Dissertation zur Erlangung des Doktorgrades der Fakultät für Chemie und Pharmazie der Ludwig-Maximilians-Universität München

Discovery of protein-stabilizing Excipient Candidates

Andreas Tosstorff

aus

Konstanz, Deutschland

2019

Erklärung

Diese Dissertation wurde im Sinne von §7 der Promotionsordnung vom 28. November 2011 von Herrn Prof. Dr. Gerhard Winter betreut.

Eidesstattliche Versicherung

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

Cambridge, GB, 17. November 2019

Andreas Tosstorff

Dissertation eingereicht am: 10.12.2019

1. Gutachter: Prof. Dr. Gerhard Winter

2. Gutachter: Prof. Dr. Günther H.J. Peters

Mündliche Prüfung am 4. Februar 2020

Department of Pharmacy - Pharmaceutical Technology and Biopharmaceutics

LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

DISCOVERY OF PROTEIN-STABILIZING EXCIPIENT CANDIDATES

Andreas Tosstorff, M.Sc.

Acknowledgements

The work presented in this thesis was made possible by Prof. Dr. Gerhard Winter. When I decided to postpone my PhD after graduating, my former supervisor advised me that it might be very difficult to get a position after once leaving university. I am therefore particularly grateful that I was given this opportunity and consider Prof. Dr. Winter's commitment to giving second chances a real inspiration.

Dr. Hristo Svilenov was not only an excellent roommate in so many PIPPI trips, but a great friend, whose excellent scientific input was crucial to this work. I want to thank Prof. Dr. Günther Peters for hosting me at his lab in Copenhagen and Dr. Sowmya Indrakumar and Ulf Mollich for making this a wonderful experience that helped us start our molecular modeling journey. I am grateful to Dr. Sasha Golovanov and Matja Zalar for hosting me at their facility and their valuable scientific input. I want to thank Simon Eisele for the great collaboration studying protein excipient interactions, for dedicating a lot of time to helping me repair HPLC systems and for being a friend to always count on. I want to thank Prof. Dr. Wolfgang Frieß for interesting discussions at the weekly seminar, acting as co-referee of this thesis and for teaching me about the "Argentinische Rückhand". I thank Dillen Augustijn for helping with the development of a robust fitting protocol for the DSF data. I am grateful to Dr. Thomas Wein and Dr. Georg Höfner for fruitful discussions on virtual and experimental screening studies. I would like to thank the European Union's Horizon 2020 research and innovation program for funding this work with a Marie Skłodowska-Curie grant (agreement No. 675074).

Sharing the lab with Alice Hirschmann was a great experience. She has been a true friend, always there when needed, but especially at lunchtime, together with Dr. Michaela Breitsamer, which was as much fun as it was delicious. Dr. Teresa Kraus was the best coach one can think of and her positive attitude was absolutely contagious. I want to thank Tobias Keil for his friendship. Lorenzo Gentiluomo's spirit made the PIPPI meetings highly memorable. I am especially thankful to Andreas Stelzl for helping with FlowCam trouble, but also want to mention his commitment to group activities together with Ivonne Seifert and Ruth Rieser. Special thanks also to Dennis Krieg, Weiwei Liu, Julian Gitter, Inas ElBialy, Carolin Berner, Bernhard Haryiadi, Sebastian Groel and Ute Rockinger for all the fun at our social events and the positive work atmosphere. Dr. Moritz Vollrath, Dr. Letícia Rodrigues Neibecker and Dr. Katharina Geh made me feel very welcome at the group. Letícia's and Pascal's wedding in Montes Claros was a night to remember. The hard work in the background by Ayla Tekbudak, Susanne Petzel, Regine Bahr, Imke Leitner and Sabine Kohler must not go unnoticed. Whether ordering new substances or dealing with complex bureaucratic issues, they always made sure things ran smoothly. My students Marcel Passon, Luis Sánchez and Dominik Brandstetter put a lot of effort in their work and were of an enormous help to complete this thesis.

This work would not have been possible without the support of my friends and family. I am grateful to my friend Dr. Ezequiel Monteagudo, whose advice was crucial in choosing the research topic for my PhD. I want to thank Christoph Mähler for his close friendship ever since we started studying. He was of great help to set up in Munich after I returned from Argentina. I was in the fortunate position to share a home with my dear brother Martin Tosstorff, who happens

to be a passionate physicist and who's scientific advice was of highest value to my work.

Publications arising from this thesis

I. Tosstorff, A., Svilenov, H., Peters, G.H.J., Harris, P. & Winter, G. Structure-based Discovery of a new Protein-Aggregation Breaking Excipient, *Eur. J. Pharm. Biopharm.* **144**, 207-216 (2019).

II. Tosstorff, A., Menzen, T. & Winter G. Exploring Chemical Space for new Substances to stabilize a therapeutic Monoclonal Antibody, *J. Pharm. Sci.* **109**, 301-307 (2020).

III. Tosstorff, A., Peters, G.H.J., & Winter, G. Study of the interaction between a novel, protein-stabilizing dipeptide and Interferon-alpha-2a by construction of a Markov state model from molecular dynamics simulations, *Eur. J. Pharm. Biopharm.* **149**, 105-112 (2020).

Table of Contents

List of abbreviations

A3D	Aggrescan3D
APR	Attach-Pull-Release
BPTI	Bovine pancreatic trypsin inhibitor
CUDA	Compute Unified Device Architecture
DMP	Dexamethasone phosphate
ff14SB	Amber protein force field
GAFF2	General Amber force field 2
GIST	Grid inhomogeneous solvation theory
GPU	Graphical processing unit
IFN	Interferon-alpha-2a
IP	Infliction point of temperature dependent fluorescence signal curve
LMW	Low molecular weight
MD	Molecular dynamics
MM-GBSA	Molecular mechanics – generalized born surface area
MMP-3	matrix metalloproteinase 3

MSA	Molecular surface area
MST	Microscale Thermophoresis
Mw	Molecular weight
N_{bb}	Backbone amide nitrogen
NTA	Nitrilotriacetic acid
PDB	Protein database
pmemd	Particle-Mesh Ewald Molecular Dynamics
RESP	Restrained electrostatic potential
SASA	Solvent accessible surface area
SLS	Static light scattering
sPRE	Solution paramagnetic relaxation enhancement
TEMPOL	4-hydroxy-2,2,6,6-tetramethylpiperidin-1-oxyl
TIMP-1	tissue inhibitor of metalloprotease
T_{onset}	Temperature of onset of aggregation

1. Introduction

In the development of liquid formulations of therapeutic drug proteins, controlling the protein's stability is of utmost importance[1]. A protein's stability can be considered as its resistance to deviate from its original, native state. One of the major pathways of protein degradation is aggregation, which for a given protein and container can only be controlled by its solution environment[2]. Protein formulations are therefore optimized in terms of their pH, ionic strength and co-solutes. The latter are typically chosen from a limited list of historically used substances generally regarded as safe and are referred to as excipients[3]. It is therefore of interest to assess the protein-stabilizing potential of substances beyond the list of currently used excipients. This requires the development of systematic approaches that can identify new substances with the potential to stabilize proteins. The discovery of such new substances is for example of interest to formulate proteins in solution otherwise only stable in the dried state. New excipients could also help to achieve protein formulations that are stable at room temperature.

As outlined in this thesis, the challenge of discovering protein-stabilizing excipients can be faced by numerous strategies. Regardless of the approach chosen, understanding of protein aggregation and protein-excipient interactions is fundamental. Underlying both phenomena are non-covalent, pairwise atomic interactions which are described in the introduction. Low molecular weight substances are added to protein formulations, among other, with the purpose of inhibiting the formation of protein aggregation. Their mechanism of action is discussed briefly. Additionally, a structural elucidation of protein aggregation and protein-excipient interaction is helpful to understand and drive decision-making in excipient discovery. Therefore, examples of relevant applications are presented. The introduction is wrapped up by describing strategies to identify small molecules with certain desired physicochemical attributes, as it is common in drug discovery.

1.1. Non-covalent molecular interactions

Non-covalent interactions of molecules are ubiquitous in nature and of central importance in the development of drugs. They drive signaling in a cell and underly the mechanism of most drugs[4]. They are also responsible for self-interaction and aggregation of molecules. Non-covalent interactions are also involved when excipients stabilize a protein in its formulation[5]. A theoretical description of non-covalent interactions is exploited to simulate atomistic processes, for example by molecular dynamics simulations.

Electrostatic interactions are the strongest of interactions involved in non-covalent binding[6]. For two single point charges it is described by Coulomb's law, which considers the magnitude and position of the interacting charges and the

presence of any dielectric material surrounding them, by taking into account the dielectric constant. A common example are salt bridges that form if oppositely charged amino acid residues are in close contact to each other.

Van der Waals interactions describe an attractive interaction between neutral atoms. They are actually the combined effect of multiple phenomena[7,8]. These are electric dipole-dipole interactions (Keesom interactions)[9], induction (Debye interactions) or dispersion interactions (London interactions)[10]. Opposed to the attractive interactions of neutral atoms is a short range repulsive effect, as, when approximated, the atoms' electrons would not comply to the Pauli principle, as they have the same spin in the same location[11]. The Lennard-Jones-potential is commonly employed to describe the repulsive and the attractive interactions between neutral atoms. The parameter r describes the distance between two nuclei, ε is the depth of the potential well and σ the zero-potential distance. Attractive interactions scale with the power of 6, repulsive interactions with the power of 12 (Equation 1.1)[12].

$$V_{LJ}(r) = 4\varepsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \qquad \text{Equation 1.1}$$

Charge-dipole interactions form between ions and molecules with a dipole as for example water or CO and NH groups from the protein backbone. A dipole can however also be induced by either another dipole or a charge, also referred to as polarization[13]. Dipole-dipole interactions can lead to the stacking of amides due to their large dipole moment, leading to an antiparallel orientation[14,15].

In the context of biochemistry and drug development, hydrogen bonds are typically encountered between hydrogen carrying oxygen or nitrogen atoms

and lone pair carrying oxygen or nitrogen atoms[16]. The O-H or N-H groups both have a strong electric dipole moment, with the hydrogen carrying a positive partial charge, making it prone to interact with the free lone pair of oxygen or nitrogen atoms. Beta-sheets and alpha-helices in proteins are examples of structural motifs that form due backbone hydrogen bonds.

Classical hydrophobic interactions are driven by a loss of entropy caused by an increased order of water molecules in the first layer of a hydrophobic molecule's hydration shell. By minimizing the solvent exposed surface area through interaction with another hydrophobic binding partner, the number of ordered water molecules is minimized[17]. Non-classical hydrophobic interactions driven by enthalpy changes have also been observed[18].

π-stacking refers to a preferential orientation of aromatic systems, which are often observed to align in an off-centered, parallel or a centered, t-shaped fashion (Figure 1.1)[19]. The attractive forces governing π-stacking have been shown to be electrostatic and dispersion[20].

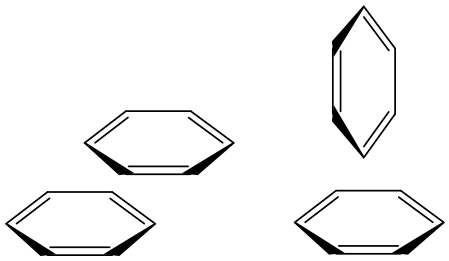


Figure 1.1: Relative configuration of benzene molecules upon favorable π-stacking interaction. Left: off-centered, parallel orientation. Right: centered, t shaped.

Cation-π interactions have been first observed between potassium and benzene[21]. Theoretical calculations showed that the cation would preferentially

be situated on top of the aromatic ring's center. Since their discovery, multiple examples involving aromatic and non-aromatic π-systems and various cations, have been described, also in biologic systems[4,22]. Cation-π interactions have long found their way into molecular dynamics force-fields, for example by adding an additional "10-12-term" to the Lennard-Jones potential[23].

Covalently bound atoms of the groups IV to VII show an anisotropic charge distribution, resulting in an area of positive electrostatic potential located opposite to the covalent bond, a so-called σ-hole. This leads to interactions with nucleophiles or Lewis bases[24]. Sigma-holes can for example lead to a preferred conformation or locking due to the interaction of Sulfur with a Nitrogen's lone pair[25].

Molecular interaction of course is more than the sum of the pairwise interactions described above. Conformational changes and solvation or desolvation occur often simultaneously, adding further complexity to the process.

1.2. Protein aggregation

Protein aggregation is a particular complex example of a molecular interaction relevant to many sectors in biotechnology, but especially to the pharmaceutical industry. While studied for long, it has been, on the one hand, a topic of recent attention due to its possible involvement in Alzheimer's disease[26]. On the other hand, with an ever-increasing amount of protein therapeutics being investigated and marketed, protein aggregation has been identified as a fundamental quality attribute[27]. Here, we will focus on the latter.

The first and most obvious reason for pharmaceutical companies to control and limit the aggregation of protein drugs is that it is imposed by regulatory agencies. The Food and Drug administration does so, for example in USP 787 and 788, by giving limits to the number of particles, which can form from aggregated proteins, while the European medicines agency specifies them in EuPh 2.9.19.[28]. Stress induced protein aggregation has been shown to cause inactivation, which can lead to a reduced drug efficacy[29]. Protein particles have been furthermore related to immune responses, leading to anti-drug antibody induced loss of efficacy and allergic reactions in the patient[30,31]. What property it is that distinguishes immunogenic from non-immunogenic aggregates remains a topic of ongoing investigation[32]. By comparing aggregates formed from different mAbs with an artificial lymph node, it was for example recently shown that the immune response depends on the aggregates' parent protein[33].

Reversible, native aggregation, i.e. native self-association, can be either driven by attractive electrostatic interactions resulting from an inhomogeneous charge distribution on the protein's surface or hydrophobic interactions that occur when the protein's net charge is low. The protein's net charge is controlled by its pH dependent protonation and charge shielding caused by ions present in the protein containing solution[34].

Non-reversible, non-native protein aggregation is driven by the previously described hydrophobic effect. Partially unfolded proteins expose hydrophobic regions that are buried in the protein core of the native conformation. The interaction of the hydrophobic regions of two partially unfolded proteins is therefore thermodynamically favored. The microscopic steps of the process of

formation of an aggregation prone, partially unfolded species and the nucleation can be elucidated by studying the kinetics of aggregation[35].

Mechanism of aggregation

Understanding the underlying mechanism in the formation of aggregates is of highest interest in the field of protein stability and formulation development, as it helps to design formulations according to the desired stability profile.

There are multiple characteristics of an aggregation mechanism, that are helpful to classify and understand a process. Aggregates may form in a reversible way, meaning that they will dissociate for example upon dilution or heating[36]. They may be constituted by monomeric units maintaining largely their native state, in which case the terms clustering, self-assembly or self-association is often preferred over the more general aggregation term. Insulin is one of the most common examples for native protein aggregation[37]. In the context of protein stability, non-native monomeric units have been recognized as being the dominant species in formed aggregates[38]. The presence of non-native protein molecules implies a conformational change in the monomeric unit which can occur either after or prior to self-association[39]. Non-native aggregates are considered to be irreversible unless exposed to elevated temperatures or chemical denaturants[40,41].

Aggregation kinetic studies help to identify the multiple elementary steps occurring in protein aggregation[42,43]. Partial unfolding, as mentioned before, is often observed in protein stability studies[44]. Here, conformations significantly different from the native state form and expose hydrophobic structural features which facilitate aggregation. During primary nucleation, two monomers form

an aggregate, either in bulk or at an interface. Aggregate growth occurs by the addition of monomers to aggregates, a process referred to as elongation in the case of fibrils[45]. In secondary nucleation, two monomers form an aggregate at the interface of another protein aggregate. It can be divided in three phases. First, two monomers adsorb to the aggregate surface. In a second step they react to form the new aggregate. Finally, the new aggregate desorbs from the other aggregate's surface[46]. As previously stated, non-native aggregates are irreversible, but aggregate decomposition can still occur to a certain extent, for example by fragmentation of filaments that leads to the exposure of additional nuclei[47]. Monomer dissociation, which is the opposite to elongation, is largely neglected in the description of aggregation processes as it is considered to be very slow, thus the irreversibility of aggregation. Its role becomes relevant only at very high aggregate to monomer ratios[45,46,48].

The described complexity in aggregation processes typically results in a non-Arrhenius behavior, making prediction of low temperature protein stability from accelerated, high temperature stability studies very challenging[49]. Isothermal chemical denaturation assays appear to be more appropriate predictors[50].

Aggregation mechanisms are susceptible to even minor changes in the protein structure, as was demonstrated for amyloid-beta 40 and 42 proteins, which differ by merely 2 amino acid residues[48]. The protein environment can also have an effect on the mechanism of aggregation. Small molecules have been demonstrated to inhibit different elementary steps of the aggregation of amyloid-beta 42[51].

Aggregation mechanisms do not only vary by protein structure and formulation conditions but of course also by the type of stress a protein is exposed to. Freezing and thawing will lead to a different pathway of aggregate formation compared to heating. Multiple simultaneously occurring stress factors are present during freezing and thawing, which are absent upon heating. These include interfacial stress from the formation of ice, cold denaturation, and up-concentration. A change in the reaction rates for the elementary steps seems therefore mandatory when going from heat stress to freeze-thaw stress. It has been shown for an antibody, that different stresses will involve different protein residues in the formation of oligomerization interfaces[52]. The interplay between structure, formulation, mechanism and stress appears to be highly complex.

Stabilization of proteins through low molecular weight co-solutes

Protein stabilization by a small molecule co-solute is aimed at increasing the free energy of unfolding, thus shifting the equilibrium from the non-native, aggregation prone, to the native, non-aggregation prone state of the protein. This can be achieved by either stoichiometric binding to the native state or preferred hydration.

Stabilization through stoichiometric binding requires that the small molecules affinity to the native state is higher than to the non-native one. This implies specificity of binding to a region that is prone to partially unfold, since binding to a region that largely retains its structure would not lead to the desired preferred affinity to the native state, but to similar affinities in both unfolded and folded states. The effect is often employed in thermal shift assays for drug discovery[53].

Instead of binding, preferential exclusion, also referred to as preferential hydration, can cause the same desired shift of equilibrium towards the native protein[54]. Preferential exclusion refers to the phenomenon of a decreased co-solute concentration at the protein surface relative to the bulk solution. This unfavorable state is proportional to the solvent accessible surface area (SASA) of the protein. Since unfolding will increase the protein's SASA, the amount of excluded co-solute will also increase. The non-native state is therefore less favored than the native state, resulting in a stabilization[5].

While for stoichiometric stabilization, the required co-solute concentration depends on its binding affinity, for preferential exclusion, the effective co-solute concentration is empirically found to be approximately 100 mM to 10 mM[55,56]. Both mechanisms of stabilization have been described for frozen and liquid protein formulations[57]. Stoichiometric binding and preferential exclusion are not mutually exclusive but can occur simultaneously for the same solute. Depending on the co-solutes concentration, one or the other phenomena is however more dominating regarding protein stability[57].

Structural biology, computational chemistry and bioinformatics in protein stability and formulation

The process of liquid formulation development for therapeutic proteins traditionally consists of an empirical screen of the different proteins or formulation parameters by high throughput stability indicating methods such as for example DSF or nanoDSF[58]. Alternatively, accelerated stability studies with multiple formulations can be performed. Based on thresholds for a stability indicating parameter such as for example the inflection point T_m or the monomer retention, the most stable formulations then proceed to long term

studies. Statistical analysis of the formulation and sequence design space by design of experiment or neural networks are now employed to accelerate the development process[59,60].

Opposed to these straightforward strategies are approaches that rely on extracting information from the protein structure. The best-known method to study protein structures is X-ray crystallography. The formation of protein crystals implies the self-association of the protein in a highly ordered manner and is typically achieved maintaining the protein's native structure. The structures are therefore not representative of those of aggregates that form as the result of protein degradation. Crystal structures are, however, ideal to study the structures of excipient molecules bound to proteins complexes (Table 1.1).

Cryo electron microscopy single particle analysis (Cryo-EM) has evolved more recently as a structure elucidation method and has been also employed to characterize protein complexes, achieving resolutions similar to those obtained by X-ray crystallography. The extreme conditions to which proteins are exposed, such as vacuum and low temperatures and the required sample preparation techniques, such as staining make it unsuitable for a direct study of formulation effects in situ[61–63]. It has however been employed to determine the structure of protein aggregates, which makes it an interesting method to study also aggregates of therapeutic proteins[64,65]. Cryo-EM was also used to resolve the structure of complexes between proteins and common excipient molecules (Table 1.1)[66].

Solution nuclear magnetic resonance (NMR) spectroscopy is another well-established method for structure elucidation. It offers a broad range of

applications designed to answer all sorts of questions related to protein structure. One of the main applications of NMR consists in the study of protein-ligand interactions, for example to identify binding sites[67]. NMR was also used to resolve the structure of protein-arginine complexes (Table 1.1).

Table 1.1: Occurrences of common excipient molecules as ligands in the PDB, determined by NMR, X-ray crystallography or Cryo electron microscopy (14.07.2019).

Excipient molecule	X-Ray	NMR	Cryo-EM
Sucrose	251	0	0
Trehalose	61	0	0
L-arginine	193	6	1
Sorbitol	19	0	0
Glycine	252	0	2

While classic NMR experiments are limited to proteins with a molecular weight of up to approximately 35kDa[68], there are numerous approaches to also study larger proteins indirectly. ^{19}F Dark-state exchange saturation experiments allow for example the characterization of antibody clusters in highly concentrated formulations. While the method can elucidate cluster populations, concentrations and sizes, it cannot identify the oligomerization interfaces[69]. For proteins below the critical size limit, oligomerization interfaces can be

identified indirectly for example by paramagnetic relaxation enhancement[70]. Another method that provides information on the oligomerization interface measures hydrogen-deuterium exchange rates[71].

Instead of measuring Hydrogen-deuterium exchange by NMR, it is also possible to do so by mass spectrometry (HDX-MS)[72]. Advantages are the higher sensitivity of HDX-MS and a higher limit regarding the size of the protein.

Based on the paradigm that specific regions or residues of a protein are involved in either native self-association or aggregation, models have been developed to predict protein-protein association and aggregation prone regions.

Protein-protein docking algorithms were developed to correctly predict protein-protein complexes from the structures of the individual proteins[73–75]. Methods are typically based on shape and chemical complementarity and are intended for strong, biochemical protein-protein interactions rather than weak self-association and aggregation. They also imply a native structure of the binding partners. Modern protein-protein docking methods account only for a limited degree of protein backbone flexibility and protonation effects[76]. Partial unfolding is not accounted for in protein-protein docking.

Molecular mechanics simulations present an alternative route to understanding processes of protein aggregation. Atomistic simulations of protein aggregation have been reported for small peptides[77]. Due to their large computational cost, atomistic simulations are not expected to play a role in the simulation of aggregation processes of larger proteins in the foreseeable future.

Instead, the computational burden of simulations can be reduced at the cost of accuracy and structural resolution by employing coarse grained models. Aggregation processes of amyloidogenic peptides mediated by surfaces were for example simulated by representing the peptides as patchy spherocylinders[78].

Instead of simulating protein-protein interactions to identify oligomerization interfaces, heuristic approaches have been developed that relate intrinsic (e.g. hydrophobicity) and extrinsic (e.g. their neighboring residues) residual properties to their aggregation propensity. The computational burden of these methods is minimal and many have been made available through webservers[79,80].

1.3. Parallels to small molecule drug discovery

While the process of discovering new excipients is barely rationalized or described, this is not the case for the discovery of small molecular drugs. Despite them serving a different purpose, multiple concepts can readily be transferred to the discovery of new small molecule excipients. In the following, the hit to lead process in drug discovery will be outlined shortly before parallels to excipient discovery are discussed.

The most common approach to small molecule drug discovery consists in identifying a substance that binds to a target identified as relevant, in order to either activate or inhibit it. Typically, the process starts with the identification of hits, meaning substances that show the desired activity in a specific assay. Strategies to discover hits are for example high throughput screening (HTS) of large compound libraries, where large compound libraries are tested experimentally for example for binding, a shift in protein stability or

biochemical activity. HTS assays are often fluorescence based as it can be detected at a high sensitivity and therefore allows for sample miniaturization[81]. Examples include microscale thermophoresis (MST), differential scanning fluorimetry (DSF), time-resolved fluorescence energy transfer (TR-FRET) or scintillation proximity assays (SPA)[82–84]. An alternative to experimentally screening compound libraries comes in the form of virtual screens (VS). In a VS, compounds in large databases are filtered by queries that span from physicochemical attributes over matching a defined pharmacophore to automated docking and scoring[85,86]. The detailed strategy selected in a virtual screen depends on the available knowledge regarding target and ligand structures and the desired hit profile.

After hits were identified, an expanded set of analogs to the hits is generated to obtain a series of related substances that differ by only a few structural features, a so-called lead series[87]. Lead series are evaluated by their properties regarding target affinity but also absorption, distribution, metabolism and excretion (ADME) in humans.

One of the central paradigms in lead series generation is the development of a structure activity relationship (SAR). While there are numerous approaches to SAR, they all share the concept of correlating a small molecule descriptor to the molecule's activity. Hansch and Dunn found for example a linear relationship between drug lipophilicity and biological activity[88]. As mentioned previously, in the so called SAR by kinetics method, authors identified physicochemical descriptors that correlate with a substance's propensity to inhibit aggregation of proteins related to Alzheimer's disease[51].

Structure-based drug discovery (SBDD) presents an alternative approach to lead series generation and affinity optimization that is based on knowledge of the target-hit complex structure[89]. Favorable and unfavorable interactions between the hit molecule and the target can be identified and the hit molecule's structure can be adapted to optimize the interaction profile accordingly.

While hit identification is centered solely on the interaction with the target protein, affinity optimization is not the only property considered in the generation and evaluation of lead series. Depending on the route of administration, small molecule drugs have to be optimized regarding absorption into the bloodstream. This implies crossing cell membranes, which can occur by diffusion or carrier transport. Numerous models exist to predict a substance's absorption[90]. Since absorption depends on physicochemical properties such as solubility and permeability, lead structures can be evaluated and optimized accordingly. After absorption, small molecules are distributed to extracellular fluid and tissues, with transfer mechanisms similar to those occurring during absorption. Volume of distribution and fraction unbound, two properties describing drug distribution, can be predicted from molecular descriptors[91–94]. Assessing a compounds metabolization profile is important before advancing it into clinical trials. Knowledge regarding its susceptibility to metabolization and resulting metabolites is crucial. A structure-metabolism relationship has been reported for the metabolic enzyme aldehyde oxidase[95]. Next to metabolization, excretion is a second major pathway of drug elimination. Also drug clearance can be predicted using physicochemical descriptors and structural features[96].

Drug toxicity typically describes adverse effects that occur at concentrations above the therapeutic dose. In drug discovery, it is desirable to increase the gap between therapeutic and toxic dose[97]. As conventional, non-clinical toxicity studies are slow and require animal experiments, several approaches to predict toxicity from the small molecule structure have been developed[98–101]. When evaluating small molecule toxicity, it is important to consider the parent molecule and its metabolites[102]. Given the multitude of parameters to be considered during lead development, QSAR models have been developed to optimize structures in terms of multiple of the aforementioned objectives, which are often in conflict with one another[103].

The most promising series of compounds is optimized and investigated in assays of higher complexity, such as animal studies, before a clinical candidate is selected[104]. It is worthwhile mentioning that organic synthesis is the key enabling technology for drug discovery. Only the continued evolution of synthesis methods allows the generation of the compounds required to build QSARs and modify hits in SBDD. Organic synthesis in drug discovery is beyond the scope of this thesis, but excellent reviews are available[105]. The field is currently moving towards increased automation of compound synthesis, testing and optimization[106].

When considering the discovery of new small molecule excipients, applying the same screen to hit to lead approach seems plausible. High throughput stability indicating methods such as nanoDSF are available. When the excipients desired mechanism is by stoichiometric binding, a structure-based approach as in SBDD is possible (Chapters 1-3). QSARs are applicable for both, stoichiometric binding and preferential exclusion stabilizers (Chapter 4). Just as drugs, also excipients

are absorbed, distributed, metabolized and excreted and they will be toxic above a certain concentration. It is thus apparent that excipients have to be optimized in terms of these parameters, but with a different target profile than that of a drug. An excipient's desired ADME profile is quite different to that of a drug. Ideally, an excipient spends as little time as possible in a patient's body. Rapid metabolization could therefore be beneficial for an excipient as long as the metabolites are harmless. As most protein therapeutics are administered intravenously, excipient absorption does not have to be considered here, but transport across membranes remains relevant to the excipient's distribution. When designing a novel excipient, its stabilizing effect is consequently a necessity but not a sufficiency.

2. Aim and outline of the thesis

The aim of this thesis is to describe strategies that can be used to successfully identify small molecules that reduce the formation of aggregates of therapeutic proteins and could potentially be used for their formulation. Consequently, this will also demonstrate the potential in the chemical space to stabilize proteins compared to standard excipients.

Besides simply expanding the parameter space available to formulation design, improved excipients are also demanded by the pharmaceutical industry due to shortcomings of those currently employed. Polysorbate has for example been linked to particle formation and protein oxidation.

The current strategy in the development of therapeutic proteins consists of sequence optimization followed by formulation optimization. Sequence optimization typically consists of replacing aggregation prone residues of the protein. However, when considering antibody-target complexes, we observe that a large majority of aggregation prone residues are also present in the mAb's target binding paratope, indicating their presence is mandatory to achieve a sufficiently high target affinity (Figure 2.1). This implies that there is a natural

limit to the stability achievable by sequence and buffer optimization. New excipients can push stability boundaries beyond these without affecting the drug's target affinity.

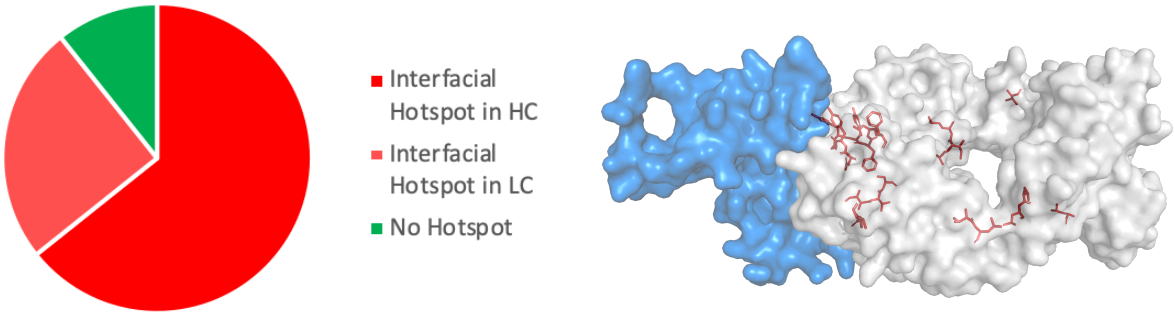


Figure 2.1: Aggregation prone residues present in drug-target interface for 30 mAb structures.

Recently, patent protection of many of the first-generation therapeutic antibodies has expired. New excipients present an opportunity to expand the patent lifetime of a drug, if the new formulation is superior to the former. This could for example be the case if the new excipient achieves a reduction in immunogenicity of the drug product, due to a lower particle load compared to the standard formulation.

The thesis starts with a target-based approach to excipient discovery that consists of three parts. First, we describe our strategy that led to the identification of a novel stabilizing dipeptide. It parts from a virtual screen that is used to identify hit molecules to be tested experimentally. Subsequent binding and forced degradation studies characterize the small molecule's effect on protein stability. A similar virtual-screen approach was also applied to the monoclonal antibody Trastuzumab. The work resulted in the discovery of the

stabilizing effects of the substance N,N,N',N'-tetrakis-(2-hydroxyethyl) adipinic acid amide. A patent application describing the use of this substance as excipient was submitted. The corresponding data can be found in the appendix of this thesis.

Next, we describe a novel method to identify aggregation prone regions by using solution paramagnetic enhancement NMR. The results are then compared to in-silico prediction tools. The target-based approach is wrapped up by a study of the mechanism of interaction between the protein and the dipeptide by molecular dynamics simulations analyzed by the Markov chain formalism.

To overcome the limitations encountered in the structure-based approach, the final chapter describes the development of a ligand-based model by combining chemoinformatic methods and machine learning with a recently developed stability indicating method.

3. Structure-based Discovery of a new Protein-Aggregation Breaking Excipient

A version of this chapter has been published in the European Journal of Pharmaceutics and Biopharmaceutics:

Tosstorff, A., Svilenov, H., Peters, G.H.J., Harris, P. & Winter, G. Structure-based Discovery of a new Protein-Aggregation Breaking Excipient, *Eur. J. Pharm. Biopharm.* **144**, 207-216 (2019).

This work was conducted in collaboration with the Department of Chemistry of the Technical University of Denmark. The manuscript was written by Andreas Tosstorff. Hristo Svilenov provided scientific advice and reviewed the manuscript. Pernille Harris reviewed the manuscript. Surface pressure measurements were performed by Luis Sánchez. Other experiments, simulations and data analysis were performed by Andreas Tosstorff under the supervision of Günther H.J. Peters and Gerhard Winter.

3.1. Abstract

Reducing the aggregation of proteins is of utmost interest to the pharmaceutical industry. Aggregated proteins are often less active than their non-aggregated form and upon administration, they can cause severe immune reactions in the patient. Despite these risks, there is an increasing demand for high concentration formulations and products that do not require refrigerated storage conditions, both of which favor the formation of aggregates. For a given protein, solution pH, ionic strength and concentration of a very limited number of excipients are the only parameters that can be varied to obtain a stable formulation. In this work, we present a structure-based approach to discover new molecules that successfully reduce the aggregation of proteins and apply the approach to the model protein Interferon-alpha-2a.

Keywords

Interferon-alpha-2a, Virtual Screen, Excipient, Protein Aggregation, Protein Formulation

3.2. Introduction

Protein aggregation

When assessing the development and production of therapeutic proteins, their aggregation is a major concern to regulatory agencies across the world. Not only can aggregation cause a decrease in biological activity, but the resulting aggregates also raise serious safety concerns as they can induce immunogenic side reactions upon parenteral injection[107]. Pharmaceutical companies

therefore strive to inhibit the formation of protein aggregates early on during drug development[1].

The process of protein aggregation is very complex, with thermodynamics and kinetics depending on formulation conditions, stress, protein sequence and structure[108]. Proteins aggregate through the interaction of exposed hydrophobic regions, which is driven by the classical hydrophobic effect. There is a variety of models suggesting different microscopic steps involved in the formation of aggregates. Typically these involve the formation of an aggregation prone species and nucleation[38]. Depending on the mechanism of aggregation, the resulting aggregates can consist of native or (partially) unfolded protein molecules. As shown by mutation experiments, hydrophobic patches on the proteins surface, so called aggregation hot-spots, are crucial to the formation of protein-protein interfaces, a key step in the formation of aggregates[109]. Various computational tools to identify aggregation hot-spots from a protein's primary sequence are available[79,110,111]. Aggrescan3D (A3D) additionally takes into account the tertiary structural information of the protein, mitigating the risk of false positive results from hydrophobic residues buried within the protein fold[80].

Excipients

Excipients reduce protein aggregation by various mechanisms of action. Computational studies suggest that arginine binds non-covalently to certain sites on a protein[112]. In combination with glutamate, the stabilizing effect of arginine could be further enhanced. The improved of stabilizing effect was attributed to the formation of arginine-glutamate clusters[113]. The small molecule drug dexamethasone phosphate (DMP) was discovered to reduce the formation

of bevacizumab aggregates when administered in a co-formulation. Docking studies of DMP to a homology model of bevacizumab found a lysine residue as binding site. The lysine residue was shown to form crystal contacts and DMP binding was concluded to sterically hinder the formation of protein-protein interfaces and thus inhibit aggregation[114–116]. In another study used hydrogen-deuterium exchange spectroscopy to identify a patch of residues in the CDR region to be involved in the formation of bevacizumab aggregates at elevated temperatures[52].

Virtual Screen

Here, we present an approach that aims at identifying new compounds that bind to a predicted aggregation hotspot of Inteferon-alpha-2A (IFN), thus inhibiting the formation of protein-protein interfaces and subsequently aggregation.

Due to the large, flat interfaces that form during protein-protein interactions, these have long been considered difficult targets for small molecules. More recently, many successful examples have been presented[117]. In order to identify small molecules that bind to a defined protein site, a common approach is running a virtual screen, where databases of millions of compounds are tested for affinity towards the specified binding site by docking algorithms[118]. The database selection is the first step critical to the success of a docking campaign. Database size and compound diversity and availability need to be considered. The ZINC15 database is one of the largest publicly accessible databases, including more than 700 million compounds, that can be filtered according to their commercial availability, reactivity or hydrophobicity[119]. Glide, Gold and Autodock Vina are some programs to perform high throughput pose prediction and scoring[120–122]. While current docking algorithms account for ligand

flexibility, the receptor is considered to be rigid, an assumption that can drastically reduce enrichment of active compounds in the highest scoring hits[123]. Docking algorithms do not account for the presence of water explicitly and may be inaccurate in predicting protonation states of the binding site, which can lead to poor predictions of poses and energy scores. Due to docking's many simplifications and limitations, its results should be considered as a starting point to suggest interesting compounds, rather than a method to elucidate detailed features of protein-ligand interaction, such as binding kinetics and free energies.

Free energy of binding

Atomistic molecular dynamics simulations present a more accurate way to calculate free energies of binding than docking. There are various approaches to calculate free energies of binding between two molecules through atomistic simulations. Unbiased simulations can give detailed information on the binding mechanism, kinetics and secondary binding sites[124]. However, they demand large amounts of computational resources. Biased simulations reduce the computational cost by introducing potentials that facilitate the sampling of unfavorable regions in the system's phase space. In the simplest case, a biasing potential can be a harmonic oscillator, restraining the distance between two atoms. Two commonly applied methods making use of biasing potentials are meta-dynamics and umbrella sampling[125,126]. Introducing biasing potentials to a system has been observed to cause dissipation of energy in umbrella sampling simulations[127]. This effect has been overcome more recently by accounting for the energy required to attach and release these potentials[128]. The resulting attach-pull-release umbrella sampling (APR-US) method has a solid theoretical

foundation and has been able to accurately predict free energies of binding in guest-host systems[129,130].

Instead of determining binding energies experimentally, they can also be measured from titration experiments using methods such as isothermal calorimetry, surface plasmon resonance, nuclear magnetic resonance or microscale thermophoresis. In all these methods, a signal is measured for a series of ligand concentrations while maintaining the concentration of the binding partner constant. The dissociation constant is then determined by fitting an appropriate model to the concentration dependent function. In microscale thermophoresis, the effect of ligand concentration on protein thermophoresis is measured. Thermophoresis refers to the directed motion of a particle in a temperature gradient and depends on the particle's mass. Upon binding of a ligand to a protein, the thermophoresis will therefore change due to the difference in molecular mass between free and bound protein. In microscale thermophoresis, the protein motion is measured through fluorescence and a temperature gradient is established by an infrared laser[131].

Additional aspects of virtual screens

Identifying a small molecule that binds to its target is crucial to achieve the desired effect in the protein. But binding is not the only aspect to be taken into consideration when selecting compounds through a virtual screen. Other physico-chemical properties such as reactivity, toxicity and solubility are equally important to obtain successful candidate compounds. A low reactivity will ensure that the compound remains stable and will not alter other substances present in the formulation. Only if the substance is of low toxicity it can be considered for the use in patients. A minimum solubility is required to

ensure that the small molecule can be employed without the use of organic co-solvents. Solubility can be easily accounted for in a virtual screen since multiple models for its prediction have been developed.

A compounds solubility is typically indicated by its $log_{10}S$ value, where S is the compounds concentration in the aqueous phase in equilibrium with the most stable form of the crystalline compound[132]. Solubility is most commonly predicted by quantitative structure-properties relationship (QSPR) methods, such as group contributions[133,134], neural networks[135] or multiple linear regression analysis[136]. A public challenge to predict the solubility of a set of 32 compounds from a training set of 100 molecules revealed the current state of prediction quality: the best performing predictions on a dataset including outliers had a coefficient of determination (R^2) of 0.6 and close to 20% of the $log_{10}S$ values were calculated correctly[137–139]. However, solubility predicting methods typically do not consider solution pH but are only trained against physiological conditions. In formulation science, where pH and ionic strength can differ strongly from this condition, pK_As should therefore also be considered when assessing solubility. A carboxylic acid will for example show different solubilities depending on its protonation state.

A property closely linked to the water solubility is the octanol-water partition coefficient as a measure of hydrophobicity for small molecules[140]. The ZINC15 database can conveniently be filtered by predicted $log_{10}P$ values[86,141].

Experimental assessment of protein stability

For a compound that passes all filters of the virtual screen, we want to test its effect on protein aggregation experimentally.

Aggregation processes are typically very slow. To predict the stability of a formulation in a reasonable time frame, one can test a formulation for surrogate endpoints such as the interaction parameter k_d or the apparent molecular weight as a measure of colloidal stability through dynamic and static light scattering respectively[142,143]. The inflection point (IP) of an unfolding experiment serves a measure of conformational stability and is also referred to as the protein's melting temperature. The onset of aggregation temperature T_{onset} describes the temperature at which aggregates start to form when exposing a protein to a temperature ramp. Alternatively, stress-studies can be performed, where the formulation is exposed to an aggregation trigger such as freezing/thawing, heat, shaking, shear or light. Chemical changes, which are incurred to the protein by light and thermal stress, are not the scope of this work[144,145]. We apply heat, freeze-thaw and shaking stress to evaluate the effect of the candidate excipients. To benchmark our compounds, we compare them against L-arginine and D(+)-trehalose, two substances commonly employed as excipients in protein formulation.

3.3. Methods

Virtual Screen

A homology model of IFN was generated based on the PDB entry 4Z5R using Modeller[146]. A potential aggregation hotspot was identified by submitting the homology model to the Aggrescan3D server[80].

The protein structure of IFN was prepared for docking using Maestro's (Schrödinger, Inc., New York, New York, USA) protein preparation wizard with pH set to 7.0. Maestro was used to generate a docking grid using the residues

that are located in the identified aggregation hotspot as grid center. The ZINC15 database tranches were selected to include only compounds with a $log_{10}P \leq$ -1, "in-stock" availability and standard reactivity. The compounds were then prepared for docking using LigPrep as implemented in Maestro. Qikprop was used to predict the compounds physicochemical properties and only compounds with a $log_{10}S$ value ≥ -1 were retained. All compounds were then docked with Glide HT. The best scoring 10 % were then redocked and scored with GlideSP. The best scoring 10 % were redocked and rescored using GlideXP and up to 3 poses per compound were generated. These poses were rescored using the Prime MM-GBSA model. We then looked manually for substances available for purchase below 200€/g.

Sample Preparation

An aqueous bulk solution of Interferon-alpha-2a (Roche, Penzberg) was dialyzed (Spectra-Por) into 50 mM sodium phosphate (di-Sodium hydrogen phosphate dihydrate: VWR Chemicals, Leuven, Sodium dihydrogen phosphate dihydrate: Grüssing GmbH, Filsum) buffer at pH 7.0. The solution was filtered using a 0.22 µm cellulose acetate filter (VWR Chemicals, Leuven), which were previously reported to be low protein binding[147]. A protein concentration of 1.4 mg/ml was obtained as determined by measuring the UV absorption at 280 nm using a NanoDrop (Thermo Fisher Scientific, Waltham, MA, USA).

Excipient stock solutions were prepared by dissolving the excipient in 50 mM sodium phosphate buffer (di-Sodium hydrogen phosphate dihydrate: VWR Chemicals, Leuven, Sodium dihydrogen phosphate dihydrate: Grüssing GmbH, Filsum) at pH 7.0 and adjusting the pH to 7.0 as required either with hydrochloric acid or concentrated sodium hydroxide. Buffer was then added to

obtain a final excipient concentration of 500 mM. The excipient stock solution was then filtered using a 0.22 µm filter (VWR Chemicals, Leuven).

Binding study

Binding affinities of the excipient candidates were determined by microscale thermophoresis (Monolith, NanoTemper, Munich, Germany). Interferon-alpha-2a was labelled fluorescently (Monolith Protein Labeling Kit RED-NHS) and excipient candidates were titrated using 50 mM phosphate buffer at pH 7.0 (di-Sodium hydrogen phosphate dihydrate: VWR Chemicals, Leuven, Sodium dihydrogen phosphate dihydrate: Grüssing GmbH, Filsum) with a polysorbate 20 (Sigma Aldrich) concentration of 0.05 %[83]. A dilution series of 16 samples of 20 µl each was prepared in triplicates from stock solution containing 500 mM small molecule by mixing it with the assay buffer through pipetting in reaction tubes. 20 µl of labelled protein was added to each sample, yielding a final protein concentration of 20 nM. Excitation-power was set to 20% and MST-power was set to “high”. Binding affinities, standard deviations and confidence intervals were calculated using MO.Affinity Analysis v2.2.7 (NanoTemper, Munich, Germany).

Molecular dynamics simulations

The best scoring pose of the MM-GBSA rescoring served as input structure to calculate free energies of binding by the APR-US approach[128–130]. The PDB structure generated by the virtual screen, containing the ligand docked to the protein, was reoriented using the z-align script from the APR suite. Restraints were gradually attached in 13 windows and the distance between the compound and its binding site was gradually increased in 46 windows. For the first window

of the attachment phase where the APR restraints are set to 0, an additional distant restraint was implemented to define the binding site and avoid the ligand leaving. The systems for each window were constructed using tleap, adding 20500 water molecules to each system, using the APR procedure. The program pmemd.CUDA as implemented in Amber16 was used along with the ff14SB, GAFF2 and TIP3P force-fields[148,149]. The ligand was parametrized using GAFF2 for bonded and non-bonded parameters. Atomic partial charges were calculated with Gaussian 16 (Gaussian Inc., Wallingford, CT, U.S.A.) and fitted with the RESP procedure in antechamber. Hydrogen mass repartitioning and the SHAKE algorithm were used to allow timesteps of 4 fs[150,151]. Pressure was regulated using a Monte Carlo barostat and a Langevin thermostat was used to keep the temperature at 298.15 K. Modifications to the APR script were implemented to allow parallel runs of the respective windows on the GPU cluster and facilitate system preparation. The simulation time in each window was 112.5 ns resulting in approximately 6.6 μs total simulation time. Calculation of the free energy of binding was performed by using the thermodynamic integration scheme as implemented in the APR script.

Toxicity Prediction

The toxicity for the candidate compound A was predicted using OpenVirtualToxLab[152].

Forced degradation studies

Each replicate sample was filled in a separate 2R vial (Fiolax, klar HGA 1/ISO 720). The vials were capped and crimped pneumatically. Excipients and buffer

were spiked into the IFN solution to obtain a final formulation of 1 mg/ml of protein, 50 mM excipient, 50 mM sodium phosphate at pH 7.0.

Samples were prepared freshly before any forced degradation experiment, without any substantial incubation time. To evaluate the stabilizing impact of the excipient candidates, samples were frozen and thawed three times in a Christ 2D-6 freeze dryer. A temperature ramp of 1 K/min and a hold time of 2 h were used. The protein was also exposed to shaking stress during 60 h using a horizontal shaker (IKA HS 260 basic, 300 rpm). Sub-visible particles were detected by flow imaging (FlowCam, Fluid Imaging Technologies, Inc., Scarborough, ME, USA). Soluble aggregates were detected by size-exclusion chromatography on a Dionex Summit HPLC system at 214 nm using a Superose 12 10/300 GL as stationary phase (GE Healthcare Life Sciences, Chalfont St Giles, UK) and 50 mM sodium phosphate (di-Sodium hydrogen phosphate dihydrate: VWR Chemicals, Leuven, Sodium dihydrogen phosphate dihydrate: Grüssing GmbH, Filsum), 200 mM NaCl, pH 7.0 as mobile phase. High molecular weight species were quantified by measuring the area under the corresponding signal of the chromatogram.

Heat induced degradation was measured with by nanoDSF and backscattering (Prometheus NT.48, NanoTemper, Munich, Germany) at a heating rate of 1 °C/min from 25 to 95 °C in standard capillaries (NanoTemper, Munich, Germany). T_{onset} and IP were extracted from backscattering and ratio of fluorescence at 350 nm and 330 nm curves respectively using the software PR.ThermControl (NanoTemper, Munich, Germany).

Apparent M_w

Apparent M_w was measured by static light scattering (DynaPro III, Wyatt Technology Europe, Dernbach, Germany) in a 1536 well plate (Aurora Microplates, Whitefish, MT, USA) with 8 µl of sample volume and 3 µl of silicon oil (Alfa Aesar, ThermoFisher GmbH, Kandel, Germany). The well plate was calibrated with a dilution series of dextran (Sigma-Aldrich Chemie GmbH, Taufkirchen, Germany). Due to the sensitivity of light scattering to larger particles, stock solutions were additionally filtered using 0.02 µm filters (Whatman, GE Healthcare UK, Buckinghamshire, UK)

Surface Pressure

Surface pressures of the protein free buffers containing the different excipients were measured as duplicates in a multiwell plate with a metal ally dyne probe (Microtrough XS, Kibron Inc., Finland).

3.4. Results

Virtual Screen

The purpose of the virtual screen was to identify small organic molecules from the ZINC database that would potentially bind to the IFN. We identified a potential aggregation hotspot at residues L26 and F27 of IFN using Aggrescan3D (Figure 3.1) [80]. The hotspot's score remained unchanged among all 25 available structures, showing little effect of protein dynamics on the calculated propensity. The highest-ranking residue patch was defined as binding site for a subsequent virtual screen. Candidate compounds would ideally bind in proximity to the hotspot, blocking it from driving the formation of a protein-protein interface.

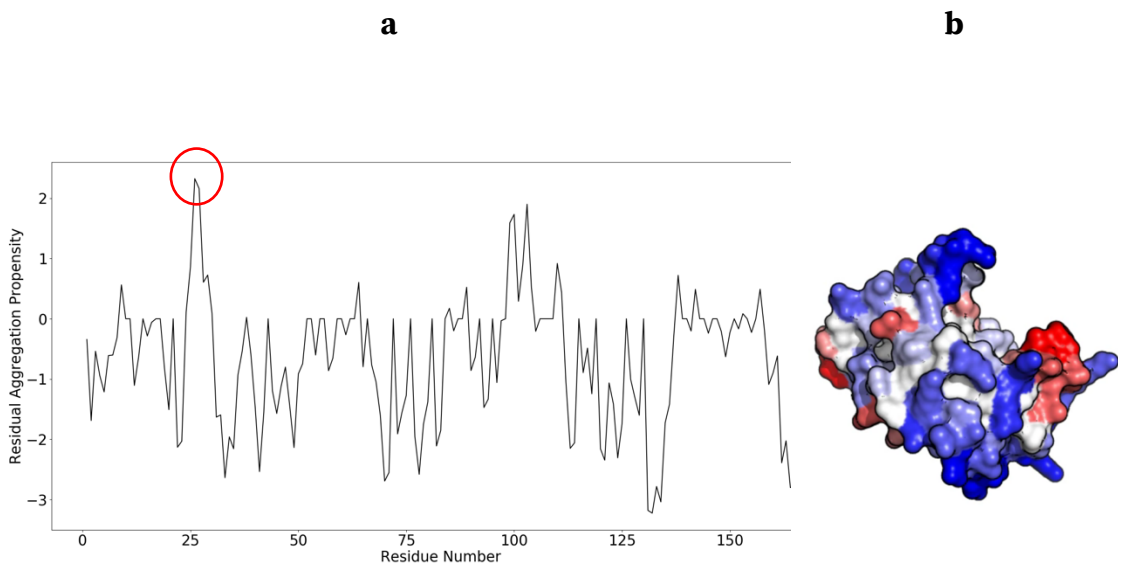


Figure 3.1: ***a****: Residual propensity for aggregation determined by Aggrescan3D. Highest scoring hotspot highlighted with a red circle.* ***b****: Visualization of residual aggregation propensity (Blue: low propensity, Red: high propensity).*

Applying a solubility filter orthogonal to the ZINC database's internal $log_{10}P$ filter showed that only 33,101 of the 52,980 had a sufficiently high solubility. These compounds were then docked with Maestro's virtual screen workflow using GlideSP and GlideXP. The best scoring compounds were then rescored using the MM-GBSA solvent model. After docking the compounds at increasing levels of precision and conformational sampling, 167 compounds were predicted to bind in the hotspot's proximity. These were inspected visually and five were purchased based on their price and availability (Figure 3.2).

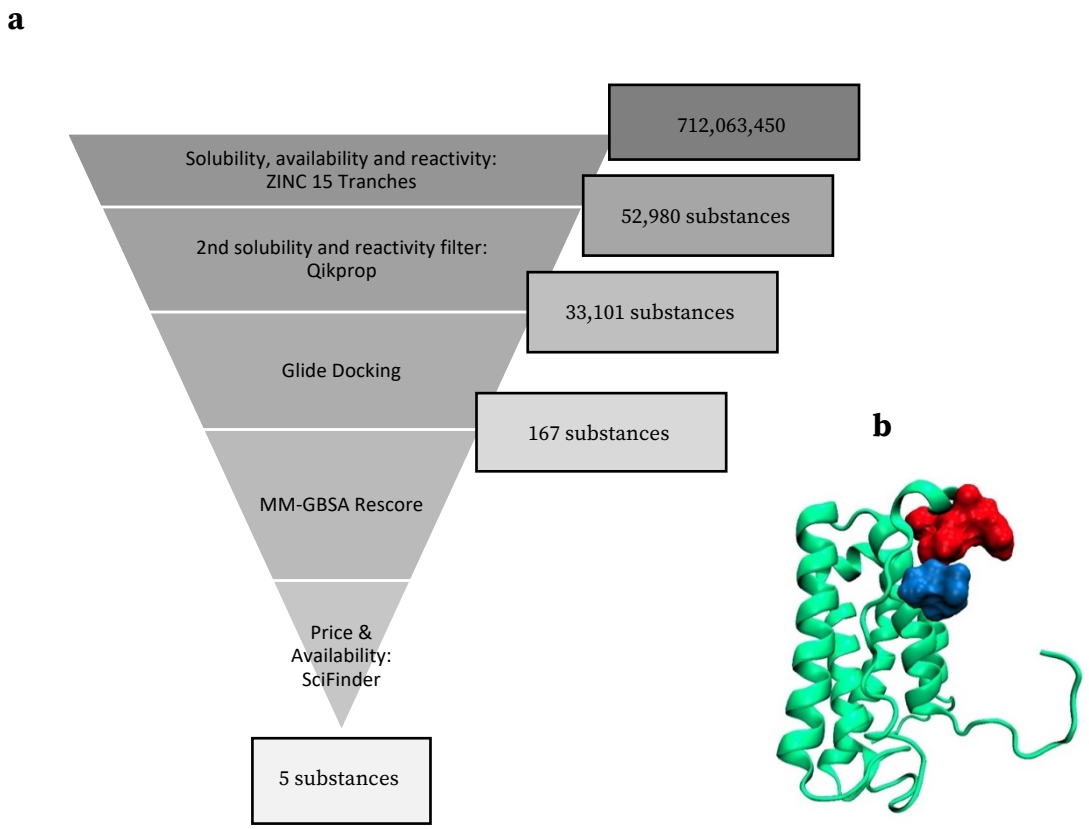
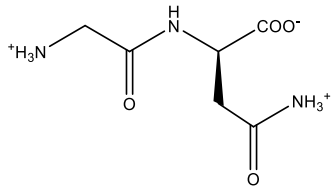
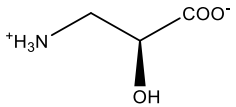
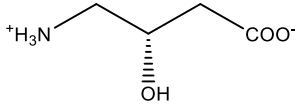


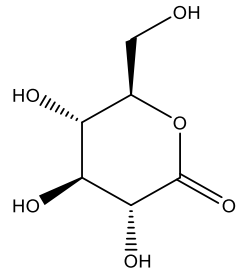
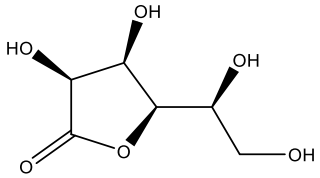
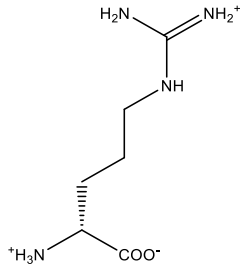
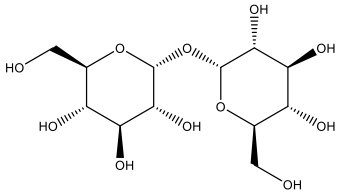
Figure 3.2: Virtual Screen. Left: Scheme of the virtual screen, designed to identify substances that possess high solubility, low reactivity and high affinity towards the defined binding site. Right: visualization of a ligand (blue) bound to IFN (green) in proximity to the aggregation hotspot predicted by Aggrescan3D (red).

Binding study

All five compounds were tested for binding by microscale thermophoresis. Due to the rigorous filters applied in the virtual screen, all compounds dissolved readily in the experimental buffer. Out of the tested compounds, only compound **A** and L-arginine were detected to bind to the target (Table 3.1).

Table 3.1: List of purchased compounds

Compound	Name	Structure	$log_{10}S$	ΔG MM-GBSA (kcal/mol)	Dissociation constant K_d (MST)	Source	Purity
A	Glycyl-D-asparagine		1.8	-18.9	108 µM ± 24 µM	abcr	98 %
B	L-isoserine		0.5	-18.9	No binding detected	abcr	98 %
C	(S)-4-Amino-3-hydroxy-butyric acid		0.4	-19.0	No binding detected	Sigma-Aldrich	97 %

D	D-(+)-glucono- 1,5-lactone		-0.9	-32.8	No binding detected	Sigma-Aldrich	>99 %
E	L-(+)-glutonic acid gammalactone		-0.7	-27.7	No binding detected	abcr	98 %
-	L-arginine (K47275343 621)		N/A	N/A	657 µM ± 211 µM	Merck KGaA	>98.5 %
-	D(+)-trehalose dihydrate		N/A	N/A	No binding detected	VWR	>98 %

In a control experiment, the fluorescent dye from the protein labelling kit (Monolith Protein Labeling Kit RED-NHS) was used as target and showed no dose response. For **A**, a dissociation constant of 108 µM ± 24 µM was determined, which corresponds to a free energy of binding of -5.44 ± 0.13 kcal/mol.

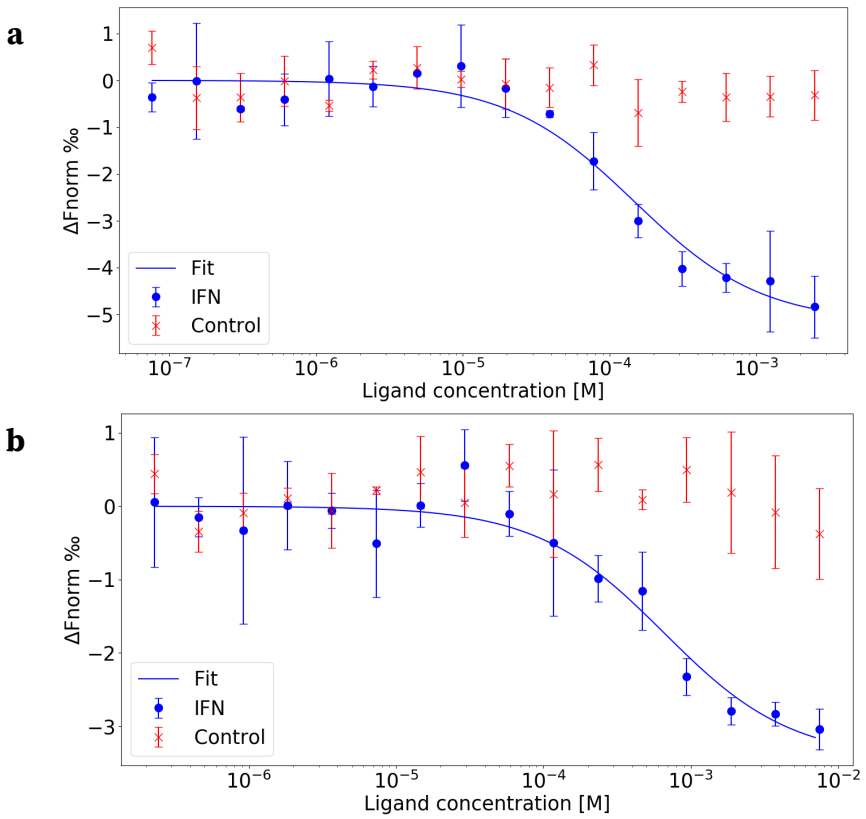


Figure 3.3: ***Experimental and calculated binding affinities.*** *a:* *Dose response curve of* ***A*** *targeting IFN (dots) and the control dye (crosses) as determined by MST: Kd=108 µM ± 24 µM. 50 mM Pi, pH 7.0, 0.05% Tween 20, N=3, IR intensity=high. Error bars represent the standard deviation of the measurement of three independent samples.* ***b:*** *Dose response curve of L-arginine targeting IFN (dots) and the control dye (crosses) as determined by MST: Kd=657 µM ± 211 µM. 50 mM Pi, pH 7.0, 0.05% Tween 20, N=3, IR intensity=high. Error bars represent the standard deviations of the measurement of three independent samples.*

The free energy of binding calculated by the APR-US method was found to be below the measured energy (Figure 3.4).

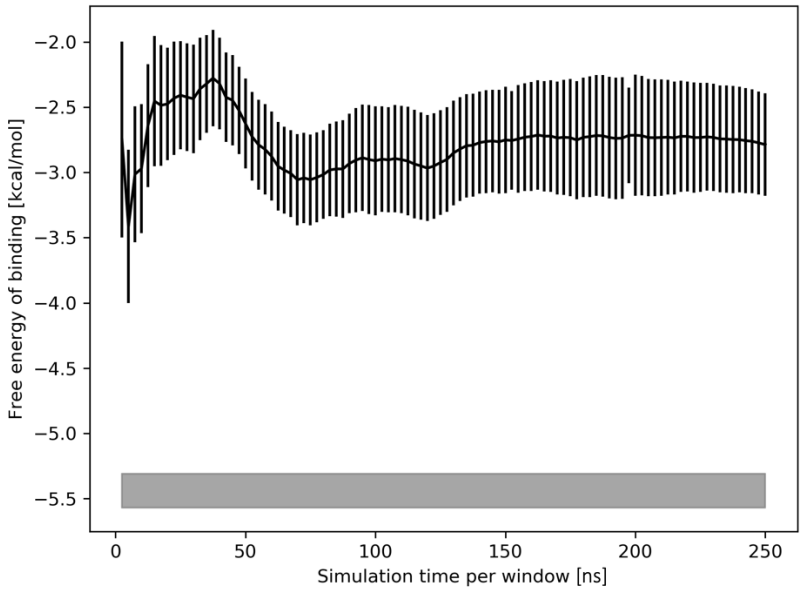


Figure 3.4: Black curve: Free energy of binding as calculated by the APR-US method. Error bars represent the standard error of the mean. Grey bar: Free energy of binding as determined by MST. The bar's thickness indicates the 68% confidence range.

Protein self-interaction

To determine colloidal stability, the apparent molecular weight (M_w) of IFN was measured in the absence and presence of compound **A** using static light scattering (SLS). As expected from the choice of pH and ionic strength, IFN forms aggregates in solution. While the aggregation is concentration dependent for low IFN concentrations, a plateau is reached at approximately 6 mg/ml. Even though the presence of compound **A** leads to slight reductions in M_w (Figure S-1) it does not quantitively break up aggregates.

Forced degradation studies

To evaluate the effect of the selected candidate compounds on protein stability, aggregation of IFN was induced by forced degradation experiments. Sub-visible particles and high molecular weight species were quantified by microflow imaging and SEC after three freeze-thaw cycles with the 5 formulations containing the excipient candidates. Additionally, a negative control was run containing only protein and buffer, but no other stabilizing agent. The only compound to significantly reduce both the formation of high molecular weight species and sub-visible particles was found to be compound **A**. While compounds **B** and **C** would slightly reduce soluble aggregate formation, they showed no benefit on sub-visible particle count compared to the excipient free control (Figure 3.6 a and b).

To further evaluate the effect of compound **A** on the stability of IFN, formulations containing different concentrations of compound **A** were exposed to horizontal shaking stress. The ligand's concentration range was chosen to be cover the mM and µM range according to the previously determined dissociation constant of 108 µM. The formation of sub-visible particles shows a strong dose response. At high ligand concentrations, where all protein molecules are bound to **A**, sub-visible particle formation is at a minimum and monomer area is at a maximum. With decreasing ligand concentration, the share of unbound protein increases and an increase in sub-visible particles and a decrease in monomer area is observed (Figure 3.6 a). When comparing the particle size distributions of the formulation with the highest and lowest content in compound **A**, no shift towards higher or lower particle sizes is apparent (Figure S 3.3).

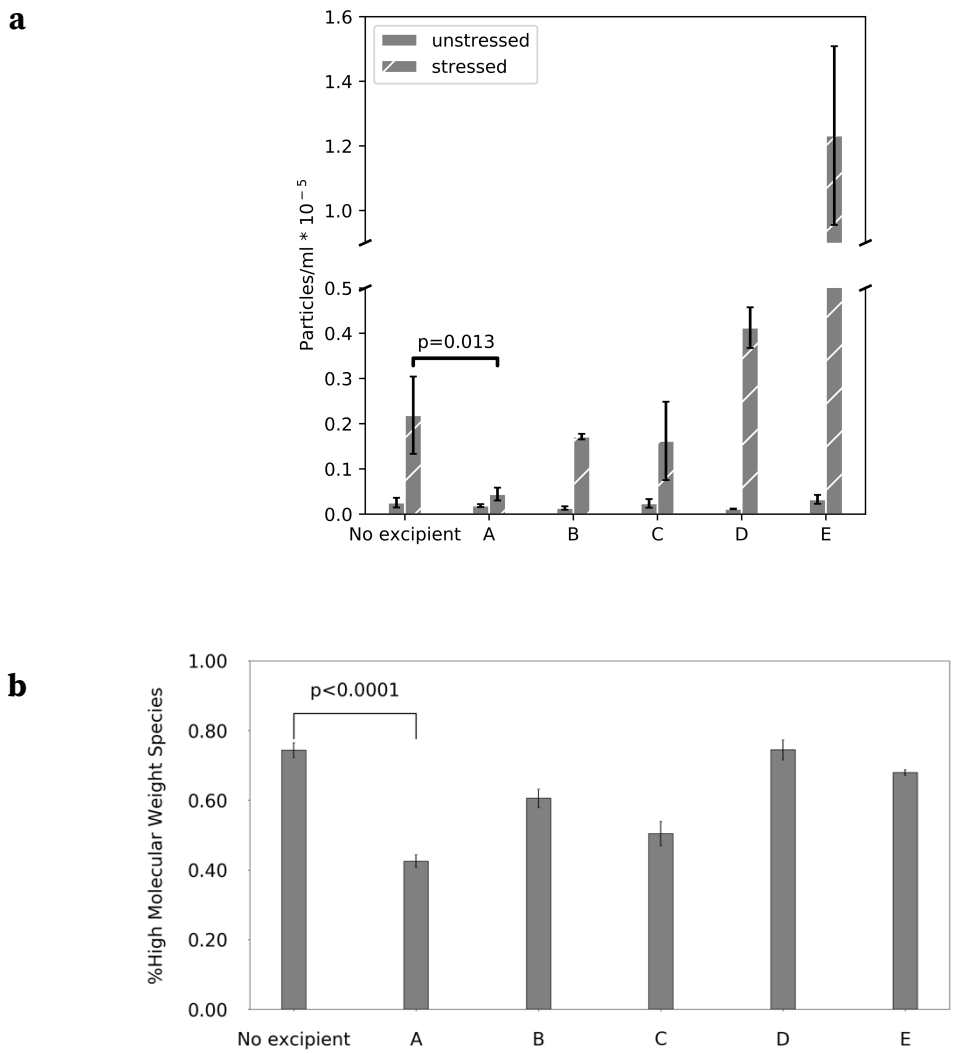


Figure 3.5: ***Forced degradation studies. a****: Count of particles ≥ 1 µm after three cycles of freezing and thawing of IFN formulations.* ***b****: Soluble high molecular weight species after three cycles of freezing and thawing of IFN formulations. A-E corresponds to the compounds from Table 3.1.*

As a benchmark test, compound **A** was compared to the standard excipients L-arginine and D(+)-trehalose at a concentration of 6.25 µM. All three compounds readily reduce the formation of sub-visible particles. However, compound **A** shows a lower particle count than the standard excipients D(+)-trehalose and L-arginine (Figure 3.6Figure 3.6 b).

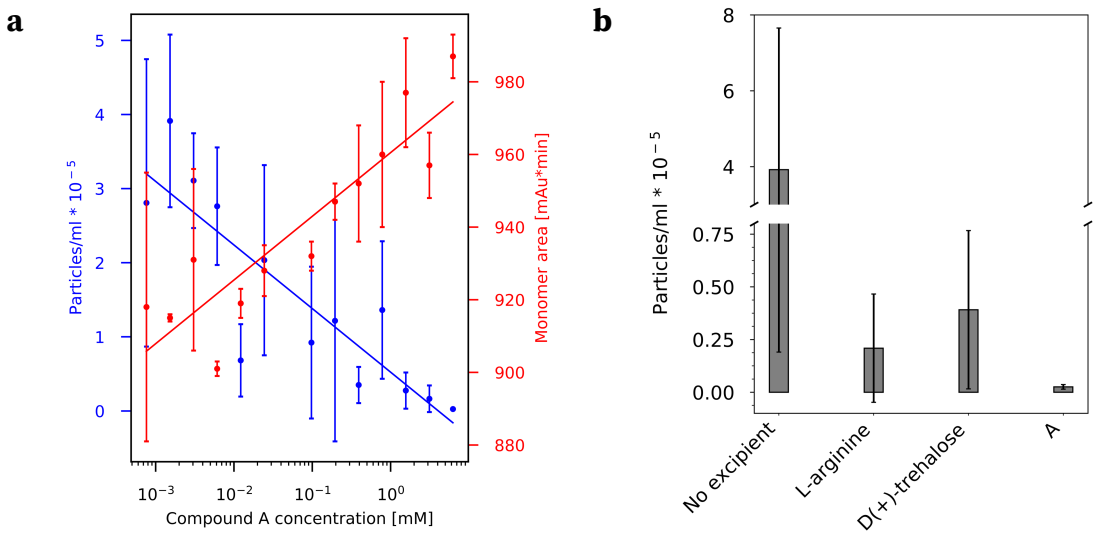


Figure 3.6: ***Forced degradation studies. a****: Dependence of sub-visible particle count on* ***A*** *concentration after horizontal shaking. The line is a guide for the eye generated by linear regression from the mean values.* ***b****: Sub-visible particle count for* ***A*** *and standard excipients at 6.25 mM after horizontal shaking.*

In order to rule out that the positive effect of compound **A** on the protein's stability is due to a non-specific effect, the surface activity (Table 3.2) of the compound was measured. While compound **A** leads to slightly higher surface pressures than the non-surfactant references, its surface activity is far below that of a typical surfactant polysorbate 20.

Table 3.2: Surface pressure data for different excipients. Excipient concentration was 50 mM, except for Tween 20, for which it was 0.005% v/v. All measurements were done twice. The errors given correspond to the standard deviations.

Excipient	Surface pressure [mN/m]
Buffer	1.7±0.2
NaCl	1.7±0
L-arginine	3.25±0.15
D(+)-trehalose	2.1±1.6
Glycerol	4.75±0.95
Polysorbate 20 [0.005%]	34.7±1
Compound A	9.0±0.5

Furthermore, the effect of compound **A**'s L-isomeric form, glycyl-L-asparagine, on particle formation was tested (Figure 3.7). Compound **A** drastically reduces sub-visible particle formation compared to all other tested molecules. Surprisingly even slightly lowering particle counts compared to the unstressed sample. Glycyl-L-asparagine does not have a beneficial effect on particle formation compared to the excipient free formulation.

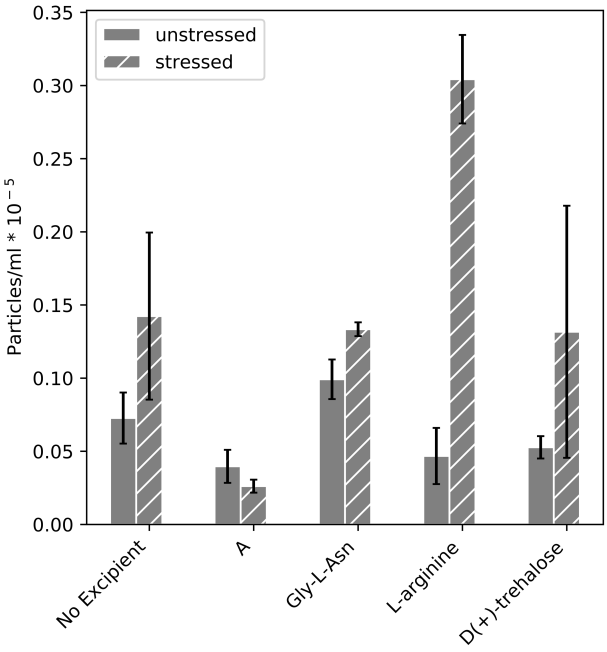


Figure 3.7: ***Forced degradation studies.*** *Sub-visible particle count after submitting a formulation of IFN to 60 h of horizontal agitation stress. Error bars represent the standard deviations of the measurements of three independent samples.*

In order to study the target specificity of compound **A**, its stabilizing effect during freezing and thawing was tested in combination with a monoclonal antibody (mAb) (Figure S-2). Here, all tested compounds reduced particle formation with compound **A** performing slightly worse than the benchmark excipients L-arginine and D(+)-trehalose.

While compound **A** showed a stabilizing effect on IFN when formulations were exposed to agitation or freezing/thawing, it had no effect on the protein's melting temperature and temperature of onset of aggregation as measured by nanoDSF, neither did any other of the examined compounds (Table S-1, Figure S 3.1, Figure S 3.2).

Toxicity Prediction

To estimate the toxicity of compound a, the VirtualToxLab tool was used. It predicts a very low toxicity of compound **A**. It was predicted not to bind to any of the toxicity related target proteins and its overall toxicity score was found to be 0.079, ranking for example below vitamin C which has a score of 0.253.

3.5. Discussion

The virtual screen was successful with a hit rate of 20% in identifying one out of five tested molecules that bind to IFN with µM affinities. Identifying substances with higher binding affinities could be achieved by allowing for more hydrophobic compounds in the screen or increasing the compound's size. Nevertheless, an increased hydrophobicity could have a negative effect on solubility, toxicity and clearance of the compound. Even though we were successful in identifying a compound that reduces particle formation, docking alone cannot be considered as proof of a structure-activity relationship. While MM-GBSA ranked affinities of compounds **C** to **E** higher than that of compound **A**, they were not detected to bind in MST measurements. This may be explained by the previously mentioned many simplifications made by the docking algorithms.

In order to obtain additional binding molecules, the same library was docked against an ensemble of IFN conformations, leading to the identification of one additional hit, which showed no increase in stability in any forced degradation study (data not shown). This finding indicates that not all protein-ligand complexes would result in a stabilization, but only specific interactions. When adding the tested compounds to formulations containing mAb-1 instead of IFN,

compound **A**, L-arginine and D(+)-trehalose would all reduce particle formation after freeze-thaw stress to the same extend. Given the structural diversity of the three compounds, stabilization of mAb-1 can be interpreted as a non-specific effect. The non-specific stabilization observed with a mAb and the non-stabilizing effect of compound **A**'s enantiomer with IFN both strongly support our initial hypothesis of a specific protein-ligand interaction leading to a stabilization against native protein aggregation of IFN. It is important to point out that the stabilizing effect of compound **A** may very well be pH dependent, especially due to its multiple titratable sites which could result in a pH dependent protein-ligand interaction profile[153].

The free energy of binding to the defined site calculated by APR-US is approximately 3 kcal/mol below the experimentally measured one (Figure 3.4, Figure S 3.4). This may indicate the presence of additional binding sites with higher affinities towards the ligand. The presence of multiple binding sites could be confirmed by unrestrained simulations (to be published by the authors). Limitations arise from using fixed protonation states for both the ligand and the protein, even though interactions between conformations, protein-ligand interactions and protonation states are well described. Taking these factors into account e.g. by constant pH MD simulations would further increase the computational cost of these simulations which is already large.

A search in the BindingDB database for compounds with binding energies between -3 and -2 kcal/mol results in multiple Guest-Host systems, with guests similar in structure and size to compound **A** (see for example BindingDB entries BDBM36112, BDBM36038, BDBM36057). Compounds in the -6 to -5 kcal/mol range tend to be more hydrophobic and/or larger (see for example BindingDB

entries BDBM50335563, BDBM23449) [154]. This indicates that the actual binding mechanism may be more complex than initially suggested.

Even though we were successful in identifying a stabilizing compound, it is important to point out that we readily relied on assumptions regarding the identification of aggregation prone regions and the binding site that have yet to be proven. A3D does not take the electrostatics surrounding hydrophobic patches into account and was only tested on a limited amount of proteins. We find that compound **A** has a stabilizing effect when exposing the formulation to freeze-thaw or shaking stress but not when exposing it to heat stress. To our knowledge, no method to predict aggregation prone regions does consider the type of forced degradation used to induce aggregates. Heat induced aggregation has been shown to induce non-native aggregation involving partial unfolding of the protein. While compound **A** was shown to bind to IFN, it would not lead to a conformational stabilization as indicated by measurements of IP and T_{onset}. After identifying the stabilizing effect of compound **A** upon freeze-thaw stress, we wanted to rule out that it was caused by a changing the process of ice formation but due to its interaction with the protein. We therefore used horizontal shaking stress as an orthogonal forced degradation method. Measurements of the compounds surface activity do not indicate a high affinity towards interfaces. Together with the observed decrease in apparent M_w from the SLS measurements in the presence of compound **A**, it supports our hypothesis of an inhibition of sub-visible particle formation by impeding the formation of specific native protein-protein contacts.

Previous studies have already shown the existence of a stress-structure interaction[52]. This poses a set-back to our approach. Since drug products have

to be stabilized against all possible stresses they could encounter during their lifetime, an excipients effect should ideally not be limited to only one type of stress. It can therefore only be considered a hypothesis that the selection of the binding site is related to the observed effects. The actual binding mechanism of compound **A** has to be determined experimentally. Due the self-association of IFN at pH 7.0, this cannot be achieved by NMR but possibly by crystallographic methods. Given these insights, it seems sensible to favor ligand-based approaches opposed to our receptor-based approach. An additional concern for the development of excipients, is the limited predictive power of forced degradation studies. Establishing relevant stability indicating assays remains a topic of ongoing research[155].

Given the proximity of the hotspot to the IFN's receptor binding site, binding kinetics and clearance of the excipient are highly relevant for an *in-vivo* application. A dissociation rate of the ligand that would limit the formation rate of the drug-target complex, i.e. a high residence time of the protein-excipient complex, will alter the drug's efficacy. From molecular dynamics simulations, we calculated the residence time $1/k_{off}$ to be below a microsecond (to be published by us). The protein-ligand complex will therefore dissociate rapidly after administration. The large size difference between small molecule excipient and protein will result in a much shorter lifetime of the excipient in the patient compared to the protein. Under these considerations, it seems plausible that the excipient will not affect the drug's efficacy.

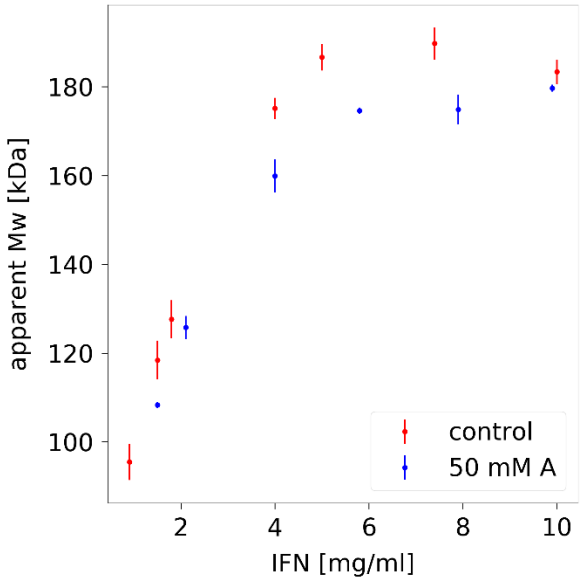
For drug products, toxicity of the excipient candidates remains a critical point. A specifically designed database containing only compounds with a proven record of low toxicity could help to overcome this problem. Considering the low

hit rate in the virtual screen, further limiting the screened chemical space might cause the elimination of any potential binders. Additional *in-silico* methods to predict toxicity can be considered, always taking resulting metabolites into consideration. Nevertheless, the discovered compound could immediately be used in diagnostic devices without the need for additional toxicity studies. While IFN is currently not a typical reagent in diagnostics, our approach can easily be transferred to any other relevant protein.

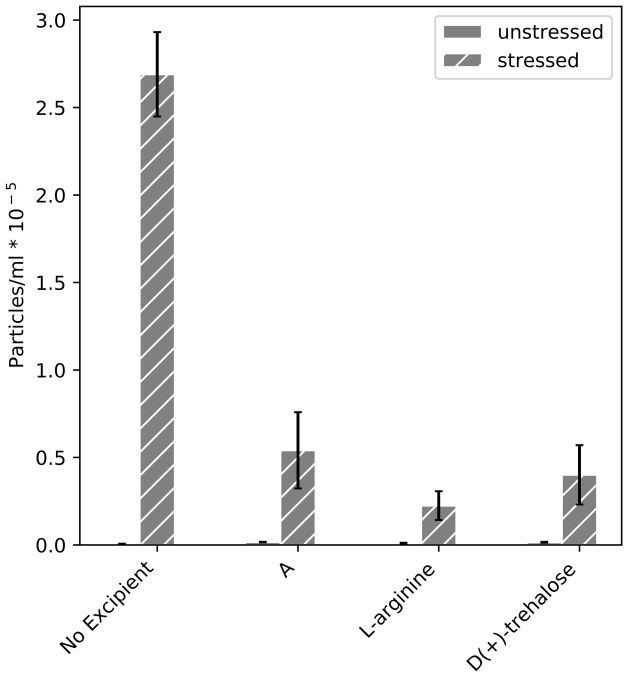
3.6. Conclusion

Here, we describe a structure-based approach that was successful in discovering a small organic molecule that stabilizes Interferon-alpha-2a and confirmed the hypothesis that the formation of a protein-ligand complex can lead to an inhibition of aggregation and particle formation. Our systematic approach helped us to narrow down a database of millions of compounds to merely 167. The compound glycyl-D-asparagine binds to IFN with an affinity of 108µM and reduces the formation of sub-visible particles and soluble aggregates after freeze-thaw and agitation stress in a concentration dependent manner. It shows higher stabilizing activity than its enantiomer glycyl-L-asparagine and the standard excipients L-arginine and D(+)-trehalose. We gave a new use to tools that are developed with small molecule drug discovery in mind and show how they can be applied to therapeutic protein formulation development.

3.7. Supplementary Data



S-Figure 1: ***Apparent Mw.*** *Measured for different IFN concentrations in presence and absence of* ***A*** *as determined by SLS. Error bars represent the standard deviations of the measurements of three independent samples.*



S-Figure 2: Sub-visible particle count before and after submitting a formulation of mAb-1 to three freeze-thaw cycles.

Table S-1: Inflection point (IP) and aggregation onset temperatures T_{onset} of IFN formulations. 1 mg/ml IFN, 50 mM excipient, 50 mM Pi, pH 7.0.

Excipient	IP [°C]	T_{onset} [°C]
A	68.0±0.0	64.2±0.1
Glycyl-L-asparagine	68.1±0.2	64.1±0.1
L-arginine	67.7±0.0	63.8±0.0
D(+)-trehalose	67.7±0.0	64.5±0.1
None	67.8±0.1	64.4±0.2

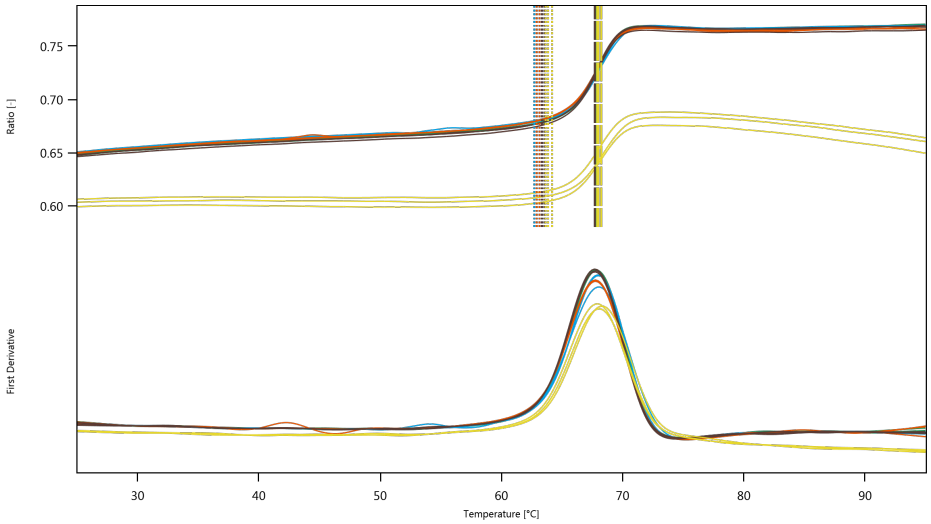


Figure S 3.1: Fluorescence curve and first derivative. Blue: Compound A, red: L-arginine, brown: D(+)-trehalose, yellow: glycyl-L-asparagine.

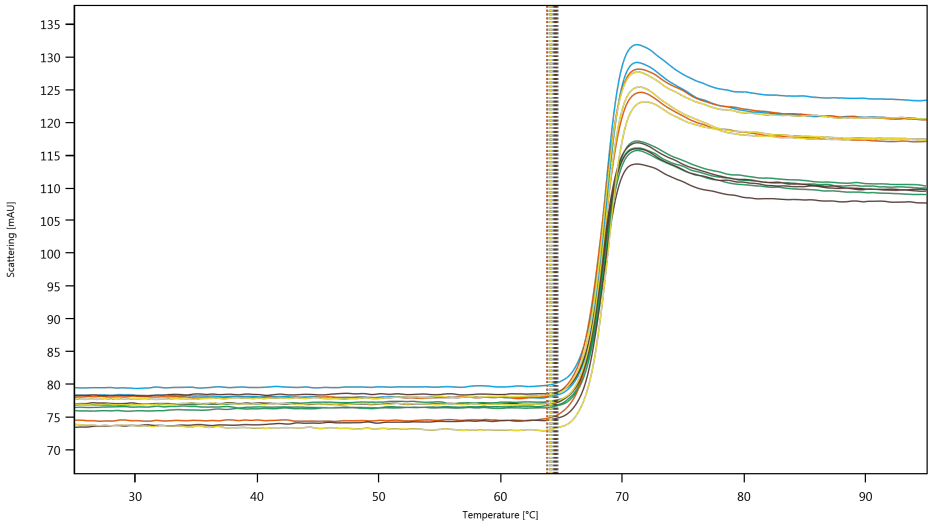


Figure S 3.2: Backscattering data. Green: Excipient free. Blue: Compound A, red: L-arginine, brown: D(+)-trehalose, yellow: glycyl-L-asparagine.

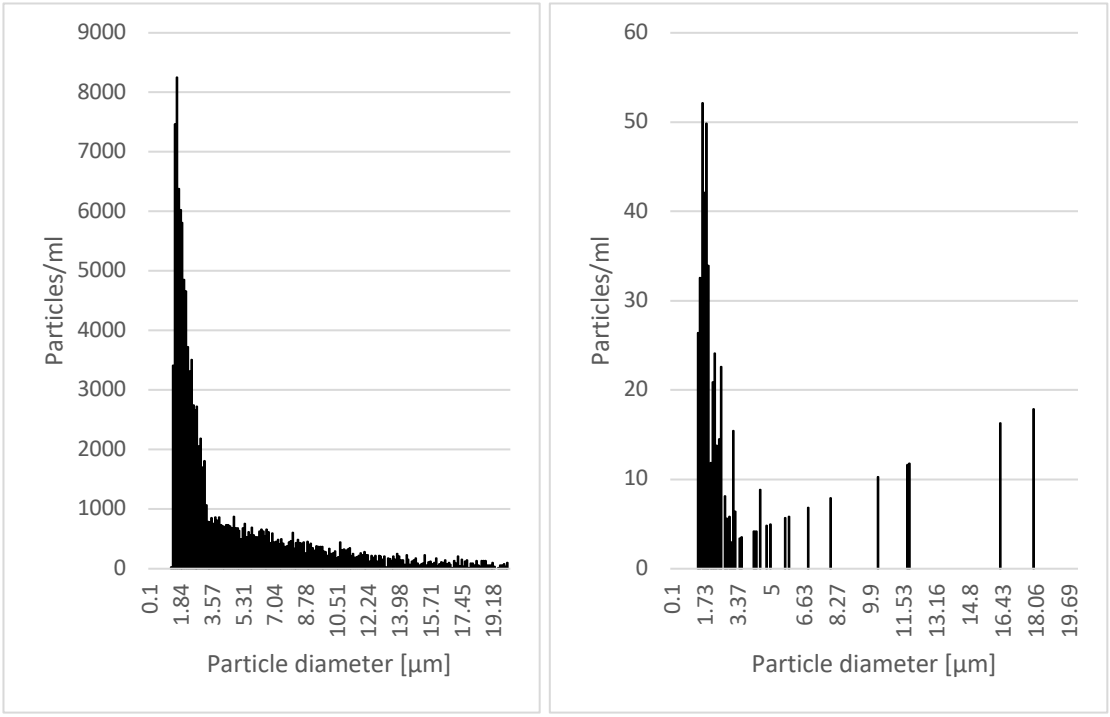


Figure S 3.3: Histogram of particle diameters after horizontal shaking in presence of 0.8 µM (left) and 6.25 mM (right) of compound A.

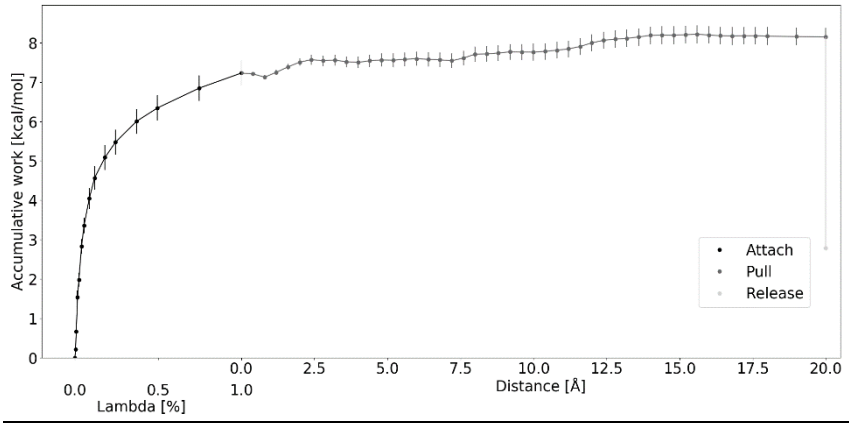


Figure S 3.4: Potential of mean force for the APR-US simulation of IFN and compound A after a simulation time of 250 ns per window. Depicted are the work required to attach the restraints, to pull the ligand from its binding site and to release the restraints. The distance for the binding site is set to 0. Error bars represent the standard error of the mean.

4. Predicted regions of protein self-interaction correlate with solution paramagnetic enhancement-NMR measurements

Andreas Tosstorff, Matja Zalar, Matthew J. Cliff, Gerhard Winter*, Alexander P. Golovanov

This work was conducted in collaboration with the Manchester Institute of Biology of the University of Manchester. The manuscript was written by Andreas Tosstorff. The experiments were performed by Andreas Tosstorff, Matja Zalar and Alexander Golovanov. The data analysis was performed by Andreas Tosstorff under the supervision of Gerhard Winter and Alexander Golovanov.

4.1. Abstract

Aggregation and self-interaction of therapeutic proteins are of great relevance for the pharmaceutical industry as they have been linked to adverse side reactions and deactivation of the active compound. While the process of oligomerization and particle formation can be monitored for example by turbidity, intrinsic fluorescence or light scattering, these methods do not resolve the underlying molecular interactions with a high level of detail. It is thus for example not possible to identify the residues that contribute to the self-interaction. This information is however crucial when it comes to assessing the developability of candidate molecules. Several computational tools have been developed to predict the contribution of individual residues to the aggregation propensity or solubility of a protein. These are typically trained or benchmarked for a certain type of protein at a specific buffer and by a forced degradation method or surrogate readout.

Here, by using solution paramagnetic relaxation enhancement (sPRE) NMR experiments, we identify self-interaction sites for Interferon-alpha-2a and compare the results to available computational tools. The sPRE method provides insights into native protein self-association and does not rely on forced degradation methods. We find that all three tested *in-silico* tools identify most regions of self-interaction correctly. However, none was able to identify all of them. While our data are limited to one protein and condition, we propose that the method could be used to enhance and supplement currently available computational tools.

4.2. Introduction

In the development of therapeutic proteins, eliminating aggregation prone candidate molecules early on is of great concern, since they increase the risk of delays and failure of the entire drug development process. Understanding protein self-interactions on a structural level will facilitate the process of identifying stable candidates and formulations.

Aggregation pathways during the lifetime of a therapeutic protein are diverse. They can be reversible or irreversible and involve partially or completely unfolded or native protein species. Irreversible aggregation is often induced by protein exposure to stress, such as heat, light, reactive molecules, freezing/thawing or shaking. It has been shown that the amino acid residues involved in the oligomerization interface depend on the type of stress the protein is exposed to, which implies a stress dependency of the aggregation mechanism[52].

The terms hotspots and aggregation prone region[111] are often used equivalently. However it is important to point out that the term “hotspot” has been coined in the context of native association of proteins[109]. Here, a hotspot is defined as the main residual contributor to the free energy of binding of a protein-protein complex. The term aggregation prone region is often applied in the context of non-native aggregation, sometimes involving the formation of beta-sheets. The term hotspot has been furthermore used in the context of druggability of a protein surface, where sites favorable for small molecule binding were termed that way[156].

In order to identify structural features of proteins that affect their physical stability, computational methods have been developed that score the protein's aggregation propensity or solubility based on either the protein's sequence or structure[157]. Exhaustive reviews on the currently available tools have been published recently[158]. Here we compare three different tools, CamSol[159], Aggrescan3D[80] and AggScore[160], with experimental data from sPRE measurements for Interferon-alpha-2a. All three tools generate structure corrected scores with different scales.

Table 4.1: Comparison of scoring tools for structurally corrected, residual aggregation propensity

Parametrization/benchmarking method:	Tools	pH	Ionic strength
Inclusion body formation	AggScore	Yes	No
Solubility	CamSol	Yes	No
Qualitative solubility from literature	Aggrescan(3D)	No	No

CamSol calculates a solubility score from a linear combination of the amino acid's hydrophobicity and electrostatic charges smoothed over a window of seven amino acids to account for their interplay. The term is structurally corrected to account for proximity in space and solvent exposure. Positive scores are interpreted as favoring solubility and negative scores as aggregation prone. Aggrescan3D scores residues by each amino acid's intrinsic aggregation

propensity and solvent accessible surface area, and those of spatially close residues weighted by their pairwise distance. In the 'Dynamic Mode', the score is calculated for various generated conformations, and the score for the most aggregation prone conformer is reported. The scale is inverse to that of CamSol, with aggregation prone regions receiving positive scores and soluble ones receiving negative scores. Just as CamSol it can be accessed through webserver. AggScore as implemented in Schrödinger's Maestro software scores residues by the intensity and relative orientation of hydrophobic and electrostatic surface patches. The program was trained on published data on inclusion body formation of 31 adnectin proteins. The scale starts at zero for low aggregation propensity and increases with aggregation propensity.

The amount of experimental methods to characterize interfaces of self-interaction on a level of residual or atomic resolution are scarce and challenging, especially when data should be recorded *in-situ*. Crystallographic methods can only be employed to determine self-interaction interfaces under conditions at which proteins form specific oligomers, which can be affected by crystal packing. While this method is of course well established and powerful, it does not serve the purpose of identifying regions of self-interaction *in-situ* at arbitrary conditions and in solution. Hydrogen-deuterium exchange mass spectrometry (HDX-MS) is an emerging and still time-consuming method. Here, the protein is exposed to deuterated water and depending on the level of residual solvent exposure, deuterium and hydrogen atoms will exchange. The resulting changes in molecular weight are measured by digesting the protein and measuring the molecular weight of the produced fragments. One can then assign the degree of deuteration and correlate it to the formation of protein-

protein interfaces. It has been successfully employed to identify for example regions of self-association of monoclonal antibodies (mAbs)[52,161].

As an alternative, here we propose the usage of sPRE-NMR to measure residual solvent exposure and thereby identify residues that are involved in self-interaction of globular proteins in solution. We show how they correlate with computational methods and discuss reasons why *in-silico* tools may produce false-negative results.

Paramagnetic relaxation enhancement (PRE) occurs due to dipolar interactions for example between an unpaired electron of a molecular probe and surrounding hydrogen atoms. In an NMR experiment, these interactions result in increased longitudinal and transverse relaxation rates of the protons, causing signal broadening and reduced intensities[162]. The effect has therefore been widely used to study protein structure and dynamics by attaching spin labels either covalently to a protein or adding them to the solution (sPRE)[163–165].

One molecular probe widely employed is 4-hydroxy-2,2,6,6-tetramethylpiperidin-1-oxyl (TEMPOL). It was used to map the protein surface of lysozyme and identify solvent exposed amide groups in NOESY and TOCSY experiments[166]. It was furthermore used to identify druggable site in bovine pancreatic RNase A[156]. Petros et al. found TEMPOL to be superior over gadolinium(III) diethylenetriaminepentaacetate which binds specifically to carboxylate groups[167]. Bovine pancreatic trypsin inhibitor (BPTI) has been characterized by measuring TEMPOL induced signal attenuation, finding a region of low attenuation, which was attributed to tightly bound water molecules preventing the residues' contact with paramagnetic probe[168].

sPRE can furthermore be used to identify the interface of protein complexes, as residues that are buried in the protein-protein interface are no longer accessible by the dissolved paramagnetic probe. Their relaxation times are therefore significantly longer than those of solvent accessible residues. This phenomenon has been exploited to identify the interface of human matrix metalloproteinase 3 (MMP-3) and the inhibitory domain of human tissue inhibitor of metalloprotease (TIMP-1)[169].

Weak self-association has been observed by PRE for example for histidine containing protein (HPr) and cytochrome c peroxidase (CcP)[70,170]. To our knowledge there is no report of the use of sPRE to study protein self-association. By comparing relaxation rates of IFN amino acid residues in the presence and absence of the paramagnetic probe TEMPOL, we measure their solvent accessibility. A comparison of the experimental solvent accessibility of each residue to its theoretical solvent accessible surface leads to the identification of residues engaging in protein self-association.

4.3. Methods

Protein Expression

1 µl of plasmid coding for Interferon-alpha-2a with a TEV cleavage site and 6xHis-tag (Genscript) was transformed into *E. coli* Origami cells by heat shock. Transformed cells were incubated at 37 °C for one hour. The medium was then spread on LB plates with ampicillin and kanamycin antibiotics and incubated overnight. Cells were grown in ^{15}N labeled M9 medium and expression was induced by Isopropyl β-D-1-thiogalactopyranoside (IPTG). Cells were harvested by centrifugation. Cell lysis occurred by ultrasound in 6 M guanidine

hydrochloride (GndHCl). IFN was added to a Ni-NTA slurry and incubated under gentle shaking for 1 h using 20 mM Tris, 6 M GndHCl, 50 mM NaCl and 5 mM β-mercaptoethanol as mobile phase. The bound sample was washed with 5 ml mobile phase. Weakly bound molecules were eluted by adding 10 mM imidazole to the mobile phase. IFN was eluted by adding 500 mM imidazole to the mobile phase. Refolding was performed in 20 mM TRIS, 150 mM NaCl, 25 mM arginine, 25 mM glutamate, 5 mM EDTA, 1 mM glutathione red, 0.25 glutathione disulfide, pH 8.5 at 4 °C. The sample was then dialyzed in acetate buffer at pH 4.0 and concentrated by ultrafiltration (Vivaspin, 5 kDa cut-off, Sartorius, Stonehouse, UK).

NMR

All NMR experiments were acquired at 25°C on 800 MHz Bruker Avance III spectrometer equipped with 5 mm triple resonance TCI cryoprobe with temperature control unit. Spectra were acquired using Bruker Topspin 3.5 (Bruker) while processing and analysis was performed in Bruker Topspin 4.0 (Bruker) and Dynamics Center 2.5.5 (Bruker). NMR samples were prepared by adding 5 % v/v $^{2}H_2O$ to 4 mg/ml ^{15}N labelled Interferon-alpha-2a in 10 mM Acetate, pH 4. The backbone assignment of Interferon-alpha-2a was based on BMRB entry 4081 [171].

Paramagnetic relaxation enhancement by soluble probes (sPRE) experiments

^{15}N longitudinal (R1) relaxation rates were measured using a pseudo-3D hsqct1etf3gpsi3d experiment from the standard Bruker library. Longitudinal (T_1) relaxation times were calculated by fitting the signal intensities to a single exponential function available in Dynamics Center 2.5.5. Errors of fit were

estimated using a 95% confidence level. R_1 relaxation rates were calculated as the inverse of the longitudinal relaxation time T_1. NMR sPRE data were obtained by determining ^{15}N R_1 relaxation rates in the absence and presence of 36 mM TEMPOL.

Computational tools

PDB entry 1ITF was submitted to the webservers of CamSol and A3D. AggScore was used within Schrödinger's Maestro program. All methods were used with their default parameters. A3D was used in the dynamic mode in order to account for protein flexibility. MSAs for N_{bb} were calculated using PyMOL. To make them comparable across the different methods, aggregation propensity scores were normalized according to Equation 4.1, where $fnorm$ is the normalized score, $f(x)$ the score, $fmin$ the lowest, yet aggregation prone, score and $fmax$ the most aggregation prone score.

$$fnorm = \frac{f(x) - fmin}{fmax - fmin} \qquad \textit{Equation 4.1}$$

4.4. Results

Spectra obtained were coherent with previously reported data[172]. No peak shifts were observed after the addition of TEMPOL, however multiple peaks broadened or vanished due to the paramagnetic nature of the compound, which also affected peak resolution (Figure 4.1). Signals from the His-tag were not assigned.

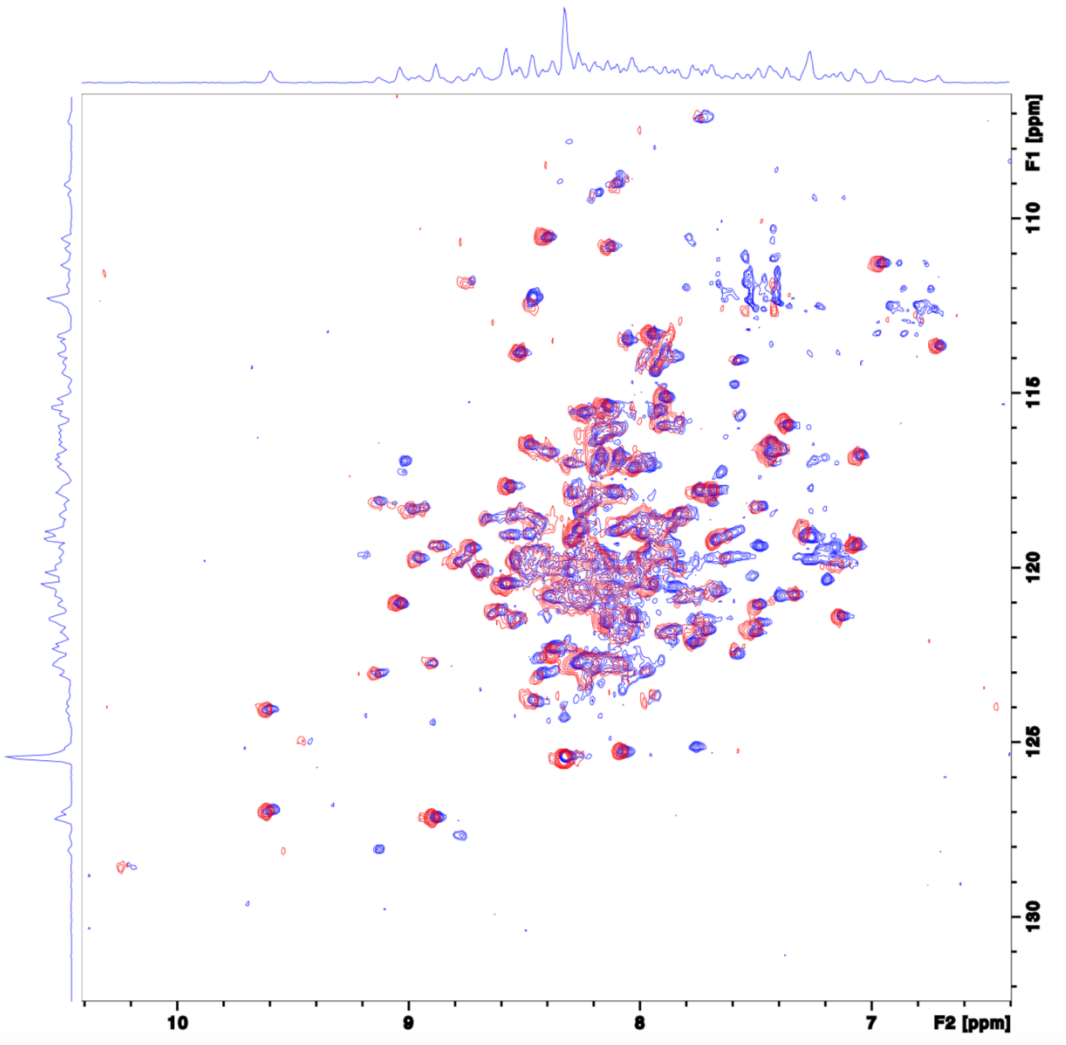


Figure 4.1: Overlay of the 10th 23 plane of the pseudo-3D hsqct1etf3gpsi3d experiment. Blue without TEMPOL, red with TEMPOL.

T_1 Relaxation rates were calculated for all resolved signals (supplementary data: Figure 4.7) and compared to molecular surface areas (MSA) of the backbone amide groups. Residues with a change in relaxation rate ΔR_1 below 0.05 upon addition of TEMPOL, an error below 0.5 and a backbone amide MSA above the median were considered to contribute to oligomerization interfaces (Figure 4.2, supplementary data). All other signals were not considered for comparison with

in-silico tools. Here, we use sPRE to evaluate false negative and true positive results from the prediction tools. The residues identified by our sPRE approach are mostly hydrophobic (Table 4.2). Most notably is a patch constituted by residues E141, A145, M148 (patch 1). A second patch was located at the highly exposed loop region G102, V105, T108 (patch 2). A third patch is constituted by residues F36, F38 and L128 (patch 3). M59, Q61 and N65 make up a fourth patch (patch 4). Isolated interacting residues are located at positions A75 and Q158.

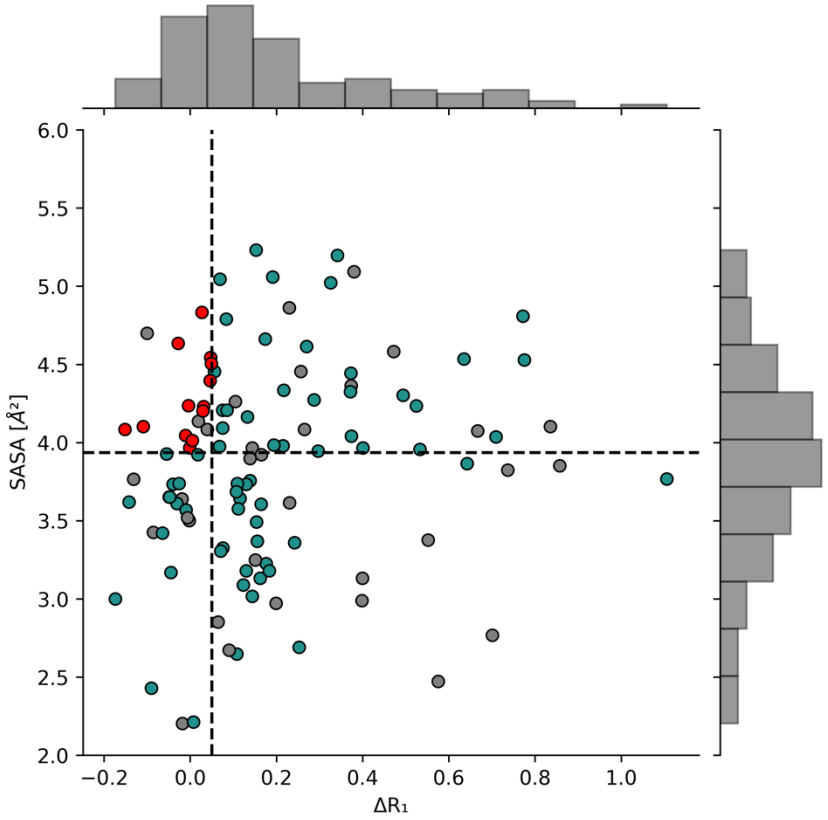


Figure 4.2: MSA for backbone amides plotted against the change in Relaxation rate ΔR_1. Grey markers indicate an error of ΔR_1>0.5, red markers indicate aggregation prone residues. Error bars are not shown for clarity.

Table 4.2: Overview on residues involved in self-interaction.

Residue Number	Amino acid	Patch	ΔR1 [1/s]	N_{bb} MSA [Å²]	Side chain
36	F	3	0.05	4.54	hydrophobic
38	F	3	-0.01	4.05	hydrophobic
59	M	4	-0.15	4.08	hydrophobic
61	Q	4	-0.11	4.10	polar, uncharged
65	N	4	0.00	3.97	polar, uncharged
75	A	-	0.00	4.01	hydrophobic
102	G	2	-0.03	4.63	no side chain
105	V	2	0.03	4.23	hydrophobic
108	T	2	0.05	4.40	polar, uncharged
128	L	3	-0.06	4.28	hydrophobic
141	E	1	0.03	4.83	negative
145	A	1	0.03	4.20	hydrophobic
148	M	1	0.00	4.24	hydrophobic
158	Q	-	0.05	4.51	polar

CamSol, A3D and AggScore all ranked F27 as highly aggregation prone. Characterization of this specific residue by sPRE is not possible, as the signal is not unambiguously resolved. The region ranging from residue 98 to 120 was flagged by all three prediction tools and also by sPRE data. The specific residues do vary however depending on the method. Only AggScore flagged L128 as aggregation prone. Patch 1, formed by residues E141, A145, and M148 is not fully captured by any of the tools. A3D signals residue E141, CamSol does so for residue A145 but both with a low score (Figure 4.3). It is also the only tool to identify two out of three residues from patch 4, Q61 and N65, again with a rather low score. Strikingly, none of the computational tools would highlight F36 and F38 as aggregation prone, despite their obvious intrinsic hydrophobicity, and neither were A75 or Q158.

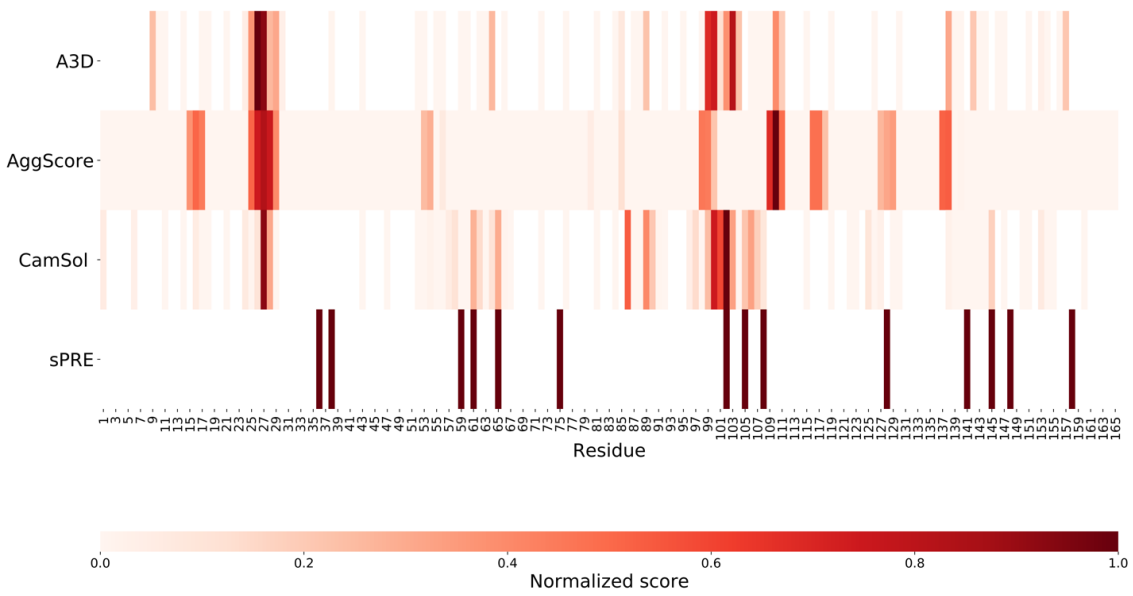
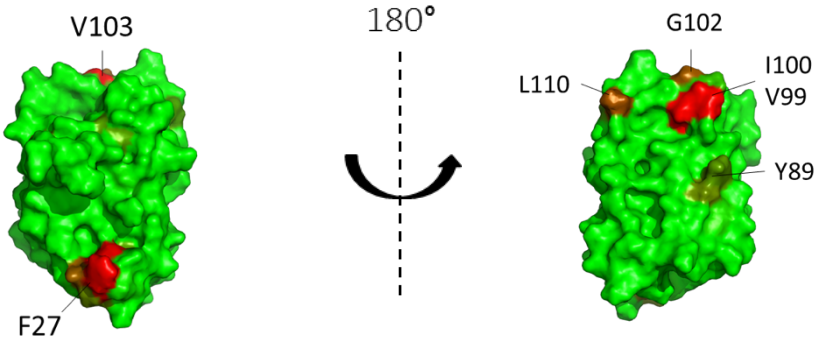


Figure 4.3: Normalized scores of residual aggregation propensity compared to residues identified as aggregation prone by sPRE. It is important to note that residues

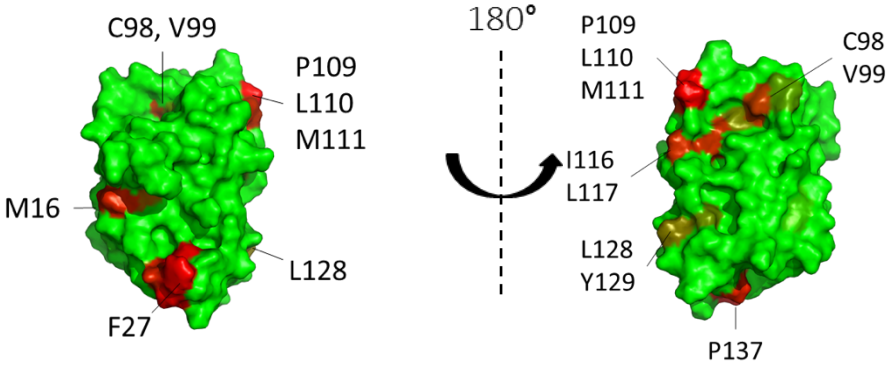
with sPRE scores equal to zero should not be interpreted as non-aggregation prone, but as unresolved.

Projecting the scores on the 3D structure of IFN shows that while the residues identified by the computational tools do not match perfectly with the residues identified by sPRE, broad regions are approximated well. For example, F27, which was highlighted by all tools and not resolved by NMR is in close neighborhood to patch 1. Furthermore, all tools identified a region surrounding V105 and T108. While CamSol and A3D identified the region around F64 as aggregation prone, AggScore did not. AggScore did signal residue L128 which was not identified by neither A3D nor CamSol (Figure 4.4, Figure 4.5). It is also the only tool identifying residues (C98, V99) in proximity to the experimentally determined Q158.

A3D



AggScore



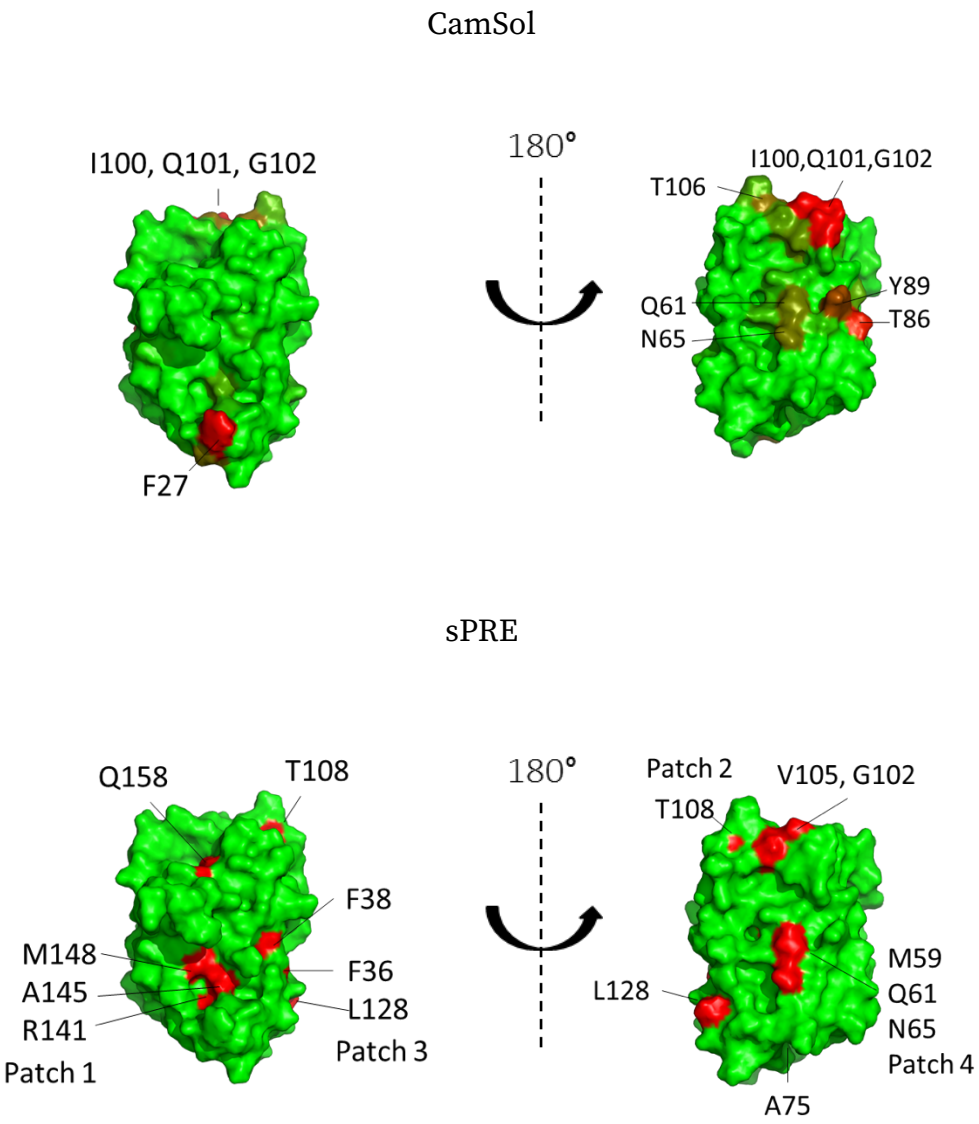


Figure 4.4: Comparison of aggregation prone regions determined by A3D, CamSol, AggScore and sPRE. Comparing experimental and predicted regions shows that most regions are identified well by the in-silico tools. Aggregation prone regions are colored in red. Left: View 1. Right: View 2.

4.5. Discussion

We were able to identify regions of self-interaction for IFN *in-situ*. The protein shows multiple regions prone to participate in self-association which goes hand

in hand with a broad coverage of the surface by hydrophobic residues, explaining its overall low solubility (Figure 4.5). Attempting to increase the protein's solubility by targeting all of these regions through mutations does therefore not appear feasible without drastically altering the structural integrity and activity profile.

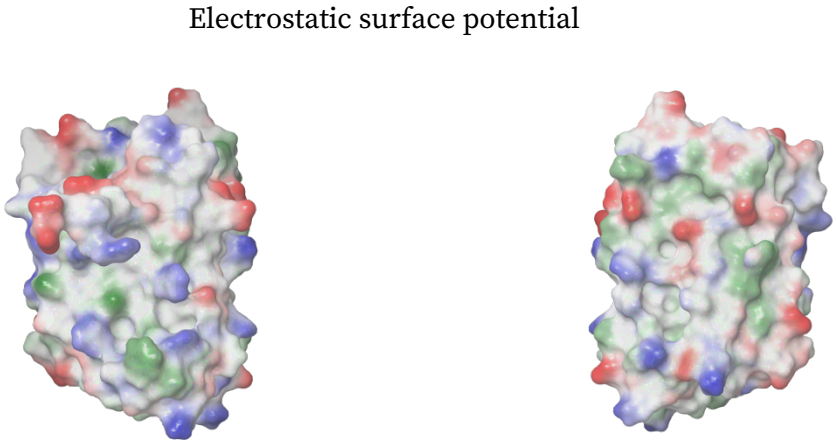


Figure 4.5: Protein surface charge distribution. Red: negative charge, blue: positive charge, green: hydrophobic. Left: View 1. Right: View 2.

Qualitatively, all three tested *in-silico* methods, perform well in closing in on most aggregation prone regions. None of the evaluated methods is however able to capture all of the regions flagged by sPRE, most notably the hydrophobic patch 3.

The false negative result by all computational methods regarding F36 and F38 could be explained by a bias introduced by using a single protein structure. Both phenyl-rings in the input structure are pointing inwards to the protein core, which could lead to an underestimation of aggregation propensity due to a low calculated MSA (Figure 4.6). Opposed to the prediction tools, which consider the solvent exposure of the entire amino acid residue, for the evaluation of the sPRE

measurements, only N_{bb} MSA was used as reference. The backbone amides of F36 and F38 are clearly solvent exposed in the input structure which is why we could identify them as aggregation prone. The procedure is plausible, as the NMR signals correspond to the N_{bb}. Solvent exposure may not be ideal as scoring parameter, given that self-interaction may cause conformational changes that are unfavorable for the free protein[173]. Even though A3D was used in "Dynamic mode", it apparently did not sample conformations in which F36 and F38 were sufficiently solvent exposed.

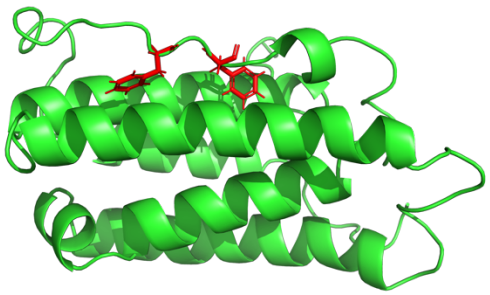


Figure 4.6: F36 and F38 marked in red. Phenyl-rings of F36 and F38 of input structure are pointing inwards.

Interestingly, patch 1 is also involved in binding the receptor IFNAR2 (3S9D), which indicates that the presented tools could also be used in a broader context to predict regions of heterogeneous protein-protein interactions[174].

Our sPRE approach is able to identify residues that lie in the interface of a protein-protein complex, however, it cannot rank the individual residue's contribution to the free energy of binding. sPRE should therefore be considered as an orthogonal method to evaluate regions prone to self-association and especially to identify false negatives. Combining both in-silico and experimental methods in order to identify regions to target either by mutation or LMW

molecules appears as a promising approach to inhibit self-association in order to increase protein stability or reduce viscosity. We previously reported how we discovered a dipeptide that increases the stability of IFN by targeting patch 1.

It is important to point out the limitations of the sPRE method, which lie for example in the ambiguous assignment of some peaks, as was for example the case for F27. NMR experiments are furthermore limited by the size of the protein to be studied. Certain buffer conditions could also lead to a degree of protein aggregation too high to be studied with the presented method.

While there is a good agreement between the *in-silico* and experimental data, it seems unlikely that the computational tools will perform as good in a scenario of non-native aggregation that involves partial unfolding. We therefore propose to develop novel methods that discriminate between the stress or mechanism that leads to protein aggregation and account for solvent pH. sPRE-NMR could be a helpful method to evaluate these prediction tools.

4.6. Conclusion

For the first time, we demonstrate the usage of sPRE-NMR to characterize regions of self-association for a protein. Multiple regions were identified. Among others, the binding site of the proteins target, IFNAR2 shows to participate in self-interaction. We used our *in-situ* data to evaluate three different computational tools designed to predict residual contribution to aggregation or solubility. Overall there is a good agreement between the three methods and experimental data, however none of the methods is able to identify all of the regions that were flagged by sPRE data.

4.7. Supplementary Data

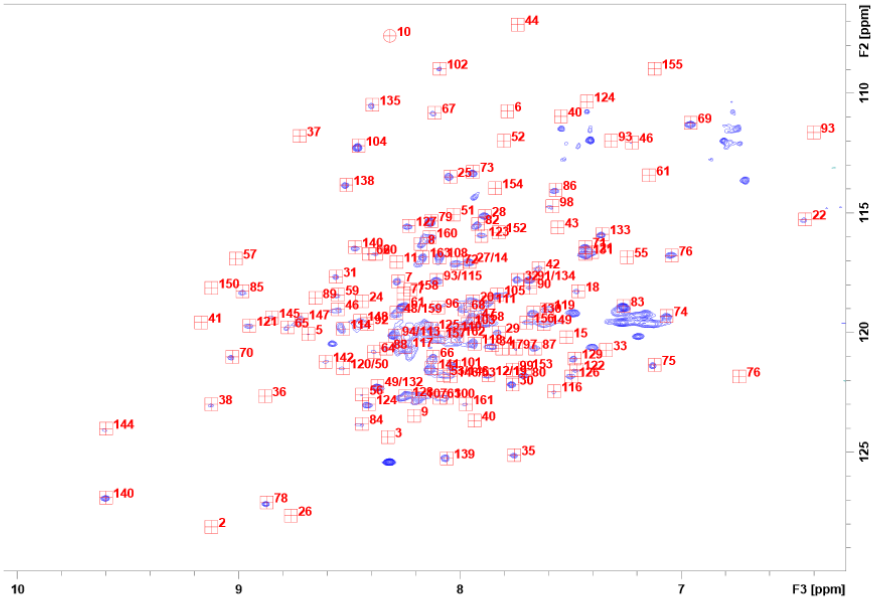


Figure S 4.1: 23 Plane, number 10, without TEMPOL

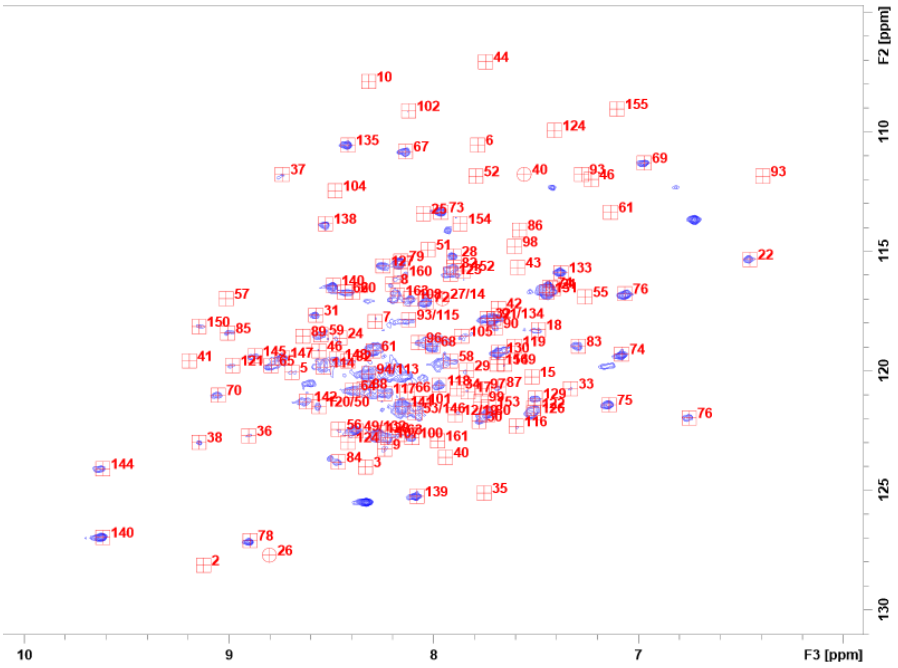
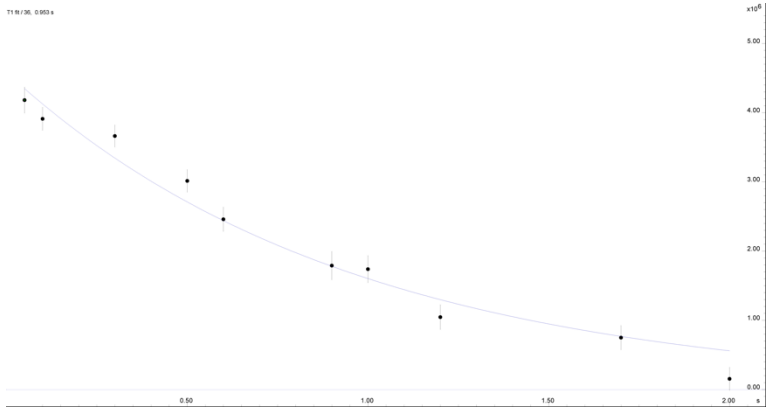
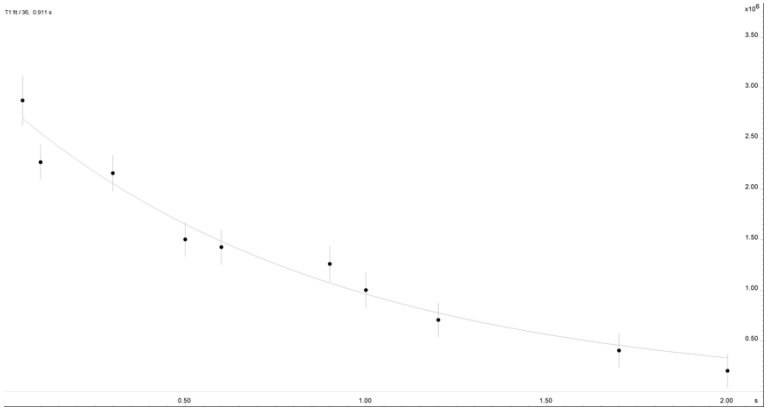
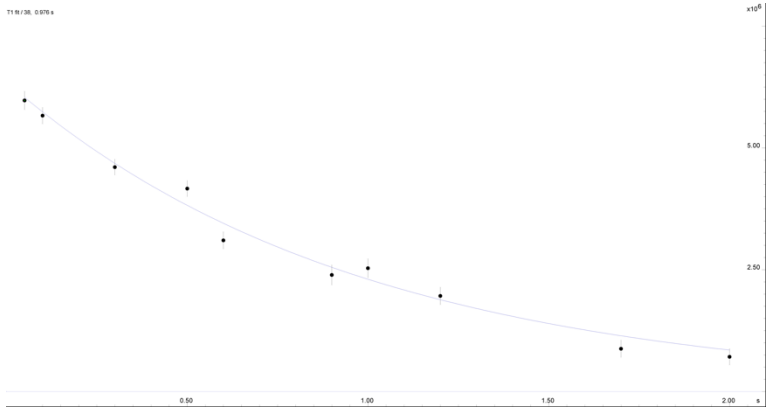
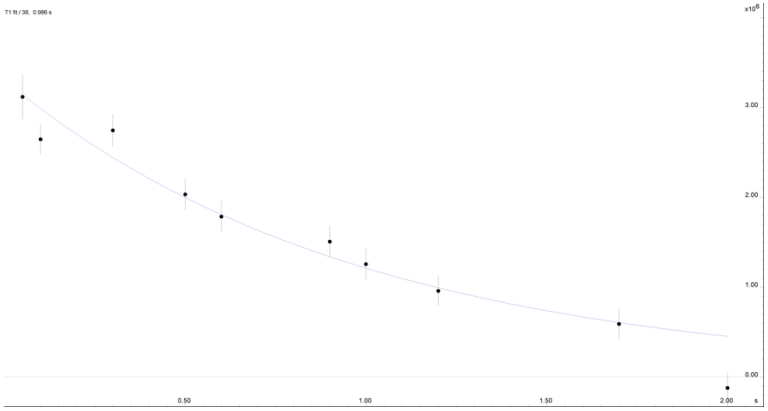
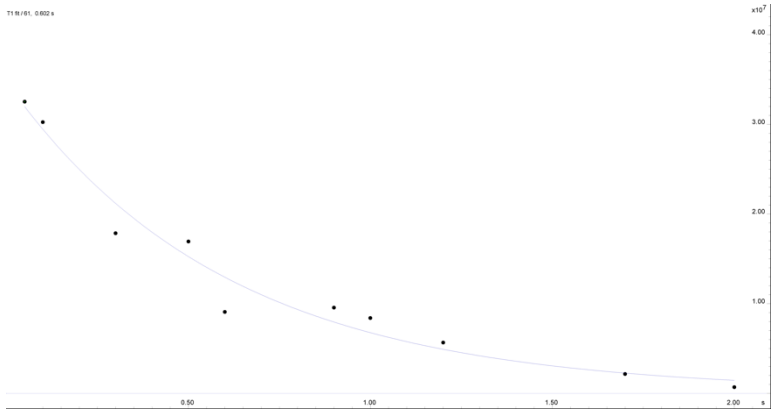
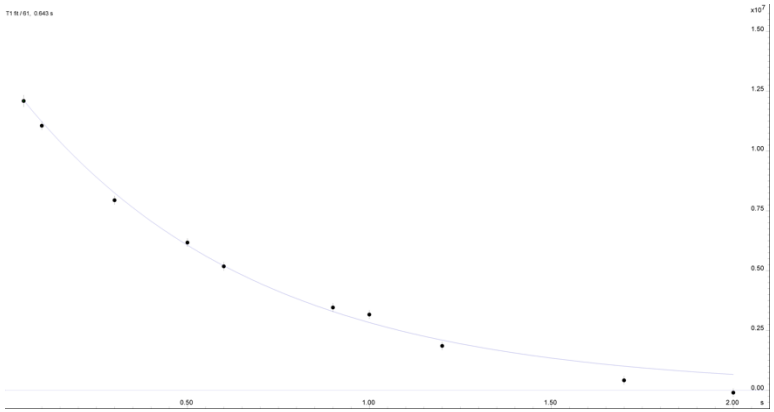


Figure S 4.2: 23 Plane, number 10, with TEMPOL

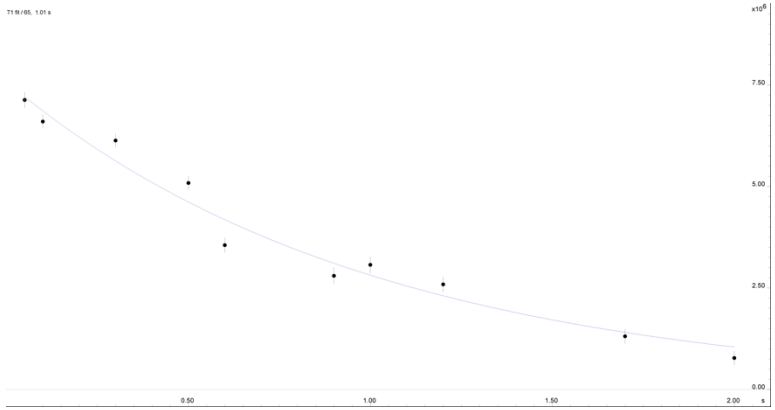
Residue	0 mM TEMPOL	36 mM TEMPOL
36		
38		

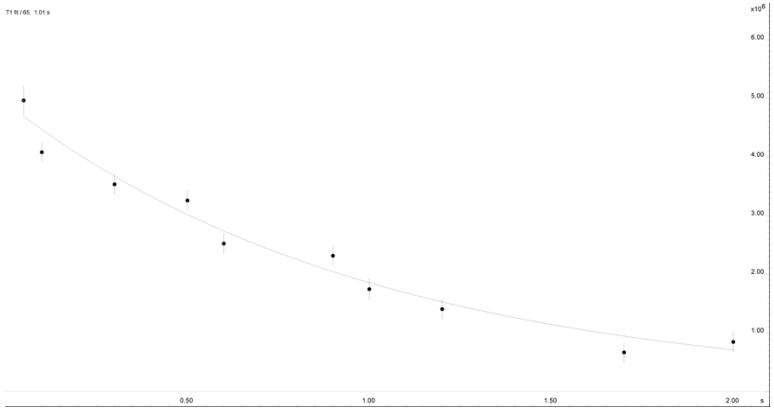
61


T1 fit / 61, 0.602 s
x10^7
4.00
3.00
2.00
1.00
0.50
1.00
1.50
2.00
s

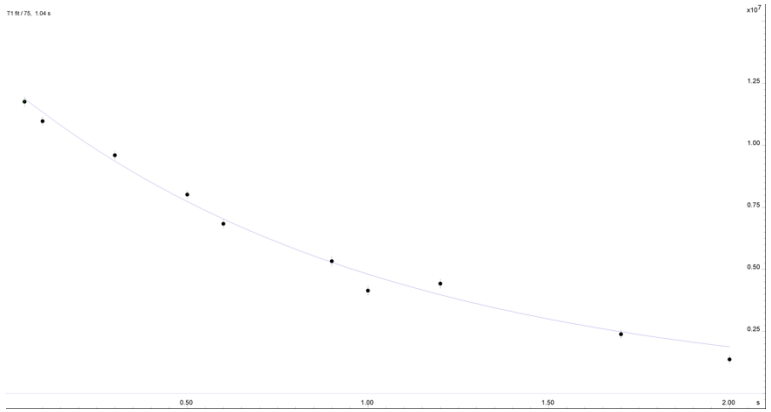

T1 fit / 61, 0.643 s
x10^7
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.50
1.00
1.50
2.00
s

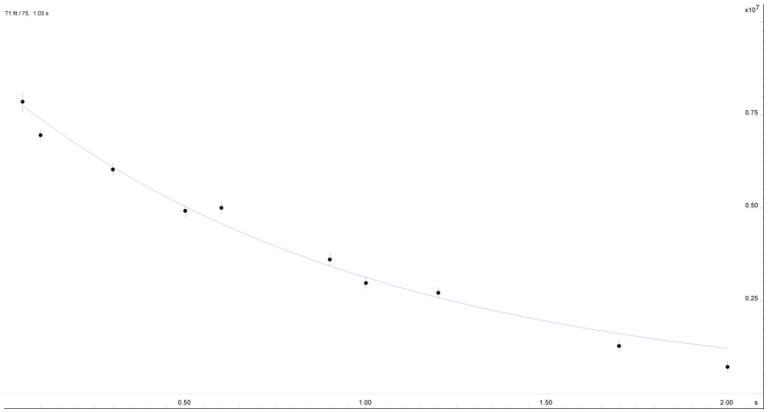
65


T1 fit / 65, 1.01 s
x10^6
7.50
5.00
2.50
0.00
0.50
1.00
1.50
2.00
s

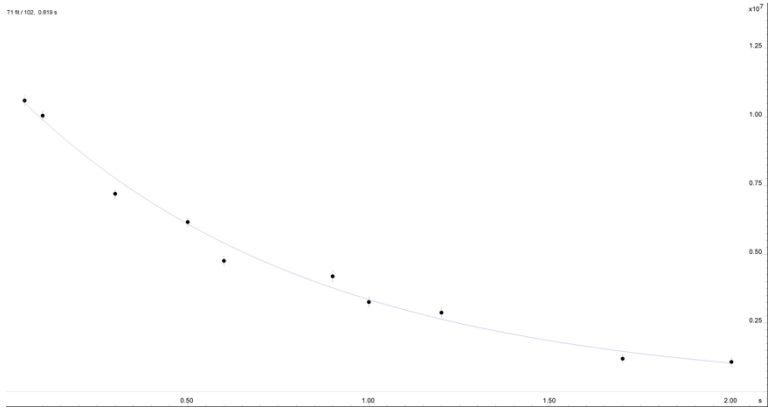

T1 fit / 65, 1.01 s
x10^6
6.00
5.00
4.00
3.00
2.00
1.00
0.50
1.00
1.50
2.00
s

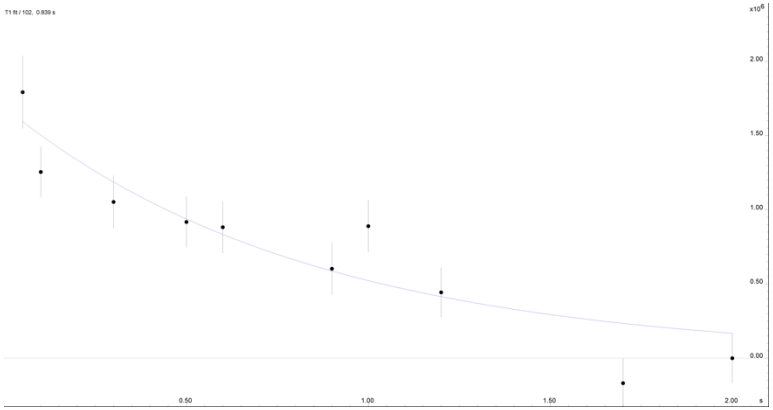
75


T1 fit / 75, 1.04 s
x10^7
1.25
1.00
0.75
0.50
0.25
0.50
1.00
1.50
2.00
s

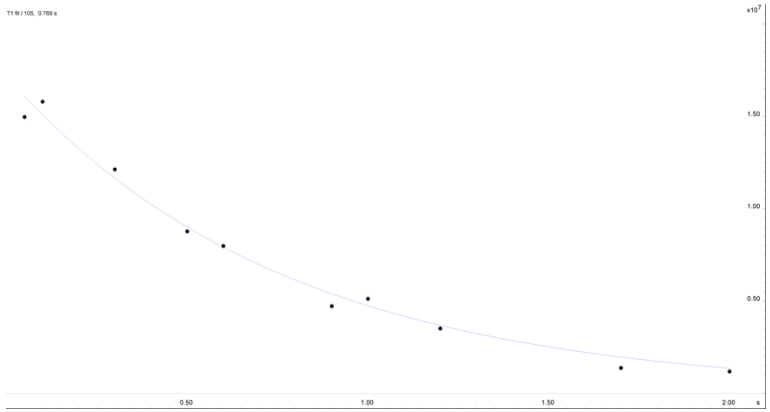

T1 fit / 75, 1.03 s
x10^7
0.75
0.50
0.25
0.50
1.00
1.50
2.00
s

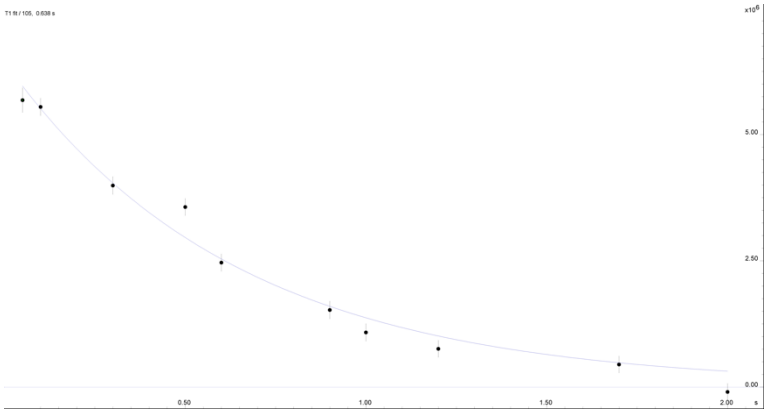
102


T1 fit / 102, 0.819 s
x10^7
1.25
1.00
0.75
0.50
0.25
0.50
1.00
1.50
2.00
s

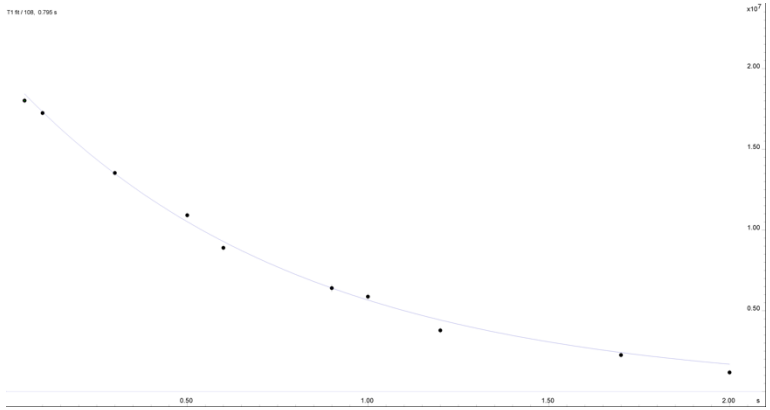

T1 fit / 102, 0.839 s
x10^6
2.00
1.50
1.00
0.50
0.00
0.50
1.00
1.50
2.00
s

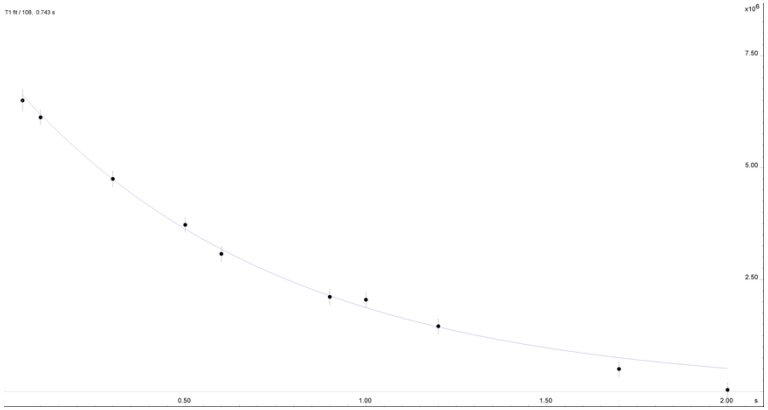
105


T1 fit / 105, 0.769 s
x10^7
1.50
1.00
0.50
0.50
1.00
1.50
2.00
s

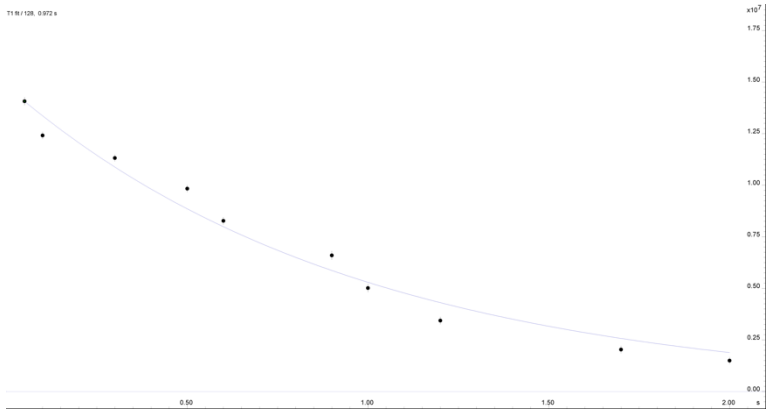

T1 fit / 105, 0.638 s
x10^6
5.00
2.50
0.00
0.50
1.00
1.50
2.00
s

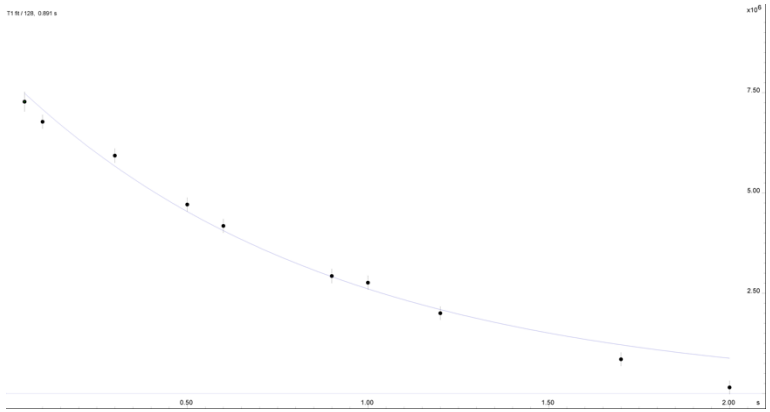
108


T1 fit / 108, 0.795 s
x10^7
2.00
1.50
1.00
0.50
0.50
1.00
1.50
2.00
s

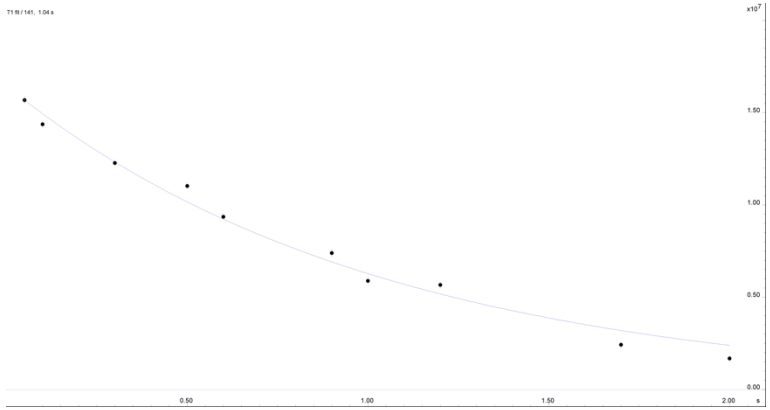

T1 fit / 108, 0.743 s
x10^6
7.50
5.00
2.50
0.50
1.00
1.50
2.00
s

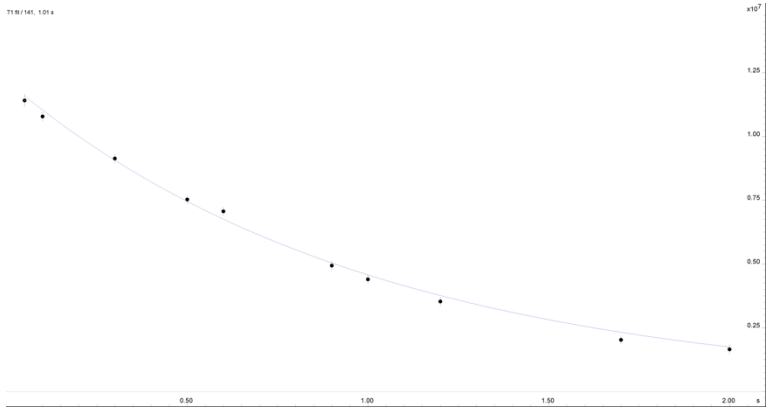
128


T1 fit / 128, 0.972 s
x10^7
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.50
1.00
1.50
2.00
s

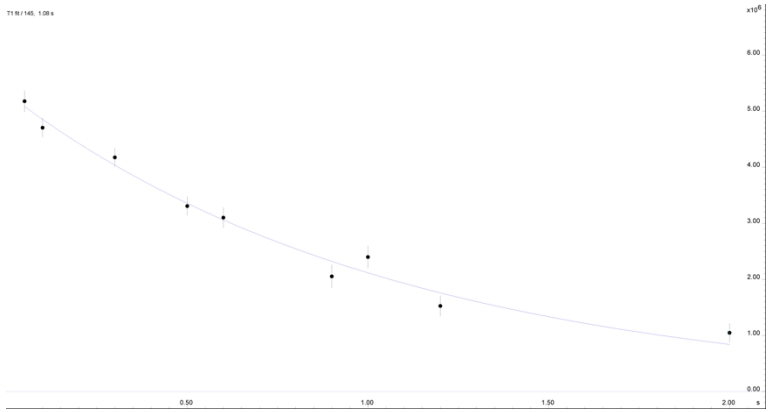

T1 fit / 128, 0.891 s
x10^6
7.50
5.00
2.50
0.50
1.00
1.50
2.00
s

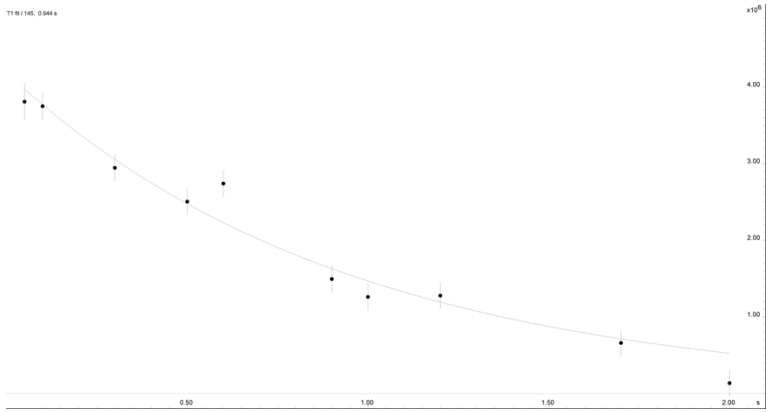
141


T1 fit / 141, 1.04 s
x10^7
1.50
1.00
0.50
0.00
0.50
1.00
1.50
2.00
s

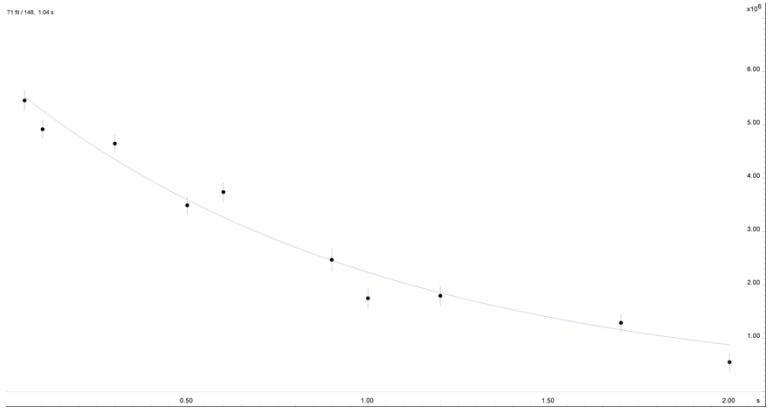

T1 fit / 141, 1.01 s
x10^7
1.25
1.00
0.75
0.50
0.25
0.50
1.00
1.50
2.00
s

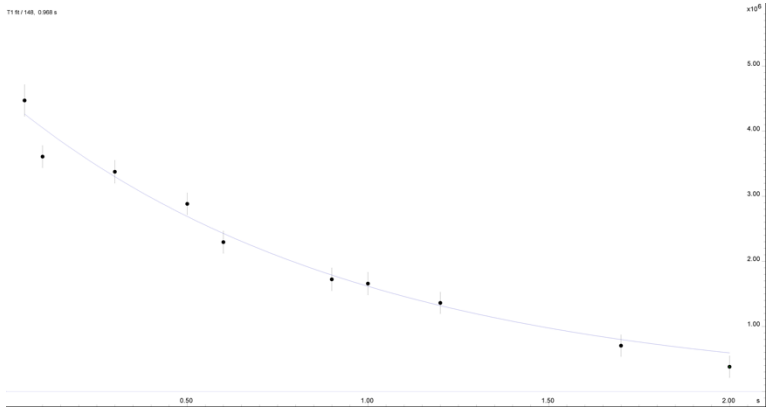
145


T1 fit / 145, 1.08 s
x10^6
6.00
5.00
4.00
3.00
2.00
1.00
0.00
0.50
1.00
1.50
2.00
s

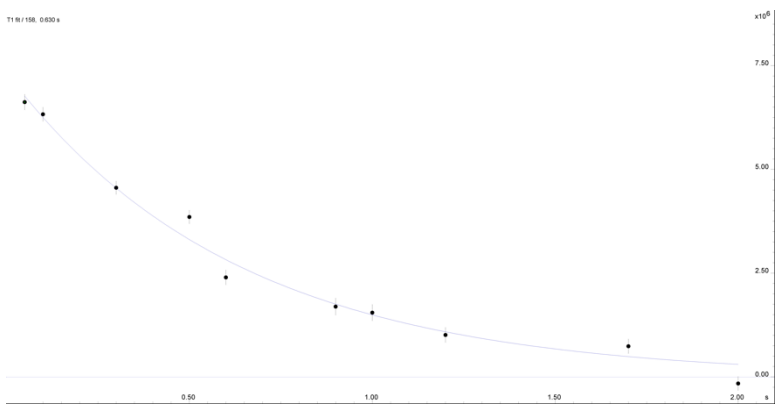

T1 fit / 145, 0.944 s
x10^6
4.00
3.00
2.00
1.00
0.50
1.00
1.50
2.00
s

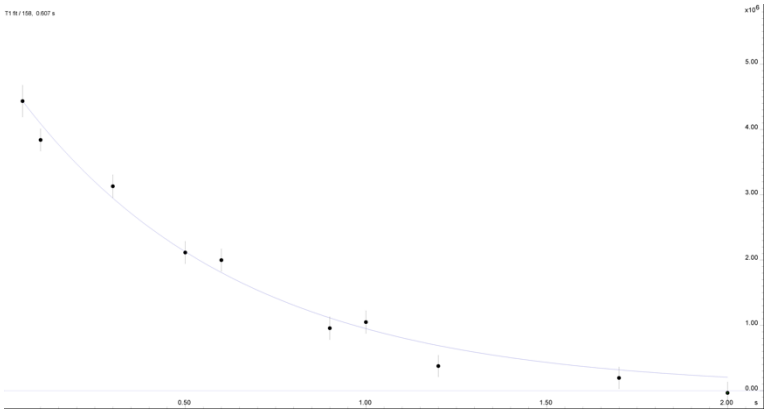
148


T1 fit / 148, 1.04 s
x10^6
6.00
5.00
4.00
3.00
2.00
1.00
0.50
1.00
1.50
2.00
s


T1 fit / 148, 0.968 s
x10^6
5.00
4.00
3.00
2.00
1.00
0.50
1.00
1.50
2.00
s

158


T1 fit / 158, 0.630 s
x10^6
7.50
5.00
2.50
0.00
0.50
1.00
1.50
2.00
s


T1 fit / 158, 0.607 s
x10^6
5.00
4.00
3.00
2.00
1.00
0.00
0.50
1.00
1.50
2.00
s

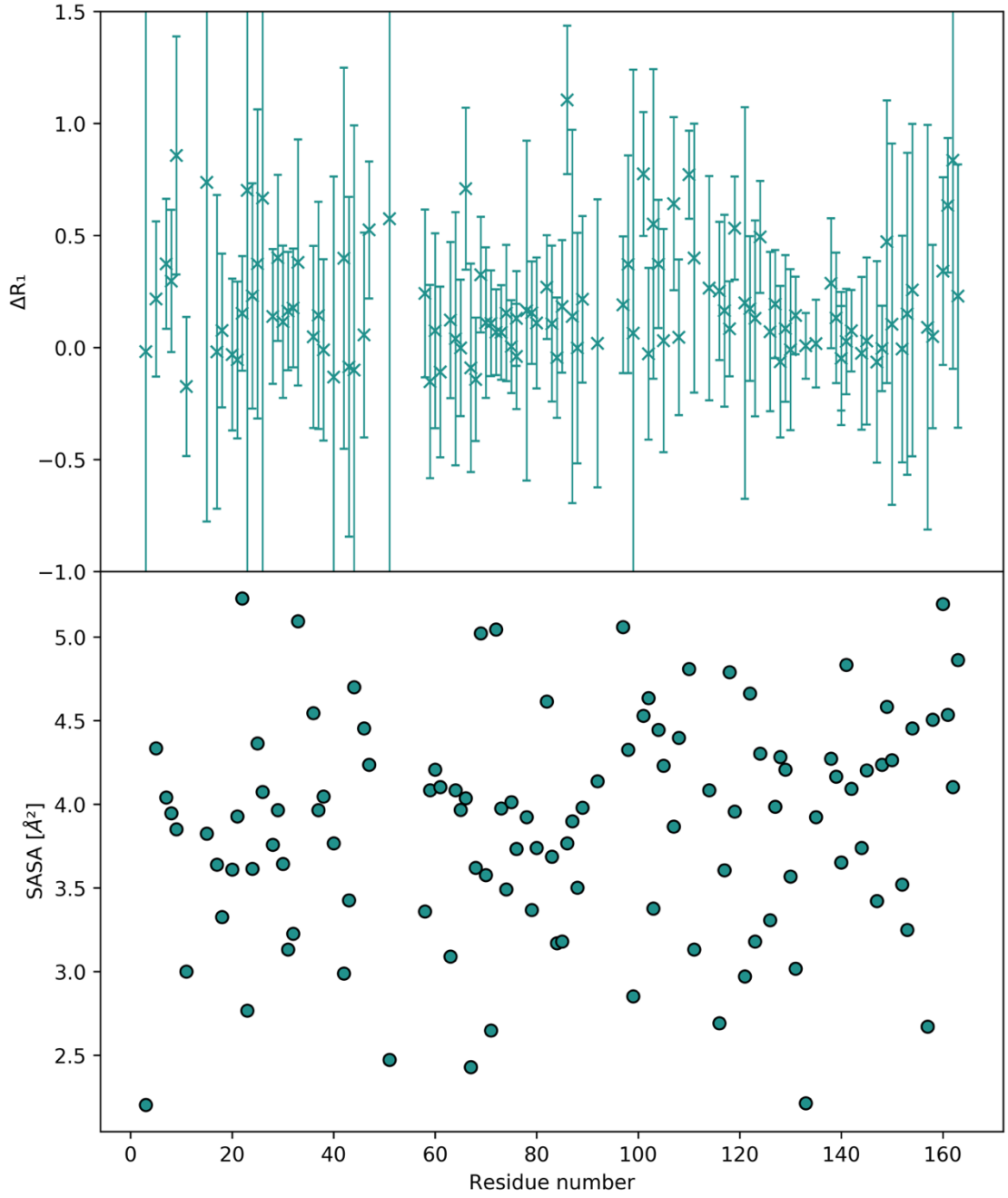


Figure 4.7: ΔR_1 and N_{bb} MSA for all resolved residues.

Error propagation was calculated by a Taylor expansion according to:

$$\sigma_{R_1} = \left|\frac{dR_1}{dT_1}\sigma_{T_1}\right| = \left|-\frac{1}{{T_1}^2}\sigma_{T_1}\right|$$

Where σ_{T_1} is the error of the relaxation time T_1 as derived from the fit and σ_{R_1} is the error of the relaxation rate R_1.

5. Study of the interaction between a novel, protein-stabilizing dipeptide and Interferon-alpha-2a by construction of a Markov State Model from Molecular Dynamics simulations

Andreas Tosstorff[1*], Günther H.J. Peters[2], Gerhard Winter[1]

A version of this chapter has been published in the European Journal of Parmaceutics and Biopharmaceutics:

Tosstorff, A., Peters, G.H.J., & Winter, G. Study of the interaction between a novel, protein-stabilizing dipeptide and Interferon-alpha-2a by construction of a Markov state model from molecular dynamics simulations, *Eur. J. Pharm. Biopharm.* **149**, 105-112 (2020).

This work was conducted in collaboration with the Department of Chemistry of the Technical University of Denmark. The manuscript was written by Andreas Tosstorff. All simulations, experiments and data analysis were performed by Andreas Tosstorff under the supervision of Günther H.J. Peters and Gerhard Winter.

5.1. Abstract

We recently reported the discovery of a novel protein-stabilizing dipeptide, glycyl-D-asparagine, through a structure-based approach. As the starting hypothesis leading to the discovery, we postulated a stabilizing effect achieved by binding of the dipeptide to an aggregation prone region on the protein's surface. Here we present a detailed study of the interaction mechanism between the dipeptide and Interferon-alpha-2A (IFN) through the construction of a Markov state model from molecular dynamics trajectories. We identify multiple binding sites and compare these to aggregation prone regions. Additionally, we calculate the lifetime of the protein-excipient complex. If the excipient remained bound to the IFN after administration, it could alter the protein's therapeutic efficacy. We establish that the lifetime of the complex between IFN and glycyl-D-asparagine is extremely short. Under these circumstances, stabilization by stoichiometric binding is consequently no impediment for a safe use of an excipient.

Keywords

Interferon-alpha-2a, Excipient, Protein Aggregation, Protein Formulation, Markov State Model

5.2. Introduction

Small molecules are commonly found in therapeutic protein drug formulations as co-solutes with the intend to stabilize the drug product among other against chemical degradation or aggregation of the therapeutic protein. Opposed to

native self-association, protein aggregation proceeds by multiple steps that among other involve a partial or complete unfolding of the protein[38].

Two commonly accepted mechanisms of stabilization of a protein against aggregation by a small molecular co-solute are preferential exclusion and stoichiometric binding[5,56,175–177]. Preferential exclusion describes an entropically driven rise of chemical potential of both, protein and co-solute molecules relative to their separate solutions. The increase in chemical potential manifests by a reduced concentration of co-solute in proximity to the protein surface relative to the bulk solution. Protein unfolding will lead to an increased exposure of protein surface, increasing the unfavorable exclusion of co-solute. The protein's native state is therefore preferred to the non-native. The stabilizing effect of a diverse group of co-solutes such as sugars, polyols, amino acids, methylamines and inorganic salts on proteins has been well established and traced back to preferential exclusion as mechanism of action[55,175]. Preferential exclusion is observed for weakly interacting co-solutes that require to be present at high concentration (above 100 mM) in order to benefit protein stability[55,56].

Stoichiometrically interacting co-solutes are known to stabilize proteins by binding preferentially to the native protein structure relative to the unfolded one. Stabilization through stoichiometric binding can for example be measured by differential scanning fluorimetry or calorimetry and results in a shift of the infliction point of the characteristic unfolding curve (T_m) to higher temperatures[53].

The large majority of pharmaceutical excipients act through the mechanism of preferential exclusion, which has the intrinsic benefit that their application is not limited to a single protein but across many if not all. Developing excipients that act as stoichiometric stabilizers has largely been neglected, despite the potential to provide a complementary mean to stabilize a protein[178].

We previously described the discovery of an outstanding stabilizing effect of the dipeptide glycyl-D-asparagine at low concentrations against aggregation of Interferon-alpha-2A upon exposure to freezing-thawing and shaking stress[179]. We found that the dipeptide would bind to the protein at a μM affinity and reduces particle formation at low concentration (6.25 mM), hinting at a stabilization through a stoichiometric interaction. The compound was discovered through a virtual screen that targeted the hydrophobic and solvent exposed residue Phe27. This residue is involved in the interaction between interferon-alpha-2 and interferon-alpha-receptor 2 (Figure 5.1, PDB entry 3S9D)[174]. A potential risk of stoichiometrically acting excipients is that the protein drug-excipient complex does not disassociate after drug administration, thus potentially altering the drug's efficacy. The lifetime of the protein-excipient complex is therefore a crucial parameter to consider when developing stoichiometrically binding excipients. A short lifetime means that the protein-excipient complex disassociates rapidly. As the excipient is much smaller than the protein, it will distribute, metabolize and clear much faster than the protein after administration. In the case of the dipeptide presented here, its metabolism is facilitated further due to the presence of a peptide bond prone to enzymatic hydrolysis[180]. Its low molecular weight compared to that of IFN will lead to a fast clearance through the kidneys[181]. A long lifetime of the protein-excipient complex would instead result in a permanent occupation of the protein surface

by one or more excipient molecules, potentially altering the proteins interaction with its target molecule, and consequently its efficacy.

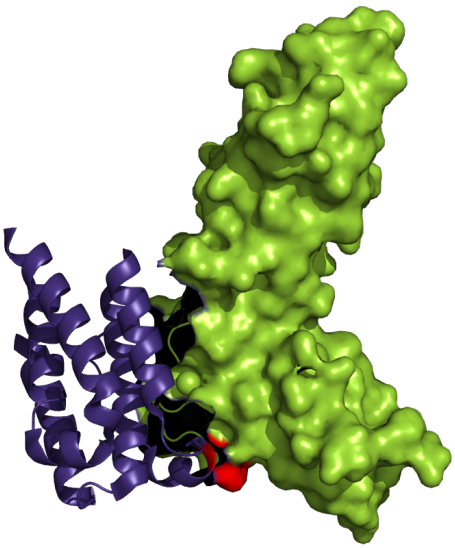


Figure 5.1: Complex between Interferon-alpha-2 (violet) and Interferon-alpha-receptor-2 (green) (PDB entry 3S9D). The aggregation prone region targeted by the excipient (red) to inhibit interferon-alpha-2a aggregation coincides with the binding site to the receptor.

The occupation of a protein by a ligand is the result of the simultaneously occurring binding and unbinding processes[182]. When considering the equilibrium reaction between Protein P and ligand L to form a complex PL (Equation 5.1), the rates of binding, r_{on}, and unbinding, r_{off} can be defined as the product of a rate constant k and the concentration of the reactants (Equation 5.2, Equation 5.3).

$$P + L \rightleftharpoons PL \qquad \textit{Equation 5.1}$$

$$r_{on} = k_{on} \cdot [P] \cdot [L] \qquad \textit{Equation 5.2}$$

$$r_{off} = k_{off} \cdot [PL] \qquad \textit{Equation 5.3}$$

Here, k_{on} and k_{off} are the rate constants for the corresponding binding and unbinding reactions and $[P]$, $[L]$, $[PL]$ are the concentration of the protein, ligand and protein ligand complex respectively.

In order to estimate the lifetime of a protein-ligand complex, the residence time τ can be calculated from the inverse of the off-binding rate constant k_{off} (Equation 5.4)[183].

$$\tau = \frac{1}{k_{off}} \qquad \textit{Equation 5.4}$$

Computational simulations are a popular mean to study protein-ligand interactions, as they allow to gain insights on the interaction with atomic detail. Interactions between the excipients mannitol, sucrose, trehalose and sorbitol and a ligase and a Fab fragment have previously been studied by docking calculations[112]. In this work a correlation between calculated binding affinity of excipients to the protein and T_m was observed. The T_m experiments were, however, conducted at excipient concentrations ranging from 145 to 220 mM, which may hint at a stabilization by preferential exclusion. A method that combines protein-protein and protein-excipient docking combined with molecular dynamics (MD) simulations to discover new excipients is described as well in a patent application[184]. It aims at identifying excipients that bind to regions involved in protein self-association in order to reduce protein

aggregation. It does not state how protein aggregation is measured experimentally and does not relate simulation data to data from experiments on protein aggregation. It does furthermore not yield any novel excipients but is limited to commonly employed stabilizing substances such as amino acids.

MD simulations are a mean to study protein ligand interactions at atomic detail, where each atom is treated as a classical particle and interactions between these particles are defined in force fields[185]. Shukla and Trout used MD simulations to determine the preferential interaction coefficient of a protein in aqueous arginine solutions of 250 to 2500 mM[186]. The study of stoichiometric binding by molecular dynamics is most commonly reported in the context of small molecule drug discovery. Analysis of molecular trajectories, which are often collected in parallel setups is challenging and can introduce errors due to biases in the starting structure and introduced restraints intended to enhance sampling of rare transitions.

Markov state theory has been used in trajectory analysis to eliminate these biases and accurately describe the mechanism of protein-ligand interactions[187]. In the Markov approach, a time dependent system is assumed to move from one discrete state to another. The transitions between these states are assumed to be memoryless, meaning that if the system is in a specific state, its future state does not depend on the system's history. The calculation of the probability of transitioning from one such state to another is the central result of a Markov state model. In the case of an MD trajectory of a protein-ligand system, the bound protein-ligand complex could correspond to one discrete state, whereas the unbound system could correspond to another discrete state. The transition

probability between these two states could then be related to the binding affinity of the complex.

The publicly available EMMA and HTMD programs drastically facilitate the construction of Markov state models from MD trajectories[188]. In order to construct a Markov state model, MD trajectories have to be discretized, which means to assign each trajectory frame to a defined state. Discretization for Markov state model generation has been shown to work best when the dimensionality of the trajectory data is reduced. In an MD simulation, each simulated atom is described by three cartesian coordinates indicating its position, and three cartesian velocities, indicating its current movement in three-dimensional space. One frame of an MD trajectory therefore consists of 6N dimensions, where N is the number of simulated atoms. By identifying a set of features of lower dimensionality, such as dihedral angles or the distance between the ligand and each protein residue, the complexity of an MD trajectory can be reduced, but the information of interest, e.g. protein conformation or protein ligand binding, is preserved. This so called featurization of an MD trajectory is typically the first step in reducing dimensions in order to construct a Markov state model.

Mathematical approaches to reduce the dimensionality of a matrix are principle component analysis (PCA), which preserves the highest degree of variance or time lagged independent component analysis (TICA) which preserves the highest degree of kinetic variance. The first principal component will therefore describe the motion of highest amplitude, while the first time lagged independent component will describe the slowest transition[189]. The term slow is used here to describe transitions that happen only after a long time, such as for

example, the unbinding of a tightly bound ligand. In the analysis of protein-ligand binding processes, the slowest transitions are those of interest and therefore TICA is typically the preferred choice of dimensionality reduction.

In order to assign each trajectory frame to a discrete state, after reduction of dimensionality, the data set is typically discretized by a clustering algorithm. Here, the trajectory frames found to be similar to each other, are grouped together into one single state. Finally, by counting the number of transitions between the discrete states, the transition probabilities between the clustered states can be calculated and experimental observables can be derived. Detailed descriptions on the workflow to construct a Markov model has been published by the developers of EMMA[190].

To our knowledge, Markov state theory has not yet been employed to describe protein-excipient interactions. Here we use a Markov state model to investigate the mechanism of interaction between the stabilizing dipeptide glycyl-D-asparagine and Interferon-alpha-2A, to elucidate interaction sites and to estimate the residence time of the formed protein-excipient complex.

5.3. Methods

System setup and simulation

Each randomized starting systems was constructed using HTMD[191]. One of 24 structures from PDB entry 1ITF was randomly selected using NumPy's random.choice function[171]. The protonation states of the protein were adjusted to pH 7.0. The protein was centered and randomly rotated. Subsequently the ligand was centered, randomly rotated and placed at a random distance between

6 to 11 Å away from the furthest protein atom along the x-axis (Figure S 5.1). The ligand was again rotated randomly around the origin. The system was then solvated with an additional 5 Å buffer. Finally, two disulfide bridges were built.

The ligand was parametrized using GAFF2 for bonded and non-bonded parameters. Atomic partial charges were calculated with Gaussian 16 (Gaussian Inc., Wallingford, CT, U.S.A.) and fitted with the RESP procedure in antechamber. Each system was minimized and equilibrated prior to the production run. Minimization was performed using pmemd on CPUs, whereas molecular dynamics simulations were performed on GPUs using pmemd.CUDA implemented in Amber 18[148,192–194]. A cutoff of 9 Å was defined for nonbonded interactions. The first 5000 cycles of minimization used the steepest descent algorithm, followed by 5000 cycles using the conjugate gradient algorithm. MD simulations were run using Langevin dynamics with a collision frequency of 1 ps^{-1} [195]. The SHAKE algorithm was used to allow for timesteps of 2 fs[196].

Equilibration followed the scheme described by Henriksen et al. and consisted of three steps[130]. For 1 ps, no pressure scaling was used and the temperature was set to 10 K. The system was then heated to 300 K within 100 ps. The last stage consisted of 50 equilibration cycles of 100 ps, each using a Monte Carlo barostat set to atmospheric pressure. Production was performed using the NVT ensemble, running 60 ns per trajectory. 600 trajectories were generated in total.

Data analysis

A Markov State Model was constructed using HTMD which builds on PyEMMA. We followed a stepwise approach based on the multiple tutorials accompanying HTMD and PyEMMA. Trajectories were first stripped of all water, sodium and

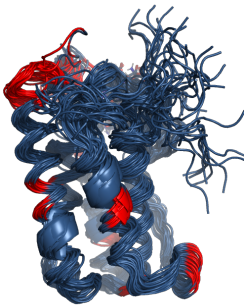
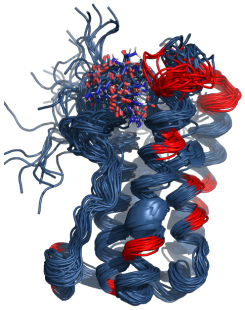
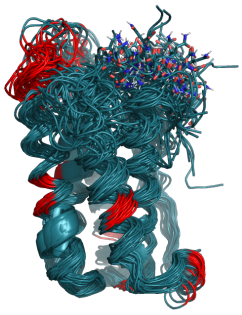
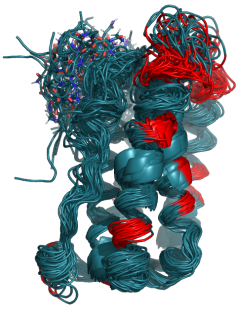
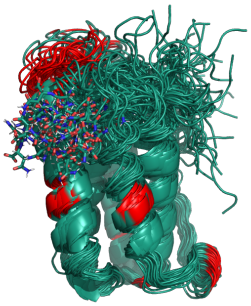
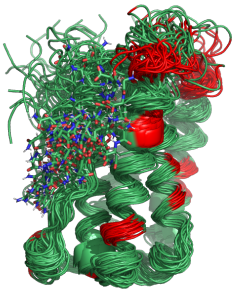
chloride. The selected featurization scheme to study the protein-ligand interaction was the pairwise, residual, minimum distance between each protein residue and the dipeptide, considering only heavy atoms. The data was then projected on the first 10 time-lagged independent components with a lag time of 1 ns. The projected data was then clustered into 60 micro-states using the k-means algorithm. A Markov state model was constructed with a lag time of 10 ns and the micro-states were clustered to 5 macro-states using PCCA++. The model was validated using the Chapman-Kolmogorov (CK) criterion. If the model fulfills the CK criterion, the occupation of future states is independent of past states, i.e. the model is markovian. (Figure S 5.2). Statistical errors of thermodynamic and kinetic quantities were obtained from 1000 bootstrapping cycles retaining 80% of the data. Structures were rendered using PyMOL.

Identification of aggregation prone regions

Three different methods were used to identify aggregation prone regions on the surface of Interferon-alpha-2A: Aggrescan3D[80], AggScore[160] and CamSol[159]. For Aggrescan3D and CamSol the scores were calculated by submitting the first frame of PDB entry 1ITF to the corresponding webserver. The aggregation propensity according to the AggScore method was calculated using Schrödinger's Maestro software using the same structure file as for the 2 other methods. Aggregation prone residues identified through any of the methods are residues 16, 27, 61,65, 86, 89, 98, 99, 100, 101, 102, 103, 106, 109, 110, 111, 116, 117, 128, 129, 137.

5.4. Results

From the constructed Markov model, 5 macro-states were identified. State 5 comprises mostly unbound and non-specifically associated structures. States 0 to 4 show specific regions of interaction between the dipeptide and INF with different degrees of fuzziness. Macro-state 0 involves interactions with residues 41, 42, 43, 46, 48, 51, 114, 115, 164. Macro-state 1 can be characterized by interactions with residues 3, 40, 41, 45-49, 155-165. For Macro-state 2, residues 5-10, 13, 90, 91, 93, 94, 96, 147 were identified. In macro-state 3, the dipeptide is in contact with residues 33-38, 40, 41, 42, 46, 114, 118, 121, 122, 125, 146, 149, 165. Macro-state 4, which is the least fuzzy one, only involves residues 22, 23, 73, 75-78 (Figure 5.2). While the study of protein conformation was not the scope of this study, we observed high flexibility in the N-terminal and the C-terminal loop region as was already described previously [171]. Interactions with the C-terminus are consequently present in multiple of the macro-states. When comparing the binding sites to aggregation prone regions identified on the protein surface, we find that macro-state 0 and 2 show an interaction close to the aggregation prone residues 98 to 100 (predicted by Aggrescan3D, AggScore, CamSol). Macro-state 3 shows an interaction in close proximity to the aggregation prone residues 27 (predicted by Aggrescan3D, AggScore, CamSol), 128 and 129 (predicted by AggScore). Macro-state 4 shows binding in proximity to aggregation prone residue 137 (predicted by AggScore).

Macro-state 0		
Macro-state 1		
Macro-state 2		
Macro-state 3		

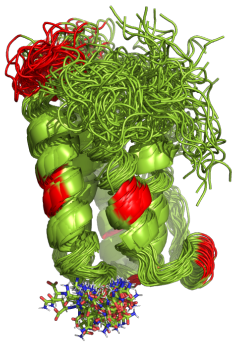
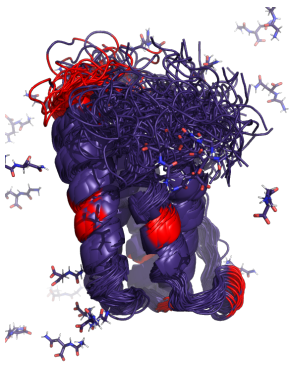
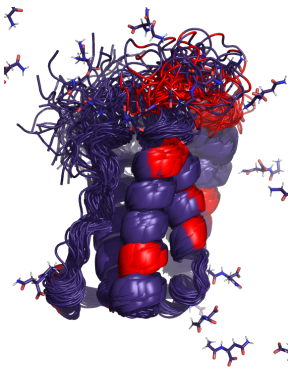
Macro-state 4		
Macro-state 5		

Figure 5.2: Representative structures from two perspectives at 180° rotation of macro-states defined by the constructed Markov model. Aggregation prone residues colored in red: 16, 27, 61,65, 86, 89, 98, 99, 100, 101, 102, 103, 106, 109, 110, 111, 116, 117, 128, 129, 137. The colors of the protein structures are meant to facilitate the correlation of the structures with Figure 5.4.

When comparing the docked structure that led to the discovery of the dipeptide as protein-stabilizing substance, one observes a similarity to macro-state 3. In both, the docked pose as well as in macro-state 3, interactions with residues 33, 34 and 146 are observed. The interaction between ARG 33 and the dipeptide in both cases consists of a salt bridge between the residue's side chain and the dipeptide's carboxyl group (not depicted for macro-state 3). In the docked pose, the interaction with residue 34 is between the backbone carbonyl group and the dipeptides N-terminal amine. In the MD simulation, the amide nitrogen of

residue 34 interacts with the dipeptide's amide carbonyl group. The docked pose suggests a hydrogen bond between the side chain carboxyl of GLU146 and the dipeptide's amine, which is also observed in the third macro-state. The docked pose shows the ASN side chain of the dipeptide forming a hydrogen bond with the backbone carbonyl of ALA145, which is not the case in the structures sampled from macro-state 3 (Figure 5.3).

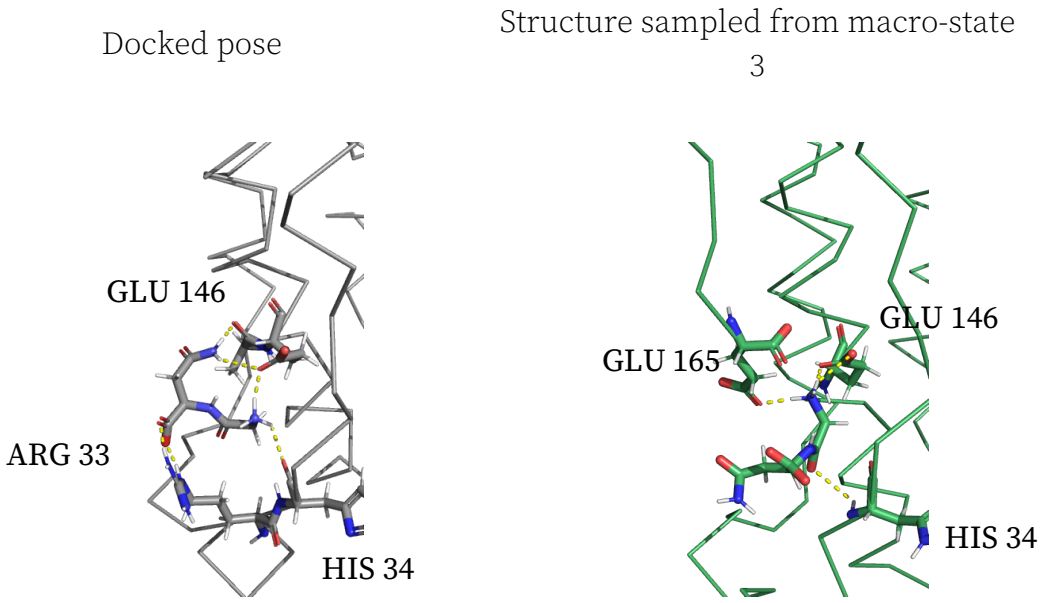


Figure 5.3: Comparison of docked pose with the most similar structure of those sampled from macro-state 3. Interacting residues are represented as sticks.

The Markov model exposes the binding path of the dipeptide, which most frequently transitions from macro-state 5 to 4, occasionally passing through state 3, which acts as an intermediate. The very infrequently occupied states 0, 1 and 2 are all connected to state 3 and are occasionally visited before the dipeptide moves along to states 3 and 4. The predicted residence time is calculated to be 0.03 µs and the equilibrium dissociation constant shows a weak binding of 29 mM compared to the µM affinity observed experimentally (Table 5.1, Figure 5.4)[179].

Table 5.1: Observables derived from the Markov model and experimentally observed dissociation constant for the interaction between IFN and Gly-D-Asp.

k_{on}	313 ± 201 $\mu M^{-1} s^{-1}$
k_{off}	30 ± 16 μs^{-1}
τ	0.03 ± 0.02 µs
K_D	29 ± 12 mM
K_D experimental	0.11 mM ± 0.02 mM

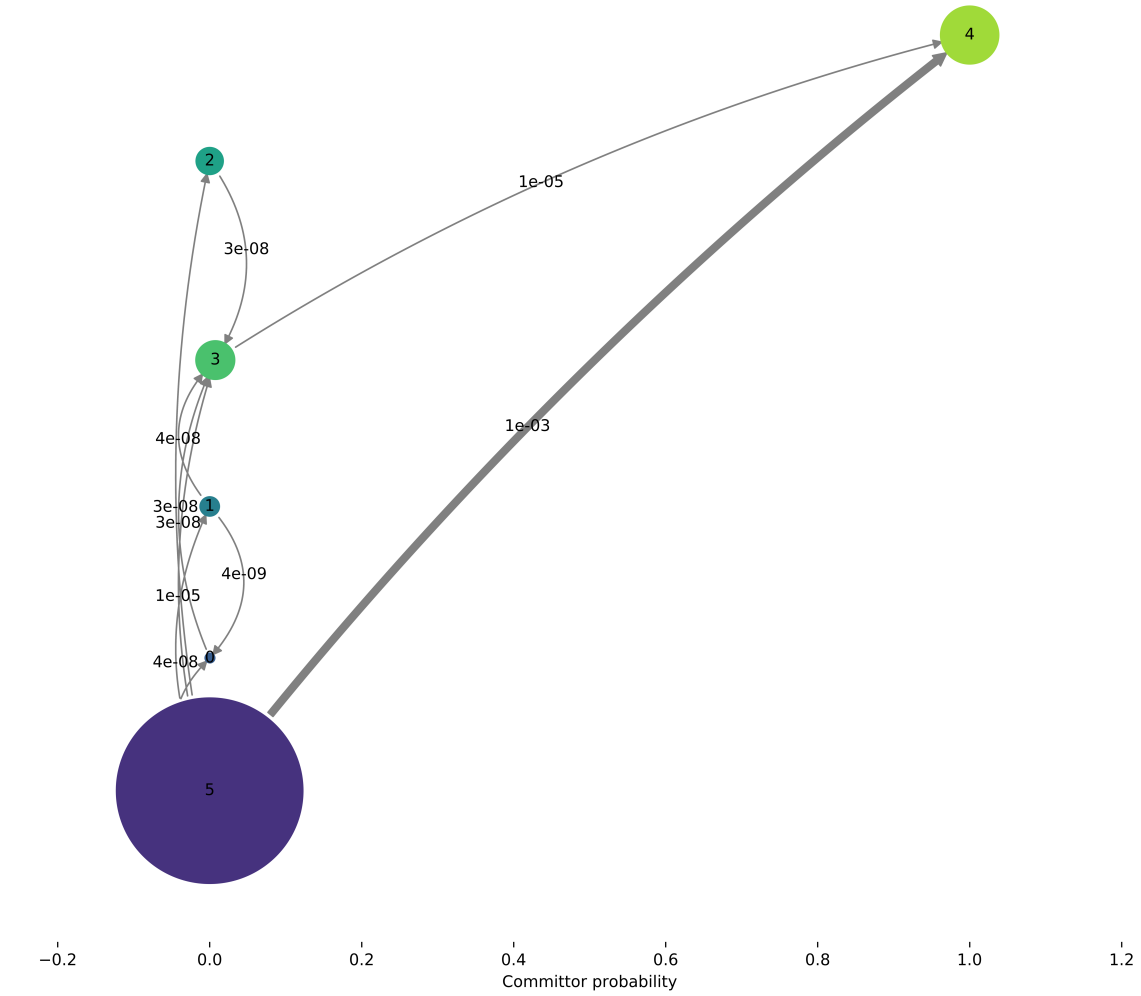


Figure 5.4: Markov processes can be visualized as a network of macro-states. Each circle represents a macro-state, which in our case corresponds to the ligand occupying a specific binding site (macro-states 0-4) or being unbound (macro-state 5). The areas of the circles are proportional to the stationary probability of the macro-state. Transitions between macro-states are visualized by arrows. Their thickness represents the probability of the transition to occur. The transition probability is also written on top of the arrows. The committor probability describes how likely it is that the system changes to the target state 4 (sink), or to the original state 5 (source). If the committor probability is close to 1, the system will move towards the sink. If it is close to 0, the system will move towards the source. One can therefore conclude that when the ligand is bound to the protein in one of the four macro-states, it will most likely unbind (i.e. transition to macro-state 5) before occupying another bound macro-state. The macro-state are colored consistently with Figure 5.2.

5.5. Discussion

Here, we use Markov theory for the first time to describe the interaction between a stabilizing small molecule and a therapeutic protein. The use of molecular dynamics simulations to study the interaction had two purposes. On the one hand, we wanted to identify the excipient's favored interaction sites and compare it to the protein's aggregation prone regions. On the other hand, we wanted to estimate the residence time of the protein-ligand complex to rule out any effect of the excipient on the drug protein's efficacy after administration.

We identified five meta-stable interaction sites showing hydrogen bonding and salt-bridges between the protein and the dipeptide, supporting the finding of stoichiometric binding between thee protein and the ligand. The protein-ligand complex formed in macro-state 3 is similar to the one that was proposed by our previously reported virtual screen[179]. In our Markov model, the macro-state 3 is, however, only a weakly populated intermediate state. Despite substantial sampling, we were not able to reproduce the experimentally observed dissociation constant. We can consequently conclude, that the simulations do not elucidate the interaction process in its entirety.

We find that in all 5 bound macro-states, the binding site is in proximity to at least one aggregation prone region. Considering the overall hydrophobicity of Interferon-alpha-2A and the implied presence of multiple of such aggregation prone regions, it seems difficult to consider this observation to be significant, since almost any binding site is likely to be close to an aggregation prone region. Therefore, the simulations on the one hand support our hypothesis of stabilization by stoichiometric binding, on the other hand it neither proves nor

disproves that the proximity to an aggregation prone region is the cause for the stabilization. Obtaining a crystal structure of the protein-ligand complex would be highly desirable to further evaluate the model.

The residence time estimated by our model is extremely low, indicating that there is no threat to an altered efficacy caused by a specific protein-excipient interaction since the complex will rapidly disassemble after administration. Since diffusion and distribution of small molecules is of orders of magnitudes faster than that of proteins, equilibrium conditions after administration are no longer given. Considering the underestimation of the dissociation constant, a higher residence time than the one calculated could nevertheless be plausible.

5.6. Conclusion

We studied the interaction between the stoichiometric stabilizer glycyl-D-asparagine and Interferon-alpha-2A through the construction of a Markov state model from MD simulations. The binding mechanism is complex and involves interaction sites in proximity to aggregation prone regions. The calculated residence time is of 0.03 μs and does therefore emphasize the improbability of a distorted efficacy of the drug protein caused by a stoichiometric stabilizer.

5.7. Supplementary Data

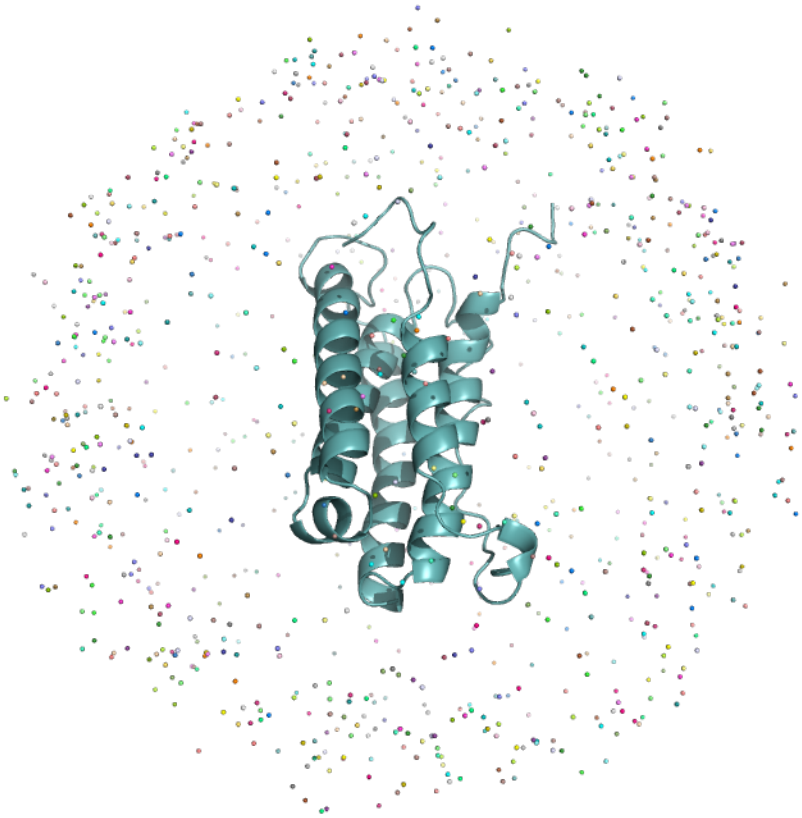


Figure S 5.1: Overlay of the position of the dipeptide in the starting structure for all 1000 simulations.

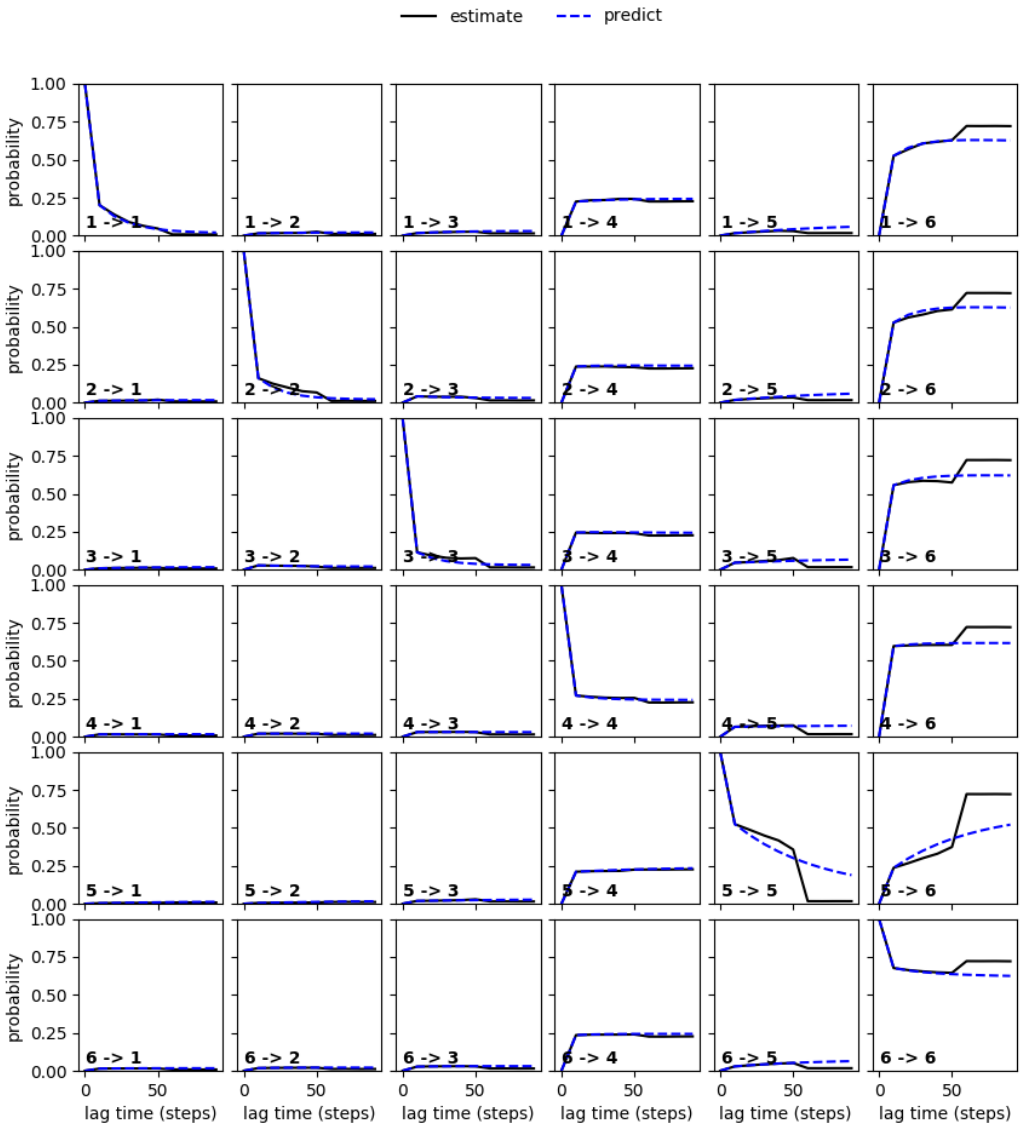


Figure S 5.2: Chapman-Kolmogorov (CK) test to confirm markovianity of the constructed model. The CK test reveals that the markovianity of the model is given for 50 time steps.

6. Exploring Chemical Space for new Substances to stabilize a therapeutic Monoclonal Antibody

A version of this chapter has been published in the Journal of Pharmaceutical Sciences:

Tosstorff, A., Menzen, T. & Winter G. Exploring Chemical Space for new Substances to stabilize a therapeutic Monoclonal Antibody, *J. Pharm. Sci.* **109**, 301-307 (2020).

The manuscript was written by Andreas Tosstorff. Tim Menzen provided scientific advice and reviewed the manuscript. nanoDSF studies using a robotic autosampler were performed by Silvia Würtenberger at NanoTemper Technologies. We are grateful to NanoTemper Technologies for providing measurement time and access to the data analysis software PR.Stability. All other experiments and data analysis were performed by Andreas Tosstorff under the supervision of Gerhard Winter.

6.1. Abstract

The physical stability of therapeutic proteins is a major concern in the development of liquid protein formulations. The number of degrees of freedom to control a given protein's stability is limited to pH, ionic strength and type and concentration of excipient. There are only very few, mostly similar excipients currently in use, restricted to the list of substances generally recognized as safe for human use by the FDA. Opposed to this limited number of available excipients, there is the vastness of chemical space which is hypothesized to consist of 10^{60} compounds. Its potential to stabilize proteins has never been explored systematically in the context of the formulation of therapeutic proteins. Here we present a screening strategy to discover new excipients to further improve an already stable formulation of a therapeutic antibody. The data are used to build a predictive model that evaluates the stabilizing potential of small molecules. We argue that prior to worrying about the hurdles of toxicity and approval of novel excipient candidates, it is mandatory to assess the actual potential hidden in the chemical space.

Keywords

mAb, excipient, protein stability, nanoDSF, DSF, chemoinformatics

6.2. Introduction

Formulation of therapeutic proteins is a field of ongoing research as the proteins can degrade in multiple ways. The process of identifying a suitable formulation occurs typically by screening solution conditions that vary by pH and ionic strength[58]. Additionally, stabilizing substances, so called excipients

are added. These can be categorized as for example surfactants, buffers, amino acids, polymers, proteins, metal ions, tonicity modifiers, sugars and polyols, salts, preservatives, antioxidants, chelators, antimicrobials. A recent review mentions 57 different substances[197]. Examples include polysorbates, polyethylene glycols, several sugars, several proteogenic amino acids or cyclodextrin[107,198,199]. The chemical space of molecules consisting of up to 30 carbon, oxygen, nitrogen or sulfur atom has been estimated to contain 10^{60} different molecules[200]. Taking into consideration that many of the aforementioned excipients are structurally very similar, the portion of the chemical space covered by currently employed excipients is next to nothing.

Hurdles in introducing new excipients to formulations of therapeutic proteins are the condition to have no pharmacological effect, the risk of their potential toxicity and the costly and time-consuming approval process, which for an excipient is as tedious as for a drug. Additionally, excipients have to be chemically stable and should have a sufficient aqueous solubility. Therefore, industry typically limits the arsenal of potential excipients during formulation development to the selection of excipients that the FDA generally recognizes as safe (GRAS list)[3,201]. However, there has been no systematic evaluation of possible benefits that may be introduced by new excipients. A better understanding of the potential of substances hidden in the chemical space to stabilize proteins could eventually provide a motivation to overcome the aforementioned hurdles.

Monoclonal antibodies (mAbs) represent the most important and best-selling class of therapeutic proteins in recent years[202]. A lot of research effort has been dedicated to optimize their sequences, in order to reduce the likelihood that

their development will negatively affect the outcome of any clinical trial[203]. One important strategy in sequence optimization consists in mutating aggregation prone regions[204]. When analyzing 28 therapeutic mAbs using Aggrescan3D[80], we found that aggregation prone moieties are present in the paratope for 20 of them (unpublished data). It seems plausible that sequence optimization is, among other factors, limited by the required affinity of the mAb to its target, often driven by hydrophobic patches in the mAb's complementarity-determining region. New excipients could therefore present a way to push the boundaries of current state formulations even with optimized protein sequences. This is desirable to achieve for example formulations that are stable at room temperature, making refrigeration and freeze-drying obsolete or to replace excipients such as polysorbates, which have a lot of critical attributes[205,206].

Besides their application in biopharmaceutical products, new excipients could easily be employed to stabilize proteins used for diagnostics or in bioprocesses, where their potential toxicity is less of a concern.

To identify excipient candidates, their effect on protein stability has to be evaluated experimentally. In long-term stability studies, formulations are stored for months or even years. The formation of aggregates and chemical changes in the formulation are monitored for example by chromatographic methods. Due to the limited throughput and time-constraints, this approach is not plausible for the purpose of screening a library of small molecules on their effect on protein stability. Instead, forced degradation studies have been developed as indicators of long-term protein stability. Differential scanning fluorimetry (DSF) measures changes in extrinsic fluorescence upon unfolding of a protein when exposed to heat. Similarly, in nanoDSF the measured changes

are of the intrinsic fluorescence of the protein's tryptophan and tyrosine residues. The inflection point (apparent protein melting temperature, T_m) of the characteristic unfolding curve serves as a surrogate to measure a protein's conformational stability. As extrinsic dye, SYPRO orange is one of the most common choices. The same method is also known as thermal shift assay in the drug discovery community, where it is used to identify new small molecular active compounds[207]. Light scattering, backscattering or optical density is often used simultaneously to monitor the formation of aggregates. The derived temperature of onset of aggregation (T_{agg}) is another common stability indicator. While DSF and nanoDSF are excellent choices regarding throughput and sample consumption, their correlation with long-time stability data is limited[155]. More recently, the ReFOLD assay has been proposed as stability indicating method, showing excellent correlation with long-term stability data[50,208]. In a first step, the protein is chemically denaturated by dialyzing against the formulation buffer containing Urea. Subsequently, the Urea is removed by dialyzing against the formulation buffer, leading to a refolding of the protein. During the process of Urea removal, the protein will be partially unfolded and not fully solubilized, making it prone to aggregate. The degree of aggregation measured for example by size exclusion chromatography can then be considered a surrogate for protein stability. As the ReFOLD assay relies on dialysis, it requires larger buffer volumes and has a lower throughput than for example DSF or nanoDSF measurements.

In this work we make use of chemoinformatic methods to classify and describe small molecule structures for multiple purposes. Very broadly speaking, there are two approaches to classify a small molecule in a machine-readable way. This is either through physicochemical descriptors, such as for example

hydrophobicity, or descriptors of structural features, such as the occurrence of a functional group. Both of these classification approaches have been implemented in a lot of different ways for numerous purposes. An excellent overview on the topic is given for example by Leach et al.[209]. One way to define hydrophobicity as physicochemical descriptor is the octanol/water partition coefficient of a substance (P). Numerous ways to predict P for a given small molecule exist[210]. Structural features of small molecules are commonly represented by binary vectors with multiple implementations. In one approach, each element of the vector corresponds to a predefined structural feature or key, as for example in the Molecular Access System keys (MACCS) method[211,212]. If for example the first MACCS key is present in the small molecule, its vector's first element will be set to 1. If the key is absent, the vector element is set to 0. In the case of so-called hashed fingerprints such as Morgan or Daylight fingerprints, the vector's elements do not directly correspond to a specific structural element. Instead they are calculated by an algorithm that considers connectivity or atom environment within a molecule.

The machine-readable description of a molecule can be exploited to build models that relate the descriptors to experimental observables, often referred to as quantitative structure activity relationship (QSAR). In QSAR, each physicochemical descriptor or vector element is considered a variable that can be fed to a machine learning algorithm in order to predict an unknown variable such as for example the biological activity of a small molecule[213]. Another example is the use of SYPRO Orange based DSF measurements of a mAb to build a QSAR model that predicts the effect of 79 osmolytes on the mAbs stability. The substances were similar to currently employed excipients, such as amino acids, methylamines and polyols[214].

Here we present an approach to identify small molecules that stabilize a mAb, starting from the selection of a suitable library by a chemoinformatic approach that focuses on compound diversity and hydrophobicity. We then screen the selected library by DSF and nanoDSF combined with backscattering to identify hit substances based on T_m and T_{agg}. After a hit expansion with analog substances we use the ReFOLD assay to identify excipient candidates and finally build a predictive QSAR model by using multiple regression.

6.3. Results

Library selection

Since there are only very few excipients commonly used in protein formulations, it is not possible to apply any general rules to the library design such as for example Lipinski's rule of five known from drug discovery[140]. We therefore opted to screen a library covering as much of the chemical space as possible. It was therefore required to be highly diverse. We quantified a library's diversity by considering its median pairwise Tanimoto coefficients calculated based on Morgan and RDkit daylight-like fingerprints. Limited lipophilicity was the only additional criterion imposed to assure sufficient solubilities. To keep time and cost of the first screening step reasonable, the library's size should be in the range of 1000-2000 compounds. Furthermore, we checked for the presence of pan-assay interfering substances (PAINS) and reducing sugars, which, however, were found to be very sparse in all cases, and thus not critical to decision making. The cost of the libraries was another key aspect since prices ranged from approx. 2000 € to 170000 € (Figure 6.1).

In total, we compared 19 different commercially available libraries from different vendors. Their median SlogP values ranged from approximately 1.5 to 3.5. Median similarities depended strongly on the type of descriptor used. The "Chemspace PPI Modulators library" (D) was found to be the least diverse and most hydrophobic library and fragment libraries from Enamine and Compound Cloud to be the most diverse and hydrophilic. Being the most cost-effective, we selected the "Enamine Golden Fragment Library" (Q). However, other selections would have also been plausible. The library consists of mostly aromatic scaffolds (Figure S 6.3), does not contain any reducing sugars and less than 1% of PAINS.

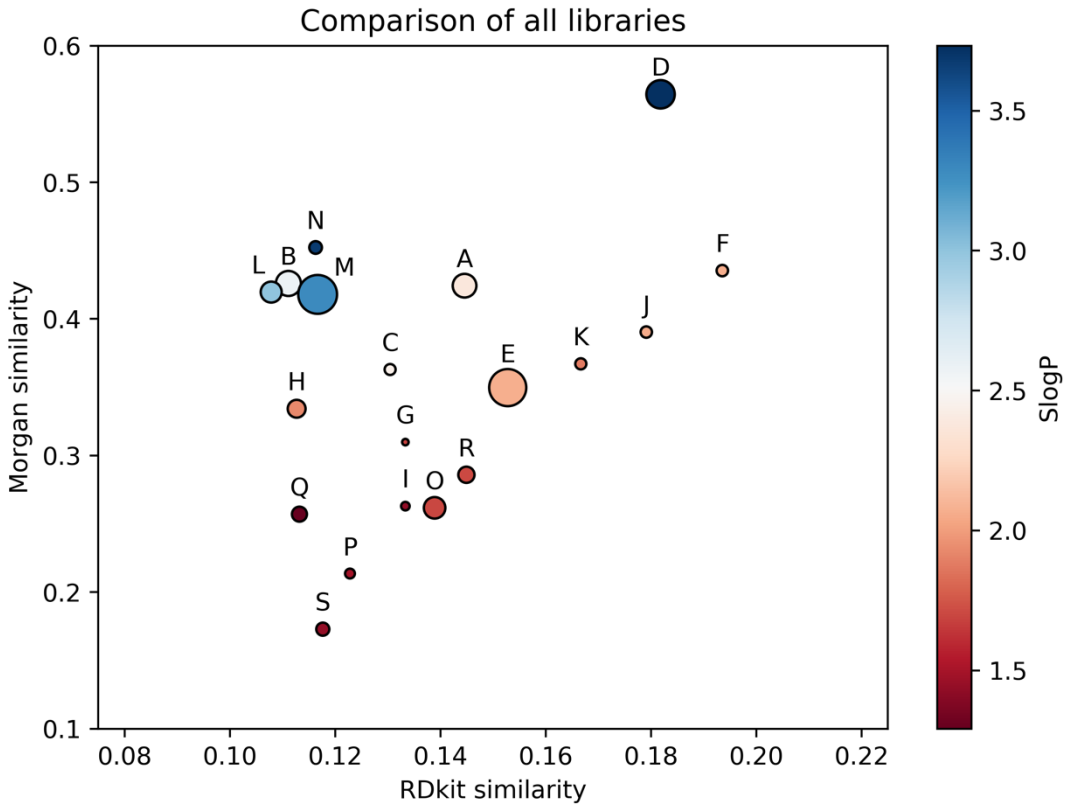


Figure 6.1: Comparison of commercially available libraries. Plotted is the median of the RDkit Tanimoto similarity vs. the median of the Morgan fingerprints Tanimoto similarity. The marker color indicates the SlogP value and the marker size corresponds to the size of the library. A: Chemspace Pre-Plated LeadLike set; B: Chemspace_Lead-Like Compounds 5000 diversity set; C: Chemspace Pre-Plated Fragment-like set; D: Chemspace PPI Modulators; E: Chemspace General Fragments; F: Chemspace Acid Fragments; G: Chemspace 3D-Shaped Fragments; H: Chemspace Singleton Fragments; I: Chemspace Selected Fragments; J: Chemspace Saturated Fragments; K: Chemspace Amine Fragments; L: Phenotypic Toolbox; M: BCCDIV14B; N: Tocriscreen; O: Enamine Cys focused covalent fragments; P: Enamine DSI poised fragment library; Q: Enamine Golden Fragment Library; R: Enamine Fluorinated Fragment Library; S: CompoundCloud Selcia. Size of library M: 12030 substances, library G: 337 substances.

Library screen

The change in thermal stability of protein induced by a small molecule, typically referred to as thermal shift, is commonly employed in drug discovery to identify active compounds. It is also an indicator of the stability of a protein in a given formulation. A shift towards higher temperature corresponds to a binding/interaction of the small molecule with the protein's native state[215,216]. Based on the same assumption that a stabilizing excipient also binds to the native state of the protein (or destabilizes the unfolded state), a positive shift is considered by us an indicator of a stabilizing protein formulation. By measuring the thermal shift of a therapeutic antibody (LMU-01) induced by all 1800 substances from the Enamine GFL we combined the rational from drug discovery and protein formulation screening (Figure 6.2). The stability of a given protein can be optimized easily and at low cost by adjusting pH and ionic strength. The use of excipients is therefore only meaningful, once these basic formulation properties have been optimized. We therefore selected an already optimized starting formulation for our excipient screen. Since our screening methods rely on temperature gradients, we limited the buffer choice to phosphate, as its pH has a low susceptibility to temperature[155]. The assay was performed at low protein concentrations to ensure an excess of small molecule, whose limited availability in the library during the screen was considered a bottleneck. The screen was performed in the following way: first all 1800 substances were tested by DSF and backscattering measurements. Hits from any of the measurements were then further evaluated by the ReFold method.

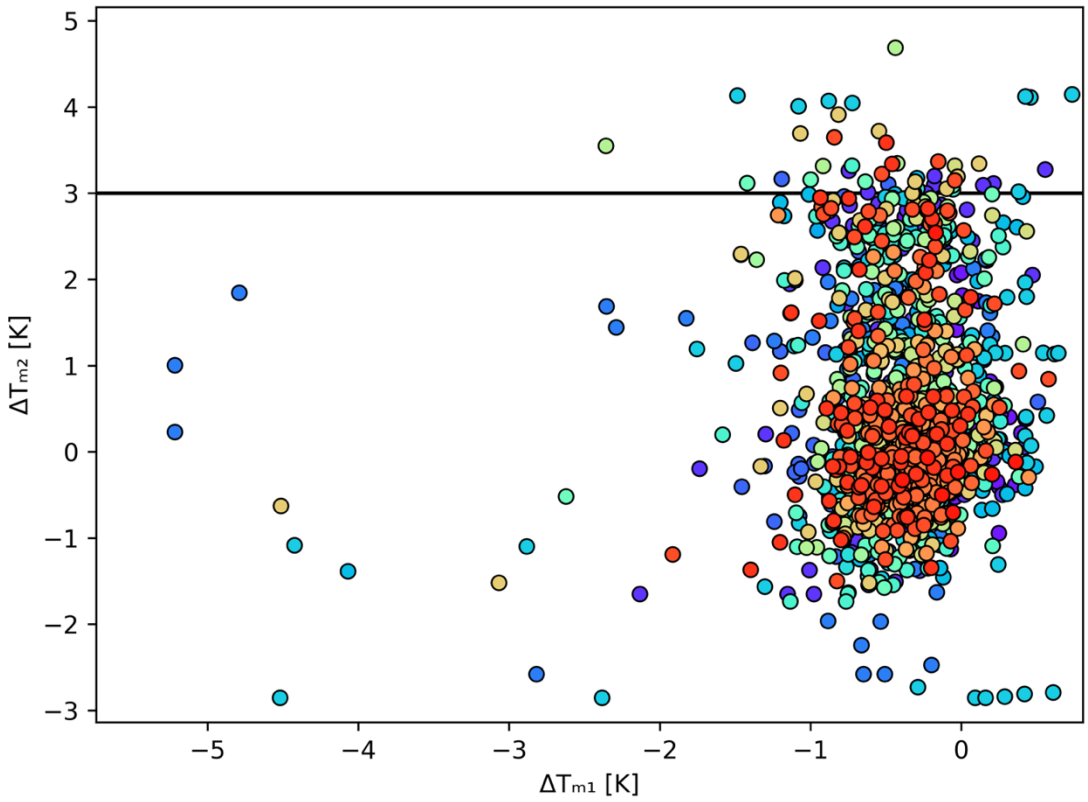
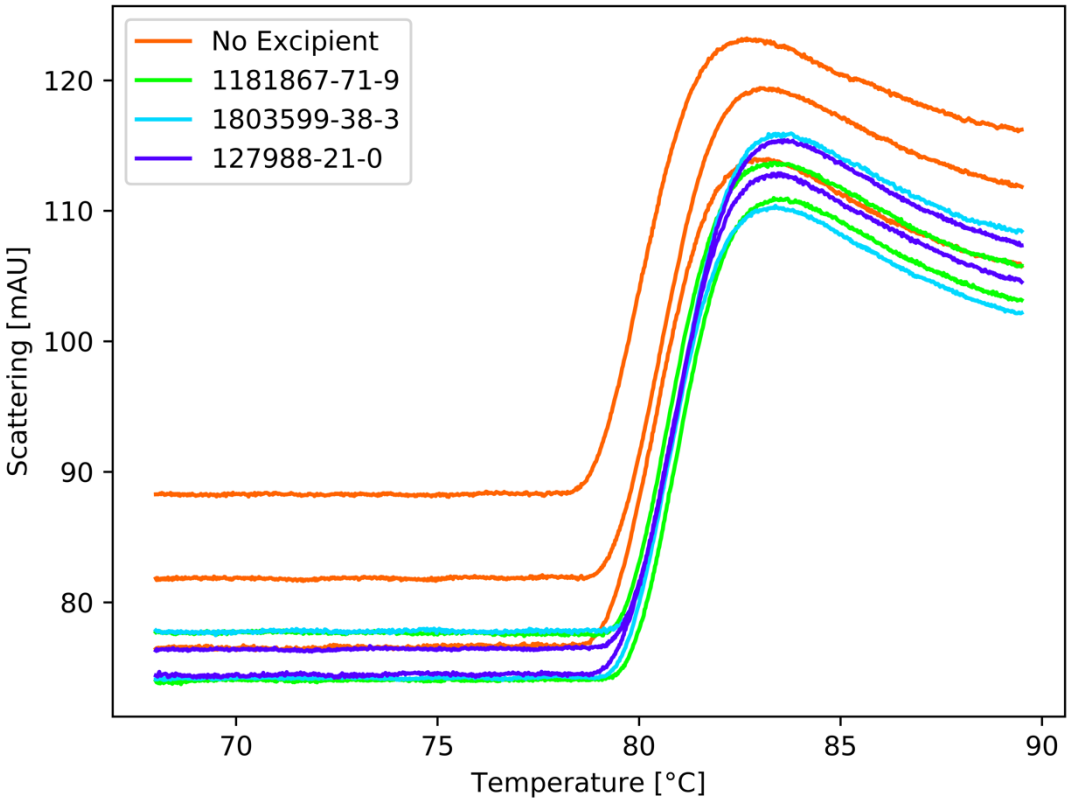


Figure 6.2: Thermal shifts relative to control samples from DSF measurements for all 1800 substances. Markers of the same color correspond to samples being on the same well plate.

For the DSF screen, Substances exceeding the threshold of 3 °C for ΔT_{m2} were considered for additional orthogonal screening. As 41 substances would exceed our capacities to measure in the ReFold assay, they were evaluated in an additional backscattering measurement by their effect on the onset of aggregation temperature T_{agg} compared to an excipient free control. Three substances exhibited a T_{agg} higher than that of all three control measurements (Figure 6.3). These were then considered for the refolding study.



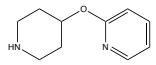
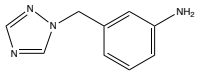
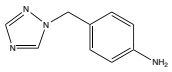
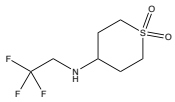
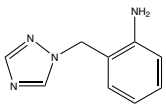
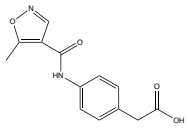
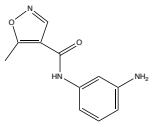
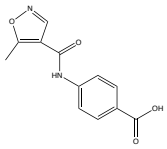
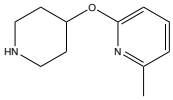
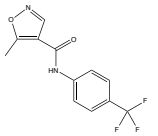
Formulation	T_{agg} [°C]		
No Excipient	78.7	78.1	78.5
1181867-71-9	79.2	79.1	-
1803599-38-3	79.2	79.0	-
127988-21-0	78.9	78.9	-

Figure 6.3: Scattering intensity from backreflection measurements and derived onset of aggregation temperature (T_{agg}) for top 3 candidate substances (n=2) and reference sample without excipient (n=3).

The backscattering screen yielded 10 substances with a T_{agg} higher than that of the control. Of these, only one substance, 380610-68-4, was affordable in price and selected for the ReFold study.

Three substances from the DSF screen and one substance from the backscattering screen and six analog substances were purchased for further evaluation in the ReFold assay (Table 6.1).

Table 6.1: Overview of candidate structures and their effect on the mAb in the ReFold assay.

CAS number	Structure	CAS number	Structure
127806-46-6		127988-22-1	
119192-10-8		1803599-38-3	
127988-21-0		953734-04-8	
1181867-71-9		67387-52-4	
380610-68-4		10170-12-06	

ReFold

The ReFold assay has previously been shown to accurately predict the long-term stability of various therapeutic mAb formulations. It is strictly orthogonal to the fluorescence-, light scattering- and temperature stress-based methods

employed in the first selection steps. It is therefore highly suitable to evaluate the candidate excipients and eliminate false positive results. Out of the 10 candidates (4 hits and 6 analogs) selected, we identified five that would increase the relative monomer area compared to the excipient free formulations and formulations containing the standard excipients sucrose, L-arginine or D(+)-trehalose. The substance 1803599-38-3 turned out to be a false positive (Figure 6.4).

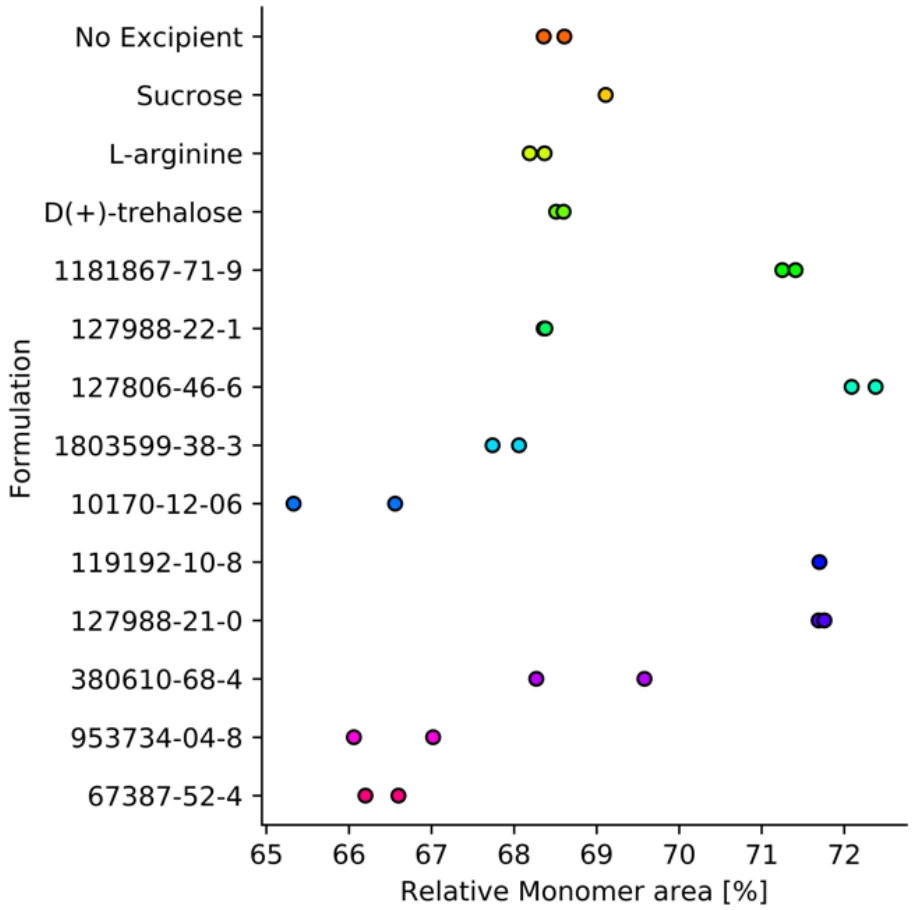


Figure 6.4: Relative monomer area after ReFold assay for formulations containing the candidate excipients, benchmark excipients and for an excipient free reference formulation (n=2).

Four out of the five stabilizing compounds show a clear interaction with the protein upon unfolding as can be seen in nanoDSF measurements (Figure 6.5). Control experiments show that the change in curve shapes are not caused by a temperature dependence of the small molecules' fluorescence signals (Figure S 6.1). A change in curve shape was also observed for compound 127988-21-0 in the initial DSF screen, but not for compound 380610-68-4 (Figure S 6.2), for the other substances no DSF data are available since they are analogs purchased after the initial library screen.

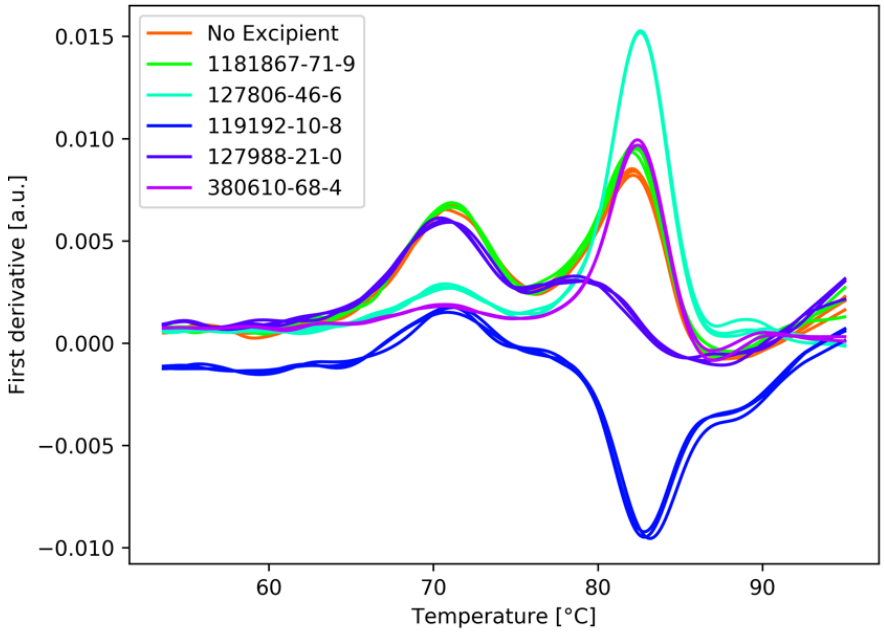


Figure 6.5: First derivative of nanoDSF data for all ReFold stabilizers. All compounds except 1181867-71-9 significantly alter the shape of the curve in the transition region (n=3).

QSAR

The data from the ReFold assay was used to evaluate the effect of structural features of a small molecule on the relative monomer area by constructing a model through multiple regression. The model is built from 8 MACCS keys and achieves an R^2 of 0.49 and RMSE of 2.13 (Figure 6.6). We found that structures containing MACCS keys 89 and 157 would lead to a decreased relative monomer area, while substances containing MACCS keys 91, 100, 117, 131, 132, 150 would increase the relative monomer area of the ReFold assay (Table 6.2).

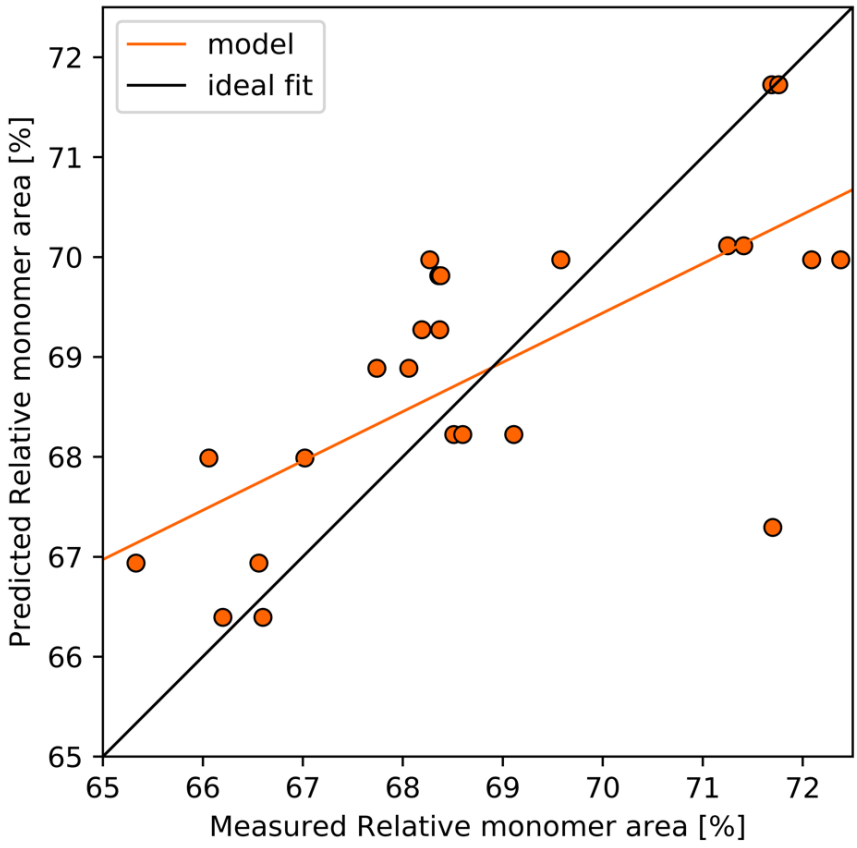
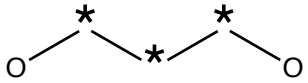
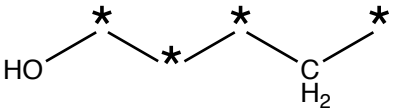
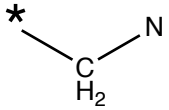
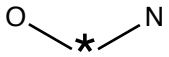
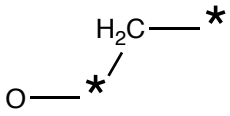
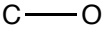


Figure 6.6: Multiple regression model to predict the effect of a small molecule on the relative monomer area determined by the ReFold assay. R^2=0.49, RMSE=2.13. MACCS keys used for the model: 89, 91, 100, 117, 131, 132, 150, 157.

*Table 6.2: Visualization and regression coefficient of MACCS keys used to build a regression model for the ReFold assay. * represent a wildcard. Unless specified, all bond representations are wildcards*

MACCS key	Structural feature	Regression coefficient
89		-3.53
91		2.52
100		3.53
117		2.28
131	OH > 1	3.17
132		1.59
150	 any atom – non ring bond - any atom - ring bond – any atom – non ring bond – any atom	4.43
157	 single bond	-3.72

6.4. Discussion

The two criteria driving the library selection, diversity and hydrophilicity, allowed us to select a compound library covering a broad part of chemical space with substances with a reasonable solubility in aqueous formulations (Figure 6.1). The libraries considered in our analysis were all from commercial vendors and designed for the purpose of drug discovery. The selected "Golden Fragment Library" has been already used in a thermal shift screen to identify inhibitors of bromodomain-containing protein 4[217]. The advantages of selecting a commercial library are that the cost per amount of substance is lower and that the libraries are curated and tested. Ideally this avoids pitfalls like PAINS, reactive or unstable substances. Substances from commercial libraries are furthermore provided pre-dissolved in well plates, allowing for an easy transfer with standard multi-pipettes. Typically, the substances found in commercial libraries can also be obtained individually at a reasonable cost, together with analogs, which makes following up on any hit molecules straightforward.

As typically observed for mAbs, the temperature dependent fluorescence signal of LMU-01 shows two transitions (T_{m1} and T_{m2}). From measurements of backscattering of light as an indicator of aggregate formation, the second transition, corresponding to the unfolding of the Fab fragment, has been identified to induce particle formation. The point density from the DSF measurements is only 1/K, which results in a considerable level of noise. We therefore selected candidates for further exploration based on thermal shifts of T_{m2} above 3 °C (Figure 6.2).

The selected compounds were then evaluated by simultaneous nanoDSF and backscattering measurements, with backscattering being a truly complementary detection method to DSF to evaluate actual particle formation. The low working volumes did not allow for pH adjustment at this stage, inevitably leading to false positive and negative measurements, since shifts to lower pH typically increase the electrostatic repulsion among mAb molecules with pI values between 7-9[218]. Selecting a higher buffer concentration may be an approach to mitigate the risk of pH shifts, however, at the cost of increased ionic strength, altering the proteins reference stability profile. The presence of DMSO as standard solvent known from drug discovery screens was an additional source of error which we considered inevitable. For the last step of the screening we adjusted pH and worked in DMSO free conditions, leading to reduced solubilities of the candidate compounds and an altered protein stability profile. Additional false positive results could therefore be identified by using the ReFold assay (Figure 6.4, Table 6.1).

In order to screen the library for its effect on protein stability, we considered three different analytical methods. DSF (in the presence of SYPRO Orange), nanoDSF/backscattering and SLS (data not shown). By using two fluorescence-based methods, two different excitation and three emission wavelengths are covered. If a compound's fluorescence happens to interfere in one of the assays, this ensured that it would not interfere in the other one. DSF measurements could be performed at a high throughput due to its well plate-based format. The use of SYPRO Orange as extrinsic fluorescent dye allowed for a very sensitive monitoring of mAb unfolding based on the exposure of hydrophobic regions, buried inside the core of the protein's native conformation. Consequently, the presence of extrinsic dye may also interfere in the interaction between the

tested, partially hydrophobic substances and the protein. Furthermore, the low resolution of the measurement introduced a significant amount of noise. The lack of dedicated software to analyze the data, required the generation of our own script. In contrast, data from nanoDSF and simultaneous backscattering measurements had a vastly higher resolution than our DSF measurements and the provided software allowed for a straightforward way to handle the large amount of data. Since the capillary based system makes sampling loading a time-consuming drawback, a capillary-chip-based version of the instrument equipped with an automated sample loading device was used in this study. SLS/DLS measurements provide a sensitive way to detect the formation of protein aggregates in a well-plate format. Here, in order to prevent evaporation of the sample either silicon oil or adhesive films have to be used. Due to the hydrophobic nature of some of our substances, only the use of films was plausible for our case. While the method requires very low sample volumes, DLS measurement require long measurement times and are therefore a limitation to throughput. We therefore tested the use of scattering intensity (SLS) as a fast and sensitive readout to detect aggregate formation in isothermal conditions. Whereas this experiment would have presented a complementary approach to the DSF and nanoDSF experiments, it did not turn out to be sufficiently robust. Possible reasons could be the formation of air bubbles during the measurement and detachment of the adhesive film during the course of the measurement. Further optimization of the assay in terms of adhesive film selection and adhesion process was not feasible in the timeframe of this work. One could also consider this method as an intermediate screening step, where the number of candidates is already narrowed down and replicate measurements can be performed in a reasonable time frame.

After candidate selection through DSF, nanoDSF and backscattering measurements, we purchased the hit substances together with analog compounds. The use of analogs provides a way to identify the substructures responsible for the stabilizing effect and provides a mean to build a robust hypothesis.

The recent development of the ReFold assay presents a straightforward, orthogonal way to evaluate the hits. While its throughput is considerably lower and its buffer consumption higher than that of the other discussed methods, it requires only a minimum amount of handling, is highly parallelizable and relies on methods established in any protein analytics lab.

We observe that the candidates that positively affect the relative monomer area also change the nanoDSF curve shape (Figure 6.5). The altered shape of the nanoDSF curves could indicate an interaction between the stabilizers and the (partially) unfolded species or a change in the unfolding mechanism, a bias that is not observed with the ReFold assay. A change in the nanoDSF curve shape could be considered an alternative principle for the selection of excipient candidates from nanoDSF screens.

In the ReFold assay, we find that several of the candidate compounds outperform the standard excipients arginine, trehalose and sucrose for the given concentration of 5 mM, which indicates a stabilization through stoichiometric binding (Figure 6.4). For a full benchmarking of the candidate compounds, a comparison with these substances at higher concentrations, resulting in preferential exclusion as dominating stabilization mechanism, is of interest. One could furthermore investigate the effect of combining the

excipient candidates at low concentration with preferential exclusion stabilizers at high concentration. This would combine two complementary stabilization mechanisms[57].

Predicting the effect of a small molecule on protein stability would be highly desirable to facilitate the discovery of new excipients. Through multiple regression, a model was constructed from the ReFold data using MACCS keys as input features to predict the effect of a substance on the assay (Figure 6.6, Table 6.2). Even though it was cross validated by the leave-one-out method, its predictive power, is of course limited to the design space. Nevertheless, it can be considered a starting point for more sophisticated models for novel stabilizing substances, as already known from drug discovery. More, high quality input data will enable the construction of more general models. While we also considered the DSF and nanoDSF screening data for model generation, we found that the signal to noise ratio was not sufficient to construct meaningful models. Algorithms other than multiple regression were tested but led to overfitting, meaning that they would also fit to the noise in the data.

In this work, we purposely left out toxicity as a factor in excipient selection, but instead we considered it the main purpose to explore the vast potential of chemical space for protein stabilization against non-native aggregation. As known from drug discovery, toxicity adds another degree of complexity to the endeavor of identifying new substances. We suggest that this factor should be accounted for in the candidate optimization stage by eliminating any entities responsible for toxicity from the structure[219]. Additional factors to be considered in the optimization stage are solubility, metabolism and the stability of the candidate substance itself. Compatibility with buffers other than phosphate is

an additional aspect to be taken into consideration. To fully assess the effect of an excipient on protein stability, long term stability and additional forced degradation studies paired with analytics covering all aspects of protein stability are necessary. It is apparent that excipient discovery is a multi-objective problem that engulfs all of the aforementioned requirements. In this study we focused on generating a single objective QSAR model to demonstrate the concept of combining modern screening methods with chemoinformatics. A multi-objective approach requires additional data generation which is only achievable in a highly automated laboratory. Conceptionally, however, the presented approach would require only slight adaptations.

6.5. Conclusion

In order to assess the potential of substances hidden in the chemical space beyond the GRAS list to stabilize a protein, we rationally selected a compound library by its lipophilicity and diversity. We screened the library to select stabilizing candidate substances for a mAb using two different, complementary, standard stability indicating methods. Both DSF and nanoDSF resulted in different hits. Subsequently, the hit substances and analogs thereof were evaluated by the ReFOLD assay, based on chemical denaturation and thus using a different physicochemical principle than the thermal screenings. This led to the identification of multiple substances outperforming standard excipients and the excipient free formulation. The candidate excipients can be developed and investigated further, for example in accelerated and long-term stability studies and additional forced degradation experiments. The stability of the excipient candidates themselves has to be tested as well as their toxicity. They could also be further optimized by structural modifications. The data was also used to

generate a MACCS keys-based model that can predict a substance's effect on the ReFold assay. The model can be used to rapidly evaluate a novel substances effect and help to identify additional compounds for further studies. Combining high-throughput screening of the chemical space with QSAR modeling enables therefore the generation of formulations with novel excipients that outperform those containing established GRAS list excipients.

6.6. Methods

Library selection

In order to select an appropriate compound library for screening, several commercially available libraries were analyzed. A KNIME workflow was set up using RDkit nodes to desalt the structure files, calculate SlogP values as a measure of solubility and a similarity matrix by querying individual entries from a library against their entire library (Figure S-1). The median values for each property was calculated using NumPy (version 1.16.2) and plotted using Matplotlib (version 3.0.3).

Sample preparation

The Enamine Golden Fragment Library was shipped in 29x 96 well plates containing stock 20 µl of 50 mM small molecule dissolved dimethyl sulfoxide (DMSO, Sigma-Aldrich). 250 µM stock solutions of small molecules were prepared in 96 well plates (Greiner Bio-One GmbH) with 50 mM sodium phosphate buffer at pH 6.0 (di-Sodium hydrogen phosphate dihydrate: VWR Chemicals, Sodium dihydrogen phosphate dihydrate: Grüssing GmbH).

Differential scanning fluorimetry

LMU-01 solutions containing SYPRO orange were prepared by adding 2 µl of SYPRO Orange stock solution to 5 ml 1 mg/ml LMU-01 stock solution. The solution was prepared daily. The apparent protein melting temperature (T_{m1} and T_{m2}) was measured with the a qTower 2.2 (Analytik Jena) in 96 well plates. Final working concentrations were 0.5 mg/ml LMU-01, 1:5000 SYPRO orange, 125 µM ligand, 0.25% DMSO in 50 mM sodium phosphate buffer at pH 6.0. The data was analyzed by calculating the unfolding curves' first derivative by using a Savitzky-Golay filter as implemented in the SciPy library[220]. The first derivative curve was fitted to a skewed gaussian by using the LMFIT module for Python[221].

Backreflection library screen

T_{agg}, were measured with the Prometheus NT.Plex, equipped with backreflection optics, in standard capillary chips (NanoTemper). Final working concentrations were 0.5 mg/ml LMU-01, 125 µM ligand, 0.25% DMSO in 50 mM sodium phosphate buffer at pH 6.0. Automated sample loading into capillary chips was performed with an NT.Robotic Autosampler (NanoTemper).

nanoDSF hit confirmation

T_{agg}, T_{m1} and T_{m2} were measured with the Prometheus NT.48, equipped with backreflection optics, in standard capillaries (Nano Temper). Final working concentrations were 0.5 mg/ml LMU-01, 2 mM ligand, 4% DMSO in 50 mM sodium phosphate buffer at pH 6.0.

ReFold assay

The ReFold assay was adapted from Svilenov et al.[50]. The refolding buffer was prepared by adding a stock solution of 50 mM sodium phosphate buffer at pH 6.0 to excipient candidate substances to yield 5 mM solutions thereof. In cases where the solubility limit was exceeded, the saturated solution was used. The same procedure was used for the unfolding buffer which contained additional 10 M of urea. pH values were adjusted to the excipient free reference buffer. The resulting buffers were centrifuged at 15000 rpm. Protein solutions were prepared by spiking 3 µl of LMU-01 stock solution to 237 µl of refolding buffer, yielding a protein concentration of 1 mg/ml. Duplicates of 100 µl of protein sample were transferred into micro-dialysis tubes with a 3.5 kDa cutoff. Dialysis was performed at room temperature and unfolding buffer was exchanged after 3 h and 7 h. Refolding commenced after 24 h with buffer exchanges after 3 h and 7 h.

QSAR

MACCS keys fingerprints of the substances tested in the ReFold assay were built using the Conda distribution of RDkit (version 02-2019). Low variance keys were eliminated. Of the remaining features, those with regression coefficients close to zero were removed to rule out overfitting and obtain a robust model using only 8 MACCS keys. Multiple regression using leave-one-out cross validation was performed using Scikit learn (version 0.20.3).

6.7. Supplementary Data

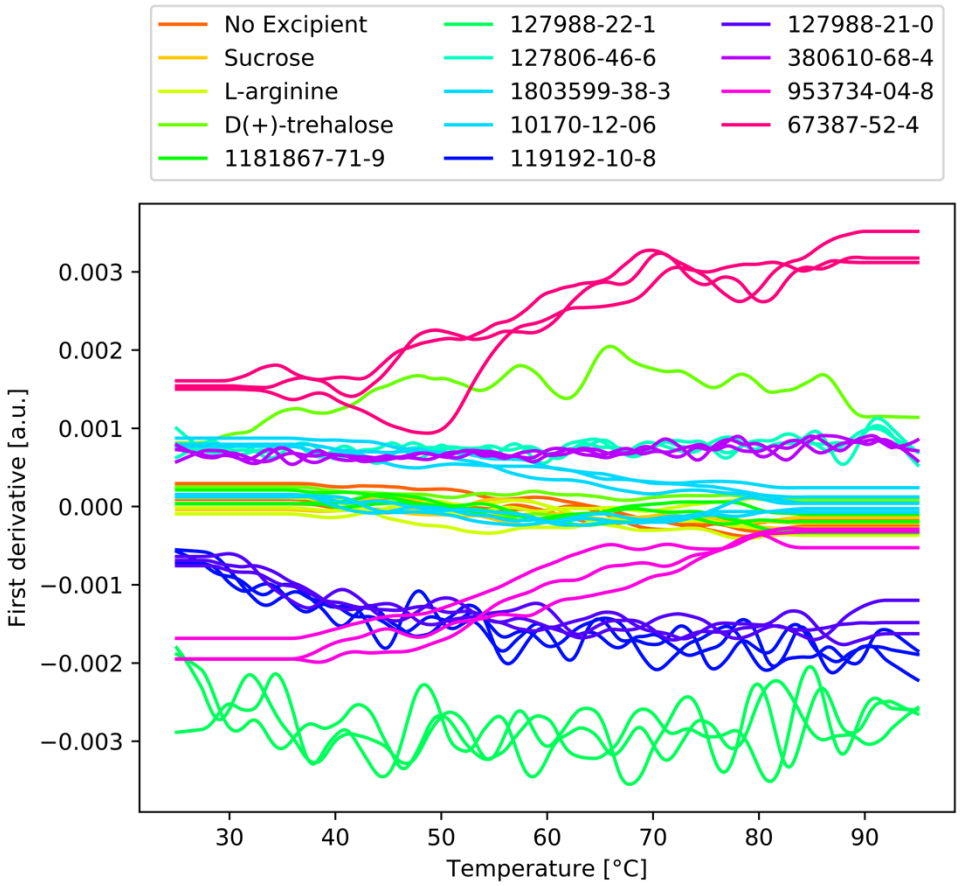


Figure S 6.1: First derivative of Temperature dependent fluorescence signal from nanoDSF measurements for protein free control samples. The 350 nm/330 nm fluorescence signal of the tested small molecules shows a neglectable temperature dependence.

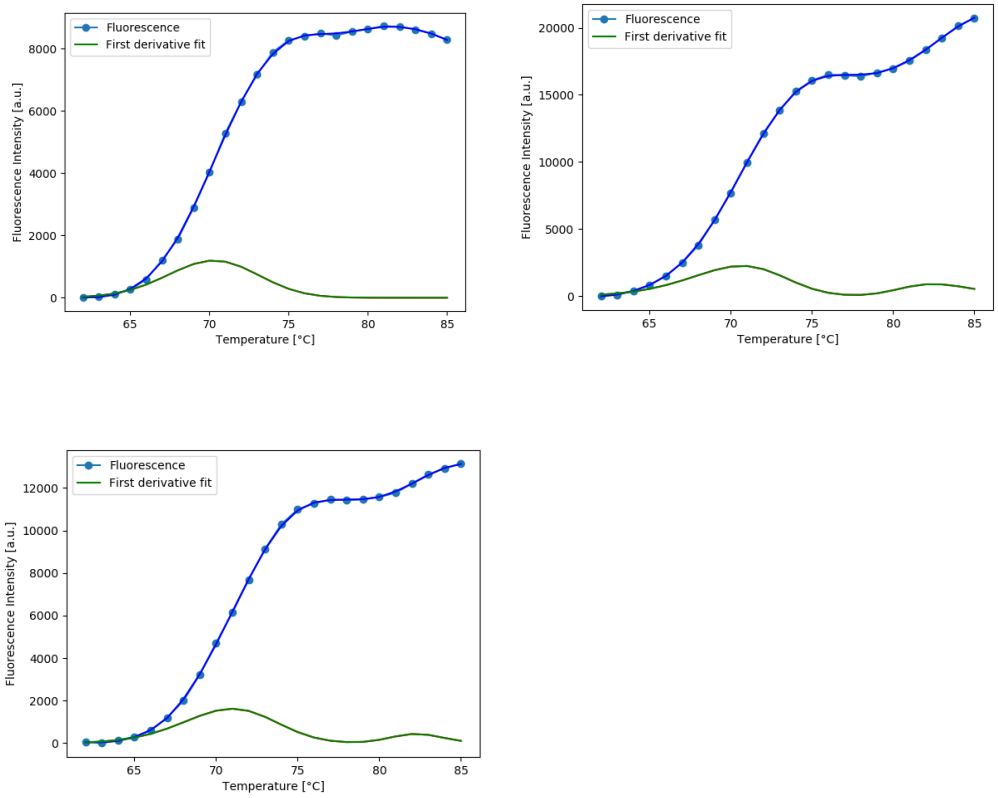


Figure S 6.2: DSF data for compounds 127988-21-0 (top left), 380610-68-4 (top right), excipient free control (bottom)

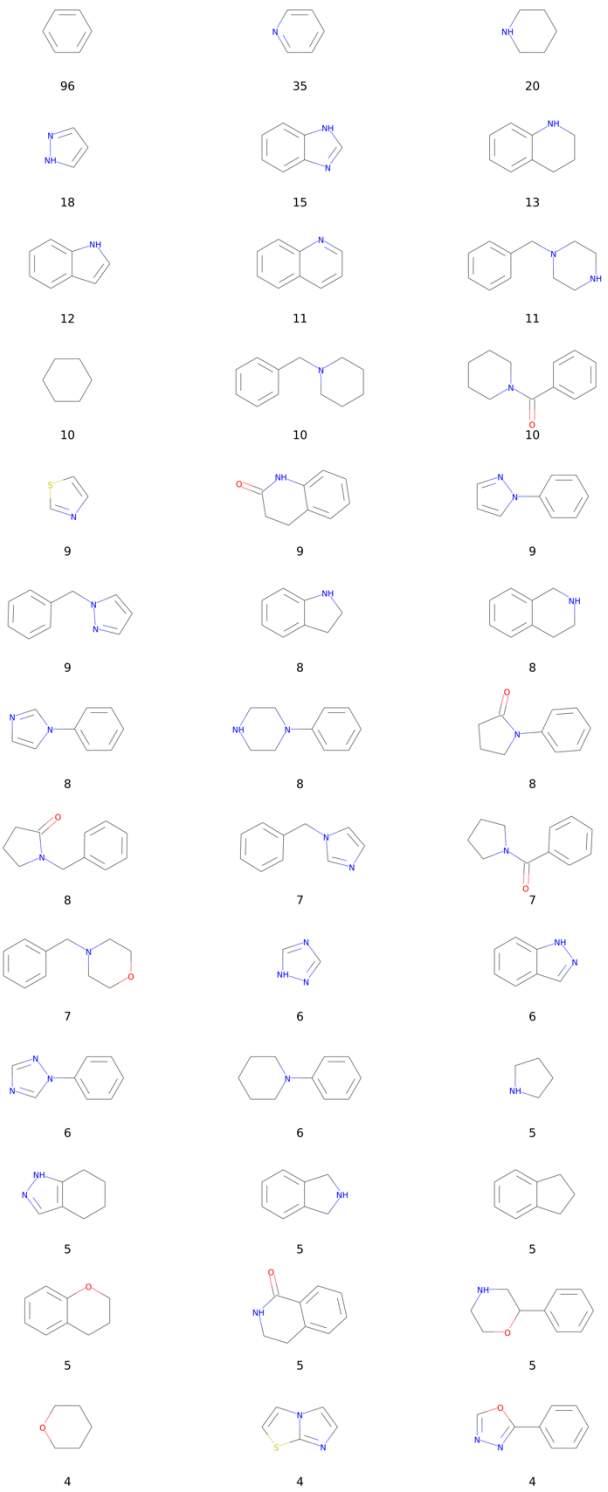


Figure S 6.3: Most common scaffolds in the Enamine "Golden Fragment Library"

7. Summary of the thesis

In this work, we give an overview of concepts, methods and applications that facilitate the discovery of new protein-stabilizing small molecules. The first, introductory chapter, describes non-covalent interactions, that are the basis for protein aggregation and protein ligand interactions. Mechanisms that lead to the aggregation of proteins are introduced and the effect of small molecules on physical protein stability is discussed. Since the problem of excipient discovery is similar to that of small molecule drug discovery, central concepts of the latter are illustrated and where applicable, parallels to the excipient discovery process are highlighted.

The second chapter describes the discovery of a new excipient candidate. The aggregation prone regions on the model protein IFN are identified using Aggrescan3D. By virtually screening the ZINC database's millions of compounds in terms of hydrophobicity, solubility and affinity towards the aggregation prone region, a set of candidate compounds was selected for experimental screening. Of the candidate compounds, the dipeptide glycyl-D-asparagine was identified to bind to IFN with a micromolar affinity. In agitation and freeze-thaw

degradation studies, the compounds stabilizing effect on IFN is demonstrated. Furthermore, evidence is provided, that the compounds stabilizing effect is due to stoichiometric binding, as it outperforms its enantiomeric counterpart. The compound also shows better stabilization than the reference excipients trehalose and L-arginine.

In the third chapter, a novel approach to validate aggregation prone regions is presented. By using solution paramagnetic enhancement NMR, residues that are buried in the oligomerization interface of the model protein Interferon-alpha-2a are identified. The hence identified regions overlap or are in close proximity to regions identified by three computational methods to predict aggregation prone regions.

In chapter four, the interaction between IFN and glycyl-D-asparagine is studied in detail by constructing a Markov state model from molecular dynamics trajectories. The model allows to calculate physical observables such as the free energy of binding or the rate constant of unbinding for the protein-excipient complex. One property of particular interest is the residence time, which is a measure of the lifetime of the formed complex. If the protein-excipient complex exhibits a long lifetime after drug administration, the drug protein's efficacy might be altered. The calculated residence time indicates that the complex is short-lived and the excipient poses no threat to the drug's efficacy. The model consists of six different macro-states, of which five correspond to the bound protein-ligand complex with different binding sites occupied. One macro-state overlaps with the structure obtained through the virtual screen that initially resulted in the discovery of the substance as described in chapter 2. The results therefore further support the hypothesis of stabilization through stoichiometric

binding, but show the presence of multiple binding sites. It is not possible to attribute the stabilizing effect to any binding site in particular nor to the proximity of a binding site to an aggregation prone region.

The chapters two to four are centered on a structure-based hypothesis which was successful in obtaining a stabilizer against surface-stress. However, it requires access to methods and expertise of structural biology that is atypical to conventional protein formulation laboratories. The presented strategy furthermore did not yield a substance that would enhance the proteins stability in accelerated stability studies at elevated temperatures. Consequently, the fifth chapter presents an alternative approach to excipient discovery, this time using a monoclonal antibody as target, that parts from a high-throughput screen. Hits are identified by nanoDSF and backscattering measurements and validated by the novel ReFOLD assay. The experimental screen yields a number of substances that outperform standard excipients at concentrations of 5 mM or below, indicating stabilization through stoichiometric binding. By correlating the experimental data with the structural MACCS keys descriptors through multiple regression analysis, a QSAR model is generated that can predict the effect of a small molecule on the relative monomer area of the ReFOLD assay only from the molecule's structure.

8. Discussion and Outlook

Due to the layered nature of this work, a discussion on the overall strategy, not only on the separate chapters seems appropriate. Despite the successful identification of multiple stabilizing compounds, certain decisions should be reconsidered in the light of the obtained results. An outlook on appropriate further studies is given.

For the purpose of identifying a novel excipient that stabilizes through stoichiometric binding, selection of the model protein is the first crucial step. Considerations that led to the choice of model proteins in this project included the availability of protein structure and availability of the protein substance. Information on mechanism of aggregation and “ligandability” of the protein were not considered. Therapeutic enzymes for example could be an interesting target for stoichiometric stabilizers as one can expect them to have a defined binding cavity and ligands may already be known.

Along with the choice of the model protein comes the identification of the purpose of the excipient. Defining the desired formulation profile beforehand can be helpful to design an appropriate screening strategy. One could for

example define the development of a surfactant free formulation as target formulation. The resulting formulation screen could then use surfactant containing formulations as benchmark and shaking or freezing stresses would be the first line stress studies in hit identification. Alternatively, if the new excipient should enable for example a formulation stable at room temperature, long-term stability indicating methods such as the ReFOLD assay could serve as first line assay. When opting for a structure-based approach for the identification of a stoichiometric stabilizer, defining the purpose of the excipient is important, due to the dependence of oligomerization interface formation on the type of applied stress. Thermal stabilization may require a binding site different to that required for interfacial stabilization.

Another impactful decision that occurred early on during the IFN project consisted in buffer pH and ionic strength selection. The main rational was to select an unstable or "bad" formulation, that would make any excipient induced improvement in stability visible easily. This decision is flawed for numerous reasons. For once, developing a novel excipient is of highest value when it outperforms the best available formulation. A good excipient in a bad formulation buffer is still likely to produce a bad formulation. Moreover, in the case of IFN, the selected buffer induces self-association due to low electrostatic repulsion. This made NMR studies at the corresponding pH impossible, which in turn hindered proving the binding hypothesis.

Molecular dynamics simulations were one of the central methods employed to study protein-excipient interactions. While initially thought of as a validation tool as part of a virtual screen by the APR-US approach, we found that simulation times to reach convergence and simulation setup times are unreasonable

compared to for example a nanoDSF experiment. While obtaining information with atomistic resolution is an attractive outcome, the results have to be considered with care. Ligand force fields are often trained on limited data (ca. 70 molecules for GAFF2) and quantum-mechanic calculations in vacuo will neither represent the solution environment of the free ligand nor the protein environment of the bound ligand. One of the most common application of MD to study protein-ligand interactions are free energy perturbation, which is successful because it relies on a known protein-ligand complex structure (including protonation states) and compares the effect of only small alterations to the ligand structure. Applications of MD that coincide with experimental values are often retrospective and anecdotical, but rarely prospective and broadly applicable, as their setup requires a lot of forehand knowledge on the system of interest.

Future work regarding the glycyl-D-asparagine-IFN complex should consist in the elucidation of the structure by X-ray crystallography as ultimate proof of the binding site. Measuring the off-binding rate by surface plasmon resonance would present a way to confirm the calculated residence time.

The substances identified to stabilize the mAb should be evaluated for example in long-term stability studies. An elucidation of their mechanism of action would be of great interest in order to design additional stabilizing molecules. Furthermore, extending the present data set would allow to further improve the QSAR model. At increasing structural diversity and complexity, alternative molecular descriptors and machine learning algorithms have to be considered for model generation. To ensure the safety of the discovered excipient

candidates, they need to be assessed in terms of their chemical stability and ADME-tox properties.

Finally, combining the stoichiometric stabilizers described in this work with conventional preferential exclusion stabilizers could present an interesting area of further studies, that combines two complementary mechanisms of stabilization.

References

1. Wang, W. Advanced protein formulations. *Protein Sci.* **24**, 1031–9 (2015).

2. Wang, W., Nema, S. & Teagarden, D. Protein aggregation-Pathways and influencing factors. *Int. J. Pharm.* **390**, 89–99 (2010).

3. Paulette M. Gaynor, Richard Bonnette, Edmundo Garcia, Jr., Linda S. Kahl, Luis G. Valerio, J. FDA's Approach to the GRAS Provision: A History of Processes. (2006). Available at: https://www.fda.gov/Food/IngredientsPackagingLabeling/GRAS/ucm094040.htm.

4. Ferreira De Freitas, R. & Schapira, M. A systematic analysis of atomic protein-ligand interactions in the PDB. *Medchemcomm* **8**, 1970–1981 (2017).

5. Timasheff, S. N. & Tsutomu Arakawa. Mechanism of Protein Precipitation and Stabilization by co-solvents. *J. Cryst. Growth* **90**, 39–46 (1988).

6. Israelachvili, J. N. Strong Intermolecular Forces: Covalent and Coulomb Interactions. in *Intermolecular and Surface Forces* 53–70 (Elsevier, 2011). doi:10.1016/B978-0-12-375182-9.10003-X

7. Van Oss, C. J., Chaudhury, M. K. & Good, R. J. Interfacial Lifshitz-van der Waals and polar interactions in macroscopic systems. *Chem. Rev.* **88**, 927–941 (1988).

8. van der Waals, J. D. De continuiteit van den gas- en vloeistoftoestand. (1873).

9. Keesom, W. H. & Kamerlingh Onnes, H. The second virial coefficient for rigid spherical molecules, whose mutual attraction is equivalent to that of a quadruplet placed at their center. *Proc. Sect. Sci.* **18**, 636–646 (1915).

10. London, F. Zur Theorie und Systematik der Molekularkräfte. *Zeitschrift für Phys.* **63**, 245–279 (1930).

11. Pauli, W. Über den Zusammenhang des Abschlusses der Elektronengruppen im Atom mit der Komplexstruktur der Spektren. *Zeitschrift für Phys.* **31**, 765–783 (1925).

12. Jones, J. E. & Chapman, S. On the Determination of Molecular Fields.-II. From the Equation of State of a Gas. **4**, 463–477 (1924).

13. Israelachvili, J. N. Interactions Involving the Polarization of Molecules. in *Intermolecular and Surface Forces* 91–106 (Elsevier, 2011). doi:10.1016/B978-0-12-391927-4.10005-2

14. James III, W. H. *et al.* Intramolecular Amide Stacking and Its Competition with Hydrogen Bonding in a Small Foldamer. *J. Am. Chem. Soc.* **131**, 14243–14245 (2009).

15. Rabinovitz, M. & Pines, A. The association of dimethylformamide molecules in carbon tetrachloride solution. *J. Chem. Soc. B Phys. Org.* 1110–1111 (1968). doi:10.1039/J29680001110

16. Werner, A. Ueber Haupt- und Nebenvalenzen und die Constitution der Ammoniumverbindungen. *Justus Liebigs Ann. Chem.* **325**, 261–296 (1902).

17. Jorgensen, W. L., Gao, J. & Ravimohan, C. Monte Carlo simulations of alkanes in water: Hydration numbers and the hydrophobic effect. *J. Phys. Chem.* **89**, 3470–3473 (1985).

18. Meyer, E. A., Castellano, R. K. & Diederich, F. Interactions with Aromatic Rings in Chemical and Biological Recognition. *Angew. Chemie Int. Ed.* **42**, 1210–1250 (2003).

19. McGaughey, G. B., Gagné, M. & Rappé, A. K. π-Stacking Interactions. *J. Biol. Chem.* **273**, 15458–15463 (1998).

20. Tsuzuki, S., Honda, K., Uchimaru, T., Mikami, M. & Tanabe, K. Origin of attraction and directionality of the π/π interaction: Model chemistry calculations of benzene dimer interaction. *J. Am. Chem. Soc.* **124**, 104–112 (2002).

21. Sunner, J., Nishizawa, K. & Kebarle, P. Ion-solvent molecule interactions in the gas phase. The potassium ion and benzene. *J. Phys. Chem.* **85**, 1814–1820 (1981).

22. Ma, J. C. & Dougherty, D. A. The cation-π interaction. *Chem. Rev.* **97**, 1303–1324 (1997).

23. Chipot, C., Maigret, B., Pearlman, D. A. & Kollman, P. A. Molecular Dynamics Potential of Mean Force Calculations: A Study of the Toluene−Ammonium π-Cation Interactions. *J. Am. Chem. Soc.* **118**, 2998–3005 (1996).

24. Politzer, P., Murray, J. S. & Clark, T. Halogen bonding and other σ-hole interactions: A perspective. *Phys. Chem. Chem. Phys.* **15**, 11178–11189 (2013).

25. Beno, B. R., Yeung, K. S., Bartberger, M. D., Pennington, L. D. & Meanwell, N. A. A Survey of the Role of Noncovalent Sulfur Interactions in Drug Design. *J. Med. Chem.* **58**, 4383–4438 (2015).

26. Irvine, G. B., El-Agnaf, O. M., Shankar, G. M. & Walsh, D. M. Protein Aggregation in the Brain: The Molecular Basis for Alzheimer's and Parkinson's Diseases. *Mol. Med.* **14**, 451–464 (2008).

27. (Rob) Aggarwal, S. What's fueling the biotech engine—2012 to 2013. *Nat. Biotechnol.* **32**, 32–39 (2014).

28. United States Pharmacopeia. <788> Particulate Matter in Injections. *USP* **34**, 326–328 (2011).

29. Colombié, S., Gaunand, A., Rinaudo, M. & Lindet, B. Irreversible lysozyme inactivation and aggregation induced by stirring: Kinetic study and aggregates characterisation. *Biotechnol. Lett.* **22**, 277–283 (2000).

30. Porter, S. Human Immune Response to Recombinant Human Proteins. *J. Pharm. Sci.* **90**, 1–11 (2001).

31. Ahmadi, M. *et al.* Small amounts of sub-visible aggregates enhance the immunogenic potential of monoclonal antibody therapeutics. *Pharm. Res.* **32**, 1383–1394 (2015).

32. Rosenberg, A. S. Effects of protein aggregates: An immunologic perspective. *AAPS J.* **8**, E501–E507 (2006).

33. Kraus, T. *et al.* Evaluation of a 3D Human Artificial Lymph Node as Test Model for the Assessment of Immunogenicity of Protein Aggregates. *J. Pharm. Sci.* **108**, 2358–2366 (2019).

34. Roberts, D. *et al.* The Role of Electrostatics in Protein–Protein Interactions of a Monoclonal Antibody. *Mol. Pharm.* **11**, 2475–2489 (2014).

35. Roberts, C. J. Non-Native Protein Aggregation Kinetics. *Biotechnol. Bioeng.* **98**, 927–938 (2007).

36. Cromwell, M. E. M., Felten, C., Flores, H., Liu, J. & Shire, S. J. *Misbehaving Proteins*. (Springer New York, 2006). doi:10.1007/978-0-387-36063-8

37. Lord, R. S., Gubensek, F. & Rupley, J. A. Insulin Self-Association. Spectrum Changes and Thermodynamics. *Biochemistry* **12**, 4385–4392 (1973).

38. Roberts, C. J. Therapeutic protein aggregation: Mechanisms, design, and control. *Trends Biotechnol.* **32**, 372–380 (2014).

39. Lumry, R. & Eyring, H. Conformation Changes of Proteins. *J. Phys. Chem.* **58**, 110–120 (1954).

40. Callahan, M. A., Xiong, L. W. & Caughey, B. Reversibility of Scrapie-associated Prion Protein Aggregation. *J. Biol. Chem.* **276**, 28022–28028 (2001).

41. Sasahara, K., Naiki, H. & Goto, Y. Exothermic effects observed upon heating of β2-microglobulin monomers in the presence of amyloid seeds. *Biochemistry* **45**, 8760–8769 (2006).

42. Meisl, G. *et al.* Molecular mechanisms of protein aggregation from global fitting of kinetic models. *Nat. Protoc.* **11**, 252–272 (2016).

43. Arosio, P., Rima, S., Lattuada, M. & Morbidelli, M. Population balance modeling of antibodies aggregation kinetics. *J. Phys. Chem. B* **116**, 7066–7075 (2012).

44. Kim, N. *et al.* Aggregation of anti-streptavidin immunoglobulin gamma-1 involves Fab unfolding and competing growth pathways mediated by pH and salt concentration. *Biophys. Chem.* **172**, 26–36 (2013).

45. Michaels, T. C. T. & Knowles, T. P. J. Role of filament annealing in the kinetics and thermodynamics of nucleated polymerization. *J. Chem. Phys.* **140**, (2014).

46. Cohen, S. I. A. *et al.* Nucleated polymerization with secondary pathways. I. Time evolution of the principal moments. *J. Chem. Phys.* **135**, 065105 (2011).

47. Knowles, T. P. J. *et al.* An Analytical Solution to the Kinetics of Breakable Filament Assembly. *Science* **326**, 1533–1537 (2009).

48. Meisl, G. *et al.* Differences in nucleation behavior underlie the contrasting aggregation kinetics of the Aβ40 and Aβ42 peptides. *Proc. Natl. Acad. Sci.* **111**, 9384–9389 (2014).

49. Wang, W. & Roberts, C. J. Non-Arrhenius Protein Aggregation. *AAPS J.* **15**, 840–851 (2013).

50. Svilenov, H. & Winter, G. The ReFOLD assay for protein formulation studies and prediction of protein aggregation during long-term storage. *Eur. J. Pharm. Biopharm.* **137**, 131–139 (2019).

51. Chia, S. *et al.* SAR by kinetics for drug discovery in protein misfolding diseases. *Proc. Natl. Acad. Sci.* **115**, 10245–10250 (2018).

52. Zhang, A., Singh, S. K., Shirts, M. R., Kumar, S. & Fernandez, E. J. Distinct aggregation mechanisms of monoclonal antibody under thermal and freeze-thaw stresses revealed by hydrogen exchange. *Pharm. Res.* **29**, 236–250 (2012).

53. Waldron, T. T. & Murphy, K. P. Stabilization of proteins by ligand binding: Application to drug screening and determination of unfolding energetics. *Biochemistry* **42**, 5058–5064 (2003).

54. Arakawa, T., Kita, Y. & Timasheff, S. N. Protein precipitation and denaturation by dimethyl sulfoxide. *Biophys. Chem.* **131**, 62–70 (2007).

55. Shikama, K. & Yamazaki, T. Denaturation of Catalase by Freezing and Thawing. *Nature* **190**, 83–84 (1961).

56. Timasheff, S. N. Thermodynamic binding and site occupancy in the light of the Schellman exchange concept. *Biophys. Chem.* **101–102**, 99–111 (2002).

57. Carpenter, J. F., Crowe, J. H. & Arakawa, T. Comparison of Solute-Induced Protein Stabilization in Aqueous Solution and in the Frozen and Dried States. *J. Dairy Sci.* **73**, 3627–3636 (1990).

58. Svilenov, H. & Winter, G. Rapid sample-saving biophysical characterisation and long-term storage stability of liquid interferon alpha2a formulations: Is there a correlation? *Int. J. Pharm.* **562**, 42–50 (2019).

59. Gentiluomo, L. *et al.* Application of interpretable artificial neural networks to early monoclonal antibodies development. *Eur. J. Pharm. Biopharm.* **141**, 81–89 (2019).

60. Feng, Y. W., Ooishi, A. & Honda, S. Aggregation factor analysis for protein formulation by a systematic approach using FTIR, SEC and design of experiments techniques. *J. Pharm. Biomed. Anal.* **57**, 143–152 (2012).

61. Costa, T. R. D., Ignatiou, A. & Orlova, E. V. *Bacterial Protein Secretion Systems*. **1615**, (Springer New York, 2017).

62. Grigorieff, N. & Harrison, S. C. Near-atomic resolution reconstructions of icosahedral viruses from electron cryo-microscopy. *Curr. Opin. Struct. Biol.* **21**, 265–273 (2011).

63. Lyumkis, D. Challenges and opportunities in cryo-EM single-particle analysis. *J. Biol. Chem.* **294**, 5181–5197 (2019).

64. Guo, Q. *et al.* In Situ Structure of Neuronal C9orf72 Poly-GA Aggregates Reveals Proteasome Recruitment. *Cell* **172**, 696–705 (2018).

65. Gruber, A. *et al.* Molecular and structural architecture of polyQ aggregates in yeast. *Proc. Natl. Acad. Sci.* **115**, E3446–E3453 (2018).

66. Yan, K., Zhang, Z., Yang, J., McLaughlin, S. H. & Barford, D. Architecture of the CBF3–centromere complex of the budding yeast kinetochore. *Nat. Struct. Mol. Biol.* **25**, 1103–1110 (2018).

67. Ziarek, J. J., Peterson, F. C., Lytle, B. L. & Volkman, B. F. Binding site identification and structure determination of protein-ligand complexes by NMR a semiautomated approach. *Methods Enzymol.* **493**, 241–75 (2011).

68. Yu, H. Extending the size limit of protein nuclear magnetic resonance. *Proc. Natl. Acad. Sci.* **96**, 332–334 (1999).

69. Edwards, J. M. *et al.* 19 F Dark-State Exchange Saturation Transfer NMR Reveals Reversible Formation of Protein-Specific Large Clusters in High-Concentration Protein Mixtures. *Anal. Chem.* **91**, 4702–4708 (2019).

70. Tang, C., Ghirlando, R. & Clore, G. M. Visualization of transient ultra-weak protein self-association in solution using paramagnetic relaxation enhancement. *J. Am. Chem. Soc.* **130**, 4048–4056 (2008).

71. Kuwata, K. *et al.* NMR-detected hydrogen exchange and molecular dynamics simulations provide structural insight into fibril formation of prion protein fragment 106-126. *Proc. Natl. Acad. Sci.* **100**, 14790–14795 (2003).

72. Hilser, V. J., Dowdy, D., Oas, T. G. & Freire, E. The structural distribution of cooperative interactions in proteins: Analysis of the native state ensemble. *Proc. Natl. Acad. Sci.* **95**, 9903–9908 (1998).

73. Dominguez, C., Boelens, R. & Bonvin, A. M. J. J. HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* **125**, 1731–1737 (2003).

74. Smith, G. R. & Sternberg, M. J. E. Prediction of protein-protein interactions by docking methods. *Curr. Opin. Struct. Biol.* **12**, 28–35 (2002).

75. Camacho, C. J., Thirumalai, D., Bryson, J., Roder, H. & DeGrado, W. Kinetics and thermodynamics of folding in model proteins. *Proc. Natl. Acad. Sci.* **90**, 6369–6372 (1993).

76. Kilambi, K. P., Reddy, K. & Gray, J. J. Protein-Protein Docking with Dynamic Residue Protonation States. *PLoS Comput. Biol.* **10**, (2014).

77. Schor, M., Vreede, J. & Bolhuis, P. G. Elucidating the locking mechanism of peptides onto growing amyloid fibrils through transition path sampling. *Biophys. J.* **103**, 1296–1304 (2012).

78. Vácha, R., Linse, S. & Lund, M. Surface effects on aggregation kinetics of amyloidogenic peptides. *J. Am. Chem. Soc.* **136**, 11776–11782 (2014).

79. Trovato, A., Seno, F. & Tosatto, S. C. E. The PASTA server for protein aggregation prediction. *Protein Eng. Des. Sel.* **20**, 521–523 (2007).

80. Zambrano, R. *et al.* AGGRESCAN3D (A3D): Server for prediction of aggregation properties of protein structures. *Nucleic Acids Res.* **43**, W306–W313 (2015).

81. Janzen, W. P. *High Throughput Screening*. (Springer New York, 2016). doi:10.1007/978-1-4939-3673-1

82. Kranz, J. K. & Schalk-Hihi, C. Protein Thermal Shifts to Identify Low Molecular Weight Fragments. in *Fragment Based Drug Design Tools, Practical Approaches, and Examples* **493**, 277–298 (2011).

83. Jerabek-Willemsen, M., Wienken, C. J., Braun, D., Baaske, P. & Duhr, S. Molecular Interaction Studies Using Microscale Thermophoresis. *Assay Drug Dev. Technol.* **9**, 342–353 (2011).

84. Von Ahsen, O., Schmidt, A., Klotz, M. & Parczyk, K. Assay Concordance between SPA and TR-FRET in High-Throughput Screening. *J. Biomol. Screen.* **11**, 606–616 (2006).

85. Meng, E. C., Shoichet, B. K. & Kuntz, I. D. Automated docking with grid-based energy evaluation. *J. Comput. Chem.* **13**, 505–524 (1992).

86. Sterling, T. & Irwin, J. J. ZINC 15 - Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **55**, 2324–2337 (2015).

87. Redl, G., Tert., R. D. C. & Berkoff, C. E. Quantitative drug design. *Chem. Soc. Rev.* **3**, 273 (1974).

88. Hansch, C. & Dunn, W. J. Linear Relationships between Lipophilic Character and Biological Activity of Drugs. *J. Pharm. Sci.* **61**, 1–19 (1972).

89. Anderson, A. C. The Process of Structure-Based Drug Design. *Chem. Biol.* **10**, 787–797 (2003).

90. Samuels, P. B. & Roedling, H. Prediction of intestinal absorption: comparative assessment of gastroplus™ and idea™. *Eur. J. Pharm. Sci.* **17**, 51–61 (2002).

91. del Amo, E. M. *et al.* Applying Linear and Non-Linear Methods for Parallel Prediction of Volume of Distribution and Fraction of Unbound Drug. *PLoS One* **8**, e74758 (2013).

92. Talevi, A. & Quiroga, P. A. M. *ADME Processes in Pharmaceutical Sciences*. (Springer International Publishing, 2018). doi:10.1007/978-3-319-99593-9

93. Eddershaw, P. J., Beresford, A. P. & Bayliss, M. K. ADME/PK as part of a rational approach to drug discovery. *Drug Discov. Today* **5**, 409–414 (2000).

94. Alqahtani, S. In silico ADME-Tox modeling: progress and prospects. *Expert Opin. Drug Metab. Toxicol.* **13**, 1147–1158 (2017).

95. Lepri, S. *et al.* Structure–metabolism relationships in human- AOX: Chemical insights from a large database of aza-aromatic and amide compounds. *Proc. Natl. Acad. Sci.* **114**, E3178–E3187 (2017).

96. Berellini, G., Waters, N. J. & Lombardo, F. In silico Prediction of Total Human Plasma Clearance. *J. Chem. Inf. Model.* **52**, 2069–2078 (2012).

97. Waller, D. G. & Sampson, A. P. Drug toxicity and overdose. in *Medical Pharmacology and Therapeutics* 659–673 (Elsevier, 2018). doi:10.1016/B978-0-7020-7167-6.00053-1

98. Banerjee, P., Eckert, A. O., Schrey, A. K. & Preissner, R. ProTox-II: A webserver for the prediction of toxicity of chemicals. *Nucleic Acids Res.* **46**, W257–W263 (2018).

99. Yang, H., Sun, L., Li, W., Liu, G. & Tang, Y. In Silico Prediction of Chemical Toxicity for Drug Design Using Machine Learning Methods and Structural Alerts. *Front. Chem.* **6**, 1–12 (2018).

100. Pritchard, J. F. *et al.* Making Better Drugs: Decision Gates in Non-Clinical Drug Development. *Nat. Rev. Drug Discov.* **2**, 542–553 (2003).

101. Food and Drug Administration (FDA) Center for Drug Evaluation Research (CDER). *Guidance for Industry: M3(R2) Nonclinical Safety Studies for the Conduct of Human Clinical Trials and Marketing Authorization for Pharmaceuticals*. (2010).

102. Kitteringham, B., Munir, N. R. & Park, K. The Role of Active Metabolites in Drug Toxicity. *Drug Saf.* **11**, 114–144 (1994).

103. Lambrinidis, G. & Tsantili-Kakoulidou, A. Challenges with multi-objective QSAR in drug discovery. *Expert Opin. Drug Discov.* **13**, 851–859 (2018).

104. Khanna, I. Drug discovery in pharmaceutical industry: Productivity challenges and trends. *Drug Discov. Today* **17**, 1088–1102 (2012).

105. Boström, J., Brown, D. G., Young, R. J. & Keserü, G. M. Expanding the medicinal chemistry synthetic toolbox. *Nat. Rev. Drug Discov.* **17**, 709–727 (2018).

106. Schneider, G. Automating drug discovery. *Nat. Rev. Drug Discov.* **17**, 97–113 (2018).

107. Ratanji, K. D., Derrick, J. P., Dearman, R. J. & Kimber, I. Immunogenicity of therapeutic proteins: Influence of aggregation. *J. Immunotoxicol.* **11**, 99–109 (2014).

108. Roberts, C. J. Protein aggregation and its impact on product quality. *Curr. Opin. Biotechnol.* **30**, 211–217 (2014).

109. Clackson, T. & Wells, J. a. A hot spot of binding energy in a hormone-receptor interface. *Science* **267**, 383–386 (1995).

110. Fernandez-Escamilla, A.-M., Rousseau, F., Schymkowitz, J. & Serrano, L. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.* **22**, 1302–1306 (2004).

111. Conchillo-Solé, O. *et al.* AGGRESCAN: a server for the prediction and evaluation of 'hot spots' of aggregation in polypeptides. *BMC Bioinformatics* **65**, (2007).

112. Barata, T. S., Zhang, C., Dalby, P. A., Brocchini, S. & Zloh, M. Identification of protein-excipient interaction hotspots using computational approaches. *Int. J. Mol. Sci.* **17**, (2016).

113. Kheddo, P. *et al.* The effect of arginine glutamate on the stability of monoclonal antibodies in solution. *Int. J. Pharm.* **473**, 126–133 (2014).

114. Veurink, M., Westermaier, Y., Gurny, R. & Scapozza, L. Breaking the aggregation of the monoclonal antibody bevacizumab (Avastin®) by dexamethasone phosphate: Insights from molecular modelling and asymmetrical flow field-flow fractionation. *Pharm. Res.* **30**, 1176–1187 (2013).

115. Westermaier, Y. *et al.* Identification of aggregation breakers for bevacizumab (Avastin®) self-association through similarity searching and interaction studies. *Eur. J. Pharm. Biopharm.* **85**, 773–780 (2013).

116. Veurink, M., Stella, C., Tabatabay, C., Pournaras, C. J. & Gurny, R. Association of ranibizumab (Lucentis®) or bevacizumab (Avastin®) with dexamethasone and triamcinolone acetonide: An in vitro stability assessment. *Eur. J. Pharm. Biopharm.* **78**, 271–277 (2011).

117. Jin, L., Wang, W. & Fang, G. Targeting Protein-Protein Interaction by Small Molecules. *Annu. Rev. Pharmacol. Toxicol.* **54**, 435–456 (2014).

118. Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **432**, 862–865 (2004).

119. Sterling, T. & Irwin, J. J. ZINC 15 – Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **55**, 2324–2337 (2015).

120. Repasky, M. P. *et al.* Docking performance of the glide program as evaluated on the Astex and DUD datasets: A complete set of glide SP results and selected results for a new scoring function integrating WaterMap and glide. *J. Comput. Aided. Mol. Des.* **26**, 787–799 (2012).

121. Trott, O. & Olson, A. Autodock vina: improving the speed and accuracy of docking. *J. Comput. Chem.* **31**, 455–461 (2010).

122. Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **267**, 727–748 (1997).

123. Korb, O. *et al.* Potential and limitations of ensemble docking. *J. Chem. Inf. Model.* **52**, 1262–1274 (2012).

124. 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.* **108**, 10184–9 (2011).

125. Barducci, A., Bonomi, M. & Parrinello, M. Metadynamics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **1**, 826–843 (2011).

126. Woo, H.-J. & Roux, B. Calculation of absolute protein-ligand binding free energy from computer simulations. *Proc. Natl. Acad. Sci.* **102**, 6825–6830 (2005).

127. Velez-Vega, C. & Gilson, M. K. Force and stress along simulated dissociation pathways of cucurbituril-guest systems. *J. Chem. Theory Comput.* **8**, 966–976 (2012).

128. Velez-Vega, C. & Gilson, M. K. Overcoming dissipation in the calculation of standard binding free energies by ligand extraction. *J. Comput. Chem.* **34**, 2360–2371 (2013).

129. Gilson, M. K., Given, J. A., Bush, B. L. & McCammon, J. A. The statistical-thermodynamic basis for computation of binding affinities: A critical review. *Biophys. J.* **72**, 1047–1069 (1997).

130. Henriksen, N. M., Fenley, A. T. & Gilson, M. K. Computational Calorimetry: High-Precision Calculation of Host-Guest Binding Thermodynamics. *J. Chem. Theory Comput.* **11**, 4377–4394 (2015).

131. Seidel, S. A. I. I. *et al.* Microscale thermophoresis quantifies biomolecular interactions under previously challenging conditions. *Methods* **59**, 301–315 (2013).

132. Jorgensen, W. L. & Duffy, E. M. Prediction of drug solubility from structure. *Adv. Drug Deliv. Rev.* **54**, 355–366 (2002).

133. Klopman, G., Wang, S. & Balthasar, D. M. Estimation of Aqueous Solubility of Organic Molecules by the Group Contribution Approach. Application to the Study of Biodegradation. *J. Chem. Inf. Comput. Sci.* **32**, 474–482 (1992).

134. Kühne, R., Ebert, R.-U., Kleint, F., Schmidt, G. & Schüürmann, G. Group contribution methods to estimate water solubility of organic chemicals. *Chemosphere* **30**, 2061–2077 (1995).

135. Tetko, I. V, Tanchuk, V. Y., Kasheva, T. N. & Villa, A. E. P. Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices. *J. Chem. Inf. Comput. Sci.* **41**, 1488–1493 (2001).

136. Duffy, E. M. & Jorgensen, W. L. Prediction of Properties from Simulations: Free Energies of Solvation in Hexadecane, Octanol, and Water. *J. Am. Chem. Soc.* **122**, 2878–2888 (2000).

137. Hewitt, M. *et al.* In silico prediction of aqueous solubility: The solubility challenge. *J. Chem. Inf. Model.* **49**, 2572–2587 (2009).

138. Llinàs, A., Glen, R. C. & Goodman, J. M. Solubility challenge: Can you predict solubilities of 32 molecules using a database of 100 reliable measurements? *J. Chem. Inf. Model.* **48**, 1289–1303 (2008).

139. Hopfinger, A. J., Esposito, E. X., Llinàs, A., Glen, R. C. & Goodman, J. M. Findings of the Challenge To Predict Aqueous Solubility. *J. Chem. Inf. Model.* **49**, 1–5 (2009).

140. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development. *Adv. Drug Deliv. Rev.* **46**, 3–26 (2001).

141. Irwin, J. J. & Shoichet, B. K. ZINC – A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **45**, 177–182 (2005).

142. He, F., Woods, C. E., Becker, G. W., Narhi, L. O. & Razinkov, V. I. High-Throughput Assessment of Thermal and Colloidal Stability Parameters for Monoclonal Antibody Formulations. *J. Pharm. Sci.* **100**, 5126–5141 (2011).

143. Goldberg, D. S., Bishop, S. M., Shah, A. U. & Sathish, H. A. Formulation Development of Therapeutic Monoclonal Antibodies Using High-Throughput Fluorescence and Static Light Scattering Techniques: Role of Conformational and Colloidal Stability. *J. Pharm. Sci.* **100**, 1306–1315 (2011).

144. Vanhooren, A., Devreese, B., Vanhee, K., Van Beeumen, J. & Hanssens, I. Photoexcitation of tryptophan groups induces reduction of two disulfide bonds in goat α-lactalbumin. *Biochemistry* **41**, 11035–11043 (2002).

145. Brange, J., Havelund, S. & Hougaard, P. Chemical Stability of Insulin. 2. Formation of Higher Molecular Weight Transformation Products During Storage of Pharmaceutical Preparations. *Pharm. Res.* **9**, 727–734 (1992).

146. Eswar, N. *et al.* Comparative Protein Structure Modeling Using Modeller. *Curr. Protoc. Bioinforma.* **15**, 5.6.1-5.6.30 (2006).

147. Mahler, H.-C. *et al.* Adsorption Behavior of a Surfactant and a Monoclonal Antibody to Sterilizing-Grade Filters. *J. Pharm. Sci.* **99**, 2620–2627 (2010).

148. Salmon-Ferrer, R., Goetz, A. W., Poole, D., Le Grand, S. & Walker, R. C. Routine microsecond molecular dynamics simulations with AMBER - Part II: Particle Mesh Ewald. *J. Chem. Theory Comput.* **9**, 3878–3888 (2013).

149. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water William. *J. Chem. Phys.* **79**, 926–935 (1983).

150. Ryckaert, J.-P., Ciccotti, G. & Berendsen, H. J. . Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **23**, 327–341 (1977).

151. Feenstra, K. A., Hess, B. & Berendsen, H. J. C. Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems. *J. Comput. Chem.* **20**, 786–798 (1999).

152. Vedani, A., Dobler, M., Hu, Z. & Smieško, M. OpenVirtualToxLab—A platform for generating and exchanging in silico toxicity data. *Toxicol. Lett.* **232**, 519–532 (2015).

153. Nuhu, M. M. & Curtis, R. Arginine dipeptides affect insulin aggregation in a pH- and ionic strength-dependent manner. *Biotechnol. J.* **10**, 404–416 (2015).

154. Gilson, M. K. *et al.* BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* **44**, D1045–D1053 (2016).

155. Svilenov, H., Markoja, U. & Winter, G. Isothermal chemical denaturation as a complementary tool to overcome limitations of thermal differential scanning fluorimetry in predicting physical stability of protein formulations. *Eur. J. Pharm. Biopharm.* **125**, 106–113 (2018).

156. Niccolai, N. *et al.* Hot spot mapping of protein surfaces with TEMPOL: Bovine pancreatic RNase A as a model system. *Biochim. Biophys. Acta* **1865**, 201–207 (2017).

157. Wang, W. & Roberts, C. J. *Aggregation of Therapeutic Proteins.* (John Wiley & Sons, Inc., 2010). doi:10.1002/9780470769829

158. Meric, G., Robinson, A. S. & Roberts, C. J. Driving Forces for Nonnative Protein Aggregation and Approaches to Predict Aggregation-Prone Regions. *Annu. Rev. Chem. Biomol. Eng.* **8**, 139–159 (2017).

159. Sormanni, P., Aprile, F. A. & Vendruscolo, M. The CamSol Method of Rational Design of Protein Mutants with Enhanced Solubility. *J. Mol. Biol.* **427**, 478–490 (2015).

160. Sankar, K., Krystek, S. R., Carl, S. M., Day, T. & Maier, J. K. X. AggScore: Prediction of aggregation-prone regions in proteins based on the distribution of surface patches. *Proteins Struct. Funct. Bioinforma.* **86**, 1147–1156 (2018).

161. Arora, J. *et al.* Hydrogen exchange mass spectrometry reveals protein interfaces and distant dynamic coupling effects during the reversible self-association of an IgG1 monoclonal antibody. *MAbs* **7**, 525–539 (2015).

162. Solomon, I. & Bloembergen, N. Nuclear magnetic interactions in the HF molecule. *J. Chem. Phys.* **25**, 261–266 (1956).

163. Pintacuda, G. & Otting, G. Identification of Protein Surfaces by NMR Measurements with a Paramagnetic Gd(III) Chelate. *J. Am. Chem. Soc.* **124**, 372–373 (2002).

164. Berliner, L. *Protein NMR: Modern Techniques and Biomedical Applications*. (Springer, 2015). doi:10.1007/978-1-4899-7621-5

165. Keizers, P. H. J. & Ubbink, M. Paramagnetic tagging for protein structure and dynamics analysis. *Prog. Nucl. Magn. Reson. Spectrosc.* **58**, 88–96 (2011).

166. Esposito, G. *et al.* Probing protein structure by solvent perturbation of nuclear magnetic resonance spectra. *J. Mol. Biol.* **224**, 659–670 (1992).

167. Kopple, K. D., Petros, A. M. & Mueller, L. NMR Identification of Protein Surfaces Using Paramagnetic Probes. *Biochemistry* **29**, 10041–10048 (1990).

168. Niccolai, N. *et al.* NMR Studies of Protein Surface Accessibility. *J. Biol. Chem.* **276**, 42455–42461 (2001).

169. Arumugam, S. *et al.* TIMP-1 Contact Sites and Perturbations of Stromelysin 1 Mapped by NMR and a Paramagnetic Surface Probe. *Biochemistry* **37**, 9650–9657 (1998).

170. Schilder, J. & Ubbink, M. Weak self-association of cytochrome c peroxidase molecules observed by paramagnetic NMR. *J. Biomol. NMR* **65**, 29–40 (2016).

171. Klaus, W., Gsell, B., Labhardt, A. M., Wipf, B. & Senn, H. The three-dimensional high resolution structure of human interferon α-2a determined by heteronuclear NMR spectroscopy in solution. *J. Mol. Biol.* **274**, 661–675 (1997).

172. Panjwani, N., Hodgson, D. J., Sauvé, S. & Aubin, Y. Assessment of the Effects of pH, Formulation and Deformulation on the Conformation of Interferon Alpha-2 by NMR. *J. Pharm. Sci.* **99**, 3334–3342 (2010).

173. Tobi, D. & Bahar, I. Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. *Proc. Natl. Acad. Sci.* **102**, 18908–18913 (2005).

174. Thomas, C. *et al.* Structural linkage between ligand discrimination and receptor activation by Type i interferons. *Cell* **146**, 621–632 (2011).

175. Carpenter, J. F. & Crowe, J. H. The mechanism of cryoprotection of proteins by solutes. *Cryobiology* **25**, 244–255 (1988).

176. Timasheff, S. N. Protein hydration, thermodynamic binding, and preferential hydration. *Biochemistry* **41**, 13473–13482 (2002).

177. Schellman, J. A. The Thermodynamic Stability of Proteins. *Annu. Rev. Biophys. Biophys. Chem.* **16**, 115–137 (1987).

178. Matulis, D., Kranz, J. K., Salemme, F. R. & Todd, M. J. Thermodynamic Stability of Carbonic Anhydrase: Measurements of Binding Affinity and Stoichiometry Using ThermoFluor. *Biochemistry* **44**, 5258–5266 (2005).

179. Tosstorff, A., Svilenov, H., Peters, G. H. J., Harris, P. & Winter, G. Structure-based discovery of a new protein-aggregation breaking excipient. *Eur. J. Pharm. Biopharm.* **144**, 207–216 (2019).

180. Lochs, H., Morse, E. L. & Adibi, S. A. Mechanism of hepatic assimilation of dipeptides. Transport versus hydrolysis. *J. Biol. Chem.* **261**, 14976–14981 (1986).

181. Ohlson, M. *et al.* Effects of filtration rate on the glomerular barrier and clearance of four differently shaped molecules. *Am. J. Physiol. Physiol.* **281**, F103–F113 (2001).

182. Corzo, J. Time, the forgotten dimension of ligand binding teaching. *Biochem. Mol. Biol. Educ.* **34**, 413–416 (2006).

183. Ferruz, N. & De Fabritiis, G. Binding Kinetics in Drug Discovery. *Mol. Inform.* **35**, 216–226 (2016).

184. Insaidoo, F. & Roush, D. In silico process for selecting protein formulation excipients. (2017).

185. Hagler, A. T., Huler, E. & Lifson, S. Energy functions for peptides and proteins. I. Derivation of a consistent force field including the hydrogen bond from amide crystals. *J. Am. Chem. Soc.* **96**, 5319–5327 (1974).

186. Shukla, D. & Trout, B. L. Preferential interaction coefficients of proteins in aqueous arginine solutions and their molecular origins. *J. Phys. Chem. B* **115**, 1243–1253 (2011).

187. Plattner, N. & Noé, F. Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and Markov models. *Nat. Commun.* **6**, (2015).

188. Scherer, M. K. *et al.* PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theory Comput.* **11**, 5525–5542 (2015).

189. Pérez-Hernández, G. & Noé, F. Hierarchical Time-Lagged Independent Component Analysis: Computing Slow Modes and Reaction Coordinates for Large Molecular Systems. *J. Chem. Theory Comput.* **12**, 6118–6129 (2016).

190. Wehmeyer, C. *et al.* Introduction to Markov state modeling with the PyEMMA software [Article v1.0]. *Living J. Comput. Mol. Sci.* **1**, 1–8 (2019).

191. Doerr, S., Harvey, M. J., Noé, F. & De Fabritiis, G. HTMD: High-Throughput Molecular Dynamics for Molecular Discovery. *J. Chem. Theory Comput.* **12**, 1845–1852 (2016).

192. Le Grand, S., Götz, A. W. & Walker, R. C. SPFP: Speed without compromise - A mixed precision model for GPU accelerated molecular dynamics simulations. *Comput. Phys. Commun.* **184**, 374–380 (2013).

193. Case, D. A. *et al.* Amber 18. (2018).

194. Pearlman, D. A. *et al.* AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comput. Phys. Commun.* **91**, 1–41 (1995).

195. Chandrasekhar, S. Stochastic Problems in Physics and Astronomy. *Rev. Mod. Phys.* **15**, 1–89 (1943).

196. Miyamoto, S. & Kollman, P. A. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* **13**, 952–962 (1992).

197. Chi, E. Y. *Excipients Used in Biotechnology Products. Pharmaceutical Excipients: Properties, Functionality, and Applications in Research and Industry* (John Wiley & Sons, Inc., 2017). doi:10.1002/9781118992432.ch4

198. Koo, O. M. Y. *Pharmaceutical Excipients*. (John Wiley & Sons, Inc., 2016). doi:10.1002/9781118992432

199. Manning, M. C., Chou, D. K., Murphy, B. M., Payne, R. W. & Katayama, D. S. Stability of Protein Pharmaceuticals: An Update. *Pharm. Res.* **27**, 544–575 (2010).

200. Bohacek, R. S., Mcmartin, C. & Guida, W. C. The Art and Practice of Structure-Based Drug Design : A Molecular Modeling Perspective. **16**, 3–50 (1996).

201. Kamerzell, T. J., Esfandiary, R., Joshi, S. B., Middaugh, C. R. & Volkin, D. B. Protein-excipient interactions: Mechanisms and biophysical characterization applied to protein formulation development. *Adv. Drug Deliv. Rev.* **63**, 1118–1159 (2011).

202. Ecker, D. M., Jones, S. D. & Levine, H. L. The therapeutic monoclonal antibody market. *MAbs* **7**, 9–14 (2015).

203. Raybould, M. I. J. *et al.* Five computational developability guidelines for therapeutic antibody profiling. *Proc. Natl. Acad. Sci.* **116**, 4025–4030 (2019).

204. Seeliger, D. *et al.* Boosting antibody developability through rational sequence optimization. *MAbs* **7**, 505–515 (2015).

205. Tomlinson, A., Demeule, B., Lin, B. & Yadav, S. Polysorbate 20 Degradation in Biopharmaceutical Formulations: Quantification of Free Fatty Acids, Characterization of Particulates, and Insights into the Degradation Mechanism. *Mol. Pharm.* **12**, 3805–3815 (2015).

206. Ha, E., Wang, W. & John Wang, Y. Peroxide formation in polysorbate 80 and protein stability. *J. Pharm. Sci.* **91**, 2252–2264 (2002).

207. Kranz, J. K. & Schalk-Hihi, C. Protein Thermal Shifts to Identify Low Molecular Weight Fragments. *Methods Enzymol.* **493**, 277–298 (2011).

208. Svilenov, H., Gentiluomo, L., Friess, W., Roessner, D. & Winter, G. A New Approach to Study the Physical Stability of Monoclonal Antibody Formulations—Dilution From a Denaturant. *J. Pharm. Sci.* **107**, 3007–3013 (2018).

209. Leach, A. R. & Gillet, V. J. *An Introduction To Chemoinformatics.* (Springer Netherlands, 2007). doi:10.1007/978-1-4020-6291-9

210. Sakuratani, Y., Kasai, K., Noguchi, Y. & Yamada, J. Comparison of predictivities of log P calculation models based on experimental data for 134 simple organic compounds. *QSAR Comb. Sci.* **26**, 109–116 (2007).

211. Polton, D. J. Installation and operational experiences with MACCS (Molecular Access System). *Online Rev.* **6**, 235–242 (1982).

212. Welford, S. M., Lynch, M. F. & Barnard, J. M. Towards simplified access to chemical structure information in the patent literature. *J. Inf. Sci.* **6**, 3–10 (1983).

213. Myint, K.-Z., Wang, L., Tong, Q. & Xie, X. Q. Molecular fingerprint-based artificial neural networks QSAR for ligand biological activity predictions. *Mol. Pharm.* **9**, 2912–2923 (2012).

214. Oyetayo, O. O., Méndez-Lucio, O., Bender, A. & Kiefer, H. Diversity selection, screening and quantitative structure–activity relationships of osmolyte-like additive effects on the thermal stability of a monoclonal antibody. *Eur. J. Pharm. Sci.* **97**, 151–157 (2017).

215. Pantoliano, M. W., Bone, R. F., Rhind, A. W. & Salemme, F. R. Microplate Thermal Shift Assay Apparatus for Ligand Development and multi-variable protein chemistry optimization. (2004).

216. Niesen, F. H., Berglund, H. & Vedadi, M. The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. *Nat. Protoc.* **2**, 2212–2221 (2007).

217. Borysko, P. *et al.* Straightforward hit identification approach in fragment-based discovery of bromodomain-containing protein 4 (BRD4) inhibitors. *Bioorganic Med. Chem.* **26**, 3399–3405 (2018).

218. Sule, S. V. *et al.* Solution pH that minimizes self-association of three monoclonal antibodies is strongly dependent on ionic strength. *Mol. Pharm.* **9**, 744–751 (2012).

219. Dunn, W. J. QSAR approaches to predicting toxicity. *Toxicol. Lett.* **43**, 277–283 (1988).

220. Jones, E., Oliphant, E., Peterson, P. & Others, A. SciPy: Open Source Scientific Tools for Python. (2001).

221. Newville, M., Stensitzki, T., Allen, D. B. & Ingargiola, A. LMFIT: Non-Linear Least-Square Minimization and Curve-Fitting for Python. (2014). doi:10.5281/ZENODO.11813

Appendix

The data presented in the following describes the stabilizing effect of N,N,N',N'-tetrakis-(2-hydroxyethyl) adipinic acid amide on the monoclonal antibody Trastuzumab and Interferon-alpha-2a. The substance was discovered by applying the same virtual screen approach described in chapter 3 to Trastuzumab. The stress studies with Trastuzumab were performed by Luis Sánchez. The stress study with IFN was performed by Andreas Tosstorff. A patent application that describes the use of this substance as excipient was filed (19186002.2 - 1116). Designated inventors are Andreas Tosstorff, Günther Peters and Gerhard Winter.

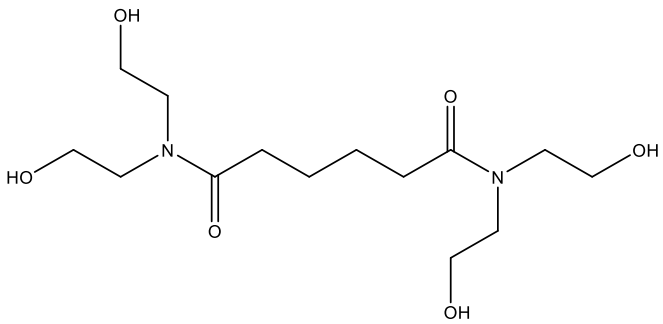


Figure A 1: Structure of N,N,N',N'-tetrakis-(2-hydroxyethyl) adipinic acid amide (compound A)

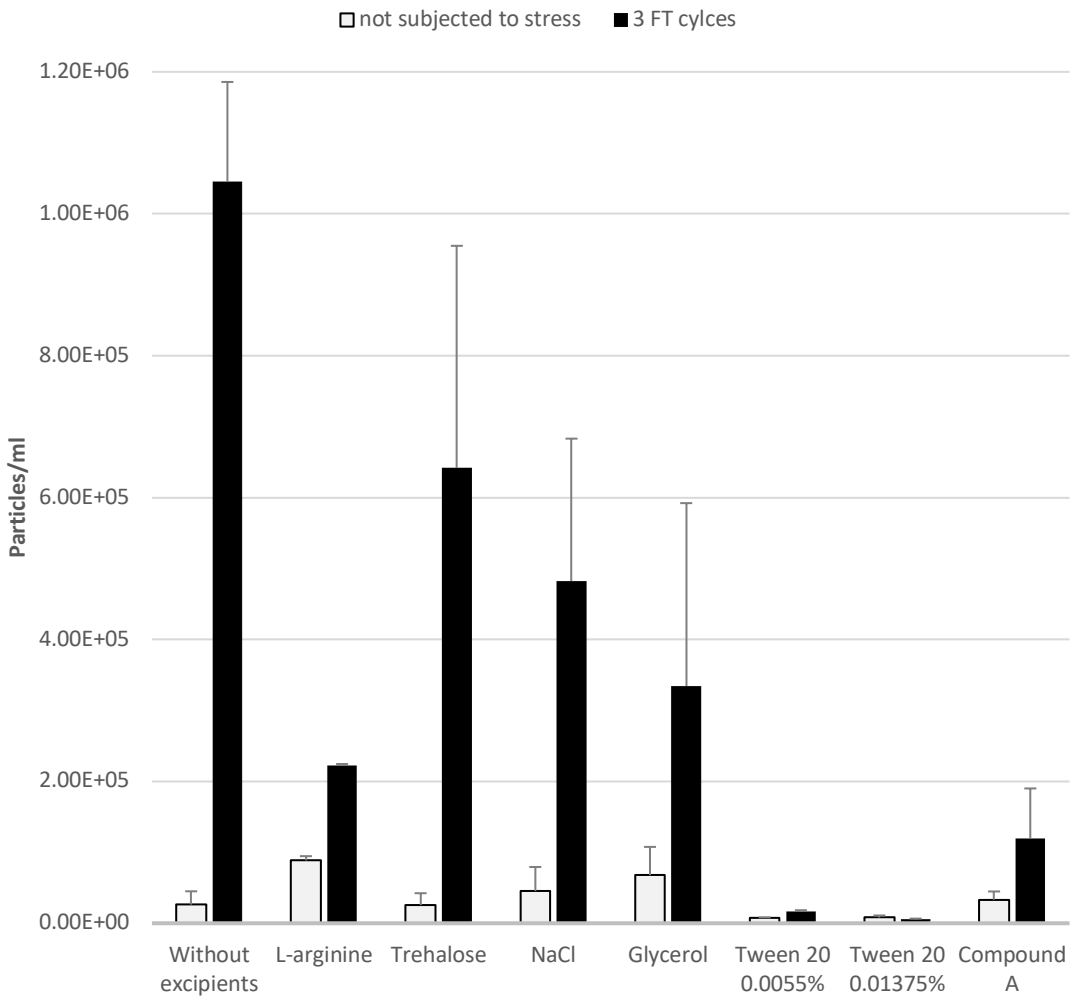


Figure A 2: Number of particles after 3 freezing/thawing cycles. 50 mM phosphate buffer, pH 7.0; 5 mg/ml antibody IgG trastuzumab.

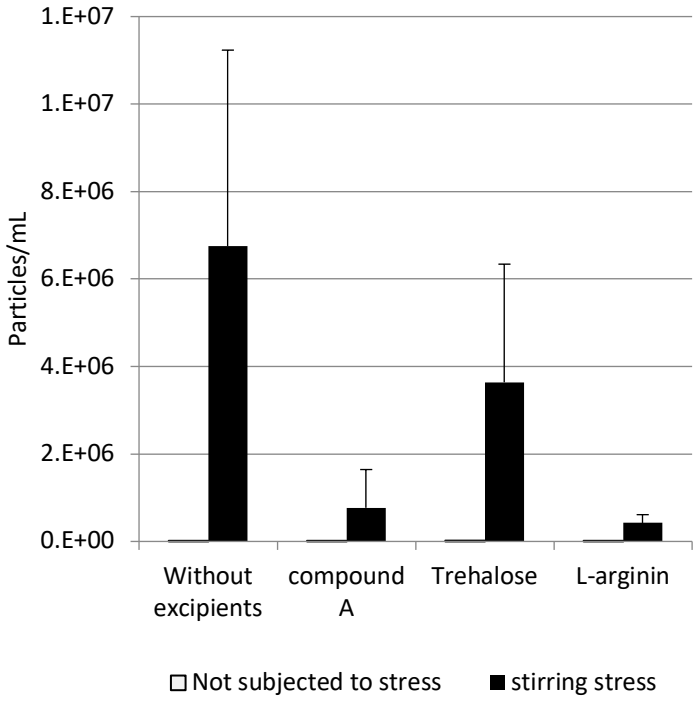


Figure A 3: Number of particles after stirring stress. 50 mM phosphate buffer, pH 7.0; 5 mg/ml antibody IgG trastuzumab.

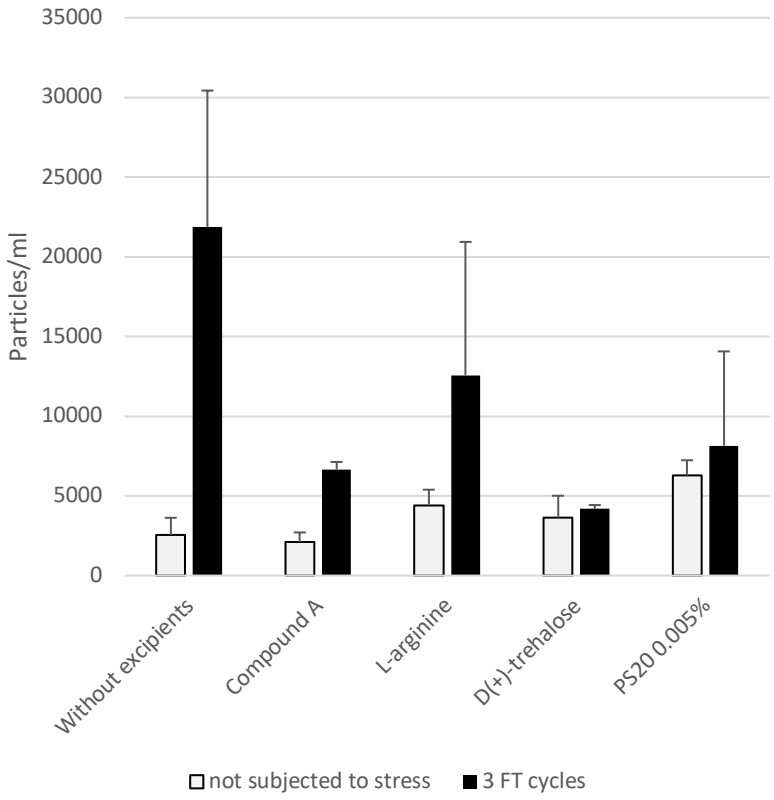


Figure A 4: Number of particles after 3 freezing/thawing cycles. 50 mM phosphate buffer, pH 7.0; 1 mg/ml interferon-alpha-2a.