be calculated by integrating over all degrees of freedom $\mathbf{x}$, except the reaction coordinate, in the

Boltzmann distribution:[67]

$$P^u(\xi) = \frac{\int \delta[\xi(\mathbf{x}) - \xi]e^{-\beta V^u(\mathbf{x})}d\mathbf{x}}{\int e^{-\beta V^u(\mathbf{x})}d\mathbf{x}}. \tag{3.19}$$

The unbiased probability distribution $P^u(\xi)$ cannot be calculated directly from umbrella sampling simulation data, because the simulation trajectories were generated under a biased potential and therefore only allow the direct calculation of the biased probability distribution $P^b(\xi)$. To get to the free energy $G^u(\xi)$ from the simulation data, a connection between $P^b(\xi)$ and $P^u(\xi)$ is needed. The unbiased probability distribution of each window i and the corresponding free energy can be calculated as:[67]

$$P_i^u(\xi) = P_i^b(\xi)e^{\beta\omega_i(\xi)}\langle e^{-\beta\omega_i(\xi)}\rangle \tag{3.20}$$

and

$$G_i^u(\xi) = -\frac{1}{\beta}ln(P_i^b) - \omega_i(\xi) + F_i \tag{3.21}$$

$$F_i = -\frac{1}{\beta}ln(\langle e^{-\beta\omega_i(\xi)}\rangle)$$

where F_i is an undetermined constant that represents the free energy associated with the introduction of $\omega_i(\xi)$.[66]

If the whole region of interest is sampled by only one window, the equation above is enough to calculate the free energy $G^u(\xi)$, as $P^b(\xi)$ can be obtained from the simulation, $\omega_i(\xi)$ can be calculated analytically and F_i in this case is just an arbitrary constant.[67] In an umbrella sampling run as described above, the reaction coordinate is sampled by multiple windows, which means the unknown constants F_i of each window have to be calculated.[66,67] This can be done by matching the various PMF of adjacent windows in the regions of ξ, where they overlap and then connecting the adjusted PMFs.[65,66] This method requires significant overlap of the simulation data of the separate windows and includes discarding the data from the region of overlap, which makes it impractical.[66] Another approach for the reweighting of the simulation windows is the Weighted Histogram Analysis Method (WHAM).[70] The basic idea of WHAM is to calculate an optimal estimate of the unbiased probability distribution as a sum of the data gathered from the simulation windows and determining the weight factors that minimize the statistical error.[66] The WHAM approach starts with expressing the unbiased probability distribution as a weighted sum of the windows:[67]

$$P^u(\xi) = \sum_i^{N_w} p_i(\xi)P_i^u(\xi). \tag{3.22}$$

Then, the statistical error of P^u is minimized:[67]

$$\frac{\partial \sigma^2(P^u)}{\partial p_i} = 0. \tag{3.23}$$

Under the condition that the weights p_i are normalized, i.e. $\sum p_i = 1$, this leads to the following solution for the weights p_i:[67]

$$p_i = \frac{a_i}{\sum_j a_j} \tag{3.24}$$

$$a_i = N_i e^{-\beta \omega_i(\xi) + \beta F_i}.$$

where N_i is the number of data points gathered in the window. The constant F_i is then calculated as:[67]

$$e^{-\beta F_i} = \int P^u(\xi) e^{-\beta \omega_i(\xi)} d\xi. \tag{3.25}$$

The weights a_i depend on F_i and F_i in turn depends on $P^u(\xi)$, which itself depends on the weights a_i. This means the equations above have to be solved iteratively in a self-consistent manner until convergence is reached.[66,67]

For simplicity, only the case of a one dimensional reaction coordinate ξ was considered above. However, WHAM is straightforwardly extendable to the case of multiple CVs. Then the unbiased probability distribution as a weighted sum over the simulation windows becomes:[66]

$$P^u(\xi_1, \xi_2, ...) = \sum_i^{N_w} p_i(\xi_1, \xi_2, ...) P_i^u(\xi_1, \xi_2, ...) \tag{3.26}$$

and F_i can be calculated from:[66]

$$e^{-\beta F_i} = \int d\xi_1 \int d\xi_2 \; ... \; P^u(\xi) e^{-\beta \omega_i(\xi_1, \xi_2, ...)}. \tag{3.27}$$

3.2.2 Metadynamics

Metadynamics is an enhanced sampling method that uses an adaptive biasing potential and is the one of the most widely used methods using such a potential.[16] Conventional metadynamics has been first introduced by Laio and Parrinello in 2002[71] and builds on earlier work in this field.[72–74] Metadynamics aims at enhancing the sampling along selected CVs, represented here by the CV vector $\mathrm{s} = [\xi_1, \xi_2, ...]$, by adding a time-dependent bias potential that counteracts the potential energy surface.[16] The time-dependent bias potential is realized by depositing Gaussian kernels in regular intervals τ_G at the current position during the simulation.[16,71,75] This causes the free energy surface along the CVs to gradually fill up and allows the system to overcome energy barriers. After simulation time t has passed, the bias potential can be described as:[16,75]

$$V^{\text{bias}}(\mathbf{s},t) = H_0 \sum_{t'=\tau_G,2\tau_G,\ldots}^{t'<t} \exp\left(-\sum_{k=1}^{d} \frac{(\xi_k - \xi_k(\mathbf{x}(t')))^2}{2\sigma_k^2}\right). \tag{3.28}$$

The parameter H_0 defines the height of the Gaussian kernels, σ the width of the Gaussian kernels and τ_G is the deposition rate. With sufficiently long simulation time, V^{bias} is intended to estimate the free energy along the CVs, in addition to an arbitrary constant C:[75]

$$V^{\text{bias}}(\mathbf{s}, t \to \infty) = -G^u(\mathbf{s}) + C. \tag{3.29}$$

In conventional metadynamics as described above, Gaussian kernels are deposited with a constant height H_0. Using this technique, the bias potential continues to change in the limit of long simulation times, even after all energy minima have been filled and the system moves diffusively in CV space.[76] The bias potential, however, retains qualitatively its shape and can be seen as an ordinary observable that fluctuates around its equilibrium value.[76] This means the free energy surface can be estimated by calculating the time average of the bias potetial.[76] Another issue with conventional metadynamics is that it may allow the system to explore unnatural states by overfilling the free energy surface.[75] This can be addressed by introducing restraining walls, i.e. additional potentials, to keep the system from exploring unnatural or uninteresting regions of the free energy surface.[16] Naturally, an approach like this is only possible with sufficient knowledge about the sampled system and its natural boundaries.

A widely used extension of the conventional method is well-tempered metadynamics,[77] which overcomes the issues of convergence and overfilling of the free energy surface.[75] Well-tempered metadynamics introduces the bias factor γ, which is used to limit the heights of the deposited Gaussian kernels:[16]

$$H_n = H_0 \exp\left[-\frac{1}{\gamma - 1} \cdot \beta \cdot V_{n-1}^{\text{bias}}(\mathbf{s},t)\right] \tag{3.30}$$

where H_n is the hight of the n-th deposited Gaussian kernel. For long simulation times, the height of the Gaussian kernels goes towards zero and the bias potential asymptotically goes towards:[16,77]

$$V^{\text{bias}} = -\left(1 - \frac{1}{\gamma}\right) \cdot G^u(\mathbf{s}) + C(t) \tag{3.31}$$

with the time-dependent constant $C(t)$. The bias potential thus only partly cancels out the free energy surface and the well-tempered metadynamics simulation can be viewed to sample an effective free energy surface, where all energy barriers along the

CVs have been scaled by the bias factor.[16] Another interpretation of the bias factor is to derive it from a temperature parameter ΔT:[16,63]

$$\gamma = \frac{T + \Delta T}{T} \tag{3.32}$$

with the simulated temperature T. This way, the dynamics of the system can be viewed as sampling an unchanged free energy surface with a temperature of $T + \Delta T$ along the CVs.[16] The free energy surface along the CVs can be recovered from a converged well-tempered metadynamics simulation with:[63]

$$G^{u}(\mathbf{s}) = V^{\text{bias}}(\mathbf{s}, t) \cdot \frac{-\gamma}{-\gamma + 1} + \frac{1}{\beta}\ln\left(\int \exp\left[\frac{\gamma}{\gamma - 1}\beta V^{\text{bias}}(\mathbf{s}, t)\right]\right) d\mathbf{s}. \tag{3.33}$$

3.2.3 Random Acceleration Molecular Dynamics

At first introduced under the name Random Expulsion Molecular Dynamics,[78] Random Acceleration Molecular Dynamics (RAMD) is a biased simulation technique that allows accelerated observation of dissociation pathways, but does not allow for the estimation of the free energy functions along a CV.[8,78] RAMD has been originally developed to study the access and exit of a protein ligand from a deeply buried binding pocket within the protein.[78] Other than umbrella sampling and metadynamics, the method is specifically tailored to the study of protein-ligand systems. The basic principle of RAMD is to introduce an additional force $\mathbf{F}$ with constant magnitude and random direction on the center-of-mass of the ligand:[78]

$$\mathbf{F} = k \cdot \mathbf{r_0} \tag{3.34}$$

with the force constant k and a unit vector in random direction $\mathbf{r_0}$. The direction of $\mathbf{r_0}$ is then kept for a specified number of timesteps, before it is re-evaluated.[78] During this simulation time interval, the ligand (more precisely its center-of-mass) needs to have moved by at least the specified distance r_{min} for the direction to not be changed.[78] Otherwise, the direction is changed randomly.[78] The idea behind this approach is that if the ligand discovers rigid parts of the binding pocket or the unbinding path, respectively, it will change its direction randomly and thereby probing different ways to exit the pocket.[78]

RAMD can be used to estimate unbinding kinetics using a protocol called τ-RAMD.[79,80] Essentially, the protocol works by running an ensemble of RAMD simulations from different starting configurations and random seeds for the force direction and by then calculating the average residence time τ that the ligand needs to dissociate from a bootstrapped distribution.[8,79] The average residence time τ is defined as the simulation time after which 50% of the ligands of the ensemble have dissociated.[79] Given the bias of the simulations, the calculated average residence times τ cannot be directly be compared to experimental residence times, however the differences in τ for selected protein-ligand systems have been shown to correlate with experimental values.[79]

3.3 Force Fields

In MD simulations, the potential energy function $V(\mathbf{x})$ is constructed from a set of parameters and functions called the force field. The use of an appropriate and accurate force field is essential for producing meaningful results. Protein systems could be studied with complex force field models like e.g. polarizable force fields,[81] but for these systems, modern atomistic additive point-charge based force fields are known to produce accurate results (with some exceptions such as intrinsically disordered proteins).[82] In this section I will survey the AMBER force field family, which is widely used for molecular simulations in the context of protein systems. Then, I will highlight some approaches to model fluorine in such force fields and systems.

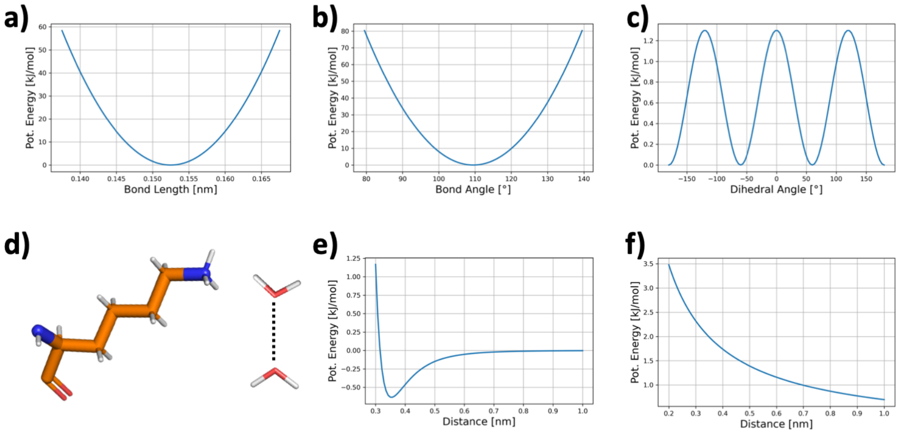


Figure 3.1: Potential energy terms of a typical MD force field. (a) Potential energy of a C-C bond. (b) Potential energy of a C-C-C angle. (c) Potential energy of a C-C-C-N dihedral. (d) Left: Molecular structure of lysine that features the bonded terms presented in this figure. Right: Molecular structures of two water molecules with their non-bonded interactions between the two water oxygen atoms indicated. (e) Lennard-Jones potential between the two oxygen atoms. (f) Coulomb potential between the two oxygen atoms.

3.3.1 AMBER Force Field Family for Proteins

The AMBER protein force field family is a series of pairwise additive force fields, which are widely used for MD simulations of proteins. The beginnings of this force field family are early works by Weiner et al.,[83,84] which were then used by Cornell et al.[85] to develop the AMBER ff94 force field for biomolecules. The ff94 can be used for simulations of proteins, but also for nucleic acids like DNA and RNA and for organic molecules.[85] Its functional form and many of its parameters are still in use in modern AMBER protein force fields, as later iterations of the AMBER force field, like ff99,[86] ff99SB[87] and ff14SB,[88] mainly re-parameterized the dihedral parameters only.

The functional form established in ff94 and used in all other AMBER protein force fields to calculate the potential energy throughout MD simulations $V(\mathbf{x})$ can be expressed as:[12,85]

$$V(\mathbf{x}) = \sum_{\text{bonds}} k_b(r - r_{\text{eq}})^2 + \sum_{\text{angles}} k_\theta(\theta - \theta_{\text{eq}})^2 + \sum_{\text{dihedrals}} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{\text{pairs}} \frac{q_i q_j}{\epsilon r_{ij}} + \sum_{\text{pairs}} 4\varepsilon \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^{6} \right]. \quad (3.35)$$

This means the potential energy is calculated by adding the bonded energy terms of bonds, angles and dihedrals and the pairwise non-bonded energy terms representing Coulomb and Lennard-Jones interactions. The potential energy terms for bonds and angles are expressed as harmonic potentials of the bond distance r or the angle θ respectively. The equilibrium distances and angles r_{eq} and θ_{eq} and the harmonic force constants k_b and k_θ are parameters that have to be provided by the force field. The potential energy terms of the dihedrals ϕ are expressed as a periodic function with the parameters V_n and γ and the multiplicity n. The potential energy terms for all non-bonded pairs are represented by the Coulomb interaction terms that model the electrostatic interaction and the 12-6 Lennard-Jones potential terms, that model the Pauli repulsion and van-der-Waals interactions. Plots of the potential energy arising from these terms are shown in fig. 3.1. In the following, I will describe how the parameters for these bonded and non-bonded terms were derived in the AMBER protein force fields.

The parameters of the AMBER force fields are derived based on atom types.[85] An atom type classifies a certain atom in a simulation based on its element and its chemical environment. E.g. the atom type "CA" would be for a sp^2 carbon atom in an aromatic system or the atom type "CT" would be for a sp^3 hybridized carbon atom.[85] The atom types are then used to determine which type of parameters should be assigned to the respective bonded and non-bonded terms.

The functional form of the non-bonded terms has mostly remained the same as that in ff94 in modern force fields like the ff14SB.[88] The Lennard-Jones parameters are atom type specific, meaning that they do not have to be determined for every individual atom in the simulated system, but instead for every atom type in the system. The Lennard-Jones parameters were derived from Monte-Carlo simulations, e.g. of aliphatic or aromatic hydrocarbons, and then adjusted to reproduce experimental observables of the simulated substances, such as the density and the enthalpy of evaporation.[85] Special attention has to be given to hydrogen atoms, as hydrogen does not have an inner shell of electrons and hence it makes physical sense for the atomic radius of hydrogen to be very sensitive to its bonding partners.[85]

Other than the Lennard-Jones parameters, the partial charges q_i, which define the Coulomb interactions, are assigned to every individual atom in the molecular system. These partial charges are based on electrostatic potentials derived from HF/6-31G* calculations.[85,89] The basic method of deriving the charges is to fit the atomic partial charges to reproduce the electrostatic potential at a large number of grid points around the molecule.[89] Atomic partial charges derived this way have been shown to be superior for molecular simulation than Mulliken population analysis derived charges, in terms of basis set dependency and multipole representation.[90] The fitting of the partial charges q_i to the quantum-mechanically calculated electrostatic potential V_j is realized with a least squares fit procedure. The electrostatic potential at the positions j, $\hat{V}_j$, which arises from the point charge model and depends on the distance to the atoms r_{ji}, is described as:[89]

$$\hat{V}_j = \sum_i \frac{q_i}{r_{ji}}. \tag{3.36}$$

The so-called figure-of-merit χ^2_{esp}, which needs to be minimized in the least squares fit is then defined as:[89]

$$\chi^2_{\mathrm{esp}} = \sum_j (V_j - \hat{V}_j)^2. \tag{3.37}$$

The minimum can then be found by setting χ^2_{esp} to zero and solving the system of equations:[89,91]

$$\frac{\partial(\chi^2_{\mathrm{esp}})}{\partial q_i} = -2 \sum_j \frac{V_j - \hat{V}_j}{r_{ji}} = 0. \tag{3.38}$$

Using atomic partial charges derived as described above has some drawbacks as the charges can be severely dependent on conformation and may lead to artifacts in the conformational energetics.[89] These drawbacks are caused by the fitting procedure, where charges of buried atoms can fluctuate heavily for only very little improvements of the quality of the fit, frequently leading to large charges.[89] To overcome these drawbacks, an additional penalty function, or restraint, χ^2_{rstr} is introduced to the fitting procedure:[89]

$$\chi^2 = \chi^2_{\mathrm{esp}} + \chi^2_{\mathrm{rstr}}. \tag{3.39}$$

The goal of the penalty function is to restrict the magnitude of the charges with a minimal cost in quality of the fit.[89] The penalty function has the form of a hyperbolic function:[89]

$$\chi^2_{\text{rstr}} = a \sum_i (\sqrt{q_i^2 + b^2} - b) \tag{3.40}$$

where a determines the asymptotic limits of the strength of the penalty function and b determines the tightness around the minimum. The resulting method is called the restrained electrostatic potential (RESP) fit.[89] The atomic partial charges obtained by this method at HF/6-31G* level of theory overestimate gas-phase dipoles in a fortuitous way that compensates for the influence of water.[89,92] In practice, the RESP fit of atomic partial charges is conducted in two stages. The reason behind this is to force the molecular systems to have charge symmetry in atoms that exchange quickly in Molecular Dynamics simulations, like e.g. the hydrogen atoms on a methyl group.[89] In the first stage, the fit is conducted with weak restraints and without forced symmetry and in the second stage, only the atoms that should have forced symmetry are refitted with strong restraints and symmetry enforced.[89] The RESP model is still in use in modern force fields like ff14SB as it has been extensively tested and is well compatible with other parameter sets.[88]

Moving on to the bonded parameters, the parameters for bonds and angles are derived to resemble experimental normal mode frequencies of molecular fragments[85] and were retained in modern protein force fields.[88] The dihedral parameters, on the other hand, were updated multiple times in the iterations of the AMBER protein force fields. In the ff94, a minimalist approach was followed, where the dihedral parameters are derived from quantum-mechanical energies of the simplest molecule possible and then applied to larger molecules.[85] Extra Fourier terms V_n were only added in case of physical necessity, like e.g. to model the Gauche effect for X-C-C-X dihedrals, and for the dipeptide ϕ and ψ dihedrals.[85] The dihedral parameters were updated in the ff99 force field by again adding a limited number of Fourier terms.[86] Given issues with protein backbone parameters in ff94 and ff99, most relevantly the over-stabilization of α helices, Hornack et al. updated the backbone dihedral parameters in the ff99SB force field based on the quantum-mechanical energies of selected conformations of tetra-peptides from *ab-initio* calculations.[87] Here it is important to note, that they not only updated the ϕ and ψ dihedrals along the protein backbone (ϕ=C-N-Cα-C, ψ=N-Cα-C-N), but also the ϕ' and ψ' dihedrals (ϕ'=C-N-Cα-Cβ, ψ'=Cβ-Cα-C-N), which influence the protein backbone properties, but also include side chain atoms.[87]

After limitations of the ff99SB with respect to side chain rotamers and backbone secondary structure preferences became apparent, Maier et al. conducted a complete refit of all amino acid side chain parameters and made empirical adjustments to the backbone dihedral parameters.[88] For the side chain re-parametrization, the fitting strategy focussed on obtaining a more diverse and consisted training dataset, aimed at representing the conformational diversity found in protein amino acids.[88] The dataset included full amino acid structures, instead of small organic fragments, and sampled a large set of conformations, including side chain and backbone rotamers.[88] The refit also employed new atom types and fitted the dihedral parameters on full rotational profiles, instead of single point energies of structures in gas phase minima only.[88] The geometries along the dihedral profiles were calculated at HF/6-31-G* level of theory and the single point energies of the obtained geometries were calculated at MP2/6-31+G** level of theory.[88]

3.3.2 General AMBER Force Field for Small Organic Molecules

The AMBER force fields, which I described in the last section, only contain parameters for proteins (and sometimes also for other bio- macromolecules like nucleic acids). Especially in a drug discovery context, it is often of interest to study protein-ligand systems, where the ligand is a small organic molecule. Hence, there is a need for force fields that model small organic molecules in a way that is compatible with protein force fields.

The General AMBER Force Field (GAFF) is an extension of the AMBER protein force fields and covers parameters for most of small organic molecules containing the elements C, H, O, N, S, P, F, Cl, Br and I.[93] GAFF uses the same functional form and modeling strategies as the AMBER force fields for proteins.[93] GAFF attempts to be economic with the number of atom types, as more atom types mean a greater parametrization burden and thus employ 35 basic atom types and an additional 22 special atom tyes, which are only used to describe specific chemical environments.[93]

The Lennard-Jones parameters in GAFF are taken over from the AMBER protein force fields.[93] The default scheme for deriving charges is stated to be the RESP fit procedure with HF/6-31G* level of theory, just like in the protein force fields.[93] However, since users of GAFF will have to parametrize the charges for their own small molecule systems and *ab initio* calculations may be too expensive, especially when handling large numbers of molecules, the AM1-BCC scheme is also recommended.[93–95] The idea behind this scheme is to first conduct a semiempirical AM1 calculation, which is then followed by a bond charge correction aimed at reproducing the electrostatic potential at HF/6-31G* level.[93, 95]

The bonded force field terms for bonds and angles are either taken over from the protein force fields, fitted to *ab initio* quantum chemical calculations or based on crystal structure data.[93] Dihedral parameters are derived from fitting to torsional scans at higher level of theory.[93]

Some of GAFF's parameters have been updated to yield the second generation force field GAFF2. These updates include bonded parameters to reproduce experimental and theoretical observables on a larger set of model compounds and better non bonded parameters.[96]

3.3.3 Force Fields for Fluorinated Amino Acids

Halogens are generally difficult to model in a point-charge based force field, because covalently bound halogens may show a highly anisotropic charge distribution.[97,98] The anisotropic distribution is composed of a negative charge belt, perpendicular to the covalent bond and a region of reduced electron density at the opposite side of the covalent bond, called the σ-hole.[99] This is a problem for point-charge based force fields, because a single point-charge for a halogen atom would model an isotropic charge distribution and therefore miss important properties of the anisotropic charge distribution, like the ability to form halogen bonds.[97,98] This problem can be solved by adding virtual extra particles to the topology of the halogenated molecule. One of such approaches is the positive-extra-point (PEP) model, that introduces an extra charge close to the halogen atom.[100] Another approach would be to add four negative extra charges to model the negative charge belt and one positive extra charge for the σ-hole[101] or to use cosine functions to describe the Coulomb and Lennard-Jones terms of the halogens.[102]

The magnitude of the charge anisotropy decreases from iodine to fluorine. The anisotropy is still relevant for fluorine in the case of hydrogen fluoride, where positive extra charges are used to model the correct angled HF..HF binding geometry.[103,104] However, for carbon bound fluorine, the charge distribution is nearly isotropic and a single point-charge representation is sufficient.[34,105] Generic parameters for fluorine are included in standard all-atom force fields like GAFF, but there have also been considerable advances to tailor the parameters for specific systems such as fluorinated amino acids.[34,35,106,107]

Particularly interesting in the context of this thesis is the parametrization of fluorinated aliphatic amino acids by Robalo et al..[34,35] In these efforts, non-aliphatic amino acids like valine, leucine or the non-natural Abu were fluorinated at one or two of their side chain carbon atoms and the parameters were derived. The bonded parameters for the fluorinated amino acids were transferred from GAFF. The Coulomb parameters were generated in two iterations of a RESP fit, as is used for the AMBER protein force fields. For clarity, this means two iterations of the whole RESP procedure, not the two stages of the standard RESP fit. First, initial charge parameters were generated based on a single structure and then, using the initial parameters, an ensemble of structures was generated from an MD simulation and the charge parameters were calculated again for all the structures of the ensemble and averaged to yield the final charges.[35] For the Lennard-Jones parameters fluorine amino acids, the authors originally attempted to optimize ε and σ against three target properties of CF_4, the hydration free energy, the enthalpy of vaporization and the molar volume. As there is no combination of ε and σ that accurately reproduces all of the above properties, the authors selected the hydration free energy and the molar volume as the target values.[35] This selection ensures that the hydrophobic effect and the packing in the hydrophobic core of proteins is accurately modeled. Partially fluorinated aliphatic amino acids additionally need Lennard-Jones parameters for the atom type "H_F". This atom type describes a hydrogen atom that is bound to a carbon that is also bound to a fluorine. The Lennard-Jones parameters were calculated using the same target properties but for CHF_3 instead of CH_4.[34]

In the first publication presented in this thesis, I write about the representation of fluorine in MD force fields in more detail.

3.4 Hydration in Protein Systems

Water is essential for the structure, function and energetics of any biological molecule and has fundamental importance in protein-ligand and protein-protein binding.[33,108,109] A wide variety of computational methods has been developed to study protein hydration, including methods based on MD simulations.[108] Of special interest are water molecules that are inside cavities of proteins. These cavities typically have a concave shape, which provides a confined and partially buried environment.[33] The water molecules in this environment are likely to be positionally ordered and precisely located, which gives them the name "structural waters".[33] Structural waters are relevant for protein binding, as they can mediate bonds and have sizable effects on binding enthalpies and entropies through rearrangement or displacement into the bulk upon binding.[33]

Here, I will first describe how MD simulations can be used to study the structure and dynamics of structural water molecules and then focus on methods based on Grid Inhomogeneous Solvation Theory (GIST) to study the energetics of such water molecules.

3.4.1 Analysis of Structural Waters

Structural waters are often interconnected in hydrogen bond networks. MD simulations can be used to detect these hydrogen bonds by assessment of geometric criteria. Wernet et al. introduced such a geometric criterion for hydrogen bonds by studying calculated X-ray absorption spectra of water molecules.[110] Their criterion is fulfilled, when the donor-acceptor distance r_{DA} (in nm) fulfills the inequality:[110]

$$r_{DA} < 0.33\mathrm{nm} - 0.00044 \cdot \theta^2 \tag{3.41}$$

with the angle between donor, hydrogen atom and acceptor θ in degrees. The factor of 0.00044 is in the unit of nm per degrees squared.

Within the duration of a typical MD simulation, it is highly unlikely that observed water networks remain static, as water molecules exchange their hydrogen bond partners rapidly.[111] In bulk water, these exchanges happen on a scale of 1 ps to 5 ps.[111] The dynamics in protein hydration shells of reorientation, exchange of hydrogen bond partners and translation into the bulk also happen on a sub-nanosecond scale.[111] These dynamics may be slower for structural waters with the extreme case of timescales of microseconds and milliseconds for deeply buried waters.[111]

MD simulations can be used to estimate mean residence times τ of water molecules in water networks. At first, one needs to define the position of the water molecule with some kind of criterion, e.g. based on distances. Then, one can define a survival function $B(t)$ that indicates whether at time t the water is at the respective position.[112] From $B(t)$, the function $C(t')$ can be calculated, that estimates the number of times, the water molecule resides at its position continuously for the lag time t':[112]

$$C(t') = \sum_{t=0}^{N_t - t'} B(t') \prod_{k=t}^{t'+t} B(k) \tag{3.42}$$

with the total number of simulation snapshots N_t.

This function can then be fitted to a double exponential decay with the parameter w to get the mean residence time:[112]

$$C(t') = w \cdot \exp\left(-\frac{t'}{\tau_1}\right) + (1 - w) \cdot \exp\left(-\frac{t'}{\tau_2}\right). \tag{3.43}$$

τ_1 and τ_2 are the mean residence times for the slow and fast component of the water dynamics, respectively.

3.4.2 Grid Inhomogeneous Solvation Theory

The basic idea of Grid Inhomogeneous Solvation Theory (GIST) is to obtain spatially resolved hydration thermodynamics by discretizing the water density, solvation energy and solvation entropy into a 3D grid, composed of cubic voxels.[33] This is realized by replacing the analytical spatial integrals of Inhomogeneous Solvation Theory (IST)[113] with sums of voxels of the 3D grid.[114] GIST based analysis uses MD to sample the equilibrium distribution of water molecules around a given solute in a pre-defined conformation.[33, 114] For every voxel k, the following thermodynamic quantities are computed:[115] the mean solute-water (sw) interaction energy of a water molecule in the voxel $\Delta E^{\mathrm{norm}}_{\mathrm{k,sw}}$, the mean water-water (ww) interaction energy of a water in the voxel with all other waters $\Delta E^{\mathrm{norm}}_{\mathrm{k,ww}}$, the single-body translational entropy of a water the voxel $\Delta S^{\mathrm{trans,norm}}_{\mathrm{k,sw}}$ and the corresponding orientational entropy $\Delta S^{\mathrm{orient,norm}}_{\mathrm{k,sw}}$. In the following, I will briefly describe how these quantities can be computed. For $\Delta E^{\mathrm{norm}}_{\mathrm{k,ww}}$, it is important to note that this quantity is sometimes defined as the full water-water mean interaction energy,[114] or one half of the water-water mean interaction energy.[115] The factor of 0.5 is customary in liquid-state theory.[115]

The solute-water interaction energy $\Delta E_{\mathrm{k,sw}}(\mathbf{r}_k)$ in the region of space $\mathbf{r}_k$ that belongs to voxel k is in practice computed as the total interaction energy of the solute with all water molecules in $\mathbf{r}_k$, averaged over the number of simulation snapshots.[114] The solute-water interaction energy is then normalized with the average number of water molecules in the voxel n_w:[114]

$$\Delta E^{\mathrm{norm}}_{\mathrm{k,sw}} = \frac{\Delta E_{\mathrm{k,sw}}(\mathbf{r}_k)}{n_w} \tag{3.44}$$

Analogously, the water-water interaction energy of the voxels $\Delta E_{\mathrm{k,ww}}(\mathbf{r}_k)$ is computed from the simulation snapshots and then normalized.[114]

Assuming the radial distribution function $g(\mathbf{r})$ is uniform within the voxel, the translational entropy $\Delta S^{\mathrm{trans}}_{\mathrm{k,sw}}$ can be calculated as:[114]

$$\Delta S^{\mathrm{trans}}_{\mathrm{k,sw}} = -k_B \rho^0 V_k \cdot g(\mathbf{r}_k) \cdot \ln\left[g(\mathbf{r}_k)\right] \tag{3.45}$$

with the number density of bulk water ρ^0 and the volume of the voxel V_k. The radial distribution function $g(\mathbf{r}_k)$ is defined as:[114]

$$g(\mathbf{r}_k) = \frac{N_w}{\rho^0 V_k N_t} \tag{3.46}$$

with the total number of water molecules in the voxel N_w summed over all simulation snapshots and the number of simulation snapshots N_t. The orientational entropy $\Delta S^{\mathrm{orient}}_{\mathrm{k,sw}}$ can be calculated as:[114]

$$\Delta S^{\mathrm{orient}} = \rho^0 V_k g(\mathbf{r}_k) S^{\omega}(\mathbf{r}_k). \tag{3.47}$$

The orientational entropy associated with the voxel $S^{\omega}(\mathbf{r}_k)$ can be estimated using a nearest neighbour method, which eventually yields:[114]

$$S^{\omega}(\mathbf{r}_k) = -\frac{k_B}{N_w}\left(\gamma + \sum_{i=0}^{N_w} \ln\left[g(\omega_i|\mathbf{r}_k)\right]\right) \tag{3.48}$$

with Euler's constant γ and the radial distribution function of the Euler angles ω given the space of the voxel $g(\omega_i|\mathbf{r}_k)$ is:[114]

$$g(\omega_i|\mathbf{r}_k) = \frac{8\pi^2}{V_i^{\omega}} \tag{3.49}$$

with

$$V_i^{\omega} = \frac{4\pi(\Delta\omega_i)^3}{3}. \tag{3.50}$$

Both entropic quantities are normalized with the average number of water molecules in the voxel n_w.[114]

3.5 Molecular Docking

Molecular docking is a well established method in structure-based drug design.[116,117] The goal of docking methods is to predict the preferred binding mode and binding affinity of a ligand, usually a drug-like organic molecule, to a receptor, usually a protein.[116] Docking protocols typically start with the receptor, the structure of the ligands of interest and the identification of the binding site.[116] Then, a search and sampling algorithm samples multiple binding poses, which are evaluated by a scoring function.[116] Here, I describe one docking method, which is named Glide.[118,119]

In the Glide docking protocol, the binding site region of the receptor structure is represented as a grid, on which different sets of fields encode the shape and properties of the receptor.[118] The first step of the sampling algorithm is an initial screening of the ligand conformations. This is done with an exhaustive search based on truncated Molecular Mechanics potential energy functions and a heuristic screening, which makes its assumptions based on data from protein-ligand crystal structures.[118] Then, ligand poses are sampled on the receptor grid. A number of the best-scoring poses are then energy minimized and scored again. The three to six best-scoring poses are then subjected to a Monte-Carlo sampling to examine torsional minima nearby.[118] The scoring function is a sum of estimated energy terms. While van-der-Waals and Coulomb terms are included as well, most of these terms are heuristic energy terms, which reward or penalize different contributions that may influence the binding affinity and which were fitted to reproduce experimental binding affinities.[118] These terms include rewards for interactions like hydrogen bonds or lipophilic contacts and metal ligation or penalize close contacts and entropy loss upon ligand binding, as estimated from rotatable bonds.[118] The protocol described above is the "standard precision" model, however Glide also comes with the option of an "extra precision" protocol. Here, a more exhaustive sampling algorithm is used, that builds on the standard precision algorithm and also uses a more complex scoring function.[119] This scoring function accounts for specific structural motifs, that are estimated to enhance the binding affinity like enclosed lipophilic atoms and special motifs for hydrogen bonds and ionic contacts.[119]

4 Results

4.1 Fluorinated Protein-Ligand Complexes Perspective

Title: Fluorinated Protein-Ligand Complexes - A Computational Perspective

Fluorine, with its exceptional capability to influence the physico-chemical properties of organic compounds, is a remarkable substituent for protein binding ligands. In this perspective article, we explore how fluorine can influence the binding properties of protein ligands by focussing on recent literature examples of specific effects of fluorination and their analysis using molecular simulations. We cover the following topics:

- Force Fields
- Fluoride Channels
- Effects Involving Aryl Groups
- Hydrogen Bonds
- Water Networks
- Entropic Effects

We highlight the importance of choosing an accurate force field for molecular simulations with fluorine and show recent force field development efforts. We then present fluoride channels as a case study of how proteins interact with fluoride anions and how computational methods can be used to study these interactions. We then focus on specific effects of fluorine on the protein binding properties of small molecule ligands. With respect to aryl groups, we show how fluorine can interrupt aromatic interactions. Regarding hydrogen bonds, we describe the controversial discussion of fluorine as an acceptor of hydrogen bond like interactions, which we underline by discussing recent examples of such interactions in protein-ligand systems. Direct hydrogen bond like interactions between proteins and fluorine seem generally unlikely to be main driving forces behind changes in binding affinity. However, in special cases, like with heavily fluorinated phosphorus as a substituent, a significant impact of fluorine on the binding affinity is probable.

The observation that organic fluorine interacts with water and can influence water dynamics in its immediate surrounding suggests that fluorinated substituents may affect the binding affinity to proteins by interacting with water networks. The rationalization of the various effects of fluorine on water networks in protein-ligand systems is difficult, but can be approached with molecular simulations, which we show with recent examples. Among these effects, the impact of fluorination on entropic effects is specifically interesting, as seemingly subtle fluorine substitution can lead to sizable differences in the entropic part of binding affinities.

Bettina Keller and I wrote and revised the manuscript. I conducted the literature research.

Fluorinated Protein-Ligand Complexes - A Computational Perspective

Leon Wehrhan and Bettina G. Keller*

Department of Chemistry, Biology and Pharmacy; Freie Universität Berlin, Arnimallee 22, 14195, Berlin, Germany

E-mail: bettina.keller@fu-berlin.de

Abstract

Fluorine is an element renowned for its unique properties. Its powerful capability to modulate molecular properties makes it an attractive substituent for protein binding ligands, however the rational design of fluorination can be challenging with effects on interactions and binding energies being difficult to predict. In this perspective, we highlight how computational methods help understand the role of fluorine in protein-ligand binding with a focus on molecular simulation. We underline the importance of an accurate force field, present fluoride channels as a showcase for biomolecular interactions with fluorine and discuss fluorine specific interactions like the ability to form hydrogen bonds and interactions with aryl groups. We put special emphasis on the disruption of water networks and entropic effects.

Introduction

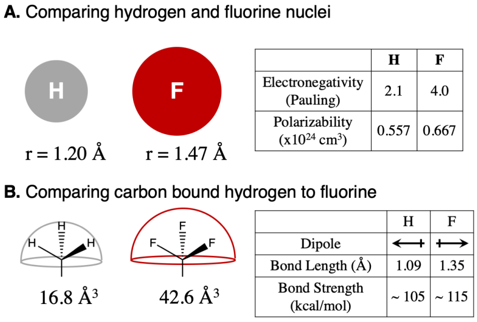
Fluorine has an exceptional standing among the main group elements due to its remarkable reactivity in its elemental form and due to its extraordinarily high electronegativity. Even though inorganic fluorine is abundant on earth, fluorine is virtually absent from the organic compounds in living systems.[1,2] Yet, fluorinated molecules are essential to medicinal chemistry, with a share of roughly 20% of fluoro-organic compounds in all globally registered pharmaceuticals.[3,4] For example, fluorinated molecules have played a major role in the pursuit of anti-COVID19 drugs.[5] Paxlovid, the first orally administered drug for the treatment for the coronavirus disease, is a combination of a fluorinated inhibitor of the coronavirus main protease and a non-fluorinated assisting compound.[5]

The importance of fluorine for medicinal chemistry can be attributed to its exceptional potential to modulate the physico-chemical properties of organic compounds.[1,6–8] Fluorine is a small atom with high electronegativity and low polarizability. With its van-der-Waals radius of 1.47 $\AA$, fluorine occupies a smaller volume than typical organic substituents, like methyl, amino, or hydroxy groups, but is larger than a hydrogen atom (Fig. 1). The C-F bond is strong (460-540 kJ/mol)[7] and slightly longer than the C-H bond. Compared to the C-H bond, the C-F bond shows a reversed dipole moment, induced by the high electronegativity of fluorine. These properties make fluorine a powerful modulator of the lipophilicity, the pKa or electrostatic potential of an organic molecule. Additionally, fluorine substituents can change the conformational preferences of organic molecules.[1,8] Another pharmaceutically relevant effect is that fluorine can increase the metabolic stability of drug molecules.[9] Finally, ^{19}F, the only stable isotope of fluorine, has a nuclear spin of $+\frac{1}{2}$ and can thus be detected in NMR spectroscopy. It is a sensitive alternative to more commonly used nuclei like ^{1}H, ^{13}C or ^{15}N in NMR spectroscopy.[10]

Because fluorine substituents influence a wide range of molecular properties, the rational design of fluorinated molecules is challenging. Ref. 11 gives an overview of non-covalent interactions of fluorine compounds. Several reviews cover the impact of fluorinated molecules in drug design and modern pharmaceuticals,[3–5,8] as well as synthetic approaches and challenges of obtaining organic fluorinated molecules.[12–14] ^{19}F NMR spectroscopy in the context of fragment screening in drug discovery is reviewed in ref. 15. Fluorinated amino acids and peptides are reviewed in Refs. 16–18.

Computational methods play an essential role in disentangling and understanding the relative magnitude of the often competing effects that a fluorine substituent can have on the molecular properties of the compound. Our goal here is to summarize how recent computational studies contributed to our understanding of fluorine's role in protein ligand binding. We particularly, focus on molecular simulations and include an overview of recent atomistic force fields for fluorinated substances. While the focus of this perspective lies on molecular simulations using classical atomistic force fields, fluorinated protein ligands have also been studied with quantum chemical calculations, like those in ref. 19 and ref. 20.

In a protein environment, fluorine can act as a hydrogen bond acceptor but can also form contacts with nearby aryl groups.[21,22] It has even been proposed that fluorine substituents do not need to interact with the protein directly but instead bind to the protein via water mediated contacts.[23,24] Closely related to these water networks is the desolvation of fluorinated ligands during the binding process, which induces surprising enthalpy-entropy compensation effects.[25,26] We hope to provide useful insight into how the relative importance of these effects can be measured using molecular simulations.


A. Comparing hydrogen and fluorine nuclei
H
F
r = 1.20 Å
r = 1.47 Å
B. Comparing carbon bound hydrogen to fluorine
H
H
H
F
F
F
16.8 Å^3
42.6 Å^3

	H	F
Electronegativity (Pauling)	2.1	4.0
Polarizability (x10^{24} cm^3)	0.557	0.667

	H	F
Dipole	⟵+	+⟶
Bond Length (Å)	1.09	1.35
Bond Strength (kcal/mol)	~ 105	~ 115

Figure 1: Comparison of basic properties of fluorine and hydrogen atoms (A) and of carbon bound fluorine vs. carbon bound hydrogen (B). Adapted in part with permission from ref. 7. Copyright 2020 Wiley.

Results and discussion

Force Fields

The accuracy of a molecular simulation critically depends on the underlying potential energy function, whose negative gradient is the molecular force field. The challenge in describing halogen substituents within the functional form of a molecular mechanics force field lies in balancing the Coulomb and the van-der-Waals interactions. Additionally, covalently bound halogen atoms may exhibit a highly anisotropic charge distribution, which consists of a negative charge belt in the direction perpendicular to the covalent bond and a region of diminished electron density on the opposite side of the covalent bond, called the σ-hole[27] (Fig. 2.a). Due this anisotropic charge distribution, an intermolecular interaction called halogen bond may form, in which a lone pair on neighboring atom binds to the σ-hole. The σ-hole decreases from iodine to fluorine (Fig. 2.b), with bromine and chlorine featuring medium-sized σ-holes.

Most atom types in molecular mechanics force fields are modeled using a single point charge, which creates an isotropic electrostatic potential around this atom. These atom types lack the ability to model the σ-hole and halogen bonds.[28,29] The anisotropic charge distribution of covalently bound halogens can, however, be modeled by introducing a positive extra point (PEP) close to the halogen atom,[30] or by a 5 pseudo-atom scheme,[31] where an extra four pseudo-atoms are introduced to model the negative charge belt around the halogen (Fig. 2.c). Another approach is to tune the Coulomb and Lennard-Jones interactions of the halogen using cosine functions[29,32]

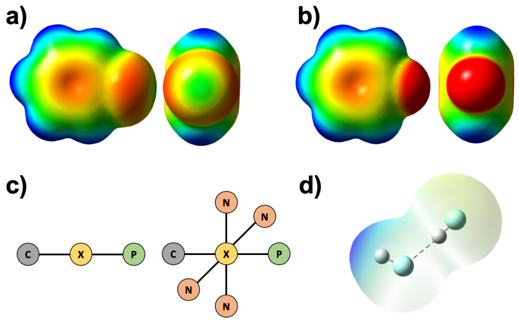


Figure 2: a) Electrostatic potential of chlorobenzene, top view of the aromatic system and view along the Cl-C bond. b) Electrostatic potential of fluorobenzene, top view of the aromatic system and view along the F-C bond. c) Molecular mechanics based models for covalently bound halogens (left) Positive extra point model. (right) 5-Pseudo-atom model with positive (P) and negative (N) extra point charges. d) HF dimer in typical angled geometry. Electrostatic potentials were calculated from QM structure optimization at the MP2/aug-cc-pVTZ level of theory in Gaussian16. The electrostatic potentials are shown on the MP2 electron density surface with an isovalue of 0.0004 au. Subfigures (a) and (b) show the electrostatic potential on a scale from -55 kJ/mol to 55 kJ/mol. Subfigure (c) shows the electrostatic potential on a scale from -50 kJ/mol to 250 kJ/mol.

The anisotropic charge distribution around fluorine is particularly relevant in hydrogen fluoride, where it leads to an angled hydrogen bond geometry. Parameters for hydrogen fluoride which make use of an positive extra point charge and correctly reproduce the hydrogen bond geometry have been proposed in recent years[33,34] (Fig. 2.d).

In fluorine bound to carbon, the electrostatic potential is almost isotropic around the fluorine substituent[35] (Fig. 2.b), and therefore, the fluorine atom in carbon-fluorine substituents may be modeled with a single point charge. In fact, generic fluorine atom types with a single point charge and corresponding van-der-Waals parameters are included in traditional all-atom force fields like GAFF/GAFF2. However, the balance between Coulomb-interacions and van-der-Waals interactions can vary significantly across different organic compounds. Parameters determined for model compounds may not seamlessly align with those of a protein force field. As a result, various customized parameter sets have been published for specific fluorinated compounds and fluorinated amino acids.

Multiple fluorinated amino acids have been added to the CHARMM36 protein force field and the CHARMM general force fields.[36] Furthermore, a set of fluorinated aliphatic amino acids[35,37] and a set of fluorinated tyrosine derivatives[38] was added to the AMBER14SB protein force field. Using the implicit polarization charge scheme, parameters for fluorinated aromatic amino acids were added to the AMBER ff15ipq protein force field.[39]

Robalo et al.[37] have focussed on balancing the thermodynamic properties of fluorinated amino acids. Their parameter sets are based on point charges and the AMBER functional form,[35,37] where the fluorine point charges are calculated based on the AMBER typical RESP fitting procedure. To account for the variations in the partial charge due to conformational changes in the amino acid, they determine a conformationally averaged partial charged by iteratively fitting point charge and sampling the conformational equilibrium. The parameters of the Lennard-Jones potential, which is used to model the van-der-Waals interactions in AMBER, are then optimized against the hydration free energy of CF_4 and the molar volume of an equimolar CF_4:CH_4 mixture, to capture the hydrophobic effect of fluorine and packing constraints in the hydrophobic core of proteins. To model hydrogen bonds to partially fluorinated carbons, the authours additionally optimize the Lennard-Jones parameters of the hydrogen atom on these partially fluorinated carbons against the hydration free energy and molar volume of HCF_3.

While Ref. 37 is concerned with fluorinated amino acids carrying one or two fluorinated carbons and focused on changing the non-bonded force field parameters of fluorine, it may sometimes be necessary to re-parametrize bonded force field terms as well. Träg and Zahn benchmarked dihedral parameters of the GAFF2 all-atom force field and found weaknesses when describing molecules with more than three adjacent fluorinated carbons.[40]

The examples above show that, while traditional point charge based force fields are sufficient to model fluorine substituents, generic fluorine parameters have a limited scope and often need to be fine-tuned to the system at hand. These re-parametrizations can be tedious. Automatic or semi-automatic processes could help to streamline the parametrization of fluorinated molecules in the future. One approach is applied in the open force field.[41–43]

The open force field iterations Parsley[42] and Sage[41] evade the use of atom types, required for traditional force fields, and use a process called native chemical perception instead, that relies on querying functional groups with SMIRKS strings.[43] This way, the force field is easily extendable without excessively inflating the complexity by adding large numbers of new atom types. Another recent promising approach is the Espaloma[44,45] force field. The authors use graph neural networks to perceive chemical environments (by learning embeddings for atoms, bonds, angles and dihedrals) and determine force field parameters based on QM calculations in a way that is completely end-to-end differentiable with respect to model parameters. In this way, developing customized parameters for fluorinated ligands can be automized by using standard neural network libraries. We note that an accurate force field parameters are not only essential for molecular simulations, but for any computational method that requires an accurate molecular-mechanics based representation of the fluorinated molecules, such as molecular docking.

Having stressed the importance of selecting an appropriate force field in general and for fluorinated compounds in particular, our focus shifts to exploring specific interactions in the following sections. Before we explore the specific interactions and effects of fluorine in protein-ligand systems, we present fluoride channels as a case study of how proteins interact with fluorine, here in the form of a fluoride anion.

Fluoride Channels (Flucs)

Fluoride channels (Flucs) are an interesting case study that showcases how proteins interact with fluoride anions. Flucs are membrane channels found in microorganisms, where they export toxic fluoride anions across the cell membrane.[46] Their structure is unique, as the Flucs are expressed as homodimers, where the monomers are aligned in an anti parallel manner. Despite the anti parallel alignment, the monomers transport fluoride ions independently in the same direction, so that inactivating one of the monomers does not stop fluoride transportation through the other one. Flucs are exceptionally selective for fluoride over chloride and small cations. How this selectivity is achieved and how fluoride ions traverse the channel is subject to current research.[47–49]

Yue et al.[47] studied the fluoride channel of *E . Coli* (Fluc-Ec2) using the polarizable Drude force field and CHARMM36 additive force field for comparison. They employed replica exchange umbrella sampling to get a potential of mean force of the fluoride position along the channel. They find energetic minima that align with fluoride positions observed in crystal structures and also calculate permeation rates that match experimental values well. The fluoride permeation relies on a network of different non bonded interactions in the channel, that compensate for the desolvation of fluoride. These interactions include hydrogen bonds between amino acid side chains, the protein backbone or water and fluoride, ionic contacts to positively charged side chains and anion-pi contacts. The non-polarizable force field overestimates the interactions and therefore leads to wrong permeation rates, which indicates that the polarizability plays an important role in the ion permeation process.

Zhang et al.[48] expand on this study by using solid state NMR and molecular dynamics simulations of the Flucs and mutated versions of the Flucs to investigate the permeation mechanism. They identify additional fluoride residence sites at the aqueous regions by the entrance/exit of the channel and discover, through molecular dynamics, that these sites can be explored by fluoride as well as by chloride. The authors also identify a structural loop which is important for channel gating. They propose a water mediated ”knock on” mechanism for the fluoride permeation (see fig. 3). In this model, one of the two internal fluoride sites within the Fluc is occupied by a fluoride while the other is occupied by water in the static state. An entering fluoride will push the water into the next site, where the residing fluoride will be expelled from the channel. The state that is now formed is short-lived and will return to the static state.

This proposed mechanism has intriguing consequences for the fluorine-protein interactions. As the fluoride ions inside the channel are likely to be unhydrated, the desolvation energy of 464 kJ/mol[47] has to be compensated by interactions of the protein with the fluorine anion. These interactions are ionic contacts and hydrogen bonds, which are likely to be particularly stable because of the negative charge of the fluoride ion. Moreover, anion-π interactions are observed, which are not as common as cation-π interactions and rely on the interaction of the anion with the positive edge of an aromatic system.

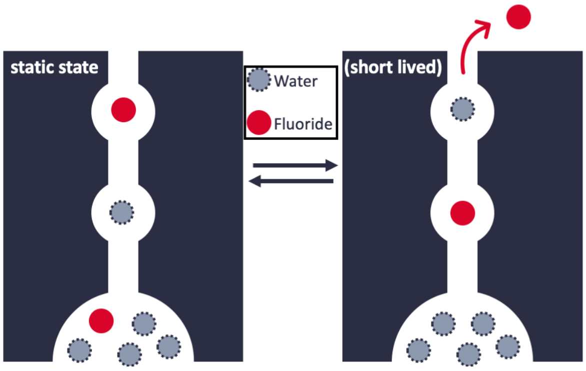


Figure 3: Fluoride ion transport in Flucs via the water mediated "knock-on" mechanism proposed by Zhang et al.[48] One of the two internal sites is occupied by a fluoride ion in the static state. A second entering fluoride will expell the first fluoride via a short lived state, that eventually returns back to the static state. Adapted in part from ref. 48. Available under a CC BY-NC license. Copyright 2023 Zhang et al.

While fluoride anions are naturally different from fluorine in small molecule protein ligands, the case study explored here demonstrates how interactions of proteins with fluorine can counteract sizable magnitudes of energy, such as the desolvation of fluoride. Moreover, this case study demonstrates how molecular simulation methods can be utilized to study the interactions between fluorine and proteins. We now shift our focus to fluorine in small molecule ligands and to its specific interactions.

Effects Involving Aryl Groups

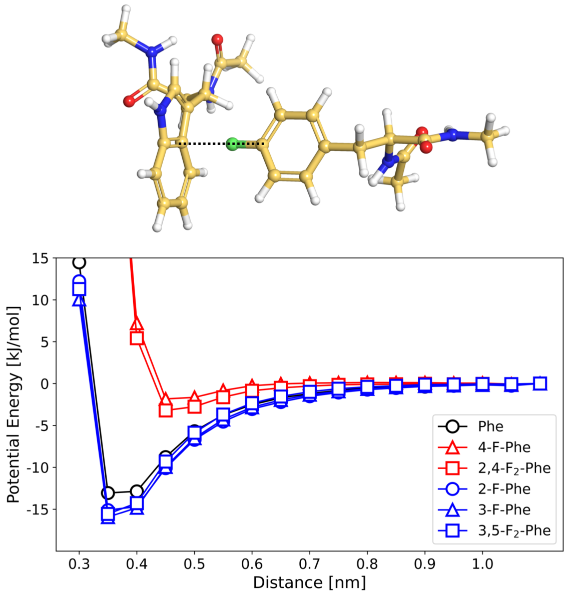


Figure 4: Potential energy of the T-shaped π interaction between variants of phenylalanine and the aromatic ring of tryptophan. The interaction energy was calculated using GAFF2 parameters. Adapted from ref. 22. Available under a CC BY-NC license. Copyright 2023 Müll et al.

Fluorine has a strong influence on conjugated or aromatic π-electron systems in its vicinity due to its high electronegativity. When the fluorine is attached to an aryl group, as in most fluorinated pharmaceuticals,[3] it acts as an electron withdrawing group and reduces the electron density in the aromatic system. This weakens any π-stacking interactions the unflourinated aryl might be involved in.[50,51] When fluorinating tyrosine, the electron withdrawing effect of the fluorine substituent leads to a decreased $\mathrm{p}K_\mathrm{a}$ of the tyrosine hydroxyl group.[52] Fluorinating anilines can have interesting behaviour on their hydrogen bonding abilities, as monofluorination leads to weaker and less frequent NH..N hydrogen bonds while fluorination of 4X-anilines can increase the strength of these bonds.[53,54]

Fluorine can also directly interrupt a T-shaped π-interaction if it is introduced as a substituent on the aryl group that represents the T-stem (Fig. 4). Müll et al.[22] discovered that that the preference of nonribosomal peptide synthetase for the natural substrate phenylalanine is 31 times higher than that for a singly fluorinated phenylalanine. But this preference is observed only when the fluorine substituent is positioned in the *para*-position, i.e. directly pointing towards the π-system of the aryl-group that represents the T-bar. If phenylalanine is fluorinated in *meta*- or *ortho*-position, the fluorinated phenylalanines are accepted as substrate with about the same likelihood as phenylalanine.

This drastic change in substrate activity can be explained through computational modeling, which reveals that the potential energy minimum of the T-shaped π-stacking interaction is almost completely erased if the partially positively charged hydrogen in *para*-position of phenylalanine is substituted by a partially negatively charged fluorine atom. By contrast, fluorination at other postitions of the phenyl ring besides the *para*-position slightly lowers the energy minimum and thereby stabilizes T-shaped π-stacking interaction. This is likely caused by the electron-withdrawing effect of the fluorine substituent in *meta*- or *ortho*-position, which increases the partial charge on the hydrogen atom in *para*-position. It is worth pointing out that the loss of the π-stacking interaction is a sizeable enthalpic effect, as it decreases the interaction strength (in the model system in Fig. 4) by about 10 kJ/mol.

Fluorine's impact on aromatic systems is particularly important in the context of drug discovery, given its frequent occurence attached to aromatic systems in pharmaceuticals. Here, we discussed the indirect effect fluorine may have on the interactions of aromatics in protein systems and we demonstrated how fluorine can directly disrupt aromatic interactions with proteins. While in this example, fluorine impacts the binding affinity unfavorably, it is reasonable to assume that fluorine may also have a favorable effect on binding affinities through direct interactions. In the next section, we will cover a particularly important type of such interactions, hydrogen bonds to fluorine substituents.

Hydrogen Bonds to Fluorine Substituents - Donor's Last Resort?

Hydrogen fluoride in the gas phase and in aqueous solution forms strong hydrogen bonds. Most notably, the hydrogen bond within the bifluoride anion, FHF–, is the strongest known hydrogen bond. With a dissociation energy of 161.5 kJ/mol,[55] it is in fact so strong that it is disputed whether the bond should be counted as hydrogen bond.[56] By contrast, the hydrogen bonds to fluorine substituents in organic molecules are much weaker and their influence on the stability of protein-ligand complexes has been discussed controversially.[57–62] Despite the high electronegativity of fluorine, fluorine subsitutents are weak hydrogen bond acceptors, which is attributed to fluorine's low polarizability and low charge capacity.[61,62] The hydrogen bond strength of C-F$\cdots$H-O is between 6 kJ/mol and 10 kJ/mol, depending on the hybridization of the C-atom.[58] This is less than half of the typical strength of a hydrogen bond O$\cdots$H-O, which is about 21 kJ/mol.[63] Dalvit et al. therefore conclude that intermolecular hydrogen bonds with fluorine as acceptor only occur in environments shielded from water and void of other competing acceptors.[61]

Nevertheless, fluorine hydrogen bonds are frequently observed in protein-ligand complexes, particularly if the fluorinated ligand is a small organic molecule. Protein targets include FXIa,[64] μ opioid receptor,[24] YAP:TEAD protein-protein interaction,[65] S1P receptor,[66] Akt1,[67] HIV protease,[68] tyrosinase,[69] Bruton's tyrosine kinase,[70] Janus kinases[71] and the SARS-CoV-2 main protease.[72–74] Thus, hydrogens bond with fluorine as acceptor are by no means a marginal phenomenon in protein-ligand systems. Whether they are a driving force for the stability of a protein-ligand complex and can thus be exploited as a design element is, however, questionable. Compared to other typical hydrogen bond acceptors in protein ligands (Fig.5), fluorine ranks rather low. Interestingly, fluorine is an even weaker acceptor than water, meaning that replacing a water molecule as acceptor from a protein donor with fluorine results in an enthalpic loss.

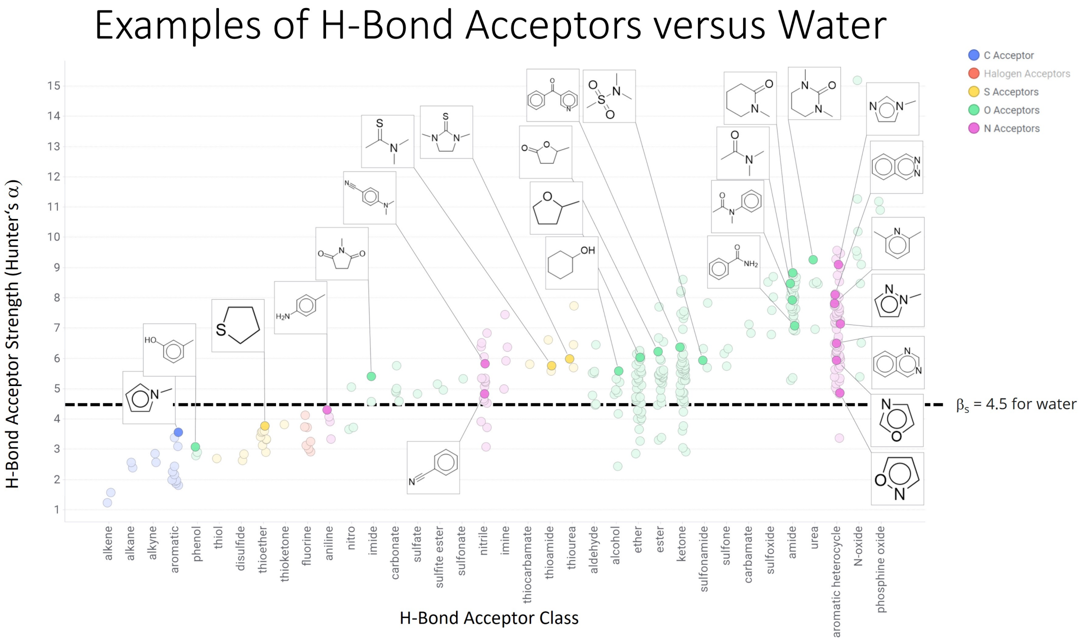


Figure 5: Hydrogen bond acceptor strength of common hydrogen bond acceptors in protein ligands found in medicinal chemistry. Reprinted with permission from Daryll McConnel.

A recent survey[21] of protein-ligand complexes in the protein data bank yielded more than 4000 complexes with fluorinated ligands. The authors identified hydrogen bond donors to the fluorine acceptor and evaluated the hydrogen bond interaction energy. When considering the difference in energy between the isolated hydrogen bonded structure and the donor and acceptor separated from each other, the hydrogen bond interaction energies range from 0 to -5.02 kJ/mol.

Fluorine was found to accept hydrogen bonds from OH, NH and CH donors, with CH being the most frequent donor. As expected, the interaction energy was depended on the donor-acceptor distance. However, it did not correlate with the F$\cdots$H-X angle, which indicates that the hydrogen bond is no necessarily aligned with a lone pair of the fluorine acceptor. Moreover, the distances and angles of the fluorine hydrogen-bonds did not coincide with the energy minima of the corresponding isolated hydrogen bonds. This leads to the conclusion, that the enthalpic gain due to the fluorine hydrogen-bonds is likely not the driving force of the binding affinity. The fluorine hydrogen bonds probably form in addition to stronger interactions that stabilize the ligand in the binding pocket. This causes the authors to taunt fluorine as ”donor’s last resort”.

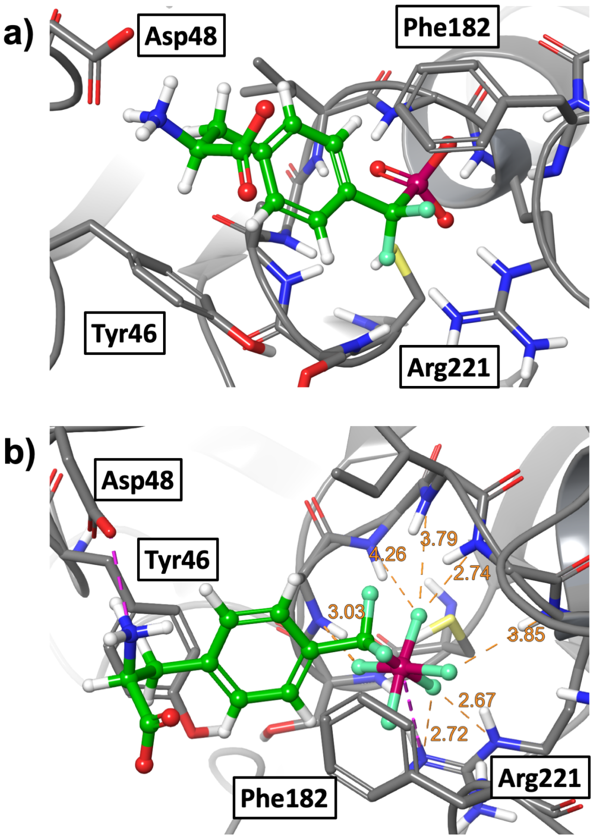


Figure 6: a) Phosphotyrosine mimetic amino acid with phosphate headgroup in the main binding pocket of PTP1B. b) Phosphotyrosine mimetic amino acid with the PF5 headgroup in the main binding pocket of PTP1B. Adapted from ref. 75. Available under a CC BY-NC license. Copyright 2022 Accorsi et al.

The acceptor strength of fluorine atoms can be boosted considerably by incorporating it into a larger functional group. For example, fluorine can replace the oxygen in a phosphate group, generating fluorinated analogues of phosphate groups. Accorsi et al.[75] used this strategy to improve the phosphotyrosine mimetic 4-phosphono-difluoromethyl-phenylalanine, which inhibits the tyrosine phosphatase PTP1B (Fig. 6). The phosphate group in the original inhibitor is replaced by pentafluorophosphato group (PF5), which improves the binding affinity from $K_I = 1555\,\mu$M to $K_I = 61\,\mu$M, where K_I is the inhibition constant measured in an enzyme inhibition assays. The pentafluorophosphato group has a valency of six, isoelectronically to a hexaflourophosphate anion, and a charge of -1. The expected charge state of a phopshate group is between -1 and -2 at pH 7. Thus, the increased binding affinity cannot simply be attributed to an overall increased Coulomb interaction between the positive binding pocket of PTP1B and the negatively charged functional group. Instead, computational docking and molecular dynamics simulations show that the PF5 headgroup sterically fits well into the binding pocket and forms several stable interactions between an arginine side chain and multiple protein backbone NH moieties. Similar fluorine specific interactions in the PTP1B binding pocket were also observed by Tiemann et al.,[76] where a trifluoromethylsulfonamide headgroup is placed into the binding pocket.

In conclusion, heavy fluorination of the phosphor headgroup draws so much negative charge density to this essential part of the inhibitor that the interaction to the very positive binding pocket is likely to be significantly enhanced by direct fluorine hydrogen bonds.

Water Networks

An interesting, yet highly controversial question is whether fluorinated ligands can bind to proteins via water-mediated hydrogen bonds. In this scenario, the fluorine substituent is in contact with water molecules in the binding pocket rather than with the protein surface. The hypothesis then is that the partial negative charge of the fluorine substituent and fluorine's capacity to accept hydrogen bonds structures the water-network and thereby stabilizes the ligand in the binding pocket. Computations of fluorinated amino acids and how they interact with water have shown that hydrogen-bond-like interactions between organic fluorine atoms and water molecules do occur and influence hydration free energies.[35,37,77] Additionally, ultrafast fluorescence spectroscopy revealed that fluorinated amino acid side chains can slow down water motion on protein surfaces.[78] These results suggest that fluorine substituents may indeed interact with proteins via water networks. However, detecting the change in the water network and disentangling the various interactions involved requires atomistic simulations and detailed computational analyses, as we will showcase in the following examples.

Van der Westhuizen et al.[79] analyzed the binding poses of a series of inhibitors of acetylcholinesterase using molecular docking and measured their activity with an enzyme inhibition assay. In one of the scaffolds, a pyridin group was replaced by a benzyl group or by a fluorobenzyl group. In the pyridine variant, the pyridine nitrogen forms a water-mediated bridge to a glycine residue deep within the binding pocket. This water-mediated contact seems to be critical for the stability of the inhibitor in the binding pocket and is present in most active compounds in this study. When a fluorobenzyl group instead of pyridin is present, the water-mediated contact can still be formed with water forming a (possible) hydrogen bond to fluorine, but the contact is expected to be much weaker. This would explain the reduced inhibitor activity of the fluorobenzyl variant compared to the pyridine variant. By contrast the benzyl variant cannot interact with the water molecule, and the absence of the water-mediated contact likely explains that all benzyl variants of the inhibitor were inactive (half maximal inhibitory concentration $> 50\ \mu$M).

The modulating effect of fluorine on a complex hydrogen bonded water network inside a protein environment can also be observed in the case of the μ opioid receptor. The μ opioid receptor is a membrane-bound G protein-coupled receptor, which exhibits a water network which stretches across the receptor. Lešnik et al.[24] studied the response of this water network to a fluorinated ligand. Specifically, they compared the unfluorinated ligand fentanyl and the fluorinated ligand NFEPP, which is identical to fentanyl except for a single fluorine substitution. They find the fluorinated ligand induces a markedly different protein-water hydrogen bond network than the unfluorinated fentanyl. These changes in the water network are relevant for drug design, since agonists of μ opioid receptor that selectively bind at low pH values are highly sought after.

Fluorination can also affect the binding affinity of protein ligands through entropic effects. Breiten et al.[26] study a series of similar ligands that differ by fluorine substitution in their binding thermodynamics to human carbonic anhydrase. The ligands show a very similar binding affinity, but the enthalpic and entropic contributions to this binding affinity widely varies across the ligands. Using Inhomogeneous Solvation Theory[80,81] based calculations, these effects can be linked to disruptions in the water network in the binding site, where the fluorine disrupts the water network in a way that causes less restricted water motion. This highlights how fluorine can be added to a protein binding ligand to specifically modulate a water network, but also how difficult it may be to predict the effect of fluorination to the binding strength of a protein ligand and that considerable efforts, involving theory, are needed to rationalize and quantify the specific effects, as they can vary in unexpected ways.

Ye et al.[23] demonstrated in inhibition assays that fluorination of the unnatural amino acid α-butyric acid at direct proximity of a water filled binding pocket in a protein-protein complex can restore inhibitor activity. Following the hypothesis, that the restoration in inhibitor strength is driven by the fluorine binding to the water network in the binding pocket, eventually establishing a water mediated bond to the protein, we investigated the water network of the complex and its interaction with the fluorinated amino acid using moelcular dynamics simulations.[82,83] While we found the water molecules in the binding pocket to be highly connected and binding to the protein receptor, we did not observe any hydrogen-bond like interactions with fluorine as acceptor.

Hydrogen bonds with fluorine as acceptor rarely seem to be the driver of ligand binding affinity. Rather fluorine substituents modulate complex molecular structures like protein-water hydrogen bond networks and thereby may stabilize or destabilize a ligand-protein complex. To disentangle and rationalize the various effects of fluorine substituents in ligand-protein system, detailed computational models are essential. This is particularly true if the fluorine substituent influences water networks at the interface between ligand and protein. Ref. 84 reviews recent computational methods to analyse protein-water interactions.

Entropic Effects

A direct consequence of the observation that the effects of fluorine substituents usually cannot be rationalized in terms of simple enthalpic effects is that entropic and in particular enthalpy-entropy compensation play a crucial role. We already touched on this topic in the context of fluorinated human carbonic anhydrase ligands and their effect on the water network, but will disucss it in more detail in this section.

A recent example for enthalpy-entropy compensation in a fluorinated system is the ligand binding to STING protein studied by Smola et al.[85] The authors compared the binding free energy of fluorinated and unfluorinated cyclic dinucleotides with isothermal calorimetry and computational QM and QM/MM calculations. The fluorinated ligands show more favourable binding entropy because of lower conformational flexibility in the unbound state and less entropic cost due to interactions with solvent and receptor. This effect is partially compensated by stronger enthalpic interactions of the unfluorinated ligands.

Fluorinated systems do not only exhibit enthalpy-entropy compensation, but also entropy-entropy compenstation, in which different kinds of entropy, i.e. protein conformational entropy, ligand conformational entropy and solvation entropy compensate each other. Wallerstein et al.[25] found entropy-entropy compensation in the protein-ligand systems of galectin-3 in complex with three different fluorinated ligands, which only differed in the position of fluorine on the phenyl ring of the ligand (o- m- or p-fluoro-phenyl, Fig. 7).

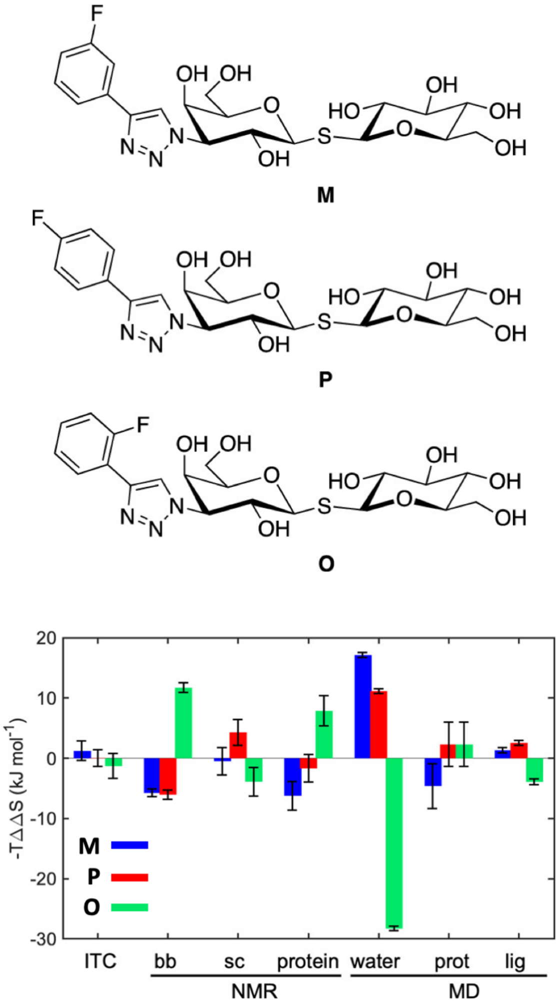


Figure 7: Entropic contributions to binding free energy of ligands **M**, **P** and **O**. $-T\Delta\Delta S$ is the difference of the entopic contribution to ligand binding $-T\Delta S$ between one complex and the average contribution of the two other complexes. Adapted from ref. 25. Available under a CC BY license. Copyright 2021 Wallerstein et al.

The authors study the binding thermodynamics of the three systems using multiple experimental and computational methods like X-ray crystallography, isothermal titration calorimetry, NMR relaxation, Molecular Dynamics and Grid Inhomogeneous Solvation Theory (GIST).[80,81] They find that the overall entropic contribution to binding is about equally strong in all three ligands. However, when the entropic contribution is decomposed into various contributions, either using NMR or MD simulations, each ligand exhibits a unique pattern of opposing entropic effects. Many of the entropic effects have a sizeable magnitude.

Specifically, the *o*-substituted ligand stands out from the other two ligands, which is almost entirely stabilized in the complex by the change in water entropy. The study showcases that detailed computational and experimental analysis may be needed to understand the influence of such a minor change in fluorination as the re-positioning of a single fluorine on the overall binding strength of a fluorinated ligand.

Finally, we want to note that fluorination can also give rise to compensating enthalpic effects and unexpected trends in physico-chemical properties like hydration free energy of fluorinated amino acids.[35] Compensating enthalpic effects may originate in changes in surface area, disruptions of backbone-water hydrogen bonds and changes in side chain polarity. Delineating and quantifying these effects separately can be achieved by alchemical free-energy perturbation.[86,87]

Conclusions

Fluorine is a powerful and versatile modulator of molecular interactions through its own unique properties. However, the effects of fluorination of such ligands can often be unexpected and difficult to rationalize. The support of computational methods can be essential for making rational predictions about the effects of fluorination. In this perspective, we covered different aspects of how computational methods help understand fluorine in protein binding ligands, including accurate force fields, hydrogen bond interactions, aryl groups, water networks and entropic effects. By showcasing the variety of fluorine interactions, we hope to provide a resource for researchers who design fluorinated protein-ligand complexes. Our objective is to provide clarity on what aspects to investigate and on computational methods to quantify these aspects.

Acknowledgements

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): Project-ID 387284271-SFB 1349.

References

(1) Purser, S.; Moore, P. R.; Swallow, S.; Gouverneur, V. Fluorine in Medicinal Chemistry. *Chemical Society Reviews* **2008**, *37*, 320–330.

(2) Berger, A. A.; Völler, J.-S.; Budisa, N.; Koksch, B. Deciphering the Fluorine Code—The Many Hats Fluorine Wears in a Protein Environment. *Accounts of Chemical Research* **2017**, *50*, 2093–2103.

(3) Inoue, M.; Sumii, Y.; Shibata, N. Contribution of Organofluorine Compounds to Pharmaceuticals. *ACS Omega* **2020**, *5*, 10633–10640.

(4) O'Hagan, D.; Young, R. J. Future Challenges and Opportunities with Fluorine in Drugs? *Medicinal Chemistry Research* **2023**, *32*, 1231–1234.

(5) Zhang, C. Fluorine in Medicinal Chemistry: In Perspective to COVID-19. *ACS Omega* **2022**, *7*, 18206–18212.

(6) Böhm, H.-J.; Banner, D.; Bendels, S.; Kansy, M.; Kuhn, B.; Müller, K.; Obst-Sander, U.; Stahl, M. Fluorine in Medicinal Chemistry. *ChemBioChem* **2004**, *5*, 637–643.

(7) Miller, M. A.; Sletten, E. M. Perfluorocarbons in Chemical Biology. *ChemBioChem* **2020**, *21*, 3451–3462.

(8) Gillis, E. P.; Eastman, K. J.; Hill, M. D.; Donnelly, D. J.; Meanwell, N. A. Applications of Fluorine in Medicinal Chemistry. *Journal of Medicinal Chemistry* **2015**, *58*, 8315–8359.

(9) Johnson, B. M.; Shu, Y.-Z.; Zhuo, X.; Meanwell, N. A. Metabolic and Pharmaceutical Aspects of Fluorinated Compounds. *Journal of Medicinal Chemistry* **2020**, *63*, 6315–6386.

(10) Gronenborn, A. M. Small, but Powerful and Attractive: 19F in Biomolecular NMR. *Structure* **2022**, *30*, 6–14.

(11) Murray, J. S.; Seybold, P. G.; Politzer, P. The many faces of fluorine: Some noncovalent interactions of fluorine compounds. *The Journal of Chemical Thermodynamics* **2021**, *156*, 106382.

(12) Caron, S. Where Does the Fluorine Come From? A Review on the Challenges Associated with the Synthesis of Organofluorine Compounds. *Organic Process Research & Development* **2020**, *24*, 470–480.

(13) Britton, R.; Gouverneur, V.; Lin, J.-H.; Meanwell, M.; Ni, C.; Pupo, G.; Xiao, J.-C.; Hu, J. Contemporary Synthetic Strategies in Organofluorine Chemistry. *Nature Reviews Methods Primers* **2021**, *1*, 1–22.

(14) Moschner, J.; Stulberg, V.; Fernandes, R.; Huhmann, S.; Leppkes, J.; Koksch, B. Approaches to Obtaining Fluorinated -Amino Acids. *Chemical Reviews* **2019**, *119*, 10718–10801.

(15) Dalvit, C.; Vulpetti, A. Ligand-Based Fluorine NMR Screening: Principles and Applications in Drug Discovery Projects. *Journal of Medicinal Chemistry* **2019**, *62*, 2218–2244.

(16) Mei, H.; Han, J.; Klika, K. D.; Izawa, K.; Sato, T.; Meanwell, N. A.; Soloshonok, V. A. Applications of Fluorine-Containing Amino Acids for Drug Design. *European Journal of Medicinal Chemistry* **2020**, *186*, 111826.

(17) Mei, H.; Han, J.; White, S.; Graham, D. J.; Izawa, K.; Sato, T.; Fustero, S.; Meanwell, N. A.; Soloshonok, V. A. Tailor-Made Amino Acids and Fluorinated Motifs as

Prominent Traits in Modern Pharmaceuticals. *Chemistry – A European Journal* **2020**, *26*, 11349–11390.

(18) Salwiczek, M.; Nyakatura, E. K.; Gerling, U. I. M.; Ye, S.; Koksch, B. Fluorinated Amino Acids: Compatibility with Native Protein Structures and Effects on Protein–Protein Interactions. *Chemical Society Reviews* **2012**, *41*, 2135–2171.

(19) Abula, A.; Xu, Z.; Zhu, Z.; Peng, C.; Chen, Z.; Zhu, W.; Aisa, H. A. Substitution Effect of the Trifluoromethyl Group on the Bioactivity in Medicinal Chemistry: Statistical Analysis and Energy Calculations. *Journal of Chemical Information and Modeling* **2020**, *60*, 6242–6250.

(20) Vulpetti, A.; Dalvit, C. Hydrogen Bond Acceptor Propensity of Different Fluorine Atom Types: An Analysis of Experimentally and Computationally Derived Parameters. *Chemistry – A European Journal* **2021**, *27*, 8764–8773.

(21) Pietruś, W.; Kafel, R.; Bojarski, A. J.; Kurczab, R. Hydrogen Bonds with Fluorine in Ligand–Protein Complexes-the PDB Analysis and Energy Calculations. *Molecules* **2022**, *27*, 1005.

(22) Müll, M.; Pourmasoumi, F.; Wehrhan, L.; Nosovska, O.; Stephan, P.; Zeihe, H.; Vilotijevic, I.; G. Keller, B.; Kries, H. Biosynthetic Incorporation of Fluorinated Amino Acids into the Nonribosomal Peptide Gramicidin S. *RSC Chemical Biology* **2023**, *4*, 692–697.

(23) Ye, S.; Loll, B.; Ann Berger, A.; Mülow, U.; Alings, C.; Christian Wahl, M.; Koksch, B. Fluorine Teams up with Water to Restore Inhibitor Activity to Mutant BPTI. *Chemical Science* **2015**, *6*, 5246–5254.

(24) Lešnik, S.; Bren, U.; Domratcheva, T.; Bondar, A.-N. Fentanyl and the Fluorinated Fentanyl Derivative NFEPP Elicit Distinct Hydrogen-Bond Dynamics of the Opioid Receptor. *Journal of Chemical Information and Modeling* **2023**,

(25) Wallerstein, J.; Ekberg, V.; Ignjatović, M. M.; Kumar, R.; Caldararu, O.; Peterson, K.; Wernersson, S.; Brath, U.; Leffler, H.; Oksanen, E.; Logan, D. T.; Nilsson, U. J.; Ryde, U.; Akke, M. Entropy–Entropy Compensation between the Protein, Ligand, and Solvent Degrees of Freedom Fine-Tunes Affinity in Ligand Binding to Galectin-3C. *JACS Au* **2021**, *1*, 484–500.

(26) Breiten, B.; Lockett, M. R.; Sherman, W.; Fujita, S.; Al-Sayah, M.; Lange, H.; Bowers, C. M.; Heroux, A.; Krilov, G.; Whitesides, G. M. Water Networks Contribute to Enthalpy/Entropy Compensation in Protein–Ligand Binding. *Journal of the American Chemical Society* **2013**, *135*, 15579–15584.

(27) Clark, T.; Hennemann, M.; Murray, J. S.; Politzer, P. Halogen Bonding: The Sigma-Hole. *Journal of Molecular Modeling* **2007**, *13*, 291–296.

(28) Zhu, Z.; Xu, Z.; Zhu, W. Interaction Nature and Computational Methods for Halogen Bonding: A Perspective. *Journal of Chemical Information and Modeling* **2020**, *60*, 2683–2696.

(29) Czarny, R. S.; Ho, A. N.; Shing Ho, P. A Biological Take on Halogen Bonding and Other Non-Classical Non-Covalent Interactions. *The Chemical Record* **2021**, *21*, 1240–1251.

(30) Ibrahim, M. A. A. AMBER Empirical Potential Describes the Geometry and Energy of Noncovalent Halogen Interactions Better than Advanced Semiempirical Quantum Mechanical Method PM6-DH2X. *The Journal of Physical Chemistry B* **2012**, *116*, 3659–3669.

(31) Franchini, D.; Dapiaggi, F.; Pieraccini, S.; Forni, A.; Sironi, M. Halogen Bonding in the Framework of Classical Force Fields: The Case of Chlorine. *Chemical Physics Letters* **2018**, *712*, 89–94.

(32) Carter, M.; Rappé, A. K.; Ho, P. S. Scalable Anisotropic Shape and Electrostatic Models

for Biological Bromine Halogen Bonds. *Journal of Chemical Theory and Computation* **2012**, *8*, 2461–2473.

(33) Orabi, E. A.; Faraldo-Gómez, J. D. New Molecular-Mechanics Model for Simulations of Hydrogen Fluoride in Chemistry and Biology. *Journal of Chemical Theory and Computation* **2020**, *16*, 5105–5126.

(34) Saito, K.; Torii, H. Hidden Halogen-Bonding Ability of Fluorine Manifesting in the Hydrogen-Bond Configurations of Hydrogen Fluoride. *The Journal of Physical Chemistry B* **2021**, *125*, 11742–11750.

(35) Robalo, J.; Verde, A. V. Unexpected Trends in the Hydrophobicity of Fluorinated Amino Acids Reflect Competing Changes in Polarity and Conformation. *Physical Chemistry Chemical Physics* **2019**, *21*, 2029–2038.

(36) Croitoru, A.; Park, S.-J.; Kumar, A.; Lee, J.; Im, W.; MacKerell, A. D. J.; Aleksandrov, A. Additive CHARMM36 Force Field for Nonstandard Amino Acids. *Journal of Chemical Theory and Computation* **2021**, *17*, 3554–3570.

(37) Robalo, J. R.; Huhmann, S.; Koksch, B.; Vila Verde, A. The Multiple Origins of the Hydrophobicity of Fluorinated Apolar Amino Acids. *Chem* **2017**, *3*, 881–897.

(38) Wang, X.; Li, W. Development and Testing of Force Field Parameters for Phenylalanine and Tyrosine Derivatives. *Frontiers in Molecular Biosciences* **2020**, *7*, 608931.

(39) Yang, D. T.; Gronenborn, A. M.; Chong, L. T. Development and Validation of Fluorinated, Aromatic Amino Acid Parameters for Use with the AMBER Ff15ipq Protein Force Field. *The Journal of Physical Chemistry A* **2022**, *126*, 2286–2297.

(40) Träg, J.; Zahn, D. Improved GAFF2 Parameters for Fluorinated Alkanes and Mixed Hydro- and Fluorocarbons. *Journal of Molecular Modeling* **2019**, *25*, 39.

(41) Boothroyd, S. et al. Development and Benchmarking of Open Force Field 2.0.0: The Sage Small Molecule Force Field. *Journal of Chemical Theory and Computation* **2023**, *19*, 3251–3275.

(42) Qiu, Y. et al. Development and Benchmarking of Open Force Field v1.0.0—the Parsley Small-Molecule Force Field. *Journal of Chemical Theory and Computation* **2021**, *17*, 6262–6280.

(43) Mobley, D. L.; Bannan, C. C.; Rizzi, A.; Bayly, C. I.; Chodera, J. D.; Lim, V. T.; Lim, N. M.; Beauchamp, K. A.; Slochower, D. R.; Shirts, M. R.; Gilson, M. K.; Eastman, P. K. Escaping Atom Types in Force Fields Using Direct Chemical Perception. *Journal of Chemical Theory and Computation* **2018**, *14*, 6076–6092.

(44) Takaba, K.; Pulido, I.; Henry, M.; MacDermott-Opeskin, H.; Chodera, J. D.; Wang, Y. Espaloma-0.3.0: Machine-learned Molecular Mechanics Force Field for the Simulation of Protein-Ligand Systems and Beyond. 2023.

(45) Wang, Y.; Fass, J.; Kaminow, B.; E. Herr, J.; Rufa, D.; Zhang, I.; Pulido, I.; Henry, M.; Macdonald, H. E. B.; Takaba, K.; D. Chodera, J. End-to-End Differentiable Construction of Molecular Mechanics Force Fields. *Chemical Science* **2022**, *13*, 12016–12033.

(46) Turman, D. L.; Cheloff, A. Z.; Corrado, A. D.; Nathanson, J. T.; Miller, C. Molecular Interactions between a Fluoride Ion Channel and Synthetic Protein Blockers. *Biochemistry* **2018**, *57*, 1212–1218.

(47) Yue, Z.; Wang, Z.; Voth, G. A. Ion Permeation, Selectivity, and Electronic Polarization in Fluoride Channels. *Biophysical Journal* **2022**, *121*, 1336–1347.

(48) Zhang, J.; Song, D.; Schackert, F. K.; Li, J.; Xiang, S.; Tian, C.; Gong, W.; Carloni, P.; Alfonso-Prieto, M.; Shi, C. Fluoride Permeation Mechanism of the Fluc Channel in Liposomes Revealed by Solid-State NMR. *Science Advances* **2023**, *9*, eadg9709.

(49) McIlwain, B. C.; Gundepudi, R.; Koff, B. B.; Stockbridge, R. B. The Fluoride Permeation Pathway and Anion Recognition in Fluc Family Fluoride Channels. *eLife* **2021**, *10*, e69482.

(50) Knox, H. J.; Rego Campello, H.; Lester, H. A.; Gallagher, T.; Dougherty, D. A. Characterization of Binding Site Interactions and Selectivity Principles in the 3β4 Nicotinic Acetylcholine Receptor. *Journal of the American Chemical Society* **2022**, *144*, 16101–16117.

(51) Shao, J.; Kuiper, B. P.; Thunnissen, A.-M. W. H.; Cool, R. H.; Zhou, L.; Huang, C.; Dijkstra, B. W.; Broos, J. The Role of Tryptophan in π Interactions in Proteins: An Experimental Approach. *Journal of the American Chemical Society* **2022**, *144*, 13815–13822.

(52) Lai, R.; Cui, Q. How to Stabilize Carbenes in Enzyme Active Sites without Metal Ions. *Journal of the American Chemical Society* **2022**, *144*, 20739–20751.

(53) Pietruś, W.; Kurczab, R.; Kafel, R.; Machalska, E.; Kalinowska-Tłuścik, J.; Hogendorf, A.; Żylewski, M.; Baranska, M.; Bojarski, A. J. How Can Fluorine Directly and Indirectly Affect the Hydrogen Bonding in Molecular Systems? – A Case Study for Monofluoroanilines. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **2021**, *252*, 119536.

(54) Pietruś, W.; Kurczab, R.; Kalinowska-Tłuścik, J.; Machalska, E.; Golonka, D.; Barańska, M.; Bojarski, A. J. Influence of Fluorine Substitution on Nonbonding Interactions in Selected Para-Halogeno Anilines. *ChemPhysChem* **2021**, *22*, 2115–2127.

(55) Larson, J.; McMahon, T. Gas-phase bihalide and pseudobihalide ions. An ion cyclotron resonance determination of hydrogen bond energies in XHY-species (X, Y= F, Cl, Br, CN). *Inorganic Chemistry* **1984**, *23*, 2029–2033.

(56) Dunning Jr, T. H.; Xu, L. T. Nature of the Bonding in the Bifluoride Anion, FHF–. *The Journal of Physical Chemistry Letters* **2021**, *12*, 7293–7298.

(57) Dunitz, J. D.; Taylor, R. Organic Fluorine Hardly Ever Accepts Hydrogen Bonds. *Chemistry – A European Journal* **1997**, *3*, 89–98.

(58) Howard, J. A. K.; Hoy, V. J.; O'Hagan, D.; Smith, G. T. How Good Is Fluorine as a Hydrogen Bond Acceptor? *Tetrahedron* **1996**, *52*, 12613–12622.

(59) Schneider, H.-J. Hydrogen Bonds with Fluorine. Studies in Solution, in Gas Phase and by Computations, Conflicting Conclusions from Crystallographic Analyses. *Chemical Science* **2012**, *3*, 1381–1394.

(60) Champagne, P. A.; Desroches, J.; Paquin, J.-F. Organic Fluorine as a Hydrogen-Bond Acceptor: Recent Examples and Applications. *Synthesis* **2015**, *47*, 306–322.

(61) Dalvit, C.; Invernizzi, C.; Vulpetti, A. Fluorine as a Hydrogen-Bond Acceptor: Experimental Evidence and Computational Calculations. *Chemistry – A European Journal* **2014**, *20*, 11058–11068.

(62) Murray, J. S.; Seybold, P. G.; Politzer, P. The Many Faces of Fluorine: Some Noncovalent Interactions of Fluorine Compounds. *The Journal of Chemical Thermodynamics* **2021**, *156*, 106382.

(63) Grinberg, N.; Carr, P. W. *Advances in Chromatography, Volume 57*; CRC Press, 2020; Vol. 57.

(64) Roehrig, S. et al. Design and Preclinical Characterization Program toward Asundexian (BAY 2433334), an Oral Factor XIa Inhibitor for the Prevention and Treatment of Thromboembolic Disorders. *Journal of Medicinal Chemistry* **2023**, *66*, 12203–12224.

(65) Sellner, H. et al. Optimization of a Class of Dihydrobenzofurane Analogs toward Orally

Efficacious YAP-TEAD Protein-Protein Interaction Inhibitors. *ChemMedChem* **2023**, *18*, e202300051.

(66) Yu, L.; He, L.; Gan, B.; Ti, R.; Xiao, Q.; Yang, X.; Hu, H.; Zhu, L.; Wang, S.; Ren, R. Structural Insights into Sphingosine-1-Phosphate Receptor Activation. *Proceedings of the National Academy of Sciences* **2022**, *119*, e2117716119.

(67) Li, G.; He, X.-H.; Li, H.-P.; Zhao, Q.; Li, D.-A.; Zhu, H.-P.; Zhang, Y.-H.; Zhan, G.; Huang, W. Design, Synthesis, and Biological Evaluation of Tetrahydro-α-carbolines as Akt1 Inhibitors That Inhibit Colorectal Cancer Cell Proliferation. *ChemMedChem* **2022**, *17*, e202200104.

(68) Ghosh, A. K. et al. Design, Synthesis and X-Ray Structural Studies of Potent HIV-1 Protease Inhibitors Containing C-4 Substituted Tricyclic Hexahydro-Furofuran Derivatives as P2 Ligands. *ChemMedChem* **2022**, *17*, e202200058.

(69) Mirabile, S.; Vittorio, S.; Paola Germanò, M.; Adornato, I.; Ielo, L.; Rapisarda, A.; Gitto, R.; Pintus, F.; Fais, A.; De Luca, L. Evaluation of 4-(4-Fluorobenzyl)Piperazin-1-Yl]-Based Compounds as Competitive Tyrosinase Inhibitors Endowed with Antimelanogenic Effects. *ChemMedChem* **2021**, *16*, 3083–3093.

(70) Kim, T.; Kim, K.; Park, I.; Hong, S.; Park, H. Two-Track Virtual Screening Approach to Identify the Dual Inhibitors of Wild Type and C481S Mutant of Bruton's Tyrosine Kinase. *Journal of Chemical Information and Modeling* **2022**, *62*, 4500–4511.

(71) Ojha, A. A.; Srivastava, A.; Votapka, L. W.; Amaro, R. E. Selectivity and Ranking of Tight-Binding JAK-STAT Inhibitors Using Markovian Milestoning with Voronoi Tessellations. *Journal of Chemical Information and Modeling* **2023**, *63*, 2469–2482.

(72) Thakur, A.; Sharma, G.; Badavath, V. N.; Jayaprakash, V.; Merz, K. M. J.; Blum, G.; Acevedo, O. Primer for Designing Main Protease (Mpro) Inhibitors of SARS-CoV-2. *The Journal of Physical Chemistry Letters* **2022**, *13*, 5776–5786.

(73) Mondal, S.; Chen, Y.; Lockbaum, G. J.; Sen, S.; Chaudhuri, S.; Reyes, A. C.; Lee, J. M.; Kaur, A. N.; Sultana, N.; Cameron, M. D.; Shaffer, S. A.; Schiffer, C. A.; Fitzgerald, K. A.; Thompson, P. R. Dual Inhibitors of Main Protease (MPro) and Cathepsin L as Potent Antivirals against SARS-CoV2. *Journal of the American Chemical Society* **2022**, *144*, 21035–21045.

(74) Khoury, L. E. et al. Computationally Driven Discovery of SARS-CoV-2 Mpro Inhibitors: From Design to Experimental Validation. *Chemical Science* **2022**, *13*, 3674–3687.

(75) Accorsi, M.; Tiemann, M.; Wehrhan, L.; Finn, L. M.; Cruz, R.; Rautenberg, M.; Emmerling, F.; Heberle, J.; Keller, B. G.; Rademann, J. Pentafluorophosphato-Phenylalanines: Amphiphilic Phosphotyrosine Mimetics Displaying Fluorine-Specific Protein Interactions. *Angewandte Chemie International Edition* **2022**, *61*, e202203579.

(76) Tiemann, M.; Nawrotzky, E.; Schmieder, P.; Wehrhan, L.; Bergemann, S.; Martos, V.; Song, W.; Arkona, C.; Keller, B. G.; Rademann, J. A Formylglycine-Peptide for the Site-Directed Identification of Phosphotyrosine-Mimetic Fragments. *Chemistry – A European Journal* **2022**, *28*, e202201282.

(77) R. Robalo, J.; Oliveira, D. M.; Imhof, P.; Ben-Amotz, D.; Verde, A. V. Quantifying How Step-Wise Fluorination Tunes Local Solute Hydrophobicity, Hydration Shell Thermodynamics and the Quantum Mechanical Contributions of Solute–Water Interactions. *Physical Chemistry Chemical Physics* **2020**, *22*, 22997–23008.

(78) Kwon, O.-H.; Yoo, T. H.; Othon, C. M.; Van Deventer, J. A.; Tirrell, D. A.; Zewail, A. H. Hydration Dynamics at Fluorinated Protein Surfaces. *Proceedings of the National Academy of Sciences* **2010**, *107*, 17101–17106.

(79) Van der Westhuizen, C. J.; Stander, A.; Riley, D. L.; Panayides, J.-L. Discovery of Novel Acetylcholinesterase Inhibitors by Virtual Screening, In Vitro Screening, and Molecular

Dynamics Simulations. *Journal of Chemical Information and Modeling* **2022**, *62*, 1550–1572.

(80) Lazaridis, T. Inhomogeneous Fluid Approach to Solvation Thermodynamics. 1. Theory. *The Journal of Physical Chemistry B* **1998**, *102*, 3531–3541.

(81) Nguyen, C. N.; Kurtzman Young, T.; Gilson, M. K. Grid Inhomogeneous Solvation Theory: Hydration Structure and Thermodynamics of the Miniature Receptor Cucurbit[7]Uril. *The Journal of Chemical Physics* **2012**, *137*, 044101.

(82) Wehrhan, L.; Leppkes, J.; Dimos, N.; Loll, B.; Koksch, B.; Keller, B. G. Water Network in the Binding Pocket of Fluorinated BPTI–Trypsin Complexes-Insights from Simulation and Experiment. *The Journal of Physical Chemistry B* **2022**, *126*, 9985–9999.

(83) Wehrhan, L.; Keller, B. G. Pre-bound State Discovered in the Unbinding Pathway of Fluorinated Variants of the Trypsin-BPTI Complex Using Random Acceleration Molecular Dynamics Simulations. *BioRxiv* **2024**, 2024–02.

(84) Mukherjee, S.; Schäfer, L. V. Spatially Resolved Hydration Thermodynamics in Biomolecular Systems. *The Journal of Physical Chemistry B* **2022**, *126*, 3619–3631.

(85) Smola, M.; Gutten, O.; Dejmek, M.; Kožíšek, M.; Evangelidis, T.; Tehrani, Z. A.; Novotná, B.; Nencka, R.; Birkuš, G.; Rulíšek, L.; Boura, E. Ligand Strain and Its Conformational Complexity Is a Major Factor in the Binding of Cyclic Dinucleotides to STING Protein. *Angewandte Chemie International Edition* **2021**, *60*, 10172–10178.

(86) Mobley, D. L.; Gilson, M. K. Predicting Binding Free Energies: Frontiers and Benchmarks. *Annual Review of Biophysics* **2017**, *46*, 531–558.

(87) Cournia, Z.; Allen, B.; Sherman, W. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *Journal of Chemical Information and Modeling* **2017**, *57*, 2911–2937.

4.2 Fluorinated BPTI-Trypsin: Water Network

Title: Water Network in the Binding Pocket of Fluorinated BPTI–Trypsin Complexes–Insights from Simulation and Experiment

The complex between bovine trypsin and the Bovine Pancreatic Trypsin Inhibitor (BPTI) is a tightly bound inhibitory protein-protein complex. The exceptionally strong inhibition of wildtype-BPTI is significantly reduced when the the crucial amino acid Lys15 is replaced by the much shorter and aliphatic non-natural amino acid α-aminobutyric-acid (Abu). This can easily be rationalized by considering that the the positively charged end of the lysine side chain is missing and that it cannot bind to the negatively charged amino acids at the bottom of the main binding pocket (S1 pocket) of trypsin. As the Abu side chain is shorter than lysine, it does not occupy as much space in the binding pocket as lysine and hence, the open space is occupied by two additional water molecules. Interestingly, when the Abu side chain is fluorinated, yielding the mono- di- and tri-fluorinated variants MfeGly, DfeGly and TfeGly, some of the inhibitor activity is restored. We followed the hypothesis that the restoration of inhibitor strength is caused by hydrogen bond like interactions between fluorine and the water molecules inside the main binding pocket of the BPTI-trypsin complex and pursued a better understanding of the water dynamics and network inside the trypsin S1 pocket and how it reacts to fluorine. The complex of trypsin with the Abu-BPTI variants has not been studied computationally, using MD simulations, before. Here, we use a new self-parameterized force field for the fluorinated amino acids and Abu, which follows the protocol of Robalo et al.[35] Our results give novel insights on the hydrogen bond network in the S1 pocket of fluorinated BPTI-trypsin complexes and about the free energy landscape of unbinding. We used the following methods:

- Unbiased MD simulations including analysis methods like RMSF, amino acid side chain dihedral analysis and water dynamics analysis
- Hydrogen bond detection based on MD simulations
- Force field parameterization
- Umbrella sampling for protein complex dissociation
- Grid Inhomogeneous Solvation Theory (GIST)

Using umbrella sampling of the unbinding process of the four complexes, we confirmed computationally a stepwise increase in binding strength by fluorinating Abu. The potential of mean force profiles show well defined minima for the fully bound complexes of trypsin with the Abu-BPTI variants and a rugged unbinding path with multiple additional minima. The unbinding path is not sufficiently sampled, so the paths for the fluorinated variants cannot be compared to each other, but the multiple minima indicate that there are likely additional metastable states along the unbinding path. Regarding the fully bound complexes, we find that the water molecules observed in our MD simulations populate the same positions as observed in X-ray crystallography structures of the Abu-BPTI-trypsin complexes, however are highly dynamic. The water molecules leave and enter the binding pocket on a sub-nanosecond scale. Not only do the water molecules leave and enter the binding pocket, but they also exchange their positions within the pocket rapidly. Our analysis of average water lifetimes shows that the mobility does not seem to be decreased by the presence of fluorine in a stepwise manner. Calculations of the RMSF of the binding pocket waters does not show any significant differences of the water molecule fluctuations with respect to fluorination. Moreover, the water molecules in the S1 binding pocket are strongly interconnected by hydrogen bonds, forming a hydrogen bond network. Despite the strong interconnection between each other, the water molecules do not form hydrogen bond like interactions to the fluorine atoms at all, which leads us to the conclusion that direct hydrogen bond like interactions between fluorine and binding pocket water is likely not the reason for the restoration of inhibitor strength by fluorination.

Bettina Keller and I developed the concept of the publication in collaboration with Beate Koksch and Jakob Leppkes. I conducted all simulations and their analysis using all the computational methods listed above. Bettina Keller and I wrote and revised the manuscript and the other co-authors commented on the manuscript. Jakob Leppkes conducted and analyzed the inhibition assays. Nicole Dimos and Bernhard Loll measured the X-ray crystal structure of the MfeGly-BPTI-trypsin complex.

The research presented here was published in: L. Wehrhan et al. *J. Phys. Chem. B* **2022**. DOI: https://doi.org/10.1021/acs.jpcb.2c05496.

4.3 Fluorinated BPTI-Trypsin: Pre-Bound State

Title: Pre-bound State Discovered in the Unbinding Pathway of Fluorinated Variants of the Trypsin-BPTI Complex Using Random Acceleration Molecular Dynamics Simulations

Given the rugged unbinding pathways observed in the previous publication,[120] it is likely that the unbinding pathways of the Abu-BPTI variants include additional metastable states, which differ from the fully bound state, which can be measured by X-ray crystallography. Possibly, some of the states are also inhibitory, which means that if fluorination has an impact on these states, it would affect the experimentally observable inhibitory activity of the Abu-BPTI variants. Here, we attempt to characterize the unbinding path of the Abu-BPTI variants and we find a pre-bound state, which is similar to the fully bound state but can be clearly differentiated from that state. We characterize this pre-bound state and analyze how fluorine impacts the transition from the fully bound state to the pre-bound state. We employed the following methods:

- Random Acceleration Molecular Dynamics (RAMD)
- Unbiased MD simulations
- Umbrella sampling
- Metadynamics

By the time of the preprint publication, this was the first application of RAMD to a protein-protein complex. The pre-bound state was unknown before, for complexes with wildtype-BPTI and also for complexes of BPTI mutants and here, we characterized this novel state.

By studying the dissociation pathway of the Abu-BPTI variants from trypsin with restrained metadynamics, we find that the BPTI variants do not re-bind after dissociation, indicating a curved binding and unbinding pathway. Using RAMD simulations, we find a metastable state along such a curved pathway, that we call the pre-bound state. This novel metastable state occurs in the dissociation trajectories of all BPTI variants and is stable for a significant amount of simulation time despite a strong biasing force. The pre-bound state differs from the fully bound state in the positional coordinates of the Abu-BPTI variants, which are shifted in their center-of-mass distance with respect to trypsin and also slightly rotated. The states also differ in their interaction pattern as three hydrogen bonds in the protein-protein interface, around Arg17 of the BPTI variants, are broken in the pre-bound state. Using unbiased MD simulations, we demonstrate that the pre-bound state is stable for at least 50 ns, but considering that we ran extensive simulations with an aggregated simulation time of 1 μs per BPTI variant its lifetime is likely to be on the scale of 100 ns. By measuring the distribution of the states in the unbiased simulations, we could quantify the shift in positional coordinates as 0.2 nm in center-of-mass distance and a 10° shift on the angular coordinates. The overlap of the two states in these coordinates is small. Moreover, we confirm the change in interaction pattern, already observed in the RAMD trajectories. As three hydrogen bonds around Arg17 between the BPTI variants and trypsin occur frequently in the simulations of the fully bound state but not in the simulations of the pre-bound state. In exchange for the hydrogen bonds, there is a new cation-pi interaction in the pre-bound state, which does not occur in the fully bound state. As there is little structural rearrangement for the amino acid residues close to the main binding pocket and the active site, it is likely that the pre-bound state is also inhibitory. We then employed umbrella sampling simulations using the distance between acceptor and donor of one of these hydrogen bonds (between Ph41 and Arg17). We find that fluorination lowers the barrier of transition between the fully bound state and the pre-bound state and also the minimum of the pre-bound state with respect to the fully bound state. We speculate that this might be due to an interaction of the negatively charged fluorine with the side chain of trypsin's Gln194, which can be found in close vicinity of the fluorine atoms in the pre-bound state.

Bettina Keller and I developed the concept for the manuscript. I conducted all simulations and analyses, including unbiased MD, RAMD, umbrella sampling and metadynamics. Bettina Keller and I wrote and revised the manuscript.

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Open Access

pubs.acs.org/jcim

Article

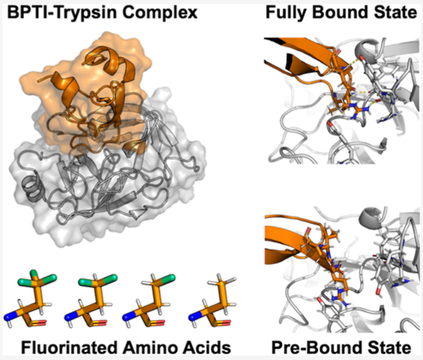
Prebound State Discovered in the Unbinding Pathway of Fluorinated Variants of the Trypsin–BPTI Complex Using Random Acceleration Molecular Dynamics Simulations

Leon Wehrhan and Bettina G. Keller*

Cite This: https://doi.org/10.1021/acs.jcim.4c00338

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: The serine protease trypsin forms a tightly bound inhibitor complex with the bovine pancreatic trypsin inhibitor (BPTI). The complex is stabilized by the P1 residue Lys15, which interacts with negatively charged amino acids at the bottom of the S1 pocket. Truncating the P1 residue of wildtype BPTI to α-aminobutyric acid (Abu) leaves a complex with moderate inhibitor strength, which is held in place by additional hydrogen bonds at the protein–protein interface. Fluorination of the Abu residue partially restores the inhibitor strength. The mechanism with which fluorination can restore the inhibitor strength is unknown, and accurate computational investigation requires knowledge of the binding and unbinding pathways. The preferred unbinding pathway is likely to be complex, as encounter states have been described before, and unrestrained umbrella sampling simulations of these complexes suggest additional energetic minima. Here, we use random acceleration molecular dynamics to find a new metastable state in the unbinding pathway of Abu-BPTI variants and wildtype BPTI from trypsin, which we call the prebound state. The prebound state and the fully bound state differ by a substantial shift in the position, a slight shift in the orientation of the BPTI variants, and changes in the interaction pattern. Particularly important is the breaking of three hydrogen bonds around Arg17. Fluorination of the P1 residue lowers the energy barrier of the transition between the fully bound state and prebound state and also lowers the energy minimum of the prebound state. While the effect of fluorination is in general difficult to quantify, here, it is in part caused by favorable stabilization of a hydrogen bond between Gln194 and Cys14. The interaction pattern of the prebound state offers insights into the inhibitory mechanism of BPTI and might add valuable information for the design of serine protease inhibitors.

INTRODUCTION

Proteases are enzymes that play a crucial role in the breakdown of peptides by catalyzing the hydrolysis of peptide bonds. Among them, serine proteases form a subgroup that catalyzes this reaction via a serine residue in their active site. Serine proteases are found in most life forms including bacteria, viruses, fungi, plants, and animals. They play essential roles in digestion, signal transduction, blood clotting, immune responses, and other cellular functions. In humans, serine proteases are important drug targets for many diseases including cardiovascular, cancer, and infectious diseases.[1,2] An example of a serine protease is trypsin, which is a mammalian digestive enzyme. It has been widely used as a model system for serine proteases since it exhibits the most prevalent fold for proteases in humans and higher organisms.[3,4]

Protein–protein complexes involving trypsin are stabilized by a long positively charged residue located on the binding protein, the P1 residue, which reaches into the deep S1 binding pocket of trypsin (Schechter and Berger notation).[5] The S1 pocket is lined with negatively charged residues which either bind directly to the positive charge or via water-mediated contacts.[6,7] Trypsin's catalytic site is located at the rim of the S1 binding pocket. Most proteins that bind to trypsin in this manner are cleaved at the C-terminal side of their P1 residue and, thus, act as substrates. Some proteins, despite binding to the S1 pocket, are not cleaved and instead act as inhibitors toward trypsin. Examples are bovine pancreatic trypsin inhibitor (BPTI), antitrypsin, or serpins. Several mechanisms have been put forward to explain why these proteins are not hydrolyzed by trypsin but instead form such a stable trypsin-inhibitor complex. Possibly, initial hydrolysis might take place, but relegation of the cleaved bond is fast and thus favored over release of the hydrolyzed product.[8] In the clogged gutter mechanism,[9] the hydrolyzed products are bound in a tight and specific orientation to trypsin, such that product release is hindered.

Received: February 27, 2024
Revised: June 2, 2024
Accepted: June 4, 2024

ACS Publications

We here focus on BPTI, which is an exceptionally well-studied protein[9−13] and inhibits trypsin with an extraordinarily high binding affinity (the binding constant is $K_i = 5 \times 10^{-24}$ M).[14] Its P1 residue is Lys15, which forms water-mediated bonds to Asp189 and Ser190 at the bottom of the S1 pocket. The importance of Lys15 for the binding process has been demonstrated by kinetic studies with BPTI mutants, where the K15A mutant BPTI shows dramatically decreased binding affinity.[14] Interestingly, the K15A mutant is also the only variant that has a significantly decreased association rate, highlighting the importance of the P1 residue for trypsin−BPTI recognition.

While for complexes of proteins with small molecules, like the trypsin−benzamidine complex, the full energy landscape of the binding and unbinding process has been calculated,[15] and the computational characterization of the binding equilibrium in protein−protein complexes, like the BPTI−trypsin complex, is much more challenging.[16,17] The reasons for this include the slow movements of macromolecules along the translational and rotational degrees of freedom. Also, the number of possible contact conformations of a protein−protein complex far exceeds that of a protein−small-molecule complex. In computational studies of protein−protein complexes, additional restraints to the relative position and orientation may be applied to increase the sampling of the binding/unbinding process.[18,19] However, this requires knowledge of the exact binding/unbinding path to obtain an accurate free-energy profile and characterize relevant intermediate states.

Kahler et al.[20] studied the binding/unbinding process of wildtype BPTI with trypsin using unbiased simulations, seeded by umbrella simulations. They describe the binding/unbinding process as a two-step mechanism, in which trypsin and BPTI recognize each other first through Coulomb interactions and form encounter states before moving on to form the fully bound protein−protein complex.

We here study the unbinding process of BPTI variants where the Lys15 residue has been mutated to α-aminobutyric acid (Abu) and its mono- (MfeGly), di- (DfeGly), and trifluorinated (TfeGly) variants. These BPTI variants are not cleaved by trypsin but instead act as moderate inhibitors with half-maximal inhibitory concentration (IC_{50}) of $IC_{50} = 4 \times 10^{-7}$ M (Lys15Abu) and $IC_{50} = 6 \times 10^{-8}$ M (Lys15TfeGly).[21,22]

Interestingly, the inhibitor strengths of the four BPTI variants systematically increase with increasing fluorination. An initial hypothesis based on crystal structures of the complexes suggested that this increase in binding affinity could be traced back to direct and specific interactions of the fluorine substituents with the water molecules in the S1 pocket.[21] In a recent computational study with molecular dynamics (MD) simulations, we did not find significant differences in the water structure or water−protein interaction strength across the four variants of the BPTI−trypsin complex[22] and thus could not confirm this hypothesis. However, a rough scan using umbrella sampling of the unbinding pathways hinted at a second free-energy minimum next to the bound state. This prebound state was closer to the bound state than encounter states,[20] which could also be identified in our scan of the unbinding pathway. The existence of a prebound state might offer insights into why BPTI acts as an inhibitor rather than a substrate to trypsin and might open up new avenues for the design of trypsin inhibitors.

In this contribution, we investigate the unbinding path of the four BPTI variants Abu-BPTI, MfeGly-BPTI, DfeGly-BPTI, and TfeGly-BPTI using random acceleration molecular dynamics (RAMD) simulations.[23−26] Additionally, we also study the unbinding process of wildtype BPTI.

RAMD is an enhanced sampling method that applies an additional biasing force, which is randomly redirected throughout the simulation, to the center of mass of a ligand and thereby facilitates the exploration of curved unbinding pathways.[23] RAMD has been used frequently for complexes of proteins with small molecules, but to the best of our knowledge, this is the first application of RAMD on a protein−protein complex. Our goal is to verify the presence of the prebound state and to explain its stability.

■ METHODS

Collective Variables. We constructed the collective variables describing the position and orientation of the BPTI variants with respect to trypsin from the positions of three reference points in trypsin (T1, T2, and T3) and three reference points in (Abu, MfeGly, DfeGly, and TfeGly)-BPTI (B1, B2, and B3), adapted from ref 18. The three reference points for the enzyme trypsin were defined to be the center of mass of the backbone of the whole enzyme (T1), the backbone of Val233-Ala241 (T2), and the backbone of Gln46-Leu67 (T3). The reference points of the ligands (Abu, MfeGly, DfeGly, and TfeGly)-BPTI were defined to be the center of mass of the backbone of the whole ligand (B1), the backbone of Ala48-Thr54 (B2), and the backbone of Cys14-Ala16 (B3). The positions of the reference points in the starting structure are shown in Figure 1. The main collective variable is the center-of-mass distance, r, between the enzyme and the ligand (T1−B1). The angle θ_o (T1−B1−B2) and dihedrals ϕ_o (T2−T1−B1−B2) and ψ_o (T1−B3−B1−B2) describe the orientation of the ligand with respect to the enzyme. The angle Θ_p (T2−T1−B1) and dihedral Φ_p (T3−T2−T1−B1) describe the position of the ligand with respect to the enzyme. The collective variables r, Θ_p, Φ_p, θ_o, ϕ_o, and ψ_o were calculated with Plumed 2.8[27,28] for all simulations.

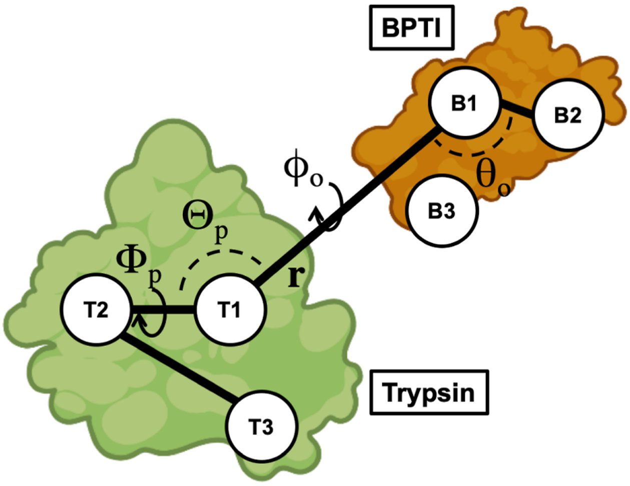


Figure 1. Method to construct collective variables that describe the position and orientation of the BPTI variants with respect to trypsin. The position relative to trypsin is described by Θ_p (T2−T1−B1) and Φ_p (T3−T2−T1−B1). The orientation is described by θ_o (T1−B1−B2), ϕ_o (T2−T1−B1−B2), and ψ_o (T1−B3−B1−B2). **r** (T1−B1) is the center-of-mass distance. Compare ref 18.

Molecular Dynamics General Methods. We ran all MD simulations using GROMACS[29−31] software and our selfparametrized Amber14SB force field.[22,32−34] Energy minimizations were conducted with the steepest descend algorithm. Equilibrations in the *NVT* ensemble were using a velocity rescaling scheme with a stochastic term[35] to keep the temperature at 300 K and harmonic restraints were applied on all protein heavy atom positions. Subsequent equilibrations in the *NPT* ensemble without restraints made use of the same velocity rescaling scheme with a stochastic term and the Parinello−Rahman barostat[36] to keep the temperature at 300 K and the pressure at 1.0 bar. Production MD simulations were run in the *NPT* ensemble at 300 K and 1.0 bar by using the same thermostat and barostat. All MD simulations were performed with the leapfrog integrator and an integration time step of 2 fs. Bond lengths involving hydrogen atoms were kept constant using the LINCS[37] algorithm. Long-range electrostatic interactions above a cutoff distance of 1.0 nm were treated using the PME[38] algorithm.

Starting Structure Preparation. Starting structures were generated from the crystal structure of the TfeGly−BPTI−trypsin complex (pdb code: 4Y11).[21] Cosolutes and ions were deleted, and appropriate hydrogen atoms were added to the crystal structure using the pdbfixer software. The three histidine side chains in the complex were protonated at N(ϵ) and N(δ). From this initial starting structure, the TfeGly residue was transformed into DfeGly, MfeGly, and Abu, respectively, to yield one initial starting structure for every BPTI variant. For the RAMD simulations, the initial starting structures were placed inside a cubic box with periodic boundary conditions with a 2.1 nm distance between the solute and the box edges and solvated in TIP3P[39] water. The systems were energy minimized and equilibrated in the *NVT* ensemble for 100 ps, followed by equilibration in the *NPT* ensemble for 1 ns. To generate two more replicas for each of the complexes, two subsequent simulations of 10 ns were run with the equilibrated starting structures to yield the starting structures for the next replicas.

Random Acceleration Molecular Dynamics. We used RAMD to explore unbinding pathways of (Abu, MfeGly, DfeGly, and TfeGly)-BPTI and wildtype BPTI from trypsin using GROMACS2020.5-RAMD-2.0. Two pull groups were defined: one included all atoms of trypsin, and the other included all atoms of (Abu, MfeGly, DfeGly, and TfeGly)-BPTI. A random force acting between the two pull groups with a magnitude of 3500 kJ/(mol nm) was applied. The appropriate force was estimated by running single RAMD simulations of the TfeGly-BPTI−trypsin complex starting with a force of 250 kJ/(mol nm) and raising the force by 250 kJ/(mol nm) every simulation until dissociation within 10 ns was achieved. Retrospectively, higher forces between 4000 kJ/(mol nm) and 5500 kJ/(mol nm) were tested with the same system to see when the RAMD simulations would fail to detect the prebound state at all. Three starting structures for each of the four complexes of trypsin with Abu-BPTI, MfeGly-BPTI, DfeGly-BPTI, and TfeGly-BPTI were generated as described above. For every one of these replicas, ten RAMD simulations were run from the same starting structure, where the random seed of the random force was changed. The simulations were stopped after dissociation was achieved, and the maximum length of the simulations was set to be 40 ns. At the beginning of the simulations, the direction of the biasing force was chosen at random. Throughout the simulations, after every 100 fs, the direction of the force was either retained, if the center of mass of the second pull group moved by more than 0.0025 nm, or changed randomly, if this was not the case. Snapshots were extracted every 2 ps.

Unbiased Molecular Dynamics of the Prebound State. We ran unbiased MD simulations to sample the fully bound state and the prebound state of the four complexes of trypsin with Abu-BPTI, MfeGly-BPTI, DfeGly-BPTI, TfeGly-BPTI, and wildtype BPTI using GROMACS2021.5,[29−31] patched with Plumed 2.8[27,28]. For every complex and state, 20 simulations of 50 ns length were run, totaling 160 simulations with an aggregated length of 8 μs. Initial starting structures for the simulations of the fully bound state were generated from the pdb structure of TfeGly-BPTI as described above. The initial starting structures were placed in a cubic box with periodic boundary conditions with a 1.5 nm distance between the solute and the box edges and solvated in TIP3P water. Then, 20 starting structures of every complex for the production MD simulations were generated by individually energy-minimizing the systems, followed by equilibration in the *NVT* ensemble for 100 ps and equilibration in the *NPT* ensemble for 1 ns. Initial starting structures for the simulations of the prebound state were generated by extracting the coordinates of all protein atoms of 20 snapshots of one RAMD simulation of the TfeGly-BPTI−trypsin complex, when the system was in the prebound state. The TfeGly residue was transformed into DfeGly, MfeGly, and Abu to yield initial starting structures for the other three complexes. The initial starting structures were individually placed in a box with periodic boundary conditions and energy minimized and equilibrated in the same way as the starting structures for the fully bound state. Production MD simulations were run for a length of 50 ns for every replica. Snapshots were extracted every 10 ps.

Analysis of Distances, Hydrogen Bonds, and SASA. We calculated atomic distances and detected hydrogen bonds in simulation snapshots using the Python package MDTraj 1.9.4[40]. Hydrogen bonds were detected using the Wernet−Nilsson criterion[41] implemented in MDTraj

$$r_{\mathrm{DA}} < 0.33\mathrm{nm} - 0.00044\cdot\delta_{\mathrm{HDA}}^{2} \tag{1}$$

with the donor−acceptor distance r_{DA} and the angle between the hydrogen atom, donor, and acceptor δ_{HDA}.

Distances to the nitrogen atoms in the guanidine moieties of arginine side chains were calculated by computing the distance of the respective interaction partner to all three nitrogen atoms of the guanidine moiety and taking the minimum of these three distances for every simulation snapshot.

Solvent-accessible surface area (SASA) was calculated using the MDTraj implementation of the Shrake Rupley algorithm.[42] The SASA for residues was calculated by summing over the atoms in each residue.

Umbrella Sampling. We conducted umbrella sampling using GROMACS2021.5,[29−31] patched with Plumed 2.8[27,28], based on the distance between the backbone oxygen of Phe41 (Phe41-O) of trypsin and the backbone nitrogen of Arg17 (Arg17-N) of the BPTI variants as the main collective variable ξ. Starting structures for the umbrella windows were generated starting from the fully bound crystal structure of the Abu-BPTI−trypsin complex, as described above. An initial harmonic restraint with a force constant of 6276 kJ/(mol nm^2) was placed at ξ = 0.25 nm. The system was equilibrated in the *NPT* ensemble with this harmonic restraint for 500 ps to yield the starting structure of the first umbrella window. Then, the harmonic potential was shifted by 0.05 nm, and a new *NPT* equilibration of 500 ps was run to yield the starting structure of the next window. This procedure was repeated until ξ reached a value of 0.70 nm. Additional windows were added in-between at ξ values of 3.75, 4.25, and 4.75 nm to achieve better sampling of the region of the free-energy barrier. Finally, there were 13 umbrella windows at the following positions of ξ (all in nm): 0.250, 0.300, 0.350, 0.375, 0.400, 0.425, 0.450, 0.475, 0.500, 0.550, 0.600, 0.650, and 0.700. In each of the umbrella windows, a production MD simulation with a harmonic restraint and a force constant of 6276 kJ/(mol nm^2) was run for a length of 30 ns. The potential of mean force profiles was calculated using binless WHAM.[43,44] Statistical uncertainty was estimated using a simplified bootstrapping scheme: The simulations of every window were separated into five parts of 6 ns length. Then, for every window, five combinations of four of these parts were constructed by combining all parts but one. The WHAM calculation was performed on all of these five combinations and the mean and standard deviation of the resulting potential of mean force profiles was calculated.

RESULTS AND DISCUSSION

Protein–Protein Complex between Trypsin and BPTI Variants. We consider BPTI variants in which the P1 residue is substituted by Abu and its MfeGly, DfeGly, and TfeGly variants, i.e., K15Abu, K15MfeGly, K15DfeGly, and K15TfeGly. The crystal structures of all four BPTI variants (pdb codes: 4Y0Z, 7PH1, 4Y10, 4Y11, and 4Y0Y) are similar to each other and to the wildtype complex.[21,22] The interaction strength and pattern between the P1 residue and the S1 pocket and water molecules within the S1 pocket did not differ significantly across BPTI variants, and thus did not explain the observed differences in the stability of the protein–protein complexes.[22] Figure 2 shows the complex and binding interface of the TfeGly-BPTI complex as a representative.

Besides the S1–P1 interactions, the complex is stabilized by hydrogen bonds throughout the entire protein–protein interface. Figure 2b shows that the P1 residue TfeGly is held in place by seven hydrogen bond-like contacts, most notably three backbone interactions holding the backbone carbonyl of the P1 residue in the oxyanion hole of the catalytic pocket. To the left in Figure 2b, Arg39 of the BPTI variant can be seen in two alternative conformations, forming interactions with either the side chain or the backbone of Asn97 in trypsin.

Figure 2c shows the other side of the interface. On this side of the interface, Arg17 (P2′ residue) of the BPTI variant forms interactions with its side chain to the backbone of His40 and with its backbone to the backbone of Phe41. The P4′ residue Ile19 forms an interaction with the side chain of Tyr39 in trypsin. This interaction has been described as important for the binding of BPTI to trypsin, as Y39A mutants of trypsin are less sensitive to BPTI.[3]

Collective Variables and Metadynamics with Restraints. To investigate the dissociation of the complexes between the BPTI variants and trypsin, we designed a set of collective variables that describe the position and orientation of the (Abu, MfeGly, DfeGly, TfeGly)-BPTI with respect to trypsin, following ref 18. The collective variables are based on the backbone center of mass of the BPTI variants (B1) and trypsin (T1) and two additional points for each of the proteins (T2, T3 and B2, B3), defined as centers-of-mass of well-structured regions inside the proteins (Figure 1). The position of the BPTI variant relative to trypsin is then given by the distance r between B1 and T1, the angle $\Theta_p = \angle T2{-}T1{-}B1$, and the dihedral angle $\Phi_p = \angle T3{-}T2{-}T1{-}B1$. The orientation of the BPTI variant relative to trypsin is given by the angle $\theta_o = \angle T1{-}B1{-}B2$ and dihedrals $\phi_o = \angle T2{-}T1{-}B1{-}B2$ and $\psi_o = \angle T1{-}B3{-}B1{-}B2$.

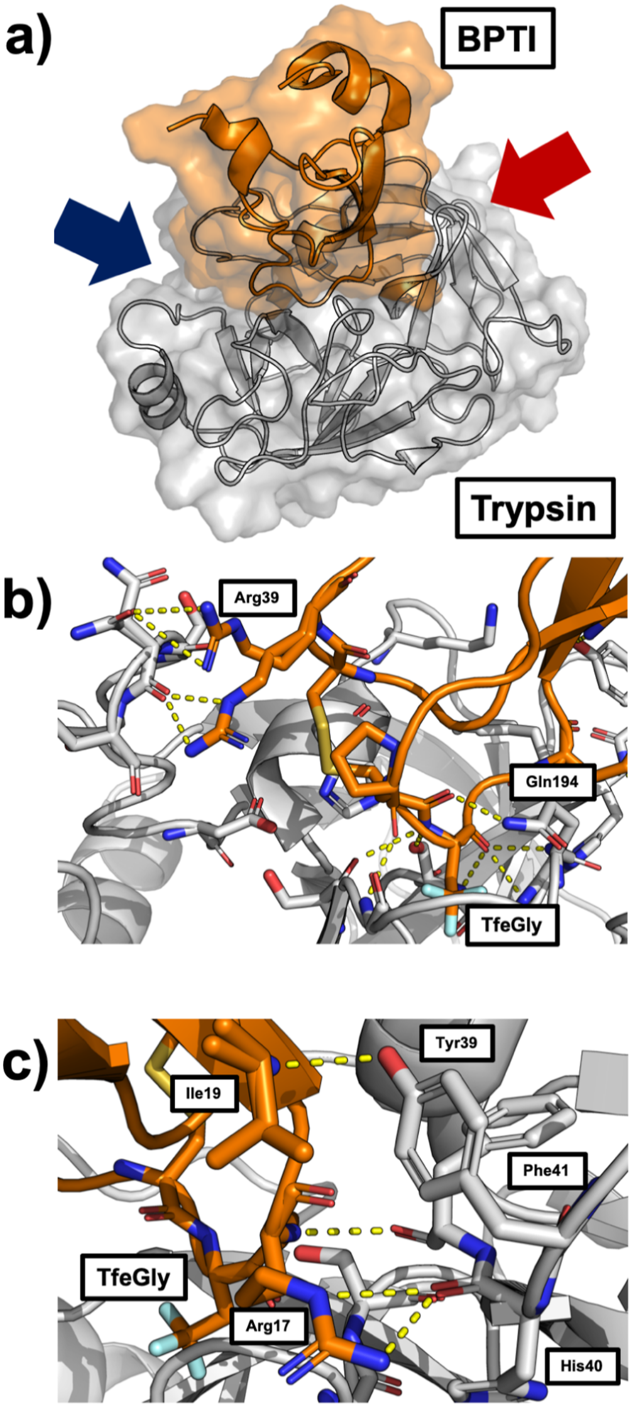


Figure 2. (a) TfeGly-BPTI–trypsin complex with surface and cartoon representation (pdb code: 4Y11). (b) Protein–protein interface of the complex seen from the perspective of the blue arrow. The interactions around the S1 pocket are to the bottom right and the interactions of Arg39 are on the top left. (c) Protein–protein interface of the complex seen from the red arrow. The interactions of Arg17 (P2′) and Ile19 (P4′) are shown.

In an initial attempt to achieve a free-energy surface of the binding and unbinding process of (Abu and TfeGly)-BPTI

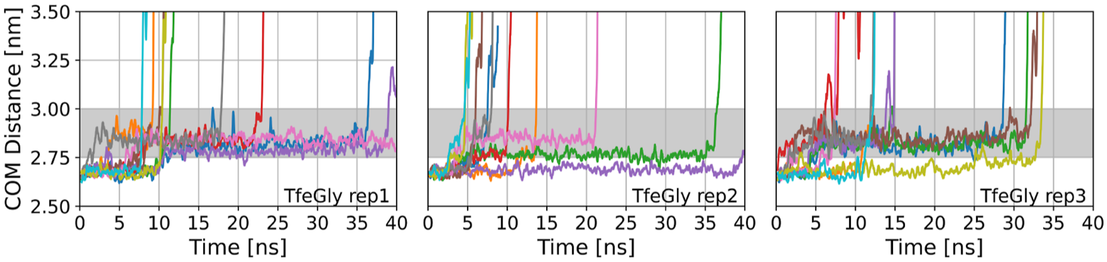


Figure 3. RAMD dissociation time series of TfeGly-BPTI. COM = center of mass. Gray area shows the center of mass of the prebound state. Left panel: replica 1, middle panel: replica 2, and right panel: replica 3. Ten RAMD runs per replica.

from trypsin, we performed restrained metadynamics simulations, where we chose the center-of-mass distance between (Abu and TfeGly)-BPTI and trypsin as the main collective variable and used harmonic restraints to restrain the other collective variables to the values of the fully bound complex, which we extracted from the X-ray crystal structure of the TfeGly-BPTI—trypsin complex (pdb code: 4Y11). Our efforts did not yield sufficient sampling of the binding and unbinding process, as after a single unbinding event, the ligand did not find back into the fully bound complex throughout 500 ns metadynamics simulations, although their orientation and movement around the receptor were restrained (see Figure S1). We conclude that the preferred binding and unbinding pathway has to be more complex than a simple movement on a straight line defined only by the center-of-mass distance and likely contains intermediate states.

Random Acceleration Molecular Dynamics. To study the unbinding pathways of (Abu, MfeGly, DfeGly, and TfeGly)-BPTI and wildtype BPTI from trypsin, we performed RAMD[23,24,45] simulations. RAMD is an enhanced sampling method that applies an additional biasing force to the center of mass of a ligand in an otherwise unbiased MD simulation.[23] If the unbinding process does not make progress despite the biasing force, the direction of this force is reoriented in a random direction at regular time intervals. The method was originally invented to discover unbinding pathways of buried protein ligands.[26]

For every four Abu-BPTI variants and wildtype BPTI, we generated three different starting structures and ran 10 simulations with a maximum length of 40 ns for each of these replicas. To achieve dissociation, we needed a force with a high magnitude of 3500 kJ/(mol nm), which is about an order of magnitude higher than for protein—small-molecule systems like benzamidine—trypsin.[23,24] This might be expected, as according to inhibition assays,[22] our systems have a binding affinity of −37 to −41 kJ/mol, while benzamidine binds to trypsin with a binding affinity of −22 to −26 kJ/mol.[15] Moreover, as the complex is held in place by many hydrogen bonds, it is likely that some, if not most, of them must be broken in a concerted way to achieve dissociation, which would result in a very steep free-energy barrier, requiring a strong force to drive the system out of the bound state. Possibly, proteins, in general, need a higher force constant to be dissociated efficiently compared to small molecules.

For the TfeGly-BPTI—trypsin complex, Figure 3 shows the time series of the center-of-mass distance (r) as a moving average with a moving window of 200 ps. The panels correspond to the three different starting structures, and we show the time series of the 10 simulations per starting structure in different colors. See the Supporting Information (Figures S2—S6) for the corresponding time series of DfeGly, MfeGly, Abu, and wildtype BPTI. The time series in Figure 3 first varies around the center-of-mass distance of the fully bound complex at around 2.65 nm. Then, they tend to transition to a state in which the center-of-mass distance fluctuates between 2.75 and 3.00 nm. The systems tend to remain in this state for tens of nanoseconds until they dissociate very rapidly. We call this intermediate state of the protein—protein complex the prebound state. We distinguish it from the fully bound state at 2.65 nm center-of-mass distance and from encounter states, which were investigated by Kahler et al.[20] and which would lie at center-of-mass distances around 3.00 nm.[22]

The prebound state occurs in dissociation trajectories of all four complexes of trypsin with (Abu, MfeGly, DfeGly, and TfeGly)-BPTI as well as in the dissociation trajectories of wildtype BPTI. While in some of the simulations, dissociation occurs without visiting the prebound state, we observe that in more than half of the trajectories for all BPTI variants, the moving average of the center-of-mass distance remains at least 1 ns between 2.75 and 3.00 nm; i.e., the prebound state is visited. Some trajectories did not dissociate after 40 ns of RAMD simulation, with some simulations ending in the prebound state and others ending in the fully bound state (see Supporting Information Table S1). The stability of the prebound state is remarkable, since throughout the RAMD simulations, a strong biasing force designed to dissociate the protein—protein complex acts on the center of mass of the BPTI variant.

Once the system leaves the prebound state toward larger center-of-mass distances, the protein—protein complex rapidly dissociates. That is, we do not observe encounter complexes around or above a 3.00 nm center-of-mass distance for any of the BPTI variants in our RAMD simulations. Encounter complexes are typically only weakly bound, and we assume that because of the strong biasing force, encounter complexes rapidly dissociated in the RAMD simulations.

Inspecting the RAMD trajectories more closely, we find that the prebound state is characterized not only by an increase in center-of-mass distance r but also by a significant shift of the system in the positional collective variables Θ_p and Φ_p compared to the fully bound state (see Figure S7). Figure 4 shows that the positional variables Θ_p or Φ_p change along with the center-of-mass distance r when transitioning from the fully bound state to the prebound state. In the orientational variables, θ_o, ϕ_o, and ψ_o, we do not find such a correlation, except for a slight shift in the θ_o angle (see Figure S8).

The stability in the presence of the biasing force and the systematic change in the positional variables indicate that the

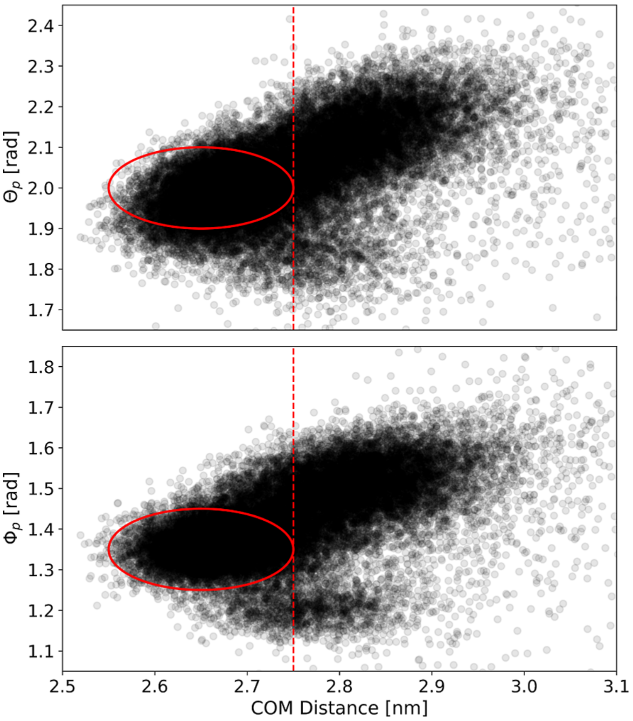


Figure 4. Scatter plots of the center-of-mass distance and Θ_p (top) and Φ_p (bottom) for all combined RAMD simulations. The red circle marks the position of the fully bound state.

prebound state might be a chemically relevant state, which is stabilized by different interactions than the fully bound state and separated by a free-energy barrier from the fully bound state. As the fully bound state is held in place tightly by a number of hydrogen bond-like interactions, it is likely that some of these interactions must be broken so that the prebound state can be reached. In the fully bound state of the complex between trypsin and (Abu, MfeGly, DfeGly, and TfeGly)-BPTI, there are 12 hydrogen bond-like contacts that can be found by visually inspecting the crystal structures which are shown in Figure 2. We calculated the frequency with which these interactions are formed as a function of the center-of-mass distance and present the histograms in Figure 5. The criterion for an interaction to be in place was a heavy atom distance of less than 0.35 nm. Note that the histograms were generated from RAMD simulations, i.e., nonequilibrium simulations, and therefore do not represent equilibrium distributions.

One of the 12 interactions is only rarely populated in the fully bound state and not populated at all in the prebound state: the hydrogen bond between the side chain hydroxyl oxygens of serine 197 (Ser197-OG) in trypsin and the amide hydrogen in the backbone of the P1 residue in the BPTI variants (X15-N), shown in orange in Figure 5 a. But since the overall change is small, this interaction is not suited to further define the prebound state. Also, in Figure 5 a, we show the histograms of eight further interactions, which are present in the fully bound state as well as in the prebound state. Their population decreases with increasing center-of-mass distance, but since the change is gradual and there is still a significant population in the interval 2.75 nm < r < 3.00 nm, it is not plausible that this change in population constitutes a clear free-energy barrier between the fully bound state and the prebound state.

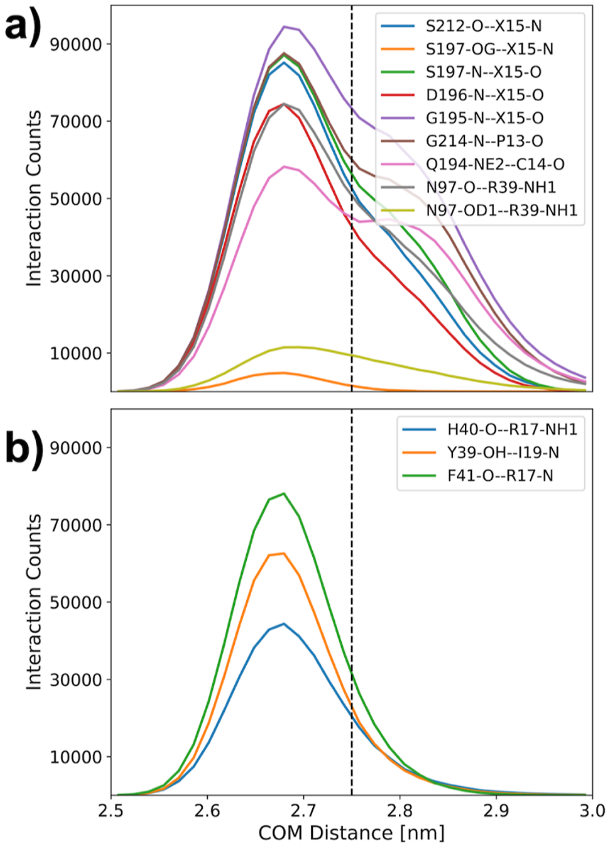


Figure 5. Interaction histogram along the center-of-mass distance in RAMD simulations of the Abu-BPTI variants and trypsin. The criterion for an interaction to be in place was that the involved heavy atoms were separated by a distance of less than 0.35 nm. The histograms were generated from the biased (nonequilibrium) RAMD simulations and therefore do not represent a Boltzmann distribution.

Figure 5 b shows the histogram of three interactions which are highly populated in the fully bound state but rarely populated in the interval 2.75 nm < r < 3.00 nm. These are the backbone—backbone interaction between Phe41 and Arg17, the interaction of the backbone of His40 with the side chain of Arg17, and the interaction between the side chain of Tyr39-OH and the backbone of Ile19 (compare Figure 2c). The breaking of these three interactions likely contributes to the free-energy barrier between the fully bound state and the prebound state.

Considering that we used a very high random force of 3500 kJ/(mol nm), we note that it is remarkable that the systems remain in the prebound state for a substantial amount of simulation time, despite the strong bias force introduced to the simulation. To retrospectively test the limits of this method, we ran sets of simulations with the TfeGly-BPTI variant, where we increased the magnitude of the random force to even higher values up to 5500 kJ/(mol nm). The trajectories are shown in Figure S9. We still observe the dissociating system to briefly stay in the region of r typical for the prebound state for some trajectories with a random force of 5000 kJ/(mol nm) but not with 5500 kJ/(mol nm). Hence, we conclude that a random force of 5000 kJ/(mol nm) is the limit to observe the prebound state for this system.

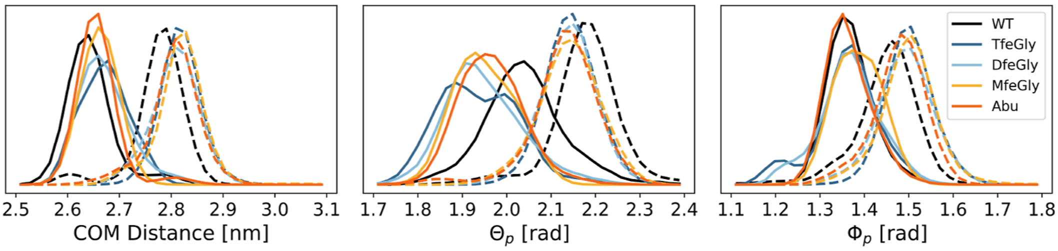


Figure 6. Position of the BPTI variants in the unbiased simulations of the fully bound state (solid lines) and the prebound state (dashed lines) as described by the center-of-mass distance r (left), Θ_p (center), and Φ_p (right). WT = wildtype.

Unbiased Simulation of the Prebound and Fully Bound State. To further characterize the difference between the fully bound state and the prebound state, we ran 20 unbiased simulations of 50 ns each (i.e., 1 μs total simulation time) of the fully bound state and prebound state in all of the four complexes between trypsin and (Abu, MfeGly, DfeGly, and TfeGly)-BPTI and wildtype BPTI. The starting structures for the fully bound state were generated from the crystal structure, and the starting structures for the simulations of the prebound state were generated from snapshots of the system in the prebound state from the RAMD simulations.

The time series of the center-of-mass distance r for all of the unbiased MD simulations can be found in the Supporting Information (Figure S10 for the Abu-BPTI variants and Figure S11 for wildtype BPTI). With very few exceptions, the systems remained in their starting state throughout the whole simulation time. This indicates that both fully bound state and the prebound state are stable on the time scale of 50 ns.

Figure 6 compares the equilibrium distributions of the positional variables, r, Θ_p, and Φ_p, of the fully bound state and prebound state. All BPTI variants have similar distributions (different colors in Figure 6), with the exception of the distributions of Θ_p of wildtype BPTI, which is shifted toward higher values, compared to the Abu-BPTI variants. However, the distribution differs significantly between the fully bound state and the prebound state (solid vs dashed lines in Figure 6). In the fully bound state, the systems adopt an average center-of-mass distance of 2.65 nm with a standard deviation of 0.04 nm, while in the prebound state, the center-of-mass distance r amounts to a mean of 2.85 nm with a standard deviation of 0.05 nm. Likewise, the coordinates Θ_p and Φ_p shift to larger values in the prebound state. In all three positional coordinates, there is little overlap between the distributions of the fully bound state and the prebound state, confirming that the positions that the BPTI variants can occupy in these two states are distinct.

The distributions of the orientational variables, θ_o, ϕ_o, and ψ_o, are included in the Supporting Information (Figure S12). In each of the three variables, we observe a systematic shift from the distributions of the fully bound state and to those of the prebound state, which is most pronounced for θ_o. However, the overlap between the fully bound state distributions and the prebound state distributions is larger than for the positional variables. This indicates that the BPTI variant does not gain (much) orientational freedom when transitioning from the fully bound state to the prebound state. We provide example snapshots from the unbiased simulations of the fully bound state and the prebound state for all of the four BPTI variants in the Supporting Information.

Figure 7 shows the relative population of all of the hydrogen bonds between the two proteins with at least 0.1 relative

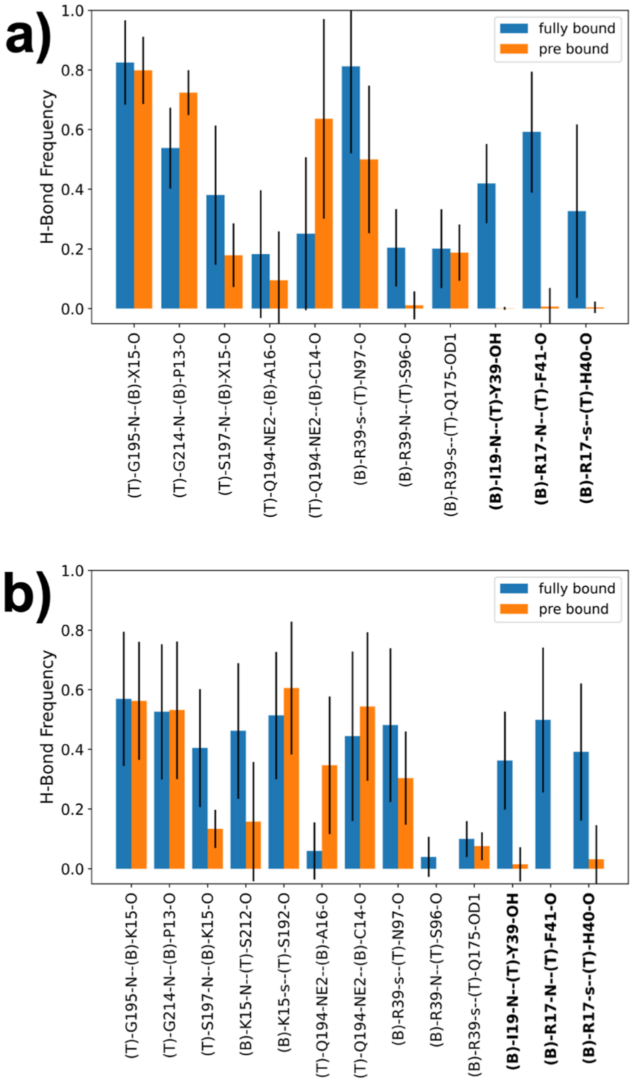


Figure 7. (a) Hydrogen bond frequencies of all combined unbiased simulations of the fully bound state and the prebound state. Residue name X = Abu, MfeGly, DfeGly, or TfeGly, T = trypsin, B = (Abu, MfeGly, DfeGly, or TfeGly)-BPTI. O and N = heteroatoms in the backbone; OD1, OH, and NE2 = heteroatoms in side chains. Hydrogen bonds are denoted as donor−acceptor. The side chain of arginine residues is denoted as "s", which means a hydrogen bond with any of the donors in the guanidine moiety. (b) Hydrogen bond frequencies of the unbiased simulations of the fully bound state and the prebound state with wildtype BPTI. The labels follow the same scheme as above. The side chain of lysine is also denoted as "s".

population. For this analysis, we merged the trajectories of the fully bound state of all four Abu-BPTI variants, and we merged the trajectories of the prebound state of all four Abu-BPTI variants (Figure 7a). At this point, this is justified, because the 12 interactions do not involve the side chain of the P1 residue and because we did not observe any significant difference in the positional and orientational variables across the four systems (Figure S13). We employed the Wernet−Nilsson criterion[41] in the MDTraj implementation to identify hydrogen bonds between trypsin and the BPTI variants.

In the simulations of the fully bound state, we observe 11 hydrogen bonds with a relative population >0.1. These are five hydrogen bonds, which are located around the S1 pocket, three hydrogen bonds of Arg39 (compare Figure 2b), and three hydrogen bonds of Arg17 and Ile19 (compare Figure 2c). We find the hydrogen bonds around the S1 pocket to also be present in the prebound state. The backbone−backbone interaction between Gly195 and the P1 residue has the same frequency in the prebound state as in the fully bound state, while the frequency of the neighboring interaction between the side chain of Gln194 and the backbone of Ala16 is lower in the prebound state, albeit with high statistical uncertainty. The frequency of two hydrogen bonds close to the S1 pocket, namely, between the side chain of Gln194 and the backbone of Cys14 and the backbone−backbone interaction between Gly214 and Pro13 is higher in the prebound state, but again with high statistical uncertainty.

In the simulations of the complex with wildtype BPTI, we find the same hydrogen bonds as for the Abu-BPTI variants (Figure 7b). Additionally, we observe a frequent hydrogen bond between the side chain of Lys15 and Ser192, which is a well-known key interaction between trypsin and wildtype BPTI at the bottom of the S1 pocket.[6] As for the Abu-BPTI variants, the hydrogen bonds around the S1 pocket are in place in the fully bound state and prebound state. Interestingly, this also applies to the interaction between Lys15 and Ser192 at the bottom of the S1 pocket, meaning that in the prebound state, this key interaction of wildtype BPTI is still in place.

Three hydrogen bonds are frequently populated in the fully bound state but are virtually nonexistent in the prebound state, making these three broken hydrogen bonds a defining property of the prebound state. These are the same three hydrogen bonds that already showed a loss of population when transitioning from the fully bound state to the prebound state in the RAMD simulations (Figure 5 b). In the fully bound state, two of the hydrogen bonds are formed between Arg17 in the BPTI variants and the backbone in trypsin, one between the side chain of Arg17 and the backbone of His40, and the other between the backbone of Arg17 and the backbone of Phe41. The third hydrogen bond is formed between the amide hydrogen of Ile19 in BPTI and the side chain of Tyr39 in trypsin. These hydrogen bonds are shown for the fully bound state in Figure 9 a. Figure 9 b shows the same region in the prebound state. Side chains of Arg17 and Tyr39 have been reoriented, and the three hydrogen bonds cannot be formed in the prebound state.

The hydrogen bonds of Arg39 in the BPTI variants with the backbone of trypsin are also more frequently populated in the fully bound state than in the prebound state (Figure 7). However, the drop in population is less pronounced than that for the three hydrogen bonds discussed above. For wildtype BPTI, the hydrogen bonds are less populated in the fully bound state and also in the prebound state, compared to the Abu-BPTI variants.

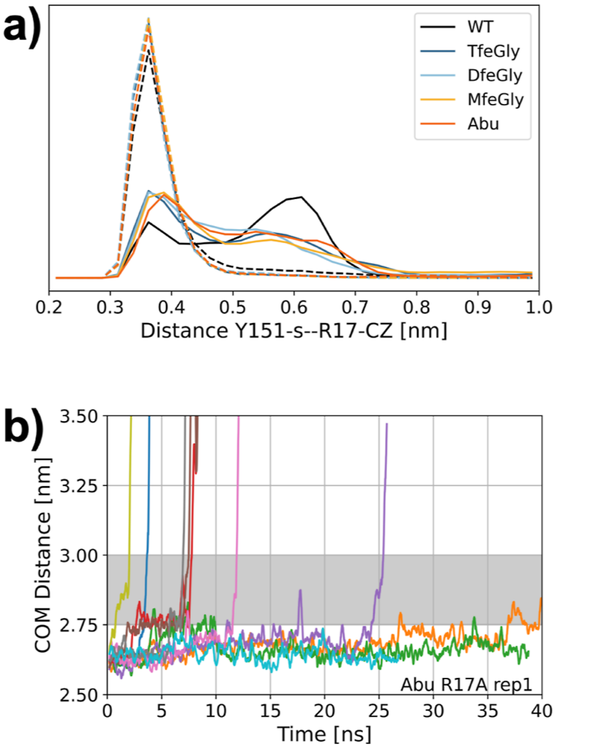


Figure 8. (a) Distance between the centroid of the Tyr151 aromatic system (Y151-s) and the carbon of the guanidine moiety of Arg17 (R17-CZ) in the unbiased simulations of the fully bound state (solid line) and the prebound state (dashed lines). (b) RAMD dissociation trajectories of one replica of the R17A mutant of the Abu-BPTI variant.

The analysis so far shows that the dissociation of the protein−protein complex between trypsin and (Abu, MfeGly, DfeGly, and TfeGly)-BPTI proceeds via a prebound state which is stable at least on the time scale of 50 ns. The prebound state is characterized by a shift in the positional variables of BPTI and, to a lesser extent, by a shift in the orientational variables. To form the prebound state, three hydrogen bonds that are highly populated in the fully bound state are broken.

Stabilizing Interactions in the Prebound State. The analysis so far does not show why the breaking of the three hydrogen bonds results in a stable state that does not immediately revert back to the fully bound state. Figure 8 a suggests that one of the factors contributing to the stability of the prebound state could be a cation−pi interaction that is formed by the now free Arg17 side chain of BPTI with the aromatic system of Tyr151 of trypsin.

We measured the distance distribution between the carbon atom of the guanidine moiety of Arg17 (Arg17-CZ) and the centroid of the aromatic ring of Tyr151 (Figure 8a). While in the fully bound state, the distance can take a range of values between 0.3 and 0.8 nm, and the distance in all simulation snapshots of the prebound state remains well below 0.45 nm. The broad distribution of the Tyr151-s-Arg17-CZ distance in the fully bound state shows that no specific bond is observed between the two residues. By contrast, the narrow distribution at low distances in the prebound state suggests the existence of a cation−pi interaction.

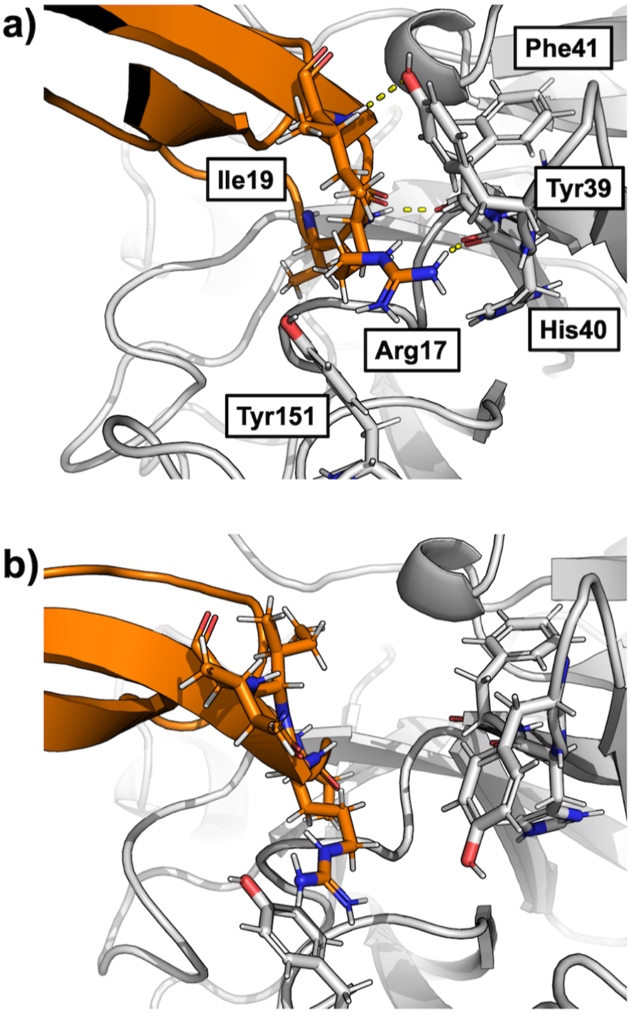


Figure 9. Example snapshots from the unbiased simulations of the (a) fully bound and (b) prebound state. The figure shows a similar region as Figure 2c.

An interaction with the aromatic system of Tyr151 has so far not been described for the BPTI−trypsin complex. It is however present in X-ray crystallography structures of other trypsin inhibitors like bdellastasin (pdb code: 1C9T), where a cationic lysine side chain at P2′ position forms a cation−pi interaction with Tyr151,[46] or microviridin (pdb code: 4KTU), where a tyrosine at P2′ position forms a t-shaped pi−pi interaction with Tyr151.[47]

To verify whether an interaction of the Arg17 side chain is indeed essential for the stabilization of the prebound state, we repeated the RAMD simulations for one replica of the Abu-BPTI−trypsin complex, where we mutated Arg17 in Abu-BPTI variants to alanine (Figure 8b). The dissociation happens roughly on the same time scale as for the nonmutated Abu-BPTI variants. However, the prebound state is traversed rapidly on all ten of the unbinding trajectories. This supports the hypothesis that Arg17 is indeed essential for the stabilization of the prebound state.

Additionally, we analyzed the SASA of the protein−protein interface amino acid residues for the fully bound state and prebound state (see Figures S14−S17). Most of the residues in the interface do not show significant differences in their SASA in the fully bound state and prebound state. A notable exception is that the SASA of residues Arg17, Ile18, and Ile19 of the BPTI variants, as well as of Tyr39 and Phe41 of trypsin, increases significantly in the prebound state. The SASA of Tyr151 decreases in the prebound state. These changes reflect the difference in binding between the fully bound state and the prebound state. This implies that the hydration shell of the fully bound state and the prebound state is similar, except for the region around Arg17. Thus, the prebound state is likely not only stabilized by the interaction of Arg17 and Tyr151 but other effects, such as hydration, play a role as well.

Influence of the Fluorine Substituents. As a last step, we were interested in how the fluorine substituents in the BPTI variants influence the stability of the prebound state. To this end, we performed umbrella sampling between the fully bound state and the prebound state, where we used the newly identified interaction between the backbone amide of Arg17 in BPTI and the backbone oxygen of Phe41 in trypsin. We selected this reaction coordinate combined with a slow growth approach for the starting structures of the umbrella windows to ensure an accurate transition path between the fully bound state and the newly discovered prebound state. We find this approach to model the transition more accurately than picking

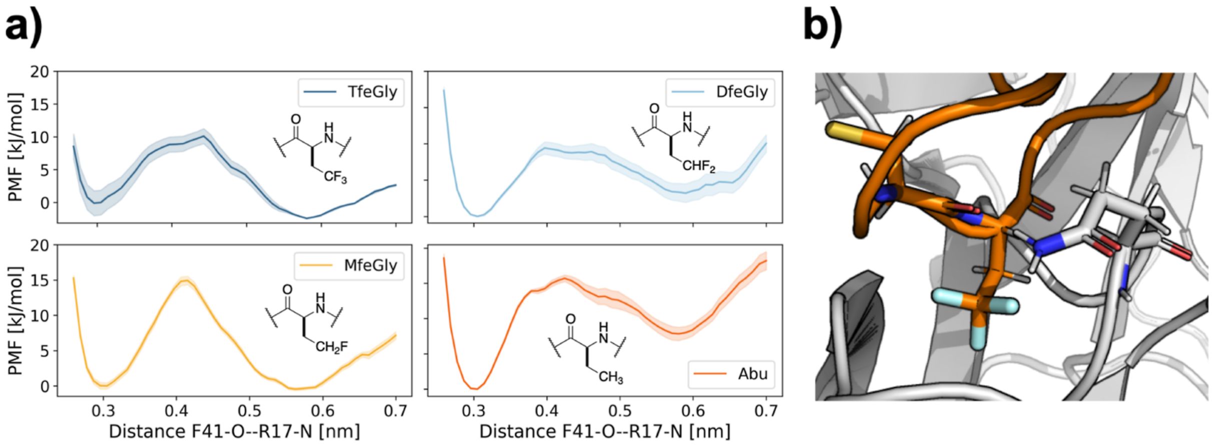


Figure 10. (a) Potential of the mean force profile of the fully bound state and prebound state from umbrella sampling over the distance between the carbonyl oxygen of Phe41 (F41-O) and the backbone nitrogen of Arg17 (R17-N). (b) Interaction between Gln194 and Cys14 in the direct proximity of the TfeGly side chain.

starting structures from our RAMD simulations and using the center-of-mass distance as the reaction coordinate, as attempts to model the transition path using the string method with swarms-of-trajectories[48,49] did not capture the transition state between the two states (see Figure S18).

Figure 10a shows the resulting potential of the mean force along this reaction coordinate derived from the newly identified interaction and the slow-growth approach. In all four systems, the potential of the mean force exhibits two minima. The minimum around the 0.3 nm corresponds to the fully bound state, whereas the minimum around 0.6 nm corresponds to the prebound state. In a previous study,[22] we investigated the interactions in the fully bound state and found no significant differences between the four BPTI variants. For the prebound state, we find that the barrier height between the two states for the unfluorinated Abu and the monofluorinated MfeGly is about 15 kJ/mol, while for the higher fluorinated DfeGly and TfeGly, it is only about 10 kJ/mol. The minimum of the prebound state for Abu lies well above the minimum for the fully bound state. By contrast, in the TfeGly-BPTI complex, the prebound state is stabilized relative to the bound state. The partially fluorinated complexes lie in between. Thus, there is a clear effect of the fluorination on the energetic landscape between the fully bound state and the prebound state.

To find a possible mechanism for this stabilization, we revisited our hydrogen bond analysis, for which we reported the aggregate statistics for all four Abu-BPTI variants in Figure 7a. We reanalyzed for each BPTI variant and found that for most interactions, the hydrogen bond populations did not differ significantly across the BPTI variants. A notable exception is the hydrogen bond between the side chain of Gln194 in trypsin and the backbone oxygen of Cys14 in the BPTI variants. This interaction can be observed in the fully bound state and also in the prebound state, but it is more frequent in the prebound state of the fluorinated variants (MfeGly, DfeGly, and TfeGly)-BPTI, while it is equally populated in the states of Abu-BPTI (see Figure S13). Gln194 and Cys14 are close to the side chain of the P1 residue (Figure 10b). When the hydrogen bond is formed, the side chain of Gln194 is in fact so close to the fluorine atoms that it appears plausible that the fluorine atoms with their negative partial charge help stabilize the NH2 end of the Gln194 side chain by providing an extra binding partner in addition to the backbone oxygen of Cys14.

■ CONCLUSIONS

We applied several MD simulation techniques to characterize the unbinding pathway of (Abu, MfeGly, DfeGly, and TfeGly)-BPTI and wildtype BPTI from trypsin. The BPTI variants likely dissociate via a curved pathway in a coordinate space that describes the relative position and orientation of the two proteins, as evidenced by restrained metadynamics simulations in which the two proteins do not rebind once they are dissociated. Using RAMD simulations[23−26] to accommodate this curved unbinding pathway, we identified a new metastable state on the unbinding pathway.

This prebound state is present on the unbinding pathway in all four variants of the BPTI−trypsin complex and also in the wildtype-BPTI−trypsin complex. In unbiased simulations, it is stable for at least 50 ns. Since in an aggregated simulation time of 1 μs per BPTI variant, the prebound state only very rarely reverted to the fully bound state, we suspect that the average lifetime of the prebound state is in fact in the order of several 100 ns.

The prebound state is clearly distinct from the fully bound state in the positional coordinates from the fully bound state. The center-of-mass distance between the two proteins in the complexes of the Abu-BPTI variants is increased by about 0.2 nm (from 2.65 to 2.85 nm) and the BPTI variants rotate by about 10° (0.2 rad) in Θ_p and by 10° (0.2 rad) around the dihedral angle Φ_p. There is little overlap between the distributions of the prebound state and fully bound state in these coordinates. We also observe a systematic shift in the orientational coordinates but less pronounced. The distribution of fully bound state and prebound state for wildtype BPTI is very similar to those of the Abu-BPTI variants, with the exception of Θ_p, which is slightly shifted toward higher values.

The interaction pattern between the two proteins changes when transitioning from the fully bound state to the prebound state. These changes particularly involve Arg17 (P2′ residue) and Arg39 in the BPTI variants. In the prebound state, the hydrogen bond of the Arg17 side chain to the backbone of trypsin is broken, but it is replaced by a cation−pi interaction between the guanidine moiety and a nearby trypsin tyrosine residue. Two further hydrogen bonds in the vicinity are also broken in this process, and the hydrogen bond between the side chain of Arg39 and the trypsin backbone becomes less populated. When we replaced Arg17 by an alanine residue in RAMD simulations, the protein−protein complex dissociated without spending time in the prebound state, which demonstrates that Arg17 is essential for the stabilization of this state.

The prebound state is likely not only stabilized by the interaction of Arg17 and Tyr151 but also due to other effects, such as hydration. The SASA is increased for the residues close to Arg17 in the prebound state, which might imply a change in hydration. This aspect should be addressed in future research, e.g., by an analysis of the water molecules in the vicinity of Arg17 similar to our analysis of the water molecules in the S1 binding pocket.[22]

The structural rearrangements that stabilize the prebound state do not involve the P1 residue in BPTI or the negatively charged residues at the bottom of the S1 pocket of trypsin. The same structural rearrangements can also be found for wildtype BPTI, which means that the unbinding of the Abu-BPTI variants proceeds via the same prebound state.

In potentials of mean force (PMF), we find that fluorination of Abu lowers the free-energy barrier between the fully bound and the prebound state and also lowers the free-energy minimum of the prebound state. However, quantitative interpretation of these one-dimensional PMFs is difficult. In particular, we suspect that the PMF might overstabilize the prebound state, as in some of the potentials, the prebound state minimum is as low as the fully bound minimum. Nonetheless, the fluorine substituents on the P1 residue clearly have an influence on the stability of the prebound state. A possible, yet speculative, explanation is that the hydrogen bond between the side chain of Gln194 and Cys14 is stabilized by fluorine substituents in the direct proximity of the side chain NH2 group of Gln194. Fluorine is known to have a wide range of possible effects on protein-inhibitor interactions, e.g., through hydrogen bonds,[50] desolvation,[33,34] or entropy,[51] whose elucidation often requires in-depth computational studies. The differences in barrier height and stability of the prebound state in the fluorinated variants of BPTI are likely not only due to a single stabilizing interaction, like the Gln194−Cys14 hydrogen bond, but instead due to a combination of enthalpic and entropic effects.

Because of the large magnitude of the biasing force in the RAMD simulations, which is necessary to dissociate the protein−protein complexes, we did not observe encounter complexes in our simulations. We expect that encounter states do play a role in the binding and unbinding process of the (Abu, MfeGly, DfeGly, and TfeGly)-BPTI−trypsin complexes.[20] However, the transition between the prebound state and these weakly bound encounter states should be characterized with other methods like weighted ensemble MD[52] and molecular rotational grids.[53]

While this manuscript was in review, D'Arrigo et al.[54] published a preprint, in which they dissociate a series of protein−protein systems, including wildtype BPTI and some of its mutants from trypsin, using RAMD with a smaller force. In the dissociation trajectories, they find that the contacts of Arg17 are cleaved first, which aligns well with our results. The authors find additional states along the dissociation trajectory, which may correspond to the encounter states mentioned above. These additional states, together with works of Kahler et al.,[20] are excellent starting points for the characterization of encounter states that we suggest above.

The existence and structure of the prebound state invite speculation on the inhibitory mechanism of BPTI and its variants. After formation of the initial Michaelis complex of a substrate with trypsin, the hydrolysis of the peptide bond proceeds via two steps. First the peptide bond is broken, and the N-terminal part of the substrate (i.e., all residues from the N-terminus up to and including P1) forms a covalently bound acyl-enzyme intermediate. The C-terminal part of the substrate (i.e., all residues from P1′ to the C-terminus) remains noncovalently bound and needs to dissociate before, in a second step, and the acyl-enzyme intermediate can be hydrolyzed. Radisky and Koshland showed that for a closely related serine protease complex, the initial formation of the acyl-enzyme intermediate is fast, but the release of the C-terminal part of the substrate is slow,[9] such that the reaction reverts back to the intact peptide bond. This "clogged gutter" mechanism is further supported by a high-resolution structure of a cleaved BPTI variant with trypsin.[13] Our analysis showed that the interface between trypsin and the BPTI variants is stabilized by hydrogen bonds primarily from the C-terminal part of the BPTI variants (Figure 7). Specifically, Arg17 which stabilizes the prebound state via a cation−pi interaction belongs to the C-terminal part. Thus, assuming that the clogged gutter mechanism applies to the BPTI−trypsin complex, these interactions likely contribute to stabilizing the C-terminal part of the protein complex.

Finally, our study shows that, to understand the stability of the wildtype-BPTI−trypsin complex or the (Abu, MfeGly, DfeGly, and TfeGly)-BPTI−trypsin complex, one needs to consider two states, the fully bound state and the prebound state, which likely are in dynamic equilibrium. By mimicking the interactions in the prebound state, one may open up additional ways to design serine-protease inhibitors.

■ ASSOCIATED CONTENT

Data Availability Statement

RAMD and unbiased MD simulations were performed with openly available GROMACS (https://www.gromacs.org), GROMACS-RAMD (https://github.com/HITS-MCM/gromacs-ramd/tree/release-2022), and Plumed (https://www.plumed.org) software. Starting structure preparation was done with the openly available pdbfixer (https://github.com/openmm/pdbfixer). Interaction distances and hydrogen bonds were analyzed using mdtraj (https://www.mdtraj.org), which is openly available. Force-field parameters for the fluorinated amino acids and Abu have been released in a recent publication,[22] and all other force-field parameters were taken from the openly available Amber14SB force field (https://ambermd.org). Protein structure starting files, MD parameter files, and analysis scripts for all simulations are published on GitHub (https://github.com/leonwehrhan/Trypsin_BPTI_-Prebound_RAMD_2024)

SI Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.4c00338.

> Center-of-mass distance time series for restrained metadynamics simulations; center-of-mass distance time series for RAMD simulations of TfeGly-, DfeGly-, MfeGly-, and Abu-BPTI and wildtype BPTI; positional and orientational collective variable histograms for the RAMD simulations; center-of-mass distance time series for unbiased MD simulations; orientational collective variable histograms for the unbiased MD simulations; hydrogen bond frequencies for TfeGly-, DfeGly-, MfeGly-, and Abu-BPTI and wildtype-BPTI simulations; SASA for protein−protein interface residues of TfeGly-, DfeGly-, MfeGly-, and Abu-BPTI; RAMD dissociation trajectories with higher forces (PDF)
>
> Example snapshots of the fully bound state and the prebound state of the four variants of the (TfeGly, DFeGly, MfeGly, and Abu)-BPTI−trypsin complex (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

Bettina G. Keller − *Department of Biology, Chemistry, and Pharmacy, Freie Universität Berlin, Berlin 14195, Germany;* orcid.org/0000-0002-7051-0888; Email: bettina.keller@fu-berlin.de

Author

Leon Wehrhan − *Department of Biology, Chemistry, and Pharmacy, Freie Universität Berlin, Berlin 14195, Germany;* orcid.org/0000-0002-9894-8013

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.4c00338

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Ariane Nunes-Alves for fruitful discussion on RAMD simulations. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): Project-ID 387284271-SFB 1349. The authors acknowledge access to high-performance computers via the Zentraleinrichtung FUB-IT of Freie Universität Berlin.

■ REFERENCES

(1) Turk, B. Targeting Proteases: Successes, Failures and Future Prospects. *Nat. Rev. Drug Discovery* **2006**, *5*, 785−799.

(2) Drag, M.; Salvesen, G. S. Emerging Principles in Protease-Based Drug Discovery. *Nat. Rev. Drug Discovery* **2010**, *9*, 690−701.

(3) Batt, A. R.; St. Germain, C. P.; Gokey, T.; Guliaev, A. B.; Baird, T., Jr Engineering Trypsin for Inhibitor Resistance. *Protein Sci.* **2015**, *24*, 1463−1474.

(4) Page, M. J.; Di Cera, E. Evolution of Peptidase Diversity. *J. Biol. Chem.* **2008**, *283*, 30010−30014.

(5) Schechter, I. Mapping of the Active Site of Proteases in the 1960s and Rational Design of Inhibitors/Drugs in the 1990s. *Curr. Protein Pept. Sci.* **2005**, *6*, 501−512.

(6) Kawamura, K.; Yamada, T.; Kurihara, K.; Tamada, T.; Kuroki, R.; Tanaka, I.; Takahashi, H.; Niimura, N. X-Ray and Neutron Protein Crystallographic Analysis of the Trypsin-BPTI Complex. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2011**, *67*, 140−148.

(7) Marquart, M.; Walter, J.; Deisenhofer, J.; Bode, W.; Huber, R. The Geometry of the Reactive Site and of the Peptide Groups in Trypsin, Trypsinogen and Its Complexes with Inhibitors. *Acta Crystallogr., Sect. B: Struct. Sci.* **1983**, *39*, 480−490.

(8) Peräkylä, M.; Kollman, P. A. Why Does Trypsin Cleave BPTI so Slowly? *J. Am. Chem. Soc.* **2000**, *122*, 3436−3444.

(9) Radisky, E. S.; Koshland, D. E. A Clogged Gutter Mechanism for Protease Inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 10316−10321.

(10) Vincent, J. P.; Lazdunski, M. Trypsin-Pancreatic Trypsin Inhibitor Association. Dynamics of the Interaction and Role of Disulfide Bridges. *Biochemistry* **1972**, *11*, 2967−2977.

(11) Warshel, A.; Russell, S. Theoretical Correlation of Structure and Energetics in the Catalytic Reaction of Trypsin. *J. Am. Chem. Soc.* **1986**, *108*, 6569−6579.

(12) Farady, C. J.; Craik, C. S. Mechanisms of Macromolecular Protease Inhibitors. *ChemBioChem* **2010**, *11*, 2341−2346.

(13) Zakharova, E.; Horvath, M. P.; Goldenberg, D. P. Structure of a Serine Protease Poised to Resynthesize a Peptide Bond. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 11034−11039.

(14) Castro, M. J. M.; Anderson, S. Alanine Point-Mutations in the Reactive Region of Bovine Pancreatic Trypsin Inhibitor: Effects on the Kinetics and Thermodynamics of Binding to β-Trypsin and α-Chymotrypsin. *Biochemistry* **1996**, *35*, 11435−11446.

(15) Buch, I.; Giorgino, T.; De Fabritiis, G. Complete Reconstruction of an Enzyme-Inhibitor Binding Process by Molecular Dynamics Simulations. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 10184−10189.

(16) Siebenmorgen, T.; Zacharias, M. Computational Prediction of Protein-Protein Binding Affinities. *WIREs Comput. Mol. Sci.* **2020**, *10*, No. e1448.

(17) Wu, Z.; Liao, Q.; Liu, B. A Comprehensive Review and Evaluation of Computational Methods for Identifying Protein Complexes from Protein-Protein Interaction Networks. *Briefings Bioinf.* **2020**, *21*, 1531−1548.

(18) Woo, H.-J.; Roux, B. Calculation of Absolute Protein-Ligand Binding Free Energy from Computer Simulations. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6825−6830.

(19) Gumbart, J. C.; Roux, B.; Chipot, C. Efficient Determination of Protein-Protein Standard Binding Free Energies from First Principles. *J. Chem. Theory Comput.* **2013**, *9*, 3789−3798.

(20) Kahler, U.; Kamenik, A. S.; Waibl, F.; Kraml, J.; Liedl, K. R. Protein-Protein Binding as a Two-Step Mechanism: Preselection of Encounter Poses during the Binding of BPTI and Trypsin. *Biophys. J.* **2020**, *119*, 652−666.

(21) Ye, S.; Loll, B.; Berger, A. A.; Mülow, U.; Alings, C.; Wahl, M. C.; Koksch, B. Fluorine Teams up with Water to Restore Inhibitor Activity to Mutant BPTI. *Chem. Sci.* **2015**, *6*, 5246−5254.

(22) Wehrhan, L.; Leppkes, J.; Dimos, N.; Loll, B.; Koksch, B.; Keller, B. G. Water Network in the Binding Pocket of Fluorinated BPTI-Trypsin ComplexesInsights from Simulation and Experiment. *J. Phys. Chem. B* **2022**, *126*, 9985−9999.

(23) Kokh, D. B.; Doser, B.; Richter, S.; Ormersbach, F.; Cheng, X.; Wade, R. C. A Workflow for Exploring Ligand Dissociation from a Macromolecule: Efficient Random Acceleration Molecular Dynamics Simulation and Interaction Fingerprint Analysis of Ligand Trajectories. *J. Chem. Phys.* **2020**, *153*, 125102.

(24) Kokh, D. B.; Amaral, M.; Bomke, J.; Grädler, U.; Musil, D.; Buchstaller, H.-P.; Dreyer, M. K.; Frech, M.; Lowinski, M.; Vallee, F.; Bianciotto, M.; Rak, A.; Wade, R. C. Estimation of Drug-Target Residence Times by -Random Acceleration Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2018**, *14*, 3859−3869.

(25) Nunes-Alves, A.; Kokh, D. B.; Wade, R. C. Recent Progress in Molecular Simulation Methods for Drug Binding Kinetics. *Curr. Opin. Struct. Biol.* **2020**, *64*, 126−133.

(26) Lüdemann, S. K.; Lounnas, V.; Wade, R. C. How Do Substrates Enter and Products Exit the Buried Active Site of Cytochrome P450cam? 1. Random Expulsion Molecular Dynamics Investigation of Ligand Access Channels and mechanisms. *J. Mol. Biol.* **2000**, *303*, 797−811.

(27) Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R. A.; Parrinello, M. PLUMED: A Portable Plugin for Free-Energy Calculations with Molecular Dynamics. *Comput. Phys. Commun.* **2009**, *180*, 1961−1972.

(28) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New Feathers for an Old Bird. *Comput. Phys. Commun.* **2014**, *185*, 604−613.

(29) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1−2*, 19−25.

(30) Páll, S.; Abraham, M. J.; Kutzner, C.; Hess, B.; Lindahl, E. Tackling Exascale Software Challenges in Molecular Dynamics Simulations with GROMACS. *Solving Software Challenges for Exascale* **2015**, *8759*, 3−27.

(31) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. GROMACS 4.5: A High-Throughput and Highly Parallel Open Source Molecular Simulation Toolkit. *Bioinformatics* **2013**, *29*, 845−854.

(32) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696−3713.

(33) Robalo, J. R.; Huhmann, S.; Koksch, B.; Vila Verde, A. The Multiple Origins of the Hydrophobicity of Fluorinated Apolar Amino Acids. *Chem* **2017**, *3*, 881−897.

(34) Robalo, J.; Vila Verde, A. Unexpected Trends in the Hydrophobicity of Fluorinated Amino Acids Reflect Competing Changes in Polarity and Conformation. *Phys. Chem. Chem. Phys.* **2019**, *21*, 2029−2038.

(35) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling through Velocity Rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.

(36) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.* **1981**, *52*, 7182−7190.

(37) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* **1997**, *18*, 1463−1472.

(38) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103*, 8577−8593.

(39) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926−935.

(40) McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **2015**, *109*, 1528−1532.

(41) Wernet, P.; Nordlund, D.; Bergmann, U.; Cavalleri, M.; Odelius, M.; Ogasawara, H.; Näslund, L. A.; Hirsch, T. K.; Ojamäe, L.; Glatzel, P.; Pettersson, L. G. M.; Nilsson, A. The Structure of the

First Coordination Shell in Liquid Water. *Science* **2004**, *304*, 995−999.

(42) Shrake, A.; Rupley, J. A. Environment and Exposure to Solvent of Protein Atoms. Lysozyme and Insulin. *J. Mol. Biol.* **1973**, *79*, 351−371.

(43) Bussi, G.; Tribello, G. A. Biomolecular Simulations: Methods and Protocols. In *Methods in Molecular Biology*; Bonomi, M., Camilloni, C., Eds.; Springer, 2019; pp 529−578.

(44) Tan, Z.; Gallicchio, E.; Lapelosa, M.; Levy, R. M. Theory of Binless Multi-State Free Energy Estimation with Applications to Protein-Ligand Binding. *J. Chem. Phys.* **2012**, *136*, 144102.

(45) Lüdemann, S. K.; Lounnas, V.; Wade, R. C.; Thornton, J. How Do Substrates Enter and Products Exit the Buried Active Site of Cytochrome P450cam? 2. Steered Molecular Dynamics and Adiabatic Mapping of Substrate pathways. *J. Mol. Biol.* **2000**, *303*, 813−830.

(46) Rester, U.; Bode, W.; Moser, M.; Parry, M. A. A.; Huber, R.; Auerswald, E. Structure of the complex of the antistasin-type inhibitor bdellastasin with trypsin and modelling of the bdellastasin-microplasmin system. *J. Mol. Biol.* **1999**, *293*, 93−106.

(47) Weiz, A. R.; Ishida, K.; Quitterer, F.; Meyer, S.; Kehr, J.-C.; Müller, K. M.; Groll, M.; Hertweck, C.; Dittmann, E. Harnessing the Evolvability of Tricyclic Microviridins To Dissect Protease-Inhibitor Interactions. *Angew. Chem., Int. Ed.* **2014**, *53*, 3735−3738.

(48) Roux, B. String Method with Swarms-of-Trajectories, Mean Drifts, Lag Time, and Committor. *J. Phys. Chem. A* **2021**, *125*, 7558−7571.

(49) Pan, A. C.; Sezer, D.; Roux, B. Finding Transition Pathways Using the String Method with Swarms of Trajectories. *J. Phys. Chem. B* **2008**, *112*, 3432−3440.

(50) Pietruś, W.; Kafel, R.; Bojarski, A. J.; Kurczab, R. Hydrogen Bonds with Fluorine in Ligand-Protein Complexes-the PDB Analysis and Energy Calculations. *Molecules* **2022**, *27*, 1005.

(51) Wallerstein, J.; Ekberg, V.; Ignjatović, M. M.; Kumar, R.; Caldararu, O.; Peterson, K.; Wernersson, S.; Brath, U.; Leffler, H.; Oksanen, E.; Logan, D. T.; Nilsson, U. J.; Ryde, U.; Akke, M. Entropy-Entropy Compensation between the Protein, Ligand, and Solvent Degrees of Freedom Fine-Tunes Affinity in Ligand Binding to Galectin-3C. *J. Am. Chem. Soc. Au* **2021**, *1*, 484−500.

(52) Zuckerman, D. M.; Chong, L. T. Weighted Ensemble Simulation: Review of Methodology, Applications, and Software. *Annu. Rev. Biophys.* **2017**, *46*, 43−57.

(53) Zupan, H.; Heinz, F.; Keller, B. G. Grid-Based State Space Exploration for Molecular Binding. *Can. J. Chem.* **2023**, *101*, 710−724.

(54) D'Arrigo, G.; Kokh, D. B.; Nunes-Alves, A.; Wade, R. C. Computational Screening of the Effects of Mutations on Protein-Protein off-Rates and Dissociation Mechanisms by RAMD. *BioRxiv* **2024**, *2024*..

4.4 Phosphotyrosine Mimetics Targeting PTP1B

Title: Pentafluorophosphato-Phenylalanines: Amphiphilic Phosphotyrosine Mimetics Displaying Fluorine-Specific Protein Interactions

Protein tyrosin phosphatases, such as PTP1B are important drug targets as overactivity is linked to diseases such as diabetes or cancer. Phosphotyrosine mimetics are essential tools for targeting these proteins. Here, we design novel phosphotyrosine mimetics, which are specifically tailored to make use of fluorine specific interactions. Specifically, the novel mimetics feature a pentafluorophosphate (PF5) moiety as a negatively charged headgroup, which should fit well into the positively charged main binding pocket of PTP1B. These mimetics need to be amphiphilic, to be able to penetrate cell membranes. Naturally, the phosphotyrosine mimetics also need to efficiently inhibit the protein target, which in this case is selected to be PTP1B. Thus, we aim to demonstrate the novel mimetic's ihibitory activity. Furthermore, as the mimetic's design is based on the novel strategy to exploit fluorine specific interactons, our objective is to rationalize how such interactions can drive the binding strength of the mimetic to PTP1B. For this rationalization, we use the following computational methods:

- Molecular docking
- MD simulations
- Force field parameterization

We realized the synthesis of two phosphotyrosine mimetic amino acids, equipped with a PF5 headgroup via a mild acidic fluorination protocol. The new mimetics were structurally characterized using NMR, IR and X-ray crystallography and the amphiphilicity was confirmed experimentally. Activity assays show one of the PF5 mimetics to bind 25-30 times stronger to PTP1B than the currently strongest biomimetic, which features a phosphate moiety as headgroup. Using molecular docking, we were able to characterize the binding pose of the novel PF5 mimetic. The PF5 headgroup is placed inside the positively charged main binding pocket, close to a series of backbone amine groups and the side chain of PTP1B's Arg221. The aromatic ring of the PF5 mimetic is placed in the close vicinity of Phe128 and Tyr48 and the amino acid backbone ends are pointed towards oppositely charged binding partners. While this preferred docking pose is similar to that of the mimetic with the phosphate headgroup, the lower docking score of the PF5 mimetic indicates a higher binding affinity of this mimetic. In MD simulations of 100 ns length, we see based on RMSF analysis of the protein backbone, that the binding pocket is in one of the most stable regions of the protein and it can thus be expected that the pocket does change its shape significantly. The PF5 headgroup stays close to the side chain of Arg221, which indicates a bond between the positively charged guanidine moiety of Arg221 and the negatively charged PF5 headgroup. Moreover, there are five hydrogen bond like interactions between the backbone amine groups in the main binding pocket and the fluorine atoms in the PF5 headgroup. These interactions are likely to contribute to the high binding affinity in a unique way, as the mimetic with the PF5 headgroup binds with higher affinity than that with phosphate, despite its charge of -1 instead of -2 for the phosphate headgroup. This means the extra negative charge is at least compensated by the fluorine specific interactions.

I conducted the molecular docking experiments and the MD simulations and their analysis. Lauren Finn parameterized the force field needed for the simulations under my supervision as part of her Master's thesis. I wrote the paragraph in the main text explaining the computational results and wrote the section about computational methods in the SI, including the corresponding figures and tables. Bettina Keller and Jörg Rademann revised my text contributions. All other tasks including synthesis, experimental methods, activity assays and text in the manuscript and SI not mentioned above were completed by the other authors.

How to cite: *Angew. Chem. Int. Ed.* **2022,** *61,* e202203579
International Edition: doi.org/10.1002/anie.202203579
German Edition: doi.org/10.1002/ange.202203579

Peptidomimetics

Pentafluorophosphato-Phenylalanines: Amphiphilic Phosphotyrosine Mimetics Displaying Fluorine-Specific Protein Interactions

Matteo Accorsi[+], *Markus Tiemann*[+], *Leon Wehrhan, Lauren M. Finn, Ruben Cruz, Max Rautenberg, Franziska Emmerling, Joachim Heberle, Bettina G. Keller, and Jörg Rademann**

Abstract: Phosphotyrosine residues are essential functional switches in health and disease. Thus, phosphotyrosine biomimetics are crucial for the development of chemical tools and drug molecules. We report here the discovery and investigation of pentafluorophosphato amino acids as novel phosphotyrosine biomimetics. A mild acidic pentafluorination protocol was developed and two PF_5-amino acids were prepared and employed in peptide synthesis. Their structures, reactivities, and fluorine-specific interactions were studied by NMR and IR spectroscopy, X-ray diffraction, and in bioactivity assays. The mono-anionic PF_5 motif displayed an amphiphilic character binding to hydrophobic surfaces, to water molecules, and to protein-binding sites, exploiting charge and H–F-bonding interactions. The novel motifs bind 25- to 30-fold stronger to the phosphotyrosine binding site of the protein tyrosine phosphatase PTP1B than the best current biomimetics, as rationalized by computational methods, including molecular dynamics simulations.

Phosphorylation of the amino acid L-tyrosine is a key regulatory mechanism controlling the function of numerous proteins, governing cellular processes such as protein expression, cell division, development, mobility, and aging.[1] Aberrant activity of tyrosine kinases (TK) or protein tyrosine phosphatases (PTP) in biological systems is linked to diseases such as diabetes[2] and cancer,[3] and thus these proteins have been identified as pharmacologically relevant targets.[4] For a profound understanding of protein tyrosine phosphorylation chemical tools are required to bind, inhibit or manipulate phosphotyrosine binding sites without being prone to enzymatic cleavage or dephosphorylation. To date the most potent "gold standard" phosphotyrosine mimetic is 4-phosphono-difluoromethyl-phenylalanine (PDFM-Phe) **1** and numerous studies have demonstrated highly potent and selective inhibitors with this structure integrated in peptide sequences (Scheme 1A).[5,6] Phosphonic acids, however, are strong acids and form highly polar di-anions resulting in low membrane permeability and thus inactivity in cells.[7]

Considering that phosphate binding sites are coated with a positively charged surface resulting from cationic arginine residues and from H-bond donors,[8] we hypothesized that fragments containing fluorine atoms with negative partial charge might act as H-bond acceptors and thus might be useful as phosphate mimetics. First aromatic fragments containing the pentafluoro-phosphato-difluoromethyl-motif (PFPDFM) were prepared and the PFPDFM-substituted benzene **2** was found to inhibit the phosphotyrosine phosphatase activity of PTP1B with low, millimolar affinity.[9] Synthesis and purification of amino acids containing the PFPDFM-motif failed using the reported conditions with basic fluoride or anhydrous HF.[9,10] Thus, a new pentafluorination protocol needed to be developed. Here, we report the refined synthesis of the PFPDFM motif resulting in the unnatural amino acid 4-pentafluoro-phosphato-difluoromethyl-phenylalanine **3** PFPDFM-Phe, Phe*, Scheme 1B) and investigate the (bio)physical, chemical and biochemical properties of the pentafluoro phosphate motif with a focus on the fluorine-specific interactions exerted by it. Starting from *O*-methyl *N*-Fmoc-4-iodo-phenylalanine **4** and subsequently di-*O*-ethyl-phosphonato-difluoromethyl derivative **5**, the yield of the pentafluorination step toward **8** was raised from traces to almost 70% under acidic conditions (Scheme 1B, b–d).

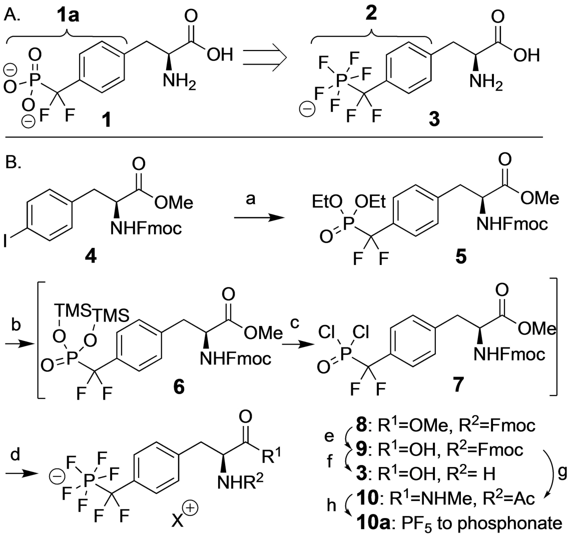
[*] M. Accorsi,[+] M. Tiemann,[+] Prof. Dr. J. Rademann
Department of Biology, Chemistry, Pharmacy, Institute of Pharmacy, Freie Universität Berlin
Königin-Luise-Str. 2+4, 14195 Berlin (Germany)
E-mail: joerg.rademann@fu-berlin.de

L. Wehrhan, L. M. Finn, Prof. Dr. B. G. Keller
Department of Biology, Chemistry, Pharmacy, Institute of Chemistry and Biochemistry, Freie Universität Berlin
Arnimallee 22, 14195 Berlin (Germany)

R. Cruz, Prof. Dr. J. Heberle
Department of Physics, Freie Universität Berlin
Arnimallee 14, 14195 Berlin (Germany)

M. Rautenberg, Dr. F. Emmerling
Bundesanstalt für Materialforschung und -prüfung (BAM)
Richard-Willstätter-Str.11, 12489 Berlin (Germany)

[+] These authors contributed equally to this work.



Scheme 1. Design (A) and synthesis (B) of PFPDFM-phenylalanine **3** as a potential phosphotyrosine mimetic. Reaction conditions: a) Cd, CuBr, $BrCF_2P(O)(OEt)_2$, DMF, 99%; b) TMSBr (5 equiv), ACN; c) oxalylchloride (10 equiv), DMF (5 equiv); d) NMe_4F (10 equiv) (68% for b–d); e) *bacillus licheniformis* protease, 50 mM NH_4HCO_3 buffer, RT, o.n., 96%; f) 20% piperidine in ACN, RT, 8 h, 97%; g) $MeNH_3Cl$, TBTU, DIPEA, ACN, RT, 1 h, 95%; 10% piperidine in ACN, RT, 7 h, 96%; Ac_2O, DIPEA, ACN, RT, 4 h, quant.; h) 2 M HCl, 72 h, quant. TBTU = *O*-(Benzotriazol-1-yl)-*N*,*N*,*N*′,*N*′-tetramethyluronium tetrafluoroborate, TMSBr = Trimethylsilylbromide, DIPEA = Di-isopropyl-ethylamine, RT = room temperature.

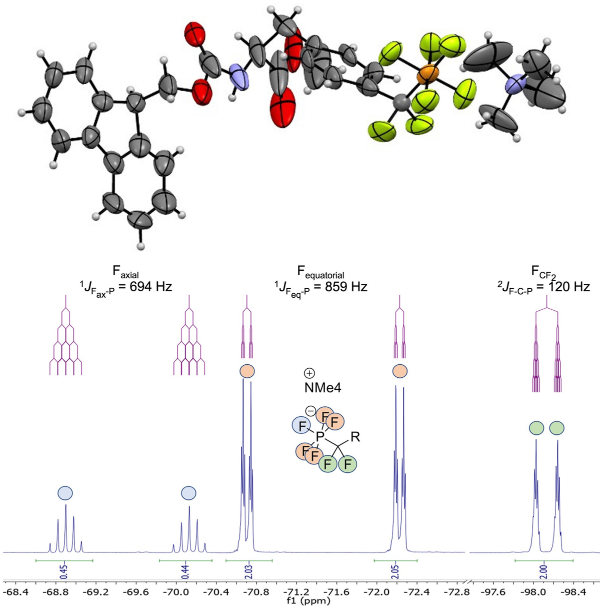


Figure 1. Crystal structure and ^{19}F NMR spectrum of *N*-Fmoc-PFPDFM-Phe-OMe **8**.

Using trimethylsilyl (TMS) bromide, **5** was converted to intermediary di-*O*-TMS-phosphonate **6**, which was transformed in situ to the di-chloro-phosphonate **7** with oxalyl chloride and DMF. Subsequently, fluorination with an excess of tetra-methyl-ammonium fluoride yielded compound **8**, which was isolated after workup in aqueous buffer by reversed phase MPLC. Purity and structure of **8** was confirmed by high-resolution mass spectrometry, ^1H, ^{13}C, ^{19}F, and ^{31}P NMR spectroscopy and by X-ray diffraction of crystalized product.[11] In solution, the phosphorus (V) center was coordinated bipyramidally resulting in a doublet-quintet splitting of the axial fluorine in the ^{19}F spectrum. In the crystal, small deviations of bond angles between axial and equatorial fluorine atoms from 90 , and a slightly elongated axial P–F bond were observed (Figure 1, Supporting Information Figure S2, Supporting Information Table S2).

The methyl ester of **8** was saponified by enzymatic cleavage with *bacillus licheniformis* protease followed by ion exchange on Amberlite yielding the sodium salt of Fmoc-amino acid **9**, the building block for solid phase peptide synthesis. **9** was further converted to the unprotected amino acid **3**. Reaction with acetic anhydride and condensation with *N*-methylamine using TBTU furnished the *N*-acetyl-*N*′-methylamide **10**.

Despite of its permanent negative charge, the PFPDFM motif displayed higher hydrophobicity than the phosphonodifluoromethyl precursor **1**.[12] The retention time of the pen-tafluorinated amino acid **3** in reversed phase C18-HPLC shifted relative to the respective phosphonate **1** by 2.2 min during a 6 min gradient of water and acetonitrile (3.8 min vs. 6 min), corresponding to an increase of from 66 % to 99 % of acetonitrile in the eluent mixture (Figure 2A). The amphiphilicity of the pentafluorophosphato residues was also reflected by the logarithmic partition coefficients (logP) of −0.23 and −0.73 for the tetramethyl ammonium salt of fragment **2** and for the sodium salt **10** measured in DCM/water, respectively. These logP values indicate that a significant portion of pentafluorophosphate salts is found in the organic phase, namely 37 % and 16 %, respectively. In contrast, phosphonates **1** and **10a** remained entirely in the water phase. The amphiphilic nature of the pentafluorophosphate was also reflected in the FTIR spectra. The difference ATR/FTIR spectrum of **2** dissolved in water, recorded against a pure water background, showed two characteristic water bands at 3630 cm^{-1} (O–H stretching mode) and at 1628 cm^{-1} (H–O–H bending mode) (Figure 2B). The peak position of the O–H stretching mode in the presence of **2** is higher and narrower as compared to the O–H stretching mode of bulk water (≈3340 cm^{-1}, full width at half maximum (FWHM) ≈420 cm^{-1}) and the corresponding bending mode also appears lower than that of bulk water (1640 cm^{-1}). These values are characteristic of water molecules lacking one of the four typical hydrogen bonds of bulk water (so-called dangling water).[13] The characteristic water bands do not appear in the IR spectrum of the phosphonate fragment **1a** suggesting that the pentafluorophosphato group forms a hydration shell with O–H–F hydrogen bonds and dangling water molecules.

The chemical stability of the PFPDFM-motif was investigated for compound **3** by HPLC-MS and by ^{19}F NMR spectroscopy. The pentafluorophosphate anion was stable in water from pH 2–12 at RT for 24 h. It tolerated organic

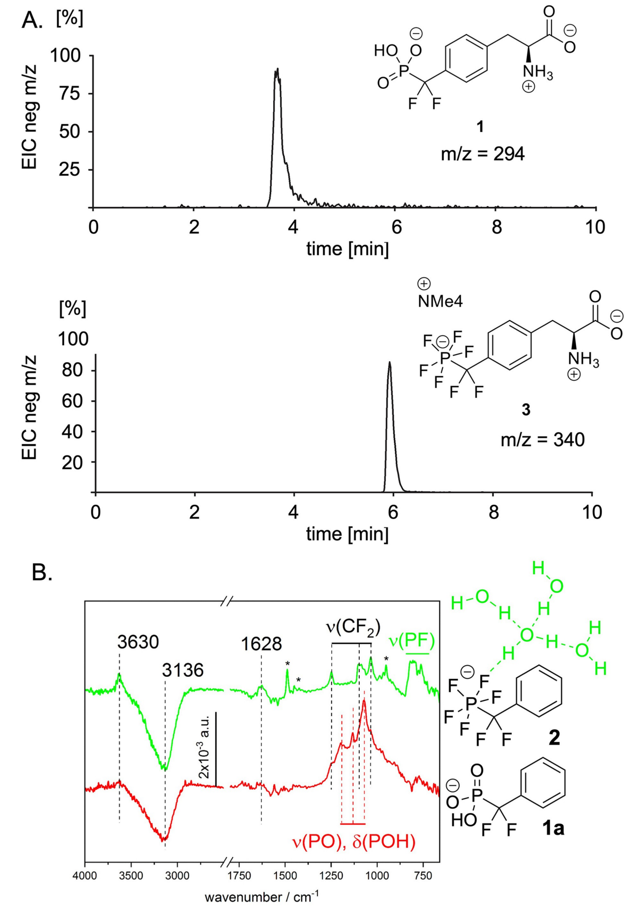


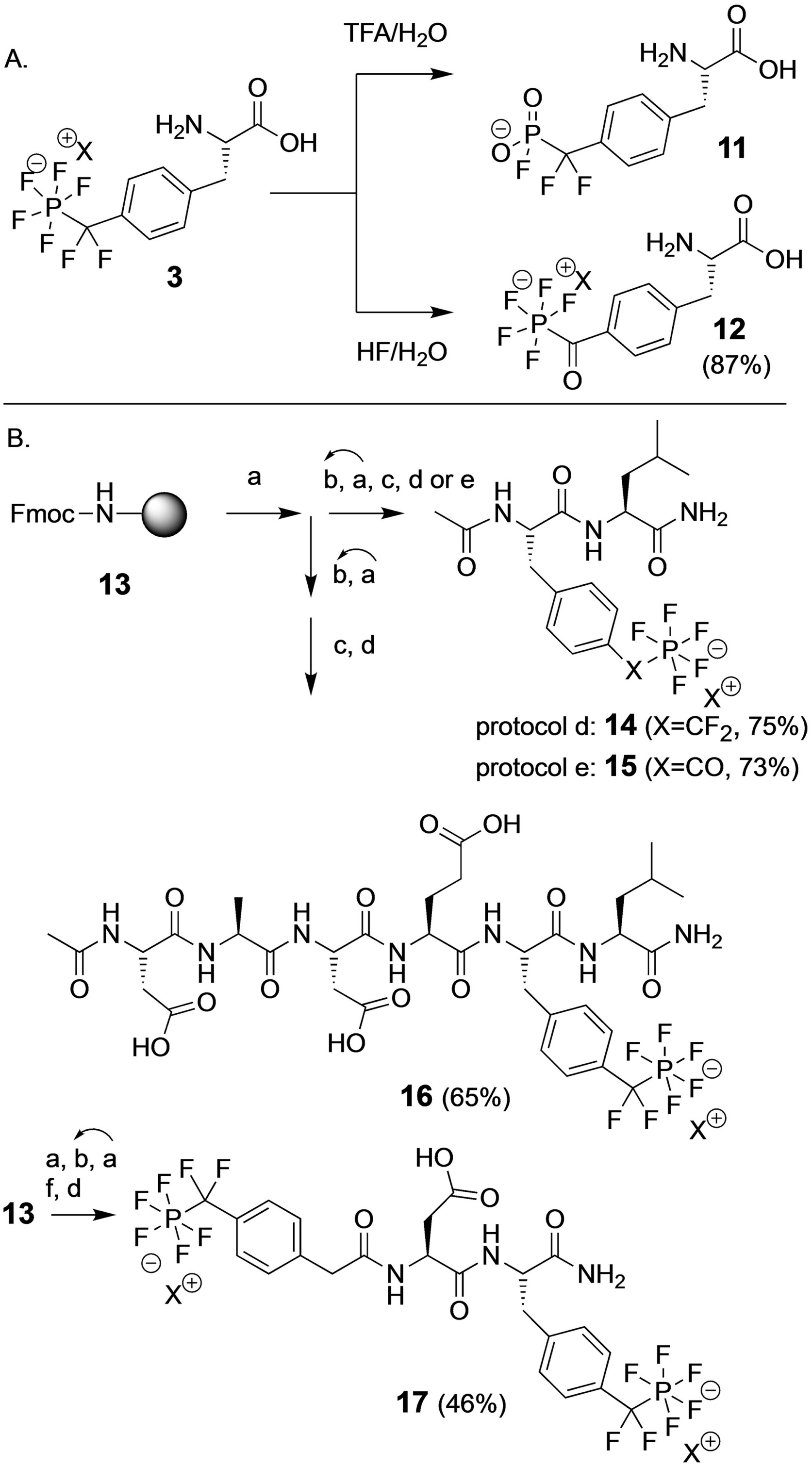
Figure 2. Amphiphilicity of the pentafluorophoshato-difluormethyl (PFPDFM) motif in amino acid **3** and in **2**. A. HPLC on RP-18 silica of **1** and **3**; B. FT-ATR-IR difference spectra of **1a** and **2** (10 mM) in comparison with the spectrum in water. Characteristic signals of the fluorinated fragments including dangling water signals (3630 and 1628 cm^{-1}) are highlighted, full peak assignment and density functional theory (DFT) calculations in Supporting Information Table S3, Supporting Information Figures S3–S5.

bases like pyridine, 20 % piperidine, and 2 % 1,8-diazabicyclo[5,4,0]-undec-8-en (DBU) in DMF or acetonitrile, reducing agents such as dithiothreitol (DTT) and Pd/C with hydrogen. No decomposition was observed by heating to 60 C for one day, by sonification, and under typical condensation conditions for peptide synthesis using *N,N'*-diisopropyl-carbodiimide (DIC)/HOBt and TBTU/DIPEA. In contrast, both aqueous acid at pH < 2 (0.1 M HCl or 0.1 % TFA) and non-aqueous acid (10 % acetic acid in DCM or hexafluoro-isopropanol (HFIP)) effected the hydrolysis of the PFPDFM motif to the monofluorophosphonate **11** (Scheme 2A). Monofluoro-phosphonate **11** was stable at neutral pH and hydrolyzed slowly to the free phosphonate under acidic conditions, e.g. with 5 % perchloric acid over 3 d.

Lability of the PFPDFM motif under acidic conditions excluded the use of standard Fmoc or Boc strategies for solid phase peptide synthesis, even with very acid-labile linkers such as 2-chlorotrityl resin. As an alternative hydrogen fluoride was investigated for cleavage assuming that an excess of HF should protect the PF$_5$-residue from acidic decomposition.[13] Indeed, compounds **2** and **3** were stable when treated with dry pyridinium poly-(hydrogen fluoride) (Olah's reagent).[14] In contrast, treatment with aqueous HF hydrolyzed the difluoromethyl position but not the pentafluorophosphato group, forming the novel amino acid 4-(pentafluorophosphato-carbonyl)-phenylalanine (PFPC-Phe) **12** in 87 % yield. Compound **12** is to our best knowledge the first example of an acyl-pentafluorophosphate formed via hydrolysis of CF$_2$ and was stable over a pH range from 2–12. Structurally related benzoyl phosphonates have been described as photoactive phosphotyrosine mimetics and have been employed in the photo-crosslinking and photo-deactivation of phosphotyrosine binding sites in proteins.[15] Amino acid **12** shows an n-π* transition at 343 nm and irradiation in isopropanol/water (70:30) resulted in the photoconversion of the PF$_5$ residue (Supporting InformationFigure S6).

Employing Rink amide linker on 1 % divinylbenzene (DVB)-polystyrene resin **13** (0.34 mmol g^{-1}), the dipeptide amide Ac-Phe*-Leu-NH$_2$ **14** was synthesized as the first peptide containing the pentafluorophosphato residue (Scheme 2B). Coupling of Fmoc-building block **9** succeeded after activation with TBTU. Following to the final Fmoc-deprotection, the resin was *N*-acetylated, washed, dried, and subsequently treated for 90 min with dry poly-HF-pyridine containing 10 % anisole. Products were washed off the resin with THF and the washing solution was neutralized with saturated aqueous sodium bicarbonate. After evaporation the resulting dipeptide **14** was isolated by reversed phase MPLC using a gradient of ammonium bicarbonate pH 7.5 and acetonitrile in a yield of 75 %. Addition of 1 % water to HF-pyridine cleavage cocktail for 6 h furnished dipeptide **15** containing the pentafluorophosphato-carbonyl (PFPC) residue instead of the PFPDFM-group in 73 % yield after MPLC purification. Applying the water-free protocol, hexapeptide amide Ac-DADEF*L-NH$_2$ **16** was prepared and isolated in 65 % yield (Scheme 2), representing the native autophosphorylation sequence of the epidermal growth factor (EGF) receptor and an established substrate of PTP1B.[16] Tripeptide mimetic **17** containing two PFPDFM residues was designed from a potent PTP1B inhibitor[17] carrying 4-PFPDFM-phenyl-acetamide as N-terminal cap. Synthesis of this peptide required the preparation of 4-PFPDFM-phenylacetic acid **18** starting from 4-iodo-phenylacetate in two steps. Peptide amide **17** could be obtained and isolated using the established protocol in a yield of 46 % (Scheme 2).

Amino acids and peptides were subsequently tested in an enzymatic assay of protein tyrosine phosphatase PTP1B using 6,8-difluoro-4-methylumbelliferyl phosphate (DiFMUP) as the substrate. While phosphonate amino acid **1** displayed a K_I value of 1.55 mM, pentafluorophosphato amino acid **3** showed a K_I of 61 μM, indicating a 25-fold increase of affinity (Figure 3A, Table 1). PF$_5$-amino acid **12** was an even stronger inhibitor of PTP1B with a K_I of 52 μM, 30-fold better than mimetic **1**.



Scheme 2. A) Conversion of amino acid **3** to products **11** and **12** (Reaction conditions to **12**: poly-HF-pyridine, 1% water, RT, 6 h). B) Fmoc-based peptide synthesis employing the PFPDFM motif in amino acid **9**. Reaction conditions: a) 20% Piperidine/DMF, RT, 10 min (twice); b) 5 equiv Fmoc-AA-OH, 4.9 equiv TBTU, DIPEA, DMF, RT, 2 h, repeat a + b for each AA; c) Ac_2O/pyridine (1:1), DMF, 1 h; d) dry poly-HF-pyridine, 10% anisole, RT, 90 min; e) poly-HF-pyridine, 10% anisole, 1% water, RT, 6 h; f) 5 equiv **18**, 4.9 equiv TBTU, HOBt, 10 equiv DIPEA, DMF, 2 h. HOBt = Hydroxybenzotriazole, Fmoc = 9-Fluorenyl-methyl-oxycarbonyl.

As protected amino acid **10** and tripeptide mimetic **17** precipitated in the assay, DMSO concentrations and dilution protocols were varied in the peptide assays. For peptide **15**, identical K_I-values (24, 25, 21 µM) were recorded at 0, 2.5,

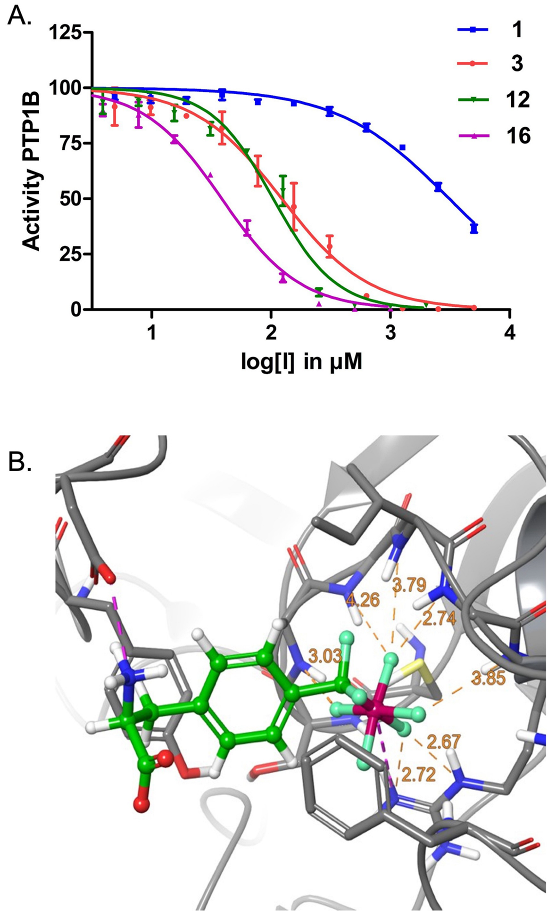


Figure 3. A) Inhibition of the protein tyrosine phosphatase PTP1B by the pentafluorophosphato-phenylalanines **3** and **12**, respectively, was 25-fold and 30-fold stronger than of the classical phosphono-difluormethyl phenylalanine **1** and was further enhanced in peptide mimetics **15** and **16**. B) Docking of **3** suggested the preferred binding pose in the phosphotyrosine binding pocket of PTP1B.

Table 1: Binding affinities of the phosphotyrosine mimetic amino acids and peptides **1**, **3**, **10**, **12**, and **14–17** to protein tyrosine phosphatase PTP1B calculated from enzyme inhibition assays.

Compound	K_I in µM[a]
1	1555 ± 183[b]
3	61 ± 8[c]
10	[f]
12	52 ± 7[d]
14	90 ± 10[c]
15	24 ± 4,[b] 25 ± 4,[c] 21 ± 3[d]
16	74 ± 13,[e] 33 ± 5,[c] 19 ± 3[d]
17	[f]

[a] Assays were performed in triplicate with DiFMUP as a substrate (see Supporting Information part for raw data). Enzyme concentration was 1.5 nM, substrate concentration 67 µM, identical with the experimentally determined K_M value of the substrate. IC_{50} values were converted into the corresponding K_I values applying the Cheng-Prusoff equation (with $[\mathrm{S}] = K_\mathrm{M}$ this results in $\mathrm{IC}_{50}/2 = K_I$). [b] 0% DMSO, [c] 2.5%, [d] 5% DMSO, [e] 2.5%, stock diluted in buffer, [f] visible precipitate.

and 5% DMSO, while the apparent affinity of peptide **16** was raised significantly from 74 µM (2.5% DMSO, dilution in buffer) to 33 µM (2.5%, dilution in DMSO) and 19 µM (5% DMSO). This observation corresponded with the higher polarity of **15** vs. **16** as shown in RP-HPLC (Supporting Information Figure S16). While **15** was soluble in aqueous buffer, dynamic light scattering revealed the aggregation of peptide **16** in DMSO/buffer (Supporting Information Figures S12–S15) which reduced the apparent affinity to the target. Both peptides **15** and **16** showed stronger inhibition than the respective PF_5 amino acids alone. These data suggest that the bioactivity of PF_5-containing compounds depends not only on their target-binding but also on their physico-chemical properties and formulation.

Docking studies of amino acids **1**, **3**, **10**, and peptide **15** to PTP1B using the commercial docking software Glide[18,19] revealed binding poses in the binding pocket and key interactions between the amino acids and the protein (Figure 3B, Supporting Information Figures S14, S15, S20–S23). In the docked pose, the ligand was held in place by two salt bridges, one between the negative charge of the phosphate residue and the sidechain of Arg221 and the other between the positively charged, protonated alpha-amino group and the negatively charged sidechain of Asp48. We derived forcefield parameters for amino acids **1** and **3** and were able to extensively sample the conformational flexibility of these ligands in the binding pocket through molecular dynamics (MD) simulations (see Supporting Information Figures S15–S19). While the salt bridge from the phosphate head group stayed in place throughout 100 ns of the MD simulation, in explicit solvent the backbone amine turned outward at the start of the simulation and preferred solvent exposure. The aromatic rings of **1** and **3**, respectively, were close to the aromatic rings of Phe182 and Tyr46, however, π-π interactions were not observed in the final docking poses and rarely found in the simulation snapshots. In the MD simulations, the backbone amide NH groups remained close to the fluorines of the PF_5 moiety and for amino acids 217–221 the closest average N⋅⋅⋅F distance was found to be generally below 3 Å, which implied the continuous presence of N–H–F interactions.

In summary, this work established the first synthesis and validation of pentafluorophosphato amino acids. High-yielding pentafluorinations were developed as a one-pot reaction under acidic conditions and yielded pentafluorophosphato-difluoromethyl (PFPDFM) phenylalanine **3**. Hydrolysis with aqueous HF provided the pentafluorophosphato-carbonyl (PFPC) phenylalanine **12**. Both PF_5-amino acids were successfully incorporated into peptides using HF for cleavage. The new chemical entities displayed remarkable structural, physico-chemical, and biochemical properties resulting from fluorine-specific interactions of the PF_5-anion. These include interactions of PF_5 residues with water, hydrophobic surfaces, organic solvents, as well as aggregation events. As a result, PF_5 amino acids bind 25 to 30 times stronger to the phosphotyrosine binding site of PTP1B than classical phosphonate biomimetics. Our studies demonstrate that developing improved PF_5 ligands requires the optimization of protein-ligand interactions but also of physicochemical properties and formulation of the molecules.

Acknowledgements

The authors acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project-ID 387284271—SFB 1349 (fluorine-specific interactions, projects A3, A5, A6, and C5). M. A. received a graduate fellowship from the SFB 765. We thank Dr. Felix Bredendiek for analytical support, the work was supported by the DFG-funded core facility BioSupraMol and the Zentraleinrichtung für Datenverarbeitung (ZEDAT) of Freie Universität Berlin. Open Access funding enabled and organized by Projekt DEAL.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available in the Supporting Information of this article.

Keywords: Chemical Biology · Drug Development · Pentafluorophosphates · Phosphotyrosine Biomimetics · Protein Tyrosine Phosphatases

[1] a) T. Hunter, *Cell* **2000**, *100*, 113–127; b) F. Ardito, M. Giuliani, D. Perrone, G. Troiano, L. Lo Muzio, *Int. J. Mol. Med.* **2017**, *40*, 271–280.

[2] S. S. Abdelsalam, H. M. Korashy, A. Zeidan, A. Agouni, *Biomolecules* **2019**, *9*, 286.

[3] a) T. Kostrzewa, J. Styszko, M. Gorska-Ponikowska, T. Sledzinski, A. Kuban-Jankowska, *Anticancer Res.* **2019**, *39*, 3379–3384; b) V. Singh, M. Ram, R. Kumar, R. Prasad, B. K. Roy, K. K. Singh, *Protein J.* **2017**, *36*, 1–6.

[4] a) M. F. Cicirelli, N. K. Tonks, C. D. Diltz, J. E. Weiel, E. H. Fischer, E. G. Krebs, *Proc. Natl. Acad. Sci. USA* **1990**, *87*, 5514–5518; b) D. Kraskouskaya, E. Duodu, C. C. Arpin, P. T. Gunning, *Chem. Soc. Rev.* **2013**, *42*, 3337–3370; c) R. J. He, Z. H. Yu, R. Y. Zhang, Z. Y. Zhang, *Acta Pharmacol. Sin.* **2014**, *35*, 1227–1246.

[5] a) T. R. Burke, Jr., M. S. Smyth, A. Otaka, M. Nomizu, P. P. Roller, G. Wolf, R. Case, S. E. Shoelson, *Biochemistry* **1994**, *33*, 6490–6494; b) I. K. Shen, Y. F. Keng, L. Wu, X. L. Guo, D. S. Lawrence, Z. Y. Zhang, *J. Biol. Chem.* **2001**, *276*, 47311–47319; c) T. Burke, *Curr. Top. Med. Chem.* **2006**, *6*, 1465–1471; d) G. Boutselis, X. Yu, Z. Y. Zhang, R. F. Borch, *J. Med. Chem.* **2007**, *50*, 856–864; e) S. Zhang, Z. Y. Zhang, *Drug Discovery Today* **2007**, *12*, 373–381; f) M. Köhn, C. Meyer, *Synthesis* **2011**, 3255–3260; g) C. Meyer, B. Hoeger, K. Temmerman, M. Tatarek-Nossol, V. Pogenberg, J. Bernhagen, M. Wilmanns, A. Kapurniotu, M. Köhn, *ACS Chem. Biol.* **2014**, *9*, 769–776; h) H. Liao, D. Pei, *Org. Biomol. Chem.* **2017**, *15*, 9595–9598.

[6] For a recent review on phosphotyrosine mimetics: N. Makukhin, A. Ciulli, *RSC Med. Chem.* **2021**, *12*, 8–23.

[7] a) R. D. Kornberg, M. G. McNamee, H. M. McConnell, *Proc. Natl. Acad. Sci. USA* **1972**, *69*, 1508–1513; b) L. Xie, S. Y. Lee, J. N. Andersen, S. Waters, K. Shen, X. L. Guo, N. P. Moller, J. M. Olefsky, D. S. Lawrence, Z. Y. Zhang, *Biochemistry* **2003**, *42*, 12792–12804; c) A. J. Wiemer, D. F. Wiemer, *Top. Curr. Chem.* **2015**, *360*, 115–160.

[8] a) J. P. Sun, L. Wu, A. A. Fedorov, S. C. Almo, Z. Y. Zhang, *J. Biol. Chem.* **2003**, *278*, 33392–33399; b) D. Zhao, L. Sun, S. Zhong, *Mol. Diversity* **2021**, https://doi.org/10.1007/s11030-021-10323-2.

[9] S. Wagner, M. Accorsi, J. Rademann, *Chem. Eur. J.* **2017**, *23*, 15387–15395.

[10] Preparation of few simple pentafluorophosphates has been described before via disproportionation under harsh conditions not tolerating functional groups like esters and carbamates: a) N. V. Pavlenko, L. A. Babadzhanova, I. I. Gerus, Y. L. Yagupolskii, W. Tyrra, D. Naumann, *Eur. J. Inorg. Chem.* **2007**, 1501–1507; b) C. Tian, W. Nie, M. V. Borzov, P. Su, *Organometallics* **2012**, *31*, 1751–1760; c) O. I. Guzyr, S. V. Zasukha, Y. G. Vlasenko, A. N. Chernega, A. B. Rozhenko, Y. G. Shermolovich, *Eur. J. Inorg. Chem.* **2013**, 4154–4158; d) N. Ignatyev, P. Barthen, K. Koppe, W. Frank, WO 2016/074757A, **2016**.

[11] Deposition Number 2157920 contains the supplementary crystallographic data for this paper. These data are provided free of charge by the joint Cambridge Crystallographic Data Centre and Fachinformationszentrum Karlsruhe Access Structures service.

[12] W. Hoffmann, J. Langenhan, S. Huhmann, J. Moschner, R. Chang, M. Accorsi, J. Seo, J. Rademann, G. Meijer, B. Koksch, M. T. Bowers, G. von Helden, K. Pagel, *Angew. Chem. Int. Ed.* **2019**, *58*, 8216–8220; *Angew. Chem.* **2019**, *131*, 8300–8304.

[13] a) D. Ben-Amotz, *J. Am. Chem. Soc.* **2019**, *141*, 10569–10580; b) L. F. Scatena, M. G. Brown, G. L. Richmond, *Science* **2001**, *292*, 908–912; c) P. N. Perera, K. R. Fega, C. Lawerence, E. J. Sundstrom, J. Tomlinson-Philips, D. Ben-Amotz, *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 12230–12234; d) M. Bonn, Y. Nagata, E. H. G. Backus, *Angew. Chem. Int. Ed.* **2015**, *54*, 5560–5576; *Angew. Chem.* **2015**, *127*, 5652–5669.

[14] S. Matsuura, C.-H. Niu, J. S. Cohen, *J. Chem. Soc. Chem. Commun.* **1976**, 451–452.

[15] a) A. Horatscheck, S. Wagner, J. Ortwein, M. Lisurek, B. G. Kim, S. Beligny, A. Schütz, J. Rademann, *Angew. Chem. Int. Ed.* **2012**, *51*, 9441–9447; *Angew. Chem.* **2012**, *124*, 9577–9583; b) S. Wagner, A. Schütz, J. Rademann, *Bioorg. Med. Chem.* **2015**, *23*, 2839–2847; c) E. L. Wong, E. Nawrotzky, C. Arkona, B. G. Kim, S. Beligny, X. Wang, S. Wagner, M. Lisurek, D. Carstanjen, J. Rademann, *Nat. Commun.* **2019**, *10*, 66.

[16] Z. Jia, D. Barford, A. J. Flint, N. K. Tonks, *Science* **1995**, *268*, 1754–1758.

[17] J.-P. Sun, A. A. Fedorov, S.-Y. Lee, X.-L. Guo, K. Shen, D. S. Lawrence, S. C. Almo, Z.-Y. Zhang, *J. Biol. Chem.* **2003**, *278*, 12406–12414.

[18] C. Lu, et al., *J. Chem. Theory Comput.* **2021**, *17*, 4291–4300.

[19] a) R. A. Friesner et al., *J. Med. Chem.* **2004**, *47*, 1739–1749; b) R. A. Friesner et al., *J. Med. Chem.* **2006**, *49*, 6177–6196; c) T. A. Halgren et al., *J. Med. Chem.* **2004**, *47*, 1750–1759.

Manuscript received: March 8, 2022
Accepted manuscript online: March 18, 2022
Version of record online: April 25, 2022

4.5 Fluorinated Phenylalanine and the GrsA A-Domain

Title: Biosynthetic incorporation of fluorinated amino acids into the nonribosomal peptide gramicidin S

Natural products are a major source for pharmaceuticals and they almost never contain fluorine. Fluorine on the other hand is an effective modulator of molecular properties, especially those of drug-like molecules, which is why we attempt here to incorporate fluorine into a natural product, namely the nonribosomal peptide gramicidin S. As we found that the A domain of the nonribosomal peptide synthetase GrsA rejects fluorinated amino acids, more specifically phenylalanine residues fluorinated at the *para*-position (4-F-Phe), we then attempt to characterize this effect. The discovery of the rejection of 4-F-Phe residues by the GrsA A domain and the computational study of this effect are novel in the field of fluorinated natural products. We used the following computational methods for studying the rejection of 4-F-Phe residues by the GrsA A domain:

- Molecular Mechanics based interaction analysis
- MD simulations
- Molecular docking

We found that the GrsA A domain rejects 4-F-Phe over the natural substrate with a 31-fold difference in *in vitro* activity assays. This rejection can be explained by considering the specific protein-substrate interactions, as a crucial T-shaped aromatic interaction between the substrate and the side chain of Trp239 of GrsA is interrupted by the introduction of fluorine. We quantified this interruption in a Molecular Mechanics based distance scan. The energy well of the interaction in our simple model system of only the interacting amino acids in vacuum is about 10 kJ more shallow for 4-F substituted Phe variants, compared to unsubstituted Phe. We also tested the interaction energy for mono- and di-substituted Phe variants that are substituted in 2-,3- or 5-position and found that, as long as the 4-position is not substituted, the energy well of the interaction is slightly deeper compared to the unsubstituted amino acid. Interestingly, the selectivity of the GrsA A domain for Phe over 4-F-Phe can be reversed by mutating the amino acid Trp239 to a serine (W239S). MD simulations show that the space that was occupied by the large Trp239 side chain is occupied by two or three water molecules in the W239 mutant. With 4-F-Phe as a substrate instead of Phe, this number reduces to zero to two water molecules. We did not observe any direct hydrogen bond like interactions from fluorine to the water molecules, which means the reason for the increase in selectivity must have a different reason. We speculate that entropic effects, which fluorine might induce on the water molecules, possibly contribute to the observed selectivity. Moreover, it is possible that the aqueous environment is a better fit for the negative fluorine instead of the protein binding pocket. A better fit of fluorine in the cavity left by the mutation could be implied by the reduced number of water molecules in the cavity. A similar effect was observed when using O-propargyl-Tyr as a substrate, which causes a substantial shift in selectivity of the substituted substrate over the natural substrate for W239S GrsA. The better fit of O-propargyl-Tyr over 4-F-Phe was confirmed using molecular docking experiments.

I conducted and analyzed the Molecular Mechanics based interaction scan and the MD simulations. I conducted and analyzed the molecular docking experiments. I wrote the paragraph in the main text explaining the results of computational modeling and created the corresponding figures. I wrote the section in the SI about computational methods and created the corresponding figures and tables. Bettina Keller and Hajo Kries revised my text contributions. All experimental methods not mentioned above and their analysis and corresponding text and figures were done by the other authors.

RSC
Chemical Biology

PAPER

View Article Online
View Journal | View Issue

Cite this: *RSC Chem. Biol.*, 2023, 4, 692

Received 30th April 2023,
Accepted 24th July 2023

DOI: 10.1039/d3cb00061c

rsc.li/rsc-chembio

Biosynthetic incorporation of fluorinated amino acids into the nonribosomal peptide gramicidin S†

Maximilian Müll,[a] Farzaneh Pourmasoumi,[a] Leon Wehrhan,[b] Olena Nosovska,[c] Philipp Stephan,[a] Hannah Zeihe,[a] Ivan Vilotijevic,[c] Bettina G. Keller[b] and Hajo Kries *[ad]

Fluorine is a key element in medicinal chemistry, as it can significantly enhance the pharmacological properties of drugs. In this study, we aimed to biosynthetically produce fluorinated analogues of the antimicrobial cyclic decapeptide gramicidin S (GS). However, our results show that the A-domain of the NRPS module GrsA rejects 4-fluorinated analogues of its native substrate Phe due to an interrupted T-shaped aromatic interaction in the binding pocket. We demonstrate that GrsA mutant W239S improves the incorporation of 4-fluorinated Phe into GS both *in vitro* and *in vivo*. Our findings provide new insights into the behavior of NRPSs towards fluorinated amino acids and strategies for the engineered biosynthesis of fluorinated peptides.

Introduction

The introduction of fluorine atoms is a widespread strategy in medicinal chemistry to fine-tune drug properties.[1,2] Fluorination can improve the fit to a protein binding pocket, or improve pharmacokinetic parameters, as in the case of the "second generation" macrolide antibiotic flurithromycin.[3] Throughout the last two decades, many fluorinated drugs transitioned from the clinical stage to the market (Fig. 1), which demonstrates the importance of fluorination. For drugs acting on protein targets, the focus has long been on using fluorine to better fit a binding pocket. This led to an understanding of fluorine–enzyme interactions,[1,4] which can be used for rational design of small molecule libraries.

Although natural products are a major source of new drugs,[5] they rarely contain fluorine.[6] Therefore, attempts have been made to biosynthetically incorporate fluorine into natural products belonging to the classes of alkaloids, nonribosomal peptides (NRPs), polyketides, and cyclic dinucleotides.[7–19] Notable successes have been achieved in engineering polyketide biosynthesis, bringing the biosynthesis of flurithromycin almost within reach.[17] While the similarity between a C–H and a C–F bond allows many non-natural, fluorinated analogues to slip through biosynthetic selectivity filters, designing biosynthetic enzymes with binding pockets selective for fluorinated substrates is largely unexplored.[18] Hence, directing fluorine incorporation into natural products often remains unpredictable and low yielding.

Incorporation of fluorinated amino acids into NRPs, which are a prolific source of various drugs and antibiotics,[20–22] has so far relied on precursor-directed biosynthesis exploiting the natural promiscuity of biosynthetic enzymes.[9,11,13–16,19] NRPs are produced by large multimodular enzymes called nonribosomal peptide synthetases (NRPSs).[21,23,24] These NRPSs are built from three core domains (Fig. 2A). Adenylation domains (A-domains) activate a specific amino acid, thiolation domains (T-domains) carry the activated thioester intermediates, and condensation domains (C-domains) catalyze peptide bond formation between substrates bound to adjacent T-domains. A-domains typically exhibit selectivity for specific amino acids and thereby determine the NRP sequence.[25–27] This selectivity correlates strongly with the identity of residues in the A-domain binding pocket,[25,26] which allows reliable prediction,[28] and design to some extent.[29] Since fluorinated amino acids have not been described as natural NRPS substrates, A-domain binding pockets specific for them are elusive.

Here, we investigate the specificity for fluorinated Phe analogues of GrsA, the first module of the gramicidin S (GS) NRPS.[30] According to adenylation kinetics, GrsA has a large preference for the native Phe substrate over 4-fluorinated analogues, which prevents incorporation of fluorine into GS.

[a] *Junior Research Group Biosynthetic Design of Natural Products, Leibniz Institute for Natural Product Research and Infection Biology, Hans Knöll Institute (HKI Jena), Jena 07745, Germany. E-mail: hajo.kries@leibniz-hki.de*
[b] *Freie Universität Berlin, Department of Biology, Chemistry, and Pharmacy, Institute of Chemistry and Biochemistry, Arnimallee 20, Berlin 14195, Germany*
[c] *Institute of Organic Chemistry and Macromolecular Chemistry, Friedrich Schiller University Jena, Humboldtstr. 10, Jena 07743, Germany*
[d] *University of Bayreuth, Organic Chemistry I, Bayreuth 95440, Germany*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d3cb00061c

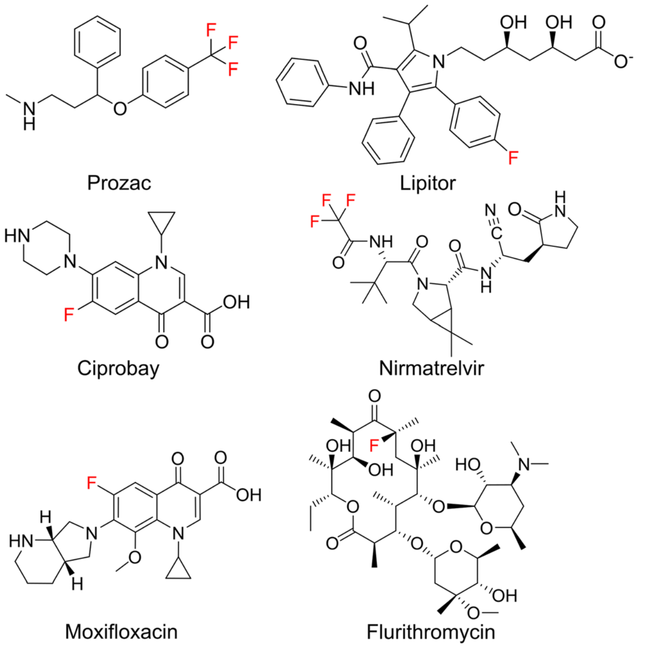


Fig. 1 Important fluorinated drug molecules.

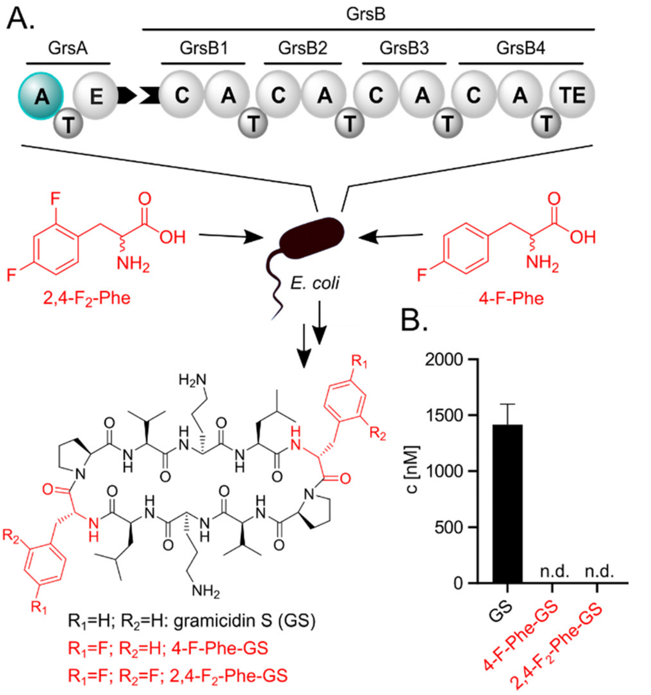


Fig. 2 (A) *E. coli* heterologously expressing the GS NRPS. Either Phe or a fluorinated analog (2,4-F_2-Phe or 4-F-Phe) is supplied to the medium. The A-domain of GrsA which selects Phe or analogs is highlighted in teal. (B) Biosynthetic products are quantified *via* UPLC-MS/MS. N.d.: not detectable. Fluorinated compounds or residues are highlighted in red.

However, mutation W239S known to enhance incorporation of *para*-substituted Phe-derivatives[31] yields a preference for the fluorinated substrates and thus allows *in vivo* production of fluorinated GS. We conclude that A-domain mutagenesis can significantly improve fluorine incorporation into NRPs, which is good news for the biosynthetic engineering of therapeutically valuable peptides.

Results

We initially hypothesized that one or two fluorine substituents would be well tolerated by the biosynthetic machinery and aimed to use precursor feeding to make fluorinated analogues of the cyclic decapeptide GS (Fig. 2). For this purpose, we used *E. coli* HM0079[32] carrying the gene cluster for GS biosynthesis on plasmid pSU18-GrsTAB as a heterologous production platform.[33] The pentamodular GS NRPS consisting of GrsA (one module) and GrsB (four modules) produces a DPhe-Pro-Val-Orn-Leu pentapeptide, which is dimerized and cyclized. Addition of racemic 4-F-Phe or 2,4-F_2-Phe did not impair the growth of the GS producing *E. coli* cultures. However, the expected products 2,4-F_2-Phe-GS and 4-F-Phe-GS (Fig. 2B), in which Phe residues should have been substituted with the fluorinated analogues, were not detectable by LC-MS/MS. We hypothesized that the A-domain of module GrsA, which incorporates Phe, rejects the fluorinated amino acids despite the seemingly small structural perturbation caused by fluorination.

To explain the rejection of fluorinated amino acids by GrsA, we investigated the adenylation specificity of this NRPS module using the MESG/hydroxylamine assay[34] and recorded saturation kinetics. For this purpose, GrsA was expressed in His-tagged form and purified (Fig. S1, ESI†).[31] A comparison of the specificity constants (k_{cat}/K_M's; Table 1 and Fig. S2, ESI†) revealed 8- and 31-fold preferences for Phe over 2,4-F_2-Phe and 4-F-Phe, respectively. The turnover rates at substrate saturation (k_{cat}) are in a similar range for all substrates and the difference is mostly caused by an increase in the Michaelis wconstant (K_M) for the non-native substrates. Interestingly, the detrimental effect of fluorination in the 2- and 4-position of the phenyl ring is not additive, as 2,4-F_2-Phe (k_{cat}/K_M = 44 mM^{-1} min^{-1}) is slightly superior to 4-F-Phe (k_{cat}/K_M = 11 mM^{-1} min^{-1}) as a substrate. These results indicate that GrsA-A is catalytically capable of adenylating fluorinated analogs of Phe, but Phe outcompetes these analogs when both substrates are present. To compare the poor fluorine selectivity of GrsA with other NRPS modules, we additionally tested the Val-specific A-domain SrfA-B1 from surfactin A biosynthesis.[35] While adenylation activity for hexafluorinated F_6-Val was not even detectable, *rac*-3-F-Val showed a similar trend as GrsA with a 39-fold selectivity for the native over the monofluorinated substrate.

To directly test the selectivity of GrsA under competition conditions, which are similarly found inside cells fed with fluorinated amino acids, we used a multiplexed hydroxamate assay (HAMA, Fig. 3a and b). HAMA is based on direct detection of amino acid hydroxamates formed from amino acyl adenylates in the presence of 150 mM hydroxylamine.[36] HAMA confirmed strong preference of wildtype GrsA for Phe over fluorinated Phe (Fig. 3a).

To understand the origins of the high K_M-values for fluorinated substrates, we performed binding studies using

Table 1 Michaelis–Menten parameters for adenylation[a]

Enzyme	Substrate[b]	k_{cat} [min^{-1}]	K_M [mM]	k_{cat}/K_M [min^{-1} mM^{-1}]	Selectivity[c]
GrsA	Phe	16 ± 2	0.05 ± 0.01	340	N.a.
	4-F-Phe	26 ± 3	2.3 ± 0.7	11	31
	2,4-F_2-Phe	42 ± 3	0.9 ± 0.2	44	7.6
	2-F-Phe	27.5 ± 0.8	0.020 ± 0.002	1400	0.24
	3-F-Phe	21 ± 2	0.014 ± 0.005	1500	0.23
	3,5-F_2-Phe	31 ± 2	0.014 ± 0.003	2200	0.15
GrsA-W239S	Phe	21 ± 3	5 ± 1	4.1	N.a.
	4-F-Phe	20 ± 10	3 ± 5	5.7	0.7
	2,4-F_2-Phe	15 ± 4	2 ± 2	7.2	0.57
	2-F-Phe	40 ± 10	3 ± 1	16	0.25
	3-F-Phe	14 ± 1	0.9 ± 0.1	15	0.27
	3,5-F_2-Phe	10 ± 2	2.2 ± 0.8	4.3	0.9
SrfA-B1	Val	8 ± 2	0.06 ± 0.02	140	N.a.
	3-F-Val	28 ± 3	8 ± 1	3.5	39
	F_6-Val	N.d.	N.d.	N.a.	N.a.

[a] Adenylation kinetics were measured using the MESG/hydroxylamine assay (Fig. S2, ESI).[34] N.a.: not applicable; n.d.: not detectable. [b] All amino acids are used in racemic form. [c] The selectivity is calculated as (k_{cat}/K_M[native substrate])/(k_{cat}/K_M[fluorinated analog]).

isothermal titration calorimetry (ITC, Fig. S3, ESI†). For ITC experiments, we expressed only the N-terminal fragment (A_{core}) of GrsA-A, which contains the amino acid binding pocket but not the catalytic lid of the A-domain. The K_D values determined by ITC were 600 ± 300 μM for 4-F-Phe, 420 ± 40 μM for 2, 4-F_2-Phe, and 60 ± 10 μM for Phe binding to GrsA-A_{core}. Apparently, the increase in K_M caused by fluorination of the substrate is due to a higher K_D (weaker binding) of the fluorinated substrates. Previously, fluorination of drug molecules has been observed to cause entropy-enthalpy trade-offs upon binding[38] resulting only in minor changes in K_D. We did not observe this effect with GrsA-A_{core}. If there are differences, these were obscured by the errors on the measured ΔH and ΔS values (Table S2, ESI†).

GrsA's surprisingly strong discrimination against 4-fluorinated Phe derivatives also suggested that this preference could perhaps be inverted through mutagenesis to allow efficient and selective biosynthesis of fluorinated peptides. According to the three-dimensional structure of GrsA-A (Fig. 3C),[37] the para-position of the substrate's phenyl side chain points towards the indole side chain of protein residue Trp239. Since the substrate side chain is accommodated in a tightly packed, hydrophobic space, the fluorinated substrates might be rejected due to the slightly larger size and higher polarity of fluorine compared to hydrogen (Fig. 3D). To better accommodate the fluorine substituent, we mutated residue Trp239 to Ser which is smaller and more polar. This mutation has previously been shown to promote activation of para-substituted Phe in GrsA and related A-domains.[31,39] Gratifyingly, adenylation kinetics show that the k_{cat} is barely influenced by mutation W239S (Table 1 and Fig. S2, ESI†). At the same time, the K_M value for Phe increases 100-fold, while it remains nearly unchanged for the 4-fluorinated analogs. As a result, the specificity constant (k_{cat}/K_M) of GrsA-W239S is higher for the fluorinated Phe analogs than for Phe. For 4-F-Phe, mutation W239S causes a 43-fold switch in specificity. To test the performance of the mutant under cell-like substrate competition, we again employed HAMA (Fig. 3B). The specificity switch in favor of the fluorinated substrates was confirmed, although Tyr turned out to be the overall preferred substrate. Screening of other side-chains in position 239 using HAMA in 96-well plate format[29] showed that mutants W239A and W239L also have increased preference for the 4-fluorinated substrates but slightly less than W239S (Fig. S5, ESI†).

The effect of a fluorine substituent on substrate binding to GrsA and GrsA-W239S was further investigated by computational modelling (Fig. 3C–E). We hypothesized that the fluorine substituent might disturb the edge-to-face (T-shaped) aromatic interaction[40] between the substrate side-chain and Trp239 (Fig. 3C). To gauge the strength of this interaction, the potential energy was calculated for a model system consisting only of capped Trp and Phe in vacuum using the Amber14SB/GAFF2 force field (Fig. 3E). With one or two fluorine substituents, the energy well for the interaction is about 10 kJ mol^{-1} shallower than with an unsubstituted phenyl ring, explaining the lower binding of the fluorinated substrates. To understand why F-Phe binds GrsA-W239S better than Phe, we considered the cavity generated by the large-to-small mutation in a 10 ns molecular dynamics simulation. While GrsA with Phe as a ligand has zero water molecules near the tip of the phenyl side-chain, two or three waters are found in GrsA-W239S (Table S3, ESI†). With F-Phe as a ligand, that number is reduced to zero to two waters, indicating a better fit into the cavity. No binding interaction between Ser239 and fluorine was detected, but the aqueous environment may accommodate the electronegative fluoro-substituent better than the hydrophobic pocket in GrsA. However, the fluoro-substituent fills the enlarged GrsA-W239S binding pocket less efficiently than the previously tested *O*-propargyl-Tyr, that was also docked into the pocket (Table S4 and Fig. S4, ESI†) and for which the mutation causes a more dramatic 200 000-fold specificity switch.[31]

After identifying the T-shaped aromatic interaction as crucial for substrate binding in GrsA, we were eager to investigate the behavior of fluorinated Phe analogues that lacked substitution at the 4-position. We expected strongly electron-withdrawing fluoro-substituents to strengthen the electrostatic

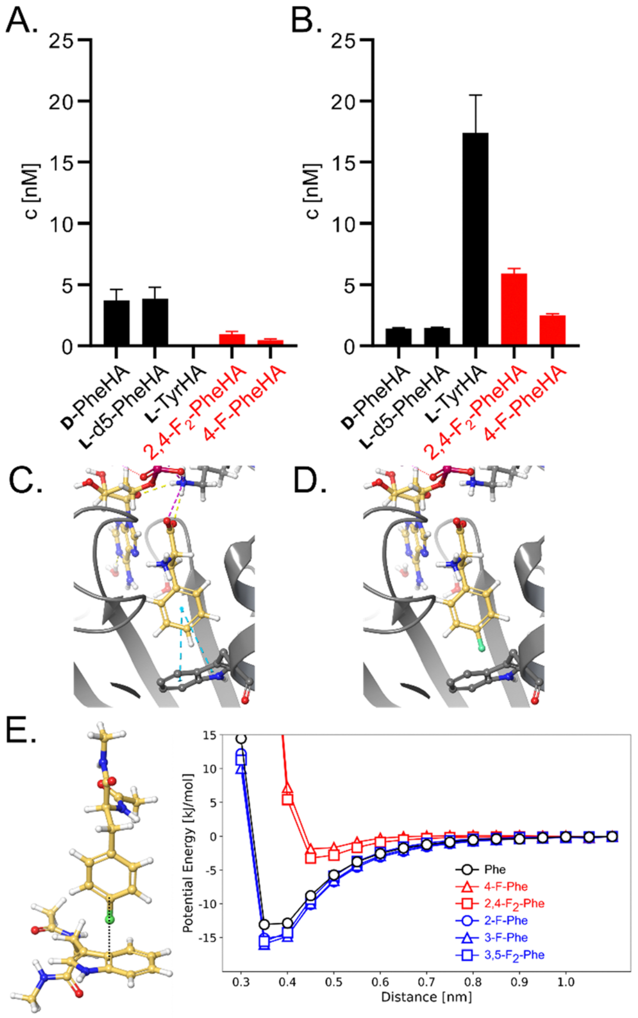


Fig. 3 Impact of mutation W239S on substrate specificity. (A) HAMA profiles of GrsA and (B) GrsA-W239S. All proteinogenic amino acids were present as substrates, but only detectable hydroxamates are shown. Deuterated Phe was used to distinguish D- and L-Phe.[36] Fluorinated substrates were added in racemic form. (C) Crystal structure of GrsA with Phe and AMP bound as ligands (PDB ID 1 amu)[37] and (D) a computational model of 4-F-Phe bound to the same active site. (E) Potential energy of a model system, calculated from classical force field parameters, as a function of the distance between Phe (carbon atom in 4-position) and the aromatic system of Trp.

interaction with the electron-rich indole side-chain when they were not interrupting the interaction at the 4-position. These expectations were substantiated by additional simulations also indicating a slightly lowered energy minimum for the T-shaped aromatic interaction (Fig. 3E). Consequently, we recorded saturation kinetics with 2-F-Phe, 3-F-Phe, and 3,5-F_2-Phe (Table 1). Indeed, GrsA prefers 2-F-Phe, 3-F-Phe, and 3,5-F_2-Phe over Phe. The best substrate was 3,5-F_2-Phe, which was adenylated with a 6.5-fold higher catalytic efficiency than the wild-type substrate Phe. In agreement with our binding model, this substantial improvement is absent in the W239S mutant where the crucial aromatic interaction is absent.

Having identified the selectivity of GrsA as critical for the biosynthesis of fluorinated GS analogs, we eliminated competition with Phe by using an *in vitro* system to produce fluorinated GS (Fig. 4A). Therefore, we expressed and purified GrsA and GrsB in His-tagged form from *E. coli* HM0079.[33,41] To a reaction with GrsA and GrsB, we supplied either L-Phe, 2,4-F_2-Phe, or 4-F-Phe in addition to ATP and all other required amino acids. *In vitro*, 4-F-Phe-GS reaches a concentration of 1100 ± 100 nM, which is 61% of the GS concentration obtained under the same conditions (Fig. 4A). Apparently, the fluorinated substrates are tolerated by all downstream domains once they have passed the selectivity filter of GrsA-A. Under competitive conditions with both 4-F-Phe and Phe added, wild-type GrsA allows less than 2% of fluorine incorporation into GS, which is in good agreement with the adenylation preference of GrsA-A. With mutant GrsA-W239S, the fraction of 4-F-Phe-GS increases to 50% (Fig. 4B). A similar biosynthetic preference is observed *in vitro* for incorporation of 2,4-F_2-Phe (Fig. 4C).

Encouraged by the successful production of fluorinated GS *in vitro*, we revisited the *in vivo* conditions, which initially yielded no detectable fluorine incorporation. *In vivo* conditions are

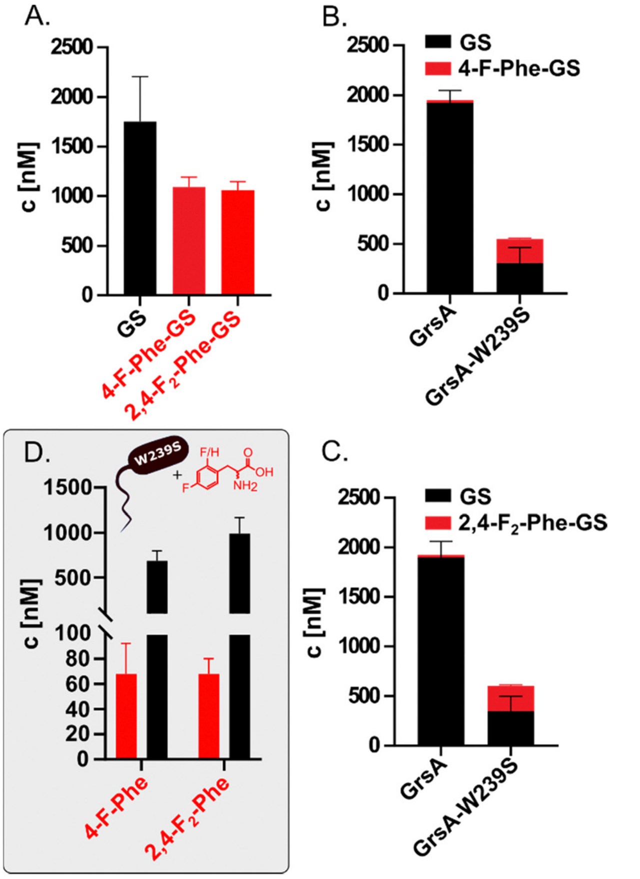


Fig. 4 Enhancement of GS formation with mutation W239S *in vitro* and *in vivo*. (A) *In vitro* production of GS and fluorinated GS analogs without competition using purified GrsA and GrsB. (B) *In vitro* production of GS and 4-F-GS under competitive conditions, using GrsA or GrsA-W239S. (C) *In vitro* production of GS and 2,4-F_2-GS under competitive conditions, using GrsA or GrsA-W239S. (D) *In vivo* production of GS and the fluorinated variants using a producer strain harboring the W239S mutation. The indicated fluorinated amino acids are supplemented to the growth medium. The error bars represent two biological replicates.

generally preferable because they do not rely on the tedious and low-yielding expression of the fragile NRPS proteins and are thus more easily scalable. To test whether mutation W239S would convey sufficient specificity for fluorinated Phe analogs *in vivo*, we generated the mutated plasmid pSU18-GrsTAB-W239S (Table S1, ESI†). Indeed, when 4 mM of 4-F-Phe or 2.5 mM 2,4-F_2-Phe were added to cultures of *E. coli* HM0079 carrying this plasmid, LC-MS/MS analysis revealed production of 68 nM 4-F-GS and 67 nM 2,4-F_2-GS, respectively. These concentrations correspond to volumetric yields of 0.079 mg L^{-1} for 4-F-GS and 0.081 mg L^{-1} for 2,4-F_2-GS from 3 mL cultures (Fig. 4D).

Conclusion

Hydrogen-fluorine exchange is one of the most subtle, yet powerful changes that can be introduced into a molecule. Our results show that nonribosomal A-domains can be unexpectedly sensitive to this change, which may prevent the incorporation of fluorinated amino acids into nonribosomal peptides. While the observed difference in $k_{\text{cat}}/K_{\text{M}}$ for Phe and 4-F-Phe in GrsA is 31-fold, the discrimination under *in vivo* conditions seems to be even stronger (Fig. 2B). Computational modelling indicates that fluorine-substitution in the 4-position of the Phe side-chain interrupts a crucial T-shaped aromatic interaction between Trp239 and the substrate. Perhaps, the poor acceptance of 4-F-Phe is compounded by poor uptake into the cell. With the wild-type GS NRPS, biosynthesis of 4-F-GS was only possible *in vitro*, where competition with the native substrate Phe is eliminated. However, for synthesis of nonribosomal peptides at scale, *in vitro* conditions are hardly viable since the NRPS proteins are not sufficiently high-yielding and robust.

For efficient *in vivo* biosynthesis of a fluorinated peptide analog, the NRPS must be able to fend off competition from the unfluorinated building block. Here we show how a single mutation within the binding pocket of GrsA-A, W239S, shifts the selectivity 43-fold in favor of 4-F-Phe. With this mutation, production of 4-fluorinated GS analogs becomes possible both under *in vitro* and *in vivo* conditions, although native GS remains the major product *in vivo*. A similar mutational effect was observed by Sirirungruang and coworkers, when they introduced mutation F190V into the *trans*-AT DszAT that they used for activation of fluoromalonyl-CoA.[18] Based on computational modelling, we speculate that the mutational effect is explained by the entropic contribution of ordered water molecules rather than specific binding interactions to the fluorine substituent, which were not found here or in a previous study investigating another protein–ligand interaction.[42] Importantly, the observed enhancement of the fluorine specificity in GrsA upon mutation augurs well for future projects aiming to further increase this specificity.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

We thank Prof. Donald Hilvert (ETH Zurich) for kindly providing pSU18-GrsA and Vanessa Rongen for technical assistance with kinetic measurements. We are thankful for the financial support of the Jena School for Microbial Communication (F. P.), Deutscher Akademischer Austauschdienst (F. P.), Konrad Adenauer Stiftung (O. N.), International Leibniz Research School (P. S.), the Daimler und Benz Stiftung (H. K.), the Fonds der Chemischen Industrie, and the Bundesministerium für Bildung und Forschung (ZIK Septomics, FKZ 03Z22JI1). This project has been supported by a Max-Buchner-Fellowship of the DECHEMA. Partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project ID 387284271 – SFB 1349 Fluorine-Specific Interactions and project ID 441781663.

References

1 K. Müller, C. Faeh and F. Diederich, *Science*, 2007, **317**, 1881–1886.

2 S. Purser, P. R. Moore, S. Swallow and V. Gouverneur, *Chem. Soc. Rev.*, 2008, **37**, 320–330.

3 T. Kaneko, T. J. Dougherty and T. V. Magee, in *Comprehensive Medicinal Chemistry II*, ed. J. B. Taylor and D. J. Triggle, Elsevier, Oxford, 2007, pp. 519–566.

4 A. A. Berger, J.-S. Völler, N. Budisa and B. Koksch, *Acc. Chem. Res.*, 2017, **50**, 2093–2103.

5 D. J. Newman and G. M. Cragg, *J. Nat. Prod.*, 2020, **83**, 770–803.

6 M. C. Walker and M. C. Y. Chang, *Chem. Soc. Rev.*, 2014, **43**, 6527–6536.

7 K. Rosenthal, M. Becker, J. Rolf, R. Siedentop, M. Hillen, M. Nett and S. Lütz, *ChemBioChem*, 2020, **21**, 3225–3228.

8 J. Rivera-Chávez, H. A. Raja, T. N. Graf, J. E. Burdette, C. J. Pearce and N. H. Oberlies, *J. Nat. Prod.*, 2017, **80**, 1883–1892.

9 S. Weist, B. Bister, O. Puk, D. Bischoff, S. Pelzer, G. J. Nicholson, W. Wohlleben, G. Jung and R. D. Süssmuth, *Angew. Chem., Int. Ed.*, 2002, **41**, 3383–3385.

10 W. Runguphan, J. J. Maresh and S. E. O'Connor, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**, 13673–13678.

11 N. K. O'Connor, D. K. Rai, B. R. Clark and C. D. Murphy, *J. Fluorine Chem.*, 2012, **143**, 210–215.

12 M. C. Walker, B. W. Thuronyi, L. K. Charkoudian, B. Lowry, C. Khosla and M. C. Y. Chang, *Science*, 2013, **341**, 1089–1094.

13 N. K. O'Connor, A. S. Hudson, S. L. Cobb, D. O'Neil, J. Robertson, V. Duncan and C. D. Murphy, *Amino Acids*, 2014, **46**, 2745–2752.

14 K. M. J. De Mattos-Shipley, C. Greco, D. M. Heard, G. Hough, N. P. Mulholland, J. L. Vincent, J. Micklefield, T. J. Simpson, C. L. Willis, R. J. Cox and A. M. Bailey, *Chem. Sci.*, 2018, **9**, 4109–4117.

15 C. S. M. Amrine, J. L. Long, H. A. Raja, S. J. Kurina, J. E. Burdette, C. J. Pearce and N. H. Oberlies, *J. Nat. Prod.*, 2019, **82**, 3104–3110.

16 A. Sester, K. Stüer-Patowsky, W. Hiller, F. Kloss, S. Lütz and M. Nett, *ChemBioChem*, 2020, **21**, 2268–2273.

17 A. Rittner, M. Joppe, J. J. Schmidt, L. M. Mayer, S. Reiners, E. Heid, D. Herzberg, D. H. Sherman and M. Grininger, *Nat. Chem.*, 2022, **14**(9), 1000–1006.

18 S. Sirirungruang, O. Ad, T. M. Privalsky, S. Ramesh, J. L. Sax, H. Dong, E. E. K. Baidoo, B. Amer, C. Khosla and M. C. Y. Chang, *Nat. Chem. Biol.*, 2022, **18**(8), 886–893.

19 M. S. Lichstrahl, L. Kahlert, R. Li, T. A. Zandi, J. Yang and C. A. Townsend, *Chem. Sci.*, 2023, **14**(14), 3923–3931.

20 L. J. Stevenson, J. G. Owen and D. F. Ackerley, *ACS Chem. Biol.*, 2019, **14**, 2115–2126.

21 R. D. Süssmuth and A. Mainz, *Angew. Chem., Int. Ed.*, 2017, **56**, 3770–3821.

22 Y. Liu, S. Ding, J. Shen and K. Zhu, *Nat. Prod. Rep.*, 2019, **36**, 573–592.

23 K. A. Bozhüyük, J. Micklefield and B. Wilkinson, *Curr. Opin. Microbiol.*, 2019, **51**, 88–96.

24 J. M. Reimer, A. S. Haque, M. J. Tarry and T. M. Schmeing, *Curr. Opin. Struct. Biol.*, 2018, **49**, 104–113.

25 G. L. Challis, J. Ravel and C. A. Townsend, *Chem. Biol.*, 2000, **7**, 211–224.

26 T. Stachelhaus, H. D. Mootz and M. A. Marahiel, *Chem. Biol.*, 1999, **6**, 493–505.

27 A. Stanišić and H. Kries, *ChemBioChem*, 2019, **20**, 1347–1356.

28 M. H. Medema, K. Blin, P. Cimermancic, V. De Jager, P. Zakrzewski, M. A. Fischbach, T. Weber, E. Takano and R. Breitling, *Nucleic Acids Res.*, 2011, **39**(suppl_2), W339–W346.

29 A. Stanišić, C.-M. Svensson, U. Ettelt and H. Kries, *bioRxiv*, 2022, preprint, DOI: 10.1101/2022.08.30.505883.

30 J. Krätzschmar, M. Krause and M. A. Marahiel, *J. Bacteriol.*, 1989, **171**, 5422–5429.

31 H. Kries, R. Wachtel, A. Pabst, B. Wanner, D. Niquille and D. Hilvert, *Angew. Chem., Int. Ed.*, 2014, **53**, 10105–10108.

32 S. Gruenewald, H. D. Mootz, P. Stehmeier and T. Stachelhaus, *Appl. Environ. Microbiol.*, 2004, **70**, 3282–3291.

33 F. Pourmasoumi, S. De, H. Peng, F. Trottmann, C. Hertweck and H. Kries, *ACS Chem. Biol.*, 2022, **17**, 2382–2388.

34 D. J. Wilson and C. C. Aldrich, *Anal. Biochem.*, 2010, **404**, 56–63.

35 A. Koglin, F. Löhr, F. Bernhard, V. V. Rogov, D. P. Frueh, E. R. Strieter, M. R. Mofid, P. Güntert, G. Wagner, C. T. Walsh, M. A. Marahiel and V. Dötsch, *Nature*, 2008, **454**, 907–911.

36 A. Stanišić, A. Hüsken and H. Kries, *Chem. Sci.*, 2019, **10**, 10395–10399.

37 E. Conti, T. Stachelhaus, M. A. Marahiel and P. Brick, *EMBO J.*, 1997, **16**, 4174–4183.

38 B. Breiten, M. R. Lockett, W. Sherman, S. Fujita, M. Al-Sayah, H. Lange, C. M. Bowers, A. Heroux, G. Krilov and G. M. Whitesides, *J. Am. Chem. Soc.*, 2013, **135**, 15579–15584.

39 D. L. Niquille, I. B. Folger, S. Basler and D. Hilvert, *J. Am. Chem. Soc.*, 2021, **143**, 2736–2740.

40 C. R. Martinez and B. L. Iverson, *Chem. Sci.*, 2012, **3**, 2191.

41 D. L. Niquille, D. A. Hansen, T. Mori, D. Fercher, H. Kries and D. Hilvert, *Nat. Chem.*, 2017, **10**, 282–287.

42 L. Wehrhan, J. Leppkes, N. Dimos, B. Loll, B. Koksch and B. G. Keller, *J. Phys. Chem. B*, 2022, **126**, 9985–9999.

5 Conclusions and Outlook

Fluorine is a unique element with the remarkable ability to change the physico-chemical properties of a molecule through seemingly subtle substitutions. This has lead to advancements in the field of medicinal chemistry and drug discovery, where fluorine is part of multiple drug molecules, already in the market or in development. Fluorine can enhance the metabolic stability of such drug molecules, but it can also modulate the binding affinity of the drug molecule to its protein target. The influence that fluorine has on the binding affinity is often difficult to predict by experimental methods only, meaning that insights from computational methods such as MD simulations are highly valuable for understanding in which way fluorine can be used to enhance key molecular properties and protein binding affinity.

In this thesis, I use analyses based on MD simulations on selected systems of proteins in complex with fluorinated amino acids. The main project was concerned with complexes of trypsin with variants of the natural inhibitor BPTI, whose crucial amino acid Lys15 is replaced with the shorter and aliphatic Abu and its fluorinated variants MfeGly, DfeGly and TfeGly. Following the hypothesis that fluorine is able to restore inhibitor strength of Abu-BPTI by binding to water molecules inside the main binding pocket of trypsin, I analyzed the mobility and interactions of the water molecules in the binding pocket and how these react to the introduction of fluorine. The water molecules are highly mobile and form a tight hydrogen bond network. Fluorination of Abu-BPTI does not lead to a stepwise decrease in mobility, nor to a change in the hydrogen bond network in a stepwise and systematic way. Moreover, fluorine does not interact with the water molecules vie hydrogen bond like interactions, making it unlikely that fluorine can restore the inhibitor activity this way.

The PMF profile of the unbinding process of the Abu-BPTI variants confirms the gain in binding strength, observed in experimental inhibition assays and shows a rugged unbinding path, hinting at multiple metastable states along this path.

Following up on these results, I investigated the unbinding path of the BPTI variants from trypsin using enhanced sampling methods, such as metadynamics and RAMD. In RAMD trajectories, I discovered a new metastable state in the unbinding pathway of all of the BPTI variants, which we call the pre-bound state. The pre-bound state is positionally and structurally clearly distinct from the fully bound state, but a large part of the binding interface, including the region around the active site, is still intact. This implies that the pre-bound state is likely inhibitory. The transition between the fully bound state and the pre-bound state is characterized by changes in hydrogen bond and cation-pi interaction patterns. Umbrella sampling simulations, based on one of the newly identified broken hydrogen bonds of the pre-bound state, revealed an influence of fluorine on the energetic minima and transition barriers of the pre-bound state. A speculative, yet logical, explanation is that fluorine supports the formation of the pre-bound state by interacting with the side chain of trypsin's Gln194.

Moving on to the impact of fluorine on other protein systems, I characterized the preferred binding pose and key interactions of a newly developed phosphotyrosine mimeticum with a PF5 moiety as its headgroup using a combination of molecular docking and MD simulations. Among the key binding interactions are hydrogen bond like interactions between amine groups of the protein and the highly fluorinated PF5 moiety, which remain stable in MD simulations. Given the higher binding affinity of the mimeticum with the PF5 headgroup over that with the phosphate headgroup despite the more negative charge, it is likely that the hydrogen bond like interactions to fluorine contribute significantly to the binding affinity. In yet another protein system, GrsA, a single fluorination at the *para*-position can lead the GrsA A domain to reject the amino acid Phe as a substrate. Molecular modeling shows that this is caused by disrupting a crucial aromatic interaction. The selectivity of the GrsA A domain for Phe over *para*-fluorinated Phe is reversed in the W239S mutant of GrsA. MD simulations reveal that while the big-to-small mutation of tryptophane to serine leaves a water-filled cavity next to the fluorine atom, there are no hydrogen bond like interactions between fluorine and the water molecules. The selectivity for the fluorinated substrate may be explained by a better fit in the cavity or a better accommodation of fluorine in the aqueous environment.

Considering the overarching research question of how fluorine influences the interactions of proteins, I draw the following conclusions from the results of specific protein systems summarized above. (1) The study of the protein systems described here in this thesis does not provide evidence that direct hydrogen bond like interactions between water and fluorine lead to increases in binding affinity. This becomes apparent by analyzing the water network and interactions of fluorinated BPTI-trypsin complexes and is supported by the absence of direct interactions between fluorine and water in W239S GrsA complexes. (2) Fluorine may act on pre-bound intermediates of protein-protein complexes. The existence of a pre-bound state in the unbinding path of BPTI variants from trypsin, which is likely to be inhibitory and affected by fluorine in its energetic minimum and transition barrier, may lead to the conclusion that the effect of fluorine on the inhibitor activity is not only caused by acting on the fully bound state only, but on an equilibrium between the fully bound state and the pre-bound state. (3) Direct interactions between fluorine and proteins can significantly alter the binding properties. This is demonstrated by the highly fluorinated PF5 headgroup, which is likely to significantly increase binding affinity by utilizing direct interactions between fluorine and the protein. In the case of GrsA A domain in complex with 4-F-Phe, an aromatic interaction is disrupted by fluorine substitution at one precise position.

The results from this thesis raise interesting questions for future research. As binding affinities in the case of fluorinated BPTI variants and trypsin are likely not affected by direct hydrogen bond like interactions between fluorine and water, it will be interesting to research other effects that fluorine might have on the binding affinities. Considering the sizable effects that fluorine can have on entropic contributions, as described in the introduction of this thesis, the entropy of hydration sites and also the conformational entropy of fluorinated proteins are an especially interesting research target. Moreover, given that fluorine influences the energetic minima and transition barriers of the pre-bound state, not only the impact of fluorine on fully bound protein complexes should be analyzed, but also how fluorine interacts with inhibitory pre-bound intermediates. Given the occurrence of encounter states, it is important to characterize these states as well and evaluate if they may be already inhibitory and how they are impacted by the presence of fluorine. Moving on to the fluorine specific interactions between the PF5 headgroup and the PTP1B binding pocket, it will be intriguing to see how this specific moiety may be applied to other protein systems, like related tyrosine phosphatases. Lastly, the precise disruption of a specific aromatic interaction by fluorine substitution may provide a tool in efforts to manipulate substrate selectivity of other protein systems.

All together, the effects of fluorine on protein-protein interactions remain a challenging, yet exciting, topic. Analysis based on computational methods, and especially MD simulations,[121] are an excellent tool to study these effects and will surely contribute to the understanding of fluorine's role in protein-protein interactions in the future.

Bibliography

[1] E. Riedel and H.-J. Meyer. *Allgemeine und Anorganische Chemie*. De Gruyter, 11. edition, **2013**.

[2] S. Purser, P. R. Moore, S. Swallow and V. Gouverneur. “Fluorine in Medicinal Chemistry”. *Chem. Soc. Rev.* **2008** *37*, 320–330.

[3] H.-J. Böhm, D. Banner, S. Bendels, M. Kansy, B. Kuhn, K. Müller, U. Obst-Sander and M. Stahl. “Fluorine in Medicinal Chemistry”. *ChemBioChem* **2004** *5*, 637–643.

[4] M. A. Miller and E. M. Sletten. “Perfluorocarbons in Chemical Biology”. *ChemBioChem* **2020** *21*, 3451–3462.

[5] E. P. Gillis, K. J. Eastman, M. D. Hill, D. J. Donnelly and N. A. Meanwell. “Applications of Fluorine in Medicinal Chemistry”. *J. Med. Chem.* **2015** *58*, 8315–8359.

[6] M. Inoue, Y. Sumii and N. Shibata. “Contribution of Organofluorine Compounds to Pharmaceuticals”. *ACS Omega* **2020** *5*, 10633–10640.

[7] D. O’Hagan and R. J. Young. “Future Challenges and Opportunities with Fluorine in Drugs?” *Med. Chem. Res.* **2023** *32*, 1231–1234.

[8] K. Ahmad, A. Rizzi, R. Capelli, D. Mandelli, W. Lyu and P. Carloni. “Enhanced-Sampling Simulations for the Estimation of Ligand Binding Kinetics: Current Status and Perspective”. *Front. Mol. Biosci.* **2022** *9*, 899805.

[9] S. Decherchi and A. Cavalli. “Thermodynamics and Kinetics of Drug-Target Binding by Molecular Simulation”. *Chem. Rev.* **2020** *120*, 12788–12833.

[10] R. Lazim, D. Suh and S. Choi. “Advances in Molecular Dynamics Simulations and Enhanced Sampling Methods for the Study of Protein Systems”. *Int. J. Mol. Sci.* **2020** *21*, 6339.

[11] O. M. H. Salo-Ahen, I. Alanko, R. Bhadane et al. “Molecular Dynamics Simulations in Drug Discovery and Pharmaceutical Development”. *Processes* **2021** *9*, 71.

[12] B. Roux. *Computational Modeling and Simulations of Biomolecular Systems*. World Scientific, 1. edition, **2022**.

[13] D. Frenkel and B. Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. Academic Press, 3. edition, **2023**.

[14] J. A. McCammon, B. R. Gelin and M. Karplus. “Dynamics of Folded Proteins”. *Nature* **1977** *267*, 585–590.

[15] D. E. Shaw, P. J. Adams, A. Azaria et al. “Anton 3: Twenty Microseconds of Molecular Dynamics Simulation before Lunch”. *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis* **2021** (Conference Paper).

[16] J. Hénin, T. Lelièvre, M. R. Shirts, O. Valsson and L. Delemotte. “Enhanced Sampling Methods for Molecular Dynamics Simulations”. *LiveCoMS* **2022** *4*.

[17] G. Bitencourt-Ferreira, M. Veit-Acosta and W. F. de Azevedo. “Hydrogen Bonds in Protein-Ligand Complexes”. In “Docking Screens for Drug Discovery”, Springer, **2019** .

[18] C. Dalvit, C. Invernizzi and A. Vulpetti. “Fluorine as a Hydrogen-Bond Acceptor: Experimental Evidence and Computational Calculations”. *Chem. Eur. J.* **2014** *20*, 11058–11068.

[19] J. S. Murray, P. G. Seybold and P. Politzer. “The Many Faces of Fluorine: Some Noncovalent Interactions of Fluorine Compounds”. *J. Chem. Thermodyn.* **2021** *156*, 106382.

[20] P. Kollman, J. McKelvey, A. Johansson and S. Rothenberg. “Theoretical Studies of Hydrogen-Bonded Dimers. Complexes Involving HF, H2O, NH3, CH1, H2S, PH3, HCN, HNC, HCP, CH2NH, H2CS, H2CO, CH4, CF3,H, C2H2, C2H4, C6H6, F- and H3O+”. *J. Am. Chem. Soc.* **1975** *97*, 955–965.

[21] S. Roehrig, J. Ackerstaff, E. Jiménez Núñez et al. “Design and Preclinical Characterization Program toward Asundexian (BAY 2433334), an Oral Factor XIa Inhibitor for the Prevention and Treatment of Thromboembolic Disorders”. *J. Med. Chem.* **2023** *66*, 12203–12224.

[22] S. Lešnik, U. Bren, T. Domratcheva and A.-N. Bondar. “Fentanyl and the Fluorinated Fentanyl Derivative NFEPP Elicit Distinct Hydrogen-Bond Dynamics of the Opioid Receptor”. *J. Chem. Inf. Model.* **2023** *63*, 4732–4748.

[23] L. Yu, L. He, B. Gan, R. Ti, Q. Xiao, X. Yang, H. Hu, L. Zhu, S. Wang and R. Ren.

"Structural Insights into Sphingosine-1-Phosphate Receptor Activation". *Proc. Natl. Acad. Sci.* **2022** *119*, e2117716119.

[24] Y. Li, D. Zhang, X. Gao, X. Wang and L. Zhang. "2′- and 3′-Ribose Modifications of Nucleotide Analogues Establish the Structural Basis to Inhibit the Viral Replication of SARS-CoV-2". *J. Phys. Chem. Lett.* **2022** *13*, 4111–4118.

[25] A. K. Ghosh, S. Kovela, A. Sharma et al. "Design, Synthesis and X-Ray Structural Studies of Potent HIV-1 Protease Inhibitors Containing C-4 Substituted Tricyclic Hexahydro-Furofuran Derivatives as P2 Ligands". *ChemMedChem* **2022** *17*, e202200058.

[26] S. Mirabile, S. Vittorio, M. Paola Germanò, I. Adornato, L. Ielo, A. Rapisarda, R. Gitto, F. Pintus, A. Fais and L. De Luca. "Evaluation of 4-(4-Fluorobenzyl)Piperazin-1-Yl]-Based Compounds as Competitive Tyrosinase Inhibitors Endowed with Antimelanogenic Effects". *ChemMedChem* **2021** *16*, 3083–3093.

[27] T. Kim, K. Kim, I. Park, S. Hong and H. Park. "Two-Track Virtual Screening Approach to Identify the Dual Inhibitors of Wild Type and C481S Mutant of Bruton's Tyrosine Kinase". *J. Chem. Inf. Model.* **2022** *62*, 4500–4511.

[28] A. A. Ojha, A. Srivastava, L. W. Votapka and R. E. Amaro. "Selectivity and Ranking of Tight-Binding JAK-STAT Inhibitors Using Markovian Milestoning with Voronoi Tessellations". *J. Chem. Inf. Model.* **2023** *63*, 2469–2482.

[29] A. Thakur, G. Sharma, V. N. Badavath, V. Jayaprakash, K. M. J. Merz, G. Blum and O. Acevedo. "Primer for Designing Main Protease (Mpro) Inhibitors of SARS-CoV-2". *J. Phys. Chem. Lett.* **2022** *13*, 5776–5786.

[30] S. Mondal, Y. Chen, G. J. Lockbaum et al. "Dual Inhibitors of Main Protease (MPro) and Cathepsin L as Potent Antivirals against SARS-CoV2". *J. Am. Chem. Soc.* **2022** *144*, 21035–21045.

[31] L. E. Khoury, Z. Jing, A. Cuzzolin et al. "Computationally Driven Discovery of SARS-CoV-2 Mpro Inhibitors: From Design to Experimental Validation". *Chem. Sci.* **2022** *13*, 3674–3687.

[32] W. Pietruś, R. Kafel, A. J. Bojarski and R. Kurczab. "Hydrogen Bonds with Fluorine in Ligand–Protein Complexes-the PDB Analysis and Energy Calculations". *Molecules* **2022** *27*, 1005.

[33] S. Mukherjee and L. V. Schäfer. "Spatially Resolved Hydration Thermodynamics in Biomolecular Systems". *J. Phys. Chem. B* **2022** *126*, 3619–3631.

[34] J. R. Robalo and A. Vila Verde. “Unexpected Trends in the Hydrophobicity of Fluorinated Amino Acids Reflect Competing Changes in Polarity and Conformation”. *Phys. Chem. Chem. Phys.* **2019** *21*, 2029–2038.

[35] J. R. Robalo, S. Huhmann, B. Koksch and A. Vila Verde. “The Multiple Origins of the Hydrophobicity of Fluorinated Apolar Amino Acids”. *Chem* **2017** *3*, 881–897.

[36] J. R. Robalo, D. M. Oliveira, P. Imhof, D. Ben-Amotz and A. Vila Verde. “Quantifying How Step-Wise Fluorination Tunes Local Solute Hydrophobicity, Hydration Shell Thermodynamics and the Quantum Mechanical Contributions of Solute–Water Interactions”. *Phys. Chem. Chem. Phys.* **2020** *22*, 22997–23008.

[37] O.-H. Kwon, T. H. Yoo, C. M. Othon, J. A. Van Deventer, D. A. Tirrell and A. H. Zewail. “Hydration Dynamics at Fluorinated Protein Surfaces”. *Proc. Natl. Acad. Sci.* **2010** *107*, 17101–17106.

[38] B. Breiten, M. R. Lockett, W. Sherman, S. Fujita, M. Al-Sayah, H. Lange, C. M. Bowers, A. Heroux, G. Krilov and G. M. Whitesides. “Water Networks Contribute to Enthalpy/Entropy Compensation in Protein–Ligand Binding”. *J. Am. Chem. Soc.* **2013** *135*, 15579–15584.

[39] J. Wallerstein, V. Ekberg, M. M. Ignjatović et al. “Entropy–Entropy Compensation between the Protein, Ligand, and Solvent Degrees of Freedom Fine-Tunes Affinity in Ligand Binding to Galectin-3C”. *J. Am. Chem. Soc. Au* **2021** *1*, 484–500.

[40] B. Turk. “Targeting Proteases: Successes, Failures and Future Prospects”. *Nat. Rev. Drug Discov.* **2006** *5*, 785–799.

[41] M. Drag and G. S. Salvesen. “Emerging Principles in Protease-Based Drug Discovery”. *Nat. Rev. Drug Discov.* **2010** *9*, 690–701.

[42] A. R. Batt, C. P. St. Germain, T. Gokey, A. B. Guliaev and T. Baird Jr. “Engineering Trypsin for Inhibitor Resistance”. *Prot. Sci.* **2015** *24*, 1463–1474.

[43] M. J. Page and E. D. Cera. “Evolution of Peptidase Diversity”. *J. Biol. Chem.* **2008** *283*, 30010–30014.

[44] M. Marquart, J. Walter, J. Deisenhofer, W. Bode and R. Huber. “The Geometry of the Reactive Site and of the Peptide Groups in Trypsin, Trypsinogen and Its Complexes with Inhibitors”. *Acta. Cryst. B* **1983** *39*, 480–490.

[45] I. Schechter. “Mapping of the Active Site of Proteases in the 1960s and Rational Design of Inhibitors/Drugs in the 1990s”. *Curr. Protein Pep. Sci.* **2005** *6*, 501–512.

[46] M. Peräkylä and P. A. Kollman. “Why Does Trypsin Cleave BPTI so Slowly?” *J. Am. Chem. Soc.* **2000** *122*, 3436–3444.

[47] E. S. Radisky and D. E. Koshland. “A Clogged Gutter Mechanism for Protease Inhibitors”. *Proc. Natl. Acad. Sci.* **2002** *99*, 10316–10321.

[48] K. Kawamura, T. Yamada, K. Kurihara, T. Tamada, R. Kuroki, I. Tanaka, H. Takahashi and N. Niimura. “X-Ray and Neutron Protein Crystallographic Analysis of the Trypsin–BPTI Complex”. *Acta. Cryst. D* **2011** *67*, 140–148.

[49] M. J. M. Castro and S. Anderson. “Alanine Point-Mutations in the Reactive Region of Bovine Pancreatic Trypsin Inhibitor: Effects on the Kinetics and Thermodynamics of Binding to β-Trypsin and α-Chymotrypsin”. *Biochemistry* **1996** *35*, 11435–11446.

[50] U. Kahler, A. S. Kamenik, F. Waibl, J. Kraml and K. R. Liedl. “Protein-Protein Binding as a Two-Step Mechanism: Preselection of Encounter Poses during the Binding of BPTI and Trypsin”. *Biophys. J.* **2020** *119*, 652–666.

[51] S. Ye, B. Loll, A. Ann Berger, U. Mülow, C. Alings, M. Christian Wahl and B. Koksch. “Fluorine Teams up with Water to Restore Inhibitor Activity to Mutant BPTI”. *Chem. Sci.* **2015** *6*, 5246–5254.

[52] L. Verlet. “Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules”. *Phys. Rev.* **1967** *159*, 98–103.

[53] R. W. Hockney, S. P. Goel and J. W. Eastwood. “Quiet High-Resolution Computer Models of a Plasma”. *J. Comput. Phys.* **1974** *14*, 148–158.

[54] A. R. Leach. *Molecular Modelling: Principles and Applications*. Pearson Education, 2. edition, **2001**.

[55] B. Hess, H. Bekker, H. J. C. Berendsen and J. G. E. M. Fraaije. “LINCS: A Linear Constraint Solver for Molecular Simulations”. *J. Comput. Chem.* **1997** *18*, 1463–1472.

[56] G. Bussi, D. Donadio and M. Parrinello. “Canonical Sampling through Velocity Rescaling”. *J. Chem. Phys.* **2007** *126*, 014101.

[57] H. J. Berendsen, J. v. Postma, W. F. Van Gunsteren, A. DiNola and J. R. Haak. “Molecular dynamics with coupling to an external bath”. *J. Chem. Phys.* **1984** *81*, 3684–3690.

[58] M. Parrinello and A. Rahman. “Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method”. *J. Appl. Phys.* **1981** *52*, 7182–7190.

[59] N. Awtrey. “Parallelizing DensityAnalysis and RMSF in PMDA”. *SPIDAL Summer REU* **2019** (Technical Paper).

[60] R. C. Bernardi, M. C. R. Melo and K. Schulten. “Enhanced Sampling Techniques in Molecular Dynamics Simulations of Biological Systems”. *Biochim. Biophys. Acta* **2015** *1850*, 872–877.

[61] J. N. Onuchic, Z. Luthey-Schulten and P. G. Wolynes. “THEORY OF PROTEIN FOLDING: The Energy Landscape Perspective”. *Annu. Rev. Phys. Chem.* **1997** *48*, 545–600.

[62] Y. I. Yang, Q. Shao, J. Zhang, L. Yang and Y. Q. Gao. “Enhanced Sampling in Molecular Dynamics”. *J. Chem. Phys.* **2019** *151*, 070902.

[63] O. Valsson, P. Tiwary and M. Parrinello. “Enhancing Important Fluctuations: Rare Events and Metadynamics from a Conceptual Viewpoint”. *Annu. Rev. Phys. Chem.* **2016** *67*, 159–184.

[64] H. Sidky, W. Chen and A. L. Ferguson. “Machine Learning for Collective Variable Discovery and Enhanced Sampling in Biomolecular Simulation”. *Mol. Phys.* **2020** *118*, e1737742.

[65] G. M. Torrie and J. P. Valleau. “Modeling Condensed Phase Reaction Dynamics”. *Chem. Phys. Lett.* **1974** *28*, 578–581.

[66] B. Roux. “The Calculation of the Potential of Mean Force Using Computer Simulations”. *Comput. Phys. Commun.* **1995** *91*, 275–282.

[67] J. Kästner. “Umbrella Sampling”. *WIREs Comput. Mol. Sci.* **2011** *1*, 932–942.

[68] S. Izrailev, S. Stepaniants, B. Isralewitz, D. Kosztin, H. Lu, F. Molnar, W. Wriggers and K. Schulten. “Steered Molecular Dynamics”. In P. Deuflhard, J. Hermans, B. Leimkuhler, A. E. Mark, S. Reich and R. D. Skeel, editors, “Computational Molecular Dynamics: Challenges, Methods, Ideas”, Springer Berlin Heidelberg, **1999** 39–65.

[69] J. G. Kirkwood. “Statistical Mechanics of Fluid Mixtures”. *J. Chem. Phys.* **1935** *3*, 300–313.

[70] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen and P. A. Kollman. “THE Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method”. *J. Comput. Chem.* **1992** *13*, 1011–1021.

[71] A. Laio and M. Parrinello. “Escaping Free-Energy Minima”. *Proc. Natl. Acad. Sci.* **2002** *99*, 12562–12566.

[72] T. Huber, A. E. Torda and W. F. van Gunsteren. “Local Elevation: A Method for Improving the Searching Properties of Molecular Dynamics Simulation”. *J. Comput. Aided Mol. Des.* **1994** *8*, 695–708.

[73] H. Grubmüller. “Predicting Slow Structural Transitions in Macromolecular Systems: Conformational Flooding”. *Phys. Rev. E* **1995** *52*, 2893–2906.

[74] A. F. Voter. “Hyperdynamics: Accelerated Molecular Dynamics of Infrequent Events”. *Phys. Rev. Lett.* **1997** *78*, 3908–3911.

[75] A. Barducci, M. Bonomi and M. Parrinello. “Metadynamics”. *WIREs Comput. Mol. Sci.* **2011** *1*, 826–843.

[76] G. Bussi and A. Laio. “Using Metadynamics to Explore Complex Free-Energy Landscapes”. *Nat. Rev. Phys.* **2020** *2*, 200–212.

[77] A. Barducci, G. Bussi and M. Parrinello. “Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method”. *Phys. Rev. Lett.* **2008** *100*, 020603.

[78] S. K. Lüdemann, V. Lounnas and R. C. Wade. “How Do Substrates Enter and Products Exit the Buried Active Site of Cytochrome P450cam? 1. Random Expulsion Molecular Dynamics Investigation of Ligand Access Channels and mechanisms11Edited by J. Thornton”. *J. Mol. Biol.* **2000** *303*, 797–811.

[79] D. B. Kokh, M. Amaral, J. Bomke et al. “Estimation of Drug-Target Residence Times by τ-Random Acceleration Molecular Dynamics Simulations”. *J. Chem. Theory Comput.* **2018** *14*, 3859–3869.

[80] D. B. Kokh, B. Doser, S. Richter, F. Ormersbach, X. Cheng and R. C. Wade. “A Workflow for Exploring Ligand Dissociation from a Macromolecule: Efficient Random Acceleration Molecular Dynamics Simulation and Interaction Fingerprint Analysis of Ligand Trajectories”. *J. Chem. Phys.* **2020** *153*, 125102.

[81] J. Huang, P. E. M. Lopes, B. Roux and A. D. J. MacKerell. “Recent Advances in Polarizable Force Fields for Macromolecules: Microsecond Simulations of Proteins Using the Classical Drude Oscillator Model”. *J. Phys. Chem. Lett.* **2014** *5*, 3144–3150.

[82] P. S. Nerenberg and T. Head-Gordon. “New Developments in Force Fields for Biomolecular Simulations”. *Curr. Opin. Struct. Biol.* **2018** *49*, 129–138.

[83] S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta and P. Weiner. “A New Force Field for Molecular Mechanical Simulation of Nucleic Acids and Proteins”. *J. Am. Chem. Soc.* **1984** *106*, 765–784.

[84] S. J. Weiner, P. A. Kollman, D. T. Nguyen and D. A. Case. “An all atom force field for simulations of proteins and nucleic acids”. *J. Comput. Chem.* **1986** *7*, 230–252.

[85] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell and P. A. Kollman. “A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules”. *J. Am. Chem. Soc.* **1995** *117*, 2309–2309.

[86] J. Wang, P. Cieplak and P. A. Kollman. “How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?” *J. Comput. Chem.* **2000** *21*, 1049–1074.

[87] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg and C. Simmerling. “Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters”. *Proteins: Struct., Funct., Bioinf.* **2006** *65*, 712–725.

[88] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser and C. Simmerling. “ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB”. *J. Chem. Theory Comput.* **2015** *11*, 3696–3713.

[89] C. I. Bayly, P. Cieplak, W. Cornell and P. A. Kollman. “A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model”. *J. Phys. Chem.* **1993** *97*, 10269–10280.

[90] U. C. Singh and P. A. Kollman. “An Approach to Computing Electrostatic Charges for Molecules”. *J. Comput. Chem.* **1984** *5*, 129–145.

[91] B. H. Besler, K. M. Merz Jr. and P. A. Kollman. “Atomic Charges Derived from Semiempirical Methods”. *JJ. Comput. Chem.* **1990** *11*, 431–439.

[92] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein. “Comparison of Simple Potential Functions for Simulating Liquid Water”. *J. Chem. Phys.* **1983** *79*, 926–935.

[93] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case. “Development and Testing of a General Amber Force Field”. *J. Comput. Chem.* **2004** *25*, 1157–1174.

[94] A. Jakalian, B. L. Bush, D. B. Jack and C. I. Bayly. “Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method”. *J. Comput. Chem.* **2000** *21*, 132–146.

[95] A. Jakalian, D. B. Jack and C. I. Bayly. “Fast, efficient generation of high-quality

atomic charges. AM1-BCC model: II. Parameterization and validation". *J. Comput. Chem.* **2002** *23*, 1623–1641.

[96] X. He, V. H. Man, W. Yang, T.-S. Lee and J. Wang. "A Fast and High-Quality Charge Model for the next Generation General AMBER Force Field". *J. Chem. Phys.* **2020** *153*, 114502.

[97] Z. Zhu, Z. Xu and W. Zhu. "Interaction Nature and Computational Methods for Halogen Bonding: A Perspective". *J. Chem. Inf. Model.* **2020** *60*, 2683–2696.

[98] R. S. Czarny, A. N. Ho and P. Shing Ho. "A Biological Take on Halogen Bonding and Other Non-Classical Non-Covalent Interactions". *Chem. Rec.* **2021** *21*, 1240–1251.

[99] T. Clark, M. Hennemann, J. S. Murray and P. Politzer. "Halogen Bonding: The Sigma-Hole". *J. Mol. Model.* **2007** *13*, 291–296.

[100] M. A. A. Ibrahim. "AMBER Empirical Potential Describes the Geometry and Energy of Noncovalent Halogen Interactions Better than Advanced Semiempirical Quantum Mechanical Method PM6-DH2X". *J. Phys. Chem. B* **2012** *116*, 3659–3669.

[101] D. Franchini, F. Dapiaggi, S. Pieraccini, A. Forni and M. Sironi. "Halogen Bonding in the Framework of Classical Force Fields: The Case of Chlorine". *Chem. Phys. Lett.* **2018** *712*, 89–94.

[102] M. Carter, A. K. Rappé and P. S. Ho. "Scalable Anisotropic Shape and Electrostatic Models for Biological Bromine Halogen Bonds". *J. Chem. Theory Comput.* **2012** *8*, 2461–2473.

[103] E. A. Orabi and J. D. Faraldo-Gómez. "New Molecular-Mechanics Model for Simulations of Hydrogen Fluoride in Chemistry and Biology". *J. Chem. Theory Comput.* **2020** *16*, 5105–5126.

[104] K. Saito and H. Torii. "Hidden Halogen-Bonding Ability of Fluorine Manifesting in the Hydrogen-Bond Configurations of Hydrogen Fluoride". *J. Phys. Chem. B* **2021** *125*, 11742–11750.

[105] P. Politzer, J. S. Murray and T. Clark. "Halogen Bonding: An Electrostatically-Driven Highly Directional Noncovalent Interaction". *Phys. Chem. Chem. Phys.* **2010** *12*, 7748–7757.

[106] A. Croitoru, S.-J. Park, A. Kumar, J. Lee, W. Im, A. D. J. MacKerell and A. Aleksandrov. "Additive CHARMM36 Force Field for Nonstandard Amino Acids". *J. Chem. Theory Comput.* **2021** *17*, 3554–3570.

[107] X. Wang and W. Li. “Development and Testing of Force Field Parameters for Phenylalanine and Tyrosine Derivatives”. *Front. Mol. Biosci.* **2020** *7*, 608931.

[108] M. L. Samways, R. D. Taylor, H. E. B. Macdonald and J. W. Essex. “Water Molecules at Protein–Drug Interfaces: Computational Prediction and Analysis Methods”. *Chem. Soc. Rev.* **2021** *50*, 9104–9120.

[109] F. Rodier, R. P. Bahadur, P. Chakrabarti and J. Janin. “Hydration of Protein–Protein Interfaces”. *Proteins: Struct., Funct., Bioinf.* **2005** *60*, 36–45.

[110] P. Wernet, D. Nordlund, U. Bergmann et al. “The Structure of the First Coordination Shell in Liquid Water”. *Science* **2004** *304*, 995–999.

[111] D. Laage, T. Elsaesser and J. T. Hynes. “Water Dynamics in the Hydration Shells of Biomolecules”. *Chem. Rev.* **2017** *117*, 10694–10725.

[112] T. Hüfner-Wulsdorf and G. Klebe. “Role of Water Molecules in Protein–Ligand Dissociation and Selectivity Discrimination: Analysis of the Mechanisms and Kinetics of Biomolecular Solvation Using Molecular Dynamics”. *J. Chem. Inf. Model.* **2020** *60*, 1818–1832.

[113] T. Lazaridis. “Inhomogeneous Fluid Approach to Solvation Thermodynamics. 1. Theory”. *J. Phys. Chem. B* **1998** *102*, 3531–3541.

[114] C. N. Nguyen, T. Kurtzman Young and M. K. Gilson. “Grid Inhomogeneous Solvation Theory: Hydration Structure and Thermodynamics of the Miniature Receptor Cucurbit[7]Uril”. *J. Chem. Phys.* **2012** *137*, 044101.

[115] C. N. Nguyen, A. Cruz, M. K. Gilson and T. Kurtzman. “Thermodynamics of Water in an Enzyme Active Site: Grid-Based Hydration Analysis of Coagulation Factor Xa”. *J. Chem. Theory Comput.* **2014** *10*, 2769–2780.

[116] J. Fan, A. Fu and L. Zhang. “Progress in Molecular Docking”. *Quant. Biol.* **2019** *7*, 83–89.

[117] L. Pinzi and G. Rastelli. “Molecular Docking: Shifting Paradigms in Drug Discovery”. *Int. J. Mol. Sci.* **2019** *20*, 4331.

[118] R. A. Friesner, J. L. Banks, R. B. Murphy et al. “Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy”. *J. Med. Chem.* **2004** *47*, 1739–1749.

[119] R. A. Friesner, R. B. Murphy, M. P. Repasky, L. L. Frye, J. R. Greenwood, T. A. Halgren, P. C. Sanschagrin and D. T. Mainz. “Extra Precision Glide: Docking

and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein-Ligand Complexes". *J. Med. Chem.* **2006** *49*, 6177–6196.

[120] L. Wehrhan, J. Leppkes, N. Dimos, B. Loll, B. Koksch and B. G. Keller. "Water Network in the Binding Pocket of Fluorinated BPTI-Trypsin Complexes-Insights from Simulation and Experiment". *J. Phys. Chem. B* **2022** *126*, 9985–9999.

[121] C. Rakers, M. Bermudez, B. G. Keller, J. Mortier and G. Wolber. "Computational Close up on Protein–Protein Interactions: How to Unravel the Invisible Using Molecular Dynamics Simulations?" *WIREs Comput. Mol. Sci.* **2015** *5*, 345–359.

Appendix

List of Abbreviations

Abu	α-aminobutyric acid
BPTI	Bovine Pancreatic Trypsin Inibitor
CV	Collective Variable
DfeGly	γ-difluoro-α-aminobutyric acid
GAFF	General AMBER Force Field
GIST	Grid Inhomogeneous Solvation Theory
GrsA	Gramicidin S Synthetase 1
IST	Inhomogeneous Solvation Theory
MD	Molecular Dynamics
MfeGly	γ-monofluoro-α-aminobutyric acid
PF5	Pentafluorophosphate
PMF	Potential of Mean Force
PTP1B	Protein Tyrosine Phosphatase 1B
RAMD	Random Acceleration Molecular Dynamics
RESP	Restrained Electrostatic Potential
RMSF	Root Mean Square Fluctuation
SASA	Solvent Accessible Surface Area
TfeGly	γ-trifluoro-α-aminobutyric acid
WHAM	Weighted Histogram Analysis Method

Supporting Information for Section 4.2

The Supporting Information for the publication:

"Water Network in the Binding Pocket of Fluorinated BPTI–Trypsin Complexes–Insights from Simulation and Experiment"
Leon Wehrhan, Jakob Leppkes, Nicole Dimos, Bernhard Loll, Beate Koksch, Bettina G. Keller
J. Phys. Chem. B **2022**
is published under the DOI: 10.1021/acs.jpcb.2c05496.

Supporting Information for Section 4.3

The following pages hold the Supporting Information for the publication:

"Pre-bound State Discovered in the Unbinding Pathway of Fluorinated Variants of the Trypsin-BPTI Complex Using Random Acceleration Molecular Dynamics Simulations"
Leon Wehrhan, Bettina G. Keller
J. Chem. Inf. Model. **2024**
DOI: 10.1021/acs.jcim.4c00338

Supplementary Material

Leon Wehrhan and Bettina G. Keller*[a]

Department of Biology, Chemistry, and Pharmacy, Freie Universität Berlin,
Institute of Chemistry and Biochemistry, Structural Biochemistry, Arnimallee 22,
Berlin, 14195 Germany

[a] Electronic mail: bettina.keller@fu-berlin.de

I. RESTRAINED METADYNAMICS

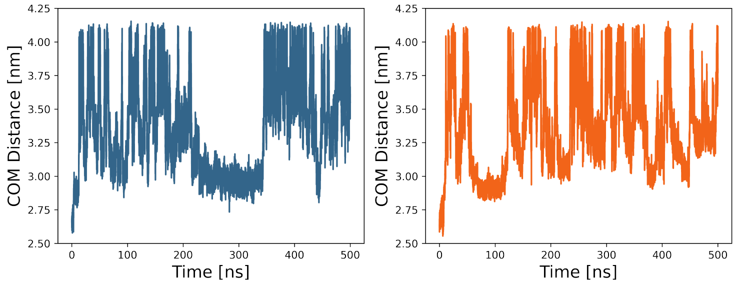


Figure S1. Center-of-mass (COM) distance between trypsin and BPTI as P1 TfeGly variant (left) and P1 Abu variant (right) throughout Metadynamics simulations with positional restraints.

A. Computational Methods

a. Restrained Metadynamics. Metadynamics simulations were run with Gromacs 2021.5 and Plumed 2.8. Well tempered Metadynamics simulations were employed for studying the undinding process of BPTI from Trypsin. The collective variable was defined to be the backbone center of mass distance between the receptor trypsin and the ligand BPTI. The repulsive gaussians were deposited at a rate of 1 ps at an initial height of 1.2 kJ/mol and σ of 0.1 nm. The biasfactor was set to be 5. An upper wall at a center of mass distance of 4.0 nm was installed with a force constant of 2500 kJ/(mol$\cdot$nm^2). The collective variables Θ_p, Φ_p, θ_1, ϕ_1 and ψ_1 were constrained with harmonic potentials of 400 kJ/(mol $\cdot$ rad^2) to remain at their respective value in the crystal structure of β-Trypsin-TfeGly-BPTI (pdb code: 4Y11). The orientational angle θ_1 and the orientational dihedrals ϕ_1 and ψ_1 were restrained at values of 2.068 rad, -0.956 rad and 0.592 rad, respectively. The positional dihedrals Θ_p and Φ_p were restrained at 1.991 rad and 1.437 rad, respectively.

II. RAMD TRAJECTORIES

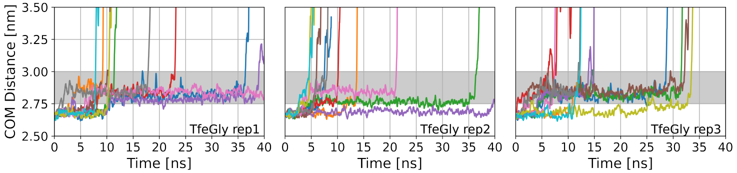


Figure S2. RAMD dissociation timeseries of TfeGly-BPTI. COM = center-of-mass. Grey area shows the center-of-mass distance of the pre-bound state. Left panel: replica 1, middle panel: replica 2, right panel:replica 3. 10 RAMD runs per replica.

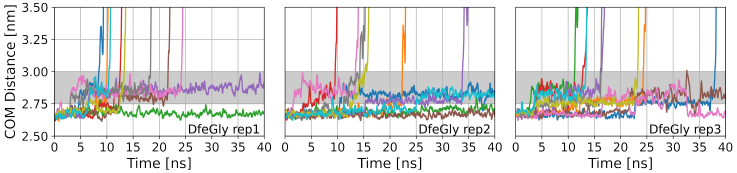


Figure S3. RAMD dissociation timeseries of DfeGly-BPTI. COM = center-of-mass. Grey area shows the center-of-mass distance of the pre-bound state. Left panel: replica 1, middle panel: replica 2, right panel:replica 3. 10 RAMD runs per replica.

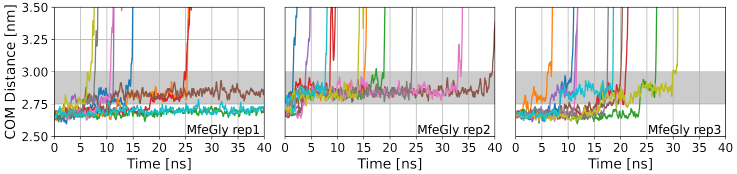


Figure S4. RAMD dissociation timeseries of MfeGly-BPTI. COM = center-of-mass. Grey area shows the center-of-mass distance of the pre-bound state. Left panel: replica 1, middle panel: replica 2, right panel:replica 3. 10 RAMD runs per replica.

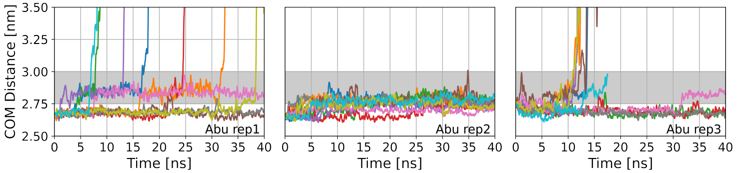


Figure S5. RAMD dissociation timeseries of Abu-BPTI. COM = center-of-mass. Grey area shows the center-of-mass distance of the pre-bound state. Left panel: replica 1, middle panel: replica 2, right panel:replica 3. 10 RAMD runs per replica.

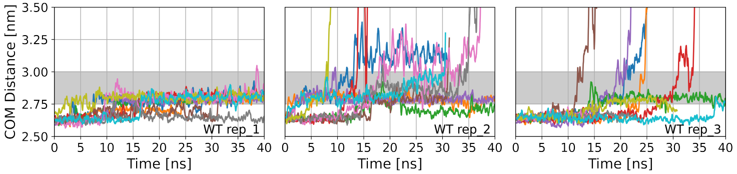


Figure S6. RAMD dissociation timeseries of wildtype-BPTI. COM = center-of-mass. Grey area shows the center-of-mass distance of the pre-bound state. Left panel: replica 1, middle panel: replica 2, right panel:replica 3. 10 RAMD runs per replica.

TABLE S1. RAMD simulations of complexes of trypsin with (fluorinated) Abu-BPTI complexes. For every replica there are 10 simulations of max. 40 ns length, which only differ by the random seed for the RAMD force. The criterion for the pre-bound state to be observed is that, throughout the whole simulation, the 200 ps moving average of the center-of-mass distance has to be at least 1 ns without interruption between 2.75 nm and 3.00 nm.

Complex	**Replica**	**Dissociated**	**Undissociated**	**Pre-bound observed**
TfeGly	1	8	2	8
TfeGly	2	9	1	6
TfeGly	3	10	0	8
DfeGly	1	8	2	7
DfeGly	2	6	4	7
DfeGly	3	8	2	10
MfeGly	1	7	2	7
MfeGly	2	9	1	9
MfeGly	3	10	0	7
Abu	1	7	3	7
Abu	2	0	10	7
Abu	3	6	4	4

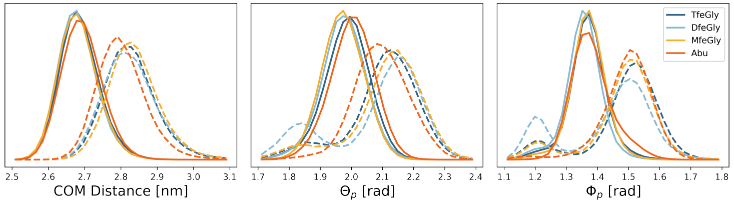


Figure S7. Center-of-mass distance histogram (left) histograms of Θ_p (middle) and Φ_p (right) for the fully bound state (solid line) and pre-bound state (dashed line) in the RAMD simulations of the Abu-BPTI variants.

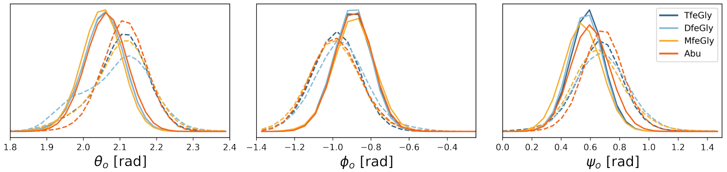


Figure S8. Histograms of θ_o (left), ϕ_o (middle) and ψ_o (right) for the fully bound state (solid line) and pre-bound state (dashed line) in the RAMD simulations of the Abu-BPTI variants.

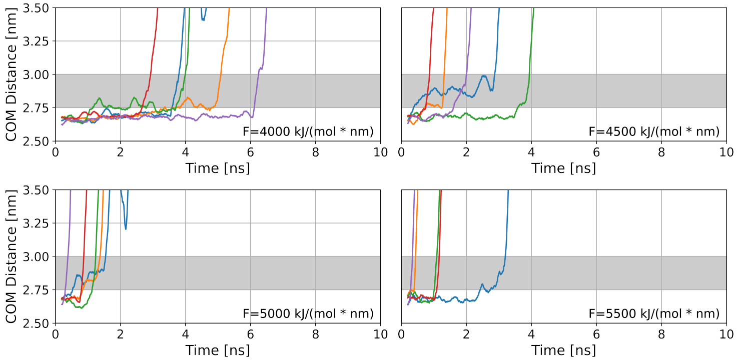


Figure S9. RAMD dissociation trajectories of TfeGly-BPTI with high random forces.

III. UNBIASED SIMULATIONS

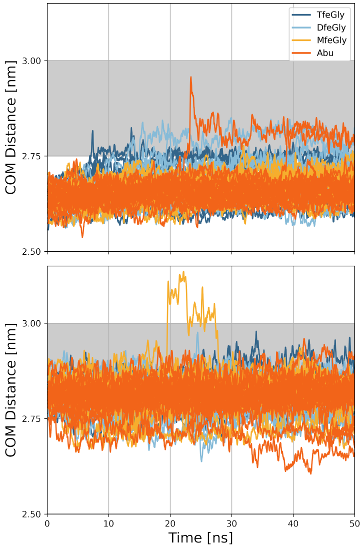


Figure S10. COM distance trajectory of all RAMD simulations of the Abu-BPTI variants overlayed. The grey area indicates the pre-bound state. The fluctuations throughout the trajectory are smoothed by calculating a moving average with a moving window of 200 ps.

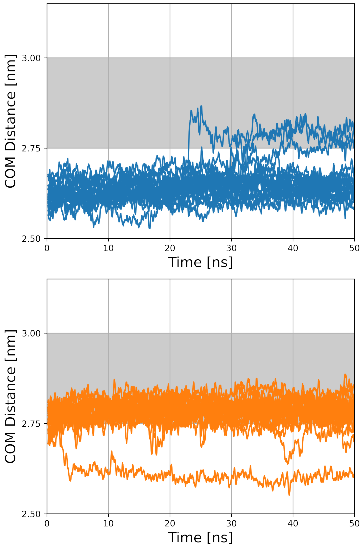


Figure S11. COM distance trajectory of all RAMD simulations of wildtype-BPTI overlayed. The grey area indicates the pre-bound state. The fluctuations throughout the trajectory are smoothed by calculating a moving average with a moving window of 200 ps. The simulations in the top panel were started in the fully bound state, the simulations in the bottom panel were started in the pre-bound state.

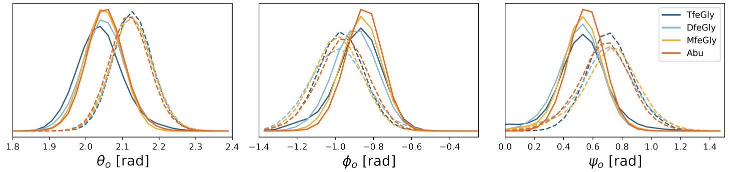


Figure S12. Histograms of θ_o (left), ϕ_o (middle) and ψ_o (right) for the fully bound state (solid line) and pre-bound state (dashed line) in the unbiased simulations of the Abu-BPTI variants.

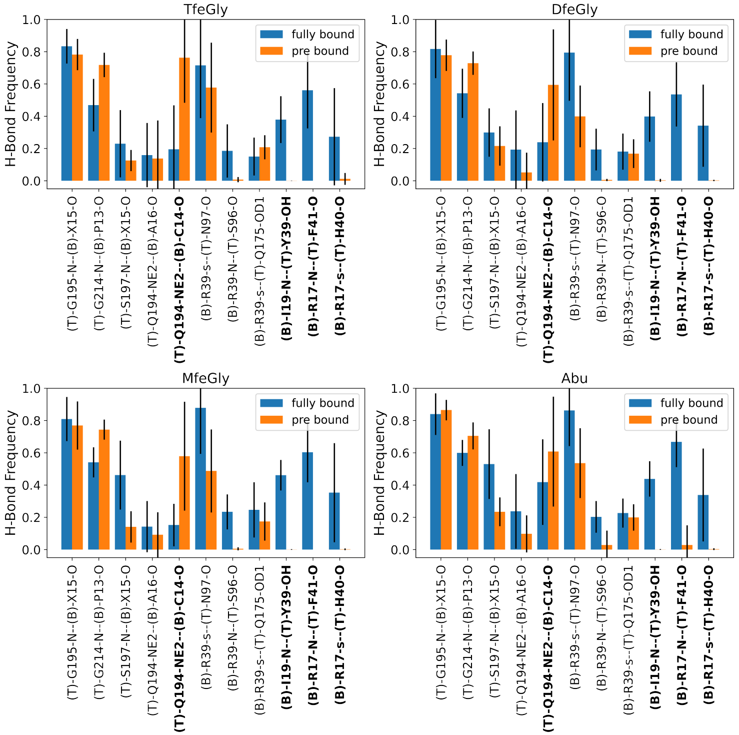


Figure S13. Hydrogen bond frequencies of all the unbiased simulations of the fully bound state and the pre-bound state, separate for the four Abu-BPTI variants. Residue name X = Abu, MfeGly, DfeGly or TfeGly, T = Trypsin, B = (Abu, MfeGly, DfeGly, TfeGly)-BPTI. O, N = heteroatoms in the backbone, OD1, OH, NE2 = heteroatoms in side chains. Hydrogen bonds are denoted as Donor- Acceptor. The side chain of arginine residues is denoted as "s", which means a hydrogen bond with any of the donors in the guanidine moiety.

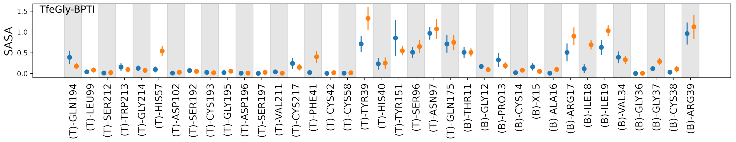


Figure S14. Solvent Accessible Surface Area (SASA) of all residues of the protein-Protein interface of trypsin and TfeGly-BPTI. (T) = trypsin, (B) = BPTI, X = TfeGly.

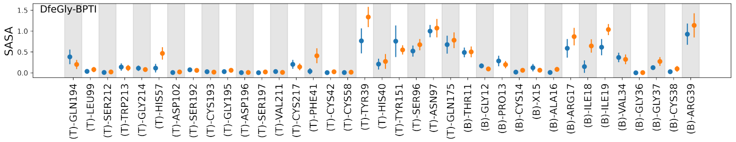


Figure S15. Solvent Accessible Surface Area (SASA) of all residues of the protein-Protein interface of trypsin and DfeGly-BPTI. (T) = trypsin, (B) = BPTI, X = DfeGly

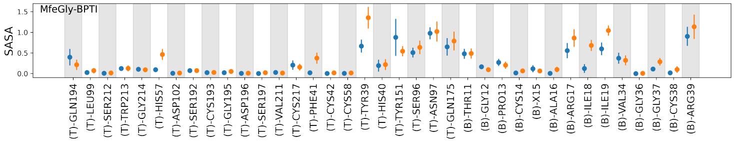


Figure S16. Solvent Accessible Surface Area (SASA) of all residues of the protein-Protein interface of trypsin and MfeGly-BPTI. (T) = trypsin, (B) = BPTI, X = MfeGly

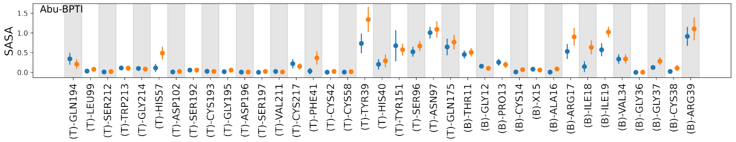


Figure S17. Solvent Accessible Surface Area (SASA) of all residues of the protein-Protein interface of trypsin and Abu-BPTI. (T) = trypsin, (B) = BPTI, X = Abu

IV. SHORT EQUILIBRATION OF RAMD SNAPSHOTS FOR SWARMS-OF-TRAJECTORIES STRING METHOD

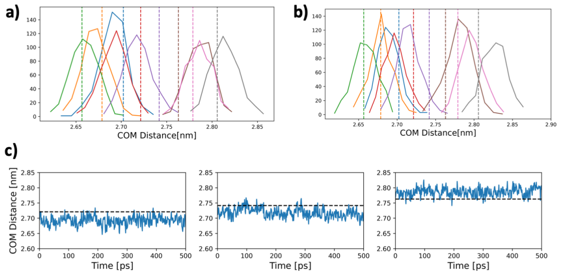


Figure S18. Short equilibration trajectories in our attempts for the swarms-of-trajectories string method. (a) Histograms of the distribution of the equilibration trajectories along the center-of-mass distance. The simulations are restrained on r, Θ_p and Φ_p. $k = 6500 kJ/mol/nm^2$ for r and $k = 400 kJ/mol/\text{rad}^2$ for Θ_p and Φ_p. The equilibrium value of the restraint = initial value along r is indicated as dashed line. (b) Histograms of the distribution of the equilibration trajectories along the center-of-mass distance. The simulations are restrained on the distance between Phe41-O and Arg17-N. $k = 6500 kJ/mol/nm^2$ (c) Trajectory timeseries of the simulations seen in (a). The equilibrium value of the restraint = initial value is indicated as dashed line.

Supporting Information for Section 4.4

The following pages hold the Supporting Information for the publication:

"Pentafluorophosphato-Phenylalanines: Amphiphilic Phosphotyrosine Mimetics Displaying Fluorine-Specific Protein Interactions"
Matteo Accorsi, Markus Tiemann, Leon Wehrhan, Lauren M. Finn, Ruben Cruz, Max Rautenberg, Franziska Emmerling, Joachim Heberle, Bettina G. Keller, Jörg Rademann
Angew. Chem. Int. Ed. **2022**
DOI: 10.1002/anie.202203579

Angewandte Chemie
Eine Zeitschrift der Gesellschaft Deutscher Chemiker

Supporting Information

Pentafluorophosphato-Phenylalanines: Amphiphilic Phosphotyrosine Mimetics Displaying Fluorine-Specific Protein Interactions

*M. Accorsi, M. Tiemann, L. Wehrhan, L. M. Finn, R. Cruz, M. Rautenberg, F. Emmerling, J. Heberle, B. G. Keller, J. Rademann**

Table of Contents

General methods

All moisture sensitive reactions were performed in glassware that was previously vacuum heat dried and flushed with Ar or N_2 using Schlenk technology.

Cadmium powder was activated with HCl (1 N) until a metallic shine was observed, washed with H_2O and acetone, dried at high vacuum and stored under inert atmosphere.

NMe_4F was purchased as the tetrahydrate from Sigma Aldrich and dried as described in the literature[1]. For this, the reagent was dissolved in dry MeOH. The solution was concentrated to a syrup at the rotary evaporator, re-dissolved 4 times in dry MeOH and concentrated again. Subsequently, the residue was heated at 130 °C for 3 d at high vacuum. The obtained white powder was stored and handled under inert argon atmosphere.

All other chemicals were purchased from Sigma (Merck), ABCR and Fluka and were used without any further purification.

Dry DMF and ACN were bought as anhydrous and stored over activated molecular sieves 4Å. All other dry solvents were obtained from a column-based solvent system (MBraun, MB-SPS-800).

Removal of volatile components was performed using rotary evaporators from Heidolph with a hot water bath of 40 °C, if not otherwise stated. The high vacuum obtained with an oil pump corresponds to 1 µbar or less. Lyophilized fractions were obtained from Christ Alpha 2-4 LD.

Product isolation was conducted on Biotage, Isolera^{TM} Spektra equipped with KP-Sil or RP-C18 SNAP Cartridges with appropriate HPLC grade solvent mixtures and deionized water, or with HPLC (Agilent Technologies, 1260 series, column Macherey-Nagel, Nucleodur 5 µm C18, 150 x 32 mm, equipped with Agilent 1260 Infinite diode array and multiple wavelength detector and fraction collector).

Melting points were measured with Büchi Melting point apparatus B-545.

Thin layer chromatography analyses were conducted on Merck Aluminum sheets pre-coated with silica gel (Merck, 60 F_{254}). Detection was carried out using 254 nm UV-Light, followed by dipping in ceric ammonium molybdate or ninhydrin stains.

The optical rotation was determined using IBZ Messtechnik Polar LµP (quartz cuvette, optic path 1 cm).

NMR spectra were measured on the following spectrometers: JEOL ECX400 (9.39 T), JEOL ECP500 (11.74 T), JEOL ECZ600 (14.09 T), Bruker Avance III 700 (16.44 T). Chemical shifts (δ) are reported in ppm and coupling constants (J) are given in Hz. ^1H and ^{13}C chemical shifts were referenced to the solvent peaks. ^{13}C and ^{31}P NMR spectra were hydrogen decoupled. Chemical shifts are given in ppm relative to the signal of the used deuterated solvent as internal standard.

HPLC chromatograms were recorded with an analytical HPLC system (Agilent Technologies, 1100 Series) equipped with a Luna column (column A), 3 µm C18 100 Å, 4.6 x 100 mm coupled with an ESI single quadrupole mass spectrometer LCMSD (Model# G1956B, Serial# US 44500857) from Agilent and a DAD detector using a gradient of water (A) and 99% ACN/water (B) both with 0.1% formic acid.

Alternatively, for shorter retention times, chromatograms were recorded with an HPLC system (Agilent Technologies, Infinity 1260), using Zorbax Eclipse plus C-18 RRHD column (column B) (2.1 x 50 mm, 1.8 µM, 95 Å) column, coupled with a DAD or mass detector (Agilent Technologies).

ESI high resolution mass spectra were recorded with an equipped with an analytical HPLC system (Agilent Technologies, Infinity II 1290), Zorbax Eclipse plus C-18 RRHD (2.1 x 50 mm, 1.8 µM, 95 Å) column, coupled with an ESI-Q-TOF iFunnel mass spectrometer (Agilent Technologies, 6550).

X-ray crystallographic analysis was performed on single crystals. Single crystals X-ray diffraction was performed on a Bruker D8 Venture system with graphite-monochromatic Mo-Kα radiation (λ = 0.71073 Å). Data reduction and structure solution were conducted as described in the section crystal structure determination below.

Dynamic light scattering experiments were performed on a Nicomp Nano DLS/ZLS system, at 25 °C and a wavelength of 660 nm. The respective viscosity and refraction index for each solvent mixture were taken from the literature.

General method of peptide synthesis

Peptide synthesis was conducted using Fmoc-strategy on Rink amide resin **13** from Merck (loading 0.34 mmol/g, 100-200 mesh, 1% divinyl-benzene/polystyrene). PP-PE syringes equipped with a PE-frit were used as reaction vessels.

Coupling of amino acids
N-Fmoc-protected amino acids with suitable side chain protection (5 equiv. with respect to the loading of the resin) were pre-activated with TBTU (4.9 equiv.) and DIPEA (10 equiv.) in a minimal volume of DMF. The solution was added to the resin pre-swollen in DMF and shaken for 2 h.

The coupling reactions were monitored using the Kaiser test.[2] Coupling effectiveness was quantified via UV-photometric determination of the dibenzofulvene product at 301 nm following to Fmoc cleavage with the following equation:

$$x = \frac{10^5 \cdot E_\lambda}{\varepsilon_\lambda \cdot m_{resin}}$$

x = Loading resin [mmol/g]
E_λ = Extinction
$\varepsilon_{\lambda 301}$ = 7800 l/mol cm

ε_λ = molar extinction coefficient
m_{resin} = mass resin [mg]

The resin was carefully vacuum dried prior to the Fmoc determination.

Washing of the resin
The resin was washed after every coupling and cleavage procedure with 5 syringe volumes of DMF.

Fmoc cleavage
The Fmoc-group was cleaved by using a mixture of piperidine (20%) in DMF. After adding the basic cocktail to the resin, the syringe was shaken for 10 min and then washed with DMF. The cleavage procedure was repeated once.

End capping of peptide
Before capping, the resin was swollen in DMF (2 ml / 100 mg of resin). The peptide N-Terminus was capped with an acetic anhydride/pyridine mixture (1:1, 1 ml/200 mg of resin) for 1 h at RT.

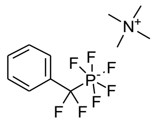
Peptide cleavage
The vacuum-dried resin was treated with Olah´s reagent (pyridinium poly-(hydrogen fluoride), ca. 70% HF and 30% pyridine) + 10% anisole for 90 min at RT. The cleavage mixture was slowly dropped into a saturated solution of $NaHCO_3$. The beads were washed with small portions of THF, followed by THF/ H_2O (1:1) and the washings dropped in the $NaHCO_3$ solution as well. The mixture was then concentrated to a minimum at the rotary evaporator. The residue was purified using an MPLC with a C18 column and a gradient of eluent A (10 mM NH_4HCO_3 in H_2O, pH 7.5) and eluent B (ACN). Collected fractions were analyzed with LCMS and lyophilized.

Scope of cleavage conditions

The cleavage conditions mentioned above (Olahs reagent with 10% anisol for 1h) were shown to successfully deprotect tert-butyl protection groups of commercially available amino acids such as glutamate and aspartate and the trityl protection group of cysteine.

Chemical synthesis

Tetramethylammonium 1-(pentafluorophosphato-difluoromethyl) benzene (2)



Diethyl difluoro-(phenyl)-methyl phosphonate (200 mg, 0.757 mmol, 1 equiv.) was dissolved in a Schlenk flask in dry ACN (5 mL). TMSBr (287 µL, 1.67 mmol 2.2 equiv.) was added dropwise under inert atmosphere. The solution was heated at 60°C for 1 hour. After disappearance of the starting material monitored via LC-MS, the vial was equipped with inert gas inlet and outlet to allow the release of the gaseous components developed after addition of dry DMF (293 µL, 3.78 mmol, 5 equiv.) and $(COCl)_2$ (636 µL, 7.57 mmol, 10 equiv.). After gas development decreased, the reaction was heated in a sealed vessel under inert atmosphere at 40°C with a water bath. After 1.5 hours, the reaction was cooled to 0°C with an ice bath. Previously weighted under inert atmosphere and dried NMe_4F (705 mg, 7.57 mmol, 10 equiv.) was then added slowly under inert atmosphere to the stirred and cooled reaction mixture. After 30 min, the mixture was slowly quenched in a cooled sat. aq. sol. of $NaHCO_3$ (25 mL/mmol starting material) and extracted with DCM (3 x 30 mL). The collected organic layer was then concentrated at the rotary evaporator, redissolved in H_2O/ACN and purified with RP-MPLC (RP c18, ACN / H_2O, 5 to 99%). After purification the product (119 mg, 62%) was isolated as white fluffy solid.

Rf= 0.5 (EtOAc)

Melting point = 107 °C

^{1}H NMR (600 MHz, Acetonitrile- *d_3*) δ = 7.42 (d, J=7.6, 2H, Ar-H), 7.35 – 7.27 (m, 3H, Ar-H), 3.02 (s, 12H, NMe_4)

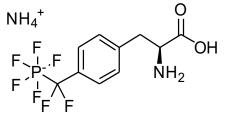
^{13}C NMR (151 MHz, Acetonitrile- *d_3*) δ = 127.96, 127.32 (Ar-C), 125.53 (t, J=7.8, CF_2), 55.71 – 54.20 (s, NMe4)

^{19}F NMR (565 MHz, Acetonitrile- *d_3*) δ = -70.07 (dp, J=696.0, 43.5, 1F, F_{ax}), -71.76 (dt, J=855.7, 42.9, 9.1, 2F, F_{eq}), -98.59 (dt, J=119.9.8, 9.5, 2F, F_{eq})

^{31}P NMR (243 MHz, Acetonitrile- *d_3*) δ = -145.27 (dtquin, J=864.3, 696.3, 125.2)

HRMS (ESI): $[M]^-$ calculated for $C_7H_5F_7P^-$: 253.0023 Da, found: 253.0033 m/z

Ammonium 4-(pentafluorophosphato-difluoromethyl)-L-phenylalanine (3)



The sodium salt of Fmoc-protected amino acid **9** (100 mg, 0.165 mmol, 1 equiv.) was dissolved in ACN (1.8 ml). Piperidine (200 µl, 2.00 mmol, 12 equiv.) was added and the resulting mixture stirred for 8 h at RT. All volatile components were removed under reduced pressure and the obtained crude product purified via MPLC using a C18 reversed phase column and a gradient of eluent A (10 mM NH_4HCO_3 in H_2O, pH 7.5) and eluent B (ACN). Fractions containing the product were concentrated at a rotary evaporator and lyophilized, yielding the product as a pale yellow solid (58 mg, 0.160 mmol, 97%).

^{1}H NMR (700 MHz, D_2O) δ = 7.47 (d, J=7.6, 2H, Ar-H), 7.32 (d, J=8.0, 2H, Ar-H), 3.95 (dd, J=8.5, 4.9, 1H, CHN), 3.30 (dd, J=14.5, 4.8, 1H, $CH_{2\alpha}$-Phe), 3.12lll – 3.04 (m, 1H, $CH_{2\beta}$-Phe)

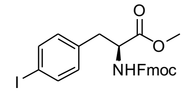
^{13}C NMR (176 MHz, D_2O) δ = 173.7 (O=C), 138.1, 136.3, 129.1, 129.0, 125.8 (6 x Ar-C), 44.5 (CHN)z, 36.1 (CH_2Phe)

^{19}F NMR (376 MHz, D_2O) δ = -68.50 (dp, J=693.6, 43.6, 1F, F_{ax}), -72.14 (ddt, J=864.2, 43.0, 8.5, 4F, F_{eq}), -98.27 (dt, J=126.2, 8.3, 2F, CF_2)

^{31}P NMR (162 MHz, D_2O) δ = -143.00 (pdt, J=864.8, 693.45, 126.3)

HRMS (ESI): $[M]^-$ calculated for $C_{10}H_{10}F_7NO_2P^-$: 340.03429 Da, found: 340.03452 m/z

Methyl N-(fluorenyl-9H-methoxy-carbonyl)-4-iodo-L-phenylalanine (4)



Fmoc-4-I-Phe-OH (from ABCR, 2988 mg, 10.26 mmol, 1 equiv.) was dissolved in dry MeOH (25 ml) in a heat- and vacuum-dried Schlenk flask under Ar atmosphere. 3 drops of dry DMF were added and the solution was cooled to 0 °C. Oxalyl chloride (2.6 ml, 30.795 mmol, 3 equiv.) was added dropwise under stirring and the reaction was allowed to reach RT for 16 h. The amber solution was then concentrated using a rotary evaporator, diluted with EtOAc and washed with H_2O, saturated $NaHCO_3$ solution, and brine. The organic layer was dried over Na_2SO_4, filtrated and concentrated in vacuo. Product **4** (4717 mg, 92%) was obtained as white solid.

R_f = 0.3 (EtOAc/Hex, 20%)

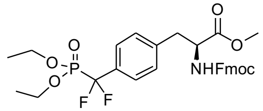
^{1}H NMR (500 MHz, $CDCl_3$) δ = 7.77 (d, J=7.6, 2H, Ar-H), 7.60 (d, J=8.1, 2H, Ar-H), 7.61 – 7.51 (m, 2H, Ar-H), 7.41 (t, J=7.4, 2H, Ar-H), 7.32 (t, J=7.5, 2H, Ar-H), 6.81 (d, J=8.1, 2H, Ar-H), 5.24 (d, J=7.9, 1H, NH), 4.64 (q, J=5.9, 1H, α-H-Phe), 4.47 (dd, J=10.5, 7.2, 1H, $CH_{2\alpha}$-Fmoc), 4.37 (dd, J=10.4, 6.9, 1H, $CH_{2\beta}$-Fmoc), 4.20 (t, J=6.7, 1H, CH-Fmoc), 3.73 (s, 3H, OMe), 3.09 (dd, J=13.9, 5.6, 1H, $CH_{2\alpha}$-Phe), 3.02 (dd, J=13.9, 5.7, 1H, $CH_{2\beta}$-Phe)

^{13}C NMR (126 MHz, $CDCl_3$) δ = 171.7 (O=C-methyl ester), 155.6 (O=C-Fmoc), 143.9, 143.8, 141.5, 141.4, 137.8, 135.5, 131.4, 127.9, 127.2, 120.1 (17 x Ar-C), 92.8 (C-I), 67.0 (CH_2-Fmoc), 54.6 (CHN), 52.6 (OCH_3), 47.3 (CH-Fmoc), 37.8 (CH_2Phe)

HRMS (ESI): $[M+H]^+$ calculated for $C_{25}H_{23}INO_4^+$: 528.06663 Da; found: 528.06571 m/z;
$[M+Na]^+$ calculated for $C_{25}H_{22}INNaO_4^+$: 550.04857 Da; found: 550.04790 m/z

Spectral data were consistent with published values.[3]

Methyl N-(fluorenyl-9H-methoxy-carbonyl)-4-(diethoxyphosphoryl-difluoromethyl)-L-phenylalanine (5)



A heat- and vacuum-dried Schlenk flask was charged with cadmium powder (2.578 g, 22.94 mmol, 6 equiv.) activated and dried as described above and dry DMF (3 ml). To the stirred suspension, diethyl bromo-difluoromethyl-phosphonate (2.261 ml, 12.73 mmol, 3.33 equiv.) was added dropwise to the reaction flask at room temperature. The slightly exothermic reaction was stirred for 3 h. In another flask, previously dried Fmoc-(4-I)Phe-OMe **4** (2.016 g, 3.82 mmol, 1 equiv.) was dissolved in dry DMF (1 ml) and CuBr (1.645 g, 11.47 mmol, 3 equiv.) was added. The solution containing the organocadmium reagent was added slowly and dropwise to this stirred mixture under Ar atmosphere and the reaction mixture was stirred for 16 h and monitored via TLC (1:4, EtOAc / Hex). After addition of EtOAc, the precipitate was filtrated off over a bed of Celite and the filtrate washed with a saturated aqueous solution of NH_4Cl (20 ml, 3x), H_2O (20 ml) and brine (20 ml). The organic layer was dried over Na_2SO_4, filtrated and concentrated under reduced pressure. After purification of the crude via column chromatography at MPLC (SiO_2, EtOAc / Hex 1:4 then 1:2), the product (2.256 g, 99%) was isolated as a colorless oil.

R_f = 0.3 (EtOAc/Hex, 50%)

^{1}H NMR (600 MHz, $CDCl_3$) δ = 7.75 (d, J=7.5, 2H, 2 x Ar-H), 7.54 (dd, J=14.7, 7.7, 4H, 2 x Ar-H), 7.38 (t, J=7.4, 2H, 2 x Ar-H), 7.30 (t, J=7.5, 2H, 2 x Ar-H), 7.17 (d, J=7.9, 2H, 2 x Ar-H), 5.31 (d, J=8.2, 1H,NH), 4.66 (q, J=6.0, 1H, α-H-Phe), 4.43 (dd, J=10.6, 7.2, 1H, CH_2-Fmoc), 4.36 (dd, J=10.8, 7.0, 1H, CH_2-Fmoc), 4.23 – 4.05 (m, 5H, 2xCH_2-ethyl, CH-Fmoc), 3.70 (s, 3H, OMe), 3.17 (dd, J=13.9, 5.8, 1H, CH_2-Phe), 3.11 (dd, J=13.9, 6.1, 1H, CH_2-Phe), 1.28 (t, J=7.1, 6H, CH_3-ethyl)

^{13}C NMR (151 MHz, $CDCl_3$) δ = 171.67 (O=C-methyl ester),155.61 (O=C-Fmoc), 143.87, 143.75, 141.41, 138.99, 131.54 (td, J=22.2, 14.0), 129.50, 127.84, 127.16, 126.58 (t, J=7.6), 125.7, 125.14, 120.08 (d, J=3.2, 18 x Ar-C), 117.21 (dd, J=262.5, 218.4, CF_2), 67.02 (CH_2-Fmoc), 64.88 (dd, J=6.8, 1.3, 2 x CH_2 Ethyl), 54.72 (CHN), 52.52 (OCH_3), 47.23 (CH-Fmoc), 38.05 (CH_2Phe), 16.40 (d, J=5.5, 2 x CH_3 Ethyl)

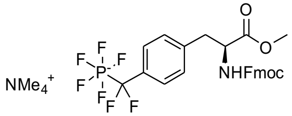
^{19}F NMR (376 MHz, $CDCl_3$) δ = -108.18 (d, J=116.0)

^{31}P NMR (162 MHz, $CDCl_3$) δ = 6.94 (t, J=116.1)

HRMS (ESI): $[M + H]^+$ calculated for $C_{30}H_{33}F_2NO_7P^+$: 588.1957 Da, found: 588.1932 m/z;
$[M+Na]^+$ calculated for $C_{30}H_{32}F_2NNaO_7P^+$: 610.1777 Da, found: 610.1765 m/z;
$[M+K]^+$ calculated for $C_{30}H_{32}F_2KNO_7P^+$: 626.1516, found: 626.1493 m/z.

Spectral data were consistent with published values.[4]

Tetramethylammonium methyl N-(fluorenyl-9H-methoxy-carbonyl)-4-(pentafluorophosphato-difluoromethyl)-L-phenylalanine (8)



Diethyl phosphonic acid ester **5** (828 mg, 1.41 mmol, 1 equiv.) was dissolved in a Schlenk flask in dry ACN (5 ml). TMSBr (930 µL, 7.05 mmol, 5 equiv.) was added dropwise under inert atmosphere. The solution was heated at 60 °C for 1.5 h. After disappearance of the starting material monitored via LC-MS, the vial was equipped with inert gas inlet and outlet to allow the release of the gaseous components developed after dropwise addition of dry DMF (545 µL, 7.05 mmol, 5 equiv.) followed by $(COCl)_2$ (1.18 ml, 14.09 mmol, 10 equiv.). After gas development ceased, the reaction was heated in a sealed vessel under inert atmosphere at 40 °C with a water bath. After 1.5 h, a small aliquot was taken and MeOH was added. The formation of the dimethyl ester was confirmed via LCMS. The reaction mixture was then cooled to 0 °C with an ice bath. Previously weighted under inert atmosphere and dried as described above, NMe_4F (1050 mg, 14.09 mmol, 10 equiv.) was then added slowly under inert atmosphere to the stirred and cooled reaction mixture. After 1 h, the mixture was quenched by slowly pouring it into an ice cooled saturated aqueous solution of $NaHCO_3$ (30 ml) and extracted with DCM (3 x 30 ml). The collected organic layers were then concentrated at the rotary evaporator, redissolved in H_2O/ACN and purified with RP-MPLC. After purification of the crude via column chromatography at MPLC (RP C18, ACN / H_2O, 5 to 99%), the fractions were analyzed at LCMS and those containing the product were concentrated at the rotary evaporator and lyophilized, yielding 619 mg (68%) of the title compound as white lyophilisate.

Melting point = 140-142 °C

$[\alpha]_D^{20}$ = - 6.9 (c = 1, MeOH)

^{1}H NMR (600 MHz, ACN-d3) δ = 7.80 (d, J=7.6, 2H, Ar-H), 7.65 – 7.53 (m, 2H, Ar-H), 7.39 (t, J=7.5, 2H, Ar-H), 7.38 – 7.23 (m, 4H), 7.15 (d, J=7.8, 2H), 4.39 (q, J=8.2, 1H, CHN), 4.32 – 4.20 (m, 2H, Fmoc CH_2), 4.18 (t, J=6.9, 1H, Fmoc CH), 3.64 (s, 3H, OMe), 3.15 – 3.09 (m, 1H, $CH_{2\alpha}$Phe), 3.02 (s, 12H, NMe_4), 2.97 – 2.88 (m, 1H, $CH_{2\beta}$Phe)

^{13}C NMR (151 MHz, ACN-d3) δ = 172.22 (O=C-methyl ester), 155.93 (O=C-Fmoc), 144.13, 141.19, 136.76, 129.45, 128.24, 127.78, 127.21 (d, J=3.2), 125.71 (t, J=7.0, CF_2), 125.27 (d, J=10.3), 120.05 (18x Ar-C), 66.39 (CH_2-Fmoc) , 55.46 (CHN), 55.24 (NMe_4), 51.89 (OMe), 47.03 (CH-Fmoc), 36.87 (CH_2Phe)

^{19}F NMR (565 MHz, ACN-d3) δ = -69.65 (p, J=42.9, F_{ax}), -70.88 (p, J=42.9, F_{ax}), -71.05 (t, J=8.8, F_{eq}), -71.12 (t, J=8.7, F_{eq}) -98.26 (dt, J=120.3, 9.1, CF_2)

^{31}P NMR (243 MHz, ACN-d3) δ = -143.43 (pdt, J=855.99, 696.35, 120.1)

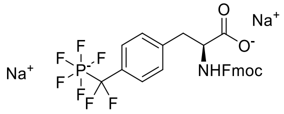
HRMS (ESI): $[M]^-$ calculated for $C_{26}H_{22}F_7NO_4P^-$: 576.11802 Da, found: 576.11809 m/z

Supplementary Table 1: Results of stability tests

Compound	Conditions	PF_5^- integrity (analyzed via NMR)
PhenylCF$_2$PF$_5^-$ **2**	HFIP neat, 1 h	50% degradation (to monofluorophosphate)
	HFIP / DCM 1:4, 1 h	stable
Fmoc-(4-PF$_5^-$CF$_2$)Phe-OH **9**	TFA 100%, 3 h	decomposition
	TFA 95% (aq.), 2 h	decomposition
	0.1 M HCl in HFIP, 2 h	decomposition
	AcOH in DCM (1:9), 1.5 h	decomposition
	5% TFA in DCM, 1 h	decomposition
	Sonication in ACN	stable
	0.1 M HCl (aq.), 1 h	stable
	0.1 M HCl (aq.), 24 h	15% decomposition
	0.01 M HCl (aq.), 1 h	stable
	0.01 M HCl (aq.), 24 h	stable
	0.1 M TFA (aq.), 1 h	10% decomposition
	0.01 M TFA (aq.), 1 h	stable
	0.1% TFA (aq.), 2 h	stable
	0.2% TFA (aq.), 2 h	35% decomposition
	TMSBr (100 equiv.) in ACN, 1 h	decomposition
	SiO$_2$ (30 µL/mg), in ACN, 1 h	stable
	piperidine (20%) in DMF, 24 h	stable
	pyridine neat, 1 h	stable
	DBU (2%) in DMF, 30 min	stable (analyzed via LCMS)

The compound (3 mg) was added to a vial containing the reagent to be tested and transferred into an NMR tube. After the specified time, a ^{19}F-NMR spectrum was recorded, or in case of the LCMS study, an aliquot was taken and analyzed.

Sodium N-(fluorenyl-9H-methoxy-carbonyl)-4-(pentafluorophosphato-difluoromethyl)-L-phenylalanine (9)



To the methyl ester **8** (60 mg, 0.098 mmol, 1 equiv.) in 50 ml of an aqueous solution of ammonium bicarbonate (50 mM, pH 7.8) were added two spatula tips of *Bacillus licheniformis* protease (from Sigma Aldrich) and 5 ml ACN. The resulting mixture was stirred at RT overnight. All volatile components were removed under reduced pressure and the obtained crude product was purified via MPLC using a C18 reversed phase column and a gradient of eluent A (10 mM NH_4HCO_3 in H_2O, pH 7.5) and eluent B (ACN). Fractions containing the product were concentrated at a rotary evaporator and lyophilized, yielding the product as an off white solid. Subsequently the pure compound was dissolved in a 1:1 mixture of H_2O/ACN and ion exchanged over Na-loaded Amberlite (IR 120). The exchange was performed in a glass column with a length of 21 cm, an inner diameter of 7 mm and a flowrate of 30 µl/s yielding **9** as a white solid (57 mg, 0.094 mmol, 96%).

[α]$_D$^{20} = + 12.1 (c = 1, MeOH)

^{1}H NMR (600 MHz, DMSO-*D*$_6$) δ = 7.88 (d, *J* = 7.5 Hz, 2H, Ar-H), 7.66 (dd, *J* = 15.9, 7.5 Hz, 2H, Ar-H), 7.44 – 7.37 (m, 2H, Ar-H), 7.37 – 7.27 (m, 2H, Ar-H), 7.21 (d, *J* = 7.9 Hz, 2H, Ar-H), 7.14 (d, *J* = 7.9 Hz, 2H, Ar-H), 6.28 (s, 1H, NH), 4.30 – 4.13 (m, 3H, Fmoc-CH$_2$, Fmoc-CH), 4.03 – 3.96 (m, 1H, αH), 3.10 – 2.87 (m, 2H, βH).

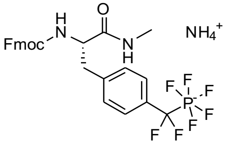
^{13}C NMR (151 MHz, DMSO-*D*$_6$) δ = 173.34 (C=O), 155.54 (C=O), 143.9, 140.67, 138.18, 128.94, 128.05, 127.58, 127.08, 125.3, 125.2, 124.9, 121.40, 120.07 (Ar- C), 65.46 (CH$_2$Fmoc), 56.13 (CHN), 46.62 (CH-Fmoc), 36.60 (CH$_2$Phe).

^{19}F NMR NMR (565 MHz, DMSO-*D*$_6$) δ = -67.57 (dp, *J* = 698.8, 45.3, 44.9 Hz, 1F, F$_{ax}$), -69.86 (ddt, *J* = 858.5, 44.6, 8.1 Hz, 4F, F$_{eq}$), -96.65 (d, *J* = 120.9 Hz, 2F, CF$_2$).

^{31}P NMR (243 MHz, DMSO-*D*$_6$) δ = -143.55 (pdt, *J* = 858.5, 698.0, 121.0 Hz).

HRMS (ESI): [M]$^-$ calculated for $C_{25}H_{20}F_7NO_4P^-$: 562.10237 Da, found: 562.10245 m/z

Ammonium N-(fluorenyl-9H-methoxy-carbonyl)-4-(pentafluorophosphato-difluoromethyl)-L-phenylalanyl-N-methylamide (S1)



To **9** (0.17 g, 0.28 mmol, 1 equiv.) in 10 ml dry ACN were added TBTU (0.27 g, 0.84 mmol, 3 equiv.) and diisopropylethylamine (0.24 ml, 1.68 mmol, 6 equiv.). The resulting solution was stirred at RT for 3 min. Methylamine hydrochloride (0.95 mg, 1.40 mmol, 5 equiv.) was added and the reaction mixture stirred for 1 h at RT. The volatile components were removed under reduced pressure and the crude product was purified via MPLC using a C18 reversed phase column and a gradient of eluent A (10 mM NH_4HCO_3 in H_2O, pH 7.5) and eluent B (ACN). Fractions containing the product were concentrated at a rotary evaporator and lyophilized, yielding the product **S1** as an off-white solid (0.16 g, 0.27 mmol, 95%).

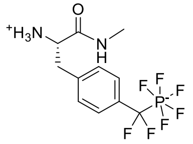
^{1}H NMR (600 MHz, ACN-d_3): δ = 7.83 (d, J = 7.6 Hz, 2H, Ar-H), 7.62 (dd, J = 14.7, 7.5 Hz, 2H, Ar-H), 7.42 (t, J = 7.5 Hz, 2H, Ar-H), 7.34 (dt, J = 15.3, 7.3 Hz, 4H, Ar-H), 7.18 (d, J = 7.8 Hz, 2H, Ar-H), 6.53 (s, 1H, MeNH), 5.91 (d, J = 8.3 Hz, 1H, FmocNH), 4.35 – 4.15 (m, 4H, αH+CH Fmoc+CH_2 Fmoc), 3.13 – 2.83 (m, 2H, βH), 2.64 (d, J = 4.7 Hz, 3H, $NHCH_3$).

^{19}F NMR (565 MHz, ACN -d_3): δ = -70.28 (dp, J = 696.3, 43.1, 42.1 Hz, 1F, F_{ax}), -71.88 (ddt, J = 856.3, 43.0, 9.1 Hz, 4F, F_{eq}), -98.27 (dp, J = 120.7, 9.4, 8.9 Hz, 2F, CF_2).

^{31}P NMR (243 MHz, ACN -d_3): δ = -143.89 (pdt, J = 856.3, 695.5, 119.8 Hz).

ESI-MS (m/z): $[M]^-$ calculated for $([C_{26}H_{23}F_7N_2O_3P])^-$: 575.1 Da; found: 575.0 m/z

4-(Pentafluorophosphato-difluoromethyl)-L-phenylalanyl-N-methylamide (S2)



(S1) (0.15 g, 0.26 mmol, 1 equiv.) was dissolved in ACN (9 ml). Piperidine (1 ml) was added and the resulting solution stirred at RT for 7 h. The volatile components were removed under reduced pressure and the crude product was purified via MPLC using a C18 reversed phase column and a gradient of eluent A (10 mM NH_4HCO_3 in H_2O, pH 7.5) and eluent B (ACN). Fractions containing the product were concentrated at a rotary evaporator and lyophilized, yielding product **S2** as a white solid (0.09 g, 0.25 mmol, 96%.).

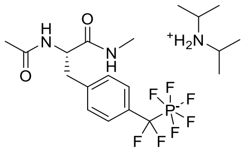
^{1}H NMR (600 MHz, D_2O): δ = 7.51 (d, J = 7.8 Hz, 2H, Ar-H), 7.31 (d, J = 7.9 Hz, 2H, Ar-H), 4.00 (t, J = 7.5 Hz, 1H, αH), 3.19 – 3.08 (m, 2H, βH), 2.63 (s, 3H, $NHCH_3$).

^{19}F NMR (565 MHz, D_2O): δ = -68.48 (dp, J = 692.4, 43.2 Hz, 1F, F_{ax}), -72.12 (ddt, J = 864.4, 43.1, 8.6 Hz, 4F, F_{eq}), -98.30 (ddt, J = 126.2, 17.6, 8.7 Hz, 2F, CF_2).

^{31}P NMR (243 MHz, D_2O): δ = -142.95 (pdt, J = 863.6, 692.9, 126.0 Hz).

ESI-MS (m/z): $[M]^-$ calculated for $([C_{11}H_{13}F_7N_2OP])^-$: 353.0 Da; found: 353.0 m/z

Diisopropylammonium N-Acetyl-4-(pentafluorophosphato-difluoromethyl)-3-(methylamino)-L-phenylalaninyl-amide (10)



To **(S2)** (0.04 g, 0.13 mmol, 1 equiv.) in 10 ml dry ACN were added diisopropylamine (0.13 ml, 0.75 mmol, 6 equiv.) and acetic anhydride (0.06 ml, 0.63 mmol, 5 equiv.). The resulting solution was stirred at RT for 4 h. The volatile components were removed under reduced pressure and the crude product was purified via MPLC using a C18 reversed phase column and a gradient of eluent A (10 mM NH_4HCO_3 in H_2O, pH 7.5) and eluent B (ACN). Fractions containing the product were concentrated at a rotary evaporator and lyophilized, yielding product **10** as a white solid (0.06 g, 0.13 mmol, quant.).

^{1}H NMR (600 MHz, D_2O): δ = 7.48 (d, J = 6.7 Hz, 2H, Ar-H), 7.30 (d, J = 8.0 Hz, 2H, Ar-H), 4.48 (t, 1H, αH), 3.51 (hept, J = 6.5 Hz, 2H, DIPA $NCHCH_3$), 3.16 – 3.00 (m, 2H, βH), 2.63 (s, 3H, $NHCH_3$), 1.95 (s, 3H, $NHCOCH_3$), 1.30 (d, J = 6.5 Hz, 12H, DIPA $NCHCH_3$).

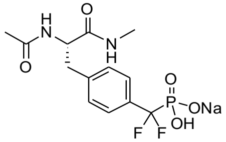
^{19}F NMR (565 MHz, D_2O): δ = -68.38 (dp, J = 692.4, 43.4 Hz, 1F, F_{ax}), -72.11 (ddt, J = 864.8, 43.1, 8.8 Hz, 4F, F_{eq}), -98.10 (dt, J = 126.4, 8.7 Hz, 2F, CF_2).

^{31}P NMR (243 MHz, D_2O): δ = -142.86 (pdt, J = 864.9, 692.4, 127.3 Hz).

^{13}C NMR (151 MHz, D_2O): δ = 174.13 ($NHCOCH_3$), 173.56 ($CONHCH_3$), 137.48 (Ar-$C_{quart.}$), 128.81 (Ar-CH), 125.40 (Ar-CH), 55.34 (αC), 47.27(DIPA $NCHCH_3$), 36.82 (βC), 25.66 ($NHCH_3$), 21.60 ($NHCOCH_3$), 18.29 (DIPA $NCHCH_3$).

ESI-HRMS (m/z): $[M]^-$ calculated for $([C_{13}H_{15}F_7N_2O_2P])^-$: 395.0759 Da; found: 395.0759 m/z

Sodium N-Acetyl-4-(phosphato-difluoromethyl)-3-(methylamino)-L-phenylalaninyl-amide (10a)



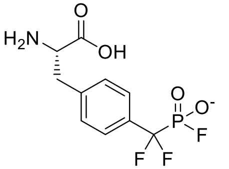
10 (0.01 g, 0.02 mmol, 1 equiv.) was stirred in 2 M HCl for 72 h. The volatile components were removed under reduced pressure and the desired product obtained as an off-white solid (0.01 g, 0.02 mmol, quant.)

^{1}H NMR (600 MHz, D_2O): δ = 7.55 (d, *J* = 5.4 Hz, 1H, Ar-H), 7.34 (d, *J* = 5.1 Hz, 2H, Ar-H), 4.49 (t, *J* = 9.3 Hz, 1H, αH), 3.26 – 2.97 (m, 2H, βH), 2.65 (s, 3H), 1.94 (s, 3H).

^{19}F NMR (565 MHz, D_2O): δ = -108.06 (d, *J* = 106.0 Hz).

^{31}P NMR (243 MHz, D_2O): δ = 4.88 (t, *J* = 105.1 Hz).

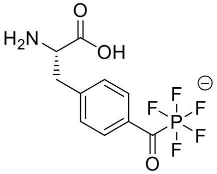
Ammonium 4-(monofluorophosphono-difluoromethyl)-L-phenylalanine (11)



The amino acid **3** (10 mg, 0.03 mmol, 1 equiv.) was stirred in a mixture of glacial acetic acid and deuterated DCM (1:9,) for 90 min. NMR analysis showed full conversion to the product **11**.

^{19}F NMR (376 MHz, DCM): δ = -70.45 (d, J = 1000.3 Hz, 1F), -106.39 (d, J = 97.9 Hz, 2F, CF_2).

Ammonium 4-(pentafluorophosphato-carbonyl)-L-phenylalanine (12)



Amino acid **3** (11 mg, 0.030 mmol, 1 equiv.) was dissolved in 200 µl Olah´s reagent (pyridinium poly-(hydrogen fluoride)) and 2 µl of water. The reaction was stirred for 6 hours and subsequently quenched with a saturated solution of $NaHCO_3$. All volatile components were removed under reduces pressure and the crude product purified via MPLC using an C18 reversed phase column and a gradient of eluent A (10 mM NH_4HCO_3 in H_2O, pH 7.5) and eluent B (ACN). Fractions containing the product were concentrated at a rotary evaporator and lyophilized. The product was obtained as a white solid (9 mg, 0.026, 87%).

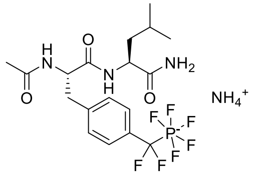
^{1}H NMR (600 MHz, D_2O): δ = 7.97 (d, J = 7.8 Hz, 2H, Ar-H), 7.38 (d, J = 7.7 Hz, 2H, Ar-H), 3.68 – 3.65 (m, 1H, αH), 3.20 – 2.89 (m, 2H, βH).

^{19}F NMR (565 MHz, D_2O): δ = -64.85 (dd, J = 891.8, 44.6 Hz, 4F, F_{eq}), -67.63 (dp, J = 709.7, 44.6, 44.1 Hz, 1F, F_{ax}).

^{31}P NMR (243 MHz, D_2O): δ = -145.71 (pd, J = 892.2, 709.9 Hz).

ESI-MS (m/z): $[M]^-$ calculated for ($[C_{10}H_{10}F_5NO_3P])^-$: 318.0319 Da; found: 318.0382 m/z

Ammonium N-acetyl-4-(pentafluorophosphato-difluoromethyl)-L-phenylalaninyl-L-leucyl-amide (14)



The dipeptide **14** was synthesized according to the general method. From 94 mg Rink amide resin, 11.6 mg (75%) of the product were obtained as white lyophilisate.

^{1}H NMR (600 MHz, D_2O): δ = 7.49 (d, J = 7.8 Hz, 2H, Ar-H), 7.33 (d, J = 8.0 Hz, 2H, Ar-H), 4.59 (t, J = 7.6 Hz, 1H, αH-Y*), 4.26 (dd, J = 10.1, 4.1 Hz, 1H, αH-Leu), 3.11 (dd, J = 7.6, 3.2 Hz, 2H, βH-Y*), 1.97 (s, 3H, $NHCOCH_3$), 1.62 – 1.49 (m, 3H, βH+γH-Leu), 0.87 (dd, J = 34.1, 5.8 Hz, 6H, δH-Leu).

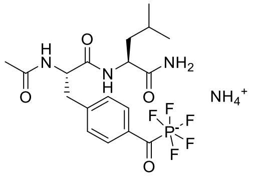
^{13}C NMR (151 MHz, D_2O): δ = 177.10 ($CONH_2$.), 174.16 ($NHCOCH_3$), 173.44 (CO-Y*), 137.70 (Ar-$C_{quart.}$), 137.21 (Ar-$C_{quart.}$), 128.94 (Ar-C), 125.59 (Ar-fC), 55.12 (αC-Y*), 52.20 (αC-Leu), 39.70 (βC-Leu), 36.50 (βC-Y*), 24.14 (γC-Leu), 22.22 (δC-Leu), 21.54 ($NHCOCH_3$), 20.43 (δC-Leu).

^{19}F NMR (565 MHz, D_2O): δ = -68.41 (dp, J = 692.0, 43.1 Hz, 1F, F_{ax}), -72.05 (ddt, J = 864.5, 42.9, 8.5 Hz, 4F, F_{eq}), -97.98 (d, J = 117.9 Hz, 2F, CF_2).

^{31}P NMR (243 MHz, D_2O): δ = -142.91 (pdt, J = 864.7, 691.5, 126.9 Hz).

ESI-HRMS (m/z): $[M]^-$ calculated for ($[C_{18}H_{24}F_7N_3O_3P])^-$: 494.1443 Da; found: 494.1445 m/z

Ammonium N-acetyl- 4-(pentafluorophosphato-carbonyl)-L-phenylalaninyl-L-leucyl-amide (15)



The dipeptide **15** was synthesized according to the general method, but cleaved from the resin over 6 h and with the addition of 1% H_2O to the cleavage mixture. From 155 mg Rink amide resin, 16.1 mg (73%) of the product were obtained as white lyophilisate.

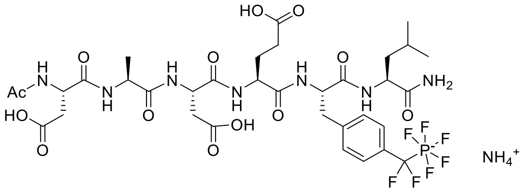
^{1}H NMR (600 MHz, D_2O): δ = 7.97 (d, J = 8.0 Hz, 2H, Ar-H), 7.37 (d, J = 8.2 Hz, 2H, Ar-H), 4.61 (t, J = 7.5 Hz, 1H, αH-Y*), 4.24 (dd, J = 10.0, 4.1 Hz, 1H, αH-Leu), 3.13 (d, J = 7.6 Hz, 2H, βH-Y*), 1.97 (s, 3H, $NHCOCH_3$), 1.64 – 1.40 (m, 3H, βH+γH-Leu), 0.84 (dd, J = 35.2, 5.8 Hz, 6H, δH-Leu).

^{19}F NMR (565 MHz, D_2O): δ = -64.76 (dd, J = 891.9, 44.8 Hz, 4F, F_{eq}), -67.65 (dp, J = 710.5, 44.5 Hz, 1F, F_{ax}).

^{31}P NMR (243 MHz, D_2O): δ = -145.76 (pd, J = 892.1, 710.1 Hz).

ESI-HRMS (m/z): $[M]^-$ calculated for $([C_{18}H_{24}F_5N_3O_4P])^-$: 472.1424 Da; found: 472.1421 m/z

N-Acetyl-L-aspartyl-L-alaninyl-L-aspartyl-L-glutamyl-4-(pentafluorophosphato-difluoromethyl)-L-phenylalaninyl-L-leucyl-amide (16)



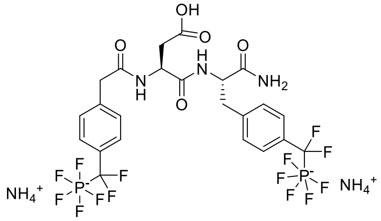
The dipeptide **16** was synthesized according to the general method. From 94 mg Rink amide resin, 15.3 mg (65%) of the product were obtained as white lyophilisate.

^{1}H NMR (600 MHz, D_2O) δ = 7.43 – 7.26 (m, 2H, Ar-H), 7.25 – 7.15 (m, 2H, Ar-H), 4.52 – 4.11 (m, 6H, CHN), 3.15 – 2.89 (m, 2H, CH_2Phe), 2.64 (dd, J=63.7, 15.9, 4H, 2xCH_2Asp), 2.17 (m, 2H, CH_2Glu), 1.91 (s, 3H, Ac), 1.81 (dd, J=29.9, 7.5, 2H, CH_2Glu), 1.56 – 1.37 (m, 3H, CH_2CH) 1.27 (m, 3H, CH_3Ala), 0.76 (d, J=36.1, 6H, 2xCH_3Leu)

^{19}F NMR (565 MHz, D_2O) δ = -67.75 (p, J=42.9, F_{ax}), -68.97 (p, J=44.0, 42.4, F_{ax}), -70.69 – -71.76 (d, J=42.9, 2F, F_{eq}), -72.78 (d, J=42.9, 2F, F_{eq}), -97.98 (dd, J=126.4, 8.7, 2F, CF_2)

HRMS (ESI): $[M]^-$ calculated for $C_{34}H_{46}F_7N_7O_{13}P^-$: 924.2785 Da, found: 924.2795 m/z

Bis-ammonium 4-(pentafluorophosphato-difluoromethyl)-phenylacetamidyl-L-aspartyl- 4-(pentafluorophosphato-difluoromethyl)-L-phenylalaninyl-amide (17)



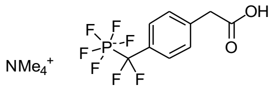
Compound **17** was synthesized applying the general methods for peptide synthesis starting from resin **13** (0.2 g, loading 0.34 mmol/g). Following to the final Fmoc cleavage, compound **18** (5 equiv.) was pre-activated with TBTU (4.9 equiv.) and DIPEA (10 equiv.) in DMF and then added to the N-unprotected dipeptide resin for 2 h. Cleavage and purification was conducted again as described in the general methods. From 159 mg of resin, 11.5 mg (46%) of the product **17** were obtained as white lyophilisate.

^{1}H NMR (600 MHz, D_2O) δ = 7.37 (t, J=9.5, 4H, Ar-H), 7.17 (d, J=7.8, 4H, Ar-H), 4.52 – 4.43 (m, 2H, CHN), 3.45 (s, 2H, CH_2Phe), 3.10 (d, J=13.3, 1H, CH_2Phe), 2.86 (dd, J=13.2, 10.2, 1H, CH_2Phe), 2.43 (ddd, J=96.3, 15.9, 7.2, 2H, CH_2Asp)

^{19}F NMR (565 MHz, D_2O) δ = -67.53 – -68.13 (m), -69.00 (h, J=43.2, 42.7, 2F, F_{ax}), -72.04 (dddd, J=864.7, 42.9, 25.4, 7.9, 8F, F_{eq}), -98.00 (dd, J=126.4, 8.7, 4F, 2xCF_2)

HRMS (ESI): $[M]^-$ calculated for $C_{23}H_{21}F_{14}N_3O_5P_2^{2-}$: 373.5372 Da, found: 373.5387 m/z

Tetramethyl 4-(pentafluorophosphato-difluoromethyl)-phenylacetic acid (18)



The methyl ester **21** (320 mg, 0.985 mmol, 1 equiv.) was added to B*acillus licheniformis* protease (from Sigma Aldrich) (20 mg, 20 mg /mmol) in aqueous buffer (NH_4HCO_3, 50 mM, pH = 7.8, 30 ml) and stirred at 50 °C overnight. The reaction mixture was concentrated under reduced pressure. The residue was purified via column chromatography at MPLC (RP C18, ACN / H_2O + 10 mM NH_4HCO_3, 5 to 99%). Fractions containing the product were concentrated at rotary evaporator and lyophilized, yielding the product **18** (255 mg, 83%) as a white lyophilisate.

^{1}H NMR (500 MHz, MeOD-d_4) δ = 7.41 (d, J=7.7, 2H, Ar-H), 7.26 (d, J=7.9, 2H, Ar-H), 3.57 (s, 2H, CH_2Phe), 3.09 (s, 12H, NMe_4^+)

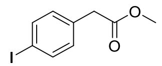
^{13}C NMR (126 MHz, MeOD-d_4) δ = 176.15 (s, C=O), 135.86, 127.86, 125.55 (s, 6 x Ar-H), 54.53 (s, $NMe_4^{+)}$, 42.27 (s, CH_2Phe)

^{19}F NMR (471 MHz, MeOD-d_4) δ = -71.18 (dp, J=696.1, 43.6, 1F, F_{eq}), -72.30 (d, J=43.6, 2F, F_{ax}), -74.13 (d, J=41.4, 2F, F_{ax}), -99.54 (d, J=124.3, 2F, CF_2)

^{31}P NMR (202 MHz, MeOD-d_4) δ = -143.37 (pdt, J=861.4, 694.0, 122.9)

HRMS (ESI): $[M]^-$ calculated for $C_9H_7F_7O_2P^-$: 311.00774 Da, found: 311.00795 m/z

Methyl 4-iodophenylacetate (19)



4-Iodophenylacetic acid (from ABCR, 5 g, 19.1 mmol, 1 equiv.) was dissolved in dry MeOH (25 ml) in a heat- and vacuum-dried Schlenk flask under Ar atmosphere. 3 drops of dry DMF were added and the solution was cooled to 0 °C. Oxalyl chloride (3.27 ml, 38.2 mmol, 2 equiv.) was added dropwise under stirring and the reaction was allowed to reach RT for 16 h. The amber solution was then concentrated using a rotary evaporator, diluted with EtOAc and washed with H_2O, saturated $NaHCO_3$ solution, and brine. The organic layer was dried over Na_2SO_4, filtrated and concentrated in vacuo. The crude was purified via column chromatography at MPLC (SiO_2, EtOAc / Hex, 5 to 100%), product **19** (3.7 g, 70%) was obtained as a colorless oil.

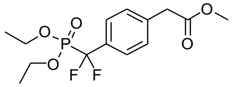
^{1}H NMR (500 MHz, $CDCl_3$) δ = 7.64 (d, J=8.4, 2H, Ar-H), 7.02 (d, J=8.5, 2H, Ar-H), 3.69 (s, 3H, OMe), 3.56 (s, 2H, CH_2)

^{13}C NMR (126 MHz, $CDCl_3$) δ = 171.54 (C=O), 137.76, 133.66, 131.37, (Ar-C), 92.73 (C-I), 52.25 (OMe), 40.73 (CH_2Phe)

HRMS (ESI): $[M+H]^+$ calculated for $C_9H_{10}IO_2^+$: 276.97200 Da, found: 276.97089 m/z;
$[M+Na]^+$ calculated for $C_9H_9INaO_2^+$: 298.95394 Da, found: 298.95356 m/z

Spectral data were consistent with published values.[5]

Methyl 4-(diethyl-phosphonato-difluoromethyl)-phenylacetate (20)



A heat- and vacuum-dried Schlenk flask was charged with metallic cadmium (1832 mg, 16.9 mmol, 6 equiv.) activated and dried as previously described and dry DMF (8 ml). To the stirred suspension, diethyl bromophosphonate (1.595 ml, 8.965 mmol, 3.33 equiv.) was added dropwise at RT. The slightly exothermic reaction was stirred for 3 h. In another flask, previously dried 4-iodophenylacetic acid methylester **19** (750 mg, 2.72 mmol, 1 equiv.) was dissolved in dry DMF (1 ml) and CuBr (1169 mg, 8.15 mmol, 3 equiv.) was added. The solution containing the organocadmium was added slowly and dropwise to this stirred mixture under Ar atmosphere and the reaction mixture was stirred for 16 h and monitored via TLC (1:3, EtOAc / Hex). After addition of EtOAc, the precipitate was filtrated off over a bed of Celite and the filtrate washed with sat. aq. sol. NH_4Cl (20 ml, 3x), H_2O (20 ml) and brine (20 ml). The organic layer was dried over Na_2SO_4, filtrated and concentrated under reduced pressure. After purification of the crude via chromatography (SiO_2, EtOAc / Hex 5 to 100%), product **20** (836 mg, 99%) was isolated as a colorless oil.

R_f = 0.4 (50% EtOAc/hexane)

^{1}H NMR (500 MHz, $CDCl_3$) δ = 7.57 (d, J=7.9, 2H, Ar-H), 7.37 (d, J=8.3, 2H, Ar-H), 4.21 – 4.09 (m, 4H, OCH_2CH_3), 3.69 (s, 3H, OMe), 3.66 (s, 2H, CH_2), 1.30 (t, J=7.1, 6H, CH_2CH_3)

^{13}C NMR (126 MHz, $CDCl_3$) δ = 171.43 (C=O), 136.91 (Ar-C), 131.64 (Ar-C), 129.48 (Ar-C), 129.47 (Ar-C), 126.58 (Ar-C), 118.97 (CF_2), 64.88 (2x CH_2 ethyl), 52.25 (OMe), 41.01 (CH_2), 16.43 (2x CH_3 ethyl)

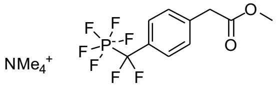
^{19}F NMR (471 MHz, $CDCl_3$) δ = -108.14 (d, J=116.5)

^{31}P NMR (202 MHz, $CDCl_3$) δ = 6.92 (d, J=234.3)

HRMS (ESI): $[M+H]^+$ calculated for $C_{14}H_{20}F_2O_5P^+$: 337.10109 Da, found: 337.10247 m/z
$[M+Na]^+$ calculated for $C_{14}H_{19}F_2NaO_5P^+$: 359.08304 Da, found: 359.08386 m/z

Spectral data were consistent with published values.[5]

Tetramethylammonium O-methyl-4-(pentafluorophosphato-difluoromethyl)-phenylacetate (21)



Diethyl phosphonic acid ester **20** (670 mg, 1.99 mmol, 1 equiv.) was dissolved in a Schlenk flask in dry ACN (10 ml). TMSBr (1.314 ml, 9.96 mmol, 5 equiv.) was added dropwise under inert atmosphere. The solution was heated at 60°C for 1.5 h. After disappearance of the starting material monitored via LC-MS, the vial was equipped with inert gas inlet and outlet to allow the release of the gaseous components developed after dropwise addition of dry DMF (766 µL, 9.96 mmol, 5 equiv.) and $(COCl)_2$ (1.67 ml, 19.92 mmol, 10 equiv.). After gas development ceased, the reaction was heated in a sealed vessel under inert atmosphere at 40 °C with a water bath. After 1.5 h, a small aliquot was taken and MeOH was added. The formation of the dimethyl ester was confirmed via LCMS. The reaction mixture was then cooled to 0°C with an ice bath. Previously weighted under inert atmosphere and dried as previously described NMe_4F (1856 mg, 19.92 mmol, 10 equiv.) was then added slowly under inert atmosphere to the stirred and cooled reaction mixture. After 1 h, the mixture was slowly quenched in a cooled sat. aq. sol. of $NaHCO_3$ (30 ml) and extracted with DCM (3 x 30 ml). The collected organic layers were then concentrated at the rotary evaporator, redissolved in H_2O/ACN and purified with RP-MPLC. After purification of the crude via chromatography (RP-C18, ACN / H_2O, 5 to 99%), the fractions were analyzed at LCMS and the one containing the product were concentrated at rotary evaporator and lyophilized, yielding the product **21** (320 mg, 49%) as white lyophilisate.

^{1}H NMR (600 MHz, ACN-d3) δ = 7.40 – 7.32 (m, 2H, Ar-H), 7.21 (d, J=8.1, 2H, Ar-H), 3.62 (s, 5H, CH_2, OMe), 3.02 (s, 12H, NMe_4)

^{13}C NMR (151 MHz, ACN-d3) δ = 172.01 (O=C), 134.40 (d, J=2.1), 129.69 (d, J=1.2), 128.41 (Ar-C), 125.72 (t, J=7.5, CF_2), 55.20 (NMe_4), 51.57 (OMe), 40.14 (CH_2Phe)

^{19}F NMR (565 MHz, ACN-d3) δ = -70.09 (dp, J=696.7, 43.4, 42.6, F_{ax}), -71.75 (ddt, J=855.5, 43.1, 9.4, 4F, F_{eq}), -98.39 (dp, J=120.6, 9.3, 2F, CF_2)

^{31}P NMR (243 MHz, ACN-d3) δ = -143.92 (pdt, J=856.2, 696.6, 120.0)

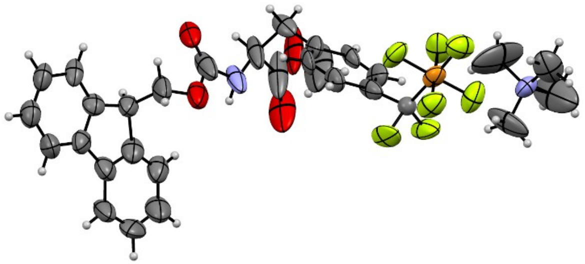
HRMS (ESI): $[M]^-$ calculated for $C_{10}H_9F_7O_2P^-$: 325.0234 Da, found: 325.0241 m/z

Crystal structure determination

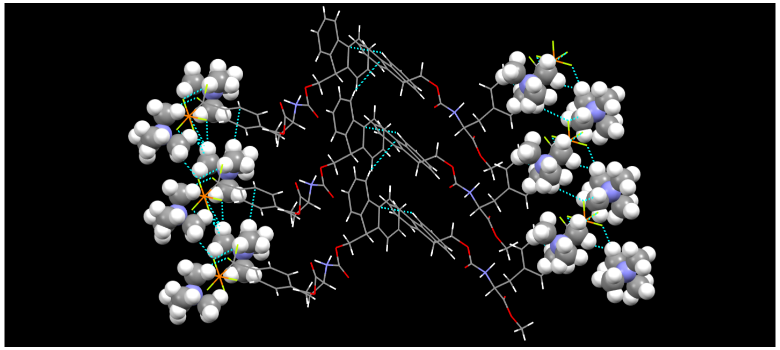
To obtain single crystals of **8**, the compound was dissolved in a 1:1 mixture of water and methanol, transferred into a beaker and sealed with parafilm. Several small holes were punctured into the film to allow slow evaporation of the methanol. After four weeks at room temperature single crystals had grown. Single crystals X-ray diffraction was performed on a Bruker D8 Venture system with graphite-monochromatic Mo-Kα radiation (λ = 0.71073 Å). Data reduction was performed with Bruker AXS SAINT[6] and SADABS[7] packages. The structure was solved by SHELXS 2018[8] using direct methods and followed by successive Fourier and difference Fourier synthesis. Full matrix least-squares refinements were performed on F2 using SHELXL 2018[8] with anisotropic displacement parameters for all non-hydrogen atoms. All other calculations were carried out using SHELXS 2018[8], SHELXL 2018[7] and WinGX (Ver-1.80)[9]. Mercury 2020.1[10] and Diamond 4.6.5[11] were used for structure visualization. Data collection, structure refinement parameters and crystallographic data of compounds **8** are summarized in Supplementary Table 2.

Supplementary Table 2. Crystallographic data of compound **8**

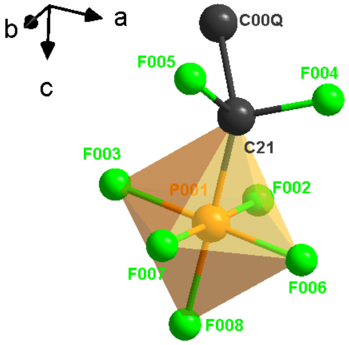
Identification code	**8**	
Empirical formula	C30 H35 F7 N2 O4 P	
Formula weight	651.57	
Temperature	293(2) K	
Wavelength	0.71073 Å	
Crystal system	Orthorhombic	
Space group	P 21 21 21	
Unit cell dimensions	a = 6.6335(10) Å	a= 90°.
	b = 13.663(3) Å	b= 90°.
	c = 33.745(5) Å	g = 90°.
Volume	3058.4(9) Å^{3}	
Z	4	
Density (calculated)	1.415 Mg/m^{3}	
Absorption coefficient	0.171 mm^{-1}	
F(000)	1356	
Crystal size	0.18 x 0.22 x 0.29 mm^{3}	
Theta range for data collection	2.345 to 23.277°.	
Index ranges	-6<=h<=7, -15<=k<=15, -37<=l<=37	
Reflections collected	24008	
Independent reflections	4391 [R(int) = 0.1564]	
Completeness to theta = 25.242°	80.5 %	
Refinement method	Full-matrix least-squares on F^{2}	
Data / restraints / parameters	4391 / 0 / 406	
Goodness-of-fit on F^{2}	1.119	
Final R indices [I>2sigma(I)]	R1 = 0.1063, wR2 = 0.2783	
R indices (all data)	R1 = 0.1932, wR2 = 0.3354	
Absolute structure parameter	0.1(6)	
Extinction coefficient	0.003(3)	
Largest diff. peak and hole	0.634 and -0.403 e.Å^{-3}	



Supplementary Figure 1. Crystal structure of compound **8**



Supplementary Figure 2. Crystal structure of compound **8**



Bond lengths:

P001	F006	1.5485(101)	
		F007	1.5774(92)
		F002	1.6109(93)
		F003	1.6191(90)
		F008	1.6237(115)
		C21	1.6833(115)

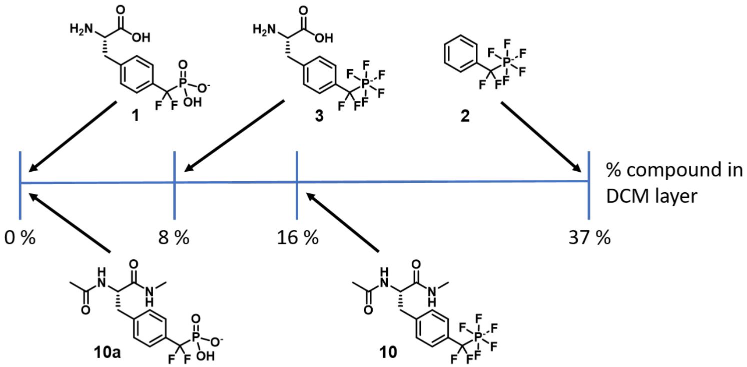
Bond angles:

P001	F006	F007	89.656(548)	
		F006	F002	91.873(529)
		F006	F003	178.168(531)
		F006	F008	90.256(574)
		F006	C21	91.517(615)
		F007	F002	178.380(514)
		F007	F003	88.599(535)
		F007	F008	89.025(584)
		F007	C21	92.528(566)
		F002	F003	89.877(516)
		F002	F008	90.432(570)
		F002	C21	87.967(565)
		F003	F008	90.273(568)
		F003	C21	88.002(594)
		F008	C21	177.649(619)

Supplementary Figure 3: Bond angles and lengths of the R-CF_2-PF_5 group of **8**

Determination of partition coefficients

Compounds were weighted and added to 50 ml round bottom flasks. 10 ml DCM and 10 ml water (MilliQ, 0.055 µS) were added and the mixture stirred for 24 h at room temperature. The phases were separated, evaporated and the remaining solids weighted to determine the concentration ratio between the two phases and thereof the logarithmic partition coefficient log P.



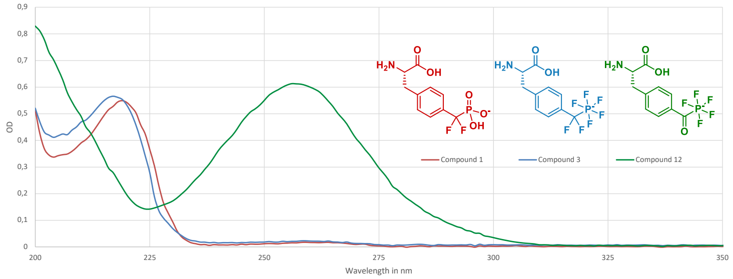
Supplementary Figure 4. Visual representation of the partition of compounds **1**, **2**, **3**, **10** and **10a** between water and DCM

Supplementary Table 3. Partition of compounds **1**, **2**, **3**, **10**, **10a** between water and DCM

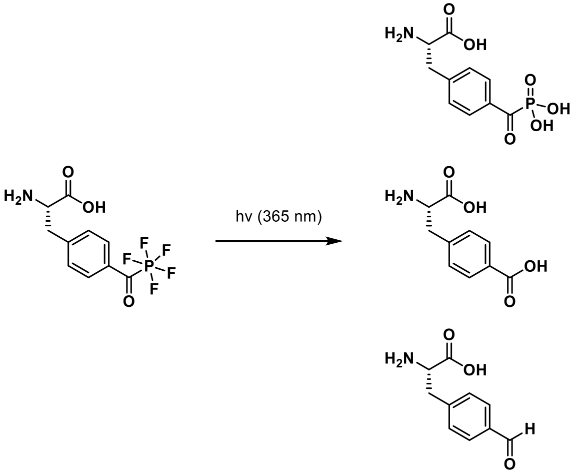
Compound	Percent compound in DCM	Log P (DCM/water)
1	< 1 %	-
2	37 %	-0.23
3	8.2 %	-1.05
10	16 %	-0.81
10a	< 1 %	-

UV spectroscopy and irradiation experiments

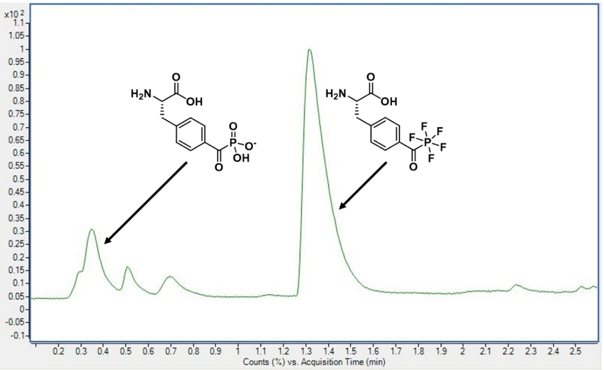
The three amino acids **1**, **3**, and **12** were dissolved in water (MilliQ, 0.055 µS) at a concentration of 0.5 mM. 50 µl of these solutions were added into a UV-star 96-well plate and an absorbance scan was performed. Compounds **1** and **3** displayed a similar absorbance pattern, with **1** giving a distinct maximum at 219 nm (ε = 17259 $M^{-1}cm^{-1}$) and **3** at 219 nm (ε = 17788 $M^{-1}cm^{-1}$). Compound **12** showed a maximum at 256 nm (ε = 19271 $M^{-1}cm^{-1}$) and a second peak at 343 nm (ε = 192 $M^{-1}cm^{-1}$) (see Supplementary Figure 71). Thus, substitution of the CF_2 group by a CO group leads to a shift of the maximum by 39 nm. The second peak can be attributed to the n-π* transition of the ketophosphonate. Accordingly, irradiation at 365 nm for 2 h led to the photoconversion of the amino acid **12**.



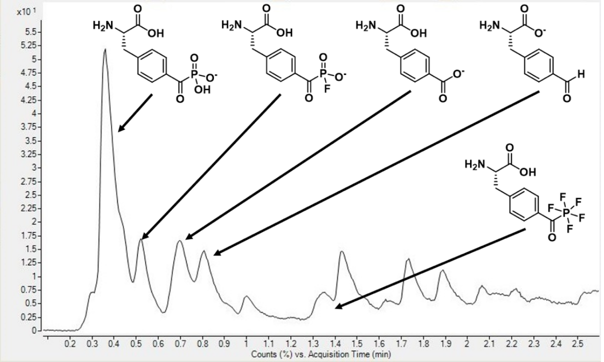
Supplementary Figure 5: UV spectra of 0.5 mM solutions of compounds **1**, **3** and **12** in water



Supplementary Figure 6: After irradiation at 365 nm for 2 h at room temperature in 70/30 iPrOH/H_2O compound **12** (7.5 mM) several derivatives of compound **12** were identified, namely the keto-phosphonate, carboxylic acid and aldehyde.



Supplementary Figure 7: HPLC chromatogram of a 7.5 mM solution of compound **12** in 70/30 iPrOH/H_2O prior to irradiation. Column B, eluent 5-95% ACN in 8 min, detector: total ion current (TIC).



Supplementary Figure 8: HPLC chromatogram of a 7.5 mM solution of compound **12** in 70/30 iPrOH/H_2O after irradiation at 365 nm for 2 h at room temperature. Column B, eluent 5-95% ACN in 8 min, detector: total ion current (TIC).

IR spectroscopy experiments

Methods

FTIR spectra were recorded on a Bruker Vertex 70 IR spectrometer, using an attenuated total reflection (ATR) element from IRubis and a custom-made PTFE cell. A background spectrum was first recorded with 300 µl of milliQ water, which were then replaced by 300 µl of a 10 mM aqueous sample solution of **1a** or **2**. The difference absorbance spectrum of the sample was recorded against the pure water background. All spectra were measured using a DTGS detector by averaging 4096 spectra at a resolution of 2 cm^{-1}. Data were baseline-corrected applying a 4th degree polynomial.

Band assignment

The theoretical IR spectra of the tetramethylammonium pentafluorophosphato-difluoromethyl-benzene **2** and the sodium phosphono-difluoromethyl-benzene fragment **1a** were calculated by density functional theory (DFT) with Gaussian 16 software using the B3LYP/6-31++G(d,p) level of theory. Spectra of **2** and **1a** samples between 4000 and 400 cm^{-1} were recorded in water solution and in transmission using KBr pellets (not shown), the latter to improve the signal to noise ratio and to avoid interference from water bands during the band assignment.

Bands between 1380 and 810 cm^{-1} were assigned to the stretching vibrations of the CF_2 spacer and are observed in both compounds. The beating vibrations of the phenyl group appear at ~1630 cm^{-1} in both spectra, overlapped by the O-H deformation mode from water (solvent). In both DFT models, CF_2 stretching modes were strongly coupled with the a1 and b2 beating vibrations of the phenyl group. As a result, multiple CF_2 stretching bands can be observed (Supporting Table 3) instead of the two expected from a single CF_2 group (symmetric and asymmetric stretching).

In **1a**, the stretching vibrations from the CF_2 appear in the experimental spectra between 1354 cm^{-1} and 824 cm^{-1} overlapped with the P-O stretching and P-O-H deformation vibrations. In the DFT simulation, CF_2 stretching modes were strongly coupled to these vibrations. Simulation was performed for different protonation degrees of the PO_3 headgroup and the monoprotonated form was found to reproduce better the experimental bands (Supplementary Table 3). The band at 1198 cm^{-1}, assigned to one of the P-O stretching and P-O-H deformation modes, did not appear in the monosodium derivative but could be observed in the disodium salt (deprotonated) at 1125 cm^{-1}. The experimental spectrum therefore suggests that the monoprotonated (monoanion) and deprotonated (dianion) forms of the phosphonate group coexist in solution under these conditions.

Supplementary Table 4. Band assignment of **2** and **1a** fragments from comparison with DFT calculations. Abbreviations: v=stretching, δ=deformation, asym=asymmetric, sym=symmetric, //=parallel to.

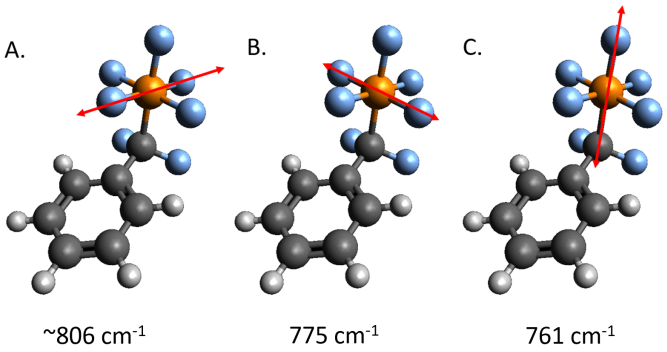
2

DFT	Exp.	Assignment
wavenumber / cm^{-1}	wavenumber / cm^{-1}	
776.5	761	v(PF) axial // b1
808.1	775	v(PF) in plane // b2
814.7	~806	v(PF) in plane // a1
1037.1	1035	v(CF_2)$_{asym}$
1090.1	1103	v(CF_2)$_{sym}$
1254.1	1249	v(CF_2)$_{sym}$

1a

DFT (monoNa)	Exp.	Assignment
wavenumber / cm^{-1}	wavenumber / cm^{-1}	
1036.4	1033	v(CF2)$_{asym}$, v(PO), δ(POH)
1059.6	1069	v(CF2)$_{sym}$, v(PO), δ(POH)
1090	1100	v(CF2)$_{sym}$, v(PO)
1116.7	1133	v(CF2)$_{asym}$
*	1198	v(PO), δ(POH)
1286.1, 1251.3	1251	v(PO), δ(POH), v(CF2)$_{sym}$

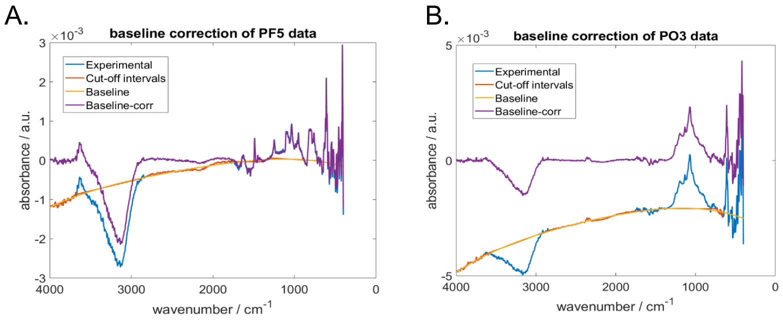
The characteristic PF stretching vibrational bands of the fragment **2** appear at lower frequencies between 900 and 710 cm^{-1}. In both, experimental and theoretical spectra, 3 peaks could be distinguished at ~806, 775 and 761 cm^{-1} (aqueous solution) and at 815, 808 and 777 cm^{-1} (DFT). These were assigned to the in-plane PF stretching vibration along the a1 direction of the phenyl group (Supplementary Figure 3, A), to the in-plane PF stretching along the b2 direction of the phenyl group (Supplementary Figure 3, B), and to the axial (terminal) PF stretching (Supplementary Figure 3, C) along the b1 direction of the phenyl group. Fluorination of the phosphorous also decouples the CF_2 stretching modes from those of the headgroup, which can be distinguished clearly in the **2** experimental spectra.



Supplementary Figure 9. The three PF stretching vibrations with experimental values

Baseline correction

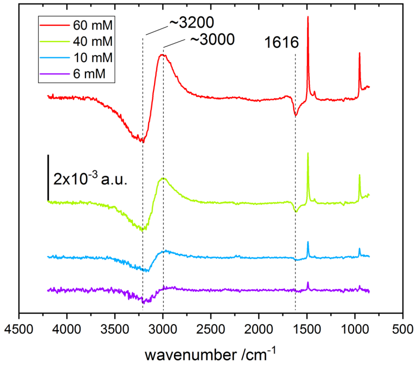
Experimental spectra were baseline corrected by manually selecting those intervals where no specific bands neither from the compound nor from water were observed. The ends of these intervals were connected by straight lines and data were averaged using a sliding window with a length of 200 data points (~1/20 of total spectral interval) over the complete spectral range. The resulting trace was fitted by a 4th degree polynomial to obtain the baseline, which was then subtracted from the experimental data. A visual representation of all steps can be found in Supplementary Figure 4.



Supplementary Figure 10. Baseline correction of experimental data from A.) **2** and B.) **1a** in 10 mM aqueous solution.

Hydration shell of tetramethylammonium counterion

The hydration shell of the tetramethylammonium counterion was evaluated in an independent experiment to discard it as the cause of the observed dangling-water specific bands. The difference spectrum of a TMA fluoride aqueous solution against a pure-water background was recorded at different concentrations (Supplementary Figure 5). The spectra showed two negative peaks at ~3200 cm^{-1} and at ~1620 cm^{-1}, caused by the reduction in water concentration, and linked to water molecules fully exposed to hydrogen bonding. No dangling-water specific bands (~3630 cm^{-1}) were observed at any concentration.

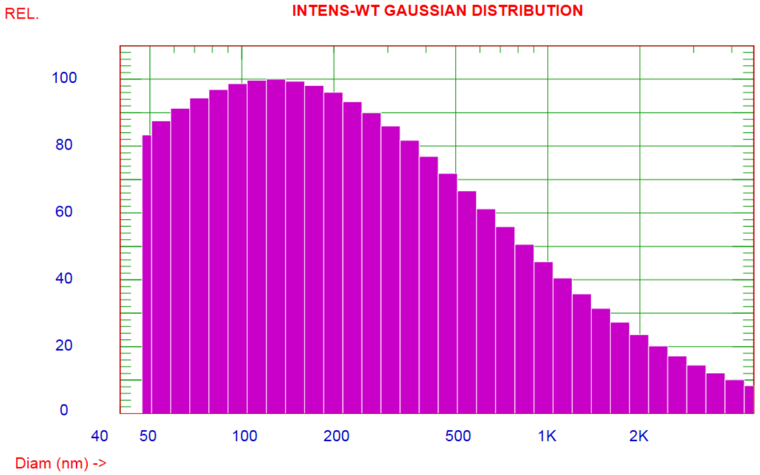


Supplementary Figure 11. Spectra of tetramethylammonium fluoride at different concentrations versus a pure water background. Spectra recorded with 512 co-additions with a LN-MCT detector

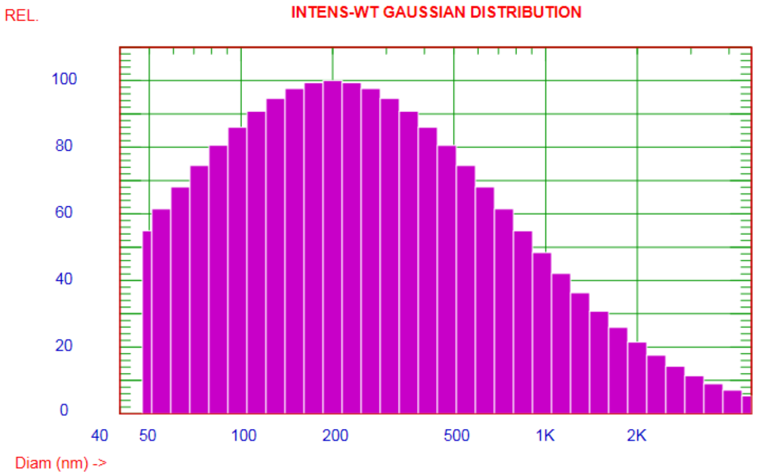
Dynamic Light Scattering

A clear solution of compound **16** in DMSO / buffer 1:1 (650 µM) was subjected to dynamic light scattering indicated the formation of nanoparticular aggregates with a mean diameter of 427 nm (median ca. 120 nm). Dilution of this sample with 1:1 buffer/DMSO to a concentration of 325 µM led to a slight change in particle size distribution, with a mean diameter of 474 nm (median 180 nm).

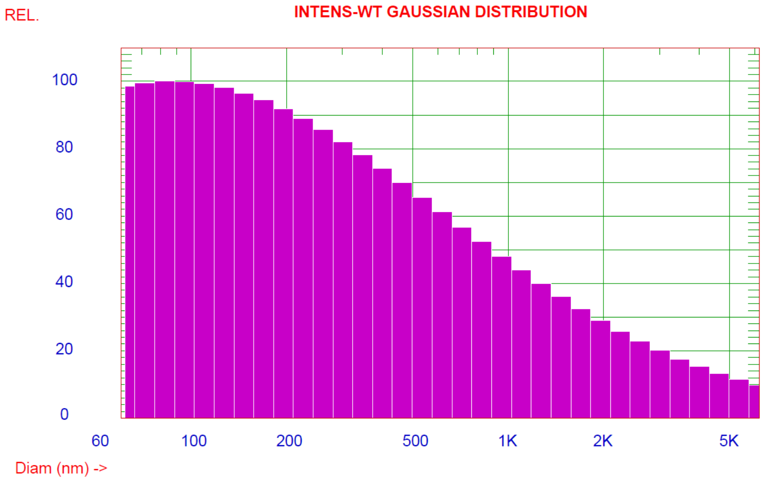
Furthermore, two samples of **16** in pure DMSO were diluted with buffer to contain 5% DMSO (assay conditions). Here, at concentrations of 250 µM and 125 µM, nano particles with a mean diameter of 619 nm and 776 nm, respectively, were observed. These results indicate that **16** was not fully dissolved under assay conditions.



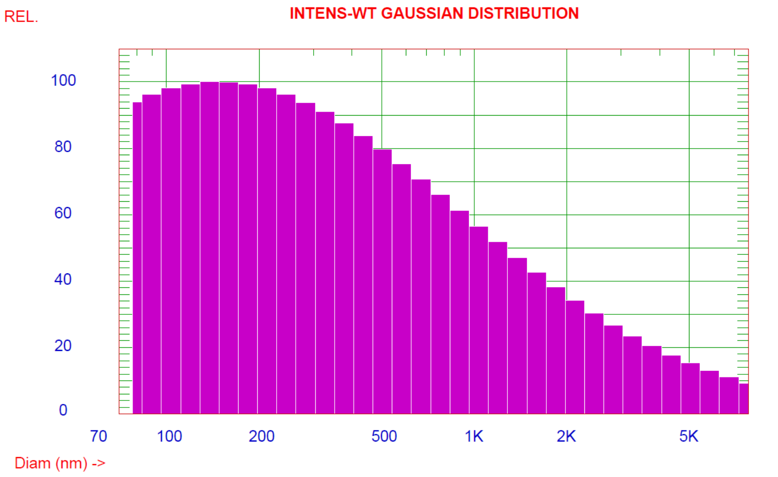
Supplementary Figure 12. DLS of compound **16** at an apparent concentration of 650 µM (50% DMSO) shows undissolved compound in the form of nano particles. These have a mean diameter of 471.7 nm.



Supplementary Figure 13. DLS of compound **16** at an apparent concentration of 325 µM (50% DMSO) shows undissolved compound in the form of nano particles. These have a mean diameter of 473.7 nm.

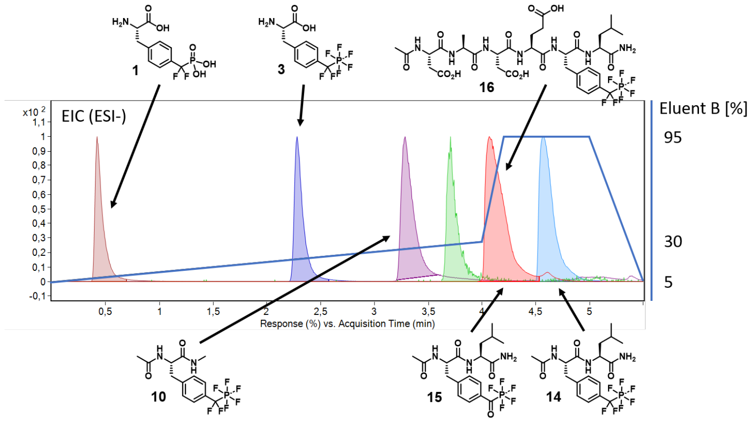


Supplementary Figure 14. DLS of compound **16** at an apparent concentration of 250 µM (5% DMSO) shows undissolved compound in the form of nano particles. These have a mean diameter of 618.5 nm.



Supplementary Figure 15. DLS of compound **16** at an apparent concentration of 125 µM (5% DMSO) shows undissolved compound in the form of nano particles. These have a mean diameter of 775.5 nm.

HPLC retention times



Supplementary Figure 16. Overlayed HPLC chromatograms of compounds **1**, **3**, **10**, **14**, **15** and **16**, column B, eluent ACN, gradient as shown, EIC detector.

Biochemical evaluation of synthesized compounds

PTP1B

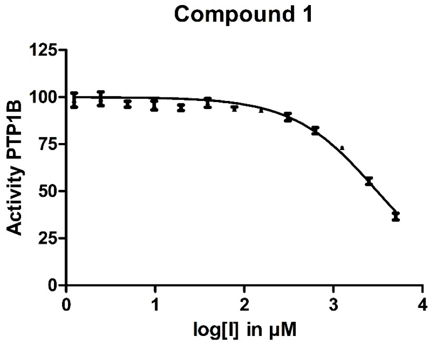
Recombinant human PTP1B was obtained from Abcam (ab51277) at a concentration of 100 µM and used as received, without further purification.

Enzymatic DiFMUP assay:

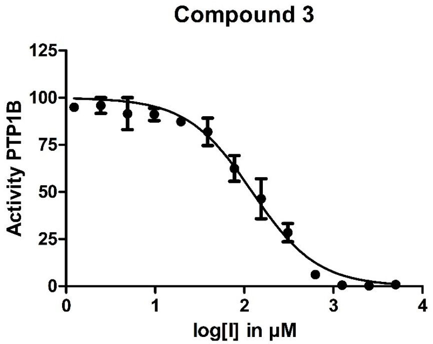
An enzyme assay with 6,8-difluoro-4-methylumbelliferyl phosphate (DiFMUP) as substrate was used to determine the activity of inhibitors toward PTP1B. In Method A, test compounds were dissolved and serially diluted in buffer (0% DMSO). In Method B, test compounds were dissolved in a 1:1 mixture of DMSO and buffer (20 mM stock) and serially diluted with the same mixture resulting in a final DMSO concentration of 2.5% in the assay. In Method C, compounds were dissolved and diluted in DMSO to a final DMSO concentration of 5% in the assay, and in method D, compounds were dissolved in DMSO and diluted with buffer to a DMSO concentration of 5%, serially diluted with 5% DMSO in buffer resulting in a final DMSO concentration of 2.5% DMSO in the buffer.

The assay buffer contained 50 mM MOPSO (pH 6.5), 200 mM NaCl, 0.03% Tween-20, 50 µM tris-(2-carboxyethyl)-phosphin (TCEP) (freshly added prior to each measurement) and 1.5 nM PTP1B (final concentration). The final assay volume was 20 µL. Enzyme and test compound in buffer solution were incubated for 30 min at RT. The reaction was started by adding DiFMUP to a final concentration of 67 µM. This substrate concentration matches the experimentally determined K_M value of the enzyme. Measurements were performed on a Genius Pro Reader (SAFIRE II, instrument serial number: 512000014) with the following settings: measurement mode: Fluorescence Top; λ_{ex}: 360 nm (bandwidth 20 nm); λ_{em}: 460 nm (bandwidth 20 nm) ; gain (manual): 60; number of scans: 8; FlashMode: high sensitivity; integration time: 40 µs; lag time: 0 µs; Z-position (manual): 13900 µM; number of kinetic cycles 10; kinetic interval: 60 s; total kinetic run time 10 min. Measurements were performed in triplicate. IC_{50} values were calculated with Prism 5 (for Windows, Version 5.01, GraphPad Software Inc.) and were converted into the corresponding K_I values applying the Cheng Prusoff equation.[12]

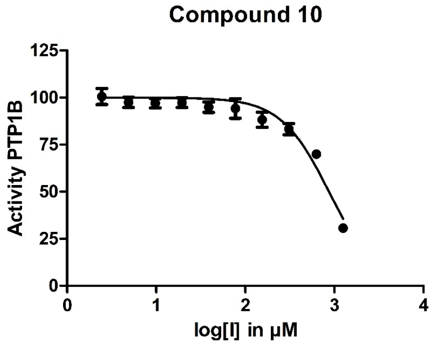
Inhibition of PTP1B



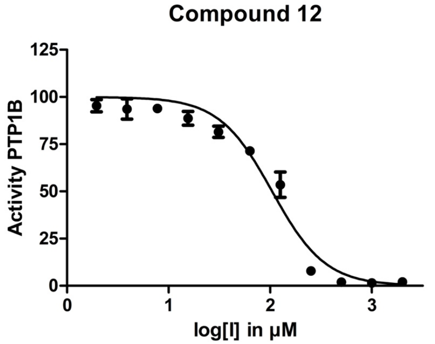
Supplementary Figure 17. Inhibition of PTP1B by compound **1**, IC_{50}: 3.1 ± 0.37 mM (Method A)



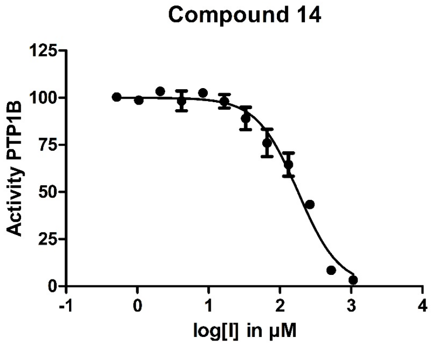
Supplementary Figure 18. Inhibition of PTP1B by compound **3**, IC_{50}: 122 ± 16 µM (Method B)



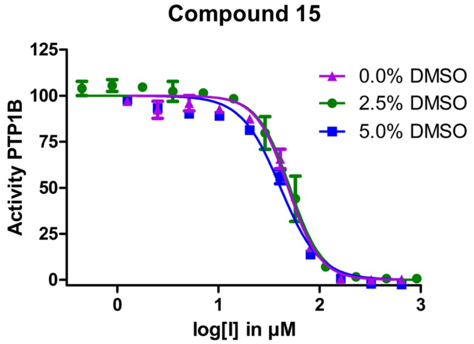
Supplementary Figure 19. Inhibition of PTP1B by compound **10**, IC_{50}: 872 ± 95 µM (Method B)



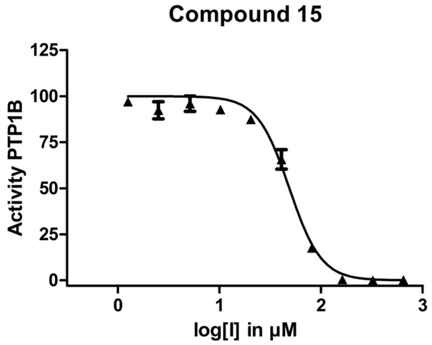
Supplementary Figure 20. Inhibition of PTP1B by compound **12**, IC_{50}: 104 ± 14 µM (Method B)



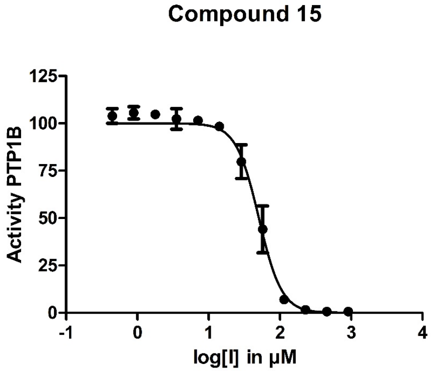
Supplementary Figure 21. Inhibition of PTP1B by compound **14**, IC_{50} = 180 ± 20 µM (Method B)



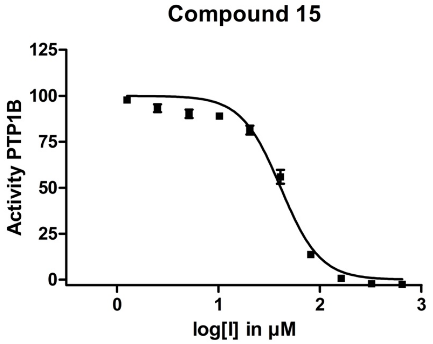
Supplementary Figure 22. Inhibition of PTP1B by compound **15** under different assay conditions. Method A (magenta): final DMSO concentration of 0%. Method B (green): final DMSO concentration of 2.5%. Method C (blue): final DMSO concentration of 5%.



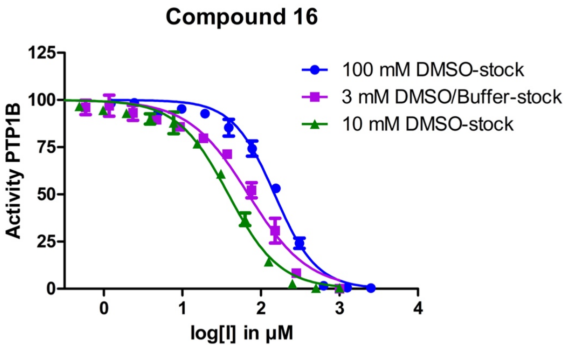
Supplementary Figure 23. Inhibition of PTP1B by compound **15**, IC_{50} = 48 ± 8 μM (Method A)



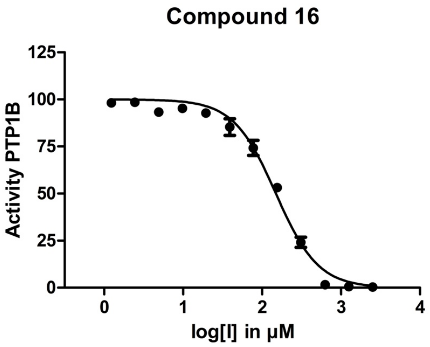
Supplementary Figure 24. Inhibition of PTP1B by compound **15**, IC_{50} = 50 ± 8 μM (Method B)



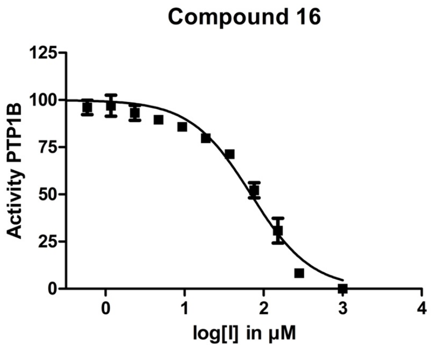
Supplementary Figure 25. Inhibition of PTP1B by compound **15**, IC_{50} = 41 ± 6 μM (Method C)



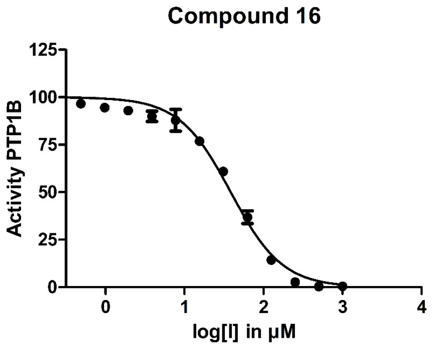
Supplementary Figure 26. Inhibition of PTP1B by compound **16** under different assay conditions. Method D (blue): final DMSO concentration of 2.5%. Method B (magenta): final DMSO concentration of 2.5%. Method C (green): final DMSO concentration of 5%.



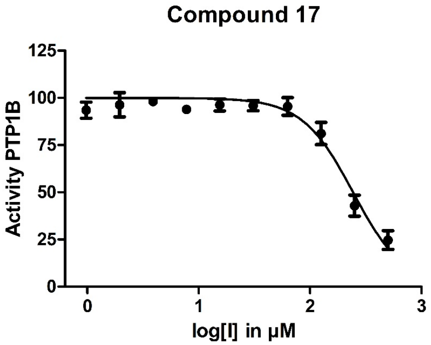
Supplementary Figure 27. Inhibition of PTP1B by compound **16**, IC$_{50}$ = 149 ± 26 µM (method D)



Supplementary Figure 28. Inhibition of PTP1B by compound **16**, IC$_{50}$ = 67 ± 10 µM (method B)



Supplementary Figure 29. Inhibition of PTP1B by compound **16**, *IC$_{50}$* = 38 ± 6 µM (method C)



Supplementary Figure 30. Inhibition of PTP1B by compound **17**, IC_{50} = 243 ± 23 µM (method B)

Computational methods

Protein preparation for docking

The protein X-ray diffraction crystal structure of PTP1B (PDB code: 4Y14)[13] was prepared for docking and simulations with Schrödinger's Protein Preparation Wizard.[14] The protonation states of amino acid sidechains were assigned with PROPKA at pH 7.0. Small molecules, crystal water and the A Chain of the dimer were deleted. The hydrogen-bond network was optimized and a brief molecular mechanics minimization using the OPLS4 force field[7] was run.

Ligand docking

The structures of 1 and 3 were docked to the binding pocket of PTP1B using Schrödinger's Glide [15] and OPLS4 force field[16]. A receptor grid was generated using the default setting with OH- and SH- groups within the binding pocket allowed to rotate. Ligand docking was performed with the XP protocol, which applies sampling based on anchors and refined growth as well as a scoring function which scores the docking poses based on physico-chemical descriptors. Non-planar amide conformations were penalized and halogens were included as weak noncovalent interaction acceptors of hydrogen bond type.

Molecular dynamics simulation

Molecular dynamics simulations with PTP1B without ligand and with ligands 1 and 3 were performed with GROMACS 2019-4[17] and our amended version of the AMBER14SB force field.[18] The protein was prepared in the same way as for docking and placed into a dodecahedric box of TIP3P[19] water with 1.1 nm distance of the box edges to the solute. The starting structure was energy minimized and equilibrated for 100 ps in the NVT ensemble, followed by a 1 ns equilibration in the NPT ensemble. Production run had a length of 100 ns, the integration timestep was 2 fs and a snapshot was saved every 1 ps. Periodic boundary conditions were used in all three directions. Covalent bonds to hydrogen atoms were treated as constraints. The applied thermostat was a velocity rescaling scheme[20] and the applied barostat the Parrinello-Rahman barostat[21]. The cut-off for Lennard-Jones and Coulomb interactions was set to 1.0 nm. For Coulomb interactions, the PME method[22] was used. The Verlet cut-off scheme was used to generate neighbor lists.

Molecular dynamics simulation analysis

Molecular dynamics simulations were analysed using the mdtraj[23] python package. Distances and RMSF were calculated using its built-in methods. For the detection of π-π interactions, the centroid of the two respective aromatic rings were calculated and their distance was measured. To be considered for a π-π interaction, the distance had to be less than 4.4 Å. Additionally, the angle between the unit vectors orthogonal to the ring planes had to be less than 30 degrees to be counted as π-π interaction.

Force field parametrization

Parameters were retained from the AMBER14SB force field where possible, however supplementary parameters were required to simulate the nonstandard amino acid structures.[18]

Missing bonded parameters were provided by the general Amber force field (GAFF) for organic molecules,[24] with the help of the acpype tool.[25] Although GAFF can describe an extensive variety of organic molecules, it does not provide a sufficient model for the sp^3d^2-hybridized phosphorous atom in **3**, necessitating the determination of additional bonded parameters using ab initio methods. The bond parameters and the angular force constant of GAFF atom type "p5" (phosphorous with four substituents) were reappropriated for phosphorous with six substituents. The geometry was approximated by an octahedral shape. This assumption was made based on a density functional theory (DFT) geometry optimization. The angle parameters 90° and 180° were chosen to enforce octahedral geometry. Missing torsional parameters were obtained via relaxed dihedral scans in 72 steps of 5° intervals at the HF/6-31G* level of theory. The torsional parameters k_ϕ, n and ϕ_s, used for proper dihedrals in the Amber force field family, were obtained by optimizing the function, $\mathrm{V_d}\left(\phi_\mathrm{ijkl}\right) = \mathrm{k}_\phi(1 + \cos(\mathrm{n}\phi - \phi_\mathrm{s})$ to fit the relaxed scan data using the SciPy module scipy.optimize.curve_fit.[26] Re-optimized Lennard Jones parameters, which better model hydrophobic properties, were used as nonbonding parameters for fluorine in place of GAFF parameters.[27]

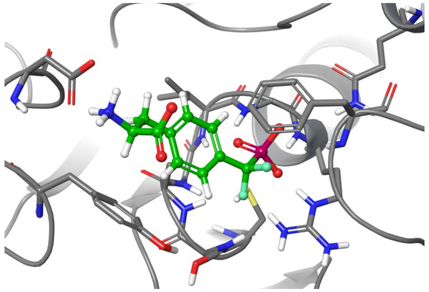
Partial atomic charges for the amino acids **1** and **3** were determined using a charge fitting procedure adapted from Robalo et al.[27] The method involves two iterations of the two stage restrained electrostatic potential (RESP) protocol,[28] in which the first iteration relies on a single conformation and the second iteration averages RESP-fitted charges over multiple conformations. The initial RESP-fitted charges were applied to simulate the free amino acid in TIP3P water for a production run of 100 ns in the NPT ensemble. Conformations at 1 ns intervals were extracted from the simulation and submitted to conventional two stage RESP fitting, such that the final partial atomic charges are based on the average values from 101 conformations. The RESP methodology was implemented using the antechamber program in the Ambertools package and ab initio calculations were performed in Gaussian 16.[29]

Docking studies

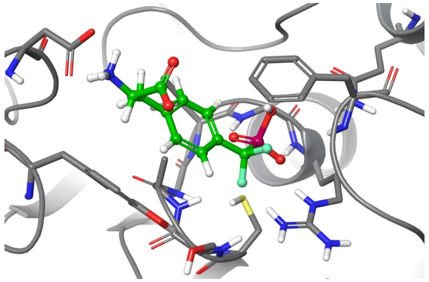
Structure and druggability of the PTP1B binding pocket

The binding site was assessed using Schrödinger's SiteMap[30], which applies a grid-based algorithm to detect and score binding pockets suitable for drug-like ligand binding based on electrostatic and geometric properties. A positively charged main pocket and a negatively charged side pocket were identified. The main pocket consists of the backbone NH-groups of residues 215-220. Also the positively charged sidechain of Arg221 can be found there. The negative charge in the side pocket can be allocated to the sidechain of Asp48 and the backbone carbonyl and sidechain C=O of Asn262. Between the main- and side pocket there are two aromatic rings of Tyr46 and Phe182. The pocket was evaluated by SiteMap's Dscore. The Dscore of the binding pocket is calculated to be 0.6, which rates it as an "undruggable" pocket, meaning it is difficult to address by drug-like ligands.[31]

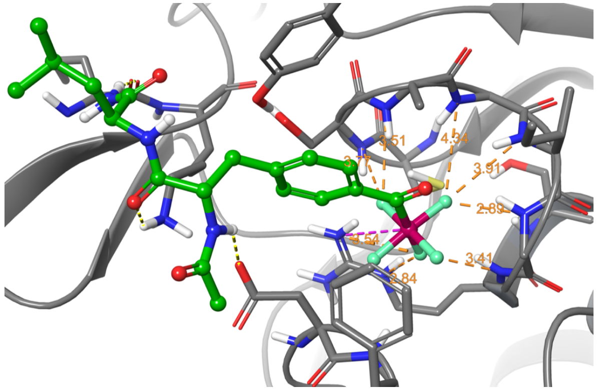
Docking poses



Supplementary Figure 31. Ligand **1** (di-anion) final docking pose in PTP1B binding pocket



Supplementary Figure 32. Protonated ligand **1** (mono-anion) final docking pose in PTP1B binding pocket



Supplementary Figure 33. Ligand **15**, final docking pose in PTP1B binding pocket

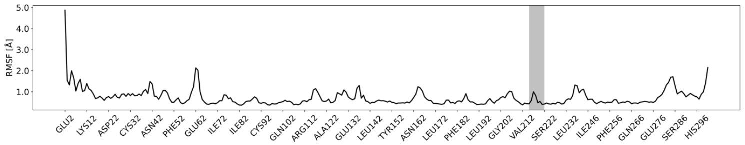
Docking score

The final docking poses of **1** and **3** were evaluated for their docking score, as seen in Table 4. Amino acid **3** scores better (more negative) than amino acid **1**. This result is expected as a better inhibition is measured for amino acid **3** as well, however the lower docking score does not necessarily imply a higher binding affinity of compound **1** as approximations in the Glide score "omit essential thermodynamics of the free energy of binding" and "accurately estimating ligand-protein affinities remains beyond the capabilities of docking scoring functions".[32]

Supplementary Table 6. Glide docking scores of amino acids **1** and **3**

Compound	Glide Docking Score
1 di-anion	-7.494
1 mono-anion	-7.793
3	-10.333

Flexibility of the PTP1B protein backbone

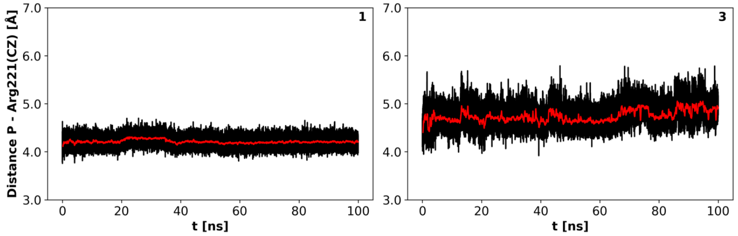


Supplementary Figure 34. RMSF of C$_\alpha$–atoms of each residue of PTP1B throughout a 10 ns MD simulation of the apo protein. The binding pocket area of residues 215-221 is highlighted in grey

The RMSF of C$_\alpha$-atoms of PTP1B were evaluated in an MD simulation to get a measure of the flexibility of the protein backbone, especially at the binding site region (Supplementary Figure 15). The backbone of the binding site region (residues 215-221) shows a little spike of flexibility but is in one of the least flexible areas of the protein.

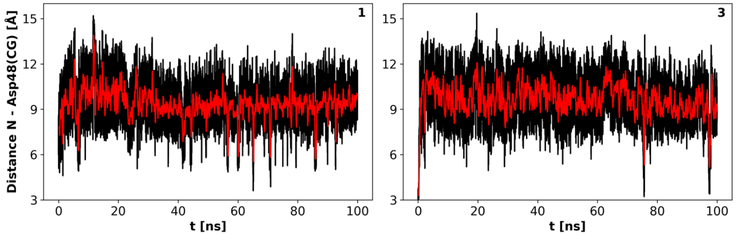
Ligand-protein key interactions throughout MD simulation

Both ligands remain stable inside the pocket throughout the 100 ns of simulation. Ligand-protein interactions found through docking were evaluated during molecular dynamics simulations of ligands **1** and **3** in complex with PTP1B.



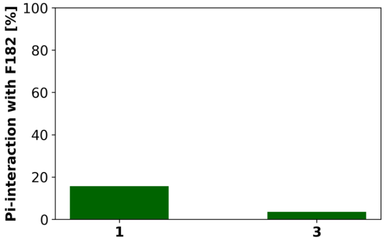
Supplementary Figure 35. Distance between phosphorus of ligand headgroup of **1** (left) and **3** (right). The moving average with a sliding window of 200 frames is shown in red

For both complexes it can be seen that the headgroup phosphorus remains close to the sidechain of Arg221, here indicated by sidechain carbon CZ. The distance is smaller and shows less fluctuation for ligand **1** than for ligand **3**.



Supplementary Figure 36. Distance between backbone nitrogen of ligands **1** (left) and **3** (right). The moving average with a sliding window of 200 frames is shown in red

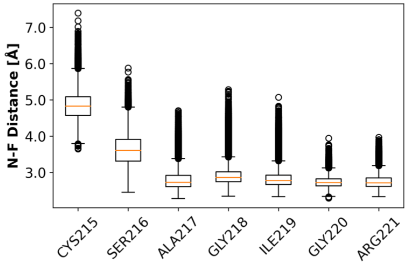
The distance between the ligands' positively charged backbone nitrogen and the Asp48 sidechain is too large for an intact salt bridge. This is already the case at the beginning of the simulation, after equilibration. It can be explained by the backbone ammonium ion turning outward and thereby preferring solvent exposure over the formation of the salt bridge.



Supplementary Figure 37. Percentage of frames in which a pi-interaction between the aromatic rings of the ligand and Phe182 of PTP1B is present

We do not observe π-π interaction of any ligand with Tyr46. We observe significantly more π-π interaction with Phe182 for **1** compared to **3**, but for both ligands this interaction occurs rarely.

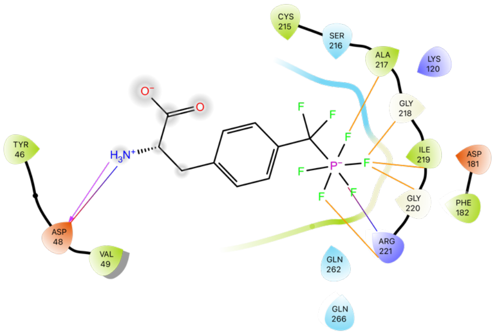
The backbone NH groups remain close to the fluorines of the PF$_5$ moiety, which implies the presence of N-H..F interactions. Supplementary Figure 19 shows the distribution of N..F distances inside the binding pocket. Specifically, the figure contains the distance of the respective backbone nitrogen to the closest fluorine of the PF_5 moiety. For residues 217-221, the closest N..F distance can be stated to be generally below 3 Å.



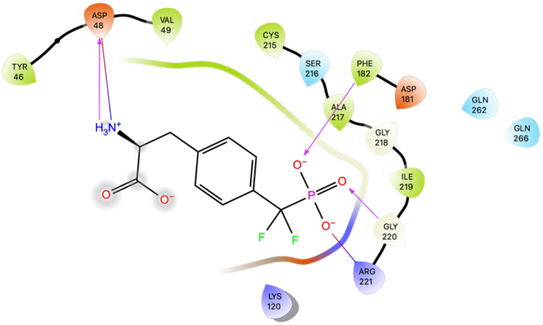
Supplementary Figure 38. Distance of backbone nitrogen of respective residue to the closest fluorine of the PF_5 moiety

Ligand-protein interaction in the PTP1B binding pocket

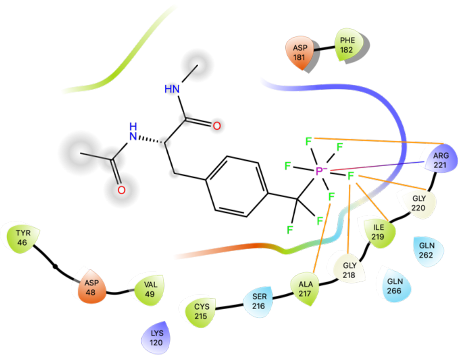
Figures 20-23 show the interactions of **1** and **3** as well as of their ACE/NME capped counterparts when bound to the pocket of PTP1B. The interaction was analysed using the final docking poses. For all ligands salt bridges are identified between phosphorus head group and Arg221, and between backbone nitrogen and Asp48. Hydrogen bonding, indicated by pink arrows, is detected to the phosphate group of **1** and between the backbone amide of the uncapped amino acids and Asp48. The capped derivative of **1** also shows hydrogen bonds between its backbone amide and Asp48.



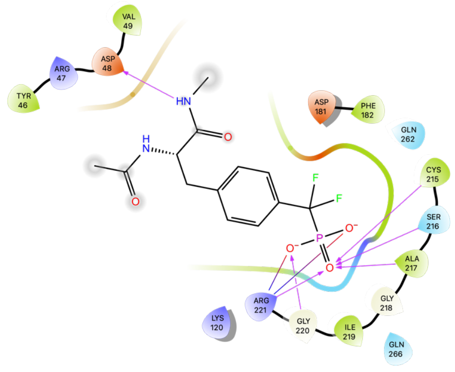
Supplementary Figure 39. Ligand interaction diagram of **3** in PTP1B binding pocket. Salt bridges are indicated by a straight line, hydrogen bonds by a pink arrow. The orange lines indicate contacts between backbone NH and F of the PF_5 moiety. Grey shaded atoms are solvent exposed. The binding pocket is represented by a line around the ligand with the color matching the closest amino acid



Supplementary Figure 40.2 Ligand interaction diagram of **1** in PTP1B binding pocket. Salt bridges are indicated by a straight line, hydrogen bonds by a pink arrow. Grey shaded atoms are solvent exposed. The binding pocket is represented by a line around the ligand with the color matching the closest amino acid

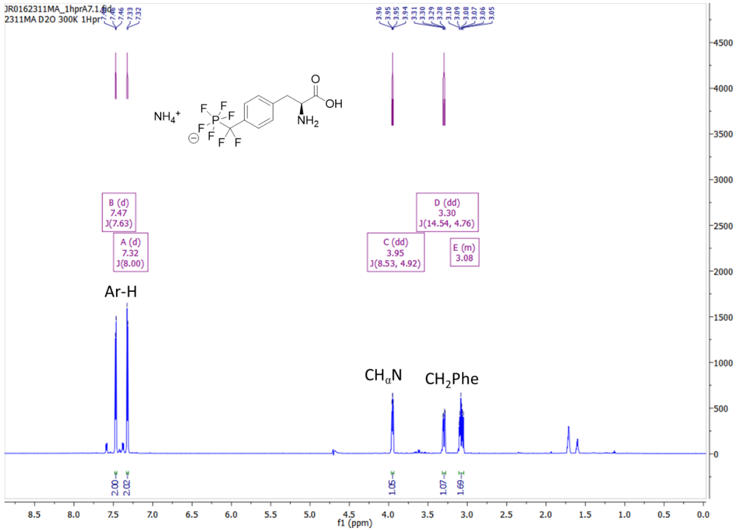


Supplementary Figure 41. Ligand interaction diagram of capped **10** in PTP1B binding pocket. Salt bridges are indicated by a straight line, hydrogen bonds by a pink arrow. The orange lines indicate contacts between backbone NH and F of the PF_5 moiety. Grey shaded atoms are solvent exposed. The binding pocket is represented by a line around the ligand with the color matching the closest amino acid

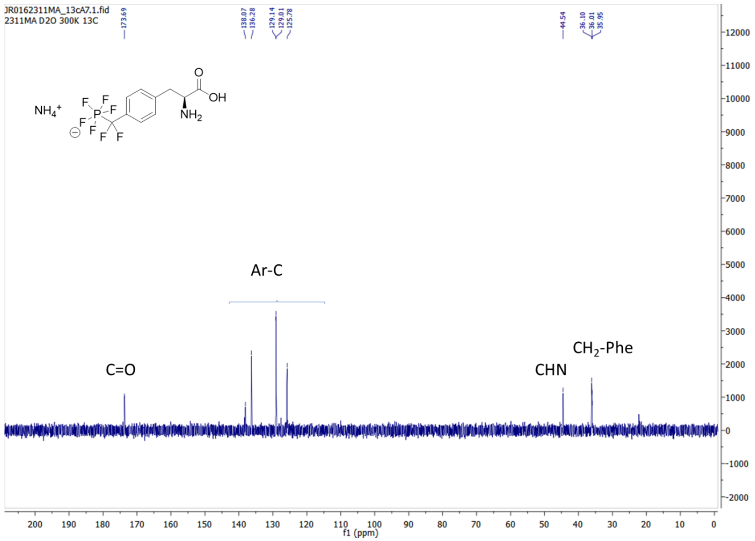


Supplementary Figure 42. Ligand interaction diagram of capped **1** in PTP1B binding pocket. Salt bridges are indicated by a straight line, hydrogen bonds by a pink arrow. Grey shaded atoms are solvent exposed. The binding pocket is represented by a line around the ligand with the color matching the closest amino acid

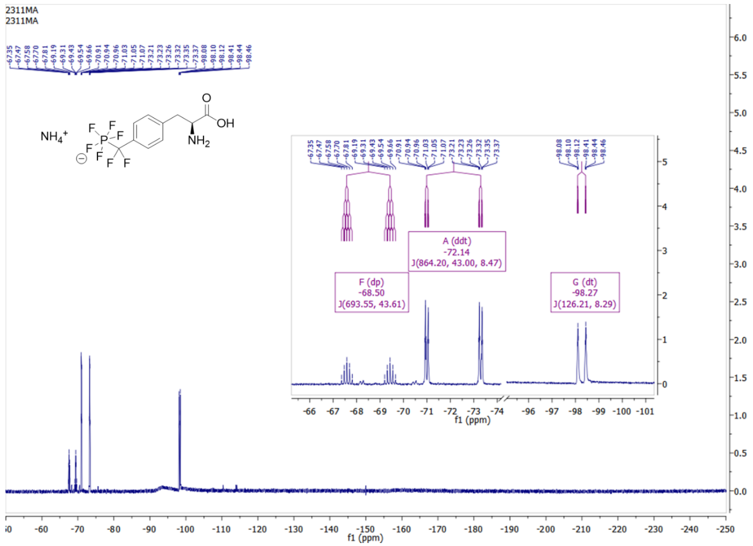
NMR and UV/vis spectra, HPLC chromatograms



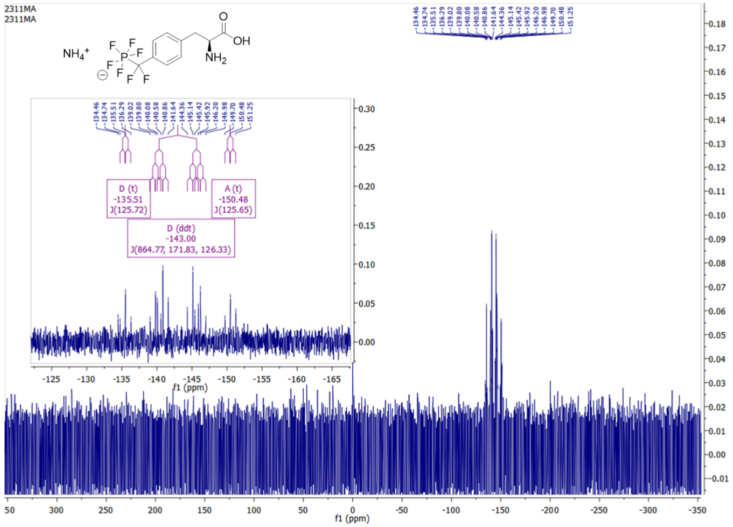
Supplementary Figure 43. ^1H NMR spectrum (700 MHz, D_2O) of **3**



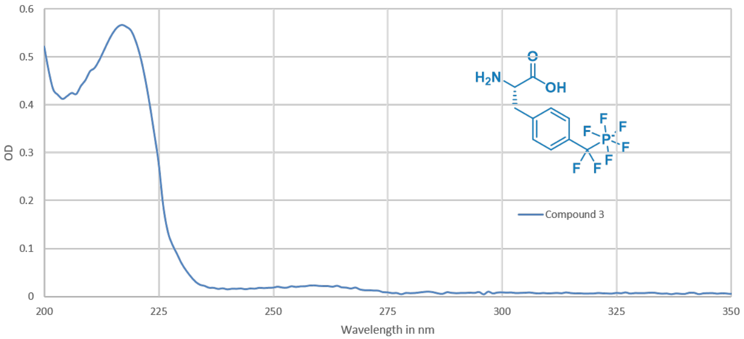
Supplementary Figure 44. ^{13}C NMR spectrum (176 MHz, D$_2$O) of **3**



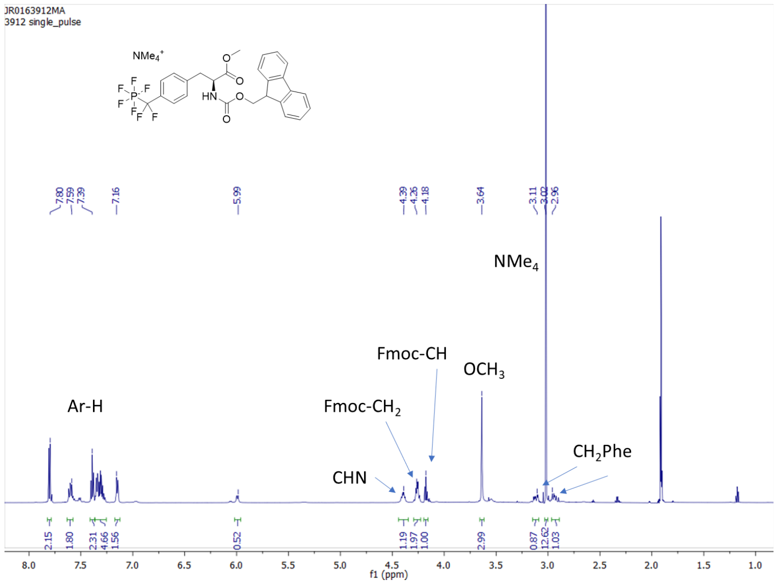
Supplementary Figure 45. ^{19}F NMR spectrum (376 MHz, D$_2$O) of **3**



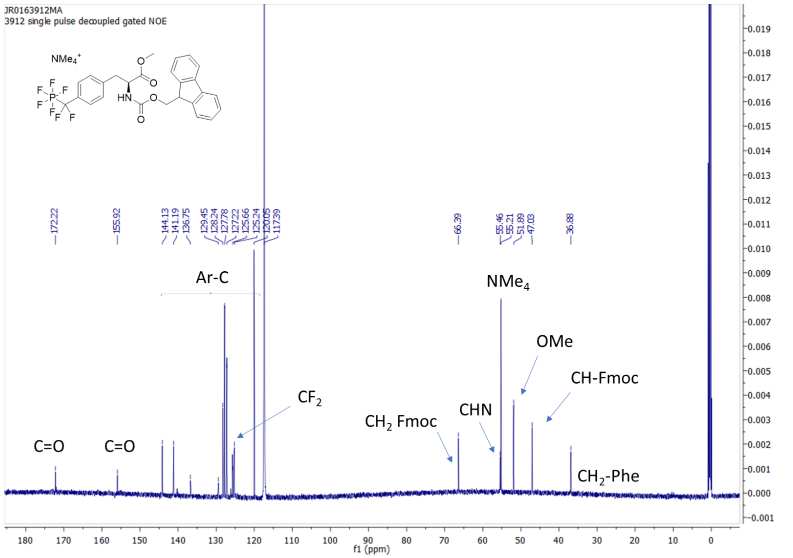
Supplementary Figure 46. ^{31}P NMR spectrum (161 MHz, D_2O) of **3**



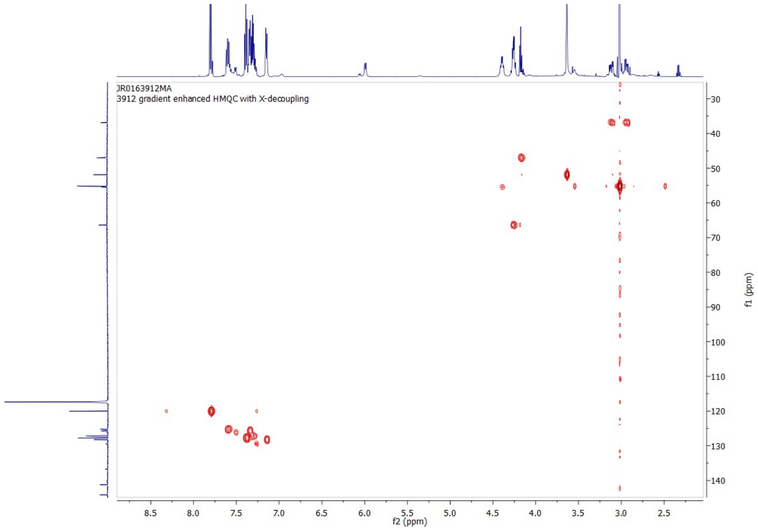
Supplementary Figure 47: UV spectrum of 0.5 mM solutions of compound **12** in water



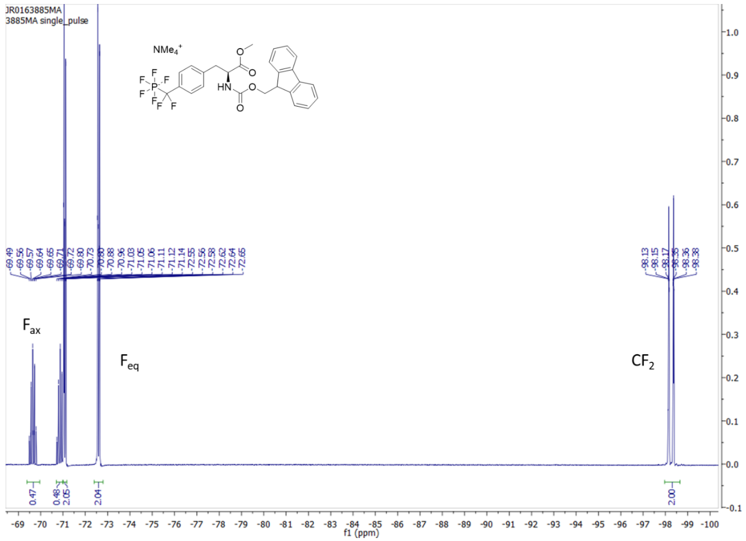
Supplementary Figure 48. ^{1}H NMR spectrum (600 MHz, ACN-d$_3$) of **8**



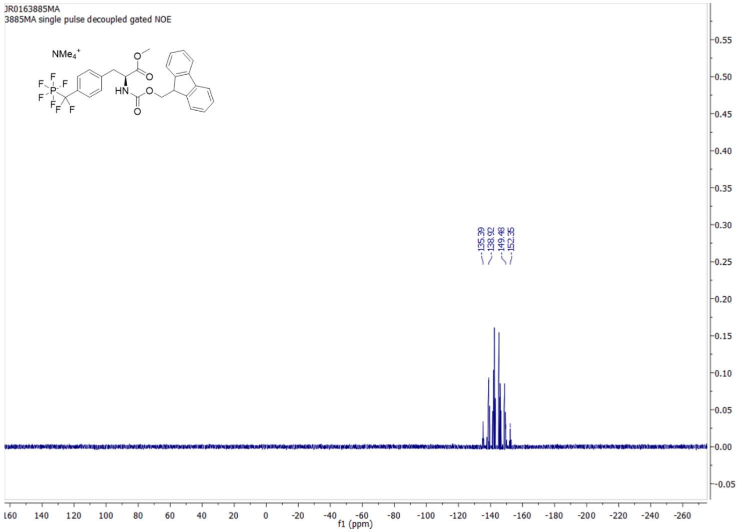
Supplementary Figure 49. ^{13}C NMR spectrum (151 MHz, ACN-d$_3$) of **8**



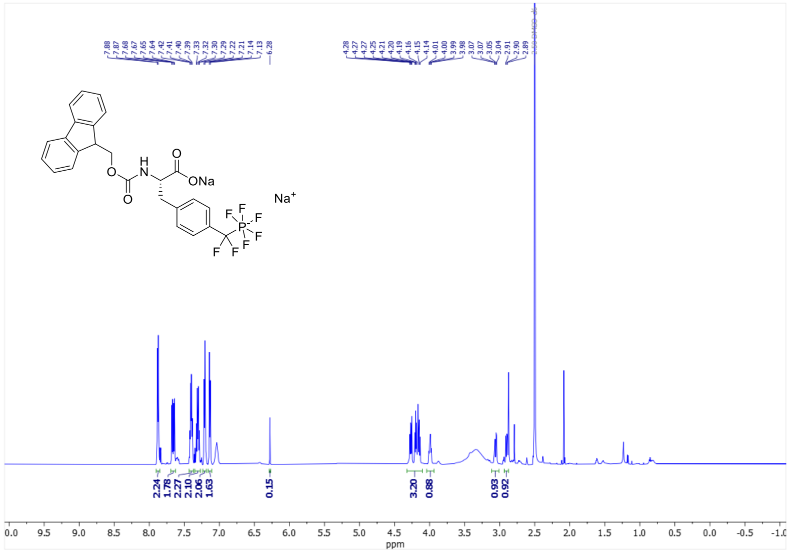
Supplementary Figure 50: HMQC spectrum (1H, 13C) in ACN-d$_3$ of **8**



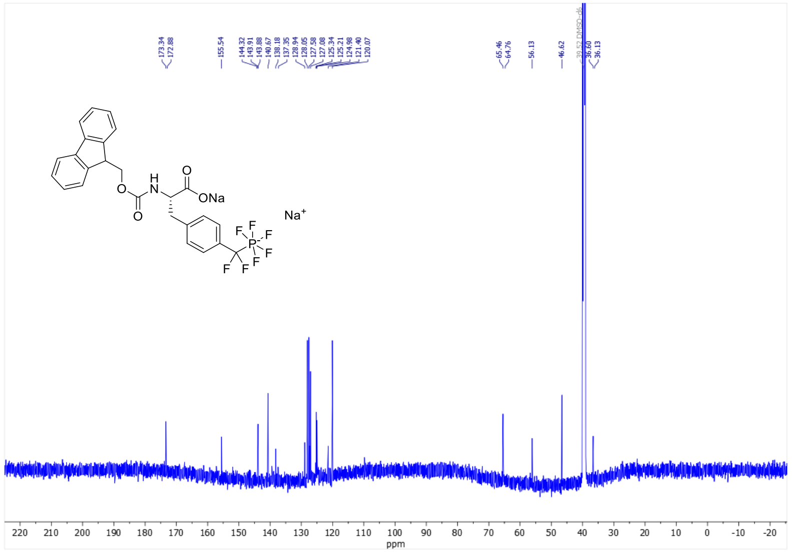
Supplementary Figure 51. ^{19}F NMR spectrum (565 MHz, ACN-d3) of **8**



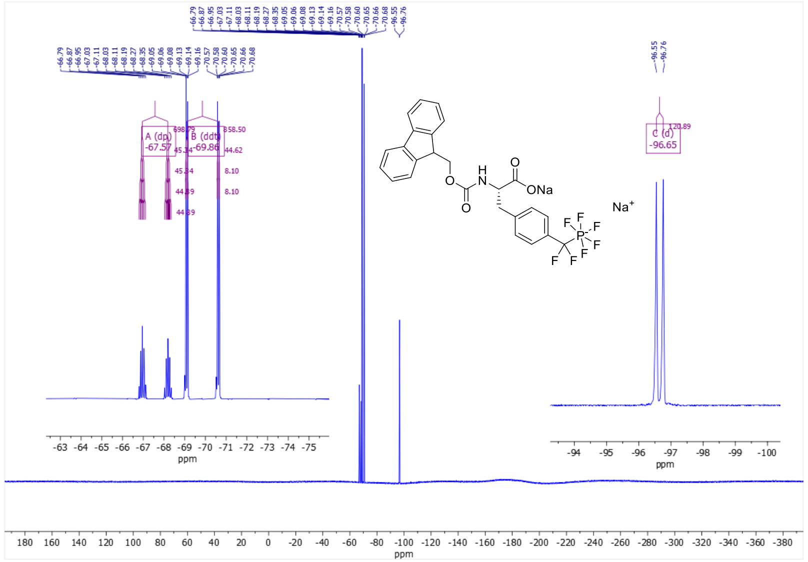
Supplementary Figure 52.3 ^{31}P NMR spectrum (243 MHz, ACN-d3) of **8**



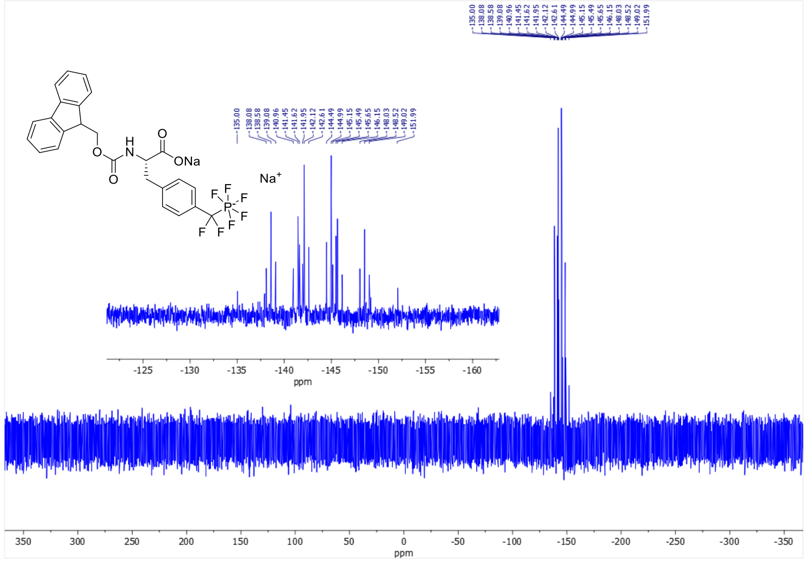
Supplementary Figure 53. ^{1}H NMR spectrum (600 MHz, DMSO-d6) of **9**



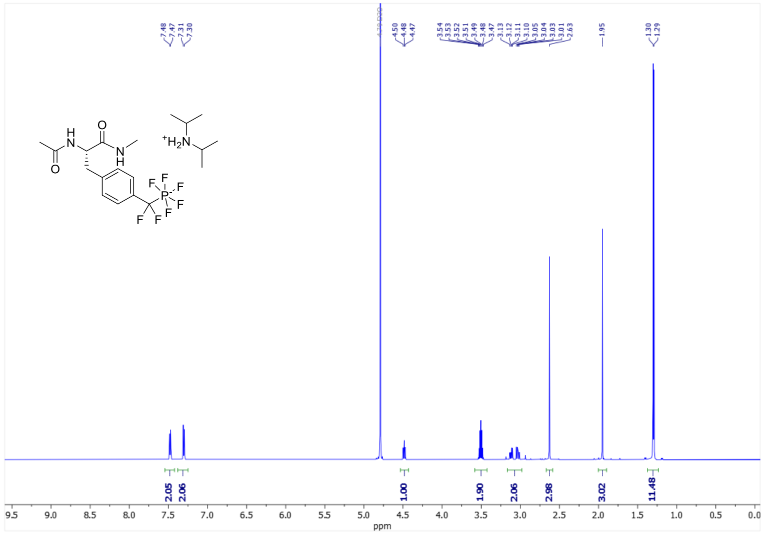
Supplementary Figure 54. ^{13}C NMR spectrum (151 MHz, DMSO-d6) of **9**



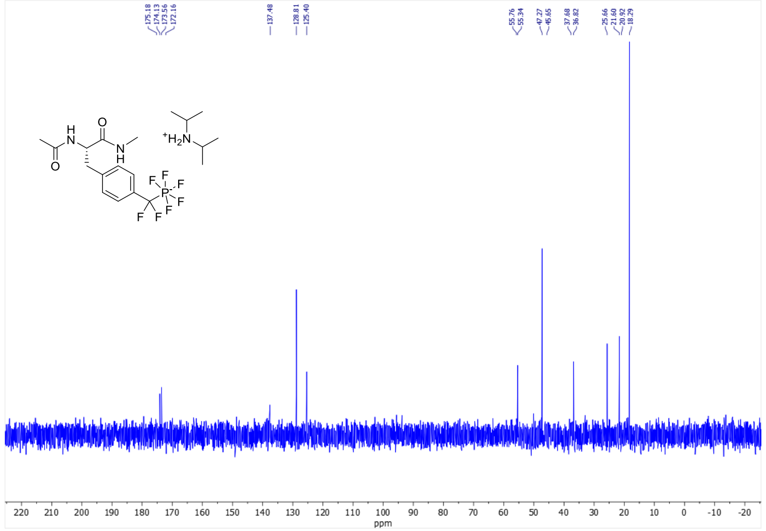
Supplementary Figure 55. ^{19}F NMR spectrum (565 MHz, DMSO-d6) of **9**



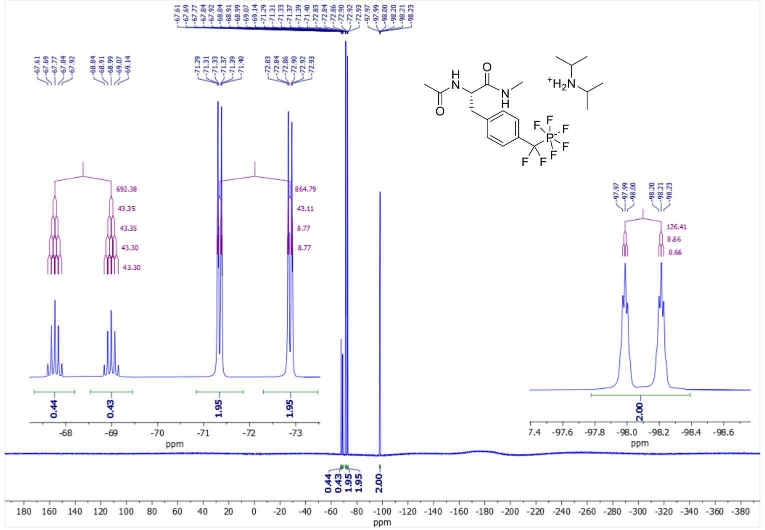
Supplementary Figure 56. ^{31}P NMR spectrum (243 MHz, DMSO-d6) of **9**



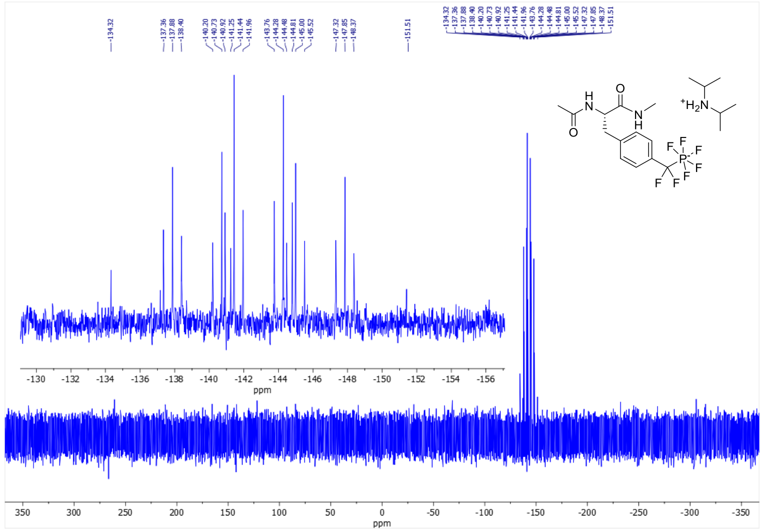
Supplementary Figure 57. ^{1}H NMR spectrum (600 MHz, D_2O) of **10**



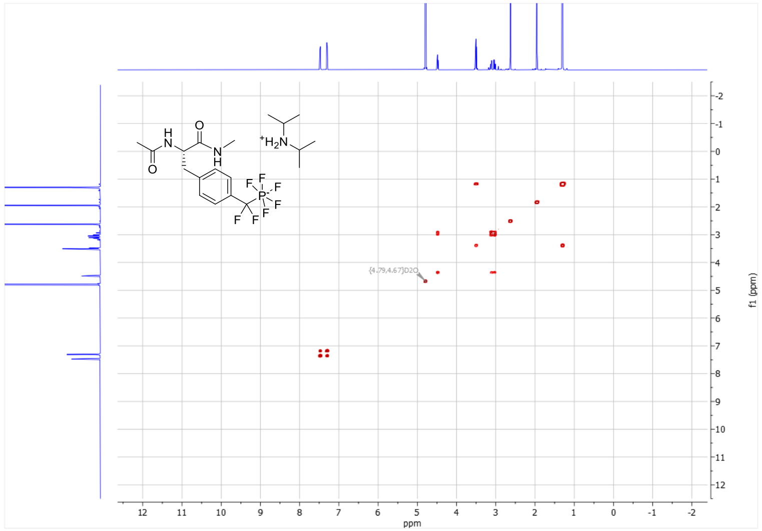
Supplementary Figure 58. ^{13}C NMR spectrum (151 MHz, ACN-d_3) of **10**



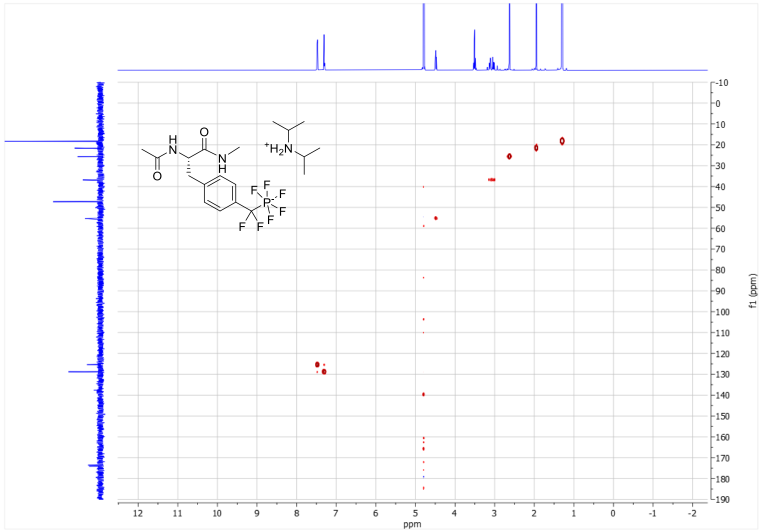
Supplementary Figure 59. ^{19}F NMR spectrum (565 MHz, D_2O) of **10**



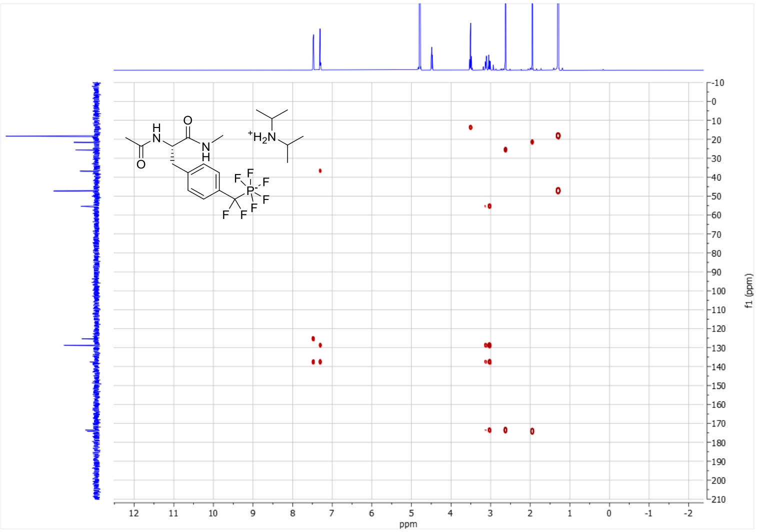
Supplementary Figure 60. ^{31}P NMR spectrum (243 MHz, D_2O) of **10**



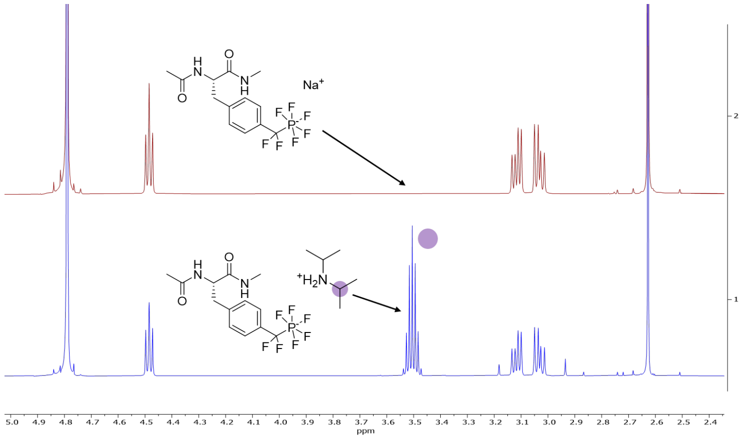
Supplementary Figure 61. H,H-COSY-NMR spectrum (600 MHz, D_2O) of **10**



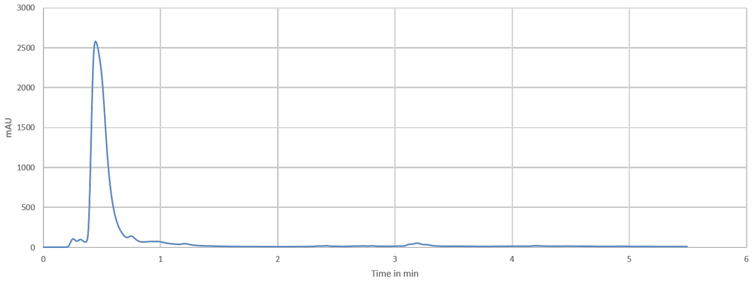
Supplementary Figure 62. HMQC spectrum (600 MHz, D_2O) of **10**



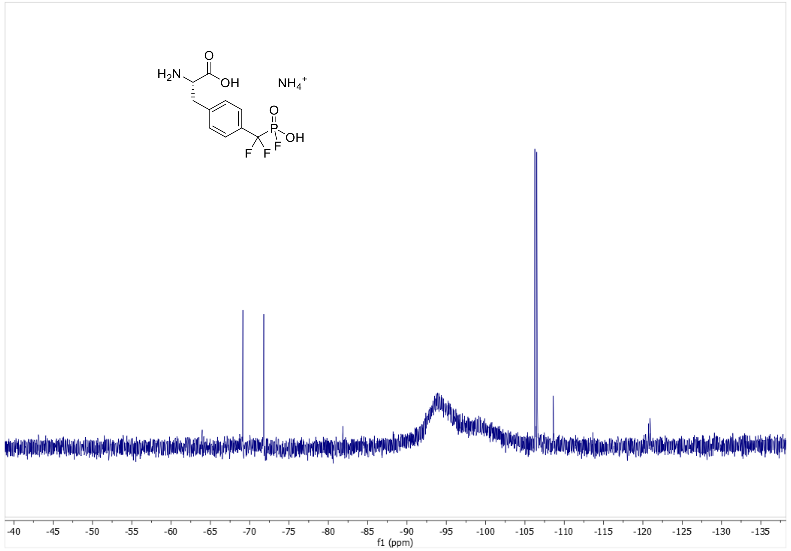
Supplementary Figure 63. HMBC (600 MHz, D_2O) of **10**



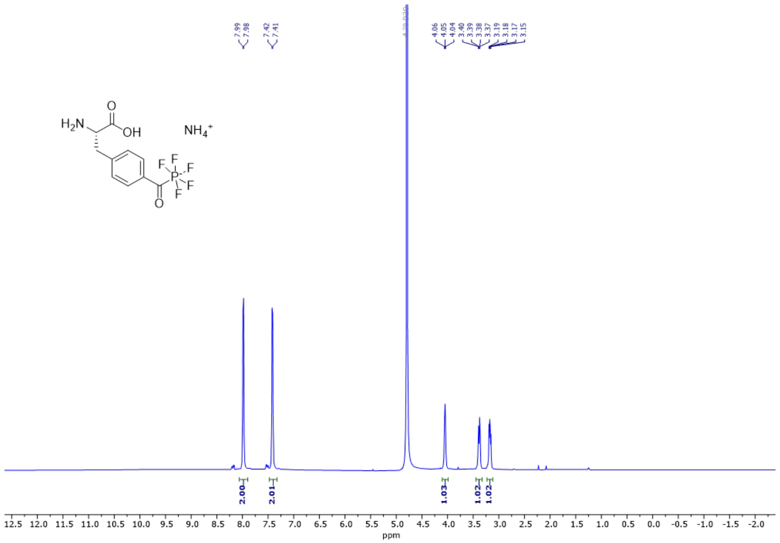
Supplementary Figure 64. ^{1}H NMR (600 MHz, D_2O) of **10**, before and after ion exchange



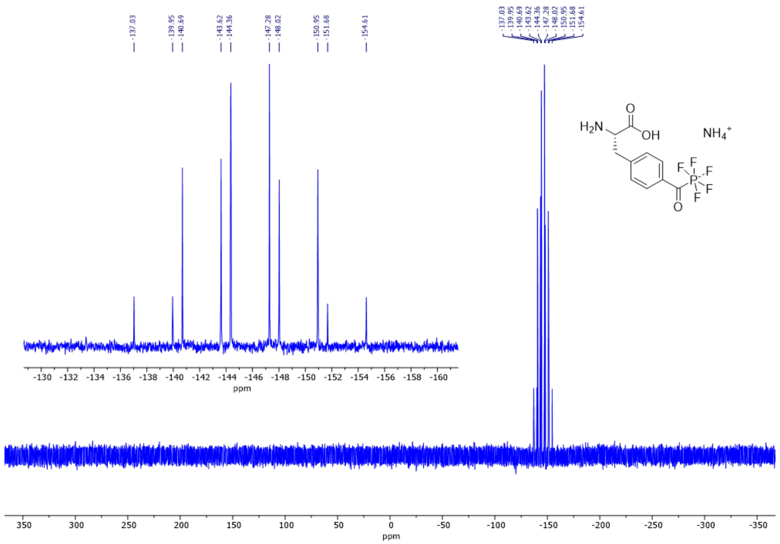
Supplementary Figure 65. HPLC chromatogram of compound **10**. Column B, eluent 15-95% ACN in 5 min, 210 nm.



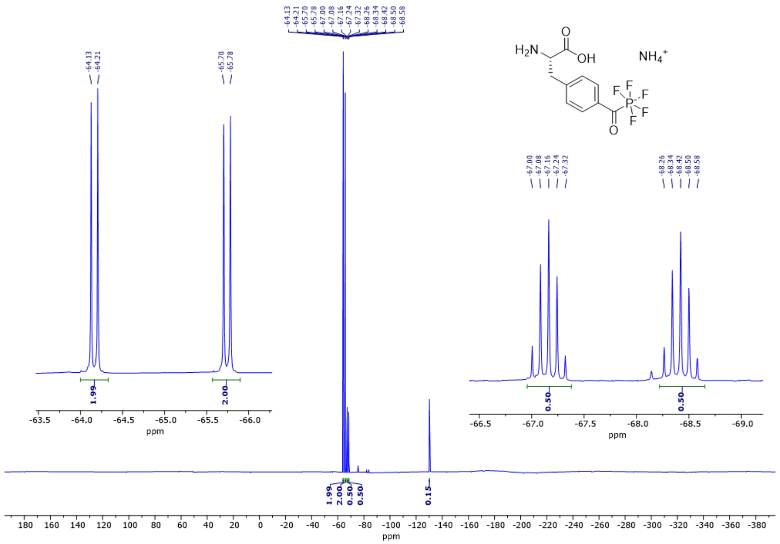
Supplementary Figure 66. ^{19}F NMR spectrum (565 MHz, D_2O) of **11**



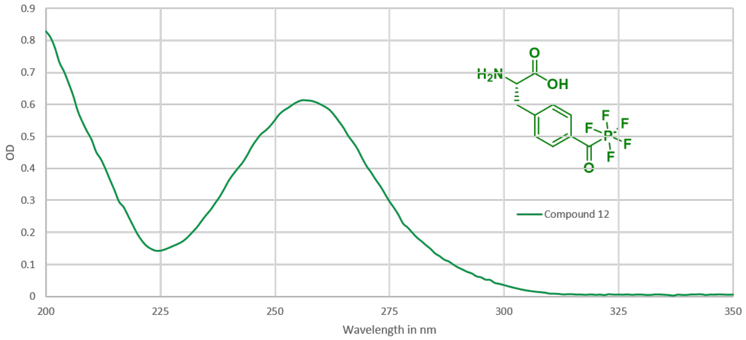
Supplementary Figure 67. ^{1}H NMR spectrum (600 MHz, D_2O) of **12**



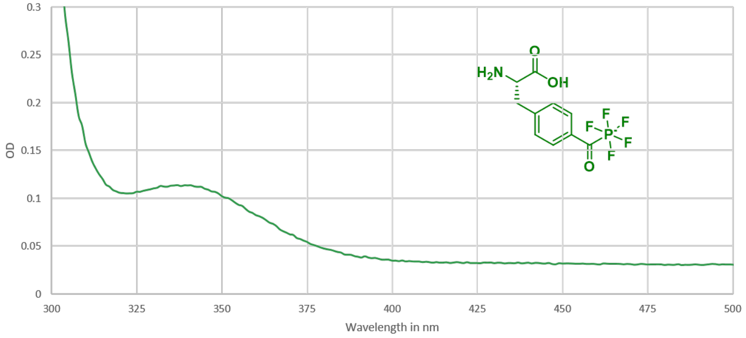
Supplementary Figure 68. ^{31}P NMR spectrum (243 MHz, D_2O) of **12**



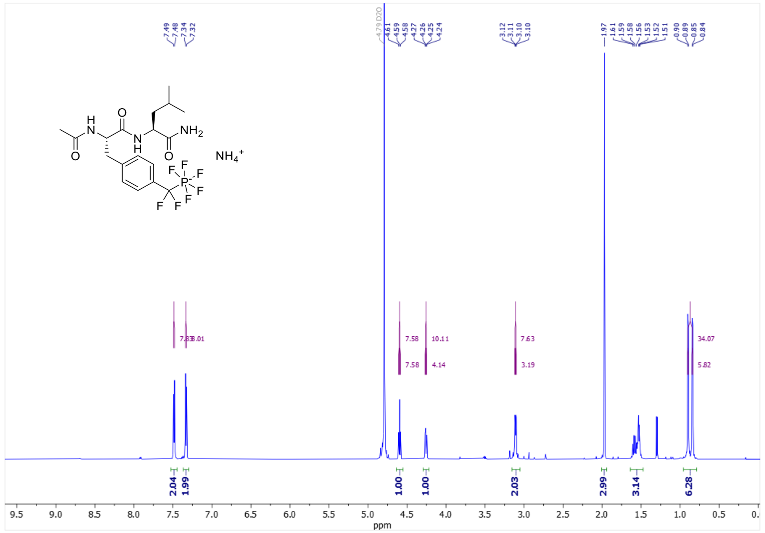
Supplementary Figure 69. ^{19}F NMR spectrum (565 MHz, D_2O) of **12**



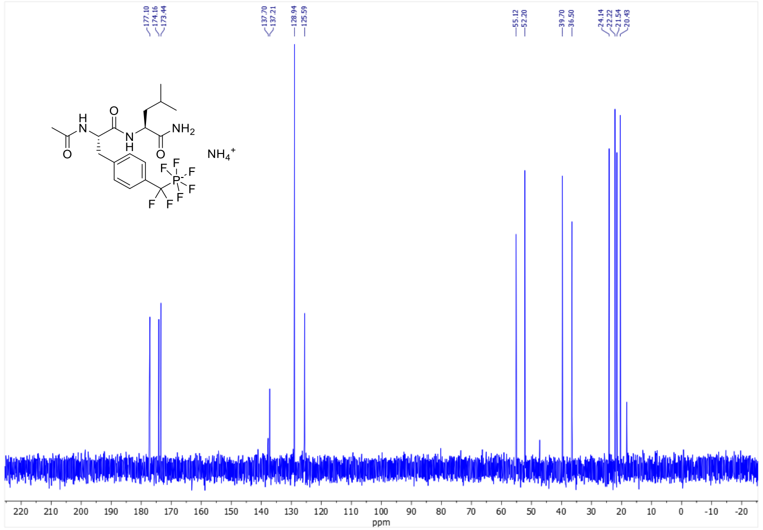
Supplementary Figure 70: UV spectrum of 0.5 mM solution of compounds **12** in water



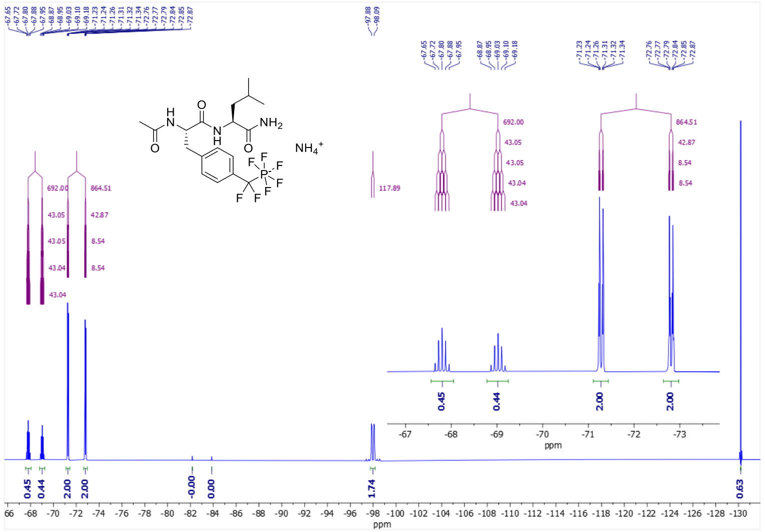
Supplementary Figure 71: UV spectrum of **12** in water displaying an n-π* transition



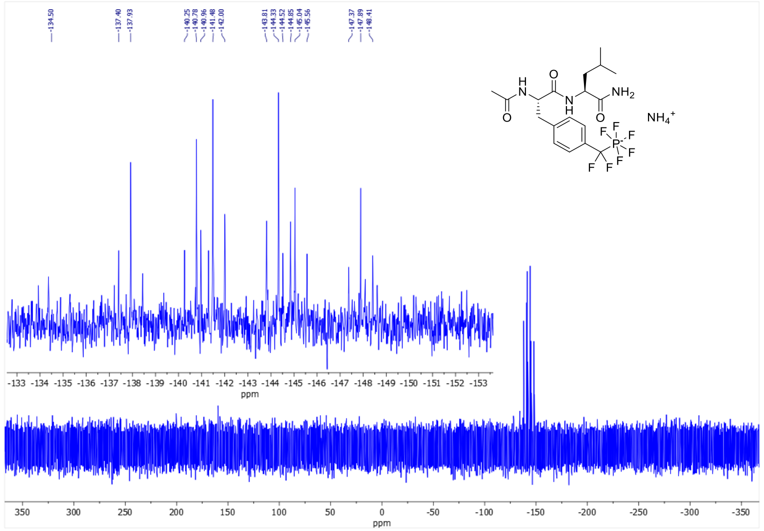
Supplementary Figure 72. ^{1}H NMR spectrum (600 MHz, D_2O) of **14**



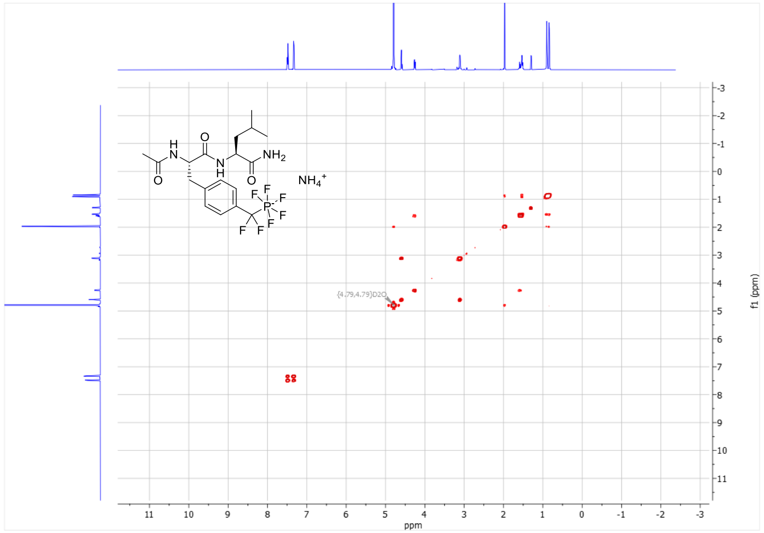
Supplementary Figure 73. ^{13}C NMR spectrum (151 MHz, D_2O) of **14**



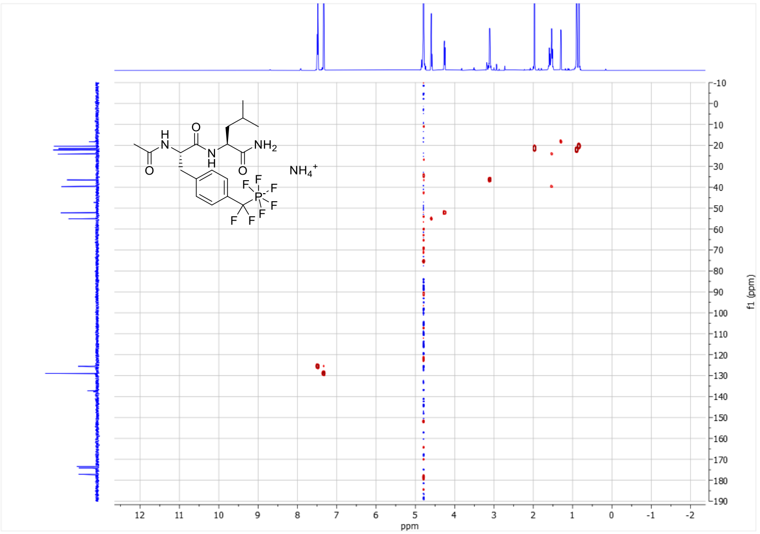
Supplementary Figure 74. ^{19}F NMR spectrum (565 MHz, D$_2$O) of **14**



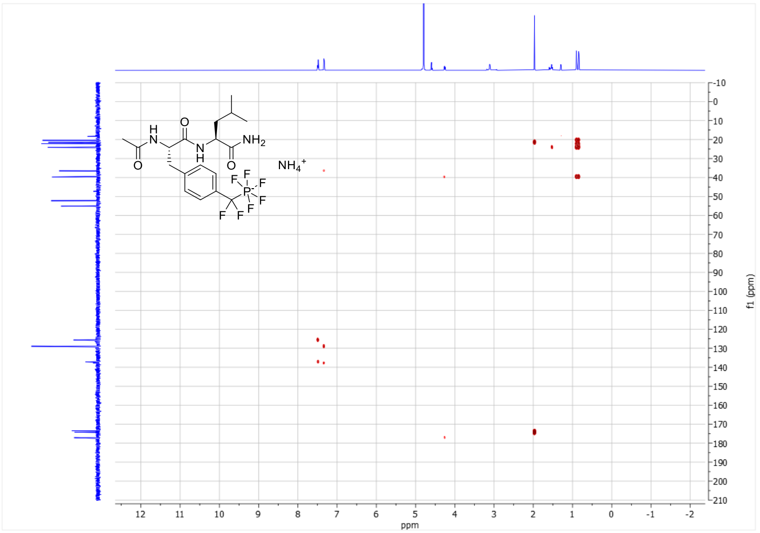
Supplementary Figure 75. ^{31}P NMR spectrum (243 MHz, D$_2$O) of **14**



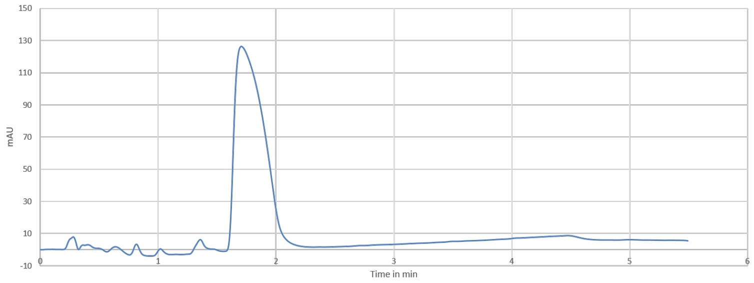
Supplementary Figure 76. H,H-COSY-NMR NMR spectrum (600 MHz, D_2O) of **14**



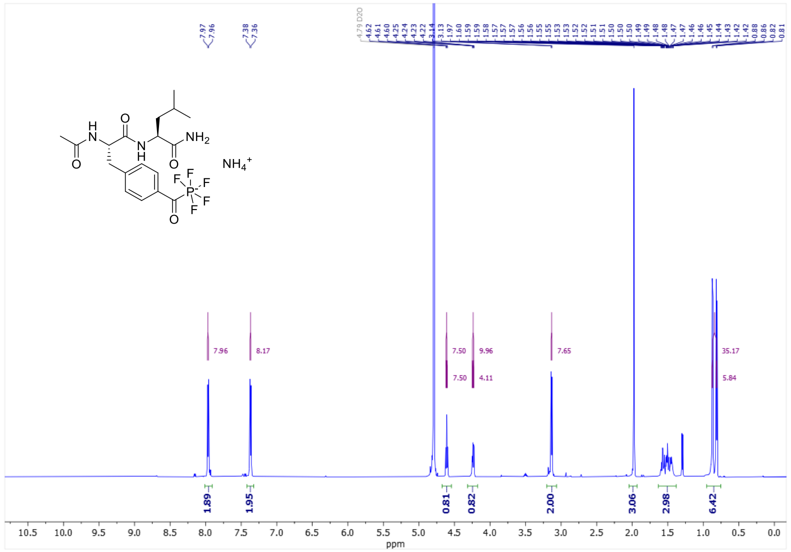
Supplementary Figure 77. HMBC NMR spectrum (600 MHz, D_2O) of **14**



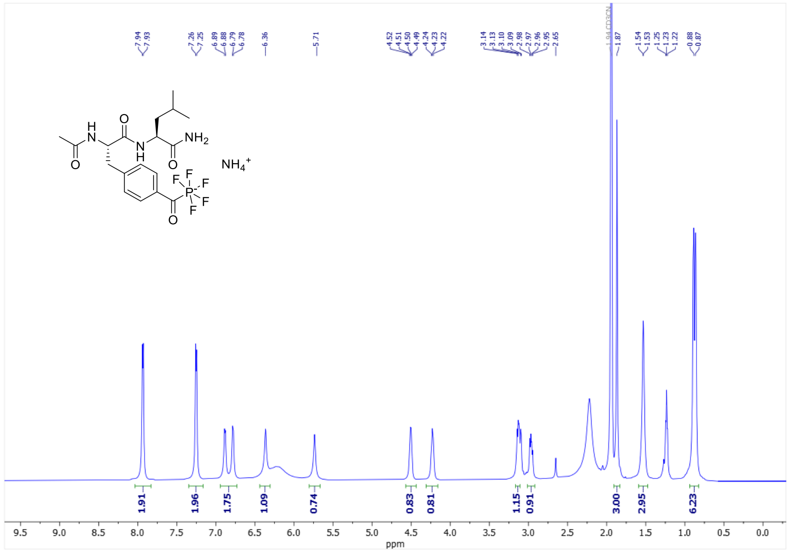
Supplementary Figure 78. HMBC NMR spectrum (600 MHz, D$_2$O) of **14**



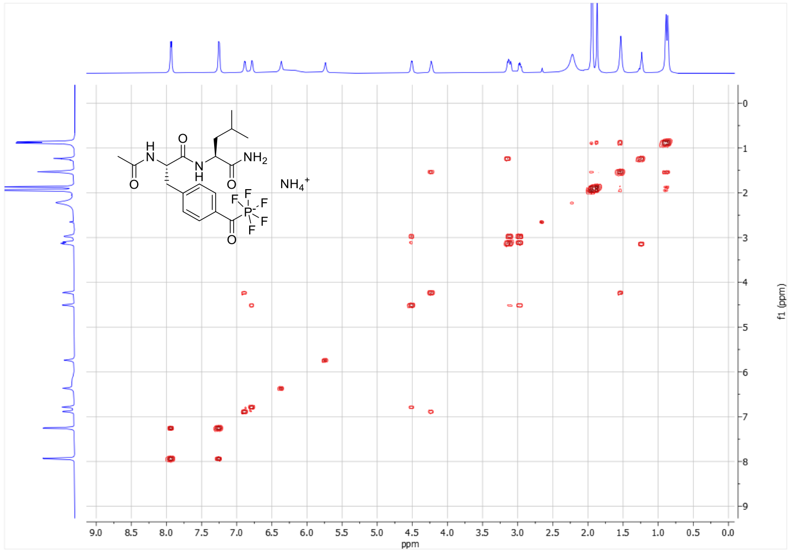
Supplementary Figure 79. HPLC chromatogram of compound **14**. Column B, eluent 15-95% ACN in 5 min, DAD 210 nm.



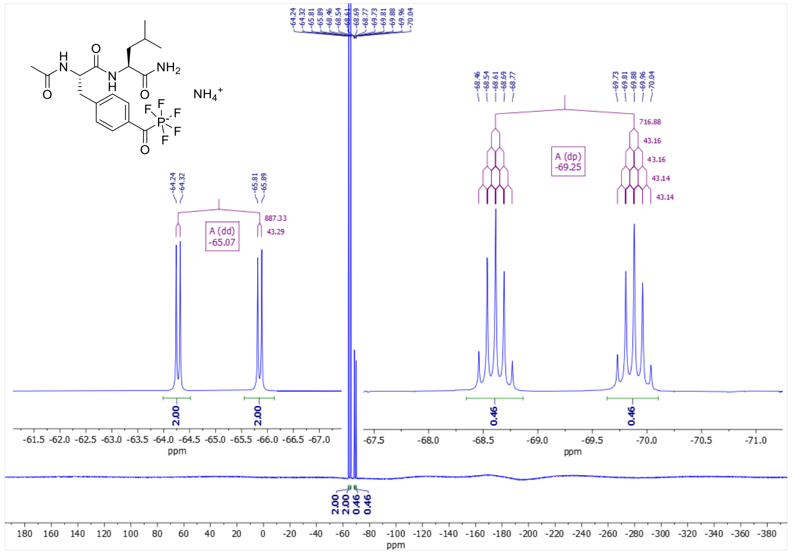
Supplementary Figure 80. ^{1}H NMR spectrum (600 MHz, D_2O) of **15**



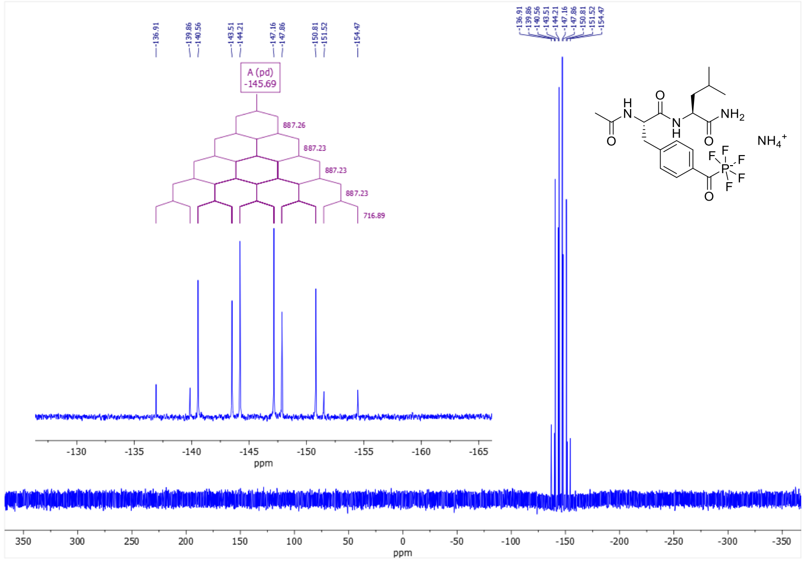
Supplementary Figure 81. ^{1}H NMR spectrum (600 MHz, ACN-D3) of **15**



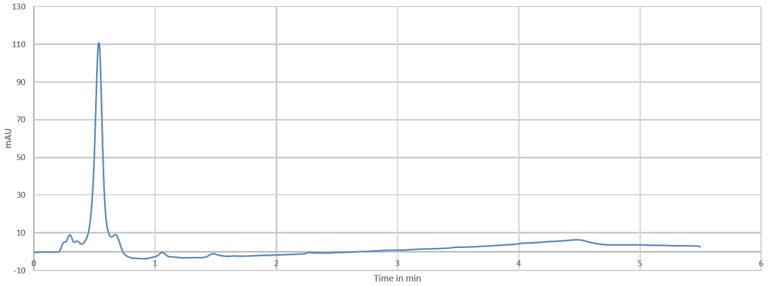
Supplementary Figure 82. H,H-COSY-NMR spectrum (600 MHz, ACN-D3) of **15**



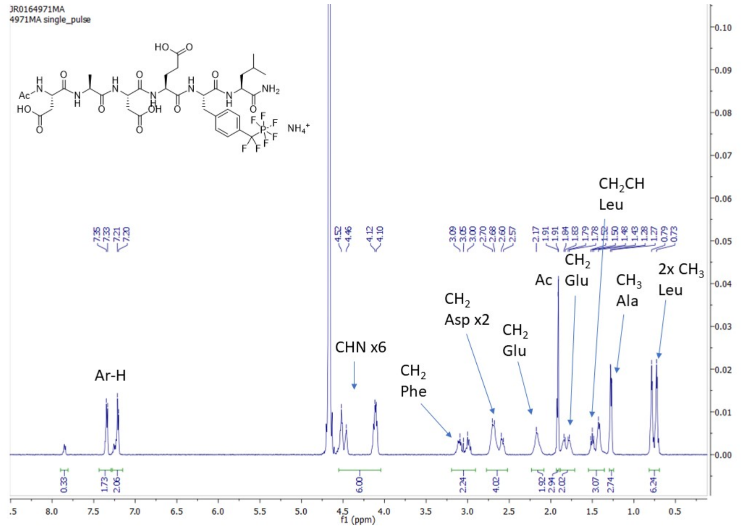
Supplementary Figure 83. ^{19}F NMR spectrum (565 MHz, ACN-D3) of **15**



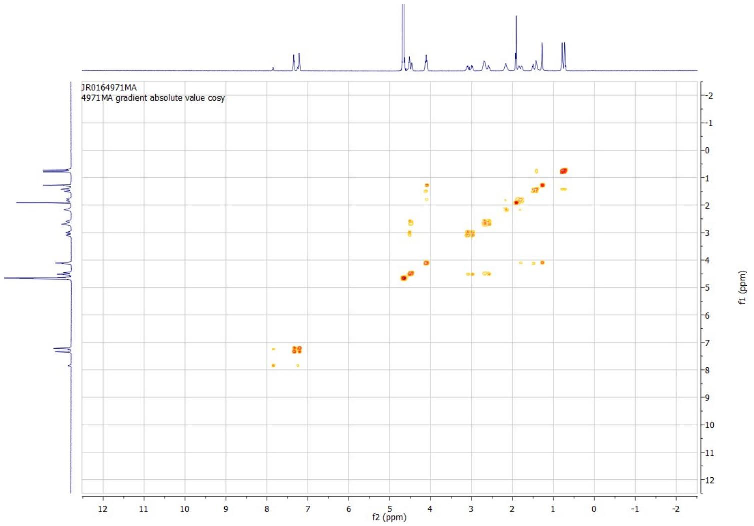
Supplementary Figure 84. ^{31}P NMR spectrum (243 MHz, ACN-D3) of **15**



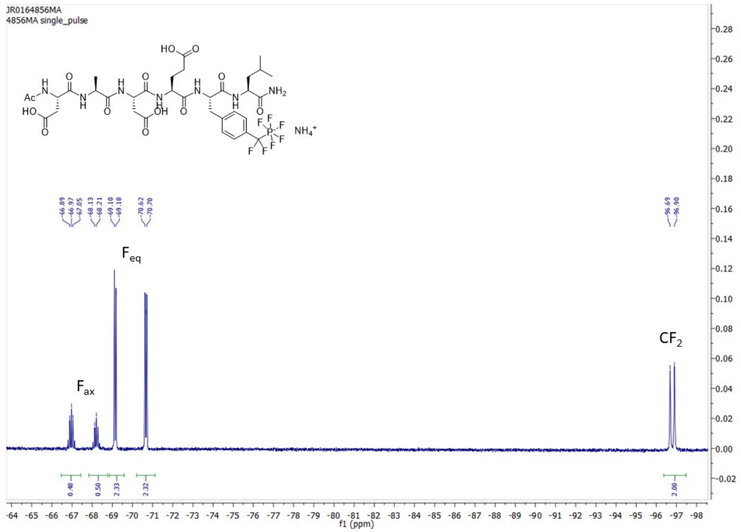
Supplementary Figure 85. HPLC chromatogram of compound **15** with UV detection at 220 nm. Column B, eluent 15-95% ACN in 5 min, DAD 210 nm.



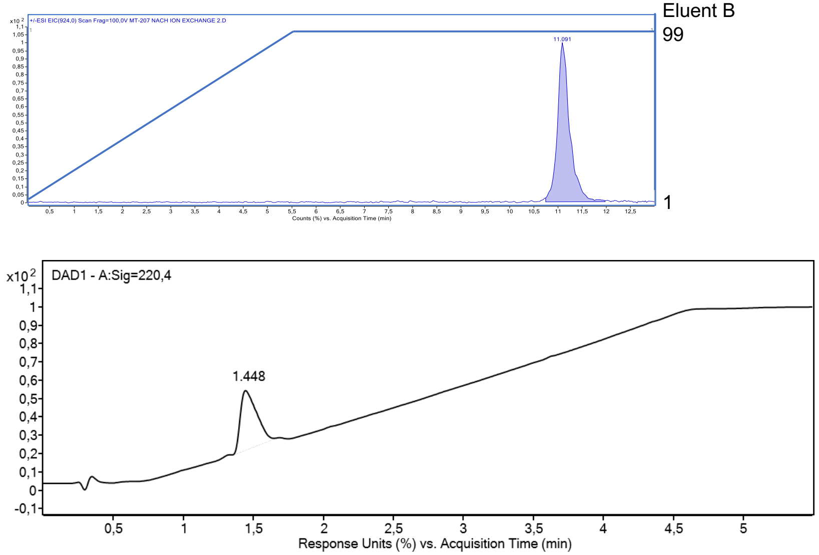
Supplementary Figure 86. ^{1}H NMR spectrum (600 MHz, D_2O) of **16**



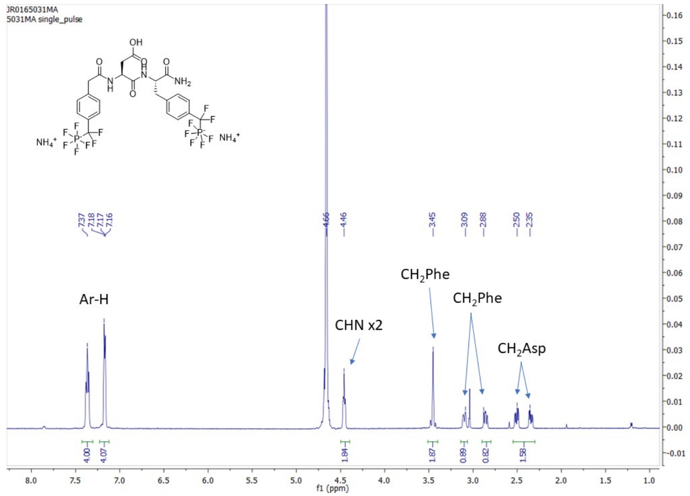
Supplementary Figure 87. H,H-COSY-NMR spectrum (600 MHz, D_2O) of **16**



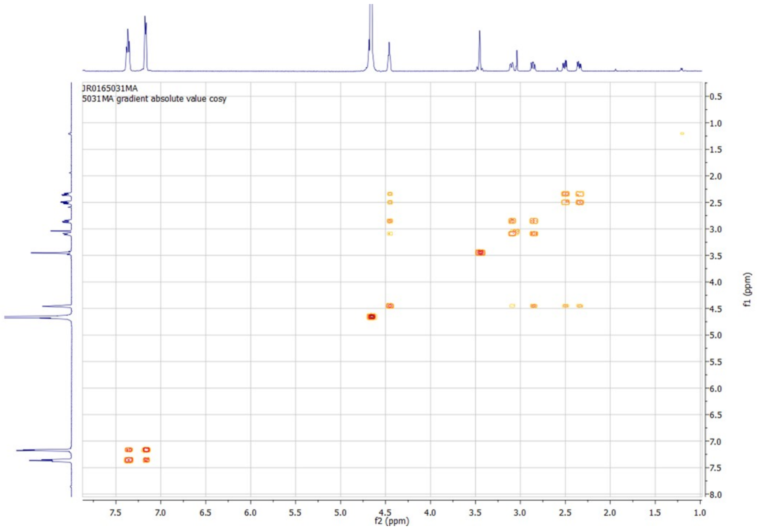
Supplementary Figure 88. ^{19}F NMR spectrum (565 MHz, D$_2$O) of **16**



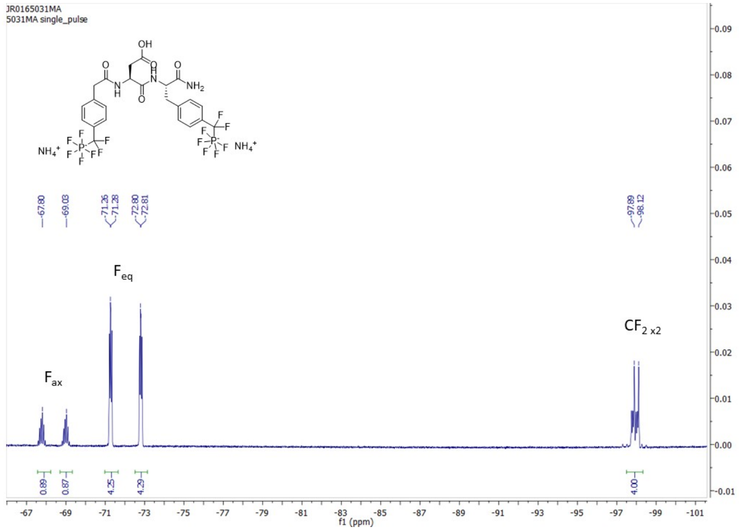
Supplementary Figure 89. HPLC chromatogram of compound **16**. Top: Column A, 1-99% eluent ACN in 5.5 min. Bottom: Column B, eluent 15-95% ACN in 4.5 min, 220 nm DAD.



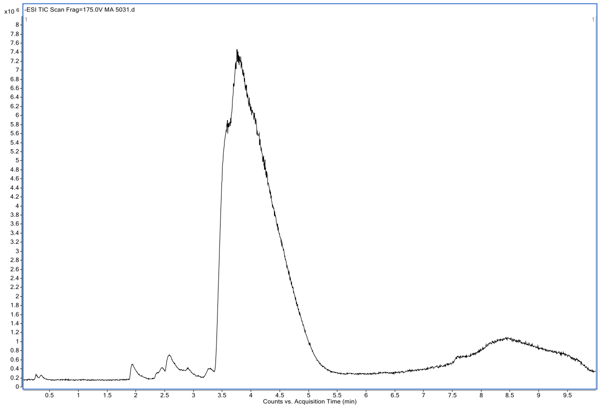
Supplementary Figure 90. ^{1}H NMR spectrum (600 MHz, D_2O) of **17**



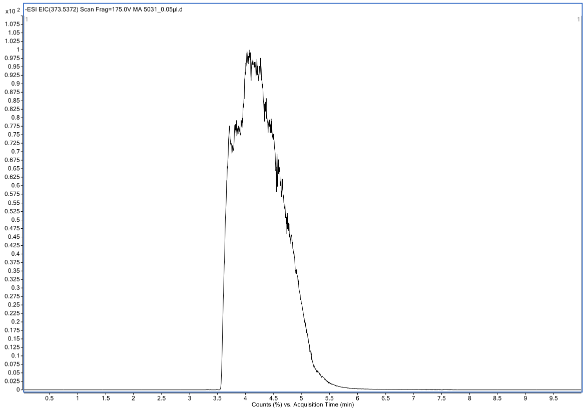
Supplementary Figure 91. H,H-COSY-NMR spectrum (600 MHz, D_2O) of **17**



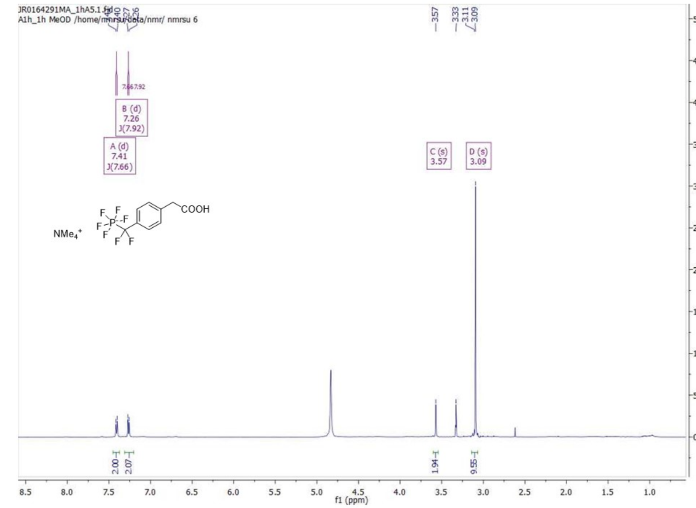
Supplementary Figure 92. ^{19}F NMR spectrum (565 MHz, D_2O) of **17**



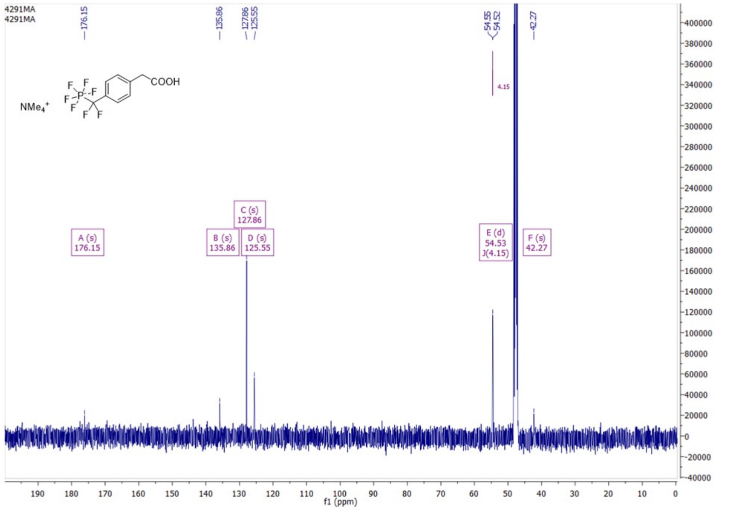
Supplementary Figure 93. Total ion chromatogram of compound **17**. Column B, eluent 15-95% ACN in 8 min



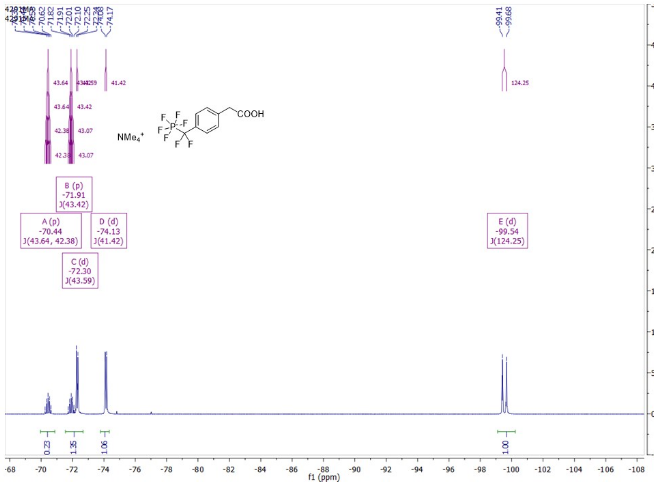
Supplementary Figure 94. Extracted ion chromatogram of compound **17**. Column B, eluent 15-95% ACN in 8 min



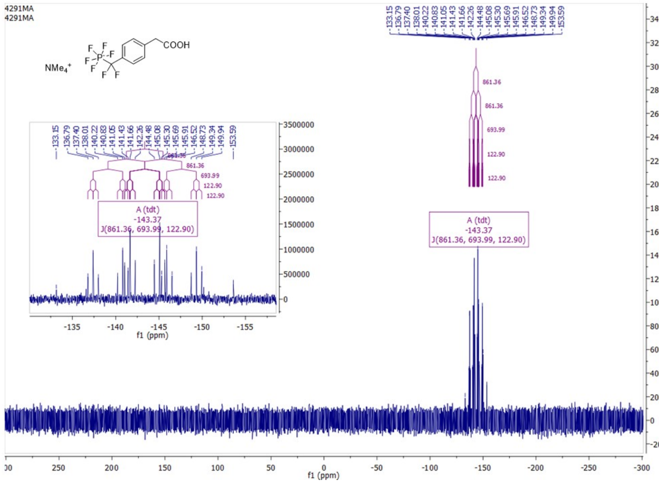
Supplementary Figure 95. ^{1}H NMR spectrum (500 MHz, MeOD) of **18**



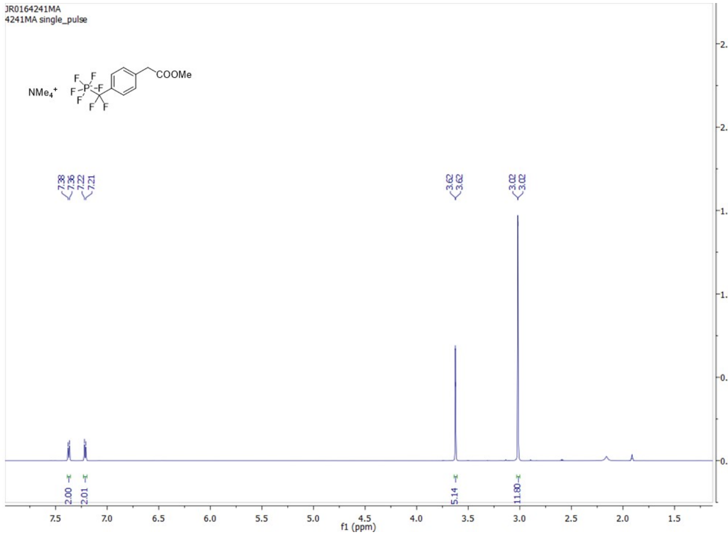
Supplementary Figure 96. ^{13}C NMR spectrum (125 MHz, MeOD) of **18**



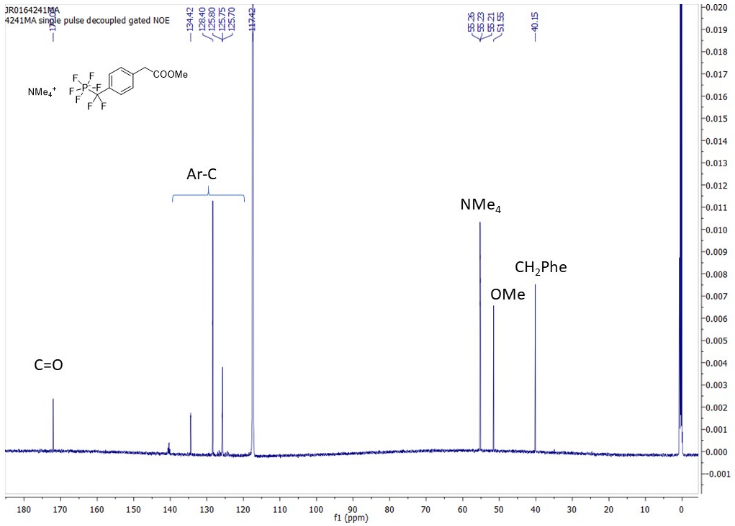
Supplementary Figure 97. ^{19}F-NMR spectrum (470 MHz, MeOD) of **18**



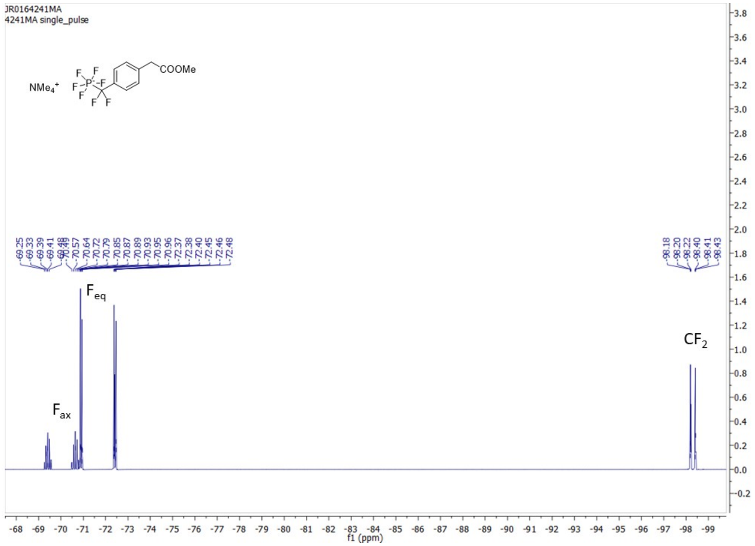
Supplementary Figure 98. ^{31}P-NMR spectrum (202 MHz, MeOD) of **18**



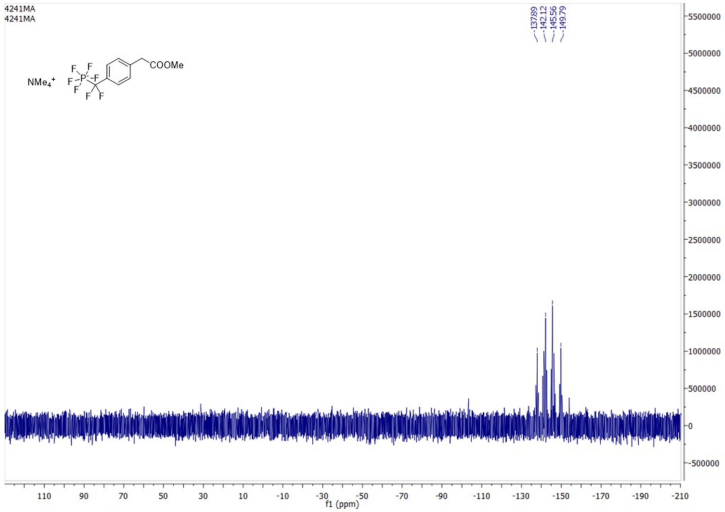
Supplementary Figure 99. ^{1}H NMR spectrum (600 MHz, ACN-d3) of **21**



Supplementary Figure 100. ^{13}C NMR spectrum (151 MHz, ACN-d3) of **21**



Supplementary Figure 101. ^{19}F NMR spectrum (565 MHz, ACN-d3) of **21**



Supplementary Figure 102. ^{31}P NMR spectrum (243 MHz, ACN-d3) of **21**

References

[1] K. O. Christe, W. W. Wilson, R. D. Wilson, R. Bau, J. A. Feng, *J. Am. Chem. Soc.* **1990**, *112*, 7619-7625.

[2] E. Kaiser, R. L. Colescott, C. D. Bossinger, P. I. Cook, *Anal. Biochem.* **1970**, *34*, 595-598.

[3] J. Ge, H. Wu, S. Q. Yao, *Chem. Commun.* **2010**, *46*, 2980-2982.

[4] M. N. Qabar, J. Urban, M. Kahn, *Tetrahedron* **1997**, *53*, 11171-11178.

[5] I. G. Boutselis, X. Yu, Z. Y. Zhang, R. F. Borch, *J. Med. Chem.* **2007**, *50*, 856-864.

[6] a) A. S. Bruker, AXS Inc., **2004** Madison, WI; b) G. M. Sheldrick, *Acta Crystallogr. A* **2008**, *64*, 112-122.

[7] G. M. Sheldrick, SADABS (Version 2.03), **2002** University of Göttingen. Göttingen, Germany.

[8] G. M. Sheldrick, *Acta Crystallogr. A Found Adv.* **2015**, *71*, 3-8.

[9] L. J. Farrugia, *Journal of Applied Crystallography* **1999**, *32*, 837-838.

[10] C. F. Macrae, I. Sovago, S. J. Cottrell, P. T. A. Galek, P. McCabe, E. Pidcock, M. Platings, G. P. Shields, J. S. Stevens, M. Towler, P. A. Wood, *J. Appl. Crystallogr.* **2020**, *53*, 226-235.

[11] a) K. Brandenburg, H. Putz, Crystal Impact 2.6.4. **2021** GbR, Bonn, Germany; b) G. Bergerhoff, M. Berndt, K. Brandenburg, *J. Res. Natl. Inst. Stand. Technol.* **1996**, *101*, 221-225.

[12] C. Yung-Chi, W. H. Prusoff, *Biochemical Pharmacology* **1973**, *22*, 3099-3108.

[13] N. Krishnan, K. Krishnan, C. R. Connors, M. S. Choy, R. Page, W. Peti, L. Van Aelst, S. D. Shea, N. K. Tonks, *J. Clin. Invest.* **2015**, *125*, 3163-3177.

[14] G. M. Sastry, M. Adzhigirey, T. Day, R. Annabhimoju, W. Sherman, *J. Comput. Aided Mol. Des.* **2013**, *27*, 221-234.

[15] a) R. A. Friesner, R. B. Murphy, M. P. Repasky, L. L. Frye, J. R. Greenwood, T. A. Halgren, P. C. Sanschagrin, D. T. Mainz, *J. Med. Chem.* **2006**, *49*, 6177-6196; b) R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry, D. E. Shaw, P. Francis, P. S. Shenkin, *J. Med. Chem.* **2004**, *47*, 1739-1749; c) T. A. Halgren, R. B. Murphy, R. A. Friesner, H. S. Beard, L. L. Frye, W. T. Pollard, J. L. Banks, *J. Med. Chem.* **2004**, *47*, 1750-1759.

[16] C. Lu, C. Wu, D. Ghoreishi, W. Chen, L. Wang, W. Damm, G. A. Ross, M. K. Dahlgren, E. Russell, C. D. Von Bargen, R. Abel, R. A. Friesner, E. D. Harder, *J. Chem. Theory Comput.* **2021**, *17*, 4291-4300.

[17] a) E. Lindahl, B. Hess, D. van der Spoel, *J. Mol. Model.* **2001**, *7*, 306-317; b) S. Pronk, S. Pall, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess, E. Lindahl, *Bioinformatics* **2013**, *29*, 845-854; c) M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, E. Lindahl, *SoftwareX* **2015**, *1-2*, 19-25.

[18] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, C. Simmerling, *J. Chem. Theory Comput.* **2015**, *11*, 3696-3713.

[19] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, *J. Chem. Phys.* **1983**, *79*, 926-935.

[20] G. Bussi, D. Donadio, M. Parrinello, *J. Chem. Phys.* **2007**, *126*, 014101.

[21] M. Parrinello, A. Rahman, *Journal of Applied Physics* **1981**, *52*, 7182-7190.

[22] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, L. G. Pedersen, *J. Chem. Phys.* **1995**, *103*, 8577-8593.

[23] R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernandez, C. R. Schwantes, L. P. Wang, T. J. Lane, V. S. Pande, *Biophys. J.* **2015**, *109*, 1528-1532.

[24] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, D. A. Case, *J. Comput. Chem.* **2004**, *25*, 1157-1174.

[25] A. W. Sousa da Silva, W. F. Vranken, *BMC Res. Notes* **2012**, *5*, 367.

[26] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, I. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, C. SciPy, *Nat. Methods* **2020**, *17*, 261-272.

[27] J. R. Robalo, S. Huhmann, B. Koksch, A. Vila Verde, *Chem* **2017**, *3*, 881-897.

[28] a) C. I. Bayly, P. Cieplak, W. Cornell, P. A. Kollman, *The Journal of Physical Chemistry* **1993**, *97*, 10269-10280; b) W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, P. A. Kollman, *J. Am. Chem. Soc.* **1995**, *117*, 5179-5197.

[29] a) D. A. Case, T. E. Cheatham, 3rd, T. Darden, H. Gohlke, R. Luo, K. M. Merz, Jr., A. Onufriev, C. Simmerling, B. Wang, R. J. Woods, *J. Comput. Chem.* **2005**, *26*, 1668-1688; b) D. A. Case, K. Belfon, I. Y. Ben-Shalom, S. R. Brozell, D. S. Cerutti, *AMBER 2020,* **2020**; c) M. J. Frisch, G. W. Trucks, H. B. Schlegel, **2016**, Gaussian 16, Revision C.01.

[30] a) T. A. Halgren, *J. Chem. Inf. Model.* **2009**, *49*, 377-389; b) T. Halgren, *Chem. Biol. Drug Des.* **2007**, *69*, 146-148.

[31] A. C. Cheng, R. G. Coleman, K. T. Smyth, Q. Cao, P. Soulard, D. R. Caffrey, A. C. Salzberg, E. S. Huang, *Nat. Biotechnol.* **2007**, *25*, 71-75.

[32] Schrödinger Knowledge Base. Available at: https://www.schrodinger.com/kb/144. (Accessed: 122nd September 2021).

Supporting Information for Section 4.5

The following pages hold the Supporting Information for the publication:

"Biosynthetic Incorporation of Fluorinated Amino Acids into the Nonribosomal Peptide Gramicidin S"
Maximilian Müll, Farzaneh Pourmasoumi, Leon Wehrhan, Olena Nosovska, Philipp Stephan, Hannah Zeihe, Ivan Vilotijevic, Bettina G. Keller, Hajo Kries
RSC Chem. Biol. **2023**
DOI: 10.1039/D3CB00061C

Supporting Information

Biosynthetic incorporation of fluorinated amino acids into the nonribosomal peptide gramicidin S

Maximilian Müll[1], Farzaneh Pourmasoumi[1], Leon Wehrhan[2], Olena Nosovska[3], Philipp Stephan[1], Hannah Zeihe[1], Ivan Vilotijevic[3], Bettina G. Keller[2], Hajo Kries[1,4*]

[1]Junior Research Group Biosynthetic Design of Natural Products, Leibniz Institute for Natural Product Research and Infection Biology, Hans Knöll Institute (HKI Jena), 07745 Jena, Germany

[2]Freie Universität Berlin, Department of Biology, Chemistry, and Pharmacy, Institute of Chemistry and Biochemistry, Arnimallee 20, 14195 Berlin, Germany

[3]Institute of Organic Chemistry and Macromolecular Chemistry, Friedrich Schiller University Jena, Humboldtstr. 10, 07743 Jena, Germany

[4]University of Bayreuth, Organic Chemistry I, 95440 Bayreuth, Germany

*hajo.kries@leibniz-hki.de

Contents

Supporting Methods

Cloning

In-Fusion cloning was performed following the supplier's instructions. Primers for PCR amplification of inserts (Table S1) were designed with 15-20 bp long overlaps complementary to the insertion position in the vector DNA and ordered as synthetic oligonucleotides (Eurofins GmbH). Competent cells were transformed by heat shock. The insert was checked by Sanger sequencing (Genewiz). Restriction enzymes for linearization of vectors were purchased from New England Biolabs (NEB). PCR reactions were performed with Phusion or Q5 High-Fidelity polymerase (NEB).

Construction of pSU18-grsTAB/W239S. For the construction of pSU18-grsTAB/W239S, the corresponding region of the *grs*A gene was amplified in two fragments with primers grsA/W239S-P1-F-b, grsA/W239S-P1-R-a, grsA/W239S-P2-F-b, and grsA/W239S-P2-R-b. The vector pSU18-grsTAB was linearized with PmlI and EcoO01091.

Construction of pTrc99a-SrfA-B1. For the construction of pTrc99a-SrfA-B1 the corresponding region of the *srf*A-B gene was amplified in two PCR steps with the primers SrfA_Nested1_f and SrfA_Nested1_r for the first step and B1_isolation_Nested2_CATCf3 and B1_isolation_Nested2_CATr2 for the second step.

Construction of pSU18-GrsA/239NNK. For the construction of the plasmid pSU18-GrsA/NNK a linker fragment of the grsA gene was amplified from pSU18-GrsA using the primer pair GrsA_f and GrsA_r. The NNK fragment was amplified from pSU18-GrsA using the primer pair GrsA_W239NNK_f and GrsA_W239NNK_r. Linearized pSU18 was used as a vector.

Construction of pOPINF-PheA-ND. For the construction of plasmid pOPINF-PheA_ND the corresponding region of the *grsA* gene was PCR amplified with the primer pair GrsA-ND_f and GrsA-ND_r. For use as vector, plasmid pOPINF was linearized with KpnI and HindIII.

Protein expression and purification

Proteins were expressed and purified as described previously.[1] Precultures of *E. coli* HM0079 carrying either the plasmid pSU18-mGrsA, pSU18-GrsBMtoL, pSU18-GrsA/W239S, pTrc99a-SrfA-B1 or *E. coli* BL21 containing pOPINE-deoD, which is coding for purine nucleoside phosphorylase (PNP), were prepared by inoculation of 3 mL LB media containing ampicillin (K029.4, Roth) or chloramphenicol (3886.2, Roth) as resistance marker. Precultures were incubated at 37 °C at 180 rpm overnight in a rotary shaker. Main cultures were prepared by inoculation of 400 mL 2xYT media in 2 L flasks, containing the corresponding resistance marker. The cultures were incubated at 37 °C at 250 rpm on a rotary shaker for approximately 4 h until they reached an OD_{600} of approximately 0.6. The cultures were cooled down to 18 °C and induced with 0.25 mM isopropyl-β-D-thiogalactoside (IPTG; BP1755-10, Fisher Scientific). Proteins were expressed overnight at 18 °C at 250 rpm. Cells were harvested by centrifugation and the supernatant was discarded. After resuspending the cell pellet in 30 mL lysis buffer (50 mM TRIS [pH 7.4], 500 mM NaCl, 20 mM imidazole, 2 mM TCEP), 100 µL protease inhibitor mix (APE-K1010, APExBio) were added, and cells were lysed by sonication while cooling on ice. The lysate was cleared by centrifugation at 19,000 g for 30 min at 4°C and the supernatant was loaded onto a column packed with 2 mL of Ni-IDA suspension (1308.2, Roth) and equilibrated with lysis buffer. After washing the column twice with 20 mL of the lysis buffer, the target protein was eluted with 6 x 0.75 mL elution buffer (50 mM TRIS [pH 7.4], 500 mM NaCl, 300 mM imidazole, 2 mM TCEP). After pooling the protein-containing fractions, they were buffer exchanged with storage buffer (100 mM TRIS [pH 7.4], 500 mM NaCL, 2 mM TCEP), using Vivaspin 6 (Sartorius) filters with a cutoff of 10 kDa for the PNP, 30 kDa for GrsA, GrsA-W239S and SrfA-B1 or 100 kDa for GrsB. Proteins were aliquoted in 10% glycerol and flash frozen in liquid nitrogen for storage at -80 °C. Protein concentrations were determined from the absorbance at 280 nm measured in Take3 plates on an epoch2 microplate reader (Biotek) using calculated extinction coefficients (www.benchling.com). PNP was stored at a concentration of 500 µM in aliquots of 1.3 mg.

Adenylation kinetics

Michaelis-Menten parameters of adenylation reactions were determined from kinetic data recorded with the MesG/hydroxylamine assay which was performed as described previously with minor modifications.[2] Reactions contained 50 mM TRIS (pH 7.6), 5 mM $MgCl_2$, 100 µM 7-methylthioguanosine (MesG; PR3790-B100, Biosearch Technologies), 150 mM hydroxylamine (adjusted to pH 7.5-8 with NaOH), 5 mM ATP (A2383, Sigma), 1 mM TCEP, 0.4 U/mL inorganic pyrophosphatase (I1643, Sigma), 50 µM of PNP purified from *E. coli* in-house, the NRPS protein of interest (0.1 µM), and a suitable amino acid substrate. The amino acids *rac*-3,5-F_2-Phe, *rac*-2-F-Phe, and *rac*-3-F-Phe were purchased from abcr GmbH, while *rac*-2,4-F_2-Phe and *rac*-4-F-Phe were purchased from Fluorochem Ltd. In flat-bottom 384-well plates (781620, Brand), reactions were started in a total volume of 100 µL by addition of substrate. Then, the absorbance was followed at 355 nm on a Synergy H1 (BioTek) microplate reader at 30 °C. Slopes for the background activity were recorded in wells containing buffer but not substrate and were subtracted. Each substrate concentration was measured as biological triplicate. Initial velocities were divided by the slope of a pyrophosphate calibration curve to obtain the pyrophosphate release rate. Michaelis-Menten parameters were extracted from the initial velocity $v_0/[E_0]$ data through nonlinear regression in the R software package version 3.4.2.[3]

Isothermal titration calorimetry

For isothermal titration calorimetry (ITC), the preparation of protein was slightly modified. Protein expression and purification was done as described above from precultures of *E. coli* BL21 carrying plasmid pOPINF-PheA_ND for expression of the core-domain of GrsA-A. Then, a Vivaspin filter with a 10 kDa cutoff was used for buffer exchange into low salt buffer (100 mM TRIS [pH 7.4], 20 mM NaCl). To achieve the required purity for ITC, the protein sample was further purified by anion exchange chromatography on an NGC chromatography system (Bio-Rad Laboratories) using a MonoQ 5/50 GL column (GE Healthcare). Protein was eluted with a gradient of 20 to 600 mM NaCl in 20 mM TRIS (pH 8). After anion exchange, buffer was exchanged into storage buffer (100 mM HEPES [pH 8], 10% glycerol) using a Vivaspin filter. As before, aliquots were flash frozen in liquid nitrogen and stored at -80 °C.

For ITC measurements, 800 µL of a 60 µM protein solution were prepared in HEPES buffer (100 mM HEPES [pH 8], 10% glycerol). Substrate stocks (*rac*-Phe, *rac*-4-F-Phe and *rac*-2,4-F_2-Phe) were prepared at 6.25 mM in HEPES buffer. Protein samples were loaded into the cell of a MicroCal PEAQ-ITC (Malvern Panalytical) and titrated against the substrate solution at 25 °C at 750 rpm stirrer speed with 15 injections of 2 µL substrate. The initial delay was 60 s, the reference power was set to 10.0 µcal/s, the feedback to high, the injection spacing to 150 s and the injection duration to 4 s. Data analysis was performed with the MicroCal PEAQ-ITC Analysis software from Malvern Panalytical.

Hydroxamate assay (HAMA)

Hydroxamate assays for adenylation specificity were performed as previously reported.[1] In brief, HAMA was conducted at room temperature in 100 µL volume containing 50 mM TRIS (pH 7.6), 5 mM $MgCl_2$, 150 mM hydroxylamine (pH 7.5-8, adjusted with NaOH), 5 mM ATP (A2383, Sigma), 1 mM TCEP, and 1 mM proteinogenic amino acids with D-Val, D-Phe, deuterated L-Leu-d7, deuterated L-Phe-d5, deuterated L-Val-d8, *rac*-4-F-Phe, and *rac*-2,4-F_2-Phe. Reactions were started by adding the NRPS protein to a final concentration of 0.1 µM. Samples were incubated for 30 min and 10 µL of the reaction was quenched in 95% ACN containing 0.1% formic acid (A117-50, Fisher Scientific), cooled down and centrifuged to remove the precipitated hydroxylamine. As control, heat denatured enzyme was used. Samples were measured as biological duplicates.

Samples were analyzed on a UPLC-MS/MS (Xevo TQ-S micro, Waters) as described previously[1] with the same mass transitions and MS conditions. The mass transitions for the detection of 4-F-PheHA and 2,4-F_2-PheHA are 199.18 → 138.08 and 217.17 → 156.13, respectively.

Screening a GrsA-W239X library using HAMA

The screening method has been adapted from a published procedure.[4]

Protein expression. *E. coli* HM0079 hosting the plasmid library pSU18-GrsA-W239/NNK was used to overexpress GrsA-W239 variants in a 96-well plate format. Precultures were prepared by inoculating the transformants picked from an agar plate into a round bottom 96-well plate (310 µl, Sarstedt) filled with 150 µl of 2xYT medium supplemented with 100 µg/ml of chloramphenicol (3886.2, Roth). The 96-well plate contained four wells with *E. coli* HM0079::pSU18-GrsA and four wells with *E. coli* HM0079::pSU18-GrsA/ W239S as controls. Plates were covered with a breathable polyurethane film (Breathe-Easy, Sigma-Aldrich) and incubated for 18 h at 30 °C and 400 rpm in an orbital shaker. The following liquid handling steps were typically performed using a Gilson Platemaster 220 µL as 96-well pipette. For protein expression, 20 µl of the preculture was inoculated into a 96-deep-well plate (2 mL, Sarstedt) containing 1 ml 2xYT medium supplemented with 100 µg/ml chloramphenicol and incubated for 5 h at 30 °C and 400 rpm. A 20 µL aliquot was taken from the culture and stored with 25% glycerol at -70 °C for sequencing. To start the induction process, the cultures were cooled down to 18 °C for 30 min, followed by addition of 0.25 mM IPTG (BP1755-10, Fisher Scientific). Cells were incubated overnight for 18 h at 18 °C at 400 rpm. Cells were harvested by centrifugation at 3,220 g and 10 °C for 5 min and the supernatant was discarded. Lysis buffer (50 mM TRIS [pH 8.0], 100 mM NaCl, 10 mM imidazole, 1.5 mg/mL lysozyme) was prepared freshly by adding 1 µL/mL of protease inhibitor mix (APE-K1010, APExBio). Per well, 400 µL lysis buffer was used for resuspension of the cells. Cells were incubated at room temperature for 30 min, followed by freezing at -20 °C. Lysis was achieved by thawing the cells for 2 h at room temperature.

Protein purification. After thawing, 100 µL of DNA removal mix (50 mM TRIS [pH 8.0], 100 mM NaCl, 10 mM imidazole, 10 mM $MgCl_2$, 10 mM TCEP, 15 U/mL Turbonuclease [Jena Bioscience]) was added to reduce the viscosity of the lysate. Cell debris was removed by centrifugation at 3,220 g and 6 °C for 30 min. In a separate, 96-well plate (1.8 mL, Sarstedt) compatible with the magnetic separation rack (S1511S, New England Biolabs), 20 µl of a 25% Ni-IDA MagBeads (PureCube) suspension was added. For equilibration of the beads, 700 µL of lysis buffer was used and supernatant was discarded. Next, 400 µL of lysate was added to the equilibrated beads. The plate was covered with a silicon lid and kept at 6 °C for 20 min. Every 5 min the plate was shaken vigorously to resuspend the beads. Beads were pelleted on a magnetic rack and supernatant was discarded. Beads were washed twice with 700 µl of wash buffer (50 mM TRIS [pH 8.0], 100 mM NaCl) using the magnetic rack for separation of beads and wash fraction.

HAMA in 96-well plate format. After the second washing step, 100 µl of freshly prepared HAMA master mix (50 mM TRIS [pH 8.0], 5 mM ATP, 5 mM $MgCl_2$, 100 mM hydroxylamine adjusted to pH 7.5-8 with NaOH, 1 mM TCEP, 1 mM amino acid mix) was added directly to the beads containing the adsorbed protein and incubated at room temperature for 1.5 h. The amino acid mix contained the proteinogenic amino acids with D-Val, D-Phe, deuterated L-Leu-d7, deuterated L-Phe-d5, deuterated L-Val-d8, *rac*-4-F-Phe and *rac*-2,4-F_2-Phe. After incubation, 6 µl of the reaction mixture was diluted with 54 µl of analysis solution (acetonitrile with 0.1% formic acid) in a 384-well plate (100 µL, Brandt). After the dilution step, the 384-well plate was immediately placed on ice and covered with aluminum foil to minimize evaporation of the solvent. The plate was analysed immediately by UPLC-MS/MS according to the general HAMA procedure.

In vitro gramicidin S formation

For *in vitro* biosynthesis of GS, reactions were performed for 2 h at 37 °C in 100 µL PCR tubes in reaction buffer (50 mM HEPES [pH 8], 1 mM $MgCl_2$, and 100 mM NaCl). Amino acids L-Leu, L-Pro, L-Orn, and L-Val were added to a final concentration of 20 mM each. The amino acids *rac*-4-F-Phe or *rac*-2,4-F_2-Phe were added to a concentration of 5 mM. As positive control, 5 mM *rac*-Phe was used. GrsA or GrsA/W239S were added to a concentration of 1 µM. GrsB was added to a concentration of 3 µM, pretending that the protein was homogenous (for SDS-PAGE, see Fig. S1). Denatured GrsB was used as negative control. Glycerol was added to a final concentration of 2%. To start the reaction, ATP was added to a concentration of 5 mM. Reactions were quenched by the addition of 100 µL MeOH. Samples were cooled down for 20 min at -20 °C and centrifuged at 12,000 g for 5 min. From the supernatant, 160 µL were taken and mixed with 240 µL of 50% EtOH in water containing 0.1% formic acid. Samples were run on a UPLC-MS/MS (Xevo TQ-S micro, Waters) with an H-class UPLC. For chromatography, a CSH C18 column (186005296, Waters) was used with a linear gradient of 40% ACN and 60% water containing 0.1% formic acid to 98% ACN and 2% water containing 0.1% formic acid over 1 min, followed by 1.2 min reequilibration. Analytes were detected in MRM mode based on the mass transitions for GS (571.696→70.099) and the fluorinated analogs (589.85→70.0991, 607.6757→70.0991). These MRMs use the pyrrolidinium fragment (70.0991) of Pro for quantification. Authentic GS purified from the natural producer *Aneurinibacillus migulanus*[5] was used as a standard for quantification assuming an identical response for GS and the fluorinated analogs.

In vivo production of GS, 4-F-Phe-GS and 2,4-F_2-Phe-GS

Samples containing complete TB (3 mL) with chloramphenicol (25 µg/mL) were inoculated with a starter culture (1/500 v/v) of *E. coli* HM0079 cells transformed with plasmids pSU18-grsTAB/W239S and incubated at 30 °C, 230 RPM until OD_{600} = 2 was reached. Samples were divided into three groups with two biological replicates each. To group one, *rac*-2,4-F_2-Phe (2.5 mM) was added. To group two, *rac*-4-F-Phe (4 mM), and to group three (control group), only deionized water was added. All cultures were sampled after 96 h (1 mL), and clarified by centrifugation (11,000 g, 4 min, RT). The supernatants were removed and the cell pellets, which contain most of the GS, were resuspended in 70% ethanol, followed by sonication for 10 min and further incubation at 60°C for 30 min. The cell debris was removed by centrifugation (19,000 g, 4 min, RT). The supernatants were further diluted 50-fold in 50% ethanol containing 0.1% formic acid and the concentrations of GS, 4-F-Phe-GS and 2,4-F_2-Phe-GS were quantified using UPLC-MS/MS in comparison with a GS standard purified from *A. migulanus.*[5] Concentrations were calculated assuming that GS, 4-F-Phe-GS, and 2,4-F_2-Phe-GS were homogenously distributed in the culture volume. A control experiment was conducted in parallel with *E. coli* HM0079 cells transformed with plasmid pSU18-grsTAB instead of the mutated variant.

UPLC-MS/MS analysis was performed on a Waters ACQUITY H-class UPLC system coupled to a Xevo TQ-S micro (Waters) tandem quadrupole instrument under optimized conditions. The injection volume was 2 µL and the flow rate was 0.5 mL min^{-1}. Acetonitrile (B) and water with 0.1% formic acid (A) were used as strong and weak eluent, respectively. Acetonitrile was used as the needle wash between the samples. Data acquisition and quantification were done using the MassLynx software (version 4.1). MS/MS analyses were performed using an ESI source in positive ion mode. Nitrogen was used as desolvation gas and argon as collision gas. The following source parameters were used: capillary voltage 0.5 kV, desolvation temperature 600 °C, desolvation gas flow 1000 L h^{-1}.

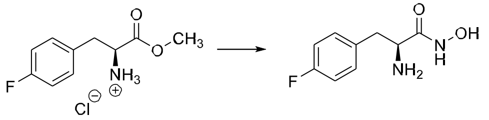
Column: ACQUITY UPLC CSH C18, 1.7 µm particle size, 2.1 × 50 mm
Elution profile: linear gradient of 40 to 98% B over 1 min followed by 1.2 min re-equilibration.
GS MRM transition: 571.696 > 70.099
4-F-Phe-GS MRM transition: 589.85 > 70.099
2,4-F_2-Phe-GS MRM transition: 607.676 > 70.099
GS calibration curve range: 0.01 µM to 10 µM

Synthesis of amino acid hydroxamates

All chemicals for synthesis were purchased from Merck, Alfa Aesar, Acros Organics, Fluorochem or TCI and used without further purification. The solvents were dried according to standard conditions if needed. The TLC-glass-plates DURASIL consisted of a 0.25 mm layer of silica 60 with fluorescence indicator UV254. TLCs were checked under UV-light (254 nm or 365 nm) and stained with an aq. $KMnO_4$ solution. ^1H, ^{13}C and ^{19}F NMR spectra were measured on BRUKER Fourier 500 or a BRUKER Avance 400 spectrometers. The chemical shift of each signal was reported in ppm. For ^1H and ^{13}C measurements, the chemical shift refers to TMS, showing a signal at 0 ppm. As an internal standard, the residual ^1H or ^{13}C nuclei of the corresponding deuterated solvents were used (DMSO-d_6, 2.50 ppm [^1H-NMR], 39.51 ppm [^{13}C-NMR]). The chemical shift of the fluorine NMR was determined indirectly. For carbon spectra, a broadband decoupling was performed. High-resolution mass spectra (HRMS) were measured using a Thermo Q-Exactive plus device with an ESI source coupled to a binary UHPLC system. IR spectra were measured using the Shimadzu IR-Affinity-1 (FTIR) device.

Synthesis of 4-F-Phe hydroxamate

Hydroxamate **2** was prepared from commercially available methyl ester **1**.



1 **2**

(*S*)-2-amino-3-(4-fluorophenyl)-N-hydroxypropanamide **2**

A suspension of 4-fluoro-L-phenylalanine methyl ester hydrochloride **1** (91.60 mg, 0.39 mmol) in dichloromethane (10 mL) was washed with saturated aqueous K_2CO_3 solution (2 x 10 mL). The combined aqueous layers were washed with dichloromethane (3 x 5 mL). The combined organic layers were dried over anhydrous K_2CO_3 and concentrated under reduced pressure to provide 4-fluoro-L-phenylalanine methyl ester free base as a colourless oil (72 mg, 0.37 mmol, 93%). It was subjected to the next transformation without further purification.

A 5.0 M solution of KOH in dry methanol (0.2 mL, 1.09 mmol, 3 equiv.) was added to a 1.0 M solution of hydroxylamine hydrochloride in dry methanol (1 mL, 1.09 mmol, 3 equiv.) at 0 °C. The resulting mixture was kept at 0 °C for 15 min and was filtered through a Teflon 2.5 µm filter to remove the precipitate. The filtrate was then added to a solution of 4-fluoro-L-phenylalanine methyl ester free base (72 mg, 0.37 mmol, 1 equiv.) in dry methanol (1 mL) and kept at -20 °C without stirring to facilitate crystallization. After 5 days, precipitate was filtered, washed with dry methanol, and dried under vacuum to afford the desired hydroxamate **2** as a colourless solid (10 mg, 0.05 mmol, 10%). The product was stored at -20 °C.

^{1}H NMR (500 MHz, DMSO-d_6) δ ppm 7.16 - 7.26 (m, 2 H), 7.08 (br t, *J* = 8.79 Hz, 2 H), 3.14 - 3.27 (m, 1 H), 2.80 (br dd, *J* = 13.27, 5.97 Hz, 1 H), 2.61 (br dd, *J* = 13.35, 7.71 Hz, 2 H).

^{13}C NMR (126 MHz, DMSO-*d*$_6$) δ ppm 170.98, 160.31 (br d, *J* = 90.76 Hz), 134.85 (d, *J* = 2.99 Hz), 131.00 (d, *J* = 7.98 Hz), 114.67 (d, *J* = 20.94 Hz), 54.39.

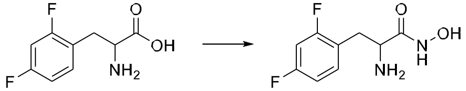
^{19}F NMR (471 MHz, DMSO-*d*$_6$) δ ppm -117.35.

HRMS [ESI]: m/z calculated for $C_9H_{11}FN_2O_2$ $[M+H]^+$ 199.0877, found 199.0875.

IR (ATR): ν = 2825 (w), 1608 (s), 1508 (s), 1489 (m), 1474 (m), 1383 (m), 1289 (w), 1161 (w), 891 (m).

Synthesis of 2,4-F$_2$-Phe hydroxamate

Hydroxamate **4** was prepared from commercially available amino acid **3.**



3 **4**

(*R,S*)-2-amino-3-(2,4-difluorophenyl)-N-hydroxypropanamide **4**

To a suspension of amino acid **3** (201 mg, 1 mmol, 1 equiv.) in dry methanol (4 mL), thionyl chloride (177 µL, 2.5 mmol, 2.5 equiv.) was added. The resulting mixture was stirred at 80 °C. After the reaction was completed, the reaction mixture was concentrated under reduced pressure. The residue was dissolved in saturated aqueous K_2CO_3 solution and extracted with dichloromethane (3 x 15 mL). The combined organic layers were dried over anhydrous K_2CO_3 and concentrated under reduced pressure to give methyl 2-amino-3-(2,4-difluorophenyl)propanoate as a colourless oil (186 mg, 0.87 mmol, 87%). The product was subjected to the next transformation without further purification.

A 5.0 M solution of KOH in dry methanol (0.5 mL, 2.5 mmol, 3 equiv.) was added to a 1.0 M solution of hydroxylamine hydrochloride in dry methanol (2.5 mL, 2.5 mmol, 3 equiv.) at 0 °C. The resulting mixture was kept at 0 °C for 15 min and was filtered through a Teflon 2.5 µm filter to remove the precipitate. The filtrate was then added to a solution of methyl 2-amino-3-(2,4-difluorophenyl)propanoate (176 mg, 0.81 mmol, 1 equiv.) in dry methanol (0.8 mL) and kept at -20 °C without stirring to facilitate crystallization. After 2 days, precipitate was filtered, washed with dry methanol, and dried under vacuum to give the desired hydroxamate **4** as a colourless solid (55.3 mg, 0.25 mmol, 31%). The product was stored at -20 °C.

^{1}H NMR (400 MHz, DMSO-*d*$_6$) δ ppm 8.04 - 10.65 (m, 1 H), 7.29 (br d, *J* = 7.31 Hz, 1 H), 7.14 (br t, *J* = 8.77 Hz, 1 H), 6.99 (br s, 1 H), 3.10 - 3.27 (m, 1 H), 2.80 (br dd, *J* = 12.42, 5.70 Hz, 1 H), 2.57 - 2.71 (m, 1 H).

^{13}C NMR (101 MHz, DMSO-*d*$_6$) δ ppm 170.87, 160.93 (dd, *J* = 244.49, 13.00 Hz), 160.63 (dd, *J* = 247.52, 13.44 Hz), 132.72 (dd, *J* = 9.10, 6.50 Hz), 121.56, 110.99 (dd, *J* = 20.37, 3.03 Hz), 103.39 (t, *J* = 26.01 Hz), 53.19, 33.81.

^{19}F NMR (376 MHz, DMSO-*d*$_6$) δ ppm -113.30, -113.52.

HRMS [ESI]: m/z calculated for $C_9H_{10}F_2N_2O_2$ $[M+H]^+$ 217.0783, found 217.0778.

IR (ATR): ν = 3186 (w), 2897 (w), 1624 (s), 1605 (m), 1541 (m), 1508 (s), 1379 (s), 1292 (m), 1265 (m), 1128 (s), 978 (m), 891 (s), 743 (w).

Computational Methods

Protein structure preparation

The X-ray protein structure of the GrsA A-domain in complex with AMP and Phe (PDB ID 1amu)[6] was loaded into Schrödinger's Maestro and chain B of the dimer was deleted. The remaining structure was prepared with the Protein Preparation Wizard[7] in default settings, which includes adding missing hydrogens and sidechains, deleting waters far from the natural ligand and running a short MM energy minimization. The W239S mutant was modeled by mutating the residue and briefly minimizing the obtained structure inside Maestro. The fluorinated ligands 4-F-Phe, 2,4-F$_2$-Phe, 2-F-Phe, 3-F-Phe, and 3,5-F$_2$-Phe were constructed by replacing the respective hydrogens of the natural ligand Phe with the Maestro Build tool.

Classical force field distance scan

The potential energy of a system of capped fluorinated and unfluorinated Phe, which is placed perpendicularly above the aromatic system of capped Trp, was calculated at various distances in vacuum. The capped Trp molecule was constructed by extracting Trp239 from the prepared protein structure of WT GrsA and then capping the N-terminal end with an acetyl cap (ACE) and the C-terminal end with an N-methyl cap (NME). The capped Phe and fluorinated Phe analogues were constructed in a similar way, by extracting the ligands from the prepared PDB structure and then applying ACE and NME caps. The Phe molecule was placed directly above the center between CD2 and CE2 of the Trp molecule. The distance between this center and CZ of the Phe molecule was adjusted to cover distances between 0.3 nm and 1.1 nm in steps of 0.05 nm. For each of the systems one structure was energy minimized with the Gromacs[8–10] 2021 software using a steepest descent algorithm and positional restraints on every heavy atom of the structure. The force field used was Amber14SB[11] with GAFF2[12] parameters for the fluorinated amino acids. All parameters were generated using Acpype.[13–15] The total potential energy of the system after energy minimization was then calculated using Gromacs energy. For each system the total potential energy for the distance of 1.1 nm was set to zero and the difference to this energy was calculated for the other distances.

Molecular dynamics simulation

Molecular dynamics (MD) simulations were run using the Gromacs[8–10] 2021 software and the Amber14SB[11] force field, combined with GAFF2[12] parameters, generated with Acpype,[13–15] for the fluorinated amino acid. The simulations were initiated with the prepared structures of GrsA (PDB ID 1amu) in complex with Phe and —F-Phe as starting structures. The three systems that were simulated were WT GrsA in complex with Phe, GrsA-W239S in complex with Phe and GrsA-W239S in complex with 4-F-Phe. The structures were solvated in a cubic box with periodic boundary conditions and with 1.1 nm between the solute and box edges in TIP3P[16] water. The systems were energy minimized using a steepest descent algorithm, followed by an equilibration in the NVT ensemble for 100 ps at 300 K with restraints on all solute heavy atoms and a subsequent unrestrained equilibration in the NPT ensemble at 300 K and 1.0 bar for 1 ns. The production simulations were run in the NPT ensemble for 10 ns. The temperature was kept constant with a velocity rescaling scheme with a stochastic term[17] and the employed barostat was the Parinello-Rahman barostat.[18]

The special region of the cavity next to Trp/Ser239 was modeled by a sphere of 0.4 nm radius. The center of the sphere was placed at the center of the straight line between the C-alpha atoms of Trp/Ser239 and Thr334 in the prepared starting structure. The position of the sphere was not moved throughout the simulation. The number of water molecules inside the sphere was counted throughout the simulation and divided by the number of simulation snapshots to obtain the percentages (Table S3).

Molecular docking of Phe analogs to GrsA-W239S

The ligands Phe, 4-F-Phe, 2,4-F_2-Phe and O-propargyl-Tyr were docked against the W239S mutant of GrsA using the GLIDE[19–21] docking software. A receptor grid was generated based on the prepared structure of GrsA-W239S. The docking was run in extra precision mode with flexible ligand sampling and penalized non-planar amide conformations. Core constraints were applied by restricting the ligands to the reference position of the natural Phe ligand in the PDB structure 1amu with a tolerance of 0.1 Å. The core atoms were defined as the maximum common substructure. The preferred docking pose is shown in Fig. S4 and the docking scores in Table S4.

Supporting Figures

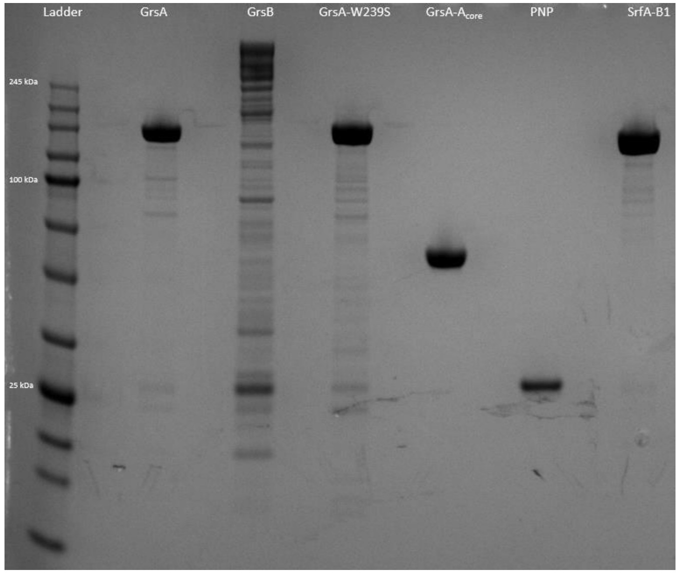


Figure S1. Analysis of protein purity by SDS-PAGE.

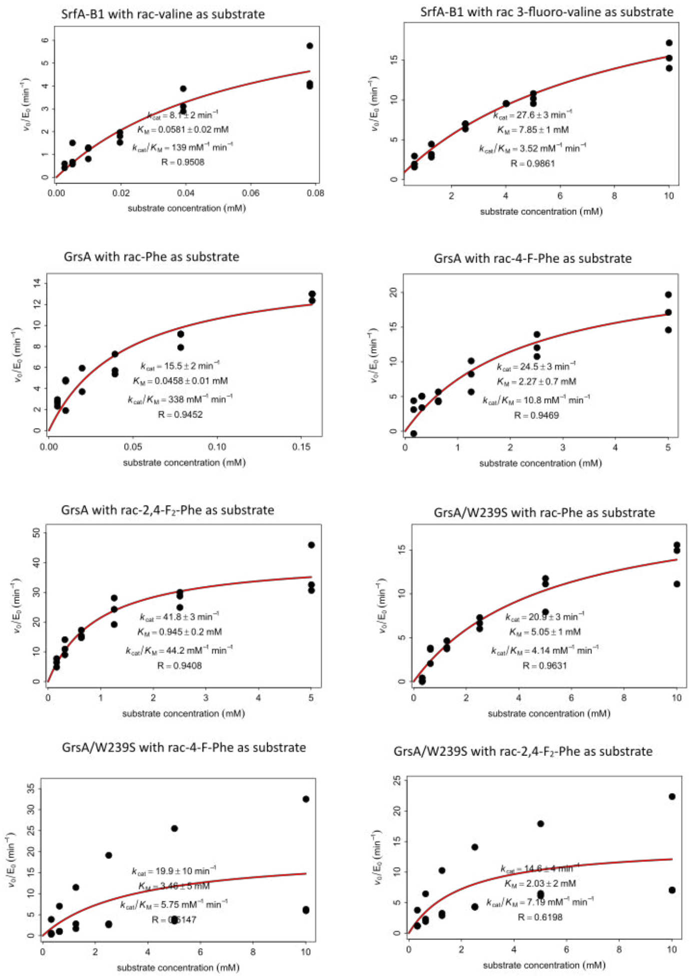


Figure S2. Adenylation kinetics (continued on next page).

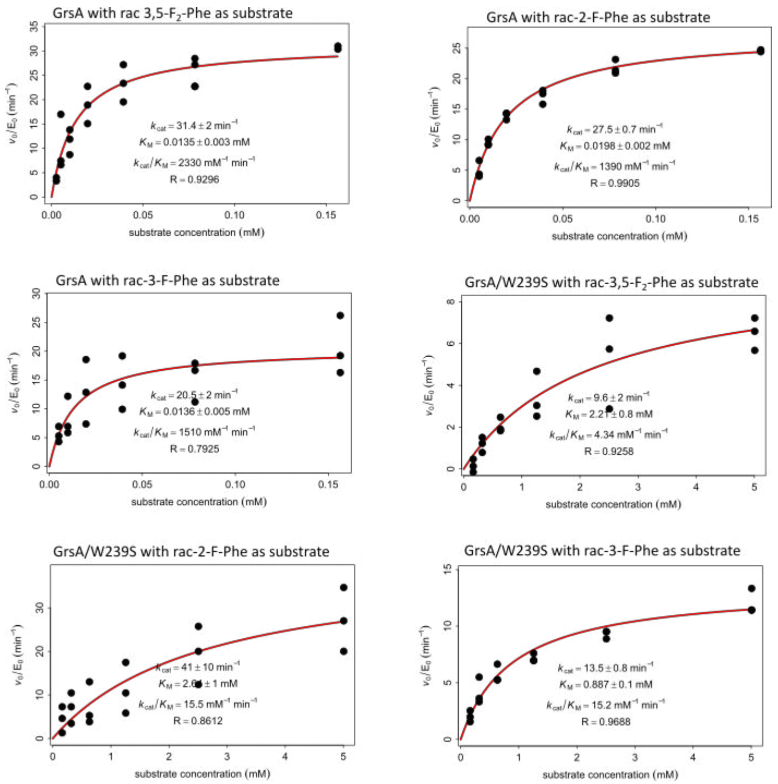


Figure S2. Continued.

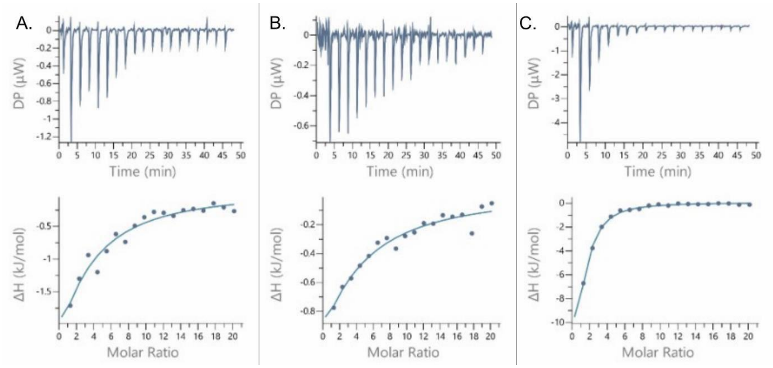


Figure S3. Representative thermograms of amino acid binding to GrsA-A$_{core}$ (60 μM) measured by ITC. A) Titration with 6.25 mM *rac*-2,4-F$_2$-Phe. Four replicates were recorded. B) Titration with 6.25 mM *rac*-4-F-Phe. To compensate for the weak signal, 13 replicates were recorded. C) Titration with 6.25 mM *rac*-Phe. Three replicates were recorded.

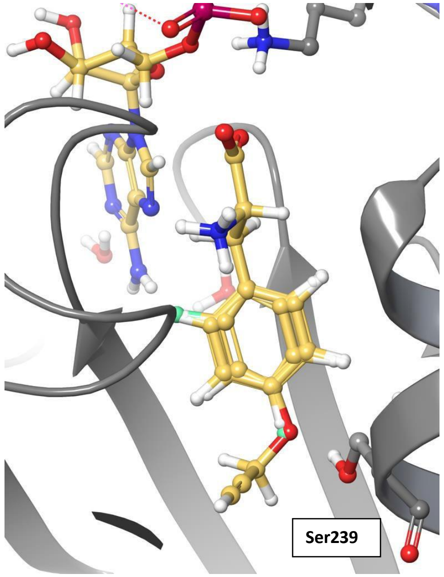


Figure S4. Preferred docking poses of the ligands 4-F-Phe, 2,4-F$_2$-Phe and *O*-propargyl-Tyr in the modelled structure of GrsA-W239S are superposed. Core constraints are applied on the docking poses based on the natural substrate in pdb structure 1amu, so the core region of the ligands has little degrees of freedom.

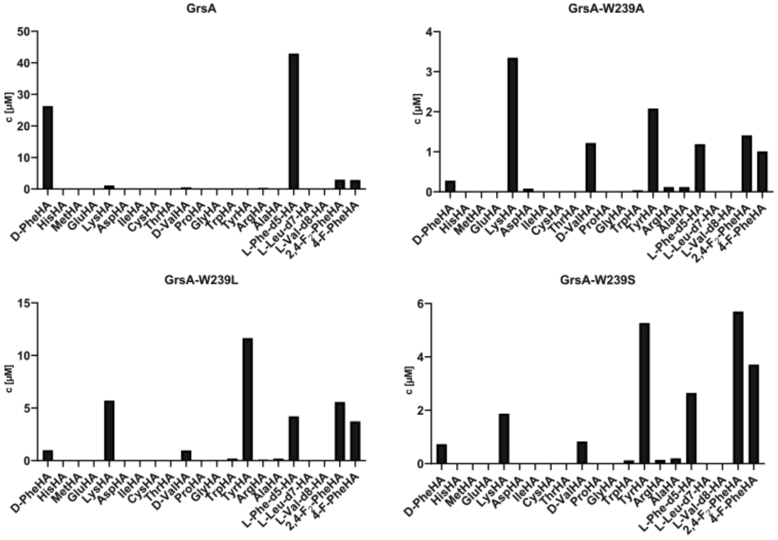


Figure S5. Selected mutants in position GrsA-W239 showing improved selectivity for 4-F-Phe in the HAMA 96 well screening experiment. The mutant W239S shows the highest selectivity towards the fluorinated substrates.

Supporting Tables

Table S1. List of primers.

Primer names	Sequence (5' - 3')
GrsA-ND_f	AAG TTC TGT TTC AGG GCC CGA TGT TAA ACA GTT CTA AAA G
GrsA-ND_r	ATG GTC TAG AAA GCT TTA TAT TCT TCC GAG ATA TTC AAT ATT TCC
grsA/W239S-P1-F-b	AAG CGG AGT ATC CAC GTG ATA AGA CGA TCC ATC AGT TAT TTG AAG AGC AG
grsA/W239S-P1-R-a	TAC AGA TGC ATC AAA AGA GAT GCT GGC AAA
grsA/W239S-P2-F-b	TCT CTT TTG ATG CAT CTG TAA GCG AGA TGT TTA TGG C
grsA/W239S-P2-R-b	TCT TTA CTA GAG GGC CTA CTT CCA AGT TTA TAC TAT TTT GTA ATC GAG CA
SrfA_Nested1_f	CTA TTT AGG TCA GTT TGA CGA AAT G
SrfA_Nested1_r	CTT GGG CAC GAA GAT GAT G
B1_isolation_Nested2_CATCf3	CAC AGG AAA CAG ACC ATG AGC AAA AAA TCG ATT CAA
B1_isolation_Nested2_CATr2	ATG GTG ATG AGA TCT CAA ATA CAG TGC CAG TTC TTG AAT A
GrsA_f	AAG AGG AGA AAT TAA CCA TGT TAA
GrsA_r	TAC AGA TGC ATC AAA AGA GAT G
GrsA_W239NNK_f	CAT CTC TTT TGA TGC ATC TGT ANN KGA GAT GTT TAT GGC T
GrsA_W239NNK_r	GAT GGT GAT GAG ATC TGG A

Table S2. Thermodynamic parameters of amino acid binding to GrsA-A$_{core}$.

Enzyme	Substrate	K_D [µM]	ΔG [kJ/mol]	ΔH [kJ/mol]	- TΔS [kJ/mol]
GrsA-A$_{core}$	Phe	60±10	-24.1±0.5	-20±2	-4±3
	4-F-Phe	600±300	-19±2	-11±6	-7±7
	2,4-F$_2$-Phe	420±40	-19.3±0.2	-15±2	-5±2

Table S3. Percentage of simulation snapshots that show the according number of water molecules inside the cavity throughout a 10 ns MD simulation of GrsA (wild type and W239S mutant) in complex with the substrates Phe and 4-F-Phe.

GrsA variant	Substrate	Number of water molecules				
		0	1	2	3	4
wild type	Phe	100%	0%	0%	0%	0%
W239S	Phe	0.1%	4.1%	41.0%	44.7%	9.6%
W239S	4-F-Phe	18.9%	39.7%	33.2%	8.1%	0.1%

Table S4. Docking scores of Phe analogs docked to GrsA-W239S.

Ligand	Docking Score
O-propargyl-Tyr	-8.425
2,4-F$_2$-Phe	-7.560
4-F-Phe	-7.205
Phe	-7.049

References

1 A. Stanišić, A. Hüsken and H. Kries, *Chem. Sci.*, 2019, **10**, 10395–10399.

2 D. J. Wilson and C. C. Aldrich, *Analytical biochemistry*, 2010, **404**, 56–63.

3 R Core Team, R: A Language and Environment for Statistical Computing http://www.r-project.org 2017.

4 A. Stanišić, C.-M. Svensson, U. Ettelt and H. Kries, *bioRxiv*, 2022, 2022.08.30.505883.

5 F. Pourmasoumi, S. De, H. Peng, F. Trottmann, C. Hertweck and H. Kries, *ACS Chem. Biol.*, 2022, **17**, 2382–2388.

6 E. Conti, T. Stachelhaus, M. A. Marahiel and P. Brick, *The EMBO journal*, 1997, **16**, 4174–83.

7 G. Madhavi Sastry, M. Adzhigirey, T. Day, R. Annabhimoju and W. Sherman, *J Comput Aided Mol Des*, 2013, **27**, 221–234.

8 M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess and E. Lindahl, *SoftwareX*, 2015, **1–2**, 19–25.

9 S. Páll, M. J. Abraham, C. Kutzner, B. Hess and E. Lindahl, in *Solving Software Challenges for Exascale*, eds. S. Markidis and E. Laure, Springer International Publishing, Cham, 2015, pp. 3–27.

10 S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess and E. Lindahl, *Bioinformatics*, 2013, **29**, 845–854.

11 J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser and C. Simmerling, *J. Chem. Theory Comput.*, 2015, **11**, 3696–3713.

12 X. He, V. H. Man, W. Yang, T.-S. Lee and J. Wang, *The Journal of Chemical Physics*, 2020, **153**, 114502.

13 A. W. Sousa da Silva and W. F. Vranken, *BMC Research Notes*, 2012, **5**, 367.

14 J. Wang, W. Wang, P. A. Kollman and D. A. Case, *Journal of Molecular Graphics and Modelling*, 2006, **25**, 247–260.

15 J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, *Journal of Computational Chemistry*, 2004, **25**, 1157–1174.

16 W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *The Journal of Chemical Physics*, 1983, **79**, 926–935.

17 G. Bussi, D. Donadio and M. Parrinello, *The Journal of Chemical Physics*, 2007, **126**, 014101.

18 M. Parrinello and A. Rahman, *Journal of Applied Physics*, 1981, **52**, 7182–7190.

19 R. A. Friesner, R. B. Murphy, M. P. Repasky, L. L. Frye, J. R. Greenwood, T. A. Halgren, P. C. Sanschagrin and D. T. Mainz, *J. Med. Chem.*, 2006, **49**, 6177–6196.

20 R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry, D. E. Shaw, P. Francis and P. S. Shenkin, *J. Med. Chem.*, 2004, **47**, 1739–1749.

21 T. A. Halgren, R. B. Murphy, R. A. Friesner, H. S. Beard, L. L. Frye, W. T. Pollard and J. L. Banks, *J. Med. Chem.*, 2004, **47**, 1750–1759.