

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

**Application of molecular dynamics simulations
for developability assessment and formulation
development of biologics**



Carolin Berner

aus

Osnabrück, Deutschland

2023

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.

München, den 19.09.2023

.....

(Carolin Berner)

| | |
|-----------------------------|--------------------------|
| Dissertation eingereicht am | 17.03.2023 |
| 1. Gutachter: | Prof. Dr. Gerhard Winter |
| 2. Gutachter: | Prof. Dr. Wolfgang Frieß |
| Mündliche Prüfung am | 31.03.2023 |

To my family

Acknowledgements

First and foremost, I would like to thank my supervisor Professor Dr. Gerhard Winter for giving me the opportunity to work on this exciting project and his support in exploring the field of molecular dynamics simulations. I appreciate his continuous encouragement, advice, trust in my work, and the provided freedom in following my ideas.

Second, I want to thank Professor Dr. Günther H. J. Peters for the great collaboration over the years and hosting me at his lab in Copenhagen. My special thanks to Dr. Sowmya Indrakumar for making this a wonderful experience that helped me start my journey with MD simulations. Thank you Dr. Suk Kyu Ko for being such a great collaboration partner and good friend, and of course for your support whenever I needed advice on the computational part of the work.

I am also grateful to Dr. Tim Menzen for his expert advice, constructive feedback, and the regular scientific discussions that greatly shaped the outcome of this thesis.

Further, thank you Prof. Dr. Wolfgang Frieß for being the co-referee of this thesis and the valuable scientific input at the weekly seminar.

I am grateful to my collaboration partners from Peptomyc S.L., Dr. Marie-Eve Beaulieu and Dr. Virginia Castillo Cano, for giving me the chance to work on their interesting molecule and for sharing their scientific expertise.

Thanks to all colleagues from AK Winter, AK Frieß, and AK Merkel for the numerous scientific and social activities inside and outside the university. Special thanks to Dr. Hristo Svilenov for his endless support, to Dr. Andreas Tosstorff for dedicating a lot of time to helping me with the computational setup and the numerous scientific discussions, and to Dr. Simon Eisele for the joint work on the HY-133 project. Many thanks also to Ms. Kohler who always took care of everyone's concerns with heart and soul! I want to furthermore thank my students Rosanna Pellicano, Imge Su Yilmaz, Clarissa Gabler, Maria Siegle, and Alina Tiepermann.

Finally, I want to thank my family and close friends for their unlimited support, encouragement and love.

Table of Content

| | |
|--|----|
| Acknowledgements | 7 |
| Table of Content..... | 9 |
| List of abbreviations | 14 |
| <i>Chapter 1 - General introduction</i> | 18 |
| 1.1. Introduction | 18 |
| 1.2. Protein stability and degradation pathways | 19 |
| 1.3. Protein Aggregation..... | 21 |
| 1.4. Developability assessment..... | 22 |
| 1.5. Formulation development..... | 23 |
| 1.6. Computational tools..... | 24 |
| 1.7. Molecular dynamics (MD) simulations | 25 |
| 1.8. Aim and outline of the thesis | 26 |
| 1.9. References | 28 |
| <i>Chapter 2 - Computational approaches to assess nanobody stabilities and aggregation propensities</i> | 33 |
| 2.1 Introduction | 34 |
| 2.2. Materials and methods..... | 36 |
| 2.2.1. Sequence-based analyses..... | 36 |
| 2.2.2. Structure prediction using AlphaFold2 | 36 |
| 2.2.3. Structure-based aggregation and stability predictions | 37 |
| 2.2.4. MD simulations | 37 |
| 2.3. Results and discussion | 38 |
| 2.3.1. Experimental dataset | 38 |
| 2.3.2. Sequence-based approaches to explain the thermal stabilities of nanobodies | 38 |
| 2.3.3. Structure-based prediction of aggregation propensity and protein stability..... | 43 |
| 2.3.4. Partially unfolded species in the aggregation process of nanobodies | 47 |
| 2.4. Conclusions | 49 |
| 2.5. Supporting Information | 50 |
| 2.6. References | 56 |
| <i>Chapter 3 - Combining unfolding reversibility studies and molecular dynamics simulations to select aggregation-resistant antibodies</i> | 59 |
| 3.1 Introduction | 60 |

| | | |
|---|--|-----|
| 3.2 | Materials and methods..... | 62 |
| 3.2.1 | Proteins and chemicals | 62 |
| 3.2.2 | Isothermal chemical denaturation (ICD)..... | 62 |
| 3.2.3 | Size exclusion chromatography with multi-angle light scattering (SEC-MALS).... | 63 |
| 3.2.4 | ReFOLD assay | 63 |
| 3.2.5 | Storage stability study | 64 |
| 3.2.6 | Intrinsic differential scanning fluorimetry (nanoDSF)..... | 64 |
| 3.2.7 | Modulated scanning fluorimetry (MSF) | 64 |
| 3.2.8 | Dynamic light scattering (DLS) | 64 |
| 3.2.9 | Statistical analysis | 65 |
| 3.2.10 | Computational characterization..... | 65 |
| 3.3 | Results | 66 |
| 3.3.1 | Isothermal chemical denaturation and the ReFOLD assay to compare antibodies.. | 66 |
| 3.3.2 | Antibody unfolding, refolding and aggregation after heating..... | 68 |
| 3.3.3 | Aggregation during storage of the antibodies | 71 |
| 3.3.4 | Comparison between ReFOLD, MSF and aggregation during storage..... | 73 |
| 3.3.5 | Preliminary <i>in silico</i> characterization of the antibodies | 74 |
| 3.3.6 | Temperature-ramped molecular dynamics (MD) simulations | 75 |
| 3.4 | Discussion..... | 78 |
| 3.5 | Supporting information..... | 83 |
| 3.6 | References | 92 |
| <i>Chapter 4 - Investigation of the pH-dependent aggregation mechanisms of G-CSF using low resolution protein characterization techniques and advanced molecular dynamics simulations ...</i> | | 98 |
| 4.1 | Introduction | 99 |
| 4.2 | Methods | 101 |
| 4.2.1 | Conventional Molecular Dynamics Simulations..... | 101 |
| 4.2.2 | Metadynamics | 102 |
| 4.2.3 | CG Simulations | 103 |
| 4.2.4 | Materials..... | 104 |
| 4.2.5 | Sample Dialysis and Preparation..... | 104 |
| 4.2.6 | Intrinsic Fluorescence Spectroscopy | 105 |
| 4.2.7 | Circular Dichroism (CD) Spectroscopy | 105 |
| 4.2.8 | Differential Scanning Fluorimetry (nanoDSF) | 105 |
| 4.2.9 | Dynamic Light Scattering (DLS) | 106 |
| 4.2.10 | Small Angle X-ray Scattering (SAXS) | 106 |

| | | |
|---|---|------------|
| 4.3 | Results and Discussion | 107 |
| 4.3.1 | pH-dependent structural differences of G-CSF | 107 |
| 4.3.2 | Effect of pH and Sodium Chloride on the Thermal Unfolding and Aggregation of G-CSF | 112 |
| 4.3.3 | Future perspectives and potential challenges | 128 |
| 4.4 | Conclusions | 129 |
| 4.5 | Supporting information..... | 130 |
| 4.6 | References | 130 |
| <i>Chapter 5 - Interaction of HEPES with the therapeutic protein G-CSF – A combination of molecular dynamics simulations and biophysical characterization</i> | | <i>136</i> |
| 5.1 | Introduction | 137 |
| 5.2 | Material and methods | 139 |
| 5.2.1 | Materials..... | 139 |
| 5.2.2 | Sample Dialysis and Preparation..... | 139 |
| 5.2.3 | Circular Dichroism (CD) spectroscopy..... | 140 |
| 5.2.4 | Differential Scanning Fluorimetry (nanoDSF) | 140 |
| 5.2.5 | Dynamic Light Scattering (DLS) | 140 |
| 5.2.6 | Reversed-Phase High-Performance Liquid Chromatography (RP-HPLC)..... | 141 |
| 5.2.7 | Molecular dynamics simulations..... | 141 |
| 5.3 | Results and discussion | 143 |
| 5.3.1 | HEPES influences thermal stability and aggregation of G-CSF in a concentration-dependent manner | 143 |
| 5.3.2 | HEPES induces minor structural changes in G-CSF | 146 |
| 5.3.3 | HEPES directly binds to G-CSF in MD simulations | 149 |
| 5.4 | Conclusions | 153 |
| 5.5 | Supporting Information | 154 |
| 5.6 | References | 158 |
| <i>Chapter 6 - Stabilization of a novel recombinant bacteriophage endolysin by protein-exciipient interaction with HEPES and other Good’s buffers</i> | | <i>160</i> |
| 6.1. | Introduction | 161 |
| 6.2. | Materials and methods | 163 |
| 6.2.1. | Protein and chemicals..... | 163 |
| 6.2.2. | Reversed phase high pressure liquid chromatography (RP-HPLC)..... | 164 |
| 6.2.3. | Ion exchange chromatography (IEX)..... | 164 |
| 6.2.4. | Size exclusion chromatography (SEC) | 164 |
| 6.2.5. | Subvisible particle analysis (FlowCam)..... | 165 |

| | | |
|--|---|-----|
| 6.2.6. | Intrinsic differential scanning fluorimetry (nanoDSF)..... | 165 |
| 6.2.7. | Circular dichroism (CD) spectroscopy..... | 165 |
| 6.2.8. | Composition-gradient multi-angle light scattering (CG-MALS)..... | 165 |
| 6.2.9. | Activity: FRET-Assay..... | 166 |
| 6.2.10. | Storage stability study | 166 |
| 6.2.11. | Molecular dynamics (MD) simulations..... | 166 |
| 6.3 | Results and discussion | 167 |
| 6.3.1. | Influence of buffer excipients on chemical stability | 167 |
| 6.3.2. | Function of HEPES as stabilizing excipient | 168 |
| 6.3.3. | Influence of different Good's buffers on protein-excipient interactions..... | 172 |
| 6.3.4. | Stability study of HY-133 with HEPES and further Good's buffers | 176 |
| 6.4. | Conclusions | 184 |
| 6.5. | References | 186 |
| <i>Chapter 7 - Mechanistic understanding of the stability of the anti-MYC mini-protein Omomyc in different formulation conditions.....</i> | | 190 |
| 7.1. | Introduction | 191 |
| 7.2. | Materials and methods..... | 193 |
| 7.2.1. | Materials..... | 193 |
| 7.2.2. | Sample Dialysis and Preparation..... | 194 |
| 7.2.3. | Circular Dichroism (CD) Spectroscopy | 194 |
| 7.2.4. | Dynamic Light Scattering (DLS) | 194 |
| 7.2.5. | Differential Scanning Fluorimetry (nanoDSF) | 195 |
| 7.2.6. | Molecular dynamics simulations..... | 195 |
| 7.2.7. | Calculation of binding free energy by MM/PBSA..... | 196 |
| 7.3. | Results and Discussion | 197 |
| 7.3.1. | Control of electrostatics is critical in homo- and heterodimerization of MYC, MAX, and Omomyc | 197 |
| 7.3.2. | Effect of pH and ionic strength on Omomyc structure and stability..... | 200 |
| 7.3.3. | Concentration-dependent influence of phosphate buffer on the Omomyc stability 204 | |
| 7.3.4. | Phosphate binds to Omomyc in a concentration-dependent manner in MD simulations | 208 |
| 7.4. | Conclusions | 212 |
| 7.5. | Supporting information..... | 213 |
| 7.6. | References | 220 |
| <i>Chapter 8 - Summary of the thesis.....</i> | | 223 |

Appendix228
I. List of publications associated with this thesis228

List of abbreviations

APRs – aggregation-prone regions

bHLHLZ – basic helix-loop-helix leucine zipper

BLI – bio-layer interferometry

CBD – cell wall binding domain

CD – circular dichroism

CDRs – complementarity-determining regions

CG – coarse-grained

CG-MALS – composition-gradient multi-angle light scattering

C_{m1} – melting denaturant concentration for the first unfolding from a three-state fit

C_{m2} – melting denaturant concentration for the second unfolding from a three-state fit

cMD – conventional molecular dynamics

COM – center of mass

CV – collective variable

D_0 – infinite dilution diffusion coefficient

DSF – differential scanning fluorimetry

DLS – dynamic light scattering

EAD – enzymatic active domain

ESRF – European Synchrotron Radiation Facility

Fab – antigen-binding fragment

FES – free energy surface

FF – force field

FI₃₅₀/FI₃₃₀ – fluorescence intensity ratio

FR – framework region

FRET – Förster resonance energy transfer

Fv – fragment variable consisting of the V_H and V_L

G-CSF – granulocyte-colony stimulating factor

GuHCl – guanidine hydrochloride

HEH – protonated HEPES molecule

HEH₂ – double protonated HEPES molecule

HEM – deprotonated HEPES molecule

HEPES – N-2-hydroxyethylpiperazine-N-2-ethane sulfonic acid

HDX – hydrogen deuterium exchange

HY-133 – recombinant bacteriophage endolysin

ICD – isothermal chemical denaturation

IDP – intrinsically disordered protein

IEX – ion exchange chromatography

k_D – diffusion interaction parameter

K_D – dissociation constant

MAX – Myc-associated factor X

MD – molecular dynamics

MM/PBSA – molecular mechanics Poisson–Boltzmann surface

MSF – modulated scanning fluorimetry

MWCO – molecular weight cut-off

MYC – Myelocytomatosis proto-oncogene transcription factor

NaAc – sodium acetate

NaP – sodium phosphate

NMR – nuclear magnetic resonance

PDB – Protein Data Bank

pI – isoelectric point

PISA – Proteins, Interfaces, Structures and Assemblies

$P(I_{\text{score}})$ – interaction score probability

pLDDT – predicted local-distance difference test

PME – particle-mesh Ewald

PPI – protein-protein interactions

RC – reentrant condensation

rcf – relative centrifugal speed

$r(H)_0$ – infinite dilution value for the hydrodynamic radius

RMSD – root mean square deviations

RMSF – root mean square fluctuations

RMY – relative monomer yield

RP-HPLC – reversed-phase high-performance liquid chromatography

SAP – spatial aggregation propensity

SASA – solvent accessible surface area

SAXS – small-angle X-ray scattering

scFV – single-chain variable fragment

SEC – size-exclusion chromatography

SEC-MALS – size exclusion chromatography with multi-angle light scattering

SIRAH – South American Initiative for a Rapid and Accurate Hamiltonian

T_{agg} – aggregation onset temperature from dynamic light scattering

T_{m1} – first melting temperature from nanoDSF

T_{m2} – second melting temperature from nanoDSF

T_{nr} – non-reversibility onset temperature from MSF

T_{on} – unfolding onset temperature from nanoDSF

UHPLC – ultra high-performance liquid chromatography

V_{H} – antibody variable domain of the heavy chain

V_{HH} – V_{H} from heavy chain antibody

V_{L} – antibody variable domain of the light chain

Chapter 1 - General introduction

1.1. Introduction

Due to their high effectivity and limited side effects, biopharmaceuticals play an increasingly important role in the treatment of a variety of diseases, including diabetes, metabolic disorders, and cancer.¹ Compared to small molecules, proteins offer complex functions and specificity but also exhibit inherent physical and chemical instabilities.² However, the protein's instabilities are not solely determined by the intrinsic property of the sequence but largely by the buffer composition.³ Therefore, not only a thorough developability assessment of the protein is of utmost importance, but also an adequate formulation development to control the protein's stability during its lifetime.⁴ Protein stability is a main quality attribute to determine the drug-like properties of a candidate molecule or a formulation. However, calculating the stability-indicating parameters of a formulation or candidate molecule can be a tedious and resource-intensive process since they are usually determined by a variety of experimental screens. So far, no specific parameter is universal to indicate stability under various stress conditions such as elevated temperatures, freeze-thaw, and agitation or to predict the long-term shelf-life. Especially protein aggregation is a long-standing issue in biopharmaceutical technology which is difficult to predict.¹ Additionally, the suitability and predictive power of certain stability-indicating parameters have been questioned in several studies.⁵⁻⁷ In general, advanced knowledge of the protein structure, dynamics, degradation pathways, and protein-excipient interactions accelerates the protein developability assessment and formulation process. In the last years, a significant improvement in instrumentation for high-throughput applications with minimal sample consumption could be observed. Additionally, recent trends to move experimental screens *in silico* are useful to limit or eliminate required experimental characterization, and to obtain new strategies and guidance to perform efficient experiments.

Hereafter, the four “types“ of protein stability, the degradation pathways with a focus on protein aggregation, as well as biophysical techniques for protein characterization are discussed. Additionally, a short description of the common approaches for developability assessment and formulation development of therapeutic proteins is given. Computational tools to predict protein stability and aggregation parameters are also introduced as well as the application of molecular dynamics (MD) simulations in the field of therapeutic protein development.

1.2. Protein stability and degradation pathways

Protein stability is determined by conformational, colloidal, interfacial, and chemical stability.⁸ The conformational stability of a protein is defined as the free energy change (ΔG) for the transition from the native (folded) state to the denatured (unfolded) state. This energy change depends on the three major stabilizing forces present in the protein – hydrophobic interactions, hydrogen bonding, and electrostatic interactions.^{9,10} Hydrophobic interactions in the protein lead to the formation of hydrophobic clusters which are especially important for the stability of larger proteins.¹⁰ However, burying charged residues in these nonpolar environments comes with large energetic penalties.¹¹ Furthermore, polar and charged residues can form hydrogen bonds which play an important role in protein stability, but the contribution is highly context-dependent. Nevertheless, hydrogen bonds by side chains and peptide groups contribute equally for both, small and large proteins.¹² Additionally, charged residues can be involved in charge–charge/ electrostatic interactions which can be strong even at a distance, therefore called “long-range” interactions.¹¹ The protonation state of ionizable residues, and thus their electrostatic interaction potential, can be controlled either by alterations of the protein conformation or by changes in the solution pH. The strength of these electrostatic interactions can be adjusted by variations in salt concentrations.¹¹ The native state is also energetically favored by the hydration of protein surface residues.¹³ Suboptimal conditions tend to weaken these stabilizing forces and favor local flexibility, leading to protein degradation. An important pathway of protein aggregation involves conformational alterations, partial or complete protein unfolding accompanied by the increased exposure of hydrophobic regions.¹⁴ Indeed, the conformational stability of many proteins has been shown to correlate with their aggregation kinetics.^{15,16} Techniques that are commonly applied to assess the protein conformational stability include differential scanning calorimetry (DSC), differential scanning fluorimetry (DSF), circular dichroism (CD) spectroscopy, ultraviolet (UV) spectroscopy, Fourier-transform infrared (FTIR) spectroscopy, fluorescence spectroscopy, and isothermal chemical denaturation (ICD).

Colloidal stability is the tendency of protein molecules to associate in response to attractive, weak forces that are caused by hydrophobic surface residues and electrostatic interactions. The weak net interactions between protein molecules in solution are the sum of long-range electrostatic, short-range attractive, and hard-sphere interactions.^{8,17} The interactions between proteins can be assessed with light scattering techniques. Static light scattering is used to determine the second virial

coefficient (B_{22}) and dynamic light scattering to determine the protein interaction parameter (k_D).¹⁸ A positive B_{22} is indicative of repulsion between molecules and thus less probability for self-association. A negative B_{22} indicates more attractive forces between the protein molecules where protein–protein interactions are favored over protein–solvent interactions. This in turn results in colloidal instability of the protein which could potentially lead to aggregation.¹⁴ By increasing protein interactions via oppositely charged regions or by interactions with ions in solution, colloidal stability can also contribute to aggregation.¹⁴ The aggregation onset temperature (T_{agg}) is a parameter that is often used to investigate the colloidal stability of proteins. Note that the T_{agg} is often related to the protein melting temperature (T_m) and does not only reflect on the net protein–protein interactions. Additionally, techniques such as analytical ultracentrifugation, small angle X-ray scattering (SAXS), nuclear magnetic resonance (NMR), precipitation with polyethylene glycol, and bio-layer interferometry (BLI) can be applied to determine the colloidal stability.

The interfacial stability is related to the behavior of proteins at air-liquid, solid-liquid, or liquid-liquid interfaces. Since many proteins are surface active, they have a high propensity to interact with and adsorb to different types of interfaces. This could lead to protein destabilization and protein loss. Due to stresses like freezing or shaking, new or larger interfaces are formed, often resulting in severe degradation of the protein.^{8,19–21} The interfacial stability is determined by subjecting the protein to stresses like shaking, stirring, pumping, or freeze-thaw. Subsequently, the samples are tested for remaining monomer in solution, aggregation, and particle formation.

Modifications of covalent bonds, e.g., deamidation, oxidation, hydrolysis, and disulfide bond exchange, determine the chemical stability of a protein. Oxidation of histidine, tryptophan, tyrosine, methionine, and cysteine groups, as well as deamidation of Asn and Gln residues to Asp and Glu, are the two most common chemical degradation pathways.²² Metal ions, oxygen and light exposure, and high temperatures are some of the factors that accelerate oxidation. While oxidation is rarely affected by solution properties, the rate of deamidation is pH-dependent and exhibits a minimum between pH 3 and 6.⁸ Hydrolysis is the non-enzymatic cleavage of solvent-exposed peptide bonds, which is also pH and temperature dependent.²³ Disulfide bond shuffling can result in an altered conformation of the protein leading to a loss of function.²⁴ The chemical modifications are usually assessed by liquid or reversed-phase chromatography coupled to mass spectrometry, ion exchange chromatography, hydrophobic interaction chromatography, isoelectric focusing, and reduced/ non-reduced gel electrophoresis.

1.3. Protein Aggregation

Protein aggregation is a major challenge in the drug development process since it may reduce the biopharmaceutical's activity and potentially lead to an immunogenic reaction in the patient.²⁵ Protein aggregation occurs through different pathways, either via self-association of the native state (native aggregation), via unfolding intermediates and unfolded states (non-native aggregation), or through chemical degradations.²⁶ Several proteins can directly interact from the native state to form aggregates. Depending on the environmental conditions, the association can be electrostatic only or a combination of electrostatic and hydrophobic. Self-association often results in reversible oligomers/aggregates which can be precursors of irreversible aggregates.²⁷ It must be noted that the association process is not limited to one specific site of the protein but can involve multiple aggregation-prone regions (APRs).²⁸ It has been shown that the self-association of the native state is mainly related to colloidal stability. Therefore, the second virial coefficient B_{22} or protein interaction parameter k_D are measures of the tendency of protein-protein self-association.²⁶

Aggregation through unfolding intermediates or unfolded states is on the other hand determined by both conformational stability, and colloidal aspects. Even at normal conditions, the native state is in equilibrium with a small population of unfolding intermediates which were found to be precursors of the aggregation process. The intermediates have more surface-exposed hydrophobic patches and higher flexibility compared to the folded state making them especially aggregation-prone. In contrast, native or completely unfolded proteins show a lower aggregation propensity since the hydrophobic side chains are either buried in the protein core or randomly scattered. Aggregates containing only a few monomers are soluble but as soon as they exceed certain size and solubility limits, start to precipitate.²⁶ In reality, it is difficult to differentiate between self-association and aggregation through unfolding intermediates. This is especially challenging since the unfolding intermediates are poorly populated and cannot be investigated separately experimentally.

For the sake of completeness, the aggregation through chemical crosslinking and degradation is also mentioned, even though it will not be further addressed in this thesis. The most observed crosslinking is the intermolecular disulfide bond formation which can further promote physical aggregation of proteins.²⁹ Additionally, many chemical degradations have been shown to increase the aggregation propensity of proteins by changing their physical properties. For example,

oxidation,³⁰ dimerization,³¹ deamidation,³² hydrolysis,³³ and glycation³⁴ have been shown to affect protein hydrophobicity, secondary and tertiary structure, and the thermodynamic barrier to protein unfolding.

The (biological) consequence of protein aggregation furthermore depends on its reversibility, meaning the ability of the aggregates to dissociate, e.g., when changing the solution conditions to non-aggregation-inducing conditions. Early-stage aggregation for example tends to be reversible while late-stage and thermally induced aggregation are mostly irreversible.²⁶

1.4. Developability assessment

Only a small percentage of drug candidates entering development reach commercialization. Most candidates fail due to issues related to their design and molecular characteristics, as well as the applied manufacturing and delivery strategies.³ Research efforts during the discovery phase typically revolve around finding the best possible binder to the desired target. Due to the now frequent use of directed evolution and *in silico* tools, even more candidates with high binding affinities are identified. This trend requires better screening for suitable drug-like biophysical properties in early stages to ensure that risk of failure is detected as early as possible and as cheaply as possible in the development life cycle.³⁵ In this context, the concept of developability assessment was introduced: Developability is a broad term covering the evaluation of the suitability of a drug candidate to be successfully developed regarding its ability to meet adequate quality, manufacturability, effectiveness, and safety requirements.³ Biophysical parameters that are more readily accessible are often used to predict these properties indirectly. However, it is impossible to measure all physicochemical and stability issues at an early development stage when large amounts of material are not available and there is no single property that can predict the success of a candidate in later clinical stages. Therefore, a combination of various methods which require low sample amounts is used to flag variants with unfavorable biophysical properties.³⁶ These properties include specificity, solubility, colloidal stability, conformational stability, resistance to aggregation, interfacial stability, and chemical stability.³⁷ Lately, several computational methods have emerged that aim to predict these properties from the sequence or the 3-dimensional structure of the protein.³⁸ These are partly discussed later in this chapter. The *in silico* approach greatly reduces the number of required lab experiments which saves time and precious material. However, it is still not possible to reliably predict all physicochemical properties from the sequence.

Especially the propensity of non-native protein aggregation and poor conformational stability are hard to predict. Additionally, the computational tools rarely take into account the influence of different formulation components. However, the predictions from sequences, including the effect of formulation conditions, are expected to improve drastically as new large experimental datasets emerge.³⁷ Particularly, new approaches combining machine learning (ML) with high-throughput experimental techniques are promising screening tools for developability assessment.³⁶

1.5. Formulation development

Formulation development is a critical step in the development of a commercial protein product to prevent damage to the protein during production, processing, and long-term storage. One primary aim is to find optimized solution conditions that ensure that all four types of protein stability are sufficiently high. It has been demonstrated over decades that there is no general approach to stabilizing proteins and that each protein requires a customized formulation.^{22,39} Therefore, a wide variety of excipients such as buffering agents, sugars, salts, amino acids, and surfactants is usually screened to identify a stable formulation.⁴⁰ These are part of a library of generally regarded as safe (GRAS) excipients.⁴¹ Their efficiency has been evaluated in various experiments, and different stabilizing mechanisms have been proposed.

The most important solution factor is pH. The pH influences protein stability by controlling the surface charge, which can affect conformational and colloidal stability. In turn, it can significantly enhance or reduce protein aggregation and the rate of chemical degradations, e.g. deamidation and hydrolysis.⁸ Additionally, the addition of salts or modifications in the ionic strength can control the protein stability by neutralization of the protein surface charge, also called charge-screening effect. This leads to enhanced hydrophobic interactions between the monomers.⁴² Excipients that enhance conformational stability shift the equilibrium from the non-native to the native state of the protein, thereby decreasing the concentration of unfolded or partly unfolded aggregation-prone species in solution.¹⁶ This can be achieved by either preferential binding to the native state or preferential exclusion from the protein surface.⁴³ For proteins prone to oxidation, the formulation often contains antioxidants such as ascorbic acid or sacrificial additives such as methionine.⁸ Nonionic surfactants such as PS 20 and PS 80 are often used to reduce interfacial stress and to prevent adsorption, protein denaturation, and aggregation at interfaces.⁸ However, if the formulation is not properly designed, excipients might even have a detrimental effect.⁴⁴ Thus, understanding protein degradation

pathways, especially aggregation, and protein-excipient interactions is fundamental to designing optimized formulations. Unfortunately, our current knowledge about the interplay between protein structure, formulation, and stability is limited.

1.6. Computational tools

In recent years, the availability of increasingly powerful computers and algorithms enabled the development of computational methods that aim to predict and optimize important protein properties. Various predictive tools are available to assess different aspects of developability, including aggregation propensity, developability potential, and solubility. Many of these tools, especially those focusing on the prediction of aggregation propensity, have recently been reviewed.^{38,45,46} *In silico* tools for protein aggregation can be classified into aggregation-prone region (APR) and aggregation propensity predictors, and aggregation kinetics predictors. The approaches to predict APRs and aggregation propensity can be further divided into tools that require only the amino acid sequence as input and tools that use the 3-dimensional structure of the protein for the prediction.⁴⁶

Sequence-based prediction tools include for example CamSol, AGGRESCAN, Zygggregator, Pafig, PAGE, WALTZ, SALSA β -strand contiguity, TANGO, SecStr, NetCSP, PASTA, and AMYLPRED2. The different algorithms are based on different aggregation propensity scales for amino acids. These are determined from *in vivo* experiments, α -helix and β -sheet propensity, hydrophobicity, charge, the presence of gatekeeper residues, physicochemical properties, hydrogen bond probabilities, residue interactions, and solvation energies, to name a few. The structure-based prediction tools include SAP, DI, AGGRESCAN3D 2.0, Protein-Sol, and structurally corrected CamSol and are based on accessible surface area and surface patches. More detailed information on the algorithms can be found in the above-mentioned reviews.

The application of such predictive methods offers a time- and cost-effective complement or even an alternative to tedious experiments. Furthermore, *in silico* tools allow a potentially unlimited throughput, and many sequences can be screened in parallel. However, computational tools are only as good as the underlying data used to build them. Therefore, it must always be considered that the predictive accuracy can be low.³ Furthermore, sequence-based predictors usually neglect the effect of the folded structure on the parameter they are predicting. Globular proteins commonly possess large aggregation-prone regions, which drive the folding process and are buried in the

protein core in the native state.⁴⁵ Structure-based predictors correct for these buried APRs; however, experimental structures are usually not available at early development stages. Therefore, structures can be modeled with different tools, e.g. MODELLER, AlphaFold, or MOE. It must be noted though, that structure-based prediction tools can be sensitive to errors in the atomic coordinates. The CamSol structurally corrected method was for example shown to only be sensitive to residue solvent exposure so that the prediction did not change substantially depending on which model was selected as input.⁴⁵

1.7. Molecular dynamics (MD) simulations

MD simulations have become an indispensable complement to experiment, as they can in principle provide molecular mechanistic insights into the (thermodynamic) properties of proteins and even protein function over limited timescales.^{47,48} They can help interpret experiments, explain unexpected results, and guide experiments. MD simulations mimic the protein's behavior in real life on a molecular level, like a molecular microscope for properties that cannot be observed experimentally due to very short time and very small spatial scales. Proteins can adopt numerous conformational states, described by atomic coordinates and energy levels, which are organized in the energy landscape.⁴⁹ Exploring this energy landscape enables the observation of structural fluctuations over time.⁵⁰ However, MD simulations are dependent on a suitable energy function to describe this energy landscape: atoms are described as charged points connected by springs corresponding to bonds. The forces that are exerted on individual atoms are defined by a molecular mechanics force field which is developed to reproduce experimental data.⁵¹ The force field contains contributions of many terms that represent different types of interactions between the atoms of the protein including bond energy, bond angles, dihedral angles, improper dihedral angles, Van der Waals, electrostatic interactions, and hydrogen bonds. Based on the positions of the other atoms, the force experienced by each atom is determined. This force is used to numerically solve the classical equation of motion (Newton's equation), calculate accelerations and velocities, and update the atom coordinates.⁵² This results in a sequence of conformations ("frames") explored over time, containing atomic positions, velocities, and single-point energies, called "trajectory". If the simulation was run long enough, the average of the trajectory that reached convergence describes a statistical ensemble of the simulated molecular system. Thus, statistical mechanics can be applied to connect the distribution and motions of atoms to macroscopic observables for

example molecular size and shape, flexibilities, interactions with other molecules, and the relative frequency of one state or conformation compared to another.^{50,53}

In this thesis, the AMBER biomolecular simulation package with AMBER force fields was used for all-atom MD simulations.⁵⁴ It must be noted that covalent bonds cannot break or form during MD simulations, even though some bonds form and break more frequently in real life, e.g. disulfide bridges. Furthermore, as the integration of movement is done numerically, time steps shorter than the fastest movements in the molecule must be used. This in turn limits the simulation time and thus full sampling of the conformational states on biologically relevant timescales.⁵³ This is also relevant for MD simulations to evaluate protein folding or aggregation processes, which exceed the commonly possible simulation times. Due to their large computational cost, atomistic simulations are not applied to evaluate these processes. Instead, the computational burden is reduced at the cost of accuracy and structural resolution by employing coarse-grained (CG) models. CG models consist of beads describing an approximated collective property of multiple atoms in the protein. For these simulations, the GROMACS software was employed in combination with the SIRAH force field.^{55,56} The explicit water model contains four beads in a tetrahedral shape, while the protein backbones are described with three beads that represent the nitrogen, α -carbon, and oxygen part. The protein side chains are described as one to five beads, depending on the physicochemical properties.⁵⁶

1.8. Aim and outline of the thesis

This thesis aimed to assess the potential of computational tools and MD simulations in the context of developability assessment and formulation development. All computational results were validated with experimental data using common biophysical characterization techniques relevant to developability assessment and formulation development. The current strategy in the development of therapeutic proteins consists of a thorough characterization of the protein structure, stability, and degradation pathways via extensive experimental screens. Recently, computational tools have gained increased importance in biotherapeutic technology. MD simulations can be used to explain observations during experimental formulation work to improve the future design of experiments and obtain new ideas for formulations. The ultimate goal is to use reliable predictive computational tools to dramatically reduce experimental excipient screens.

In *Chapter 2* and *Chapter 3*, available web servers were applied for developability assessment and the results were compared to experimentally determined stability data of 13 monoclonal antibodies and 68 nanobodies. Additionally, structural models of the proteins were generated and used for structure-based predictions which were critically compared to the experimental results. Since these web servers only consider aggregation from the native state and several studies revealed that the most critical factor for long-term stability is non-native aggregation, MD simulations at elevated temperatures were applied to generate the partially unfolded, reactive species responsible for this degradation mechanism. These altered conformations were then examined for their aggregation propensity.

In *Chapter 4*, the pH-dependent aggregation mechanism of granulocyte colony-stimulating factor (G-CSF) was elucidated by combining experimental characterization of the protein at three pH values on a structural level and all-atom simulations, metadynamics simulations, and coarse-grained MD simulations.

In *Chapter 5*, the different effects of the zwitterionic buffer HEPES on G-CSF at various concentrations and under different stress conditions were elucidated by biophysical characterization and the evaluation of HEPES binding to the protein in MD simulations.

In *Chapter 6*, the stabilization of a novel recombinant bacteriophage lysin by HEPES and other Good's buffers was explained experimentally and by examining excipient binding and subsequent conformational changes in MD simulations.

In *Chapter 7*, the stability of the anti-MYC mini-protein Omomyc at different pH values and ionic strengths was assessed employing molecular mechanics Poisson–Boltzmann surface area (MM-PBSA) calculations and structural characterization. Additionally, the influence of phosphate anions on protein folding and reentrant condensation was explained on a molecular level.

Finally, *Chapter 8* provides a summary of the presented results on the application of MD simulations in developability assessment and formulation development in the context of the current state of the art and an outlook on the future benefit of using computational methods to reduce experimental effort.

1.9. References

1. Roberts, C. J. Therapeutic protein aggregation: Mechanisms, design, and control. *Trends Biotechnol* **32**, 372–380 (2014).
2. Leader, B., Baca, Q. J. & Golan, D. E. Protein therapeutics: A summary and pharmacological classification. *Nat Rev Drug Discov* **7**, 21–39 (2008).
3. Zurdo, J. Developability assessment as an early de-risking tool for biopharmaceutical development. *Pharm Bioprocess* **1**, 29–50 (2013).
4. Wang, W. Advanced protein formulations. *Protein Science* **24**, 1031–1039 (2015).
5. Youssef, A. M. K. & Winter, G. A critical evaluation of microcalorimetry as a predictive tool for long term stability of liquid protein formulations: Granulocyte Colony Stimulating Factor (GCSF). *European Journal of Pharmaceutics and Biopharmaceutics* **84**, 145–155 (2013).
6. Thiagarajan, G., Semple, A., James, J. K., Cheung, J. K. & Shameem, M. A comparison of biophysical characterization techniques in predicting monoclonal antibody stability. *MAbs* **8**, 1088–1097 (2016).
7. Goldberg, D. S. et al. Utility of High Throughput Screening Techniques to Predict Stability of Monoclonal Antibody Formulations During Early Stage Development. *J Pharm Sci* **106**, 1971–1977 (2017).
8. Manning, M. C., Liu, J., Li, T. & Holcomb, R. E. Rational Design of Liquid Formulations of Proteins. *Adv Protein Chem Struct Biol* **112**, 1–59 (2018).
9. Pace, C. N., Shirley, B. A. & Thomson, J. A. Measuring the conformational stability of a protein. in *Protein structure: A practical approach* **2** 299–321 (1997).
10. Pace, C. N. et al. Contribution of hydrophobic interactions to protein stability. *J Mol Biol* **408**, 514–528 (2011).
11. Zhou, H. X. & Pang, X. Electrostatic Interactions in Protein Structure, Folding, Binding, and Condensation. *Chem Rev* **118**, 1691–1741 (2018).
12. Pace, C. N. et al. Contribution of hydrogen bonds to protein stability. *Protein Science* **23**, 652–661 (2014).
13. Huggins, D. J. Studying the role of cooperative hydration in stabilizing folded protein states. *J Struct Biol* **196**, 394–406 (2016).
14. 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).

15. Bai, S., Manning, M. C., Randolph, T. W. & Carpenter, J. F. Aggregation of recombinant human botulinum protein antigen serotype c in varying solution conditions: Implications of conformational stability for aggregation kinetics. *J Pharm Sci* **100**, 836–848 (2010).
16. Costanzo, J. A. et al. Conformational stability as a design target to control protein aggregation. *Protein Engineering, Design and Selection* **27**, 157–167 (2014).
17. Chi, E. Y., Krishnan, S., Randolph, T. W. & Carpenter, J. F. Physical stability of proteins in aqueous solution: Mechanism and driving forces in nonnative protein aggregation. *Pharm Res* **20**, 1325–1336 (2003).
18. le Brun, V., Friess, W., Bassarab, S. & Garidel, P. Correlation of protein-protein interactions as assessed by affinity chromatography with colloidal protein stability: A case study with lysozyme. *Pharm Dev Technol* **15**, 421–430 (2010).
19. Lee, H. J., McAuley, A., Schilke, K. F. & McGuire, J. Molecular origins of surfactant-mediated stabilization of protein drugs. *Adv Drug Deliv Rev* **63**, 1160–1171 (2011).
20. Koepf, E., Eisele, S., Schroeder, R., Brezesinski, G. & Friess, W. Notorious but not understood: How liquid-air interfacial stress triggers protein aggregation. *Int J Pharm* **537**, 202–212 (2018).
21. Arsiccio, A. & Pisano, R. The Ice-Water Interface and Protein Stability: A Review. *J Pharm Sci* **109**, 2116–2130 (2020).
22. Capelle, M. A. H., Gurny, R. & Arvinte, T. High throughput screening of protein formulation stability: Practical considerations. *European Journal of Pharmaceutics and Biopharmaceutics* **65**, 131–148 (2007).
23. Pan, B., Ricci, M. S. & Trout, B. L. A molecular mechanism of hydrolysis of peptide bonds at neutral pH using a model compound. *Journal of Physical Chemistry B* **115**, 5958–5970 (2011).
24. Manning, M. C., Patel, K. & Borchartd, R. T. Stability of Protein Pharmaceuticals. *Pharm Res* **6**, 903–918 (1989).
25. Frokjaer, S. & Otzen, D. E. Protein drug stability: A formulation challenge. *Nat Rev Drug Discov* **4**, 298–306 (2005).
26. Wang, W., Nema, S. & Teagarden, D. Protein aggregation - Pathways and influencing factors. *Int J Pharm* **390**, 89–99 (2010).
27. Wang, W. & Roberts, C. J. Protein aggregation – Mechanisms, detection, and control. *Int J Pharm* **550**, 251–268 (2018).

28. Kanai, S., Liu, J., Patapoff, T. W. & Shire, S. J. Reversible Self-Association of a Concentrated Monoclonal Antibody Solution Mediated by Fab–Fab Interaction That Impacts Solution Viscosity. *J Pharm Sci* **97**, 4219–4227 (2008).
29. Cabra, V., Vázquez-Contreras, E., Moreno, A. & Arreguin-Espinosa, R. The effect of sulfhydryl groups and disulphide linkage in the thermal aggregation of Z19 α -zein. *Biochim Biophys Acta Proteins Proteom* **1784**, 1028–1036 (2008).
30. Rosenfeld, M. A., Leonova, V. B., Konstantinova, M. L. & Razumovskii, S. D. Self-assembly of fibrin monomers and fibrinogen aggregation during ozone oxidation. *Biochemistry (Moscow)* **74**, 41–46 (2009).
31. Roostae, A., Côté, S. & Roucou, X. Aggregation and amyloid fibril formation induced by chemical dimerization of recombinant prion protein in physiological-like conditions. *Journal of Biological Chemistry* **284**, 30907–30916 (2009).
32. Takata, T., Oxford, J. T., Demeler, B. & Lampi, K. J. Deamidation destabilizes and triggers aggregation of a lens protein, β A3-crystallin. *Protein Science* **17**, 1565–1575 (2008).
33. van Buren, N., Rehder, D., Gadgil, H., Matsumura, M. & Jacob, J. Elucidation of two major aggregation pathways in an IgG2 antibody. *J Pharm Sci* **98**, 3013–3030 (2009).
34. Wei, Y., Chen, L., Chen, J., Ge, L. & He, R. Q. Rapid glycation with D-ribose induces globular amyloid-like aggregations of BSA with high cytotoxicity to SH-SY5Y cells. *BMC Cell Biol* **10**, 1–15 (2009).
35. Pritchard, J. F. et al. Making better drugs: Decision gates in non-clinical drug development. *Nat Rev Drug Discov* **2**, 542–553 (2003).
36. Narayanan, H. et al. Machine Learning for Biologics: Opportunities for Protein Engineering, Developability, and Formulation. *Trends Pharmacol Sci* **42**, 151–165 (2021).
37. Svilenov, H. L., Arosio, P., Menzen, T., Tessier, P. & Sormanni, P. Approaches to expand the conventional toolbox for discovery and selection of antibodies with drug-like physicochemical properties. *MAbs* **15**, 2164459 (2023).
38. Khetan, R. et al. Current advances in biopharmaceutical informatics: guidelines, impact and challenges in the computational developability assessment of antibody therapeutics. *MAbs* **14**, (2022).
39. Jaenicke, R. Stability and stabilization of globular proteins in solution. *J Biotechnol* **79**, 193–203 (2000).
40. Cheng, W. et al. Comparison of high-throughput biophysical methods to identify stabilizing excipients for a model IgG2 monoclonal antibody: Conformational stability and kinetic aggregation measurements. *J Pharm Sci* **101**, 1701–1720 (2012).

41. Bhambhani, A. et al. Formulation design and high-throughput excipient selection based on structural integrity and conformational stability of dilute and highly concentrated IgG1 monoclonal antibody solutions. *J Pharm Sci* **101**, 1120–1135 (2012).
42. Wang, W. & Ohtake, S. Science and art of protein formulation development. *Int J Pharm* **568**, (2019).
43. Timasheff, S. N. Control of protein stability and reactions by weakly interacting cosolvents: The simplicity of the complicated. *Adv Protein Chem* **51**, 355–432 (1998).
44. Falconer, R. J. Advances in liquid formulations of parenteral therapeutic proteins. *Biotechnol Adv* **37**, (2019).
45. Wolf Perez, A. M., Lorenzen, N., Vendruscolo, M. & Sormanni, P. Assessment of Therapeutic Antibody Developability by Combinations of In Vitro and *In Silico* Methods. *Therapeutic Antibodies: Methods and Protocols*, 57-113 (2022).
46. Prabakaran, R., Rawat, P., Thangakani, A. M., Kumar, S. & Gromiha, M. M. Protein aggregation: in *silico* algorithms and applications. *Biophys Rev* **13**, 71–89 (2021).
47. van Gunsteren, W. F. et al. Biomolecular modeling: Goals, problems, perspectives. *Angewandte Chemie - International Edition* **45**, 4064–4092 (2006).
48. Lindorff-Larsen, K. & Ferkinghoff-Borg, J. Similarity measures for protein ensembles. *PLoS One* **4**, e4203 (2009).
49. Frauenfelder, H. et al. A unified model of protein dynamics. *Proceedings of the National Academy of Sciences* **106**, 5129–5134 (2009).
50. Adcock, S. A. & McCammon, J. A. Molecular dynamics: Survey of methods for simulating the activity of proteins. *Chem Rev* **106**, 1589–1615 (2006).
51. Jorgensen, W. L. & Tirado-Rives, J. Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proc Natl Acad Sci USA* **102**, 6665–6670 (2005).
52. Caflisch, A. & Paci, E. Molecular Dynamics Simulations to Study Protein Folding and Unfolding. *Proteins* **5**, 1143–1169 (2005).
53. Leimkuhler, B. & Matthews, C. Molecular Dynamics. *Interdisciplinary applied mathematics* vol. **39** (2015).
54. Case, D. A. et al. The Amber biomolecular simulation programs. *J Comput Chem* **26**, 1668–1688 (2005).
55. Abraham, M. J. et al. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2**, 19–25 (2015).

56. Darré, L. et al. SIRAH: A structurally unbiased coarse-grained force field for proteins with aqueous solvation and long-range electrostatics. *J Chem Theory Comput* **11**, 723–739 (2015).

Chapter 2 - Computational approaches to assess nanobody stabilities and aggregation propensities

Berner, C.¹, Menzen, T.², Winter, G.¹

¹Department of Pharmacy, Ludwig-Maximilians-Universität München, Munich, Germany

²Coriolis Pharma Research GmbH, Martinsried, Germany

Author contributions:

CB., T.M. and G.W. conceived the idea. C.B. performed the analyses, evaluated the data, and wrote the manuscript. G.W. and T.M. provided conceptual guidance. G.W. corrected the manuscript.

2.1 Introduction

Nanobodies are derived from the variable binding domains of camelid heavy-chain antibodies (VHH) and exhibit unique properties, making them attractive research tools, diagnostics, and therapeutics.¹ Several nanobodies against various targets can be identified through animal immunization and in vitro methods like phage display. Screening for enriched phages is typically sufficient for the initial selection of binders, a more thorough characterization is however necessary to identify nanobodies with critical features like high thermal stability, low aggregation propensity, and high specificity.^{2,3} Nanobodies have been reported to generally have high solubility and high resistance to thermal unfolding with melting temperatures ranging from 60 to 80 °C.^{4,5} While several nanobodies have shown reversible thermal unfolding,⁵ others are unable to refold to their native conformations. Instead, the partially unfolded species tend to aggregate, which is known as non-native aggregation.⁶ The absence of aggregation below the onset temperature of unfolding furthermore highlights that protein unfolding represents the rate-limiting step in the aggregation process of these nanobodies.⁷ To identify aggregation-resistant molecules, strategies often focus on selecting candidates with high melting temperatures, which indicate a small population of unfolded species. However, not all partially unfolded species are equally prone to aggregation.⁸ To answer the question, if and how these partially unfolded species are involved in the aggregation mechanisms, it is necessary to understand the key features of nanobodies that show reversible thermal unfolding and of those that aggregate.

Compared to conventional mAbs, nanobodies have a convex paratope shape due to different CDR1 and CDR2 conformations, and CDR3 loops of unusual lengths.⁹ These features allow them to target hidden epitopes in small cavities, e.g., active sites of enzymes, with high affinity and specificity. In exchange for the divergent CDR conformations, nanobodies sacrifice effective packing of the upper core and therefore some of their thermodynamic stability.⁵ However, non-canonical disulfide bridges formed between Cys residues of CDR1 and CDR3, have been reported to contribute to high thermodynamic stability and good solubility.^{3,7} Increased solubility is further achieved by substitution of hydrophobic residues in the FR2 region (corresponding to the V_H-V_L interface in mAbs) to more hydrophilic residues in nanobodies.⁵ Moreover, the long CDR3 loops shield hydrophobic residues in FR2 and help to mask Trp118, which is key for V_H-V_L interaction.³ All these specific features reveal that, despite their small size of ~15 kDa, nanobodies are highly complex molecules, with every amino acid potentially having a direct or indirect impact on the

stability, structural integrity, and antigen-binding. This complex interplay of residues restricts the extent to which engineering can be tolerated.¹ The development of nanobodies with optimal stability profiles thus far involved shielding of aggregation-prone regions, increasing charge repulsion, and fine-tuning structural dynamics to prevent the formation of aggregation-prone conformations.⁷ However, the molecular determinants responsible for favorable thermal properties, reversible unfolding or aggregation are not yet fully understood. We aim to get insights into the underlying mechanisms by thorough *in silico* characterization. Kunz et al. kindly provided the sequences and experimental data of a dataset of 68 nanobodies including melting temperatures (T_m), onset temperatures of aggregation (T_{agg}), and aggregation kinetics for selected nanobodies. In this study, 78 % of the investigated nanobodies showed aggregation during thermal unfolding.^{7,10}

To potentially identify sequence features that could discriminate well- from poor-behaving nanobodies, we first performed a multiple sequence alignment and compared the amino acid compositions. In a second step, we selected publicly available aggregation prediction tools, which are easily applicable to our dataset for a comparative study, to predict aggregation propensities and aggregation-prone regions (APRs) of the proteins. We applied a sequence-based consensus approach with outputs from AGGRESCAN, NetCSSP, AmyloidMutants, Pafig, Amyloidogenic Pattern, SecStr, Average Packing Density, TANGO, β -strand contiguity, WALTZ, Hexapeptide Conformational Energy, and the intrinsic solubility profiles calculated with CamSol.¹¹ Additionally, sequence-based solubility scores from Tango, AGGRESCAN, CamSol and the Protein-Sol webserver¹² were compared to the experimental data. In the third step, we used AlphaFold2 to predict the structures of all 68 nanobodies from their sequences. For a structure-based aggregation prediction we applied AGGRESCAN3D 2.0¹³ and the structurally corrected CamSol score. Furthermore, the ratio of hydrophobic to total solvent accessible surface area (SASA) and the number of hydrophobic clusters were compared.

It is important to note that the differences captured with these *in silico* techniques are features of the native folded proteins. As discussed earlier, unfolding of the nanobodies is the rate-limiting step in the aggregation process. In the fourth step, we therefore estimated the overall protein stability with the Rosetta all-atom energy function.¹⁴ Nevertheless, not only the resistance to unfolding but also the aggregation propensity of the partially unfolded species must be considered. As a result, the final step in our approach was to perform molecular dynamics (MD) simulations to obtain partially unfolded species of the nanobodies and to assess their aggregation propensities.

2.2. Materials and methods

2.2.1. Sequence-based analyses

The sequences and experimental data were provided by Kunz et al.⁷ A multiple sequence alignment was generated with Clustal Omega.¹⁵ The theoretical pIs were calculated with the ProtParam tool of ExPASy (<http://www.expasy.ch/tools/protparam.html>).¹⁶ The hydrophathy of each nanobody was calculated based on the values proposed by Kyte and Doolittle.¹⁷ The web tool AMYLPRED2¹⁸ (<http://biophysics.biol.uoa.gr/AMYLPRED2>) was applied, including the following methods: AGGRESCAN, AmyloidMutants, Amyloidogenic Pattern, Average Packing Density, Beta-strand contiguity, Hexapeptide Conformational Energy, NetCSSP, Pafig, SecStr, Tango, and Waltz. In addition to the AMYLPRED2 results, intrinsic solubility profiles calculated with CamSol¹¹ (<https://www-cohsoftware.ch.cam.ac.uk/index.php/camsolintrinsic>) at pH 7.0 were considered in a consensus approach. Each amino acid residue was assigned a score of 1 if the algorithm predicted that residue to be aggregation-prone and a score of 0 if not. The sum of these scores per residue was calculated as the total score. Since 11 prediction algorithms were used, the maximum score for a given residue was 11. Any residue with a total score of ≥ 6 was defined to be aggregation-prone in this study. Additionally, sequence-based solubility scores from AGGRESCAN, Tango, CamSol, and the Protein-Sol webserver¹² (<https://protein-sol.manchester.ac.uk>) were determined.

2.2.2. Structure prediction using AlphaFold2

The structures of the 68 nanobodies are not available in the PDB. We therefore used the AlphaFold2 notebook on Google Collaboratory (ColabFold)¹⁹ to predict their 3D structures. The notebook requires FASTA-sequences as input, generates a multiple-sequence alignment (MSA) and uses this to predict the models. We used the default settings of three recycles, meaning that the prediction is fed through the model three times, and relaxation of the predicted structures using amber force fields. The notebook generates five models and ranks them based on two measures of confidence. The local reliability (predicted local-distance difference test, pLDDT) and the reliability of pairwise interactions between different residues in the chain.¹⁹ We selected the best ranked model, relaxed the structure with the Rosetta²⁰ FastRelax mover and the scoring function ref2015 on Google Collaboratory, and used the output for further analyses. For comparison, the RMSDs for each pair of nanobodies was calculated using PyMOL.

2.2.3. Structure-based aggregation and stability predictions

We used AGGRESCAN3D 2.0¹³ and the structurally corrected CamSol score at pH 7 and with a patch radius of 10 Å. Furthermore, the ratio of hydrophobic to total solvent accessible surface area (SASA) was calculated with the molecular visualization program VMD.²¹ Hydrophobic residues are packed inside the protein core and form clusters, which drive protein folding and stability. Additionally, hydrogen bond networks between multiple sidechains impact protein stability. The hydrophobic clusters were determined with ProteinTools.²² The number of hydrophobic clusters was counted, and the total area calculated. Scoring of the structures with the Rosetta score function ref2015 on Google Collaboratory gave an estimation of the protein's stability. While a lower scoring structure is more likely to be stable, the scores do not have a direct conversion to physical energy units like kcal/mol. The energies are instead represented in Rosetta Energy Units (REU).

2.2.4. MD simulations

All-atom simulations were performed with the Amber19 program in a periodic box with explicit solvent.²³ The ff14SB force field for proteins was employed in combination with the TIP3P water model. The protonation states of ionizable residues at pH 7.2 were adjusted using the H ++ server.²⁴ All bonds involving hydrogen atoms were constrained using the SHAKE algorithm. The nonbonded electrostatic interactions were treated using the particle mesh Ewald algorithm with a direct space cut-off of 10 Å. The models of the nanobodies were solvated in a truncated octahedral water box with a layer of at least 20 Å from the protein surface and neutralized with chloride or sodium ions. The system was energy minimized with the steepest descent algorithm for the first 5000 cycles, followed by 5000 cycles, using the conjugate gradient method. Subsequently, the system was heated to 450 K in an NVT ensemble. System equilibration was carried out for 1 ns in NVT ensemble to stabilize the specified temperature using the Langevin thermostat, and subsequently for 1 ns in NPT ensemble to adjust the density of the system using the Berendsen barostat. The simulations were performed for 100 ns with a time step of 1 fs. The coordinates were saved every 5 ps. Trajectories were visually inspected in VMD to verify that there are no dimensional constraints due to the chosen box size and shape. All trajectories were analyzed using the CPPTRAJ module of Amber19. The last frame of each simulation was saved as a PDB file and analyzed with A3D 2.0 and CamSol. The root-mean square fluctuations per residue were compared with the predicted sequence-based APRs.

2.3. Results and discussion

2.3.1. Experimental dataset

The dataset in our study includes 68 nanobodies with a wide range of thermal stabilities with T_m s ranging from ~ 48 °C to ~ 86 °C (Figure S1). The T_m s have been measured with two techniques: thermal shift assay with SYPRO[®] orange and intrinsic fluorescence-based nanoDSF. The determined values from the two techniques show a strong correlation with a Pearson's r of 0.88. We therefore only consider the T_m Sypro in the following comparisons with the *in silico* data. nanoDSF measurements simultaneously determined the onset temperatures of unfolding (T_{on}) and onset temperatures of aggregation (T_{agg}) detected by the backscattering optics of the device. Additionally, the ΔT_m shift has been determined, which is specified as the difference in T_m between low (13.1 μ M) and high (32.72 μ M) protein concentration. The aim of our study was to explain the different thermal stabilities and aggregation propensities of the nanobodies on a molecular level and potentially guide further engineering efforts with our findings. We therefore used a sequence- and structure-based approach and additionally included MD simulations to account for the impact of partially unfolded species.

2.3.2. Sequence-based approaches to explain the thermal stabilities of nanobodies

First, we performed a multiple sequence alignment of all 68 nanobodies and compared their sequence homologies. We found homologies ranging from ~ 53 % to ~ 99 % for the full sequences (FS) (Figure S2) and ~ 62 % to ~ 99 % for the framework regions (FR) only (Figure S3). These findings show that, besides the naturally high variability of the CDRs, the usually conserved framework regions contain several variations in this dataset. Interestingly, high sequence homology to an experimentally well behaving nanobody does not necessarily result in the same high thermal stability. An example for this is the nanobody pair NbPep39 and NbD8, sharing a sequence homology of ~ 82 % (FS) and ~ 91 % (FR) but a difference in T_m of ~ 30 °C. Strikingly, NbD7 has a very low sequence similarity to all other nanobodies in the dataset. Overall, there is no trend in sequence homology and stability revealing a diverse set of nanobodies in this study.

Next, we compared the frequencies of cysteines and charged residues within the nanobodies to find a correlation to their respective stabilities (Figure 1). As already described by Kunz et al., there is a weak correlation between the number of cysteines, thus disulfide bridges, and the thermal stabilities of the nanobodies.⁷ Furthermore, electrostatic interactions have been shown to play a

key role in protein aggregation. “Supercharging” proteins to have an increased net charge was reported to prevent aggregation of partially unfolded states and to promote refolding. For example, after high temperature exposure of a supercharged scFV variant, it refolds and retains epitope binding.²⁵ We therefore calculated the number of positively and negatively charged residues but only find weak correlations with the thermal stabilities. It must be noted that the experimental data is largely influenced by the pH and ionic strength. The relatively high ionic strength of the PBS buffer may have screened the electrostatic interactions. Additionally, the net charge, rather than the number of charged residues seems to be a better determinant of aggregation propensity, according to literature.²⁶ Thus, we conducted more thorough analyses of the nanobody sequences in the next step.

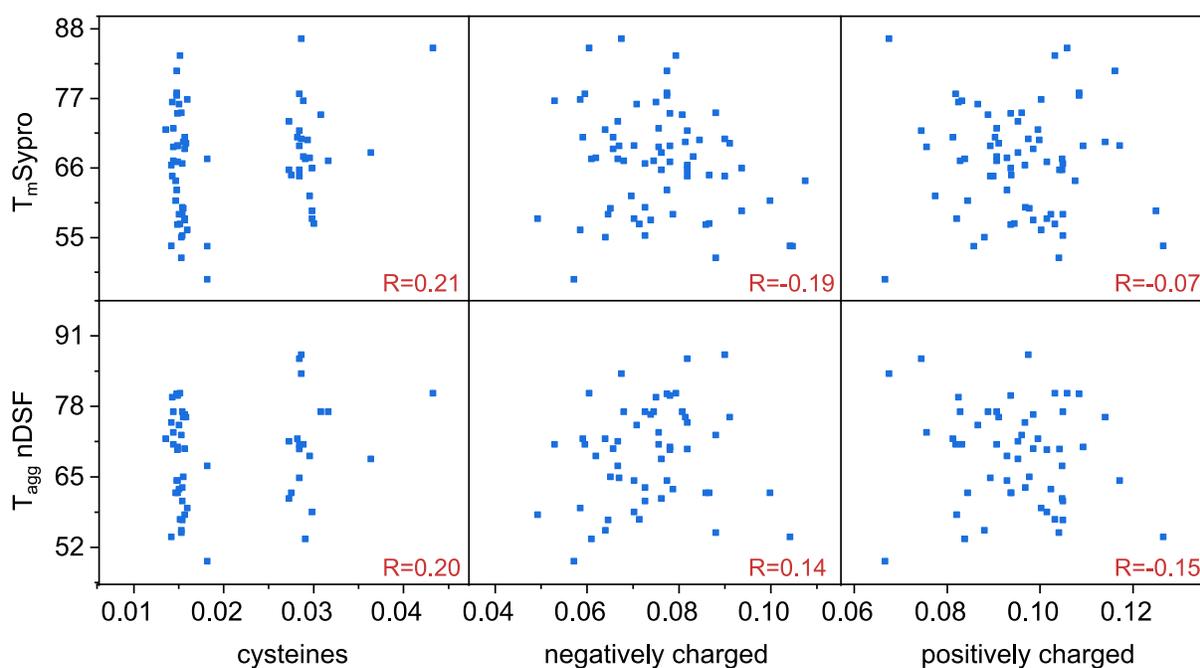


Figure 1: Fraction of cysteines, negatively charged and positively charged residues within each nanobody sequence compared to the experimentally determined T_m and T_{agg} values. The Pearson correlation coefficients were calculated with Origin 2019.

We evaluated sequence-derived metrics including the isoelectric points (pI), hydropathies and β -sheet propensities. To predict aggregation propensities and APRs of proteins, several computational methods have been developed. Depending on the input required for the prediction, these methods can be classified as sequence- or structure-based. While certain APR predictors are based on common features, e.g., amino acid hydrophobicity and β -sheet propensity, others differ

in their underlying principles. Detailed explanations of each of these predictors can be found in the review by Prabakaran et al.²⁷ Here, we used the ProteinSol, Tango, AGGRESKAN and CamSol webserver and compared the calculated data to the experimental data (Figure 2).

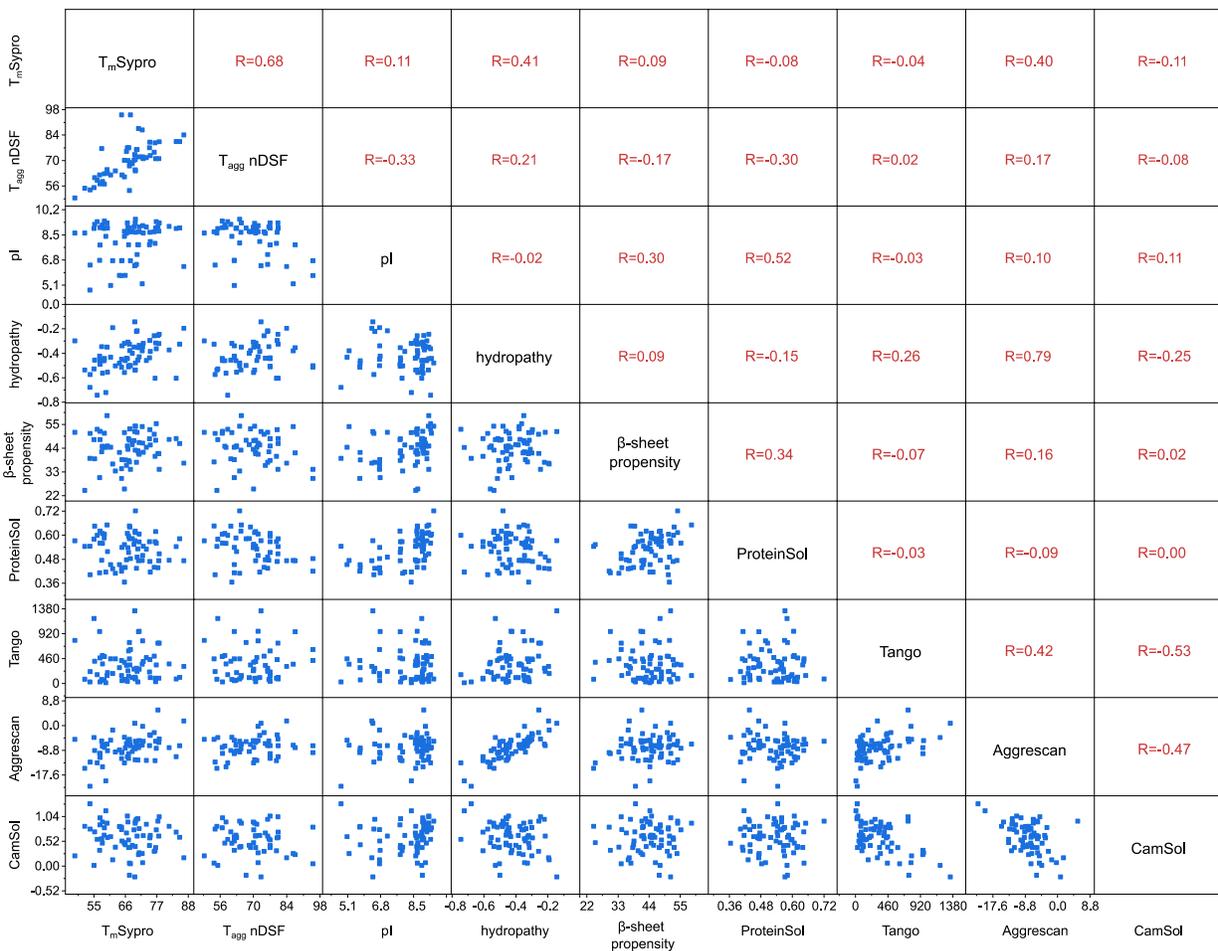


Figure 2: Correlation of sequence-derived molecular features including pI, hydropathy, β -sheet, and aggregation propensities with the thermal stability data of the 68 nanobodies. The Pearson correlation coefficients were calculated with Origin 2019.

It is known that proteins are least soluble at their pIs, where they have a net charge of zero. Most of the nanobodies in the dataset have a basic pI ≥ 8.5 showing a wide range of T_{ms} and T_{agg} . Surprisingly, the nanobodies with a pI close to the buffer pH of 7.5, still show relatively good behavior. There is no obvious correlation between the pI and thermal stability in this dataset. Next, the total hydropathy is calculated based on the hydropathy indices of each amino acid representing the hydrophobic or hydrophilic properties of the sidechains, as proposed by Kyte and Doolittle.¹⁷

The larger the hydropathy score, the more hydrophobic the molecule. Interestingly, we observe a moderate positive correlation of the hydropathy with the T_m values ($R=0.41$). This can potentially be explained by the formation of hydrophobic clusters and a better packing of the hydrophobic core, which in turn stabilizes the overall fold of the nanobody. On the other hand, more hydrophobic proteins tend to have a higher aggregation propensity. However, we again see a weak positive correlation of the hydropathy with the T_{agg} values ($R=0.21$) indicating that a high hydropathy leads to protein stabilization and thus prevents protein aggregation. Additionally, β -sheet propensity is another leading determinant of protein stability and aggregation.²⁶ The propensity of the amino acid residues to adopt α -helical or β -structural conformations has been used to calculate the overall propensity of the sequence to form a β -sheet structure. In this dataset, the β -sheet propensity does not correlate with the thermal stabilities.

To evaluate the aggregation propensities of the sequences, we selected publicly available aggregation prediction tools that are easily applicable to our dataset for a comparative study. Again, we compared the calculated overall aggregation scores with the experimental data. A ProteinSol solubility value >0.45 is predicted to have a higher solubility than the average soluble *E. coli* protein, meaning the higher the score, the more soluble the protein. As expected, there is no correlation with the T_m values but to our surprise a weak negative correlation with the T_{agg} values. Considering that a high ProteinSol score indicates higher solubility, we expected a positive correlation. However, the ProteinSol scores show a moderate correlation with the pI and a weak correlation with the β -sheet propensity, reflecting on the underlying algorithm, but no correlation with any of the other aggregation predictors. Since a higher Tango score indicates a higher aggregation propensity, a positive correlation is expected. However, there is no correlation with the experimental data. The AGGRESCAN score is calculated based on the hydropathy of the molecule which explains the strong correlation ($R=0.79$) of these values. A positive AGGRESCAN score per residue indicates aggregation propensity, whereas residues with no aggregation propensity are assigned a negative score. Overall, soluble molecules have a highly negative AGGRESCAN score whereas more aggregation prone proteins have higher AGGRESCAN scores. In contrast to the expectations, a positive correlation with the T_m ($R=0.40$) values is detected. This might indicate that the predicted APRs reside within the β -sheets in the core, and thus stabilize the structure. At last, we calculated the intrinsic CamSol scores for each nanobody. Negative scores reflect on low solubility and high scores on good solubility which should result in a positive

correlation with the T_{agg} values. Again, no correlation can be observed with the experimental data. Nevertheless, the predictors AGGRESCAN, Tango, and Camsol show good correlations among each other.

Since the total aggregation scores could not explain the different aggregation behaviors in the dataset, a more detailed analysis was conducted. We used a consensus-approach with 11 APR predictors to determine the aggregation-prone residues in each sequence (Figure 3). This approach might lead to more reliable predictions, since the individual predictors were trained against different data sets. A residue is either predicted as aggregation-prone (value of 1) or not aggregation-prone (value of 0) by the different algorithms. A cumulative representation of these APR predictions per residue for NbD5 is shown in Figure 3a. If a residue reaches the threshold of 6 positive predictions, it is counted as aggregation prone. The total APR score is the sum of residues predicted as aggregation prone within the sequence. These scores were compared to the experimental data (Figure 3b, c). Nanobodies with high APR scores should show lower aggregation onset temperatures, resulting in a negative correlation. However, there is again no correlation with the experimental data. Previous work has shown that a significant fraction of sequence-based APRs is predicted in β -strands.²⁸ Thus, in the native folded state, several of the APRs are not solvent exposed but could potentially contribute to non-native aggregation upon unfolding. To better understand the position of the APRs in the native protein, we analyzed the 68 nanobodies on a structure-based level.

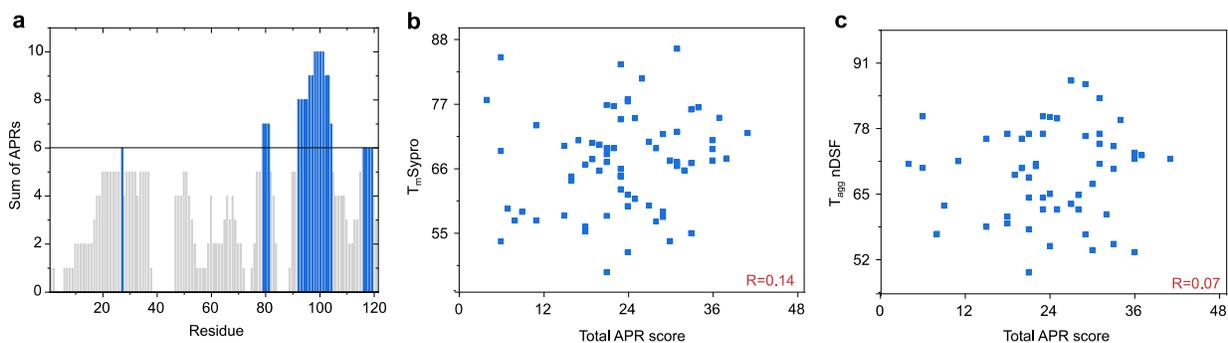


Figure 3: Consensus APR prediction on a residue level with 11 publicly available aggregation prediction tools. (a) Example of a cumulative representation of residues in NbD5 predicted as either aggregation-prone or not aggregation-prone by the different algorithms with those reaching the threshold of 6 (horizontal line) colored in blue. **(b)** Correlation of the total APR score per nanobody with the T_m and **(c)** the T_{agg} values. The Pearson's R is calculated with Origin 2019.

2.3.3. Structure-based prediction of aggregation propensity and protein stability

Since the structures of the nanobodies are not available in the PDB, we used AlphaFold2 to predict the protein structures from their sequences. The bottleneck of structure prediction of mAbs and nanobodies has thus far been the low accuracy of CDR loop modeling, especially CDR-H3 due to its increased diversity in sequence and length.²⁹ AlphaFold2 has been successful in addressing this challenge. In a comparison of four methods for nanobody structure prediction, AlphaFold showed highest accuracy, especially for CDR3 loops with an average RMSD of 2.9 Å. In contrast to the other evaluated methods, AlphaFold could even predict structured loops.³⁰ Here, five models were generated per nanobody and ranked based on two measures of confidence: the local reliability (predicted local-distance difference test, pLDDT) and the reliability of pairwise interactions between different residues in the chain.¹⁹ The highest ranking structure for each nanobody was selected and subsequently energy minimized with Rosetta.²⁰ As expected, the framework regions for all nanobodies have been modeled with very high confidence whereas there are varying confidence levels in the modeling of the CDRs (Figure 4). For 15 of the 68 nanobodies, the CDRs could be confidently modeled while for the remaining 53 nanobodies the CDRs show low confidence. Clearly, shorter CDR loops could be modeled with more accuracy, e.g., in NbD4, than the longer loops in e.g., NbPep4. However, NbPep14 contains a long CDR3 loop which shows a confidence between 70 and 90 %. In total, 59 of the 68 nanobodies show an overall pLDDT >90, the remaining 9 nanobodies are still predicted confident with the lowest pLDDT of 87.8 % (Table S1). Several of the nanobodies with long CDR3s showed the above-mentioned shielding of the former V_H-V_L interface, e.g., NbPep4. This in turn should reduce their hydrophobic surface area and thus aggregation propensity.

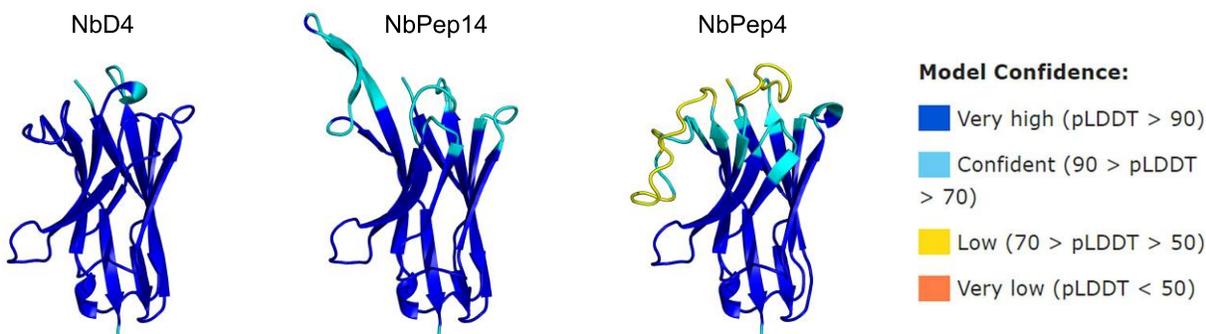


Figure 4: Examples of AlphaFold2 models of nanobodies NbD4, NbPep14, and NbPep4. AlphaFold produces a per-residue score (pLDDT) between 0 and 100. The confidence of the pLDDT is indicated by the color code. The framework regions have a very high model confidence, whereas the CDRs have been modeled confidently for 15 of the nanobodies in the dataset and with low confidence for the remaining 53 nanobodies.

To investigate whether structure homology can indicate the stability of a nanobody, we determined the RMSD between each pair of nanobodies based on the full structure and the framework region only (Figure S4, Figure S5). As for the sequence homologies, a high structure similarity does not necessarily result in the same stability behavior. A few well-performing nanobodies share a similar structure which is however also adopted by less well-behaving molecules in the dataset. Overall, there is again no visible trend. Interestingly, NbD7 which had the lowest sequence homology within the dataset, has a relatively high structure homology.

As explained above, buried APRs cannot contribute to aggregation of the native state. We highlighted the predicted APRs in the example structure of NbPep39 to better reflect on the positions within the 3D structure (Figure 5a). In this case, it becomes obvious that the APRs are all located in β -strands which typically form the core of the nanobody. Next, we calculated the solvent exposure of each residue in the structure and performed an overlay with the predicted APRs (Figure 5b). The APRs perfectly match with the least solvent exposed regions of the molecule. This proves that the aggregation propensity calculated based on the primary sequence alone, does not reflect on the aggregation of the native state. Consequently, the total APR solvent exposure per nanobody was calculated as the sum of the solvent exposure of each APR. Again, a comparison with the experimental T_{agg} values does not show any correlation (Figure 5c).

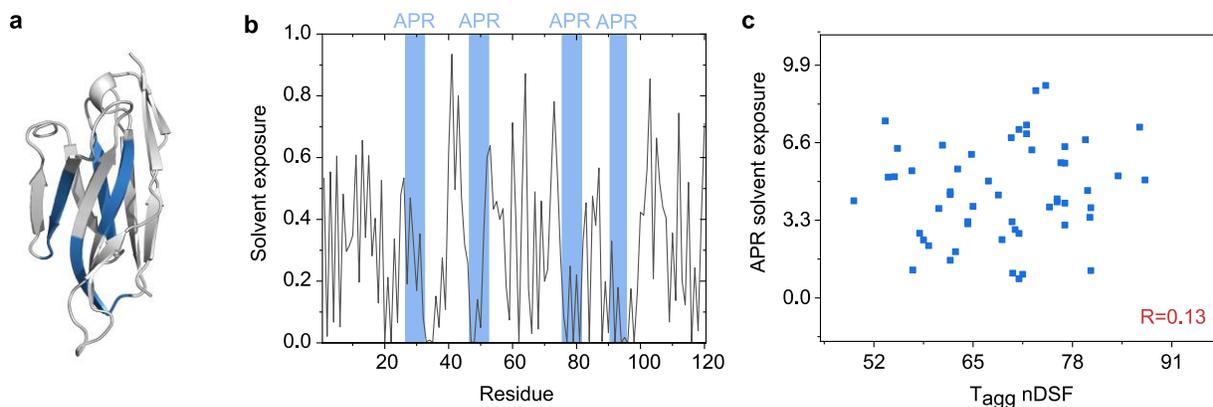


Figure 5: Solvent exposure of sequence-based APRs. (a) APRs predicted from the primary sequence of NbPep39 (colored in blue) comprise residues in the β -strands. (b) The predicted APRs of NbPep39 in the β -strands are not solvent exposed in the native state. (c) Total solvent exposure of sequence-based APRs for each nanobody does not correlate with T_{agg} .

Several computational methods have been applied to predict the aggregation propensities and stabilities of proteins on a structure-based level (Figure 6). We used AGGRESCAN3D 2.0 and the structurally corrected CamSol webserver to predict the aggregation propensities. These webserver have previously successfully been used in the selection of aggregation-resistant antibodies and other proteins.³¹ As already seen for the sequence-based analysis with AGGRESCAN and CamSol, the two techniques are not able to explain the different thermal stabilities of the nanobodies. Again, CamSol does not show any correlation while AGGRESCAN 3D shows a moderate positive correlation with the T_{ms} and a weak positive correlation with the T_{agg} values, which was expected to be negative. Furthermore, the ratio of hydrophobic to total SASA was calculated to give an estimate of the hydrophobicity of the surface. Surprisingly, we again observe a weak positive correlation with the experimental data. In theory, more hydrophobic surfaces would lead to more association of monomers, thus aggregation. Usually, hydrophobic residues are tightly packed into the protein core and form clusters, which drive protein folding and stability. We therefore evaluated the number of hydrophobic clusters within each structure but could not find an explanation for increased stability. Since the experimental data showed that aggregation only occurs after unfolding of the molecules, we were interested if the stability of the protein can be a predictor for aggregation-resistance. For an estimate of the protein stability, we scored the structures with a Rosetta energy function. It must be noted that a lower scoring structure is more likely to be stable, but the scores do not have a direct conversion to physical energy units.¹⁴ However, for a relative

comparison of the molecules in this dataset, the determined energy values can be applied. Here, we find a weak negative correlation with the T_m s but no correlation with the T_{agg} values. As expected, the Rosetta score does not correlate with any of the aggregation prediction methods. In the end, none of the approaches thus far led to a satisfying explanation or prediction of the thermal stability or aggregation behavior of the nanobodies in the dataset.

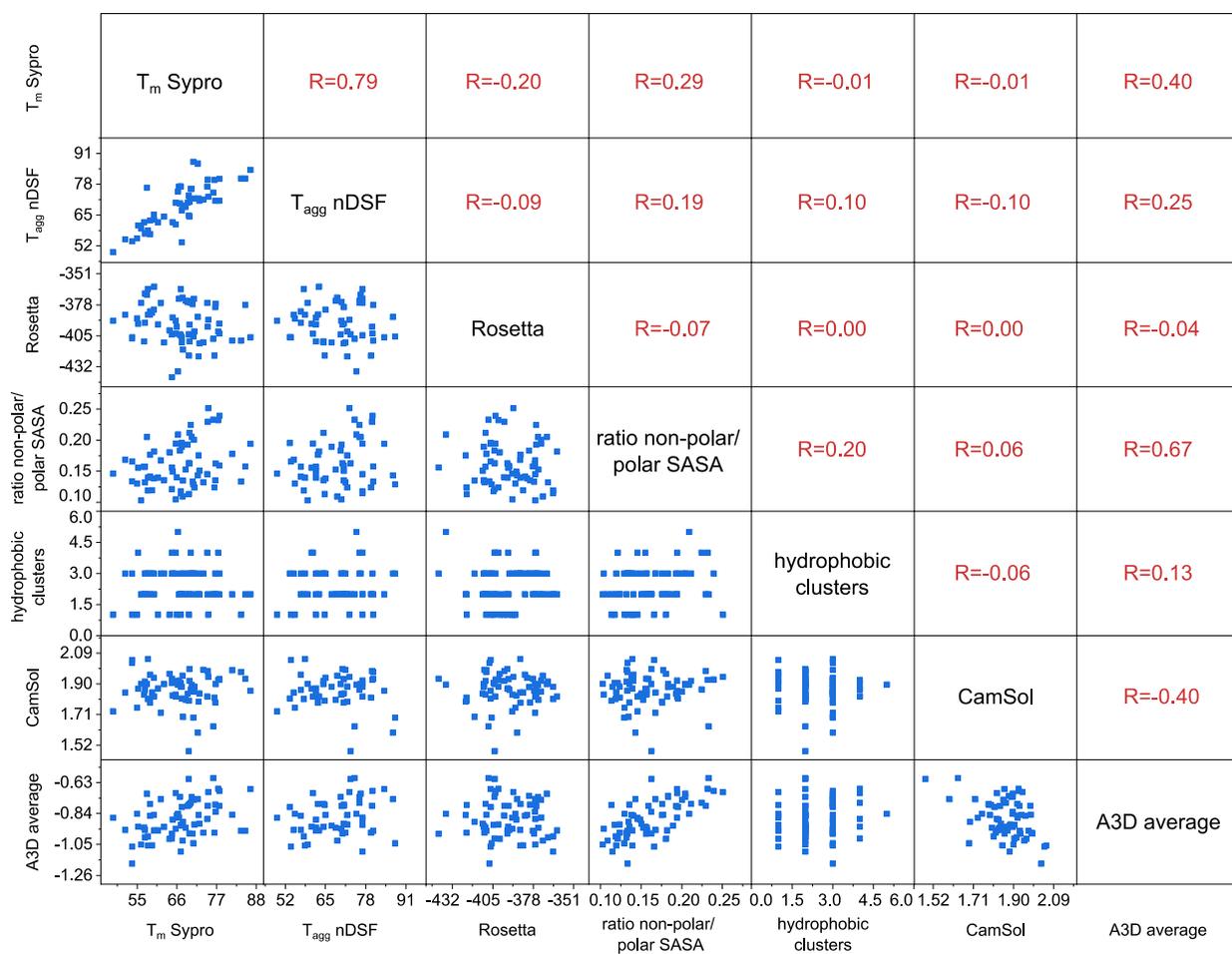


Figure 6: Correlation of the structure-derived features including the Rosetta energy score, the ratio of non-polar/polar solvent accessible surface area (SASA), the number of hydrophobic clusters, and the aggregation propensities determined with the structurally corrected CamSol and the AGGRESCAN3D 2.0 webserver. The Pearson correlation coefficients were calculated with Origin 2019.

2.3.4. Partially unfolded species in the aggregation process of nanobodies

Since the nanobodies do not show native but non-native aggregation in the experiments, the impact of partially unfolded species must be evaluated. Therefore, we performed MD simulations at elevated temperatures (450 K) to obtain partially unfolded species *in silico* and analyzed their aggregation potential with the previously applied approaches (Figure 7).

A slight decrease in β -sheet structure but no significant unfolding events can be observed after the 100 ns simulation at 450 K. The final frame of each simulation was submitted to the CamSol and AGGRESCAN 3D 2.0 webservers, and the solvent exposure of the consensus APRs was calculated for the partially unfolded species. A comparison with the experimental data again showed no correlation with the CamSol scores and a moderate positive correlation with the AGGRESCAN 3D 2.0 scores, which was expected to be negative. As could be seen before, the *in silico* tools show a better correlation among each other. Interestingly, the solvent exposure of the consensus APRs even decreased for some of the nanobodies, indicating that the structures were potentially not fully energy minimized. Nevertheless, this approach could not explain the different stability behaviors of the nanobodies. This could be because the simulations were not conducted long enough to achieve significant unfolding of the structures. Since we could not find any correlation between the experimental data and the *in silico* tools while these show good correlation among each other, a hypothesis is that the experimental conditions may influence the T_m and T_{agg} results, making a direct comparison impossible. First, the investigated nanobodies in the dataset were produced with different tags (HA-tag and Myc-tag) which add additional 9 or 10 amino acid residues to each sequence. Depending on the length of the nanobody, these additional residues comprise between 6.7 and 8.8 % of the full sequence which can have a significant impact on the stability and aggregation propensity. Especially since the tag sequences contain charged and polar residues which can be involved in intra- and intermolecular interactions. These tags have not been accounted for in the computational assessment. Furthermore, the experimental work has been performed in PBS buffer, which has a high ionic strength and can screen potential electrostatic effects which are included in the *in silico* scores. Additionally, phosphate anions can directly interact with residues on the surface of proteins and thus alter their self-association behavior.

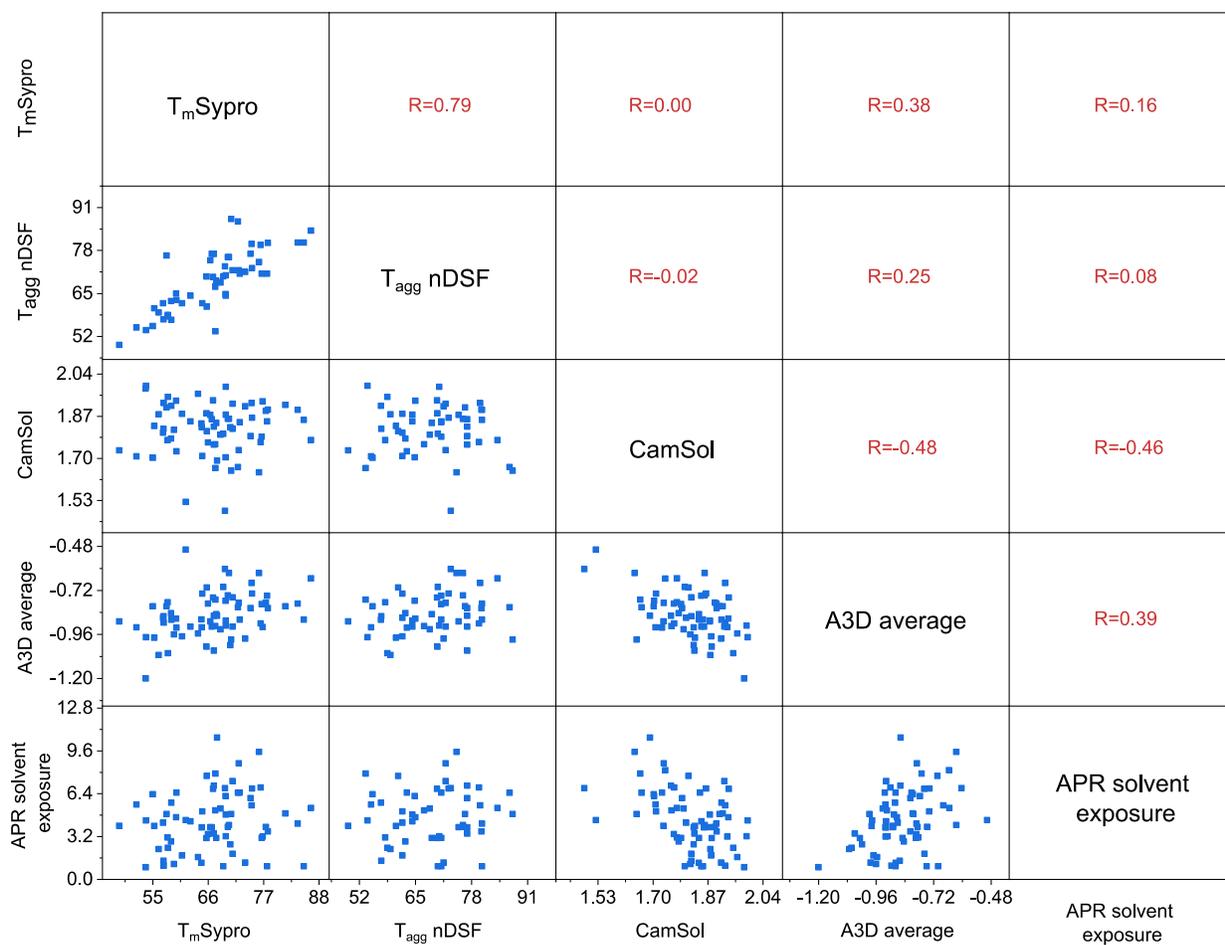


Figure 7: Aggregation propensities of partially unfolded species generated by MD simulations at 450 K do not correlate with the experimental data. The Pearson correlation coefficients were calculated with Origin 2019.

2.4. Conclusions

In this work we carried out a quantitative comparison between various computational methods of assessing nanobody thermal stability and aggregation propensity, including selected *in silico* solubility predictors and MD simulations. Our dataset consisted of 68 nanobodies that spanned a broad range of T_{agg} and T_m values to allow the differentiation between well and poor behaving molecules *in silico*. We can conclude that none of the algorithms has been proven to have clearly superior performance over another. While the computational tools show relatively good correlation among each other, the correlation with the T_m and T_{agg} values was either weak or not existing. Since the webservers do not account for non-native protein aggregation, we performed MD simulations at elevated temperature to induce partial unfolding and monitored the aggregation propensity of these altered conformations. Again, we were not able to accurately rank the molecules according to the measured T_m and T_{agg} values.

It must be noted that the performance of the structure-based tools can greatly be impeded by the quality of the input structure. Furthermore, the computational screens did not account for the different tag-sequences of the nanobodies and the potential effects of the formulation buffer in the experimental results. Some of the applied webservers indeed account for the ionic strength in the system but not for specific buffer interactions, which could have stabilizing or destabilizing effects on the protein.^{32–35} Phosphate could have indeed been added to the MD simulations, however, this would increase the computational burden and would not be accounted for in further analyses steps. Next to that, it has been shown that thermal denaturation techniques in some cases result in misleading stability rankings, showing poor correlation of the T_m and T_{agg} values with the long-term stability of proteins. Instead, the investigation of reversibility of unfolding and colloidal stability, e.g. via the interaction parameter k_D or the second virial coefficient A_2 , resulted in better correlations with storage stability data.^{28,36–39} It could therefore be that the experimentally determined T_m and T_{agg} values in this study do not accurately reflect the intrinsic stability of the nanobody structures relevant for developability assessment.

2.5. Supporting Information

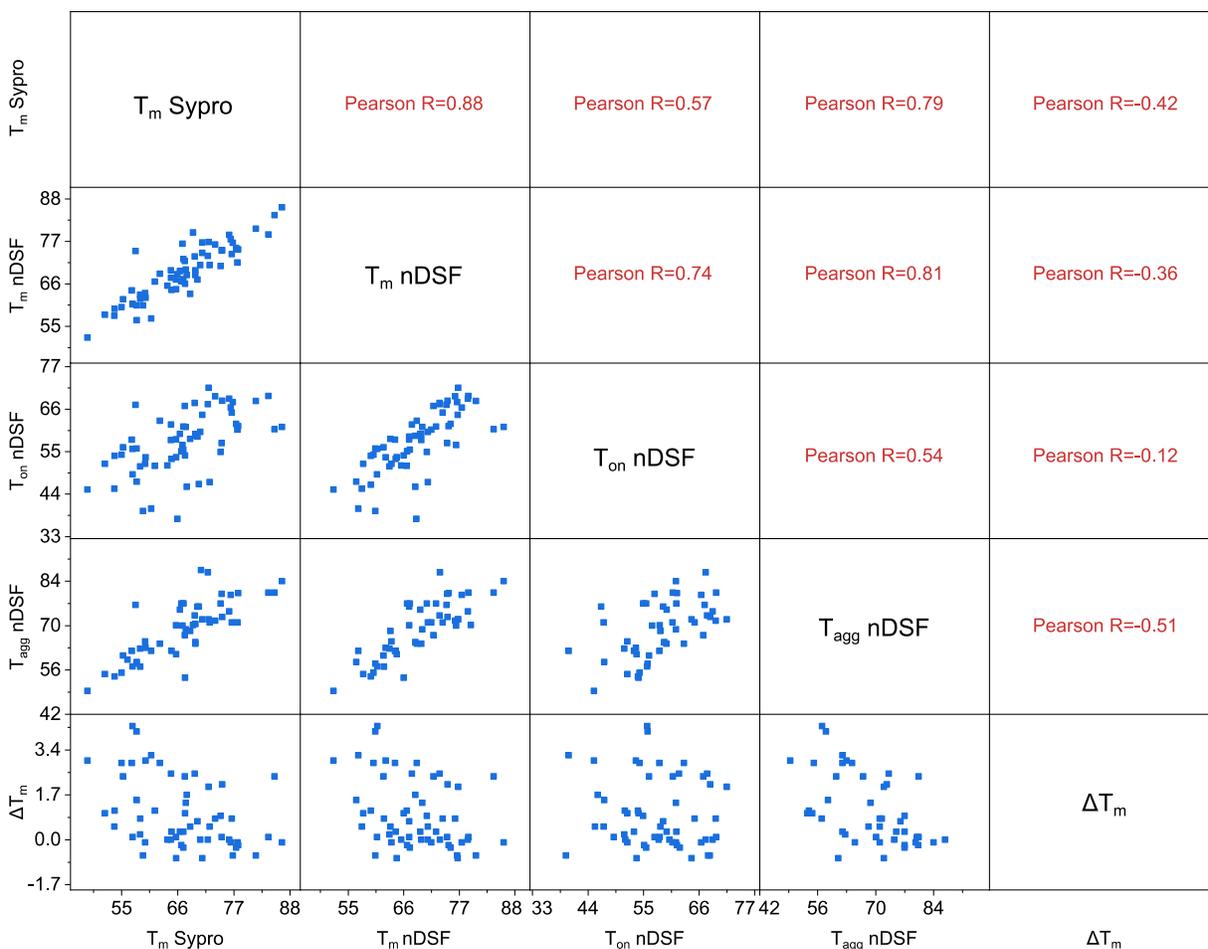


Figure S1: Correlation matrix of melting temperatures (T_m) measured via thermal shift assay with SYPRO® Orange and via nanoDSF, onset temperature of unfolding (T_{on}) and onset of aggregation (T_{agg}) as well as ΔT_m , specified as the difference in T_m between low (13.1 μ M) and high (32.72 μ M) protein concentration.

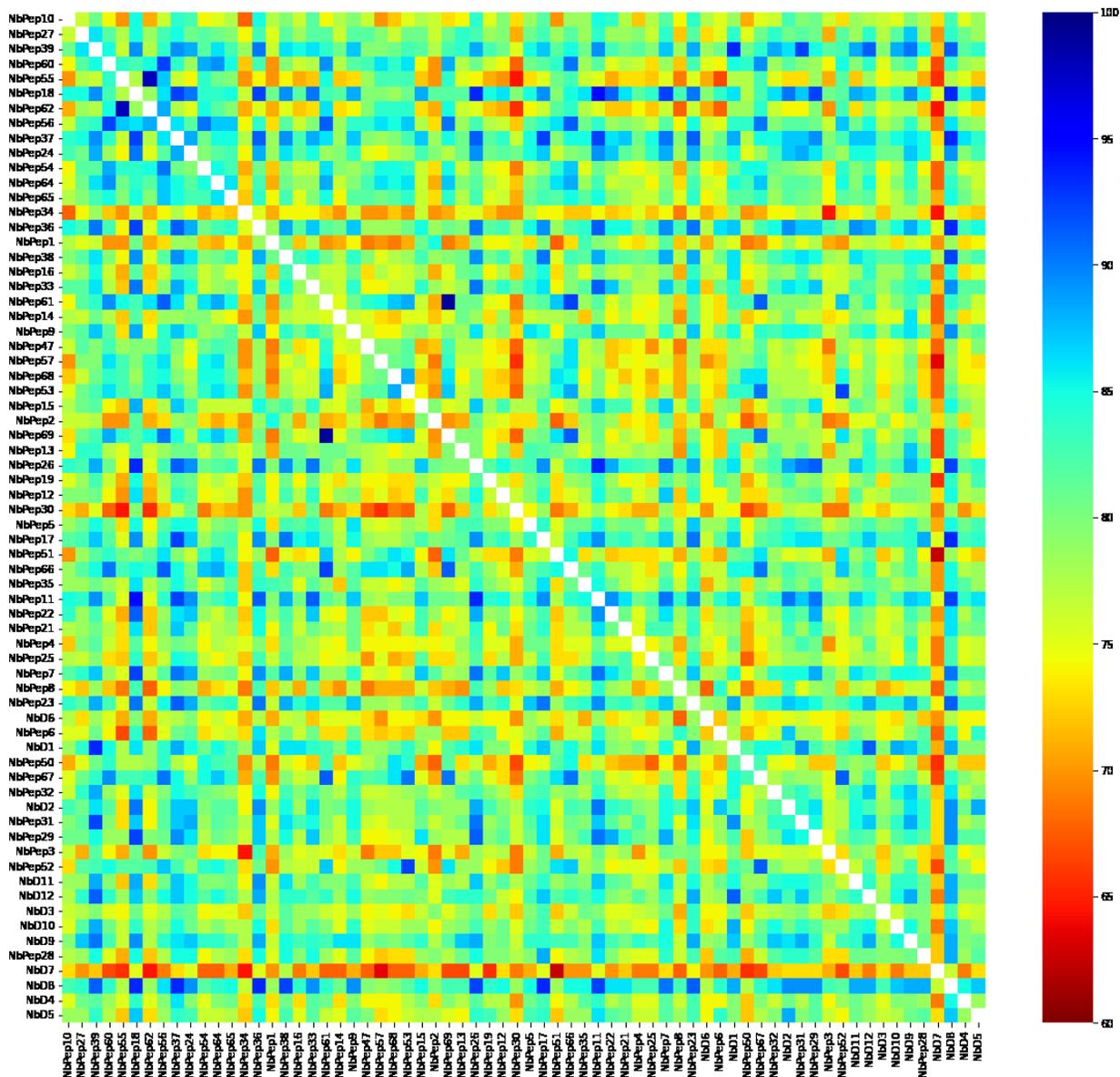


Figure S2: Heatmap of the sequence homologies between all 68 nanobodies in the dataset ranging from $\sim 50\%$ in red to $\sim 90\%$ in blue. The nanobodies are ordered based on their T_m s from highest to lowest (top to bottom/ left to right).

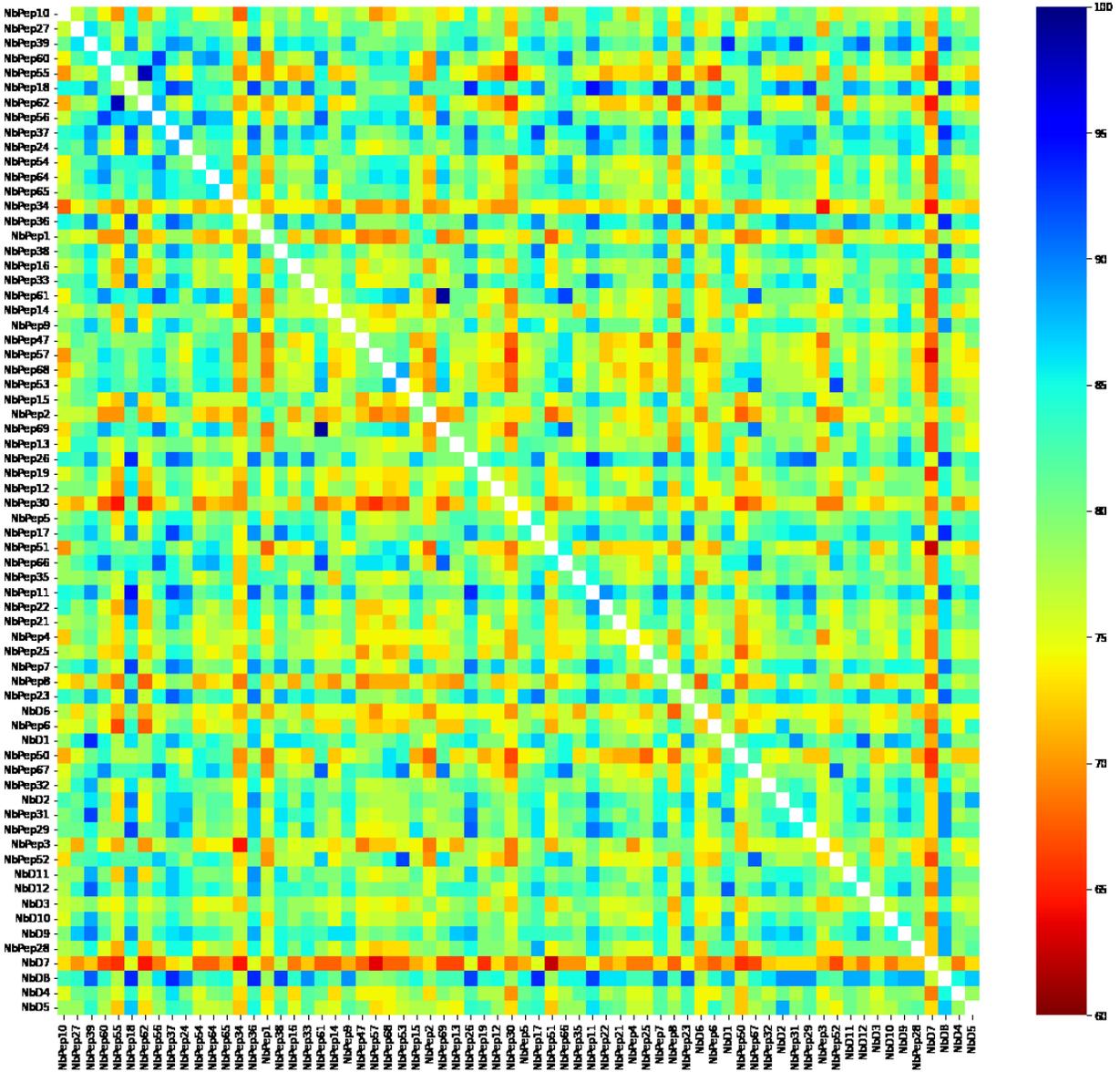


Figure S3: Heatmap of the sequence homologies of the framework regions only between all 68 nanobodies in the dataset ranging from ~ 60% in red to ~ 99% in blue. The nanobodies are ordered based on their T_m s from highest to lowest (top to bottom/ left to right).

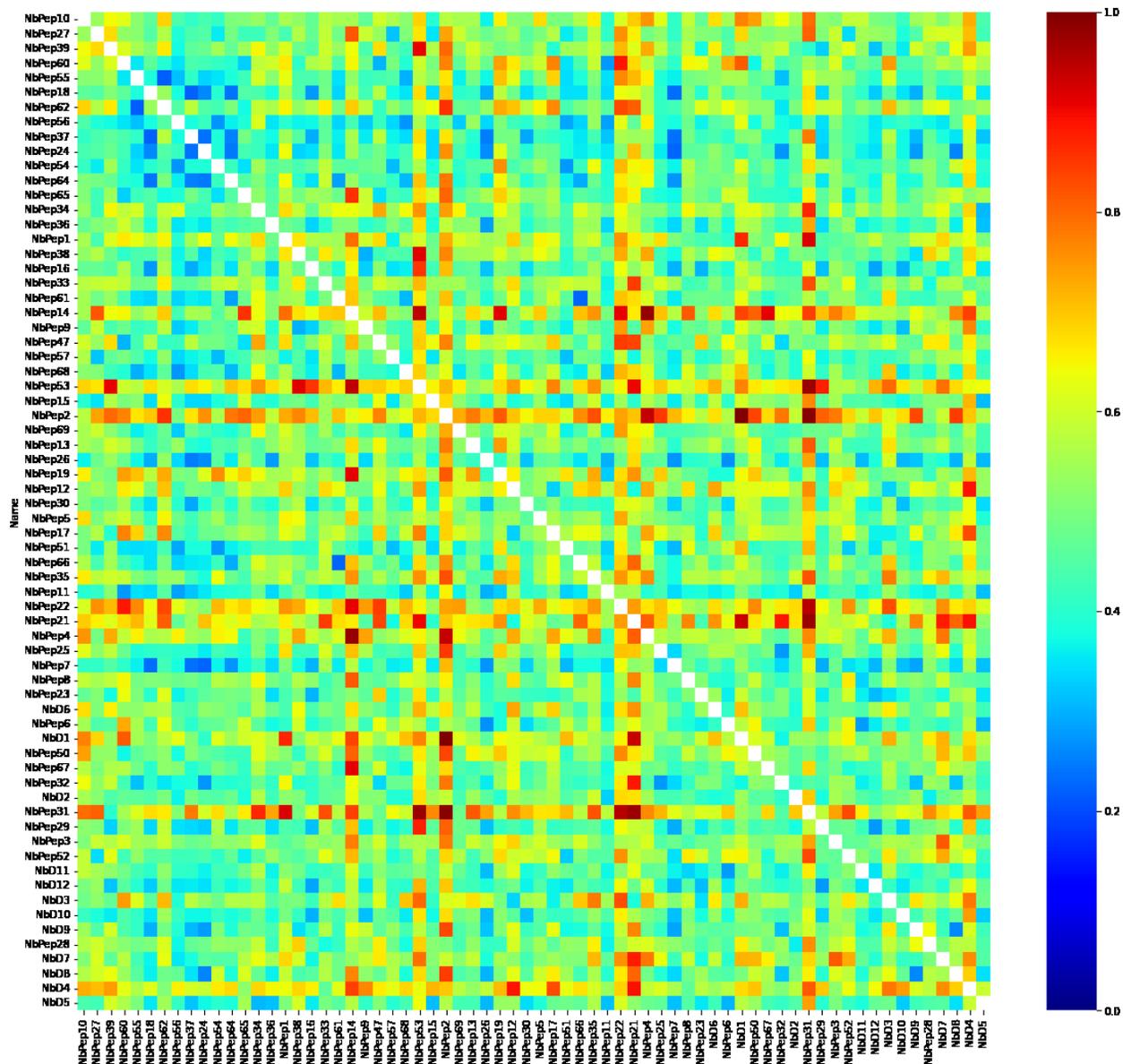


Figure S4: Heatmap of the root-mean square deviations (RMSD) between all 68 nanobodies in the dataset ranging from ~ 0.2 Å in blue to ~ 1.5 Å in red. The nanobodies are ordered based on their $T_{m,s}$ from highest to lowest (top to bottom/ left to right).

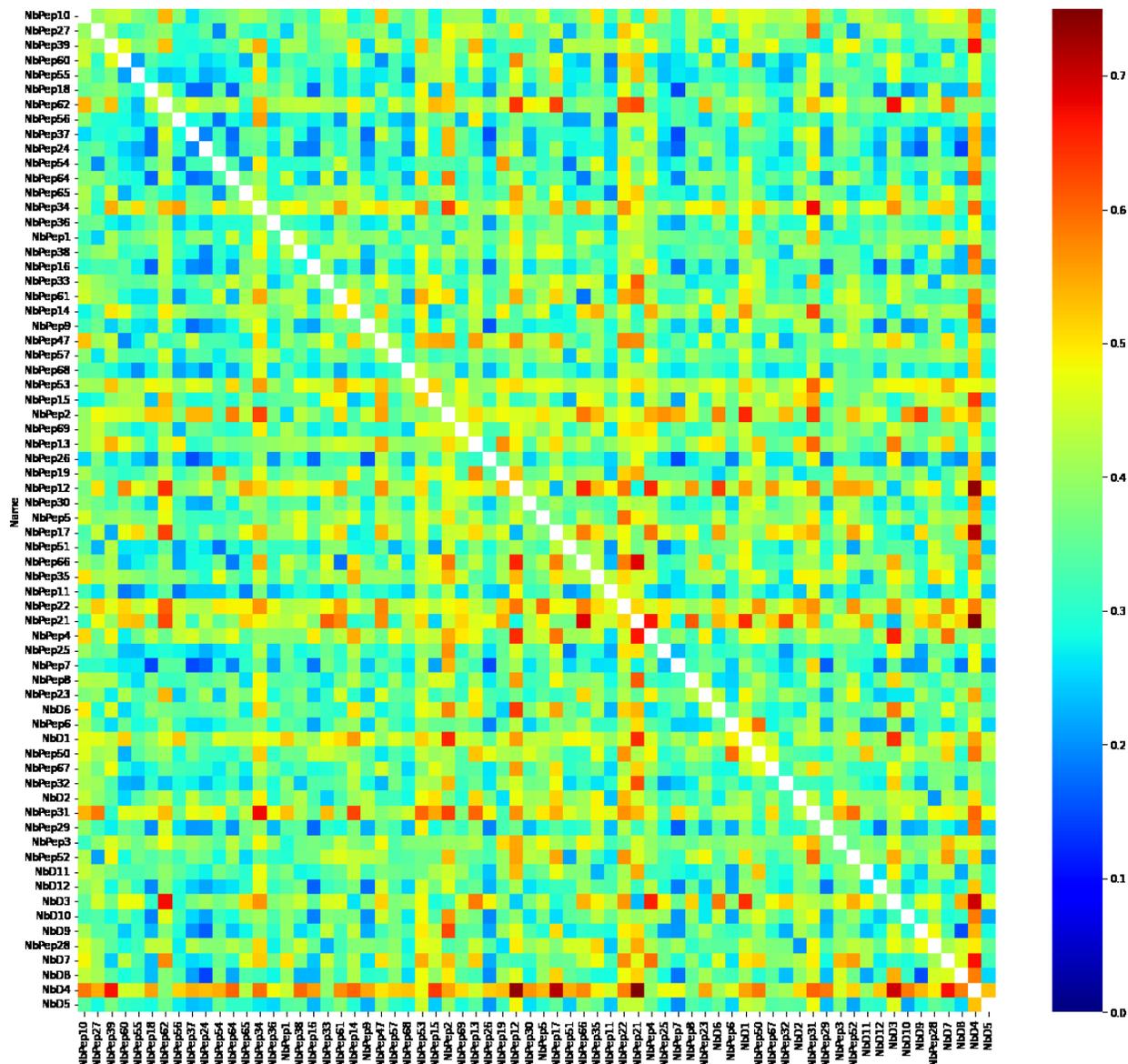


Figure S5: Heatmap of framework RMSDs between all 68 nanobodies in the dataset ranging from ~ 0.1 Å in blue to ~ 0.75 Å in red. The nanobodies are ordered based on their T_m s from highest to lowest (top to bottom/ left to right).

Table S1: pLDDT scores of the nanobody AlphaFold2 models.

| Name | pLDDT score | Name | pLDDT score |
|-------------|--------------------|-------------|--------------------|
| NbD1 | 92.66 | NbPep24 | 91.61 |
| NbD2 | 92.7 | NbPep25 | 94.51 |
| NbD3 | 91.52 | NbPep26 | 93.22 |
| NbD4 | 95.84 | NbPep27 | 91.66 |
| NbD5 | 93.99 | NbPep28 | 94.94 |
| NbD6 | 89.94 | NbPep29 | 90.35 |
| NbD7 | 90.63 | NbPep30 | 94.81 |
| NbD8 | 92.79 | NbPep31 | 93.38 |
| NbD9 | 93.31 | NbPep32 | 94.12 |
| NbD10 | 93.4 | NbPep33 | 92.13 |
| NbD11 | 89.79 | NbPep34 | 93.92 |
| NbD12 | 95.52 | NbPep35 | 94.75 |
| NbPep1 | 89.01 | NbPep36 | 93.01 |
| NbPep2 | 87.79 | NbPep37 | 90.74 |
| NbPep3 | 93.38 | NbPep38 | 90.28 |
| NbPep4 | 87.8 | NbPep39 | 95.25 |
| NbPep5 | 93.52 | NbPep47 | 94.29 |
| NbPep6 | 90.82 | NbPep50 | 92.05 |
| NbPep7 | 90.92 | NbPep51 | 92.86 |
| NbPep8 | 92.39 | NbPep52 | 94.43 |
| NbPep9 | 91.4 | NbPep53 | 90.46 |
| NbPep10 | 89.92 | NbPep54 | 92.67 |
| NbPep11 | 92.89 | NbPep55 | 95.5 |
| NbPep12 | 91.34 | NbPep56 | 93.08 |
| NbPep13 | 95.8 | NbPep57 | 94.32 |
| NbPep14 | 93.33 | NbPep60 | 95.41 |
| NbPep15 | 92.63 | NbPep61 | 95.53 |
| NbPep16 | 90.98 | NbPep62 | 95.55 |
| NbPep17 | 91.04 | NbPep64 | 96.04 |
| NbPep18 | 89.32 | NbPep65 | 93.66 |
| NbPep19 | 88.67 | NbPep66 | 94.45 |
| NbPep21 | 91.39 | NbPep67 | 94.17 |
| NbPep22 | 88.73 | NbPep68 | 96.04 |
| NbPep23 | 93.44 | NbPep69 | 95.41 |

2.6. References

1. Arbabi-Ghahroudi, M. Camelid single-domain antibodies: Historical perspective and future outlook. *Front Immunol* **8**, 1–8 (2017).
2. Rabia, L. A., Desai, A. A., Jhajj, H. S. & Tessier, P. M. Understanding and overcoming trade-offs between antibody affinity, specificity, stability and solubility. *Biochem Eng J.* **137**, 365–374 (2018).
3. Gonzalez-Sapienza, G., Rossotti, M. A. & Tabares-da Rosa, S. Single-domain antibodies as versatile affinity reagents for analytical and diagnostic applications. *Front Immunol* **8**, 1–12 (2017).
4. Dumoulin, M. et al. Single-domain antibody fragments with high conformational stability. *Protein Science* **11**, 500–515 (2009).
5. Ewert, S., Cambillau, C., Conrath, K. & Plückthun, A. Biophysical properties of camelid VHH domains compared to those of human VH3 domains. *Biochemistry* **41**, 3628–3636 (2002).
6. Roberts, C. J. Non-Native Protein Aggregation Kinetics. *Biotechnol Bioeng* **98**, 927–938 (2007).
7. Kunz, P. et al. The structural basis of nanobody unfolding reversibility and thermoresistance. *Sci Rep* **8**, 1-10 (2018).
8. Famm, K., Hansen, L., Christ, D. & Winter, G. Thermodynamically Stable Aggregation-Resistant Antibody Domains through Directed Evolution. *J Mol Biol* **376**, 926–931 (2008).
9. de Genst, E. et al. Molecular basis for the preferential cleft recognition by dromedary heavy-chain antibodies. *Proc Natl Acad Sci U S A* **103**, 4586–4591 (2006).
10. Kunz, P., Ortale, A., Mücke, N., Zinner, K. & Hoheisel, J. D. Nanobody stability engineering by employing the ΔT_m shift; A comparison with apparent rate constants of heat-induced aggregation. *Protein Engineering, Design and Selection* **32**, 241–249 (2019).
11. Sormanni, P., Aprile, F. A. & Vendruscolo, M. The CamSol Method of Rational Design of Protein Mutants with Enhanced Solubility. *J Mol Biol* **427** (2), 478-490 (2014).
12. Hebditch, M., Carballo-Amador, M. A., Charonis, S., Curtis, R. & Warwicker, J. Protein-Sol: A web tool for predicting protein solubility from sequence. *Bioinformatics* **33**, 3098–3100 (2017).
13. Kuriata, A. et al. Aggrescan3D (A3D) 2.0: Prediction and engineering of protein solubility. *Nucleic Acids Res* **47**, W300–W307 (2019).
14. Alford, R. F. et al. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J Chem Theory Comput* **13**, 3031–3048 (2017).

15. Sievers, F. & Higgins, D. G. Clustal Omega, accurate alignment of very large numbers of sequences. *Methods Mol Biol* **1079**, 105–116 (2014).
16. Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S. E., Wilkins, M. R., Appel, R. D., & Bairoch, A. Protein identification and analysis tools on the ExPASy server. The Proteomics Protocols Handbook, *Humana press*, 571-607 (2005).
17. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol* **157**, 105–132 (1982).
18. Tsolis, A. C., Papandreou, N. C., Iconomidou, V. A. & Hamodrakas, S. J. A Consensus Method for the Prediction of ‘Aggregation-Prone’ Peptides in Globular Proteins. *PLoS One* **8**, 1–6 (2013).
19. Mirdita, M. et al. ColabFold: making protein folding accessible to all. *Nat Methods* **19**, 679–682 (2022).
20. Leaver-Fay, A. et al. ROSETTA3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules. *Methods in enzymology* **487**, 545-574 (2011).
21. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *J Mol Graph* **14**, 33–38 (1996).
22. Ferruz, N., Schmidt, S. & Höcker, B. ProteinTools: A toolkit to analyze protein structures. *Nucleic Acids Res* **49**, W559–W566 (2021).
23. H. D.A. Case, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, D. Ghoreishi, G. Giambasu, T. Giese, M.K. Gilson, H. Gohlke, A.W. Goetz, D. Greene, R Harris, N. Homeyer, Y. Huang, S. Izadi, A. Kovalenko, C. L. Nguyen, A. Onufriev, F. Pan, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, Y. Simmerling, J. Smith, J. Swails, R.C. Walker, J. Wang, H. Wei, L. Wilson, R.M. Wolf, X. Wu, L. Xiao, D. M. Y. Xiong, and P. A. Kollman, AMBER 2019. University of California, San Francisco, (2019).
24. Gordon, J. C. et al. H⁺⁺: A server for estimating pK_as and adding missing hydrogens to macromolecules. *Nucleic Acids Res* **33**, 368–371 (2005).
25. Miklos, A. E. et al. Structure-based design of supercharged, highly thermoresistant antibodies. *Chem Biol* **19**, 449–455 (2012).
26. Der, B. S. et al. Alternative Computational Protocols for Supercharging Protein Surfaces for Reversible Unfolding and Retention of Stability. *PLoS One* **8**, e64363 (2013).
27. Prabakaran, R., Rawat, P., Thangakani, A. M., Kumar, S. & Gromiha, M. M. Protein aggregation: in silico algorithms and applications. *Biophys Rev* **13**, 71–89 (2021).

28. Berner, C., Menzen, T., Winter, G. & Svilenov, H. L. Combining Unfolding Reversibility Studies and Molecular Dynamics Simulations to Select Aggregation-Resistant Antibodies. *Mol Pharm* **18**, 2242–2253 (2021).
29. Ruffolo, J. A., Guerra, C., Mahajan, S. P., Sulam, J. & Gray, J. J. Geometric potentials from deep learning improve prediction of CDR H3 loop structures. *Bioinformatics* **36**, I268–I275 (2020).
30. Ruffolo, J. A. & Gray, J. J. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Biophys J* **121**, 155a–156a (2022).
31. Wolf Pérez, A. M. et al. In vitro and in silico assessment of the developability of a designed monoclonal antibody library. *MAbs* **11**, 388–400 (2019).
32. Pavani, P., Kumar, K., Rani, A., Venkatesu, P. & Lee, M. J. The influence of sodium phosphate buffer on the stability of various proteins: Insights into protein-buffer interactions. *J Mol Liq* **331**, 115753 (2021).
33. Brudar, S. & Hribar-Lee, B. Effect of buffer on protein stability in aqueous solutions: A simple protein aggregation model. *Journal of Physical Chemistry B* **125**, 2504–2512 (2021).
34. Salis, A., Cappai, L., Carucci, C., Parsons, D. F. & Monduzzi, M. Specific Buffer Effects on the Intermolecular Interactions among Protein Molecules at Physiological pH. *Journal of Physical Chemistry Letters* **11**, 6805–6811 (2020).
35. Bye, J. W. & Curtis, R. A. Controlling Phase Separation of Lysozyme with Polyvalent Anions. *Journal of Physical Chemistry B* **123**, 593–605 (2019).
36. Svilenov, H. L. & Winter, G. Formulations That Suppress Aggregation During Long-Term Storage of a Bispecific Antibody are Characterized by High Refoldability and Colloidal Stability. *J Pharm Sci* **109**, 2048–2058 (2020).
37. 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. *European Journal of Pharmaceutics and Biopharmaceutics* **125**, 106–113 (2018).
38. Svilenov, H. L., Arosio, P., Menzen, T., Tessier, P. & Sormanni, P. Approaches to expand the conventional toolbox for discovery and selection of antibodies with drug-like physicochemical properties. *MAbs* **15**, 2164459 (2023).
39. Gentiluomo, L. et al. Advancing Therapeutic Protein Discovery and Development through Comprehensive Computational and Biophysical Characterization. *Mol Pharm* **17**, 426–440 (2020).

Chapter 3 - Combining unfolding reversibility studies and molecular dynamics simulations to select aggregation-resistant antibodies

This chapter is published as:

Berner, C.¹, Menzen, T.², Winter, G.¹, & Svilenov, H. L.^{1,3} (2021). Combining unfolding reversibility studies and molecular dynamics simulations to select aggregation-resistant antibodies. *Molecular Pharmaceutics*, 18(6), 2242-2253.

¹Department of Pharmacy, Ludwig-Maximilians-Universität München, Munich, Germany

²Coriolis Pharma Research GmbH, Martinsried, Germany

³Department of Chemistry, Technische Universität München, Garching, Germany

Author contributions:

H.S., C.B., T.M. and G.W. conceived the idea. H.S. performed most of the experimental work and evaluated the data. C.B. performed the *in silico* work, the CD measurements, and the analysis of the 12 months stability time point. H.S. and C.B. wrote the manuscript. G.W. and T.M. provided conceptual guidance. G.W. and T.M. corrected the manuscript.

The published article can be accessed online via:

<https://doi.org/10.1021/acs.molpharmaceut.1c00017>

3.1 Introduction

Antibodies have become the most successful class of therapeutic proteins.¹ Animal immunization and in vitro techniques like phage display can identify dozens of monoclonal antibodies that bind to a pharmacological target. However, not all antibodies with desired antigen-binding properties will possess features that make them suitable to become drugs. The drug-like properties of an antibody are related to various characteristics like low aggregation propensity, high thermal stability, high specificity, and low viscosity at high protein concentrations.²⁻⁷ Identifying the candidates with drug-like properties at a very early stage is a part of a developability assessment strategy that aims to mitigate the risk that an antibody fails to pass checkpoints en route to becoming a marketed product. The informed early selection of the most promising candidates can save resources and help avoid project delays.

A sound developability assessment is based on extensive characterization with at least several biophysical methods.^{4,7-11} The issue with this approach is that the number of antibody candidates is usually large while the sample amounts are very scarce during the discovery and early development phases. Thus, only analytical methods that require small protein amounts and can be applied to dozens of samples per day are appropriate at this stage. It is important that these techniques provide orthogonal information about the various biophysical characteristics of therapeutic antibody candidates.

The developability assessment aims to select antibodies that do not form aggregates during processing and storage. Particularly the aggregation of partially unfolded proteins (known as non-native aggregation)¹² is a severe issue that can endanger the successful development of an antibody. The propensity for non-native aggregation of different antibodies is difficult to predict because the partially unfolded species are usually present in exceptionally low concentrations at conditions relevant for the storage of therapeutic proteins.

Some developability assessment programs aim to identify aggregation-resistant antibodies by selecting molecules with high protein melting temperatures and high Gibbs free energy of unfolding. Such strategies are based on the rationale that higher thermal and conformational protein stabilities indicate a smaller population of partially unfolded species in solution at storage temperature. However, not all unfolded proteins are equally prone to non-native aggregation. For example, aggregation-resistant antibody domains with low conformational and thermal stabilities

were reported.^{13,14} Interestingly, these antibody domains share one special feature – they exhibit reversible thermal unfolding. This earlier work indicates that the ability of a protein to refold without aggregating can be at least as important as the conformational and thermal stabilities for the selection of aggregation-resistant molecules.

Contrary to antibody domains, it is unlikely that a full-length monoclonal antibody refolds reversibly without forming aggregates after exposure to high temperatures (e.g. to 90 °C). However, this does not mean that all antibodies show the same ability to refold after exposure to different temperatures. In contrast, antibodies likely exhibit specific unfolding reversibility as a unique biophysical characteristic. In this context, we recently reported an approach to study the reversibility of thermal unfolding by modulated scanning fluorimetry (MSF).¹⁵ MSF employs incremental heating and cooling cycles to identify the temperatures that start causing irreversible protein unfolding. For example, by using MSF we showed that the non-reversibility onset temperature of trastuzumab is not related to the structural perturbations in the C_{H2} domain that occur at lower temperatures compared to the antigen-binding fragment (Fab) and the C_{H3} domain.¹⁵ A piece of information that is not directly evident from other methods that assess thermal protein unfolding.¹⁵

Different from heat-induced denaturation, isothermal chemical denaturation with urea or guanidine hydrochloride (GuHCl) more often results in reversible protein unfolding, even when the protein is completely unfolded.¹⁶ As the denaturant concentration is reduced, the protein refolding will compete with the aggregation of the partially unfolded species.¹⁷ This offers an excellent opportunity to study the aggregation propensity of the partially unfolded antibodies isothermally at storage temperatures. For example, dilution refolding experiments can provide insights into the aggregation propensity of antibodies in different solution conditions.^{18,19} Furthermore, dialysis refolding experiments with the ReFOLD assay demonstrated a link between the ability of an antibody to remain monomeric after refolding from urea and the aggregation during long-term storage in different formulations.^{20–22}

Here we investigate whether unfolding reversibility studies can select the aggregation-resistant antibodies from a set with candidates for further development. We observed vastly different unfolding reversibility of the thirteen antibodies in MSF and ReFOLD experiments. Aggregation of the unfolded proteins was the main reason for poor unfolding reversibility. We therefore

performed MD simulations to obtain partially unfolded variable domains from the antibodies. The aggregation potential of the unfolded domains was evaluated with CamSol *in silico*. We found that antibodies with higher non-reversibility onset temperature (T_{nr}) and higher relative monomer yield (RMY) after refolding from denaturants are well-behaved molecules with low aggregation propensity during storage. Furthermore, the *in silico* characterization showed that the variable domains of the aggregation-prone antibodies unfold and expose buried aggregation-prone regions (APRs) at lower temperatures compared to aggregation-resistant molecules.

3.2 Materials and methods

3.2.1 Proteins and chemicals

Eight of the antibodies (PPI01, PPI02, PPI03, PPI04, PPI10, PPI11, PPI13, PPI17) and one bispecific antibody fusion protein (PPI08) were provided by the PIPPI consortium. Detailed information about the structure and purity of all PPI proteins except PPI11 is published elsewhere.⁸ The remaining four mAbs (adalimumab, bevacizumab, rituximab and trastuzumab) were obtained from marketed products. All antibodies except PPI11 (IgG4) and PPI17 (IgG2) belong to the IgG1 class. All antibodies except PPI01, PPI04 and PPI11 (lambda) have a kappa light chain. Cation exchange chromatography was used to separate the antibodies from the excipients in the marketed formulations. Finally, the buffer of the proteins was exchanged by extensive dialysis to 50 mM histidine/histidine hydrochloride with pH 6.0 at 25 °C. The protein concentration was 1 mg/mL unless otherwise stated. All chemicals were of molecular biology or multi-compendial grade and were purchased from Sigma or Thermo Fisher Scientific (Germany).

3.2.2 Isothermal chemical denaturation (ICD)

The ICD experiments were performed as earlier described.²³ Briefly, a protein stock solution, the histidine buffer and denaturant stock solution (10 M urea or 6 M GuHCl) in the same buffer were combined in a 384-multiwell plate with non-binding surface (Corning, USA). A Vialflo Assist (Integra Biosciences, Germany) was used for dispensing and mixing of the solutions. The final protein concentration was 0.1 mg/mL. The denaturant concentration varied in 24 steps from 0 to 9 M or 0 to 5.4 M for the experiments with urea and GuHCl, respectively. After the samples were mixed, the well plate was sealed with adhesive foil and incubated for 24 hours at room temperature. Afterwards, the intrinsic protein fluorescence at 330 nm and 350 nm after excitation at 280 nm was measured in each well with a FLUOstar Omega microplate reader (BMG Labtech, Germany). The

ratio between the fluorescence intensities (FI_{350}/FI_{330}) was plotted against the protein concentration to obtain the isothermal chemical denaturation curves. CDpal was used to fit the curves using a three-state model to obtain the first (C_{m1}) and second (C_{m2}) apparent melting denaturant concentration.²⁴

3.2.3 Size exclusion chromatography with multi-angle light scattering (SEC-MALS)

A Dionex UltiMate 3000 UHPLC system with a UV-Vis absorbance detector (Thermo Fisher Scientific) and a DAWN HELEOS multi-angle static light scattering detector (Wyatt Technology) were used. The column was a Superdex 200 Increase 10/300 GL. The running buffer contained 50 mM sodium phosphate pH 7.2 with 400 mM sodium chloride and 0.05% (w/w) sodium azide. The system flow was 1 mL/min. The volume of the injected sample was 50 μ L. The elution of the samples was monitored by the absorption at 280 nm. The Astra v7.1 software (Wyatt Technology) was used to calculate the molecular mass. The integration of the chromatograms was done with Chromeleon V7 (Thermo Fisher Scientific). This SEC-MALS method was used to analyze the samples from the ReFOLD assay and from the storage stability study.

3.2.4 ReFOLD assay

A previously reported ReFOLD assay was used for the isothermal unfolding/refolding of the antibodies. More experimental details can be found elsewhere.²⁰ Briefly, the antibody samples were extensively dialyzed for 24 hours against a denaturant (10 M urea or 8 M GuHCl) dissolved in 50 mM histidine buffer pH 6.0. Next, the unfolded antibodies in denaturants were extensively dialyzed against the denaturant-free 50 mM histidine buffer pH 6.0 for 24 hours to reduce the denaturant concentration. The entire procedure was performed at room temperature in 96-deep well plates and Pierce™ microdialysis devices (3.5 kDa MWCO) during agitation at 700 rpm on a Thermomixer Comfort (Eppendorf AG, Germany). After all dialysis steps, the samples were collected, weighed on a microbalance, adjusted to the same weight with 50 mM histidine pH 6.0, centrifuged for 10 minutes at 10,000 *rcf* and analyzed by SEC-MALS. The monomer peak area after refolding was divided by the monomer peak area of the antibodies before unfolding to obtain a relative monomer yield (RMY). The RMY shows the fraction of the monomer which did not aggregate during the unfolding and refolding from denaturants.

3.2.5 Storage stability study

The formulated antibodies in 50 mM histidine buffer pH 6.0 were filled (1 mL) into 1.5-mL polypropylene microcentrifuge tubes with a screw cap and an O-ring (Corning). The protein concentration was 1 mg/mL. The samples were stored at three different temperatures (4, 25 and 40 °C). The storage time was 12 months at 4 °C and 3 months at 25 and 40 °C. The samples were analyzed by SEC-MALS at the beginning of the stability study and at the storage end at each temperature. The relative amount of aggregates was calculated with Chromeleon V7 (Thermo Fisher Scientific) from the area of the aggregate peak related to all protein peaks in the chromatogram.

3.2.6 Intrinsic differential scanning fluorimetry (nanoDSF)

The nanoDSF measurements were performed with a Prometheus NT.48 device (NanoTemper Technologies) and a ramp of 1 °C/min. The PR. ThermControl V2.1 software was used to determine the unfolding onset temperatures (T_{on}) and the first melting temperature (T_{m1}) from the change in the fluorescence intensity ratio (FI_{350}/FI_{330}).

3.2.7 Modulated scanning fluorimetry (MSF)

MSF was performed on a Prometheus NT.48 with the PR.TimeControl software (NanoTemper Technologies) as previously described.¹⁵ The samples were heated to 1 °C higher temperature and held for 1 min in each consecutive incremental heating cycle, followed by a cooling phase and hold for 5 min at 25 °C. The heating and cooling ramps (7 °C/min) are fixed in the PR.TimeControl software. The MSF analyser software was used to process the raw data.¹⁵ The onset temperature of non-reversible unfolding (T_{nr}) was determined from the 10% threshold in the increase of the fluorescence intensity signal at 350 nm.

3.2.8 Dynamic light scattering (DLS)

Prior to analysis, the samples were centrifuged for 10 minutes at 10,000 rcf. Afterwards, 50 μ L/well were filled in a 384-multiwell plate (High Content Imaging, Corning). The wells were capped with 10 μ L silicon oil and centrifuged shortly. The measurements were performed with a DynaPro DLS plate reader III (Wyatt Technology). A heating ramp was applied to the samples where the temperature was increased stepwise by 1 °C from 25 to 85 °C. The DLS plate reader collected 5 acquisitions of 3 seconds for each well at every temperature step. The number of wells in each run

was kept constant (14 wells) to ensure consistency of the data from the replicates. One full run with these settings took about 18 hours and the heating ramp between the steps was 0.07 °C/min (fixed by the Dynamics V7.8 software). The aggregation onset temperature (T_{agg}) was determined from the increase in the hydrodynamic radius (R_h) using the onset fit in the Dynamics V7.8 software.

3.2.9 Statistical analysis

Calculation of mean values and standard deviations were performed with Origin 2019 (OriginLab Corporation, USA).

3.2.10 Computational characterization

The sequences of ten of the antibodies are available in Protein Data Bank or in published work.⁸ No sequences were available for PPI04, PPI11 and PPI08, which precluded us from doing computational characterization with these three proteins. The primary sequences of the variable (V_H and V_L) domains were uploaded to the ABodyBuilder tool from the SAbPred toolbox to obtain homology models of each Fv.²⁵ Using the homology models, all-atom simulations were performed with the Amber19 program in a periodic box with explicit solvent.²⁶ The ff14SB force field for proteins was employed in combination with the TIP3P water model. The protonation states of ionizable residues at pH 6.0 were adjusted using the H++ server.²⁷ All bonds involving hydrogen atoms were constrained using the SHAKE algorithm. Non-bonded electrostatic interactions were treated using the particle mesh ewald algorithm with a direct space cut-off of 8 Å. The native-state structures of the antibody variable domains were solvated in a truncated octahedral water box with a layer of at least 20 Å from the protein surface and neutralized with chloride ions. The system was energy minimized with the steepest descent algorithm for the first 5000 cycles, followed by 5000 cycles using the conjugate gradient method. We applied incremental heating with steps of 20 K from 300 K to 400 K and steps of 10 K from 400 K to 540 K by using the final coordinates and velocities of the previous temperature simulation as input for the next temperature step. System equilibration was carried out for 1 ns in NVT ensemble to stabilize the specified temperature using the Langevin thermostat, and subsequently for 1 ns in NPT ensemble to adjust the density of the system using the Berendsen barostat. Initial box sizes at 300 K are between 708325 and 867295 Å³ but increase due to the density adjustment in each NPT run to values between 1160140 and 1433494 Å³ at 540 K. The simulations at each temperature were performed for 20 ns with a time step of 1 fs. The coordinates were saved every 5 ps. Trajectories were visually inspected in VMD²⁸

to verify that there are no dimensional constraints due to the chosen box size and shape. All trajectories were analysed using the CPPTRAJ module of Amber19; three frames of each temperature simulation were exported as PDB files and the beta sheet fraction was calculated using the DSSP method²⁹ by Kabsch and Sander and averaged for each temperature simulation. The structurally corrected solubility scores of the Fv regions were determined by submitting the three frames of each temperature simulation to the CamSol webserver 8, setting the pH to 6.0 and the patch radius to 10 Å.³⁰ The obtained solubility scores were averaged for each temperature, respectively. The resulting beta sheet fractions and the solubility scores from CamSol were plotted against the simulation temperatures and fitted to a Boltzmann function. Five CamSol solubility thresholds were defined and the temperatures at which the solubility of the Fv regions reach these thresholds determined from the Boltzmann fit. The obtained temperatures were then plotted for each antibody Fv region.

3.3 Results

3.3.1 Isothermal chemical denaturation and the ReFOLD assay to compare antibodies

At first, we asked whether the thirteen antibodies have different resistance to unfolding caused by two commonly used chemical denaturants. We therefore obtained isothermal chemical denaturation (ICD) curves of all proteins in GuHCl and urea (Figure S6). We then fitted the curves to a three-state model to obtain the melting denaturant concentrations (C_m1 and C_m2) (Figure S6) (Table S2). In this context, C_m is an apparent value showing the resistance to denaturant-induced unfolding. An earlier study has shown that an apparent C_m correlated with the aggregation rate of antibody formulations at 50 °C,³¹ thus, it was worth exploring the correlation of C_m to other variables in the context of developability assessment.

Higher concentrations of urea induced protein unfolding compared to GuHCl (Figure S6). However, the same plateau of the intrinsic protein fluorescence ratio of the unfolded state was reached in both denaturants. The only exceptions to this observation were PPI02, PPI10 and PPI13, which unfolded completely in 5 M GuHCl but only partially in up to 9 M urea.

Subsequently, we used a microdialysis-based (ReFOLD) assay to study if the antibodies can refold to their monomeric state after isothermal unfolding with GuHCl or urea.²⁰ The proteins exhibited very different RMYs and aggregate distributions after refolding (Figure 8a,b and Figure S7). These

observations revealed that we have a panel with antibodies that show considerable differences regarding their ability to refold as monomers after unfolding with denaturants. Some proteins like trastuzumab, PPI03, PPI13, PPI10, exhibited high RMYs in both denaturants, while others like PPI01, bevacizumab, PPI11, PPI04 had low RMYs after refolding from both GuHCl and urea (Figure 8c). There were also pronounced differences in the size distribution of the aggregates detected after refolding of different antibodies (Figure S7). For example, trastuzumab formed $15.6 \pm 0.2\%$ dimers and $15.8 \pm 0.7\%$ aggregates with larger sizes, while bevacizumab formed $78.4 \pm 0.7\%$ aggregates larger than dimers (Figure 8a,b).

In general, refolding from 8 M GuHCl resulted in lower RMYs compared to refolding from 10 M urea (Figure 8c). The high ionic strength when using GuHCl screens repulsive electrostatic interactions, which explains the lower RMY values measured compared to urea.³² The only exception from this trend is PPI01. Low ionic strength causes oligomerization of PPI01 around pH 6.0.³³ Therefore, the high ionic strength of the GuHCl solutions favoured the monomeric state and is most probably the reason for the higher RMY of PPI01 after refolding from GuHCl compared to urea.

Notably, the RMY did not correlate with the C_{m1} and C_{m2} when we used GuHCl (Figure 8d). Therefore, the resistance to GuHCl-induced unfolding (C_{m1} and C_{m2}) provides different information than the refoldability (RMY) of these antibodies. There was a weak correlation between the RMY and C_{m2} from experiments with urea ($R^2 = 0.56$), which indicates that such refoldability experiments could also carry some information about the resistance to unfolding induced by urea.

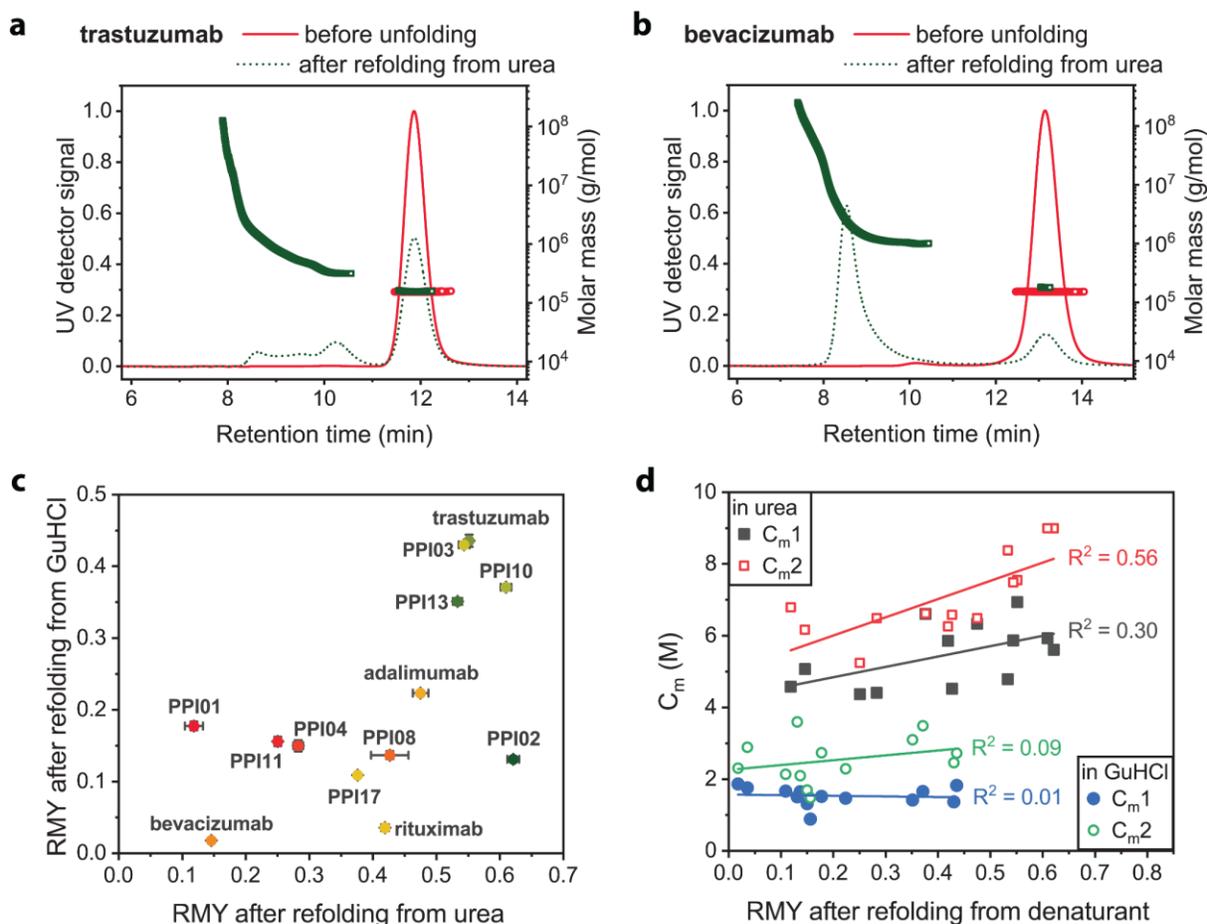


Figure 8: Analysis of the antibodies with the ReFOLD assay and ICD. SEC-MALS chromatograms of (a) trastuzumab and (b) bevacizumab before unfolding and after refolding from 10 M urea. (c) Comparison between the relative monomer yield of 13 antibodies after refolding from GuHCl and urea. Mean values of triplicates with standard deviations. (d) Correlations between the RMY after refolding from denaturants and the respective C_{m1} and C_{m2} from the fits to the ICD data in Figure S6.

3.3.2 Antibody unfolding, refolding and aggregation after heating

In addition to the characterization shown above, we wanted to compare the antibodies with orthogonal techniques that do not employ chemical denaturants. Besides using chemical denaturants, the other common approach to unfold proteins is to increase sample temperature. Furthermore, alternating heating and cooling cycles provide information on whether the protein unfolding is reversible.

We therefore used MSF to find which temperatures cause unfolding of the antibodies and whether the observed unfolding is reversible.¹⁵ The different domains of an antibody can unfold at different

temperatures and can potentially exhibit different unfolding reversibility. Indeed, MSF showed that the antibodies in our dataset behave very differently during heating and cooling. For example, PPI01 started unfolding at around 55 °C and this early unfolding is not reversible as evident from the changes in the non-reversibility curve (Figure 9a). In contrast, PPI13 showed an early unfolding transition at around 60 °C but the baseline of the non-reversibility curve did not change around that temperature (Figure 9b). The non-reversibility onset of PPI13 occurred when the protein was exposed to temperatures around 80 °C that induced a second unfolding transition (Figure 9b).

Furthermore, we derived the non-reversibility onset temperature (T_{nr}) from MSF (Figure S8) and used T_{nr} to rank the thirteen antibodies (Figure 9c). Circular dichroism measurements confirmed that the exposure of the proteins to temperatures below T_{nr} does not lead to substantial structural changes, while exposure to temperatures above T_{nr} severely impacts the tertiary protein structure (Figure S9). Looking for a mechanistic explanation of the non-reversibility onset temperature, we compared the T_{nr} values to other biophysical variables.

First, we used nanoDSF to measure the unfolding onset temperature (T_{on}) of each antibody (Table S2). The T_{on} is a parameter indicating the temperature where the protein starts to unfold. This has previously been demonstrated in various studies, for example, by a comparison of the unfolding curves of antibodies measured with nanoDSF and calorimetry.³⁴ The T_{nr} s of the antibodies did not correlate strongly with their T_{on} s (Figure 10a).

Second, we determined the first melting temperature (T_{m1}) of the proteins with nanoDSF (Table S2). The T_{nr} s of our antibodies did not correlate with their T_{m1} s (Figure 10b). Six of the antibodies also showed a second unfolding transition that can be used to determine T_{m2} . Since the unfolding at higher temperature could also be relevant for stability, we tested the correlation between T_{nr} and the highest melting temperature of each antibody (Table S2); in other words, we used the only T_m for proteins with one unfolding and the T_{m2} for proteins with two unfolding transitions. In this case, the correlation between T_{nr} and T_{m2} was better ($R^2 = 0.75$).

Third, we used dynamic light scattering (DLS) to measure the aggregation onset temperature (T_{agg}). DLS was used to measure the T_{agg} since it is more sensitive towards small aggregates that can be missed by the backscattering approach of the Prometheus NT.48.³⁵ The T_{agg} did not correlate with the T_{on} (Figure 10c), showing that the start of the unfolding does not always cause aggregation of these antibodies, which is in good agreement with previous findings.⁸

Fourth, we compared the T_{agg} with the T_{nr} and observed a strong correlation between these two variables (Figure 10d). Therefore, we concluded that aggregation is the main reason for the non-reversibility onset in MSF experiments with the antibodies in this work.

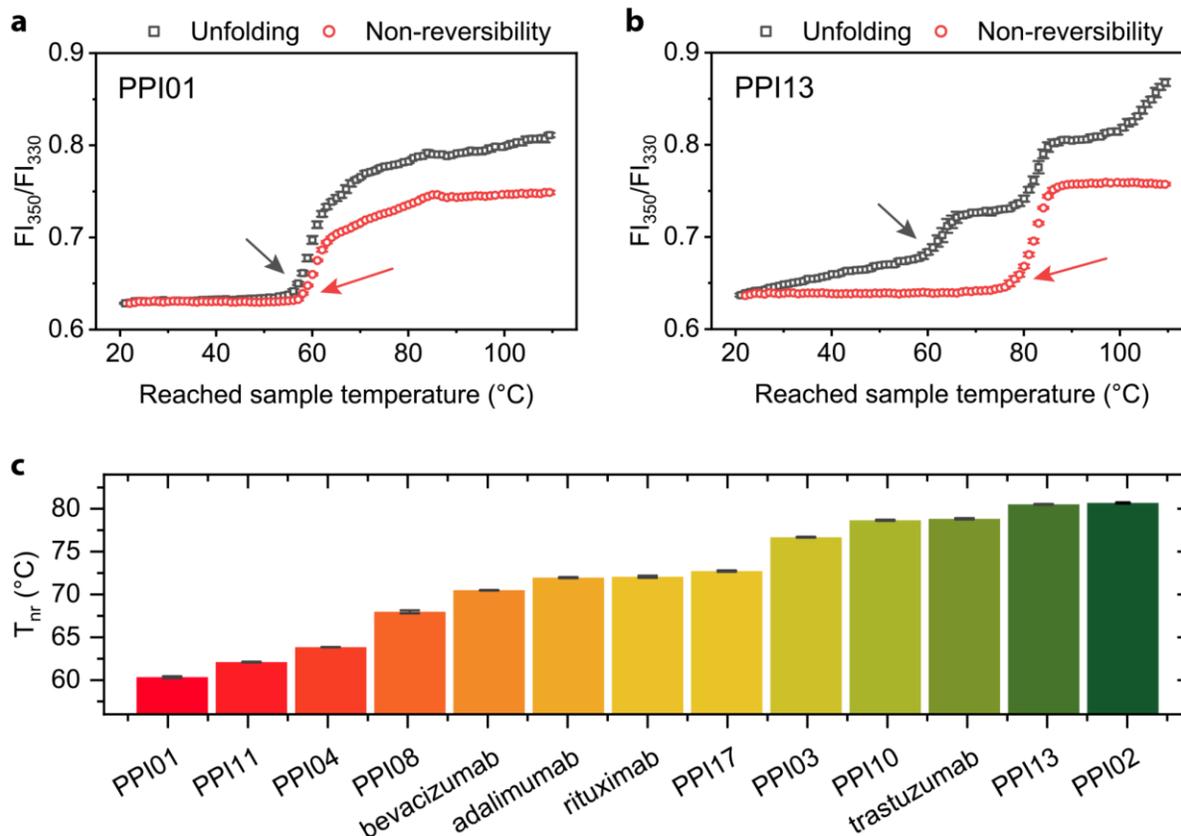


Figure 9: Analysis of the antibodies with MSF. Unfolding and non-reversibility curves of (a) PPI01 and (b) PPI13 (the start of the unfolding is indicated by gray arrows, T_{nr} values are indicated by red arrows). (c) T_{nr} values of the antibodies in this work. Mean values of triplicates with standard deviations. The color code of the antibodies is the same as in Figure 8c.

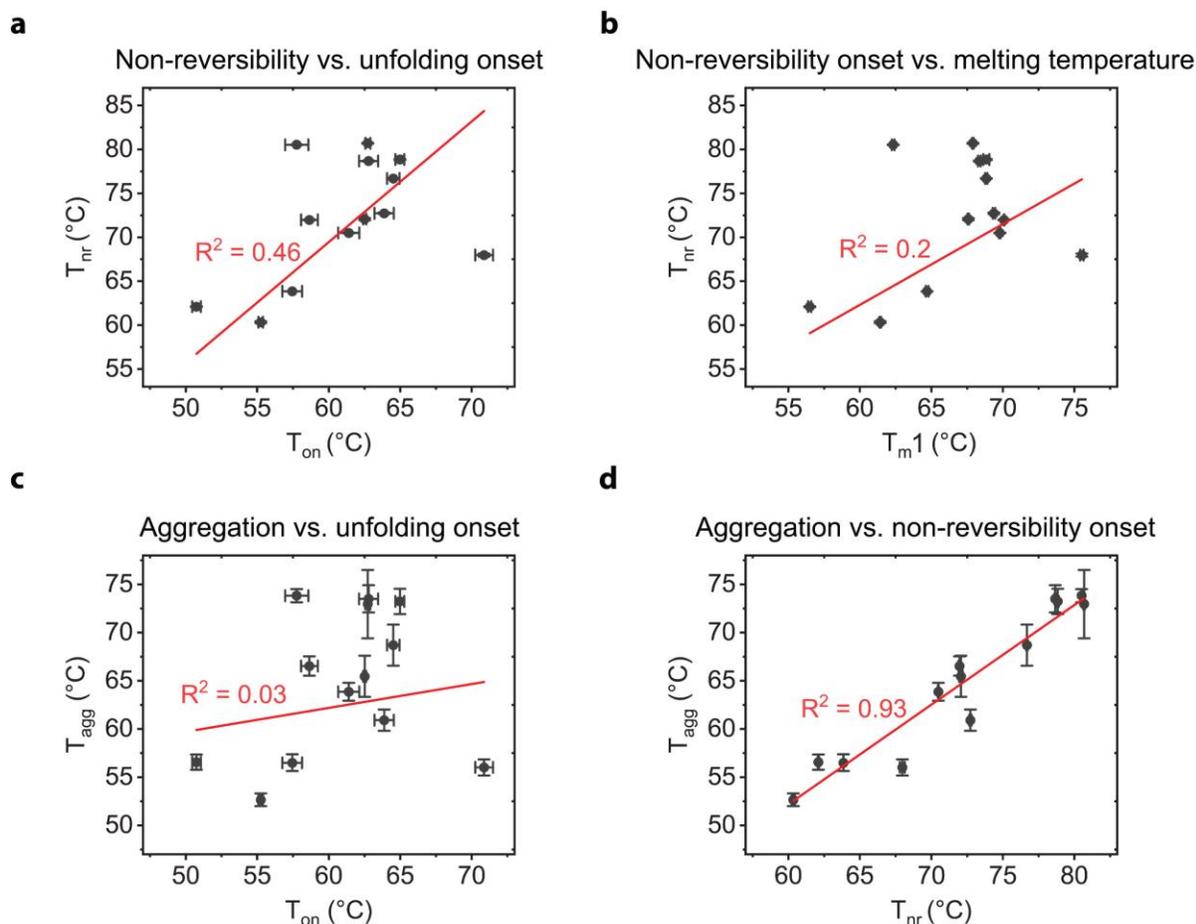


Figure 10: Comparison of biophysical variables obtained from thermal denaturation techniques. Weak correlation (a) between T_{nr} and T_{on} , (b) between T_{nr} and T_{m1} , and (c) between T_{agg} and T_{on} . (d) Strong correlation between T_{agg} and T_{nr} . Mean values of triplicates with standard deviations.

3.3.3 Aggregation during storage of the antibodies

We were interested in whether the high RMY from the ReFOLD assay and high T_{nr} from MSF are features of antibodies that aggregate less during storage. We therefore analyzed the antibodies with SEC to determine the relative area of aggregates (Table S2). Subsequently, we performed storage stability studies for 3 months at 25 and 40 °C, and for 12 months at 4 °C and analyzed the samples again with SEC (Table S2). The largest differences were observed after storage at 40 °C. We therefore calculated the change in the relative area of aggregates after storage at 40 °C (Figure 11).

The SEC analysis of the fresh samples and the samples after storage at different temperatures revealed two important pieces of information. First, we obtained the relative area of aggregates that were present already at the beginning of the stability study. The exact origin of these aggregates is

outside the scope of this work, but they are likely formed during earlier processing steps like protein expression, purification, freezing, and buffer exchange. Antibodies with drug-like properties will ideally have low aggregate content after processing. Second, we could observe whether the relative area of the aggregates increased after storage. Such an increase indicates that an antibody is prone to aggregation.

The SEC data revealed that there were several antibodies with more than 1% relative area of aggregates already at the start of the stability study (Table S2). These antibodies were PPI11 (8.5% aggregates), PPI10 (3.2% aggregates), PPI17 (3% aggregates), PPI04 (2.4% aggregates), PPI13 (1.9% aggregates), bevacizumab (1.2% aggregates), rituximab (1.1% aggregates) and adalimumab (1% aggregates).

After storage for 3 months at 40 °C, the following antibodies exhibited an increase in the relative area of aggregates of $\geq 0.5\%$ – PPI04, PPI08, bevacizumab, PPI01, rituximab, and PPI11 (Figure 11). After storage for 3 months at 25 °C or storage for 12 months at 4 °C, only PPI11 showed an increase in the aggregate area of more than 0.5% (Table S2).

An interesting observation is that the relative area of aggregates decreased by more than 0.5% after storage of PPI10 at all three temperatures and for PPI13 after storage at 4 and 25 °C. Thus, the aggregates that were present in the bulk solutions of PPI10 and PPI13 were reversible and dissociated slowly upon dialysis, dilution, and storage of the antibodies in the histidine buffer used in this work. Aggregate dissociation upon buffer exchange and dilution was previously reported but is outside the scope of our work.³⁶

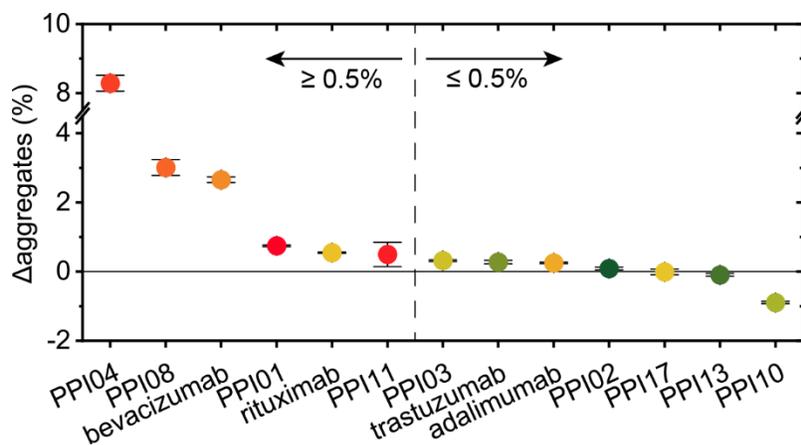


Figure 11: Aggregate increase after storage of the proteins. Change in the relative area of aggregates detected by SEC after storage for 3 months at 40 °C. Mean values of triplicates with standard deviations. The color code of the antibodies is the same as in Figure 8c and Figure 9c.

3.3.4 Comparison between ReFOLD, MSF and aggregation during storage

We can divide the antibodies in our work into three groups based on their performance in the biophysical assays and the storage stability study.

The first group comprises four antibodies (trastuzumab, PPI03, PPI10 and PPI13) that showed both high RMYs after refolding from denaturants (Figure 8c) and T_{nr} values above 75 °C (Figure 9c). None of these four antibodies exhibited an increase ($\geq 0.5\%$) in the relative area of aggregates after storage (Figure 11).

The second group includes PPI02, PPI17 and adalimumab. These three proteins had suboptimal performance in some of the assays. For example, PPI02 had the highest T_{nr} (Figure 9c) but a low RMY after refolding from GuHCl (Figure 8c). PPI17 and adalimumab had RMYs and T_{nr} that were average compared to the other antibodies (Figure 8c and Figure 9c). We did not observe an increase ($\geq 0.5\%$) in the relative area of aggregates after storage of PPI02, PPI17 and adalimumab (Figure 11).

The third group contains the remaining antibodies (PPI01, PPI04, PPI08, PPI11, bevacizumab and rituximab) that had the lowest RMYs after refolding from denaturants (Figure 8c) and the lowest T_{nr} s measured with MSF (Figure 9c). The antibodies with low RMYs and T_{nr} s showed an increase ($\geq 0.5\%$) in the relative area of aggregates after storage (Figure 11). Moreover, PPI11 also formed aggregates during storage at 4 and 25 °C (Table S2).

3.3.5 Preliminary *in silico* characterization of the antibodies

We were interested in whether the different unfolding reversibility and aggregation of the antibodies can be explained by structural differences that can be studied *in silico*.

We had access to sequences for 10 of the 13 proteins, some of them available in the DrugBank (<https://go.drugbank.com/>) or published literature.⁸ Previous work showed that the PPI antibodies differed in their isoelectric points, predicted scale solubility, hydrophobic patch score and aggregation score (see Table 1 and supplementary data in Ref⁸). However, these characteristics did not explain the aggregation propensity of the antibodies.⁸ We therefore adopted a different unanimous approach to study the commercial and PPI antibodies in this work.

We used the TANGO³⁷ and AggreScan³⁸ web servers to look for APRs in the primary structures that could provide mechanistic insights into the different aggregation of the antibodies. The V_H and V_L domains contain multiple APRs that differ between the proteins (Figure S10). Contrary, the constant domains show the same APRs, which is because most of the proteins are of the IgG1 subclass and have a kappa light chain. One exception is PPI01 that has a lambda light chain and has APRs in its C_L domain that differ from the APRs in the C_L domains of the kappa light chains in this work (Figure S10). Noteworthy, the frameworks of kappa and lambda light chains are also different in respect to V_L residues contributing to their fold stability.³⁹

To assess the position of the APRs in the three-dimensional protein structure, we created homology models of the Fv regions. We focused on the Fvs because these are the parts of the proteins where the sequences differ most. A significant part of the determined APRs in the Fv regions comprises residues in the beta strands that are not solvent exposed in the native folded state but could contribute to non-native aggregation upon unfolding. To reflect better on the position of aggregation-prone regions in the native protein, we also analyzed the homology models with the structure-based CamSol method that accounts for the residue proximity and solvent exposure. The CamSol scores revealed differences in the solubility of the Fv regions (Figure S11). However, these differences capture features of the native folded proteins. In the context of unfolding and non-native aggregation, the risk that buried APRs are exposed should also be considered. Therefore, *in silico* experiments to cause unfolding and assess the aggregation potential of partially unfolded species were the next step in our work.

3.3.6 Temperature-ramped molecular dynamics (MD) simulations

Since our experimental characterization focused mostly on the aggregation that is driven from partially unfolded species, we were interested in obtaining partially unfolded species *in silico* and studying their aggregation potential with available webservers like CamSol.

We first tried to cause unfolding of the Fvs by including chemical denaturants in the simulation box, but this approach proved very resource-consuming and was therefore abandoned. The other traditional approach to unfold proteins *in silico* is to increase the temperature in the simulation.^{40,41} To investigate the effect of different temperatures on the unfolding of the Fvs, we performed MD simulations with incremental heating steps. At each step, the simulation continued for 20 ns to allow for the unfolding to take place. Three snapshots of Fv structures at each step/temperature were analyzed with the CamSol webserver.

The folded Fv regions show some differences in exposed APRs at 300 K (Figure 12a and b). As the simulation temperature is increased, we observed protein-specific unfolding of beta strands. Noteworthy, the unfolding did not always lead to the exposure of APRs. For example, at 460 K the first beta strands to unfold in the trastuzumab Fv are near the C-terminus in V_H and the N-terminus in the V_L, together with a short beta strand near CDR-H2. Unfolding of these strands does not cause exposure of APRs in trastuzumab Fv (Figure 12a). In contrast, at 460 K the unfolding of bevacizumab Fv is characterized by a loss of structure in the N-terminal beta strands in the V_H and in a short beta strand close to the CDR-H2. The unfolding of the latter causes a disruption of a network of intramolecular interactions and exposes a considerable number of APRs (Figure 12b). At temperatures of 500 K, both Fvs have lost almost all their secondary structure with only a few beta strands intact in the V_H domain of trastuzumab and in the V_L domain of bevacizumab. At this state of unfolding, several APRs are solvent exposed in both Fv regions (Figure 12a and b).

To visualize better the interplay between unfolding and exposure of APRs, we plotted the beta sheet fraction and the solubility score from CamSol for the partially unfolded species obtained at different temperatures (Figure 12c and d). The Fv region of trastuzumab lost beta sheet structure at high temperatures and the drop in solubility occurred only after a notable unfolding (Figure 12c). In contrast, the Fv of bevacizumab exhibited lower solubility already with the unfolding of the first beta strands that occurred at lower temperatures compared to trastuzumab Fv (Figure 12d).

To compare the differences in the solubility changes of the Fv regions, we defined five CamSol solubility thresholds (Figure 12e). We then plotted the temperature of the MD simulation that resulted in partially unfolded species with a solubility score below these thresholds (Figure 12f). From this data we saw that some Fv regions, like from PPI13, PPI02, adalimumab and trastuzumab, exhibited a solubility loss over a narrow temperature range and at higher temperatures than bevacizumab and rituximab.

The Fv region of PPI01 behaved similarly to the Fv regions of stable antibodies in the dataset (Figure 12f). PPI01 is the only antibody in the dataset that has a different C_L domain. Both the TANGO and AggreScan analysis indicated that the C_L domain of PPI01 has different APRs compared to the C_L domains of the other antibodies (Figure S10). We therefore performed the MD simulations combined with CamSol to compare the unfolding and solubility drop of C_L domains from PPI01 or the other antibodies. The C_L domain of PPI01 lost beta sheet structure and showed a drop in solubility at lower temperatures than the C_L domain of trastuzumab (Figure S12a).

Noteworthy, the cut-off for electrostatic and van der Waals interactions in the MD simulations is important.⁴² A longer cut-off might improve the quality of the simulation, but increases the computational costs. To test whether the chosen cut-off (8 Å) leads to artefacts in our simulations, we conducted simulations on trastuzumab Fv and bevacizumab Fv with a cut-off of 10 Å and observed that the proteins unfolded at slightly higher temperature but with the same trend (Figure S12b and c).

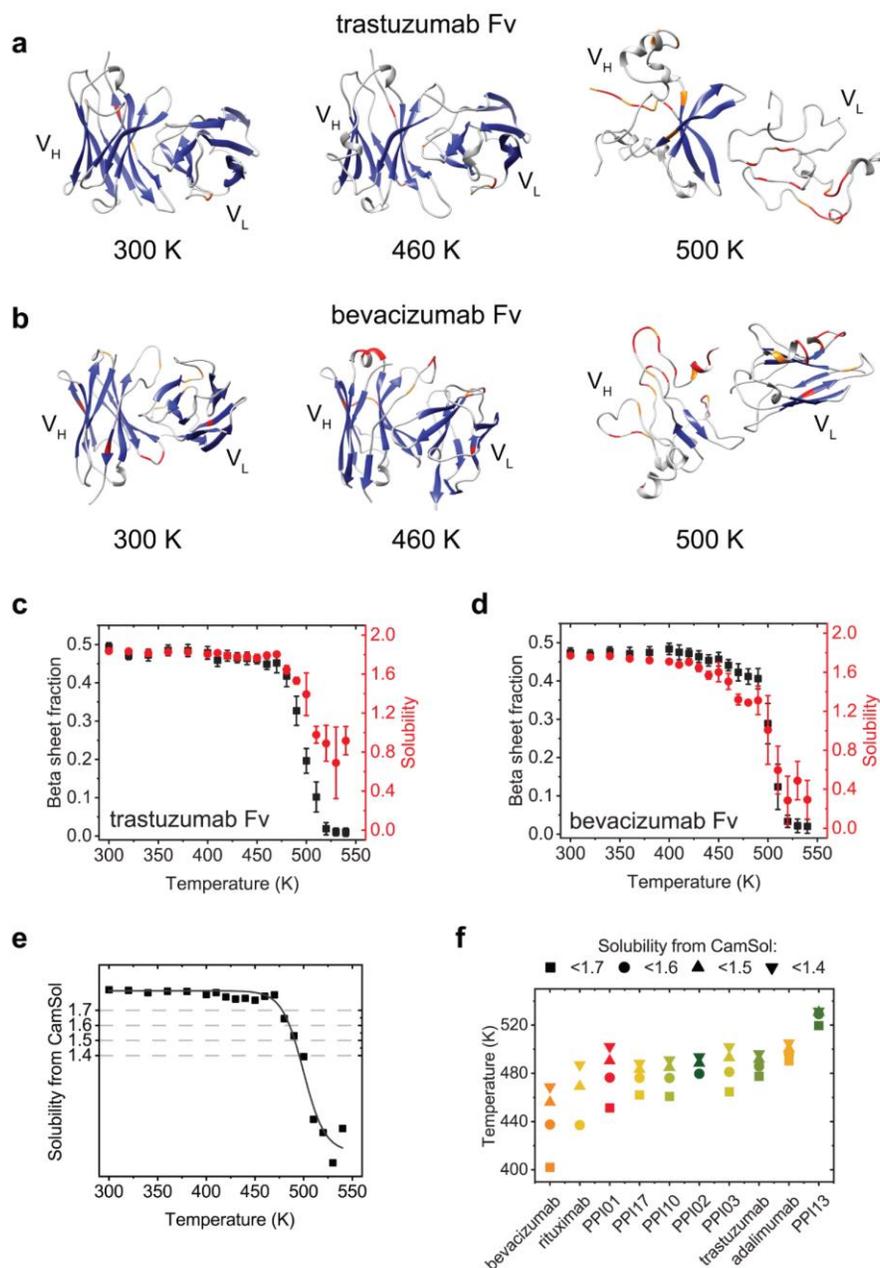


Figure 12: *In silico* characterization of antibody Fv regions. (a, b) MD simulations at different temperatures were used to obtain partially unfolded species. The beta sheets are colored in blue. The coloring of the Fv strands is based on CamSol scores of the residues - orange: values from -0.8 to -1, red: values below -1. Thus, red indicates APRs. (c, d) Plots of the loss in beta sheet fraction (black squares) and the solubility from CamSol (red circles) of the partially unfolded species reveal differences between the antibody Fv regions. (e) Different thresholds were set to report the temperature at which the solubility of the partially unfolded Fvs from CamSol decreases. (f) The temperatures at which the CamSol solubility of the Fv regions fall below a certain threshold. The color code is like in Figure 8c, Figure 9c and Figure 11.

3.4 Discussion

Here we investigated whether unfolding reversibility studies can identify the aggregation-resistant antibodies in a group with thirteen molecules. Two earlier findings inspired our work. First, the thermal unfolding reversibility of antibody domains is a feature of aggregation-resistant molecules.^{13,14,43–45} Second, antibodies aggregate less during storage in formulations favouring refolding to monomers after isothermal unfolding with chemical denaturants.^{20,21}

Using the ReFOLD assay, we observed that some antibodies have a higher tendency to refold to monomers after unfolding with urea and GuHCl (Figure 8c). The high RMYs indicate that the unfolded states of these antibodies are less prone to aggregation than antibodies with low RMYs. The RMY from the ReFOLD assay did not correlate strongly with the C_m s obtained with ICD (Figure 8d). This poor correlation underlines the difference between the resistance to denaturant-induced unfolding and aggregation via the unfolded state caused by denaturants.

In comparison to several other approaches to study protein aggregation after unfolding with denaturants,^{18,19,32,46} the ReFOLD assay is relatively simple to perform. It requires only the protein sample, the sample buffer and one denaturant stock solution. All steps are performed on standard laboratory equipment with commercially available microdialysis devices. An HPLC-SEC system is sufficient for basic sample analysis. Compared to thermal denaturation techniques, the ReFOLD protocol induces more moderate protein aggregation at ambient temperatures, revealing differences overlooked by other techniques.²⁰ In contrast to denaturant-induced unfolding, substantial antibody unfolding caused by heat usually leads to complete monomer loss.⁴⁷

Our MSF experiments showed that the non-reversibility effects of the thermal unfolding of our antibodies are protein-specific (Figure 9a and b). Compared to calorimetric techniques, MSF provides the advantage of high throughput to apply incremental heating and cooling cycles. MSF can therefore be used as a technique during the developability assessment of antibodies.

The T_{nr} from MSF did not correlate with the T_{on} or T_{m1} from nanoDSF; however, T_{nr} correlated better with the unfolding transition detected at higher temperatures (T_{m2}). It has been shown that the unfolding of the Fab can vary a lot in nanoDSF experiments,³⁴ and sometimes overlaps with the C_{H2} domain or the C_{H3} domain. The vastly different unfolding temperatures of Fabs present a complication for stability predictions based on melting temperatures. An assumption that antibodies that unfold at lower temperatures are prone to aggregation will not hold in cases where

the T_{on} and T_{m1} are not related to irreversible unfolding. Therefore, the determination of T_{nr} is rational since this variable is related to the domains that exhibit irreversible structural changes. The T_{nr} could be subsequently assigned to a specific part of the antibody in MSF experiments on individual domains.

We found that the T_{nr} from MSF correlates well with the T_{agg} determined with DLS. Therefore, we concluded that non-native aggregation was the main reason for the non-reversibility onset in MSF experiments with the antibodies in this work. At this point, it might appear that the information from MSF and DLS is redundant. However, MSF and temperature-ramped DLS should be considered orthogonal techniques due to the completely different heating profiles (incremental cycles versus linear ramp) and different physical observable (fluorescence intensity and light scattering fluctuations).

Considering the urgency of understanding protein aggregation in a larger context,⁴⁸ the reversibility of both aggregation and protein unfolding should be studied. On the one hand, MSF is fluorescence-based and can detect irreversible structural changes that might or might not be related to aggregation;¹⁵ on the other hand, DLS can detect aggregation that does not necessarily include irreversible structural changes.³³ These different perspectives on the behaviour of a protein during heating and cooling could dissect distinct processes that were previously overlooked. In the context of the current study, we observed that the T_{nr} and T_{agg} provided similar information because aggregation is the main reason for the irreversible thermal unfolding of these antibodies, which was not known before. In future, we aim to provide more specific examples that demonstrate when the combination of MSF and DLS provides a better mechanistic understanding of protein stability.

Different rules and guides for the selection of antibodies with drug-like properties were proposed recently. These approaches focus on key antibody features like folding stability, aggregation propensity, solubility, viscosity, charge in the CRDs and specificity.⁴⁹⁻⁵² Here we propose that refoldability studies can also be included in the developability assessment programs for therapeutic antibodies. In such a case, one would ideally use a combination of at least two methods to study refoldability – one that causes unfolding by an increase in temperature (MSF) and one isothermal method in which the unfolding is caused by chemical denaturants like urea and GuHCl (the ReFOLD assay). Depending on the performance of the antibodies in these two orthogonal methods to study refoldability, one can assign molecules to four groups (Figure 13a). Group A is the ideal

case comprising aggregation-resistant proteins that have high RMYs and high T_{nr} s. Examples of antibodies in Group A are trastuzumab, PPI03, PPI10, and PPI13. Group B includes proteins with high T_{nr} but low RMYs, while Group C encompasses proteins with low T_{nr} but high RMYs. Antibodies belonging to Group B or C can still be aggregation-resistant (e.g. PPI02), but they will most likely require characterization with more techniques to understand their stability. Finally, Group D includes aggregation-prone proteins with low RMYs and low T_{nr} s. Examples for antibodies in Group D are PPI01, PPI04, PPI08, and PPI11. The molecules in Group D will be problematic during development and storage. It is important to say that the cut-offs between the groups are for illustrating the idea and probably do not present values that will be true for each antibody. However, by applying the presented concept to a large antibody library (e.g. 100s to 1000s of candidates) one could provide cut-off values for T_{nr} and RMY with more widespread validity in future.

The antibodies in this work were relatively stable and most of them aggregated only during storage at 40 °C. The differences in the aggregation behaviour at different temperatures is a common issue with antibodies that obstructs the use of accelerated stability data to predict the storage stability at 2-8 °C.⁵³⁻⁵⁶ One of the antibodies (PPI11) also aggregated during storage at 4 °C and 25 °C. It is encouraging to see that both ReFOLD (Figure 8c) and MSF (Figure 9c) indicated that PPI11 is prone to aggregate. In future, we will test our approach on more antibodies that aggregate during storage at refrigerated temperatures.

In respect to storage stability prediction, one hypothesis says that the amounts of aggregates at the beginning of a stability study indicate which molecules are aggregation-prone. While there might be cases where this is true, our storage stability data shows that there is no correlation between the amount of aggregates before storage and the percentage of aggregates formed after 3 months at 40 °C (Figure S13a). Thus, storage stability prediction should be based on more comprehensive hypotheses beyond the initial amount of aggregates.

A limitation of the current study is that we used only SEC to monitor aggregate formation during storage. In principle, it is possible that larger aggregates (subvisible and visible particles) that are filtered by the SEC column are also formed during storage stability studies; however, we did not observe visible particles in the samples or a reduction in the area of the soluble antibody that was found in the SEC chromatograms (Figure S13b).

Interestingly, one can see connections between our unfolding reversibility studies and previous work on stability and developability assessment of antibodies. For example, bevacizumab and rituximab exhibited poor unfolding reversibility in our work and a suboptimal performance in biophysical assays for developability assessment in earlier studies.⁷ Indeed, bevacizumab is an antibody that required reformulation during clinical studies due to poor physical stability in histidine buffer.⁵⁷ Like bevacizumab, rituximab is also known to aggregate under weak thermal stress.⁵⁸ Both bevacizumab and rituximab were developed before the concepts for developability assessment of antibodies emerged, which can explain why they were selected for further development despite a poor performance in some biophysical assays. In contrast to bevacizumab and rituximab, trastuzumab and adalimumab showed higher unfolding reversibility and no issues in developability assessment studies with various techniques.⁷ In addition, antibodies with good unfolding reversibility like PPI02, PPI03 and PPI13 were identified as well-behaved molecules by various biophysical assays, while antibodies with poor unfolding reversibility like PPI01, PPI04 and PPI08 raise stability concerns during extensive characterization.^{8,33}

We characterized the antibodies in one buffer during this study. One could obtain more accurate predictions when each candidate is characterized in at least several different formulations.⁸ If one has the analytical capability and protein material, the assessment of T_{nr} and RMY can be done in different formulation conditions, thus combining developability assessment and pre-formulation development; this would most probably provide more accurate candidate selection, but at the expense of more tedious characterization at very early stages of product development.

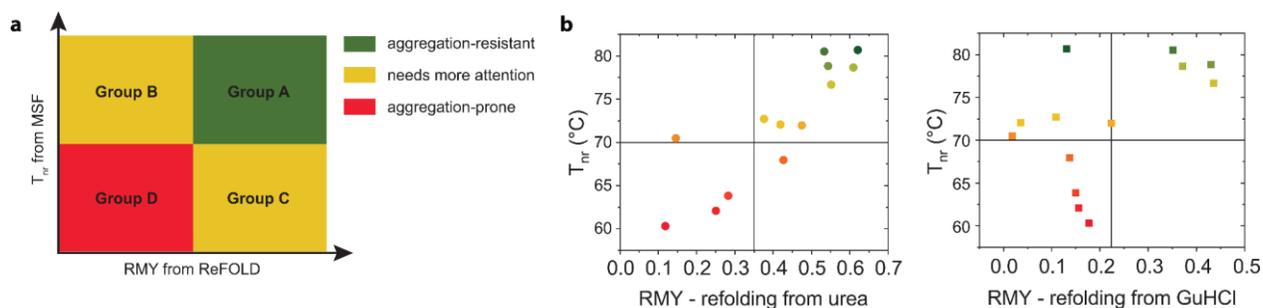


Figure 13: Proposed groups for classification of therapeutic antibodies based on their performance in the ReFOLD assay and MSF. (a) Four groups of proteins can be defined from T_{nr} and RMY. **(b)** Application of the proposed classification to the 13 antibodies in this work. The color code is like in Figure 8c, Figure 9c, and Figure 11.

In future, our goal will be to predict the refoldability and aggregation propensity of therapeutic protein candidates *in silico* before expression. Previously, different structural traits were suggested as main drivers for the thermal unfolding reversibility or the aggregation through partially unfolded species induced by denaturants. Two notable traits are the balance between positive/negative residues (overall negative charge is a feature of some antibody domains that exhibit reversible thermal unfolding) and to the arginine/lysine ratio (replacing arginine with lysine reduces the aggregation through non-native contacts of a single-chain antibody variable domain).^{14,46} The colloidal stability of the unfolded species is a third trait that seems to be related to the ability of a protein to refold.⁵⁹

In the context of developability assessment of therapeutic proteins, it will be important to have *in silico* approaches for the selection of molecules resistant to non-native aggregation. Noteworthy, most servers used for developability assessment of antibodies are focusing on the native folded protein structure, on the primary structure, or only on the complementarity-determining regions (CDRs). Inspired by earlier studies,^{40,41,60} our approach to induce partially unfolded species by MD simulations and assess the solubility of these species with web servers is a good starting point for future *in silico* studies that aim to predict the refoldability and aggregation propensity of partially unfolded proteins. Doing MD simulation on the Fvs saves computational resources and allows the screening of many antibodies; however, it is possible that potential destabilizing effects from the constant domains are overlooked. For example, if some of the antibody constant domains like C_H2 unfold before the variable domains, this could potentially have a negative effect on the stability of the Fv. To account for this, the goal in future is to perform *in silico* unfolding studies on the full-length antibodies.

In summary, our work shows that unfolding reversibility studies with the ReFOLD assay and MSF can be used to select antibodies that resist aggregation at 40 °C from a group with candidates for further development. To facilitate the selection process, we proposed to classify the antibodies into four groups based on RMYs and T_{nr} values. In addition, temperature-ramped MD simulations showed that the variable domains of aggregation-prone antibodies unfold and expose APRs at lower temperatures compared to aggregation-resistant counterparts. The proposed approaches and concepts herein can be readily included in the developability assessment programs in pharmaceutical companies.

3.5 Supporting information

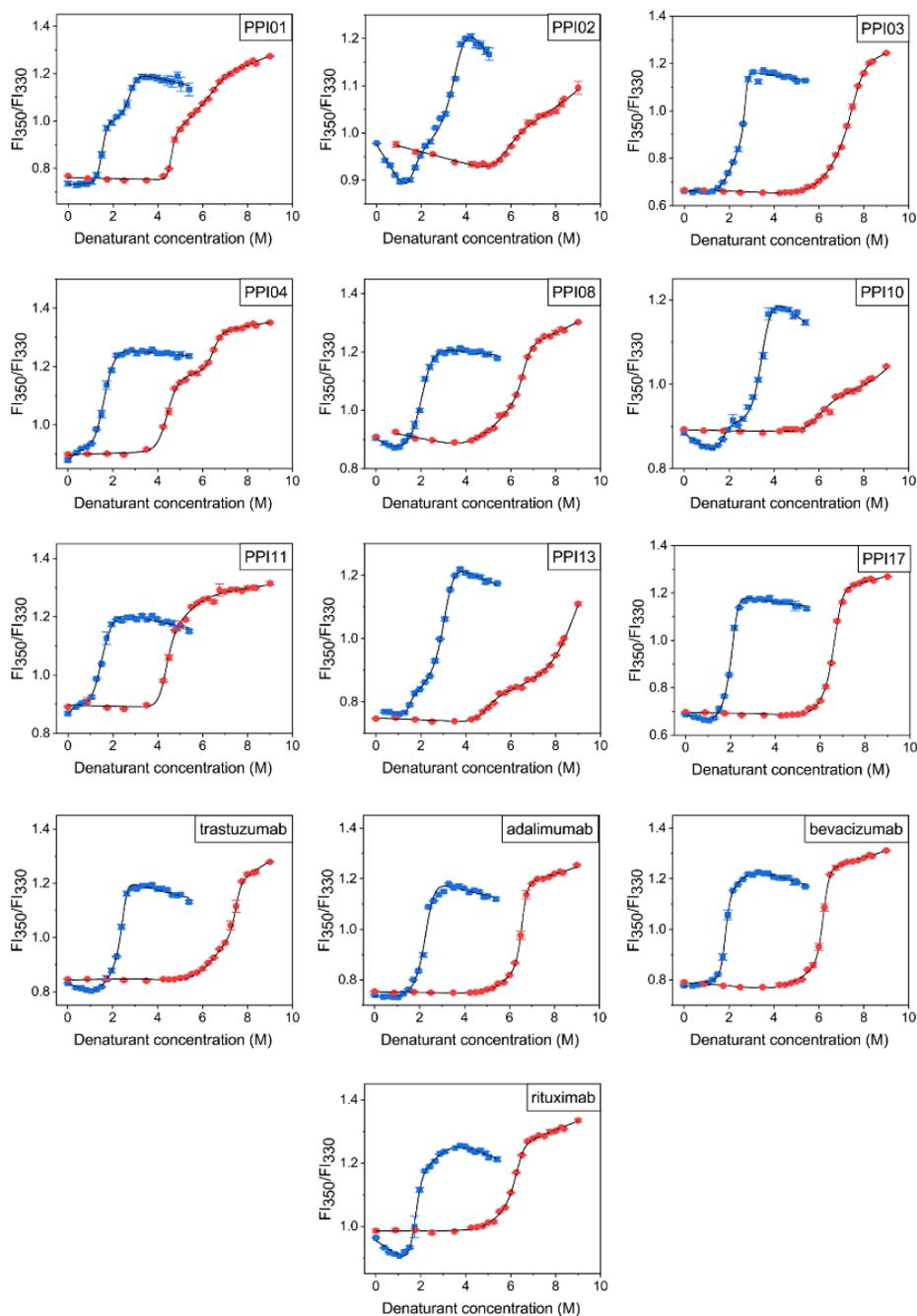


Figure S6: Isothermal chemical denaturation curves of the antibodies. The curves are obtained with GuHCl (blue squares) or urea (red circles). The lines present a fit to a three-state unfolding model with CDpal to determine the apparent melting denaturant concentration for the first (C_{m1}) and the second (C_{m2}) unfolding.

Table S2: Overview of the variables determined in this work. The values are mean of triplicates with standard deviations, except for the C_m values where the error is the fit error in CDpal. The values are rounded to the larger number. The protein concentration is 1 mg/mL in all assays except ICD where the final protein concentration in the samples was 0.1 mg/mL.

| Protein | ICD | | | | ReFOLD | | MSF | nanoDSF | | DLS | % aggregates by SEC before and after storage | | | | | |
|-------------|--------------|--------------|--------------|--------------|-------------|-------------|---------------|---------------|---------------|---------------|--|-----------|------------|-----------|------------|-------------|
| | urea | | GuHCl | | RMY | RMY | T_{nr} (°C) | T_{on} (°C) | T_{m1} (°C) | T_{m2} (°C) | T_{agg} (°C) | T_0 | 3 m | 3 m | 12 m | 3 m @ 40 °C |
| | C_{m1} (M) | C_{m2} (M) | C_{m1} (M) | C_{m2} (M) | urea | GuHCl | | | | | | | @ 40 °C | @ 25 °C | @ 4 °C | minus T_0 |
| PPI01 | 4.6 ± 0.1 | 6.8 ± 0.2 | 1.5 ± 0.1 | 2.7 ± 0.2 | 0.12 ± 0.01 | 0.18 ± 0.01 | 60.3 ± 0.1 | 55.2 ± 0.2 | 61.4 ± 0.1 | - | 52.7 ± 0.7 | 0.4 ± 0.1 | 1.2 ± 0.1 | 0.4 ± 0.1 | 0.3 ± 0.1 | 0.8 ± 0.1 |
| PPI02 | 5.6 ± 0.1 | > 9.0 | 1.5 ± 0.1 | 3.6 ± 0.1 | 0.62 ± 0.01 | 0.13 ± 0.01 | 80.7 ± 0.1 | 62.6 ± 0.2 | 67.9 ± 0.1 | - | 73.0 ± 3.5 | 0.9 ± 0.1 | 1.0 ± 0.1 | 0.7 ± 0.1 | 0.8 ± 0.1 | 0.1 ± 0.1 |
| PPI03 | 6.9 ± 0.5 | 8.0 ± 0.2 | 1.8 ± 0.1 | 2.7 ± 0.1 | 0.55 ± 0.01 | 0.44 ± 0.01 | 76.7 ± 0.1 | 64.5 ± 0.4 | 68.8 ± 0.1 | 77.6 ± 0.1 | 68.7 ± 2.1 | 0.2 ± 0.1 | 0.5 ± 0.1 | 0.1 ± 0.1 | 0.1 ± 0.1 | 0.3 ± 0.1 |
| PPI04 | 4.4 ± 0.1 | 6.5 ± 0.1 | 1.3 ± 0.3 | 1.7 ± 0.2 | 0.28 ± 0.01 | 0.15 ± 0.01 | 63.8 ± 0.1 | 57.4 ± 0.7 | 64.7 ± 0.1 | - | 56.5 ± 0.9 | 2.4 ± 0.1 | 10.7 ± 0.2 | 2.3 ± 0.1 | 2.4 ± 0.1 | 8.3 ± 0.2 |
| PPI08 | 4.5 ± 0.5 | 6.6 ± 0.1 | 1.7 ± 0.2 | 2.1 ± 0.3 | 0.43 ± 0.03 | 0.14 ± 0.01 | 68.0 ± 0.2 | 70.9 ± 0.6 | 75.5 ± 0.1 | - | 56.0 ± 0.8 | 0.2 ± 0.1 | 3.2 ± 0.2 | 0.2 ± 0.1 | 0.3 ± 0.1 | 3.0 ± 0.2 |
| PPI10 | 5.9 ± 0.1 | > 9.0 | 1.7 ± 0.1 | 3.5 ± 0.1 | 0.61 ± 0.01 | 0.37 ± 0.01 | 78.7 ± 0.1 | 62.8 ± 0.7 | 68.4 ± 0.1 | 82.3 ± 0.1 | 73.5 ± 1.4 | 3.2 ± 0.1 | 2.3 ± 0.1 | 1.9 ± 0.1 | 1.9 ± 0.1 | -0.9 ± 0.1x |
| PPI11 | 4.4 ± 0.1 | 5.2 ± 0.8 | 0.9 ± 0.2 | 1.5 ± 0.1 | 0.25 ± 0.01 | 0.16 ± 0.01 | 62.1 ± 0.1 | 50.7 ± 0.3 | 56.5 ± 0.1 | 68.4 ± 0.1 | 56.6 ± 0.8 | 8.5 ± 0.2 | 9.0 ± 0.4 | 9.8 ± 0.2 | 10.2 ± 0.1 | 0.5 ± 0.4 |
| PPI13 | 4.8 ± 0.1 | > 9.0 | 1.4 ± 0.3 | 3.1 ± 0.1 | 0.52 ± 0.01 | 0.35 ± 0.01 | 80.5 ± 0.1 | 57.8 ± 0.8 | 62.3 ± 0.1 | 83.4 ± 0.1 | 73.8 ± 0.7 | 1.9 ± 0.1 | 1.8 ± 0.1 | 1.3 ± 0.1 | 1.4 ± 0.1 | -0.1 ± 0.1x |
| PPI17 | 6.6 ± 1.3 | 6.6 ± 0.2 | 1.7 ± 0.4 | 2.1 ± 0.2 | 0.38 ± 0.01 | 0.11 ± 0.01 | 72.7 ± 0.1 | 63.9 ± 0.7 | 69.4 ± 0.1 | 78.3 ± 0.1 | 60.9 ± 1.1 | 3.0 ± 0.1 | 3.0 ± 0.1 | 2.5 ± 0.3 | 2.8 ± 0.1 | 0.0 ± 0.1 |
| trastuzumab | 5.9 ± 0.8 | 7.5 ± 0.1 | 1.4 ± 0.1 | 2.5 ± 0.1 | 0.54 ± 0.01 | 0.43 ± 0.01 | 78.8 ± 0.1 | 65.0 ± 0.3 | 68.9 ± 0.2 | 80.4 ± 0.3 | 73.2 ± 1.3 | 0.4 ± 0.1 | 0.7 ± 0.1 | 0.3 ± 0.1 | 0.3 ± 0.1 | 0.3 ± 0.1 |
| adalimumab | 6.3 ± 0.6 | 6.5 ± 0.1 | 1.5 ± 0.8 | 2.3 ± 0.2 | 0.47 ± 0.01 | 0.22 ± 0.01 | 72.0 ± 0.1 | 58.7 ± 0.6 | 70.1 ± 0.1 | 82.0 ± 0.1 | 66.5 ± 1.0 | 1.0 ± 0.1 | 1.3 ± 0.1 | 0.8 ± 0.1 | 0.8 ± 0.1 | 0.3 ± 0.1 |
| bevacizumab | 5.0 ± 0.7 | 6.2 ± 0.1 | 1.9 ± 0.1 | 2.3 ± 0.9 | 0.15 ± 0.01 | 0.02 ± 0.01 | 70.5 ± 0.1 | 61.4 ± 0.7 | 69.8 ± 0.1 | - | 63.9 ± 0.9 | 1.2 ± 0.1 | 3.9 ± 0.1 | 1.3 ± 0.1 | 1.3 ± 0.1 | 2.7 ± 0.1 |
| rituximab | 5.9 ± 1.3 | 6.3 ± 0.2 | 1.8 ± 0.1 | 2.9 ± 0.3 | 0.42 ± 0.01 | 0.04 ± 0.01 | 72.1 ± 0.1 | 62.5 ± 0.1 | 67.6 ± 0.1 | - | 65.5 ± 2.1 | 1.1 ± 0.1 | 1.7 ± 0.1 | 1.0 ± 0.1 | 1.0 ± 0.1 | 0.6 ± 0.1 |

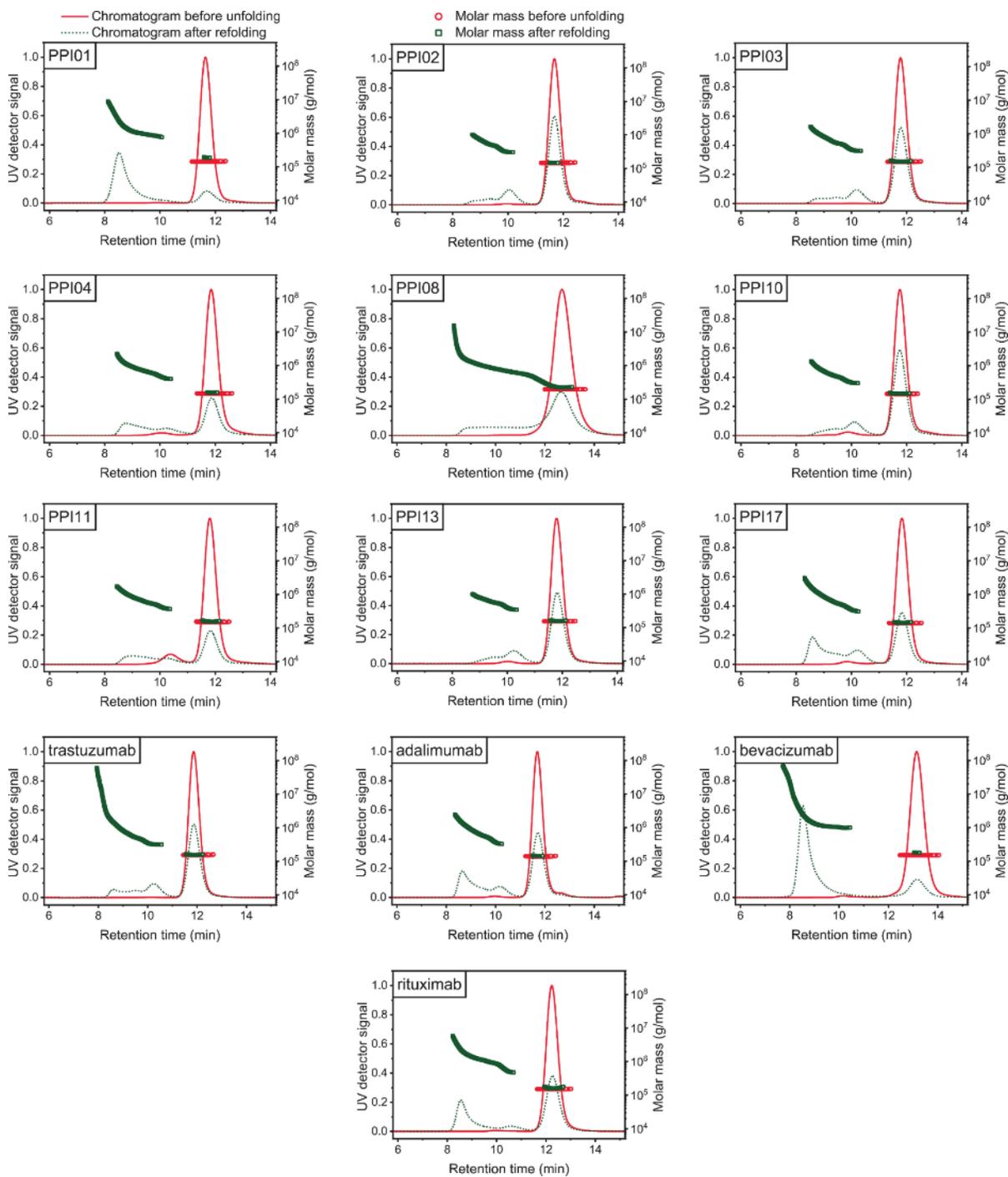


Figure S7: Exemplary data from the analysis of the proteins with the ReFOLD assay and SEC-MALS. Chromatograms of the proteins before unfolding (red lines) or after refolding from 10 M urea (green dots).

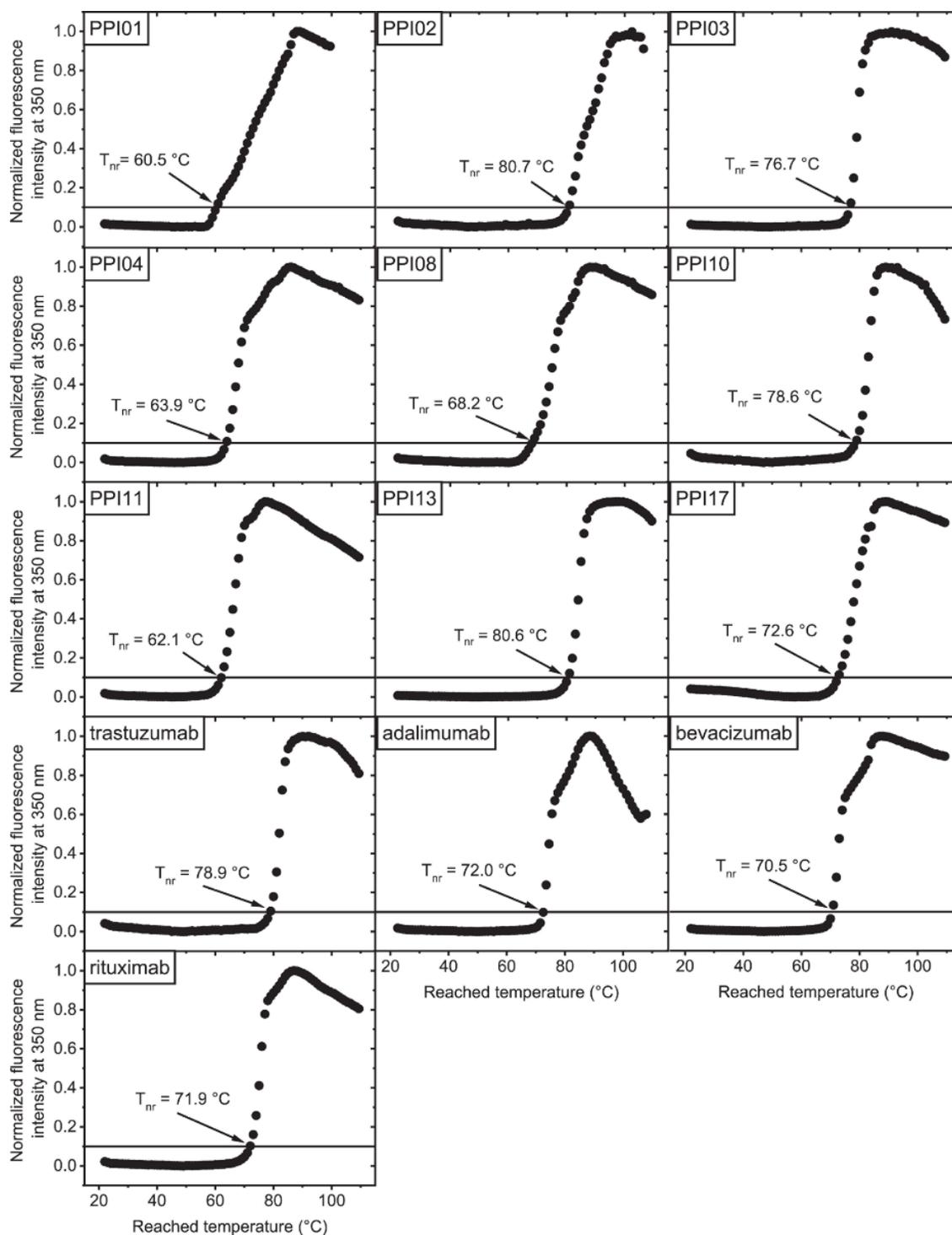


Figure S8: Non-reversibility curves of the antibodies from modulated scanning fluorimetry. Exemplary curves with the normalized fluorescence intensity signal at 350 nm and 10% threshold from which the temperature of non-reversibility onset (T_{nr}) was determined. The T_{nr} mean values from the triplicates can be found in Table S2.

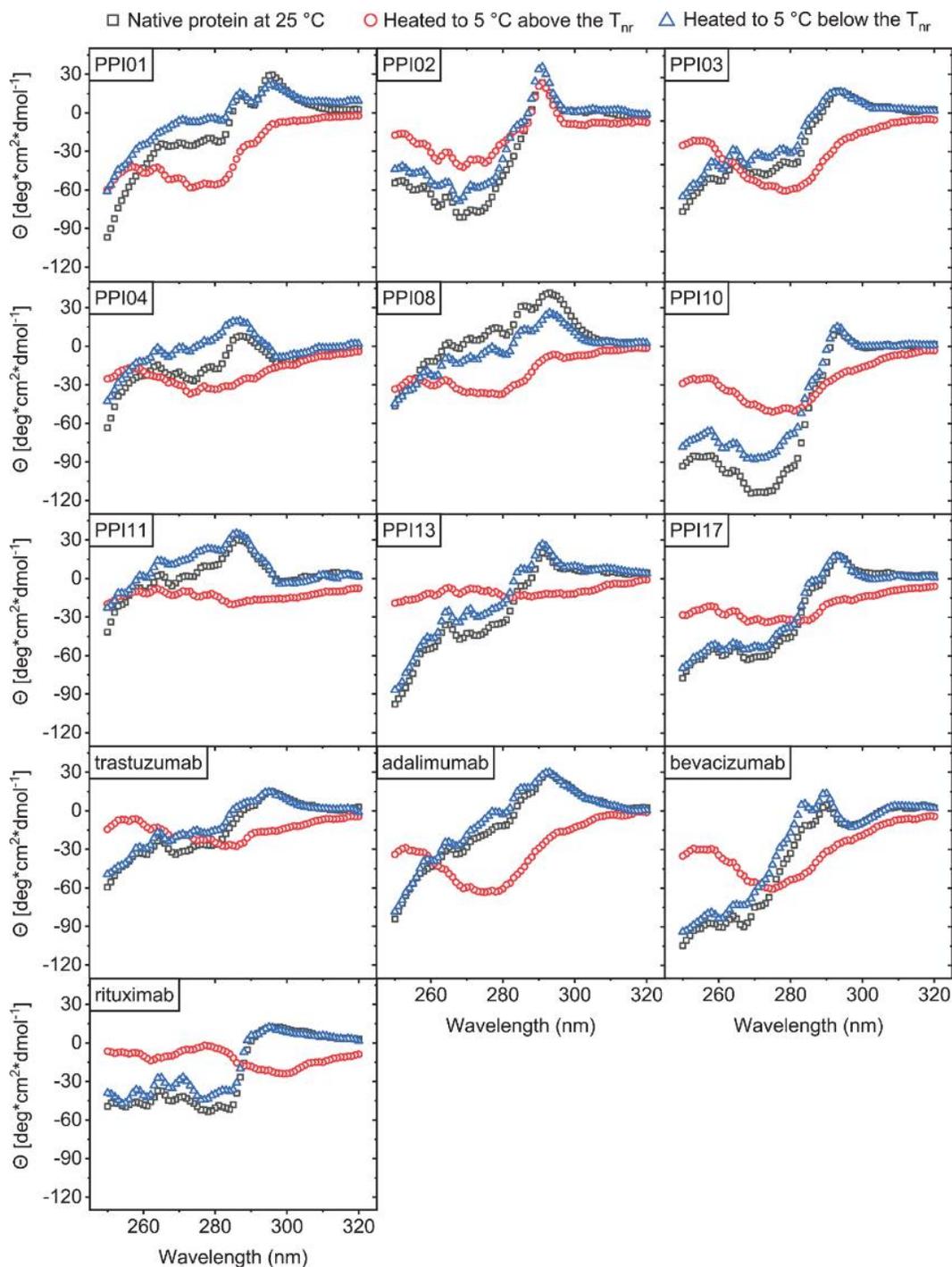


Figure S9: Impact on tertiary antibody structure after exposure to temperatures below and above the non-reversibility onset temperature T_{nr} from MSF. Near-UV CD spectra of the proteins at room temperature (black squares) or after heating for 15 minutes to 5 °C above the T_{nr} (red circles) or to 5 °C below the T_{nr} (blue triangles). The proteins were cooled back down to 25 °C before the measurement. The spectra show that heating to temperatures above the T_{nr} causes severe structural changes. A Jasco J-810 spectropolarimeter was used for the measurements. The protein concentration was 1 mg/mL.

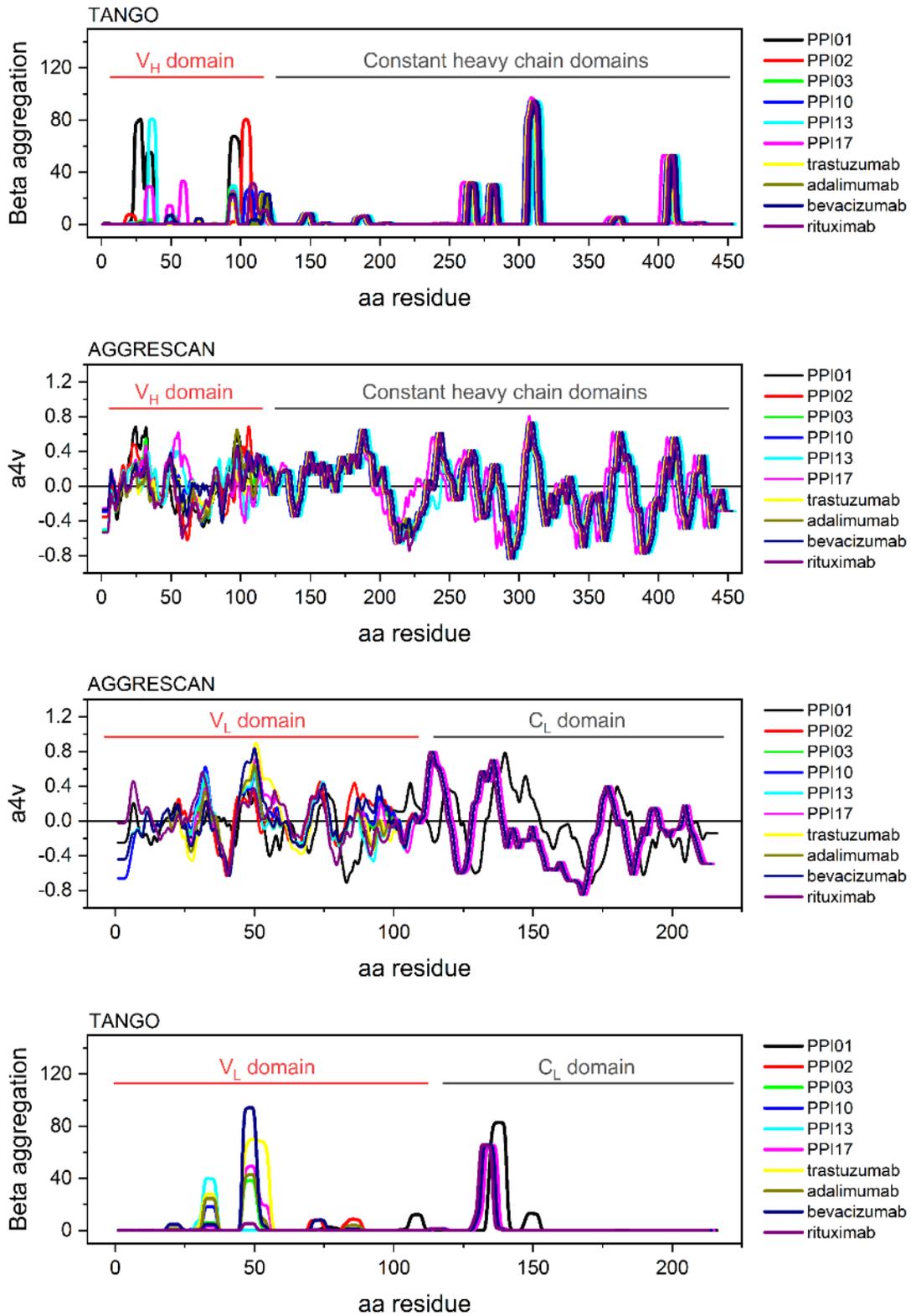


Figure S10: TANGO and AGGRESCAN analysis on the primary structures of the antibodies.

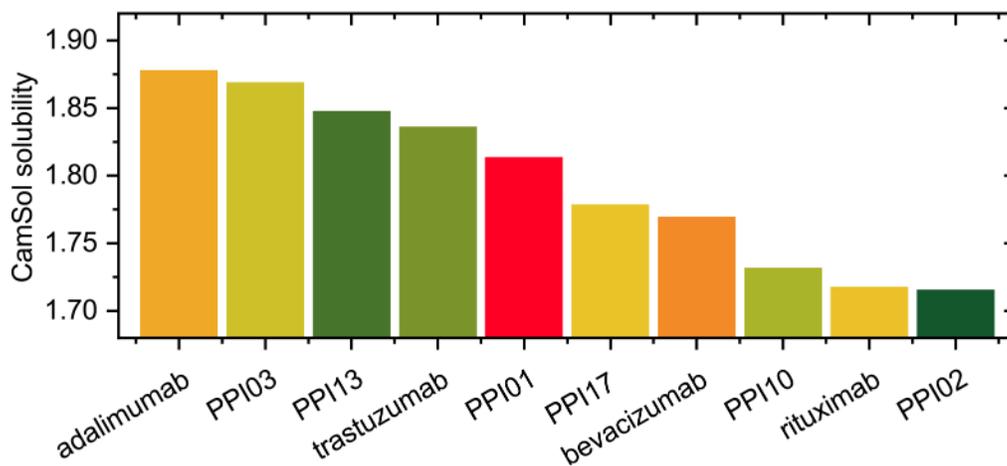


Figure S11: Antibody ranking based on the CamSol analysis on the native Fv regions. The color code is like in Figure 8c, Figure 9c, Figure 11 and Figure 12f.

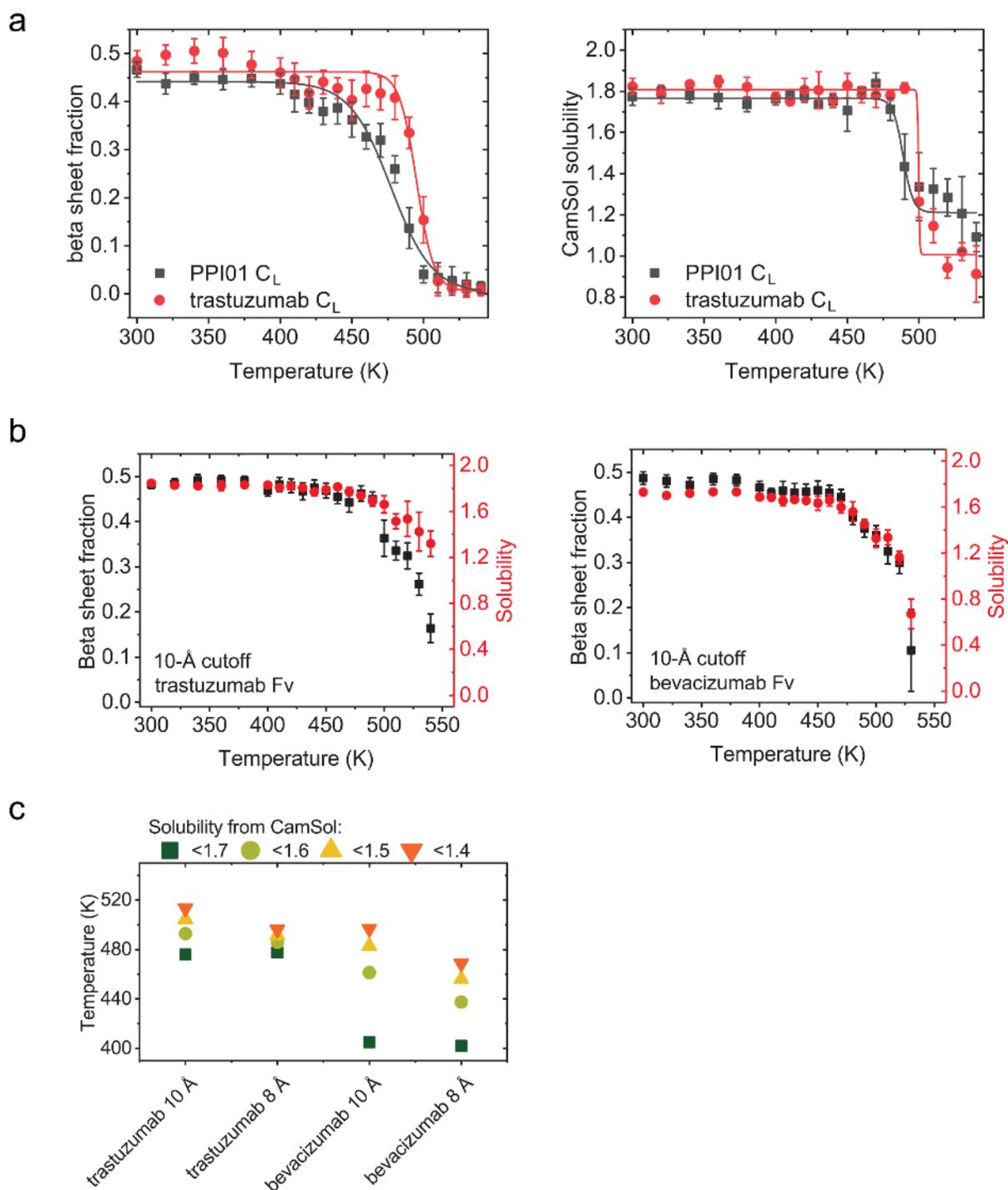


Figure S12: Additional characterization with MD and CamSol. (a) Analysis on the C_L domains from PPI01 (λ) and trastuzumab (κ). The lines are Boltzmann fits to the data. (b) MD simulations using 10 Å cut-off. (c) A comparison between solubility drop temperatures when the MD simulations are performed with 8 or 10 Å cut-off.

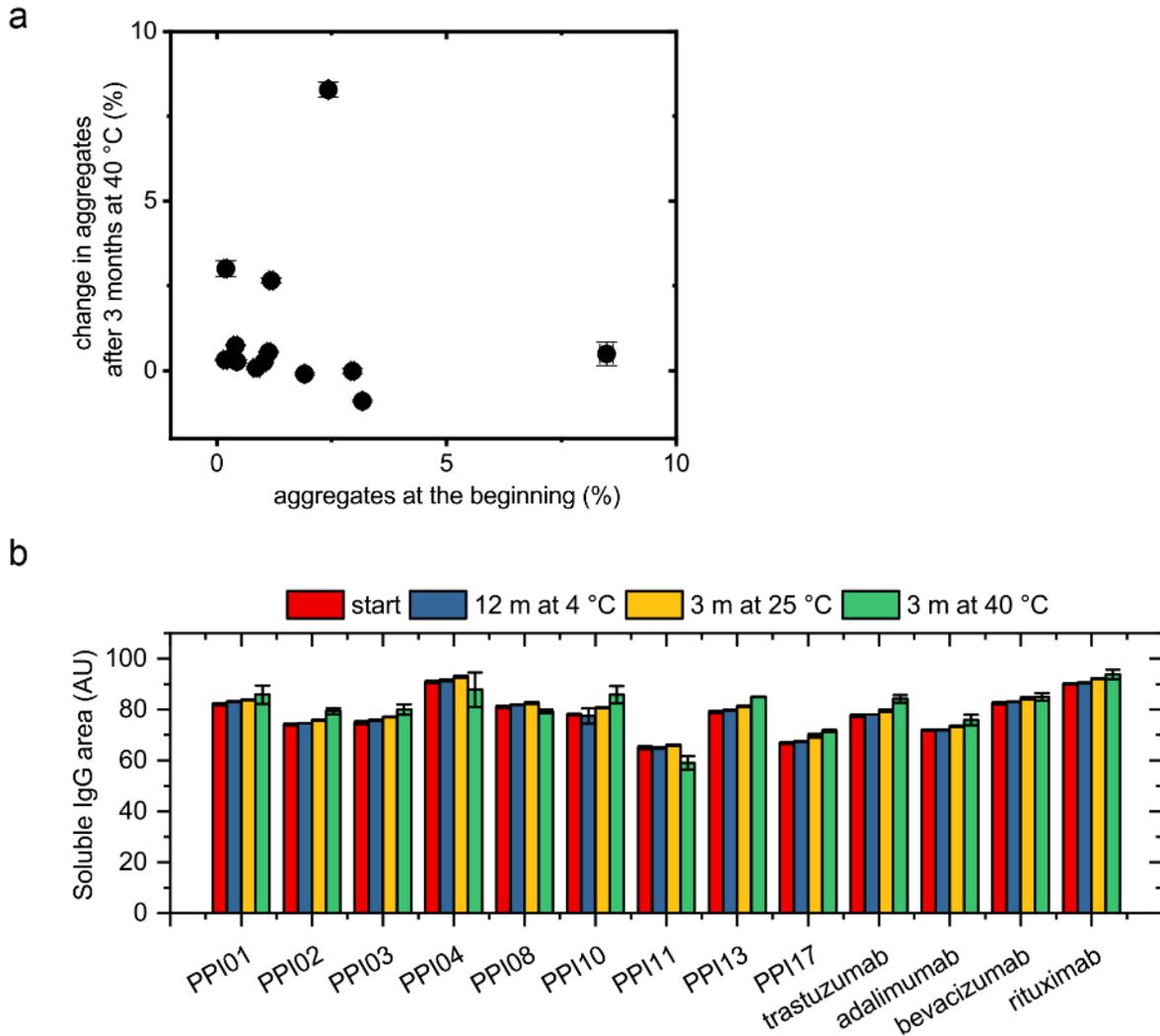


Figure S13: Additional storage stability data. (a) The amount of aggregates at the beginning of the stability study does not correlate with the aggregation during storage at 40 °C. (b) The area of soluble antibody detected in the SEC chromatograms indicates that there was no precipitation during storage.

3.6 References

1. Elvin, J. G.; Couston, R. G.; van der Walle, C. F. Therapeutic Antibodies: Market Considerations, *Disease Targets and Bioprocessing. Int. J. Pharm.* **440** (1), 83–98 (2013).
2. Rabia, L. A.; Desai, A. A.; Jhajj, H. S.; Tessier, P. M. Understanding and Overcoming Trade-Offs between Antibody Affinity, Specificity, Stability and Solubility. *Biochem. Eng. J.* **137**, 365–374 (2018).
3. Bailly, M.; Mieczkowski, C.; Juan, V.; Metwally, E.; Tomazela, D.; Baker, J.; Uchida, M.; Kofman, E.; Raoufi, F.; Motlagh, S.; et al. Predicting Antibody Developability Profiles Through Early Stage Discovery Screening. *MAbs* **12** (1), 1743053 (2020).
4. Zurdo, J. Developability Assessment as an Early De-Risking Tool for Biopharmaceutical Development. *Pharm. Bioprocess.* **1** (1), 29–50 (2013).
5. Xu, Y.; Wang, D.; Mason, B.; Rossomando, T.; Li, N.; Liu, D.; Cheung, J. K.; Xu, W.; Raghava, S.; Katiyar, A.; et al. Structure, Heterogeneity and Developability Assessment of Therapeutic Antibodies. *MAbs* **11** (2), 239–264 (2019).
6. Willis, L. F.; Kumar, A.; Jain, T.; Caffry, I.; Xu, Y.; Radford, S. E.; Kapur, N.; Vásquez, M.; Brockwell, D. J. The Uniqueness of Flow in Probing the Aggregation Behavior of Clinically Relevant Antibodies. *Eng. Reports* (2020).
7. Jain, T.; Sun, T.; Durand, S.; Hall, A.; Rewa, N.; Nett, J. H.; Sharkey, B. Biophysical Properties of the Clinical-Stage Antibody Landscape. *Proc. Natl. Acad. Sci. U. S. A.* **114** (5), 944–949 (2017).
8. Gentiluomo, L.; Svilenov, H. L.; Augustijn, D.; El Bialy, I.; Greco, M. L.; Kulakova, A.; Indrakumar, S.; Mahapatra, S.; Morales, M. M.; Pohl, C.; et al. Advancing Therapeutic Protein Discovery and Development through Comprehensive Computational and Biophysical Characterization. *Mol. Pharm.* **17** (2), 426–440 (2020).
9. Wolf Pérez, A.-M. M.; Sormanni, P.; Andersen, J. S.; Sakhnini, L. I.; Rodriguez-Leon, I.; Bjelke, J. R.; Gajhede, A. J.; De Maria, L.; Otzen, D. E.; Vendruscolo, M.; et al. In Vitro and in Silico Assessment of the Developability of a Designed Monoclonal Antibody Library. *MAbs* **11** (2), 388–400 (2019).
10. Shan, L.; Mody, N.; Sormanni, P.; Rosenthal, K. L.; Damschroder, M. M.; Esfandiary, R. Developability Assessment of Engineered Monoclonal Antibody Variants with a Complex Self-Association Behavior Using Complementary Analytical and in Silico Tools. *Mol. Pharm.* **15** (12), 5697–5710 (2018).
11. Jarasch, A.; Koll, H.; Regula, J. T.; Bader, M.; Papadimitriou, A.; Kettenberger, H. Developability Assessment during the Selection of Novel Therapeutic Antibodies. *J. Pharm. Sci.* **104** (6), 1885–1898 (2015).
12. Roberts, C. J. Non-Native Protein Aggregation Kinetics. *Biotechnol. Bioeng.* **98** (5), 927–938 (2007).

13. Famm, K.; Hansen, L.; Christ, D.; Winter, G. Thermodynamically Stable Aggregation-Resistant Antibody Domains through Directed Evolution. *J. Mol. Biol.* **376** (4), 926–931 (2008).
14. Jespers, L.; Schon, O.; Famm, K.; Winter, G. Aggregation-Resistant Domain Antibodies Selected on Phage by Heat Denaturation. *Nat. Biotechnol.* **22** (9), 1161–1165 (2004).
15. Svilenov, H. L.; Menzen, T.; Richter, K.; Winter, G. Modulated Scanning Fluorimetry Can Quickly Assess Thermal Protein Unfolding Reversibility in Microvolume Samples. *Mol. Pharm.* **17** (7), 2638–2647 (2020).
16. Freire, E.; Schön, A.; Hutchins, B. M.; Brown, R. K. Chemical Denaturation as a Tool in the Formulation Optimization of Biologics. *Drug Discov. Today* **18** (19–20), 1007–1013 (2013).
17. Kiefhaber, T.; Rudolph, R.; Kohler, H.-H. H.; Buchner, J. Protein Aggregation in Vitro and in Vivo: A Quantitative Model of the Kinetic Competition between Folding and Aggregation. *Bio/Technology* **9** (9), 825–829 (1991).
18. 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** (12), 3007–3013 (2018).
19. Rowe, J. B.; Flynn, R. P.; Wooten, H. R.; Noufer, H. A.; Cancel, R. A.; Zhang, J.; Subramony, J. A.; Pechenov, S.; Wang, Y. Submicron Aggregation of Chemically Denatured Monoclonal Antibody. *Mol. Pharm.* **15** (10), 4710–4721 (2018).
20. 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).
21. Svilenov, H. L.; Kulakova, A.; Zalar, M.; Golovanov, A. P.; Harris, P.; Winter, G. Orthogonal Techniques to Study the Effect of PH, Sucrose and Arginine Salts on Monoclonal Antibody Physical Stability and Aggregation during Long-Term Storage. *J. Pharm. Sci.* **109** (1), 584–594 (2019).
22. Svilenov, H. L.; Winter, G. Formulations That Suppress Aggregation During Long-Term Storage of a Bispecific Antibody Are Characterized by High Refoldability and Colloidal Stability. *J. Pharm. Sci.* **109** (6), 2048–2058 (2020).
23. 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).
24. Niklasson, M.; Andresen, C.; Helander, S.; Roth, M. G. L.; Zimdahl Kahlin, A.; Lindqvist Appell, M.; Mårtensson, L.-G.; Lundström, P. Robust and Convenient Analysis of Protein Thermal and Chemical Stability. *Protein Sci.* **24** (12), 2055–2062 (2015).

25. Dunbar, J.; Krawczyk, K.; Leem, J.; Marks, C.; Nowak, J.; Regep, C.; Georges, G.; Kelm, S.; Popovic, B.; Deane, C. M. SAbPred: A Structure-Based Antibody Prediction Server. *Nucleic Acids Res.* **44** (W1), W474–W478 (2016).
26. D.A. Case, K. Belfon, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, G. Giambasu, M.K. Gilson, H. Gohlke, A.W. Goetz, R. Harris, S. Izadi, S.A. Izmailov, K. Kasavajhala, A. Kovalenko, R. Krasny, T, D. M. Y. and P. A. K. AMBER 2019. Univ. California, San Fr. 2019.
27. Gordon, J. C.; Myers, J. B.; Folta, T.; Shoja, V.; Heath, L. S.; Onufriev, A. H⁺⁺: A Server for Estimating PKas and Adding Missing Hydrogens to Macromolecules. *Nucleic Acids Res.* **33** (SUPPL. 2), W368–W371 (2005).
28. Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **14** (October 1995), 33–38 (1996).
29. Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **22** (12), 2577–2637 (1983).
30. Sormanni, P.; Aprile, F. A.; Vendruscolo, M. The CamSol Method of Rational Design of Protein Mutants with Enhanced Solubility. *J. Mol. Biol.* **427** (2), 478–490 (2015).
31. Rizzo, J. M.; Shi, S.; Li, Y.; Semple, A.; Esposito, J. J.; Yu, S.; Richardson, D.; Antochshuk, V.; Shameem, M. Application of a High-Throughput Relative Chemical Stability Assay to Screen Therapeutic Protein Formulations by Assessment of Conformational Stability and Correlation to Aggregation Propensity. *J. Pharm. Sci.* **104** (5), 1632–1640 (2015).
32. Holloway, L.; Roche, A.; Marzouk, S.; Uddin, S.; Ke, P.; Ekizoglou, S.; Curtis, R. Determination of Protein-Protein Interactions at High Co-Solvent Concentrations Using Static and Dynamic Light Scattering. *J. Pharm. Sci.* **109** (9), 2699–2709 (2020).
33. Gentiluomo, L.; Roessner, D.; Streicher, W.; Mahapatra, S.; Harris, P.; Frieß, W. Characterization of Native Reversible Self-Association of a Monoclonal Antibody Mediated by Fab-Fab Interaction. *J. Pharm. Sci.* **109** (1), 443–451 (2020).
34. Melien, R.; Garidel, P.; Hinderberger, D.; Blech, M. Thermodynamic Unfolding and Aggregation Fingerprints of Monoclonal Antibodies Using Thermal Profiling. *Pharm. Res.* **37** (4), 1–13 (2020).
35. 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** (2019).

36. Giannos, S. A.; Kraft, E. R.; Zhao, Z.-Y.; Merkley, K. H.; Cai, J. Formulation Stabilization and Disaggregation of Bevacizumab, Ranibizumab and Aflibercept in Dilute Solutions. *Pharm. Res.* **35** (4), 78 (2018).
37. 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** (10), 1302–1306 (2004).
38. Conchillo-Solé, O.; de Groot, N. S.; Avilés, F. X.; Vendrell, J.; Daura, X.; Ventura, S. AGGRESCAN: A Server for the Prediction and Evaluation of “Hot Spots” of Aggregation in Polypeptides. *BMC Bioinformatics* **8**, 65 (2007).
39. van der Kant, R.; Bauer, J.; Karow-Zwick, A. R.; Kube, S.; Garidel, P.; Blech, M.; Rousseau, F.; Schymkowitz, J. Adaption of Human Antibody λ and κ Light Chain Architectures to CDR Repertoires. *Protein Eng. Des. Sel.*, **32** (3), 109–127 (2019).
40. Wang, T.; Wade, R. C. On the Use of Elevated Temperature in Simulations to Study Protein Unfolding Mechanisms. *J. Chem. Theory Comput.* **3** (4), 1476–1483 (2007).
41. Buck, P. M.; Kumar, S.; Singh, S. K. Insights into the Potential Aggregation Liabilities of the B12 Fab Fragment via Elevated Temperature Molecular Dynamics. *Protein Eng. Des. Sel.* **26** (3), 195–206 (2013).
42. Piana, S.; Lindorff-Larsen, K.; Dirks, R. M.; Salmon, J. K.; Dror, R. O.; Shaw, D. E. Evaluating the Effects of Cutoffs and Treatment of Long-Range Electrostatics in Protein Folding Simulations. *PLoS One* **7** (6), e39918 (2012).
43. Arbabi-Ghahroudi, M.; To, R.; Gaudette, N.; Hiramata, T.; Ding, W.; MacKenzie, R.; Tanha, J. Aggregation-Resistant VHs Selected by in Vitro Evolution Tend to Have Disulfide-Bonded Loops and Acidic Isoelectric Points. *Protein Eng. Des. Sel.* **22** (2), 59–66 (2008).
44. Kim, D. Y.; To, R.; Kandalaf, H.; Ding, W.; Van Faassen, H.; Luo, Y.; Schrag, J. D.; St-Amant, N.; Hefford, M.; Hiramata, T.; et al. Antibody Light Chain Variable Domains and Their Biophysically Improved Versions for Human Immunotherapy. *MAbs* **6** (1), 219–235 (2014).
45. Pérez, J. M. J.; Renisio, J. G.; Prompers, J. J.; Van Platerink, C. J.; Cambillau, C.; Darbon, H.; Frenken, L. G. J. Thermal Unfolding of a Llama Antibody Fragment: A Two-State Reversible Process. *Biochemistry* **40** (1), 74–83 (2001).
46. Austerberry, J. I.; Thistlethwaite, A.; Fisher, K.; Golovanov, A. P.; Pluen, A.; Esfandiary, R.; Van Der Walle, C. F.; Warwicker, J.; Derrick, J. P.; Curtis, R. Arginine to Lysine Mutations Increase the Aggregation Stability of a Single-Chain Variable Fragment through Unfolded-State Interactions. *Biochemistry* **58** (32), 3413–3421 (2019).

47. Sahin, E.; Weiss, W. F.; Kroetsch, A. M.; King, K. R.; Kessler, R. K.; Das, T. K.; Roberts, C. J. Aggregation and PH-Temperature Phase Behavior for Aggregates of an IgG2 Antibody. *J. Pharm. Sci.* **101** (5), 1678–1687 (2012).
48. Saad, S.; Cereghetti, G.; Feng, Y.; Picotti, P.; Peter, M.; Dechant, R. Reversible Protein Aggregation Is a Protective Mechanism to Ensure Cell Cycle Restart after Stress. *Nat. Cell Biol.* **19** (10), 1202–1213 (2017).
49. Zhang, Y.; Wu, L.; Gupta, P.; Desai, A. A.; Smith, M. D.; Rabia, L. A.; Ludwig, S. D.; Tessier, P. M. Physicochemical Rules for Identifying Monoclonal Antibodies with Drug-like Specificity. *Mol. Pharm.* **17** (7), 2555–2569 (2020).
50. Sawant, M. S.; Streu, C. N.; Wu, L.; Tessier, P. M. Toward Drug-Like Multispecific Antibodies by Design. *Int. J. Mol. Sci.* **21** (20), 7496 (2020).
51. Raybould, M. I. J. J.; Marks, C.; Krawczyk, K.; Taddese, B.; Nowak, J.; Lewis, A. P.; Bujotzek, A.; Shi, J.; Deane, C. M. Five Computational Developability Guidelines for Therapeutic Antibody Profiling. *Proc. Natl. Acad. Sci. U. S. A.* **116** (10), 4025–4030 (2019).
52. Starr, C. G.; Tessier, P. M. Selecting and Engineering Monoclonal Antibodies with Drug-like Specificity. *Curr. Opin. Biotechnol.* **60**, 119–127 (2019).
53. Wälchli, R.; Vermeire, P.-J.; Massant, J.; Arosio, P. Accelerated Aggregation Studies of Monoclonal Antibodies: Considerations for Storage Stability. *J. Pharm. Sci.* **109**.1, 595-602 (2020).
54. Roberts, C. J.; Das, T. K.; Sahin, E. Predicting Solution Aggregation Rates for Therapeutic Proteins: Approaches and Challenges. *Int. J. Pharm.* **418** (2), 318–333 (2011).
55. Brummitt, R. K.; Nesta, D. P.; Roberts, C. J. Predicting Accelerated Aggregation Rates for Monoclonal Antibody Formulations, and Challenges for Low-Temperature Predictions. *J. Pharm. Sci.* **100** (10), 4234–4243 (2011).
56. Wang, W.; Roberts, C. J. Non-Arrhenius Protein Aggregation. *AAPS J.* **15** (3), 840–851 (2013).
57. Avastin : EPAR - Scientific Discussion. Eur. Med. Agency (2006).
58. Andersen, C. B.; Manno, M.; Rischel, C.; Thóroúlfsson, M.; Martorana, V. Aggregation of a Multidomain Protein: A Coagulation Mechanism Governs Aggregation of a Model IgG1 Antibody under Weak Thermal Stress. *Protein Sci.* **19** (2), 279–290 (2010).
59. Ho, J. G. S.; Middelberg, A. P. J.; Ramage, P.; Kocher, H. P. The Likelihood of Aggregation during Protein Renaturation Can Be Assessed Using the Second Virial Coefficient. *Protein Sci.* **12** (4), 708–716 (2003).

60. Childers, M. C.; Daggett, V. Validating Molecular Dynamics Simulations against Experimental Observables in Light of Underlying Conformational Ensembles. *J. Phys. Chem. B* **122** (26), 6673–6689 (2018).

Chapter 4 - Investigation of the pH-dependent aggregation mechanisms of G-CSF using low resolution protein characterization techniques and advanced molecular dynamics simulations

This chapter is published as:

Ko, S. K.¹, Berner, C.², Kulakova, A.³, Schneider, M.⁴, Antes, I.⁴, Winter, G.², Harris, P.³ & Peters, G. H.¹ (2022). Investigation of the pH-dependent aggregation mechanisms of G-CSF using low resolution protein characterization techniques and advanced molecular dynamics simulations. *Computational and Structural Biotechnology Journal*, 20, 1439-1455.

¹Technical University of Denmark, Department of Chemistry, Kongens Lyngby, Denmark

²Ludwig Maximilian University of Munich, Department of Pharmacy Munich, Germany

³University of Copenhagen, Department of Chemistry, Copenhagen, Denmark

⁴Technical University of Munich, TUM School of Life Sciences, Freising, Germany

Author contributions:

S.K.K., C.B., G.W. and G.H.P. conceived the idea. S.K.K. performed the *in silico* work. C.B. performed the experimental work. A.K. and P.H. provided guidance with the analysis of the SAXS data. M.S. and I.A. provided guidance with the CG simulations. S.K.K. and C.B. wrote the manuscript. G.W. and P.H. and G.H.P. provided conceptual guidance and corrected the manuscript.

The published article can be accessed online via:

<https://doi.org/10.1016/j.csbj.2022.03.012>

4.1 Introduction

Proteins are widely applied as medicines due to their high specificity compared to small chemicals.^{1,2} However, protein drugs exhibit additional challenges when it comes to the development of formulations that can preserve their stability.^{3,4} Protein aggregation is a commonly encountered problem in the development of biopharmaceuticals that can affect the efficacy of the product and cause undesired immune reactions in patients.⁵ Both protein colloidal and conformational stability have been related to protein aggregation.⁵ The colloidal stability of a protein is related to weak net interactions between the protein molecules in solution, which can either be attractive or repulsive. The conformational stability is defined by the equilibrium between folded and unfolded states of a protein, and a slight deviation from these optimal conditions may shift the equilibrium towards unfolded protein species, which are often prone to form aggregates.^{6,7} Various factors including the protein's amino acid sequence and environmental factors such as pH, buffers, protein concentration, ionic strength, and storage conditions have an impact on protein aggregation. Therefore, the prevention of protein aggregation is a major challenge in the formulation development process in the biopharmaceutical industry. Due to the lack of a complete molecular understanding and predictability of protein aggregation, formulation development is still done in a trial and error approach.^{8,9}

In this study, we investigated the aggregation mechanism of the therapeutic model protein granulocyte-colony stimulating factor (G-CSF). Native G-CSF is a 19.6 kDa glycoprotein with 174 amino acid residues,¹⁰ which mediates the proliferation of granulocytes through receptor binding. Filgrastim, the non-glycosylated, recombinant form of G-CSF with an additional N-terminal methionine group is a licensed drug to treat neutropenia.¹¹ Filgrastim, hereafter referred to as G-CSF, is a hydrophobic cytokine with a molecular weight of 18.7 kDa as a result of the removal of the glycosylation.¹² The structure of G-CSF is characterized by a four-helix-bundle fold with two long loops connecting helices α A and α B as well as α C and α D. An additional short 3_{10} -helix is located within the loopAB and is perpendicular to the four anti-parallel α -helices (Figure 14).

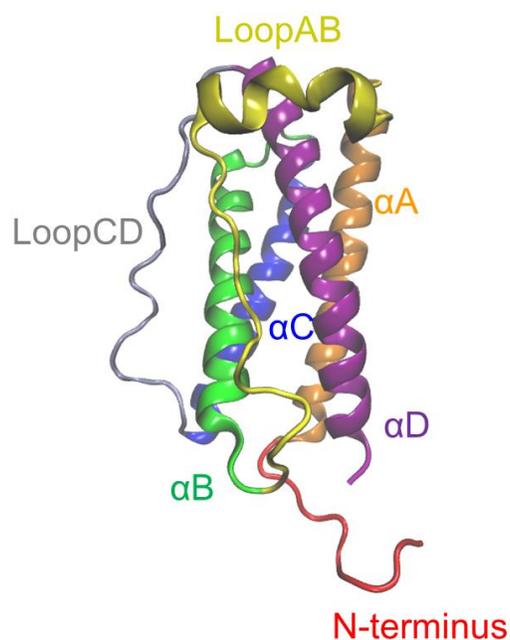


Figure 14: The structure of G-CSF obtained from Protein Data Bank (PDB) (PDB code: 1CD9¹³). MODELLER was used to generate the first five disordered residues.¹⁴ The secondary structure of G-CSF is shown with different color schemes: N-terminus (Met1-Pro11), α A (Gln12-Tyr40), loopAB (Lys41-Gln71), α B (Leu72-Leu93), α C (Leu100-Leu125), loopCD (Gly126-Ser143), α D (Ala144-Pro175).

The stability of G-CSF is highly pH-dependent with maximum stability at pH 4.0 and low stability and fast aggregation at physiological pH.¹⁵⁻¹⁷ With an isoelectric point of around 6.1,¹⁶ G-CSF is highly positively charged at acidic pH, resulting in electrostatic repulsion between the protein molecules. Furthermore, the addition of salt at pH 3.5 causes aggregation¹⁶ indicating a strong impact of electrostatic interactions on the aggregation of G-CSF. Nevertheless, the pH-dependent behavior of G-CSF is still discussed in the literature. Narhi et al. reported an increase of α -helical content of G-CSF at pH 4.0 compared to neutral pH using circular dichroism (CD) measurements.¹⁸ Another study used hydrogen deuterium exchange (HDX) mass spectrometry to compare the local changes of relative uptake difference between pH 4.0 and 7.0 and could not observe a change in α -helical content. Narhi et al. observed quenching of Trp residue(s) at pH 4.0 in fluorescence intensity measurements.¹⁸ Similarly, a Nuclear Magnetic Resonance (NMR) study performed by Aubin et al. showed that Trp configuration is pH-dependent.¹⁹

To provide a detailed molecular understanding of the pH-dependent aggregation mechanisms of G-CSF, we performed a multi-scale modeling approach using full atomic and coarse-grained (CG) molecular dynamics (MD) simulations. The aggregation mechanism was explored by CG simulations of G-CSF using the SIRAH force field (FF)^{20,21} developed to simulate proteins in explicit solvent conditions. WT4 models describe the CG water molecules, where one WT4 model consists of four beads that are connected in a tetrahedral form. The protein backbone in the SIRAH FF is defined by 3 beads representing nitrogen, alpha carbon, and oxygen atoms and thereby allowing for movement of the secondary structure since no constraint is applied to fix the protein backbone. Each side chain was modeled specifically based on a combination of physicochemical characteristics. The SIRAH FF is a relatively new force field that was recently used to study the process of seeding peptide aggregation.²² SIRAH was chosen as an alternative to MARTINI since it has been shown that the MARTINI FF overestimates PPI for membrane proteins.²³ The conformational stability of G-CSF at varying pH values was studied by carrying out full atomic MD simulations in the pH range of 4.0 to 7.5. We could observe that the conformational state of G-CSF is very similar at varying pH values in unbiased systems. To ensure that the system is not trapped in a local minimum, we carried out metadynamics simulations. We compared our *in silico* results with experimental data obtained from fluorescence intensity, CD spectroscopy, nanoDSF, and DLS measurements as well as modeling based on small-angle X-ray scattering (SAXS).

4.2 Methods

4.2.1 Conventional Molecular Dynamics Simulations

The structure of G-CSF is available from X-ray (1CD9¹³, 1RHG¹⁰, and 2D9Q²⁴) and NMR (1GNC²⁵) studies, of which 1CD9 has been widely used as the G-CSF model structure in various MD simulation studies.^{26–28} The initial structure of G-CSF for the conventional molecular dynamics simulation (cMD) study was prepared using PDB entry 1CD9 (solved at pH 7.5).¹³ The missing five residues were added using Modeller software 9.21.¹⁴ The PDB2PQR server was used to protonate the titratable residues at pH 4.0, 5.0, and 7.5.²⁹ The full atomic conventional MD (cMD) simulations were carried out using the AMBER software 20³⁰ and G-CSF was parametrized using the force field FF14sb.³¹ The protein was inserted into a cubic periodic boundary box, where the minimum distance between the protein and the edge of the box was set to 15 Å. The TIP4P Ewald water model was used to solvate the system.³² The system was neutralized by adding either

sodium or chloride ions. The initial structures were minimized using 10000 cycles. The first 5000 cycles were computed using the steepest descend algorithm. The remaining 5000 cycles were carried out using the conjugate gradient algorithm. The cut-off distance of the non-bonding interaction was set to 12 Å. The electrostatic long-range interactions were evaluated using the particle-mesh-Ewald (PME) method.³³ The SHAKE algorithm was applied to fix the bonds involving hydrogen.^{34,35} The system was heated to 300 K in the NVT ensemble (constant N = Number of atoms, V = Volume, T = Temperature) for 0.3 ns, using the Langevin thermostat³⁶ with a collision frequency of 5 ps⁻¹. The system was then subjected to a short equilibration run for 2.2 ns in the NPT (N= Number of atoms, P= Pressure, T=temperature) ensemble, while the pressure was kept at 1 bar using Monte Carlo barostat.³⁷ The final production run was carried out using the NPT ensemble for 400 ns and the last 200 ns were used for analyses.

4.2.2 Metadynamics

To ensure that G-CSF conformation is not trapped in the local minima during the simulations, we have carried out metadynamics simulations using AMBER software 20³⁰ and PLUMED 2³⁸. The initial structures for the metadynamics simulations were obtained from the final frame of the cMD simulations. All metadynamics simulations were carried out in the NVT ensemble for 400 ns using the Langevin thermostat with a collision frequency of 5 ps⁻¹.³⁶ The well-tempered metadynamics scheme was used to ensure a smooth convergence of the free energy landscape.³⁹ The collective variables (CVs) were chosen based on the experimental observations,^{18,19,40,41} and included the center of the mass distance (COM) between Trp and His residues to monitor the interactions between, Trp59-His157 (d1) and His80-Trp119 (d2), and the α -helical content (α) (Table 1).

Table 1: List of the metadynamics simulation conditions. The following CVs were investigated in the study: α -helical content (α), the COM distance between Trp59-His157 (d1), and the COM distance between His80-Trp119 (d2). The initial height and width of the Gaussian hills are also provided. Biasfactor is defined to perform the simulations in a well-tempered manner.

| Simulation Label | Input pH | CVs | Height [kJ/mol] | Width | Biasfactor | Deposition Rate [hill/ps] |
|------------------|----------|-------------------|-----------------|----------|------------|---------------------------|
| 1 | pH 4.0 | α , d1, d2 | 1 | 0.5, | 15 | 1 |
| 2 | pH 5.0 | | | 0.05 nm, | | |
| 3 | pH 7.5 | | | 0.05 nm | | |

4.2.3 CG Simulations

The CG simulations were carried out using the Gromacs software 2018⁴² with the SIRAH 2 force field.^{20,21,43} The CG model of G-CSF at pH 4.0, 5.0, and 7.5 was obtained by coarse-graining the full atomic G-CSF models that were obtained from the PDB2PQR²⁹ web server using the SIRAH toolbox.⁴³ For each simulation, 8 G-CSF monomers were added to the system. The initial G-CSF monomer was translated and duplicated along the x-, y-, and z-axes where the center of the mass distance between replicates was set to 7.5 nm. An alternative approach could have been to sample the initial structures from a population-density of structures determined from single monomer metadynamics simulations. However, without any input from experimental results, this will give rise to a large number of combinations, and we decided therefore to use the final frame of cMD simulations.

The distance between solute and box was set to 0.75 nm resulting in a concentration of ~ 30 mg/mL. Note that a too small simulation box will cause an immediate aggregation of the proteins while too large box sizes will increase the simulation time. The optimal protein-protein and protein-box distances were chosen empirically to reduce the computational burden for sampling the aggregation. The system was solvated by adding SIRAH based WT4 molecules.⁴⁴ After the solvation, the system was neutralized by adding either sodium or chloride ions. In addition to the pH study, the effect of salt was monitored by adding 150 mM of NaCl (in CG mode) to the systems at different pH conditions. The initial minimization was conducted using the steepest descent

algorithm, followed by the conjugate gradient algorithm. The maximum number of each minimization scheme was set to 50,000. The heating was performed for 2 ns where the system was coupled to the Berendsen thermostat and barostat.⁴⁵ After heating, the system was equilibrated for 500 ns using a time step of 10 fs. To accurately sample the NPT ensemble, the system was coupled to the stochastic velocity rescaling thermostat⁴⁶ and the Parrinello-Rahman barostat.⁴⁷ The production run was performed for 3 μ s. For each condition, 5 replicate simulations were carried out which corresponds to 15 μ s per condition.

4.2.4 Materials

The bulk G-CSF solution contained 4.0 g/L protein and was provided from Wacker Chemie, Germany. The protein concentration was measured spectrophotometrically using a NanoDrop 2000 (Thermo Fisher Scientific, Wilmington, USA) and an extinction coefficient at 280 nm of 0.86 (mg/mL)⁻¹cm⁻¹. All chemicals were of molecular biology or multicompendial grade and were purchased either from Sigma or Thermo Fisher Scientific (Germany). All solutions were prepared with ultrapure water from a Sartorius arium[®] pro system (Sartorius Corporate Administration GmbH, Göttingen, Germany). All buffers used had a concentration of 10 mM and the pH after preparation was \pm 0.1 of the target value.

4.2.5 Sample Dialysis and Preparation

The buffer was exchanged by extensive dialysis to the respective buffer at the given pH (10 mM sodium acetate at pH 4.0 and pH 5.0, 10 mM potassium phosphate at pH 7.5) for 24 hours at 2-8 °C using a Spectra/Por[®] dialysis membrane (cutoff 6-8 kDa, Spectrum Laboratories, Rancho Dominguez, CA, USA) or a Slide-A-Lyzer[™] MINI Dialysis Device (cutoff 3.5 kDa, Thermo Fisher Scientific, Germany). The samples were collected in microcentrifuge tubes and centrifuged at 10,000 x g for 10 minutes and subsequently filtered with 0.02 μ m Anotop[®] membrane filters (Whatman, FP 30/0.2 CA-S, GE Healthcare, Buckinghamshire, UK). Stock solutions of sodium chloride were prepared in the respective buffer and spiked into the dialysed protein stock to prepare samples containing 100 mM of sodium chloride. For measurements that required higher protein concentrations, the protein solutions were upconcentrated using Vivaspin 20 5 MWCO PES centrifugal concentrators (Sartorius Lab Instruments, Goettingen, Germany). The concentration was measured again, and the solutions were sterile filtered with 0.02 μ m Anotop[®] membrane filters.

4.2.6 Intrinsic Fluorescence Spectroscopy

Fluorescence emission measurements of the samples with a protein concentration of 0.5 g/L were performed using a Jasco FP-6500 Fluorescence Spectrophotometer. Emission spectra were recorded from 300 to 450 nm with an excitation wavelength of 280 nm, steps of 0.01 nm, and a scan speed of 100 nm min⁻¹. A 3 nm slit width was used both in excitation and emission monochromators. Buffer spectra were subtracted from the sample spectra.

4.2.7 Circular Dichroism (CD) Spectroscopy

Near- and far-UV circular dichroic spectra were collected at 25 °C with a Jasco J-810 spectropolarimeter (JASCO Deutschland GmbH, Pfungstadt, Germany). All samples contained 1 g/L of protein. Quartz cuvettes (Hellma GmbH, Muellheim, Germany) with 10 mm and 0.1 mm wavelength path were used for the measurements, respectively. Five accumulations of each sample were taken at a speed of 20 nm/min. The spectrum of the respective buffer was subtracted for each sample and smoothing of the spectra was performed using the Savitzky-Golay algorithm with 9 smoothing points. The mean residue ellipticity (MRE) of the protein at each wavelength was calculated as described elsewhere.⁴⁸

4.2.8 Differential Scanning Fluorimetry (nanoDSF)

nanoDSF was used to study the thermal unfolding and aggregation of G-CSF as a function of pH and ionic strength. Samples with 1 g/L of protein were filled in standard nanoDSF™ grade capillaries, and the capillaries were sealed. A temperature ramp of 1 °C/min from 20 to 100 °C was applied with the Prometheus NT.48 (NanoTemper Technologies, Munich, Germany) system that measures the intrinsic protein fluorescence intensity at 330 and 350 nm after excitation at 280 nm. Simultaneously, the device detects aggregation/precipitation of the samples by measuring the back-reflection intensity of a light beam that passes through the capillary. The apparent protein melting temperatures (T_m) were determined with the PR. ThermControl software V2.1 (NanoTemper Technologies, Munich, Germany) from the maximum of the first derivatives of the thermal unfolding curves. The same software was used to determine the aggregation onset temperature (T_{agg}) from the increase in the signal from the aggregation detection optics. T_m and T_{agg} are mean of triplicates with standard deviations calculated with Origin.

4.2.9 Dynamic Light Scattering (DLS)

Samples with protein concentrations from 1 to 5 g/L were prepared and 10 μ L of each sample were pipetted in triplicates into a 1536 well plate (Aurora Microplates, Whitefish, USA). The plate was centrifuged at 2000 rpm for 2 min using a Heraeus Megafuge 40 centrifuge equipped with an M-20 well plate rotor (Thermo Fisher Scientific, Wilmington, USA). Two microliter of silicon oil was added to seal each well. The plate was centrifuged again and placed in a DynaPro DLS plate reader III (Wyatt Technology, Santa Barbara, USA). All measurements were performed at 25 °C with 10 acquisitions per well and an acquisition time of 5 s. The data was analyzed with the Dynamics V7.10 software (Wyatt Technology, Santa Barbara, USA). The diffusion interaction parameter (k_D) was determined according to the method that is described in the literature.^{49,50}

4.2.10 Small Angle X-ray Scattering (SAXS)

For SAXS measurements, samples with initial protein concentrations of 2, 5, and 7 mg/mL were prepared and shipped to the ESRF (The European Synchrotron Radiation Facility, Grenoble, France) on dry ice. Before measurements, the samples were thawed at room temperature and centrifuged at 10000 rpm for 10 min. Data collection was performed at the ID02 beamline. Data collection is summarized in the supplementary information. The DOI for the data is 10.15151/ESRF-ES-404440738.

The data processing and analysis were performed using ATSAS 2.8.2 software package.⁵¹ Before modeling, the low-q region was removed to avoid fitting on aggregation/repulsion; the high-q region was removed to avoid fitting on noisy data. The monomer structure of G-CSF was fitted to the SAXS curves using CRY SOL.⁵² We have carried out rigid body modeling of the G-CSF dimer on a mixture using SASREFMX.⁵³ The dimer structures with high occurrence were manually extracted from the CG simulations. The extracted CG dimers were backmapped using the SIRAH toolbox,⁴³ i.e. resulting in full atomic structures. The backmapped structures were converted to the OLIGOMER compatible input files using FFMAKER.⁵⁴ The output from the CG simulations was fitted to the SAXS data using OLIGOMER. The dimer with the best χ^2 value (the value close to 1) was selected to estimate the monomer/dimer fractions in the mixture.

4.3 Results and Discussion

4.3.1 pH-dependent structural differences of G-CSF

We have investigated the effect of the pH on the secondary and tertiary structure of G-CSF using a combination of modeling and biophysical techniques. The characteristic far-UV CD spectra with two minima at 209 and 222 nm confirm the presence of alpha-helical protein structure at all conditions and showed no difference between pH 4.0 and 5.0 and only a slight decrease of helical content when increasing the pH to 7.5 (Figure 15a). In agreement with our findings, the G-CSF structure solved at pH 3.5 (PDB code: 1GNC) has a similar helix content to the G-CSF structure solved at pH 7.5 (PDB code: 1CD9). A similar trend could be observed from HDX-measurement performed by Wood et al., who could not find any clear evidence for a change of helical contents between pH 4.25 and 7.4.⁴¹ In contrast, Narhi et al. used CD spectroscopy and showed that the helical content is noticeably higher at low pH (pH 4.5: helical content 75 % vs. pH 7.5: helical contents 66 %).¹⁸

The near-UV CD spectra of G-CSF at pH 5.0 and 7.5 are very similar in the wavelength region from 250 to 295 nm (Figure 15b). Surprisingly, G-CSF at pH 7.5 shows a negative CD signal at wavelengths from 300 to 340 nm which is very unusual for a protein in inorganic buffer but has been previously observed for filgrastim.⁵⁵ It is presumably caused by aggregates in the sample. The near-UV CD spectrum at pH 4.0 slightly deviates from the other spectra determined at pH 5.0 and 7.5 in the wavelength region 250 to 295 nm, but the characteristic features of the spectra remain the same. Therefore, G-CSF has a well-defined tertiary structure with only little difference between the three tested pH values.

The tryptophan fluorescence of G-CSF is significantly quenched when the pH is decreased from pH 7.5 to pH 4.0 indicating that the Trp residues are in different conformational states at the different pH values (Figure 15c) This observation is in accordance with the findings of Narhi et al.¹⁸ G-CSF contains two Trp residues: Trp59 and Trp119 which are located close to His157 and His80, respectively (Figure 15d). The change in pH causes a conformational change of Trp that promotes interactions between Trp and positively charged His leading to the quenching of Trp. Furthermore, the pH-dependent change of the Trp residues is observed in the available PDB structures. The NMR structure of G-CSF at pH 3.5 (PDB code: 1GNC) revealed that the Trp residues can interact with the neighboring His residues.²⁵ On the other hand, the X-ray structure

obtained at pH 7.5 (PDB code: 1CD9), shows that Trp59 points away from His157. In contrast to 1CD9, Trp59 is pointing upwards in 1GNC (Figure 16), indicating that the conformation of the Trp residues is dependent on pH. Based on these findings, we hypothesized that the Trp configuration is an important factor influencing the conformational stability of G-CSF.

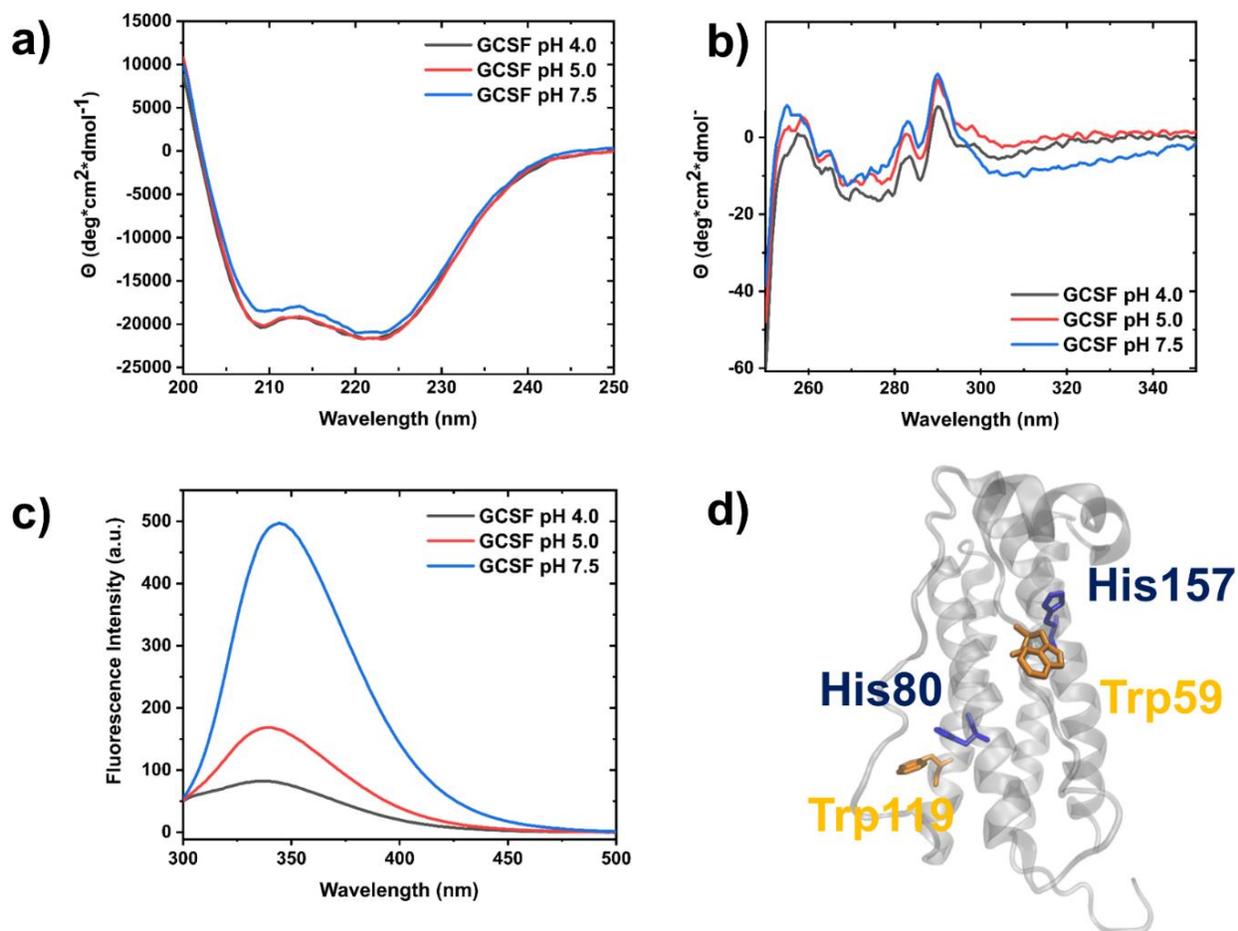


Figure 15: Effect of pH on the G-CSF secondary structure studied with (a) far-UV circular dichroism; and on the G-CSF tertiary structure studied with (b) near-UV circular dichroism. (c) fluorescence intensity measurements which indicate that the Trp residues in G-CSF are quenched at pH 4.0. (d) The location of Trp and His residues in G-CSF (PDB code: 1CD9). The protein is shown in a transparent cartoon structure. The investigated Trp and His residues are shown as sticks and colored in orange and blue, respectively.

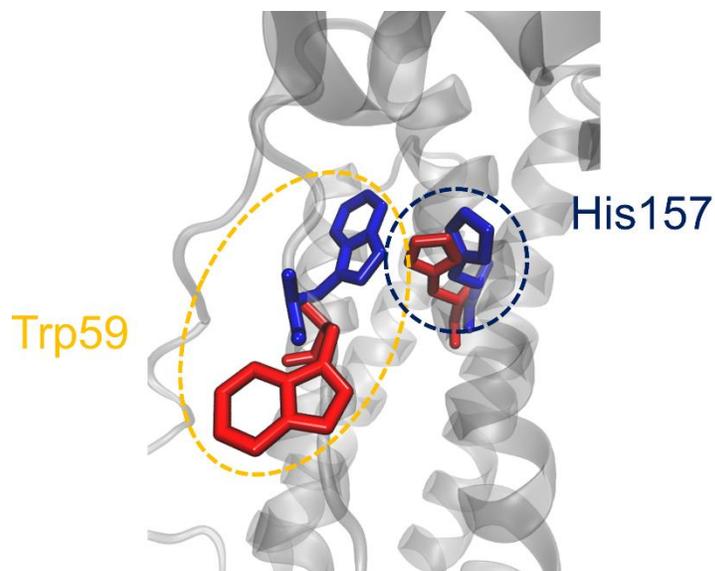


Figure 16: The conformational change of Trp59. The sidechain structure of G-CSF at pH 4.0 (PDB code: 1GNC) is colored blue. The sidechain structure at pH 7.0 (PDB code: 1CD9) is colored red. Note, Trp59 forms an upward configuration at pH 4.0. The protein is shown using in transparent cartoon structure.

To further investigate the effect of Trp configurations on the G-CSF structural integrity, we performed cMD simulations with the crystal structure 1CD9 as starting structure. The simulations were carried out for 400 ns. During the simulations, no significant conformational changes of the Trp residues could be observed. Presumably, 400 ns cMD simulations were not sufficient to induce noticeable structural changes. Therefore, we continued with well-tempered metadynamics simulations where bias potentials are added as a function of the center of mass (COM) distances between Trp and His side chains. To check the overall conformational stability, the alpha-helical content was chosen as the third CV. The 2D and 3D free energy surfaces (FES) of the CVs are shown in Figure 17. The time evolution of the FES is provided in the supplementary information.

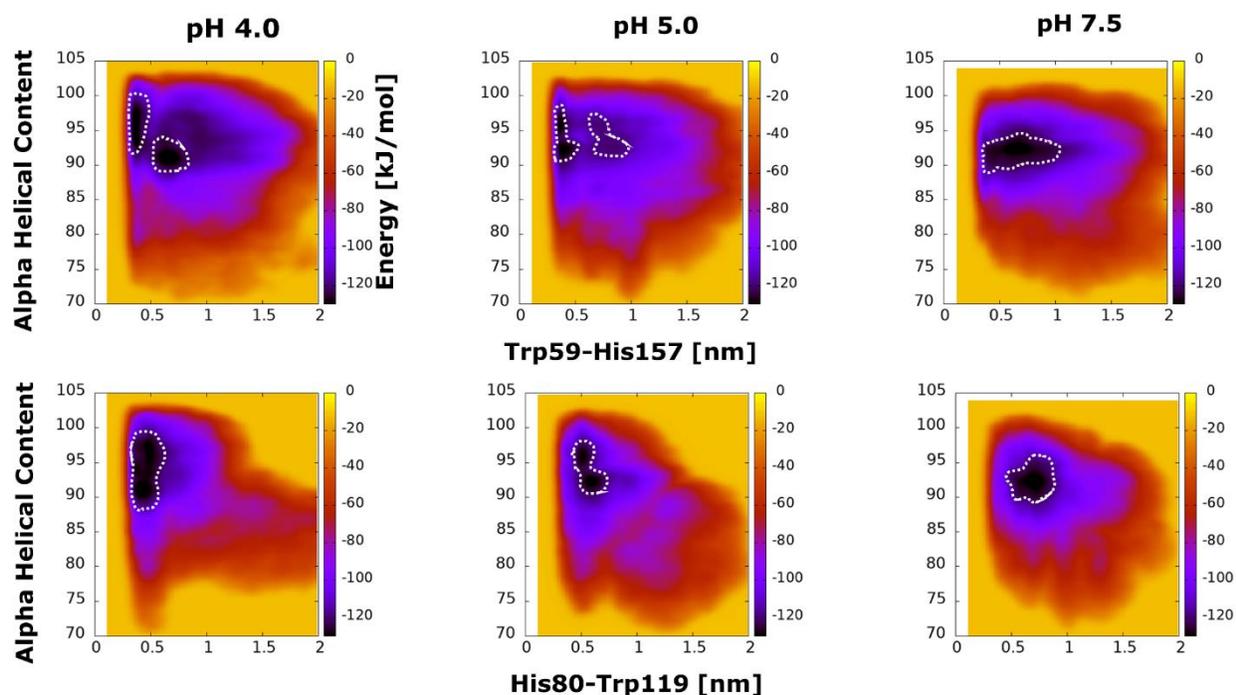


Figure 17: Estimate of the FES of G-CSF at different pH values. Each energy surface is obtained as a function of the α -helical content and the distance between His and Trp residues. The local minima are highlighted with dashed white circles. Top panel: The COM distance between Trp59 and His157 is on the x-axis. Bottom panel: The COM distance between His80 and Trp119 is on the x-axis.

Interestingly, the Trp59(loopAB)-His157(α D) pair and His80(α B)-Trp119(α C) pair show different behavior. Since Trp59 is located in the loopAB, it has much higher flexibility and can move away from His157 easier than Trp119(α C) from His80(α B). Therefore, the FES of Trp59-His157 can be sampled at a COM distance larger than 1.5 nm. In addition, all three pH conditions could reproduce the upward state of Trp59, where an energetic minimum could be estimated at a Trp59-His157 distance of around 0.4 nm. However, the Trp59 residues at pH 4.0 and pH 5.0 have an energy barrier between the up and down position corresponding to a breakage of the cation- π -interactions between Trp59 and protonated His157, while Trp59 can freely move between the two configurations at pH 7.5.

Contrarily, it is difficult to separate His80-Trp119 more than 1 nm and only one local minimum could be found from the COM distance. At pH 7.5, the COM distance of His80-Trp119 remained around 0.7 nm. At lower pH, the COM distance of His80-Trp119 remained around 0.5 nm. In addition, relatively larger fluctuation could be observed at pH 7.5, indicating that His80-Trp119

are located close to each other but they are not able to form a strong cation- π -interaction, since there is no cation at pH 7.5.

The histidine residues located closely to the Trp residues in G-CSF will be protonated at low pH. The FES has shown that the interaction between Trp and its neighboring His residue is much more favorable at pH 4.0. This interaction between Trp and His residues may stabilize G-CSF at pH 4.0 compared to pH 7.5 by clamping loopAB to helix α D and helix α B to helix α C, therefore making the structure locally less flexible. Aubin et al. investigated the interactions between Trp and His residues at pH 4.3, 5.0, and 6.4 using NMR.¹⁹ Based on chemical shift analysis, the authors could show that changes in Trp-His interactions affect the conformational stability of G-CSF.¹⁹ In addition, Ghasriani et al. have determined the relaxation parameters of G-CSF using NMR spectroscopy and assessed the protein flexibility from the calculated order parameters.⁴⁰ The authors found that the main difference between pH 4.0 and 6.0 was due to the change in the loop and helical flexibility. The authors observed that the flexibility of loopCD was increased at pH 4.0, whereas a very slight increase of flexibility was seen for loopAB at pH 4.0. The π -cation interaction between Trp59 and His157 can prevent an increase of loopAB mobility. In contrast, a decrease of flexibility could be observed for the helical packing at pH 4.0, and the authors suggested that the His80-Trp119 interaction can be the factor that is involved in reducing the flexibility. On the other hand, Wood et al. have reported an increase of the loopCD deuterium uptake at pH 7.4.⁴¹ One of the challenging parts of the experimental characterization of G-CSF above pH 6.2 is that an extensive aggregation can occur in the sample.⁴¹ Compared to experiments, the metadynamics simulations (performed on a single G-CSF molecule) provide an option to study protein conformation in highly aggregation-prone physicochemical conditions without the interference of protein-protein interactions. The current FES study has been focused on the CVs that can be directly observed in the fluorescence intensity and CD measurements (Figure 15), since the FES are based on a few local CVs the magnitude of flexibility may be dependent on the choice of CVs.

The observation made from HDX⁴¹ and NMR¹⁹ experiments is in good agreement with our metadynamics simulations, where the overall α -helical content is not significantly affected by adjusting the pH. The MD study indicates that the interactions between Trp and His residues may affect the local structural conformation and loop mobility. The interactions between loopAB- α D (Trp50-His157) and α B- α C (His80-Trp119) are lost at pH 7.5, suggesting that G-CSF will be more flexible at pH 7.5.

4.3.2 Effect of pH and Sodium Chloride on the Thermal Unfolding and Aggregation of G-CSF

The structural changes of G-CSF due to different pH values affect the thermal unfolding and aggregation of the protein. We furthermore aimed to evaluate the influence of sodium chloride on the stability of G-CSF to elucidate the importance of electrostatic interactions. Therefore, we performed nanoDSF measurements and determined T_m and T_{agg} (Table 2). At pH 4.0, G-CSF unfolds significantly later ($T_m \sim 65$ °C) than at pH 5.0 ($T_m \sim 52$ °C) and pH 7.5 ($T_m \sim 55$ °C) and does not form detectable aggregates in contrast to higher pH. Addition of 100 mM sodium chloride at pH 4.0 causes a significant shift in the unfolding transition of G-CSF to a lower temperature ($T_m \sim 53$ °C). The same trend was found for the onset of aggregation (T_{agg}). This shift cannot be seen at pH 5.0 and is less pronounced at pH 7.5. This shows that sodium chloride has a more detrimental effect on the thermal stability of G-CSF at low pH. At pH 5.0, close to the isoelectric point of the protein, sodium chloride has only a small effect on the aggregation behavior. The lower thermal stability at pH 7.5 compared to pH 4.0 is decreased even more upon addition of sodium chloride.

Table 2: T_m and T_{agg} of G-CSF were determined with the PR. ThermControl software from the thermal unfolding curves and the increase in the signal from the backreflection of the nanoDSF measurements. T_m and T_{agg} are mean of triplicates with standard deviations. (NA- no detection of aggregates.)

| | T_m [C°] (\pm error) | T_{agg} [C°] (\pm error) |
|----------------------|---------------------------|-------------------------------|
| pH 4.0 | 64.95 ± 0.02 | NA |
| pH 4.0 + 100 mM NaCl | 53.34 ± 0.05 | 53.51 ± 0.05 |
| pH 5.0 | 52.27 ± 0.07 | 50.61 ± 0.18 |
| pH 5.0 + 100 mM NaCl | 52.25 ± 0.06 | 50.75 ± 0.22 |
| pH 7.5 | 54.87 ± 0.06 | 55.1 ± 0.00 |
| pH 7.5 + 100 mM NaCl | 51.08 ± 0.03 | 47.46 ± 0.09 |

The colloidal stability of G-CSF in all tested conditions was assessed by means of the interaction parameter k_D which describes the interaction of proteins in solution (Figure 18, Table 3).⁵⁶ k_D is commonly used as a surrogate parameter for the osmotic second virial coefficient B_{22} , which is directly related to PPIs, whereas k_D provides a less direct relationship. In general, positive k_D values indicate net repulsive PPIs and negative values correspond to net attractive interactions. However, the reversal does not occur exactly at zero. The excluded volume contribution to k_D is smaller than for B_{22} and therefore values of k_D can be negative when B_{22} values are still positive.

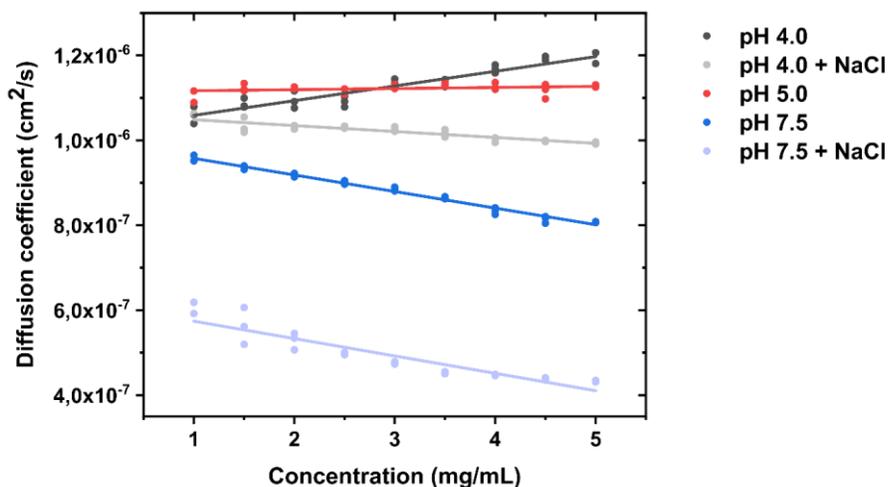


Figure 18: Diffusion coefficients at increasing protein concentrations assessed with DLS at pH 4.0, 5.0, and 7.5 with and without the addition of 100 mM NaCl.

There is considerable variation in the y-intercept, i.e. the diffusion coefficient at infinite dilution, for the measured conditions which could be due to the formation of irreversible species which do not dissociate upon dilution or due to protein conformational changes. Since we could not observe large conformational changes but a different tendency to form aggregates in the other methods, we assume that the samples contained irreversible aggregates. To confirm this hypothesis, we used the Stokes-Einstein relation to calculate the infinite dilution values for the hydrodynamic radius $r(H)_0$ from the infinite dilution diffusion coefficients (D_0) for each condition (Table 3). The $r(H)_0$ values range from 2.1 up to 4.0 nm, whereas the reported value is 2.0 nm,⁵⁷ which confirms the presence of larger species in our samples. This in turn impedes the correct determination of the diffusion interaction parameter k_D . Additionally, the partial specific volume of the protein is expected to be a function of pH and could significantly contribute to differences in k_D . However, the partial

specific volume should only change upon unfolding which could neither be observed in CD measurements nor MD simulations. Therefore, we do not expect the partial specific volume to have a drastic effect on the k_D values. To support this hypothesis, we submitted the last frames of the conventional all-atom MD simulations at the respective pH values to the HullRad webserver, which calculates the partial specific volume of a protein from a PDB structure. A partial specific volume of 0.75 mL/g was calculated for all three pH values.

Table 3: k_D and $r(H)_0$ derived from DLS measurements performed at different pH and ionic strength values. Due to strong aggregation, k_D could not be determined at pH 5.0 with salt.

| Buffer | k_D [mL/mg] | $r(H)_0$ [nm] |
|----------------------------------|-----------------------|---------------|
| 10 mM NaAc pH 4.0 | $3.3 \cdot 10^{-2}$ | 2.3 |
| 10 mM NaAc pH 4.0 + 100 mM NaCl | $-1.2 \cdot 10^{-2}$ | 2.1 |
| 10 mM NaAc pH 5.0 | $4.18 \cdot 10^{-3}$ | 2.2 |
| 10 mM NaAc pH 5.0 + 100 mM NaCl | NA | NA |
| 10 mM KPhos pH 7.5 | $-3.49 \cdot 10^{-2}$ | 2.5 |
| 10 mM KPhos pH 7.5 + 100 mM NaCl | $-6.36 \cdot 10^{-2}$ | 4.0 |

G-CSF shows a positive k_D and is assumed to show repulsion at pH 4.0 which is in agreement with the proposed highly positive electrostatic surface at low pH (Figure 19). The addition of salt screens the surface charges of the protein and results in a negative k_D . These observations are in correlation with the strongly decreased thermal stability at low pH upon addition of salt. A k_D of almost zero could be observed at pH 5.0, which indicate no strong attractive nor repulsive forces between the protein monomers. This behavior is expected at a pH close to the isoelectric point where the protein has (almost) no net charge. Due to the very high level of the aggregation, the k_D could be not measured when salt was added to the pH 5.0 formulation. This result is in accordance with the observations from Chi et al.¹⁶ The authors used static light scattering experiments to obtain the osmotic second virial coefficient (B_{22}) value. A positive and negative B_{22} value could be determined at pH 3.5 and pH 6.1, respectively.¹⁶ Aggregation of G-CSF occurred when 150 mM

of NaCl was added to the formulation, and the B_{22} value could not be determined due to precipitation.¹⁶ The negative k_D value at pH 7.5 suggests that the G-CSF monomers attract each other.

This is in accordance with the surface potential of G-CSF which is highly pH-dependent (Figure 19) as G-CSF contains a relatively large number of charged residues. The net charge of G-CSF at pH 4.0 is estimated to +13 e using PDB2PQR.²⁹ The electrostatic surface of the helical bundle is highly positively charged. Therefore, it is expected that G-CSF will be repulsive at pH 4.0. At pH 5.0, the net charge of G-CSF is decreased to +1 e, and it becomes -4 e at pH 7.5. Hence, electrostatic interactions play a substantial role in the aggregation process of G-CSF.

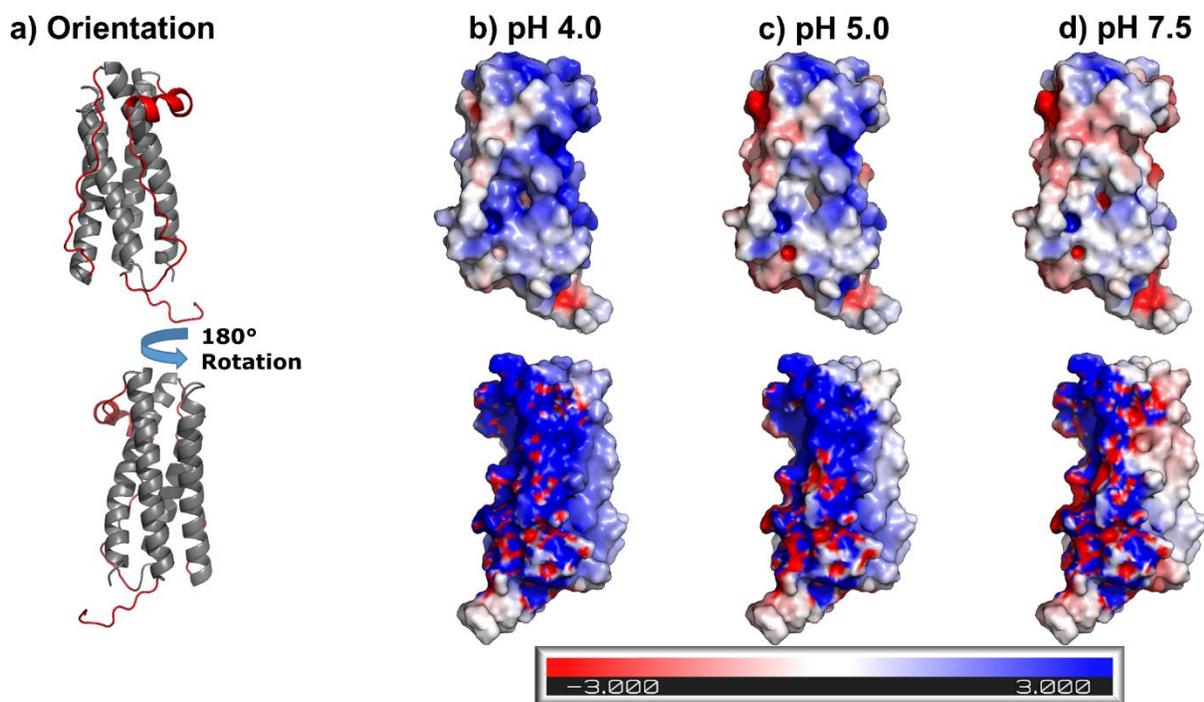


Figure 19: Electrostatic surface properties of G-CSF at different pH values. (a) Orientation of the electrostatic surface. The flexible N-terminus, loopAB, and loopCD are colored in red. Top: The region containing loopAB and loopCD is on the front view. Bottom: The helical bundle without any long loop structures is on the front view. (b-d) The electrostatic potential surface at different pH was calculated using the APBS electrostatics plugins⁵⁸ in PyMOL⁵⁹ and PDB2PQR²⁹.

In order to simulate the aggregation behavior of G-CSF at the different pH values, we performed CG simulations with eight monomers in a pH series. The snapshots of the first 500 ns simulation before the aggregation are shown in the supporting information. The aggregation behavior of G-

CSF at different conditions was estimated by tracking the number of protein-protein interactions (PPI) during the simulations (Figure 20). The number of PPI during the 3 μs of the production run was defined as the number of observed intermolecular residue pairs with pair distance less than 4 \AA .

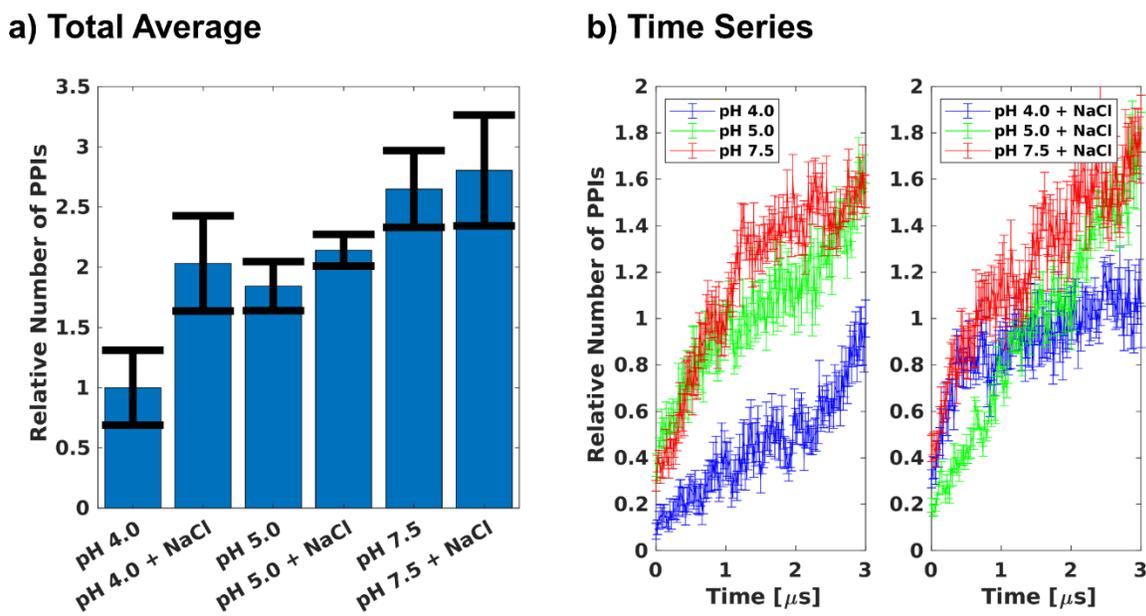


Figure 20: The relative number of protein-protein interactions (PPIs) obtained from the CG simulations. The number of PPIs during the 3 μs of the production run was defined as the number of observed intermolecular residue pairs with pair distance less than 4 \AA . The number of PPIs is normalized to the number of interactions observed in the pH 4.0 simulations. **(a)** The total relative number of PPIs during the simulations. Each bar represents the mean and the standard error of the mean of the five trajectories performed at each pH. **(b)** Time series of the relative number of PPIs. Each bar represents the mean and the standard error of the mean of the five trajectories that are observed in the current MD frame.

Since the total number of the PPIs is highly dependent on the simulation time and the size of the simulation box, the number of interactions is normalized by the number of interactions obtained at pH 4.0. Addition of NaCl or increasing the pH value to 5.0 resulted in a 2-fold increase of the sampled PPI compared to pH 4.0. A 2.5-fold increase of the PPI could be observed at pH 7.5 (Figure 20a). The 2-fold increase is following the trend that was observed for the k_D data (Table 3), and the SIRAH FF model could reproduce the increase of PPIs at the aggregating conditions. Figure 20b shows the time evolution of the relative number of PPIs during the simulations. It is interesting to note that the total number of PPIs increases with simulation time, indicating that the

overall tendency is an aggregation (irreversible oligomer formation) rather than an association (reversible oligomer formation). An increase in the relative number of PPIs can also be observed at pH 4.0. However, this is not completely surprising since the CG simulations have been performed at relatively high protein concentration (approximately 30 mg/mL) to reduce the computational time for sampling PPI. Interestingly, the slope of the time evolution of the PPI is different at each pH. At pH 7.5, a much faster increase of the relative number of PPIs is observed when compared to pH 5.0. Since the number of monomers is limited to 8, the relative number of PPIs at 3 μ s is very similar for pH 5.0 and 7.5, indicating that the difference in the observed relative number of PPIs will also be dependent on the simulation time. Accuracy and performance will always be a trade-off when sampling PPIs between diffusing protein monomers. In an ideal case, very long CG sampling with a relatively large box with many protein monomers will give more accurate sampling at the expense of computational time, but it is expected that the results will show a similar tendency observed here.

In the CG simulations, no significant increase of aggregation propensity could be sampled between pH 4.0 + NaCl, pH 5.0, and pH 5.0 + NaCl. The highest aggregation behavior could be observed at pH 7.5. Note that the degree of increase in the PPI may be dependent on the size of the simulation box and the number of the G-CSF monomers. It is implied that the relative number of PPI might change if the simulation condition is changed. However, the overall trend is expected to be the same. The results from the CG simulations suggest three different aggregation states: 1) weak aggregation at pH 4.0, 2) moderate aggregation at pH 4.0 + NaCl, pH 5.0, and pH 5.0 + NaCl, and 3) strong aggregation at pH 7.5 and pH 7.5 + NaCl.

To characterize the region of the aggregation-prone residues, interacting residue pairs in the CG simulations were depicted in heatmaps. The pixels of the heatmap are assigned by the number of frames with the pair-distance less than 4 Å that was registered from all five trajectories. The color bar is scaled to the strongest interactions that could be observed in the pH 4.0 simulations. Residue pairs were only included in the heatmap if the interactions occurred for more than 25 % of the strongest interaction observed in the pH 4.0 simulations. An example of the heatmap is shown in Figure 21. To visualize the residues that are involved in the PPI, the residues in the G-CSF structure were colored in a similar color scheme as in the heatmaps (Figure 22).

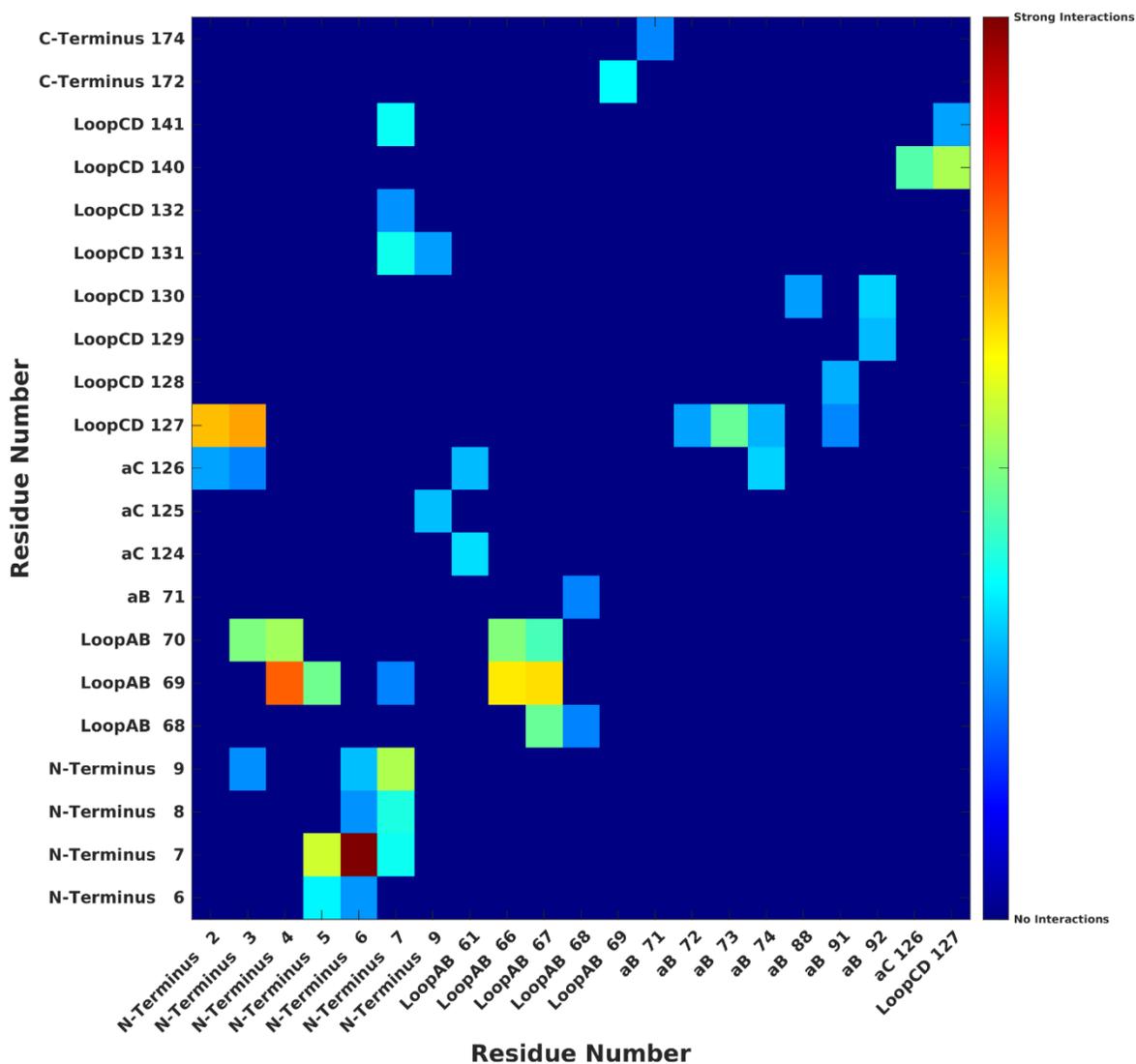


Figure 21: Protein-protein interaction (PPI) heatmap at pH 4.0. The x- and y-axes describe the residue number and their secondary structural localization of the interacting residue pair. The interacting residue pair between different monomers from all five simulations are collected into one data set. The color scale indicates the occurrence of the interactions between specific residue pairs. The color bar is scaled to the strongest interaction that occurred in the pH 4.0 simulations; here N-Terminus 6 - N-Terminus 7 interaction.

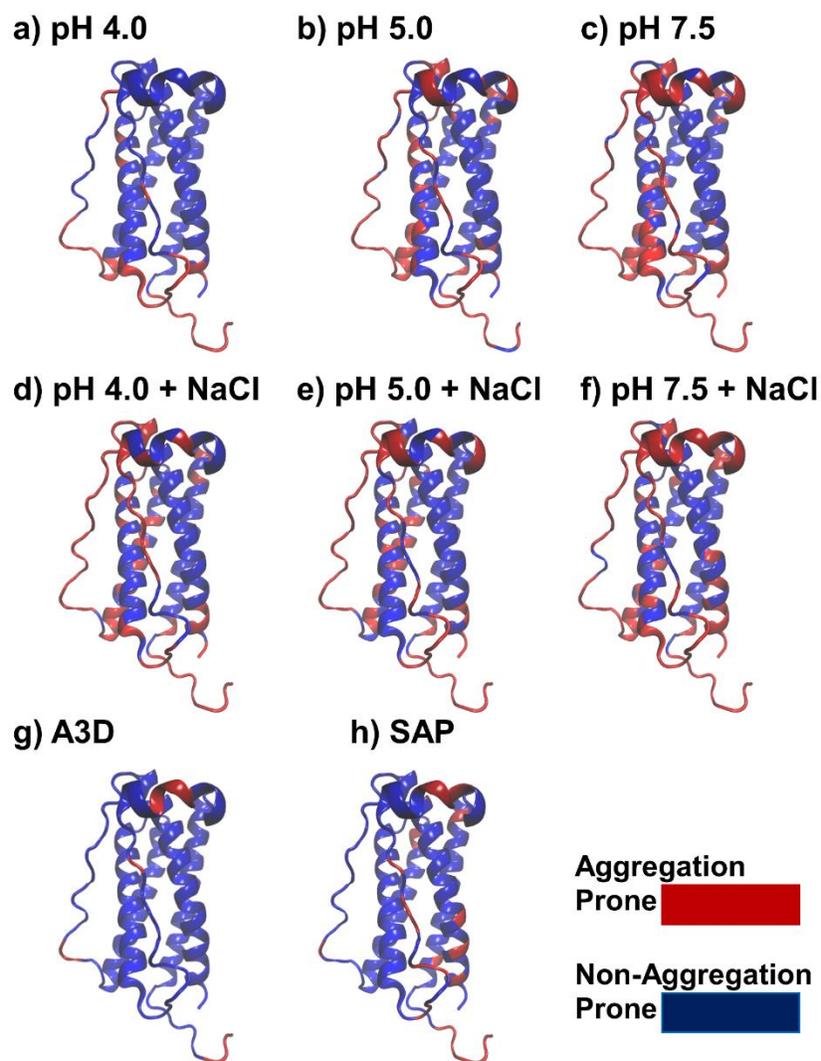


Figure 22: The aggregation-prone residues determined from the CG simulations and prediction algorithms based on the PDB structure. (a-f) Interacting residues determined from the CG simulations. Aggregation-prone residues that were involved in the PPI in the CG simulations are colored red. Residues not prone to participate in aggregation (blue) interacted 25 % or less compared to the strongest interaction at pH 4.0. **(g-h)** Aggregation-prone residues are predicted from the initial PDB structure (PDB code: 1CD9) using aggrescan3d (A3D) and spatial aggregation propensity (SAP). Red residues represent aggregation-prone residues. Blue residues represent non-aggregation prone residues based on A3D-score or SAP lower than 25 % of the strongest A3D / SAP score from the 1CD9.

At pH 4.0, the N-terminal part of G-CSF is the main region participating in aggregation (Figure 21 and Figure 22a). Since G-CSF is highly charged at pH 4.0, it is expected that the G-CSF monomers will repel each other. However, since the N-terminal part of G-CSF does not contain any charged residues, is very flexible and exposed to the solvent, it can still interact with other G-CSF monomers. Therefore, it appears that N- to N-terminus interactions may be one of the dominant PPI at pH 4.0 (Figure 21 and Figure 22a). Shibuya et al. studied the colloidal stability of the backbone circularized G-CSF, i.e. the N- and C-termini of G-CSF are connected. Their study revealed that backbone circularization of G-CSF at pH 4.0 leads to a more aggregation-resistant G-CSF when a protein denaturant is added.⁶⁰ When NaCl is added to the simulations at pH 4.0, both loopAB and loopCD are participating in the PPI (Figure 22d), which indicates that the electrostatic repulsion between the G-CSF monomers is the main limiting factor of the intermolecular long loop interactions. The aggregation-prone residues at pH 5.0 and pH 5.0 + NaCl are following a similar pattern as seen for pH 4.0 + NaCl (Figure 22b, d, and e). The relative number of interactions is very similar in these conditions (Figure 20). Therefore, this suggests that the aggregation behavior at these conditions mainly originated from the loss of the repulsion between the G-CSF molecules, i.e. that colloidal stability plays a larger role than conformational stability. At pH 7.5, the short helix in loopAB and the bottom part of the helix bundle located close to the N- and C-termini become more prone to aggregation. Addition of NaCl at pH 7.5 has a minimal effect on aggregation which indicates the screening of electrostatic interactions does not have a noticeable effect on G-CSF aggregation at this pH which is in accordance with our experimental data.

We also determined the aggrescan3d (A3D) score and the spatial aggregation propensity (SAP)⁶¹ using PBD structure 1CD9 (Figure 22g) to compare the results obtained from relatively fast prediction algorithms with results from computational demanding CG simulations. Interestingly, the A3D/SAP calculations were able to predict the N-terminus and large area of the loopAB region as aggregation-prone regions (Figure 22g-h). Those aggregation-prone residues follow a similar pattern as seen from the CG simulations. However, CG simulations have the advantage that aggregation-prone regions can be determined in a pH-dependent manner revealing additional aggregation-prone regions.

Observing the overall pattern of the aggregation-prone residues from the CG simulations, it becomes clear that the aggregation mechanism of G-CSF is non-specific, e.g. more than one aggregation site exists in G-CSF. Previously, Meric et al. used multiple aggregation prediction

algorithms and suggested that Leu83 located at ob is the most aggregation-prone residue.⁶² However, the authors found that the point mutation Leu83->Ala did not improve the aggregation propensity of G-CSF,⁶² which is in line with our results that indicate a non-specific aggregation mechanism.

To extract the important PPI of the G-CSF oligomers, we further analyzed the heatmap (Figure 21). For each condition, two to four dimers that contain several different strong interaction clusters were manually extracted from the CG simulations. The extracted G-CSF dimers are used to provide an aggregation ensemble of G-CSF oligomers. The first chain of the G-CSF dimers was aligned to a reference PDB structure (1CD9). After the alignments, only the second chain is shown together with the reference structure, mapping the different protein-protein interfaces in G-CSF aggregates (Figure 23).

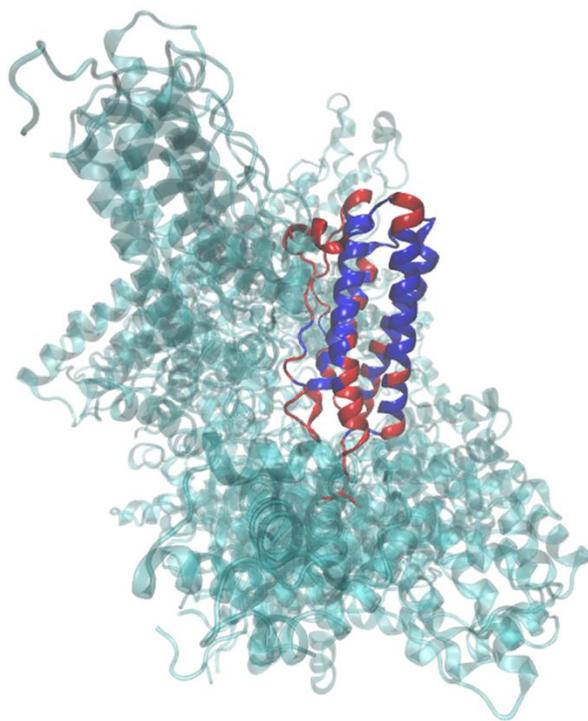


Figure 23: G-CSF aggregation ensemble. For each condition, multiple dimer structures containing several different strong interaction clusters were manually extracted from the CG simulations. Two to four dimers were extracted from each condition. In total 18 dimers were extracted from the CG simulations. The first chain of the extracted dimer was aligned to the reference structure (1CD9). The second chains of the ensemble are shown in transparent and cyan colored structures. The reference structure is colored according to the scheme used for aggregation-prone residues of pH 7.5 + NaCl (see Figure 22).

Diverse types of dimers can be observed from the aggregation ensemble. Both aggregation ensemble and the simulated aggregation-prone residues suggest that the exposed long loop regions are highly prone to aggregation. Interestingly, the exposed helical structures are not prone to aggregate, suggesting that a combination of electrostatic repulsion and compactness of the helical bundle prevents aggregation of helices. At pH 4.0, the long loop regions show a positively charged electrostatic surface (Figure 19 top). Since loopAB and loopCD become aggregation-prone when sodium chloride is added or the pH is increased to 5.0, one may argue that electrostatic repulsion of the long loop region is one of the most important factors to avoid the aggregation of G-CSF. Our metadynamics simulations suggest an increase of flexibility at pH 7.5 due to the loss of the Trp-His interactions. Since the probability to obtain unfolding of an alpha-helix at standard conditions is low without adding protein denaturants or heating the system, it suggests that the increase of loop flexibility initiates the aggregation of G-CSF.

Since the CG model has a limited atomic resolution, careful consideration is required when interpreting CG simulation results. In an attempt to validate and inspect the aggregation mechanism of G-CSF, we compared the CG simulation results to SAXS measurements of G-CSF, at different pH and NaCl concentrations (Figure 24). SAXS can be applied to investigate the inter-particle interactions of therapeutic protein.⁶³ However, it is extremely challenging to model the protein-protein interactions in the irreversible aggregating condition using SAXS data.⁶⁴ On other hand, SAXS data still provides valuable information when it is combined with the CG simulations since the combination of SAXS and CG simulations enables the direct comparison between computational and experimentally determined aggregation behavior.

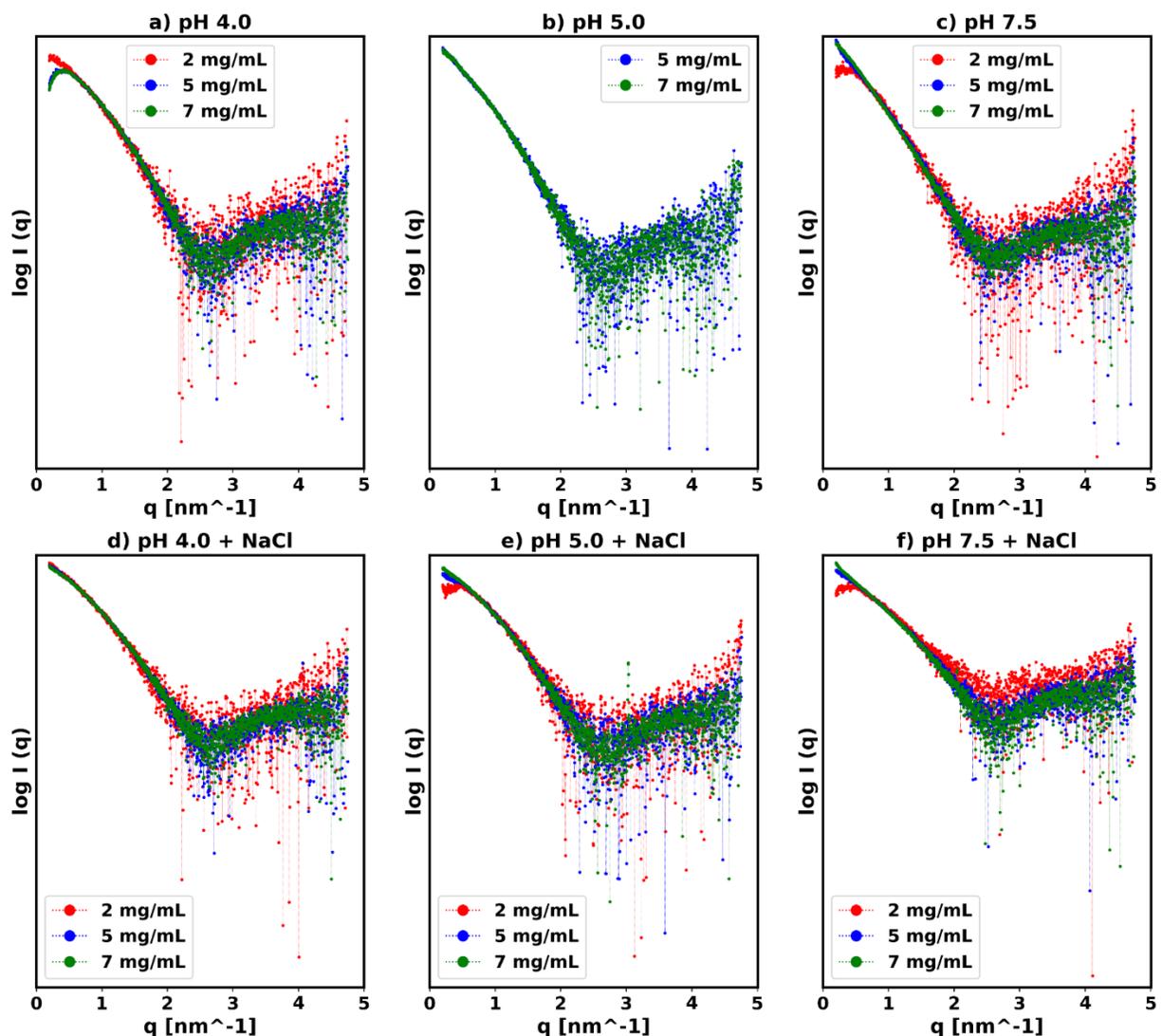


Figure 24: SAXS scattering curves of G-CSF at (a) pH 4.0 (b) pH 5.0, (c) pH 7.5, (d) pH 4.0 + 100 mM NaCl, (e) pH 5.0 + 100 mM NaCl, and (f) pH 7.5 + 100 mM NaCl. The protein concentration range: 2-7 mg/mL. The data set for pH 5.0 2 mg/mL was not included due to technical problems occurring during the measurements.

The SAXS data indicate that the only non-aggregating condition of G-CSF is at pH 4.0, where repulsion between G-CSF molecules is observed (Figure 24a). Since the data measured at the highest concentration is less noisy, it was used for the modeling process.

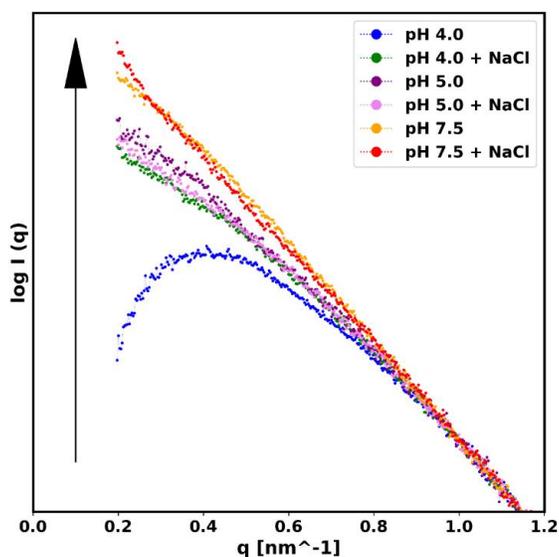


Figure 25: SAXS scattering profile plotted at the low-q region. The arrow illustrates the aggregation of G-CSF with increasing pH. The protein concentrations of the samples are 7 mg/mL.

The aggregation of G-CSF was initially investigated by inspecting the low-q region of the SAXS data. According to the data shown in Figure 25, increasing pH and the addition of NaCl lead to an increase in aggregation, which is in agreement with the aggregation profile deduced from the CG simulations (Figure 20).

In order to investigate the fraction of higher order species of the SAXS data, the dimer fraction of the considered samples was calculated. The obtained molecular weight of G-CSF in the aggregating conditions was between the molecular weight of monomer and dimer. Therefore, we decided that the modeling would be focused on the dimer only. The following modeling approaches were applied: i) fitting a dimer structure that was obtained from rigid-body modeling using SASREFMX⁵³ with two high-resolution monomer structures (1CD9) as an input (Figure 26a) and ii) fitting of the monomer (1CD9) and the dimer structures that were extracted from the CG simulations (Figure 23) using OLIGOMER⁵⁴ (Figure 26b). Both models assume that the scattering data are from the mixture, meaning that both monomer and dimer structures will be fitted to the SAXS data simultaneously. Furthermore, CRY SOL⁵² was used to fit the stand-alone monomer structure that is obtained from 1CD9. To validate the dimer structure, the χ^2 value of the monomer fitting was compared to the outcomes of the dimer fitting. The obtained dimer fraction is shown in Figure 26.

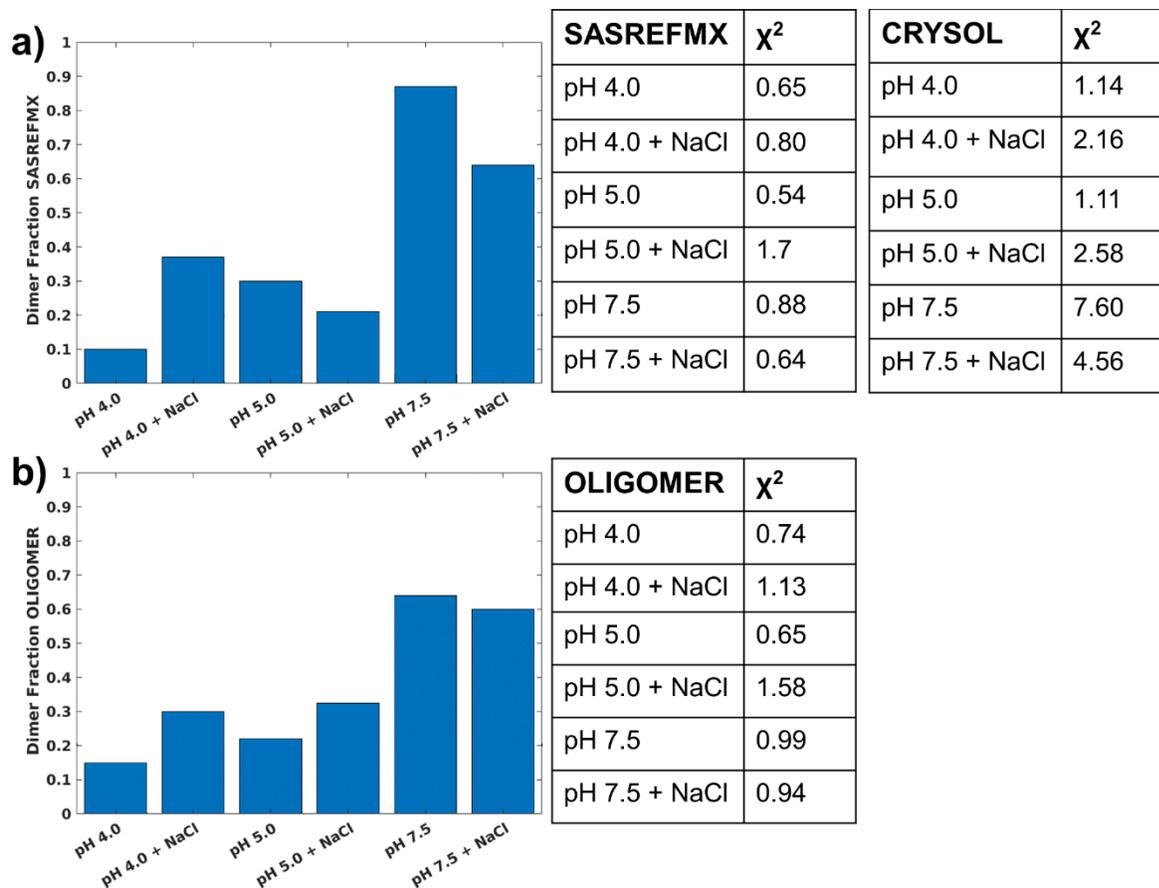


Figure 26: Analysis of the dimer fractions of G-CSF at different pH conditions. (a) Dimer fitting using SASREFMX⁵³, and (b) Dimer fitting using OLIGOMER⁵⁴. The χ^2 of CRY SOL⁵² is generated from the fitting of the G-CSF monomer structure. The modeling was performed with SAXS data obtained for protein concentration 7 mg/mL.

Overall, monomer + dimer has a better fit on the experimental data than the monomer only fit, meaning that both are present at all studied conditions (Figure 27). The rigid-body modeling approach had higher freedom to create the dimer structure to obtain an optimal fitting. Therefore, the result obtained from SASREFMX usually gave a better fit (Figure 27b). However, one must note that the dimer structure that is generated from the rigid body modeling may not have a physically realistic protein-protein interface. The structures obtained from the rigid body modeling can be found in the supporting information.

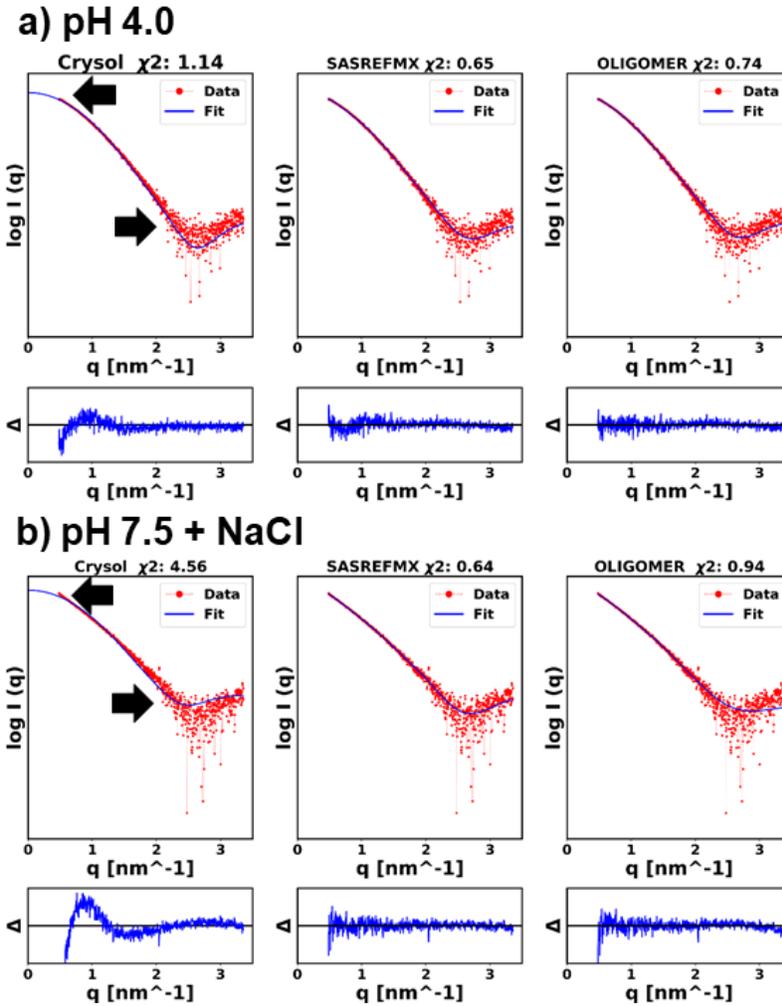


Figure 27: The results of the different modeling techniques. Two conditions are chosen as examples: (a) non-aggregating pH 4.0, and (b) highly aggregating pH 7.5 + NaCl. The arrows indicate a misfit at the low- q region. The difference between the fit and scattering profile (Δ) is plotted below the fittings. The horizontal line indicates $\Delta=0$. All y-axes are scaled to the same arbitrary chosen range. The modeling was performed with SAXS data obtained for protein concentration 7 mg/mL.

One interesting outcome is that the dimer fraction from the SAXS modeling (Figure 26) follows a similar trend as obtained from the CG simulations (Figure 20). Two entirely different modeling approaches could separate PPI at 3 different levels: 1) at pH 4.0, 2) pH 4.0+NaCl, pH 5, pH 5.0+NaCl, and 3) pH 7.5 and pH 7.5 + NaCl.

However, the modeling based on SAXS data (Figure 26) provided a more pronounced increase of dimer fraction at pH 7.5 compared to our results from the CG simulations (Figure 20). Note that the increase of dimer fraction can also indicate bigger aggregation species. One must note that the

CG simulation results will be dependent on the size of the system, the number of included monomers, and the simulation time and size. Thus, the CG simulations alone do not provide an accurate description of the level of aggregation compared to experimental data. Therefore, the SAXS data was included to predict the level of aggregation in a more physically correct manner. The SAXS data could validate the trend that was obtained from CG simulations. Furthermore, it was possible to propose a possible G-CSF dimer structure by combining both CG simulations and SAXS measurements (Figure 28). The proposed G-CSF oligomer structure contains realistic PPI, and the structure could be directly related to the experimental data.

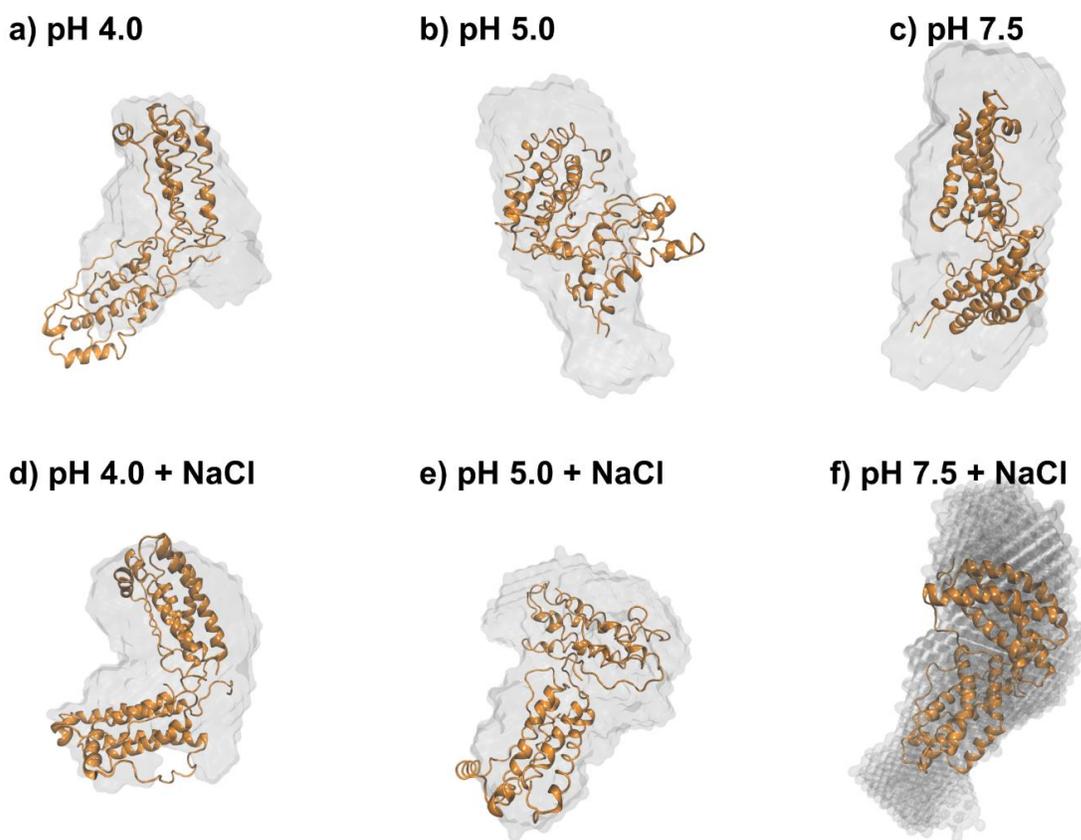


Figure 28: Dimer models of G-CSF at different conditions combining dimers extracted from CG simulations and SAXS measurements. Note, the fitting has been performed using monomer and dimer mixtures. The simulated CG dimer structures were back-mapped and then fitted to the SAXS data at the corresponding condition. The dimer with the χ^2 value closest to 1 was selected. All dimer models selected from the CG simulations are aligned to the ab-initio model of the corresponding SAXS data (gray envelope) using DAMMIF⁶⁵. However, the G-CSF SAXS data showed a certain fraction of higher order species, and therefore, the interpretation of the particle shape derived from the SAXS data must be assessed with caution.

It is worthwhile to mention that the ab-initio models in Figure 28 have different shapes, which indicates that the amount of aggregating species is different at the different pH conditions, since the ab-initio model describes an averaged protein shape in solution. The selected dimer structure suggests that the N- to N-terminus interactions are dominant at pH 4.0, and that the long loops are involved in the aggregation at pH 5.0 and pH 7.5. We propose that the SAXS models can serve as an extension to the CG aggregation model of G-CSF, where the SAXS models can be used to provide a bridge between the CG modeled G-CSF and the real system.

4.3.3 Future perspectives and potential challenges

Our study on G-CSF demonstrates that the application of orthogonal techniques can provide a molecular understanding of the driving forces for PPIs. Since soluble aggregates are usually transient, heterogeneous and present at very low concentrations, an ideal experimental technique would be able to simultaneously quantify the size and conformation of each species in a sample, without immobilization or labeling. However, the applied experimental techniques in this work are measuring averages of all species present so that species with a low population are being neglected. In some cases, the presence of (irreversible) aggregates even impedes the data analysis, for example in the determination of k_D and the modeling of the SAXS data. The MD simulations allowed us to determine pH and ionic strength-dependent changes on conformation and PPIs at an atomistic level, yet needed experimental validation. Applying multiple techniques, the shortcomings of each technique can be compensated. Due to its distinct pH-dependent behavior, G-CSF is particularly suited as a model protein for this study. However, not all proteins show such dramatic pH-dependent differences and it has to be seen if the computational approaches are sensitive enough to distinguish the aggregation behavior. Nevertheless, we propose that our approach could be extended to other proteins/systems. The challenge will be as the protein size and complexity of the system increase (e.g., glycosylation or by including excipients in the simulations), the computational cost will increase. However, with the increasing processor and network technology performance, it will become feasible to simulate systems with higher complexity. Including excipients in simulations may also require additional force field development of these molecules in the coarse-grained presentation. CG simulations have been applied to investigate self-interactions of antibodies,^{66,67} where the antibody is usually coarse-grained to much smaller beads (6-12 beads),⁶⁶ or the self-association is monitored by simulating only two antibodies.⁶⁷

4.4 Conclusions

We have demonstrated that multiple approaches are required to shed light on the aggregation process of G-CSF. Full atomic simulations have shown that it was very difficult to capture the conformations of G-CSF in an unbiased system. The metadynamics study revealed that one of the most distinct conformational changes of G-CSF at the different pH values occurs due to the loss of cation- π interactions between Trp and the neighboring His residues. The α -helix stability is not noticeably affected by pH, indicating that helical destabilization is not the main issue in standard formulations. However, the results show that the loss of Trp59-His157 and His80-Trp119 interactions will cause a local perturbation, that may contribute to the increased flexibility of G-CSF at higher pH values. The CG simulations could provide the pH-dependent aggregation-prone regions of G-CSF, which was in accordance with the SAP results. The predicted aggregation-prone regions are the N-terminal region and the two long loops parts of G-CSF. We have inspected the electrostatic surface to explain the pH-dependent change in the aggregation promoting regions, and found that the long loop regions are repulsive at pH 4.0 due to the positively charged surface potential. The addition of salt or increase in pH will make G-CSF more aggregation-prone since it will reduce the electrostatic charge located closely to the highly aggregation-prone loop regions. Inspecting the dimer structure from the CG simulations, we observe that the α -helical structures are not participating in the aggregation, and that the aggregation of G-CSF is highly unspecific where multiple forms of G-CSF dimer can exist. The CG simulations lack atomic resolution, and it is not feasible to simulate the G-CSF aggregation in a physically realistic size scale. To overcome this problem, we have included SAXS data for validation and interpretation of the simulation outcomes. We could obtain a reasonable fitting by including the dimer structures extracted from CG simulations during the modeling based on the SAXS data. The dimer fraction from SAXS data and the number of interactions from CG simulations followed a similar trend. Since both modeling and experiments of the protein aggregation process are extremely challenging, it requires a combination of multiple approaches to compensate for the weakness of each. We found that the combination of various modeling approaches could shed light on the complex pH-dependent aggregation process of G-CSF.

4.5 Supporting information

Due to the volume of the supporting information and since the manuscript is published, the information can be found here: <https://doi.org/10.1016/j.csbj.2022.03.012>

4.6 References

1. de la Torre BG, Albericio F. The pharmaceutical industry in 2019. An analysis of FDA drug approvals from the perspective of molecules. *Molecules* **25**, 745 (2020).
2. H Tobin P, H Richards D, A Callender R, J Wilson C. Protein engineering: a new frontier for biological therapeutics. *Curr Drug Metab* **15**, 743–56 (2014).
3. Manning MC, Patel K, Borchardt RT. Stability of protein pharmaceuticals. *Pharm Res* **6**, 903–18 (1989).
4. Manning MC, Chou DK, Murphy BM, Payne RW, Katayama DS. Stability of protein pharmaceuticals: an update. *Pharm Res* **27**, 544–75 (2010).
5. Roberts CJ. Therapeutic protein aggregation: mechanisms, design, and control. *Trends Biotechnol* **32**, 372–80 (2014).
6. Chi EY, Krishnan S, Randolph TW, Carpenter JF. Physical stability of proteins in aqueous solution: mechanism and driving forces in nonnative protein aggregation. *Pharm Res* **20**, 1325–36 (2003).
7. Clarkson BR, Schön A, Freire E. Conformational stability and self-association equilibrium in biologics. *Drug Discov Today* **21**, 342–7 (2016).
8. Wang W. Protein aggregation and its inhibition in biopharmaceutics. *Int J Pharm* **289**, 1–30 (2005).
9. Zalar M, Svilenov HL, Golovanov AP. Binding of excipients is a poor predictor for aggregation kinetics of biopharmaceutical proteins. *Eur J Pharm Biopharm* **151**, 127–36 (2020).
10. Hill CP, Osslund TD, Eisenberg D. The structure of granulocyte-colony-stimulating factor and its relationship to other growth factors. *Proc Natl Acad Sci* **90**, 5167-5171 (1993).
11. Weiss M, Voglic S, Harms-Schirra B, Lorenz I, Lasch B, Dumon K, et al. Effects of exogenous recombinant human granulocyte colony-stimulating factor (filgrastim, rhG-CSF) on neutrophils of critically ill patients with systemic inflammatory response syndrome depend on endogenous G-CSF plasma concentrations on admission. *Intensive Care Med* **29**, 904-914 (2003).
12. Welte K, Gabrilove J, Bronchud MH, Platzer E, Morstyn G. Filgrastim (r-metHuG-CSF): The first 10 years. *Blood*, 1907-1929 (1996).

13. Aritomi M, Kunishima N, Okamoto T, Kuroki R, Ota Y, Morikawa K. Atomic structure of the GCSF-receptor complex showing a new cytokine- receptor recognition scheme. *Nature* **401**, 713-717 (1999).
14. Šali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234**, 779–815 (1993).
15. Krishnan S, Chi EY, Webb JN, Chang BS, Shan D, Goldenberg M, et al. Aggregation of granulocyte colony stimulating factor under physiological conditions: Characterization and thermodynamic inhibition. *Biochemistry* **41** (20), 6422-6431 (2002).
16. Chi EY, Krishnan S, Kendrick BS, Chang BS, Carpenter JF, Randolph TW. Roles of conformational stability and colloidal stability in the aggregation of recombinant human granulocyte colony-stimulating factor. *Protein Sci* **12** (5), 903-913 (2003).
17. Robinson MJ, Matejtschuk P, Bristow AF, Dalby PA. T m-values and unfolded fraction can predict aggregation rates for granulocyte colony stimulating factor variant formulations but not under predominantly native conditions. *Mol Pharm* **15**, 256–67 (2018).
18. Narhi LO, Kenney WC, Arakawa T. Conformational changes of recombinant human granulocyte-colony stimulating factor induced by pH and guanidine hydrochloride. *J Protein Chem* **10**, 359-367 (1991).
19. Aubin Y, Hodgson DJ, Thach WB, Gingras G, Sauvé S. Monitoring effects of excipients, formulation parameters and mutations on the high order structure of filgrastim by NMR. *Pharm Res* **32**, 3365–75 (2015).
20. Darré L, Machado MR, Brandner AF, González HC, Ferreira S, Pantano S. SIRAH: a structurally unbiased coarse-grained force field for proteins with aqueous solvation and long-range electrostatics. *J Chem Theory Comput* **11**, 723–39 (2015).
21. Machado MR, Barrera EE, Klein F, Sónora M, Silva S, Pantano S. The SIRAH 2.0 Force Field: Altius, Fortius, Citius. *J Chem Theory Comput* **15** (4), 2719-2733 (2019).
22. Barrera EE, Zonta F, Pantano S. Dissecting the role of glutamine in seeding peptide aggregation. *Comput Struct Biotechnol J* **19**, 1595–602 (2021).
23. Javanainen M, Martinez-Seara H, Vattulainen I. Excessive aggregation of membrane proteins in the Martini model. *PLoS One* **12**, e0187936 (2017).
24. Tamada T, Honjo E, Maeda Y, Okamoto T, Ishibashi M, Tokunaga M, et al. Homodimeric cross-over structure of the human granulocyte colony-stimulating factor (GCSF) receptor signaling complex. *Proc Natl Acad Sci* **103**, 3135–40 (2006).

25. Zink T, Ross A, Lüers K, Cieslar C, Holak TA, Rudolph R. Structure and Dynamics of the Human Granulocyte Colony-Stimulating Factor Determined by NMR Spectroscopy. Loop Mobility in a Four-Helix-Bundle Protein. *Biochemistry* **33** (28), 8453-8563 (1994).
26. Chu J-W, Yin J, Wang DIC, Trout BL. Molecular Dynamics Simulations and Oxidation Rates of Methionine Residues of Granulocyte Colony-Stimulating Factor at Different pH Values. *Biochemistry* **43** (4), 1019-1029 (2004).
27. Singh SK, Mishra A, Goel G, Chirmule N, Rathore AS. Modulation of granulocyte colony stimulating factor conformation and receptor binding by methionine oxidation. *Proteins Struct Funct Bioinforma* **89**, 68–80 (2021).
28. Rospiccio M, Arsiccio A, Winter G, Pisano R. The Role of Cyclodextrins against Interface-Induced Denaturation in Pharmaceutical Formulations: A Molecular Dynamics Approach. *Mol Pharm* **18** (6), 2322-2333 (2021).
29. Dolinsky TJ, Nielsen JE, McCammon JA, Baker NA. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res* **32**, W665-W667 (2004).
30. Case DA, Belfon K, Ben-Shalom I, Brozell SR, Cerutti D, Cheatham T, et al. Amber 2020 (2020).
31. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J Chem Theory Comput* **11**, 3696–713 (2015).
32. Horn HW, Swope WC, Pitner JW, Madura JD, Dick TJ, Hura GL, et al. Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *J Chem Phys* **120**, 9665–78 (2004).
33. Darden T, York D, Pedersen L. Particle mesh Ewald: An N log(N) method for Ewald sums in large systems. *J Chem Phys*. **98**, 10089–92 (1993).
34. Ryckaert J-P, Ciccotti G, Berendsen HJC. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J Comput Phys* **23**, 327–41 (1977)
35. Miyamoto S, Kollman PA. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J Comput Chem* **13**, 952–62 (1992).
36. Pastor RW, Brooks BR, Szabo A. An analysis of the accuracy of Langevin and molecular dynamics algorithms. *Mol Phys* **65**, 1409–19 (1988).
37. Åqvist J, Wennerström P, Nervall M, Bjelic S, Brandsdal BO. Molecular dynamics simulations of water and biomolecules with a Monte Carlo constant pressure algorithm. *Chem Phys Lett* **384**, 288–94 (2004).

38. Tribello GA, Bonomi M, Branduardi D, Camilloni C, Bussi G. PLUMED 2: New feathers for an old bird. *Comput Phys Commun* **185**, 604–13 (2014).
39. Barducci A, Bussi G, Parrinello M. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Phys Rev Lett* **100**, 20603 (2008).
40. Ghasriani H, Frahm GE, Johnston MJW, Aubin Y. Effects of Excipients on the Structure and Dynamics of Filgrastim Monitored by Thermal Unfolding Studies by CD and NMR Spectroscopy. *ACS Omega* **5**, 31845–57 (2020).
41. Wood VE, Groves K, Cryar A, Quaglia M, Matejtschuk P, Dalby PA. HDX and In Silico Docking Reveal that Excipients Stabilize G-CSF via a Combination of Preferential Exclusion and Specific Hotspot Interactions. *Mol Pharm* **17**, 4637–51 (2020).
42. Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1**, 19–25 (2015).
43. Machado MR, Pantano S. SIRAH tools: mapping, backmapping and visualization of coarse-grained models. *Bioinformatics* **32**, 1568–70 (2016).
44. Darré L, Machado MR, Dans PD, Herrera FE, Pantano S. Another coarse grain model for aqueous solvation: WAT FOUR? *J Chem Theory Comput* **6**, 3793–807 (2010).
45. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. *J Chem Phys* **81**, 3684–90 (1984).
46. Bussi G, Donadio D, Parrinello M. Canonical sampling through velocity rescaling. *J Chem Phys* **126**, 14101 (2007).
47. Parrinello M, Rahman A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J Appl Phys* **52**, 7182–90 (1981).
48. Kelly SM, Price NC. The use of circular dichroism in the investigation of protein structure and function. *Curr Protein Pept Sci* **1**, 349–84 (2000).
49. Harding SE, Johnson P. The concentration-dependence of macromolecular parameters. *Biochem J* **231**, 543–7 (1985).
50. Connolly BD, Petry C, Yadav S, Demeule B, Ciaccio N, Moore JMR, et al. Weak interactions govern the viscosity of concentrated antibody solutions: high-throughput analysis using the diffusion interaction parameter. *Biophys J* **103**, 69–78 (2012).

51. Franke D, Petoukhov M V, Konarev P V, Panjkovich A, Tuukkanen A, Mertens HDT, et al. ATSAS 2.8: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. *J Appl Crystallogr* **50**, 1212–25 (2017).
52. Svergun D, Barberato C, Koch MH. CRY SOL - A program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *J Appl Crystallogr* **28**, 768–73 (1995).
53. Petoukhov M V, Franke D, Shkumatov A V, Tria G, Kikhney AG, Gajda M, et al. New developments in the ATSAS program package for small-angle scattering data analysis. *J Appl Crystallogr* **45**, 342–50 (2012).
54. Konarev P V, Volkov V V, Sokolova A V, Koch MHJ, Svergun DI. PRIMUS: A Windows PC-based system for small-angle scattering data analysis. *J Appl Crystallogr* **36**, 1277–82 (2003).
55. Nikravesht FY, Shirkhani S, Bayat E, Talebkhan Y, Mirabzadeh E, Sabzalinejad M, et al. Extension of human GCSF serum half-life by the fusion of albumin binding domain. *Sci Rep* **12**, 1–13 (2022).
56. Zhang J, Liu XY. Effect of protein–protein interactions on protein aggregation kinetics. *J Chem Phys* **119**, 10972–6 (2003).
57. Jing W, Roberts JW, Green DE, Almond A, DeAngelis PL. Synthesis and characterization of heparosan-granulocyte-colony stimulating factor conjugates: a natural sugar-based drug delivery system to treat neutropenia. *Glycobiology* **27**, 1052–61 (2017).
58. Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA. Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proc Natl Acad Sci* **98**, 10037–41 (2001).
59. DeLano WL. Pymol: An open-source molecular graphics tool. *CCP4 Newsl Protein Crystallogr* **40**, 82–92 (2002).
60. Shibuya R, Miyafusa T, Imamura H, Ooishi A, Honda S. Effect of backbone circularization on colloidal stability: Compaction of unfolded structures improves aggregation resistance of granulocyte colony-stimulating factor. *Int J Pharm* **605**, 120774 (2021).
61. Chennamsetty N, Voynov V, Kayser V, Helk B, Trout BL. Prediction of aggregation prone regions of therapeutic proteins. *J Phys Chem B* **114**, 6614–24 (2010).
62. Meric G, Naik S, Hunter AK, Robinson AS, Roberts CJ. Challenges for design of aggregation-resistant variants of granulocyte colony-stimulating factor. *Biophys Chem* **277**, 106630 (2021).
63. Pohl C, Mahapatra S, Kulakova A, Streicher W, Peters GHJ, Nørgaard A, et al. Combination of high throughput and structural screening to assess protein stability—A screening perspective. *Eur J Pharm Biopharm* **171**, 1–10 (2022).

64. Kikhney AG, Svergun DI. A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Lett* **589**, 2570–7 (2015).
65. Franke D, Svergun DI. DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering. *J Appl Crystallogr* **42**, 342–6 (2009).
66. Izadi S, Patapoff TW, Walters BT. Multiscale coarse-grained approach to investigate self-association of antibodies. *Biophys J* **118**, 2741–54 (2020).
67. Mahapatra S, Polimeni M, Gentiluomo L, Roessner D, Frieß W, Peters GHJ, et al. Self-Interactions of Two Monoclonal Antibodies: Small-Angle X-ray Scattering, Light Scattering, and Coarse-Grained Modeling. *Mol Pharm* **19** (2), 508-519 (2021).

Chapter 5 - Interaction of HEPES with the therapeutic protein G-CSF **– A combination of molecular dynamics simulations and biophysical characterization**

Berner, C.¹, Menzen, T.², Winter, G.¹

¹Department of Pharmacy, Ludwig-Maximilians-Universität München, Munich, Germany

²Coriolis Pharma Research GmbH, Martinsried, Germany

Author contributions:

CB., T.M. and G.W. conceived the idea. C.B. performed the analyses, evaluated the data, and wrote the manuscript. G.W. and T.M. provided conceptual guidance. G.W. corrected the manuscript.

5.1 Introduction

Maintaining a drug product's stability is essential for the effective development of therapeutic proteins since reduced stability in many cases leads to a loss of efficacy and potency as well as increased immunogenicity.¹ Therapeutic proteins are prone to changes in their secondary and tertiary structure (physical instabilities such as unfolding and aggregation) as well as alterations in their primary structure by e.g. oxidation, deamidation, disulfide scrambling, and proteolysis.² The risk of such changes is enhanced when the environmental conditions differ from the physiological conditions and when the proteins are exposed to various external stresses in their lifetime until final administration.³ The identification of physicochemical properties and possible degradation mechanisms of the protein is crucial to find a stable formulation and to prevent problems with stability.⁴ The inherent chemical and physical instabilities of the protein are addressed by optimizing the solution conditions (i.e. pH, ionic strength) and identifying stabilizing excipients during drug product development. To promote conformational stability, a variety of excipients, including specific ligands, amino acids, buffers, surfactants, and sugars can be used.⁵ High concentrations of excipients are usually applied when the excipient is preferentially excluded from the protein, whereas excipients that preferentially bind to the protein's native state are applied in lower concentrations.^{6,7} Therefore, excipients in protein formulations must be optimized based on their concentration-dependent effects which could be both, positive and negative.⁸

This study focuses on the stabilization of the protein granulocyte-colony stimulating factor (G-CSF) by addition of the zwitterionic buffer HEPES. G-CSF is a hematopoietic growth factor containing 174 amino acid residues, that promotes the proliferation and differentiation of blood progenitor cells and the activation of neutrophils.^{9,10} Filgrastim (hereafter referred to as G-CSF), the non-glycosylated recombinant form of G-CSF, contains an additional N-terminal methionine and is a licensed drug to treat and prevent neutropenia.¹¹ G-CSF is characterized by a four-helix-bundle fold and an additional short 3_{10} -helix located in loopAB, perpendicular to the four antiparallel α -helices (Figure 29).

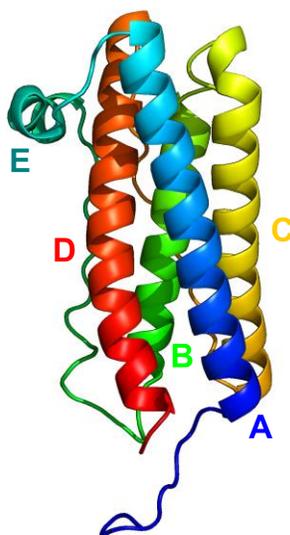


Figure 29: Structure of G-CSF (PDB code: 1CD9). The first four disordered residues which were missing in the PDB file were generated with MODELLER.

It has been shown that hydrophobicity and low solubility are stability concerns, and that methionine oxidation is an important degradation pathway of G-CSF.¹² G-CSF conformational stability and aggregation have been studied extensively. Its conformational stability is maximal at pH 4.0, while the protein self-associates as dimers and rapidly forms aggregates at neutral pH.^{13,14} High ionic strength neutralizes the charge-charge repulsions and induces aggregation also at low pH.^{15,16} Despite challenges in engineering to stabilize G-CSF, Bishop et al. identified variants with enhanced conformational stability and sustained biological activity.¹⁷ The design of an aggregation-resistant variant, however, remained unsuccessful.¹⁸ Interestingly, the presence of the buffer substance 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES) in high concentrations was shown to sustain the *in vivo* and *in vitro* activity of bovine G-CSF.¹⁹ HEPES is a zwitterionic molecule that provides buffer capacity between pH 2.5 to 3.5 and pH 6.8 to 8.2. Despite being widely used in protein expression, HEPES is not frequently found in therapeutic formulations.^{20,21}

The aim of this study was not the development of a new formulation that is more stable than those on the market, but the exemplary application of molecular dynamics (MD) simulations to explain experimental observations, here the previously reported effect of HEPES on the stability of G-CSF. This in turn can help with future experimental planning, to get new ideas for formulations, and to dramatically reduce classic excipient screenings. The molecular mechanism of the effect of a well-

known organic buffer molecule on a well-understood, relevant protein is investigated in atomistic detail at varying concentrations from 1 to 1000 mM and at two different pH values (pH 4.0 and pH 7.5). We compare experimental data obtained from near-UV and far-UV CD spectroscopy, nanoDSF and DLS measurements, RP-HPLC measurements after short-term stresses, and BLI binding studies with MD simulations at increasing concentrations of HEPES. We could identify an effect of HEPES on the G-CSF stability at both pH values with different mechanisms at low and high concentrations of HEPES and under different stress conditions. HEPES affected not only the protein conformation but also the aggregation propensity of the protein. We confirmed direct binding of HEPES to specific interaction sites at low concentrations and a crowding of HEPES molecules on the protein surface at high concentrations.

5.2 Material and methods

5.2.1 Materials

The bulk G-CSF solution contained 4.0 g/L protein in acetate buffer at pH 4.0. The protein concentration was measured spectrophotometrically using a NanoDrop 2000 (Thermo Fisher Scientific, Wilmington, USA) and an extinction coefficient at 280 nm of $0.86 \text{ (mg/mL)}^{-1}\text{cm}^{-1}$. All chemicals were of molecular biology or multicompendial grade and were purchased either from Sigma or Thermo Fisher Scientific (Germany). All solutions were prepared with ultrapure water from a Sartorius arium[®] pro system (Sartorius Corporate Administration GmbH, Göttingen, Germany).

5.2.2 Sample Dialysis and Preparation

Slide-A-Lyzer[™] MINI Dialysis Devices (cutoff 3.5 kDa, Thermo Fisher Scientific, Germany) were used for direct dialysis into the respective HEPES buffers for 24 hours at 2-8 °C. The respective HEPES concentrations were added to 10 mM sodium acetate buffer and subsequently adjusted to pH 4.0 ± 0.1 for the low pH samples whereas for pH 7.5 no additional buffer system was used. During dialysis, the samples with high HEPES concentrations were significantly upconcentrated, which led to precipitation of the protein. Therefore, in the next step the respective HEPES buffers at pH 4.0 were spiked into the G-CSF samples before dialysis for all conditions to avoid too high concentration gradients across the dialysis membrane. Afterwards, the samples were collected in microcentrifuge tubes and centrifuged at $10,000 \times g$ for 10 minutes and subsequently filtered with 0.02 μm Anotop[®] membrane filters (Whatman, FP 30/0.2 CA-S, GE Healthcare,

Buckinghamshire, UK). The pH after preparation was ± 0.1 of the target value. It must be noted that, in contrast to all marketed formulations of G-CSF, no polysorbate was used in this study which led to a faster degradation of the protein.

5.2.3 Circular Dichroism (CD) spectroscopy

Near- and far-UV circular dichroic spectra were collected at 25 °C with a Jasco J-810 spectropolarimeter (JASCO Deutschland GmbH, Pfungstadt, Germany). All samples contained 1g/L of protein. Quartz cuvettes (Hellma GmbH, Muellheim, Germany) with 10 mm and 0.1 mm wavelength path were used for the measurements, respectively. 5 accumulations of each sample were taken at a speed of 20 nm/min. The spectrum of the respective buffer was subtracted for each sample and smoothing of the spectra was performed using the Savitzky-Golay algorithm with 9 smoothing points. The mean residue ellipticity (MRE) of the protein at each wavelength was calculated as described elsewhere.²⁷ Far-UV spectra could only be recorded for samples with HEPES concentrations up to 50 mM due to the signal of the HEPES molecule itself.

5.2.4 Differential Scanning Fluorimetry (nanoDSF)

nanoDSF was used to study the thermal unfolding and aggregation of G-CSF as a function of the HEPES concentration. Samples with 1 g/L of protein were filled in standard nanoDSF™ grade capillaries, and the capillaries were sealed. A temperature ramp of 1 °C/min from 20 to 100 °C was applied with the Prometheus NT.48 (NanoTemper Technologies, Munich, Germany) system that measures the intrinsic protein fluorescence intensity at 330 and 350 nm after excitation at 280 nm. Simultaneously, the device detects aggregation/precipitation of the samples by measuring the back-reflection intensity of a light beam that passes through the capillary. The apparent protein melting temperatures (T_m) were determined with the PR. ThermControl software V2.1 (NanoTemper Technologies, Munich, Germany) from the maximum of the first derivatives of the thermal unfolding curves. The same software was used to determine the aggregation onset temperatures (T_{agg}) from the increase in the signal from the aggregation detection optics.

5.2.5 Dynamic Light Scattering (DLS)

DLS was used to determine the hydrodynamic radius (R_h) and the %Polydispersity (%PD) of the G-CSF samples. 10 μ L of each sample were pipetted in triplicates into a 1536 well plate (Aurora Microplates, Whitefish, USA). The plate was centrifuged at 2000 rpm for 2 min using a Heraeus

Megafuge 40 centrifuge equipped with an M-20 well plate rotor (Thermo Fisher Scientific, Wilmington, USA). Two microliter of silicon oil was added to seal each well. The plate was centrifuged again and placed in a DynaPro DLS plate reader III (Wyatt Technology, Santa Barbara, USA). All measurements were performed at 25 °C with 10 acquisitions per well and an acquisition time of 5 s. The data was analyzed with the Dynamics V7.10 software (Wyatt Technology, Santa Barbara, USA).

5.2.6 Reversed-Phase High-Performance Liquid Chromatography (RP-HPLC)

A Dionex Summit 2 system equipped with a UVD170U UV/Vis detector (Thermo Fisher, Dreieich, Germany) was used for the reversed-phase high-performance liquid chromatography. 20 μ L were injected on a BioBasic C18, 250 x 2.1, 5 μ m column (Thermo Fisher, Dreieich, Germany) after centrifugation at 10000 x g for 10 min. Detection was performed by UV spectrometry at 280 nm. Eluent A consisted of 10 % (w/v) acetonitrile and 0.1 % (w/v) trifluoroacetic acid in ultrapure water. Eluent B consisted of 0.1 % (w/v) trifluoroacetic acid in acetonitrile. The flow rate was 0.2 mL/min. The column oven temperature was set to 37 °C.

5.2.7 Molecular dynamics simulations

The structure of G-CSF is available in the PDB (1CD9) with the first four disordered residues missing. These were added with MODELLER. The protonation states of ionizable residues of the protein at the respective pH were adjusted using the H++ server.²⁸ The three protonation states of HEPES were parametrized with antechamber using GAFF2 for bonded and non-bonded parameters (Figure 30). Atomic partial charges were calculated with the AM1-BCC charge model in antechamber.

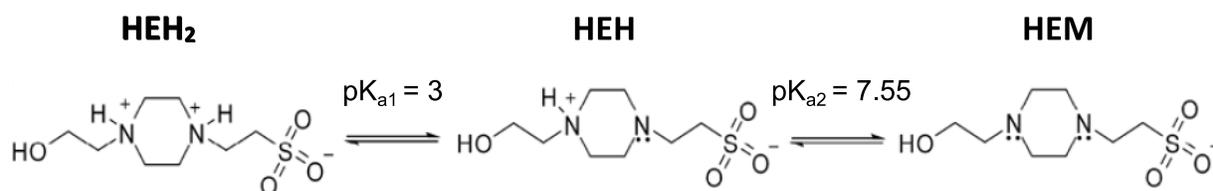


Figure 30: Structures of the different HEPES protonation states.

All-atom simulations were performed with the Amber18 program²⁵ in a periodic box with explicit solvent. The ff14SB force field for proteins was employed in combination with the TIP3P water model. PACKMOL and tleap were used to solvate G-CSF at pH 4.0 and 7.5 with the respective

number of HEPES molecules in a cubic water box. The numbers of unprotonated (HEM), protonated (HEH), and double protonated HEPES (HEH₂) were calculated via the Henderson-Hasselbalch equation for each condition. All bonds involving hydrogen atoms were constrained using the SHAKE algorithm. Non-bonded electrostatic interactions were treated using the particle mesh Ewald algorithm with a direct space cut-off of 8 Å. The system was energy minimized with the steepest descent algorithm for the first 5000 cycles, followed by 5000 cycles using the conjugate gradient method. System equilibration was carried out for 1 ns in NVT ensemble to stabilize the temperature of 300 K using the Langevin thermostat, and subsequently for 1 ns in NPT ensemble to adjust the density of the system using the Berendsen barostat. The simulations were performed for 100 ns with a time step of 2 fs. The coordinates were saved every 10 ps. All trajectories were analyzed using the CPPTRAJ module of Amber19 and VMD.²⁹ To estimate the binding of HEPES molecules to the protein, we calculated the interaction score probability ($P(I_{\text{score}})$) defined as:

$$P(I_{\text{score}}) = n / (N \cdot a)$$

where n is the number of frames HEPES was in contact with a certain residue over simulation time, N is the total number of frames, and a is a normalization factor to convert the interaction score into a probability.

5.3 Results and discussion

5.3.1 HEPES influences thermal stability and aggregation of G-CSF in a concentration-dependent manner

Previous work revealed that HEPES significantly affects the stability of G-CSF when used as a buffer component.¹⁹ We therefore investigated the effect of varying HEPES concentrations on the protein with experimental and computational techniques. Since G-CSF is a highly pH-sensitive protein which shows good stability at pH 4.0 but tends to aggregate at neutral pH, we include both pH values in this study to evaluate if HEPES could increase the stability not only at neutral pH but potentially even further enhance the protein stability at acidic conditions. HEPES only provides buffer capacity between pH 2.5 to 3.5 and pH 6.8 to 8.2 so that samples at pH 4.0 additionally contained 10 mM sodium acetate buffer.

First, we investigated the effect of HEPES at the two pH values on the stability of G-CSF. The thermal stability at different HEPES concentrations was measured with nanoDSF from 20 °C – 90 °C. Figure 31a shows the calculated melting temperatures (T_m) and aggregation onset temperatures (T_{agg}) for samples at pH 4.0 and 7.5 with 0, 1, 5, 10, 20, 50, 100, 500, and 1000 mM HEPES. Interestingly, even the low HEPES concentrations increase the melting temperature (T_m) compared to the reference sample without excipient for both pH values. At pH 4.0 the highest thermal stability was found at 10 mM HEPES ($T_m = 69.4$ °C) while increasing HEPES concentrations lead to a reduction in T_m even below the T_m of the reference without excipient. At 500 and 1000 mM HEPES, aggregation started to occur in the pH 4.0 samples. At pH 7.5, the melting temperature increases from ~51 °C without excipient to ~54 °C from 10 to 1000 mM. As could be seen for pH 4.0, the T_{agg} significantly decreases at high HEPES concentrations. These results indeed indicate a stabilizing effect of HEPES on G-CSF as was previously shown by Kasraian et al. but at lower concentrations. Furthermore, we performed dynamic light scattering (DLS) measurements on all samples at 25 °C to determine the hydrodynamic radius (R_h) (Figure 31b). The R_h of G-CSF at pH 4.0 is slightly smaller for all measured conditions compared to pH 7.5. A small drop in R_h can be seen upon the addition of HEPES compared to the reference for both pH values, whereas for the high HEPES concentrations a significant increase of R_h is observable. At 1000 mM HEPES, the R_h has even doubled compared to the references. This agrees with our nanoDSF measurements which indicate a higher aggregation propensity for the samples with high HEPES concentrations.

On the other hand, the increase in R_h could also be due to significant binding of HEPES molecules to the G-CSF monomer. To further evaluate if HEPES stabilizes G-CSF against various stresses, we performed short-term forced degradation studies on G-CSF at both pH values with increasing HEPES concentrations. The samples were subjected to temperature stress at 25 and 40 °C as well as shaking stress at 25 °C and 1400 rpm for 24 and 48 hours. All samples were analyzed with RP-HPLC, and the monomer recovery compared to the unstressed samples was calculated (Figure 31c, d; Figure S14). Note, that after 24 hours at pH 4.0 no severe degradation could be observed so that here samples after 48 hours are shown. After 24 hours at pH 7.5 degradation was already severe so that the results after the shorter stress time are presented. For reference, the results for pH 4.0 after 24 hours can be found in Figure S14a. At 25 °C only the sample without HEPES at pH 4.0 shows a decreased monomer recovery of about 86% while all other samples have a recovery of about 100%. Shaking stress however has a strong effect on the monomer recovery, especially at pH 7.5. At pH 4.0, again the sample without HEPES has a slightly lower recovery (83%), while samples with 1 to 50 mM still show a recovery >90%. As could already be seen in the nanoDSF measurements, the stability of G-CSF drops severely at HEPES concentrations of 500 and 1000 mM, where only 37% and 1% monomer were left after the stress, respectively. For pH 7.5, the highest monomer recovery was found between 5 and 20 mM HEPES (~30%) while the sample without HEPES shows only ~9% recovery and samples with 100, 500 and 1000 mM HEPES have less than 1% monomer left. These findings support the stabilizing effect of HEPES on G-CSF. The results for the 40 °C samples are however not as clear (Figure S14b). Samples at pH 4.0 after 48 hours at 40 °C remain stable, except for the sample without HEPES (80% recovery). However, the trend cannot be found at pH 7.5. Here, we first observe an increase in recovery with increasing HEPES concentration from 51% without HEPES to 75% recovery with 5 mM HEPES but a subsequent decrease to 19% for 100 mM HEPES. Surprisingly, the recovery drastically improves with 500 and 1000 mM HEPES at pH 7.5 (82 and 94%). This contrasts with our previous results but supports the findings of Kasraian et al. Altogether, G-CSF shows very different behavior at low (1 – 10 mM), intermediate (20 – 100 mM), and high (500 – 1000 mM) HEPES concentrations as well as in response to temperature or shaking stress. We propose that the different concentration-dependent effects of HEPES are due to direct interactions with the protein surface and thus potential changes in the conformation or electrostatics of the protein. Since HEPES is a zwitterionic molecule, it can both interact with positively and negatively charged residues on the protein surface.

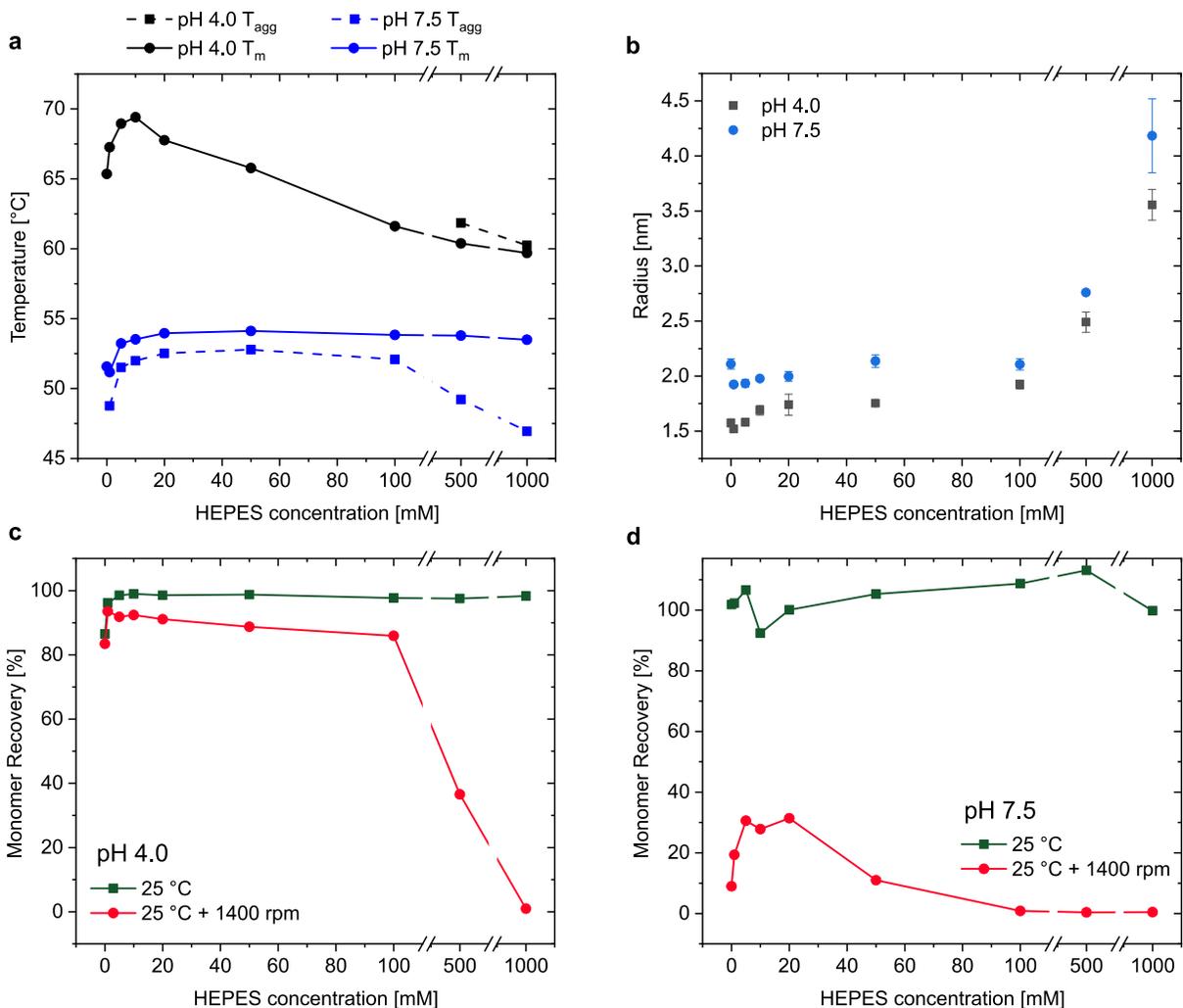


Figure 31: Effect of different HEPES concentrations on the stability of G-CSF at pH 4.0 and 7.5. (a) Melting temperatures (T_m) and aggregation onset temperatures (T_{agg}) determined by nanoDSF at varying HEPES concentrations. **(b)** Hydrodynamic radii of G-CSF at the two pH values with increasing HEPES concentrations measured by dynamic light scattering (DLS). **(c)** Monomer recovery (%) measured with RP-HPLC after 48 hours at 25 °C and shaking stress at 25 °C with 1400 rpm at pH 4.0. **(d)** Monomer recovery (%) measured with RP-HPLC after 24 hours at 25 °C and shaking stress at 25 °C with 1400 rpm at pH 7.5.

To investigate the effect on a molecular basis, we performed molecular dynamics (MD) simulations of G-CSF at pH 4.0 and pH 7.5 without and with increasing HEPES concentrations (10, 50, 100, 500, 1000 mM). Since MD simulations offer molecular mechanistic insights into conformational dynamics, protein-protein interactions, and protein stabilization by excipients, they have found broad application as a supplement to experiments.²²⁻²⁴ Here, we added the corresponding number of the different HEPES protonation states (double protonated: HEH_2 , protonated: HEH ,

deprotonated: HEM; see Figure 30) at the respective pH and concentration to the simulation box and performed 100 ns simulations at room temperature with the Amber2018 molecular dynamics software.²⁵ Due to limitations in box size, we were not able to simulate concentrations of HEPES lower than 10 mM.

In a first step, we analyzed the simulations visually and could observe that the structure of G-CSF was retained during the simulation and that HEPES molecules directly interacted with the protein. Since the direct binding could influence the stability, aggregation propensity and structure of the protein, we furthermore investigated the effects in more detail. The potential influence of HEPES on the protein's aggregation propensity has been assessed via the solvent accessible surface area (SASA) for G-CSF in presence and absence of HEPES (Figure 32). As can be clearly seen, the SASA decreases at high HEPES concentrations for both pH values due to binding of HEPES molecules to the protein. These results suggest a different mechanism of action for HEPES concentrations below and above 100 mM as it is also reflected in the nanoDSF and shaking stress results.

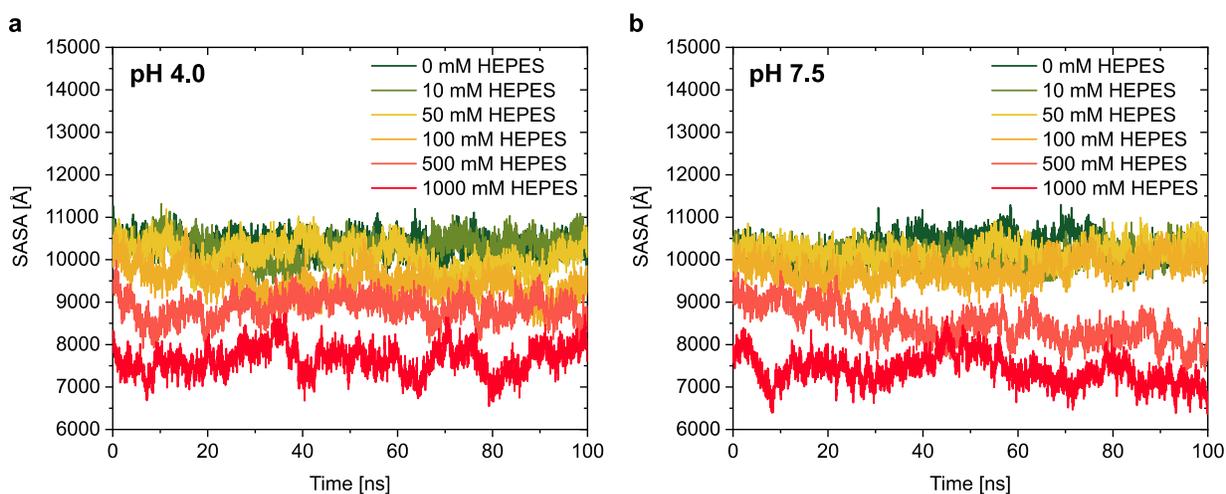


Figure 32: Solvent-accessible surface area (SASA) calculated from the MD simulations. (a) SASA for G-CSF at pH 4.0 without HEPES and with HEPES concentrations between 10 and 1000 mM. **(b)** SASA for simulations at pH 7.5.

5.3.2 HEPES induces minor structural changes in G-CSF

To further elucidate on this phenomenon, the impact of HEPES on the structure of G-CSF was determined via CD spectroscopy (Figure 33). The characteristic near-UV CD spectrum is not affected by HEPES at any concentration compared to the reference sample without HEPES at pH

4.0. However, at pH 7.5 the abnormal signal until 320 nm without excipient disappears upon the addition of HEPES. In our previous work we discuss that the reason for this signal could be self-interaction or aggregation of G-CSF.¹⁶ This finding suggests that HEPES reduces this self-interaction at pH 7.5 even at concentrations of 1 mM. Nevertheless, the shape of the spectrum remains the same for all HEPES concentrations showing that the overall tertiary structure of G-CSF is not affected by HEPES. The far-UV CD spectra could only be determined up to 50 mM HEPES, due to signal interference at higher concentrations. Two typical minima at 209 and 222 nm confirm that G-CSF is α -helical in all conditions. A slight increase of helical content when increasing the HEPES concentration was observable at pH 7.5. The spectra at pH 4.0 are all comparable. We can therefore conclude that HEPES induces minor structural changes in G-CSF at pH 7.5 but not at pH 4.0.

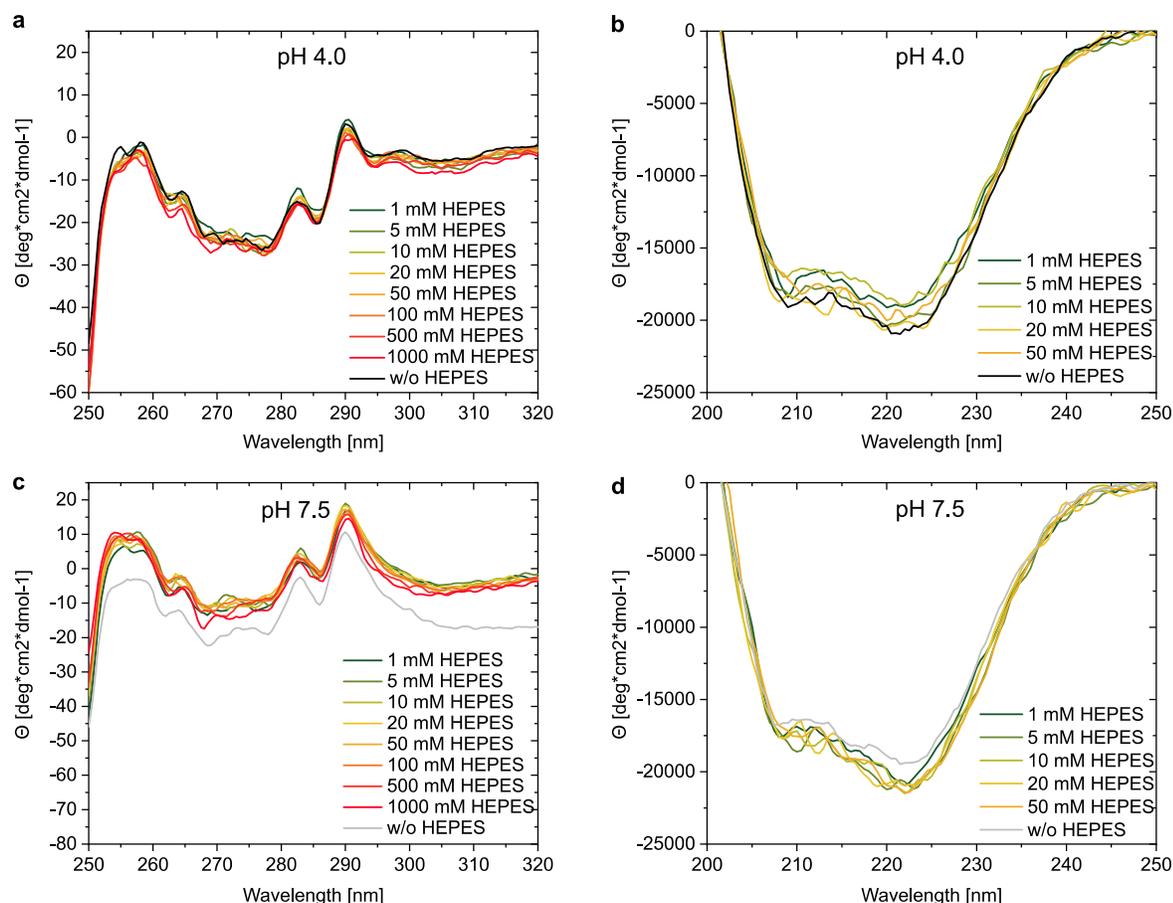


Figure 33: Influence of HEPES on the secondary and tertiary structure of G-CSF measured by CD spectroscopy. (a,c) Near-UV CD spectra of G-CSF at varying HEPES concentrations at pH 4.0 and pH 7.5, respectively. (b,d) Influence of HEPES on the far-UV CD spectra of G-CSF at pH 4.0 and pH 7.5, respectively.

From visually inspecting the MD simulations, we could further observe an influence of HEPES binding on helix E (residues 45-55) of G-CSF at pH 7.5. Therefore, the secondary structure was assigned for these residues for each condition at both pH values from the MD simulations (Figure 34). As expected, no changes were found for pH 4.0 where most of the residues in the region 45-55 form either α -helical or 3_{10} -helical structure. The same region at pH 7.5 without any excipient on the other hand shows a significant amount of turn and bend, and less α -helical structure. Upon addition of HEPES, the turn and bend structure partly decrease and forms an α -helix leading to a similar distribution as at pH 4.0. Since also the far-UV CD measurements at pH 7.5 showed the slight increase in helical context, we can conclude that HEPES binding induced the formation of helical content of residues 45-55, corresponding to helix E.

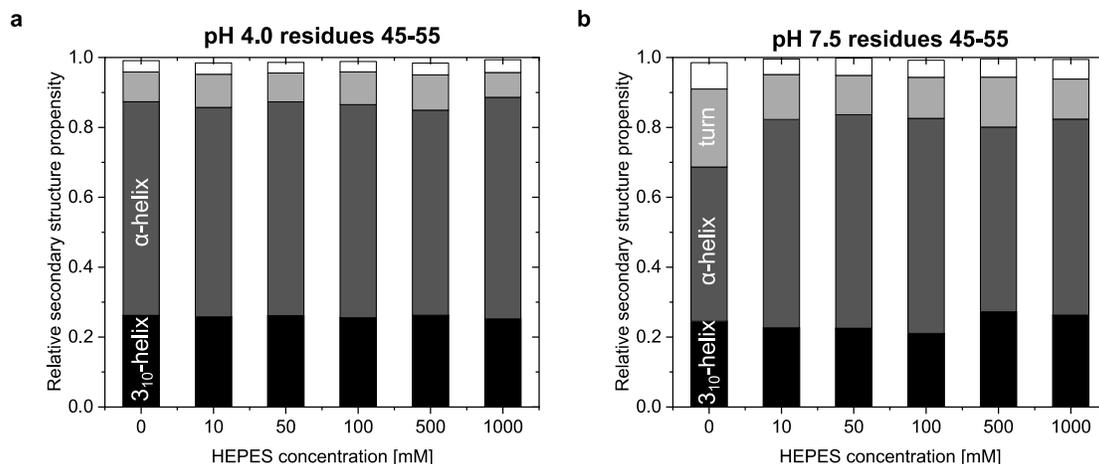


Figure 34: Secondary structure assignment of residues 45-55 of G-CSF from MD simulations with different HEPES concentrations (3₁₀-helix: black, α-helix: dark gray, turn: light gray, bend: white). **(a)** Relative secondary structure propensity at pH 4.0. **(b)** Relative secondary structure propensity at pH 7.5.

5.3.3 HEPES directly binds to G-CSF in MD simulations

In the last step, we wanted to understand the mechanisms by which HEPES causes either stabilization or destabilization of G-CSG at pH 4.0 and pH 7.5 at low, intermediate, and high concentrations. HEPES contains two hydrogen bond donor and six hydrogen bond acceptor sites and can therefore have contact with multiple residues at once.²⁰ The negatively charged sulfonic group ($-\text{SO}_3^-$) and the positively charged amino group ($-\text{NH}_3^+$) in the piperazine ring in HEH and HEH₂ can participate in salt bridges with positive or negative residues on the protein surface, respectively.²⁰ Additionally, the large size of the HEPES molecule enables it to bridge gaps between residues that are far apart in sequence but also in the structure, for example on two different helices. This mechanism was suggested by Kasraian et al. in their study of HEPES and G-CSF. We evaluated the binding of the HEPES molecules to the protein at different conditions in MD simulations. Trajectories were analyzed in terms of binding probabilities ($P(I_{\text{score}})$) of the different HEPES species to the individual residues of the protein at pH 4.0 and pH 7.5 (Figure 35a-d; Figure S15). The $P(I_{\text{score}})$ is a measure of how long a contact between a HEPES molecule and a residue of G-CSF was present during the simulation and is calculated based on a simple distance cut-off. We used a $P(I_{\text{score}})$ cut-off of 20 % to distinguish transient contacts from actual binding. It must be noted that the binding frequency and strength of binding for the three species HEH₂, HEH and HEM cannot directly be compared between the two pH values since they are included to the simulation boxes at different ratios. However, at pH 7.5 HEH and HEM are present with a 1:1 ratio.

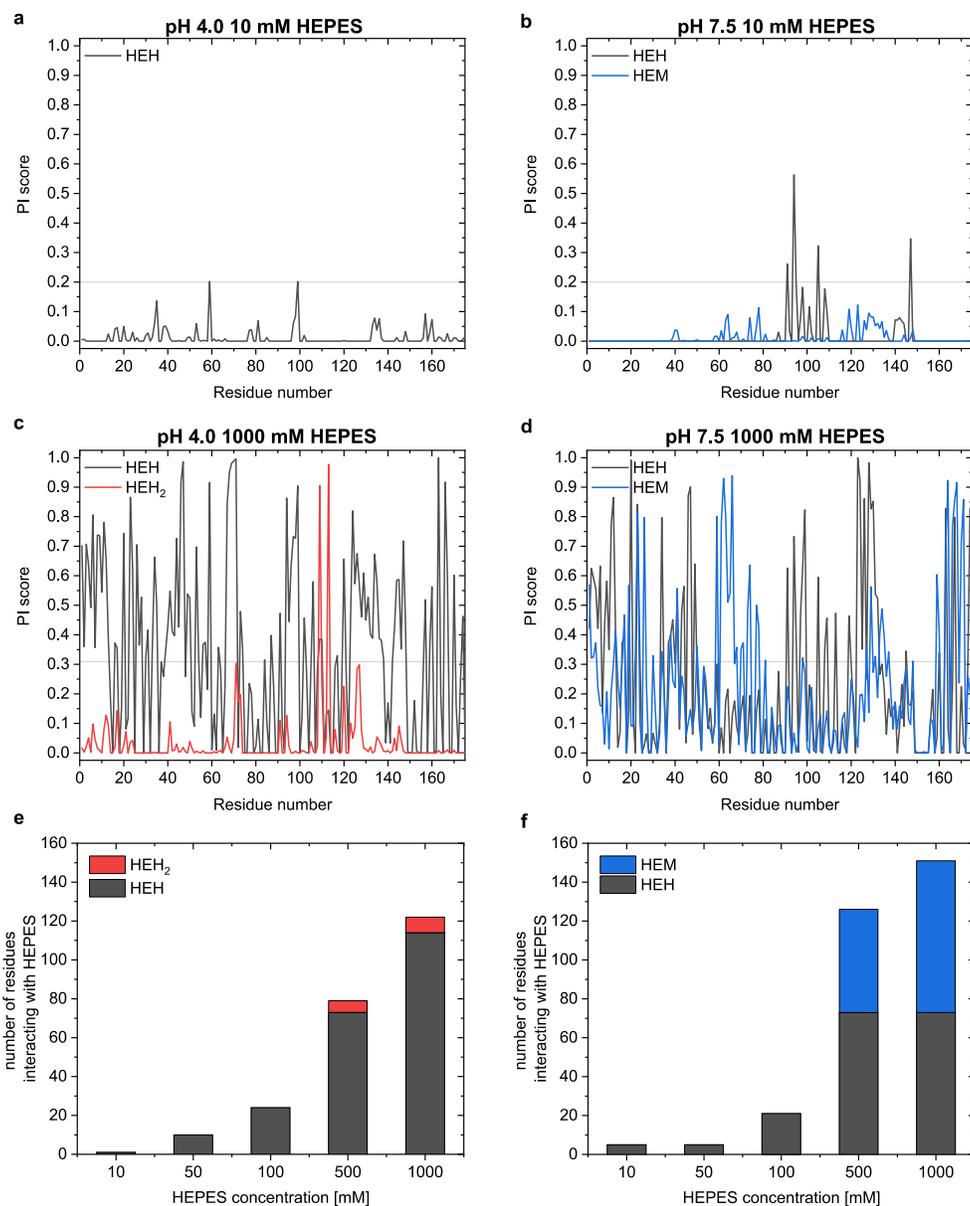


Figure 35: Binding of HEPES to G-CSF in MD simulations at pH 4.0 and pH 7.5. (a, b) Interaction score probabilities $P(I_{\text{score}})$ per residue for HEH, HEH₂, and HEM with 10 mM HEPES for pH 4.0 and 7.5, respectively. The $P(I_{\text{score}})$ gives a percentage of how long an interaction of phosphate with each residue was present during the simulation. We used a $P(I_{\text{score}})$ cut-off of 0.2 to distinguish transient contacts from actual binding (indicated by a gray line). (c, d) $P(I_{\text{score}})$ for 1000 mM HEPES at pH 4.0 and 7.5, respectively. (e, f) The number of residues interacting with either HEH, HEH₂, or HEM for $\geq 20\%$ of the simulation show a concentration-dependence for pH 4.0 (e) and for pH 7.5 (f).

At 10 mM HEPES, the interactions of HEPES with the protein are rarely present and mostly not stable throughout the simulation. The $P(I_{\text{scores}})$ at pH 7.5 with 50 mM HEPES are unexpected in comparison to the other conditions, since HEH does seem to interact less with the protein compared to lower HEPES concentrations. Interestingly, the binding sites at 10 and 50 mM HEPES compared for pH 4.0 and pH 7.5 are similar; especially residues in the region around residue 100 ± 10 interact with HEH, corresponding to the C-terminal part of helix B, loopBC, and the N-terminal part of helix C. HEH₂ and HEM, however, have transient contacts with residues at completely different regions of the protein. Further increasing the HEPES concentration leads to an even higher level of HEH binding, especially at pH 7.5. We can observe a new interaction of HEPES between residues 40 and 60 (helix E) at pH 4.0 and residues 20 and 50 (half of helix A and helix E) for pH 7.5. Additionally, the N- and C-terminus show interaction with HEH at both pH values. Even at 100 mM HEPES, HEH₂ and HEM show no interactions above the 20 % threshold. At 500 and 1000 mM HEPES, HEH molecules seem to be equally distributed on the protein surface and no distinct interaction sites can be identified. However, at these higher concentrations, also HEH₂ and HEM show interaction probabilities exceeding the threshold. To allow for an easier comparison, the number of G-CSF residues interacting with HEPES above the 20 % threshold are counted and presented in Figure 35e and f. As already discussed, at low HEPES concentrations, only very few interactions of HEH are observed which increase at higher HEPES concentrations. Interestingly, even though the total number of HEPES molecules in the simulation boxes for each concentration are the same for both pH values, at high concentrations more HEPES molecules interact at pH 7.5 compared to pH 4.0. At pH 7.5 there seems to be a limit of HEH binding since the number does not increase from 500 to 1000 mM HEPES while the number of HEM on the protein surface still increases. To better visualize the HEPES binding sites, examples of low and high concentration conditions at pH 7.5 are illustrated in Figure 36. Noteworthy, the HEPES molecules bridge residues from different helices and even the N- and C-terminus. Furthermore, the interaction sites at the low concentrations correspond to known aggregation sites of G-CSF,^{16,18,26} which support an aggregation breaking effect of HEPES. High HEPES concentrations on the other hand, do not show distinct binding sites but rather a crowding of HEPES molecules on the surface which could either improve electrostatic repulsion between G-CSF monomers or disrupt the hydration shell around the protein molecules which could in turn expose hydrophobic regions of the molecule. Taha et al. could also observe a decrease in hydrogen bonds between HEPES and water molecules at high

buffer concentrations due to buffer-buffer interactions.²⁰ This would support a mechanism of water entrapment between the protein and the layer of HEPES molecules around it.

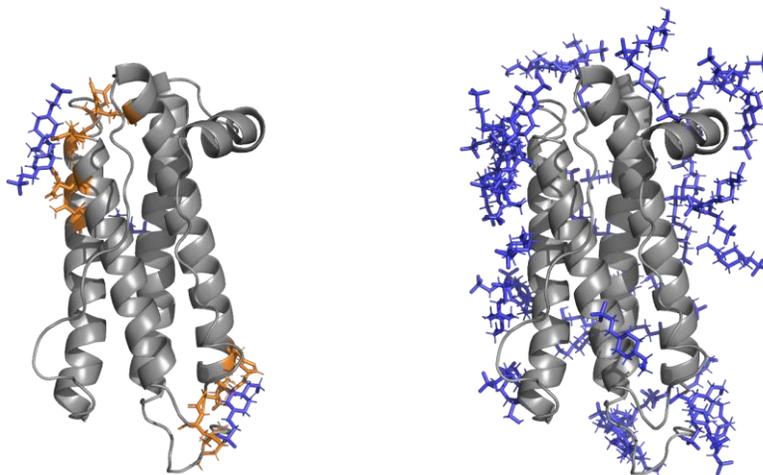


Figure 36: Binding sites of HEPES to G-CSF at pH 7.5 at low concentrations (left) and distribution of HEPES on the protein surface at high concentrations (right). HEPES molecules are colored in blue, the interaction sites are colored in orange.

The concentration-dependent binding of HEPES at pH 7.5 could also be confirmed in bio-layer interferometry (BLI) measurements up to 200 mM HEPES. The baseline aligned and double referenced sensorgram is presented in Figure S16. A very weak K_D of $51.1 \text{ mM} \pm 8.4 \text{ mM}$ was determined which is in agreement with our MD simulations where the binding was rarely stable throughout the whole simulation. However, it must be noted that the BLI-system has the capacity to measure weak affinities up to 1 mM with certainty. Nevertheless, the dose dependency of HEPES binding to G-CSF could be observed.

5.4 Conclusions

We present an extensive study on the stability of the G-CSF in the absence and presence of HEPES at pH 4.0 and pH 7.5 and elucidate the molecular mechanisms by which different HEPES concentrations influence the protein structure and stability. We revealed a stabilizing effect of HEPES on G-CSF in nanoDSF experiments and against shaking stress at low concentrations whereas high concentrations of HEPES (≥ 500 mM) resulted in aggregation of the protein. Interestingly, temperature-induced degradation of G-CSF at 40 °C could be prevented by HEPES at pH 4.0 at all concentrations whereas a different outcome could be observed at pH 7.5. Here, low HEPES concentrations stabilized the protein, while at intermediate HEPES concentrations the recovery dropped tremendously. However, at high HEPES concentrations, the highest recovery could be observed which supports the previously reported finding of Kasraian et al. We can conclude that there are different mechanisms of action for HEPES at low and high concentrations and at different stress conditions. Next, we could observe HEPES-induced helix formation of residues corresponding to helix E resulting in a more similar structure of G-CSF at pH 4.0 and 7.5. Since both, conformational and colloidal stability are superior at pH 4.0, this conformational change could be the reason for the stabilizing effect of HEPES at pH 7.5. This however does not explain the stabilization of G-CSF at pH 4.0. With MD simulations and BLI measurements we could confirm direct binding of HEPES to G-CSF in a concentration-dependent manner. Binding sites include residues of helix E as well as aggregation hot spots at low concentrations whereas an almost equal distribution of HEPES molecules on the protein surface could be found at high concentrations. In summary, HEPES binding not only influences the protein conformation at pH 7.5 but also bridges residues on different helices and blocks aggregation sites of the protein at both pH values. We can therefore conclude that HEPES indeed has a stabilizing effect on G-CSF and, depending on the concentration, can be used as an excipient in a G-CSF formulation. Our work suggests that HEPES is worth to be studied more as an excipient with a potential dual action as buffer and specific stabilizer. Additionally, we could show that the application of MD simulations should be further included in formulation development to make formulations even more rational than they are today.

5.5 Supporting Information

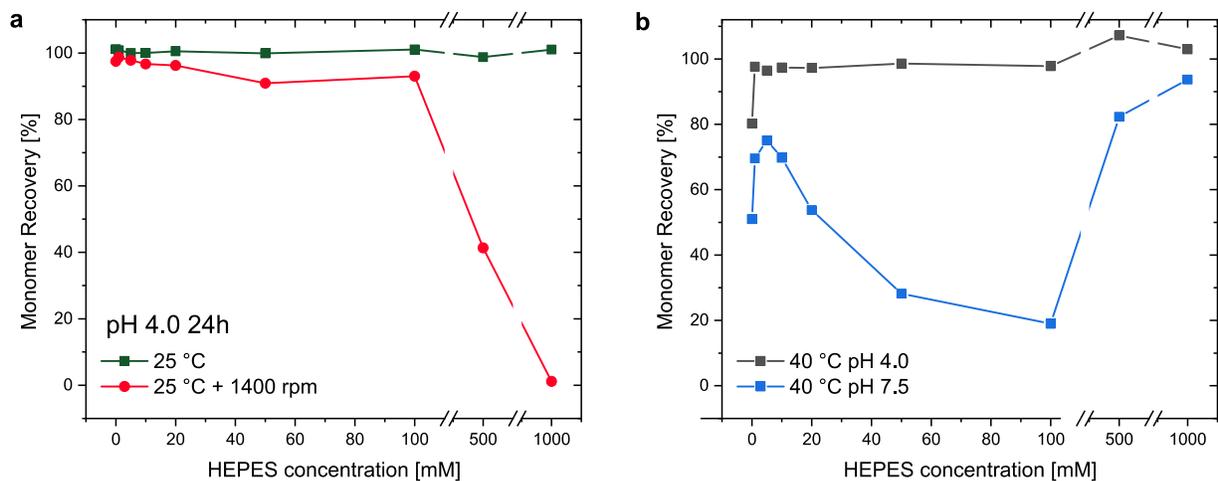


Figure S14: Effect of HEPES on the stability of G-CSF at pH 4.0 and 7.5. (a) Monomer recovery (%) measured with RP-HPLC after 24 hours at 25 °C and shaking stress at 25 °C with 1400 rpm at pH 4.0. **(b)** Monomer recovery (%) measured with RP-HPLC after 24 hours and 48 hours at pH 7.5 and pH 4.0, respectively, at 40 °C.

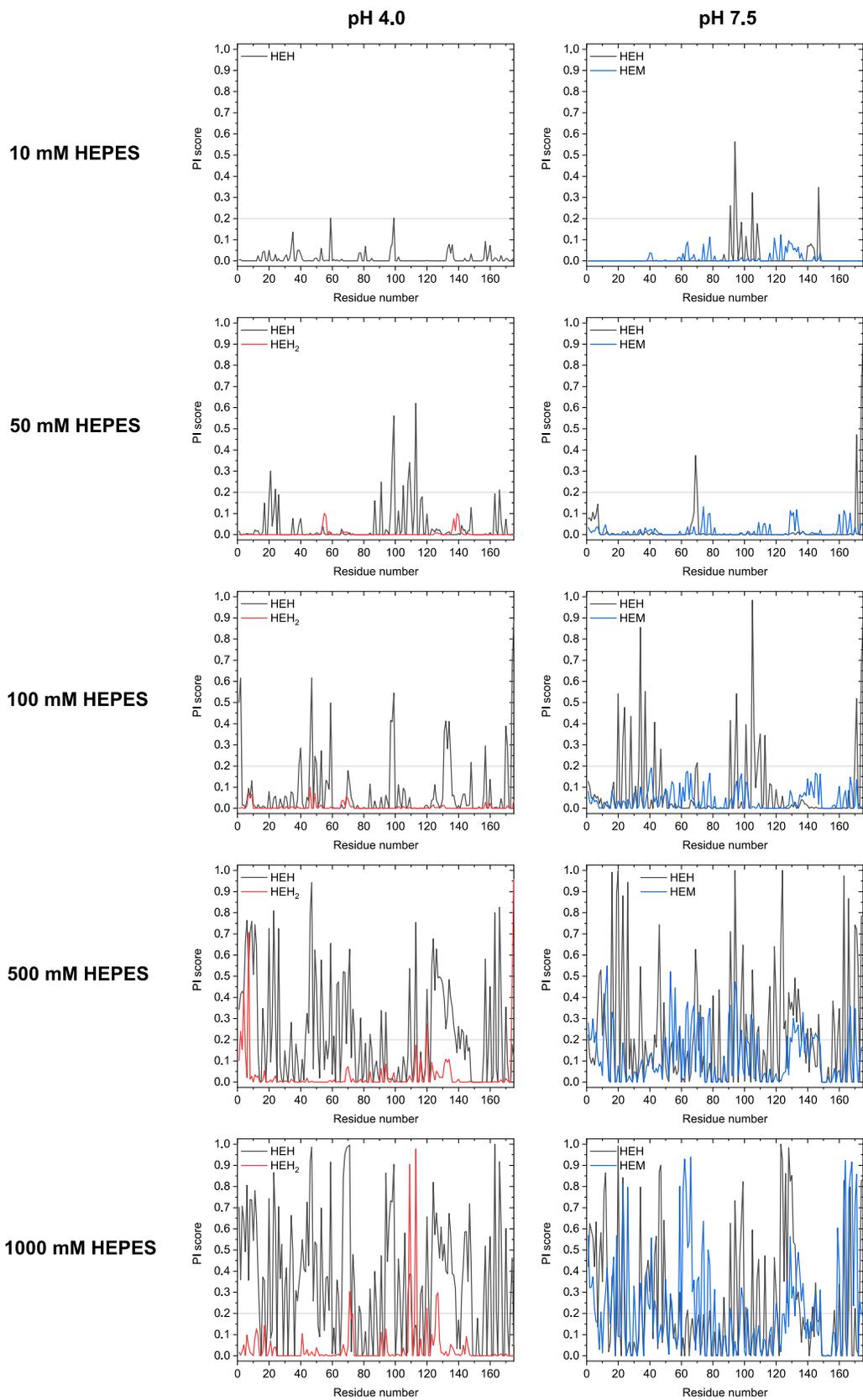


Figure S15: P(I_{scores}) of the different HEPES protonation states to G-CSF from 10 to 1000 mM HEPES.

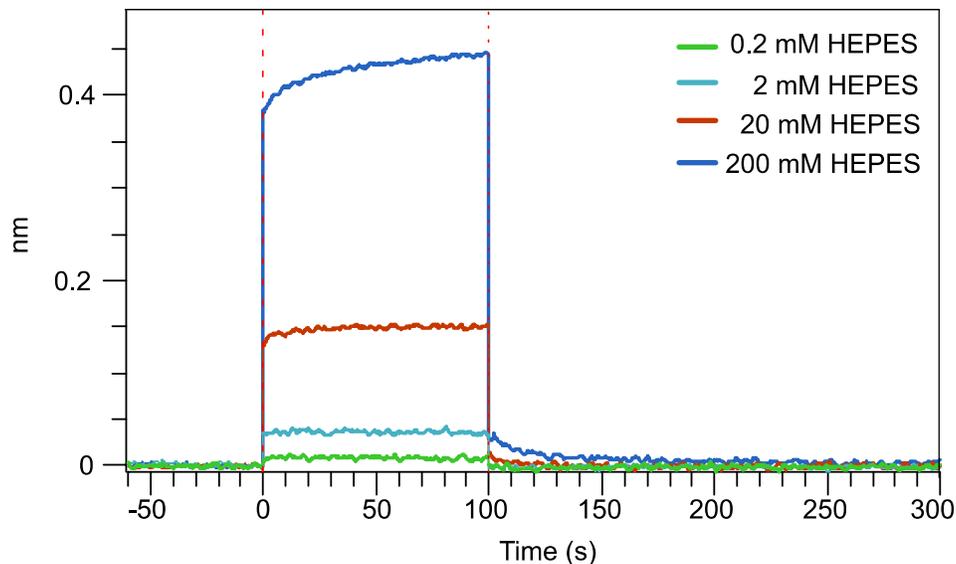


Figure S16: Baseline aligned and double referenced sensorgram of the bio-layer interferometry (BLI) measurements up to 200 mM HEPES.

Table S3: Overview of the experimental and computational parameters determined at pH 4.0.

| pH 4.0 HEPES [mM] | T_m [$^{\circ}\text{C}$] | T_{agg} [$^{\circ}\text{C}$] | R_h [nm] | Monomer Recovery [%] (48 h) | | | SASA [\AA] | # interacting HEPES molecules | |
|-------------------------|------------------------------|----------------------------------|------------|-----------------------------|----------------------------------|-----------------------|-----------------------|-------------------------------|-----|
| | | | | 25 $^{\circ}\text{C}$ | 25 $^{\circ}\text{C}$ + 1400 rpm | 40 $^{\circ}\text{C}$ | | HEH ₂ | HEH |
| 0 | 65.4 | -- | 1.6 | 86.5 | 83.4 | 80.2 | 10347 | -- | -- |
| 1 | 67.3 | -- | 1.5 | 96.2 | 93.5 | 97.6 | | | |
| 5 | 69.0 | -- | 1.6 | 98.5 | 91.8 | 96.5 | | | |
| 10 | 69.4 | -- | 1.7 | 99.0 | 92.4 | 97.4 | 10305 | -- | 1 |
| 20 | 67.8 | -- | 1.7 | 98.6 | 91.1 | 97.2 | | | |
| 50 | 65.8 | -- | 1.8 | 98.8 | 88.8 | 98.6 | 10101 | -- | 10 |
| 100 | 61.6 | -- | 1.9 | 97.7 | 85.9 | 97.8 | 9450 | -- | 24 |
| 500 | 60.4 | 61.9 | 2.5 | 97.5 | 36.5 | 107.2 | 8857 | 6 | 73 |
| 1000 | 59.7 | 60.2 | 3.6 | 98.3 | 1.0 | 103.0 | 7645 | 8 | 114 |

Table S4: Overview of the experimental and computational parameters determined at pH 7.5.

| pH 7.5 | | | | Monomer Recovery | | | | | # interacting HEPES molecules | |
|------------|---------------------|-----------------------|---------------------|------------------|------------------|-------|------------------------|----|-------------------------------|-----|
| | | | | [%] (24 h) | | | | | HEH | HEM |
| HEPES [mM] | T _m [°C] | T _{agg} [°C] | R _h [nm] | 25 °C | 25 °C + 1400 rpm | 40 °C | SASA [Å ²] | | | |
| 0 | 51.6 | -- | 2.1 | 101.8 | 9.0 | 51.0 | 10281 | -- | -- | |
| 1 | 51.2 | 48.8 | 1.9 | 102.3 | 19.4 | 69.6 | | | | |
| 5 | 53.2 | 51.5 | 1.9 | 106.6 | 30.6 | 75.1 | | | | |
| 10 | 53.5 | 52.0 | 2.0 | 92.4 | 27.8 | 69.8 | 10093 | 5 | -- | |
| 20 | 54.0 | 52.5 | 2.0 | 100.1 | 31.4 | 53.8 | | | | |
| 50 | 54.1 | 52.8 | 2.1 | 105.3 | 11.0 | 28.2 | 10153 | 5 | -- | |
| 100 | 53.8 | 52.1 | 2.1 | 108.7 | 0.9 | 19.0 | 9773 | 21 | -- | |
| 500 | 53.8 | 49.2 | 2.8 | 113.1 | 0.4 | 82.3 | 8466 | 73 | 53 | |
| 1000 | 53.5 | 47.0 | 4.2 | 99.8 | 0.5 | 93.7 | 7396 | 73 | 78 | |

5.6 References

1. Roberts, C. J. Therapeutic protein aggregation: Mechanisms, design, and control. *Trends Biotechnol.* **32**, 372–380 (2014).
2. Arakawa, T., Prestrelski, S. J., Kenney, W. C. & Carpenter, J. F. Factors affecting short term and long-term stabilities of proteins. *Adv. Drug Deliv. Rev.* **46**, 307–326 (2001).
3. Hawe, A. et al. Forced degradation of therapeutic proteins. *J. Pharm. Sci.* **101** (3), 895–913 (2012).
4. 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).
5. 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. *Advanced Drug Delivery Reviews* **63** (13), 1118–1159 (2011).
6. Timasheff, S. N. The Control of Protein Stability and Association by Weak Interactions with Water: How Do Solvents Affect These Processes? *Annual Review of Biophysics and Biomolecular Structure*, **22** (1), 67–97 (1993).
7. Zbacnik, T. J. et al. Role of Buffers in Protein Formulations. *J. Pharm. Sci.* **106**, 713–733 (2017).
8. Rani, A. & Venkatesu, P. Insights into the interactions between enzyme and co-solvents: Stability and activity of stem bromelain. *Int. J. Biol. Macromol.* **73**, 189–201 (2015).
9. Hill, C. P., Osslund, T. D. & Eisenberg, D. The structure of granulocyte-colony-stimulating factor and its relationship to other growth factors. *Proc. Natl. Acad. Sci.* **90**, 5167–5171 (1993).
10. Weiss, M. et al. Effects of exogenous recombinant human granulocyte colony-stimulating factor (filgrastim, rhG-CSF) on neutrophils of critically ill patients with systemic inflammatory response syndrome depend on endogenous G-CSF plasma concentrations on admission. *Intensive Care Med.* **29**, 904–914 (2003).
11. Welte, K., Gabrilove, J., Bronchud, M. H., Platzer, E. & Moratyn, G. Filgrastim (r-metHuG-CSF): the first 10 years. *Blood* **88**, 1907–1929 (1996).
12. Holzmann, J., Hausberger, A., Rupprechter, A. & Toll, H. Top-down MS for rapid methionine oxidation site assignment in filgrastim. *Anal. Bioanal. Chem.* **405**, 6667–6674 (2013).
13. Krishnan, S. et al. Aggregation of granulocyte colony stimulating factor under physiological conditions: Characterization and thermodynamic inhibition. *Biochemistry* **41**, 6422–6431 (2002).
14. Chi, E. Y. E. et al. Roles of conformational stability and colloidal stability in the aggregation of recombinant human granulocyte colony-stimulating factor. *Protein Sci.* **12**, 903–913 (2003).

15. Yamazaki, K., Murayama, K., Ishikawa, R. & Ozaki, Y. An infrared spectroscopy study of acid stability and thermal unfolding process of granulocyte-colony stimulating factor. *J. Biochem.* **137**, 265–271 (2005).
16. Ko, S. K. et al. Investigation of the pH-dependent aggregation mechanisms of GCSF using low resolution protein characterization techniques and advanced molecular dynamics simulations. *Comput. Struct. Biotechnol. J.* **20**, 1439–1455 (2022).
17. Bishop, B., Koay, D. C., Sartorelli, A. C. & Regan, L. Reengineering Granulocyte Colony-stimulating Factor for Enhanced Stability. *J. Biol. Chem.* **276**, 33465–33470 (2001).
18. Meric, G., Naik, S., Hunter, A. K., Robinson, A. S. & Roberts, C. J. Challenges for design of aggregation-resistant variants of granulocyte colony-stimulating factor. *Biophys. Chem.* **277**, (2021).
19. Kasraian, K. et al. Sustained in vivo activity of recombinant bovine granulocyte colony stimulating factor (rbG-CSF) using HEPES buffer. *Pharm. Dev. Technol.* **6**, 441–447 (2001).
20. Taha, M., Khoiroh, I. & Lee, M.-J. Phase Behavior and Molecular Dynamics Simulation Studies of New Aqueous Two-Phase Separation Systems Induced by HEPES Buffer. *J. Phys. Chem. B* **117**, 563–582 (2013).
21. Gupta, B. S., Taha, M. & Lee, M. J. Buffers more than buffering agent: Introducing a new class of stabilizers for the protein BSA. *Phys. Chem. Chem. Phys.* **17**, 1114–1133 (2015).
22. Salo-Ahen, O. M. H. et al. Molecular dynamics simulations in drug discovery and pharmaceutical development. *Processes* **9**, 1–63 (2021).
23. Prabakaran, R., Rawat, P., Thangakani, A. M., Kumar, S. & Gromiha, M. M. Protein aggregation: in silico algorithms and applications. *Biophys. Rev.* **13**, 71–89 (2021).
24. Indrakumar, S. et al. Development of a fast screening method for selecting excipients in formulations using MD simulations, NMR and microscale thermophoresis. *Eur. J. Pharm. Biopharm.* **158**, 11–20 (2021).
25. Case, D. A. Amber 18. Univ. California, San Fr. (2018).
26. Wood, V. E. et al. Protein Engineering and HDX Identify Structural Regions of G-CSF Critical to Its Stability and Aggregation. *Mol. Pharm.* **19**, 616–629 (2022).
27. Kelly, S. & Price, N. The Use of Circular Dichroism in the Investigation of Protein Structure and Function. *Curr. Protein Pept. Sci.* **1**, 349–384 (2000).
28. Gordon, J. C. et al. H⁺⁺: A server for estimating pK_as and adding missing hydrogens to macromolecules. *Nucleic Acids Res.* **33**, 368–371 (2005).
29. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38 (1996).

Chapter 6 - Stabilization of a novel recombinant bacteriophage endolysin by protein-excipient interaction with HEPES and other Good's buffers

Eisele, S., Berner, C., Winter, G.

Department of Pharmacy, Ludwig-Maximilians-Universität München, Munich, Germany

Author contributions:

S.E., C.B., and G.W. conceived the idea. S.E. performed the experimental work and evaluated the data. C.B. performed the in silico work, S.E. and C.B. wrote the manuscript. G.W. provided conceptual guidance and corrected the manuscript.

6.1. Introduction

The successful development of therapeutic proteins depends on the design of stable and robust formulations to sustain stability of the drug product during shelf life. Reduced protein stability can lead to a loss of efficacy and potency, and also to harmful effects like enhanced immunogenicity.¹ Therefore, potential chemical and physical protein instabilities are addressed by the adaptation of the formulation composition during drug product development.² Oxidation of proteins is often reduced by the addition of antioxidants, such as methionine.³ Deamidation is another typical degradation pathway of proteins and can often be reduced by pH adaptation.⁴ In addition, the pH has an impact on the protein colloidal stability by changing the charge distribution on the protein surface and thereby affecting the protein-protein interactions in solution.⁵ These can further be mitigated by the addition of salts, specifically anions of the Hofmeister salt series.^{6,7}

Furthermore, changes in the conformation or partial unfolding of the protein can lead to loss in activity and non-native aggregation.⁸ Various excipients like specific ligands, amino acids, buffers and osmolytes are commonly used to increase the conformational stability.⁷ The selection is usually limited to excipients that are well characterized to avoid further safety and efficacy study examinations.⁹ The excipients either stabilize the conformation of the protein by preferential exclusion or by preferential binding to the native state of the protein.^{2,10} Preferential exclusion is usually achieved with high concentrations of sugars, polyols or amino acids.⁵ In contrast, excipients that preferentially bind to the protein with a certain affinity are applied in lower concentrations.¹⁰ While protein-excipient binding can enhance protein stability when the interaction is weak and transient, stronger-interacting excipients were found to mostly destabilize proteins.¹¹ Moreover, the binding needs to be reversible to assure the proteins therapeutic efficacy after administration.

Since excipient effects are protein specific, finding the optimal excipients to achieve a stable drug product is generally done using a trial-and-error approach. However, there have been recent approaches to apply computational tools in the context of formulation design. A virtual screen has been presented that could identify new compounds which bind to predicted aggregation hotspots, thus inhibiting protein-protein interactions.¹² Furthermore, molecular dynamics (MD) simulations are now commonly used to study proteins in the presence of excipients to identify the local effects on the proteins. Thus, the mechanisms by which certain excipients, for example a dipeptide,¹³

arginine,¹⁴ arginine-salts,¹⁵ cyclodextrins,¹⁶ and surfactants¹⁷ stabilize proteins have been elucidated.

Whereas larger proteins like monoclonal antibodies and antibody-drug-conjugates (ADCs) are rather well characterized, the stabilization mechanisms for new, structurally variable proteins must be examined on an individual basis. Here, we present a combined study of experimental and computational approaches on the stabilization of the novel antibacterial drug candidate HY-133, a bacteriophage lysin which specifically targets various *S. aureus* strains. The drug substance is a recombinant chimeric lysin with two functional modules: an enzymatic active domain (EAD) and a cell wall binding domain (CBD) which are linked via a synthetic peptide linker (Figure 37). Analysis of the electrostatic and hydrophobic surface properties reveal negatively and positively charged surface patches in both domains and a large hydrophobic patch in the CB domain. These properties make HY-133 not only prone to intermolecular interactions but also to inter-domain interactions which brings additional challenges to the development. Both domains can be varied according to the drug substance's intended specificity. Topical administrations of HY-133 in the nasal cavity and on the skin is intended.

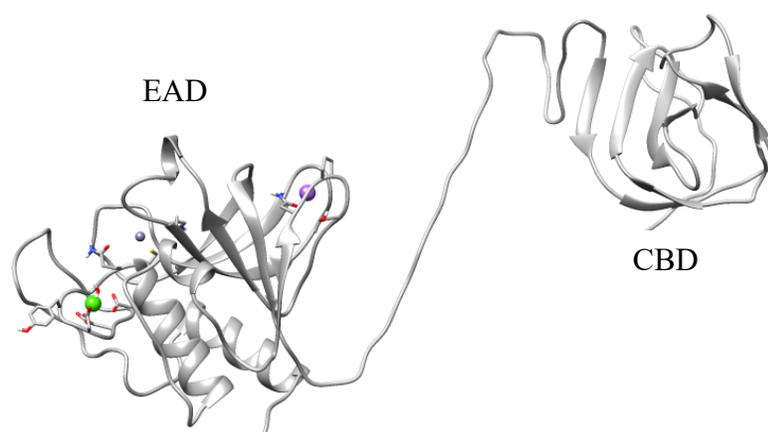


Figure 37: Homology model of HY-133 with the enzymatic active domain (EAD) from the endolysin of phage K (PDB code: 4CSH) and the cell wall binding domain (CBD) from lysostaphin (PDB code: 4LXC) with bound ions.

Early development phases revealed chemical degradation as a critical parameter of HY-133. However, a pH of 6.0 and methionine as antioxidant could maintain the protein's chemical stability in combination with a HEPES buffer. In addition, Poloxamer 188 effectively increased the colloidal

stability. We furthermore found that the buffer substance N-2-hydroxyethylpiperazine-N'-ethanesulfonic acid (HEPES) had an advantage over the commonly used buffer systems citrate and histidine at the same pH. HEPES is part of the Good's buffers family, which was developed in 1966 by Good et al. with the background of providing highly water soluble buffer systems in the physiological pH range.¹⁸ It belongs to the piperazinic subgroup and has wide structural similarities with piperazine-N,N'-bis(2-hydroxypropanesulfonic acid) (POPSO). Whereas 2-(N-morpholino)-ethanesulfonic acid (MES) and 3-(N-morpholino)propanesulfonic acid (MOPS) belong to the morpholinic subgroup, they all share the N-substituted aminosulfonic acid. The zwitterionic character of these substances provides adequate buffer capacities between pH 5.5 and 8.5.¹⁹ Interestingly, it has been shown that the addition of these buffers in high concentrations, in particular HEPES, can stabilize the native structure of BSA,²⁰ sustain the activity of G-CSF,²¹ and increase the stability of RecA.²² However, HEPES is not widely used in drug product formulations, despite being frequently used in protein expression. As a rare example, the smallpox vaccine ACAM200 contains HEPES as a buffering agent.²³

To gain a deeper understanding of the stabilizing effect, the interaction of the protein and HEPES was analyzed by CG-MALS. Moreover, we performed MD simulations to elucidate the stabilizing effect of HEPES on HY-133 at atomistic detail. Here, we determined HEPES-interaction sites, the strength of these interactions and what impact the binding has on the protein structure. HEPES and histidine were both found to interact with the charged and aromatic side chains of surface exposed residues via salt-bridge formation and cation- π interaction. The simulations also showed that the EAD and CBD tend to interact in the absence of the excipient which is prevented by the binding of HEPES but not histidine. This self-interaction gives an indication about the mechanism of protein-protein interaction and aggregation.

6.2. Materials and methods

6.2.1. Protein and chemicals

The bacteriophage lysin HY-133 was produced by HyPharm GmbH, Bernried, Germany, in a concentration of 5 mg/mL at pH 8.0. The protein was dialyzed using a Spectra/Por dialysis membrane with a molecular weight cut-off of 10 kDa (SpectrumLabs) and a 4-fold buffer exchange to achieve the basic, buffer-free formulation. The formulation contained 0.5 mg/mL HY-133, 10 mM CaCl₂ and 10 mM methionine (both from Merck KGaA, Darmstadt, Germany), 150 mM NaCl

(Bernd Kraft, Duisburg, Germany), 300 mM arginine hydrochloride (AppliChem PanReac) and 0.05 % Poloxamer 188 (Kolliphor 188, BASF, Ludwigshafen, Germany) at a pH of 6.0. HEPES (VWR International, Ismaning, Germany), MES, MOPS, POPSO, PIPES, EPPS (all from Sigma Life Science) citrate and histidine were spiked to the buffer-free formulation to achieve a final concentration of 25 mM, respectively, and the pH was adjusted thereafter. All chemicals were of molecular biology or multi compendial grade. Prior to use, all formulations were filtered with 0.2 µm PVDF membrane syringe filters (VWR International).

6.2.2. Reversed phase high pressure liquid chromatography (RP-HPLC)

An Ultimate 3000 system (Thermo Fisher, Dreieich, Germany) and a Jupiter C4, 5 µm 300A, 250 x 4.6 mm column (Phenomenex, Torrance, USA) were used for RP-HPLC. Protein detection was performed by fluorometric detection at 280/343 nm. The column oven temperature was set to 37 °C. Mobile phase A consisted of 10% acetonitrile and 0.1% trifluoroacetic acid in highly purified water. 100% acetonitrile with 0.1% trifluoroacetic acid were used as mobile phase B. A stepwise gradient with a flowrate of 1 mL/min for 22 min was applied, 10 µL of each sample was injected.

6.2.3. Ion exchange chromatography (IEX)

A Dionex Summit system (Thermo Fischer) with a ProPac WCX-10 BioLC Analytical 4x250 mm column with an attached column guard ProPac WCX-10G (Thermo Fisher) was used for ion exchange chromatography (IEX). 50 mM Tris buffer pH 8.0 (mobile phase A) and mobile phase A plus 300 mM NaCl (mobile phase B) were used as a stepwise gradient with a flowrate of 1 mL/min for 60 min. All samples were diluted 1:10 with highly-purified water prior to injection, 100 µL of each sample was injected. A fluorescence detector (280/343 nm) was used to detect the analyte variants.

6.2.4. Size exclusion chromatography (SEC)

An Ultimate 3000 system (Thermo Fisher) with a GE Superdex 75 Increase 10/300 GL column (General Electric, Boston, MA, USA) was used for size exclusion chromatography. A filtered buffer solution containing 50 mM Na₃PO₄ and 300 NaCl in highly purified water at pH 7.0 was used as a mobile phase at a flow rate of 0.5 mL/min for 45 min. Detection was performed with a

UV-detector at 280 nm. The injection volume was 50 μ L. A gel filtration standard (#1511901, Bio-Rad Laboratories, Hercules, CA, USA) was used for regular column tests.

6.2.5. Subvisible particle analysis (FlowCam)

Subvisible particles were analyzed using a FlowCam 8000 with an attached 10x magnification cell (Fluid Imaging, Scarborough, USA). Triplicates of 200 μ L sample were analyzed at a flowrate of 0.15 mL/min. Particles were defined by a dark segmentation threshold of 10 and a light segmentation threshold of 13.

6.2.6. Intrinsic differential scanning fluorimetry (nanoDSF)

NanoDSF was used to study the thermal unfolding of HY-133 with different excipients. The samples were filled in standard glass capillaries. A Prometheus[®] NT.48 (NanoTemper Technologies, Munich, Germany) was used with a ramp of 1 $^{\circ}$ C/min from 25 $^{\circ}$ C to 95 $^{\circ}$ C. Intrinsic protein fluorescence intensity at 330 nm and 350 nm was measured after excitation at 280 (\pm 10) nm. Back-reflection intensity was measured to detect protein aggregation and precipitation. The fluorescence ratio (FI₃₅₀/FI₃₃₀) was used to determine protein thermal unfolding, calculated by PR.ThermControl V2.1 software (NanoTemper).

6.2.7. Circular dichroism (CD) spectroscopy

A Jasco J-810 spectrometer (Jasco Deutschland GmbH, Pfungstadt, Germany) was used to obtain near-UV circular dichroic spectra. Quartz cuvettes (Hellma GmbH, Muellheim, Germany) with 10 mm pathlengths were installed and measurements were performed with 10 accumulations per sample at a scanning speed of 20 nm/min. Each spectrum was buffer subtracted and curve smoothing was performed using the Savitzky-Golay algorithm with 7 smoothing points. For each sample, the molar ellipticity was calculated as described elsewhere.²⁴ The molecular weight of the HY-133 was 31,045.8 Da.

6.2.8. Composition-gradient multi-angle light scattering (CG-MALS)

Homo-association experiments by CG-MALS were performed with 0 – 10 mg/mL HY-133 in the basic formulation at pH 6.0 with and without 25 mM HEPES. Before use, all samples were filtered using 0.2 μ m PES filters. Light scattering was detected with an automated CG-MALS instrument which was equipped with a dual syringe-pump (Calypso-II) sample and preparation unit, a Dawn Heleos-II multi-angle laser light scattering detector, and an OptiLab[®] T-rEX dRI detector (all

Wyatt Technologies, Santa Barbara, CA, USA). The Calypso 2.1.5 software was used to obtain Zimm Plots, K_D and A_2 values.²⁵

In addition, hetero-association experiments were performed by varying both the HY-133 and the HEPES concentration from 0 – 10 mM and 0 – 25 mM, respectively. The experimental settings were similar to the ones described above.

6.2.9. Activity: FRET-Assay

Activity of HY-133 was determined with an enzymatic assay (HY-133 Activity Assay, Microcoat, Bernried, Germany). A short peptide sequence with attached fluorophore and quencher was used as a substrate. The substrate mimics the murein cell wall of the bacteria and is specific to *S. aureus*. The substrate is hydrolyzed by HY-133, which results in elimination of the fluorescence quenching. The fluorescence signal was obtained at an excitation wavelength of 340 nm and detected at an emission wavelength of 440 nm. The specific activity can be calculated according to the protocol.

6.2.10. Storage stability study

The basic, buffer-free HY-133 formulation and the formulations containing 25 mM of the respective buffers were sterile filtered with 0.22 μ m PVDF syringe filters (Millex-GV, Merck Millipore Ltd, Ireland), filled into sterilized 2R type I glass vials (MGlass AG, Germany) and closed with rubber chlorobutyl stoppers with FluroTec coating (West Pharmaceutical Services, USA). The samples were stored at 4 °C, 30 °C and 40 °C for subsequent analysis for up to 4 months.

6.2.11. Molecular dynamics (MD) simulations

A homology model of HY-133 was generated with the EAD from the endolysin of phage K (PDB code: 4CSH) and the CBD from lysostaphin (PDB code: 4LXC) using the software MODELLER 9.21.²⁶ The protonation states of ionizable residues at pH 6.0 were adjusted using the H++ server.²⁷ The protonated form of HEPES as well as the protonated and deprotonated forms of histidine (HIP/HIS) were parametrized with antechamber using GAFF2 for bonded and non-bonded parameters. Atomic partial charges were calculated with the AM1-BCC charge model in antechamber.

All-atom simulations were performed with the Amber19 program in a periodic box with explicit solvent.²⁸ The ff14SB force field for proteins was employed in combination with the TIP3P water

model. Packmol and tleap were used to solvate the homology model of HY-133 in a cubic water box including 150 mM NaCl, 10 mM CaCl₂ without and with 25 mM of HEPES and histidine, respectively. All bonds involving hydrogen atoms were constrained using the SHAKE algorithm. Non-bonded electrostatic interactions were treated using the particle mesh ewald algorithm with a direct space cutoff of 9 Å. The system was energy minimized with the steepest descent algorithm for the first 5,000 cycles, followed by 5,000 cycles using the conjugate gradient method. System equilibration was carried out for 1 ns in NVT ensemble to stabilize the temperature of 300 K using the Langevin thermostat, and subsequently for 1 ns in NPT ensemble to adjust the density of the system using the Berendsen barostat. The simulations were performed for 300 ns with a time step of 2 fs. The coordinates were saved every 10 ps. Each system was simulated in duplicates starting from a random seed number to estimate the statistical uncertainty of the results. All trajectories were analyzed using the CPPTRAJ module of Amber19 and VMD.²⁹ In order to identify effects of the buffer molecules on the conformation of HY-133, each frame of a trajectory was grouped into one of five discrete clusters by an agglomerative hierarchical clustering approach based on the conformational similarity defined by the root-mean-square deviation (RMSD). To estimate the binding of the HEPES and histidine molecules to the protein residues, we calculated the interaction score probability ($P(I_{\text{score}})$) defined as

$$P(I_{\text{score}}) = n / (N \cdot a)$$

where n is the number of frames either HEPES or histidine was in contact with a certain residue over simulation time, N is the total number of frames, and a is a normalization factor to convert the interaction score into a probability.

6.3 Results and discussion

6.3.1. Influence of buffer excipients on chemical stability

At first, the protein's chemical stability with different buffer excipients was assessed. The 25 mM HEPES buffer was compared to a 25 mM citrate buffer and a 25 mM histidine buffer, respectively. In addition, combinations of 12.5 mM HEPES with 12.5 mM citrate or with 12.5 mM histidine were tested. All three formulations containing HEPES showed a constant native protein level upon storage for 16 weeks at 4 °C, irrespective of whether HEPES was used alone or in combination with a second buffer substance (Figure 38). In contrast, the formulations containing only citrate or

only histidine showed a reduced native protein recovery of either 96% (citrate) or 84% (histidine). Storage at an elevated temperature of 40 °C led to a decrease of the native protein content over time in all formulations. This was most pronounced in formulations without HEPES, where only 20% protein recovery was achieved after 16 weeks. The highest native protein content (> 60%) was determined in the formulation containing solely HEPES. Both buffer combinations, HEPES + citrate and HEPES + histidine, resulted in native protein contents between the single buffer compositions. These results show that HEPES could not be replaced by commonly used buffer systems which would be more suitable for this pH range. Therefore, the stabilizing effect is independent of the buffering capacity and HEPES is not acting only as a buffering agent. A specific stabilizing effect of HEPES is clearly indicated.

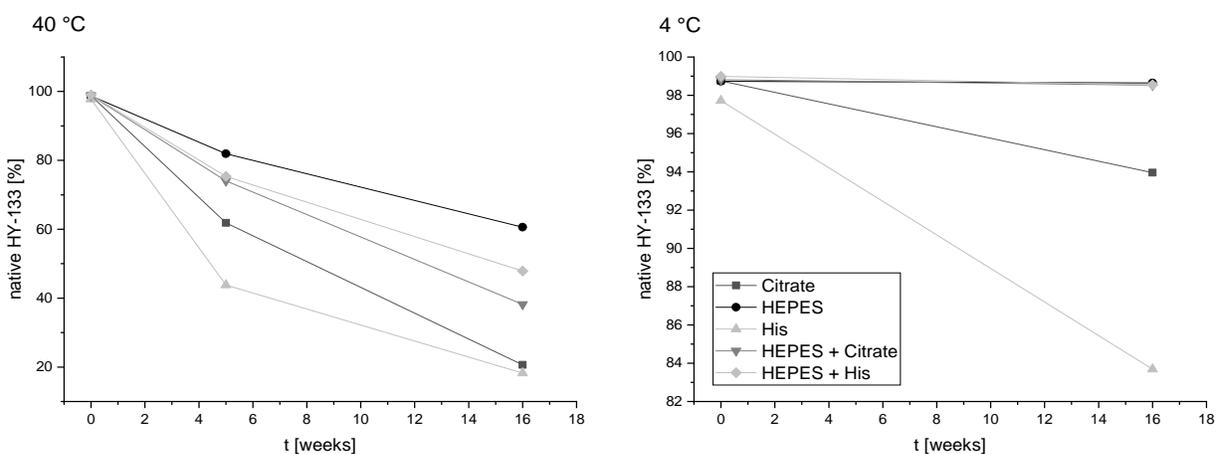


Figure 38: Native HY-133 amount over time dependent on different excipients. The pre-test was performed with each a citrate, histidine and HEPES buffer or the respective HEPES combinations and analyzed with RP-HPLC.

6.3.2. Function of HEPES as stabilizing excipient

To elucidate the role of HEPES in the HY-133 formulation, the interaction of the protein and HEPES was analyzed by CG-MALS. CG-MALS provides quantification of both self- and hetero-association of either a protein or a protein and another molecule and allows the determination of A_2 , K_D , and M_w . First, protein self-association was measured via the light scattering second virial coefficient (A_2) in the presence of HEPES and without HEPES (Figure 39). The typical step-wise decrease in light scattering signal resulted from the stop-flow injection of the different

compositions. A visible difference between the two experiments was the delay until a stable light scattering signal was reached, which was more pronounced at higher protein concentrations. The delay appeared as an initially small increase in light scattering followed by a major signal decrease until the signal stabilized. Normalized A_2 values of -2.52 and -2.38 were determined without HEPES and with HEPES, respectively. A negative A_2 is associated with attractive forces between two protein molecules,³⁰ which were prevalent in this formulation, irrespective of the presence of HEPES.

However, HEPES influenced the light scattering signal. The rapid set of the plateau and the initially observed peak at each step indicated a rapid binding of HEPES to protein hotspots. To gain a deeper understanding of the binding of HEPES to the protein, a hetero-association experiment was performed.

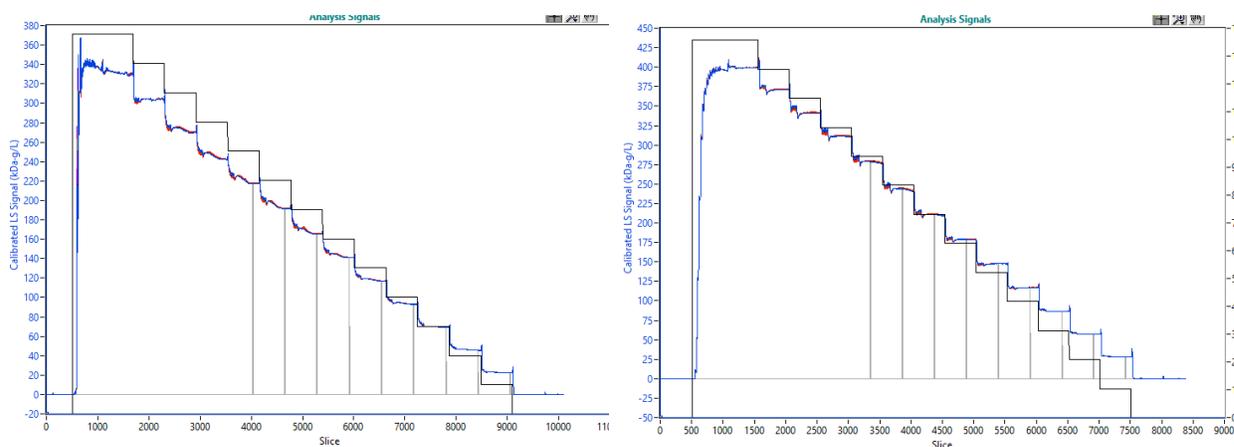


Figure 39: CG-MALS data for HY-133 with HEPES (left) and without HEPES (right). Light scattering data of HY-133 dependent on the protein concentration with and without HEPES. The black line shows the concentration signal, the blue line the light scattering signal. Grey bars represent measurement points.

In hetero-association experiments, a series of different composition ratios of HY-133 and HEPES were fitted to determine K_D and kinetic data. During the fitting process, the monomer molecular mass of both analytes and the associations stoichiometries can be modelled.³¹ The best fit of HY-133 with HEPES was found to be a 1:1 stoichiometry model, also represented by the plateau with the highest amount of complex formation. The overlap of the light scattering data in the stepwise decrease in HY-133 concentration and the stepwise increase in HEPES concentration resulted in a K_D of $0.7 \cdot 10^{-3}$ mM. The data indicated an almost immediate binding of a HEPES molecule to a HY-133 molecule.

We were interested in whether we could confirm HEPES-binding, identify the local effects of HEPES on the protein and explain the stabilization mechanism. We therefore built a homology model of HY-133 and performed MD simulations to study the protein in the absence and presence of HEPES and histidine at atomistic detail. The interaction sites of HEPES and histidine were determined by calculating the interaction score $P(I_{\text{score}})$ per residue (Figure 40A). The $P(I_{\text{score}})$ reveals that both HEPES and histidine bind to charged and aromatic side chains of surface exposed residues of HY-133 via salt-bridge formation and cation- π interactions. Moreover, hydrogen bonds between HEPES and polar residues occurred. Considering the zwitterionic character of HEPES and histidine, this behavior was expected. We, however, observed different binding sites for the two molecules. Histidine mainly interacted with charged residues in the EA domain and with the termini of HY-133. HEPES showed some interaction with residues in the EA domain but also interacted with the more hydrophobic CB domain and linker residues. Histidine therefore altered the electrostatic surface properties of the protein to a higher extent than HEPES which may influence the inter-domain and intermolecular interactions. Nevertheless, this finding alone does not explain why HEPES is superior to histidine in stabilizing the protein.

Next, the effects of HEPES and histidine binding on the conformation of HY-133 were examined by grouping each frame of the trajectory into one of five clusters, based on the conformational similarity defined by the RMSD. A representative structure of the highest occupied cluster for each condition is shown in Figure 40C. Comparing these structures, it became apparent that the orientation and the distance of the domains substantially differed between the three conditions. For the simulations of HY-133 in water only, the initially elongated linker collapsed and formed a hydrogen bond network to residues in the CB domain. Thus, the two domains were in close contact. Due to the sequence composition of the linker, this behavior is not surprising. Studies on intrinsically disordered peptides in water uncovered more compact structures for sequences with a high glycine content, supposedly because intrapeptide interactions out-compete backbone-solvent interactions.^{32,33} The simulations with histidine also revealed a collapse of the linker but with less hydrogen bonding to residues of the EA or CB domain than in the simulation with water only. Again, there was an interaction of the two domains but with a different orientation. The CB domain was rotated in a way that a hydrophobic region at its C-terminus is exposed to the surface. In the two other conditions, this region was located at the interface between the domains and mostly shielded from the surface. Only in simulations containing HEPES, the linker almost maintained the

initial extended configuration with only a short bend around a HEPES molecule. Hydrogen bonds of this HEPES molecule to residues in the linker and a salt-bridge to residue 246 in the CB domain seemed to prevent the collapse of the linker. Thus, the linker was able to keep the domains separated. For a better understanding of intramolecular domain-domain interactions and the influence of the solvent conditions, we compared fluctuations in the distance between the two domains over simulation time (Figure 40B). In water only, the distance between the two domains rapidly decreased. Histidine molecules delayed the collapse of the linker but after 150 ns of simulation time, a similar distance as for the water simulations was reached. Only the simulations with HEPES showed an almost stable domain distance of approximately 30 Å.

In summary, the ability of HEPES to maintain the extended linker configuration and saturate hydrogen bond sites in the linker was the reason for the increased stability of HY-133. In the two other conditions, the linker collapsed which could even lead to the exposure of a hydrophobic region, as seen for histidine. Notably, also the changes in electrostatic properties of the protein surface upon binding of a zwitterion could influence the stability.

Due to the several binding sites identified, we cannot confirm a 1:1 stoichiometry as shown by the CG-MALS data. However, several of the binding events were only transient throughout the simulation and only a few were stable until the end of the simulation. Based on a previous study showing that weak and transient protein-excipient binding could enhance protein stability, whereas strong interactions were found to mostly destabilize proteins,¹¹ we could conclude that the lack of strong HEPES-binding did not hinder a stabilizing effect but rather promotes it.

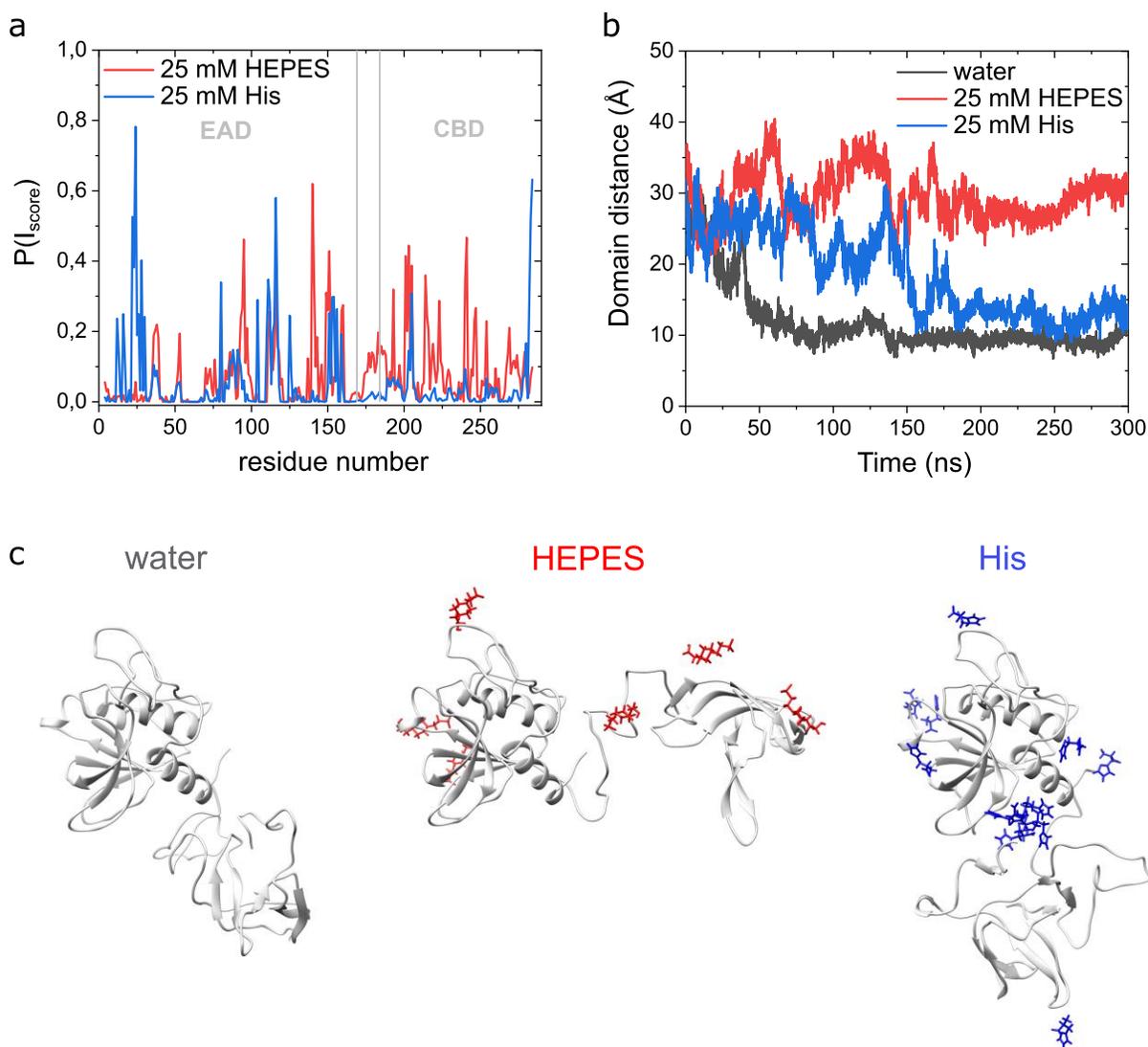


Figure 40: MD simulations of HY-133 without excipient, with 25 mM HEPES and 25 mM histidine. (a) Interaction scores $P(I_{\text{score}})$ calculated per residue for HEPES and histidine averaged over the two replicates. (b) Distance of the last residue of the EA domain and the first residue of the CB domain over simulation time. (c) Representative structures of the highest occupied clusters for the three studied conditions (water: 61.1%. HEPES 62.2%, His: 74.6%) with aligned EAD domains. HEPES molecules are colored in red, protonated and deprotonated histidine molecules are colored in blue.

6.3.3. Influence of different Good's buffers on protein-excipient interactions

HEPES is a zwitterionic N-substituted aminosulfonic acid and part of the Good's buffer series. This series can be separated into several subgroups. The most prominent ones are the morpholinic, piperazinic and the TRIS family.¹⁹ The structural similarities of buffers from the morpholinic and

piperazinic family are most pronounced. All side chains of the buffers in these families contain sulfonic acid groups and have specific lengths. The piperazinic buffers contain two side chains instead of the one side chain in morpholinic buffers. Due to structural similarities, these buffers might be able to sufficiently stabilize protein to a similar extent as HEPES.

Six different Good's buffers of the above-mentioned groups were selected to examine their ability to replace HEPES and provide comparable stability of HY-133: four piperazine derivatives, HEPES, POPSO, PIPES and EPPS, and two morpholine derivatives, MES and MOPS. These buffers were characterized as non-metal binding and therefore fully compatible with CaCl_2 , which is included in the formulation buffer and directly attached to one binding site at the protein. PIPES, HEPES, and MOPS are also reported to be non-metal binding.^{18,34} Only MES is potentially Ca^{2+} binding, a fact that is contrarily discussed in the literature.^{18,19}

The thermal stability of proteins can be expressed by their melting temperature (T_m). The T_m values of the different formulations could be ranked, with a higher T_m representing a thermally more stable product.⁷ The different buffer excipients were screened with nDSF to evaluate the thermal stability of HY-133 depending on the respective excipient. Without an additional excipient, a T_m of 46.4 °C was obtained (Table 4). Citrate and histidine buffers showed only minor changes in T_m , whereas the Good's buffers increased the respective T_m values. The highest value of 52.6 °C was found for PIPES which, along with HEPES, EPPS, and POPSO, belongs to the group of piperazine derivatives. Thus, the Good's buffers indicated a conformational protein stabilizing efficiency.³⁵

As shown before, T_m values of various proteins can depend on different N-based buffers.^{5,36} Both 20 mM MES and HEPES increased the conformational stability of RecA. Thermal stability was reported to be higher when the pH was above the pK_a of the respective buffer substance. This effect was stronger at lower pH, when the protein showed higher positive net charge, indicating a direct interaction of the buffer with the protein under these conditions.²² In addition, MOPS showed a higher propensity to increase the thermal stability of BSA than MES when used in high concentrations of 1 M. It was shown that these buffers interact with the peptide backbone leading to net stabilization.³⁷ Furthermore, HEPES and EPPS were also stabilizing the conformational stability of BSA by interacting with the hydration layers of the peptide backbone.²⁰ In contrast, HY-133 already showed an increase in T_m at much lower concentrations, which indicates direct interaction of the excipient with the protein rather than stabilization due to preferential exclusion.

Table 4: T_m values of HY-133 dependent on the added excipient (excipient concentration = 25 mM).

| Excipient | T_m [°C] |
|-------------------|----------------------------|
| Without excipient | 46.4 |
| HEPES | 51.5 |
| Citrate | 47.9 |
| Histidine | 45.0 |
| MES | 48.1 |
| MOPS | 51.1 |
| EPPS | 51.5 |
| POPSO | 52.4 |
| PIPES | 52.6 |

As described above, the T_m increase depends on the respective excipient at a fixed concentration of 25 mM. To determine the necessary ligand concentration, a thermal shift assay was performed. 12 different concentrations of HEPES, EPPS, POPSO, PIPES, MES, MOPS, citrate and histidine, respectively, ranging from 0 to 50,000 μM were spiked to the buffer-free formulation. The T_m of HY-133 in each buffer and excipient concentration was determined and plotted as a function of the respective excipient concentration (Figure 41). Formulations containing PIPES and POPSO resulted in higher overall T_m values and a faster increase of T_m. These two curves are followed by HEPES, EPPS, and MOPS, which all showed a similar curve profile. An overall higher concentration of the respective excipient was needed to achieve similar T_m values. The MES buffer system showed only a minor increase in T_m. Both citrate and histidine buffered formulations showed a non-sigmoidal shaped curve. The histidine buffered formulation resulted in an atypical curve with a local T_m minimum of about 34 °C. Citrate resulted in a local T_m maximum at 3,000 μM with decreasing T_m values at higher citrate concentrations.

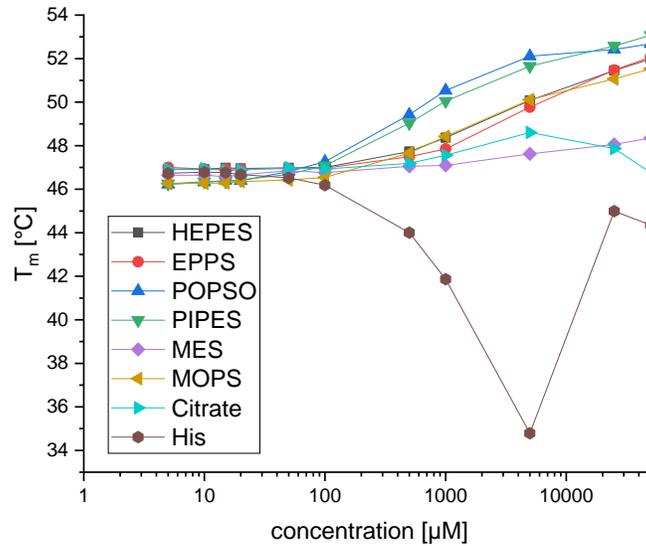


Figure 41: Overlay of the T_m values of the different excipients depending on their concentration.

For better comparison, the thermal shift curves were fitted with the Hill1-function (Figure 42). The inflection point k is given in μM and equals the concentration, which is necessary for a significant increase in T_m . As described elsewhere, the inflection point equals K_d .^{36,38,39}

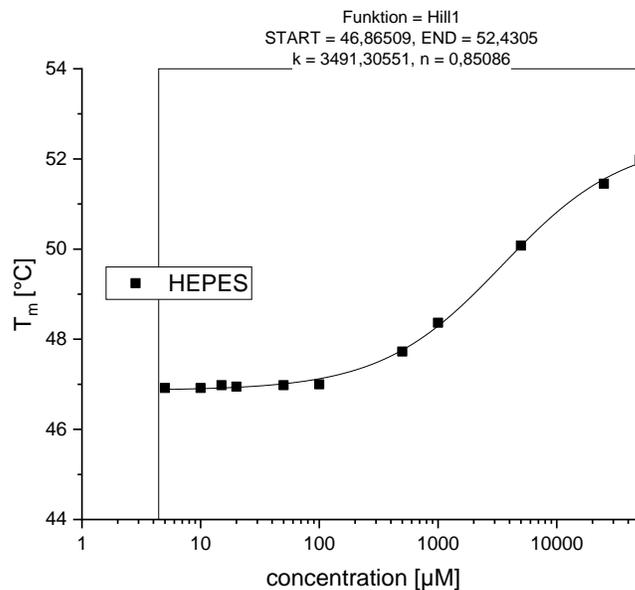


Figure 42: T_m values as a function of HEPES concentration. The curve was fitted with the Hill1 equation; the inflection point was calculated as $k=3491 \mu\text{M}$.

The K_d values were calculated for the different excipients and are shown as $c(\text{Ligand})_k$ in [mM] in Table 5. As histidine and citrate buffers did not show a sigmoidal shaped curve, fitting and calculation of K_d was not possible. The lowest $c(\text{Ligand})_k$ was obtained for POPSO (0.5 mM) and PIPES (0.8 mM), followed by MOPS (1.6 mM). HEPES (3.5 mM) and EPPS (5.1 mM) were comparable, whereas MES (30.5 mM) showed a substantially higher $c(\text{Ligand})_k$.

These results were used to select four different ligands for further evaluation of the physical and chemical stability of HY-133 in formulations containing different buffer excipients. HEPES and POPSO were selected representing the piperazine buffer excipients, differing in their respective $c(\text{Ligand})_k$. MES and MOPS were selected as morpholine derivatives. The 4 different excipients were compared to the basic formulation with only arginine-HCl, methionine, NaCl, CaCl₂, and Poloxamer 188.

Table 5: K_d values of the different excipients calculated in the thermal shift assay.

| Ligand | $c(\text{Ligand})_k$ |
|---------------|--|
| POPSO | 0.5 mM |
| PIPES | 0.8 mM |
| MOPS | 1.6 mM |
| HEPES | 3.5 mM |
| EPPS | 5.1 mM |
| MES | 30.5 mM |

6.3.4. Stability study of HY-133 with HEPES and further Good's buffers

Subsequent to the determination of the influence of Good's buffers on the conformational stability, chemical degradation, protein aggregation (soluble and insoluble), and structural changes of the protein depending on the different formulations were analyzed over the course of up to 16 weeks at different storage conditions. Different Good's buffers were compared to a buffer free formulation. The formulations were stored at 2-8 °C, 30 °C, and 40 °C.

All formulations exhibited a characteristic near-UV CD-spectrum with a large negative peak at around 285 nm (Figure 43). In addition, three less prevalent peaks at 268 nm, 276 nm and 290 nm could be observed in all spectra. However, a decrease in molar ellipticity in the near-UV region was noted in the formulation without buffer excipient at T0. This structural difference can be linked to the missing additional excipient. The presence of HEPES and its structural similar entities affect

the tertiary structure of the protein. Similar to the aforementioned CG-MALS experiments and MD simulations, where rapid binding of HEPES to the protein was shown, structural changes of the protein are caused by ligand binding to the protein.²⁴ In general, CD is a very precise tool to determine even minor changes in the protein's conformation⁷ and changes in the higher order structure caused by the addition of sucrose were reported before.⁴⁰

After 2 weeks storage at 30 °C, this initially shifted CD-spectrum was further altered, which resulted in a spectrum without a large, distinct peak. Minor changes were observed in the POPSO formulation in this storage condition, whereas the three other formulations remained unchanged. In contrast, 40 °C storage for two weeks led to a loss in molar ellipticity in all formulations. However, the characteristic near UV CD-spectrum was maintained for all formulations containing excipient. The formulation without an additional excipient showed further alteration.

After 16 weeks storage at 4 °C, the characteristic near-UV CD-spectrum were maintained in all formulations.

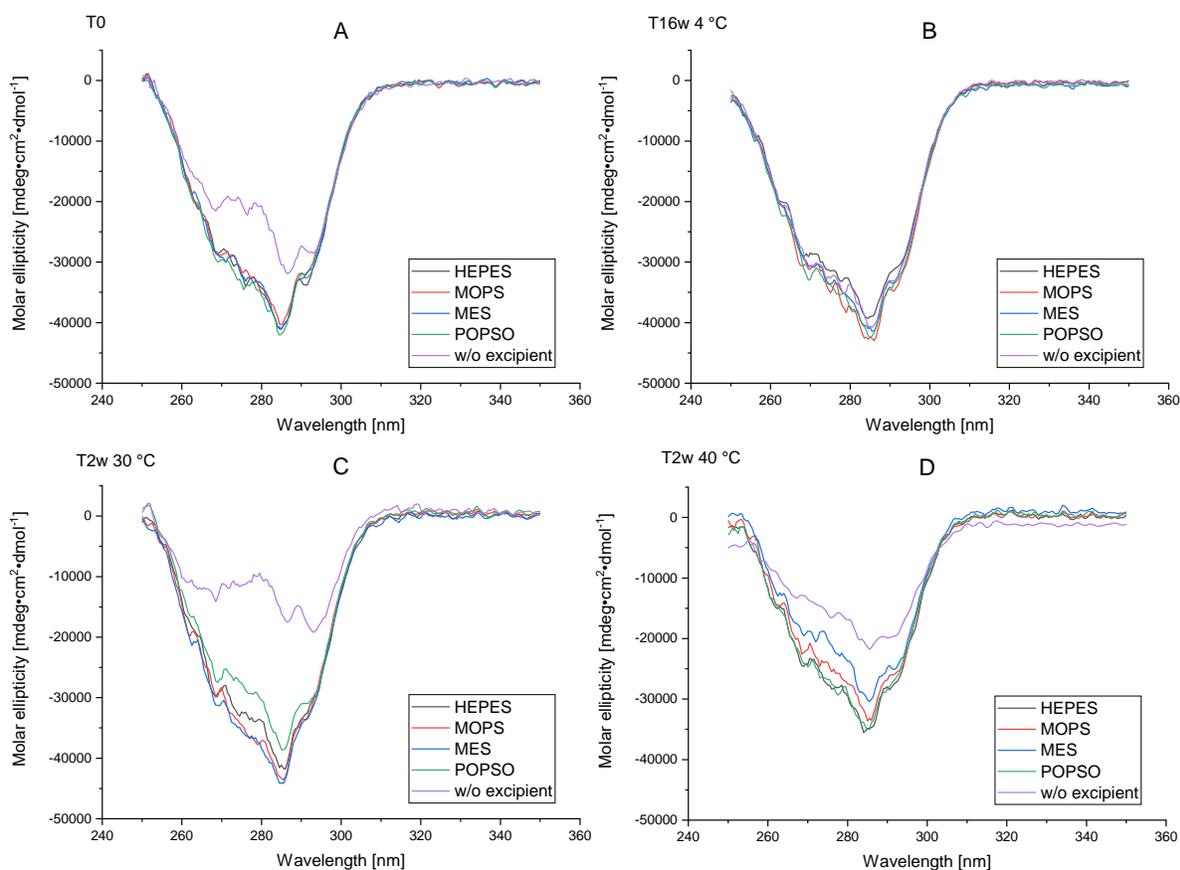


Figure 43: Molar ellipticity of formulations incl. HEPES, MOPS, MES and POPSO compared to a formulation without an additional buffer excipient. (A) T0, (B) after 16 weeks storage at 4 °C, (C) 2 weeks storage at 30 °C (D) 2 weeks storage at 40 °C.

Subvisible particle counts remained low throughout the stability study (Figure 44). Storage at 4 °C did not lead to a major increase in particle counts, irrespective of the formulation. Elevated storage temperatures resulted in a slight increase in particle counts. Higher particle numbers were observed in formulations containing MOPS and MES at 30 °C storage. Storage at 40 °C led to an increase in particle numbers in the formulations containing MES and in the formulation without an additional excipient at the 9-week time point. Yet, this increase could not be confirmed at the following time points, indicating no clear stability trend. The structural differences in protein folding shown in the CD experiment was apparently not sufficient to cause particle formation through the formation of larger oligomers.⁴¹ The presence of the surfactant Poloxamer 188 in a suitable concentration effectively inhibited subvisible particle formation by improving the colloidal stability of the protein.⁴²

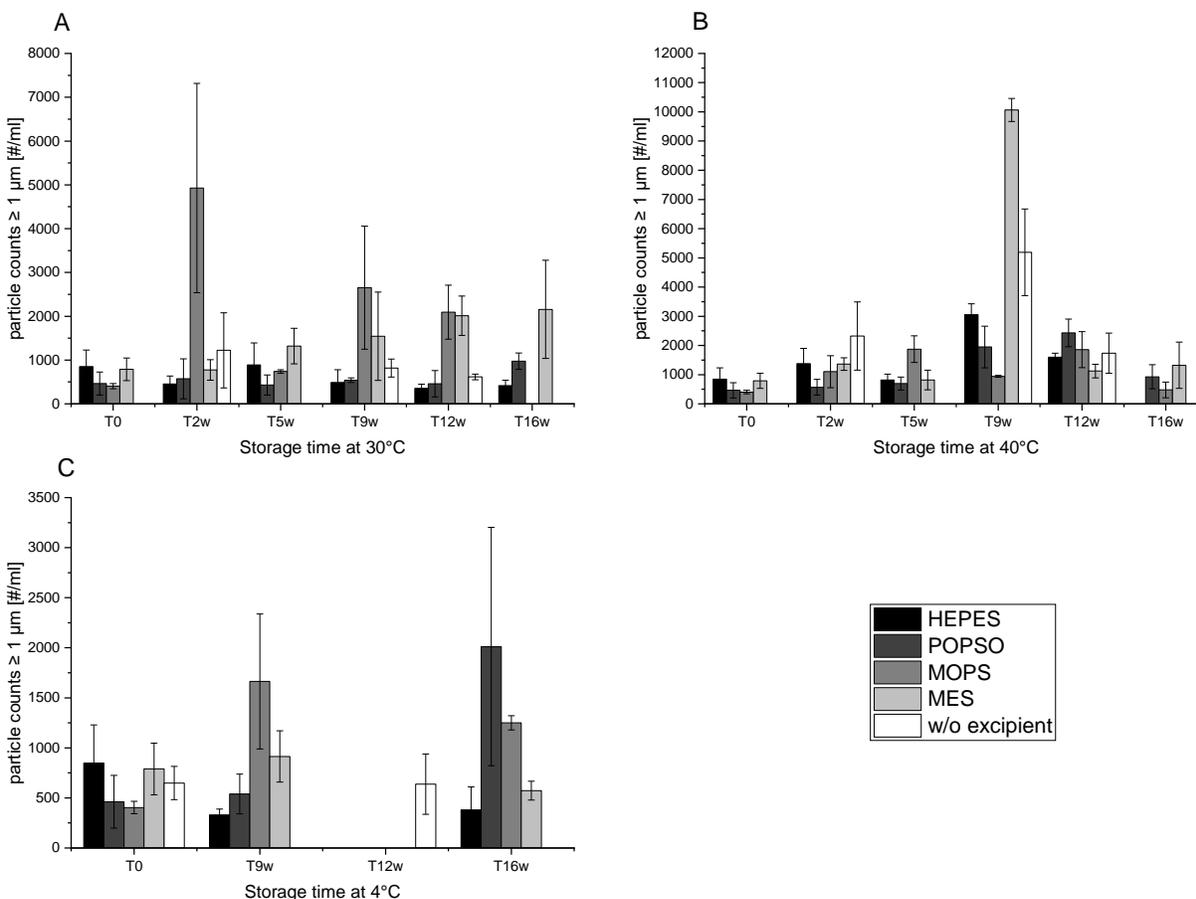


Figure 44: Particle counts of the different formulations upon storage at (A) 30 °C, (B) 40 °C, and (C) 4 °C.

Size exclusion chromatography was used to determine the monomer content of each HY-133 formulation (Figure 45). An initially lower monomer content was determined in the formulation without excipient. Storage at 4 °C provided a stable monomer content over the course of the study, as no further decrease was observed in any of the formulations. However, storage at elevated temperatures led to a loss in relative monomer content. A comparable, small decrease in monomer content was determined in all formulations stored at 30 °C, differing only in the initially lower monomer content of the formulation without an additional excipient. Storage at 40 °C led to substantial loss in monomer content over time, which was more pronounced in the formulation without an additional excipient. HEPES performed best and its similar excipients stabilized the protein to a comparable extend.

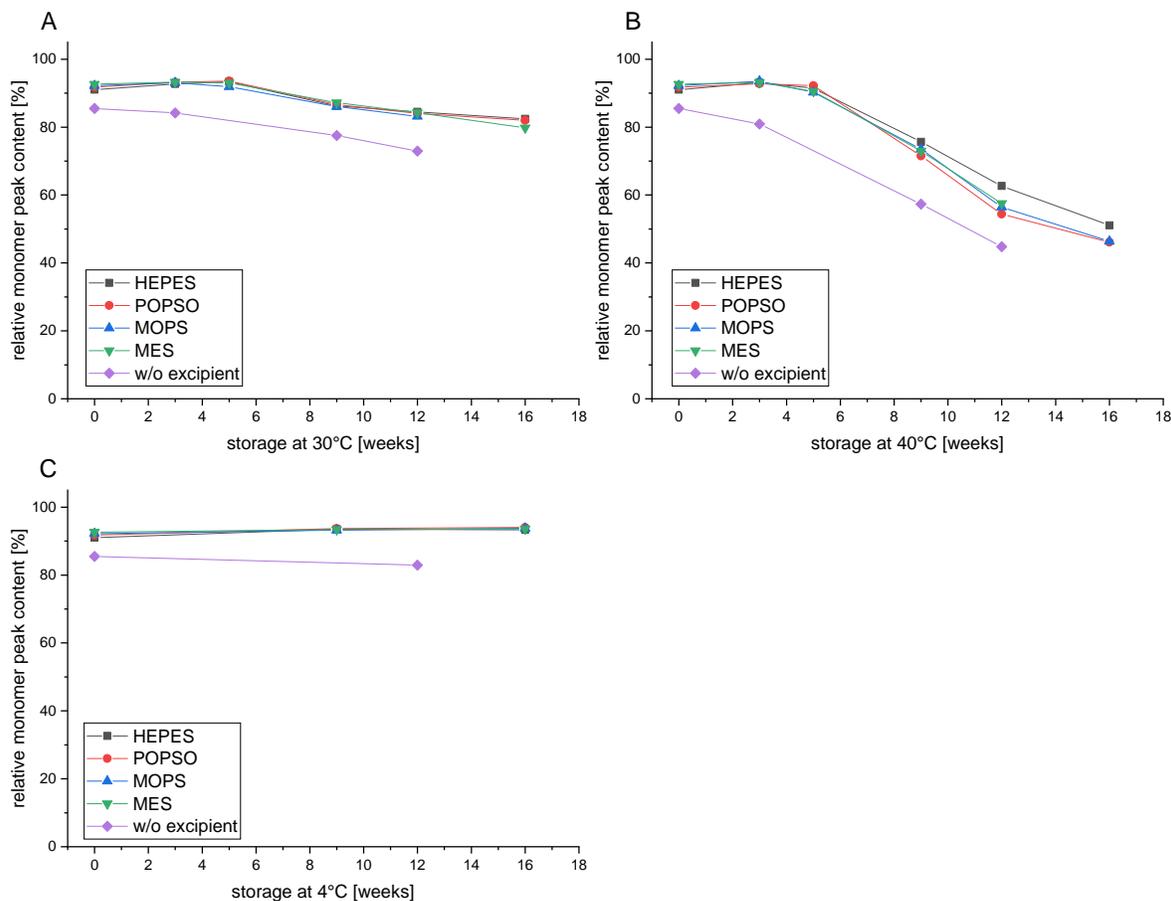


Figure 45: SEC data of the different HY-133 formulations upon storage at (A) 30 °C, (B) 40 °C, and (C) 4 °C.

Chemical changes, such as oxidation and deamidation, of HY-133 were detected with RP-HPLC. The relative protein content is displayed in Figure 46. The formulation without an additional excipient showed an immediate loss in native protein content after dialysis at T0, followed by an accelerated degradation at early time points. Chemical changes were further highly dependent on storage temperature. While only minor changes of the protein were observed in any of the formulations during storage at 4 °C and 30 °C for up to 16 weeks, storage at 40 °C led to a drastic degradation in all the formulations. These changes were the least pronounced in formulations containing HEPES and MES, followed by POPSO and MOPS and most pronounced in the formulation without an additional excipient. Thus, all the derivatives and in particular HEPES and MOPS were able to stabilize HY-133 against chemical denaturation.

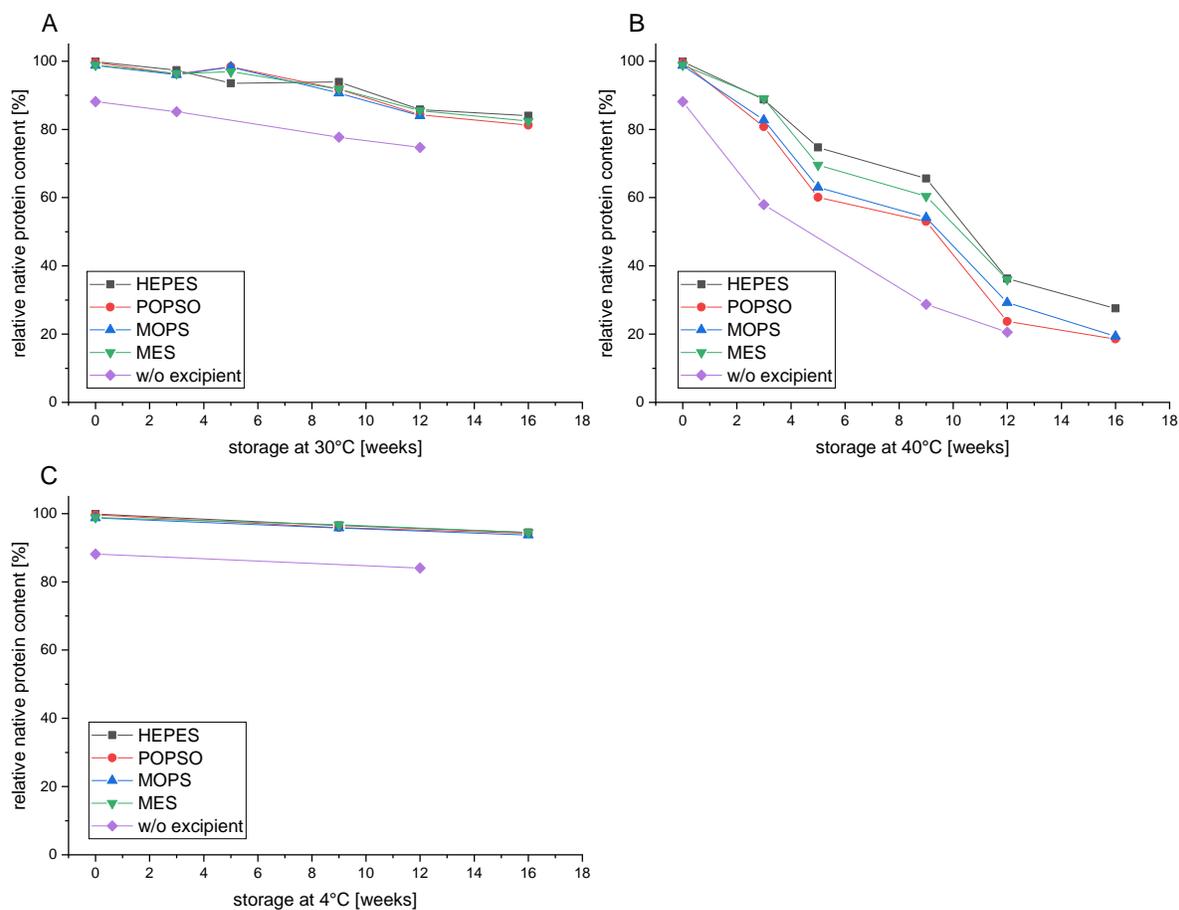


Figure 46: RP-HPLC data of the different HY-133 formulations upon storage at (A) 30 °C, (B) 40 °C, and (C) 4 °C.

Charge variant formation, i.e. the formation of specific acidic and basic species of HY-133 upon storage, was determined by cation exchange chromatography (CEX) as shown in Figure 47. In general, the loss in main peak content of HY-133 is connected to an increase in pre-peak content, reflecting more acidic species of the protein. No change in the protein main peak was determined after storage at 4 °C in any of the formulations over the course of the study. Storage at 30 °C led to only minor changes in protein main peak with about 90% remaining after 16 weeks of storage, irrespective of the formulation. Storage at 40 °C led to a major loss in main peak content in all formulations over time. Larger loss was observed in the formulation without an additional excipient, indicating lower chemical stability compared to formulations with an additional HEPES-like excipient. For example, after 12 weeks of storage, formulations with an additional excipient

resulted in more than 80% remaining main peak area, contrary to less than 30% main peak area in the HEPES-free formulation.

Several chemical alterations are known that are predominantly connected to the development of more acidic species. The most common one is deamidation of asparagine (Asn) residues, which is reflected by higher acidic species rates.⁴³ In addition to Asn residues, glutamine (Glu) residues can also be affected. Deamidation is connected to possible changes in the secondary and tertiary structure and therefore changes in the functionality of the protein.^{4,44}

Thus, our results show that an HEPES-like excipient is crucial to increase the stability of HY-133 compared to a formulation without one of these excipients.

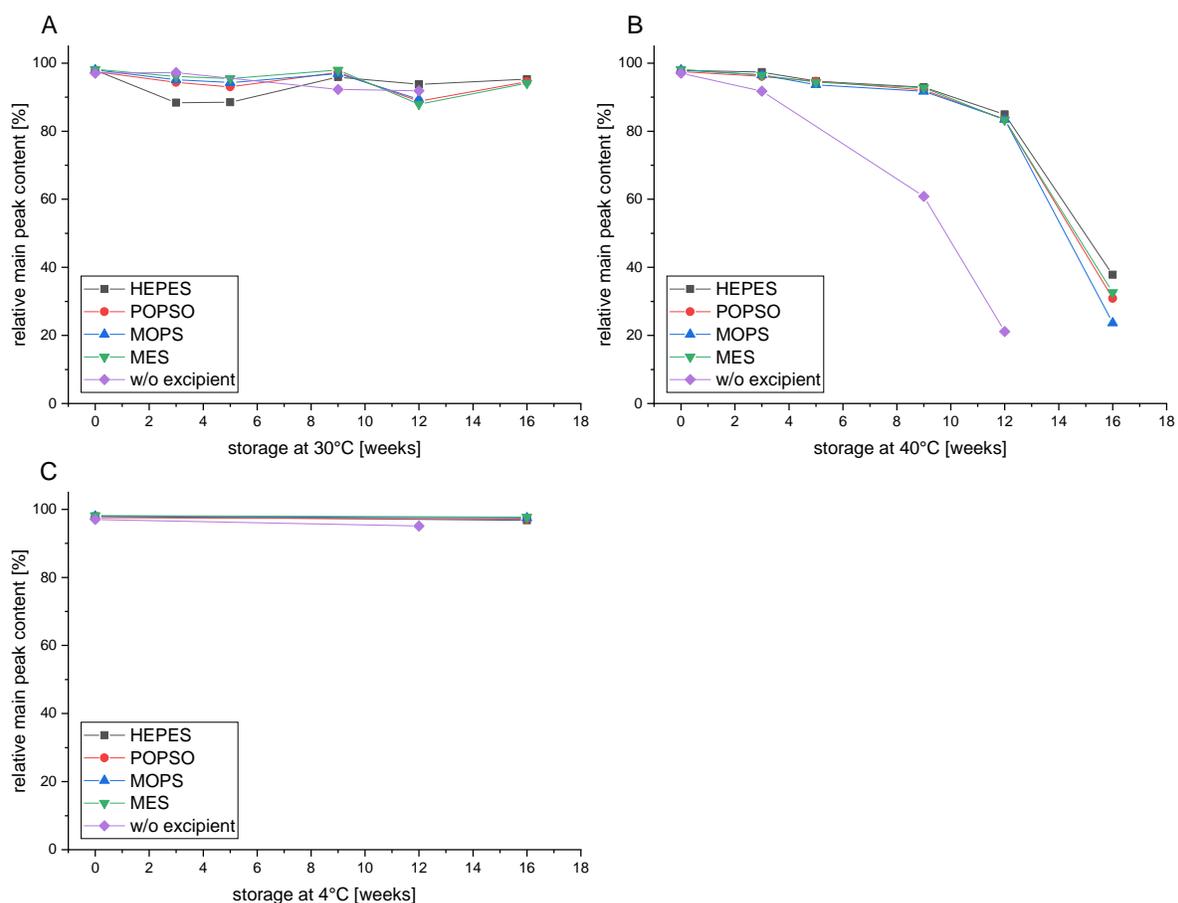


Figure 47: CEX data of the different HY-133 formulations upon storage at (A) 30 °C, (B) 40 °C, and (C) 4 °C.

Finally, an enzymatic FRET-assay was applied as an activity assay (Figure 48). Storage at 4 °C and 30 °C did not lead to a loss in activity in any of the formulations after 16 weeks. In contrast, storage at 40 °C led to a decrease in activity in all formulations with differences between the formulations and storage durations. The highest and fastest loss in activity was determined in the formulation without an additional excipient. A decrease in specific activity of more than 50% was observed after 2 weeks and it further decreased over the course of the study until inactivity.

With the addition of HEPES or a HEPES-like buffer substance, the decrease in activity during storage at 40 °C was decelerated until 9 weeks. However, after 16 weeks at 40 °C, all formulations showed a substantial loss in activity.

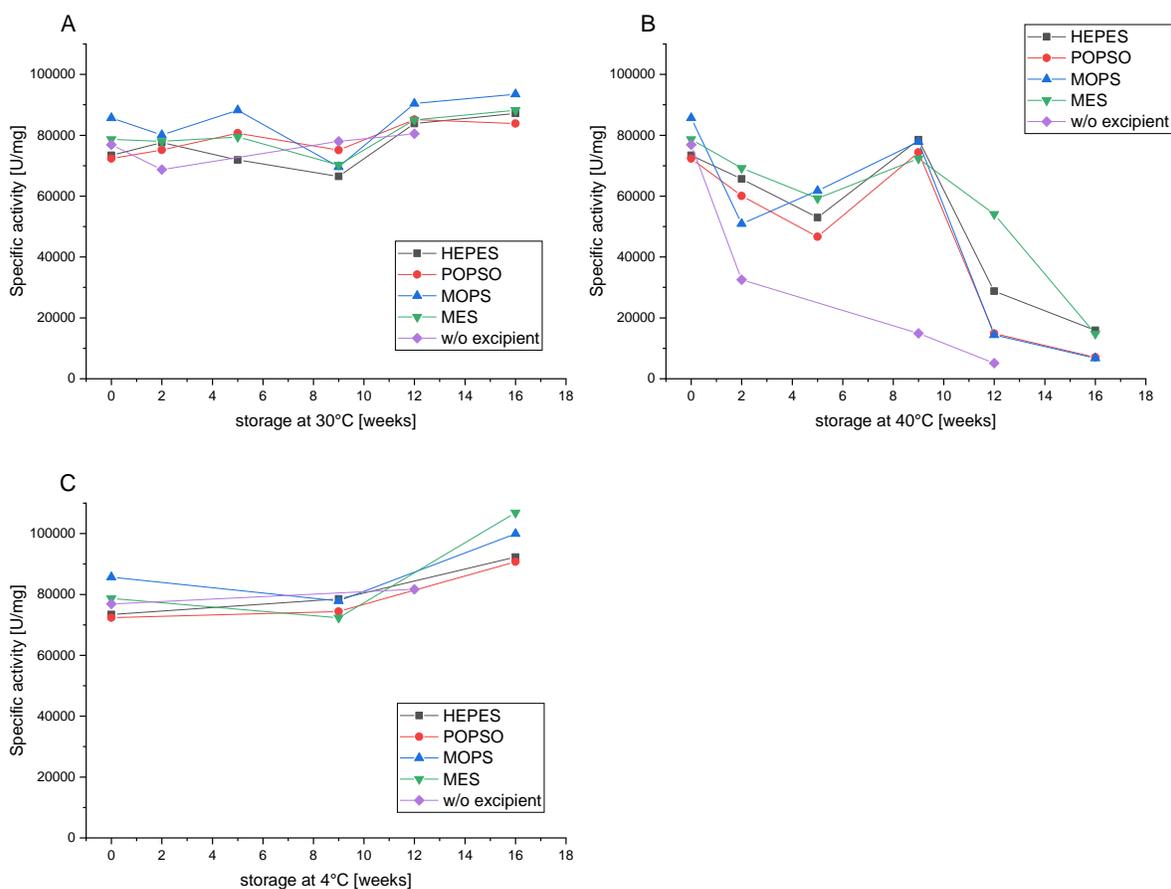


Figure 48: Activity of the different HY-133 formulations upon storage at (A) 30 °C, (B) 40 °C, and (C) 4 °C.

6.4. Conclusions

The aim of this study was to identify and further examine the influence of HEPES on the protein stability of HY-133, a novel antibacterial drug specifically targeting various *S. aureus* strains. Furthermore, HEPES was compared to other Good's buffers used as excipients in a liquid biopharmaceutical formulation. So far, HEPES is mainly used as a buffer in protein expression. However, here we investigated the potential advantage of HEPES in formulation development.

Compared to commonly used histidine and citrate buffer systems, HEPES was clearly superior in stabilizing the protein HY-133 regarding its chemical and conformational stability. While HEPES did not change the second virial coefficient A_2 , a rapid binding of HEPES to the protein was observed in CG-MALS. MD simulations could confirm the binding of HEPES to HY-133, thereby preventing the collapse of the linker and inter-domain interactions. This effect could not be reproduced with histidine, even though binding was also detected.

In the second part of the study, HEPES was substituted by various other Good's buffers with morpholinic and piperazinic ring systems: MOPS, POPSO, MES, EPPS, and PIPES. Whereas histidine and citrate buffers were not able to increase the T_m of HY-133, HEPES and the respective derivatives increased the T_m by up to 6 °C.

Furthermore, thermal shift assays with the examined formulations were used to calculate K_d based on the Hill1-equation. POPSO showed the lowest concentration to increase the T_m of HY-133 with 0.5 μM , followed by PIPES (0.8 μM), MOPS (1.5 μM), and HEPES (3.5 μM). K_d could not be calculated for citrate and histidine buffers.

Near-UV CD spectroscopy revealed that the conformational stability is sustained in formulations containing Good's buffers. Differences in the tertiary structure of the protein in the formulation without a Good's buffer was already determined at T_0 , indicating structural changes during the dialysis. Looking at different chemical and physical degradations, similar trends were found for different LC methods, such as SEC, RP, and CEX. The loss in protein activity was also reduced by the addition of HEPES and related buffer substances.

Overall, HY-133 highly benefits from the addition of HEPES or a similar Good's buffers due to sustained activity and substantially reduced chemical degradation. The studied Good's buffer substances performed similar. However, this study showed that the selection of an optimal buffer

system is not limited to the buffer capacity at the perfect buffering range of the compounds. Instead, direct effects on the protein stability must also be considered. We showed that buffer substances could directly interact with a protein and could lead to improved chemical and conformational stabilization for new and pharmaceutical highly relevant protein classes.

6.5. References

1. C. J. Roberts, Therapeutic protein aggregation: Mechanisms, design, and control, *Trends in Biotechnology* **32** (7), 372-380 (2014).
2. M. C. Manning, D. K. Chou, B. M. Murphy, R. W. Payne, and D. S. Katayama, Stability of protein pharmaceuticals: An update, *Pharm. Res.* **27** (4), 544–575 (2010).
3. W. Wang, S. Singh, David I. Zeng, K. R. King, and S. Nema, Antibody Structure, Instability, and Formulation, *J. Pharm. Sci.* **96**, 1–26, (2007).
4. D. Gervais, Protein deamidation in biopharmaceutical manufacture: understanding, control and impact, *J. Chem. Technol. Biotechnol.* **91**, 569–575 (2016).
5. T. J. Zbacnik et al., Role of Buffers in Protein Formulations, *J. Pharm. Sci.* **106** (3), 713–733 (2017).
6. W. Wang, S. Nema, and D. Teagarden, Protein aggregation-Pathways and influencing factors, *Int. J. Pharm.* **390** (2), 89–99 (2010).
7. T. J. Kamerzell, R. Esfandiary, S. B. Joshi, C. R. Middaugh, and D. B. Volkin, Protein-excipient interactions: mechanisms and biophysical characterization applied to protein formulation development., *Adv. Drug Deliv. Rev.* **63** (13), 1118–59 (2011).
8. H. J. Lee, A. McAuley, K. F. Schilke, and J. McGuire, Molecular origins of surfactant-mediated stabilization of protein drugs, *Adv. Drug Deliv. Rev.* **63** (13), 1160–1171 (2011).
9. D. P. Elder, M. Kuentz, and R. Holm, Pharmaceutical excipients — quality, regulatory and biopharmaceutical considerations, *Eur. J. Pharm. Sci.* **87** (25), 88–99, (2015).
10. S. N. Timasheff, Control of protein stability and reactions by weakly interacting cosolvents: The simplicity of the complicated, *Adv. Protein Chem.* **51**, 355–432 (1998).
11. M. Zalar, H. L. Svilenov, and A. P. Golovanov, Binding of excipients is a poor predictor for aggregation kinetics of biopharmaceutical proteins, *Eur. J. Pharm. Biopharm.* **151**, 127–136 (2020).
12. A. Tosstorff, H. Svilenov, G. H. J. Peters, P. Harris, and G. Winter, Structure-based discovery of a new protein-aggregation breaking excipient, *Eur. J. Pharm. Biopharm.* **144**, 207–216 (2019).
13. A. Tosstorff, G. H. J. Peters, and G. Winter, 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).
14. D. Shukla and B. L. Trout, Interaction of Arginine with Proteins and the Mechanism by Which It Inhibits Aggregation, *J. Phys. Chem. B* **114** (42), 13426–13438 (2010).

15. C. P. Schneider, D. Shukla, and B. L. Trout, Arginine and the Hofmeister Series: The Role of Ion–Ion Interactions in Protein Aggregation Suppression, *J. Phys. Chem. B* **115** (22), 7447–7458, (2011).
16. M. Rospiccio, A. Arsiccio, G. Winter, and R. Pisano, The Role of Cyclodextrins against Interface-Induced Denaturation in Pharmaceutical Formulations: A Molecular Dynamics Approach, *Mol. Pharm.*, **18** (6), 2322–2333 (2021).
17. A. Arsiccio, J. McCarty, R. Pisano, and J.-E. Shea, Effect of Surfactants on Surface-Induced Denaturation of Proteins: Evidence of an Orientation-Dependent Mechanism, *J. Phys. Chem. B* **122** (49), 11390–11399 (2018).
18. N. E. Good, G. D. Winget, W. Winter, T. N. Connolly, S. Izawa, and R. M. M. Singh, Hydrogen Ion Buffers for Biological Research, *Biochemistry* **5** (2), 467–477 (1966).
19. C. M. H. Ferreira, I. S. S. Pinto, E. V. Soares, and H. M. V. M. Soares, (Un)suitability of the use of pH buffers in biological, biochemical and environmental studies and their interaction with metal ions—a review, *RSC Adv.* **5** (39), 30989–31003 (2015).
20. B. S. Gupta, M. Taha, and M. J. Lee, Buffers more than buffering agent: Introducing a new class of stabilizers for the protein BSA, *Phys. Chem. Chem. Phys.* **17** (2), 1114–1133 (2015).
21. K. Kasraian et al., Sustained In Vivo Activity of Recombinant Bovine Granulocyte Colony Stimulating Factor (rbG-CSF) Using HEPES Buffer, *Pharm. Dev. Technol.* **6** (3), 441–447 (2001).
22. M. A. Metrick, J. E. Temple, and G. Macdonald, The effects of buffers and pH on the thermal stability, unfolding and substrate binding of RecA, *Biophys. Chem.* **184**, 29–36 (2013).
23. B. W. Petersen, T. J. Harms, M. G. Reynolds, and L. H. Harrison, Use of Vaccinia Virus Smallpox Vaccine in Laboratory and Health Care Personnel at Risk for Occupational Exposure to Orthopoxviruses — Recommendations of the Advisory Committee on Immunization Practices (ACIP), *MMWR. Morb. Mortal. Wkly. Rep.* **65** (10), 257–262 (2016).
24. S. M. Kelly, T. J. Jess, and N. C. Price, How to study proteins by circular dichroism, *Biochim. Biophys. Acta - Proteins Proteomics* **1751** (2), 119–139 (2005).
25. J. Arora et al., Charge-mediated Fab-Fc interactions in an IgG1 antibody induce reversible self-association, cluster formation, and elevated viscosity, *MAbs*, **8** (8), 1561–1574 (2016).
26. B. Webb and A. Sali, Comparative Protein Structure Modeling Using MODELLER, *Curr. Protoc. Bioinforma.* **54** (1), 5-6 (2016).

27. R. Anandkrishnan, B. Aguilar, and A. V Onufriev, H⁺⁺ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations, *Nucleic Acids Res.* **40**, W537–W541 (2012).
28. H. D.A. Case, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, D. Ghoreishi, G. Giambasu, T. Giese, M.K. Gilson, H. Gohlke, A.W. Goetz, D. Greene, R Harris, N. Homeyer, Y. Huang, S. Izadi, A. Kovalenko, C. L. Nguyen, A. Onufriev, F. Pan, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, Y. Simmerling, J. Smith, J. Swails, R.C. Walker, J. Wang, H. Wei, L. Wilson, R.M. Wolf, X. Wu, L. Xiao, D. M. Y. Xiong, and P. A. Kollman, AMBER 2019. University of California, San Francisco, (2019).
29. W. Humphrey, A. Dalke, and K. Schulten, VMD: Visual molecular dynamics, *J. Mol. Graph.*, **14** (1), 33–38 (1996).
30. Y. Ma et al., Determination of the second virial coefficient of bovine serum albumin under varying pH and ionic strength by composition-gradient multi-angle static light scattering, *J. Biol. Phys.*, **41** (1), 85–97 (2015).
31. D. Some and S. Kenrick, Characterization of Protein-Protein Interactions via Static and Dynamic Light Scattering. *Protein Interactions*, InTech, 401–426 (2012).
32. D. P. Teufel, C. M. Johnson, J. K. Lum, and H. Neuweiler, Backbone-Driven Collapse in Unfolded Protein Chains, *J. Mol. Biol.*, **409** (2), 250–262 (2011).
33. S. Doose, Importance of Backbone and Solvent Properties for Conformational Dynamics in Polypeptides, *Chem Phys Chem*, **9** (18), 2687–2689 (2008).
34. N. E. Good and S. Izawa, Hydrogen Ion Buffers, *Methods Enzymol.* **24** (C), 53–68 (1972).
35. R. L. Remmele Jr., N. S. Nightlinger, S. Srinivasan, and W. R. Gombotz, Interleukin-1 receptor (IL-1R) liquid formulation development using differential scanning calorimetry, *Pharmaceutical Research*, **15** (2), 200–208 (1998).
36. K. Huynh and C. L. Partch, Analysis of protein stability and ligand interactions by thermal shift assay, *Curr. Protoc. protein Sci.* **79** (1), 28–9 (2015).
37. M. Taha, B. S. Gupta, I. Khoiroh, and M. J. Lee, Interactions of biological buffers with macromolecules: The ubiquitous ‘smart’ polymer PNIPAM and the biological buffers MES, MOPS, and MOPSO, *Macromolecules*, **44** (21), 8575–8589 (2011).
38. A. Zubriene et al., Measurement of nanomolar dissociation constants by titration calorimetry and thermal shift assay - Radicol binding to Hsp90 and ethoxzolamide binding to CAII, *Int. J. Mol. Sci.*, **10** (6), 2662–2680 (2009).

39. M. Vivoli, H. R. Novak, J. A. Littlechild, and N. J. Harmer, Determination of protein-ligand interactions using differential scanning fluorimetry, *J. Vis. Exp.* **91**, 1–13 (2014).
40. S. V. Thakkar et al., Excipients differentially influence the conformational stability and pretransition dynamics of two IgG1 monoclonal antibodies, *J. Pharm. Sci.*, **101** (9), 3062–3077 (2012).
41. S. Amin, G. V. Barnett, J. A. Pathak, C. J. Roberts, and P. S. Sarangapani, Protein aggregation, particle formation, characterization & rheology, *Curr. Opin. Colloid Interface Sci.*, **19** (5), 438–449 (2014).
42. T. A. Khan, H.-C. Mahler, and R. S. K. Kishore, Key interactions of surfactants in therapeutic protein formulations: A review., *Eur. J. Pharm. Biopharm.* **97**, 60–67 (2015).
43. S. Fekete, A. Beck, J. Veuthey, and D. Guillarme, Ion-exchange chromatography for the characterization of biopharmaceuticals, *J. Pharm. Biomed. Anal.* **113**, 43–55 (2015).
44. Y. Du et al., Chromatographic analysis of the acidic and basic species of recombinant monoclonal antibodies, *MAbs* **4** (5), 578–585 (2012).

Chapter 7 - Mechanistic understanding of the stability of the anti-MYC mini-protein Omomyc in different formulation conditions

Berner, C.¹, Castillo Cano, V.², Beaulieu, M.E.², Winter, G.¹

¹Department of Pharmacy, Ludwig-Maximilians-Universität München, Munich, Germany

²Peptomyc S.L., Barcelona, Spain

Author contribution:

CB., V.C.C., M.E.B. and G.W. conceived the idea. C.B. performed the experiments, evaluated the data, and wrote the manuscript. G.W., M.E.B and V.C.C. provided conceptual guidance. G.W. and M.E.B. corrected the manuscript.

7.1. Introduction

Peptide- and protein-based therapeutics have become increasingly important for cancer therapy due to their high efficacy and limited side effects.¹ The c-MYC oncoprotein (hereafter referred to as MYC) is overexpressed in most human cancers which makes it a promising target for therapeutic intervention. MYC is a transcription factor regulating several genes important for cell growth, differentiation, metabolism, and apoptosis upon association with its partner MAX (Figure 49).² Both MYC and MAX are intrinsically disordered and contain a basic helix-loop-helix leucine zipper (bHLHLZ) domain that allows their dimerization and binding of DNA.³ Upon dimerization the individual monomers undergo coupled folding and binding which is a widely observed feature of intrinsically disordered proteins (IDPs) enabling dynamic interactions with multiple partners with modest affinities.⁴ The MYC-MAX heterodimer interacts with Ebox DNA via its highly positive charged basic region and thus activates or represses gene transcription.⁵ Since intrinsically disordered proteins lack an “active site” that could be efficiently targeted by a small molecule drug, the development of an inhibitor with high specificity has been challenging. However, a previously designed dominant-negative form of MYC called Omomyc has been shown to inhibit MYC function both in vitro and in vivo.⁶ Omomyc is a 91-amino acid mini-protein that mimics the MYC bHLHLZ domain and contains four amino acid mutations (E57T, E64I, R70Q, and R71N) in the leucine zipper domain to remove electrostatic clashes and thereby enable homodimerization as well as heterodimerization with MYC and MAX. Omomyc homodimers as well as Omomyc-MAX heterodimers can bind DNA, competing with the DNA-binding of MYC-MAX heterodimers. Additionally, direct binding of Omomyc to MYC results in its sequestration from DNA. This in turn leads to the repression of MYC-dependent gene transcription (Figure 49).⁶⁻⁸

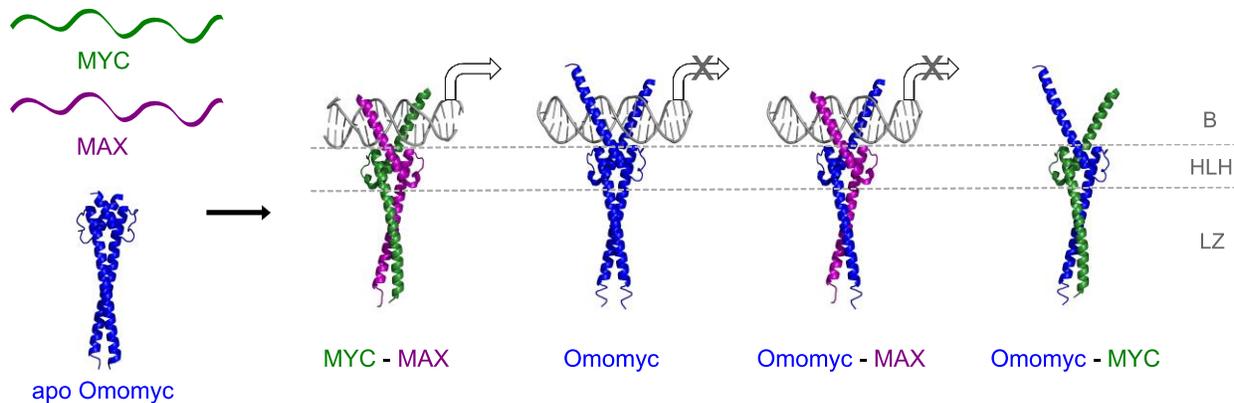


Figure 49: Omomyc dimerizes with MYC and MAX and represses MYC-dependent gene transcription. Representation of the crystal structures of apo Omomyc (5I4Z), the MYC-MAX heterodimer (1NKP), Omomyc bound to DNA (5I50), and models of the Omomyc-MYC and Omomyc-Max heterodimers. The models of Omomyc-MYC and Omomyc-Max were constructed by superposition of the crystal structures of Omomyc and MYC-MAX. All structures contain a basic region (B), a helix-loop-helix region (HLH), and a leucine zipper (LZ), which are crucial for dimerization and DNA binding. The basic region of apo Omomyc was not resolved in the crystal structure. Disordered MYC and MAX fold upon dimerization, bind DNA and recruit co-regulators of transcription. Omomyc homodimers and heterodimers are capable of binding DNA but do not recruit regulators necessary for gene transcription.⁶⁻⁸

Based on its significant *in vivo* activity and safety,⁹ Omomyc is currently being evaluated in clinical trials. However, for a successful translation into a safe and efficacious drug product, an understanding of its physicochemical behavior in different conditions is essential. So far, a detailed biophysical characterization has not been described.³ Due to the structural flexibility of IDPs, both their physical and chemical properties are unique compared to ordered proteins.¹⁰ The conformational integrity of ordered proteins is usually maintained by three major stabilizing forces, namely electrostatic interactions, hydrophobic interactions, and hydrogen bonding.¹¹ In IDPs these forces play a key role in intermolecular protein-protein interactions which are essential for their biological function.¹² The Omomyc homodimer and heterodimers with MYC and MAX for example are reportedly stabilized by a combination of multiple electrostatic as well as hydrophobic interactions.⁸ However, these interactions can potentially also lead to the formation of protein aggregates, which may reduce the therapeutic activity and result in an immunogenic reaction in the patient.¹³ Protein aggregation and other degradation pathways are however not exclusively based on the protein's intrinsic properties but largely depend on solution conditions such as pH, ionic

strength, and added excipients.¹⁴ The solution pH controls the type and distribution of surface charges, which affect both conformational and colloidal stability of a protein.¹⁵ Furthermore, ionic strength can influence the protein stability by charge-screening, which could lead to enhanced hydrophobic interactions, or by altering the protein's conformational state.¹¹ Therefore, the first step in formulation development is to find the optimal pH and ionic strength. Since buffering agents may have a diverse effect on protein stability, the choice of an ideal buffer system is another critical factor. Buffers not only maintain the pH but can increase the conformational stability and alter the colloidal stability. This effect can be due to either change in the ionic strength or direct binding to the protein.¹⁶ Preliminary studies revealed a substantial influence of phosphate buffer on the stability of Omomyc. Due to their high charge density, phosphate anions can bind to positively charged amino acid residues and consequently alter the charge distribution on the protein surface which has been shown to either stabilize proteins or cause protein aggregation.^{17–19}

In this work, we investigate the conformational and colloidal stability of Omomyc in different formulation conditions with a combination of protein characterization techniques and MD simulations. We evaluate the role of electrostatics in homo- and heterodimerization and the effect of pH, ionic strength, and phosphate on the conformational ensemble as well as the self-association behavior of Omomyc. We found that electrostatics are the main contributor to differences in protein-protein binding free energies of Omomyc to its interaction partners. We observed ion-induced folding of Omomyc and reentrant condensation of the protein in the presence of phosphate. High phosphate concentrations lead to Omomyc stabilization.

7.2. Materials and methods

7.2.1. Materials

The bulk Omomyc solution contained 10.0 g/L protein and was kindly provided from Peptomyc S.L.. The protein concentration was measured spectrophotometrically using a NanoDrop 2000 (Thermo Fisher Scientific, Wilmington, USA) and an extinction coefficient at 280 nm of $0.138 \text{ (mg/mL)}^{-1}\text{cm}^{-1}$. All chemicals were of molecular biology or multicompendial grade and were purchased either from Sigma or Thermo Fisher Scientific (Germany). All solutions were prepared with ultrapure water from a Sartorius arium[®] pro system (Sartorius Corporate Administration GmbH, Göttingen, Germany).

7.2.2. Sample Dialysis and Preparation

The buffer was exchanged by extensive dialysis to the respective buffer at the given pH (10 mM sodium acetate at pH 4.8, pH 5.8, and pH 6.8; or 1, 10, 25, 50, 100, 150, and 200 mM sodium phosphate buffer at pH 6.8) for 24 hours at 2-8 °C using Slide-A-Lyzer™ MINI Dialysis Devices (cutoff 3.5 kDa, Thermo Fisher Scientific, Germany). The samples were collected in microcentrifuge tubes and centrifuged at 10,000 x g for 10 minutes and subsequently filtered with a 0.2 µm cellulose acetate membrane filter (Whatman, FP 30/0.2 CA-S, GE Healthcare, Buckinghamshire, UK). Stock solutions of sodium chloride were prepared in the respective buffer and spiked into the dialyzed protein stock to prepare samples with an ionic strength of 100 mM. For measurements that required higher protein concentrations, the protein solutions were upconcentrated using Vivaspin 20 5 MWCO PES centrifugal concentrators (Sartorius Lab Instruments, Goettingen, Germany). The concentration was measured again, and the solutions were sterile filtered with 0.2 µm cellulose acetate membrane filters.

7.2.3. Circular Dichroism (CD) Spectroscopy

Far-UV circular dichroic spectra were collected at 25 °C with a Jasco J-810 spectropolarimeter (JASCO Deutschland GmbH, Pfungstadt, Germany). All samples contained 0.5 g/L of protein. Quartz cuvettes (Hellma GmbH, Muellheim, Germany) with a 1 mm wavelength path were used for the measurements. 5 accumulations of each sample were taken at a speed of 20 nm/min. The spectrum of the respective buffer was subtracted for each sample and smoothing of the spectra was performed using the Savitzky-Golay algorithm with 9 smoothing points.²⁰ The mean residue ellipticity (MRE) of the protein at each wavelength was calculated. To elucidate the thermal unfolding of Omomyc under different conditions, a thermal ramp of 1 °C/min was applied from 20 °C – 85 °C while the ellipticity at 222 nm was monitored. The data were fitted to a two-state unfolding model using the CDpal software.²¹

7.2.4. Dynamic Light Scattering (DLS)

To determine the interaction parameter k_D , samples with protein concentrations from 1 to 7 g/L were prepared. 10 µL of each sample were pipetted in triplicates into a 1536 well plate (Aurora Microplates, Whitefish, USA). The plate was centrifuged at 2000 rpm for 2 min using a Heraeus Megafuge 40 centrifuge equipped with an M-20 well plate rotor (Thermo Fisher Scientific, Wilmington, USA). Two microliters of silicon oil were added to seal each well. The plate was

centrifuged again and placed in a DynaPro DLS plate reader III (Wyatt Technology, Santa Barbara, USA). All measurements were performed at 25 °C with 10 acquisitions per well and an acquisition time of 5 s. The data were analyzed with the Dynamics V7.10 software (Wyatt Technology, Santa Barbara, USA).

7.2.5. Differential Scanning Fluorimetry (nanoDSF)

NanoDSF was used to study the thermal unfolding and aggregation of Omomyc at varying phosphate concentrations. Samples with 1 g/L of protein were filled in standard nanoDSF™ grade capillaries, and the capillaries were sealed. A temperature ramp of 1 °C/min from 20 to 80 °C was applied with the Prometheus NT.48 (NanoTemper Technologies, Munich, Germany) system that measures the intrinsic protein fluorescence intensity at 330 and 350 nm after excitation at 280 nm. Simultaneously, the device detects aggregation/precipitation of the samples by measuring the back-reflection intensity of a light beam that passes through the capillary. The apparent protein melting temperatures (T_m) were determined with the PR. ThermControl software V2.1 (NanoTemper Technologies, Munich, Germany) from the maximum of the first derivatives of the thermal unfolding curves. The same software was used to determine the aggregation onset temperature (T_{agg}) from the increase in the signal from the aggregation detection optics.

7.2.6. Molecular dynamics simulations

The full structure of the Omomyc homodimer is available from the PDB code 5i50, and the structure of the MYC-MAX heterodimer was obtained from the PDB code 1NKP. The double-stranded DNA was removed from the PDB files and the models of Omomyc-MYC and Omomyc-Max were constructed by superposition of the crystal structures of Omomyc and MYC/MAX. The protonation states of ionizable residues at the respective pH were adjusted using the H++ server.²² Phosphate was parametrized with antechamber using GAFF2 for bonded and non-bonded parameters. Atomic partial charges were calculated with the AM1-BCC charge model in antechamber.

All-atom simulations were performed with the Amber20 program²³ in a periodic box with explicit solvent. The ff14SB force field for proteins was employed in combination with the TIP3P water model. Tleap was used to solvate the Omomyc homodimers at different pH in a cubic water box with and without 100 mM sodium chloride and at pH 6.8 with different phosphate concentrations, respectively. All bonds involving hydrogen atoms were constrained using the SHAKE algorithm.

Non-bonded electrostatic interactions were treated using the particle mesh Ewald algorithm with a direct space cutoff of 9 Å. The system was energy minimized with the steepest descent algorithm for the first 5000 cycles, followed by 5000 cycles using the conjugate gradient method. System equilibration was carried out for 1 ns in NVT ensemble to stabilize the temperature of 300 K using the Langevin thermostat, and subsequently for 1 ns in NPT ensemble to adjust the density of the system using the Berendsen barostat. The simulations were performed for 100 or 200 ns with a time step of 2 fs. The coordinates were saved every 10 ps. All trajectories were analyzed using the CPPTRAJ module of Amber20 and VMD.²⁴ To estimate the binding of phosphate molecules to the protein, we calculated the interaction score probability ($P(I_{\text{score}})$) defined as

$$P(I_{\text{score}}) = n / (N \cdot a)$$

where n is the number of frames phosphate was in contact with a certain residue over simulation time, N is the total number of frames, and a is a normalization factor to convert the interaction score into a probability.

7.2.7. Calculation of binding free energy by MM/PBSA

The binding free energy for each system was calculated with the MMPBSA.py script in AMBER with the default parameters. The last 100 frames were extracted from the 100 ns MD trajectories described above. The simulations must be long enough for the system to reach convergence which was measured using RMSD (Figure S17b). The binding free energy ($\Delta G_{\text{bind,solv}}$) is calculated by:

$$\Delta G_{\text{bind,solv}} = \Delta G_{\text{bind,vacuum}} + \Delta G_{\text{solv,complex}} - (\Delta G_{\text{solv,ligand}} + \Delta G_{\text{solv,receptor}})$$

The different contributions to the binding free energy above are calculated as follows:

$$\Delta G_{\text{solv}} = \Delta G_{\text{electrostatic},\epsilon=80} - \Delta G_{\text{electrostatic},\epsilon=1} + \Delta G_{\text{hydrophobic}}$$

$$\Delta G_{\text{vacuum}} = \Delta E_{\text{molecular mechanics}} - T \cdot \Delta S_{\text{normal mode analysis}}$$

The conformational entropy upon binding ($-T\Delta S$) is omitted from the overall binding free energy estimate, since the entropic contribution calculation is associated with high computational costs.²⁵

7.3. Results and Discussion

7.3.1. Control of electrostatics is critical in homo- and heterodimerization of MYC, MAX, and Omomyc

To develop a stable drug product, it is crucial to maintain the intrinsic stabilizing forces and even enhance the stability of the therapeutic protein by finding optimal formulation conditions. To accomplish this task, the responsible forces and the effects of environmental conditions must first be evaluated. We therefore elucidate which forces control the homo- and heterodimerization of Omomyc, MYC, and MAX and determine the effect of the pH on Omomyc homodimers in MD simulations. MD simulations have found widespread application as a complement to experiments since they provide molecular mechanistic insights into conformational dynamics, protein-protein interactions, and protein stabilization by excipients.^{26–28} Here, we use energy-based features to reveal the binding mechanism of the individual monomers. We perform all-atom MD simulations and calculate the protein-protein binding free energy with the molecular mechanics/Poisson-Boltzmann surface area (MM/PBSA) method. MM/PBSA allows the comparison of energy contributions from different types of interactions, namely electrostatics ($\Delta E_{\text{electrostatic}}$) and van der Waals interactions (ΔE_{vdw}), the desolvation free energy (ΔG_{sol}), and the conformational entropy ($-T\Delta S$).²⁹ MYC, MAX, and Omomyc become structured upon binding to their respective partners. In this case, the energy from specific interactions compensates for the entropic penalty from ordering.³⁰ Since the time scale of our MD simulations would not allow folding of the proteins, simulations started from the fully folded homo- and heterodimers (without DNA). Therefore, the calculations of binding free energies do not account for the additional entropic contribution from protein folding. Furthermore, standard MM/PBSA does not consider the screening effect of ions on the electrostatic interactions between proteins to evaluate the binding free energy.²⁹ We were therefore not able to determine the influence of ionic strength on the homo- and heterodimerization *in silico*.

The MM/PBSA calculations discriminate between the strongly binding dimers and the MYC homodimer ($\Delta G = 6.9 \pm 9.9$ kcal/mol), which was previously shown to be unstable and bind DNA poorly.⁷ The method reveals different binding affinities for the Omomyc homodimers and heterodimers with MYC and MAX, but their statistical uncertainties overlap (Figure 50a, Table S5). Our calculations suggest that the Omomyc-MYC heterodimer and the Omomyc homodimer

at physiological pH are the most stable with average binding free energies of -43.2 ± 10.9 and -42.3 ± 3.7 kcal/mol, respectively. The MYC-MAX and Omomyc-MAX heterodimers show similar binding affinities of -33.0 ± 1.6 and -32.3 ± 1.0 kcal/mol. The determined binding free energies for the Omomyc homodimer and MYC-MAX heterodimer are in line with previously published ΔG values estimated with the PISA tool ($\Delta G = -41$ kcal/mol and -33 kcal/mol, respectively).³¹ In contrast to our findings, Jung et al. also proposed that heterodimers of Omomyc with MYC and MAX would contain repulsive interactions or lack stabilizing interactions that would lower their stability. Another study could not determine any heterodimerization of MYC and Omomyc at all but suggested that the cellular effect of Omomyc is rather due to competition of Omomyc homodimers with the MYC-MAX heterodimer for DNA binding.³ Interestingly, we observe high statistical uncertainties in the ΔG calculations for the Omomyc-MYC heterodimer suggesting the probability of less optimal stabilizing interactions. Furthermore, lowering the pH destabilizes the Omomyc homodimer in our MD simulations, as average binding free energies decrease from -42.3 ± 3.7 kcal/mol at pH 6.8 to -35.6 ± 2.8 kcal/mol at pH 5.8 to -26.7 ± 6.4 kcal/mol at pH 4.8. This finding suggests that the charge of the protein has a strong impact on the dimer stability. Hence, we examined the individual contributions to the total binding free energies. The van der Waals energy is a stabilizing contribution to the total binding affinity of each dimer but does not significantly differ between the interaction partners (Figure S17a). The only exception is the MYC homodimer, which shows slightly less favorable van der Waals energies compared to the other dimers. Since the dimerization interfaces of leucine zipper domains are enriched in hydrophobic residues that pack in a regular ‘knobs and holes’ pattern, we expected the van der Waals energies to contribute most of the stabilizing energy.³² Not only did the mutation of residue E64 of MYC to isoleucine in Omomyc remove an electrostatic clash but also introduced an additional hydrophobic interaction in the leucine zipper domain which leads to more stabilizing van der Waals contributions in the Omomyc dimers. Furthermore, we calculated the electrostatic contributions as the sum of the electrostatic energy ($\Delta E_{\text{electrostatic}}$) and the electrostatic contributions to the desolvation free energy (ΔG_{PB}). Although compensated for by the desolvation free energy, the electrostatic energy remains unfavorable in all cases. The same trend as for the total binding free energies can be observed for the electrostatics alone. This shows that electrostatics are the main contributor to differences in binding free energies of Omomyc homo- and heterodimers. Our results agree with previous reports that the specificity of leucine zipper interactions is dictated by the charged residues that flank the core hydrophobic interface.³² Electrostatic potential maps of MYC,

MAX, and Omomyc monomers reveal the cause for the observed differences (Figure 50c). Naturally, the basic region comprises high positive potential to allow binding of the negatively charged DNA backbone, but previous work showed that the N-terminal α -helix is not involved in the specificity of dimerization.³³ MYC contains a high number of both positive and negative charges in the dimerization interface. These high charges impede efficient homodimerization due to electrostatic clashes. The MAX monomer lacks the interfacial electrostatic repulsion characteristic for MYC but instead presents an optimal complementarity to the MYC interface region.⁹ The interaction site of Omomyc was designed to remove the electrostatic clashes occurring in the MYC homodimer and therefore has less charged residues. Some long charged amino acids (glutamic acid, lysine, and arginine) are however necessary in the dimerization interface to prevent the formation of higher order oligomers and to ensure specificity of dimerization.³² Omomyc can consequently homodimerize and bind to both MYC and MAX without detrimental effects of electrostatics. Previous work showed that Omomyc homodimers and heterodimers with MYC and MAX are indeed stabilized by a combination of both electrostatic and hydrophobic interactions.⁸ We can confirm that hydrophobic interactions positively contribute to the binding energies while electrostatic interactions are unfavorable in sum. It is therefore essential to control the electrostatics of the proteins to enable efficient dimerization and stability.

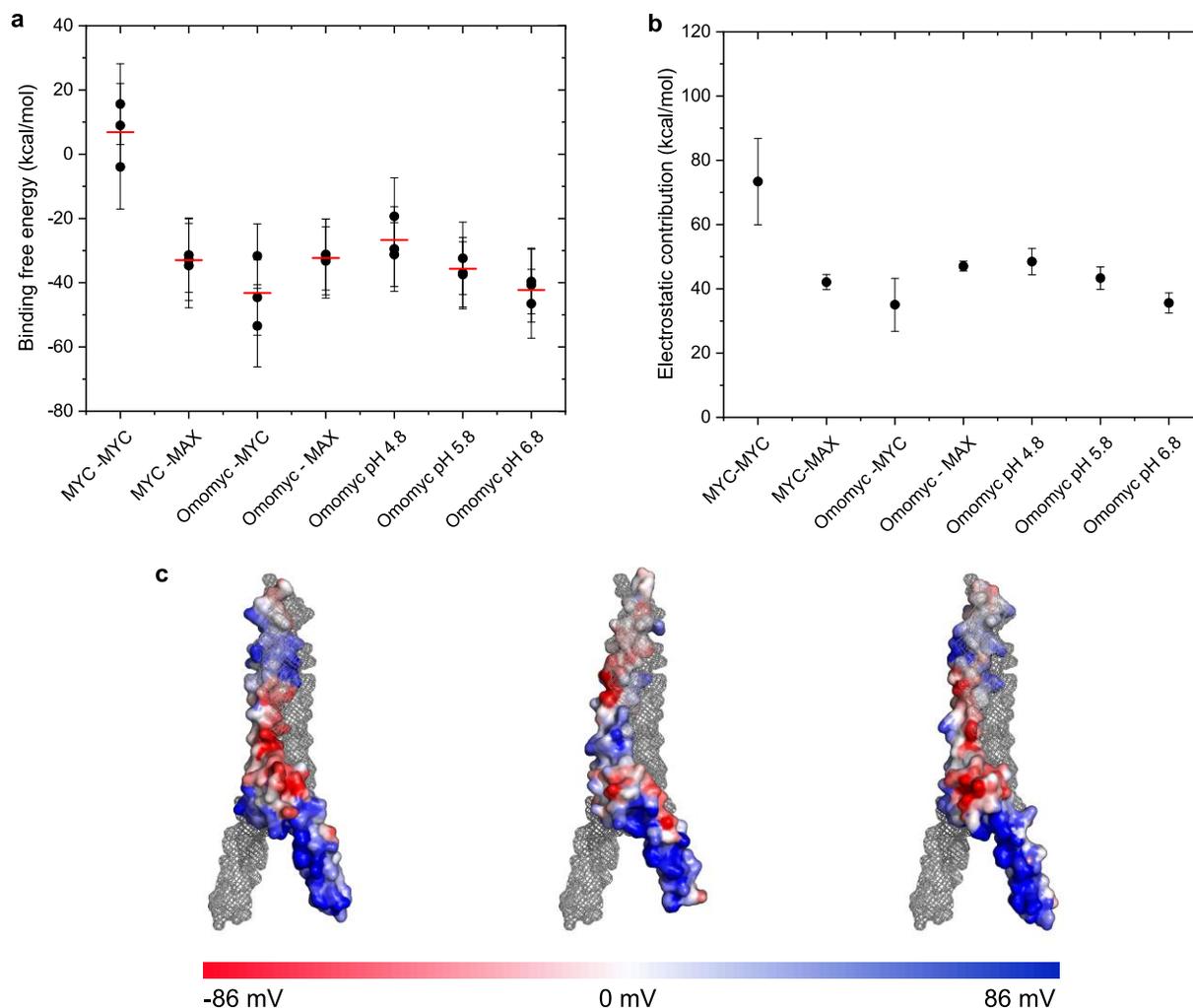


Figure 50: Binding free energies for MYC, MAX, and Omomyc homo- and heterodimers. (a) Protein-protein binding free energies for MYC, MAX, and Omomyc homo- and heterodimers were calculated with MM/PBSA from three independent trajectories each. The mean of the triplicates is shown in red. (b) Average of the electrostatic contributions to the binding free energies is calculated as the sum of the electrostatic energy ($\Delta E_{\text{electrostatic}}$) and the electrostatic contributions to the solvation free energy (ΔG_{PB}). (c) Electrostatic potential maps for monomers of MYC (left), MAX (center), and Omomyc (right). The interaction partners are shown in mesh representation to visualize the binding interfaces.

7.3.2. Effect of pH and ionic strength on Omomyc structure and stability

Since it was shown that pH and ionic strength may result in opposite changes in conformational and colloidal stabilities, these parameters need to be evaluated simultaneously.³⁴ We therefore examined the structure as well as thermal and colloidal stability of the protein using a combination of biophysical techniques at pH 4.8, 5.8, and 6.8 at low and high ionic strength (adjusted with

NaCl). To exclude potential buffer effects on the protein, we used sodium acetate buffer for all pH values considered in this part of the work even though the buffering capacity is not given at pH 6.8.

First, we investigated the effect of pH and ionic strength on the secondary structure of Omomyc. The characteristic far-UV CD spectra with two minima at 209 and 222 nm confirm that Omomyc is α -helical in all conditions and shows a slight decrease of helical content when increasing the pH from 4.8 to pH 5.8 and 6.8 (Figure 51a). Noteworthy, the ratios of the mean residue ellipticities at 222 nm and 209 nm $[\theta]_{222}/[\theta]_{209}$ are in the range of 0.81 to 0.84, which are typical for non-interacting α -helices.³⁵ Increasing the ionic strength to a total of 100 mM at each pH produces CD spectra with ratios of $[\theta]_{222}/[\theta]_{209}$ in the range of 0.97 to 0.99, which indicate the presence of two-stranded α -helical coiled-coils.³⁵ The presence of NaCl also leads to an increased α -helical content, which is reflected by the more negative ellipticity. Similar behavior has been described for MYC in the absence and presence of MAX. The MYC bHLH domain is unstructured in the monomeric form but forms an α -helical structure upon dimerization with MAX. The low ionic strength spectra of Omomyc resemble the reported spectrum of MYC monomers, whereas the high ionic strength conditions result in spectra like the one reported for the MYC-MAX complex.³ Furthermore, Lavigne et al. observed that the stability and helicity of the MYC-MAX complex were also modulated by pH, with a maximum helicity around pH 4.5.³⁶ Since the basic region could not be resolved in the apo Omomyc crystal structure (Figure 49), probably due to high flexibility, we assume that it is also unstructured without DNA interaction. However, our results show that Omomyc undergoes ion-induced folding which again highlights the importance of favorable interfacial electrostatic interactions in stabilizing the homodimer.

Next, the stability of Omomyc under the tested conditions was monitored by the decrease in ellipticity at 222 nm during thermal denaturation from 20 °C – 85 °C (Figure 51b). Omomyc at low ionic strength shows a very shallow transition in negative ellipticity with increasing temperature, while there is no difference between the different pH values. This gradual, less cooperative denaturation has already been demonstrated for the MYC bHLHLZ domain.³⁷ In contrast, the high ionic strength conditions show a cooperative decrease in α -helical content with no significant difference in melting temperature (T_m) between pH 5.8 and 6.8 (36.3 ± 1.2 °C and 34.5 ± 1.8 °C) but a shift to a higher T_m at pH 4.8 (41.8 ± 0.9 °C). NanoDSF was used as a complementary tool

to measure the protein melting temperatures and the aggregation onset temperatures (T_{agg}) and shows the same trend (Figure S18a, Table S6).

Furthermore, the colloidal stability of Omomyc was assessed by means of the interaction parameter k_D derived from DLS measurements. The k_D was determined from the concentration dependence of the translational diffusion coefficient D_t of the protein (Figure 51c). In all low ionic strength conditions, the D_t increases with increasing protein concentration, reflecting a positive k_D , regardless of the pH. The high k_D s can be attributed to repulsive double-layer forces under low ionic strength conditions.³⁸ This is in agreement with the relatively high positive charge of the protein. It is estimated to be +17 e at pH 4.8, +14 e at pH 5.8, and +11 e at pH 6.8. Upon addition of NaCl, the translational diffusion coefficient decreases with increasing protein concentration. NaCl was previously shown to compress the double layer and reduce the range of the force causing more attractive protein-protein interactions.^{38,39} Correspondingly, the interaction parameter k_D is highest at low ionic strength and lowest pH and decreases with increasing pH and ionic strength (Figure 51d). This trend can also be observed in the aggregation during heating determined with nanoDSF backscattering (Figure S18b). The strong electrostatic repulsion inhibits the aggregate growth in all low ionic strength samples so that the aggregates do not reach a size that is detectable by the backscattering detector. Only the conditions at pH 5.8 and 6.8 with an ionic strength of 100 mM show aggregation with T_{agg} values of 78.19 ± 0.70 °C and 75.55 ± 0.63 °C, respectively.

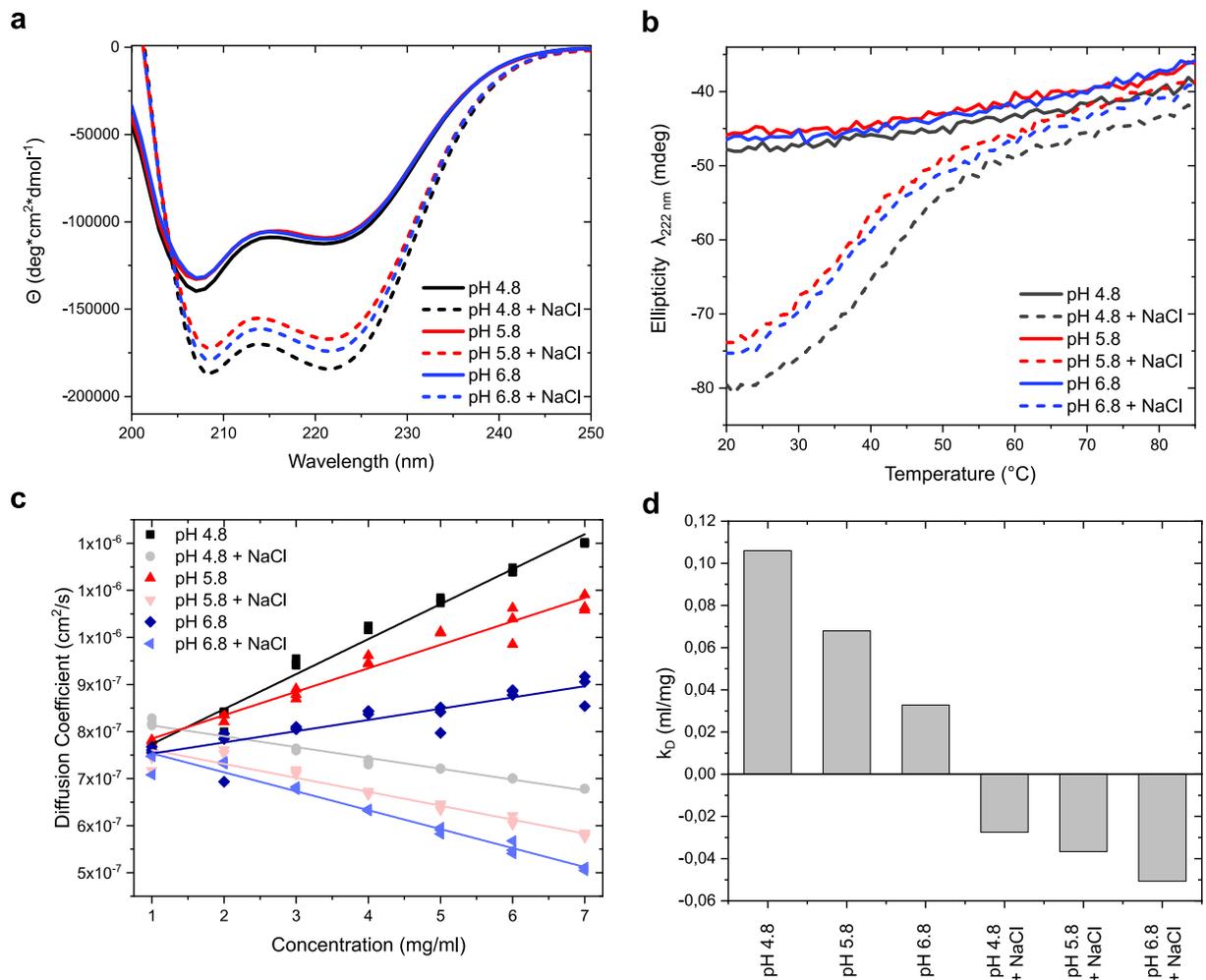


Figure 51: Structure, thermal and colloidal stability of Omomyc at different pH and ionic strength. (a) Far-UV CD spectra of Omomyc at pH 4.8, 5.8, and 6.8 with and without addition of NaCl. (b) Thermal unfolding curves of Omomyc at the different conditions monitored by the ellipticity at 222 nm. (c) Effect of pH and ionic strength on the translational diffusion coefficient D_t at different protein concentrations assessed with DLS. (d) Diffusion interaction parameters k_D derived from the slope of the concentration dependence of D_t .

Additional MD simulations of fully folded Omomyc homodimers at pH 4.8, 5.8, and 6.8 with NaCl were conducted to explore the stability and dynamics of the protein in different ionic strength conditions. The root mean square fluctuations (RMSF) show high flexibility of the basic region for all pH values both at low and high ionic strength whereas the leucine zipper domain remains relatively stable (Figure S18c). High flexibility is usually correlated to lower protein stability. As mentioned above, the basic region could not be resolved in the crystal structure of apo Omomyc

due to its high flexibility. Therefore, this behavior was expected from the basic region without DNA interaction and additional repulsion of the positively charged residues in each monomer. A comparison of the RMSF between the different systems indicates that the addition of NaCl only slightly affects the flexibility of the basic region and the helix-loop-helix region. However, we are not accounting for the ion-induced folding of the basic region in our simulations. We can only conclude that the basic region, the termini, and the loop are very dynamic in all conditions and seem to be an important factor in protein stability.

In summary, at low ionic strength, Omomyc displays only residual helicity, which slightly increases under acidic conditions but significantly extends at high ionic strength. Additionally, low pH leads to charge repulsion of the Omomyc molecules and less self-association compared to physiological pH. The addition of NaCl screens the positive charges on the protein surface causing self-association.

7.3.3. Concentration-dependent influence of phosphate buffer on the Omomyc stability

Preliminary data revealed that phosphate buffer significantly affects Omomyc's structure and stability. We therefore investigated the effect of varying phosphate concentrations on the protein with experimental and computational techniques. Already during dialysis into phosphate buffer, protein precipitation occurred at 10 and 25 mM phosphate.

First, the effect of phosphate on the secondary structure was examined by CD spectroscopy. Here, we can again observe characteristic far-UV CD spectra with two minima at 209 and 222 nm that show a significant increase in negative ellipticity with increasing sodium phosphate concentrations (Figure 52a). Even at a concentration of 1 mM phosphate, the α -helical content increases to a similar extent as could be seen for the high ionic strength condition in acetate buffer (10 mM acetate + NaCl to achieve total ionic strength of 100 mM). This indicates that it is indeed a specific buffer effect and not only a matter of ionic strength. The highest helical content was reached at 200 mM sodium phosphate. Additionally, the ratios of the mean residue ellipticities at 222 nm and 209 nm $[\theta]_{222}/[\theta]_{209}$ shift from 0.84 in 10 mM sodium acetate buffer to 0.88 in 1 mM sodium phosphate and a range of 0.97 – 1.00 for phosphate concentrations from 10 to 200 mM. These results suggest that phosphate is inducing an extension of the α -helix and facilitates the assembly of the two monomers into a two-stranded α -helical coiled-coil.

Next, we used CD spectroscopy, nanoDSF, and dynamic light scattering to measure T_m , T_{agg} , and the interaction parameters k_D to elucidate the impact of phosphate on the Omomyc stability (Figure 52b, Figure S19, Table S7). The thermal unfolding of Omomyc in different phosphate concentrations from 20 °C – 85 °C measured by CD reveals significant differences between the conditions. Omomyc in 1 mM sodium phosphate shows an almost linear decrease in negative ellipticity with increasing temperature similar to the low ionic strength conditions at pH 4.8, 5.8, and 6.8. Since the unfolding curve for 10 mM sodium phosphate does not show a sigmoidal transition, it could not be fitted with the CDpal software to determine the T_m . It can still be seen that the temperature at which unfolding starts, is lowest for 10 mM sodium phosphate. Higher sodium phosphate concentrations result in a cooperative decrease in α -helical content with a significant increase in melting temperature with increasing concentration (from 46.4 to 59.2 °C). The complementary nanoDSF measurements show the same trend. It must be noted though, that the 10 and 25 mM conditions were measured at lower protein concentrations in nanoDSF. Since the protein concentration influences heat-induced protein aggregation, this in turn affects the melting transitions.⁴⁰ If aggregation occurs around the melting temperature of the protein, the T_m will shift to lower values.⁴¹ Therefore, it is not possible to directly compare the T_m s given by nanoDSF and CD measurements at 10 and 25 mM phosphate, as the melting transitions may be affected by concentration-dependent protein aggregation. The same applies to the T_{agg} values measured with nanoDSF (Figure S19).

Additionally, the influence of phosphate on the colloidal stability of Omomyc was assessed by means of the interaction parameter k_D (Figure 52c, d). The k_D was not measured for 10 and 25 mM phosphate buffer because the required protein concentration could not be reached due to precipitation. The translational diffusion coefficients in 1 mM phosphate could not be fitted to determine the k_D due to non-linearity. However, the slope is positive indicating protein-protein repulsion at this 1 mM phosphate. The repulsive protein-protein interactions decrease quickly from low to intermediate phosphate concentrations resulting in negative k_D s. With increasing phosphate concentrations, the k_D s become less negative and almost plateau at concentrations of 150 to 200 mM phosphate. Noteworthy, the translational diffusion coefficient is dependent on the hydrodynamic radius of the protein. Since Omomyc is partly unstructured at low ionic strength conditions and folds in the presence of NaCl or phosphate, the determined k_D s must be interpreted with caution. However, the D_0 values for all conditions are almost identical for the measurements.

This indicates that the hydrodynamic radius at infinite dilution is the same which means that the structural differences are probably neglectable.

Altogether, Omomyc shows very different behavior at low (1 mM), intermediate (10 – 25 mM), and high (50 – 200 mM) phosphate concentrations. At low phosphate concentration, the protein is not fully structured and protein-protein repulsion between the positively charged Omomyc molecules is the reason for the high colloidal stability. Intermediate phosphate concentrations result in precipitation of the protein. Further increasing the phosphate concentration causes the protein precipitate to resolubilize. This phenomenon is a characteristic feature of reentrant condensation, which typically occurs between two critical anion concentrations.^{38,42} We propose that the different concentration-dependent effects of phosphate are due to interactions with protein side chains, at low, intermediate, and high concentrations. Phosphate binding to positively charged residues will cause neutralization of the long-ranged repulsive electrostatic interactions which in turn allows the formation of short-ranged ion-bridging attractive forces between the proteins. Subsequently, these attractive forces could lead to precipitation.⁴² Further binding of phosphate anions to the protein surface will induce charge inversion and thus restore protein-protein repulsion.

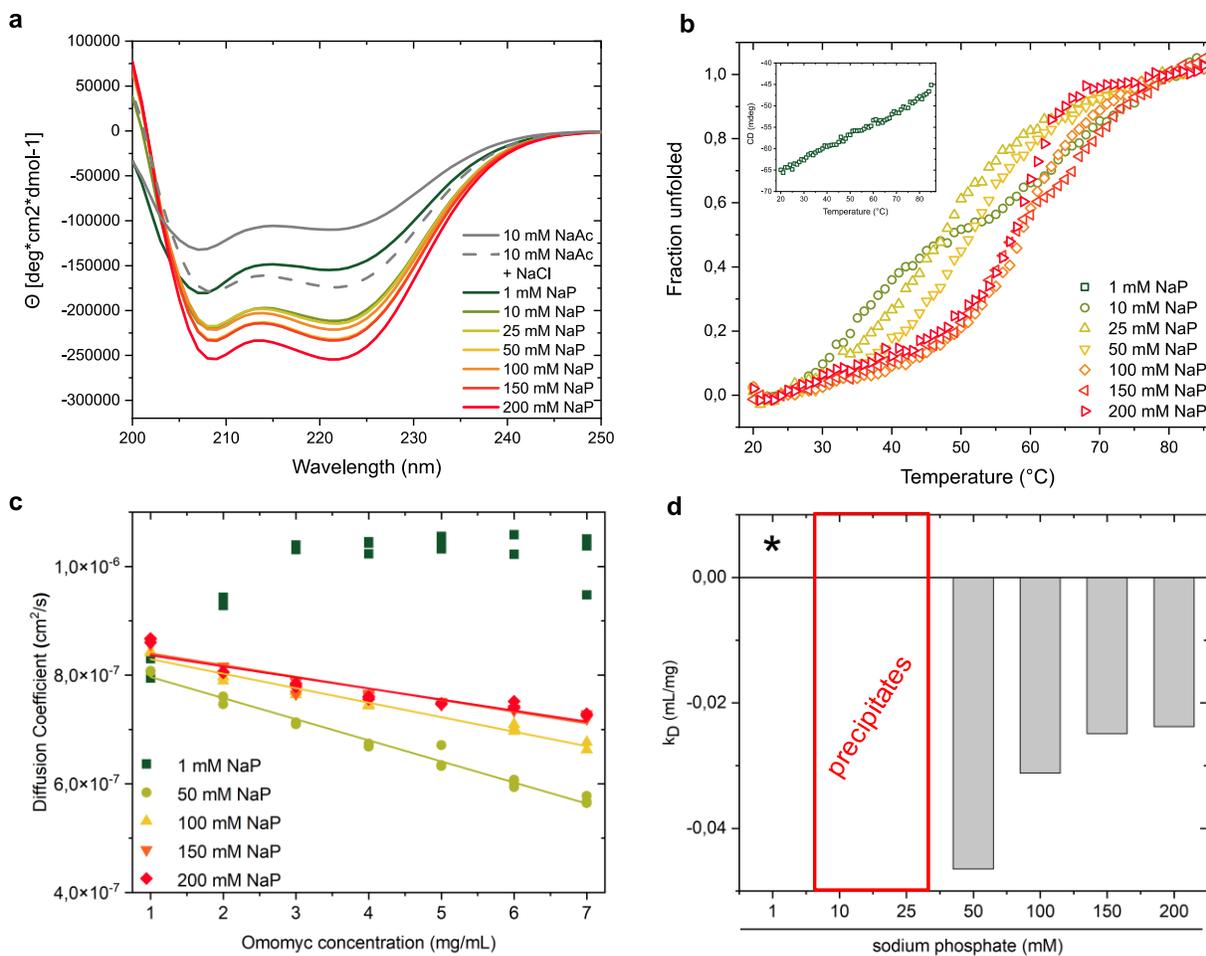


Figure 52: Effect of different phosphate concentrations on the structure and stability of Omomyc at pH 6.8. (a) Far-UV CD spectra of Omomyc at pH 6.8 in 10 mM sodium acetate buffer with and without NaCl or sodium phosphate buffer with increasing concentrations. (b) Normalized thermal unfolding curves of Omomyc in different sodium phosphate concentrations were monitored by the ellipticity at 222 nm. The insert shows the raw CD signal for the unfolding of Omomyc in 1 mM sodium phosphate which could not be normalized. (c, d) Diffusion interaction parameter k_D assessed with DLS. The asterisk indicates that the k_D could not be determined from the slope of the translational diffusion coefficients. The k_D could not be measured for samples with 10 and 25 mM sodium phosphate due to precipitation of the protein during dialysis.

7.3.4. Phosphate binds to Omomyc in a concentration-dependent manner in MD simulations

To understand the mechanisms by which phosphate causes reentrant condensation of Omomyc at intermediate concentrations and stabilization at higher concentrations, we conducted 100 ns simulations of the protein in simulation boxes containing 1, 10, 25, 50, 100, 150, and 200 mM sodium phosphate, respectively. Dihydrogen phosphate (H_2PO_4^-) and hydrogen phosphate (HPO_4^{2-}) were added to the simulation with a ratio of 1 to allow for comparison between the binding frequency and strength of the two species. Trajectories were analyzed in terms of binding probabilities ($P(I_{\text{score}})$) of phosphate to the individual residues of the protein and resulting structural changes in Omomyc (Figure S20, Table S7). The $P(I_{\text{score}})$ is a measure of how long contact between a phosphate molecule and a residue of Omomyc was present during the simulation and is calculated based on a simple distance cut-off. Although one phosphate molecule typically interacts with one or two protein residues, it can have contact with up to seven residues. The average number of contacts per phosphate molecule was reported to be 3.5 ± 2.3 .⁴³ In combination with the applied distance cut-off approach, this results in more interaction sites than phosphate molecules present in the simulation.

In particular, the basic region, the first helix, the loop, and the first residues of helix 2 show relatively high $P(I_{\text{scores}})$ already at intermediate phosphate concentrations (Figure 53a, d). In contrast, the leucine zipper domain only exhibits a few interaction sites even at high concentrations (Figure 53b, d). This is expected since the MYC/MAX/DNA structure confirmed the critical role of the several positively charged residues in the basic region, the loop, and helix 2 for the interaction with the DNA backbone phosphates in the biological context.⁵ The trend in all conditions is that residues R9, R10, R17, R19, and R20 of the basic region, residue K24 in helix 1, and residue K45 in the loop are common residues in both monomers forming salt bridges with phosphate ions (interaction in at least 5 of the 7 conditions). The positively charged guanidinium group of arginine can form multiple hydrogen bonds and is therefore well suited to bind phosphate. The ϵ -amino group of the lysine contains a positive charge and can participate in salt bridges and hydrogen bonds. It is therefore not surprising that most phosphate interactions occur with arginine and lysine residues. Additionally, the N-terminal methionine residue binds phosphate in the higher concentration range via its free amino group. Hydrogen bonding of phosphate to asparagine and glutamate residues was also detected, however less frequently. The large size of the phosphate molecule enables it to bridge gaps between chains. Thus, residues in contact with phosphate can

be quite distant in the protein sequence and even distant in space since residues commonly involved in binding phosphate (arginine and lysine) are bulky as well.⁴³ Consequently, phosphate seems to mediate the interaction between the basic and HLH region of the two Omomyc monomers but also has the potential to cross-link several monomers together.³⁸

Noteworthy, once phosphate binds to one of the residues at intermediate concentrations, the interaction seems to be stable throughout the simulations as indicated by the relatively high $P(I_{\text{scores}})$. In contrast, the interaction probability rarely exceeds 70 % at 150 or 200 mM phosphate but shows several interactions below the 20 % threshold (Figure S20f, g). Since the probability of a contact is higher due to the larger number of phosphates in the simulation box, it is unlikely that it takes longer for a phosphate ion to reach the interaction site. It is more plausible that the interactions are not as long-lasting which might be due to crowding of the phosphate anions on the protein surface and resulting repulsion.

The increasing number of residues in contact with phosphate with increasing concentration shows that the surface is not saturated in the evaluated concentration range (Figure 53c). We however observe different binding affinities of dihydrogen phosphate and hydrogen phosphate. Dihydrogen phosphate binds to Omomyc significantly less compared to hydrogen phosphate, especially at low and intermediate phosphate concentrations. This is expected based on the different charges of the molecules. While the two negative charges of hydrogen phosphate can each form a salt bridge to a positive charge on the protein surface, dihydrogen phosphate with only one negative charge can participate in just one electrostatic interaction. Noteworthy, at 200 mM phosphate the difference is considerably smaller. The number of Omomyc residues in contact with hydrogen phosphate barely increases from 150 to 200 mM phosphate (from 34 to 38) whereas the number of interactions with dihydrogen phosphate more than doubles (from 14 to 31). This again supports our hypothesis that crowding of the hydrogen phosphate anion with its higher charge density on the Omomyc surface impedes the approach of additional hydrogen phosphate ions due to charge repulsion. Comparing the RMSF of the simulations with different phosphate concentrations shows no influence of phosphate on the mobility of the different regions of Omomyc (Figure S20h). We again observe generally high flexibility of the basic region and a stable leucine zipper domain for all conditions. We propose that phosphate binding only alters the electrostatic interactions between the proteins, which also explains that anion-specific differences are reduced at higher concentrations. This effect has also been shown when comparing the sodium salts of phosphate, sulfate, and citrate.³⁸

Moreover, it has already been reported that the MYC and MAX dimer formation was favored by the presence of negatively charged ions such as poly-l-glutamic acid.⁴⁴

In summary, the MD simulations support the hypothesis that phosphate induces reentrant condensation of Omomyc. The process is driven by neutralization of the protein charge by predominant binding of hydrogen phosphate to arginine and lysine residues on the protein surface which allows the formation of ion-bridging attractive forces. Increasing the phosphate concentration results in an accumulation of hydrogen phosphate and to some extent dihydrogen phosphate molecules on the protein surface causing charge inversion and overcharging of the protein, thereby restoring repulsion between the protein molecules. Previous studies revealed that the presence of multivalent ions causes RC not only of globular proteins such as lysozyme but also of IDPs.^{38,42} The multivalent ions do not only regulate solution properties of IDPs but influence their conformational ensemble, for example, causing compaction of the IDP histatin 5.⁴² We can observe both phenomena for Omomyc in the presence of phosphate anions.

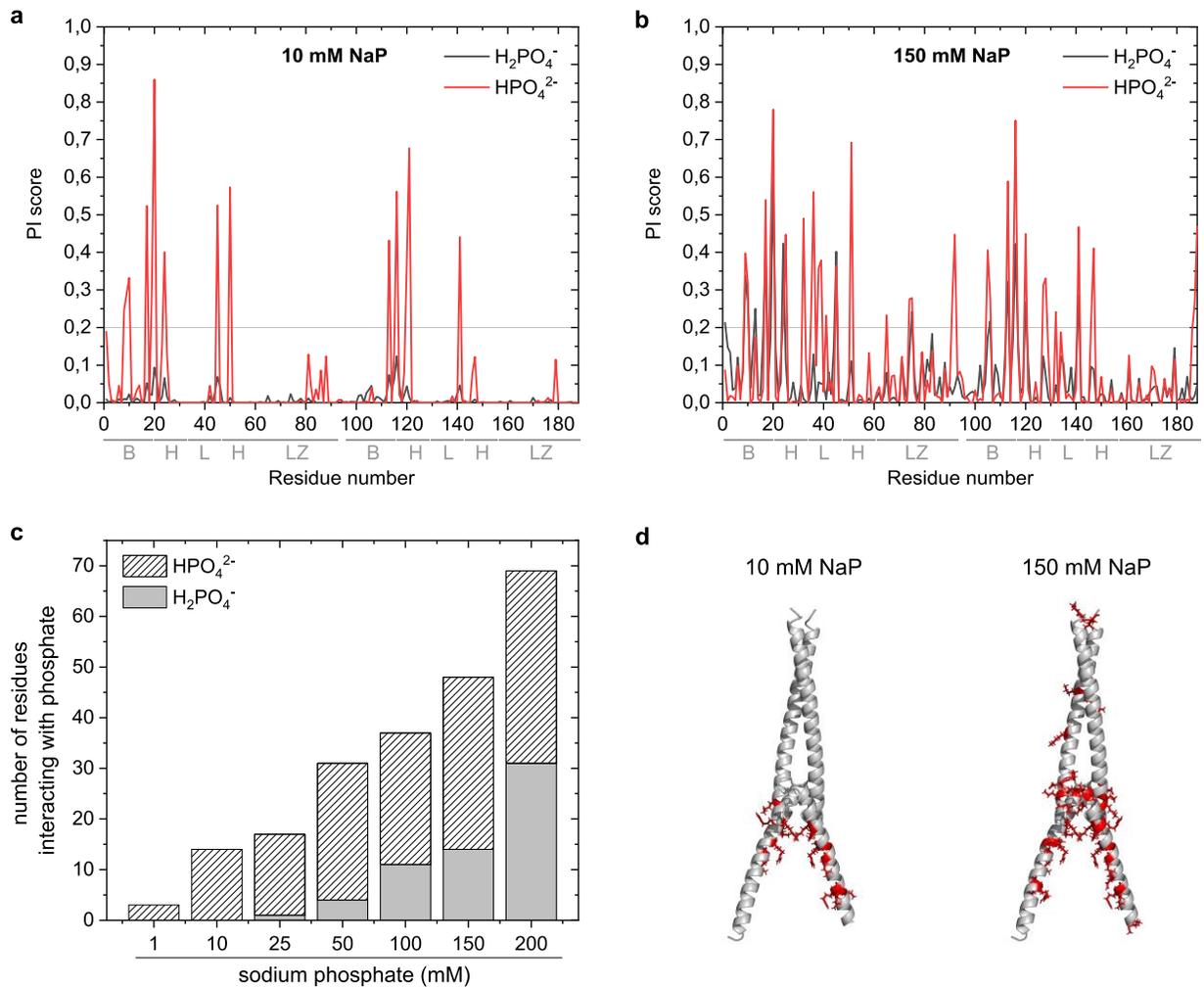


Figure 53: Binding of hydrogen phosphate and dihydrogen phosphate to Omomyc in MD simulations.

(a, b) Interaction score probabilities $P(I_{\text{score}})$ per residue for H_2PO_4^- and HPO_4^{2-} averaged over three independent MD simulations with 10 and 150 mM sodium phosphate, respectively, reveal the binding regions. The $P(I_{\text{score}})$ gives a percentage of how long an interaction of phosphate with each residue was present during the simulation. We used a $P(I_{\text{score}})$ cut-off of 0.2 to distinguish transient contacts from actual binding (indicated by a gray line). The basic (B), helix-loop-helix (HLH), and leucine zipper (LZ) domains are shown for both chains of the homodimer. (c) The number of residues interacting with either H_2PO_4^- or HPO_4^{2-} for $\geq 20\%$ of the simulation show a concentration-dependent effect. The numbers include the residues of both monomers. (d) Representation of the phosphate binding sites (both H_2PO_4^- and HPO_4^{2-}) at 10 and 150 mM highlighted in red in the Omomyc structure.

7.4. Conclusions

We present an extensive study on the stability of the anti-MYC mini-protein Omomyc and elucidate the driving forces of dimerization as well as the molecular mechanisms by which the pH, ionic strength, and phosphate influence the protein structure and self-association tendency. MM/PBSA calculations revealed that van der Waals energies contributed most of the stabilizing energy whereas electrostatics are the main contributor to differences in binding free energies of Omomyc, MYC, and MAX homo- and heterodimerization. Omomyc undergoes ion-induced folding in the presence of sodium chloride and phosphate anions. Additionally, at acidic conditions and low ionic strength, repulsion of the Omomyc molecules prevent self-association. Sodium chloride screens the positive charges on the protein surface leading to more attractive interactions. Our findings reveal that electrostatics must be precisely controlled to achieve conformational as well as colloidal stability of Omomyc since high ionic strength induces folding but also diminishes repulsive intermolecular forces. We furthermore found that phosphate induces reentrant condensation of Omomyc. Omomyc precipitates at concentrations of 10 and 25 mM phosphate due to binding of phosphate anions to arginine and lysine residues on the protein surface and subsequent neutralization of the repulsive electrostatic forces. Higher phosphate concentrations in contrast stabilize Omomyc by overcharging of the molecules which in turn leads to protein-protein repulsion. Even though Omomyc shows higher conformational and colloidal stability at acidic pH, the use of high concentrations of phosphate at pH 6.8 induces enhanced folding of the unstructured bHLH domain and increases the colloidal stability.

7.5. Supporting information

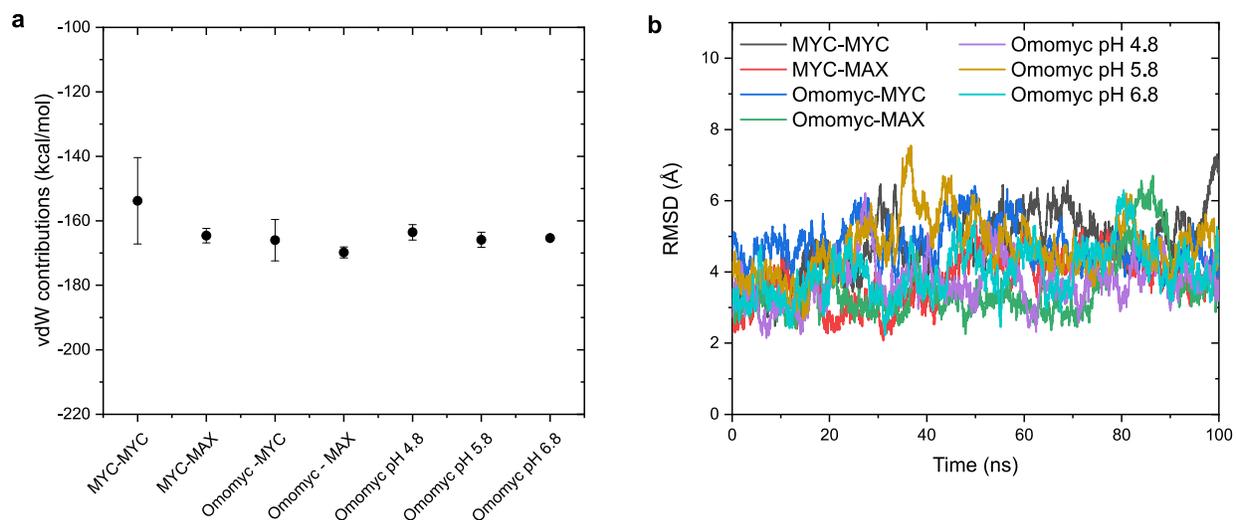


Figure S17: MM/PBSA binding free energy calculations. (a) Average of the van der Waals contributions to the binding free energies. (b) The average RMSD of the triplicates was used to confirm that the MD simulations have converged. The RMSDs for all systems fluctuate over simulation time but the average remains in a narrow range. The fluctuation is probably due to the flexibility of the basic region without DNA-binding.

Table S5: Binding free energies for MYC, MAX, and Omomyc homo- and heterodimers. Individual contributions of the different energy contributions and the resulting total ΔG values are determined with the MM/PBSA method. Calculations are shown separately for each of the triplicates.

| Energy component | MYC-MYC | | | MYC-MAX | | | Omomyc_MYC | | | Omomyc-MAX | | |
|---------------------|-----------|-----------|-------------------|-----------|-----------|-------------------|------------|-----------|-------------------|------------|-----------|-------------------|
| | Average | Std. Dev. | Std. Err. of Mean | Average | Std. Dev. | Std. Err. of Mean | Average | Std. Dev. | Std. Err. of Mean | Average | Std. Dev. | Std. Err. of Mean |
| E_{vdw} | -144.0067 | 7.2390 | 0.7203 | -162.2802 | 7.5321 | 0.7495 | -159.3676 | 6.6635 | 0.6630 | -167.9284 | 6.4996 | 0.6467 |
| $E_{electrostatic}$ | 587.7248 | 60.2313 | 5.9932 | 195.1954 | 59.8414 | 5.9544 | 556.2185 | 63.2778 | 6.2964 | 285.9861 | 39.2178 | 3.9023 |
| E_{PB} | -528.7803 | 56.2631 | 5.5984 | -155.3539 | 58.1739 | 5.7885 | -516.2222 | 59.4367 | 5.9142 | -237.5097 | 38.4214 | 3.8231 |
| $E_{nonpolar}$ | -120.4982 | 4.3684 | 0.4347 | -131.5212 | 4.4503 | 0.4428 | -131.0352 | 4.0147 | 0.3995 | -127.5408 | 3.4271 | 0.3410 |
| E_{disper} | 201.6006 | 5.6872 | 0.5659 | 221.1480 | 6.2033 | 0.6173 | 218.7245 | 4.8828 | 0.4859 | 215.7790 | 3.6536 | 0.3635 |
| ΔG_{gas} | 443.7181 | 59.3478 | 5.9053 | 32.9152 | 60.7234 | 6.0422 | 396.8510 | 63.0498 | 6.2737 | 118.0577 | 40.4438 | 4.0243 |
| ΔG_{solv} | -447.6779 | 54.9936 | 5.4721 | -65.7270 | 59.3585 | 5.9064 | -428.5329 | 58.9774 | 5.8685 | -149.2714 | 38.1832 | 3.7994 |
| ΔG_{total} | -3.9598 | 13.0515 | 1.2987 | -32.8119 | 12.7321 | 1.2669 | -31.6819 | 10.0133 | 0.9964 | -31.2137 | 11.0737 | 1.1019 |
| E_{vdw} | -148.2921 | 8.5829 | 0.8540 | -166.7391 | 7.6418 | 0.7604 | -172.1751 | 7.1986 | 0.7163 | -170.2578 | 6.4144 | 0.6383 |
| $E_{electrostatic}$ | 812.3670 | 65.3322 | 6.5008 | 138.0476 | 52.9679 | 5.2705 | 442.5023 | 60.5644 | 6.0264 | 294.5054 | 39.6530 | 3.9456 |
| E_{PB} | -736.6426 | 59.0118 | 5.8719 | -96.1686 | 48.4871 | 4.8246 | -402.9182 | 56.3655 | 5.6086 | -249.1144 | 36.1074 | 3.5928 |
| $E_{nonpolar}$ | -120.4061 | 4.8152 | 0.4791 | -132.8198 | 3.7952 | 0.3776 | -143.1123 | 3.4491 | 0.3432 | -129.1865 | 3.5809 | 0.3563 |
| E_{disper} | 208.5723 | 5.8941 | 0.5865 | 223.0364 | 4.0236 | 0.4004 | 231.1262 | 3.9155 | 0.3896 | 220.8312 | 4.1481 | 0.4127 |
| ΔG_{gas} | 664.0748 | 64.0839 | 6.3766 | -28.6915 | 52.5416 | 5.2281 | 270.3272 | 61.8135 | 6.1507 | 124.2476 | 39.2961 | 3.9101 |
| ΔG_{solv} | -648.4764 | 57.5752 | 5.7289 | -5.9520 | 48.3135 | 4.8074 | -314.9043 | 56.4570 | 5.6177 | -157.4697 | 36.3297 | 3.6149 |
| ΔG_{total} | 15.5985 | 12.5914 | 1.2529 | -34.6435 | 13.1223 | 1.3057 | -44.5771 | 11.7567 | 1.1698 | -33.2221 | 10.5831 | 1.0531 |
| E_{vdw} | -169.0609 | 7.8255 | 0.7787 | -164.7739 | 9.1157 | 0.9070 | -166.4176 | 7.6167 | 0.7579 | -171.2053 | 7.0071 | 0.6972 |
| $E_{electrostatic}$ | 786.0268 | 64.5411 | 6.4221 | 208.9222 | 80.1524 | 7.9755 | 429.8543 | 65.0353 | 6.4713 | 354.4666 | 49.8897 | 4.9642 |
| E_{PB} | -700.4934 | 61.2399 | 6.0936 | -164.3624 | 76.0523 | 7.5675 | -404.3098 | 59.9060 | 5.9609 | -307.1685 | 46.7303 | 4.6498 |
| $E_{nonpolar}$ | -135.9174 | 3.8659 | 0.3847 | -131.8799 | 4.2100 | 0.4189 | -139.9882 | 3.4352 | 0.3418 | -129.2205 | 4.0509 | 0.4031 |
| E_{disper} | 228.4056 | 4.7859 | 0.4762 | 220.6932 | 5.6244 | 0.5596 | 227.4272 | 3.6615 | 0.3643 | 220.6507 | 4.5591 | 0.4536 |
| ΔG_{gas} | 616.9660 | 65.1044 | 6.4781 | 44.1484 | 78.0419 | 7.7655 | 263.4368 | 64.3920 | 6.4072 | 183.2613 | 49.7839 | 4.9537 |
| ΔG_{solv} | -608.0052 | 61.0592 | 6.0756 | -75.5490 | 75.4031 | 7.5029 | -316.8708 | 59.0944 | 5.8801 | -215.7383 | 46.6272 | 4.6396 |
| ΔG_{total} | 8.9608 | 13.0639 | 1.2999 | -31.4006 | 11.5458 | 1.1489 | -53.4340 | 12.7567 | 1.2693 | -32.4770 | 12.2936 | 1.2233 |

| Energy component | Omomyc pH 4.8 | | | Omomyc pH 5.8 | | | Omomyc pH 6.8 | | |
|---------------------|---------------|-----------|-------------------|---------------|-----------|-------------------|---------------|-----------|-------------------|
| | Average | Std. Dev. | Std. Err. of Mean | Average | Std. Dev. | Std. Err. of Mean | Average | Std. Dev. | Std. Err. of Mean |
| E_{vdW} | -162.3683 | 7.7244 | 0.7686 | -167.4883 | 6.0592 | 0.6029 | -164.7744 | 7.1714 | 0.7136 |
| $E_{electrostatic}$ | 846.5921 | 49.6997 | 4.9453 | 682.2781 | 61.6954 | 6.1389 | 572.9592 | 33.9660 | 3.3797 |
| EPB | -802.1789 | 43.8820 | 4.3664 | -638.2144 | 56.2930 | 5.6014 | -534.9262 | 30.6031 | 3.0451 |
| $E_{nonpolar}$ | -130.2669 | 4.4495 | 0.4427 | -132.4847 | 3.9590 | 0.3939 | -131.3588 | 3.0475 | 0.3032 |
| E_{disper} | 214.6358 | 5.1065 | 0.5081 | 216.5205 | 4.0860 | 0.4066 | 216.3146 | 4.1762 | 0.4155 |
| ΔG_{gas} | 686.5766 | 48.5063 | 4.8266 | 517.1862 | 62.3545 | 6.2045 | 410.3683 | 32.3034 | 3.2143 |
| ΔG_{solv} | -717.8100 | 44.0061 | 4.3788 | -554.1786 | 56.2257 | 5.5947 | -449.9704 | 30.3767 | 3.0226 |
| ΔG_{total} | -31.2335 | 9.9625 | 0.9913 | -36.9924 | 11.0992 | 1.1044 | -39.6020 | 10.0656 | 1.0016 |
| E_{vdW} | -166.3204 | 8.1749 | 0.8134 | -163.2150 | 6.4993 | 0.6467 | -165.5059 | 7.5690 | 0.7531 |
| $E_{electrostatic}$ | 872.8189 | 42.3182 | 4.2108 | 530.5877 | 37.6968 | 3.7510 | 602.3643 | 42.1464 | 4.1937 |
| EPB | -824.4452 | 37.8656 | 3.7678 | -491.0493 | 33.5036 | 3.3337 | -565.6904 | 38.2154 | 3.8026 |
| $E_{nonpolar}$ | -133.3267 | 4.6518 | 0.4629 | -133.8103 | 3.7843 | 0.3765 | -133.7519 | 3.8806 | 0.3861 |
| E_{disper} | 219.0308 | 4.9215 | 0.4897 | 217.6254 | 4.8361 | 0.4812 | 219.5995 | 5.1693 | 0.5144 |
| ΔG_{gas} | 709.2357 | 42.7345 | 4.2522 | 369.7785 | 36.8512 | 3.6668 | 439.1209 | 42.2342 | 4.2025 |
| ΔG_{solv} | -738.7412 | 37.5816 | 3.7395 | -407.2342 | 33.1541 | 3.2990 | -479.8428 | 38.0091 | 3.7820 |
| ΔG_{total} | -29.5054 | 13.1739 | 1.3109 | -37.4557 | 10.1764 | 1.0126 | -40.7220 | 11.5122 | 1.1455 |
| E_{vdW} | -162.0029 | 6.9079 | 0.6874 | -166.9645 | 6.9517 | 0.6917 | -165.7336 | 6.6057 | 0.6573 |
| $E_{electrostatic}$ | 885.6308 | 58.3998 | 5.8110 | 595.3236 | 43.4254 | 4.3210 | 554.8949 | 44.4020 | 4.4182 |
| EPB | -833.0491 | 50.6563 | 5.0405 | -548.8966 | 40.6924 | 4.0490 | -522.8448 | 40.2843 | 4.0084 |
| $E_{nonpolar}$ | -132.0798 | 4.2469 | 0.4226 | -135.8470 | 3.9150 | 0.3896 | -136.7948 | 3.9577 | 0.3938 |
| E_{disper} | 219.6634 | 5.4542 | 0.5427 | 221.8664 | 4.4183 | 0.4396 | 221.4723 | 4.3239 | 0.4302 |
| ΔG_{gas} | 726.1442 | 56.1277 | 5.5849 | 430.4480 | 44.1728 | 4.3954 | 391.6196 | 43.9836 | 4.3765 |
| ΔG_{solv} | -745.4655 | 49.4203 | 4.9175 | -462.8772 | 40.3820 | 4.0182 | -438.1673 | 39.3684 | 3.9173 |
| ΔG_{total} | -19.3213 | 11.9767 | 1.1917 | -32.4292 | 11.3050 | 1.1249 | -46.5477 | 10.6933 | 1.0640 |

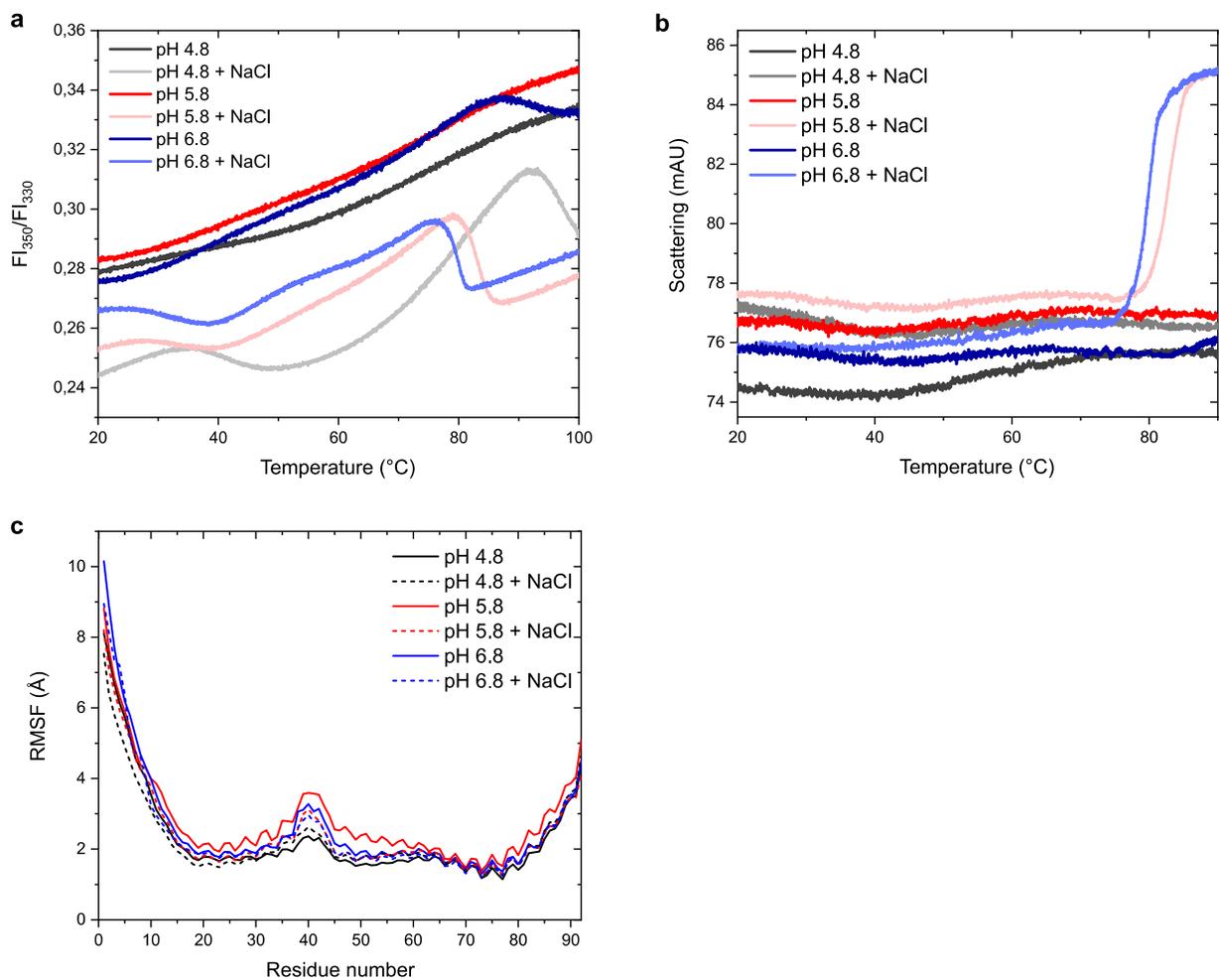


Figure S18: Stability of Omomyc at different pH and ionic strength. (a) Thermal unfolding and (b) backscattering traces from the aggregation detection optics at different pH and ionic strength measured with nanoDSF. NanoDSF is based on the protein's intrinsic fluorescence. It must therefore be noted that Omomyc does not contain any tryptophan but only tyrosine residues. (c) Root mean square fluctuations per residue of the 100 ns MD simulations indicate high flexibility of the basic and loop region of Omomyc.

Table S6: Overview of the experimental parameters determined at different pH and ionic strength.

The values are the mean of triplicates with standard deviations, except for the T_m values from CD where the error is the fit error in CDpal. The low ionic strength conditions did not show a sigmoidal transition and could not be fitted in CDpal.

| | CD | | nanoDSF | | DLS | |
|---------------|---------------------------------|----------------|------------------|------------------|---------------|----------------------------|
| | $[\theta]_{222}/[\theta]_{209}$ | T_m (°C) | T_m (°C) | T_{agg} (°C) | k_D (mL/mg) | D_0 (cm ² /s) |
| pH 4.8 | 0.84 | - | - | - | 0.106 | $6.99 \cdot 10^{-7}$ |
| pH 4.8 + NaCl | 0.99 | 41.8 ± 0.9 | 41.14 ± 0.08 | - | -0.028 | $8.36 \cdot 10^{-7}$ |
| pH 5.8 | 0.85 | - | 78.32 ± 0.76 | - | 0.068 | $7.35 \cdot 10^{-7}$ |
| pH 5.8 + NaCl | 0.97 | 36.3 ± 1.2 | 32.98 ± 0.15 | 78.19 ± 0.70 | -0.037 | $7.86 \cdot 10^{-7}$ |
| pH 6.8 | 0.86 | - | 75.04 ± 0.49 | - | 0.033 | $7.29 \cdot 10^{-7}$ |
| pH 6.8 + NaCl | 0.98 | 34.5 ± 1.8 | 31.70 ± 0.20 | 75.55 ± 0.63 | -0.051 | $7.94 \cdot 10^{-7}$ |

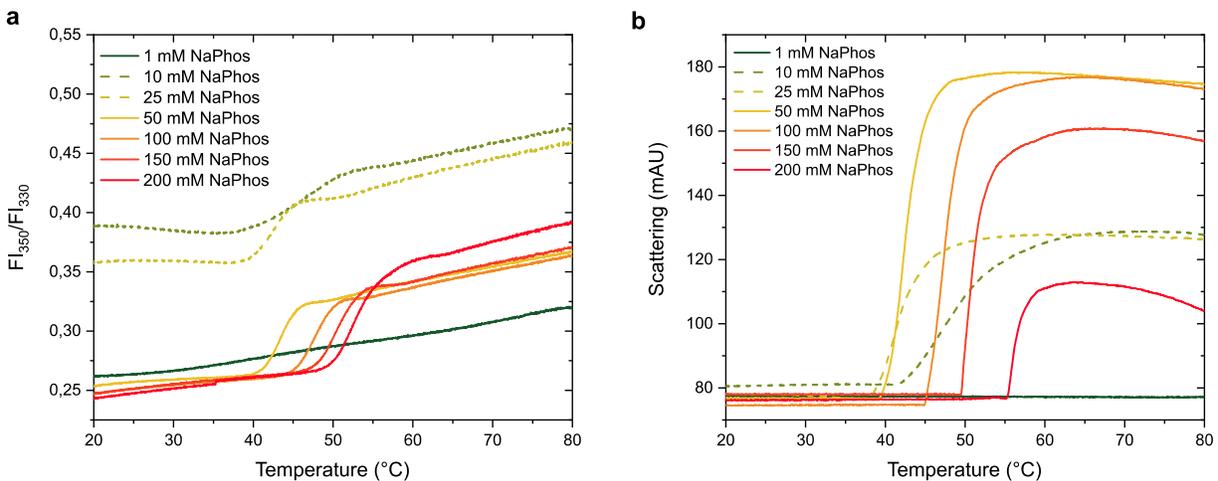


Figure S19: Thermal unfolding and aggregation of Omomyc in varying phosphate concentrations. (a)

Thermal unfolding and (b) backscattering traces from the aggregation detection optics at 1 to 200 mM phosphate with nanoDSF. The conditions with 10 and 25 mM phosphate were measured at lower protein concentrations and can therefore not directly be compared with the other conditions as well as CD measurements.

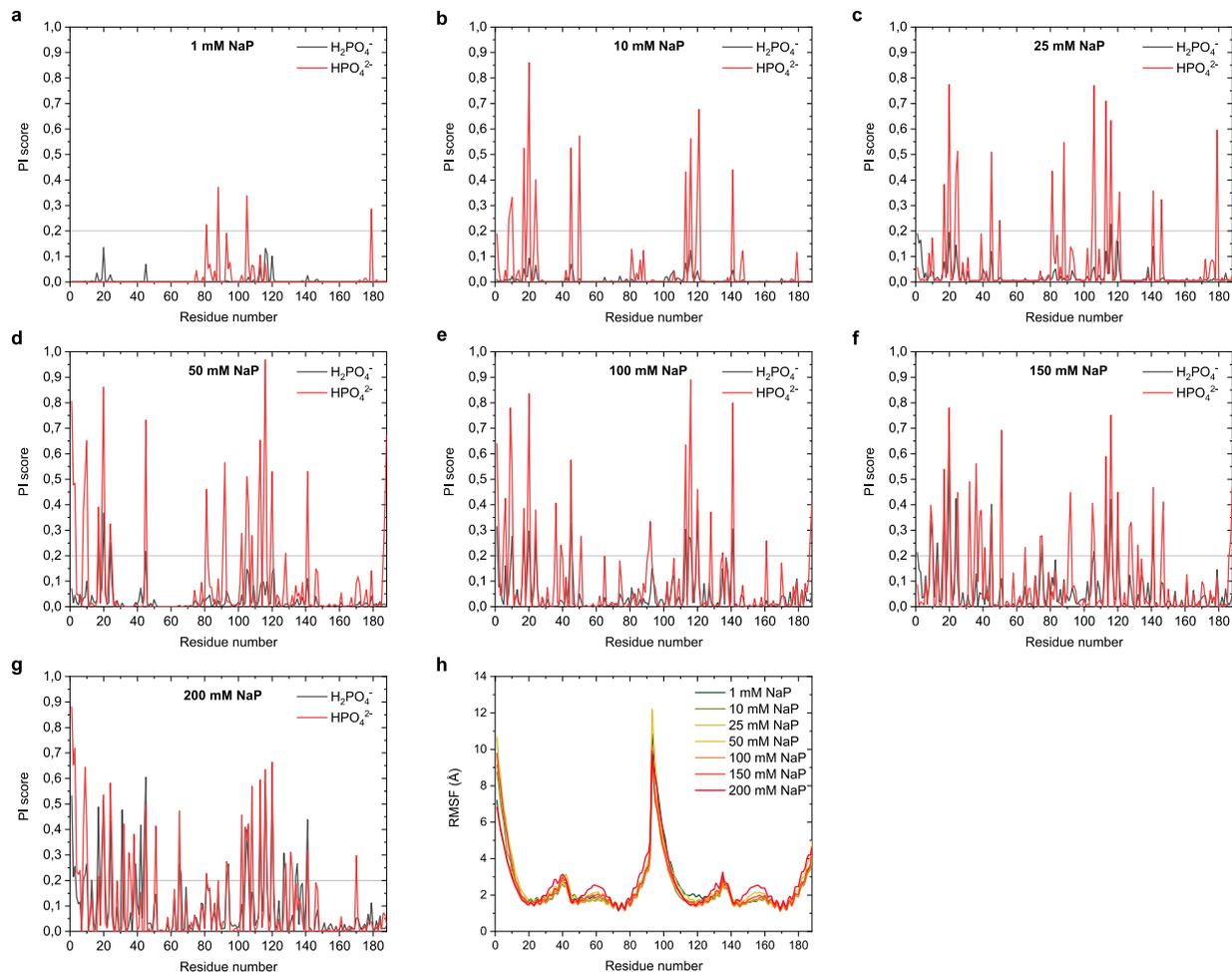


Figure S20: P(I_{score}) of hydrogen phosphate and dihydrogen phosphate and phosphate-dependent dynamics of Omomyc. (a-g) Interaction score probabilities P(I_{score}) per residue for H_2PO_4^- and HPO_4^{2-} averaged over three independent MD simulations with phosphate concentrations from 1 to 200 mM. (h) Root mean square fluctuations per residue at varying phosphate concentrations do not show large differences in flexibility upon phosphate binding.

Table S7: Overview of the experimental parameters determined at different phosphate concentrations and residues interaction with phosphate in MD simulations. The values are the mean of triplicates with standard deviations, except for the T_m values from CD where the error is the fit error in CDpal. The low ionic strength conditions did not show a sigmoidal transition and could not be fitted in CDpal. The k_D could not be determined for 10 and 25 mM phosphate due to precipitation of the protein. The slope of the diffusion coefficient at 1 mM phosphate could not be fitted due to non-linearity of the measured data. The number of residues interacting with either dihydrogen phosphate or hydrogen phosphate is calculated from the $P(I_{score})$ and are in contact with phosphate for more than 20 % of the simulation.

| | CD | | nanoDSF | | DLS | | | MD simulations | |
|------------|---------------------------------|------------|--------------|----------------|----------------|---------------|----------------------------|---|--|
| | $[\theta]_{222}/[\theta]_{209}$ | T_m (°C) | T_m (°C) | T_{agg} (°C) | T_{agg} (°C) | k_D (mL/mg) | D_0 (cm ² /s) | # residues (H ₂ PO ₄ ⁻) | # residues (HPO ₄ ²⁻) |
| 1 mM NaP | 0.88 | - | 73.62 ± 0.67 | - | 69.4 ± 1.2 | - | 8.91 · 10 ⁻⁷ | - | 3 |
| 10 mM NaP | 0.97 | - | 46.75 ± 0.18 | 41.66 ± 0.10 | 37.5 ± 0.02 | - | - | - | 14 |
| 25 mM NaP | 0.99 | 46.4 | 42.23 ± 0.05 | 37.59 ± 0.04 | 39.3 ± 0.2 | - | - | 1 | 16 |
| 50 mM NaP | 1 | 50.7 | 43.38 ± 0.03 | 39.20 ± 0.02 | 40.1 ± 0.1 | -0.047 | 8.36 · 10 ⁻⁷ | 4 | 27 |
| 100 mM NaP | 1 | 59.2 | 47.71 ± 0.05 | 44.62 ± 0.03 | 45.3 ± 0.07 | -0.031 | 8.56 · 10 ⁻⁷ | 11 | 26 |
| 150 mM NaP | 1 | 60.5 | 50.58 ± 0.02 | 49.09 ± 0.29 | 49.1 ± 0.02 | -0.025 | 8.61 · 10 ⁻⁷ | 14 | 34 |
| 200 mM NaP | 1 | 57.8 | 52.71 ± 0.02 | 53.58 ± 2.30 | 53.9 ± 0.3 | -0.024 | 8.57 · 10 ⁻⁷ | 31 | 38 |

7.6. References

1. Aldeghaither, D., Smaglo, B. G. & Weiner, L. M. Beyond peptides and mAbs - Current status and future perspectives for biotherapeutics with novel constructs. *J Clin Pharmacol* **55**, S4–S20 (2015).
2. Dang, C. v. c-Myc Target Genes Involved in Cell Growth. *Mol Cell Biol* **19**, 1–11 (1999).
3. Macek, P. et al. Myc phosphorylation in its basic helix-loop-helix region destabilizes transient-helical structures, disrupting Max and DNA binding. *Journal of Biological Chemistry* **293**, 9301–9310 (2018).
4. Metallo, S. J. Intrinsically disordered proteins are potential drug targets. *Curr Opin Chem Biol* **14**, 481–488 (2010).
5. Nair, S. K. & Burley, S. K. X-Ray Structures of Myc-Max and Mad-Max Recognizing DNA: Molecular Bases of Regulation by Proto-Oncogenic Transcription Factors. *Cell* **112** (2), 193–205 (2003).
6. Beaulieu, M. E. et al. Intrinsic cell-penetrating activity propels omomyc from proof of concept to viable anti-myc therapy. *Sci Transl Med* **11**, 1–14 (2019).
7. Soucek, L. et al. Omomyc, a potential Myc dominant negative, enhances Myc-induced apoptosis. *Cancer Res* **62**, 3507–3510 (2002).
8. Jung, L. A. et al. OmoMYC blunts promoter invasion by oncogenic MYC to inhibit gene expression characteristic of MYC-dependent tumors. *Oncogene* **36**, 1911–1924 (2017).
9. Beaulieu, M. E., Castillo, F. & Soucek, L. Structural and Biophysical Insights into the Function of the Intrinsically Disordered Myc Oncoprotein. *Cells* **9**, 1038 (2020).
10. Wang, J., Bian, Y., Cao, X. & Zhao, N. Understanding diffusion of intrinsically disordered proteins in polymer solutions: A disorder plus collapse model. *AIP Adv* **7**, 115120 (2017).
11. Wang, W. & Ohtake, S. Science and art of protein formulation development. *Int J Pharm* **568**, 118505 (2019).
12. Huang, Y. & Liu, Z. Do intrinsically disordered proteins possess high specificity in protein-protein interactions? *Chemistry - A European Journal* **19**, 4462–4467 (2013).
13. Moussa, E. M. et al. Immunogenicity of Therapeutic Protein Aggregates. *J Pharm Sci* **105**, 417–430 (2016).
14. Narayanan, H. et al. Machine Learning for Biologics: Opportunities for Protein Engineering, Developability, and Formulation. *Trends Pharmacol Sci* **42**, 151–165 (2021).
15. Wang, W., Nema, S. & Teagarden, D. Protein aggregation-Pathways and influencing factors. *Int J Pharm* **390**, 89–99 (2010).
16. Zbacnik, T. J. et al. Role of Buffers in Protein Formulations. *J Pharm Sci* **106**, 713–733 (2017).

17. Pavani, P., Kumar, K., Rani, A., Venkatesu, P. & Lee, M. J. The influence of sodium phosphate buffer on the stability of various proteins: Insights into protein-buffer interactions. *J Mol Liq* **331**, 115753 (2021).
18. Brudar, S. & Hribar-Lee, B. Effect of buffer on protein stability in aqueous solutions: A simple protein aggregation model. *Journal of Physical Chemistry B* **125**, 2504–2512 (2021).
19. Salis, A., Cappai, L., Carucci, C., Parsons, D. F. & Monduzzi, M. Specific Buffer Effects on the Intermolecular Interactions among Protein Molecules at Physiological pH. *Journal of Physical Chemistry Letters* **11**, 6805–6811 (2020).
20. Savitzky, A. & Golay, M. J. E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal Chem* **36**, 1627–1639 (1964).
21. Niklasson, M. et al. Robust and convenient analysis of protein thermal and chemical stability. *Protein Science* **24**, 2055–2062 (2015).
22. Gordon, J. C. et al. H⁺⁺: A server for estimating pK_as and adding missing hydrogens to macromolecules. *Nucleic Acids Res* **33**, 368–371 (2005).
23. Case, D. A. et al. The Amber biomolecular simulation programs. *J Comput Chem* **26**, 1668–1688 (2005).
24. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *J Mol Graph* **14**, 33–38 (1996).
25. Wong, S., Amaro, R. E. & Andrew McCammon, J. MM-PBSA captures key role of intercalating water molecules at a protein-Protein interface. *J Chem Theory Comput* **5**, 422–429 (2009).
26. Salo-Ahen, O. M. H. et al. Molecular dynamics simulations in drug discovery and pharmaceutical development. *Processes* **9**, 1–63 (2021).
27. Prabakaran, R., Rawat, P., Thangakani, A. M., Kumar, S. & Gromiha, M. M. Protein aggregation: in silico algorithms and applications. *Biophys Rev* **13**, 71–89 (2021).
28. Indrakumar, S. et al. Development of a fast screening method for selecting excipients in formulations using MD simulations, NMR and microscale thermophoresis. *European Journal of Pharmaceutics and Biopharmaceutics* **158**, 11–20 (2021).
29. Sheng, Y. J., Yin, Y. W., Ma, Y. Q. & Ding, H. M. Improving the Performance of MM/PBSA in Protein-Protein Interactions via the Screening Electrostatic Energy. *J Chem Inf Model* **61**, 2454–2462 (2021).
30. Dyson, H. J. & Wright, P. E. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* **6**, 197–208 (2005).
31. Jung, L. A. Targeting MYC Function as a Strategy for Tumor Therapy. (Julius-Maximilians-Universität Würzburg, 2016).

32. Krylov, D. & Vinson, C. R. Leucine Zipper. e LS (John Wiley & Sons, Ltd., 2001).
33. Marchetti, A., Abril-Marti, M., Illi, B., Cesareni, G. & Nasi, S. Analysis of the Myc and Max Interaction Specificity with 1 Repressor-HLH Domain Fusions. *J Mol Biol* **248**, 541–550 (1995).
34. Kim, N. A. et al. Effects of pH and Buffer Concentration on the Thermal Stability of Etanercept Using DSC and DLS. *Biol Pharm Bull* **37**, 808–816 (2014).
35. Monera, O. D., Zhou, N. E., Kay, C. M. & Hodges, R. S. Comparison of antiparallel and parallel two-stranded α -helical coiled- coils. Design, synthesis, and characterization. *Journal of Biological Chemistry* **268**, 19218–19227 (1993).
36. Lavigne, P. et al. Insights into the Mechanism of Heterodimerization from the 1 H-NMR Solution Structure of the c-Myc-Max Heterodimeric Leucine Zipper. *J Mol Biol* **281**, 165–181 (1998).
37. McDuff, F. O., Naud, J. F., Montagne, M., Sauv e, S. & Lavigne, P. The Max homodimeric b-HLH-LZ significantly interferes with the specific heterodimerization between the c-Myc and Max b-HLH-LZ in absence of DNA: A quantitative analysis. *Journal of Molecular Recognition* **22**, 261–269 (2009).
38. Bye, J. W. & Curtis, R. A. Controlling Phase Separation of Lysozyme with Polyvalent Anions. *Journal of Physical Chemistry B* **123**, 593–605 (2019).
39. Roberts, D. et al. The role of electrostatics in protein-protein interactions of a monoclonal antibody. *Mol Pharm* **11**, 2475–2489 (2014).
40. Sanchez-Ruiz, J. M. Theoretical analysis of Lumry-Eyring models in differential scanning calorimetry. *Biophys J* **61**, 921–935 (1992).
41. Roberts, C. J. Therapeutic protein aggregation: Mechanisms, design, and control. *Trends Biotechnol* **32**, 372–380 (2014).
42. Lenton, S. et al. Impact of Arginine-Phosphate Interactions on the Reentrant Condensation of Disordered Proteins. *Biomacromolecules* **22**, 1532–1544 (2021).
43. Copley, R. R. & Barton, G. J. A Structural Analysis of Phosphate and Sulphate Binding Sites in Proteins. *J Mol Biol* **242**, 321–329 (1994).
44. Banerjee, A., Hu, J. & Goss, D. J. Thermodynamics of protein-protein interactions of cMyc, Max, and Mad: Effect of polyions on protein dimerization. *Biochemistry* **45**, 2333–2338 (2006).

Chapter 8 - Summary of the thesis

The primary aim of this thesis was to investigate the potential of MD simulations in developability assessment to accurately rank and select therapeutic protein candidates, and in formulation development to gain a mechanistic insight into effects of pH, ionic strength, and excipient interactions on the stability of therapeutic proteins.

In *Chapter 1*, I started the thesis with a brief overview of protein stability, degradation pathways, and biophysical characterization techniques relevant for developability assessment and formulation development. Additionally, commonly applied approaches in lead candidate selection and formulation development are shortly described and emerging computational tools and MD simulations are introduced.

In *Chapter 2*, I applied available sequence- and structure-based webservers for aggregation and stability prediction on a dataset of 68 nanobodies and compared the results to the experimentally determined stability-indicating parameters T_m and T_{agg} . The computational results showed relatively good correlation among each other but neither the sequence- nor the structure-based predictions were able to rank the nanobodies according to their experimental behavior. Since the webservers do not account for non-native protein aggregation, we performed MD simulations at elevated temperature to induce partial unfolding and monitored the aggregation propensity of these altered conformations. Again, we were not able to accurately rank the molecules according to the measured T_m and T_{agg} values.

Chapter 3 is based on a similar approach as *Chapter 2* with a dataset of 13 monoclonal antibodies but includes more orthogonal biophysical characterization techniques and long-term stability data. This study focuses on the assessment of refoldability in combination with temperature-ramped MD simulations of the variable domains to select aggregation resistant antibodies. Available webservers for aggregation prediction were not able to accurately rank the molecules according to their long-term stability. We could show that the ReFOLD assay and MSF can be used to select aggregation resistant antibodies within this dataset and could explain that the higher propensity for non-native aggregation of some of the antibodies is caused by unfolding and exposure of APRs at lower temperatures compared to the aggregation-resistant counterparts. The MD simulations

showed a better correlation with the temperature of non-reversibility and relative monomer yield after refolding than with the T_m and T_{agg} values.

In *Chapter 4*, the complex pH-dependent aggregation mechanism of G-CSF was investigated by low-resolution biophysical characterization techniques and advanced MD simulations. The differences in conformational stability between pH 4.0 and pH 7.5 could be explained with metadynamics simulations which showed that one of the most distinct conformational changes at different pH values occurs due to the loss of cation- π -interactions between Trp and neighboring His residues. The loss of this interaction leads to increased protein flexibility and thus decreased conformational stability. CG simulations revealed that the pH-dependent aggregation-prone regions of G-CSF are located at the N-terminus and in the two long loops. Electrostatic interactions were found to be a main reason for the different aggregation behavior at acidic and neutral pH. This finding was supported with the increased aggregation at higher ionic strength due to charge-shielding effects.

In *Chapter 5*, I investigated the effect of the zwitterionic buffer molecule HEPES on G-CSF at acidic and neutral conditions. I revealed a stabilizing effect of HEPES on G-CSF in nanoDSF experiments and against shaking stress at low concentrations whereas high concentrations resulted in aggregation of the protein. I concluded different mechanisms of action for HEPES at low and high concentrations and at different stress conditions. With MD simulations and BLI measurements I could confirm direct binding of HEPES to G-CSF in a concentration-dependent manner. Binding sites were found in helix E as well as aggregation hot spots at low concentrations whereas an almost equal distribution of HEPES molecules on the protein surface could be found at high concentrations.

Chapter 6 also focused on the stabilizing effect of HEPES and other Good's buffers, but here on a novel two-domain recombinant bacteriophage lysin, HY-133. Compared to other commonly used buffer systems, HEPES showed a clear stabilizing effect on the protein. While HEPES did not change the intermolecular interaction, MD simulations showed that binding of HEPES to HY-133 prevents the collapse of the linker and inter-domain interactions. Other Good's buffers showed a similar stabilizing effect experimentally but were not further assessed in MD simulations.

In *Chapter 7*, a combination of biophysical techniques, molecular dynamics (MD) simulations, and molecular mechanics/ Poisson-Boltzmann surface area (MM/PBSA) free energy calculations was

applied to elucidate the driving forces of dimerization of Omomyc, MYC, and MAX as well as the molecular mechanisms by which the pH, ionic strength, and phosphate influence the conformational and colloidal stability of Omomyc. I discovered that electrostatics are the main contributor to dimerization. Additionally, sodium chloride and phosphate cause ion-induced folding of disordered regions of Omomyc. I could furthermore show that Omomyc undergoes reentrant condensation in phosphate buffer, as intermediate phosphate concentrations result in precipitation, whereas higher concentrations lead to stabilization of the protein. MD simulations revealed binding of phosphate ions to arginine and lysine residues on the protein surface with concentration-dependent binding affinities of dihydrogen phosphate and hydrogen phosphate ions.

The findings in this thesis show that the application of computational tools has great potential in providing molecular mechanistic insights into the properties of different therapeutic proteins from monoclonal antibodies to mini proteins containing intrinsically disordered regions.

By moving expensive experimental screens *in silico*, the efficiency of developability assessment and formulation development can be improved significantly. Computational tools, particularly the sequenced-based methods, have a much greater throughput than experimental screens. However, we found that sequence-based predictions applied in this thesis could not accurately rank candidates according to their experimental performance. The structure-based approaches showed a better correlation; we could however see that the performance of these structure-based tools can greatly be impeded by the quality of the input structure. For monoclonal antibodies and nanobodies, modelling of the V_H/V_L pairing and the CDR-H3 loop are particularly challenging. Nevertheless, the recent advances in protein structure prediction by, e.g., AlphaFold and RoseTTAFold have made it possible to achieve high accuracy protein structures from amino acid sequences. Due to the tremendous progress in computer hardware and algorithms, *in silico* tools start applying a more realistic description of proteins as conformational ensembles. However, APR prediction is complex and currently available methods are still not 100% accurate. A further step of improvement of these tools would be the consideration of environmental conditions, besides the pH, in the predictions. Nevertheless, molecular properties that can be derived from the sequence and structure can be very useful to define formulation strategies in the next steps of development. In summary, our results highlight the need for more robust methods to be able to apply them as reliable screening tools during developability assessment.

MD simulations are a powerful technique to investigate the conformational behavior of proteins, which cannot be captured experimentally. The MD simulations in our work considering the developability assessment, allowed the ranking and selection of promising candidates only in the case of the monoclonal antibodies but not for the nanobodies. Here, a probable reason is that the conditions in the MD simulations were not appropriate to achieve significant unfolding of the nanobodies to accurately rank the aggregation potential of the unfolded species. However, required simulation times to reach convergence and larger conformational changes are unreasonable compared to a nanoDSF or MSF experiment, especially for larger datasets.

Next, we could show that a better understanding of a protein aggregation mechanism can be obtained through MD simulation studies, where CG simulations are particularly useful. The loss in resolution of the model is compensated by increased computational efficiency, but validation with SAXS experiments was still necessary to verify the aggregation prone regions. However, CG simulations so far do not allow the inclusion of more complex features such as temperature, buffer components, excipients, and agitating conditions. This would tremendously help to predict rates of aggregation with better accuracy.

The use of MD simulations to study protein-excipient interactions was a successful application in this thesis. The (de-)stabilizing mechanisms could be explained for the buffer HEPES for two very different proteins, and for phosphate for a highly charged mini protein. It is important to note that the results obtained from MD simulations can be used to interpret the phenomena observed in experiments, but the setup of the computational work required a lot of knowledge on the system of interest beforehand. These results show the value of MD simulations in complement to experiments to explain the underlying molecular mechanisms. A completely computational high-throughput screening method for formulation development is however not (yet) realistic due to the high number of simulations, and thus setup and computation times, that would be required. Here, machine-learning approaches are emerging and promising tools to aid the drug development process. The advantage of machine-learning models is that they can be predictive even when the underlying biophysical mechanisms are not well understood.

Despite all improvements over the past years, MD simulations have limitations in reproducing a realistic protein ensemble. One of the major limitations is the accuracy of force fields used to calculate interatomic interactions. Especially force fields for small molecules, such as excipients,

are often trained on limited data (ca. 70 molecules for GAFF2). Hence, experimental confirmation is necessary in any case. The bottom line of this thesis is that MD simulations combined with high-throughput experimental techniques represent a very promising screening solution to improve efficiency during the early stages of the development process of therapeutic proteins.

Appendix

I. List of publications associated with this thesis

Berner, C., Menzen, T., Winter, G., & Svilenov, H. L. (2021). Combining unfolding reversibility studies and molecular dynamics simulations to select aggregation-resistant antibodies. *Molecular Pharmaceutics*, 18(6), 2242-2253.

Ko, S. K., **Berner, C.**, Kulakova, A., Schneider, M., Antes, I., Winter, G., Harris, P. & Peters, G. H. (2022). Investigation of the pH-dependent aggregation mechanisms of G-CSF using low resolution protein characterization techniques and advanced molecular dynamics simulations. *Computational and Structural Biotechnology Journal*, 20, 1439-1455.

Berner, C., Menzen, T., Winter, G., Computational approaches to assess nanobody stabilities and aggregation propensities. (*Manuscript ready for submission*)

Eisele, S., **Berner, C.**, Winter, G., Stabilization of a novel recombinant bacteriophage endolysin by protein-excipient interaction with HEPES and other Good's buffers. (*Manuscript ready for submission*)

Berner, C., Castillo Cano, V., Beaulieu M.E., Winter, G., Mechanistic understanding of the stability of the anti-MYC mini-protein Omomyc in different formulation conditions. (*Manuscript ready for submission*)