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

Prediction and Characterization of Therapeutic Protein Aggregation

---

Lorenzo Gentiluomo

aus

Rom, Italien

2020

ERKLÄRUNG

Diese Dissertation wurde im Sinne von §7 der Promotionsordnung vom 18. Juni 2016 von Herrn Prof. Dr. Wolfgang Frieß betreut.

EIDESSTATTLICHE VERSICHERUNG

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, 27.01. 2020

_____

Lorenzo Gentiluomo

Datum der Einreichung: 31.01.2020

Dissertation eingereicht am:

1. Gutachter: Prof. Dr. Wolfgang Frieß
2. Gutachter: Prof. Dr. Gerhard Winter

Mündliche Prüfung am: 27.04.2020

Tutto quello che sono lo devo ai miei genitori, a mio fratello, e a mia moglie. A loro dedico questo lavoro.

Acknowledgements

Most of all, I want to express my deepest gratitude to my supervisors Prof. Dr. Wolfgang Frieß and Dr. Dierk Roessner. I highly appreciate their valuable advice, guidance, and inspiring discussions.

Thanks to Dr. Dierk Roessner for providing the greatest working environment. I am deeply grateful for numerous opportunities to express myself and develop in most interesting scientific projects and collaborations.

Thanks to Prof. Dr. Wolfgang Frieß for tutoring me throughout the entire PhD. I started this project coming from a completely different field, and I have highly appreciated all the effort he spent to make sure I had the best scientific support.

Thanks to Prof. Dr. Gerhard Winter for all the inspiring discussions. He has always reserved for me wise words and encouragement.

Thanks to all the Wyatt technology colleagues for the many contributions to this thesis and for the nice time in Dernbach. I was delighted from all the support I have received. I know I have been extremely lucky in my PhD to work without any downtime and always in the best working environment. Thanks to Dr. Roger Scherrers and his team to always provide me with all the instruments and support I needed. A special mention goes to Thomas Davis who taught me all the technicality of the Wyatt toolbox. Thanks to his Stakhanovism I always had a solution to my issues with an instrument.

Thanks to Christoph Johann for all the inspiring conversation on FFF and all the friendly interactions. I have enjoyed all the time together with his family and the time spent in his house.

Thanks to Felix Gloge for all the inspiring conversations on DLS and CG-MALS. His suggestions were always of the highest value. I have especially enjoyed all the nice friendly time we spent together.

Thanks to all the colleagues in Santa Barbara and from around the world that put their trust in me. A special thanks goes to Michelle Chen, Daniel Some and Steve Trainoff for the inspiring conversations. Least but not last, thanks to the Wyatt family, Philip, Geoffrey and Clifford, who have created and managed such an amazing company.

Thanks to all my colleagues at the Ludwig Maximilian University of Munich for the many contributions to this thesis and for all the nice time we spent together. I have always felt home in Munich thanks to them. Special thanks go to my PIPPI colleagues from Munich, Hristo Svilenov, Inas El Bialy and Andres Tosstorff, for all the scientific input and all the friendly time together; you all have been for me a source of never ending inspiration.

Thanks to all my PIPPI colleagues around Europe, Marcello Morales, Maria Laura Greco, Matja Zalar, Aisling Roche, Christin Pohl, Dillen Augustijn, Marco Polimeni, Sujata Mahapatra, Sowmya Indrakumar,

Alina Kulakova, Stefan Hansen, for the many contributions to this thesis and for all the crazy time together.

During much of the PhD I was a wandering student. At each institution and company I visited I widened the cirle of people whom I am indebted for suggestion and comments. Thus, I would like to thank all the PIPPI consortium members and companies for the great experience and the support they provided throughout the project. As too many contributions and names would need to be acknowledging, I would acknowledge instead the head of this consortium, Pernille Harris, for creating such amazing international team.

Thanks to Åsmund Rinnan and Dillen Augustijn to introduce and support me in the world of data science. Their support has been of paramount importance for the success of my work.

Thanks to Werner Streicher for the nice time together at Novozymes and for all the support and training I received for AUC experiments and data analysis.

Thanks to Vanessa Schneider for the excellent work on the RP-MALS development and for all the nice time together in Dernbach.

My greatest thanks go to Valentina, my wife. I am exceedingly grateful for your never ending encouragement, for exceptionally motivating and for supporting me. We have grown up together. We have been facing life together. Without you I would not be the man I am today.

Finally, thanks from the deepest of my heart to my mother, father and brother. No words will be ever enough to express the love we reserve for each other.

Table of contents

CHAPTER I: Introduction

1 A general overview on protein formulation development

Protein therapeutics are used against multiple severe diseases.[1,2] Their success lies in the specificity for therapeutic targets, which is rooted in the diversity and the complexity of protein structures. Promising candidates have to exhibit desirable biophysical properties that allow for sufficient stability during manufacturing, shipping, storage, handling and administration. The most successful biologics, such as monoclonal antibody (mAbs) and insulin variants, benefit from extensive literature. Whereas little is known for other protein formats. The special challenges a formulation scientist faces include physical and chemical heterogeneity and instability, conformational flexibility and the possibility of specific and non-specific self-association.[3] Phase separation, opalescence and undesirable rheological properties, can result from attractive protein self-interaction. In addition, self-association can trigger formation of protein aggregates which come over a broad size range and with at least partial conformational rearrangement. Aggregation is considered a major risk factor in terms of drug immunogenicity.[4,5] The complexity of the aggregation process on a molecular level, and the interplay between the molecular structure, association, aggregation, unfolding and formulation conditions are far from being fully understood. Furthermore, many publications investigating protein aggregation are case-studies concerning individual proteins or a group of strictly related proteins, of which the primary sequence or more advanced structural information are in most cased not disclosed. Therefore the compilation of a systematic data analysis to reach a good understanding of the molecular mechanism behind protein drug aggregation is yet lacking. As a consequence, the development of a stable liquid formulation with the typical shelf-life of two or more years is often very challenging.[9] Lyophilization or freeze drying is typically used for biopharmaceuticals to overcome instability of labile drug candidates. On the other hand, lyophilized products are not convenient to administer as a sterile, ready to use solution and have high manufactory costs.[6,7] Thus, liquid formulation for biologics are typically preferred over lyophilizes as the aqueous solutions are ready to use and do not require a rehydration step.[8]

Although there are different ways to develop a therapeutic protein, most of them share common steps[10,11], i.e. identification of a druggable target; generation of a library of proteins which could bind to that target; selection and optimization of lead candidates; formulation development; decision on one biologically active, safe and stable protein which will continue to clinical trials. In general, the failure of a drug candidate becomes increasingly expensive as the development process advances. For this reason, pharmaceutical companies aim to adopt strategies for selecting the most promising molecules at early stages.[10,12–14]

The main aim of this thesis was the development of a toolkit for the prediction of protein aggregation in early stages. Therefore, the physical stability of proteins in solution is reviewed in the following, with a focus on the factors that induce protein aggregation. In order to generate a comprehensive general dataset a systematic characterization has to be conducted by extensive screening of therapeutic proteins. Such datasets can be investigated by multiple approaches which fall in the field of data mining. Therefore, additionally a broad general overview of the most important computational methods for this purpose (e.g. machine learning, multivariate data analysis) is given. Finally, a general overview of the most widely applied experimental method, light scattering, is given.

2 Proteins´ physical stability in solution

Proteins are macromolecules and as such they present colloidal properties. Their size is similar to the range of forces that exist between them and the scale of diffusive motion of the macromolecules is similar to that at which we are aware of changes.[15] The linear combination of the dispersion force contribution to the pair potential with the electrostatic repulsion gave the first comprehensible model framework for the stability of colloidal dispersions, the DLVO theory.[16,17] Similarly, the proximity energy framework can be applied to partially understand protein colloidal stability.[18] In fact, protein colloidal stability is influenced by interactions with other molecules (proteins, excipients, preservatives, metal ions, salts, etc.) as well as other external factors such as surfaces.[19–21] Important physical properties such as solubility, viscosity, surface interaction and aggregation, are influenced by the colloidal properties of proteins solutions,[19–25] as schematically depicted in **Fig. 1**

The colloidal properties of the solution, ultimately defined by the molecular properties of the protein, characterize the protein-protein interactions and association. A series of parameters have been developed to characterize the colloidal stability of proteins such as preferential interaction parameters and excluded volume effects,[26,27] as well as the second virial coefficient.[19,28,29] Using these parameters, the physical stability of various proteins has been successfully correlated with different formulation conditions.[19,30] Differently to macromolecules and particles, from which classical colloidal theory was derived, proteins are naturally folded in a conformation of low free energy, the native state. This may correspond to the free energy global minimum or to local minima with extremely slow rate of conversion due to high energy barriers towards the absolute minimum. Further, conformation flexibility is an intrinsic protein property which is required in nature to exert functions.[32–39] This property can be rationalized into an energy landscape which is defined by both thermodynamic and kinetic properties of a protein.[40] The conformational stability of a protein can be described as the protein´s ability to remain in the monomeric native form. Thus, handling and processing of proteins could prompt the formation of conformational variants. [41,42] These variants differ in molecular properties, including their colloidal stability, and may have a higher tendency to aggregate or be more surface active. Furthermore, sparsely populated states of

partially folded protein, which have a greater tendency to self-associate, may be an ubiquitous intermediate on the road to non-native aggregates.[43] Such small soluble aggregates may be reversible or irreversible.[44]

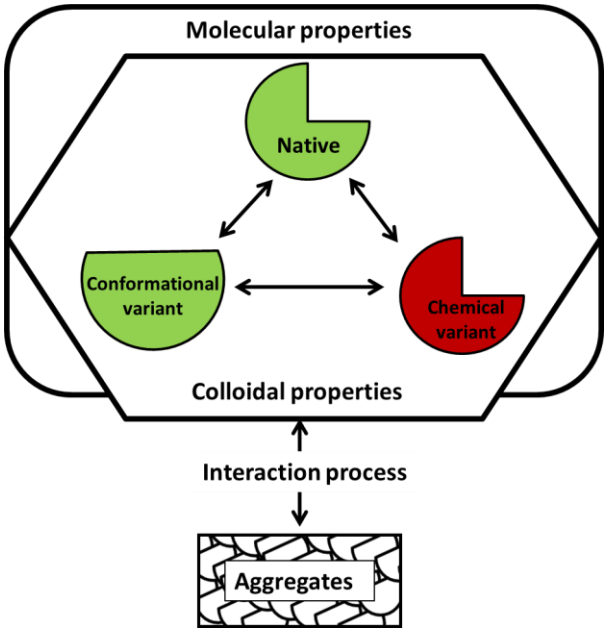


Figure 1. *Schematic illustration of the role of colloidal protein properties in aggregate formation. Re-adapted from Laue T., 2019.*[31]

2.1 Effect of chemical stability on physical stability

Therapeutic proteins contain multiple functional groups and as such they are prone to various chemical reactions.[45] An in depth understanding of these degradation pathways is necessary in order to stabilize proteins. Extensive reviews are available on the chemical instability of mAbs.[46,47] Typical chemical changes include i) hydrolytic reactions such as deamidation and proteolysis; ii) N-terminal cyclization reactions such as diketopiperazine and pyroglutamic acid formation; iii) oxidation through metal-catalyzed, photoinduced, and free-radical cascade pathways; iv) condensation reactions, particularly with sugars. Among these reactions deamidation [45,48–53] and oxidation [54–61] are the ones most often connected to physical instabilities. Deamidation not only affects charge based interaction but may also lead to alteration of the local structure of the peptide chain,[62] prompting an increase rate of aggregation and/or higher conformational instability.[63–67] Oxidation can produce a wide range of degradation products and may decrease the conformational stability.[68–70] Typically oxidation increases the propensity of proteins to aggregate.[69,71–74] Methionine oxidation has been linked to an increased rate of aggregation[75–78] but some studies have suggested an aggregation inhibition.[76,77] Currently the link between chemical degradation and aggregation is still not well understood.

2.2 Assessment of protein solution behavior in early stages

During early stage formulation development or preformulation for a protein drug candidate the solution behavior of the molecule as a function of multiple variables such as pH, ionic strength, and temperature is evaluated. This enables to identify significant modalities of instability and forms the basis for strategies to minimize such instabilities. Several biophysical assays can be used to assess protein solution behavior in early stages with limited sample consumption. For instance, it has been shown that with only 360 µg of protein it would be possible to investigate the first apparent temperature of unfolding (T_m) and the onset temperature of aggregation (T_{agg}) in 24 different fomulations.[79] Furthermore, such methods might also individuate complex aggregation behavior.[80] Nonetheless, as product development move forward the application of several methods including accelerated stability studies is necessary for a clear understanding of the candidates stability.[79] This would allow to better define the work space for the molecules in study. As usually limited amount of material is available at this stage of the development, this work space needs to be explored with robust mathematical methods, i.e. design of experiment (DoE). Experiments designed for optimization can be investigated by response surface methodology, which explores the relationship between variables. An approximated second-degree polynomial fit, which is an approximation including two-ways interaction between variables, is usually preferred as it is easy to estimate and apply.

In parallel with the understanding of the candidate's solution behavior, the development of stability-indicating analytical assays is of the highest importance. Due to the complex nature of proteins, generic approaches to early stage formulation development may be of limited value. Even mAbs with extensive structural similarity may drastically differ in their solution behavior. Thus, candidate-specific biophysical and thermodynamic analyses are necessary.

3 Protein aggregation

Proteins tend to aggregate and this is the most typical and troubling manifestation of protein instability during the development of biotherapeutics.[81] The rate and type of aggregation depends on several factors that can be broadly classified as intrinsic (primary, secondary, tertiary or quaternary structure) or extrinsic (formulation, processing conditions, etc.).[82]

As the understanding of aggregation mechanisms and kinetics are of the high importance, it has been studied for many decades.[83] Currently, it is accepted that proteins aggregate through multiple pathways but still many aspect are not fully understood. This is partially because in the typical experimental assays key intermediates of aggregation cannot be tracked,[84] which is the key for the development of general models and for control of aggregation.[85,86]

3.1 Protein aggregation pathways

The main aggregation pathways are not mutually exclusive and can be broadly divided into native and non-native processes[43,87] (**Fig. 2**): i) non-native aggregation through formation of unfolding intermediates (pathway NI), ii) native aggregation by protein self-association or chemical bonding (pathway NA), and iii) aggregation through chemical degradations (pathway ND or UD). Native proteins in solution are in an equilibrium with their unfolding intermediates (pathways NI), which are themselves in equilibrium with the completely unfolded states (pathways IU). Significant evidence suggests that under real-time storage of therapeutic proteins the sparsely populated unfolded intermediates (state "I" in **Fig. 2**) prompt the aggregation process, where the unfolding is the rate-limiting step.[88] The higher aggregation tendency of partially unfolded proteins is due to the exposure of hydrophobic patches and the higher flexibility.[89] Proteins in their native state have typically a lower rate of aggregation as the hydrophobic patches are either buried or randomly scattered respectively.[81,90] Further, partially unfolded species may be key intermediates even if the observed kinetic constant is not equal to the kinetic of unfolding.[81] The soluble aggregates (state "A" in **Fig. 2**) can gradually turn into insoluble aggregates as they exceed certain size and solubility limits (state "P" in **Fig. 2**). Depending on the protein and its environment the precipitates can be amorphous or ordered structures (e.g. fibrils).[91-93]

Self-association can be driven by the native state (pathway NA) through electrostatic and hydrophobic and van der Waals forces.[94,95] Association limited aggregation often drives the formation of reversible oligomers and irreversible aggregates and/or precipitates.[88,96,97] An important biophysical parameter to characterize the native self-association tendency is the second virial coefficient, B_{22}, which indicates whether protein-protein interactions are favored over protein-solvent interactions.[98] Protein self-association is mainly related to colloidal stability, while formation of partially unfolded intermediates is mainly related to conformational stability. Notably, either conformational or colloidal stability could be potentially rate limiting, depending on the solution conditions[95] and it is often not possible to differentiate between the two pathways (i.e. NA vs NI). A general rule of thumb is that the higher the surface charge of a protein the lower the tendency to aggregate (to the limit of extreme chemical activity which will drive phase separation), regardless of the specific amino acid sequence.[99] In certain cases it is possible to observe direct protein cross-linking, leading to aggregation (pathway NA). Intermolecular disulfide bond formation/exchange is the most common process in this regard.[100–103]

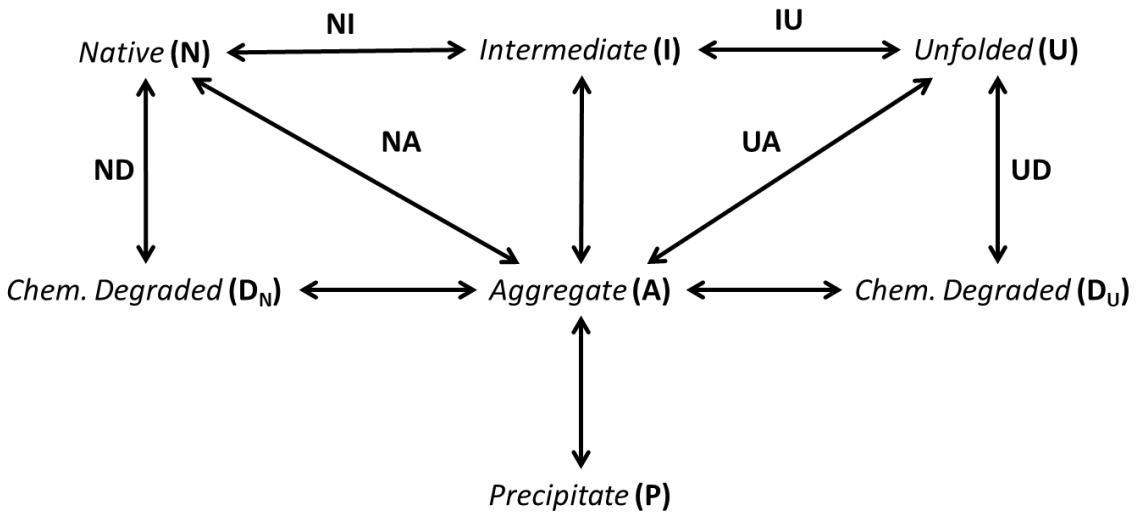


Figure 2. *Schematic illustration of the major protein aggregation pathways. Re-adapted from Wang 2010.*[82]

Not surprisingly, surface-exposed cysteines are more prone to form intermolecular disulfide bond formation than buried cysteins.[104] Due to the disulfide exchanges via β - elimination aggregation through this process is possible even for proteins without free cysteine on the surface.[105] Other cross-linking pathways leading to protein aggregation include formaldehyde-mediated cross-linking,[106,107] dityrosine formation,[106] oxidation,[71] and Maillard reaction.[108,109] Finally, chemical degradation (pathway ND) is the last major aggregation mechanism. Chemical degradations often change the physical properties of a protein as described in section 2.1. The processes NI or NA are considered as nucleation steps and will be rate limiting in nucleation dependent mechanisms.[88,110-116] Multiple aggregation pathways can occur for a single protein. For example, ovalbumin aggregates were found to be both disulfide bonded (pathway ND) and physically linked, i.e. pathway NI and/or NA.[117] Insulin can form soluble hexamers or insoluble fibrils via pathway NA and NI, soluble dimers via cyclic anhydride intermediate or insoluble disulfide-bonded aggregates via pathway ND.[105,118–121] IgG2 molecules can form dimers and high-molecular-weight (HMW) aggregates with altered secondary and tertiary structures (pathway NI), or form dimers through both disulfide and nondisulfide linkage without structural changes (pathway ND).[122] Furthermore, different forms of physical aggregation are possible for a single protein as a function of the environment. For example, amyloid fibrils or particulates can be formed respectively under high and low net charge environmental conditions.[99] Therefore, to prevent protein aggregation it is necessary to consider all the relevant factors of this process.[123]

3.2 External factors affecting protein aggregation

As mentioned above, inhibition of protein aggregation requires the understanding of many different aspects. In the next section the effect of the main solution factors are reviewed, namely temperature, pH, ionic strength, excipients (e.g. small molecules, polymers, surfactants, preservatives, and antioxidants), protein concentration. Further, it is important to highlight that the interaction between these factors may also play a critical role, which prompt the use of multivariate data analysis to properly characterize protein aggregation. Finally, it is worth noting that other factors may result critical in the development of commercial products (e.g. metal ions, reducing agents, impurities, organic solvents, container and light).

3.2.1 The effect of the temperature

Brownian diffusion is the physical principle causing two protein molecules to approach each other, which is a necessary step for aggregation. Diffusion is directly proportional to the temperature and higher temperature increases the frequency of collision promoting protein aggregation.[102,127–140] Furthermore, the chemical reactivity of the molecules is also a function of the temperature, e.g. for insulin.[121] The conformational stability of a protein can be measured thermodynamically by its unfolding free energy change (ΔG_{unf}), which is usually in the range of only 5 to 20 kcal/mol.[128–132] ΔG_{unf} is a function of temperature and can be derived at constant pressure, yielding a negative parabolic curve.[133,134] The maximum of ΔG_{unf} resides in a narrow temperature range and proteins are usually relatively stable in this range. If the temperature is outside this range the conformation stability is low, which has the potential to promote aggregation. Certain proteins can aggregate easily at slightly elevated temperature due to their narrow ΔG_{unf} maximum at room temperature such as insulin[135] or a therapeutic immunoglobulin.[136] Not surprisingly, increased temperature has a marked effect on thermodynamic stability, hydrophobic interaction, protein diffusion and chemical reactivity, leading to aggregation on much shorter timescales. Thus, high temperature is a common parameter to be selected for accelerated stability studies, although Arrhenius behavior needs to be assumed, which can lead to wrong predictions.[88] Typically, thermally induced aggregation is irreversible,[137–139] depending on the stage of the process.[44] At temperatures equal to the first unfolding temperature (T_m), ΔG_{unf} is equal to 0, and therefore proteins rapidly aggregate due to the exposure of hydrophobic patches.[140] In fact, thermally induced unfolding is typically concomitant with aggregation,[141–146] and further increase of temperature above T_m can lead to higher aggregation rates.[147–153] Temperature may affect the multiple steps of the aggregation (e.g. nucleation and growth) to a different degree[154,155] and change the pathways of aggregation depending on the solution conditions.[156,157] As for high temperature, low temperature is expected to yield thermodynamic protein instability due to the negative parabolic shape of $\Delta G_{unf}=f(T)$.[158–163] However, experimental evidence of protein denaturation is more challenging to achieve as it is usually observed below 0°C. Different approaches have been applied such as high presussures,[164] cryo-solvents, denaturants, emulsions and

super-cooled aqueous solution.[160] Nevertheless, aggregation rates at low temperature may be reduced due to weakened hydrophobic interactions, which may also lead to reversibility of the low-temperature-induced aggregation.[165–167] Aggregation at low temperature may be promoted by changes of the physical properties of the solution, e.g. due to phase separation or crystallization of excipients or buffers.[168–171] Overall, temperature is arguably the most critical factor during the entire development and commercialization processes of therapeutic protein.

3.2.2 The effect of pH

Proteins are amphoteric and the pH exerts a dominating influence on the surface charges distribution of proteins.[172] These charges affect both intramolecular folding interactions and intermolecular protein-protein interactions, and consequentially aggregation.[173,174] Thus, aggregation can be dictated by a pH effect via altered charge-charge interactions and/or pH-induced protein partial unfolding.[95,175] For example, at extreme pHs the dense charges on the protein surface significantly increase repulsive intra/inter-molecular interactions, which leads to partial protein unfolding and potentially increasing aggregation rates thanks to the increased hydrophobicity.[176] The rate of aggregation is often maximal at the proteins isoelectric point (pI) due to the low charge-charge repulsion.[173,174] Nevertheless, for some proteins the aggregation rate is lower close to their pI value.[177] Finally, protein's chemical stability is also affected by the pH, which in turn can lead to protein aggregation. As an instance, asparagine deamidation presents complex pH dependent mechanisms.[178]

Thus, proteins usually show a narrow pH stability optimum, as the solution pH conformational, colloidal and chemical stability.[179-181] In order to properly control protein stability buffering agents are necessary. However, protein stability varies also with the buffer system[88,100,139,182] and its concentration.[100,182–186] Interestingly, one of the most commonly used buffer, phosphate buffers, showed an increased aggregation rates in several studies.[186,187] Finally, it is necessary to consider that protein interaction with excipients, which affect protein aggregation, may additionally be pH dependent.[188,189]

3.2.3 The effect of ionic strength

Both positively and negatively charged species can electrostatically interact with proteins, and the type of interacting ions can have significant impact on the proteins' propensity to aggregate. [190] Such interactions can even dictate a change in the protein conformational state [191] or increase the rate of chemically formed aggregates.[104] Increasing the ionic strength reduce intra- and intermolecular charge-charge interactions, by electrostatic screening.[18] This in turn typically leads to increased protein aggregation since mostly

repulsive charge-charge interactions counteract with attractive forces.[192–194] The formed aggregates can be reversible or irreversible, native or nonnative. However, if the aggregation is led by short range interaction, e.g. dipole-dipole attraction,[18] the ionic strength screening effect may inhibit the aggregation.[141,177,191] Therefore, a complex relationship between ionic strength and aggregation has to be expected for some proteins.[97,135] Further, different inorganic salts can exert a different effect on aggregation, possibly by binding to the protein. In fact, in concentrated salt solution protein solubility depends on the cation´s or anion´s position in the Hofmeister series.[195] This cannot be explained by considering salt ions as charged hard spheres and it has been shown that protein-protein interactions in solutions are convincingly explained in terms of protein-salt interaction.[196]

3.2.4 The effect of protein concentration

Protein concentration can mainly affect the aggregation in three ways i) a higher aggregation rate due to an increased probability of association,[94,96,139] ii) a decreased aggregation rate due to mobility reduced by crowding,[197] and iii) precipitation when exceeding reaching the solubility limit. The crowding theory suggests that high concentrations should prompt the formation of compact protein structures, while favoring both specific associations and nonspecific association.[24,198,199] The formation of compact structures may increase protein stability, whereas an increased association tendency could yield higher aggregation rates.[200,201] More often the balance is in favor of increased aggregation with higher protein concentration,[121,150,167,174,180,183,202-213] as predicted by the mean field lattice model.[214] Furthermore, nucleation could be also prompted when exceeding a certain critical protein concentration.[114] It is also worth notice that if protein oligomers, such as dimers, are less competent in aggregation than the monomer, a lower rate of aggregation is to be expected.[215]

3.2.5 The effect of excipients

Various excipients to control protein aggregation have been investigated. These additives include small neutral and charged small molecules, surfactants, preservatives, polymers, and antioxidants.

A large amount of empirical findings on the effect of weakly interacting molecules will be discussed in the following. As pointed out by Timasheff, it is possible to group all this apparently unrelated observations.[27,216] Briefly, cosolvents can be classified as preferentially excluded or preferentially bound. This is determined by exchange constant of water and cosolvent molecules to the protein. The difference in values between the interactions with two generic state of a protein molecule in equilibrium between, e.g. native and unfolded, state are the one determining the direction in which the cosolvent drives the

equilibrium. For example, if a cosolvent preferentially binds less to the unfolded state to a smaller degree than in to the folded state, stabilization is to be expected. On the other hand, if a cosolvent is less excluded to the protein in the unfolded state, it will act as a denaturant.

3.2.5.1 The effect of small neutral molecules

Protein stability or aggregation can be profoundly affected by numerous small non-charged molecules,[217] major categories of which are sugars and polyols. These normally exert a stabilization effect and inhibit protein aggregation by forcing the native state conformation of proteins.[217,218] For this reason those molecules are typically called chemical chaperones and are recognized as osmolytes. The effects have been observed even under processing and culturing condition,[217,219] and it has been shown that sugars may suppress aggregation in different phases of the aggregation process.[220] Their effect on protein aggregation usually positively correlates with their concentration.[187,218,221,222] However, exceptions have been observed where sugars/polyols destabilize proteins and promote protein aggregation.[223,224] Interestingly, in certain cases excluded solutes (e.g. sugars), which tend to force the native state ensemble to its most compact form, protect exposed residues from oxidation by reducing the solvent exposed area.[225–227]

3.2.5.2 The effect of small charged molecules

Small charged molecules may affect protein aggregation,[228] either via an increase of ionic strength or by specific association. Their net effect on protein aggregation depends on their concentration, other environmental conditions like pH value and on the protein. A major category includes natural amino acid and their derivatives, which may have multiple effects.[229] Positively charged amino acid (such as histidine, lysine and arginine) can inhibit aggregation.[91,187,230–234] The mechanism of action of arginine has received special attention. It has been suggested that arginine i) increases the solubility of protein by weakening protein-protein interactions,[231,233] ii) slows down aggregation through preferential exclusion,[110] iii) reduces aggregation propensity by masking protein hydrophobic surfaces through arginine clusters possessing aligned methylene groups.[234] Similarly, proline has been reported to stabilize proteins.[217,229,235] However, the effect of proline is presumed to be nonspecific involving a combination of unfavorable and favorable side-chain interactions.[229,235,236] Nevertheless, the negative effect of the counter anions such as Cl^- can overtake the positive effect of amino acids.[237] Similarly to amino acids, several amine compounds e.g. spermidine and imidazole have been shown to inhibit protein aggregation.[238,239] But both amino acids and amines can also trigger aggregation.[232,240,241] Few organic acids e.g. phytic acid, myristic acid or linoleic acid have also been shown to reduce protein aggregation, probably due to a weak anion binding to the

protein,[242–244] while other polyanions e.g. inositol hexaphosphate prompted an increased rate of aggregation.[148]

3.2.5.3 The effect of surfactants

Ionic and nonionic surfactants can protect proteins against aggregation induced by different stresses. Surfactants stabilize proteins almost exclusively by preferentially locating at an interface (i.e. precluding protein adsorption).[245] For such reasons nonionic surfactants are very effective to prevent aggregation induced by shaking and shearing. Nonionic surfactants, such as polysorbates 20 and 80,[246] have hydrophobic tails, which may cover hydrophobic patches on the protein.[247–249] This interaction would inhibit protein aggregation by preventing hydrophobic protein-protein interaction.[249,250] For example, polysorbate 20 blocks the aggregation of rhGh induced by shaking at a molar ration of 4:1, which is close to the binding stoichiometry.[250] Further, such interaction could increase ΔG_{unf} and therefore increase the protein conformational stability.[249] Nonionic surfactants are known to bind rather weakly to proteins and may stabilize proteins by covering the surfaces and effectively hindering the adsorption of proteins.

The effect on thermally induced aggregation is inconsistent.[251–253] Various stability studies have demonstrated a negative effect of nonionic surfactants on protein stability potentially due to protein binding and partial denaturation.[183,194,201,254]

Ionic surfactants (e.g. sodium dodecyl sulfate) and octanoic acid have a more pronounced effect on aggregation than the nonionic type as they can bind rather strongly to both polar and nonpolar groups. Similarly to nonionic surfactants the binding can screen aggregation hotspot preventing non-specific protein-protein interaction. In addition, the higher charge density arising from bounding surfactant molecules could increase the overall protein colloidal stability.[255] To summarize, ionic surfactants reduce protein aggregation[239,253,255–257] but their strong binding may also induce denaturation and aggregation.[258]

3.2.5.4 The effect of preservatives

In case of multi-dose systems, preservatives, e.g. benzyl alcohol or phenol, need to be included in the formulation. These often contain hydrophobic moieties e.g. aromatic rings and weakly bind to proteins.[126,259] Consequently preservatives may accelerate aggregation[260–263] and can lead to a perturbation of the tertiary structure.[126,261]

3.2.5.5 The effect of polymers

Polymers may reduce protein aggregation propensity due to surface coverage, preferential exclusion, weak binding, steric hindrance of protein–protein interactions, and increased viscosity, limiting protein structural movement. Large solution additives, or neutral crowders, can increase the free energy barrier for association or disassociation.[264] Simulation studies showed polymers with adequate hydrophobicity may wrap around proteins with surface exposed hydrophobic patches and can thus prevent protein aggregation.[90] This is the case for polyethylene glycols (PEGs).[180,265,266] On the other hand, as previously mentioned, strong hydrophobic interaction may cause denaturation.[90] Therefore PEGs can positively, negatively or not at all affect protein aggregation depending on the experimental condition.[267] Polyionic polymers, e.g. heparin and dextran sulfate, can bind to proteins, positively influencing the aggregation propensity.[135,242,265,268,269] The effect of polyions can strongly depend on pH, which defines the affinities and binding mechanism between the two types of macromolecules.[270] As in the case of neutral polymers, if binding is strong, polyions can increase the rate of aggregation.[148,155,270] Similarly to neutral and polyionic polymers, amphoteric polymers like proteins can decrease[180,271–273] or increase[267,274] aggregation of the protein of interest. Some protein can decrease the aggregation propensity of other proteins, possibly working as chaperones e.g. heat shock proteins.[275–279] The mechanism of inhibition seems to involve the neutralizing binding of the denatured monomers[275] or coverage of hydrophobic sites.[279] Protein aggregation can also be effectively inhibited by rationally designed peptides.[280,281]

4 Brief overview on data mining, multivariate data analysis and machine learning

Systematic analysis of large data sets to reach an understanding of the molecular mechanism behind protein drug instability has become attractive with the advent of high-throughput techniques and the big data era.[79,282] In the following a brief and broad overview of data mining, multivariate data analysis and machine learning is provided.

Data mining[283] involves methods at the intersection of machine learning, statistics, and database systems. In general terms, data mining is the process of discovering valuable information in large volume of data. However, there is an omnipresent gap between large dataset and our understanding of it since we are not well-adapted to think in more than a few dimensions at once. Therefore, the development of approaches is essential to facilitate integration of multiple variables in order to prompt efficient use of experimental data. Any technique that simultaneously analyzes more than two variables may be loosely considered multivariate data analysis.[284] Some of these approaches have been developed only recently as they need the computational capacity of modern computers. This encompasses a wide range of

techniques. These approaches provide an empirical method for information extraction, regression, or classification.

Machine learning[285] is a field which makes use of a group of multivariate methods. Machine learning can be broadly defined as computational methods using past information available to the learner, i.e. the collected data, to improve its performance or to make predictions. In general terms, the most common machine leaning algorithms are supervised, unsupervised or semi-supervised, where the data points for learning are labeled or unlabeled to make prediction on unseen points. Supervised learning is the most common scenario associated with classification, regression, and ranking problems, while clustering and dimensionality reduction are examples of unsupervised learning approaches. In supervised learning the data is partitioned into training sample, validation and/or test sample. The size of these samples depends on the problem at hand. For example, if there are relatively few labeled examples, the size of the training set is usually larger than the size of the test data as the learning performance depends on the training sample. Then, relevant features (i.e. factors) are associated with the examples. This is critical as useful features can guide learning, while poor feature can mislead learning. The feature choice reflects the prior knowledge of the algorithms developer about the learning task, which in practice can have a dramatic effect on the performance results (critical factors related to protein aggregation were reviewed in the previous sections). The selected features are used to train the learning algorithm by tuning its hyperparameters i.e. free parameters. Different hypothesis out of the hypothesis set are selected for each hyperparameter. The ones leading to the best performance on the validation and/or test set are then chosen and used to predict new examples. As machine learning is all about generalization it is crucial to evaluate the performance on the validation and/or test set. The trade-off between sample size and model complexity plays also a critical role in the generalization power of the developed model. As a rule of thumb, the Occam´s razor rule can be applied to select the right model complexity, which states that when a series of hypotheses achieve the same prediction, the one with the fewest assumptions should be selected. In fact, complex model applied to sample of small size may lead to poor generalization, i.e. over-fitting. On the other hand, if the algorithm is too simple its accuracy may be poor, i.e. under-fitting. Therefore, the aim is to develop the simplest algorithm possible to achieve an accurate prediction.

As the quality of a learning algorithm depends on the data used, machine learning is inherently related to data analysis and statistics. It is not easy to draw a strict dividing line between statistic and machine learning as some analysis techniques derive from statistic and others are more closely associated to machine learning. A gross oversimplification would be that statistic has a major emphasis with testing hypothesis while machine learning is focused on formulating generalization processes as a search through possible hypothesis.

4.1 The problem of inferring proteins behavior in solution

Most multivariate techniques are based on the statistical inference of a population's values or relationships among variables from a randomly drawn sample of that population. Naturally census of the entire population makes statistical inference unnecessary. However, it is often impossible to use a census, as in the case of therapeutic drug development. Furthermore, often a protein dataset cannot represent the whole population as the protein structures included are limited. The central limit theorem,[286] CLT, is one of the most used mathematical results in science which relies on large samples, and usually we refer at its results as asymptotic. The CLT indicates that if the sample size is sufficiently large, the means of samples obtained using a random sampling with replacement are normally distributed with the mean and the variance regardless of the population distribution. In reality, however, the lack of a known population variance prevents a determination of the probability density distribution. Notably, the Student's t-distribution was developed to use a sample variance instead of a population variance.[287] The assumption for the Student's t-test is that samples should be obtained from a normally distributed population. The underlying problem of protein samples to infer protein behavior in solution relies on the population definition and its normality. We usually assume that a class of proteins, such as mAbs, or subclasses such as IgG1s, is the population we are trying to predict from our sample.

As an example, we can imagine the "protein" population as the world human population, the subset "mAb" as the female world population, the subset "IgG1" as the female world population between 18 and 25 years old. If we try to infer the voting behavior from a random sample of 1000 girls from 18 to 25 year old from all over the world we immediately can see the bias: they come from very different social and economic backgrounds. Therefore we need to restrict our target population to one "social environment", e.g. Italy. Still prediction will be poor because the social and economic condition of the girls is various. Therefore we further need to restrict our target population, e.g. young girls from Southern Italy coming from families with medium income. Once we gather a sample representative of the sub-population we could predict the voting behavior of such population. In order to understand the Italian complex voting behavior, we should collect and infer data for each relevant sub-population. Similarly, in the case of protein behavior in solution we may have to restrict the population we are aiming at. This means that some of the correlations found in literature will often have a weak statistical meaning as they refer to very small subsets of not well determined populations. As an instance, it has been shown that linear correlations inferred in literature do not hold when multiple protein and solution conditions are taken into consideration.[79]

4.2 Artificial neural networks

Artificial neural networks[288] (ANNs) have been applied in pharmaceutical research for many different purposes.[289-298] ANNs are algorithms which grossly attempt to simulate the decision process in the networks of neurons of the biological central nervous system. The main advantage of ANNs is that they allow solving a wide range of complex problems in a self-organizing manner with a relatively low level of programming. Therefore, ANNs are highly suitable to solve problems that are complex, ill-defined, highly nonlinear, of many and different variables, and/or stochastic, such as protein aggregation.

The fundamental unit of a neural network is the neuron, which is its basic information-processing unit. A neuron is composed by four elements: the synapses, a summing junction (i.e. linear combiner), an activation function and bias (**Fig. 3**).

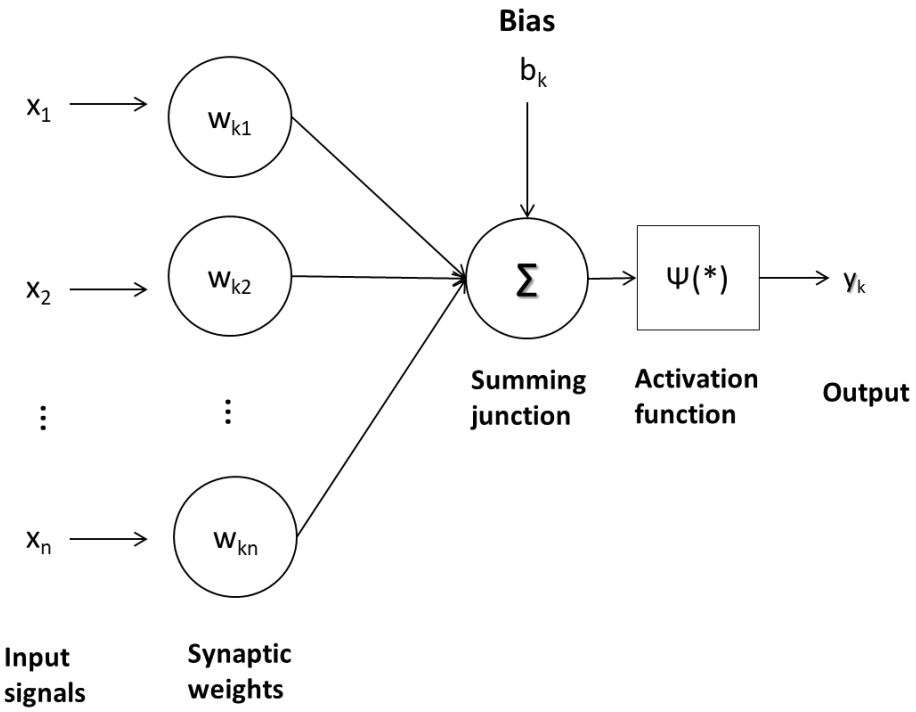


Figure 3. Model of a neuron. *x_n represent the inputs connected to the neuron, k, by the weights, w_{kn}, which multiply the corresponding input signal. All the weighted signals are summed by a summing junction Σ. An external bias b_k can be applied to Σ, to increase or lower the output signal. Finally, Σ is connected to an activation function, ψ(*), which limits the amplitude of a signal to the output, y_k. Picture reproduced from:* Gentiluomo, L., et al (2019) – **Chapter III**.[289]

The synapse is a set of connecting links each characterized by a weight. The neuron processes an arriving signal by multiplying it by the synaptic weight, which, differently from the brain, may have a negative or positive value. The linear combiner sums up the input signals, weighted by the respective synapse of the neuron. The added signal is limited by an activation function. This limits the permissible amplitude range of the output signal to some finite value. An external bias is applied to increase or decrease the net input to the activation function. The latter may assume different forms, most commonly a sigmoidal one. The neurons are then structured accordingly to the learning rules used to design the selected type of network. Three different classes of architectures can be differentiated, single-layer feedforward networks, multilayer feedforward networks, and recurrent networks. As for other machine learning algorithms the ability of ANNs to learn from their environment and to improve their performance is of primary importance. The interactive process of learning from the environment prompts adjustments of synaptic weights and bias level. A set of well-defined rules for the solution of a learning problem is called learning algorithm.

An important class of ANNs is the multilayer feedforward networks (MFNs). These kinds of networks consist of an input layer, one or more hidden layers of computational nodes and an output layer. The input signal propagates “forward” from the input to the hidden layers and finally to the out layer. MFNs are used to solve complex problems by training them in a supervised manner. A highly popular algorithm used for training MFNs known as the “error-back propagation” is based on the error-correction learning rule. This learning algorithm consists of a forward and a backward pass through the different layers of the network. In the forward pass, the signal is propagated through the network. Subsequently, a set of output is produced as the actual response of the network. During the forward pass the synaptic weights of the networks are all fixed. Then, in the backward pass, the synaptic weights are all adjusted by subtracting the actual response of the network from the desired response to produce an error signal. This error signal is then propagated back through the network and the synaptic weights are adjusted to reduce the error in a statistical sense. Overall MFNs present three characteristics: i) the neurons have a non-linear smooth activation function (e.g. sigmoidal) ii) the network contains one or more layers of hidden neurons that are not part of the input or output network iii) the network exhibits a high degree of connectivity. From these three characteristics and its ability to learn from experience through training MFNs derive their computing powers. A major drawback of MFNs is the presence of nonlinearity, the use of hidden neurons and the high connectivity which yield a model with hard theoretical interpretability, often referred to as a “black box”. To compensate such drawback, in chapters III and V we applied methods to generated surrogated models or “white boxes”.

5 Light scattering techniques and their application to protein characterization

Various biophysical techniques are employed in the characterization of protein behavior in solution in order to guide formulation development. The most common methods include spectroscopic methods e.g. second-derivative ultraviolet spectroscopy, circular dichroism, Fourier transform infrared spectroscopy, and fluorescence spectroscopy, thermal analysis e.g. differential scanning calorimetry, nano differential scanning fluorimetry and size-based analysis e.g. analytical ultracentrifugation and light scattering. Less frequently more sophisticated methods include hydrogen–deuterium exchange, mass spectrometry, nuclear magnetic resonance spectroscopy, X-ray crystallography, and electron paramagnetic spin resonance spectroscopy. As light scattering is a key technology, intensively applied in this thesis, a brief overview of this technique is given in the following.

Light scattering[300] is a natural phenomenon resulting from the interaction of light with matter related to the heterogeneity of the system. If an obstacle, which could be a single electron, an atom, a molecule or a solid or liquid particle, encounters an electromagnetic wave of the right wavelength it will generate a secondary radiation generated by the oscillatory motion of the incident wave. For example, a neutral molecule interacting with the oscillating electric field of light will form a dipole, which, due to the oscillation of the incident radiation, oscillates as well. The oscillating dipole becomes a source of new radiation, the scattered light. The tendency of the electron cloud of a molecule to be displaced by an external field is directly correlated to the ability of such molecule to scatter light. This tendency is called polarizability and it is directly proportional to the specific refractive index increment (dn/dc). In addition to the scattering phenomenon the excited elementary charges may transform part of the incident light in other forms, like thermal energy (i.e. adsorption). As everything is heterogeneous (even pure gases) all media scatter light. The theory of light scattering from macromolecular solution developed by Einstein,[301] Raman,[302] Debye,[303] Zimm[304,305] and others, represents one of the major successes of chemical physics.

The phenomenon of light scattering can be classified depending on how we collect and process the data. Multiangle static light scattering (MALS), or more generally static light scattering (SLS), refers to experiments in which the scattered light intensity is determined at given scattering angles by averaging the fluctuating intensity at the same wavelength as the incident light over a long time scale compared with the time scale of the intensity fluctuation. SLS yields the apparent weight average molecular mass (M_w), the apparent root mean square radius (R_{rms}), and the apparent second virial coefficient (A_2 or B_{22}). Dynamic light scattering (DLS) also occurs at the same wavelength as that of the incident light, but the fluctuations of the scattered light intensity over extremely short intervals are collected and processed. DLS yields the apparent diffusion coefficient (D) and by the Einstein-Stroke relation the apparent hydrodynamic radius (R_h), the solution polydispersity, and interaction diffusion parameters (k_D). Modern instruments allow performing DLS and SLS simultaneously. Many detailed reviews on MALS[300,304,306,307] and DLS[308–310] can be found in literature. Other types of light scattering include Raman scattering, which

occurs at a wavelength different from that of the incident light and can provide structural information, and phase analysis light scattering (PALS), which allows the determination of the electrophoretic mobility of the scattering particles.

5.1 Recent applications of light scattering for protein characterization

The mass determination of biomolecules and their aggregates in peaks eluting from size exclusion chromatography (SEC-MALS)[306] or field flow fractions (FFF-MALS)[311] is a common application of light scattering. SEC relies on column calibration based on size exclusions standards, which yields an empiric correlation between elution time and M_w. M_w calculated by SEC-MALS in turn provides values which are free from errors arising from non-sphericity of a protein or protein interaction with the column material. SEC-MALS was employed extensively to characterize native and non-native oligomers in several landmark studies.[7,312–316] SEC-MALS is also a primary tool in the investigation of protein aggregation kinetics.[87,317–319] Addition of multiple concentration detectors, e.g. UV and RI, allows the analysis of conjugated substances that contain a significant mass fraction of UV and non-UV absorbing material, e.g. glycosylated proteins.[320] Recently the advantages of coupling MALS with different kind of fractionation techniques, namely IEX and RP, has been demonstrated.[321,322]

Batch measurements with both SLS and DLS are used extensively for analysis of proteins in solution.[323-348] High-throughput DLS instruments have been developed to facilitate colloidal stability screening with low material consumption.[349–351] Due to the simplicity of use, DLS pitfalls are often underestimated. Care in the interpretation of DLS data needs to be taken, especially when the aim is to address protein-protein interaction by k_D.[352] Furthermore, the underlying physics of light scattering and the strict mathematical assumptions for the quantitative interpretation of multimodal polydisperse samples impose certain limitations.[353] On the other hand, DLS is very sensitive to the presence of aggregates and can be used to estimate the aggregate weight fraction. It is also an excellent tool to assess whether a sample is monodisperse or has a significant degree of polydispersity. The presence of large aggregates (e.g. 1 µm) can result in significant number fluctuation.[354] Also multiple scattering has to be considered for highly concentrated or turbid solutions.[355] This yields an apparent decrease in apparent size measured by DLS, which should not be misinterpreted as protein-protein interactions. Finally, low throughput light scattering techniques, such as composition gradient MALS (CG-MALS) allow to quantitatively characterize binding affinity and stoichiometry of homo- and hetero- bimolecular protein interactions as a function of solution conditions such as pH, ionic strength, and buffer salts.[332,256-376]

6 References

1. Dimitrov, D. S. (2012). Therapeutic proteins. In *Therapeutic Proteins* (pp. 1-26). Humana Press, Totowa, NJ.

2. Elvin, J. G., Couston, R. G., & van der Walle, C. F. (2013). Therapeutic antibodies: market considerations, disease targets and bioprocessing. *International journal of pharmaceutics*, *440*(1), 83-98.

3. Otvos Jr, Laszlo, and John D. Wade. "Current challenges in peptide-based drug discovery." *Frontiers in chemistry* 2 (2014): 62.

4. Hermeling, S., Crommelin, D. J., Schellekens, H., & Jiskoot, W. (2004). Structure-immunogenicity relationships of therapeutic proteins. *Pharmaceutical research*, *21*(6), 897-903.

5. Braun, A., Kwee, L., Labow, M. A., & Alsenz, J. (1997). Protein aggregates seem to play a key role among the parameters influencing the antigenicity of interferon alpha (IFN-α) in normal and transgenic mice. *Pharmaceutical research*, *14*(10), 1472-1478.

7. Assegehegn, G., Brito-de la Fuente, E., Franco, J. M., & Gallegos, C. (2019). The importance of understanding the freezing step and its impact on freeze-drying process performance. *Journal of pharmaceutical sciences*, *108*(4), 1378-1395.

6. Jameel, F., & Hershenson, S. (Eds.). (2010). *Formulation and process development strategies for manufacturing biopharmaceuticals*. John Wiley & Sons.

8. Geiger, G., Bernhagen, J., Wagner, E., Bisswanger, H., Brunner, H., & Vitzthum, F. (2001). Standardized measurements and differential spectroscopy in microplates. *Analytical biochemistry*, *296*(1), 29-40..

9. Bye, J. W., Platts, L., & Falconer, R. J. (2014). Biopharmaceutical liquid formulation: a review of the science of protein stability and solubility in aqueous environments. *Biotechnology letters*, *36*(5), 869-875.

10. Jarasch, A., Koll, H., Regula, J. T., Bader, M., Papadimitriou, A., & Kettenberger, H. (2015). Developability assessment during the selection of novel therapeutic antibodies. *Journal of pharmaceutical sciences*, *104*(6), 1885-1898.

11. Carter, P. (2001). Improving the efficacy of antibody-based cancer therapies. *Nature Reviews Cancer*, *1*(2), 118.

12. Liu, Y., Caffry, I., Wu, J., Geng, S. B., Jain, T., Sun, T., ... & Vásquez, M. (2014, March). High-throughput screening for developability during early-stage antibody discovery using self-interaction nanoparticle spectroscopy. In *MAbs* (Vol. 6, No. 2, pp. 483-492). Taylor & Francis.

13. Zurdo, J. (2013). Developability assessment as an early de-risking tool for biopharmaceutical development. *Pharmaceutical Bioprocessing*, *1*(1), 29-50.

14. Wolf Pérez, A. M., Sormanni, P., Andersen, J. S., Sakhnini, L. I., Rodriguez-Leon, I., Bjelke, J. R., ... & Lorenzen, N. (2019, February). In vitro and in silico assessment of the developability of a designed monoclonal antibody library. In *Mabs* (Vol. 11, No. 2, pp. 388-400). Taylor & Francis.

15. Goodwin, J. (2009). *Colloids and interfaces with surfactants and polymers*. John Wiley & Sons.

16. Derjaguin, B. V., & Landau, L. (1993). Theory of the stability of strongly charged lyophobic sols and of the adhesion of strongly charged particles in solutions of electrolytes. *Progress in Surface Science*, *43*(1-4), 30-59.

17. Verwey, E. J. W. (1947). Theory of the stability of lyophobic colloids. *The Journal of Physical Chemistry*, *51*(3), 631-636.

18. Laue, T. (2012). Proximity energies: a framework for understanding concentrated solutions. *Journal of Molecular Recognition*, *25*(3), 165-173.

19. Valente, J. J., Payne, R. W., Manning, M. C., Wilson, W. W., & Henry, C. S. (2005). Colloidal behavior of proteins: effects of the second virial coefficient on solubility, crystallization and aggregation of proteins in aqueous solution. *Current pharmaceutical biotechnology*, *6*(6), 427-436.

20. Ruckenstein, E., & Shulgin, I. L. (2006). Effect of salts and organic additives on the solubility of proteins in aqueous solutions. *Advances in colloid and interface science*, *123*, 97-103.

21. Winzor, D. J., Deszczynski, M., Harding, S. E., & Wills, P. R. (2007). Nonequivalence of second virial coefficients from sedimentation equilibrium and static light scattering studies of protein solutions. *Biophysical chemistry*, *128*(1), 46-55.

22. Leandro, P., & Gomes, C. M. (2008). Protein misfolding in conformational disorders: rescue of folding defects and chemical chaperoning. *Mini reviews in medicinal chemistry*, *8*(9), 901-911.

23. Gianni, S., Ivarsson, Y., Jemth, P., Brunori, M., & Travaglini-Allocatelli, C. (2007). Identification and characterization of protein folding intermediates. *Biophysical chemistry*, *128*(2-3), 105-113.

24. Guo, J., Harn, N., Robbins, A., Dougherty, R., & Middaugh, C. R. (2006). Stability of helix-rich proteins at high concentrations. *Biochemistry*, *45*(28), 8686-8696.

25. Chi, E. Y., Kendrick, B. S., Carpenter, J. F., & Randolph, T. W. (2005). Population balance modeling of aggregation kinetics of recombinant human interleukin-1 receptor antagonist. *Journal of pharmaceutical sciences*, *94*(12), 2735-2748.

26. Timasheff, S. N. (1993). 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.

27. Timasheff, S. N. (1998). Control of protein stability and reactions by weakly interacting cosolvents: the simplicity of the complicated. *Adv. Protein Chem*, *51*(51), 355-432.

28. Neal, B. L., Asthagiri, D., Velev, O. D., Lenhoff, A. M., & Kaler, E. W. (1999). Why is the osmotic second virial coefficient related to protein crystallization?. *Journal of Crystal Growth*, *196*(2-4), 377-387.

29. Alford, J. R., Kendrick, B. S., Carpenter, J. F., & Randolph, T. W. (2008). Measurement of the second osmotic virial coefficient for protein solutions exhibiting monomer–dimer equilibrium. *Analytical biochemistry*, *377*(2), 128-133.

30. Chi, E. Y., Krishnan, S., Randolph, T. W., & Carpenter, J. F. (2003). Physical stability of proteins in aqueous solution: mechanism and driving forces in nonnative protein aggregation. *Pharmaceutical research*, *20*(9), 1325-1336.

31. Laue, T. M., & Shire, S. J. (2019). The molecular interaction process. *Journal of pharmaceutical sciences*.

32. Boehr, D. D., Nussinov, R., & Wright, P. E. (2009). The role of dynamic conformational ensembles in biomolecular recognition. *Nature chemical biology*, *5*(11), 789.

33. Csermely, P., Palotai, R., & Nussinov, R. (2010). Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends in biochemical sciences*, *35*(10), 539-546.

34. Vogt, A. D., & Di Cera, E. (2012). Conformational selection or induced fit? A critical appraisal of the kinetic mechanism. *Biochemistry*, *51*(30), 5894-5902.

35. Hirokawa, N., Noda, Y., Tanaka, Y., & Niwa, S. (2009). Kinesin superfamily motor proteins and intracellular transport. *Nature reviews Molecular cell biology*, *10*(10), 682.

36. Conde, C., & Cáceres, A. (2009). Microtubule assembly, organization and dynamics in axons and dendrites. *Nature Reviews Neuroscience*, *10*(5), 319.

37. Panne, D., Maniatis, T., & Harrison, S. C. (2007). An atomic model of the interferon-β enhanceosome. *Cell*, *129*(6), 1111-1123.

38. Morris, G. P., & Allen, P. M. (2012). How the TCR balances sensitivity and specificity for the recognition of self and pathogens. *Nature immunology*, *13*(2), 121.

39. Ribrault, C., Sekimoto, K., & Triller, A. (2011). From the stochasticity of molecular processes to the variability of synaptic transmission. *Nature Reviews Neuroscience*, *12*(7), 375.

40. Jahn, T. R., & Radford, S. E. (2005). The Yin and Yang of protein folding. *The FEBS journal*, *272*(23), 5962-5970.

41. Santucci, R., Sinibaldi, F., & Fiorucci, L. (2008). Protein folding, unfolding and misfolding: role played by intermediate States. *Mini reviews in medicinal chemistry*, *8*(1), 57-62.

42. Ferreira, S. T., De Felice, F. G., & Chapeaurouge, A. (2006). Metastable, partially folded states in the productive folding and in the misfolding and amyloid aggregation of proteins. *Cell biochemistry and biophysics*, *44*(3), 539-548.

43. Roberts, C. J. (2007). Non-native protein aggregation kinetics. *Biotechnology and bioengineering*, *98*(5), 927-938.

44. Calamai, M., Canale, C., Relini, A., Stefani, M., Chiti, F., & Dobson, C. M. (2005). Reversal of protein aggregation provides evidence for multiple aggregated states. *Journal of molecular biology*, *346*(2), 603-616.

45. Manning, M. C., Chou, D. K., Murphy, B. M., Payne, R. W., & Katayama, D. S. (2010). Stability of protein pharmaceuticals: an update. *Pharmaceutical research*, *27*(4), 544-575.

46. Wang, W., Singh, S., Zeng, D. L., King, K., & Nema, S. (2007). Antibody structure, instability, and formulation. *Journal of pharmaceutical sciences*, *96*(1), 1-26.

47. Liu, H., Gaza-Bulseco, G., Faldu, D., Chumsae, C., & Sun, J. (2008). Heterogeneity of monoclonal antibodies. *Journal of pharmaceutical sciences*, *97*(7), 2426-2447.

48. Aswad, D. W. (1995). *Deamidation and isoaspartate formation in peptides and proteins*. CRC Press.

49. Robinson, N. E., & Robinson, A. (2004). *Molecular clocks: deamidation of asparaginyl and glutaminyl residues in peptides and proteins*. Althouse press.

50. Tonie Wright, H., & Urry, D. W. (1991). Nonenzymatic deamidation of asparaginyl and glutaminyl residues in protein. *Critical Reviews in Biochemistry and Molecular Biology*, *26*(1), 1-52.

51. Lai, M. C., & Topp, E. M. (1999). Solid-state chemical stability of proteins and peptides. *Journal of pharmaceutical sciences*, *88*(5), 489-500.

52. Reissner, K. J., & Aswad, D. W. (2003). Deamidation and isoaspartate formation in proteins: unwanted alterations or surreptitious signals?. *Cellular and Molecular Life Sciences CMLS*, *60*(7), 1281-1295.

53. Wakankar, A. A., & Borchardt, R. T. (2006). Formulation considerations for proteins susceptible to asparagine deamidation and aspartate isomerization. *Journal of pharmaceutical sciences*, *95*(11), 2321-2336.

54. Schöneich, C., Zhao, F., Yang, J., & Miller, B. L. (1997). Mechanisms of methionine oxidation in peptides.

55. Hovorka, S. W., & Schöneich, C. (2001). Oxidative degradation of pharmaceuticals: theory, mechanisms and inhibition. *Journal of pharmaceutical sciences*, *90*(3), 253-269.

56. Volkin, D. B., Mach, H., & Middaugh, C. R. (1997). Degradative covalent reactions important to protein stability. *Molecular biotechnology*, *8*(2), 105-122.

57. Li, S., Schöneich, C., & Borchardt, R. T. (1995). Chemical instability of protein pharmaceuticals: mechanisms of oxidation and strategies for stabilization. *Biotechnology and Bioengineering*, *48*(5), 490-500.

58. Stadtman, E. R. (1990). Metal ion-catalyzed oxidation of proteins: biochemical mechanism and biological consequences. *Free Radical Biology and Medicine*, *9*(4), 315-325.

59. Stadtman, E. R. (1993). Oxidation of free amino acids and amino acid residues in proteins by radiolysis and by metal-catalyzed reactions. *Annual review of biochemistry*, *62*(1), 797-821.

60. Nguyen, T. H. (1994). Oxidation degradation of protein pharmaceuticals.

61. Schey, K. L., & Finley, E. L. (2000). Identification of peptide oxidation by tandem mass spectrometry. *Accounts of chemical research*, *33*(5), 299-306.

62. Wearne, S. J., & Creighton, T. E. (1989). Effect of protein conformation on rate of deamidation: ribonuclease A. *Proteins: Structure, Function, and Bioinformatics*, *5*(1), 8-12.

63. Nilsson, M. R., Driscoll, M., & Raleigh, D. P. (2002). Low levels of asparagine deamidation can have a dramatic effect on aggregation of amyloidogenic peptides: implications for the study of amyloid formation. *Protein Science*, *11*(2), 342-349.

64. Harms, M. J., Wilmarth, P. A., Kapfer, D. M., Steel, E. A., David, L. L., Bächinger, H. P., & Lampi, K. J. (2004). Laser light-scattering evidence for an altered association of βB1-crystallin deamidated in the connecting peptide. *Protein science*, *13*(3), 678-686.

65. Harn, N. R., Jeng, Y. N., Kostelc, J. G., & Middaugh, C. R. (2005). Spectroscopic analysis of highly concentrated suspensions of bovine somatotropin in sesame oil. *Journal of pharmaceutical sciences*, *94*(11), 2487-2495.

66. Flaugh, S. L., Mills, I. A., & King, J. (2006). Glutamine deamidation destabilizes human γD-crystallin and lowers the kinetic barrier to unfolding. *Journal of Biological Chemistry*, *281*(41), 30782-30793.

67. Silva, T., Kirkpatrick, A., Brodsky, B., & Ramshaw, J. A. (2005). Effect of deamidation on stability for the collagen to gelatin transition. *Journal of agricultural and food chemistry*, *53*(20), 7802-7806.

68. Fisher, M. T., & Stadtman, E. R. (1992). Oxidative modification of Escherichia coli glutamine synthetase. Decreases in the thermodynamic stability of protein structure and specific changes in the active site conformation. *Journal of Biological Chemistry*, *267*(3), 1872-1880.

69. Liu, D., Ren, D., Huang, H., Dankberg, J., Rosenfeld, R., Cocco, M. J., ... & Remmele Jr, R. L. (2008). Structure and stability changes of human IgG1 Fc as a consequence of methionine oxidation. *Biochemistry*, *47*(18), 5088-5100.

70. Gao, J., Yin, D. H., Yao, Y., Sun, H., Qin, Z., Schöneich, C., ... & Squier, T. C. (1998). Loss of conformational stability in calmodulin upon methionine oxidation. *Biophysical journal*, *74*(3), 1115-1134.

71. Khossravi, M., Shire, S. J., & Borchardt, R. T. (2000). Evidence for the involvement of histidine A (12) in the aggregation and precipitation of human relaxin induced by metal-catalyzed oxidation. *Biochemistry*, *39*(19), 5876-5885.

72. Li, S., Nguyen, T. H., Schoneich, C., & Borchardt, R. T. (1995). Aggregation and precipitation of human relaxin induced by metal-catalyzed oxidation. *Biochemistry*, *34*(17), 5762-5772.

73. Gaudiano, M. C., Colone, M., Bombelli, C., Chistolini, P., Valvo, L., & Diociaiuti, M. (2005). Early stages of salmon calcitonin aggregation: effect induced by ageing and oxidation processes in water and in the presence of model membranes. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, *1750*(2), 134-145.

74. Hu, D., Qin, Z., Xue, B., Fink, A. L., & Uversky, V. N. (2008). Effect of methionine oxidation on the structural properties, conformational stability, and aggregation of immunoglobulin light chain LEN. *Biochemistry*, *47*(33), 8665-8677.

75. Krishnan, S., Chi, E. Y., Wood, S. J., Kendrick, B. S., Li, C., Garzon-Rodriguez, W., ... & Citron, M. (2003). Oxidative dimer formation is the critical rate-limiting step for Parkinson's disease α-synuclein fibrillogenesis. *Biochemistry*, *42*(3), 829-837.

76. Palmblad, M., Westlind-Danielsson, A., & Bergquist, J. (2002). Oxidation of methionine 35 attenuates formation of amyloid β-peptide 1–40 oligomers. *Journal of Biological Chemistry*, *277*(22), 19506-19510.

77. Watson, A. A., Fairlie, D. P., & Craik, D. J. (1998). Solution Structure of Methionine-Oxidized Amyloid β-Peptide (1− 40). Does Oxidation Affect Conformational Switching?. *Biochemistry*, *37*(37), 12700-12706.

78. Steinmetz, M. O., García-Echeverría, C., & Kammerer, R. A. (2005). Design of a coiled-coil-based model peptide system to explore the fundamentals of amyloid fibril formation. *International Journal of Peptide Research and Therapeutics*, *11*(1), 43-52.

79. Gentiluomo, L., Svilenov, H. L., Augustijn, D., El Bialy, I., Greco, M. L., Kulakova, A., ... & Frieß, W. (2019). Advancing therapeutic protein discovery and development through comprehensive computational and biophysical characterization. *Molecular Pharmaceutics*.

80. Gentiluomo, L., Roessner, D., Streicher, W., Mahapatra, S., Harris, P., & Frieß, W. (2020). Characterization of native reversible self-association of a monoclonal antibody mediated by Fab-Fab interaction. *Journal of Pharmaceutical Sciences*, *109*(1), 443-451.

81. Wang, W. (2005). Protein aggregation and its inhibition in biopharmaceutics. *International journal of pharmaceutics*, *289*(1-2), 1-30.

82. Wang, W., & Roberts, C. J. (Eds.). (2010). *Aggregation of therapeutic proteins*. John Wiley & Sons.

83. Oosawa, F., Asakura, S., Hotta, K., Imai, N., & Ooi, T. (1959). G-F transformation of actin as a fibrous condensation. *Journal of Polymer Science*, *37*(132), 323-336.

84. Philo, J. S. (2006). Is any measurement method optimal for all aggregate sizes and types?. *The AAPS journal*, *8*(3), E564-E571.

85. Morris, A. M., Watzky, M. A., & Finke, R. G. (2009). Protein aggregation kinetics, mechanism, and curve-fitting: a review of the literature. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, *1794*(3), 375-397.

86. Frieden, C. (2007). Protein aggregation processes: in search of the mechanism. *Protein Science*, *16*(11), 2334-2344.

87. Li, Y., & Roberts, C. J. (2009). Lumry− Eyring nucleated-polymerization model of protein aggregation kinetics. 2. Competing growth via condensation and chain polymerization. *The Journal of Physical Chemistry B*, *113*(19), 7020-7032.

88. Weiss IV, W. F., Young, T. M., & Roberts, C. J. (2009). Principles, approaches, and challenges for predicting protein aggregation rates and shelf life. *Journal of pharmaceutical sciences*, *98*(4), 1246-1277.

89. Damodaran, S., & Song, K. B. (1988). Kinetics of adsorption of proteins at interfaces: role of protein conformation in diffusional adsorption. *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology*, *954*, 253-264.

90. Zhang, L., Lu, D., & Liu, Z. (2008). How native proteins aggregate in solution: A dynamic Monte Carlo simulation. *Biophysical chemistry*, *133*(1-3), 71-80.

91. Ecroyd, H., & Carver, J. A. (2008). The effect of small molecules in modulating the chaperone activity of αB-crystallin against ordered and disordered protein aggregation. *The FEBS journal*, *275*(5), 935-947.

92. Necula, M., Kayed, R., Milton, S., & Glabe, C. G. (2007). Small molecule inhibitors of aggregation indicate that amyloid β oligomerization and fibrillization pathways are independent and distinct. *Journal of Biological Chemistry*, *282*(14), 10311-10324.

93. Dobson, C. M. (2003). Protein folding and misfolding. *Nature*, *426*(6968), 884.

94. Liu, J., Nguyen, M. D., Andya, J. D., & Shire, S. J. (2005). Reversible self-association increases the viscosity of a concentrated monoclonal antibody in aqueous solution. *Journal of pharmaceutical sciences*, *94*(9), 1928-1940.

95. Chi, E. Y., Krishnan, S., Kendrick, B. S., Chang, B. S., Carpenter, J. F., & Randolph, T. W. (2003). Roles of conformational stability and colloidal stability in the aggregation of recombinant human granulocyte colony-stimulating factor. *Protein Science*, *12*(5), 903-913.

96. Kanai, S., Liu, J. U. N., Patapoff, T. W., & Shire, S. J. (2008). Reversible self-association of a concentrated monoclonal antibody solution mediated by Fab–Fab interaction that impacts solution viscosity. *Journal of pharmaceutical sciences*, *97*(10), 4219-4227.

97. Saluja, A., & Kalonia, D. S. (2008). Nature and consequences of protein–protein interactions in high protein concentration solutions. *International journal of pharmaceutics*, *358*(1-2), 1-15.

98. Chi, E. Y., Krishnan, S., Randolph, T. W., & Carpenter, J. F. (2003). Physical stability of proteins in aqueous solution: mechanism and driving forces in nonnative protein aggregation. *Pharmaceutical research*, *20*(9), 1325-1336.

99. Krebs, M. R., Devlin, G. L., & Donald, A. M. (2007). Protein particulates: another generic form of protein aggregation?. *Biophysical journal*, *92*(4), 1336-1342.

100. Wang, Y. J., Shahrokh, Z., Vemuri, S., Eberlein, G., Beylin, I., & Busch, M. (2002). Characterization, stability, and formulations of basic fibroblast growth factor. In *Formulation, Characterization, and Stability of Protein Drugs: Case Histories* (pp. 141-180). Springer, Boston, MA.

101. Shahrokh, Z., Eberlein, G., Buckley, D., Paranandi, M. V., Aswad, D. W., Stratton, P., ... & Wang, Y. J. (1994). Major degradation products of basic fibroblast growth factor: Detection of succinimide and iso-aspartate in place of aspartate 15. *Pharmaceutical research*, *11*(7), 936-944.

102. Yoshioka, S., Aso, Y., Izutsu, K. I., & Terao, T. (1993). Aggregates formed during storage of β-galactosidase in solution and in the freeze-dried state. *Pharmaceutical research*, *10*(5), 687-691.

103. Jordan, G. M., Yoshioka, S., & Terao, T. (1994). The aggregation of bovine serum albumin in solution and in the solid state. *Journal of pharmacy and pharmacology*, *46*(3), 182-185.

104. Livney, Y. D., Verespej, E., & Dalgleish, D. G. (2003). Steric effects governing disulfide bond interchange during thermal aggregation in solutions of β-lactoglobulin B and α-lactalbumin. *Journal of agricultural and food chemistry*, *51*(27), 8098-8106.

105. Costantino, H. R., Langer, R., & Klibanov, A. M. (1994). Moisture-induced aggregation of lyophilized insulin. *Pharmaceutical research*, *11*(1), 21-29.

106. Malencik, D. A., & Anderson, S. R. (2003). Dityrosine as a product of oxidative stress and fluorescent probe. *Amino acids*, *25*(3-4), 233-247.

107. Schwendeman, S. P., Costantino, H. R., Gupta, R. K., Siber, G. R., Klibanov, A. M., & Langer, R. (1995). Stabilization of tetanus and diphtheria toxoids against moisture-induced aggregation. *Proceedings of the National Academy of Sciences*, *92*(24), 11234-11238.

108. Liang, J. N., & Rossi, M. T. (1990). In vitro non-enzymatic glycation and formation of browning products in the bovine lens α-crystallin. *Experimental eye research*, *50*(4), 367-371.

109. Zarina, S., Zhao, H. R., & Abraham, E. (2000). Advanced glycation end products in human senile and diabetic cataractous lenses. *Molecular and cellular biochemistry*, *210*(1-2), 29-34.

110. Baynes, B. M., Wang, D. I., & Trout, B. L. (2005). Role of arginine in the stabilization of proteins against aggregation. *Biochemistry*, *44*(12), 4919-4925.

111. Gsponer, J., & Vendruscolo, M. (2006). Theoretical approaches to protein aggregation. *Protein and peptide letters*, *13*(3), 287-293.

112. Jarrett, J. T., & Lansbury Jr, P. T. (1992). Amyloid fibril formation requires a chemically discriminating nucleation event: studies of an amyloidogenic sequence from the bacterial protein OsmB. *Biochemistry*, *31*(49), 12345-12352.

113. Finke, J. M., Roy, M., Zimm, B. H., & Jennings, P. A. (2000). Aggregation events occur prior to stable intermediate formation during refolding of interleukin 1β. *Biochemistry*, *39*(3), 575-583.

114. Lomakin, A., Teplow, D. B., Kirschner, D. A., & Benedek, G. B. (1997). Kinetic theory of fibrillogenesis of amyloid β-protein. *Proceedings of the National Academy of Sciences*, *94*(15), 7942-7947.

115. Szabo, Z., Klement, E., Jost, K., Zarandi, M., Soós, K., & Penke, B. (1999). An FT-IR study of the β-amyloid conformation: Standardization of aggregation grade. *Biochemical and biophysical research communications*, *265*(2), 297-300.

116. Weiss IV, W. F., Hodgdon, T. K., Kaler, E. W., Lenhoff, A. M., & Roberts, C. J. (2007). Nonnative protein polymers: structure, morphology, and relation to nucleation and growth. *Biophysical journal*, *93*(12), 4392-4403.

117. Sun, Y., & Hayakawa, S. (2002). Heat-induced gels of egg white/ovalbumins from five avian species: thermal aggregation, molecular forces involved, and rheological properties. *Journal of agricultural and food chemistry*, *50*(6), 1636-1642.

118. Sluzky, V., Tamada, J. A., Klibanov, A. M., & Langer, R. (1991). Kinetics of insulin aggregation in aqueous solutions upon agitation in the presence of hydrophobic surfaces. *Proceedings of the National Academy of Sciences*, *88*(21), 9377-9381.

119. Sluzky, V., Klibanov, A. M., & Langer, R. (1992). Mechanism of insulin aggregation and stabilization in agitated aqueous solutions. *Biotechnology and bioengineering*, *40*(8), 895-903.

120. Darrington, R. T., & Anderson, B. D. (1995). Evidence for a common intermediate in insulin deamidation and covalent dimer formation: effects of pH and aniline trapping in dilute acidic solutions. *Journal of pharmaceutical sciences*, *84*(3), 275-282.

121. Brange, J., Havelund, S., & Hougaard, P. (1992). Chemical stability of insulin. 2. Formation of higher molecular weight transformation products during storage of pharmaceutical preparations. *Pharmaceutical research*, *9*(6), 727-734.

122. Van Buren, N., Rehder, D., Gadgil, H., Matsumura, M., & Jacob, J. (2009). Elucidation of two major aggregation pathways in an IgG2 antibody. *Journal of pharmaceutical sciences*, *98*(9), 3013-3030.

123. Rousseau, F., Schymkowitz, J., & Serrano, L. (2006). Protein aggregation and amyloidosis: confusion of the kinds?. *Current opinion in structural biology*, *16*(1), 118-126.

124. Speed, M. A., King, J., & Wang, D. I. (1997). Polymerization mechanism of polypeptide chain aggregation. *Biotechnology and bioengineering*, *54*(4), 333-343.

125. Vetri, V., & Militello, V. (2005). Thermal induced conformational changes involved in the aggregation pathways of beta-lactoglobulin. *Biophysical chemistry*, *113*(1), 83-91.

126. Roy, S., Katayama, D., Dong, A., Kerwin, B. A., Randolph, T. W., & Carpenter, J. F. (2006). Temperature dependence of benzyl alcohol-and 8-anilinonaphthalene-1-sulfonate-induced aggregation of recombinant human interleukin-1 receptor antagonist. *Biochemistry*, *45*(12), 3898-3911.

127. Einstein, A. (1905). Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Annalen der physik*, *322*(8), 549-560.

128. Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry*, *29*(31), 7133-7155.

129. Jaenicke, R. (2000). Stability and stabilization of globular proteins in solution. *Journal of Biotechnology*, *79*(3), 193-203.

130. Pace, C. N., Shirley, B. A., McNutt, M., & Gajiwala, K. (1996). Forces contributing to the conformational stability of proteins. *The FASEB journal*, *10*(1), 75-83.

131. Chen, Y., Ding, F., Nie, H., Serohijos, A. W., Sharma, S., Wilcox, K. C., ... & Dokholyan, N. V. (2008). Protein folding: then and now. *Archives of biochemistry and biophysics*, *469*(1), 4-19.

132. Jaenicke, R. (1991). Protein stability and molecular adaptation to extreme conditions. In *EJB Reviews 1991* (pp. 291-304). Springer, Berlin, Heidelberg.

133. Talla-Singh, D., & Stites, W. E. (2008). Refinement of noncalorimetric determination of the change in heat capacity, ΔCp, of protein unfolding and validation across a wide temperature range. *Proteins: Structure, Function, and Bioinformatics*, *71*(4), 1607-1616.

134. Rees, D. C., & Robertson, A. D. (2001). Some thermodynamic implications for the thermostability of proteins. *Protein Science*, *10*(6), 1187-1194.

135. Giger, K., Vanam, R. P., Seyrek, E., & Dubin, P. L. (2008). Suppression of insulin aggregation by heparin. *Biomacromolecules*, *9*(9), 2338-2344.

136. Demeule, B., Lawrence, M. J., Drake, A. F., Gurny, R., & Arvinte, T. (2007). Characterization of protein aggregation: the case of a therapeutic immunoglobulin. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, *1774*(1), 146-153.

137. Boctor A. M., & Mehta S. C. (1992). Enhancement of the stability of thrombin by polyols: microcalorimetric studies. *Journal of pharmacy and pharmacology*, *44*(7), 600-603.

138. Charman, S. A., Mason, K. L., & Charman, W. N. (1993). Techniques for assessing the effects of pharmaceutical excipients on the aggregation of porcine growth hormone. *Pharmaceutical research*, *10*(7), 954-962.

139. Narhi, L. O., Philo, J. S., Sun, B., Chang, B. S., & Arakawa, T. (1999). Reversibility of heat-induced denaturation of the recombinant human megakaryocyte growth and development factor. *Pharmaceutical research*, *16*(6), 799-807.

140. Broersen, K., Weijers, M., de Groot, J., Hamer, R. J., & de Jongh, H. H. (2007). Effect of protein charge on the generation of aggregation-prone conformers. *Biomacromolecules*, *8*(5), 1648-1656.

141. Høiberg-Nielsen, R., Fuglsang, C. C., Arleth, L., & Westh, P. (2006). Interrelationships of glycosylation and aggregation kinetics for Peniophora lycii phytase. *Biochemistry*, *45*(15), 5057-5066.

142. Yan, Y. B., Wang, Q., He, H. W., & Zhou, H. M. (2004). Protein thermal aggregation involves distinct regions: sequential events in the heat-induced unfolding and aggregation of hemoglobin. *Biophysical journal*, *86*(3), 1682-1690.

143. Gaikwad, S. M., & Khan, M. I. (2003). pH-dependent aggregation of oligomeric Artocarpus hirsuta lectin on thermal denaturation. *Biochemical and biophysical research communications*, *311*(2), 254-257.

144. Barzegar, A., Moosavi-Movahedi, A. A., Rezaei-Zarchi, S., Saboury, A. A., Ganjali, M. R., Norouzi, P., ... & Tsai, F. Y. (2008). The mechanisms underlying the effect of α-cyclodextrin on the aggregation and stability of alcohol dehydrogenase. *Biotechnology and applied biochemistry*, *49*(3), 203-211.

145. Benjwal, S., Verma, S., Röhm, K. H., & Gursky, O. (2006). Monitoring protein aggregation during thermal unfolding in circular dichroism experiments. *Protein Science*, *15*(3), 635-639.

146. Golub, N. V., Markossian, K. A., Kasilovich, N. V., Sholukh, M. V., Orlov, V. N., & Kurganov, B. I. (2008). Thermal inactivation, denaturation and aggregation of mitochondrial aspartate aminotransferase. *Biophysical chemistry*, *135*(1-3), 125-131.

147. McGuffey, M. K., Epting, K. L., Kelly, R. M., & Foegeding, E. A. (2005). Denaturation and aggregation of three α-lactalbumin preparations at neutral pH. *Journal of agricultural and food chemistry*, *53*(8), 3182-3190.

148. Derrick, T., Grillo, A. O., Vitharana, S. N., Jones, L., Rexroad, J., Shah, A., ... & Middaugh, C. R. (2007). Effect of polyanions on the structure and stability of repifermin™(keratinocyte growth factor-2). *Journal of pharmaceutical sciences*, *96*(4), 761-776.

149. Arnaudov, L. N., & de Vries, R. (2005). Thermally induced fibrillar aggregation of hen egg white lysozyme. *Biophysical Journal*, *88*(1), 515-526.

150. Matsuoka, T., Tomita, S., Hamada, H., & Shiraki, K. (2007). Amidated amino acids are prominent additives for preventing heat-induced aggregation of lysozyme. *Journal of bioscience and bioengineering*, *103*(5), 440-443.

151. Holm, N. K., Jespersen, S. K., Thomassen, L. V., Wolff, T. Y., Sehgal, P., Thomsen, L. A., ... & Otzen, D. E. (2007). Aggregation and fibrillation of bovine serum albumin. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, *1774*(9), 1128-1138.

152. Purohit, V. S., Middaugh, C. R., & Balasubramanian, S. V. (2006). Influence of aggregation on immunogenicity of recombinant human Factor VIII in hemophilia A mice. *Journal of pharmaceutical sciences*, *95*(2), 358-371.

153. Livney, Y. D., & Dalgleish, D. G. (2004). Specificity of disulfide bond formation during thermal aggregation in solutions of β-lactoglobulin B and κ-casein A. *Journal of agricultural and food chemistry*, *52*(17), 5527-5532.

154. Andrews, J. M., Weiss IV, W. F., & Roberts, C. J. (2008). Nucleation, growth, and activation energies for seeded and unseeded aggregation of α-chymotrypsinogen A. *Biochemistry*, *47*(8), 2397-2403.

155. Boeris, V., Spelzini, D., Salgado, J. P., Picó, G., Romanini, D., & Farruggia, B. (2008). Chymotrypsin–poly vinyl sulfonate interaction studied by dynamic light scattering and turbidimetric approaches. *Biochimica et Biophysica Acta (BBA)-General Subjects*, *1780*(9), 1032-1037.

156. Fedurkina, N. V., Belousova, L. V., Mitskevich, L. G., Zhou, H. M., Chang, Z., & Kurganov, B. I. (2006). Change in kinetic regime of protein aggregation with temperature increase. Thermal aggregation of rabbit muscle creatine kinase. *Biochemistry (Moscow)*, *71*(3), 325-331.

157. Vermeer, A. W., & Norde, W. (2000). The thermal stability of immunoglobulin: unfolding and aggregation of a multi-domain protein. *Biophysical journal*, *78*(1), 394-404.

158. Pace, C. N., & Tanford, C. (1968). Thermodynamics of the unfolding of β-lactoglobulin A in aqueous urea solutions between 5 and 55. *Biochemistry*, *7*(1), 198-208.

159. Privalov, P. L., Griko, Y. V., Venyaminov, S. Y., & Kutyshenko, V. P. (1986). Cold denaturation of myoglobin. *Journal of molecular biology*, *190*(3), 487-498.

160. Jonas, J. (1997). Cold denaturation of proteins.

161. Lazar, K. L., Patapoff, T. W., & Sharma, V. K. (2010, January). Cold denaturation of monoclonal antibodies. In *MAbs* (Vol. 2, No. 1, pp. 42-52). Taylor & Francis.

162. Sanfelice, D., & Temussi, P. A. (2016). Cold denaturation as a tool to measure protein stability. *Biophysical chemistry*, *208*, 4-8.

163. Zhang, J., Peng, X., Jonas, A., & Jonas, J. (1995). NMR study of the cold, heat, and pressure unfolding of ribonuclease A. *Biochemistry*, *34*(27), 8631-8641.

164. Tanaka, N., Nishizawa, H., & Kunugi, S. (1997). Structure of pressure-induced denatured state of human serum albumin: a comparison with the intermediate in urea-induced denaturation. *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology*, *1338*(1), 13-20.

165. Ferri, C., Zignego, A. L., & Pileri, S. A. (2002). Cryoglobulins. *Journal of Clinical Pathology*, *55*(1), 4-13.

166. Ramsland, P. A., & Farrugia, W. (2002). Crystal structures of human antibodies: a detailed and unfinished tapestry of immunoglobulin gene products. *Journal of Molecular Recognition*, *15*(5), 248-259.

167. Sukumar, M., Doyle, B. L., Combs, J. L., & Pekar, A. H. (2004). Opalescent appearance of an IgG1 antibody at high concentrations and its relationship to noncovalent association. *Pharmaceutical research*, *21*(7), 1087-1093.

168. Piedmonte, D. M., Summers, C., McAuley, A., Karamujic, L., & Ratnaswamy, G. (2007). Sorbitol crystallization can lead to protein aggregation in frozen protein formulations. *Pharmaceutical research*, *24*(1), 136-146.

169. Heller, M. C., Carpenter, J. F., & Randolph, T. W. (1997). Manipulation of lyophilization-induced phase separation: Implications for pharmaceutical proteins. *Biotechnology progress*, *13*(5), 590-596.

170. Heller, M. C., Carpenter, J. F., & Randolph, T. W. (1999). Protein formulation and lyophilization cycle design: Prevention of damage due to freeze-concentration induced phase separation. *Biotechnology and bioengineering*, *63*(2), 166-174.

171. Pikal-Cleland, K. A., Rodríguez-Hornedo, N., Amidon, G. L., & Carpenter, J. F. (2000). Protein denaturation during freezing and thawing in phosphate buffer systems: monomeric and tetrameric β-galactosidase. *Archives of Biochemistry and Biophysics*, *384*(2), 398-406.

172. Schein, C. H. (1990). Solubility as a function of protein structure and solvent components. *Bio/technology*, *8*(4), 308.

173. Chiti, F., Stefani, M., Taddei, N., Ramponi, G., & Dobson, C. M. (2003). Rationalization of the effects of mutations on peptide andprotein aggregation rates. *Nature*, *424*(6950), 805.

174. DuBay, K. F., Pawar, A. P., Chiti, F., Zurdo, J., Dobson, C. M., & Vendruscolo, M. (2004). Prediction of the absolute aggregation rates of amyloidogenic polypeptide chains. *Journal of molecular biology*, *341*(5), 1317-1326.

175. Bajaj, H., Sharma, V. K., Badkar, A., Zeng, D., Nema, S., & Kalonia, D. S. (2006). Protein structural conformation and not second virial coefficient relates to long-term irreversible aggregation of a monoclonal antibody and ovalbumin in solution. *Pharmaceutical research*, *23*(6), 1382.

176. Barteri, M., Gaudiano, M. C., Rotella, S., Benagiano, G., & Pala, A. (2000). Effect of pH on the structure and aggregation of human glycodelin A. A comparison with β-lactoglobulin A. *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology*, *1479*(1-2), 255-264.

177. Majhi, P. R., Ganta, R. R., Vanam, R. P., Seyrek, E., Giger, K., & Dubin, P. L. (2006). Electrostatically driven protein aggregation: β-lactoglobulin at low ionic strength. *Langmuir*, *22*(22), 9150-9159.

178. Peters, B., & Trout, B. L. (2006). Asparagine deamidation: pH-dependent mechanism from density functional theory. *Biochemistry*, *45*(16), 5384-5392.

179. Kamat, M. S., Tolman, G. L., & Brown, J. M. (2002). Formulation development of an antifibrin monoclonal antibody radiopharmaceutical. In *Formulation, Characterization, and Stability of Protein Drugs: Case Histories* (pp. 343-364). Springer, Boston, MA.

180. Vrkljan, M., Foster, T. M., Powers, M. E., Henkin, J., Porter, W. R., Staack, H., ... & Manning, M. C. (1994). Thermal stability of low molecular weight urokinase during heat treatment. II. Effect of polymeric additives. *Pharmaceutical research*, *11*(7), 1004-1008.

181. Fatouros, A., Österberg, T., & Mikaelsson, M. (1997). Recombinant factor VIII SQ—influence of oxygen, metal ions, pH and ionic strength on its stability in aqueous solution. *International journal of pharmaceutics*, *155*(1), 121-131.

182. Pikal, M. J., Dellerman, K. M., Roy, M. L., & Riggin, R. M. (1991). The effects of formulation variables on the stability of freeze-dried human growth hormone. *Pharmaceutical Research*, *8*(4), 427-436.

183. Won, C. M., Molnar, T. E., McKean, R. E., & Spenlehauer, G. A. (1998). Stabilizers against heat-induced aggregation of RPR 114849, an acidic fibroblast growth factor (aFGF). *International journal of pharmaceutics*, *167*(1-2), 25-36.

184. Eberlein, G. A., Stratton, P. R., & Wang, Y. J. (1994). Stability of rhbFGF as determined by UV spectroscopic measurements of turbidity. *PDA journal of pharmaceutical science and technology*, *48*(5), 224-230.

185. Paborji, M., Pochopin, N. L., Coppola, W. P., & Bogardus, J. B. (1994). Chemical and physical stability of chimeric L6, a mouse− human monoclonal antibody. *Pharmaceutical research*, *11*(5), 764-771.

186. Raibekas, A. A., Bures, E. J., Siska, C. C., Kohno, T., Latypov, R. F., & Kerwin, B. A. (2005). Anion binding and controlled aggregation of human interleukin-1 receptor antagonist. *Biochemistry*, *44*(29), 9871-9879.

187. Katayama, D. S., Nayar, R., Chou, D. K., Valente, J. J., Cooper, J., Henry, C. S., ... & Manning, M. C. (2006). Effect of buffer species on the thermally induced aggregation of interferon-tau. *Journal of pharmaceutical sciences*, *95*(6), 1212-1226.

188. Chung, K., Kim, J., Cho, B. K., Ko, B. J., Hwang, B. Y., & Kim, B. G. (2007). How does dextran sulfate prevent heat induced aggregation of protein?: The mechanism and its limitation as aggregation inhibitor. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, *1774*(2), 249-257.

189. Long, X., Zhang, C., Cheng, J., & Bi, S. (2008). A novel method for study of the aggregation of protein induced by metal ion aluminum (III) using resonance Rayleigh scattering technique. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *69*(1), 71-77.

190. Kita, Y., & Arakawa, T. (2002). Salts and glycine increase reversibility and decrease aggregation during thermal unfolding of ribonuclease-A. *Bioscience, biotechnology, and biochemistry*, *66*(4), 880-882.

191. Alford, J. R., Kendrick, B. S., Carpenter, J. F., & Randolph, T. W. (2008). High concentration formulations of recombinant human interleukin-1 receptor antagonist: II. Aggregation kinetics. *Journal of pharmaceutical sciences*, *97*(8), 3005-3021.

192. Baussay, K., Le Bon, C., Nicolai, T., Durand, D., & Busnel, J. P. (2004). Influence of the ionic strength on the heat-induced aggregation of the globular protein β-lactoglobulin at pH 7. *International Journal of Biological Macromolecules*, *34*(1-2), 21-28.

193. Arnaudov, L. N., & de Vries, R. (2006). Strong impact of ionic strength on the kinetics of fibrilar aggregation of bovine β-lactoglobulin. *Biomacromolecules*, *7*(12), 3490-3498.

194. Kerstens, S., Murray, B. S., & Dickinson, E. (2005). Confocal microscopy of heat-induced aggregation and gelation of β-lactoglobulin in presence of non-ionic surfactant. *Food Hydrocolloids*, *19*(3), 625-633.

195. Hofmeister, F. (1888). Zur lehre von der wirkung der salze. *Archiv für experimentelle Pathologie und Pharmakologie*, *25*(1), 1-30.

196. Curtis, R. A., Ulrich, J., Montaser, A., Prausnitz, J. M., & Blanch, H. W. (2002). Protein–protein interactions in concentrated electrolyte solutions. *Biotechnology and bioengineering*, *79*(4), 367-380.

197. Ellis, R. J. (2001). Macromolecular crowding: obvious but underappreciated. *Trends in biochemical sciences*, *26*(10), 597-604.

198. Minton, A. P. (2006). Macromolecular crowding. *Current Biology*, *16*(8), R269-R271.

199. Jiménez, M., Rivas, G., & Minton, A. P. (2007). Quantitative characterization of weak self-association in concentrated solutions of immunoglobulin G via the measurement of sedimentation equilibrium and osmotic pressure. *Biochemistry*, *46*(28), 8373-8378.

200. Golub, N., Meremyanin, A., Markossian, K., Eronina, T., Chebotareva, N., Asryants, R., ... & Kurganov, B. (2007). Evidence for the formation of start aggregates as an initial stage of protein aggregation. *FEBS letters*, *581*(22), 4223-4227.

201. Treuheit, M. J., Kosky, A. A., & Brems, D. N. (2002). Inverse relationship of protein concentration and aggregation. *Pharmaceutical research*, *19*(4), 511-516.

202. Shukla, A. A., Gupta, P., & Han, X. (2007). Protein aggregation kinetics during Protein A chromatography: case study for an Fc fusion protein. *Journal of Chromatography A*, *1171*(1-2), 22-28.

203. Raso, S. W., Abel, J., Barnes, J. M., Maloney, K. M., Pipes, G., Treuheit, M. J., ... & Brems, D. N. (2005). Aggregation of granulocyte-colony stimulating factor in vitro involves a conformationally altered monomeric state. *Protein science*, *14*(9), 2246-2257.

204. Elshereef, R., Budman, H., Moresoli, C., & Legge, R. L. (2008). Fluorescence-based soft-sensor for monitoring β-lactoglobulin and α-lactalbumin solubility during thermal aggregation. *Biotechnology and bioengineering*, *99*(3), 567-577.

205. Saluja, A., Badkar, A. V., Zeng, D. L., Nema, S., & Kalonia, D. S. (2007). Ultrasonic storage modulus as a novel parameter for analyzing protein-protein interactions in high protein concentration solutions: correlation with static and dynamic light scattering measurements. *Biophysical journal*, *92*(1), 234-244.

206. Pease III, L. F., Elliott, J. T., Tsai, D. H., Zachariah, M. R., & Tarlov, M. J. (2008). Determination of protein aggregation with differential mobility analysis: application to IgG antibody. *Biotechnology and bioengineering*, *101*(6), 1214-1222.

207. Roy, S., Henderson, I., Nayar, R., Randolph, T. W., & Carpenter, J. F. (2008). Effect of pH on stability of recombinant botulinum serotype A vaccine in aqueous solution and during storage of freeze-dried formulations. *Journal of pharmaceutical sciences*, *97*(12), 5132-5146.

208. Rezaei-Ghaleh, N., Ramshini, H., Ebrahim-Habibi, A., Moosavi-Movahedi, A. A., & Nemat-Gorgani, M. (2008). Thermal aggregation of α-chymotrypsin: role of hydrophobic and electrostatic interactions. *Biophysical chemistry*, *132*(1), 23-32.

209. Ramani, K., Purohit, V., Middaugh, C. R., & Balasubramanian, S. V. (2005). Aggregation kinetics of recombinant human FVIII (rFVIII). *Journal of pharmaceutical sciences*, *94*(9), 2023-2029.

210. Roefs, S. P., & De Kruif, K. G. (1994). A model for the denaturation and aggregation of β-lactoglobulin. *European Journal of Biochemistry*, *226*(3), 883-889.

211. Gu, L. C., Erdös, E. A., Chiang, H. S., Calderwood, T., Tsai, K., Visor, G. C., ... & Foster, L. C. (1991). Stability of interleukin 1β (IL-1β) in aqueous solution: analytical methods, kinetics, products, and solution formulation implications. *Pharmaceutical research*, *8*(4), 485-490.

212. Clark, A. C. (2008). Protein folding: Are we there yet?. *Archives of biochemistry and biophysics*, *469*(1), 1.

213. Brange, J., Andersen, L., Laursen, E. D., Meyn, G., & Rasmussen, E. (1997). Toward understanding insulin fibrillation. *Journal of pharmaceutical sciences*, *86*(5), 517-525.

214. Fields, G. B., Alonso, D. O., Stigter, D., & Dill, K. A. (1992). Theory for the aggregation of proteins and copolymers. *The Journal of Physical Chemistry*, *96*(10), 3974-3981.

215. Zlateva, T., Boteva, R., Salvato, B., & Tsanev, R. (1999). Factors affecting the dissociation and aggregation of human interferon gamma. *International journal of biological macromolecules*, *26*(5), 357-362.

216. Arakawa, T., & Timasheff, S. N. (1985). The stabilization of proteins by osmolytes. *Biophysical journal*, *47*(3), 411-414.

217. Xia, Y., Park, Y. D., Mu, H., Zhou, H. M., Wang, X. Y., & Meng, F. G. (2007). The protective effects of osmolytes on arginine kinase unfolding and aggregation. *International journal of biological macromolecules*, *40*(5), 437-443.

218. Goyal, K., Walton, L. J., & Tunnacliffe, A. (2005). LEA proteins prevent protein aggregation due to water stress. *Biochemical Journal*, *388*(1), 151-157.

219. Béranger, F., Crozet, C., Goldsborough, A., & Lehmann, S. (2008). Trehalose impairs aggregation of PrPSc molecules and protects prion-infected cells against oxidative damage. *Biochemical and biophysical research communications*, *374*(1), 44-48.

220. Vilasi, S., Iannuzzi, C., Portaccio, M., Irace, G., & Sirangelo, I. (2008). Effect of trehalose on W7FW14F apomyoglobin and insulin fibrillization: new insight into inhibition activity. *Biochemistry*, *47*(6), 1789-1796.

221. Petersen, S. B., Jonson, V., Fojan, P., Wimmer, R., & Pedersen, S. (2004). Sorbitol prevents the self-aggregation of unfolded lysozyme leading to an up to 13 C stabilisation of the folded form. *Journal of biotechnology*, *114*(3), 269-278.

222. Bagger, H. L., Øgendal, L. H., & Westh, P. (2007). Solute effects on the irreversible aggregation of serum albumin. *Biophysical chemistry*, *130*(1-2), 17-25.

223. Schüle, S., Schulz-Fademrecht, T., Garidel, P., Bechtold-Peters, K., & Frieß, W. (2008). Stabilization of IgG1 in spray-dried powders for inhalation. *European Journal of Pharmaceutics and Biopharmaceutics*, *69*(3), 793-807.

224. Rondeau, P., Armenta, S., Caillens, H., Chesne, S., & Bourdon, E. (2007). Assessment of temperature effects on β-aggregation of native and glycated albumin by FTIR spectroscopy and PAGE: relations between structural changes and antioxidant properties. *Archives of biochemistry and biophysics*, *460*(1), 141-150.

225. DePaz, R. A., Barnett, C. C., Dale, D. A., Carpenter, J. F., Gaertner, A. L., & Randolph, T. W. (2000). The excluding effects of sucrose on a protein chemical degradation pathway: methionine oxidation in subtilisin. *Archives of biochemistry and biophysics*, *384*(1), 123-132.

226. Joo, H. S., Koo, Y. M., Choi, J. W., & Chang, C. S. (2005). Stabilization method of an alkaline protease from inactivation by heat, SDS and hydrogen peroxide. *Enzyme and microbial technology*, *36*(5-6), 766-772.

227. Li, S., Patapoff, T. W., Nguyen, T. H., & Borchardt, R. T. (1996). Inhibitory effect of sugars and polyols on the metal-catalyzed oxidation of human relaxin. *Journal of pharmaceutical sciences*, *85*(8), 868-872.

228. Maclean, D. S., Qian, Q., & Middaugh, C. R. (2002). Stabilization of proteins by low molecular weight multi-ions. *Journal of pharmaceutical sciences*, *91*(10), 2220-2229.

229. Ignatova, Z., & Gierasch, L. M. (2006). Inhibition of protein aggregation in vitro and in vivo by a natural osmoprotectant. *Proceedings of the National Academy of Sciences*, *103*(36), 13357-13361.

230. Chen, B., Bautista, R., Yu, K., Zapata, G. A., Mulkerrin, M. G., & Chamow, S. M. (2003). Influence of histidine on the stability and physical properties of a fully human antibody in aqueous and solid forms. *Pharmaceutical research*, *20*(12), 1952-1960.

231. Tsumoto, K., Ejima, D., Kita, Y., & Arakawa, T. (2005). Why is arginine effective in suppressing aggregation?. *Protein and Peptide Letters*, *12*(7), 613-619.

232. Hamada, H., Takahashi, R., Noguchi, T., & Shiraki, K. (2008). Differences in the effects of solution additives on heat-and refolding-induced aggregation. *Biotechnology progress*, *24*(2), 436-443.

233. Lyutova, E. M., Kasakov, A. S., & Gurvits, B. Y. (2007). Effects of arginine on kinetics of protein aggregation studied by dynamic laser light scattering and tubidimetry techniques. *Biotechnology progress*, *23*(6), 1411-1416.

234. Das, U., Hariprasad, G., Ethayathulla, A. S., Manral, P., Das, T. K., Pasha, S., ... & Chandrayan, S. K. (2007). Inhibition of protein aggregation: supramolecular assemblies of arginine hold the key. *PloS one*, *2*(11), e1176.

235. Ignatova, Z., & Gierasch, L. M. (2007). Effects of osmolytes on protein folding and aggregation in cells. In *Methods in enzymology* (Vol. 428, pp. 355-372). Academic Press.

236. Kar, K., & Kishore, N. (2007). Enhancement of thermal stability and inhibition of protein aggregation by osmolytic effect of hydroxyproline. *Biopolymers: Original Research on Biomolecules*, *87*(5-6), 339-351.

237. Unterhaslberger, G., Schmitt, C., Sanchez, C., Appolonia-Nouzille, C., & Raemy, A. (2006). Heat denaturation and aggregation of β-lactoglobulin enriched WPI in the presence of arginine HCl, NaCl and guanidinium HCl at pH 4.0 and 7.0. *Food Hydrocolloids*, *20*(7), 1006-1019.

238. Rezaei-Ghaleh, N., Ebrahim-Habibi, A., Moosavi-Movahedi, A. A., & Nemat-Gorgani, M. (2007). Effect of polyamines on the structure, thermal stability and 2, 2, 2-trifluoroethanol-induced aggregation of α-chymotrypsin. *International journal of biological macromolecules*, *41*(5), 597-604.

239. Schlieben, N. H., Niefind, K., & Schomburg, D. (2004). Expression, purification, and aggregation studies of His-tagged thermoalkalophilic lipase from Bacillus thermocatenulatus. *Protein expression and purification*, *34*(1), 103-110.

240. Eronina, T. B., Chebotareva, N. A., & Kurganov, B. I. (2005). Influence of osmolytes on inactivation and aggregation of muscle glycogen phosphorylase b by guanidine hydrochloride. Stimulation of protein aggregation under crowding conditions. *Biochemistry (Moscow)*, *70*(9), 1020-1026.

241. Moussa, C. E. H., Mahmoodian, F., Tomita, Y., & Sidhu, A. (2008). Dopamine differentially induces aggregation of A53T mutant and wild type α-synuclein: insights into the protein chemistry of Parkinson's disease. *Biochemical and biophysical research communications*, *365*(4), 833-839.

242. Joshi, S. B., Kamerzell, T. J., McNown, C., & Middaugh, C. R. (2008). The interaction of heparin/polyanions with bovine, porcine, and human growth hormone. *Journal of pharmaceutical sciences*, *97*(4), 1368-1385.

243. Arakawa, T., Wen, J., & Philo, J. S. (1993). Densimetric determination of equilibrium binding of sucrose octasulfate with basic fibroblast growth factor. *Journal of protein chemistry*, *12*(6), 689-693.

244. Considine, T., Patel, H. A., Singh, H., & Creamer, L. K. (2007). Influence of binding conjugated linoleic acid and myristic acid on the heat-and high-pressure-induced unfolding and aggregation of β-lactoglobulin B. *Food chemistry*, *102*(4), 1270-1280.

245. Lee, H. J., McAuley, A., Schilke, K. F., & McGuire, J. (2011). Molecular origins of surfactant-mediated stabilization of protein drugs. *Advanced drug delivery reviews*, *63*(13), 1160-1171.

246. Kerwin, B. A. (2008). Polysorbates 20 and 80 used in the formulation of protein biotherapeutics: structure and degradation pathways. *Journal of pharmaceutical sciences*, *97*(8), 2924-2935.

247. Randolph, T. W., & Jones, L. S. (2002). Surfactant-protein interactions. In *Rational design of stable protein formulations* (pp. 159-175). Springer, Boston, MA.

248. Bam, N. B., Randolph, T. W., & Cleland, J. L. (1995). Stability of protein formulations: investigation of surfactant effects by a novel EPR spectroscopic technique. *Pharmaceutical research*, *12*(1), 2-11.

249. Chou, D. K., Krishnamurthy, R., Randolph, T. W., Carpenter, J. F., & Manning, M. C. (2005). Effects of Tween 20® and Tween 80® on the stability of Albutropin during agitation. *Journal of pharmaceutical sciences*, *94*(6), 1368-1381.

250. Bam, N. B., Cleland, J. L., Yang, J., Manning, M. C., Carpenter, J. F., Kelley, R. F., & Randolph, T. W. (1998). Tween protects recombinant human growth hormone against agitation-induced damage via hydrophobic interactions. *Journal of pharmaceutical sciences*, *87*(12), 1554-1559.

251. Arakawa, T., & Kita, Y. (2000). Protection of bovine serum albumin from aggregation by Tween 80. *Journal of Pharmaceutical Sciences*, *89*(5), 646-651.

252. Katakam, M., & Banga, A. K. (1997). Use of poloxamer polymers to stabilize recombinant human growth hormone against various processing stresses. *Pharmaceutical development and technology*, *2*(2), 143-149.

253. Rozema, D., & Gellman, S. H. (1996). Artificial chaperone-assisted refolding of carbonic anhydrase B. *Journal of Biological Chemistry*, *271*(7), 3478-3487.

254. Panyukov, Y. V., Nemykh, M. A., Dobrov, E. N., & Drachev, V. A. (2008). Surfactant-induced amorphous aggregation of tobacco mosaic virus coat protein: A physical methods approach. *Macromolecular bioscience*, *8*(2), 199-209.

255. Collins, T., D'Amico, S., Georlette, D., Marx, J. C., Huston, A. L., & Feller, G. (2006). A nondetergent sulfobetaine prevents protein aggregation in microcalorimetric studies. *Analytical biochemistry*, *352*(2), 299-301.

256. Holm, N. K., Jespersen, S. K., Thomassen, L. V., Wolff, T. Y., Sehgal, P., Thomsen, L. A., ... & Otzen, D. E. (2007). Aggregation and fibrillation of bovine serum albumin. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, *1774*(9), 1128-1138.

257. Yazdanparast, R., Esmaeili, M. A., & Khodagholi, F. (2007). Control of aggregation in protein refolding: Cooperative effects of artificial chaperone and cold temperature. *International journal of biological macromolecules*, *40*(2), 126-133.

258. Panyukov, Y. V., Nemykh, M. A., Rafikova, E. R., Kurganov, B. I., Yaguzhinsky, L. S., Arutyunyan, A. M., ... & Dobrov, E. N. (2006). Low cetyltrimethylammonium bromide concentrations induce reversible amorphous aggregation of tobacco mosaic virus and its coat protein at room temperature. *The international journal of biochemistry & cell biology*, *38*(4), 533-543.

259. Calamai, M., Taddei, N., Stefani, M., Ramponi, G., & Chiti, F. (2003). Relative influence of hydrophobicity and net charge in the aggregation of two homologous proteins. *Biochemistry*, *42*(51), 15078-15083.

260. Chi, E. Y., Kendrick, B. S., Carpenter, J. F., & Randolph, T. W. (2005). Population balance modeling of aggregation kinetics of recombinant human interleukin-1 receptor antagonist. *Journal of pharmaceutical sciences*, *94*(12), 2735-2748.

261. Zhang, Y., Roy, S., Jones, L. S., Krishnan, S., Kerwin, B. A., Chang, B. S., ... & Carpenter, J. F. (2004). Mechanism for benzyl alcohol-induced aggregation of recombinant human interleukin-1 receptor antagonist in aqueous solution. *Journal of pharmaceutical sciences*, *93*(12), 3076-3089.

262. Roy, S., Jung, R., Kerwin, B. A., Randolph, T. W., & Carpenter, J. F. (2005). Effects of benzyl alcohol on aggregation of recombinant human interleukin-1-receptor antagonist in reconstituted lyophilized formulations. *Journal of pharmaceutical sciences*, *94*(2), 382-396.

263. Thirumangalathu, R., Krishnan, S., Brems, D. N., Randolph, T. W., & Carpenter, J. F. (2006). Effects of pH, temperature, and sucrose on benzyl alcohol-induced aggregation of recombinant human granulocyte colony stimulating factor. *Journal of pharmaceutical sciences*, *95*(7), 1480-1497.

264. Baynes, B. M., & Trout, B. L. (2004). Rational design of solution additives for the prevention of protein aggregation. *Biophysical journal*, *87*(3), 1631-1639.

265. Zhang, M. Z., Wen, J., Arakawa, T., & Prestrelski, S. J. (1995). A new strategy for enhancing the stability of lyophilized protein: the effect of the reconstitution medium on keratinocyte growth factor. *Pharmaceutical research*, *12*(10), 1447-1452.

266. Vemuri, S., Yu, C. T., & Roosdorp, N. (1993). Formulation and stability of recombinant α 1-antitrypsin. In *Stability and Characterization of Protein and Peptide Drugs* (pp. 263-286). Springer, Boston, MA.

267. Munishkina, L. A., Ahmad, A., Fink, A. L., & Uversky, V. N. (2008). Guiding protein aggregation with macromolecular crowding. *Biochemistry*, *47*(34), 8993-9006.

268. Chen, B. L., Arakawa, T., Morris, C. F., Kenney, W. C., Wells, C. M., & Pitt, C. G. (1994). Aggregation pathway of recombinant human keratinocyte growth factor and its stabilization. *Pharmaceutical research*, *11*(11), 1581-1587.

269. Tsai, A. M., van Zanten, J. H., & Betenbaugh, M. J. (1998). II. Electrostatic effect in the aggregation of heat-denatured RNase A and implications for protein additive design. *Biotechnology and bioengineering*, *59*(3), 281-285.

270. Chung, K., Kim, J., Cho, B. K., Ko, B. J., Hwang, B. Y., & Kim, B. G. (2007). How does dextran sulfate prevent heat induced aggregation of protein?: The mechanism and its limitation as aggregation inhibitor. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, *1774*(2), 249-257.

271. Giese, A., Bader, B., Bieschke, J., Schaffar, G., Odoy, S., Kahle, P. J., ... & Kretzschmar, H. (2005). Single particle detection and characterization of synuclein co-aggregation. *Biochemical and biophysical research communications*, *333*(4), 1202-1210.

272. Song, C., Wang, Q., & Li, C. C. H. (2007). Characterization of the aggregation-prevention activity of p97/valosin-containing protein. *Biochemistry*, *46*(51), 14889-14898.

273. Manning, M. C., Matsuura, J. E., Kendrick, B. S., Meyer, J. D., Dormish, J. J., Vrkljan, M., ... & Sheftert, E. (1995). Approaches for increasing the solution stability of proteins. *Biotechnology and bioengineering*, *48*(5), 506-512.

274. Kehoe, J. J., Morris, E. R., & Brodkorb, A. (2007). The influence of bovine serum albumin on β-lactoglobulin denaturation, aggregation and gelation. *Food hydrocolloids*, *21*(5-6), 747-755.

275. Pivovarova, A. V., Chebotareva, N. A., Chernik, I. S., Gusev, N. B., & Levitsky, D. I. (2007). Small heat shock protein Hsp27 prevents heat-induced aggregation of F-actin by forming soluble complexes with denatured actin. *The FEBS journal*, *274*(22), 5937-5948.

276. Pivovarova, A. V., Mikhailova, V. V., Chernik, I. S., Chebotareva, N. A., Levitsky, D. I., & Gusev, N. B. (2005). Effects of small heat shock proteins on the thermal denaturation and aggregation of F-actin. *Biochemical and biophysical research communications*, *331*(4), 1548-1553.

277. Yamashita, H., Kawamata, J., Okawa, K., Kanki, R., Nakamizo, T., Hatayama, T., ... & Shimohama, S. (2007). Heat-shock protein 105 interacts with and suppresses aggregation of mutant Cu/Zn superoxide dismutase: clues to a possible strategy for treating ALS. *Journal of neurochemistry*, *102*(5), 1497-1505.

278. Lee, S., Carson, K., Rice-Ficht, A., & Good, T. (2006). Small heat shock proteins differentially affect Aβ aggregation and toxicity. *Biochemical and biophysical research communications*, *347*(2), 527-533.

279. Kubota, S., Kubota, H., & Nagata, K. (2006). Cytosolic chaperonin protects folding intermediates of Gβ from aggregation by recognizing hydrophobic β-strands. *Proceedings of the National Academy of Sciences*, *103*(22), 8360-8365.

280. Estrada, L. D., & Soto, C. (2006). Inhibition of protein misfolding and aggregation by small rationally-designed peptides. *Current pharmaceutical design*, *12*(20), 2557-2567.

281. Etienne, M. A., Aucoin, J. P., Fu, Y., McCarley, R. L., & Hammer, R. P. (2006). Stoichiometric inhibition of amyloid β-protein aggregation with peptides containing alternating α, α-disubstituted amino acids. *Journal of the American Chemical Society*, *128*(11), 3522-3523.

282. Zhang, L., Tan, J., Han, D., & Zhu, H. (2017). From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug discovery today*, *22*(11), 1680-1685.

283. Witten, I. H., Frank, E., & Mark, A. (2011). Hall. 2011. *Data Mining: Practical machine learning tools and techniques*, *3*.

284. Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (1998). *Multivariate data analysis* (Vol. 5, No. 3, pp. 207-219). Upper Saddle River, NJ: Prentice hall.

285. Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of machine learning*. MIT press.

286. Pólya, G. (1920). Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentenproblem. *Mathematische Zeitschrift*, *8*(3), 171-181.

287. Student. (1908). The probable error of a mean. *Biometrika*, 1-25.

288. Haykin, S. (1994). *Neural networks: a comprehensive foundation*. Prentice Hall PTR.

289. Gentiluomo, L., Roessner, D., Augustijn, D., Svilenov, H., Kulakova, A., Mahapatra, S., ... & Harris, P. (2019). Application of interpretable artificial neural networks to early monoclonal antibodies development. *European Journal of Pharmaceutics and Biopharmaceutics*, *141*, 81-89.

290. Hussain, A. S., Yu, X., & Johnson, R. D. (1991). Application of neural computing in pharmaceutical product development. *Pharmaceutical research*, *8*(10), 1248-1252.

291. Takayama, K., Fujikawa, M., & Nagai, T. (1999). Artificial neural network as a novel method to optimize pharmaceutical formulations. *Pharmaceutical research*, *16*(1), 1-6.

292. Achanta, A. S., Kowalski, J. G., & Rhodes, C. T. (1995). Artificial neural networks: implications for pharmaceutical sciences. *Drug Development and Industrial Pharmacy*, *21*(1), 119-155.

293. King, A. C., Woods, M., Liu, W., Lu, Z., Gill, D., & Krebs, M. R. (2011). High-throughput measurement, correlation analysis, and machine-learning predictions for pH and thermal stabilities of Pfizer-generated antibodies. *Protein Science*, *20*(9), 1546-1557.

294. Yang, Y., Ye, Z., Su, Y., Zhao, Q., Li, X., & Ouyang, D. (2019). Deep learning for in vitro prediction of pharmaceutical formulations. *Acta pharmaceutica sinica B*, *9*(1), 177-185.

295. Ye, Z., Yang, Y., Li, X., Cao, D., & Ouyang, D. (2018). An Integrated Transfer Learning and Multitask Learning Approach for Pharmacokinetic Parameter Prediction. *Molecular pharmaceutics*, *16*(2), 533-541.

296. Bourquin, J., Schmidli, H., van Hoogevest, P., & Leuenberger, H. (1997). Application of artificial neural networks (ANN) in the development of solid dosage forms. *Pharmaceutical development and technology*, *2*(2), 111-121.

297. Plumb, A. P., Rowe, R. C., York, P., & Doherty, C. (2002). The effect of experimental design on the modeling of a tablet coating formulation using artificial neural networks. *European journal of pharmaceutical sciences*, *16*(4-5), 281-288.

298. Aksu, B., Paradkar, A., de Matas, M., Özer, Ö., Güneri, T., & York, P. (2012). Quality by design approach: application of artificial intelligence techniques of tablets manufactured by direct compression. *AAPS PharmSciTech*, *13*(4), 1138-1146.

299. Gambe-Gilbuena, A., Shibano, Y., Krayukhina, E., Torisu, T., & Uchiyama, S. (2020). Automatic Identification of the Stress Sources of Protein Aggregates Using Flow Imaging Microscopy Images. *Journal of pharmaceutical sciences*, *109*(1), 614-623.

300. Bohren, C. F., & Huffman, D. R. (2008). *Absorption and scattering of light by small particles*. John Wiley & Sons.

301. Einstein, A. (1910). Theorie der Opaleszenz von homogenen Flüssigkeiten und Flüssigkeitsgemischen in der Nähe des kritischen Zustandes. *Annalen der Physik*, *338*(16), 1275-1298.

302. Krishnan, R. S., & Shankar, R. K. (1981). Raman effect: History of the discovery. *Journal of Raman Spectroscopy*, *10*(1), 1-8.

303. Debye, P. (1944). Light scattering in solutions. *Journal of Applied Physics*, *15*(4), 338-342.

304. Zimm, B. H. (1948). The scattering of light and the radial distribution function of high polymer solutions. *The Journal of Chemical Physics*, *16*(12), 1093-1099.

305. Zimm, B. H. (1945). Molecular theory of the scattering of light in fluids. *The Journal of Chemical Physics*, *13*(4), 141-145.

306. Wyatt, P. J. (1993). Light scattering and the absolute characterization of macromolecules. *Analytica chimica acta*, *272*(1), 1-40.

307. Debye, P. (1947). Molecular-weight determination by light scattering. *The Journal of Physical Chemistry*, *51*(1), 18-32.

308. Berne, B. J., & Pecora, R. (2000). *Dynamic light scattering: with applications to chemistry, biology, and physics*. Courier Corporation.

309. Pecora, R. (Ed.). (2013). *Dynamic light scattering: applications of photon correlation spectroscopy*. Springer Science & Business Media.

310. Carpenter, D. K. (1977). Dynamic light scattering with applications to chemistry, biology, and physics (Berne, Bruce J.; Pecora, Robert).

311. Wittgren, B., & Wahlund, K. G. (1997). Fast molecular mass and size characterization of polysaccharides using asymmetrical flow field-flow fractionation-multiangle light scattering. *Journal of Chromatography A*, *760*(2), 205-218.

312. Saio, T., Guan, X., Rossi, P., Economou, A., & Kalodimos, C. G. (2014). Structural basis for protein antiaggregation activity of the trigger factor chaperone. *Science*, *344*(6184), 1250494.

313. Simon, A. C., Zhou, J. C., Perera, R. L., van Deursen, F., Evrin, C., Ivanova, M. E., ... & Labib, K. (2014). A Ctf4 trimer couples the CMG helicase to DNA polymerase α in the eukaryotic replisome. *Nature*, *510*(7504), 293.

314. Brunner, J. D., Lim, N. K., Schenck, S., Duerst, A., & Dutzler, R. (2014). X-ray structure of a calcium-activated TMEM16 lipid scramblase. *Nature*, *516*(7530), 207.

315. Peisley, A., Wu, B., Xu, H., Chen, Z. J., & Hur, S. (2014). Structural basis for ubiquitin-mediated antiviral signal activation by RIG-I. *Nature*, *509*(7498), 110.

316. Shi, H., Singh, N., Esselborn, F., & Blobel, G. (2014). Structure of a myosin adaptor complex and pairing by cargo. *Proceedings of the National Academy of Sciences*, *111*(12), E1082-E1090.

317. Li, Y. I., Ogunnaike, B. A., & Roberts, C. J. (2010). Multi-variate approach to global protein aggregation behavior and kinetics: Effects of pH, NaCl, and temperature for α-chymotrypsinogen A. *Journal of pharmaceutical sciences*, *99*(2), 645-662.

318. Brummitt, R. K., Nesta, D. P., Chang, L., Kroetsch, A. M., & Roberts, C. J. (2011). Nonnative aggregation of an IgG1 antibody in acidic conditions, part 2: nucleation and growth kinetics with competing growth mechanisms. *Journal of pharmaceutical sciences*, *100*(6), 2104-2119.

319. Sahin, E., Grillo, A. O., Perkins, M. D., & Roberts, C. J. (2010). Comparative effects of pH and ionic strength on protein–protein interactions, unfolding, and aggregation for IgG1 antibodies. *Journal of pharmaceutical sciences*, *99*(12), 4830-4848.

320. Wen, J., Arakawa, T., & Philo, J. S. (1996). Size-exclusion chromatography with on-line light-scattering, absorbance, and refractive index detectors for studying proteins and their interactions. *Analytical biochemistry*, *240*(2), 155-166.

321. Amartely, H., Avraham, O., Friedler, A., Livnah, O., & Lebendiker, M. (2018). Coupling multi angle light scattering to ion exchange chromatography (IEX-MALS) for protein characterization. *Scientific reports*, *8*(1), 6907.

322. Gentiluomo, L., Schneider, V., Roessner, D., & Frieß, W. (2019). Coupling Multi-Angle Light Scattering to Reverse-Phase Ultra-High-Pressure Chromatography (RP-UPLC-MALS) for the characterization monoclonal antibodies. *Scientific reports*, *9*(1), 1-8.

323. Svilenov, H., Gentiluomo, L., Friess, W., Roessner, D., & Winter, G. (2018). A New Approach to Study the Physical Stability of Monoclonal Antibody Formulations—Dilution From a Denaturant. *Journal of pharmaceutical sciences*, *107*(12), 3007-3013.

324. Faria, T. Q., Almeida, Z. L., Cruz, P. F., Jesus, C. S., Castanheira, P., & Brito, R. M. (2015). A look into amyloid formation by transthyretin: aggregation pathway and a novel kinetic model. *Physical Chemistry Chemical Physics*, *17*(11), 7255-7263.

325. Henderson, J. N., Hazra, S., Dunkle, A. M., Salvucci, M. E., & Wachter, R. M. (2013). Biophysical characterization of higher plant Rubisco activase. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, *1834*(1), 87-97.

326. Canon, F., Paté, F., Cheynier, V., Sarni-Manchado, P., Giuliani, A., Pérez, J., ... & Cabane, B. (2013). Aggregation of the salivary proline-rich protein IB5 in the presence of the tannin EgCG. *Langmuir*, *29*(6), 1926-1937.

327. Drenski, M. F., Brader, M. L., Alston, R. W., & Reed, W. F. (2013). Monitoring protein aggregation kinetics with simultaneous multiple sample light scattering. *Analytical biochemistry*, *437*(2), 185-197.

328. Owczarz, M., Motta, A. C., Morbidelli, M., & Arosio, P. (2015). A Colloidal Description of Intermolecular Interactions Driving Fibril–Fibril Aggregation of a Model Amphiphilic Peptide. *Langmuir*, *31*(27), 7590-7600.

329. Gladytz, A., Lugovoy, E., Charvat, A., Häupl, T., Siefermann, K. R., & Abel, B. (2015). Intermediates caught in the act: tracing insulin amyloid fibril formation in time by combined optical spectroscopy, light scattering, mass spectrometry and microscopy. *Physical Chemistry Chemical Physics*, *17*(2), 918-927.

330. Nicoud, L., Lattuada, M., Yates, A., & Morbidelli, M. (2015). Impact of aggregate formation on the viscosity of protein solutions. *Soft matter*, *11*(27), 5513-5522.

331. Barnett, G. V., Qi, W., Amin, S., Lewis, E. N., & Roberts, C. J. (2015). Aggregate structure, morphology and the effect of aggregation mechanisms on viscosity at elevated protein concentrations. *Biophysical chemistry*, *207*, 21-29.

332. Wu, H., Truncali, K., Ritchie, J., Kroe-Barrett, R., Singh, S., Robinson, A. S., & Roberts, C. J. (2015, November). Weak protein interactions and pH-and temperature-dependent aggregation of human Fc1. In *MAbs* (Vol. 7, No. 6, pp. 1072-1083). Taylor & Francis.

333. Streets, A. M., Sourigues, Y., Kopito, R. R., Melki, R., & Quake, S. R. (2013). Simultaneous measurement of amyloid fibril formation by dynamic light scattering and fluorescence reveals complex aggregation kinetics. *PloS one*, *8*(1), e54541.

334. Smirnova, E., Safenkova, I., Stein-Margolina, V., Shubin, V., Polshakov, V., & Gurvits, B. (2015). pH-responsive modulation of insulin aggregation and structural transformation of the aggregates. *Biochimie*, *109*, 49-59.

335. Barnett, G. V., Razinkov, V. I., Kerwin, B. A., Laue, T. M., Woodka, A. H., Butler, P. D., ... & Roberts, C. J. (2015). Specific-ion effects on the aggregation mechanisms and protein–protein interactions for anti-streptavidin immunoglobulin gamma-1. *The Journal of Physical Chemistry B*, *119*(18), 5793-5804.

336. Rakel, N., Bauer, K. C., Galm, L., & Hubbuch, J. (2015). From osmotic second virial coefficient (B22) to phase behavior of a monoclonal antibody. *Biotechnology progress*, *31*(2), 438-451.

337. Garidel, P., Blume, A., & Wagner, M. (2015). Prediction of colloidal stability of high concentration protein formulations. *Pharmaceutical development and technology*, *20*(3), 367-374.

338. Rubin, J., Sharma, A., Linden, L., Bommarius, A. S., & Behrens, S. H. (2014). Gauging colloidal and thermal stability in human IgG1–sugar solutions through diffusivity measurements. *The Journal of Physical Chemistry B*, *118*(11), 2803-2809.

339. Rubin, J., Linden, L., Coco, W. M., Bommarius, A. S., & Behrens, S. H. (2013). Salt-induced aggregation of a monoclonal human immunoglobulin G1. *Journal of pharmaceutical sciences*, *102*(2), 377-386.

340. Carvalho, F. A., Alves, F. R., Carvalho, J. W., & Tabak, M. (2015). Guanidine hydrochloride and urea effects upon thermal stability of Glossoscolex paulistus hemoglobin (HbGp). *International journal of biological macromolecules*, *74*, 18-28.

341. Mehta, S. B., Bee, J. S., Randolph, T. W., & Carpenter, J. F. (2014). Partial unfolding of a monoclonal antibody: role of a single domain in driving protein aggregation. *Biochemistry*, *53*(20), 3367-3377.

342. Arosio, P., Rima, S., & Morbidelli, M. (2013). Aggregation mechanism of an IgG2 and two IgG1 monoclonal antibodies at low pH: from oligomers to larger aggregates. *Pharmaceutical research*, *30*(3), 641-654.

343. Mohr, B. G., Dobson, C. M., Garman, S. C., & Muthukumar, M. (2013). Electrostatic origin of in vitro aggregation of human γ-crystallin. *The Journal of chemical physics*, *139*(12), 09B614_1.

344. Nicoud, L., Arosio, P., Sozo, M., Yates, A., Norrant, E., & Morbidelli, M. (2014). Kinetic analysis of the multistep aggregation mechanism of monoclonal antibodies. *The Journal of Physical Chemistry B*, *118*(36), 10595-10606.

345. Zhou, C., Qi, W., Lewis, E. N., & Carpenter, J. F. (2015). Concomitant Raman spectroscopy and dynamic light scattering for characterization of therapeutic proteins at high concentrations. *Analytical biochemistry*, *472*, 7-20.

346. Lewis, E., Qi, W., Kidder, L., Amin, S., Kenyon, S., & Blake, S. (2014). Combined dynamic light scattering and Raman spectroscopy approach for characterizing the aggregation of therapeutic proteins. *Molecules*, *19*(12), 20888-20905.

347. Yu, Z., Reid, J. C., & Yang, Y. P. (2013). Utilizing dynamic light scattering as a process analytical technology for protein folding and aggregation monitoring in vaccine manufacturing. *Journal of pharmaceutical sciences*, *102*(12), 4284-4290.

348. Rocco, M., Molteni, M., Ponassi, M., Giachi, G., Frediani, M., Koutsioubas, A., ... & Ferri, F. (2014). A comprehensive mechanism of fibrin network formation involving early branching and delayed single-to double-strand transition from coupled time-resolved X-ray/light-scattering detection. *Journal of the American Chemical Society*, *136*(14), 5376-5384.

349. Yadav, S., Shire, S. J., & Kalonia, D. S. (2012). Viscosity behavior of high-concentration monoclonal antibody solutions: correlation with interaction parameter and electroviscous effects. *Journal of pharmaceutical sciences*, *101*(3), 998-1011.

350. He, F., Becker, G. W., Litowski, J. R., Narhi, L. O., Brems, D. N., & Razinkov, V. I. (2010). High-throughput dynamic light scattering method for measuring viscosity of concentrated protein solutions. *Analytical biochemistry*, *399*(1), 141-143.

351. Minton, A. P. (2016). Recent applications of light scattering measurement in the biological and biopharmaceutical sciences. *Analytical biochemistry*, *501*, 4.

352. Sorret, L. L., DeWinter, M. A., Schwartz, D. K., & Randolph, T. W. (2016). Challenges in predicting protein-protein interactions from measurements of molecular diffusivity. *Biophysical journal*, *111*(9), 1831-1842.

353. Provencher, S. W. (1979). Inverse problems in polymer characterization: direct analysis of polydispersity with photon correlation spectroscopy. *Die Makromolekulare Chemie: Macromolecular Chemistry and Physics*, *180*(1), 201-209.

354. Nijman, E. J., Merkus, H. G., Marijnissen, J. C., & Scarlett, B. (2001). Simulations and experiments on number fluctuations in photon-correlation spectroscopy at low particle concentrations. *Applied optics*, *40*(24), 4058-4063.

355. Meyer, W. V., Cannell, D. S., Tin, P., Cheung, H. M., Mann Jr, J. A., Taylor, T. W., ... & Smart, A. E. (1999). *U.S. Patent No. 5,956,139*. Washington, DC: U.S. Patent and Trademark Office.

356. Saito, S., Hasegawa, J., Kobayashi, N., Tomitsuka, T., Uchiyama, S., & Fukui, K. (2013). Effects of ionic strength and sugars on the aggregation propensity of monoclonal antibodies: influence of colloidal and conformational stabilities. *Pharmaceutical research*, *30*(5), 1263-1280.

357. Ma, Y., Acosta, D. M., Whitney, J. R., Podgornik, R., Steinmetz, N. F., French, R. H., & Parsegian, V. A. (2015). Determination of the second virial coefficient of bovine serum albumin under varying pH and ionic strength by composition-gradient multi-angle static light scattering. *Journal of biological physics*, *41*(1), 85-97.

358. Rakel, N., Galm, L., Bauer, K. C., & Hubbuch, J. (2015). Influence of macromolecular precipitants on phase behavior of monoclonal antibodies. *Biotechnology progress*, *31*(1), 145-153.

359. Herhut, M., Brandenbusch, C., & Sadowski, G. (2016). Inclusion of mPRISM potential for polymer-induced protein interactions enables modeling of second osmotic virial coefficients in aqueous polymer-salt solutions. *Biotechnology journal*, *11*(1), 146-154.

360. Arzensek, D., Kuzman, D., & Podgornik, R. (2015). Hofmeister effects in monoclonal antibody solution interactions. *The Journal of Physical Chemistry B*, *119*(33), 10375-10389.

361. Dieterle, M., Blaschke, T., & Hasse, H. (2013). Second osmotic virial coefficients and aggregation of monoclonal antibodies by static laser light scattering. *Zeitschrift für Physikalische Chemie*, *227*(2-3), 333-344.

362. Raut, A. S., & Kalonia, D. S. (2015). Opalescence in monoclonal antibody solutions and its correlation with intermolecular interactions in dilute and concentrated solutions. *Journal of pharmaceutical sciences*, *104*(4), 1263-1274.

363. Li, W., Persson, B. A., Morin, M., Behrens, M. A., Lund, M., & Zackrisson Oskolkova, M. (2015). Charge-induced patchy attractions between proteins. *The Journal of Physical Chemistry B*, *119*(2), 503-508.

364. Soraruf, D., Roosen-Runge, F., Grimaldo, M., Zanini, F., Schweins, R., Seydel, T., ... & Schreiber, F. (2014). Protein cluster formation in aqueous solution in the presence of multivalent metal ions–a light scattering study. *Soft Matter*, *10*(6), 894-902.

365. Roberts, D., Keeling, R., Tracka, M., Van Der Walle, C. F., Uddin, S., Warwicker, J., & Curtis, R. (2014). Specific ion and buffer effects on protein–protein interactions of a monoclonal antibody. *Molecular pharmaceutics*, *12*(1), 179-193.

366. Blanco, M. A., Perevozchikova, T., Martorana, V., Manno, M., & Roberts, C. J. (2014). Protein–protein interactions in dilute to concentrated solutions: α-chymotrypsinogen in acidic conditions. *The journal of physical chemistry B*, *118*(22), 5817-5831.

367. Härtl, E., Dixit, N., Besheer, A., Kalonia, D., & Winter, G. (2013). Weak antibody–cyclodextrin interactions determined by quartz crystal microbalance and dynamic/static light scattering. *European Journal of Pharmaceutics and Biopharmaceutics*, *85*(3), 781-789.

368. Neergaard, M. S., Kalonia, D. S., Parshad, H., Nielsen, A. D., Møller, E. H., & van de Weert, M. (2013). Viscosity of high concentration protein formulations of monoclonal antibodies of the IgG1 and IgG4 subclass–Prediction of viscosity through protein–protein interaction measurements. *European Journal of Pharmaceutical Sciences*, *49*(3), 400-410.

369. Koepf, E., Schroeder, R., Brezesinski, G., & Friess, W. (2018). The missing piece in the puzzle: Prediction of aggregation via the protein-protein interaction parameter A∗ 2. *European Journal of Pharmaceutics and Biopharmaceutics*, *128*, 200-209.

370. Some, D. (2013). Light-scattering-based analysis of biomolecular interactions. *Biophysical reviews*, *5*(2), 147-158.

371. Attri, A. K., & Minton, A. P. (2005). New methods for measuring macromolecular interactions in solution via static light scattering: basic methodology and application to nonassociating and self-associating proteins. *Analytical biochemistry*, *337*(1), 103-110.

372. Halling, D. B., Kenrick, S. A., Riggs, A. F., & Aldrich, R. W. (2014). Calcium-dependent stoichiometries of the KCa2. 2 (SK) intracellular domain/calmodulin complex in solution. *The Journal of general physiology*, *143*(2), 231-252.

373. Mitchell, S. L., Ismail, A. M., Kenrick, S. A., & Camilli, A. (2015). The VieB auxiliary protein negatively regulates the VieSA signal transduction system in Vibrio cholerae. *BMC microbiology*, *15*(1), 59.

374. Zhao, M., Wu, S., Zhou, Q., Vivona, S., Cipriano, D. J., Cheng, Y., & Brunger, A. T. (2015). Mechanistic insights into the recycling machine of the SNARE complex. *Nature*, *518*(7537), 61.

375. Ahijado-Guzmán, R., Alfonso, C., Reija, B., Salvarelli, E., Mingorance, J., Zorrilla, S., ... & Rivas, G. (2013). Control by potassium of the size distribution of Escherichia coli FtsZ polymers is independent of GTPase activity. *Journal of Biological Chemistry*, *288*(38), 27358-27365.

376. Some, D., Hanlon, A., & Sockolov, K. (2008). Characterizing protein-protein interactions via static light scattering: reversible heteroassociation. *American Biotechnology Laboratory®*, *26*(4), 18.

AIM AND OUTLINE OF THE THESIS

The main aim of this thesis was the development of a toolkit for the prediction and characterization of protein aggregation in early stages of drug product development. In small molecule drug discovery some developability rules are widely accepted (e.g. the Lipinski rule of five). In contrast, guiding principles for selecting proteins with stability properties which qualify the molecule for drug product development are not yet well established. In order to fill this gap an extensive comprehensive experimental and computational screening of therapeutic proteins in pharmaceutically relevant formulation was conducted. Part of the results of such screenings is covered in **Chapter II**. In this chapter, we share a study on a diverse group of proteins, including their primary sequences, purity data, and computational and biophysical characterization at different pH and ionic strengths. This data is the foundation of a tailor-made database to support biotherapeutic discovery and development in the future.

Due to the large amount of data produced in the screenings we aimed to explore the datasets with different data mining approaches. In **Chapter III** we developed artificial neural networks (ANNs) to compute, ab-priori, relevant biophysical parameters. This approach would address a currently lack of accurate computational methods to predict protein stability as a function of the formulation, which in turn would allow the screening of thousands of molecules even before expression. In **Chapter III**, we further address the issue of ANNs interpretability by a knowledge transfer process, which was to be developed to shine light on the decision making process of this “black-box” by means of surrogate “white-box” models.

Thanks to the screening and data mining of **Chapter II** and **III** a mAb, named PPI-01, showing unusual aggregation behavior was individuated. Therefore, in **Chapter IV** we addressed the reversible native self-association of this molecule. We investigated PPI-01 in multiple conditions by several orthogonal methods to rationalize the self-aggregation behavior. The locus driving self-association of the mAb was investigated by additional characterization of mAb digests. The case-study of **Chapter IV** provides insights on the analytical challenges to characterize the reversible self-association of mAbs. More importantly, **Chapter IV** delves deep into the nature of native reversible self-association which has been often associated with phase separation, precipitation and high viscosity.

An important aspect of initial developability assessments of therapeutic proteins is the evaluation of data obtained by stressing the candidates under several stressing conditions, i.e. accelerated stability studies, which were extensively investigated in **Chapter II**. However, only real-time stability testing permits the selection of the final formulation and protein candidates, and the establishment of the recommended storage conditions and shelf life. In order to address the behavior of the protein investigated in **Chapters II**, **III** and **IV** under real storage conditions, long term stability studies were executed. Then, in **Chapter V** we applied ANNs algorithms to predict the monomer retention upon real-time storage. The aim was to

provide a way to connect accelerated stability studies and other biophysical parameters with the real-time aggregation of protein drugs in several pharmaceutically relevant conditions. Further, similarly to **Chapter III**, human friendly surrogate machine learning models were developed.

Throughout the thesis the physical stability of the investigated molecules were extensively investigated. However, protein´s chemical stability is also of pivotal importance in the development of stable products. This is especially true when proteins are investigated after long term storage. Therefore, in **Chapter VI** we aimed to develop the coupling of ultra high pressure reverse phase chromatography to multi-angle light scattering (RP-UPLC-MALS) as a new tool to investigate the chemical and physical stability of mAbs. The different principle of separation used in RP-UPLC-MALS provides an additional critical level of protein characterization.

CHAPTER II: Advancing therapeutic protein discovery and development through comprehensive computational and biophysical characterization

Lorenzo Gentiluomo[1,2,10], Hristo Svilenov[2], Dillen Augustijn[3], Inas El Bialy[2], Maria L. Greco[4], Alina Kulakova[5], Sowmya Indrakumar[5] ,Sujata Mahapatra[6], Marcello Morales[4], Christin Pohl[6], Aisling Roche[7], Andreas Tosstorff [2], Robin Curtis[7], Jeremy P. Derrick,[8] Allan Nørgaard[6], Tarik A. Khan[9], Günther H.J. Peters[5], Alain Pluen,[7] Åsmund Rinnan[3], Werner Streicher[6], Christopher F. van der Walle[4], Shahid Uddin[4,11], Gerhard Winter[2], Dierk Roessner[1], Pernille Harris[5],* Wolfgang Frieß[2]

[1] Wyatt Technology Europe GmbH, Hochstrasse 18, 56307 Dernbach, Germany

[2] Department of Pharmacy: Pharmaceutical Technology and Biopharmaceutics; Ludwig-Maximilians-Universitaet Muenchen, Butenandtstrasse 5, 81377 Munich, Germany

[3] Copenhagen University, Department of Food Science, Rolighedsvej 26, 1958 Frederiksberg, Denmark

[4] Dosage Form Design and Development, AstraZeneca, Sir Aaron Klug Building, Granta Park, Cambridge CB21 6GH, UK

[5] Technical University of Denmark, Department of Chemistry, Kemitorvet 207, 2800 Kongens Lyngby, Denmark

[6] Novozymes A/S, Krogshoejvej 36, 2880, Bagsvaerd, Denmark

[7] School of Chemical Engineering and Analytical Science, Manchester Institute of Biotechnology, The University of Manchester, 131 Princess Street, Manchester, M1 7DN, UK

[8] School of Biological Sciences, Faculty of Biology, Medicine and Health, Manchester Academic Health Science Centre, The University of Manchester, Oxford Road, Manchester M13 9PT U.K.

[9] Pharmaceutical Development & Supplies, Pharma Technical Development Biologics Europe, F. Hoffmann-La Roche Ltd., Grenzacherstrasse 124, 4070, Basel, Switzerland

[10] Present address: Coriolis Pharma Research GmbH, Fraunhoferstraße 18B, 82152, Planegg, Germany

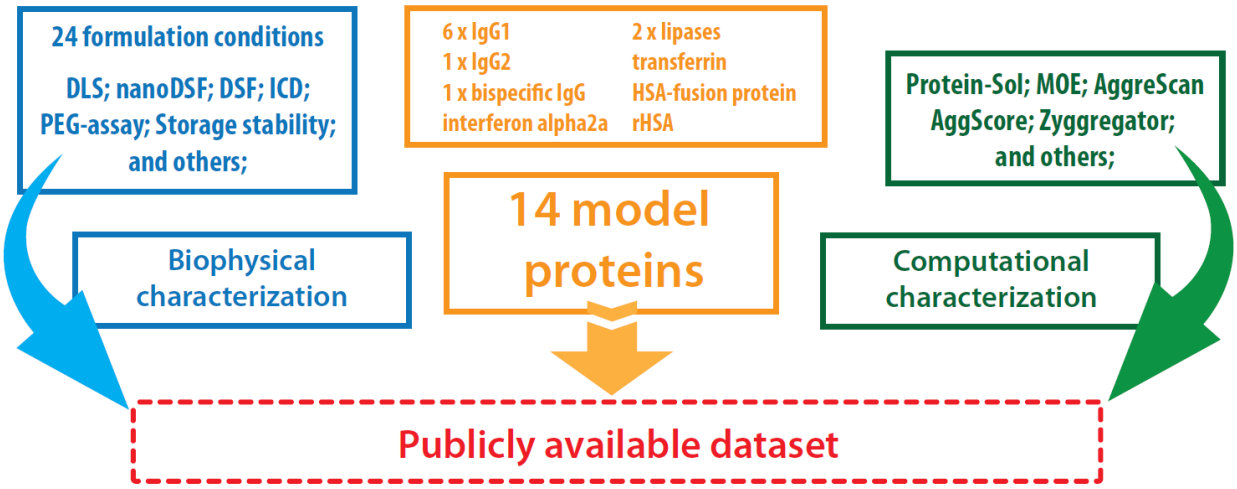
[11] Present address:Immunocore Limited, 101 Park Dr, Milton, Abingdon, OX14 4RY, UK

** denotes corresponding author.*

CONTRIBUTION: **Lorenzo Gentiluomo**, Hristo Svilenov, Gerhard Winter, Wolfgang Frieß wrote the paper. Robin Curtis, Jeremy P. Derrick, Allan Nørgaard, Günther H.J. Peters, Alain Pluen, Åsmund Rinnan, Werner Streicher, Christopher van der Walle, Shahid Uddin, Gerhard Winter, Dierk Roessner, Pernille Harris, Wolfgang Frieß planned, designed and supervised the study. **Lorenzo Gentiluomo** performed data mining. **Lorenzo Gentiluomo** and Dillen Augustijn evaluated produced models. **Lorenzo Gentiluomo**, calculated response surface, pairwise correlations and the stability risk values**. Lorenzo Gentiluomo** performed and analyzed accelerated stress stability studies, DLS, k_D, T_{agg} and SEC-MALS on the 100% of the protein library. Hristo Svilenov, performed and analyzed nanoDSF on 20% of the protein library. Inas El Bialy, performed and analyzed DSF on the 100% of protein library. Maria Laura Greco, performed PEG-assay on 20% of protein library. Alina Kulakova and Sujata Mahapatra, performed and analyzed nanoDSF and ICD on 90% of protein library. Sujata Mahapatra and Alina Kulakova performed the purification of 15% of the protein library. Marcello Morales, performed and analyzed PEG-assay on 80% of protein library. Christin Pohl, performed and analyzed nanoDSF and ICD 10% of the protein library. Christin Pohl and Sujata Mahapatra performed DLS and Tagg on 15% of the protein library. Aisling Roche, performed z-potential on 80% of protein library. Sowmya Indrakumar and Andreas Tosstorff, performed homology modeling. Tarik A. Khan, Sowmya Indrakumar, Andreas Tosstorff and **Lorenzo Gentiluomo** calculated molecular descriptors. All authors corrected and approved the final manuscript.

Abstract

Therapeutic protein candidates should exhibit favorable properties that render them suitable to become drugs. Nevertheless, there are no well-established guidelines for the efficient selection of proteinaceous molecules with desired features during early-stage development. Such guidelines can emerge only from a large body of published re-search that employs orthogonal techniques to characterize therapeutic proteins in different formulations. In this work, we share a study on a diverse group of proteins, including their primary sequences, purity data, and computational and biophysical characterization at different pH and ionic strength. We report weak linear correlations between many of the biophysical parameters. We suggest that a stability comparison of diverse therapeutic protein candidates should be based on a computational and biophysical characterization in multiple formulation conditions, as the latter can largely determine whether a protein is above or below a certain stability threshold. We use the presented dataset to calculate several stability risk scores obtained with an increasing level of analytical effort and show how they correlate with protein aggregation during storage. Our work highlights the importance of developing combined risk scores that can be used for early-stage developability assessment. We suggest that such scores can have high prediction accuracy only when they are based on protein stability characterization in different solution conditions.



Graphical abstract. *Advancing therapeutic protein discovery and development through comprehensive computational and biophysical characterization*

1 Introduction

Protein therapeutics are an essential part of the treatment plan for many patients suffering from severe diseases.[1] Proteins can bind to various drug targets with high specificity and affinity, thus improving both therapeutic efficacy and safety profiles compared to small molecule drugs. Alongside these benefits, therapeutic proteins also bring drawbacks like high costs and complexity of their discovery, development and production.[2]

Although there are different ways to develop a therapeutic protein, some of the most widely-used approaches share common steps,[3,4] i.e. identification of a druggable target; generation of a library of proteins which could bind to that target; selection and optimization of lead candidates; formulation development; and decision on one biologically active, safe and stable protein which will continue to clinical trials. In general, the failure of a drug candidate becomes increasingly expensive as the development process advances. For this reason, pharmaceutical companies aim to adopt strategies for selecting the most promising molecules at early stages.[3,5–7] Such strategies have to identify whether a molecule exhibits suitable biological and biophysical properties, i.e. drug-like properties.[8–10] Contrary to small molecule drug discovery where some developability rules are widely accepted (e.g. the Lipinski rule of five),[11] guiding principles for selecting proteins with drug-like properties are not yet well established. This gap has stimulated researchers to create rules for protein developability assessment based on computational and biophysical characterization.[7,10,12–15] Although significant progress has been made on this topic, published work is currently limited to monoclonal antibodies (mAbs) and disregards the impact of formulation conditions on the thresholds of parameters that will flag a molecule as developable or not.

Biological activity and low toxicity are essential pre-requisites for molecules to be selected for further studies. However, protein drug candidates should also exhibit desirable biophysical properties that ensure sufficient stability during manufacturing, shipping, storage, handling and administration.[7] The proper assessment of these biophysical properties requires the application of multiple orthogonal methods. Historically, most widely used methods for protein characterization required large sample amounts and suffered from low throughput, e.g. differential scanning calorimetry or circular dichroism. Since both sample amount and time are scarce during early-stage development, different candidates were usually compared in only one solution condition with a limited set of biophysical techniques that were considered to be predictive for the overall protein stability.[15,16] With the recent rapid advance in technology, new methods have emerged that require miniature sample amounts and can measure hundreds of samples per day. However, the most efficient way of how these methods can be combined and the value of the measured parameters for selecting promising candidate molecules are still a matter of debate.[12,15,17] Furthermore, most of the published data addressing the biophysical parameters and their predictive power for protein stability is based on biomolecules with un-published primary sequences and purity

data.[18,19] That makes the data reproducibility, interpretation and use for follow-up studies difficult for the scientific and industrial communities.

Here, we present a study on 14 diverse therapeutically relevant proteins, including most of the primary sequences and purity data. We show work based on computational analysis, as well as biophysical characterization and storage stability data of these proteins in 24 conditions with different pH and ionic strength. The full dataset will be available for download from a tailor-made database (https://pippi-data.kemi.dtu.dk/). We use the data to look for linear correlations between different biophysical parameters and elucidate whether some of the latter carry similar information that can be used for developability assessment of proteins. Next, we emphasize that protein stability largely depends on solution conditions. Therefore, a proper assessment of whether a drug candidate molecule is stable, and suitable for further development, requires characterization in several formulations at the earliest development stages. Such characterization is nowadays possible due to the large increase in the throughput of many biophysical assays. Finally, we discuss an early risk assessment approach based on stability risk score (SRS) values obtained from datasets of increasing size and show how these scores correlate with the amount of aggregates formed after 6-month storage at different temperatures.

2 Material and methods

2.1 Sample preparation

Six IgG1 antibodies (PPI-01, PPI-02, PPI-03, PPI-04, PPI-10, PPI-13), one IgG2 (PPI-17), one bispecific mAb (PPI-08), and one HSA-neprilysin fusion protein (PPI-18) were provided by AstraZeneca (Cambridge, UK). Interferon alpha-2a (PPI-30) was provided from Roche Diagnostics GmbH. Recombinant human transferrin[20] (PPI-44) and Recombumin® native sequence human serum albumin (PPI-49) were from Albumedix Ltd. Two lipases (PPI-45 and PPI-46) were from Novozymes A/S. Primary protein sequences can be found in Supplementary information – SI 1. The proteins in the bulk solutions were dialyzed overnight using Slide-A-Lyzer™ cassettes (Thermo Fisher Scientific, USA) with suitable membrane cut-off against excess of 10 mM of histidine/histidine hydrochloride buffer with pH 5.0, 5.5, 6.0, 6.5, 7.0, 7.5 or against 10 mM of tris(hydroxymethyl)aminomethane/tris(hydroxymethyl)aminomethane hydrochloride buffer with pH 8.0 and 9.0. Protein concentration was measured on a Nanodrop 2000 (Thermo Fisher Scientific, USA) using the respective protein extinction coefficient at 280 nm. In total, the proteins were formulated at eight different pH values mentioned above with the presence of 0, 70, or 140 mM sodium chloride accounting for 24 different formulation conditions for each of the 14 proteins. NaCl was added to the samples from a stock solution in the respective buffer. All the materials were of analytical or multi-compendial grade from J. T. Baker. After preparation, the formulations were sterile filtered with 0.22 µm cellulose acetate filters from VWR International (Germany).

2.2 In silico modeling of monoclonal antibodies and estimation of molecular descriptors

The template for modelling is identified using a BLAST search (www.ncbi.nlm.nih.gov/blast/)[21] against the PDB database.[22] In the case of low query coverage, multiple template sequences were considered. The atomic coordinate set corresponding to the crystal structure of the homologue (template) was obtained from the PDB database. Sequence alignment of the template and query sequence was generated using ClustalW2 (www.ebi.ac.uk/Tools/msa/clustalw2/)[23] alignment tool. The modeling of the three-dimensional structure was performed by Modeller9.19 software.[24] The quality of the models produced depended on the quality of the sequence alignment by ClustalW2 and template structure. In the case of antibodies (PPI-01, PPI-02, PPI-03, PPI-10, PPI-13, PPI-17), the generated Fab and Fc structural units were subsequently aligned to the full mAb structure with PDB ID 1IGT5[25] using PyMol6. The aligned Fab and Fc parts were then merged using Modeller. Disulphide bonds were added accordingly. No primary sequence was available for PPI-04 and PPI-08 and therefore these were not modelled. For PPI-18, a model accounting for the orientation of the two fragments was generated based on SAXS data (not shown). The fragments were connected using Modeller and disulphide bonds were added where suitable. The homology models for PPI-30, PPI-44, PPI-45, PPI-46, and PP-49 were generated using as templates PDB entries 4Z5R[26], 3V83[27], 5TGL[28], 1GT6[29], and 4BKE[30] respectively, and using the tools mentioned above for sequence alignment and model generation. The modelled protein structures were prepared in the desired pH environment to account for the right protonation states of residues using Glide. Further, prepared structures were energy minimized prior to structure validation to make sure the target protein has the least energy conformation without any steric clashes. The protein structure was energy minimized using the Glide software. The quality of the modelled structures was checked by examining the extent of occurrence of conformations in disallowed regions of the Ramachandran plot using Maestro[31]. In addition, Z-score values were calculated using the standalone version of Prosa2003[32]. The generated models have an overall negative Z-score indicating a good quality of built structures (Supplementary information - SI 2).

The protein homology models and primary sequences were used for computational protein characterization. The recently developed Protein-Sol server[33] was used to study the behaviour of the model proteins as a function of pH and ionic strength. Further, the molecular operating environment (MOE) software was used to calculate various molecular descriptors. Topographic, thermodynamics and structural indices were calculated from ProtDCal.[34] Aggregation scores of the proteins were calculated with the Schrödinger's Surface Analyzer command-line tool (Schrödinger Inc., USA) using previously generated homology models. The tool generates scores based on three different algorithms: AggScore, Zyggregator and Aggrescan.[35–37] Mean scores per residue were calculated for each method and protein.

2.3 Dynamic light scattering (DLS)

DynaPro® II plate reader (Wyatt Technology, USA) was used for the dynamic light scattering experiments. The measurements were performed in 1536 LoBase Assay Plates (Aurora Microplates, USA) in triplicates using 4 µL of sample sealed with a few µL of silicone oil. The plate was centrifuged for 1 min at 2000 rpm before placed in the plate reader. Data was collected and processed with the DYNAMICS® software V7.8 (Wyatt Technology, USA). The coefficient of self-diffusion, *D*, and the polydispersity index (PDI) were calculated from the obtained autocorrelation functions using cumulant analysis. The Stokes-Einstein equation was used to calculate the hydrodynamic radius (R_h) from *D*. The increase in R_h after storage at different temperatures was calculated with the following equation:

$$R_{R,X} = \frac{R_{h,X}}{R_{h,0}}$$

where $R_{h,0}$ is the hydrodynamic radius before stress and $R_{R,X}$ is the one after stress. The aggregation onset temperature (T_{agg}) was determined using protein at a concentration of 1 mg/mL. A temperature ramp of 0.1 °C/min was applied from 25 °C to 80 °C. One measurement included 3 acquisitions of 3 s. T_{agg} was calculated by the DYNAMICS® software V7.8 from the increase in R_h during heating. The interaction parameter (k_D) was determined at 25 °C from the slope of the protein concentration dependence of *D* studied with at least six dilutions between 1 and 10 mg/mL for each formulation. Every measurement was performed with 10 acquisitions of 5 s.

2.4 High throughput fluorimetric analysis of thermal protein unfolding with nanoDSF®

Samples containing 1 mg/mL protein in the respective formulations were filled in standard nanoDSF capillaries (NanoTemper Technologies, Germany). Measurements were performed using the Prometheus NT.48 (NanoTemper Technologies, Germany) system that measures the intrinsic protein fluorescence intensity at 330 and 350 nm after excitation at 280 nm (±10 nm). A temperature ramp of 1 °C/min was applied from 20 to 95 °C. The fluorescence intensity ratio (F350/F330) was plotted against the temperature, the onset and inflection points of the unfolding transitions were determined from the first derivative of each measurement using the PR.Control software V1.12 (NanoTemper Technologies, Germany). The onset temperature of the first unfolding was reported as $T_{on,\ int}$. The inflection points of the unfolding transitions were reported as $T_{m1,int}$ and $T_{m2,int}$ for the unfolding at lower and higher temperature respectively. For proteins with one thermal unfolding, only $T_{on,\ int}$ and $T_{m1,int}$ were reported.

2.5 Differential scanning fluorimetry (DSF)

The DSF measurements were performed using Sypro® Orange as an extrinsically fluorescent dye using a previously published procedure.[38] Briefly, 1 µl of the freshly prepared working solution (1:5000 of stock solution in highly purified water) of Sypro® Orange was added and mixed with 20 µl sample in MicroAmp optical 96-well reaction plate (Applied Biosystems; USA) in triplicates. The samples consisted of 1 mg/ml protein in the respective formulation. A protein-free placebo was also included for each condition and later used for background subtraction. A temperature ramp was applied from 20 to 96 °C at a rate of 1 °C/min using the qTower 2.2 RT-PCR (Jena Analytik AC; Germany). The $T_{on,ext}$ and $T_{m1,ext}$ were calculated from the fluorescence intensity data at 578 nm as described in Supplementary information SI 3.

2.6 Isothermal chemical denaturation (ICD)

All ICD studies were performed on Unchained Labs HUNK system (Unchained Labs, USA).[39] Guanidine hydrochloride (GuHCl) and urea were used as denaturants. 6 M GuHCl stock solutions were prepared in each formulation condition and mixed in different ratios with the formulation buffer by the instrument. Protein stock solutions were prepared at 1 mg/ml and diluted 12.5 times by addition to different denaturant concentrations. In total, 48-points linear denaturant gradient was automatically generated for each condition. The incubation time varied depending on the protein studied. The samples were measured using an excitation wavelength of 285 nm and emission intensities were recorded from 300 nm to 450 nm. The data analysis was performed using the software Formulator V3.02 (Unchained Labs, USA). For the native protein, the fluorescence emission maximum $\lambda_{max(native)}$ was selected from the spectrum of the sample containing no denaturant. For the samples in denaturants, the fluorescence emission maximum $\lambda_{max(den)}$ was determined in a similar way. The ratio $\lambda_{max(den)}/\lambda_{max(native)}$ was plotted against denaturant concentration to obtain the chemical denaturation curves. Apparent free energy of unfolding (ΔG), C_m and m-values were calculated for the different transitions.[40,41] Different unfolding models (e.g. two-state, three-state) were tested for each protein to find the best fit. For proteins exhibiting a three-state unfolding, C_{m1}, m_1 and dG_1 were reported for the unfolding at lower denaturant concentration, while C_{m2}, m_2 and dG_2 were reported for the unfolding at higher denaturant concentration. In cases of two-state unfolding, only C_{m1}, m_1 and dG_1 were derived.

2.7 PEG-assay

PEG 8000 was purchased from Alfa Aesar (USA). To save material, 15 different conditions were selected for the PEG-assay solubility screen including pH 5.0, 6.0, 7.0, 8.0 and 9.0 with 0, 70 and 140 mM NaCl. Proteins were buffer exchanged, formulated and their concentrations measured as described earlier.

40 % (w/v) PEG stock solutions were prepared in both the acidic and basic buffer components (with either 0, 70 or 140 mM NaCl) and titrated to achieve the desired pH as dissolving PEG directly into the buffer resulted in a shift in pH. Final sample preparation to 1 mg/mL protein concentration and increasing amounts of PEG (0-16 % (w/v)), as well as loading into a clear flat-bottom 96 well plate, was performed using a liquid handling system (Freedom-EVO 150, Tecan, Germany). Turbidity was measured using a NEPHELOstar Plus plate reader (BMG Labtech, Germany) after an incubation time of 48 hours. Non-linear regression analysis using a 4-parameter fit equation was performed for the transition region using GraphPad Prism version 7.1 (GraphPad Software, USA) to obtain the point of inflection, defined as PEG-assay turbidity midpoint (PEG_{TMP}).

2.8 Electrophoretic mobility and zeta potential

Electrophoretic mobility measurements were performed by the Zetasizer Nano ZSP (Malvern, UK). In order to extract the most reliable results from this method, which can be buffer ion-specific and of low quality at high ionic strength,[42,43] the screening conditions were changed and the effect of pH alone on the zeta potential was investigated. All measurements were performed in triplicate in a 1 mL DTS1070 folded capillary cell (Malvern, UK) at 25 °C. Proteins were measured in 25 mM NaCl solution with no buffer components added, and pH adjusted dropwise using 0.01 M HCl and 0.1 M NaOH. The relation of the electrophoretic mobility to the zeta potential is described by the Henry Equation:

$$U_E = \frac{2\,\epsilon_0 \varepsilon_m\, \zeta f(\kappa a)}{3\eta}$$

where U_E is the electrophoretic mobility, ϵ_0 is the permittivity in a vacuum, ε_m is the dielectric constant of the solvent, ζ is the zeta potential in volts, f(κa) is Henry's function calculated using the Ohshima approximation[44] and the hydrodynamic radius for each protein and η is the viscosity of water at 25 °C.

2.9 Capillary isoelectric focusing (cIEF)

Maurice system suitability kit, Maurice pI markers, Maurice cIEF 500 mM arginine, Maurice cIEF separation cartridges, 0.5 % methyl cellulose solution and 1 % methyl cellulose solution, were purchased from Protein Simple (USA). Pharmalyte pH 3-10 was purchased from GE Healthcare (Germany). Urea was obtained from Sigma-Aldrich (USA). Samples were first diluted to a final concentration of 1 mg/mL in water. Subsequently, samples were mixed with a solution containing a broad-range ampholyte (pH 3-10), methylcellulose 1 %, 500 mM of arginine and appropriate pI markers and pipetted into a 96 well-plate. Urea (final concentration of 4 M) was added to solutions containing PPI-49 to reduce self-association.

cIEF experiments were run on a MaurICE system (Protein Simple, USA). The separation cartridge was loaded with electrolyte solutions (80 mM phosphoric acid in 0.1 % methyl cellulose and 100 mM sodium hydroxide in 0.1 % methyl cellulose). Experiments were conducted with a pre-focusing time of 1 minute at 1500 V, followed by a focusing time of 5 minutes at 3000 V. Data was processed and analyzed using Compass Software for ICE (Protein Simple, USA).

2.10 Size exclusion chromatography coupled to multi-angle light scattering (SEC-MALS)

Size exclusion chromatography combined with multi-angle light scattering (SEC–MALS) was performed using a Vanquish Horizon™ UPLC with a variable wavelength UV detector (Thermo Fischer Scientific, USA). The separation was performed with a Superdex 200 Increase 10/300 GL column (GE Healthcare, USA). The aqueous mobile phase consisted of 38 mM NaH_2PO_4, 12 mM Na_2HPO_4, 150 mM NaCl and 200 ppm NaN_3 at pH 7.4 dissolved in HPLC-grade water. The mobile phase was filtered with Durapore VVPP 0.1 m membrane filters (Millipore Corporation, USA). Prior analysis, the samples were centrifuged. The autosampler was used to inject 25 or 50 µl in duplicates. The elution of the protein was monitored by the UV signal at 280 nm and by a MALS TREOS II detector (Wyatt Technology, USA). In addition, differential refractive index detector Optilab T-rEX (Wyatt Technology, USA) was used for concentration verification. Data collection and processing were performed using the ASTRA® software V7.1 (Wyatt Technology, USA). Three different parameters $m_{25,rec}$, $m_{40,rec}$ and $m_{50,rec}$ were calculated, which represent the monomer mass recovery from the theoretical calculated protein mass in percent after two weeks of stress at 25°C, 40 °C and 50 °C respectively. This value also takes into account the loss of monomer that can occur due to precipitation or due to the SEC method (e.g. adsorption of the protein on the column material). In addition, the mass fraction of the monomer compared to all peaks in the chromatograms is shown in percentage as M_{25}, M_{40} and M_{50} in the Supplementary Table SI 13. Thanks to the MALS detection, it was also possible to assess the relative amount of small population of aggregates usually not visible by normal SEC-UV. The LSA parameter was calculated from the following equation:

$$LSA_X = \frac{\frac{LSA_{X,mon}}{LSA_{X,tot}}}{\frac{UVA_{X,mon}}{UVA_{X,tot}}}$$

where LSA and UVA represent the light scattering and UV peak area after two weeks at the temperature X respectively, the subscript „mon“ indicates the monomer peak area while the subscript „tot“ indicates the sum of all defined peak areas. Due to the different sensitivity of the MALS and UV detector, an LSA_X value lower than one means that a population of aggregates is present. A decrease of LSA_X highlights an

increase of the light scattering signal which indicates an increase in the percentage of high molecular weight species.

2.11 Stress study

Protein samples with concentration of 1 mg/ml in each respective formulation condition were sterile-filtered, and 0.2 mL was filled in 0.5 mL sterile non-coated PP Eppendorf tubes. The samples were incubated at 4 °C, 25 °C, 40 °C and 50 °C for two weeks, and in a separate study at 4 °C and 25 °C for 6 months. After storage, the samples were quenched on ice, stored at 4 °C and measured within two weeks.

2.12 Response surface methodology (RSM)

We adopted a design of experiments (DoE) approach and a robust RSM to establish the dependence of 27 biophysical parameters on pH and NaCl concentration. Using those dependencies, we determined the range of optimal formulation conditions based on the desired values of the different parameters. The method of ordinary least squares was used in the regression models for data fitting. Both full and reduced models, considering the main effects of factors along with two-way interactions, were employed. A curvature response was allowed by assessing the quadratic term, also considering two-way interactions. The reduced model was obtained using a backward stepwise regression. The F-statistic approach was used to perform the effect test, considering a value of 0.05 or less as statistically significant. The fitting results are shown in Supplementary information – SI 4. All the results were calculated using the statistical software JMP® v 14.0 (SAS Institute Inc., USA), and all the analysis details can be found in the software manual.[45]

2.13 Tests for statistical significance of linear correlations

Pearson's correlation coefficient *R* was calculated to determine whether two quantities are linearly correlated and to which extent. The outliers in the dataset were detected and eliminated before calculating the pairwise correlation. Outlier detection was based on the quartiles as a method, where samples outside the outer quartiles ± 1.5 times interquartile distances were removed using MATLAB®. A Student *t*-test was carried out to test the statistical significance of *R*. The *t*-test was performed to investigate whether an *R* between two biophysical parameters will hold in general populations.

The null hypothesis of no correlation was tested using the following formula:[46,47]

$$t - value = |R| \frac{\sqrt{n-2}}{\sqrt{1-R^2}}$$

where *n* is the number of data points used to obtain *R*, and therefore it is dependent upon the biophysical parameters of interests in our study because some biophysical parameters were not measured in all conditions due to experimental hurdles (e.g. precipitation). For a given t-value and n, the value of cumulative distribution function for Student's *t*-distribution is the confidence-level of the *t*-test and was calculated in MATLAB (MathWorks, USA). The selected confidence level for the *t*-test was 95 % (*p*-value <0.05). The same procedure was applied multiple times for different subsets to assess differences in the *R* values due to the different samples. The data points of the whole dataset are also provided in Supplementary information – SI 5.

2.14 Principal component analysis (PCA)

In order to get a quick overview of all the data collected, a PCA was run with unit-variance scaling of the data to let all the parameters influence the model equally (much like calculating the Pearson's correlation). There are several entries in the data table that do not include a number due to reasons mentioned above. It was therefore necessary to calculate the PCA solution taking into account these missing values through imputation.[48] This also takes into account the actual unit-variance scaling of the data. The data analysis was performed in MATLAB (MathWorks, USA) with in-house codes based on well-known algorithms.

3 Results

3.1 Generating a dataset including computational and biophysical parameters of diverse proteins

The dataset investigated in this study consists of 14 diverse model proteins. Each protein has an assigned code made of the "PPI" letters and a number (**Table 1**). Protein primary sequences, except for PPI-04 and PPI-08, are provided in Supplementary information – SI 1. The dataset roughly represents the heterogenic group of therapeutic proteins today – mostly mAbs, a bispecific mAb, a fusion protein, a cytokine, albumin and enzymes. Some key biophysical properties and the purity of the provided proteins were investigated at the start of the study with orthogonal techniques (**Table 1**). The separations obtained with SEC-MALS and cIEF are presented in Supplementary information – SI 6. All proteins show a relative

monomer mass fraction > 98 % with two exceptions: PPI-10 contains 96 % monomer and 4 % dimer, while PPI-44 contains 85 % monomer and 15 % aggregates. The protein molecular mass from SEC-MALS matches the theoretical values closely within an experimental error of ±3 %. Two exceptions are PPI-30 that shows a deviation of about 13 % and PPI-46 with a difference close to 6 %. We hypothesize that these inconsistencies arise from the small protein molecular mass (M_m). Further, the M_m of PPI-30 showed a concentration dependency, which suggests an effect of the second osmotic virial coefficient in the running buffer used for SEC-MALS. Earlier, we reported for PPI-30 that the protein forms weak oligomers around pH 7.5 which also supports the theory for strong attractive protein-protein interaction in similar conditions.[49] In addition, we provided the retention time of the monomer peak, which can provide further insights on whether non-specific interactions occur with the chromatographic column (**Table 1**). The measured isoelectric points of the main peaks correspond well to the theoretical values calculated with Protein-Sol. The main and neighboring peaks detected by cIEF are in most cases within a narrow pH range. In addition, we calculated the predicted scale solubility from the amino acid sequences, using the Protein-Sol server. The general information and parameters presented in **Table 1** are assessed and shown for two reasons: i) they provide a good overview of the protein properties in the dataset; and ii) they can be a good starting point to explain the results from the biophysical characterization that we present below. We then selected a set of computational and biophysical methods that often find application in protein drug development to study the stability of the proteins at different pH and ionic strength. In general, we aimed to use popular techniques which are often used in published work on the characterization of therapeutic proteins. Although this selection might be subjective, it is based on our experience and on the availability of the techniques in the consortium. The type of molecular descriptors calculated with MOE and ProDCal are summarized in Supplementary information SI 7. The parameters from AggScore, Zyggregator and Aggrescan are presented in Supplementary information SI 8. The experimental dataset included information on the stability of the 14 proteins in 24 different solution conditions, including 8 pH values ranging from 5 to 9 and three concentrations of sodium chloride, 0 mM, 70 mM and 140 mM, to vary the ionic strength. In general, most of the experimental measurements were possible with several exceptions due to formulation issues (for example, precipitation of PPI-30 when dialyzed at pH close to 6); insufficient sample amount (for example, to do some of the k_D measurements); or when the method did not allow measurements of all the 24 formulation conditions (e.g. electrophoretic mobility measurements that are performed at specific ionic strength). The full dataset including the mean values of measured biophysical parameters can be found in a separate table attached as Supplementary information SI 13. Most measurements were run in technical triplicates, except, e.g. for the stress studies measured by SEC-MALS and ICD which were run as a single replicate. Selected experiments were also repeated in different laboratories. Comparisons between cross-laboratory experiments showed high consistency, indicating robustness of the standard operating procedures. In the near future, the expanded dataset, including the replicates and most of the raw data, will be available for download via a tailor-made database (https://pippi-data.kemi.dtu.dk/).

Table 1. Calculated and measured properties of the proteins in the presented dataset. *Protein primary sequences are provided in SI1.*

Protein code	Protein type	Protein-Sol		Electrophoretic mobility	cIEF		Theoretical	SEC-MALS		
		Predicted scale solubility	Calculated isoelectric point	Point of zero ζ	Main peak	Peaks range	Calculated monomer M_m (kDa)	Measured monomer M_m (kDa)	Monomer mass fraction (%)	Monomer retention volume (mL)
PPI-01	IgG1	0.366	8.37	6.94	7.2	7.1-7.3	144.8	147.7	99.7	11.8
PPI-02	IgG1	0.354	9.09	8.21	9.3	9.1-9.4	148.2	147.9	98.3	11,9
PPI-03	IgG1	0.404	9.4	8.77	9.4	9.1-9.4	144.8	147.1	99.8	12.0
PPI-04	IgG1	-*	-*	8.31	8.95	8.7-9.0	146.2	150.3	99.1	12.1
PPI-08	IgG1 + scFv	-*	-*	8.90	9.2	8.9-9.4	204.4	206.2	99.7	12.4
PPI-10	IgG1	0.378	9.15	8.87	9.2	8.8-9.3	144.2	147.8	96.3	12.0
PPI-13	IgG1	0.397	9.08	8.26	8.9	8.5-9.0	148.9	150.1	99.4	12.0
PPI-17	IgG2	0.334	8.89	8.21	9.05	8.7-9.3	145.1	148.4	98.5	12.0
PPI-18	HSA-NEP	0.431	5.68	5.01	5.6	4.5-6.0	146.7	149.4	98.3	11.2
PPI-30	IFN-α2a	0.451	6.19	5.96	6.2	6.0-6.5	19.2	22.0	100	16.2
PPI-44	transferrin	0.330	7.06	5.85	5.5	4.9-5.8	74.9	76.1	85.1	13.9
PPI-45	lipase	0.413	4.95	- †	4.7	4.5-4.9	29.5	29.8	100	16.1
PPI-46	lipase	0.391	4.99	- †	4.35	4.1-5.1	29	30.8	100	16.0
PPI-49	rHSA	0.450	6.13	- †	4.9	4.1-5.0	66.4	66.7	98.1	13.6

**No primary sequence available. † The electrophoretic mobility measurements could not accurately define this*

3.2 Linear correlation in the biophysical parameters, and similarities between the proteins

We used the obtained dataset to search for pairwise linear correlations between 27 experimental biophysical parameters that are often assessed during protein discovery and development. The Student *t*-test was applied to determine the statistical significance of the pairwise correlations evaluated by the Pearson`s correlation coefficient *R*. **Figure 1a** presents the *R* values with statistically significant correlations between the biophysical parameters at 95 % confidence level (p-values < 0.05) for all 14 studied proteins.

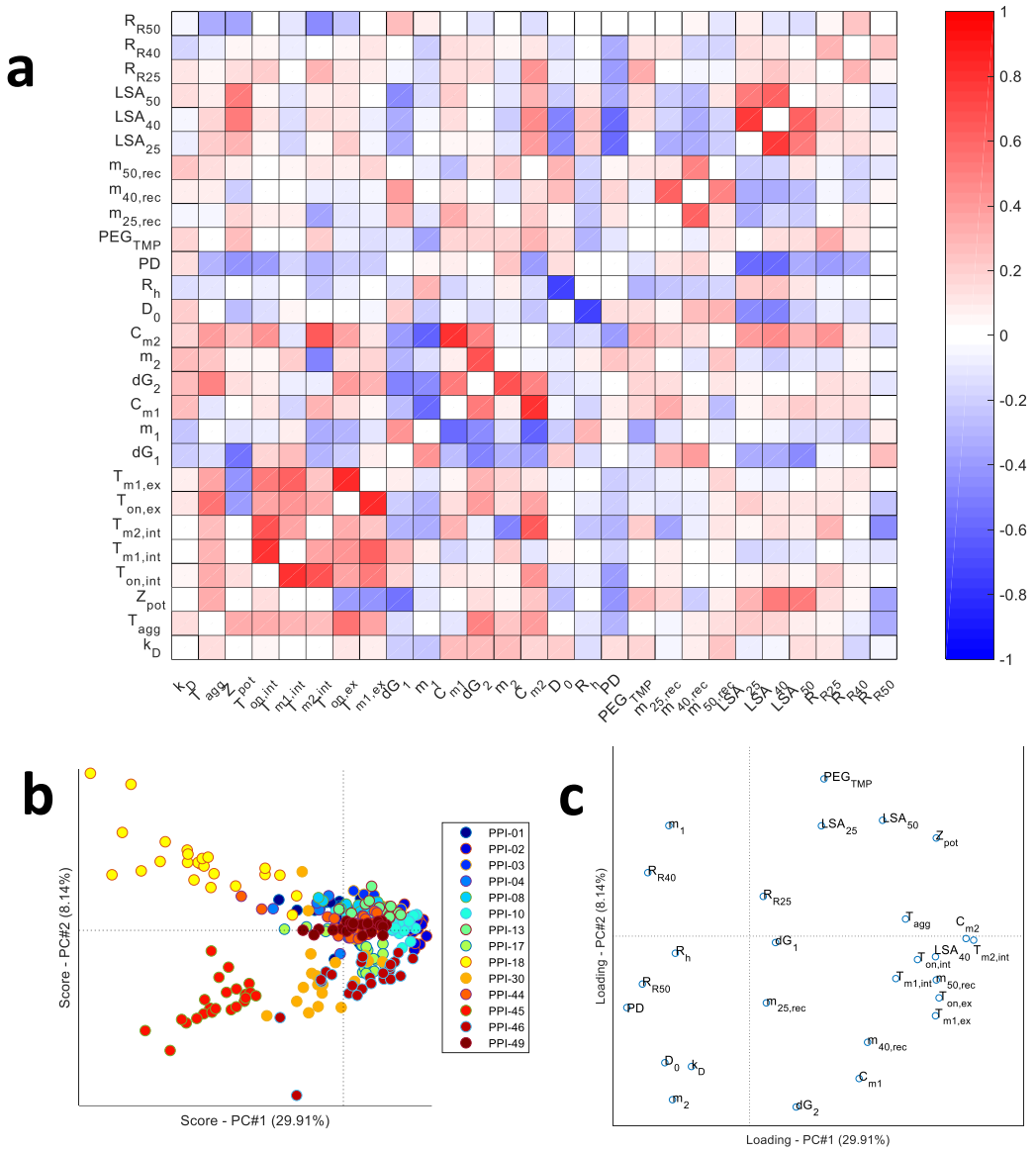


Figure 1. (a) - Pairwise linear correlations between measured biophysical parameters in the entire dataset including 14 proteins and 24 different formulation conditions. *The strength of these correlations was assessed using t-tests. p-values <0.05 are statistically significant at a 95 % confidence level. White cells represent no correlation with a p-value higher than 0.05. Blue and red cells express negative or positive correlation, respectively. No or weak correlations were observed between most biophysical parameters; (b) the PCA score-plot and (c) the corresponding loading plot. The PCA is based on all 27 parameters and all 14 proteins in all formulations. The dotted lines refer to the zero-line along both axes. (Please note that this is the result from a two-component PCA with imputed missing values).*

In general, weak linear correlations exist between some of the investigated biophysical parameters, like closely related parameters such as $T_{on,int}$ and $T_{m1,int}$, or $T_{m1,int}$ and $T_{m1,ex}$. We also tested the strength of the correlations in subsets of proteins in the dataset. For example, the analogous pairwise correlation analysis for the subset including only the 8 mAbs, each in 24 solution conditions, is shown in Supplementary information SI 9. Also, in SI10 the correlations in other subsets are shown. In general, the strength of the correlations observed in **Figure 1a** can slightly change when only a subset of the proteins like the one in SI 9 is selected, but the general trend that weak correlations exist is still present. We did not observe significant correlations between single experimental biophysical parameters and the molecular descriptors listed in SI 7 (data not shown).

In addition to the pairwise linear correlation it was decided to perform a PCA on the data to get an overview of both the similarities between the 14 different proteins, as well as a different view on the similarities between all the 27 parameters measured. As can be seen from **Figure 1b** most of the proteins are gathered around the origin, except for PPI-18 and PPI-45, clearly indicating that these proteins behave differently from the remaining proteins. By investigating the loading plot, **Figure 1c**, it becomes evident that this corresponds very well with the results from **Figure 1a**, e.g. all "*T*" parameters are grouped (indicating a high correlation), with variables such as R_{R50} and R_h on the opposite side of the origin (negative correlated). By inspecting both figures in **Figure 1b** and **1c** it is clear that PPI-18 especially has high values of m_1 and R_{R40} compared to the other proteins.

3.3 Biophysical parameters that flag proteins as developable or not are largely determined by the formulation conditions

We studied how pH and ionic strength affect the various computational and biophysical parameters often used as indicators that a protein has suitable properties for further development. The charge per amino acid calculated *in silico* with the Protein-Sol server greatly depends on the protein structure, pH and ionic strength. As an example, the dependency of charge per amino acid residue on pH and ionic strength for two antibodies (PPI-01, PPI-03), interferon α2a (PPI-30), and one lipase (PPI-45) are represented in **Figure 2**. The same server can provide similar contour plots for the effect of pH and ionic strength on conformational stability. Such computational characterization cannot immediately predict what will be the most stable condition for a given protein, but it is very important since it indicates what would be the expected trade-off between colloidal and conformational stability at different pH and ionic strength. Understanding such trade-offs is critical to determine the overall molecule stability.

Due to the volume and complexity of the data, response surface methodology (RSM) was applied to study how multiple biophysical parameters change as a function of pH and ionic strength. An example of two proteins, a bispecific antibody PPI-08 and an IgG1 PPI-03, is presented in **Figure 3**. The first

apparent melting temperature $T_{m1,int}$ from nanoDSF, the aggregation onset temperature T_{agg} from DLS, the interaction parameter k_D and the monomer mass recovery $m_{40,rec}$ after 2-week storage at 40 ºC are considered in this example. The borders of the contour plots are determined by the following cut-off values: $T_{m1,int}$ > 65 °C, T_{agg} > 55 °C, k_D > 0 mL/g, $m_{40,rec}$ > 80 %.

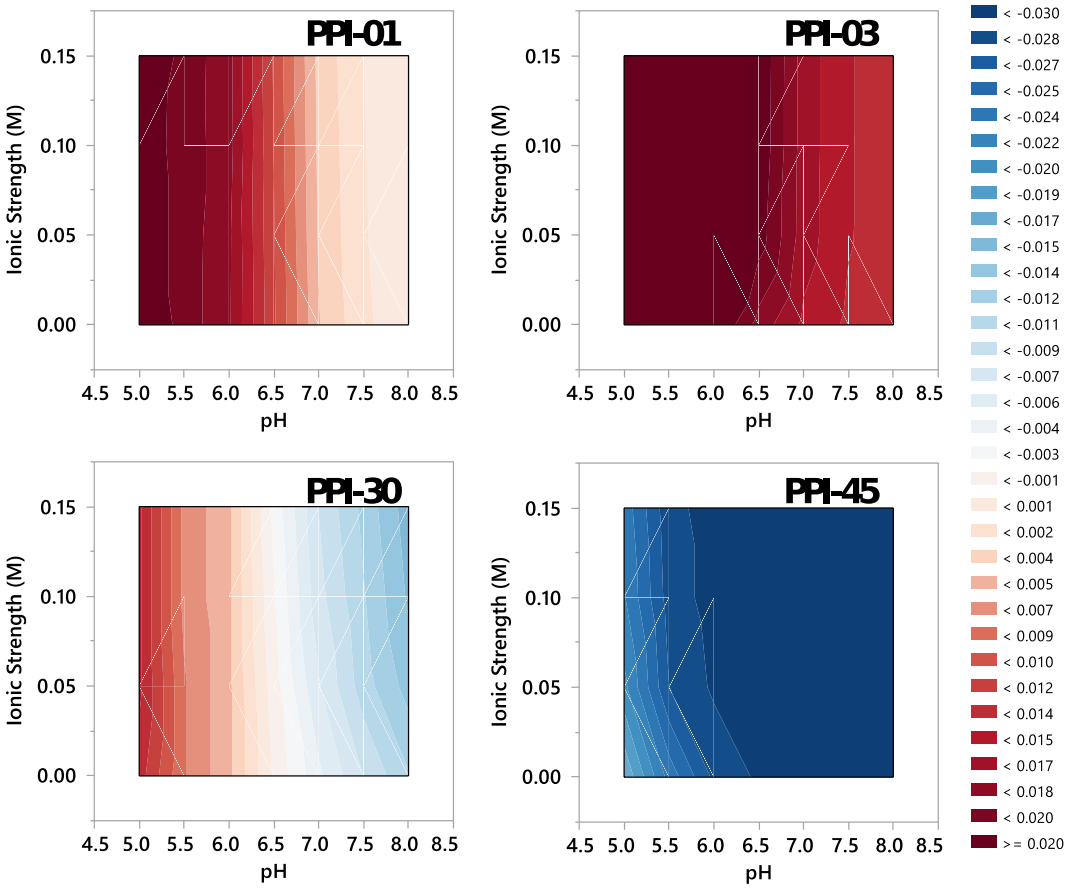


Figure 2. *Calculated charge per amino acid as a function of pH value and ionic strength for two antibodies (PPI-01, PPI-03), interferon α2a (PPI-30) and one lipase (PPI-45).*

The colored zones represent areas where the parameters are below the cut-off values mentioned above. Respectively, white areas indicate pH and ionic strength where all the parameters are above the cut-off values. Although such cut-off values are subjective and their definition may vary between labs, they are often used during developability assessment. In our case, we selected the cut-offs based on our experience, as explained in the discussion section below. Interestingly, a formulation “sweet spot” can be found for some of the proteins, but not for others. This “sweet spot” represents an area or a value in the RSM surfaces where all the selected biophysical parameters are above the defined cut-off values. Examples of proteins with a formulation “sweet spot” in our dataset are PPI-03, PPI-13, PPI-17, PPI-44 and PPI-46 (**Figure 3** and SI7). A common practice for selecting developable proteins is that the stability

of different candidates is compared in only one formulation condition. Noteworthy, if the proteins in our dataset had been assessed in only the commonly used phosphate buffered-saline (similar conditions of which are represented by a red square in **Figure 3**), all molecules but PPI-46 would have failed to be classified as developable according to the defined cut-off values. The arrow (in **Figure 3**) indicates that by using other formulation conditions, PPI-03 will move to a formulation "sweet spot" and actually meets all four cut-off criteria that would make it a good candidate for further development. On the other hand, PPI-08 presents a satisfactory $T_{m,int}$ in all the formulation conditions, while T_{agg}, k_D and especially $m_{40,rec}$ present critical values. This highlights the importance of a multi-parameter approach.

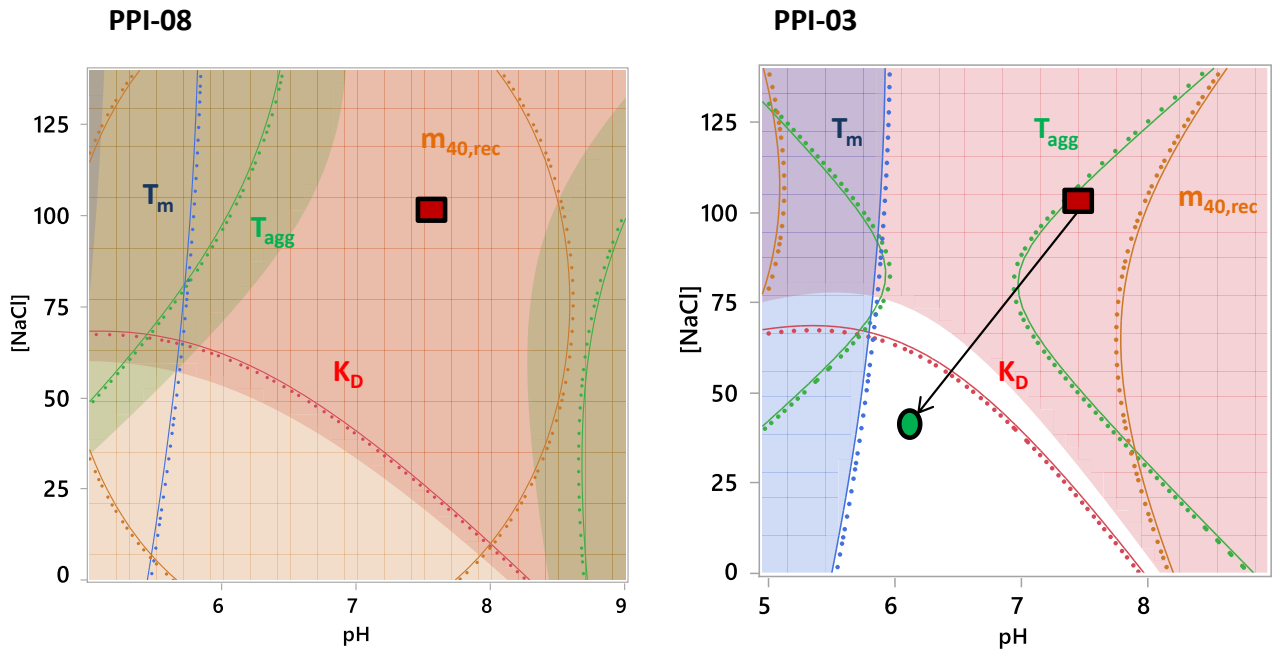


Figure 3. *Contour plot representing the change of several critical biophysical parameters with pH and salt concentration for PPI-08 and PPI-03 - white areas present regions where all four parameters are above the defined cut-off value. Areas colored in red, green, blue and yellow represent areas where respectively k_D, T_{agg}, $T_{m,int}$, and $m_{40,rec}$ are below the cut-off values. The dots highlight which part of the surface comes towards the reader, while the lines indicate a curvature of the surface. All the surfaces are superimposed.*

Of course, the example we present is very specific and changing the type of parameters and cut-offs can make molecules appear developable or not. However, **Figure 3** depicts something very important, which is often overlooked during developability assessment, i.e. the formulation conditions largely determine whether certain biophysical parameters will be above a certain stability threshold or not. Therefore, a proper assessment and comparison of therapeutic protein candidates can only be based on multiple parameters obtained in several formulation conditions. Otherwise, we risk a scenario where a generally stable molecule is not selected for further development only because it exhibits low stability in one assay buffer.

To tackle this issue and to rank the stability of the proteins based on data from multiple biophysical parameters and formulation conditions, one should focus on the existence and area of a formulation "sweet spot" area like the one for PPI-03 in **Figure 3**. We suggest that a larger cumulative "sweet spot" area of multiple biophysical properties will correspond to higher intrinsic stability of a protein molecule. Such data can be used to determine the "robustness" of the proteins across a broad formulation space, which is essential for both lead selection and formulation development. Based on this concept, we propose the calculation and use of stability risk values, as explained below.

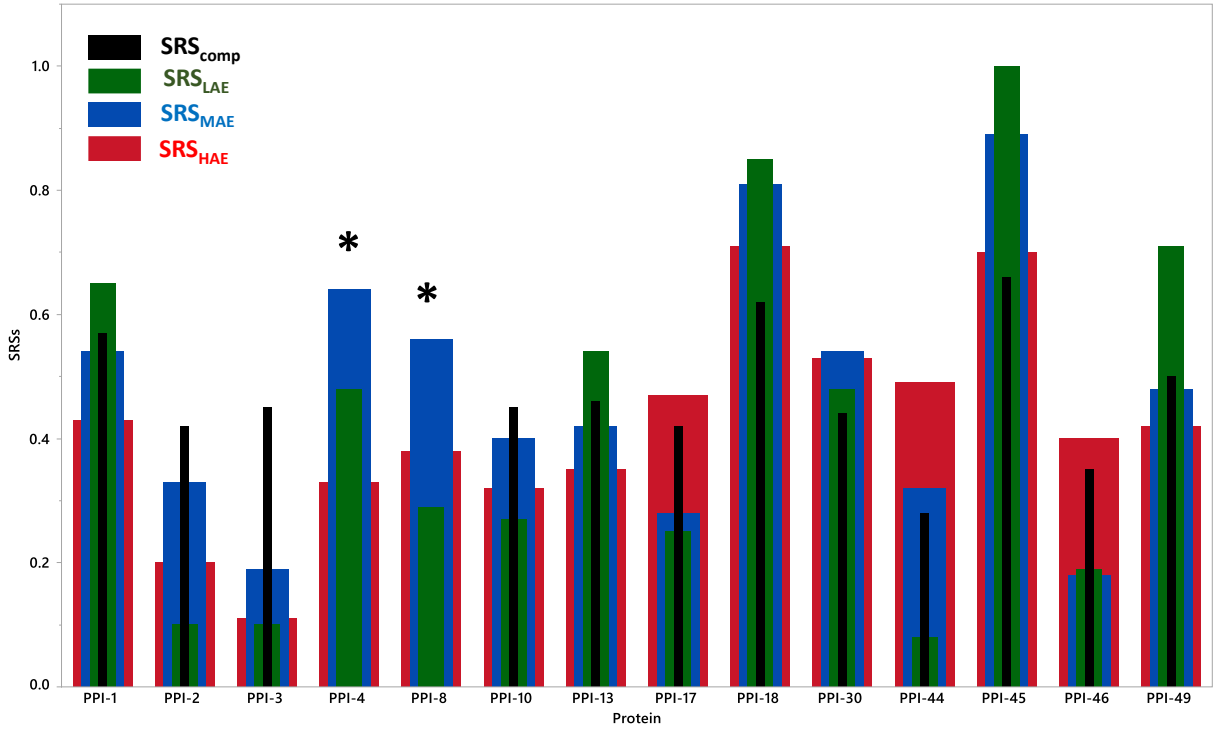
3.4 Datasets of various size can be used to generate stability risk scores for developability assessment

As shown above, a change in the formulation conditions, like pH and ionic strength, can result in a protein appearing suitable or unsuitable for development. Consequently, a more comprehensive characterization is required to understand whether a protein exhibits desirable biophysical properties or not. At the same time, the biophysical characterization is a trade-off among analytical efforts, time and sample consumption. To assess what analytical effort is needed to rank protein drug candidates based on their stability accurately, we calculated stability risk scores, ranging from 0 to 1, where higher values indicate a higher stability risk. The first stability risk score requires low analytical effort (SRS_{LAE}) and is calculated from parameters determined from high-throughput methods that require smaller protein quantities, namely T_{agg} and $T_{m1,int}$ (**Figure 4**, green bars). More advanced and labor-intensive characterization, including T_{agg}, T_m, k_D and $m_{40,rec}$, was added to the high-throughput characterization results to obtain a stability risk score obtained with medium analytical effort (SRS_{MAE}) (**Figure 4**, blue bars). Finally, many of the parameters measured in this work, namely T_{agg}, $T_{m1,int}$, k_D ,ζ , m_1, C_{m1}, PD, $m_{25,rec}$, $m_{40,rec}$, $m_{50,rec}$, LSA_{25}, LSA_{40} and LSA_{50}, were combined to obtain a stability risk score based on high analytical effort (SRS_{HAE}) (**Figure 4**, red bars).

To calculate the SRSs values, a risk region (i.e the reverse of the formulation "sweet spot") is defined by a series of cut-off parameter values. When the biophysical property value is in the risk region (below or above the cut-off value depending on the biophysical property) a value of 1 is assigned to that condition; otherwise, 0 is assigned. This procedure is repeated for all the biophysical properties and formulation conditions. Then, the nominal values are grouped, as shown in **Figure 4**.

The SRSs are calculated by calculating the mean of each group. Thus, SRS values between 0 and 1 are obtained for each protein as a function of all formulation conditions tested. The experimental SRSs are protein-dependent and calculated using multiple parameters assessed in different formulation conditions. The selection of the respective cut-off values presented in this work relies on: i) values reported in literature, e.g. many marketed antibodies have a T_{agg} greater than 55 °C;[3] ii) well-established principles,

e.g. highly positive k_D indicates high colloidal stability;[50,51] and iii) informed judgment selection, e.g $R_{R,25}$ > 1 indicates the formation of aggregates. Adjusting the cut-off values results in different slices of the surface and changes the size of the SRS region. For example, changing the T_{agg} cut-off from 55 °C to 25 °C for SRS_{LAE} will result in decreasing the risk values for all proteins.



Computational Stability Risk Score (SRS_{comp})

- No protein consumed
- Low experimental effort
- Fast
- Method used – Aggrescan, AggScore, Zyggregator
- * no primary sequence is available

Medium experimental effort Stability Risk Score (SRS_{MAE})

- ≈2 mg protein consumed to test all 24 formulations
- High experimental effort
- Slow
- Biophysical parameters: T_{agg}, k_D, T_m, $m_{40,rec}$

Low experimental effort Stability Risk Score (SRS_{LAE})

- 360 ug protein consumed to test all 24 formulations
- Moderate experimental effort
- Fast
- Biophysical parameters: T_{agg}, T_m

High experimental effort Stability Risk Score (SRS_{HAE})

- Several hundreds of mg of protein consumed to test all 24 formulations
- Very High experimental effort
- Very Slow
- Biophysical parameters: T_{agg}, k_D, T_m, $m_{40,rec}$, ζ , m_1, C_{m1}, PD% , $m_{25,rec}$, $m_{50,rec}$, LSA_{25}, LSA_{40}, LSA_{50}, R_{R25}, R_{R40}, R_{R50}

Figure 4. Stability risk score values for the proteins in the presented dataset. *SRScomp, SRSLAE, SRSMAE, SRSHAE are represented in black, green, blue and red, respectively. A higher SRS indicates an increased risk of stability issues. The asterisk (*) indicates proteins with no primary sequence available and therefore without a SRScomp in this study.*

Shifting the SRSs cut-off to an upper or lower limit, thus forcing the SRSs to 0 or 1 for all proteins, would result in a loss of information content. It is therefore important to select the values in an appropriate range such that a substantial portion of tested conditions falls on both sides of the threshold. A summary of the cut-off values to calculate the presented SRSs is also provided in Supplementary information – SI 12. Although the exact definition of the cut-off values for each biophysical parameter will still be a matter of discussion, we believe that our suggestion is a pragmatic and good starting point.

The computational SRS value, SRS_{comp} (**Fig. 4**, black line), is based on computational work only, and calculated using a different approach. The results of the total hydrophobic patch score and the mean aggregation tendency from Aggrescan, AggScore and Zyggregator were normalized from 0 to 1 and a mean value was calculated. Other variants of the SRS_{comp} were investigated, including a combination of several computational parameters and molecular descriptors (e.g. hydrophobicity index) yielding results that were generally poorer than the combined SRS_{comp} that we present in this example (data not shown).

Subsequently, we investigated the correlations between the SRSs values obtained with different analytical efforts. Interestingly, the SRS_{comp} correlates well with the SRS_{LAE} (**Figure 5**).

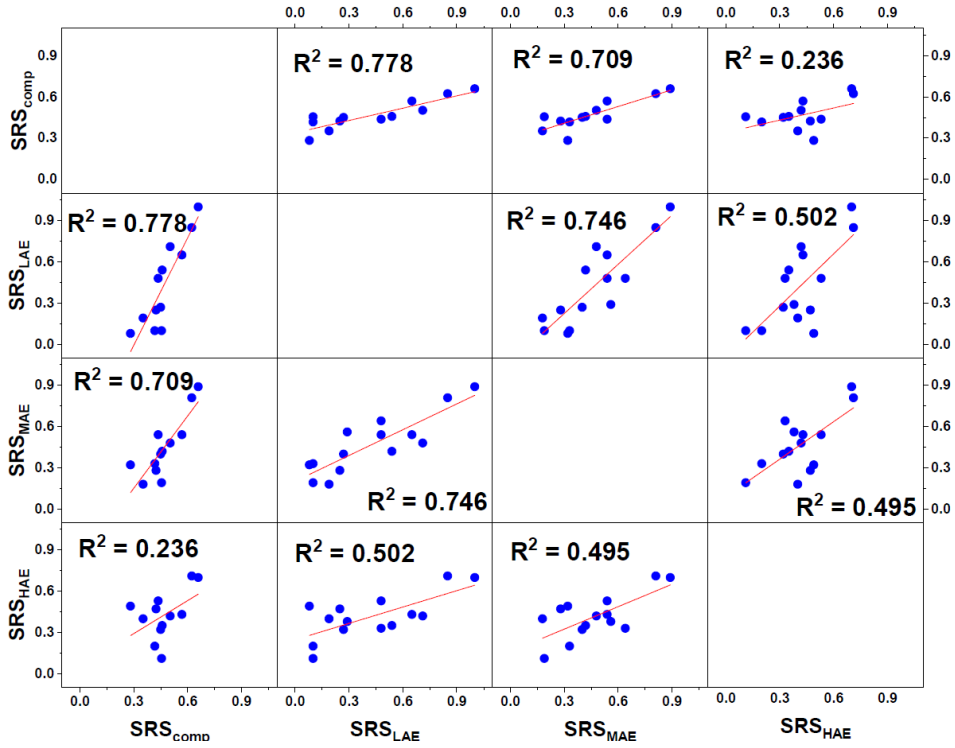


Figure 5. *Linear correlation between the protein rankings based on four stability risk scores. For more information regarding the calculation and definition of the different stability risk scores refer to the main manuscript text and Figure 4.*

However, when the size and complexity of the experimental dataset is increased, the correlation with the computational risk score decreases. The stability risk score based on the largest amount of experimental data (SRS_{HAE}) showed only a weak correlation with SRS_{comp}, but a moderate correlation with the SRS_{LAE} and SRS_{MAE}. Also, no or weak correlation among single computational parameters and experimental SRSs was observed (data not shown). In general, most of the molecular descriptors calculated from the homology models or primary sequences are either weakly or not influenced by pH and ionic strength which might explain the low correlation to stability risk scores obtained from characterization in different formulation conditions.

3.5 Stability risk scores obtained from larger datasets exhibit better correlation with the amount of aggregates formed during storage

The obtained stability risk scores are validated by correlating the values with the amount of aggregates formed during storage for 6 months at 4 and 25 ºC. This storage stability data is generated for all proteins in four different formulations. The linear correlations between SRS_{HAE} and the percentage of aggregates after six months of storage at refrigerated and room temperature are shown in **Figure 6**.

This percentage is calculated using the relative UV area of high molecular weight species, after size exclusion chromatography (SEC), and corrected for the missing mass from the total column recovery. The correction is necessary to adjust for big and/or insoluble aggregates which are filtered out by the column or lost by sedimentation before injection. Similar data can be derived from the light scattering area. These results demonstrate a strong correlation between the experimental SRSs for physical stability risk assessment and the percentage of aggregates formed during storage at temperatures relevant for therapeutic proteins. A summary of the correlation coefficients between the SRSs and the percentage of aggregation is shown in **Figure 7**.

The Pearson's correlation coefficient is calculated similarly as described earlier. These values were averaged over all proteins, formulations and temperatures of stress studied. SRS_{comp} present the lowest mean correlation and highest variability. As expected, by increasing the analytical effort the correlations become stronger and the predictions more reliable. SRS_{HAE} strongly correlates with protein stability with a very low variability, making this value the most robust for protein ranking.

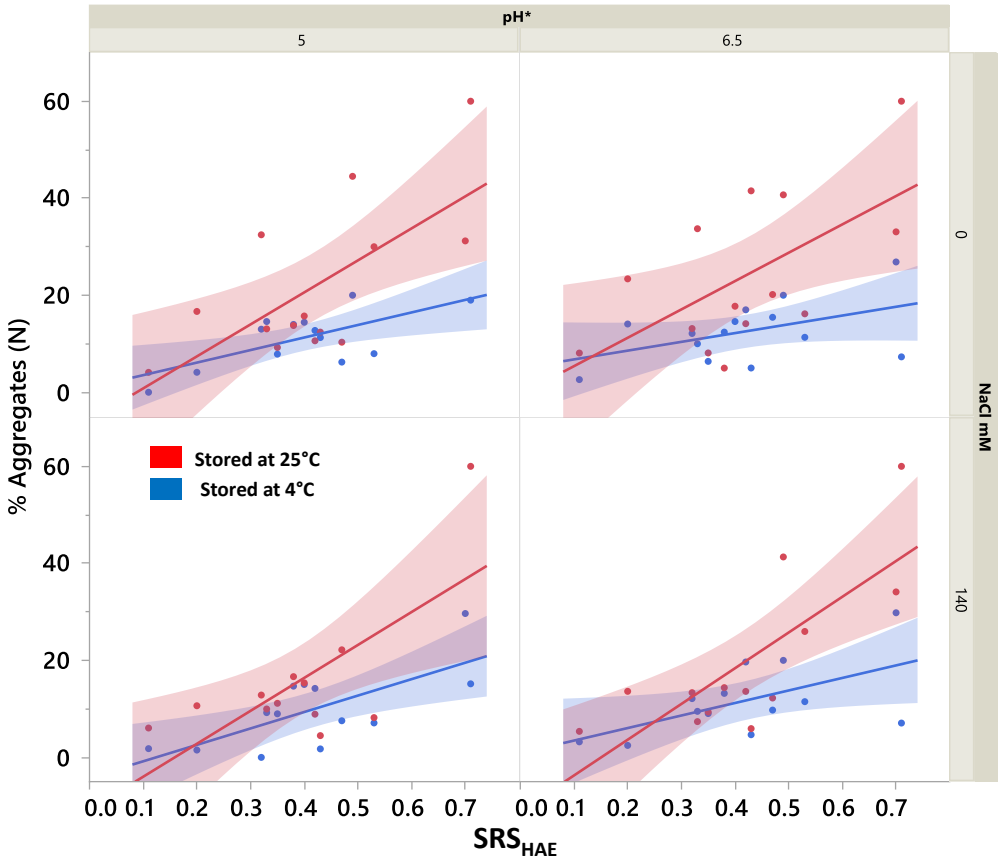


Figure 6. Linear correlation between SRSHAE and the percentage of aggregates after 6 months of storage at 4 °C (in blue) and 25 °C (in red). *A total of four formulations were studied i) 10 mM His at pH 5 ii) 10 mM His and 140 mM NaCl at pH 5, iii) 10 mM His at pH 6, iv) 10 mM His and 140 mM NaCl at pH 5. The filled area represents 95% confidence intervals. *PPI-30, PPI-45, PPI-46 were formulated at pH 7.5 instead of pH 6.5. PPI-45 and PPI-46 were formulated at pH 5.5 instead of pH 5. The pHs were selected to include a "good" and a "bad" formulation in a pharmaceutically relevant pH range.*

Interestingly SRS_{LAE} and SRS_{MAE} present similar prediction power which confirms that an early rough ranking by using few high throughput biophysical parameters, namely T_{agg} and $T_{m1,int}$, assessed in various solution conditions, is possible in cases where sample volume is very limited. Finally, we suggest that, based on the SRSs, the proteins can be classified as having a low (SRS < 0.3), medium (0.3 > SRS > 0.6) or high developability risk (SRS > 0.6).

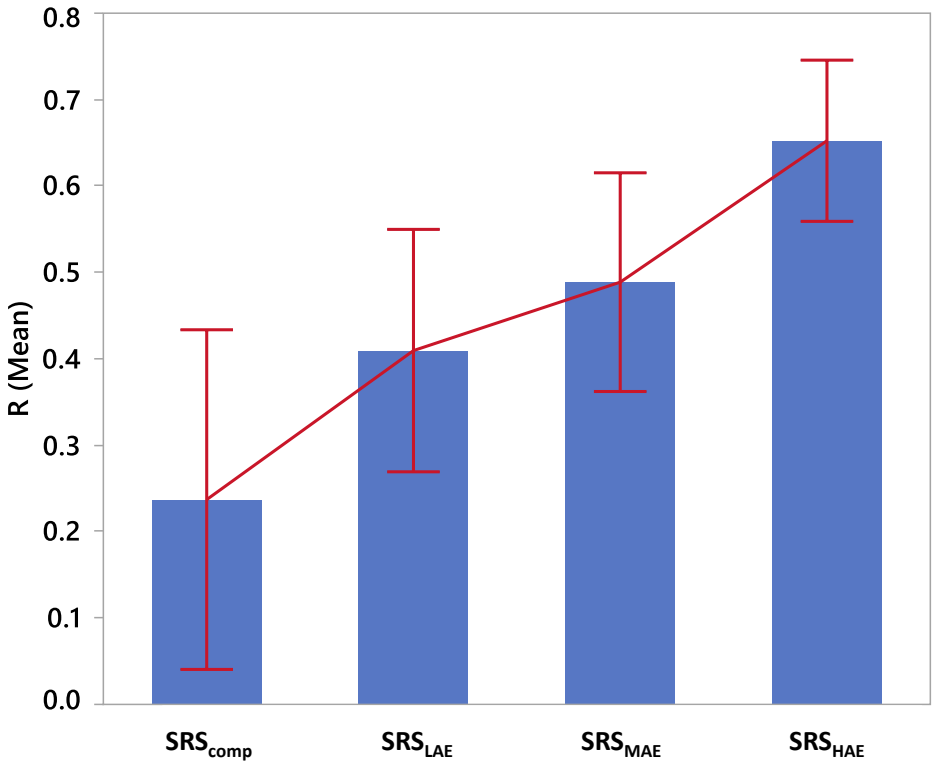


Figure 7. Averaged Pearson's correlation coefficient R between SRSs and the percentage of aggregates after 6 months of storage at 4 °C and 25 °C. *The mean and error bars are calculated as a standard variation of the R value between SRSs and the amount of aggregates after 6 months of storage over all proteins, formulations and temperatures of stress studied.*

4 Discussion

Therapeutic protein candidates should exhibit a set of desirable biophysical parameters which indicate sufficient stability and drug-like properties.[6,8,10,15,52] These properties are assessed at very early stages and serve as the decision basis for which molecules will be selected for further development.[6,53,54] For over two decades, the community has striven to find the most efficient approaches to select proteins with drug-like properties. However, generally accepted guidelines that are applicable to the diverse landscape of therapeutic proteins have not yet emerged. Defining rules and strategies for this selection can only be based on a large body of published research that employs orthogonal techniques to characterize therapeutic proteins in different formulations. Although a great progress has been made by several landmark studies, work showing the feasibility of certain selection approaches is often based on i)

molecules with unknown primary sequences and purity; ii) on protein datasets including only structurally similar molecules, e.g. antibodies; and/ or iii) assessment of biophysical parameters in only one formulation condition.[5,12,15–17,52] For example, such studies report correlations among parameters related to protein thermal stability, colloidal stability, solubility and viscosity.[55–59] Another correlation that is often reported is the one between the onset temperature of protein unfolding or protein melting temperature and the aggregation rate during accelerated stability studies.[17,60] Although we do not question the existence of such correlations in a particular case study, we show here that the relationships between some biophysical parameters cannot be generalized for a heterogeneous population of proteins in a diverse set of formulation conditions. These findings highlight that "protein stability" cannot be well described by using a single biophysical parameter, nor by studying a protein in a single solution condition.

Probably the biggest advance to understand which features make a protein developable has been made for therapeutic monoclonal antibodies. However, the next generation of therapeutic proteins will be more diverse, including fusion proteins, enzymes and cytokines, among others. Understanding what exactly indicates intrinsic stability of a protein molecule requires that more information on various therapeutically-relevant proteins, including their primary sequences, purity data, and comprehensive computational and biophysical characterization in different solution conditions is made publicly available.

In this work, we present a dataset which includes comprehensive computational and biophysical stability characterization of 14 diverse therapeutically relevant proteins in 24 different formulation conditions. We use the data to look for linear pairwise correlations among a variety of biophysical parameters that are considered to be indicative for protein stability. We find linear correlations among some biophysical parameters, but not among others. Future work will focus on more complex analyses of the presented dataset to find whether the connection between some computational and biophysical parameters can be described by more advanced models. For example, we are currently focusing on multivariate data analysis, while some machine learning approaches based on the presented data are already published.[61]

Since the presented biophysical parameters often have a complex non-linear dependence as a function of pH and ionic strength, we adopted an RMS approach to describe this behaviour. This allowed us to visualize and define boundaries which show whether a biophysical parameter will be above or below a certain stability cut-off that will flag a protein with desirable or undesirable features. The RMS methodology shows that some proteins in the dataset exhibit a formulation "sweet spot", i.e. a range of pH and ionic strength where all biophysical parameters are above the desired threshold. Interestingly, if we perform comparison among different proteins by using only one formulation condition (e.g. having pH and ionic strength close to phosphate-buffered saline), we should put a flag on many of the proteins that actually have a broad formulation "sweet spot". This raises the question whether the developability assessment of proteins based on assays performed in only one buffer are less reliable than a comparison based on data in several formulation conditions. Indeed, studying a protein in different conditions would

increase the analytical effort, but thanks to the technological advancement, it is now possible to perform developability assessment in dozens of solution conditions with only minimal protein consumption. In this paper there is, for example, a study of the thermal unfolding and aggregation of proteins in 24 different formulation conditions which consumes only a total of 360 µg of protein (i.e. for nanoDSF and DLS with temperature ramp).

Here, we also present how the multiple parameters can be combined into stability risk scores (SRS). These scores are based on the two considerations mentioned above: i) the biophysical parameters carry unique information and ii) the formulation condition substantially influences those parameters. The SRSs are protein-specific values that are calculated from multiple parameters, assessed for multiple formulation conditions. The calculations are simple and only based on critical limits for each parameter. We show how these SRS values are related to each other. Interestingly, the computational SRS ranking better correlates to the SRS ranking based only on few basic biophysical parameters. However, if the stability risk score is based on a larger set of experimental data, the correlations with the computational ranking become weaker. This does not mean that the computational characterization is not important since it still provided good predictions for the first round of characterization. Also, we have already demonstrated that other *in silico* approaches can be applied to proteins for a structure-based discovery of aggregation breaking excipient of PPI-30[62] or characterization of peptides[63]. In addition, *in silico* approaches have been developed to predict whether certain features in the complementarity-determining regions in mAbs can lead to stability problems.[12] We validate the different SRS values by showing how they correlate with the amount of aggregates formed by the different proteins during storage for 6 months at 4 °C and 25 °C. Intuitively, an SRS calculated from more biophysical parameters correlates better with the storage stability of the proteins, and thus can be used for more reliable prediction of developable candidates. Besides this, we expect that a protein having a high SRS calculated from various formulation conditions will be less challenging during formulation development. In the near future, data used in this study will be available for download from a tailor-made database (https://pippi-data.kemi.dtu.dk/). This public database will be the basis for novel insights into the complex connection between therapeutic protein structure, formulation conditions, biophysical properties and storage stability.

5 Acknowledgments

This study was funded by a project part of the EU Horizon 2020 Research and Innovation program under the Marie Skłodowska-Curie grant agreement No 675074. We thank the whole PIPPI consortium (http://www.pippi.kemi.dtu.dk) for the continuous support. The first author thanks Wyatt Technology staff members for their many contributions.

6 List of abbreviations

cIEF – capillary isoelectric focusing; C_{m1} – melting denaturant concentration from the first unfolding in GuHCl; C_{m2} – melting denaturant concentration from the second unfolding in GuHCl; D_0 – protein diffusion coefficient at infinite dilution; dG_1 – apparent Gibbs free energy of the first unfolding in GuHCl; dG_2 – apparent Gibbs free energy of the second unfolding in GuHCl; DLS – dynamic light scattering; DoE – design of experiments; DSF – differential scanning fluorimetry; ICD – isothermal chemical denaturation; k_D – interaction parameter from DLS; LSA_{25} – empirical parameter indicating the presence of particles determined after 2 weeks at 25 °C; LSA_{40} – empirical parameter indicating the presence of particles determined after 2 weeks at 40 °C; LSA_{50} – empirical parameter indicating the presence of particles determined after 2 weeks at 50 °C; m_1 – empirical parameter describing the cooperativity of the first unfolding in GuHCl; m_2 – empirical parameter describing the cooperativity of the second unfolding in GuHCl; M_{25} – mass fraction of monomer compared to all peaks after 2 weeks at 25 °C; $m_{25,rec}$ – monomer mass recovery after 2 weeks at 25 °C; M_{40} – mass fraction of monomer compared to all peaks after 2 weeks at 40 °C; $m_{40,rec}$ – monomer mass recovery after 2 weeks at 40 °C; M_{50} – mass fraction of monomer compared to all peaks after 2 weeks at 50 °C; $m_{50,rec}$ – monomer mass recovery after 2 weeks at 50 °C; mAb – monoclonal antibody; MOE - molecular operating environment software; nanoDSF® - fluorimetric method based on intrinsic protein fluorescence; PD – polydispersity from DLS; PDB – Protein Data Bank; PEG_{TMP} – inflection point of the fit to the PEG titration curve; R_h – protein hydrodynamic radius at 1 mg/ml from DLS; $R_{R,25}$ – relative increase in the hydrodynamic radius after 2 weeks at 25 °C; $R_{R,40}$ – relative increase in the hydrodynamic radius after 2 weeks at 40 °C; $R_{R,50}$ – relative increase in the hydrodynamic radius after 2 weeks at 50 °C; RSM – response surface methodology; SEC-MALS – size exclusion chromatography coupled to multi-angle light scattering; SRS_{comp} – stability risk score from computational parameters; SRS_{HAE} – stability risk score from experimental parameters with high analytical effort; SRS_{LAE} – stability risk score from experimental parameters with low analytical effort; SRS_{MAE} – stability risk score from experimental parameters with medium analytical effort; T_{agg} – aggregation onset temperature from DLS; $T_{m1,ex}$ – first apparent melting temperature from DSF with extrinsic dye; $T_{m1,int}$ – first apparent melting temperature from nanoDSF®; $T_{m2,int}$ – second apparent melting temperature from nanoDSF®; $T_{on,ex}$ – onset of the first thermal protein unfolding from DSF with extrinsic dye; $T_{on,int}$ – onset of the first thermal protein unfolding from nanoDSF®; ζ – zeta potential.

7 References

1. Dimitrov, D. S. Therapeutic Proteins; Humana Press, Totowa, NJ, 2012.

2. Strohl, W. R.; Knight, D. M. Discovery and Development of Biopharmaceuticals: Current Issues. Curr. Opin. Biotechnol. 2009, 20 (6), 668–672.

3. 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. 2015, 104 (6), 1885–1898.

4. Carter, P. J. Potent Antibody Therapeutics by Design. Nat. Rev. Immunol. 2006, 6 (5), 343–357.

5. Liu, Y.; Caffry, I.; Wu, J.; Geng, S. B.; Jain, T.; Sun, T.; Reid, F.; Cao, Y.; Estep, P.; Yu, Y.; et al. High-Throughput Screening for Developability during Early-Stage Antibody Discovery Using Self-Interaction Nanoparticle Spectroscopy. MAbs 2014, 6 (2), 483–492.

6. Zurdo, J. Developability Assessment as an Early De-Risking Tool for Biopharmaceutical Development. Pharm. Bioprocess. 2013, 1 (1), 29–50.

7. 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 2019, 11 (2), 388–400.

8. Yang, Y.; Velayudhan, A.; Thornhill, N. F.; Farid, S. S. Multi-Criteria Manufacturability Indices for Ranking High-Concentration Monoclonal Antibody Formulations. Biotechnol. Bioeng. 2017, 114 (9), 2043–2056.

9. Chennamsetty, N.; Voynov, V.; Kayser, V.; Helk, B.; Trout, B. L. Design of Therapeutic Proteins with Enhanced Stability. Proc. Natl. Acad. Sci. U. S. A. 2009, 106 (29), 11937–11942.

10. Starr, C. G.; Tessier, P. M. Selecting and Engineering Monoclonal Antibodies with Drug-like Specificity. Curr. Opin. Biotechnol. 2019, 60, 119–127.

11. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. Adv. Drug Deliv. Rev. 2001, 46 (1–3), 3–26.

12. 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. 2019, 116 (10), 4025–4030.

13. Rabia, L. A.; Zhang, Y.; Ludwig, S. D.; Julian, M. C.; Tessier, P. M. Net Charge of Antibody Complementarity-Determining Regions Is a Key Predictor of Specificity. Protein Eng. Des. Sel. 2018, 31 (11), 409–418.

14. 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. 2018, 137, 365–374.

15. Jain, T.; Sun, T.; Durand, S.; Hall, A.; Houston, N. R.; Nett, J. H.; Sharkey, B.; Bobrowicz, B.; Caffry, I.; Yu, Y.; et al. Biophysical Properties of the Clinical-Stage Antibody Landscape. Proc. Natl. Acad. Sci. 2017, 114 (5), 944–949.

16. Shan, L.; Mody, N.; Sormanni, P.; Rosenthal, K. L.; Damschroder, M. M.; Esfandiary, R.; Sormani, P.; Rosenthal, K. L.; Damschroder, M. M.; Esfandiary, R.; et al. Developability Assessment of Engineered Monoclonal Antibody Variants with a Complex Self-Association Behavior Using Complementary Analytical and in Silico Tools. Mol. Pharm. 2018, 15 (12), 5697–5710.

17. Brader, M. L.; Estey, T.; Bai, S.; Alston, R. W.; Lucas, K. K.; Lantz, S.; Landsman, P.; Maloney, K. M. Examination of Thermal Unfolding and Aggregation Profiles of a Series of Developable Therapeutic Monoclonal Antibodies. Mol. Pharm. 2015, 12 (4), 1005–1017.

18. Thiagarajan, G.; Semple, A.; James, J. K.; Cheung, J. K.; Shameem, M. A Comparison of Biophysical Characterization Techniques in Predicting Monoclonal Antibody Stability. MAbs 2016, 8 (6), 1088–1097.

19. Goldberg, D. S.; Lewus, R. A.; Esfandiary, R.; Farkas, D. C.; Mody, N.; Day, K. J. K.; Mallik, P.; Tracka, M. B.; Sealey, S. K.; Samra, H. S. Utility of High Throughput Screening Techniques to Predict Stability of Monoclonal Antibody Formulations During Early Stage Development. J. Pharm. Sci. 2017, 106 (8), 1971–1977.

20. Finnis, C. J. A.; Payne, T.; Hay, J.; Dodsworth, N.; Wilkinson, D.; Morton, P.; Saxton, M. J.; Tooth, D. J.; Evans, R. W.; Goldenberg, H.; et al. High-Level Production of Animal-Free Recombinant Transferrin from Saccharomyces Cerevisiae. Microb. Cell Fact. 2010, 9, 87.

21. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T. L. BLAST+: Architecture and Applications. BMC Bioinformatics 2009, 10 (1), 421.

22. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank, 1999–. In International Tables for Crystallography; International Union of Crystallography: Chester, England, 2006; pp 675–684.

23. Thompson, J. D.; Higgins, D. G.; Gibson, T. J. CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice. Nucleic Acids Res. 1994, 22 (22), 4673–4680.

24. Eswar, N.; Webb, B.; Marti-Renom, M. A.; Madhusudhan, M. S.; Eramian, D.; Shen, M.; Pieper, U.; Sali, A. Comparative Protein Structure Modeling Using Modeller. Curr. Protoc. Bioinforma. 2006, Chapter 5 (Unit 5.6), 1–30.

25. Harris, L. J.; Larson, S. B.; Hasel, K. W.; McPherson, A. Refined Structure of an Intact IgG2a Monoclonal Antibody. Biochemistry 1997, 36 (7), 1581–1597.

26. Maurer, B.; Bosanac, I.; Shia, S.; Kwong, M.; Corpuz, R.; Vandlen, R.; Schmidt, K.; Eigenbrot, C. Structural Basis of the Broadly Neutralizing Anti-Interferon-α Antibody Rontalizumab. Protein Sci. 2015, 24 (9), 1440–1450.

27. Noinaj, N.; Easley, N. C.; Oke, M.; Mizuno, N.; Gumbart, J.; Boura, E.; Steere, A. N.; Zak, O.; Aisen, P.; Tajkhorshid, E.; et al. Structural Basis for Iron Piracy by Pathogenic Neisseria. Nature 2012, 483 (7387), 53–58.

28. Brzozowski, A. M.; Derewenda, U.; Derewenda, Z. S.; Dodson, G. G.; Lawson, D. M.; Turkenburg, J. P.; Bjorkling, F.; Huge-Jensen, B.; Patkar, S. A.; Thim, L. A Model for Interfacial Activation in Lipases from the Structure of a Fungal Lipase-Inhibitor Complex. Nature 1991, 351 (6326), 491–494.

29. Yapoudjian, S.; Ivanova, M. G.; Brzozowski, A. M.; Patkar, S. A.; Vind, J.; Svendsen, A.; Verger, R. Binding of Thermomyces (Humicola) Lanuginosa Lipase to the Mixed Micelles of Cis-Parinaric Acid/NaTDC: Fluorescence Resonance Energy Transfer and Crystallographic Study. Eur. J. Biochem. 2002, 269 (6), 1613–1621.

30. Sivertsen, A.; Isaksson, J.; Leiros, H.-K. S.; Svenson, J.; Svendsen, J.-S.; Brandsdal, B. Synthetic Cationic Antimicrobial Peptides Bind with Their Hydrophobic Parts to Drug Site II of Human Serum Albumin. BMC Struct. Biol. 2014, 14 (1), 4.

31. Madhavi Sastry, G.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W. Protein and Ligand Preparation: Parameters, Protocols, and Influence on Virtual Screening Enrichments. J. Comput. Aided. Mol. Des. 2013, 27 (3), 221–234.

32. Wiederstein, M.; Sippl, M. J. ProSA-Web: Interactive Web Service for the Recognition of Errors in Three-Dimensional Structures of Proteins. Nucleic Acids Res. 2007, 35 (Issue suppl_2), W407–W410.

33. Hebditch, M.; Carballo-Amador, M. A.; Charonis, S.; Curtis, R.; Warwicker, J. Protein–Sol: A Web Tool for Predicting Protein Solubility from Sequence. Bioinformatics 2017, 33 (19), 3098–3100.

34. Ruiz-Blanco, Y. B.; Paz, W.; Green, J.; Marrero-Ponce, Y. ProtDCal: A Program to Compute General-Purpose-Numerical Descriptors for Sequences and 3D-Structures of Proteins. BMC Bioinformatics 2015, 16 (1), 162.

35. Sankar, K.; Krystek, S. R.; Carl, S. M.; Day, T.; Maier, J. K. X. AggScore: Prediction of Aggregation-Prone Regions in Proteins Based on the Distribution of Surface Patches. Proteins Struct. Funct. Bioinforma. 2018, 86 (11), 1147–1156.

36. Tartaglia, G. G.; Vendruscolo, M. The Zyggregator Method for Predicting Protein Aggregation Propensities. Chem. Soc. Rev. 2008, 37 (7), 1395.

37. 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 2007, 8, 65.

38. Menzen, T.; Friess, W. High-Throughput Melting-Temperature Analysis of a Monoclonal Antibody by Differential Scanning Fluorimetry in the Presence of Surfactants. J. Pharm. Sci. 2013, 102 (2), 415–428.

39. Freire, E.; Schön, A.; Hutchins, B. M.; Brown, R. K. Chemical Denaturation as a Tool in the Formulation Optimization of Biologics. Drug Discov. Today 2013, 18 (19–20), 1007–1013.

40. Myers, J. K.; Pace, C. N.; Scholtz, J. M. Denaturant m Values and Heat Capacity Changes: Relation to Changes in Accessible Surface Areas of Protein Unfolding [Published Erratum Appears in Protein Sci 1996 May;5(5):981]. Protein Sci 1995, 4 (10), 2138–2148.

41. Wafer, L.; Kloczewiak, M.; Polleck, S. M.; Luo, Y. Isothermal Chemical Denaturation of Large Proteins: Path-Dependence and Irreversibility. Anal. Biochem. 2017, 539, 60–69.

42. Filoti, D. I.; Shire, S. J.; Yadav, S.; Laue, T. M. Comparative Study of Analytical Techniques for Determining Protein Charge. J. Pharm. Sci. 2015, 104 (7), 2123–2131.

43. Roberts, D.; Keeling, R.; Tracka, M.; van der Walle, C. F.; Uddin, S.; Warwicker, J.; Curtis, R. Specific Ion and Buffer Effects on Protein–Protein Interactions of a Monoclonal Antibody. Mol. Pharm. 2015, 12 (1), 179–193.

44. Ohshima, H. A Simple Expression for Henry's Function for the Retardation Effect in Electrophoresis of Spherical Colloidal Particles. J. Colloid Interface Sci. 1994, 168 (1), 269–271.

45. Lehman, A.; O'Rourke, N.; Hatcher, L.; Stepanski, E. J. JMP for Basic Univariate and Multivariate Statistics: A Step-by-Step Guide; SAS Institute. Inc.: Cary, North Carolina, USA, 2005.

46. Kumar, S.; Tsai, C. J.; Nussinov, R. Temperature Range of Thermodynamic Stability for the Native State of Reversible Two-State Proteins. Biochemistry 2003, 42 (17), 4864–4873.

47. Tomar, D. S.; Li, L.; Broulidakis, M. P.; Luksha, N. G.; Burns, C. T.; Singh, S. K.; Kumar, S. In-Silico Prediction of Concentration-Dependent Viscosity Curves for Monoclonal Antibody Solutions. MAbs 2017, 9 (3), 476–489.

48. Grung, B.; Manne, R. Missing Values in Principal Component Analysis. Chemom. Intell. Lab. Syst. 1998, 42 (1–2), 125–139.

49. 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. 2019, 562, 42–50.

50. Tomar, D. S.; Kumar, S.; Singh, S. K.; Goswami, S.; Li, L. Molecular Basis of High Viscosity in Concentrated Antibody Solutions: Strategies for High Concentration Drug Product Development. MAbs 2016, 8 (2), 216–228.

51. Minton, A. P. Recent Applications of Light Scattering Measurement in the Biological and Biopharmaceutical Sciences. Anal. Biochem. 2016, 501, 4–22.

52. Lauer, T. M.; Agrawal, N. J.; Chennamsetty, N.; Egodage, K.; Helk, B.; Trout, B. L. Developability Index: A Rapid in Silico Tool for the Screening of Antibody Aggregation Propensity. J. Pharm. Sci. 2012, 101 (1), 102–115.

53. Rogers, R. S.; Abernathy, M.; Richardson, D. D.; Rouse, J. C.; Sperry, J. B.; Swann, P.; Wypych, J.; Yu, C.; Zang, L.; Deshpande, R. A View on the Importance of "Multi-Attribute Method" for Measuring Purity of Biopharmaceuticals and Improving Overall Control Strategy. AAPS J. 2018, 20 (1), 7.

54. 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 2019, 11 (2), 239–264.

55. Connolly, B. D.; Petry, C.; Yadav, S.; Demeule, B.; Ciaccio, N.; Moore, J. M. R.; Shire, S. J.; Gokarn, Y. R. Weak Interactions Govern the Viscosity of Concentrated Antibody Solutions: High-Throughput Analysis Using the Diffusion Interaction Parameter. Biophys. J. 2012, 103 (1), 69–78.

56. Yadav, S.; Laue, T. M.; Kalonia, D. S.; Singh, S. N.; Shire, S. J. The Influence of Charge Distribution on Self-Association and Viscosity Behavior of Monoclonal Antibody Solutions. Mol. Pharm. 2012, 9 (4), 791–802.

57. Rubin, J.; Sharma, A.; Linden, L.; Bommarius, A. S.; Behrens, S. H. Gauging Colloidal and Thermal Stability in Human IgG1-Sugar Solutions through Diffusivity Measurements. J. Phys. Chem. B 2014, 118 (11), 2803–2809.

58. George, A.; Wilson, W. W. Predicting Protein Crystallization from a Dilute Solution Property. Acta Crystallogr. Sect. D Biol. Crystallogr. 1994, 50 (4), 361–365.

59. Tomar, D. S.; Singh, S. K.; Li, L.; Broulidakis, M. P.; Kumar, S. In Silico Prediction of Diffusion Interaction Parameter (KD), a Key Indicator of Antibody Solution Behaviors. Pharm. Res. 2018, 35 (10), 193.

60. Burton, L.; Gandhi, R.; Duke, G.; Paborji, M. Use of Microcalorimetry and Its Correlation with Size Exclusion Chromatography for Rapid Screening of the Physical Stability of Large Pharmaceutical Proteins in Solution. Pharm. Dev. Technol. 2007, 12 (3), 265–273.

61. Gentiluomo, L.; Roessner, D.; Augustijn, D.; Svilenov, H.; Kulakova, A.; Mahapatra, S.; Winter, G.; Streicher, W.; Rinnan, Å.; Peters, G. H. J.; et al. Application of Interpretable Artificial Neural Networks to Early Monoclonal Antibodies Development. Eur. J. Pharm. Biopharm. 2019, 141, 81–89.

62. Tosstorff, A.; Svilenov, H.; Peters, G. H. J.; Harris, P.; Winter, G. Structure-Based Discovery of a New Protein-Aggregation Breaking Excipient. Eur. J. Pharm. Biopharm. 2019, 144, 207–216.

63. Indrakumar, S.; Zalar, M.; Pohl, C.; Nørgaard, A.; Streicher, W.; Harris, P.; Golovanov, A. P.; Peters, G. H. J. Conformational Stability Study of a Therapeutic Peptide Plectasin Using Molecular Dynamics Simulations in Combination with NMR. J. Phys. Chem. B 2019, 123 (23), 4867–4877.

8 Supplementary information

List of supplementary information

SI 1. Primary sequences of the studied proteins

SI 2. Zscore values for the homology model structures

SI 3. Details on the calculation used for extrinsic DSF

SI 4. Fitting from the response surface methodology (RSM)

SI 5. Multivariate matrix including all datapoints

SI 6. Separations obtained with SEC-MALS and cIEF for the proteins in the dataset

SI 7. List of the molecular descriptor calculated by MOE and ProDCal

SI 8. Parameters from AggScore, Zyggregator and Aggrescan

SI 9. Pairwise correlations among biophysical parameters in a subset including only mAbs.

SI 10. Pairwise correlations among biophysical parameters in subsets including different proteins.

SI 11. Surface profiles of the investigated proteins relative to a reduced subset of several biophysical parameters

SI 12. Cut off values used for the calculation of the different stability risk scores (SRSs)

SI 13. Biophysical parameters table

Supplementary information – SI 1. Primary sequences of the studied proteins

SI 1.1 PPI-01 (IgG1)

Heavy chain

EVQLVQSGAEVKKPGATVKISCKVYGYIFTDYNIYWVRQAPGKGLEWMGLIDPDNGETFYAEKFQGRAT
MTADTSSDRAYMELSSLRFEDTAVYYCATVMGKWIKGGYDYWGRGTLVTVSSASTKGPSVFPLAPSSK
STSGGTAALGCLVKDYFPEPVTVSWNSGALTSGVHTFPAVLQSSGLYSLSSVVTVPSSSLGTQTYICNVN
HKPSNTKVDKKVEPKSCDKTHTCPPCPAPELLGGPSVFLFPPKPKDTLMISRTPEVTCVVVDVSHEDPEV
KFNWYVDGVEVHNAKTKPREEQY**N**STYRVVSVLTVLHQDWLNGKEYKCKVSNKALPAPIEKTISKAKGQ
PREPQVYTLPPSRDELTKNQVSLTCLVKGFYPSDIAVEWESNGQPENNYKTTPPVLDSDGSFFLYSKLTV
DKSRWQQGNVFSCSVMHEALHNHYTQKSLSLSPGK

N-glycosylation site

Light chain

QSVLTQPPSVSGAPGQRVTISCTGSSSNIGAGYDVHWYQQLPGTAPKLLIYDNFNRPSGVPPRFSGSKS
GTSASLAITGLQAEDEADYYCQSYDSPTLTSPFGTGTLTVLGQPKAAPSVTLFPPSSEELQANKATLVCLI
SDFYPGAVTVAWKADSSPVKAGVETTTPSKQSNNKYAASSYLSLTPEQWKSHRSYSCQVTHEGSTVEK
TVAPTECS

SI 1.2 PPI-02 (IgG1)

Heavy chain (by peptide digest)

QVTLRESGPALVKPTQTLTLTCTFSGFSLSTAGMSVGWIRQPPGKALEWLADIWWDDKKHYNPSLKDRL
TISKDTSKNQVVLKVTNMDPADTATYYCARDMIFNFYFDVWGQGTTVTVSSASTKGPSVFPLAPSSKSTS
GGTAALGCLVKDYFPEPVTVSWNSGALTSGVHTFPAVLQSSGLYSLSSVVTVPSSSLGTQTYICNVNHKP
SNTKVDKRVEPKSCDKTHTCPPCPAPELLGGPSVFLFPPKPKDTLMISRTPEVTCVVVDVSHEDPEVKFN
WYVDGVEVHNAKTKPREEQY**N**STYRVVSVLTVLHQDWLNGKEYKCKVSNKALPAPIEKTISKAKGQPRE
PQVYTLPPSREEMTKNQVSLTCLVKGFYPSDIAVEWESNGQPENNYKTTPPVLDSDGSFFLYSKLTVDKS
RWQQGNVFSCSVMHEALHNHYTQKSLSLSPG

N-glycosylation site

Light chain (by peptide digest)

DIQMTQSPSTLSASVGDRVTITCSASSRVGYMHWYQQKPGKAPKLLIYDTSKLASGVPSRFSGSGSGTE
FTLTISSLQPDDFATYYCFQGSGYPFTFGGGTKVEIKRTVAAPSVFIFPPSDEQLKSGTASVVCLLNNFYP
REAKVQWKVDNALQSGNSQESVTEQDSKDSTYSLSSTLTLSKADYEKHKVYACEVTHQGLSSPVTKSFN
RGEC

SI 1.3 PPI-03 (IgG1)

Heavy chain

QVNLRESGGGLVQPGGSLRLSCAASGFTFGSYAMSWVRQAPGKGLEWVSAISGSGGSTYYADSVKGR
FTISRDNSKNSLYLQMNSLRAEDTAVYYCARRSIYGGNYYFDYWGRGTLVTVSSASTKGPSVFPLAPSSK
STSGGTAALGCLVKDYFPEPVTVSWNSGALTSGVHTFPAVLQSSGLYSLSSVVTVPSSSLGTQTYICNVN
HKPSNTKVDKKVEPKSCDKTHTCPPCPAPELLGGPSVFLFPPKPKDTLMISRTPEVTCVVVDVSHEDPEV
KFNWYVDGVEVHNAKTKPREEQY**N**STYRVVSVLTVLHQDWLNGKEYKCKVSNKALPAPIEKTISKAKGQ
PREPQVYTLPPSRDELTKNQVSLTCLVKGFYPSDIAVEWESNGQPENNYKTTPPVLDSDGSFFLYSKLTV
DKSRWQQGNVFSCSVMHEALHNHYTQKSLSLSPGK

N-glycosylation site

Light chain

DIQMTQSPSSLSASVGDRVTITCRASQSISSYLNWYQQKPGKAPKLLIYAASSLQSGVPSRFSGSGSGTD
FTLTISSLQPEDFATYYCQQSYSTPLTFGGGSKVEIKRTVAAPSVFIFPPSDEQLKSGTASVVCLLNNFYPR
EAKVQWKVDNALQSGNSQESVTEQDSKDSTYSLSSTLTLSKADYEKHKVYACEVTHQGLSSPVTKSFNR
GEC

SI 1.4 PPI-04 (IgG1) - Sequence not available.

SI 1.5 PPI-08 (Bispecific mAb) - Sequence not available.

SI 1.6 PPI-10 (IgG1)

Heavy chain

EVQLLESGGGLVQPGGSLRLSCAASGFTFGNSWMSWVRQAPGKGLEWVSAISGSGGSTYYADSVKGR
FTISRDNSKNTLYLQMNSLRAEDTAVYYCTRDLPGIAVAGYWGQGTLVTVSSASTKGPSVFPLAPSSKST
SGGTAALGCLVKDYFPEPVTVSWNSGALTSGVHTFPAVLQSSGLYSLSSVVTVPSSSLGTQTYICNVNHK
PSNTKVDKKVEPKSCDKTHTCPPCPAPELLGGPSVFLFPPKPKDTLMISRTPEVTCVVVDVSHEDPEVKF
NWYVDGVEVHNAKTKPREEQY**N**STYRVVSVLTVLHQDWLNGKEYKCKVSNKALPAPIEKTISKAKGQP
REPQVYTLPPSRDELTKNQVSLTCLVKGFYPSDIAVEWESNGQPENNYKTTPPVLDSDGSFFLYSKLTVD
KSRWQQGNVFSCSVMHEALHNHYTQKSLSLSPGK

N-glycosylation site

Light chain

DTQMTQSPSTLSASVGDRVTITCRASEGIYHWLAWYQQKPGKAPKLLIYKASSLASGVPSRFSGSGSGT
EFTLTISSLQPDDFATYYCQQYSNYPLTFGGGTKLEIKRTVAAPSVFIFPPSDEQLKSGTASVVCLLNNFYP
REAKVQWKVDNALQSGNSQESVTEQDSKDSTYSLSSTLTLSKADYEKHKVYACEVTHQGLSSPVTKSFN
RGEC

SI 1.7 PPI-13 (IgG1)

Heavy chain

QVQLQESGPGLVKPSETLSLTCTVSGGSISADGYYWSWIRQPPGKGLEWIGSLYYSGSTYYNPSLKGRV
TISGDTSKNQFSLKLSSVTAADTAVYYCARTPAYFGQDRTDFFDVWGRGTLVTVSSASTKGPSVFPLAPS
SKSTSGGTAALGCLVKDYFPEPVTVSWNSGALTSGVHTFPAVLQSSGLYSLSSVVTVPSSSLGTQTYICN
VNHKPSNTKVDKRVEPKSCDKTHTCPPCPAPEFEGGPSVFLFPPKPKDTLMISRTPEVTCVVVDVSHED
PEVKFNWYVDGVEVHNAKTKPREEQY**N**STYRVVSVLTVLHQDWLNGKEYKCKVSNKALPASIEKTISKA
KGQPREPQVYTLPPSREEMTKNQVSLTCLVKGFYPSDIAVEWESNGQPENNYKTTPPVLDSDGSFFLYS
KLTVDKSRWQQGNVFSCSVMHEALHNHYTQKSLSLSPGK

N-glycosylation site

Light chain

DIQMTQSPSTLSASVGDRVTITCRASQGISSWLAWYQQKPGKAPKVLIYKASTLESGVPSRFSGSGSGTE
FTLTISSLQPDDFATYYCQQSHHPPWTFGQGTKLEIKRTVAAPSVFIFPPSDEQLKSGTASVVCLLNNFYP
REAKVQWKVDNALQSGNSQESVTEQDSKDSTYSLSSTLTLSKADYEKHKVYACEVTHQGLSSPVTKSFN
RGEC

SI 1.8 PPI-17 (IgG2)

Heavy chain (by peptide digest)

QVQLVESGGGLVKPGGSLRLSCAASGFTFSDYYMNWIRQAPGKGLEWVSYISSSGSIIYYADSVKGRFTI
SRDNAKNSLYLQMNSLRAEDTAVYYCAREGRIAARGMDVWGQGTTVTVSSASTKGPSVFPLAPCSRST
SESTAALGCLVKDYFPEPVTVSWNSGALTSGVHTFPAVLQSSGLYSLSSVVTVPSSNFGTQTYTCNVDH
KPSNTKVDKTVERKCCVECPPCPAPPVAGPSVFLFPPKPKDTLMISRTPEVTCVVVDVSHEDPEVQFNW
YVDGVETKPREEQFNSTFREEQFNSTFRVVSVLTVVHQDWLNGKEYKCKGLPAPIEKTISKTKGQPREP
QVYTLPPSREEMTKNQVSLTCLVKGFYPSDIAVEWESNGQPENNYKTTPPMLDSDGSFFLYSKLTVDKS
RWQQGNVFSCSVMHEALHNHYTQKSLSLSPGK

Light chain (by peptide digest)

DIQMTQSPSSLSASVGDRVTITCRPSQSFSRYINWYQQKPGKAPKLLIYAASSLVGGVPSRFSGSGSGTD
FTLTISSLQPEDFATYYCQQTYSNPPITFGQGTRLEIKRTVAAPSVFIFPPSDEQLKSGTASVVCLLNNFYP
REAKVQWKVDNALQSGNSQESVTEQDSKDSTYSLSSTLTLSKADYEKHKVYACEVTHQGLSSPVTKSFN
RGEC

SI 1.9 PPI-18 (human serum albumin-neprilysin fusion protein)

DAHKSEVAHRFKDLGEENFKALVLIAFAQYLQQSPFEDHVKLVNEVTEFAKTCVADESAENCDKSLHTLF
GDKLCTVATLRETYGEMADCCAKQEPERNECFLQHKDDNPNLPRLVRPEVDVMCTAFHDNEETFLKKYL
YEIARRHPYFYAPELLFFAKRYKAAFTECCQAADKAACLLPKLDELRDEGKASSAKQRLKCASLQKFGER
AFKAWAVARLSQRFPKAEFAEVSKLVTDLTKVHTECCHGDLLECADDRADLAKYICENQDSISSKLKECC
EKPLLEKSHCIAEVENDEMPADLPSLAADFVESKDVCKNYAEAKDVFLGMFLYEYARRHPDYSVVLLLRL
AKTYETTLEKCCAAADPHECYAKVFDEFKPLVEEPQNLIKQNCELFEQLGEYKFQNALLVRYTKKVPQVS
TPTLVEVSRNLGKVGSKCCKHPEAKRMPCAEDYLSVVLNQLCVLHEKTPVSDRVTKCCTESLVNRRPCF
SALEVDETYVPKEFNAETFTFHADICTLSEKERQIKKQTALVELVKHKPKATKEQLKAVMDDFAAFVEKCC
KADDKETCFAEEGKKLVAASQAALGL**GGGG**SYDDGICKSSDCIKSAARLIQNMDATTEPCTDFFKYAC

GGWLKRNVIPETSSRYGNFDILRDELEVVLKDVLQEPKTEDIVAVQKAKALYRSCINESAIDSRGGEPLLKL
LPDIYGWPVATENWEQKYGASWTAEKAIAQLNSKYGKKVLINLFVGTDDKNSVNHVIHIDQPRLGLPSRD
YYECTGIYKEACTAYVDFMISVARLIRQEERLPIDENQLALEMNKVMELEKEIANATAKPEDRNDPMLLYN
KMTLAQIQNNFSLEINGKPFSWLNFTNEIMSTVNISITNEEDVVVYAPEYLTKLKPILTKYSARDLQNLMSW
RFIMDLVSSLSRTYKESRNAFRKALYVTTSETATWRRCANYVNGNMENAVGRLYVEAAFAGESKHVVED
LIAQIREVFIQTLDDLTWMDAETKKRAEEKALAIKERIGYPDDIVSNDNKLNNEYLELNYKEDEYFENIIQNL
KFSQSKQLKKLREKVDKDEWISGAAVVNAFYSSGRNQIVFPAGILQPPFFSAQQSNSLNYGGIGMVIGHEI
THGFDDNGRNFNKDGDLVDWWTQQSASNFKEQSQCMVYQYGNFSWDLAGGQHLNGINTLGENIADN
GGLGQAYRAYQNYIKKNGEEKLLPGLDLNHKQLFFLNFAQVWCGTYRPEYAVNSIKTDVHSPKNFRIIGTL
QNSAEFSEAFHCRKNSYMNPEKKCRVW

GGGG-Linker

SI 1.10 PPI-30 (interferon alfa-2a)

CDLPQTHSLGSRRTLMLLAQMRKISLFSCLKDRHDFGFPQEEFGNQFQKAETIPVLHEMIQQIFNLFSTKD
SSAAWDETLLDKFYTELYQQLNDLEACVIQGVGVTETPLMKEDSILAVRKYFQRITLYLKEKKYSPCAWEV
VRAEIMRSFSLSTNLQESLRSKE

SI 1.11 PPI-44 (recombinant human transferrin)

VPDKTVRWCAVSEHEATKCQSFRDHMKSVIPSDGPSVACVKKASYLDCIRAIAANEADAVTLDAGLVYDA
YLAPNNLKPVVAEFYGSKEDPQTFYYAVAVVKKDSGFQMNQLRGKKSCHTGLGRSAGWNIPIGLLYCDL
PEPRKPLEKAVANFFSGSCAPCADGTDFPQLCQLCPGCGCSTLNQYFGYSGAFKCLKDGAGDVAFVKH
STIFENLANKADRDQYELLCLDNTRKPVDEYKDCHLAQVPSHTVVARSMGGKEDLIWELLNQAQEHFGK
DKSKEFQLFSSPHGKDLLFKDSAHGFLKVPPRMDAKMYLGYEYVTAIRNLREGTCPEAPTDECKPVKWC
ALSHHERLKCDEWSVNSVGKIECVSAETTEDCIAKIMNGEADAMSLDGGFVYIAGKCGLVPVLAENYNKA
DNCEDTPEAGYFAVAVVKKSASDLTWDNLKGKKSCHTAVGRTAGWNIPMGLLYNKINHCRFDEFFSEGC
APGSKKDSSLCKLCMGSGLNLCEPNNKEGYYGYTGAFRCLVEKGDVAFVKHQTVPQNTGGKNPDPWA
KNLNEKDYELLCLDGTRKPVEEYANCHLARAPNHAVVTRKDKEACVHKILRQQQHLFGSNVADCSGNFC
LFRSETKDLLFRDDTVCLAKLHDRNTYEKYLGEEYVKAVGNLRKCSTSSLLEACTFRRP

SI 1.12 PPI-45 (Lipase)

SIDGGIRAATSQEINELTYYTTLSANSYCRTVIPGATWDCIHCDATEDLKIIKTWSTLIYDTNAMVARGDSEK
TIYIVFRGSSSIRNWIADLTFVPVSYPPVSGTKVHKGFLDSYGEVQNELVATVLDQFKQYPSYKVAVTGHS

LGGATALLCALDLYQREEGLSSSNLFLYTQGQPRVGNPAFANYVVSTGIPYRRTVNERDIVPHLPPAAFG
FLHAGSEYWITDNSPETVQVCTSDLETSDCSNSIVPFTSVLDHLSYFGINTGLCT

SI 1.13 PPI-46 (Lipase)

EVSQDLFNQFNLFAQYSAAAYCGKNNDAPAGTNITCTGNACPEVEKADATFLYSFEDSGVGDVTGFLAL
DNTNKLIVLSFRGSRSIENWIGNLNFDLKEINDICSGCRGHDGFTSSWRSVADTLRQKVEDAVREHPDYR
VVFTGHSLGGALATVAGADLRGNGYDIDVFSYGAPRVGNRAFAEFLTVQTGGTLYRITHTNDIVPRLPPR
EFGYSHSSPEYWIKSGTLVPVTRNDIVKIEGIDATGGNNQPNIPDIPAHLWYFGLIGTCL

SI 1.14 PPI-49 (human serum albumin)

MKWVTFISLLFLFSSAYSRGVFRRDAHKSEVAHRFKDLGEENFKALVLIAFAQYLQQCPFEDHVKLVNEV
TEFAKTCVADESAENCDKSLHTLFGDKLCTVATLRETYGEMADCCAKQEPERNECFLQHKDDNPNLPRL
VRPEVDVMCTAFHDNEETFLKKYLYEIARRHPYFYAPELLFFAKRYKAAFTECCQAADKAACLLPKLDELR
DEGKASSAKQRLKCASLQKFGERAFKAWAVARLSQRFPKAEFAEVSKLVTDLTKVHTECCHGDLLECAD
DRADLAKYICENQDSISSKLKECCEKPLLEKSHCIAEVENDEMPADLPSLAADFVESKDVCKNYAEAKDVF
LGMFLYEYARRHPDYSVVLLLRLAKTYETTLEKCCAAADPHECYAKVFDEFKPLVEEPQNLIKQNCELFE
QLGEYKFQNALLVRYTKKVPQVSTPTLVEVSRNLGKVGSKCCKHPEAKRMPCAEDYLSVVLNQLCVLHE
KTPVSDRVTKCCTESLVNRRPCFSALEVDETYVPKEFNAETFTFHADICTLSEKERQIKKQTALVELVKHK
PKATKEQLKAVMDDFAAFVEKCCKADDKETCFAEEGKKLVAASQAALGL

Table SI 2. Zscore values for the homology model structures

Protein name	Amino acids number	Zscore
PPI-01	451	-9.38
PPI-02	898	-12.66
PPI-03	902	-11.08
PPI-10	898	-13.03
PPI-13	910	-13.44
PPI-17	892	-12.53
PPI-18	1289	-12.43
PPI-30	165	-6.42
PPI-44	Missing	Missing
PPI-45	269	-6.89
PPI-46	269	-8.47
PPI-49	609	-11.25

SI 3. Details on the calculation used for extrinsic DSF

The raw fluorescence intensity data at 578 nm were exported for data processing using Origin 8® SR6 (OriginLab Corportation; Northampton, MA). Background fluorescence was corrected by subtracting the corresponding placebo curves from each sample curve. The melting curves were then differentiated, smoothed (polynomial order = 1, number of points = 5), and splined (cubic spline with 99 interpolated points between two data points). Peak centers of the resulting first derivative were used as melting temperatures. To determine the onset temperature ($T_{on,ext}$), the curves were fitted using Origin's Boltzmann function where the local minimum and maximum of the fluorescence transition were used as input values. The $T_{on,ext}$ was calculated using the equation below:

$$T_{on,ext} = \frac{8dT\ (A_1 - \frac{A_2 + A_1}{2} + \frac{T_{m1,ext}\ (A_2 - A_1)}{4dT})}{A_2 - A_1} - T_{m1,ext}$$

with $T_{m1,ext}$ being the inflection point, dT the slope factor, A1 the lower fluorescence intensity, and A2 the upper fluorescence intensity of the sigmoidal curve. $T_{on,ext}$ resembles twice the distance on the x-scale of $T_{m1,ext}$, and the point of intersection between the tangents through $T_{m1,ext}$ and A1.

For the high throughput screening, all the previous Origin® data processing steps were automated using the Open TM script, which applies the LabTalk code function that can be found in the following reference: Menzen, T. A. Temperature-Induced Unfolding, Aggregation, and Interaction of Therapeutic Monoclonal Antibodies. *PhD Thesis, LMU Munich* (2014).

Figure SI 4. Fitting from the response surface methodology (RSM)

Note: Each protein is represented with a different color. The R^2 and the root mean square error (RMSE) are calculated from all the datapoints, whose number is showed in the picture (n).

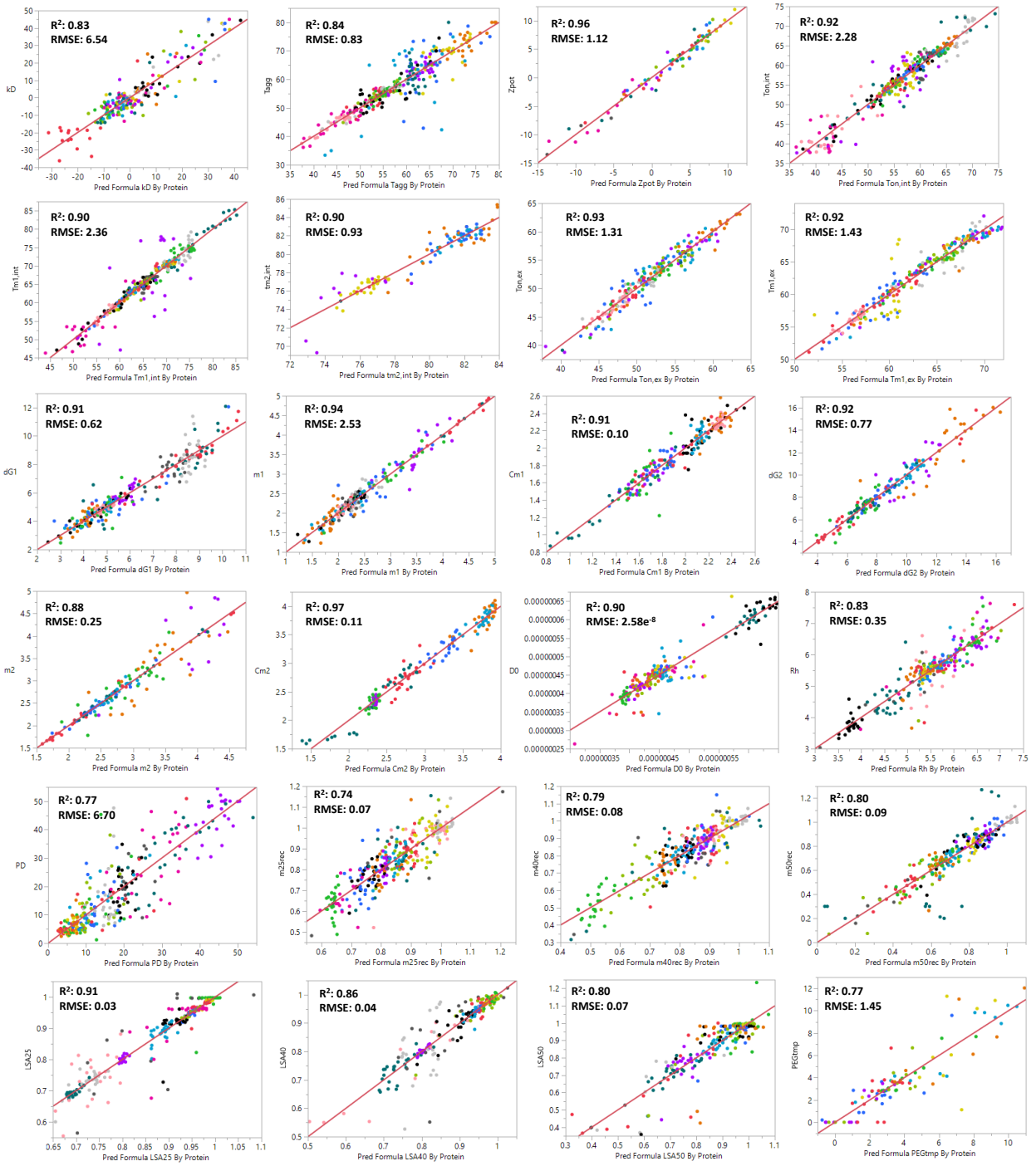


Figure SI 5. Multivariate matrixes including all datapoints. (Part 1 of 2)

Note: Each protein is represented with a different color. The two matrixes are relative to Figure 1 in the main manuscript (i.e. outliners are excluded).

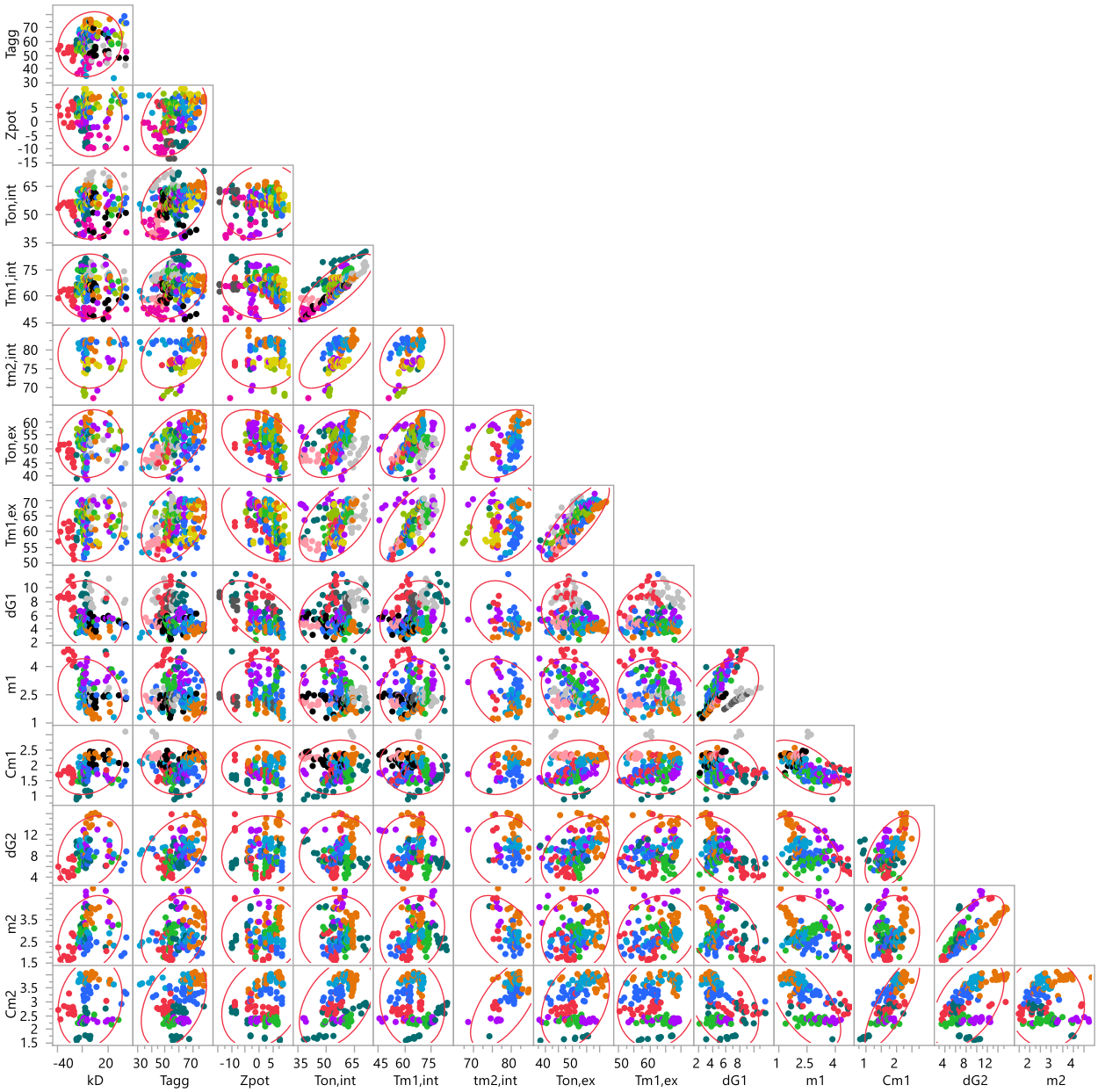


Figure SI 5. Multivariate matrixes including all datapoints. (Part 2 of 2)

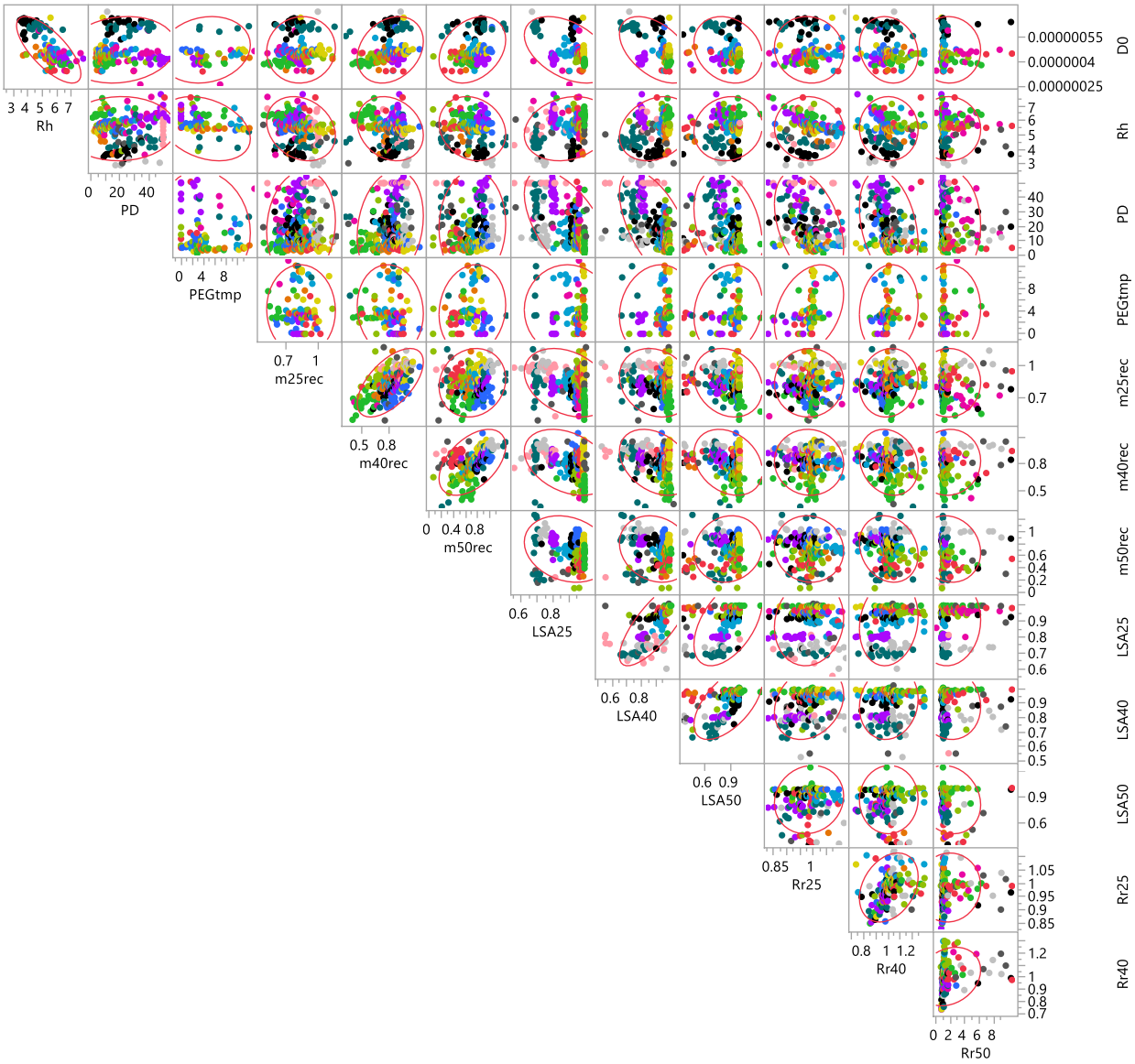


Figure SI 6. Separations obtained with cIEF and SEC-MALS for the proteins in the dataset (Part 1 of 2)

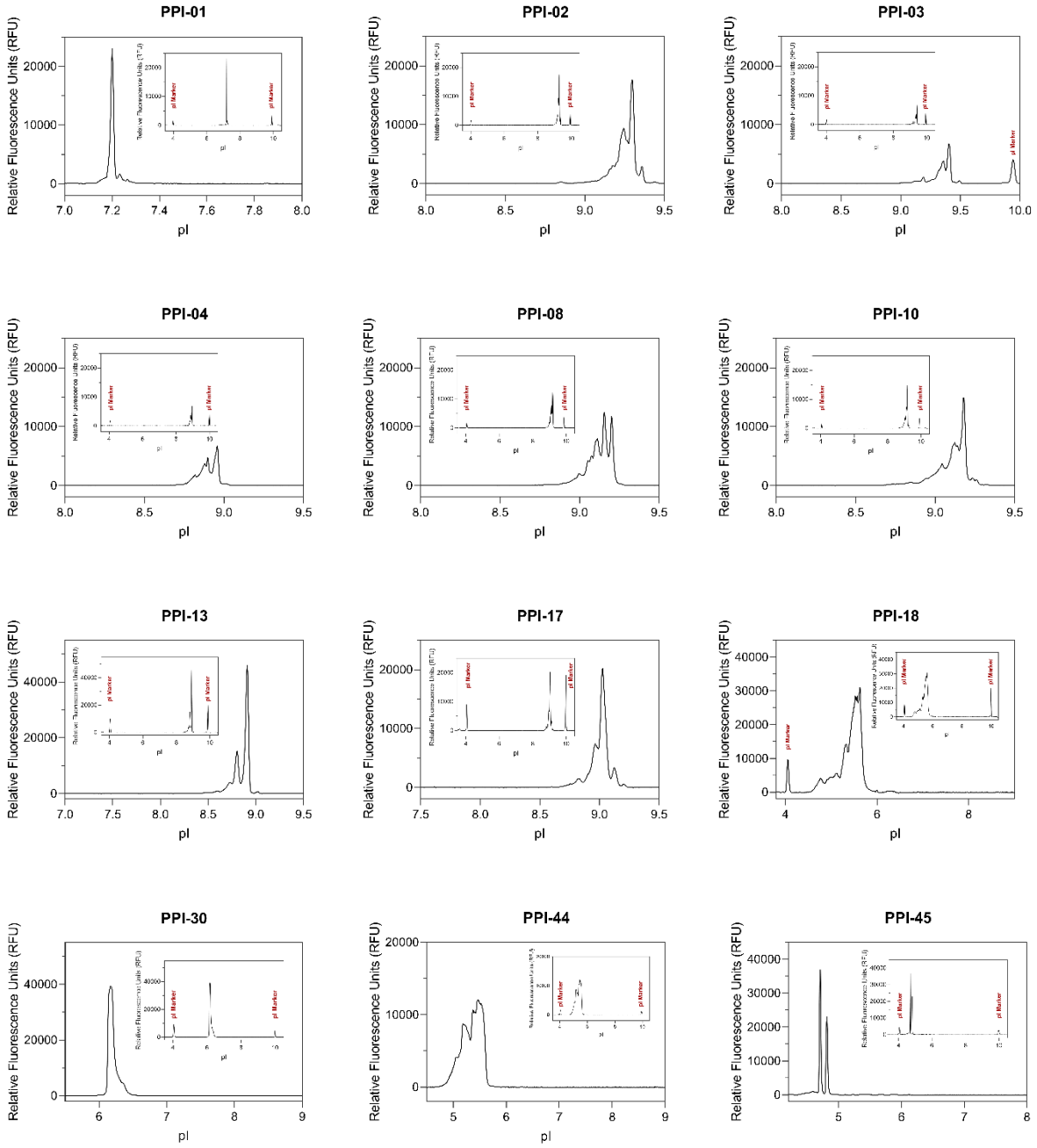


Figure SI 6. Separations obtained with cIEF and SEC-MALS for the proteins in the dataset (Part 2 of 2)

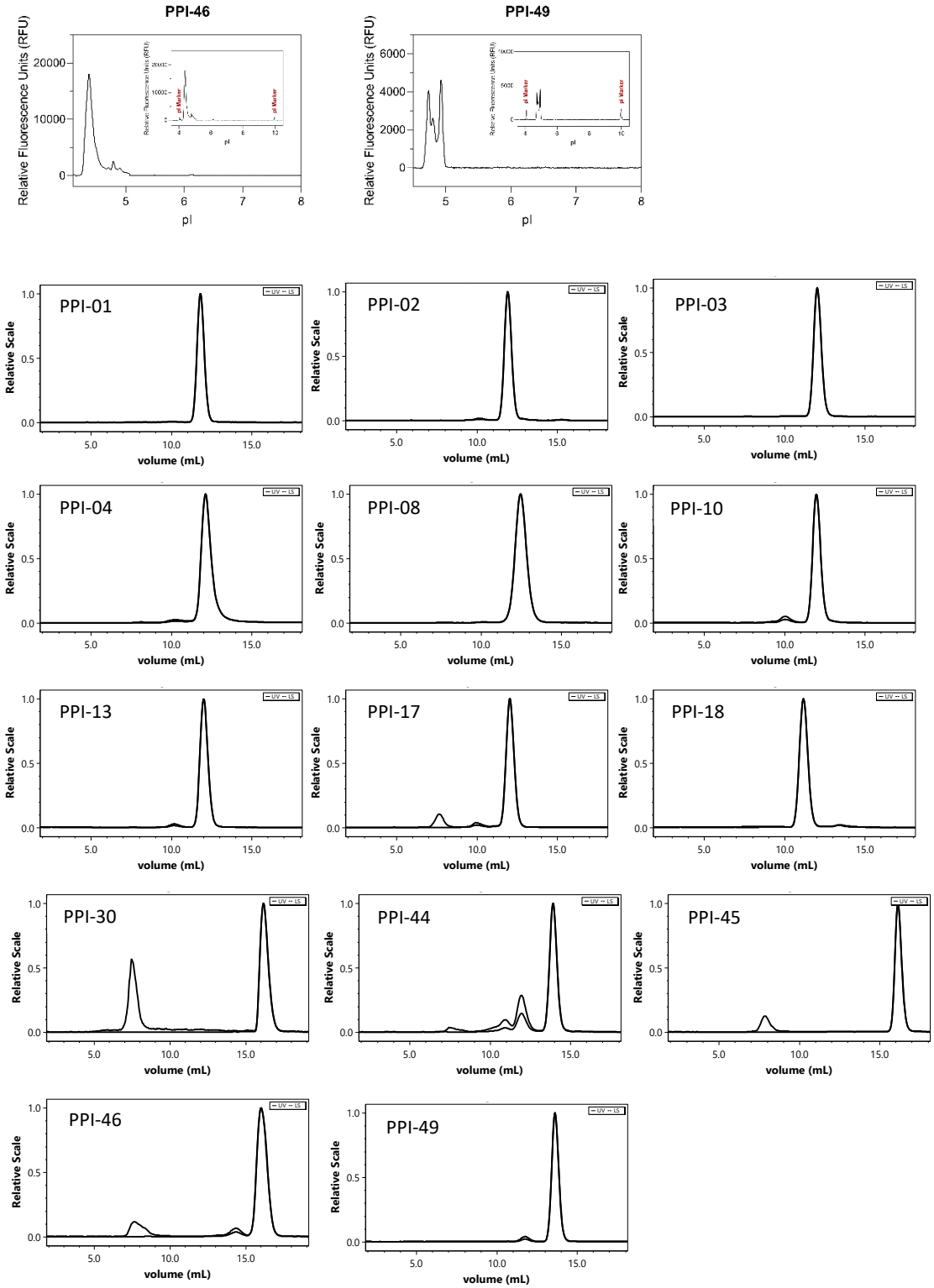


Table SI 7. List of the molecular descriptors calculated by MOE and ProDCal (Part 1 of 2)

Note: The description of the molecular indices can be found in the relative software manual.

MOE			ProtDCal	
pro_app_charge	pro_patch_cdr_pos_2	pro_zdipole	dGc(F)	wPjiH
pro_asa_hph	pro_patch_cdr_pos_3	pro_zeta	dGw(F)	wPhiS
pro_asa_hyd	pro_patch_cdr_pos_4	pro_zquadrupole	Gs(F)	wPhil
pro_asa_vdw	pro_patch_cdr_pos_5	pro_patch_neg_3	W(F)	Phi
pro_coeff_280	pro_patch_cdr_pos_n	pro_patch_neg_4	HBd	LnFD
pro_coeff_diff	pro_patch_hyd	pro_patch_neg_5	dGs	wCLQ
pro_coeff_fric	pro_patch_hyd_1	pro_patch_neg_n	dGw	wCTP
pro_debye	pro_patch_hyd_2	pro_patch_pos	dGel	wSP
pro_dipole_moment	pro_patch_hyd_3	pro_patch_pos_1	dGLJ	WNc
pro_eccen	pro_patch_hyd_4	pro_patch_pos_2	dGtor	Ap
pro_helicity	pro_patch_hyd_5	pro_patch_pos_3	Gs(U)	dA
pro_henry	pro_patch_hyd_n	pro_patch_pos_4	Gw(U)	dAnp
pro_hyd_moment	pro_patch_ion	pro_patch_pos_5	W(U)	WNLC
pro_mass	pro_patch_ion_1	pro_patch_pos_n	Mw	wFLC
pro_mobility	pro_patch_ion_2	pro_pI_3D	Ap	wR2
pro_net_charge	pro_patch_ion_3	pro_pI_seq	Ecl	lnFD
pro_patch_cdr_hyd	pro_patch_ion_4	pro_r_gyr	HP	Pb
pro_patch_cdr_hyd_1	pro_patch_ion_5	pro_r_solv	IP	Pa
pro_patch_cdr_hyd_2	pro_patch_ion_n	pro_sed_const	ISA	Pt
		pro_volume	Pa	z1

Table SI 7. List of the molecular descriptors calculated by MOE and ProDCal (Part 2 of 2)

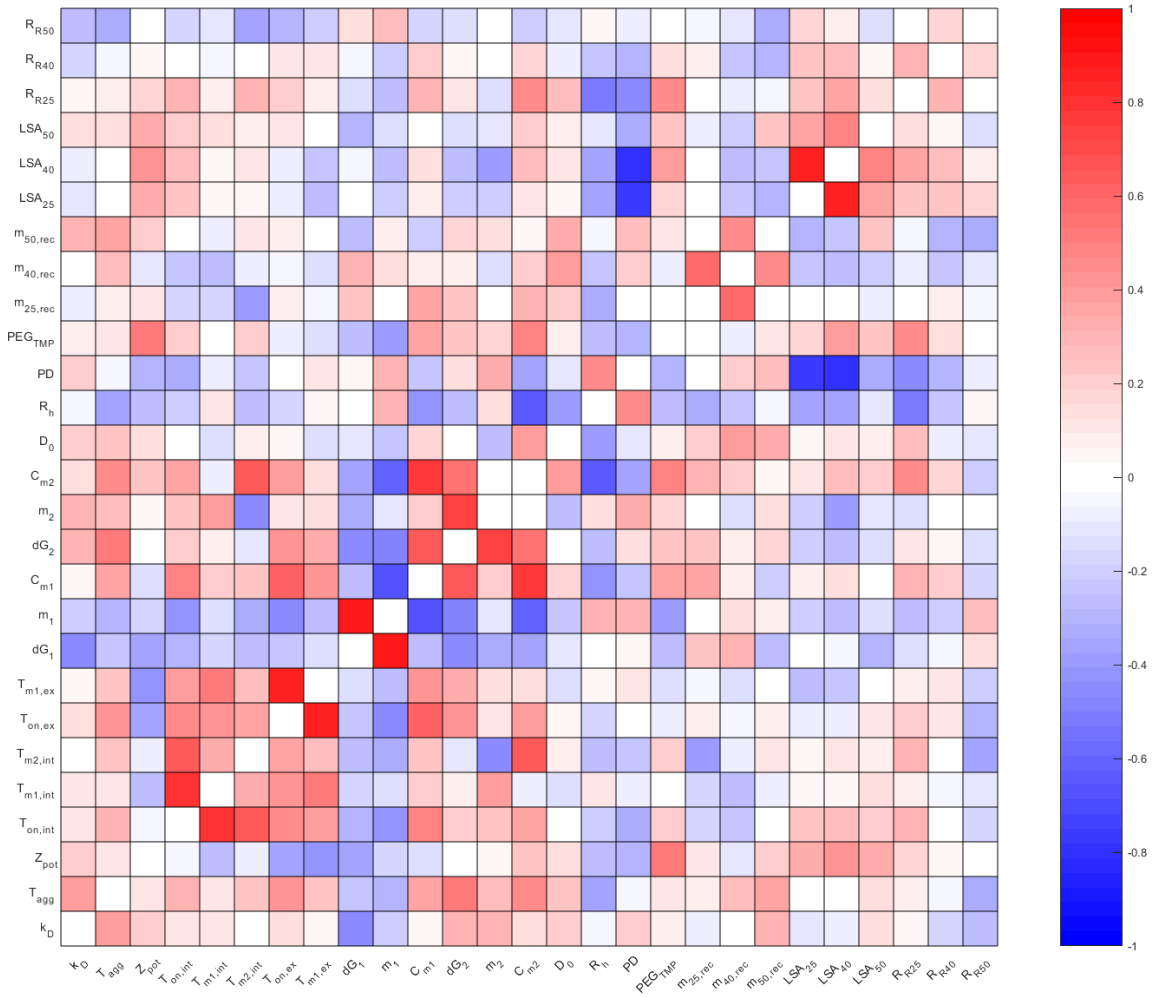
MOE	ProtDCal
pro_patch_cdr_hyd_4	z2
pro_patch_cdr_hyd_5	z3
pro_patch_cdr_hyd_n	dHf
pro_patch_cdr_ion	Xi
pro_patch_cdr_ion_1	L1-9
pro_patch_cdr_ion_2	wRWCO
pro_patch_cdr_ion_3	wdHBd
pro_patch_cdr_ion_4	wLCO
pro_patch_cdr_ion_5	wCo
pro_patch_cdr_ion_n	wFLC
pro_patch_cdr_neg	wPsiH
pro_patch_cdr_neg_1	wPsiS
pro_patch_cdr_neg_2	wPSil
pro_patch_cdr_neg_3	Psi
pro_patch_cdr_neg_4	wR2
pro_patch_cdr_neg_5	
pro_patch_cdr_neg_n	
pro_patch_cdr_pos	
pro_patch_cdr_pos_1	
pro_patch_neg_1	
pro_patch_neg_2	

Table SI 8. Parameters from Schrödinger

Protein	Total hydrophobic patch score	Mean AggScore	Mean Aggrescan	Mean Zyggregator
PPI-01	7511.6	1.14	-0.48	-0.11
PPI-02	4281.3	0.68	-0.46	-0.11
PPI-03	5493.3	0.82	-0.48	-0.13
PPI-10	5074.9	0.86	-0.48	-0.12
PPI-13	5647	0.78	-0.48	-0.12
PPI-17	4488.2	0.69	-0.46	-0.11
PPI-18	8699.9	0.98	-0.41	-0.12
PPI-30	966	1.35	-0.46	-0.01
PPI-44	1953.063	0.39	-0.18	-0.56
PPI-45	1842.4	2.33	-0.39	0.06
PPI-46	1230.3	1.56	-0.62	-0.08
PPI-49	4211.6	1.16	-0.42	-0.1

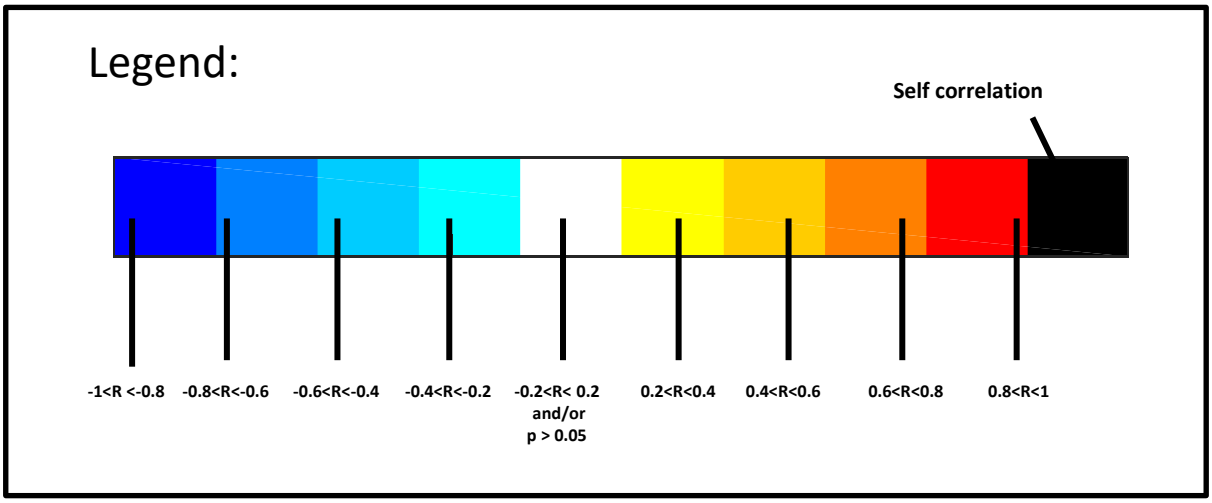
Figure SI 9. Pairwise correlations among biophysical parameters in a subset including only mAbs.

Note: The strength of these correlations was assessed using t-tests. p-values <0.05 are statistically significant at a 95 % confidence level. White cells represent 0 correlation of a p-value higher than 0.05. Blue and red cells express negative or positive correlation, respectively. No or only weak correlations were observed between most biophysical properties.



Figures SI 10. Pairwise correlations among biophysical parameters in subsets including different proteins. (Part 1 of 9)

Note: The strength of these correlations was assessed using t-tests. p-values <0.05 are statistically significant at a 95% confidence level. To allow a simpler visualization the cells are colored as showed in the legend. Below each figure description the subset restrictions are listed and separated by a comma, e.g. mAb, pharmaceutical relevant pH (5-7.5), low ionic strength (0 mM NaCl) indicates that the picture includes data only of mAb formulated in the pH range of 5-7.5 with no addition of NaCl.



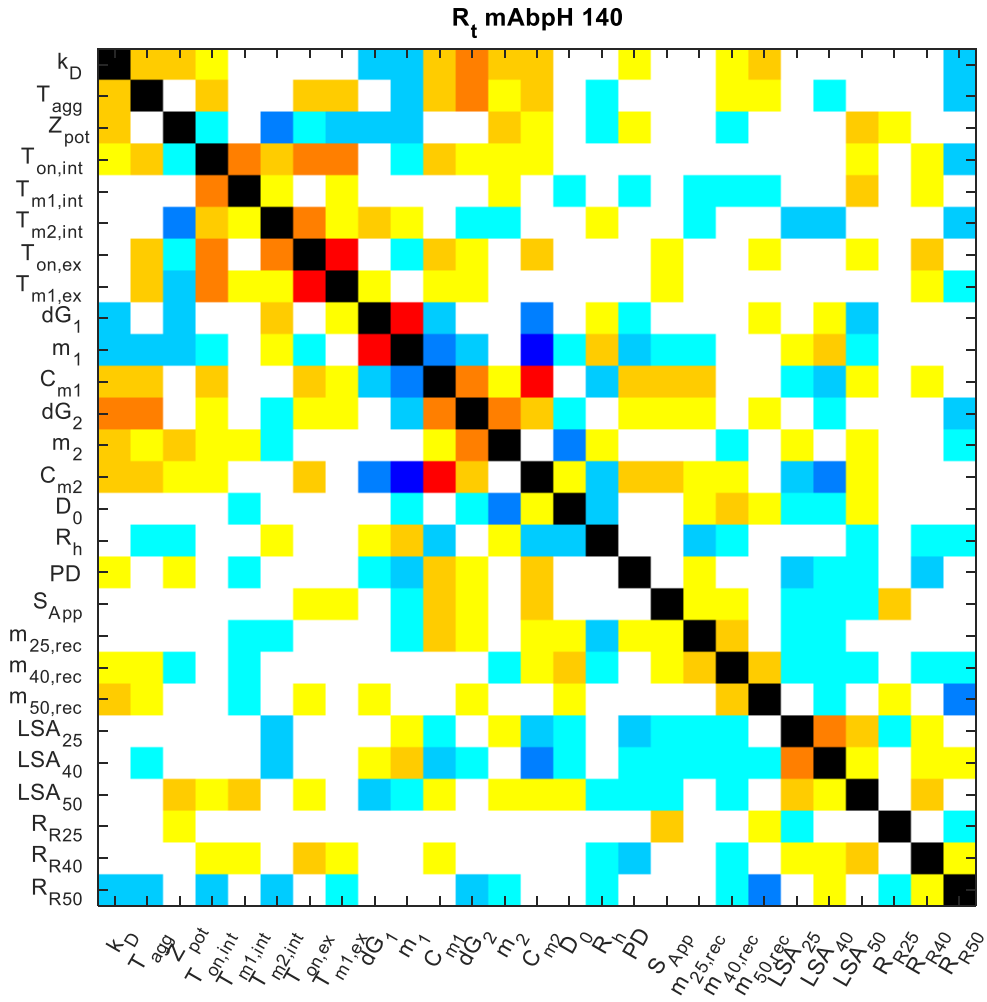


Figure SI 10. Pairwise correlations among biophysical parameters in subsets including different proteins. (Part 2 of 9). *Subset including only mAbs and formulations with pharmaceutical relevant pH values (5.0 - 7.5) and high ionic strength (140 mM NaCl):*

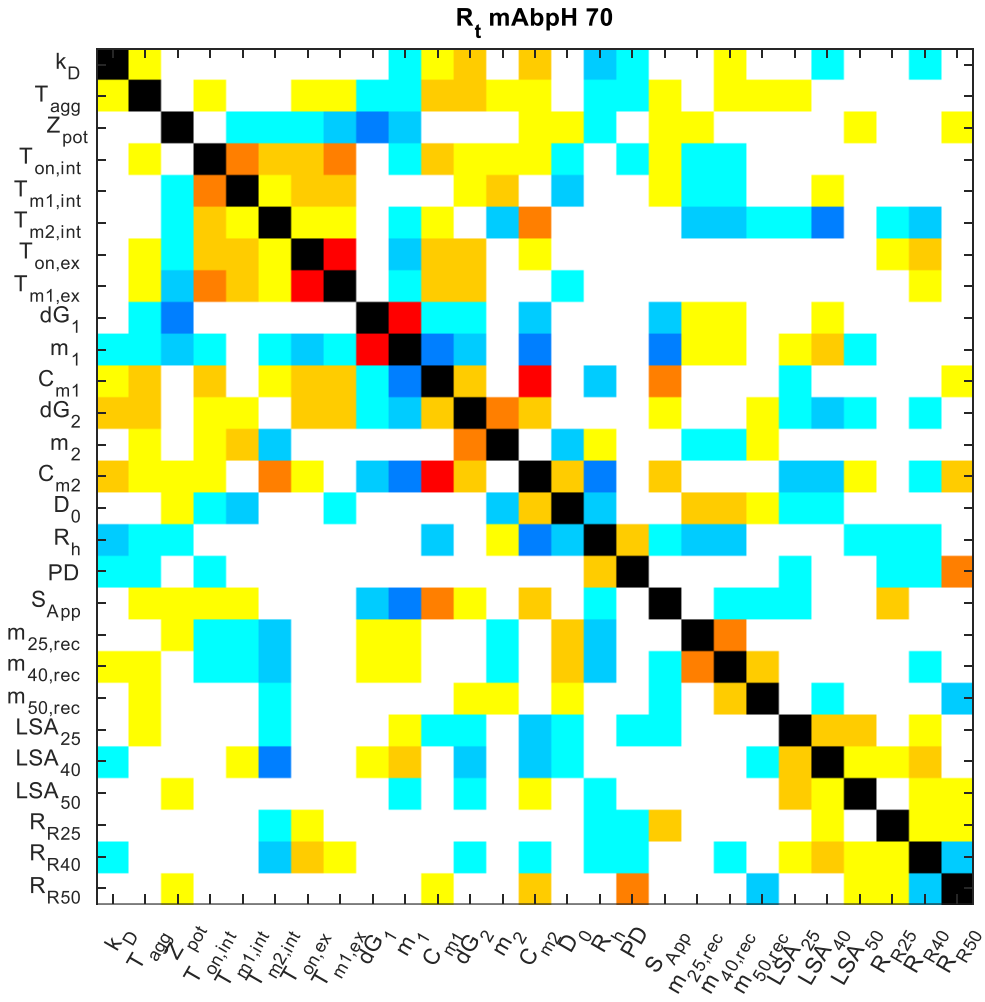


Figure SI 10. Pairwise correlations among biophysical parameters in subsets including different proteins. (Part 3 of 9). *Subset including only mAbs and formulations with pharmaceutical relevant pH values (5.0 - 7.5) and medium ionic strength (70 mM NaCl):*

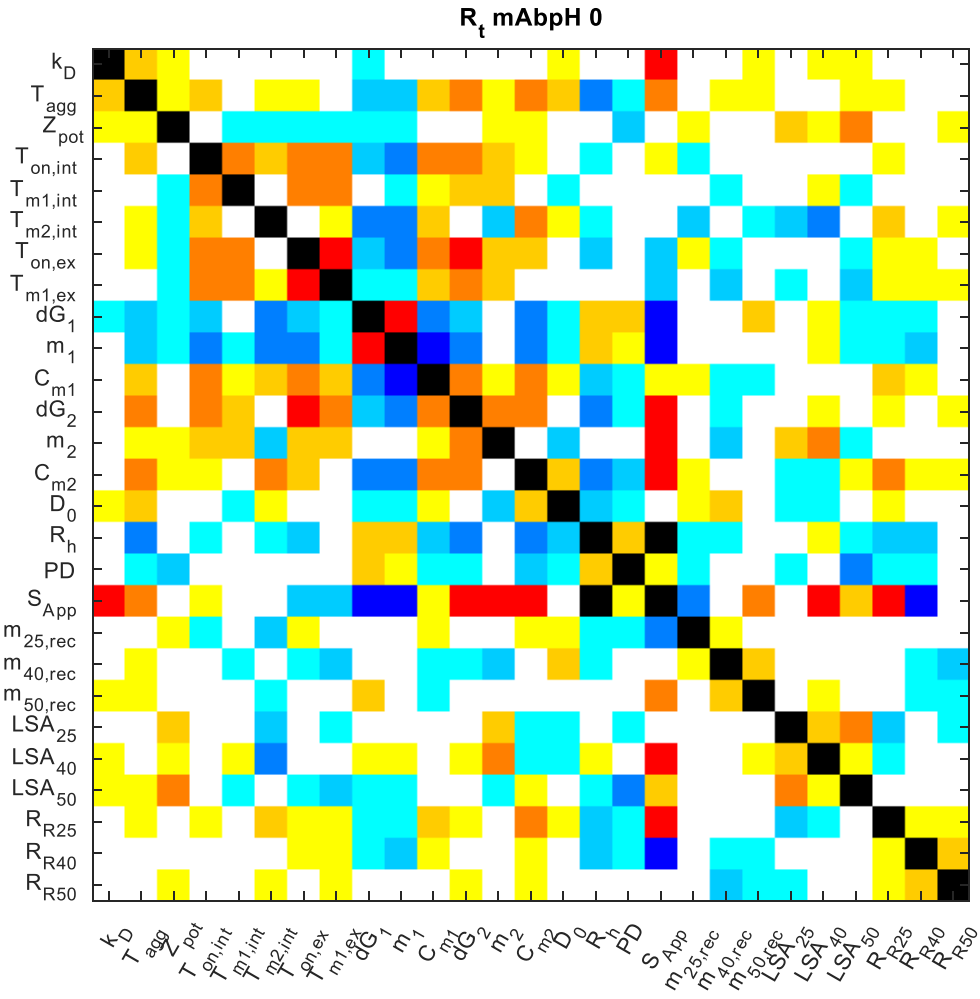


Figure SI 10. Pairwise correlations among biophysical parameters in subsets including different proteins. (Part 4 of 9). *Subset including only mAbs and formulations with pharmaceutical relevant pH values (5.0 - 7.5) and low ionic strength (0 mM NaCl):*

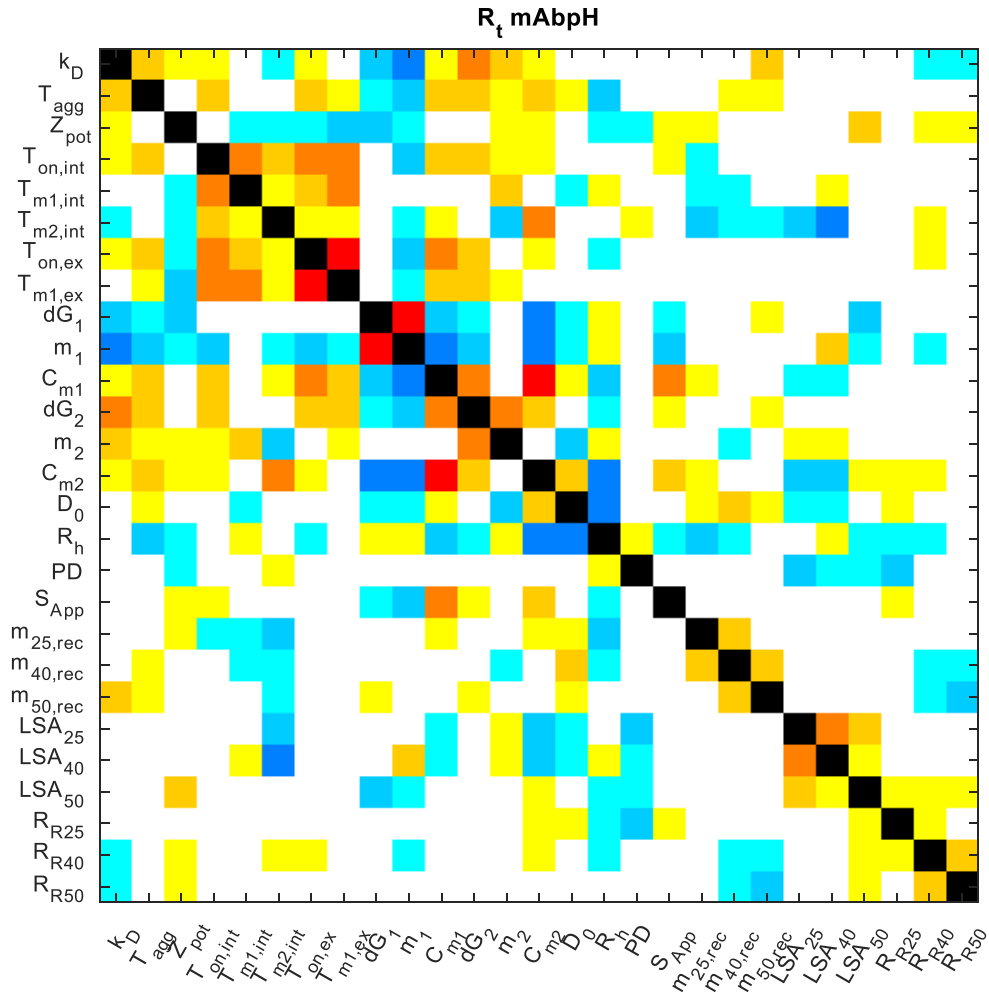


Figure SI 10. Pairwise correlations among biophysical parameters in subsets including different proteins. (Part 5 of 9). *Subset including only mAbs and formulations with pharmaceutical relevant pH values (5.0 - 7.5) with all three levels of ionic strength (0, 70 and 140 mM NaCl):*

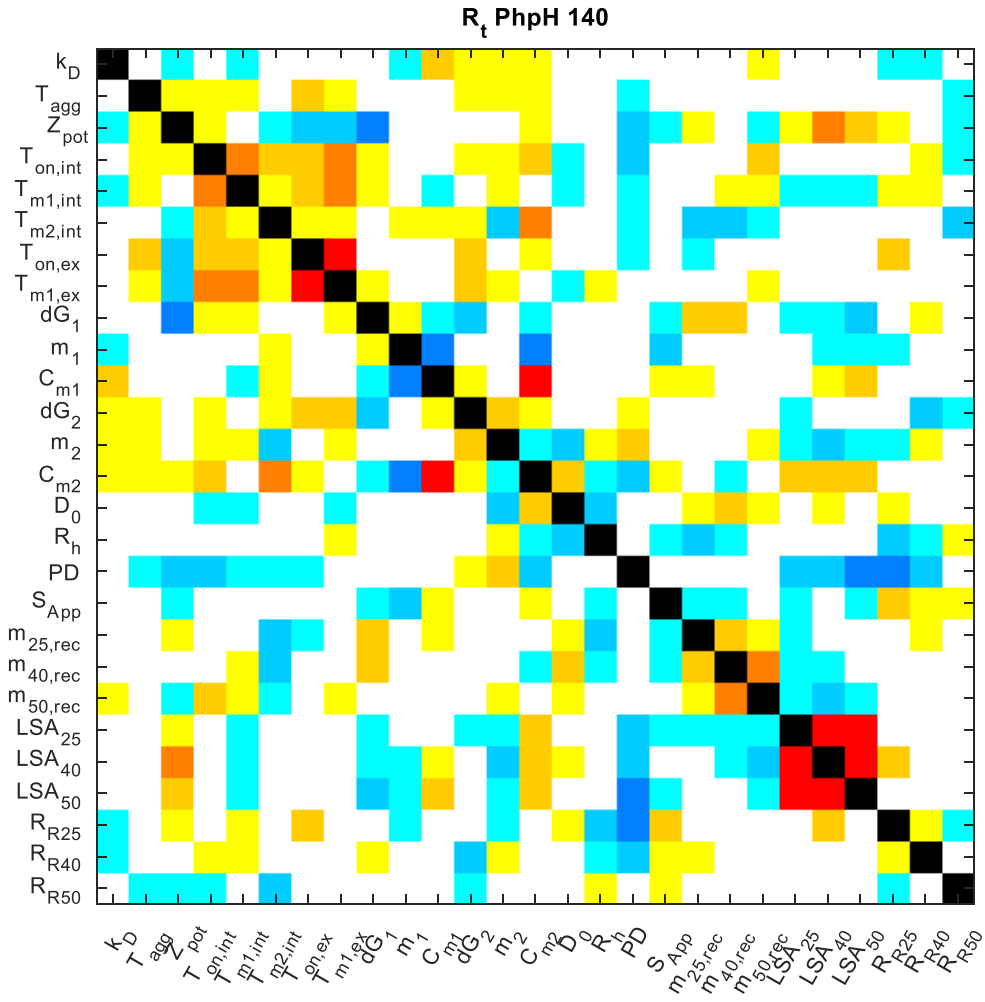


Figure SI 10. Pairwise correlations among biophysical parameters in subsets including different proteins. (Part 6 of 9). *Subset including all 14 proteins and formulations with pharmaceutical relevant pH values (5.0 - 7.5) and high ionic strength (140 mM NaCl):*

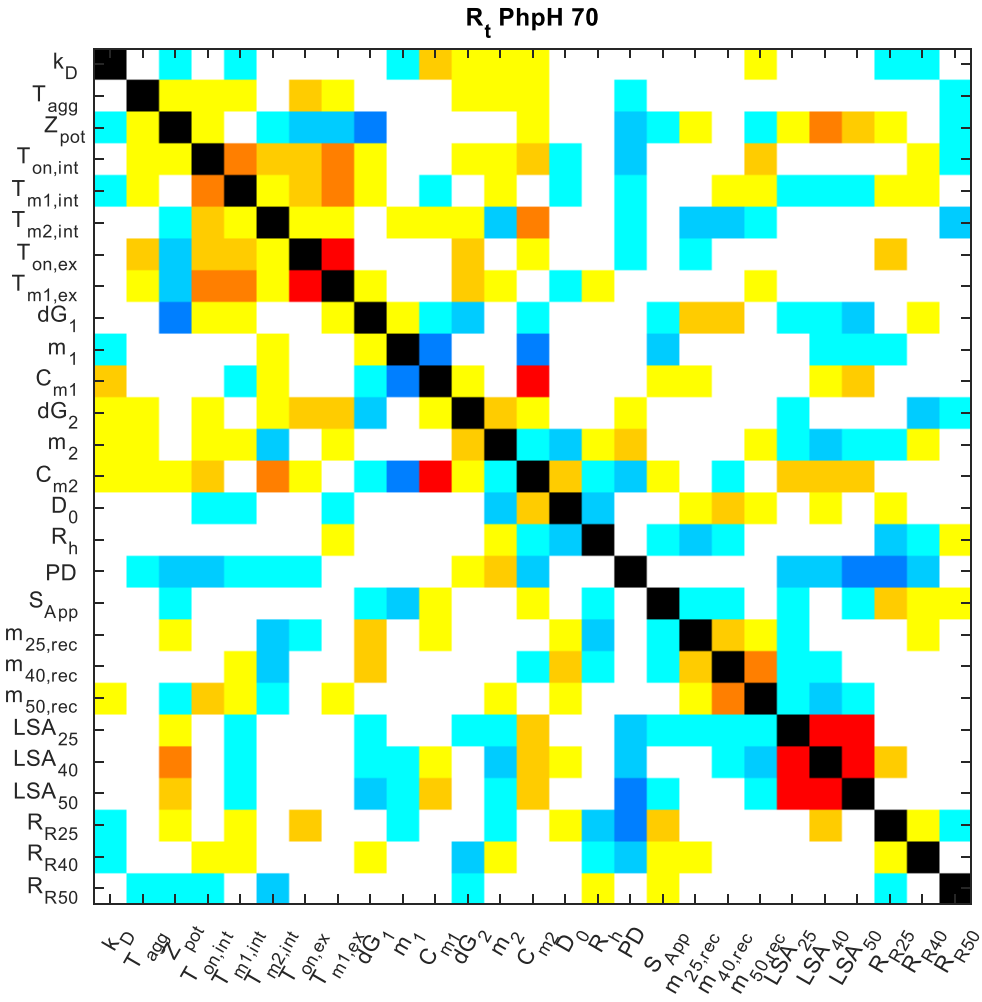


Figure SI 10. Pairwise correlations among biophysical parameters in subsets including different proteins. (Part 7 of 9). *Subset including all 14 proteins and formulations with pharmaceutical relevant pH values (5.0 - 7.5) and medium ionic strength (70 mM NaCl):*

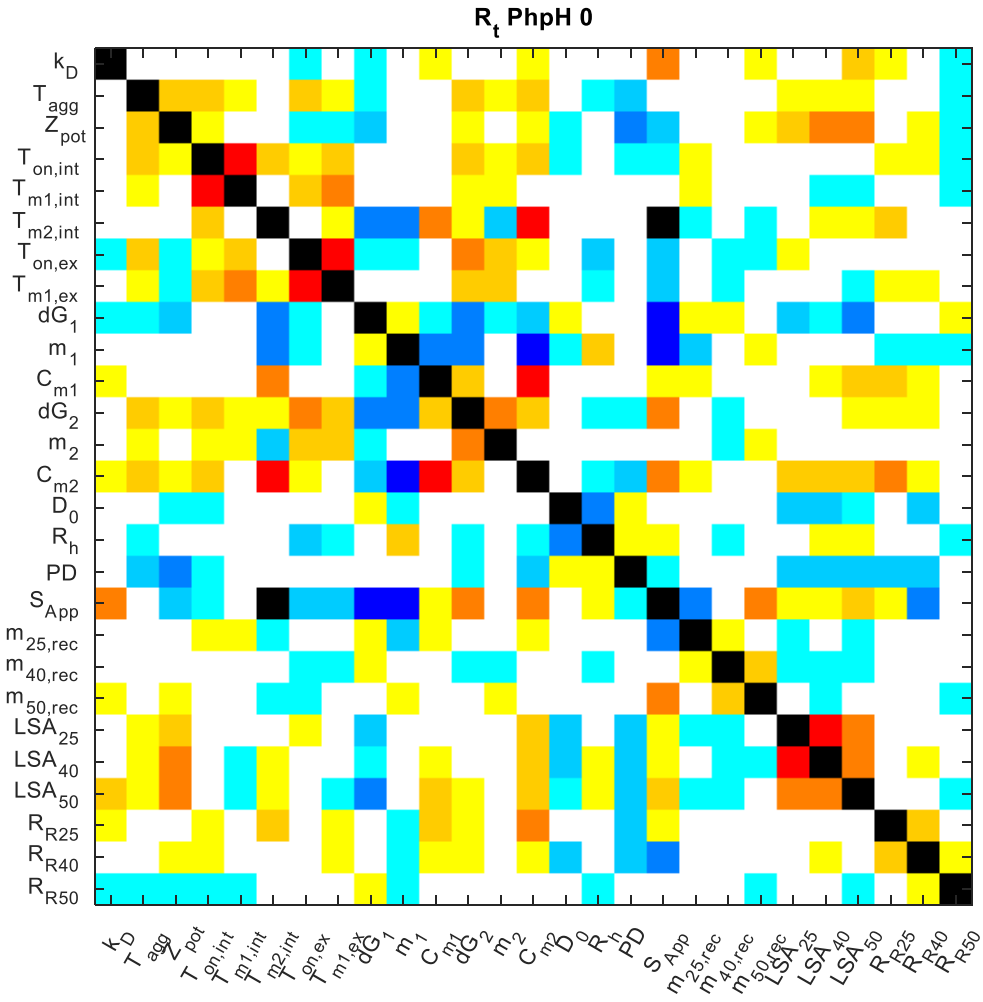


Figure SI 10. Pairwise correlations among biophysical parameters in subsets including different proteins. (Part 8 of 9). *Subset including all 14 proteins and formulations with pharmaceutical relevant pH values (5.0 - 7.5) and low ionic strength (0 mM NaCl):*

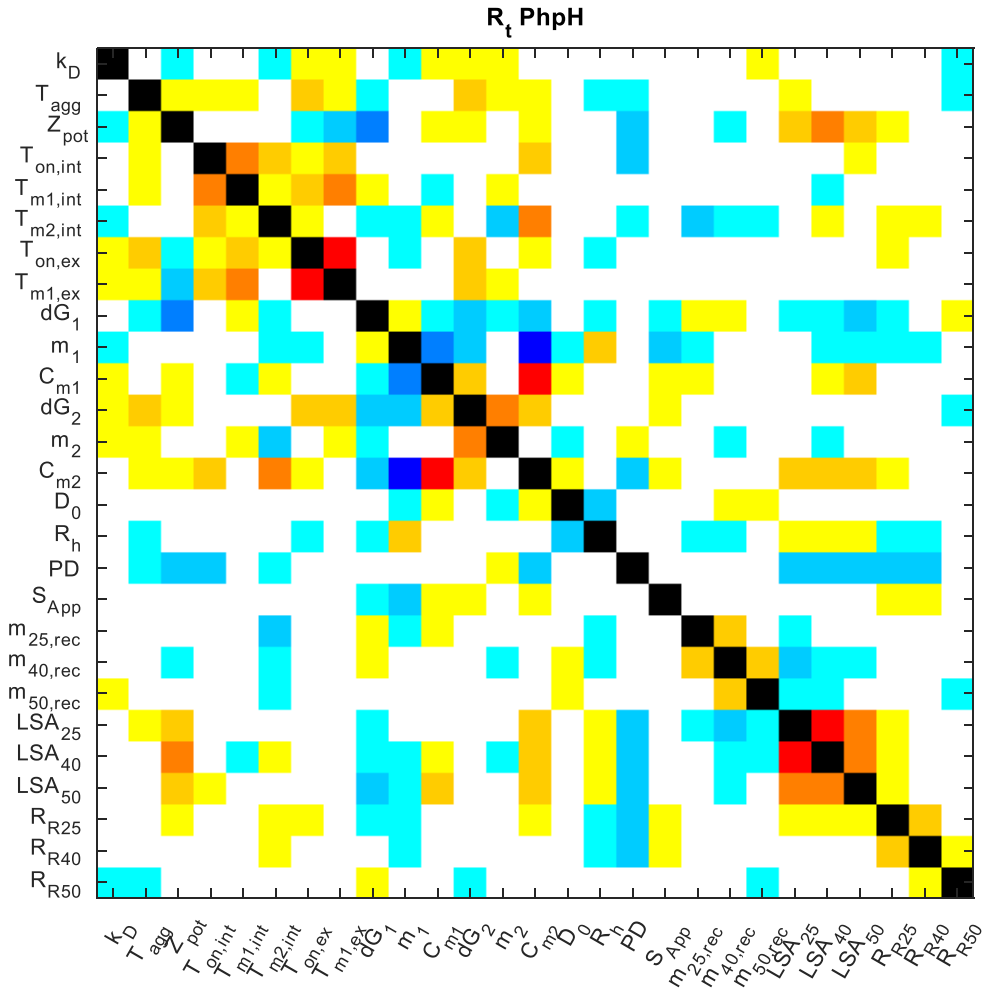


Figure SI 10. Pairwise correlations among biophysical parameters in subsets including different proteins. (Part 9 of 9). *Subset including all 14 proteins and formulations with pharmaceutical relevant pH values (5.0 - 7.5) with all three levels of ionic strength (0, 70 and 140 mM NaCl):*

Figure SI 11. Surface profiles of the investigated proteins relative to a reduced subset of several biophysical parameters

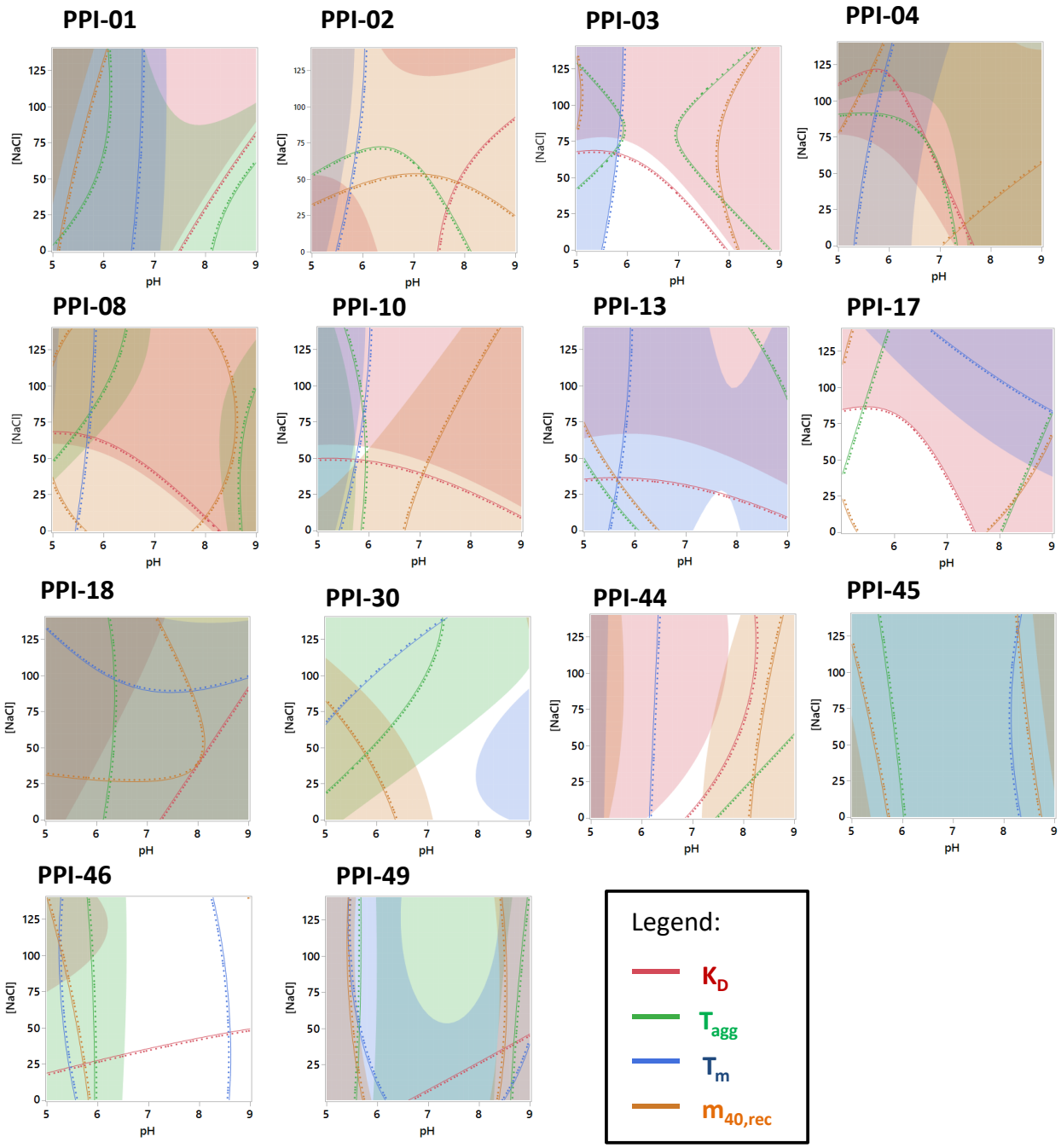


Table SI 12. **Cut off values used for the calculation of the different stability risk scores (SRSs)**

List of cut off values for the calculation of SRSs		Biophysical properties used in the calculation of the SRSs		
Biophysical property	**Cut off value**	**SRS_{LAE}**	**SRS_{MAE}**	**SRS_{HAE}**
T_{agg} (°C)	< 55	YES	YES	YES
$T_{m1,int}$ (°C)	< 65	YES	YES	YES
k_D (mL/mg)	< 0	NO	YES	YES
$m_{40,rec}$ (%)	< 80	NO	YES	YES
Zpot (mV)	< 0	NO	NO	YES
m_1	< 3	NO	NO	YES
C_{m1} (M)	< 2	NO	NO	YES
PD (%)	> 15	NO	NO	YES
$m_{25,rec}$ (%)	< 80	NO	NO	YES
$m_{50,rec}$ (%)	< 50	NO	NO	YES
LSA_{25}	< 0.9	NO	NO	YES
LSA_{40}	< 0.9	NO	NO	YES
LSA_{50}	< 0.9	NO	NO	YES

SI 13. Biophysical parameters table

The data is listed in Appendix 2.

CHAPTER III: Application of interpretable artificial neural networks to early monoclonal antibodies development

Lorenzo Gentiluomo[1,2,6*], Dierk Roessner[2], Dillen Augustijn[3], Hristo Svilenov[1], Alina Kulakova[4], Sujata Mahapatra[5], Gerhard Winter[1], Werner Streicher[5], Åsmund Rinnan[3], Günther H.J. Peters[4], Pernille Harris[4], Wolfgang Frieß[1]

[1]Ludwig Maximilians-Universität München, Department of Pharmacy, Pharmaceutical Technology and Biopharmaceutics, Butenandtstrasse 5, 81377 Munich, Germany

[2]Wyatt Technology Europe GmbH, Hochstrasse 12a, 56307 Dernbach, Germany

[3]Copenhagen University, Department of Food Science, Rolighedsvej 26, 1958 Frederiksberg, Denmark

[4]Technical University of Denmark, Department of Chemistry, Kemitorvet 207, 2800 Kongens Lyngby, Denmark

[5]Novozymes A/S, Krogshoejvej 36, Bagsvaerd, Denmark

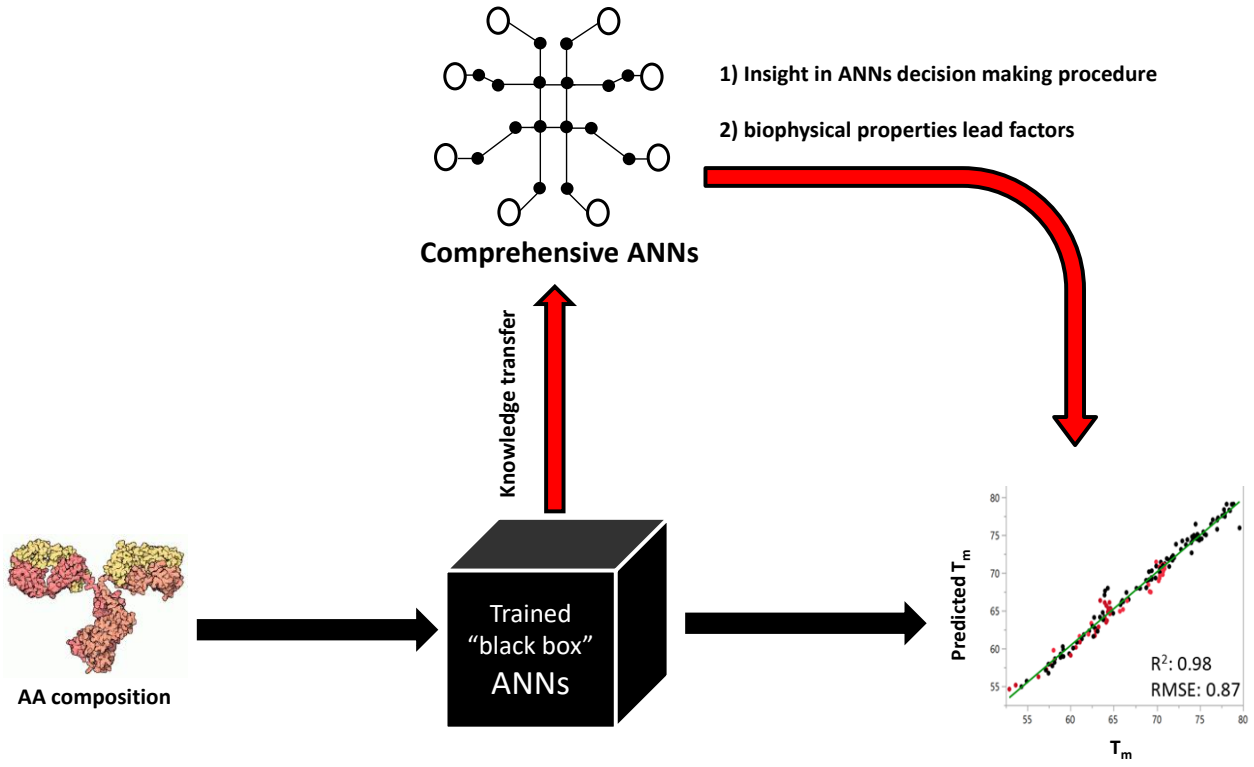
[5] Present address: Coriolis Pharma Research GmbH, Fraunhoferstraße 18B, 82152, Planegg, Germany

*denotes corresponding author

CONTRIBUTIONS: **Lorenzo Gentiluomo** and Wolfgang Frieß conceived and designed the study. **Lorenzo Gentiluomo** designed and computed the artificial neural network models. Dillen Augustijn and Åsmund Rinnan provided support in the development of models. **Lorenzo Gentiluomo** designed the surrogate models from the knowledge transfer process. **Lorenzo Gentiluomo** calculated the response surface analysis. Dillen Augustijn computed partial least squares calculations. Dierk Roessner, Gerhard Winter, Werner Streicher, Åsmund Rinnan, Günther H.J. Peters, Pernille Harris, Wolfgang Frieß supervised the students and provided support to write the manuscript. **Lorenzo Gentiluomo** conducted the T_{agg} and k_D experiments and analyzed the data**.** Hristo Svilenov, Alina Kulakova, Sujata Mahapatra conducted the nanoDSF experiments and analyzed the data**. Lorenzo Gentiluomo** and Wolfgang Frieß contributed to the writing of the manuscript.

Abstract

The development of a new protein drug typically starts with the design, expression and biophysical characterization of many different protein constructs. The initially high number of constructs is radically reduced to a few candidates that exhibit the desired biological and physicochemical properties. This process of protein expression and characterization to find the most promising molecules is both expensive and time-consuming. Consequently, many companies adopt and implement philosophies, e.g. platforms for protein expression and formulation, computational approaches, machine learning, to save resources and facilitate protein drug development. Inspired by this, we propose the use of interpretable artificial neuronal networks (ANNs) to predict biophysical properties of therapeutic monoclonal antibodies i.e. melting temperature T_m, aggregation onset temperature T_{agg}, interaction parameter k_D as a function of pH and salt concentration from the amino acid composition. Our ANNs were trained with typical early-stage screening datasets achieving high prediction accuracy. By only using the amino acid composition, we could keep the ANNs simple which allows for high general applicability, robustness and interpretability. Finally, we propose a novel "knowledge transfer" approach, which can be readily applied due to the simple algorithm design, to understand how our ANNs come to their conclusions.



Graphical abstract. *Application of interpretable artificial neural networks to early monoclonal antibodies development*

1 Introduction

Therapeutic proteins play a crucial role in the treatment of various diseases.[1-3] There are currently over 660 biologics with market approval worldwide. Due to the recent advances in protein engineering, it is nowadays possible to fine-tune desirable protein characteristics to find the optimal balance among efficacy, safety, stability and manufacturability. The development of a protein drug is an extremely complex process involving around 5000 critical steps[4]. During the whole development process the stability of a protein drug is a major concern. The choice of the formulation can drastically affect the conformational, the colloidal and the chemical stability and all three have to be controlled in the final product. The high number of formulation parameters and conditions to be screened requires a significant investment of resources and time. In addition, it has been shown that only 8% of the initially investigated new drug candidates reach license application.[5] It is therefore of significant importance to efficiently use the limited resources and finally to improve the drug-candidate success rate. Nowadays, high-throughput methods are commonly used during the early stage of protein development to select promising candidates and their formulations that will be put forward to undergo forced degradation studies and real-time stability tests.[6-11] In this work we applied Artificial Neural Networks (ANNs) to the most successful class of therapeutic proteins, the monoclonal antibodies (mAbs). ANNs are biologically inspired computer programs designed to simulate how an animal brain processes information, gathering knowledge by detecting the patterns and relationships through a trial and error procedure. There has lately been an increasing interest in ANNs since computers can now process complex shallow ANNs in minutes. The speed at which ANNs can be computed and the fact that big databases are readily available makes this approach very attractive. In recent years, this method has been applied in the pharmaceutical research area for different purposes.[12-19] Supervised ANNs were used as an alternative to response surface methodology[20] while unsupervised networks are an alternative to principal component analysis. Analysis of design of experiments is also possible by ANNs.[21] The great advantage of ANNs over classical statistical modeling is that the former can solve highly non-linear problems often encountered in pharmaceutical processes. However, when the complexity of the ANNs has increased, results from ANNs become increasingly difficult to interpret. A further drawback of ANNs is that a sufficiently big data set is usually required for the learning process.

Combined, our ANNs models provide a tool that is capable of predicting important biophysical properties commonly measured in studying protein physical stability in high throughput, namely the (melting) temperature of unfolding, T_m, the diffusion interaction parameter, k_D, and the onset temperature of aggregation, T_{agg}. These biophysical properties capture different characteristics which, taken together, define significant attributes that can be used to eliminate, or continue with, the development of a candidate. T_m values frequently correlate with the aggregation rate in accelerated stability studies. [22-24] k_D is used to characterize nonspecific protein-protein interactions in diluted solutions and is a good indicator of the solution viscosity at high protein concentrations.[25] Furthermore, the rate of aggregation upon

heating a protein solution is highly correlated to k_D.[26-27] Since the aggregation needs to be kept to a minimum level, T_{agg} is an important biophysical property. The majority of marketed antibodies have T_{agg} greater than 55°C.[28] Even though the aforementioned properties alone will not always correlate with long term stability studies, their knowledge as a function of basic formulation conditions (i.e. pH and ionic strength) allows for a high-throughput way to assess the developability for protein drug candidates in high-throughput and with minimal material consumption. This approach is still very labor and time intensive. Therefore, *in-silico* approaches are of high interest, one of them being the use of ANNs. More importantly, our trained models are based on amino acid composition only. This would allow selecting among thousands of mAbs sequences with good predicted physical stability. The selected protein could then be expressed and purified for going into the next step of the developability assessment.

As pointed out by Ali Rahimi, a researcher in artificial intelligence at Google, machine learning has become a form of alchemy.[29] Therefore our aim was to avoid black-box algorithms. We designed networks that are manageable, and give the user an understanding of their decision-making process. The number and complexity of inputs was reduced by the use of the amino acid composition only. This simple input layer allowed a simple network design which is, compared to complex networks, more general and robust, less prone to overfitting and easier to interpret. As in most cases we achieved accurate predictions, we confirmed that this design was suitable for our purpose. To interpret our models we design a novel “knowledge transfer” process which leads to interpretable ANNs. Additionally, Partial Least Squares Regression (PLS) was performed, and the results were compared with ANNs showing that only ANNs achieve accurate predictions.

2 Material and methods

2.1 Protein and sample preparation

Five IgG1, namely PPI-1, PPI-2, PPI-3, PPI-10, PPI-13 and one IgG2 named PPI-17, were selected based on the availability of the primary sequence, were provided by the PIPPI consortium (http://www.pippi.kemi.dtu.dk). The mAbs were dialyzed overnight using 10 kDa Slide-A-Lyzer™ cassettes (Thermo Fisher Scientific, USA) against an excess of buffer containing 10 mM Histidine at pH 5.0, 5.5, 6.0, 6.5, 7.0, 7.5. Similarly, a buffer containing 10 mM tris(hydroxymethyl)aminomethane (Tris) was used at pH 8.0 and 9.0. Sodium chloride stock solutions were prepared in the respective buffers and diluted to a final concentration of 0, 70 and 140 mM. Protein concentration was measured on a Nanodrop 2000 (Thermo Fisher Scientific, USA) using the respective extinction coefficients calculated from the primary sequence. Reagent chemicals were of analytical grade and were purchased from Sigma Aldrich (Germany) or VWR International (Germany). Highly purified water (HPW, Purelab Plus, USF Elga, Germany) was used for the preparation of all buffers. Formulations including sodium chloride were

prepared by mixing mAbs stock solution in the respective buffer with a stock solution of sodium chloride dissolved in the same buffer. Finally, the formulations were sterile filtered with 0.22 µm cellulose acetate filters from VWR International (Germany). The mAbs' difference in primary structures was investigated using identity and similarity scores as shown in **Table SI 1**.

2.2 Dynamic light scattering

Dynamic light scattering was conducted on a DynaPro Plate Reader II (Wyatt Technology, USA) to obtain the interaction diffusion parameter, k_D, the onset temperature of aggregation, T_{agg}, and the apparent hydrodynamic radius, R_h. 4 µL of each sample per well were pipetted in triplicates into Aurora 1536 Lobase Assay Plates (Aurora Microplates, USA). The samples were overlayed with Silicone oil and centrifuged at 2000 rpm for 1 minute. Data was processed by the DYNAMICS software V7.7 (Wyatt Technology, USA). From the relative autocorrelation function, the coefficient of self-diffusion, D, and the polydispersity index (PDI) were calculated. R_h, was calculated by means of the Stokes-Einstein equation.

k_D was determined using at least six different concentrations (from 1 to 10 mg/mL) in triplicates for each formulation. The samples were filtered using a Millex® 0.22 µm filter from Merk Millipore (USA) and equilibrated at 25 °C for 10 minutes in the Plate reader. Each measurement included 20 acquisitions, each for a duration of 5 s. k_D was determined according to:

$$D = D_0(1 + k_D \cdot c)$$

where D_0 denotes the diffusion coefficient of an isolated scattering solute molecule in the solvent and c is the protein concentration.

For the determination of T_{agg}, the filtered samples at 1 mg/mL were analyzed in duplicates. To achieve high throughput while keeping a suitable point density, 48 wells were filled, and a temperature ramp rate of 0.1°C/min from 25°C to 80°C was applied. One measurement included 3 acquisitions, each with a duration of 3 s. T_{agg} was calculated by the DYNAMICS software V7.7 onset algorithm from the increase in R_h.

2.3 Differential scanning fluorimetry with intrinsic protein fluorescence detection (nanoDSF)

Samples containing 1 mg/mL protein in the respective formulations were filled in standard nanoDSF capillaries (NanoTemper Technologies, Germany). Measurements were performed using the Prometheus NT.48 (NanoTemper Technologies, Germany) system that measures the intrinsic protein fluorescence

intensity change at 330 and 350 nm (after excitation at 280 nm) as a function of temperature. A temperature ramp of 1°C/min was used from 20 °C to 95 °C. The fluorescence intensity ratio (F350/F330) was plotted against the temperature, and the first apparent melting temperature (T_m) was derived from the maximum of the first derivative of each measurement using the PR Control software V1.12 (NanoTemper Technologies, Germany).

2.4 Artificial neural networks

Artificial Neural Networks have been extensively reviewed in the literature, and they have been successfully used in the pharmaceutical industry.[12-21, 30-36] The various applications of ANNs relevant to the pharmaceutical field are classification or pattern recognition, prediction and modeling. Theoretical details can be found elsewhere.[37] The network's fundamental parts are the neurons, also called nodes, and their connections. The diagram in **Fig. 1** shows the model of a neuron. The neuron is an information-processing unit, which is constituted of a set of connection links characterized by their weight, w_{kn}, a linear combiner, Σ, and an activation function, ψ. An externally applied bias, $b_{k,}$ is used to modify the net input received for each neuron in the network. An often used simplified description of the network is the architectural graph, depicted in **Fig. 2**. ANNs solve problems by training, a trial and error process for optimizing the synaptic weight values. During the training, the squared error between the estimated and the experimental values is minimized by reinforcing the synaptic weights, w_{kn}. ANNs have robust performance in dealing with noisy or incomplete data sets, the ability to generalize from input data and a high fault tolerance.[38]

ANNs have a series of known limitations, namely overfitting, chance effects, overtraining, and difficult interpretability.[39-41] The first three limitations were extensively reviewed in the literature and can be prevented by using various methodologies. The interpretation of ANNs is not straightforward, and it is still an open field of research. Our primary goal was therefore to build an algorithm through which it was possible to follow how the networks have come to a particular conclusion. To achieve this, we used the simplest input related to the mAbs giving an accurate prediction, namely the amino acid composition. In order to comprehend the artificial decision-making procedure a novel "knowledge transfer" process was designed, which is described in section 3.7.

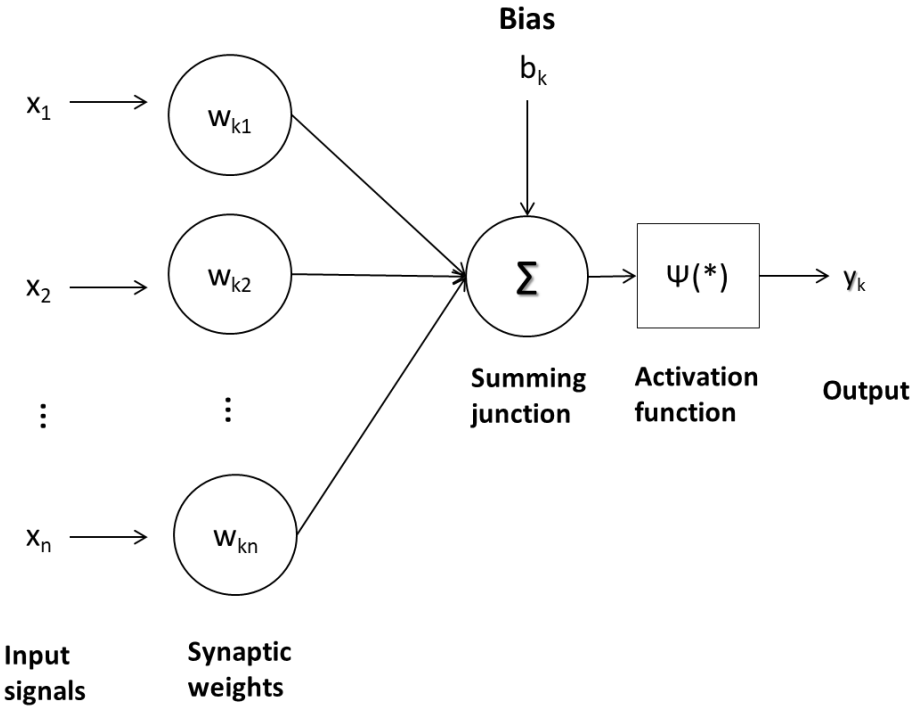


Figure 1. Model of a neuron. *x_n represent the inputs connected to the neuron, k, by the weights, w_{kn}, which multiply the corresponding input signal. All the weighted signals are summed by a summing junction Σ. An external bias b_k can be applied to Σ, to increase or lower the output signal. Finally, Σ is connected to an activation function, ψ(*), which limits the amplitude of a signal to the output, y_k. Picture modified from: Neural networks: a comprehensive foundation, S. Haykin.*[45]

Our multilayer feed-forward back-propagation networks present one hidden layer, which is usually sufficient to provide adequate predictions even when continuous variables are adopted as units in the output layer. [43-45] Equation 1 (described by Carpenter[44]) was used to estimate the optimal number of neurons in the hidden layer:

Eq. 01
$$N_{hidden} = \left(\frac{N_{sample}}{\beta} - N_{output}\right) \Big/ \left(N_{input} + N_{output} + 1\right)$$

where β, N_{hidden}, N_{output} and N_{sample} are the determination parameter, the number of hidden units, the number of output units and the number of training data pairs, respectively. Overdetermined, underdetermined and determined parameters will be reflected by $\beta>1$, $\beta<1$ and $\beta=1$, respectively. The β value to adopt depends on the degree of quality of the data set in terms of the degree of independency among other factors. Our dataset consisted of 144 instances (24 conditions per protein) for each biophysical parameter and seven neurons were estimated to provide a β of 1. In general terms, simpler

models are more general and easier to interpret. Since our aim was to have the most general and easiest to interpret model possible, we selected the minimum number of neurons, 5, which provided the same result as 7 neurons. In **Table SI 2** the list of input parameters relative to each model is shown, while in **Fig. SI 3** an exemplary scheme of the model's architecture is presented.

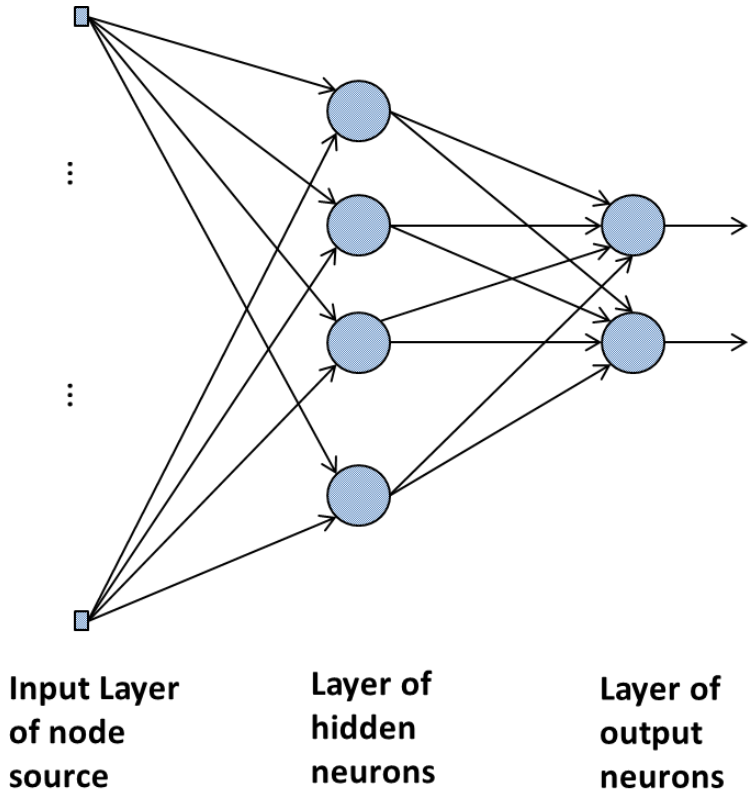


Figure 2. Signal-flow graph of a fully connected feedforward network with one hidden layer and one output layer. *The signal-flow graph provides a neat description of the neural networks describing the links among the various nodes of the model. Picture adapted from: Neural networks: a comprehensive foundation, S. Haykin.*[45]

All the input parameters were normalized before the training phase by subtracting the mean and then dividing by the standard deviation. The learning rate was selected on a trial and error basis in such a way so as to keep the minimum distance between the actual and predicted value. The validation method is described in section 4.1. JMPpro® (SAS Institute Inc., USA), MATLAB® (MathWorks, USA) and Weka (Waikato University; New Zealand) were used to generate ANNs. These networks yielded highly similar results and JMPpro® v.13 was selected for its user-friendly interface and subsequently potentially easier implementation in a drug development department.[47]

2.5 Knowledge transfer to explain ANNs results

In order to understand the decision-making process of our ANN models, a novel knowledge transfer process, implying response surface methodology (RSM), was applied by evaluating the weights of the trained network to transfer the acquired knowledge of ANNs to linear models. Parameters deemed important by the networks were selected and the interpretation of ANNs was then assessed by RSM of the linear least square regression of these "leading parameters". The scheme of this process, named "knowledge transfer", is depicted in **Fig. 3**.

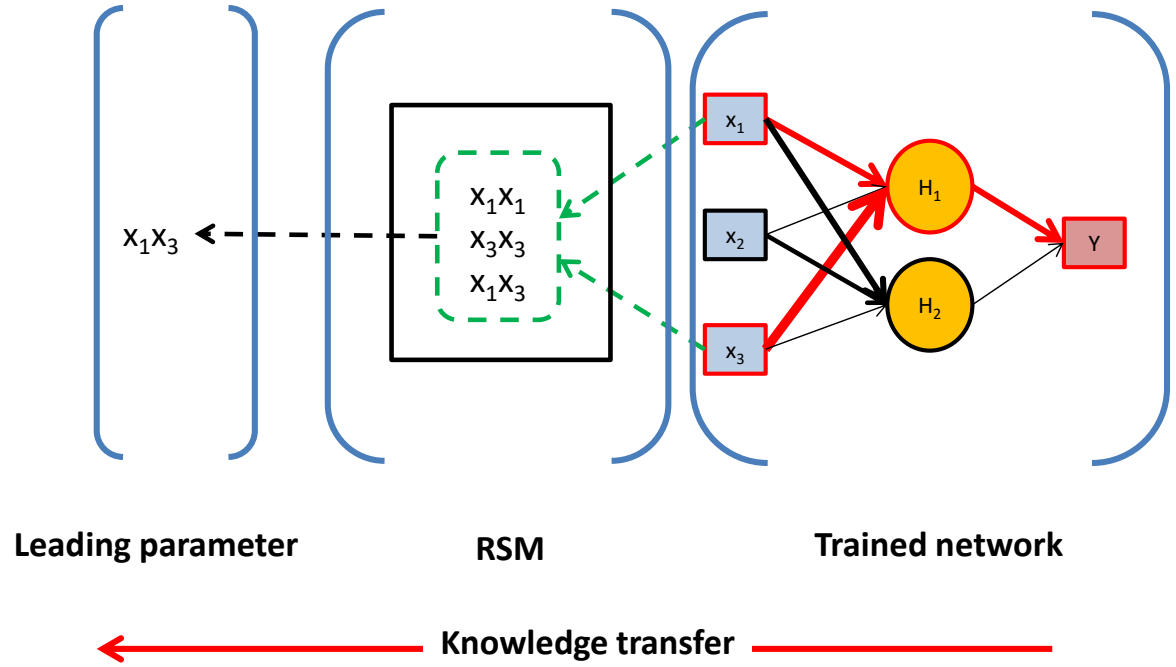


Figure 3. Scheme of the knowledge transfer procedure. *On a trained network, where the arrow thickness represents the weight value (i.e. smaller arrow present lower weights), the input parameters with the higher impact, in red, are selected. These inputs are used for a least square linear regression where the RSM is applied considering only two-way interactions. From the analysis, leading parameters are selected and discussed to interpret the network decision-making process.*

None of the hidden nodes in the ANNs' prediction formulas has a weight close to zero, which means that all nodes contribute to the final output. However, around 5% of the weights of the output layer presented values which were at least twice the average mean of all the network weights. From these 5% we selected the input parameters from the activation functions whose coefficients were at least twice the average values.

We assessed the full model using all the selected "leading parameters" from the networks, and then reduced the model to only the terms that were deemed statistically relevant. A curved response was

allowed by assessing the quadratic term considering also two-way interactions. The reduced model was obtained using a backward stepwise regression. The F-statistic approach was used to perform the effect test considering a value of 0.05 or less as statistically significant. All the results were calculated using the statistical software JMP® v 13.0 (SAS Institute Inc., Cary, USA)[47], and all the analysis details can be found in the software manual.

3 Result and discussion

A general flow diagram of our approach is shown in **Fig. 4**. At first, the power of our ANNs for prediction of the biophysical parameters T_m, T_{agg} and k_D at different pH as well as salt concentration was evaluated. Only the number of each amino acid species of the proteins was used as protein-related input parameters.

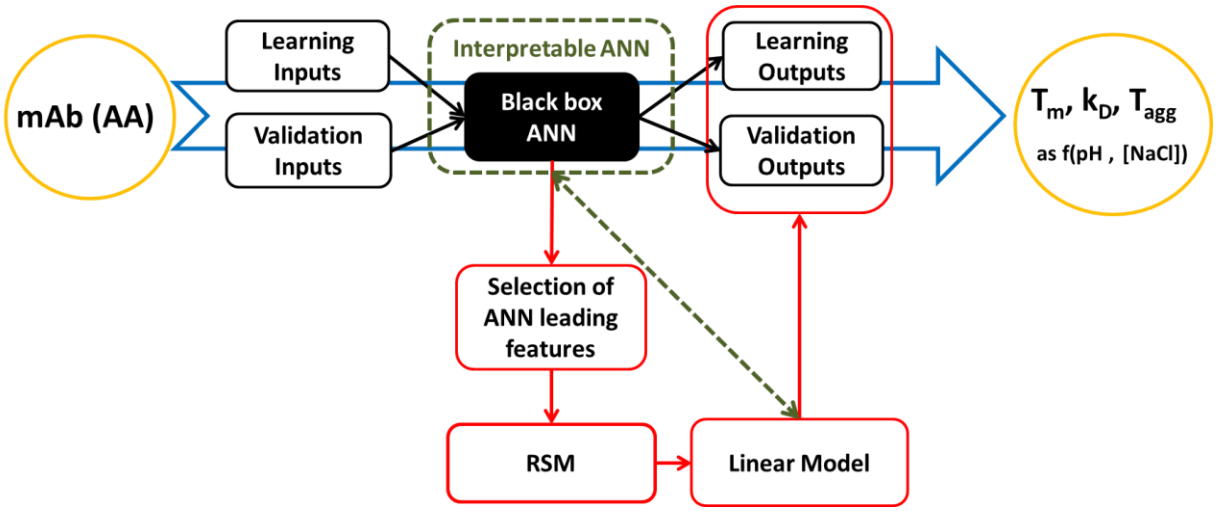


Figure 4. Diagram describing the process applied to achieve an interpretable prediction by ANNs. *The knowledge transfer process is highlighted in red. The model explanation (dashed green lines) is aimed at understanding the overall logic behind the black box. Once trained and validated the interpretable ANN can be applied to new mAb candidates, even before cell expression. This allows to predict important biophysical parameters (i.e. T_m, k_D and T_{agg}) as a function of pH and salt concentration.*

The primary sequence was not used as an input parameter, neither were other typical molecular descriptors included e.g. charge distribution, dipole moments or solvent exposure. However, we are currently working together with other members of the PIPPI consortium (http://www.pippi.kemi.dtu.dk) to

create a publicly available protein formulation database. Such a database may be used in future to build on our findings and to generate more sophisticated deep learning models based on the amino acid sequence. We avoided the use of formulation dependent molecular descriptors (e.g. net charge) to reduce redundancy, as the formulation is always included as input. Moreover, it has been proven that even net charge cannot be accurately calculated.[48] Further, we investigated a series of molecular indices which are only protein dependent, calculated by ProtDCal,[49] listed in **Table** SI 6. However, we could not find a subset of these indices that would yield an accuracy similar to the number of amino acids. As machine learning models describe correlation and not causation - highlighted by George E. P. Box: *"Essentially, all models are wrong, but some are useful"*[50] - we selected the minimum number of input parameters to achieve high accuracy and interpretability. The number of amino acids can easily be described by only 20 input values, whereas thousands of inputs are necessary to describe the primary sequence (depending on the size of the molecule). This would drastically increase the complexity of the algorithms requiring a deep neural network with thousands to millions of data points, which are nowadays not publicly available. Such a complex approach makes the algorithm difficult to interpret and interpretability was one of our goals. As we managed to reach accurate predictions we found our model useful for its purpose: an *in-silico* tool for the selection of mAbs with predicted high physical stability from a vast number of possible candidates, which is interpretable, which is independent from other calculations (e.g. solvent exposure), and which can output experimentally accessible biophysical properties in early stage (i.e. low volume, high throughput). An additional advantage of a simple design is that such models are usually more general and robust.

In order to gain insight from the ANNs decision making procedure we introduce a novel knowledge transfer process (depicted in red in **Fig. 4**). As the outputs (e.g. T_m) of our models are easily accessible in early stage, once the selected candidates are expressed and purified, it is possible to continuously re-train the network and to double check its validity. One disadvantage of such approach is that it is only suitable to predict closely related protein structures to the one used for the training phase, e.g. IgG1 and IgG2.

3.1 Prediction of T_m, T_{agg} and the sign of k_D

The ability of the model to predict T_m, T_{agg} and k_D from the numbers of each different amino acid in each mAb and the formulation conditions (i.e. pH and salt concentration) was cross-validated. Data from two mAbs were selected and held back in a validation set during the training phase. Applying the model to the validation data allows an unbiased comparison between the predicted and measured values. Thus, the estimation of the prediction error for potential new mAb samples is based on the results of the validation set. This validation method was deemed superior to the random data splitting. The latter yielded better fitting and prediction. However, the model would have experienced all the molecules during the training

phase. Therefore, we discarded the random data splitting as our aim was to validate a model capable of predict biophysical parameters of unknown mAbs. Using this cross-validation strategy, a total of fifteen models were built, each of them based on a different training and validation set, for each studied biophysical property. As the investigated mAbs presented different stability (i.e. different biophysical properties values) the point distribution varies depending on the validation mAbs. The models were characterized by the name of the withheld proteins (e.g., the model called PPI-1&2 is based on the validation data set of PPI-1 and PPI-2, and trained on the PPI-3, PPI-10 , PPI-13 and PPI-17 data).

In **Fig. 5**, the predicted T_m, T_{agg} and the sign of k_D of the PPI-3&13 models are shown. T_m and the sign of k_D were fitted to a very high degree of accuracy. The T_m model presented an R^2 of 0.98 and a root mean squared error (RMSE) of around 0.8°C from the reference T_m while the sign of the k_D model was classified with no false negative or false positives. The T_{agg} model presented an R^2 of 0.94 but with a higher RMSE value of around 2°C. The higher error is probably due to the high throughput fashion of the screening, which stretched the limit of necessary high data density for the determination of the onset. In other words, the input data has higher uncertainty that is reflected in the prediction error. In **Figs. SI 4-5** the predicted data point from the T_m and T_{agg} models are presented.

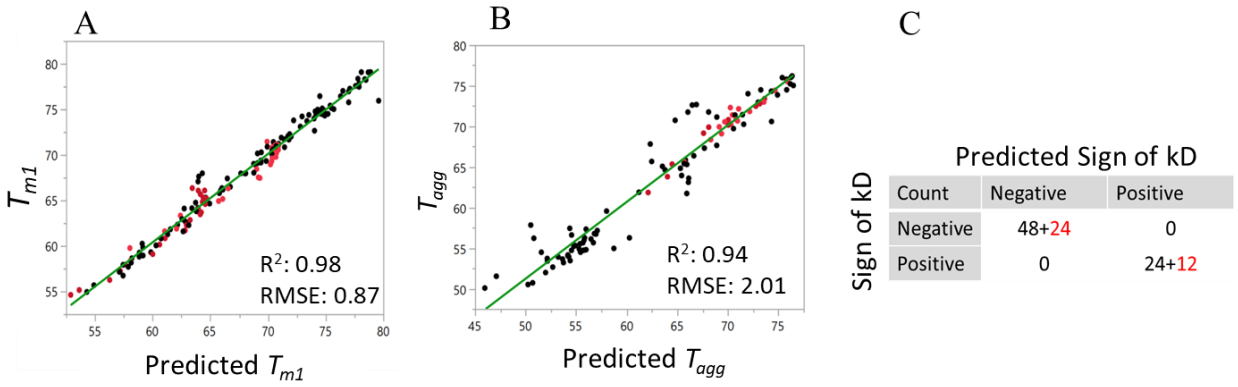


Figure 5. Results from PPI-13&3 models. *The predictions of T_m, T_{agg} and the sign of k_D are shown in graphs A, B and C respectively. Black dots and numbers represent the training set, while red dots and numbers represent the validation set.*

The robustness of the ANNs regressions was evaluated based on R^2, shown in **Fig. 6 (A)**, and RMSE values of the training and validation set. The latter was in the range of ca. 1 °C to 3 °C from the reference T_{agg} or T_m, with no particular trend or direction with respect to the measured values. The robustness of the classification problem, the sign of k_D, was evaluated on the misclassification rate, shown in **Fig. 6 (B)**.

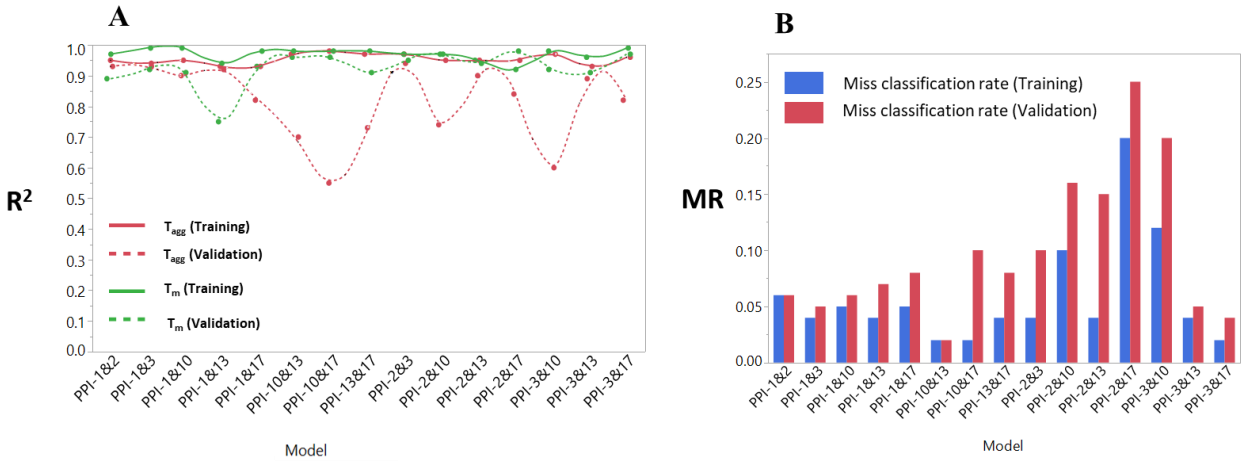


Figure 6. ANN robustness study of ANNs. *In graph A, the R^2 values for the T_m and T_{agg} models are shown. In graph B, the misclassification rate (MR) of the sign of k_D models are shown. Blue bars represent the validation set while red bars represent the validation set. The models were classified by the name of the proteins used for the validation.*

Regarding the T_m models, we observe broad robustness without significant influence of the different training sets. The colloidal stability parameters, T_{agg} and sign of k_D, appear to be more sensitive to the selected training sets. Two T_{agg} models show serious deviation in prediction both involving PPI-17 and/or PPI-10. These two proteins showed extreme aggregation during temperature ramps, compared to the other mAbs. Consequently, the ANNs can easily fit PPI-17 and PPI-10 data, but in order to predict their aggregation propensity, the network would require more data representative of this kind of aggregation behavior.

The k_D data consists for ca. 70% of negative values. This unbalanced data set is caused by the charge screening effect of the added salt that occurs in two-thirds of the formulations and therefore the number of positive values is not enough to solve an ANN regression problem. One such occurrence is shown in **Fig. 7** for the PPI-13&3 model, where all the negative values are fit well, while the positive values are not well calculated and broadly distributed. Despite this, the sign of k_D was always predicted to a high degree of accuracy as shown in **Fig. 6(B)**.

The studies on the robustness allowed us to conclude that well defined and simpler properties, such as the temperature of unfolding, are not greatly influenced by the training set. In contrast, the colloidal properties need more attention in the selection of the training set.

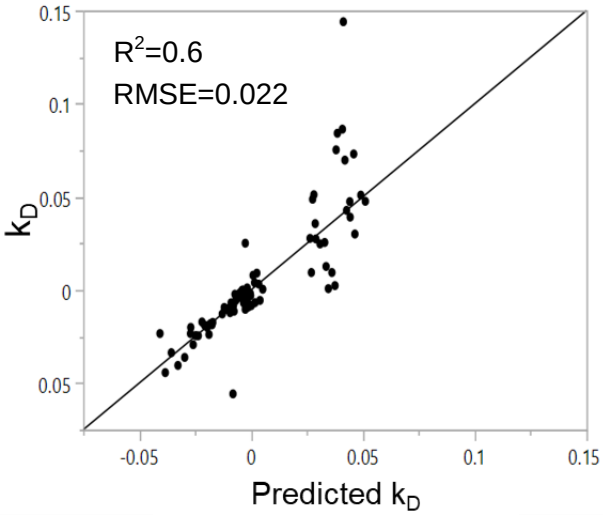


Figure 7. *Correlation between experimentally determined and predicted k_D values for the PPI-13&3 model.*

3.2 ANNs knowledge transfer

The scientific community has been investigating the problem of explaining machine learning decision models and a comprehensive survey of methods for explaining black box models has been redacted.[51] In order to understand the thought process of our ANNs, a novel knowledge transfer process, depicted in **Fig**. **3**, was applied. **Fig. 8** shows the results from the RSM relative to T_m, T_{agg}, k_D, while **Table 1** summarizes the effective test statistics which can be used as an indication of the relative impact of the parameters. Quadratic terms (e.g. Cys·Cys) were assessed to model potential curvature in the response. These linear models allow to understand the logic of the relative ANNs model and to follow the reasoning of the outcomes, i.e. each leading amino acid has a specific role in the physical process related to the output parameters.

The T_m linear model is primarily affected by pH, salt concentration, and the number of tryptophan, cysteine and tyrosine residues. Therefore, the main protein related contributors to the unfolding process are two hydrophobic amino acids residues and cysteine. It is known that the unfolding process is mainly guided by hydrophobic interactions,[52] while cysteine is involved in disulfide bonds, stabilizing the protein structure. Interestingly charged residues are of minor importance.

The T_{agg} linear model is mainly affected by pH, salt concentration, and the number of aspartic acid, glutamic acid and methionine residues. Therefore, the main protein related contributors to the aggregation process were charged amino acid residues and methionine. It is known that the oxidation of

methionine is a critical pathway of aggregation under accelerated thermal stability stress[53]. Moreover, methionine oxidation is practically pH independent[54], which could partially explain the minor impact of pH on the models. However, during a temperature ramp, the time of stress is relatively short and hence, the oxidation of methionine should have a minor impact. Consequently, during a temperature ramp, charged amino acids have a higher impact on the linear model.

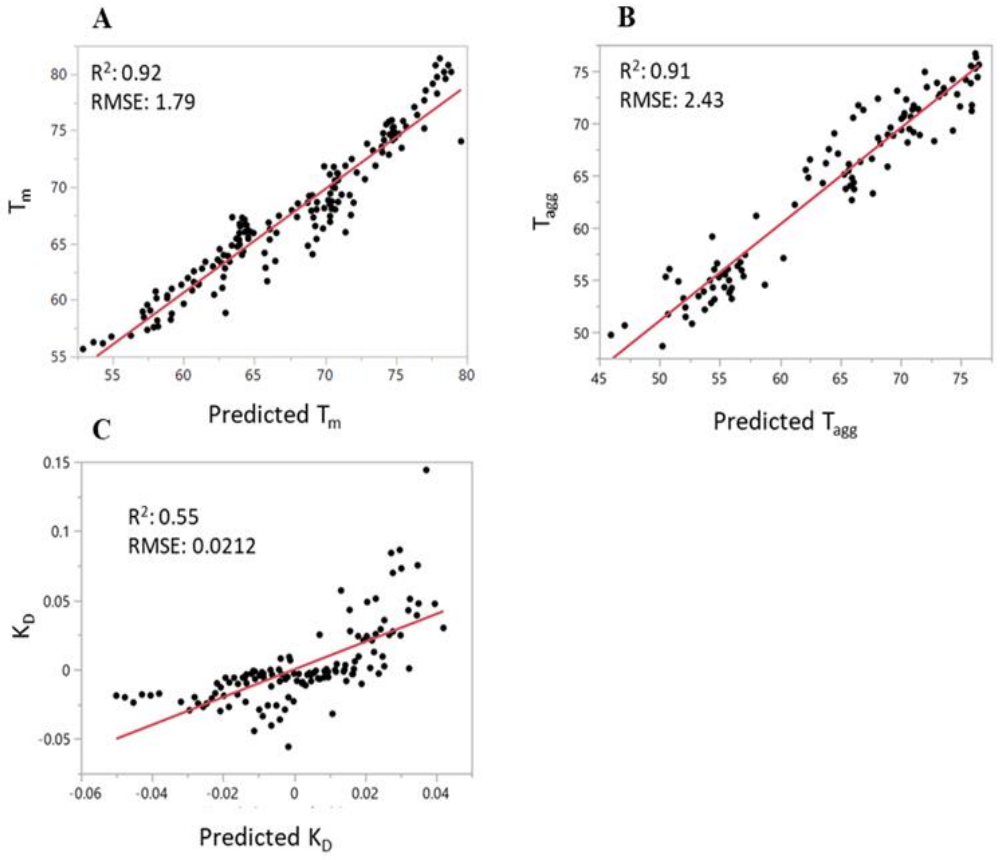


Figure 8. Results of T_m, T_{agg}, k_D linear models. *The T_m, T_{agg}, k_D results from the network knowledge transfer process are shown respectively in graph A, B and C. The 3 graphs are generated by RSM using the selected leading parameter. The relative effect test is presented in Table 1.*

The k_D linear model is affected by pH, salt concentration, and the number of glutamic acid, histidine, and tryptophan residues. Thus, both charged and hydrophobic amino acids are important. k_D is used to evaluate pairwise protein-protein nonspecific interactions, which can be rationalized by means of the DLVO[55,56] or proximity energy theory[57]. Both theories highlight the fact that protein-protein interactions depend heavily on hydrophobic and charged patches on the protein surface. Moreover, histidine plays a particular role in protein-protein interactions. This amino acid has a pK_a of 6.0 i.e. histidine changes

charge state under relevant formulation pH conditions. Therefore histidine doping is a common method in engineering stable proteins[58-62] and the presence of histidine residues can mediate structural transitions in binding or folding of the interacting proteins.[63-65]

Table 1. Effect tests results of the RSM applied to the linear least square regression from the knowledge transfer of ANNs' models. *In Fig. 6 the relative graphs are shown. Information on the inputs can be found in Table SI 6. The quadratic terms (e.g. Cys·Cys) and the cross terms (e.g. pH·Cys) from the RSM were selected by reducing the full model using a backward stepwise regression where a value of p<0.05 is deemed statistically significant. LogWorth is defined as -log$_{10}$(p-value).*

T_m		T_{agg}		k_D	
Input	LogWorth	Input	LogWorth	Input	LogWorth
Trp	27.942	Glu	36.173	[NaCl]	11.608
pH	25.425	Met·Met	26.675	Glu	9.529
pH·Cys	13.701	Met	19.023	Trp	9.151
pH·pH	13.256	Asp	6.996	His	8.828
Cys·Cys	8.528	pH	6.084	pH	2.490
Cys	4.024	pH·pH	4.881		
Tyr·Tyr	3.813	Asp·Asp	4.199		
Tyr	3.284	[NaCl]	2.474		
[NaCl]	2.753				

Taken together, our ANN knowledge transfer process allows us to interpret the factors behind the decision-making process of the ANN when predicting T_m, T_{agg} the sign of k_D. This process provided a global explanation of the black box through an interpretable and transparent model. By this, we build trust into our approach and are not left with a black box. As an agnostic process can explain unrelated algorithm only indifferently, our approach is not to be considered agnostic as it is tied to simple ANNs.

3.3 Prediction comparison with partial least square models

The main reason to apply ANNs comes from their prediction power using data sets with highly non-linear relationships. To demonstrate the necessity for a non-linear model, a linear regression analysis using the partial least square regression (PLS) method was performed. PLS is probably the strongest competitor of ANNs in terms of robustness and predictive power and can be extremely powerful in fitting data and for this reason it was compared to ANN. In fact, PLS was the only model we tested capable of fitting the dataset. As we aimed to develop an interpretable model, we also tested models usually considered readily interpretable (e.g. decision tree) without success. A detailed discussion about modeling alternatives can be found in an article by Frank and Friedmann.[66] The optimal number of latent variables was selected based on the minimum of the RMSE of the cross-validation. The same cross-validation method was applied as in the ANNs in order to make the models comparable. In **Fig. 9** the prediction for all the proteins is shown. The results demonstrate that PLS cannot be used for our dataset and we can conclude that ANN is a far better methodology than PLS to construct models that predict the formulation behavior of unknown proteins under the conditions that we have used.

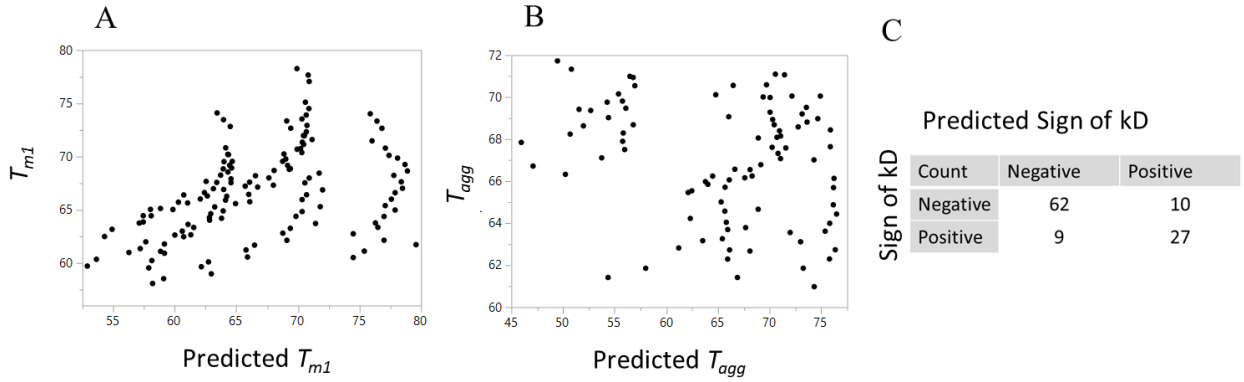


Figure 9. Results of the validation sets from the PLS model. *T_m, T_{agg} and the sign of k_D results are shown respectively in graphs A, B and C. The graphs show that the models cannot accurately predict protein properties that were not involved in the training set.*

4 Conclusions

ANNs represent an interesting alternative to the classical statistical methodologies when applied to highly non-linear data sets that are frequently encountered in the pharmaceutical industry. We successfully developed interpretable models for a set of mAbs to predict important biophysical properties as a function of pH and salt concentration. In the field of mAbs development, ANNs could be a highly valuable tool to

predict important biophysical properties and to support development risk assessment. This approach would allow the selection of mAbs with good physicochemical properties already before expression in cells. The only information required for our approach is the amino acid composition of each mAb. Due to the accuracy of the predictions, there was no reason to increase the complexity of the model since it would hamper the interpretability and robustness. Thanks to our design a novel knowledge transfer process allows to understand the decision-making process of our algorithm. In contrast, PLS models did not work demonstrating that a non-linear algorithm is required to analyze a data set like the one used in our study. The knowledge gathered with simpler ANNs can be used to build even more impressive systems in the future, to confirm the reliability of ANNs and finally to highlight which factors may impact protein stability the most.

5 Acknowledgements

This study was funded by a project part of the EU Horizon 2020 Research and Innovation program under the Marie Skłodowska-Curie grant agreement No 675074. The first author would like to thank Nanotemper Technologies GmbH for kindly proving support for the NanoDSF data, SAS Institute for proving JMPpro® V 13.0, and the whole PIPPI consortium (http://www.pippi.kemi.dtu.dk) for the continuous support offered and for reviewing the manuscript.

6 References

1. R. Gong, W. Chen, D.S. Dimitrov, Expression, purification, and characterization of engineered antibody CH2 and VH domains, Methods Mol. Biol. 899 (2012) 85–102.

2. D.S. Dimitrov, Therapeutic antibodies, vaccines and antibodyomes, MAbs 2 (3) (2010) 347–356.

3. J.G. Elvin, R.G. Couston, C.F. van der Walle, Therapeutic antibodies: market considerations, disease targets and bioprocessing, Int. J. Pharm. 440 (1) (2013) 83–98.

4. H.A. Lagassé, A. Alexaki, V.L. Simhadri, N.H. Katagiri, W. Jankowski, Z.E. Sauna, C. Kimchi-Sarfaty, Recent advances in (therapeutic protein) drug development, F1000Research 6 (2017) 113.

5. US Department of Health and Human Services. Innovation or stagnation: Challenge and opportunity on the critical path to new medical products. Available at: http:// wayback.archive-it.org/7993/20180125032208/https://www.fda.gov/ScienceResearch/SpecialTopics/CriticalPathInitiative/CriticalPathOpportunitiesReports/ucm077262.htm. Accessed July 3, 2017.

6. M.A. Capelle, R. Gurny, T. Arvinte, High throughput screening of protein formulation stability: practical considerations, J. Pharm. Biopharm. 65 (2) (2007) 131–148.

7. F. He, S. Hogan, R.F. Latypov, L.O. Narhi, V.I. Razinkov, High throughput thermostability screening of monoclonal antibody formulations, J. Pharm. Sci. 99 (4) (2010) 1707–1720.

8. D.S. Goldberg, S.M. Bishop, A.U. Shah, H.A. Sathish, Formulation development of therapeutic monoclonal antibodies using high-throughput fluorescence and static light scattering techniques: Role of conformational and colloidal stability, J. Pharm. Sci. 100 (4) (2011) 1306–1315.

9. D.S. Goldberg, R.A. Lewus, R. Esfandiary, D.C. Farkas, N. Mody, K.J. Day, P. Mallik, M.B. Tracka, S.K. Sealey, H.S. Samra, Utility of high throughput screening techniques to predict stability of monoclonal antibody formulations during early stage development, J. Pharm. Sci. 106 (8) (2017) 1971–1977.

10. R. Chaudhuri, Y. Cheng, C.R. Middaugh, D.B. Volkin, High-throughput biophysical analysis of protein therapeutics to examine interrelationships between aggregate formation and conformational stability, AAPS J. 16 (1) (2014) 48–64.

11. N.R. Maddux, V. Iyer, W. Cheng, A.M. Youssef, S.B. Joshi, D.B. Volkin, J.P. Ralston, G. Winter, C.R. Middaugh, High throughput prediction of the long-term stability of pharmaceutical macromolecules from short-term multi-instrument spectroscopic data, J. Pharm. Sci. 103 (3) (2014) 828–839.

12. A.S. Hussain, X.Q. Yu, R.D. Johnson, Application of neural computing in pharmaceutical product development, Pharm. Res. 8 (10) (1991) 1248–1252.

13. E. Murtoniemi, P. Merkku, P. Kinnunen, K. Leiviskae, J. Yliruusi, Effect of neural network topology and training end point in modelling the fluidized bed granulation process, Int. J. Pharm. 110 (2) (1994) 101–108.

14. M. Gasperlin, L. Tusar, M. Tusar, J. Smid-Korbar, J. Zupan, J. Kristl, Lipophilic semisolid emulsion systems: viscoelastic behaviour and prediction of physical stability by neural network modelling, Int. J. Pharm. 196 (1) (2000) 37–50.

15. K. Takayama, M. Fujikawa, T. Nagai, Artificial neural network as a novel method to optimize pharmaceutical formulations, Pharm. Res. 16 (1) (1999) 1–6.

16. A.S. Achanta, J.G. Kowalski, C.T. Rhodes, Artificial neural networks: implications for pharmaceutical sciences, Drug Dev. Ind. Pharm. 21 (1) (2008) 119–155.

17. A.C. King, M. Woods, W. Liu, Z. Lu, D. Gill, M.R. Krebs, High-throughput measurement, correlation analysis, and machine-learning predictions for pH and thermal stabilities of Pfizer-generated antibodies, Protein Sci. 20 (9) (2011) 1546–1557.

18. Y. Yang, Z. Ye, Y. Su, Q. Zhao, X. Li, D. Ouyang, Deep learning for in vitro prediction of pharmaceutical formulations, Acta Pharm Sin B. 9 (1) (2019 Jan) 177–185.

19. Z. Ye, Y. Yang, X. Li, D. Cao, D. Ouyang, An integrated transfer learning and multitask learning approach for pharmacokinetic parameter prediction, Mol. Pharm. 16 (2) (2019 Feb 4) 533–541.

20. J. Bourquin, H. Schmidli, P. van Hoogevest, H. Leuenberger, Application of artificial neural networks (ANN) in the development of solid dosage forms, Pharm. Dev. Technol. 2 (2) (1997) 111–121.

21. A.P. Plumb, R.C. Rowe, P. York, C. Doherty, The effect of experimental design on the modeling of a tablet coating formulation using artificial neural networks, Eur. J. Pharm. Sci. 16 (4–5) (2002) 281–288.

22. L. Burton, R. Gandhi, G. Duke, M. Paborji, Use of microcalorimetry and its correlation with size exclusion chromatography for rapid screening of the physical stability of large pharmaceutical proteins in solution, Pharm. Dev. Technol. 12 (3) (2007) 265–273.

23. M.L. Brader, T. Estey, S. Bai, R.W. Alston, K.K. Lucas, S. Lantz, P. Landsman, K.M. Maloney, Examination of thermal unfolding and aggregation profiles of a series of developable therapeutic monoclonal antibodies, Mol. Pharm. 12 (4) (2015) 1005–1017.

24. V. Kumar, N. Dixit, L.L. Zhou, W. Fraunhofer, Impact of short range hydrophobic interactions and long range electrostatic forces on the aggregation kinetics of a monoclonal antibody and a dual-variable domain immunoglobulin at low and high concentrations, Int. J. Pharm. 421 (1) (2011) 82–93.

25. M.S. Neergaard, D.S. Kalonia, H. Parshad, A.D. Nielsen, E.H. Møller, M. van de Weert, Viscosity of high concentration protein formulations of monoclonal antibodies of the IgG1 and IgG4 subclass–Prediction of viscosity through protein–protein interaction measurements, Eur. J. Pharm. Sci. 49 (3) (2013) 400–410.

26. J. Rubin, L. Linden, W.M. Coco, A.S. Bommarius, S.H. Behrens, Salt-induced aggregation of a monoclonal human immunoglobulin G1, J. Pharm. Sci. 102 (2) (2013) 377–386.

27. J. Rubin, A. Sharma, L. Linden, A.S. Bommarius, S.H. Behrens, Gauging colloidal and thermal stability in human IgG1–sugar solutions through diffusivity measurements, J. Phys. Chem. B 118 (11) (2014) 2803–2809.

28. A. Jarasch, H. Koll, J.T. Regula, M. Bader, A. Papadimitriou, H. Kettenberger, Developability assessment during the selection of novel therapeutic antibodies, J. Pharm. Sci. 104 (6) (2015) 1885–1898.

29. M. Hutson, Has artificial intelligence become alchemy? Science 360 (6388) (2018) 478.

30. A. Ghaffari, H. Abdollahi, M.R. Khoshayand, I.S. Bozchalooi, A. Dadgar, M. Rafiee-Tehrani, Performance comparison of neural network training algorithms in modeling of bimodal drug delivery, Int. J. Pharm. 327 (1–2) (2006) 126–138.

31. A. Hussain, P. Shivanand, R.D. Johnson, Application of neural computing in pharmaceutical product development: computer aided formulation design, Drug Dev. Ind. Pharm. 20 (10) (2008) 1739–1752.

32. E. Murtoniemi, J. Yliruusi, P. Kinnunen, P. Merkku, K. Leiviskae, The advantages by the use of neural networks in modelling the fluidized bed granulation process, Int. J. Pharm. 108 (2) (1994) 155–164.

33. S. Agatonovic-Kustrin, R. Beresford, Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research, J. Pharm. Biomed. Anal. 22 (5) (2000) 717–727.

34. J. Bourquin, H. Schmidli, P. van Hoogevest, H. Leuenberger, Advantages of Artificial Neural Networks (ANNs) as alternative modelling technique for data sets showing non-linear relationships using data from a galenical study on a solid dosage form, Eur. J. Pharm. Sci. 7 (1) (1998) 5–16.

35. Y. Chen, S.S. Thosar, R.A. Forbess, M.S. Kemper, R.L. Rubinovitz, A.J. Shukla, Prediction of drug content and hardness of intact tablets using artificial neural network and near-infrared spectroscopy, Drug Dev. Ind. Pharm. 27 (7) (2001) 623–631.

36. S.S. Haykin, Neural Networks: A Comprehensive Foundation, second ed., Prentice Hall PTR, 1998.

37. D.W. Patterson, Artificial Neural Networks: Theory and Applications, Prentice Hall Asia, 1998.

38. D.J. Livingstone, D.T. Manallack, I.V. Tetko, Data modelling with neural networks: advantages and limitations, J. Comput. Aided Mol. Des. 11 (2) (1997) 135–142.

39. D.T. Manallack, D.J. Livingstone, Artificial neural networks: application and chance effects for QSAR data analysis, Med. Chem. Res. 2 (1992) 181–190.

40. D.J. Livingstone, D.T. Manallack, Statistics using neural networks: chance effects, J. Med. Chem. 36 (9) (1993) 1295–1297.

41. D.T. Manallack, D.D. Ellis, D.J. Livingstone, Analysis of linear and nonlinear QSAR data using neural networks, J. Med. Chem. 37 (22) (1994) 3758–3767.

42. R.P. Lippman, An introduction to computing with neural nets, IEEE Assp Mag. 4 (2) (1987) 4–22.

43. D.G. Bunds, P.J. Lloyd, A multilayer perceptron netwotk for the diagnosis of low back pain, IEEE Int. Conf. Neur. Net 2 (1988) 481–489.

44. G. Cybenko, Approximation by superpositions of a sigmoidal function, Math Control Signals Syst. 2 (4) (1989) 303–314.

45. W.C. Carpenter, Understanding Neural network approximations and polynomial approximations helps neural network performance, AI Expert March (1995) 31–33.

46. A. Lehman, JMP for Basic Univariate and Multivariate Statistics: A Step-By-Step Guide, SAS Institute, 2005.

47. D.I. Filoti, S.J. Shire, S. Yadav, T.M. Laue, Comparative study of analytical techniques for determining protein charge, J. Pharm. Sci. 104 (7) (2015 Jul) 2123–2131.

48. Y.B. Ruiz-Blanco, W. Paz, J. Green, Y. Marrero-Ponce, ProtDCal : A program to compute general-purpose - numerical descriptors for sequences and 3D-structures of proteins, BMC Bioinf. 16 (16) (2015) 162.

49. G. Box, Science and statistic, J. Am. Stat. Assoc. 05 (791) (1976) 799.

50. R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, Pedreschi Dino, A survey of methods for explaining black box models, ACM Comput. Surv. (CSUR) 51 (5) (2018) 93.

51. C. Pratt, K. Cornely, Essential Biochemistry, third ed., Wiley, 2004.

52. W. Vogt, Oxidation of methionyl residues in proteins: tools, targets, and reversal, Free Radic. Biol. Med. 18 (1) (1995) 93–105.

53. S.C. Devanaboyina, S.M. Lynch, R.J. Ober, S. Ram, D. Kim, A. Puig-Canto, S. Breen, S. Kasturirangan, S. Fowler, L. Peng, H. Zhong, L. Jermutus, H. Wu, C. Webster, E.S. Ward, C. Gao, The effect of pH dependence of antibody-antigen interactions on subcellular trafficking dynamics, MAbs 5 (6) (2013) 851–859.

54. J.N. Israelachvili, Intermolecular and Surface Forces, third ed., Elsevier, 2011.

55. L. Nicoud, M. Owczarz, P. Arosio, M. Morbidelli, A multiscale view of therapeutic protein aggregation: A colloid science perspective, Biotechnol. J. 10 (3) (2015) 367–378.

56. T. Laue, Proximity energies: a framework for understanding concentrated solutions, J. Mol. Recognit. 25 (3) (2012) 165–173.

57. C. Schroeter, R. Guenther, L. Rhiel, S. Becker, L. Toleikis, A. Doerner, J. Becker, A. Schoenemann, D. Nasu, B. Neuteboom, H. Kolmar, B. Hock, A generic approach to engineer antibody pH-switches using combinatorial histidine scanning libraries and yeast display, MAbs 7 (1) (2015) 138–151.

58. J. Chaparro-Riggers, H. Liang, R.M. DeVay, L. Bai, J.E. Sutton, W. Chen, T. Geng, K. Lindquist, M.G. Casas, L.M. Boustany, C.L. Brown, J. Chabot, B. Gomes, P. Garzone, A. Rossi, P. Strop, D. Shelton, J. Pons, A. Rajpal, Increasing serum halflife and extending cholesterol lowering in vivo by engineering antibody with pHsensitive binding to PCSK9, J. Biol. Chem. 287 (14) (2012) 11090–11097.

59. N. Gera, A.B. Hill, D.P. White, R.G. Carbonell, B.M. Rao, Design of pH sensitive binding proteins from the hyperthermophilic Sso7d scaffold, PLoS One 7 (11) (2012) e48928.

60. T. Igawa, S. Ishii, T. Tachibana, A. Maeda, Y. Higuchi, S. Shimaoka, C. Moriyama, T. Watanabe, R. Takubo, Y. Doi, T. Wakabayashi, A. Hayasaka, S. Kadono, T. Miyazaki, K. Haraya, Y. Sekimori, T. Kojima, Y. Nabuchi, Y. Aso, Y. Kawabe, K. Hattori, Antibody recycling by engineered pH-dependent antigen binding improves the duration of antigen neutralization, Nat. Biotechnol. 28 (11) (2010) 1203–1207.

61. M.V. Kulkarni, M.C. Tettamanzi, J.W. Murphy, C. Keeler, D.G. Myszka, N.E. Chayen, E.J. Lolis, M.E. Hodsdon, Two independent histidines, one in human prolactin and one in its receptor, are critical for pH-dependent receptor recognition and activation, J. Biol. Chem. 285 (49) (2010) 38524–38533.

62. K. Maeda, Y. Kato, Y. Sugiyama, pH-dependent receptor/ligand dissociation as a determining factor for intracellular sorting of ligands for epidermal growth factor receptors in rat hepatocytes, J. Control Release 82 (1) (2002) 71–82.

63. D.C. Roopenian, S. Akilesh, FcRn: the neonatal Fc receptor comes of age, Nat. Rev. Immunol. 7 (9) (2007) 715–725.

64. D.B. Tesar, P.J. Bjoerkman, An intracellular traffic jam: Fc receptor-mediated transport of immunoglobulin G, Curr. Opin. Struct. Biol. 20 (2) (2010) 226–233

65. F.E. Ildiko, J.H. Friedman, A statistical view of some chemometrics regression tools, Technometrics 35 (2) (1993) 109–135.

7 Supplementary information

List of supplementary information

SI 1. Table of Identity and similarity scores

SI 2. List of the input parameters with corresponding statistics.

SI 3. Exemplary picture of applied network architectures.

SI 4. ANNs' Tm models results of the 15 different training sets.

SI 5. ANNs' Tagg models results of the 15 different training sets.

SI 6. List of the molecular descriptors calculated by ProDCal.

Table SI 1. Table of identity and similarity scores. *Identity and similarity scores, respectively in red and yellow cells, from the primary sequences of the heavy chains, light chains, and the complete mAb with the relative statistics. The similarity is considered as: GAVLI, FYW, CM, ST, KRH, DENQ, P, where the single letter represents the standard single letter amino acid code. The identity scores were calculated by the Sequence Manipulation Suite (Stothard P (2000) The Sequence Manipulation Suite: JavaScript programs for analyzing and formatting protein and DNA sequences. Biotechniques 28:1102-1104).*

Score Legend: Similarity identity

Heavy Chain (HC)						
PPI-1	100%	15.36%	88.69%	29.94%	17.29%	23.09%
PPI-2	8.46%	100%	16.70%	14.69%	25.16%	42.15%
PPI-3	86.25%	10.24%	100%	32.73%	17.73%	29.14%
PPI-10	18.04%	9.27%	27.39%	100%	16.03%	29.14%
PPI-13	7.98%	18.48%	8.86%	9.35%	100%	16.14%
PPI-17	15.47%	37.21%	23.76%	23.31%	11.21%	100%
	PPI-1	PPI-2	PPI-3	PPI-10	PPI-13	PPI-17
Light chain (LC)						
PPI-1	100%	11.73%	13.08%	13.55%	12.61%	11.62%
PPI-2	8.45%	100%	23.94%	23.00%	23.94%	24.88%
PPI-3	7.94%	18.30%	100%	94.39%	95.79%	48.59%
PPI-10	8.41%	18.43%	91.58%	100%	94.85%	44.85%
PPI-13	7.94%	18.77%	92.05%	93.92%	100%	45.79%
PPI-17	7.90%	16.43%	44.39%	38.78%	38.78%	100%
	PPI-1	PPI-2	PPI-3	PPI-10	PPI-13	PPI-17
mAb						
PPI-1	100%	14%	51%	22%	15%	17%
PPI-2	8.46%	100%	20%	19%	25%	34%
PPI-3	47.10%	14.27%	100%	64%	57%	39%
PPI-10	13.23%	13.85%	59.49%	100%	55%	37%
PPI-13	7.96%	18.63%	50.46%	51.64%	100%	31%
PPI-17	11.69%	26.82%	34.08%	31.05%	25.00%	100%
	PPI-1	PPI-2	PPI-3	PPI-10	PPI-13	PPI-17

Statistic	HC	LC	mAb	HC	LC	mAb
Minimum	7.98%	7.90%	7.96%	15%	12%	14%
Maximum	86.25%	93.92%	59.49%	89%	96%	64%
Mean	21.02%	34.14%	27.58%	28%	39%	28%
Std deviation	19%	31%	17%	18%	31%	16%
Variance	4%	10%	3%	4%	10%	3%

Table SI 2. *List of the input parameters with corresponding statistics. Input considered as discrete are only listed and no statistics are applied. To the right it is highlighted if the input is implemented to predict the corresponding protein stability indicator.*

Input parameters relative to the mAbs						
Amino acid	**Code**	**Minimum**	**Maximum**	**Standard deviation**	**Variance**	**Mean**
Alanine	Ala	64	80	5.62	31.56	69.33
Cysteine	Cys	30	38	2.75	7.56	32.67
Aspartic acid	Asp	52	62	3.54	12.56	54.33
Glutamic Acid	Glu	58	68	3.77	14.22	62.67
Phenylalanine	Phe	38	54	5.22	27.22	45.67
Glycine	Gly	82	98	5.63	31.67	91.00
Histidine	His	18	26	2.75	7.56	23.33
Isoleucine	Ile	28	36	2.52	6.33	31.00
Lysine	Lys	76	96	6.30	39.67	89.00
Glutamine	Glu	88	108	6.26	39.22	97.67
Methionine	Met	8	16	3.06	9.33	12.00
Asparagine	Asn	44	52	2.69	7.22	48.33
Proline	Pro	88	106	5.85	34.22	94.67
Glutammine	Gln	54	66	4.23	17.89	59.67
Arginine	Arg	30	50	6.43	41.33	38.00
Serine	Ser	158	188	10.13	102.67	172.00
Threonine	Thr	98	120	7.61	57.89	109.67
Valine	Val	110	120	3.14	9.89	115.67
Tryptophan	Trp	20	26	2.24	5.00	23.00
Tyrosine	Tyr	52	64	4.27	18.22	58.67
Input parameters relative to the formulation	**List**					
pH	5, 5.5, 6, 6.5, 7, 7.5, 8, 9	-	-	-	-	-
[NaCl] (mM)	0, 70 , 140	-	-	-	-	-

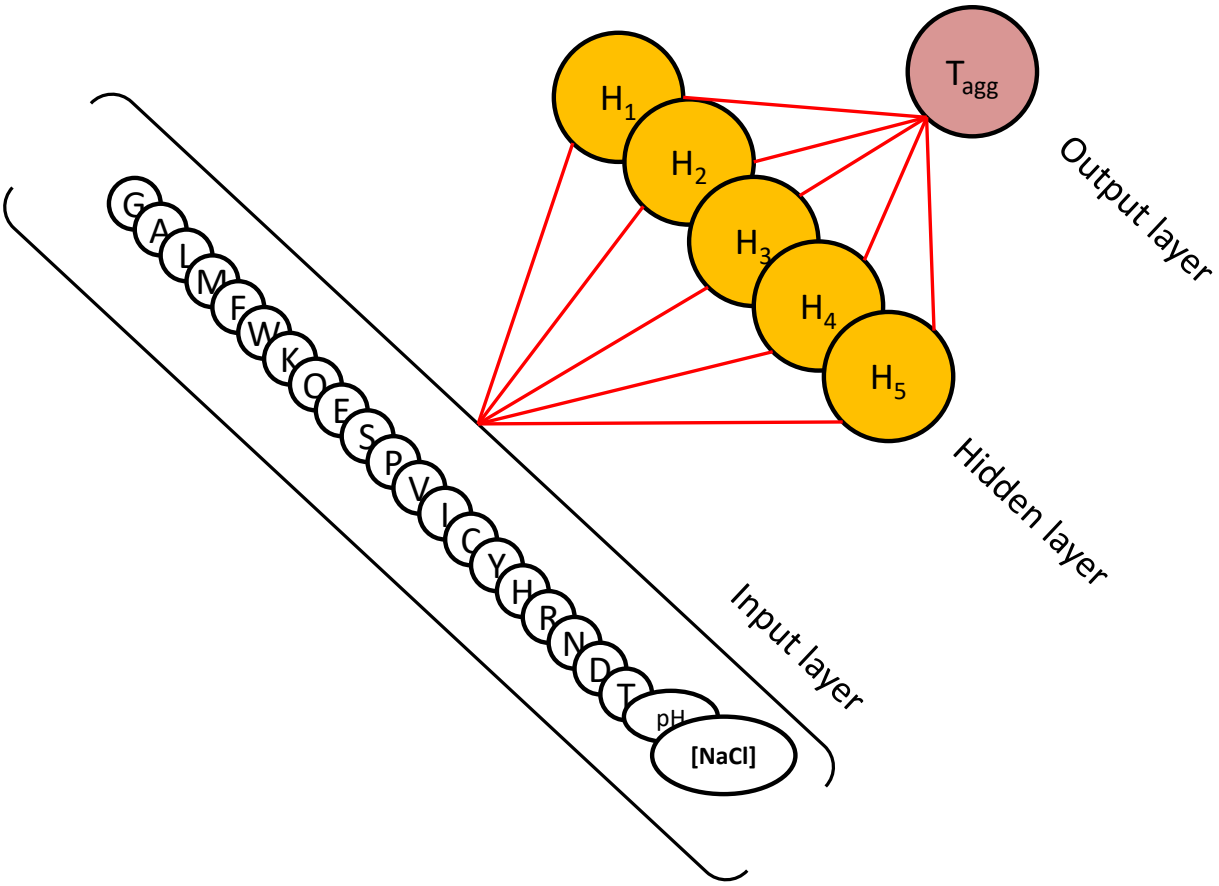


Figure SI 3. Exemplary picture of applied network architectures. *The brackets containing the input layer represent a complete connection of the input layer with the hidden one (i.e. each input is connected with all the neurons of the hidden layer).*

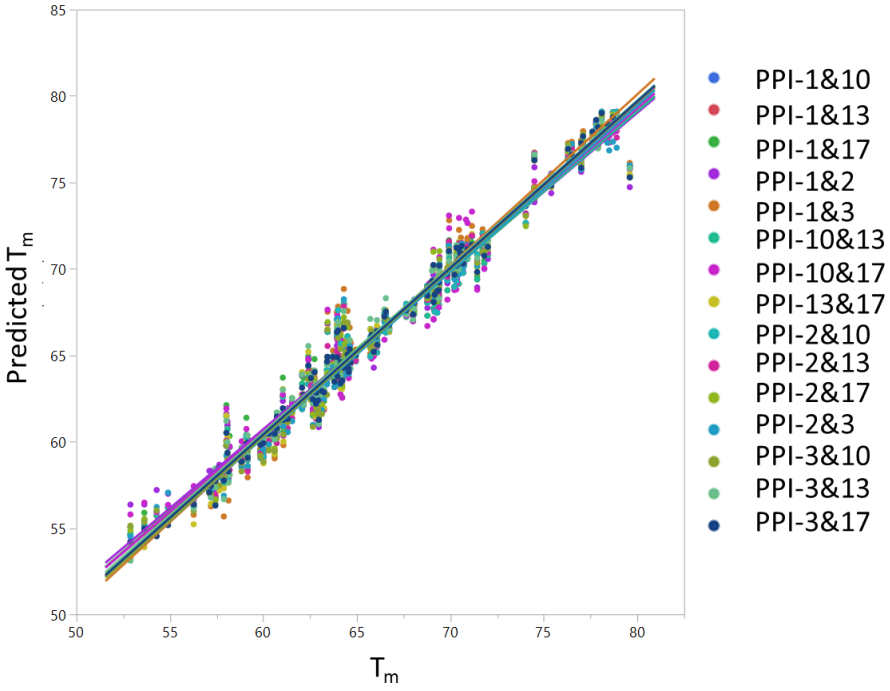


Figure SI 4. ANNs' T_m models results of the 15 different training sets.

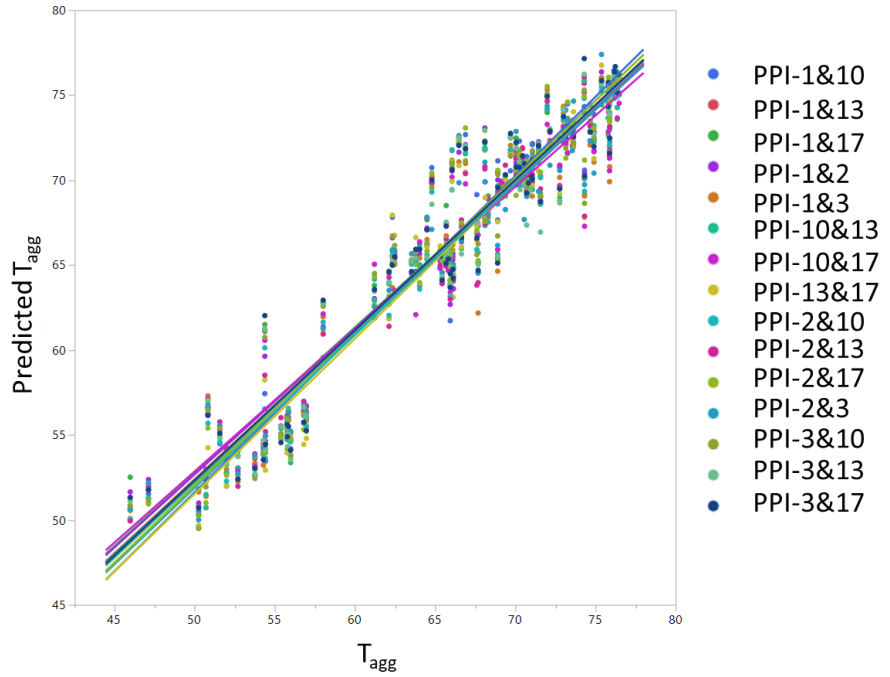


Figure SI 5. ANNs' T_{agg} models results of the 15 different training sets.

Table SI 6. List of the molecular descriptors calculated by ProDCal. *The description of the molecular indices can be found in the relative software manual.*

ProtDCal	ProtDCal
dGc(F)	wRWCO
dGw(F)	wdHBd
Gs(F)	wLCO
W(F)	wCo
HBd	wFLC
dGs	wPsiH
dGw	wPsiS
dGel	wPSil
dGLJ	Psi
dGtor	wR2
Gs(U)	wPjiH
Gw(U)	wPhiS
W(U)	wPhil
Mw	Phi
Ap	LnFD
Ecl	wCLQ
HP	wCTP
IP	wSP
ISA	WNc
Pa	Ap
Pb	dA
Pa	dAnp
Pt	WNLC
z1	wFLC
z2	wR2
z3	lnFD
dHf	
Xi	
L1-9	

CHAPTER IV: Characterization of native reversible self-association of a monoclonal antibody mediated by Fab-Fab interaction

Lorenzo Gentiluomo[1,2,5*], Dierk Roessner[1], Werner Streicher,[3] Sujata Mahapatra, [3] Pernille Harris,[4] Wolfgang Frieß[2]

[1] Wyatt Technology Europe GmbH, Hochstrasse 18, 56307 Dernbach, Germany

[2] Ludwig-Maximilians-Universitaet Muenchen, Department of Pharmacy, Pharmaceutical Technology and Biopharmaceutics, Butenandtstrasse 5, 81377 Munich, Germany

[3] Novozymes A/S, Krogshoejvej 36, 2880, Bagsvaerd, Denmark

[4] Technical University of Denmark, Department of Chemistry, Kemitorvet 207, 2800 Kongens Lyngby, Denmark

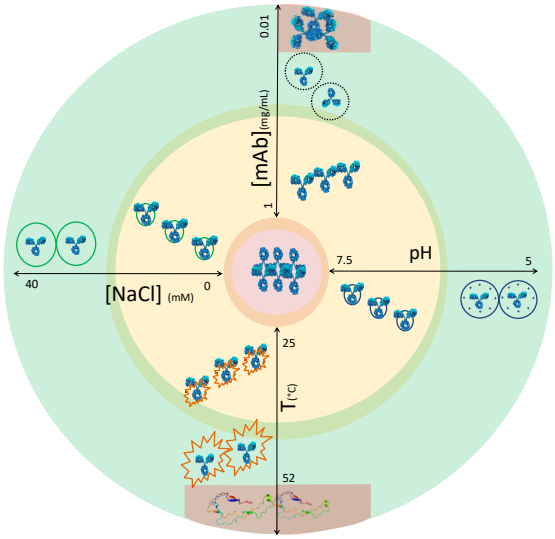
[5] Present address: Coriolis Pharma Research GmbH, Fraunhoferstraße 18B, 82152 Planegg, Germany

*denotes corresponding author

CONTRIBUTION: **Lorenzo Gentiluomo**, Wolfgang Frieß and Dierk Roessner conceived and designed the study. **Lorenzo Gentiluomo** conducted and analyzed the AUC measurements with the supervision of Werner Streicher. Sujata Mahapatra conducted the SAXS experiments. Sujata Mahatrapa and Pernille Harris analyzed SAXS data. **Lorenzo Gentiluomo** conducted SEC-MALS, AF4-MALS, DLS, nanoDSF, SLS, RP-HPLC experiments and analyzed the data. **Lorenzo Gentiluomo** performed RSM and pKa calculations. **Lorenzo Gentiluomo** and Wolfgang Frieß rationalized the data. **Lorenzo Gentiluomo** and Wolfgang Frieß contributed to the writing of the manuscript.

Abstract

The native reversible self-association of monoclonal antibodies has been associated with high viscosity, liquid-liquid, and liquid-solid phase separation. We investigated the native reversible self-association of an IgG1, which exerts this association even at low protein concentrations, in detail to gain further understanding of this phenomenon by extensive characterization of the association as a function of multiple factors, namely pH, temperature, salt concentration, and protein concentration. The nature of the self-association of the full-length IgG1 as well as the corresponding Fab and Fc fragment was studied by viz. size exclusion chromatography combined with multiangle light scattering, batch dynamic and static light scattering, analytical ultracentrifugation, small angle X-ray scattering, asymmetric flow field flow fractionation coupled with multiangle light scattering, and intrinsic fluorescence. We rationalized the self-association as a combination of hydrophobic and electrostatic interactions driven by the Fab fragments. Finally, we investigated the long-term stability of the IgG1 molecule. The native reversible self-association of monoclonal antibodies has been associated with high viscosity, liquid-liquid and liquid-solid phase separation. We investigated the native reversible self-association of an IgG1, which exerts this association even at low protein concentrations, in detail to gain further understanding of this phenomenon by extensive characterization of the association as a function of multiple factors, namely pH, temperature, salt concentration and protein concentration. The nature of the self-association of the full-length IgG1 as well as the corresponding Fab and Fc fragment was studied by viz. SEC-MALS, DLS, SLS, AUC, SAXS, AF4-MALS and intrinsic fluorescence. We rationalized the self-association as a combination of hydrophobic and electrostatic interactions driven by the Fab fragments. Finally, we investigated the long-term stability of the IgG1 molecule.



Graphical abstract. *Characterization of native reversible self-association of a monoclonal antibody mediated by Fab-Fab interaction*

1 Introduction

Humanized monoclonal antibodies (mAbs) have become major biopharmaceutical products in the last decades for the treatment of cancer and autoimmune diseases,[1-3] with over 50 of such molecules approved for therapeutic use.[4] The development of therapeutic antibodies is not straightforward, because, like other proteins, they are prone to physical instabilities and chemical degradation that need to be overcome.[5] Among these degradation pathways, protein aggregation presents one of the most common and troubling challenges. It has been shown that aggregates have reduced activity and potential for greater immunogenicity.[6-8] Moreover, it has been proven that nonnative protein aggregation is at least partially the cause of several diseases,[9] among which there are Alzheimer's disease, Parkinson disease, prion diseases, Huntington's disease, Down's syndrome, cataract, and sickle cell disease. Owing to this prevalence in pathology and therapeutic protein manufacturing, the study of the kinetics and mechanisms of protein aggregation is vital to future treatments. The critical review by Roberts[10] presents an excellent resource on the detailed steps of protein aggregation. In this study, a clear distinction is delineated between nonnative and native protein aggregation. Native self-association has been investigated in several studies,[11-13] and hydrophobic intermolecular interactions are considered to be the main driver.[14] Native reversible self-association has been associated with unusually high viscosity of concentrated mAb solutions in low ionic strength buffers which is one of the greatest challenges when developing protein formulations at higher concentrations.[11,15] It has been shown that the self-association of human myeloma protein causes the increase in viscosity through a combination of hydrophobic interactions.[16] The self-association of monoclonal serum immunoglobulins has been connected with hyperviscosity syndromes.[17-19] However, there are limited numbers of analytical methods that provide information at high protein concentrations.[20-22] This hampers an in depth characterization of the association. In this study, we characterized and rationalized the behavior of an IgG1, named PPI-1, which shows native reversible self-association at low ionic strength. PPI-1 exerts this association even at low protein concentrations, which allowed us to apply an extended analytical toolbox. We investigated how pH, salt concentration, protein concentration, and temperature impact the degree of oligomerization. Additionally, PPI-1 was digested into its Fab and Fc fragments to identify which regions of the mAb are involved in the oligomer formation. Previous studies have localized the origin of native self-association in either the Fab and the Fc fragments.[11,12] Accelerated and long-term stability studies were executed to understand the impact of the native reversible self-association on protein aggregation upon storage. Thus, this study leads to a better understanding of the underlying mechanism of self-association that in turn helps to design antibodies that are less prone to association.[23]

2 Material and methods

2.1 Sample preparation

The behavior of one IgG1, namely PPI-1, provided by the PIPPI consortium (http://www.pippi.kemi.dtu.dk) was characterized. PPI-1 was dialyzed using 10 kDa Slide-A-Lyzer™cassettes (Thermo Fisher Scientific, MA) into 10 mM histidine buffer pH 6.0 to obtain a 35 mg/mL solution. Then, PPI-1 was dialyzed into 10mM histidine buffer pH 5.0, 5.5, 6.5, 7.0, and 7.5. NaCl stock solution in the respective buffer was added to reach 0, 70, or 140 mM. Protein concentration was measured on a Nanodrop 2000 (Thermo Fisher Scientific) using the extinction coefficient calculated from the primary sequence. All chemicals were of analytical grade and were purchased from Sigma Aldrich (Germany) or VWR International (Germany). Highly purified water (HPW, Purelab Plus; USF Elga, Germany) was used for the preparation of all buffers. Finally, the formulations were sterile filtered with a low protein binding Millex® 0.22 µm filter (Merck Millipore, Burlington, MA).

2.2 Preparation and purification of Fab and Fc fragments

Immobilized Papain (Thermo Fisher Scientific, USA) was used to digest PPI-1 into its Fab and Fc fragments. PPI-1 at 20 mg/mL was pipetted into 15 mL glass vial, and the vial was capped with the resin separator provided with the kit to remove all the air-liquid interface. The vial was gently rotated by a Sunlab rotator SU1100 for 5 h at 37° C. An ÄKTA purifier 10 (GE Healthcare, Sweden) equipped with a Pierce Protein A chromatography cartridge (Thermo Fisher Scientific, USA) (column volume, CV = 5 mL) was used to separate Fc (and undigested mAb) from the Fab fragments. The binding buffer was composed of 100 mM sodium phosphate with 150 mM NaCl at pH 7.2. The column was equilibrated with 2 column volumes (CV) of binding buffer with a flow of 2 mL/min. Fractions were collected in 15-mL PP tubes using a Frac 920 fraction collector (GE Healthcare, Sweden) capturing any unbound species (e.g. Fab). The elution buffer was kept at 100% over 7 CV. The eluting protein was collected in 15-mL PP tubes using the fraction collector and was immediately neutralized with a 1 M sodium phosphate buffer pH 8.5. Ultrafiltration was performed using Vivaspin® tubes with a 10 kDa molecular weight (Mw) cut-off PES membrane (Sartorius Stedim Biotech, Germany). Success of the purification was monitored by HP-SEC (see section 2.4). Finally, different formulations of Fab and Fc were prepared as described for the mAb in 2.1.

2.3 Long term stability study

0.2 mL of each protein solution was aliquoted at a concentration of 1 mg/mL and filtered through a 0.22 um Miller® GV filter (Merck Millipore, USA) under a laminar flow hood into 0.5 mL sterile non-coated PP Eppendorf tubes. The samples were incubated at 4 °C, 25 °C, and 40 °C for 6 months.

2.4 Size exclusion chromatography coupled with multi-angle light scattering

Samples were investigated using size exclusion chromatography combined with multiangle light scattering (SEC-MALS). The system consisted of a Vanquish Horizon™ UPLC with a variable wavelength UV detector operated at 280 nm (Thermo Fischer Scientific, MA). The temperature controlled autosampler was kept at 4°C. The separation was performed with a Superdex 200 increased 10/300 GL column. Concentration was determined immediately after the column by a UV detector operated at 280nm which was followed by a static light scattering (SLS) apparatus, a TREOS II detector (Wyatt Technology, USA) followed by additional concentration verification by a differential refractive index detector (Optilab T-rEX; Wyatt Technology, USA). Data collection and processing were performed using the ASTRA software V7.2 (Wyatt Technology, USA). The mobile phase consisted of 38 mM NaH_2PO_4, 12 mM Na_2HPO_4, 150 mM NaCl, and 200 ppm NaN_3 at pH 7.4 and was filtered through Durapore VVPP 0.1 mm membrane filters (Merck Millipore, USA). The samples were injected at a volume of 25 or 50 µL.

2.5 Asymmetric flow field flow fractionation coupled with multi-angle light scattering

Asymmetric flow field flow fractionation coupled with multiangle light scattering (AF4-MALS) was used to study PPI-1 oligomers using the respective sample formulations as carrier. The system consisted of a pump (Agilent 1260 Infinity II, Agilent Technologies, Germany) with an online degasser and a temperature-controlled autosampler kept at 4°C. The separation was performed with an AF4 system (ECLIPSE; Wyatt Technology, USA) using a short channel with 490 mm spacer and a precut regenerated cellulose membrane with 10 kDa cut-off (Wyatt Technology, USA). A focus flow of 1.5 mL/min was applied for 2 min before injection. The samples were injected with a flow of 0.2 mL/min maintaining the focus flow. Then, the samples were focused at 1.5 mL/min for 5 min. A constant detector flow of 1 mL/min was used during the separation process, which included 3 stages: (1) 3 mL/min cross flow, hold constant for 20 min; (2) linear ramped flow from 3 to 0.1 mL/min in 10 min; and (3) 0.0 mL/min flow, hold constant for 5 min. A TREOS II MALS detector (Wyatt Technology, USA), a variable wavelength detector operated at 280 nm (Agilent 1260 Infinity II Agilent Technologies, Germany), and a differential refractive index detector, Optilab T-rEX (Wyatt Technology, USA) were connected to the system. Data collection and processing were performed using the ASTRA software, V 7.2 (Wyatt Technology, USA). The respective

formulation of the sample (e.g., His 10 mM at pH 5) filtered through Durapore VVPP 0.1 mm membrane filters (Merck Millipore, USA) was used as aqueous mobile phase.

2.6 Dynamic and static light scattering

High throughput dynamic light scattering (DLS) and SLS were conducted on a DynaPro® III Plate Reader (Wyatt Technology, USA) to obtain the hydrodynamic radius (Rh) and the Mw. Four mL per well of each sample, filtered using a Millex®0.22 µm filter (Merck Millipore, USA), were pipetted in triplicates into Aurora 1536 Lobase Assay Plates (Aurora Microplates, USA). Wells were sealed with silicone oil and then centrifuged at 2000 rpm for 1 min. Data was processed by the DYNAMICS software V 7.8 (Wyatt Technology, USA). To calculate the Mw, the plate was calibrated with dextran 35-45 kDa(Sigma Aldrich, USA, Lot number: SLBQ5973V). Composition gradient MALS (CG-MALS) and DLS (CG-DLS) were used to determine the diffusion interaction parameter (k_D) using at least 10 different concentrations (from 1 to 10 mg/mL) in triplicate. The samples were equilibrated at 25° C for 10 min in the plate reader before any measurement. Temperature ramps were conducted at 1 mg/mL. Temperature of aggregation, T_{agg}, was processed by the DYNAMICS software V7.8 onset algorithm from the increase in the total scattering intensity. Detailed static and dynamic light scattering studies were conducted on a DynaPro® Nanostar (Wyatt Technology, USA). Two microliters of sample were pipetted into a quartz cuvette and sealed silicone oil. Measurements were conducted in triplicate.

2.7 Differential scanning fluorimetry

Differential scanning fluorimetry (DSF) was conducted using the Prometheus NT.48 (NanoTemper Technologies, Germany). Samples containing 1 mg/mL protein were filled in nanoDSF capillaries and analyzed. A temperature ramp of 1 °C/min from 20 °C to 95 °C was applied. The fluorescence intensity ratio (F350/F330) was plotted against the temperature, and the first apparent melting temperature (T_m) was derived from the maximum of the first derivative using the PR Control software V1.12 (NanoTemper Technologies, Germany). All measurements were performed in triplicate.

2.8 Analytical ultracentrifugation

Sedimentation velocity experiments were conducted in a Beckman XLI ultracentrifuge (Beckman Coulter Inc., USA) at 40,000 rpm at 20°C using the charcoal-filled Epon 12-mm double-sector centerpieces. The moving boundary was monitored by repetitive radial scanning at a constant step size of 0.003 cm at 280

nm using a UV absorption optical system. Sedimentation velocity data was analyzed, and simulation data was created using the software program SEDFIT (National Institutes of Health, USA)[24] to generate the sedimentation coefficient distribution of protein samples.

2.9 Small angle X-ray scattering (SAXS)

35 mg/mL of PPI-1 in 10 mM histidine buffer at pH 5.0, 5.5, 6.0 and 6.5 without salt was dialyzed over three shifts. The dialysate from the final buffer exchange was sterile filtered using a 0.22 um Miller® GV filter (Merck, Millipore, USA) and used for sample dilution and buffer measurements. SAXS experiments were performed at the ESRF synchrotron, BM29 bioSAXS beamline at Grenoble, France. Measurements of pure water were used to get the data on an absolute scale. Buffers were measured both before and after each sample and averaged before subtraction. A concentration range from 0.5mg/ml – 17mg/ml was measured for each formulation. Data collection parameters are listed in **Table SI 1**. Calibrations and corrections of SAXS data collected at ESRF were carried out by an automated pipeline.[25] Buffer averaging and subsequent subtraction prior to data analyses were performed in Primus.[26] The ATSAS program package version 2.8.4[27] was used for further data analysis. Primus was also used to perform Guinier region analysis and GNOM[28] was used for pair distribution, $p(r)$, analysis. The intensity, $I(q)$, is measured as a function of scattering vector $q = 4\pi\, sin\theta/\lambda$, where λ is the wavelength, and 2θ the scattering angle. Molecular weight calculations were performed using $M_w = [N_A I(0)/c]/\Delta\rho_M^2$ where N_A is Avogadro constant, $I(0)/c$ is concentration normalized forward scattering and $\Delta\rho_M$ is the scattering contrast per mass. $\Delta\rho_M$ was calculated using proteins average partial specific volume, $0.7425\ \mathrm{cm^3 g^{-1}}$.[29]

2.10 Reversed-phase ultra-high-performance liquid chromatography (RP-UPLC)

Reversed-Phase Ultra-High-Performance liquid chromatography (RP-UPLC) was conducted on an ACQUITY UPLC H-Class system (Waters, USA) equipped with a quaternary pump, an autosampler, and UV detector operated at 280 nm. The separation was performed with a Acquity BEH-300 C4 (Waters, USA). Ten microliters of sample were injected at a concentration of 0.7 mg/mL. Eluent A consisted of 10% w/v acetonitrile and 0.1% w/v trifluoracetic acid in ultrapure water. Eluent B consisted of 0.1% w/v trifluoracetic acid in acetonitrile. The flow rate was 0.2 mL/min. The column oven temperature was set at 75° C. A preheater was included before the column. A gradient of 25% to 40% eluent B in A in 20 min was used. The chromatograms were integrated in Empower V3 (Waters, USA). A mass recovery of 100% was always achieved.

3 Results

A full factorial design of experiments was applied to study PPI-1 (pI = 7.9). Response surface methodology was then used to study the interactions of the investigated factors. We assessed the full model and then reduced it to only the terms that were deemed statistically relevant. A curvature response was allowed by assessing the quadratic term considering also two-way interactions. The reduced model was obtained using a backward stepwise regression. The F-statistic approach was used to perform the effect test considering a value of 0.05 or less as statistically significant. These calculations were performed by the statistical software JMP® v 14.0 (SAS Institute Inc., Cary, NC). This approach was not successful to quantitatively separate the factors' effects related to the colloidal stability (e.g., degree of oligomerization, k_D, T_{agg}). This is due to an intrinsic nonlinearity of PPI-1 association, which yields to very high fit errors and low accuracy and therefore poor interpretability. Hence, we discuss the investigated factors separately in sections 4.1-4.4. Differently, the apparent T_m measured by intrinsic DSF (**Table SI 2**) present a good fit with R^2 and root mean square error (RMSE) of respectively 0.93 and 1.4. The pH presents a significant effect on the DSF profile resulting in higher apparent T_m at higher pHs, whereas NaCl concentration showed an almost insignificant effect in the concentration range investigated (p value > 0.05). Nonlinear methodologies have been proved successful in predicting biophysical properties of PPI-01 and other mAbs based on the amino acid composition, pH, and NaCl concentration.[30]

3.1 The pH effect

The sedimentation coefficient (S) of PPI-1 between pH 5 and 7.5 at 0.1, 0.5, and 1 mg/mL (**Fig. 1a**) was calculated from analytical ultracentrifugation (AUC) measurements. We observed an increasing amount and size of oligomers with increasing pH. A table with the weight-average sedimentation coefficient (S_w) is reported in **Table SI 3**. Sw is reported for all the peaks and calculated across the entire distribution. The S_w is a critical parameter used to understand protein self-association.[31] AUC results correlate well with both the Rh calculated by DLS (**Figs. 1b** and **1c**) and the Rg calculated by SAXS (**Fig. SI 4**). DLS results also indicate a step decrease of the oligomer size at pH 8 and 9 (**Fig. 1c**). More information regarding SAXS results can be found in **SI 3**. All the techniques confirm the presence of monodispersed solution of the monomer at pH 5. Physical separation of the oligomers was attempted by AF4-MALS. This technique was selected as the mobile phase can be matched to the exact formulation of the measured sample. AF4 chromatograms are shown in **Fig. SI 5**. By AF4 theory, R_h was calculated from the maximum of the eluting peak. The obtained R_h values correspond to the ones measured by DLS within the experimental error (**Fig. 1c**). However, the Mw calculated from MALS yields a consistent molecular weight of around 154 kDa indicating the presence of monomer Mw. This effect is probably due to the AF4 separation process. The molecules are pushed towards the membrane during separation and then diluted before

reaching the detector. Therefore, PPI-1 probably separates in its oligomeric form but rapidly equilibrates back to the monomer due to the dilution at the end of the channel. Further, PPI-1 self-interaction was investigated by the apparent k_D (**Table SI 2**). It was not possible to differentiate the interaction effect from the oligomerization effect on the light scattering signal. This is due to the fact that the abundance of the complexes will increase with increasing solute concentration, leading to an apparent negative value of kD.[32]

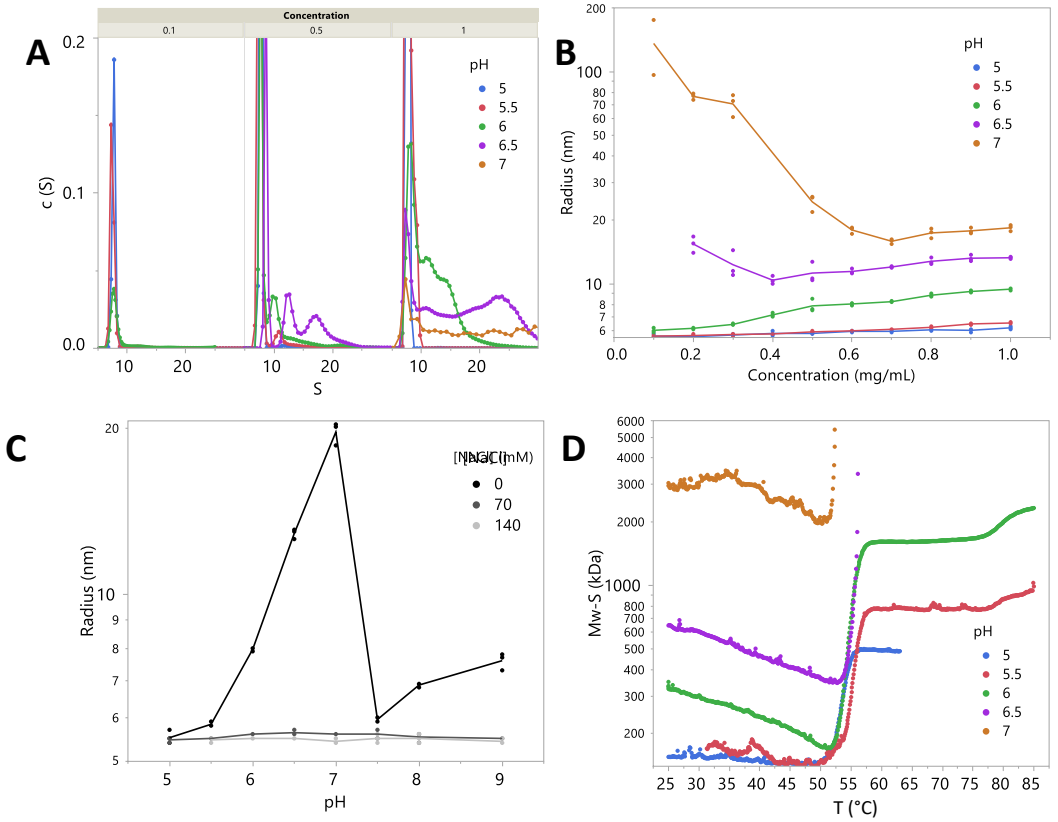


Figure 1. ***Selected studies on PPI-1 solution behavior.*** *Graph "a" shows the AUC sedimentation velocity results at protein concentrations of 0.1, 0.5, and 1 mg/mL. Missing data at lower concentrations are owing to the formation of very large, rapidly sedimenting particles. All formulations were investigated in 10 mM His from pH 5 to 7. The data points are depicted as shown in the legends. The same formulations and color codes are used for graph "b" and "d". Graph "b" shows the apparent Rh of PPI-1 as a function of protein concentration (logarithmic scale). Graph "c" shows the apparent Rh of PPI-1 at 1 mg/mL as a function of the pH. Different NaCl concentrations are depicted in scale of grays as shown on the legend. Graph "d" shows the temperature ramp curves of PPI-1 at 1 mg/mL measured by SLS as a function of pH. Similar curves for the apparent R_h were generated by DLS.*

3.2 The salt effect

Formulation with high salt content always yielded the apparent size of the monomer around 5.3 nm (**Fig. 1c**). Titration with sodium chloride was performed, and the sedimentation coefficient of PPI-1 at 1 mg/mL was analyzed by AUC at pH 6 and 6.5 (**Fig. 2, Table SI 3**). A broad distribution of oligomers was observed at zero NaCl concentration. A lower distribution of oligomers was observed at higher concentration of NaCl (**Fig. 2, Table SI 3**).

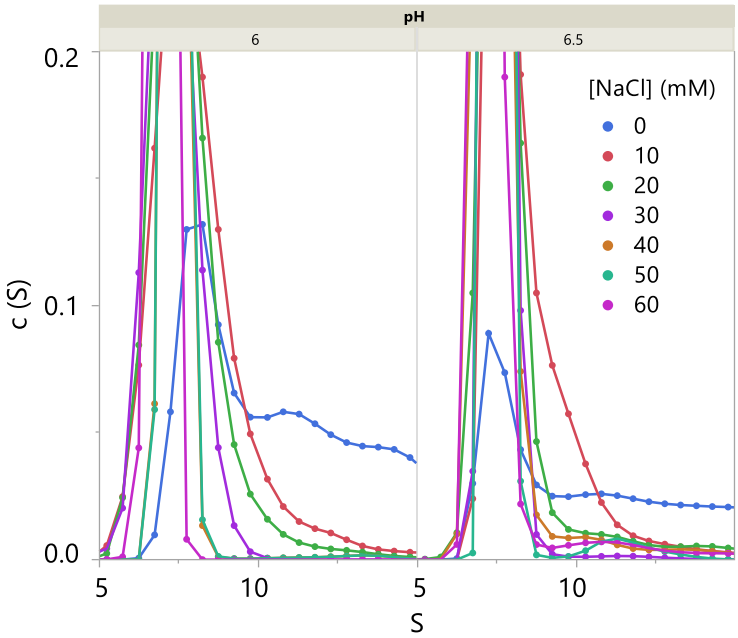


Figure 2. ***The effect of NaCl on the sedimentation coefficients of PPI-1 by sedimentation velocity AUC.*** *All formulations were investigated in 10 mM His at pH 6 and 6.5, which are showed on the left and right of the graph, respectively. The final concentration of NaCl is shown in different colors as depicted in the legend.*

The AUC results correlate well with other experimental techniques, which show a reduction of oligomers at increasing concentration of salt. However, at low ionic strength (<10 mM), the apparent radius of protein will increase (i.e., decreasing the sedimentation coefficient)[33] and may lead to quantitatively inaccurate interpretation of the results. SEC-MALS yielded no elution from the column when low salt eluents were used. However, high salt eluents (i.e., PBS as described in materials and methods) yielded 99.9% of monomer fraction for all the formulations studied. This behavior is due to the buffer exchange over the column, which shifts the equilibrium from oligomers to monomers. Further, the salt could influence the interactions between the SEC stationary phase and PPI-1, regardless of its oligomerization state. Nonetheless, column mass recovery correlates with the formation of oligomers, as highlighted by the R_h measured by DLS (**Fig. 1c**), where lower recovery (**Fig. SI 6**) is observed in formulation with higher R_h measured by DLS (**Fig. 1c**). Further, we investigated the stability of PPI-1 under isothermal stress

after long-term storage by SEC-MALS. PPI-1 in 4 different formulations was stored for 6 months at 4 °C, 25 °C, and 40 °C (**Fig. 3**). Taking into consideration only soluble aggregates, the lowest aggregate content is detected at pH 6.5. The addition of salt has a minor impact. On the other hand, at pH 5 a steep increase of aggregates is observed after 6 months of storage at 40 °C. The monomer loss, which includes both soluble aggregates and the mass lost to insoluble aggregates, indicates formation of the latter (observed also by visual inspection) at pH 5. As previously mentioned, the monomer loss in conditions where PPI-1 presents native self-association is possibly due to the precipitation of the oligomers (**Fig. SI 6**), which is also observed under unstressed conditions.

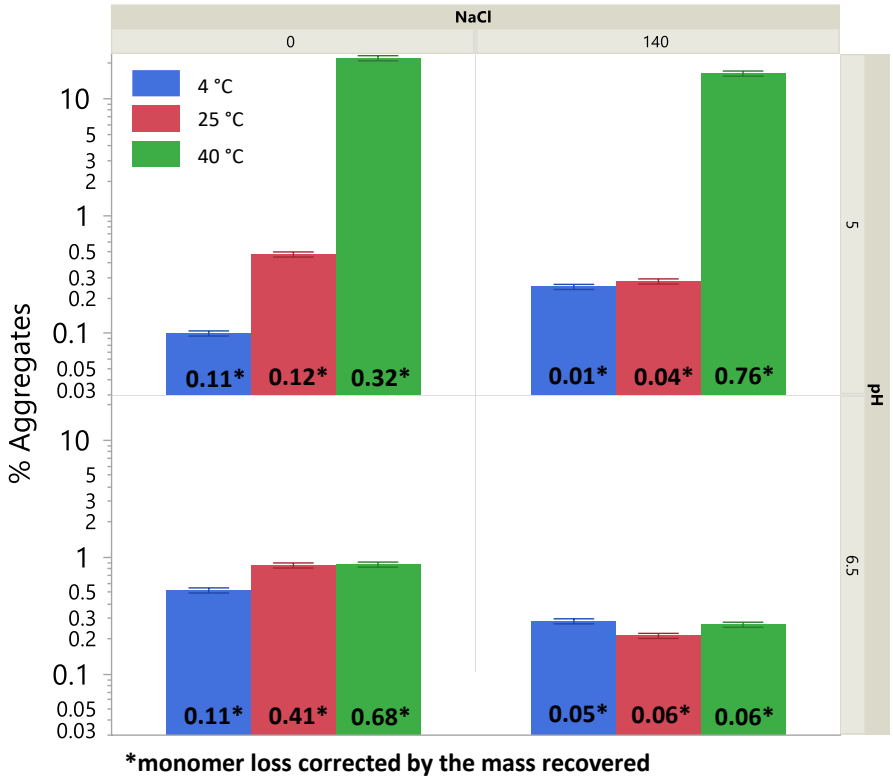


Figure 3. ***Long-term stability results.*** *Each graph represents 1 of the 4 different formulations stored for 6 months at 4 °C, 25 °C, and 40 °C depicted in blue, red, and green, respectively. The percentage of aggregates is calculated by the UV signal at 280 nm. The error bars are calculated from the analytical error. The numbers on each bar represent the calculated monomer loss corrected by the recovered mass, where 1 stands for complete monomer loss and 0 stands for no monomer loss. This is calculated including into the calculation the initial recovered mass (i.e., before stress) divided by the calculated recovered mass.*

3.3 The temperature effect

The temperature effect on PPI-1 self-association was studied by means of light scattering experiments with temperature ramps (**Fig. 1d**). We observed 2 mechanisms of aggregation: (1) irreversible nonnative aggregation (e.g. pH 5), (2) reversible native disassociation with subsequent unfolding, which leads to irreversible nonnative aggregation. Formulation including high salt concentration (**Fig. SI 7**) always presented nonnative aggregation. We confirmed the reversibility of the first step of the second mechanism by temperature cycles between 0 °C and 45 °C (**Fig. 4**). The association/dissociation is fully reversible between 0 °C and 20 °C. Upon ramping from 25 °C to 45 °C, a very small amount of PPI-1 of around 0.5% appears to irreversibly aggregate in each cycle, which is probably due to a small population of partially unfolded molecules formed with each cycle. Similarly, we confirmed the irreversibility of the second step by cycling the temperature in the range of 45 °C to 55 °C (**Fig. SI 8**). We noticed that PPI-1 irreversibly aggregates if the temperature reached the onset temperature of unfolding measured by nanoDSF (**Table SI 2**).

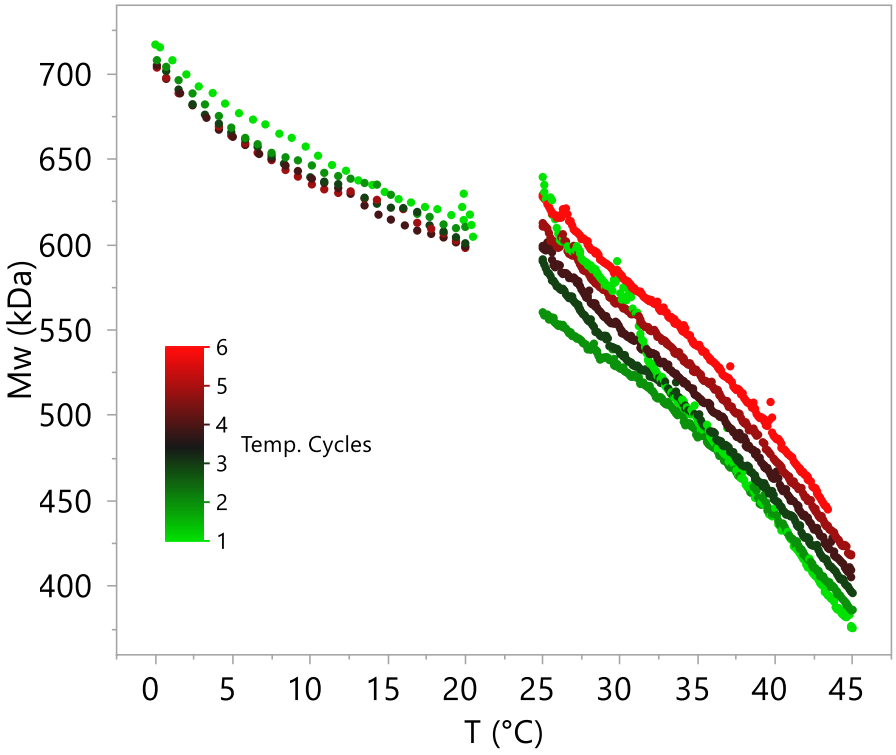


Figure 4. ***An example of PPI-1 temperature cycle data generated by SLS for 1 mg/mL of PPI-1 in 10 mM His at pH 6.5.*** *The figure shows 2 temperature cycle experiments performed between 0° C and 20° C and between 25 °C and 45 °C respectively on the left and right of the graph. 1. The Mw color is based on the cycle number. The first temperature ramp is depicted in bright green while the last in bright red. At the end of each cycle, PPI-1 was rapidly cooled (data was not collected during the cooling phase).*

3.4 The protein concentration effect

PPI-1 formed more and larger oligomers with increasing concentration in the range of 1 to 20 mg/mL as detected by DLS (**Fig. 1b**), SLS (**Fig. 1d**), AUC (**Fig. 1a**) and SAXS (**Fig. SI 3**). DLS indicates the formation of large particles at low concentration at pH 6.5 and 7 (**Figs. 1b-1d**). These particles rapidly sediment during AUC experiments. Therefore, we further investigated PPI-1 at pH 6 without salt in diluted samples by DLS (**Fig. 5**). Three different slopes for the linear correlation between the apparent diffusion and PPI-1 concentration were observed. Between 0.1 and 0.4 mg/mL, the diffusion coefficient linearly increases with concentration. From 0.4 to 2 mg/mL, the diffusion coefficient shows a steep negative linear dependency on concentration followed by a more gradual decrease with concentration above 3 mg/mL.

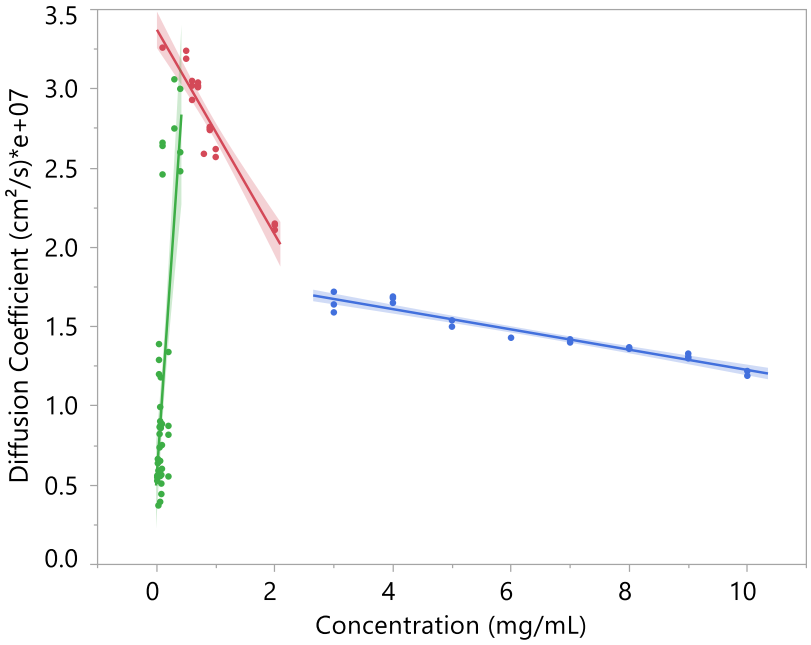


Figure 5. PPI-1 Diffusion coefficient as a function of the concentration. *Three linear regions are identified for PPI-1 formulated in 10 mM Histidine at pH 6. The corresponding fits and confidence intervals are shown in different colors.*

3.5 Fab and Fc fragments studies

A fractional design of experiment was applied to PPI-1 fragments due to material limitations. We focused on solutions where oligomers were present, which is low ionic strength. The absence of oligomers in formulation including salt for PPI-1 fragments was confirmed by DLS and SLS. pKa-based calculations of the pI of the whole mAb, its Fab, and its Fc fragment yielded 7.9, 6.2, and 8.4, respectively. Therefore, we hypothesized that the native reversible self-association may be caused by hydrophobic patches of the Fab fragment. Other low volume techniques (e.g. CG-DLS) were not successful to investigate Fab-Fab,

Fab-Fc, and Fc-Fc association. In fact, as previously mentioned in the case of the whole mAb, it is not possible to distinguish the interaction effect from the oligomerization effect on the light scattering signal. However, oligomers were observed only in the presence of the Fab fragment. To prove that the reversibility of the process is equal to the one of the whole mAb, we exploited PPI-1 behavior by temperature cycling as described in section 4.3. We confirmed our hypothesis studying the whole mAb, the Fc, and Fab fragments by SLS at pH 5 and 6 (**Fig. 6**).

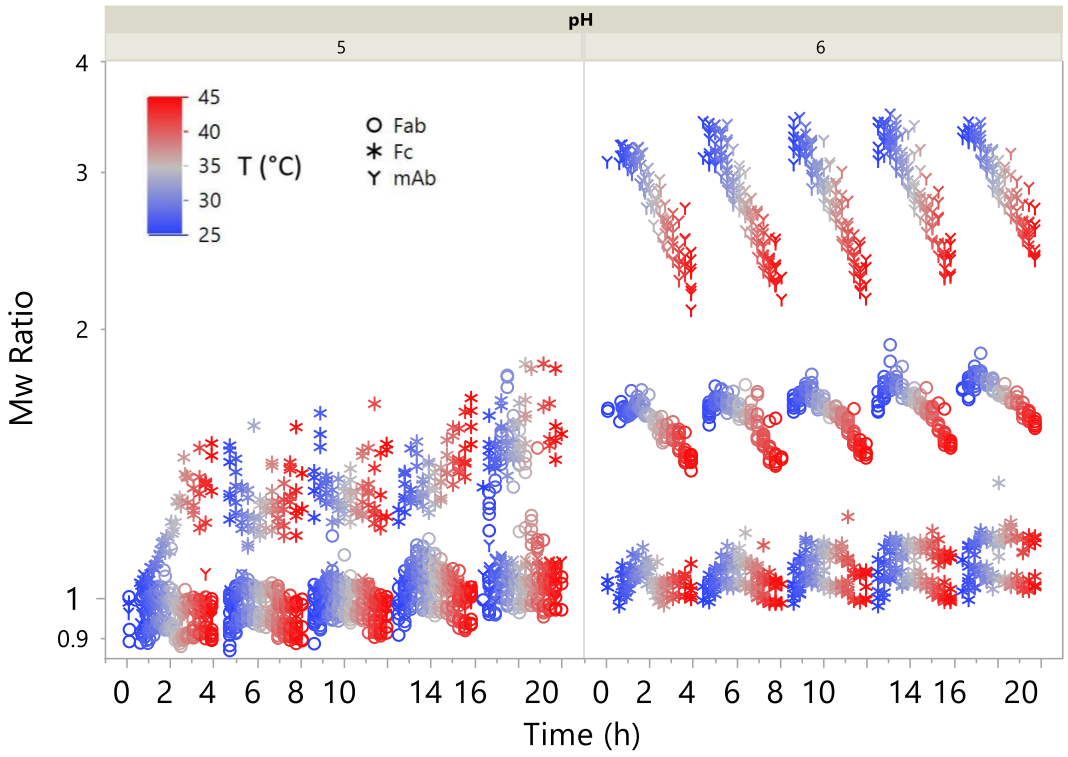


Figure 6. ***Temperature cycles investigated by SLS of the intact mAb, Fab, and Fc.*** *Temperature cycles from 25 °C to 45 °C are shown for all the samples at pH 5 and 6 (10 mM His) on the left and right of the graph, respectively. The Mw ratio is calculated dividing the measured Mw by the Mw detected at 25C. mAb, Fab, and Fc are respectively depicted as shown in the legend.*

Upon a series of temperature cycles at pH 6, self-(dis)associating oligomers for both the whole mAb and its Fab fragment were observed. On the other hand, the Fc fragment was present in its monomeric form over the cycles. At pH 5 no self-association is observed. The mixtures of Fab-Fc, Fc-PPI-1, and Fab-PPI-1 were similarly investigated; however, as the Fab or PPI-1 oligomers are responsible for most of the light scattering signal, results from this experiment are similar to the one showed in **Fig. 6**. Further, nonnative irreversible aggregation starts once that $T_{m,on}$ is reached (**Fig. SI 8**). Finally, we observed a considerable difference in the retention time measured by RP-UPLC among the whole mAb, the Fc, and the Fab fragment, which eluted at 2.9 mL, 2.58 mL, and 3.2 mL, respectively (**Fig. 7**). Thus, the Fab fragment showed a rather high degree of hydrophobicity compared to the Fc fragment and the whole mAb.

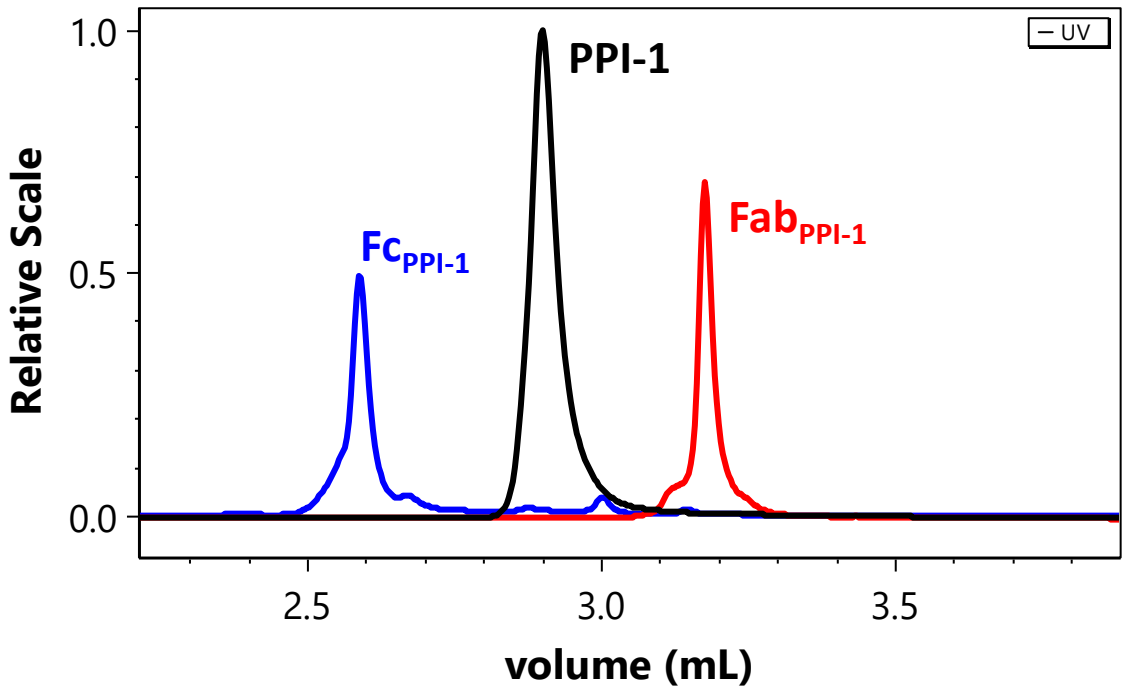


Figure 7. Reverse phase chromatograms. *In black, blue and red the results from the whole mAb, the Fc and the Fab fragments are respectively shown*

4 Discussion

4.1 Rationalization of PPI-1 native reversible self-association

Protein aggregation can occur through a number of distinct pathways that are not mutually exclusive. PPI-1 presents a reversible association of the native monomer that is intrinsic under certain solution conditions. In these conditions the surface of PPI-1 in the native structure is self-complementary and readily self-associates to form oligomers. Insulin is a typical example of therapeutic protein which forms reversible oligomers.[34] Conversely, the first step of irreversible aggregation is due to partial unfolding of the monomer, which acts as precursor of disordered oligomers.[35-42] We carried out long-term stability studies (**Fig. 3**) to differentiate these 2 mechanisms. We observed that PPI-1 does not form critical percentage of aggregates at room temperature or if refrigerated after 6 months of storage. On the other hand, high temperatures induce formation of nonnative aggregates at acidic pHs. It is known that acidic pHs lower the conformational stability of mAbs,[43] which most likely prompts the nonnative aggregation of PPI-1 stored at 40 °C. In fact, PPI-1 was demonstrated to have a lower apparent temperature of unfolding (T_m) at lower pHs (**Table SI 2**). Further, the addition of salt does not influence (at the low concentration used) the conformational stability of PPI-1. Therefore, the formulation with 10 mM histidine and 140 mM NaCl at pH 6.5 allows to both minimize the nonnative aggregation and eliminate PPI-1 native oligomers.

The latter has been observed to induce phase separation, precipitation, and high viscosity.[12,15] Several mAbs are known to have a tendency of intrinsically self-interacting, which prompts phase separation at high concentrations.[44-46] Therefore, we focused our investigation into the characterization of PPI-1 native reversible self-association. This process is schematically summarized in **Fig. 8**, and hereafter, the aim of the discussion is to rationalize this behavior.

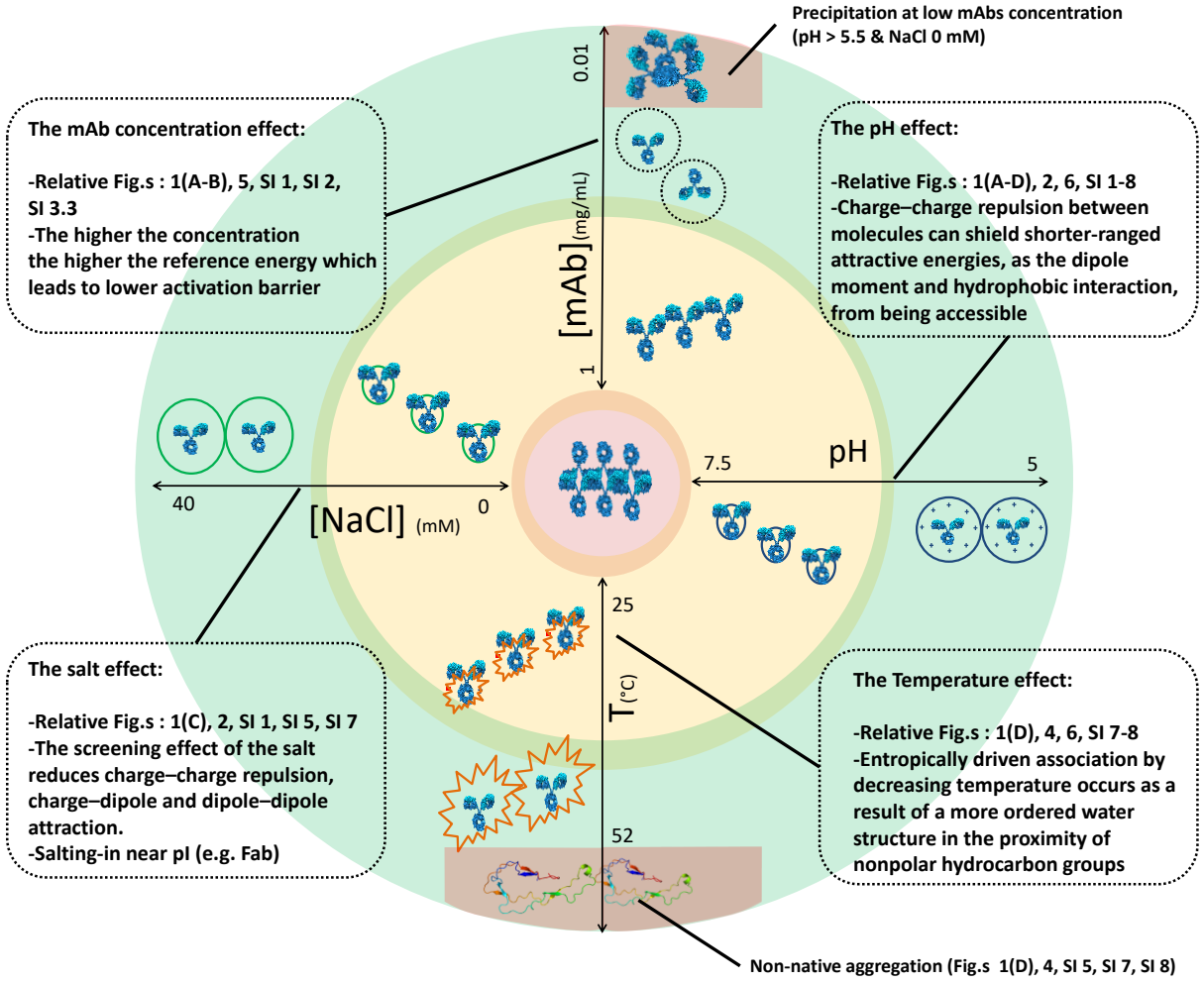


Figure 8. Graphical representation of PPI-1 self-association as a function of 4 factors: pH, Temperature, salt concentration and protein concentration. *The graph is indicative. The red, yellow and green areas represent respectively the presence of irreversible aggregates, the presence and the absence of native oligomers.*

It has been suggested that self-association at low ionic strength is due to electrostatic interaction.[47] We observed for PPI-1 that the association process is weakened at low pH values (**Fig. 1**), which is due to the increasing mAb net charge. High net charge prevents short range interactions from being accessible as described by the Derjaguin-Landau-Verwey-Overbeek (DLVO) or proximity energy theory.[48,49] PPI-1 salting-in effect (**Figs. 1c** and **3**), that is as the ionic strength increases, protein solubility increases, can

be rationalized by DLVO or proximity energy theory only if a very strong dipole moment is assumed. In fact, high ionic strength depletes electrostatic interactions such as charge-charge repulsion, charge-dipole, and dipole-dipole attraction. PPI-1 shows, at low ionic strength, an increase of R_h and Mw with increasing pH up to 7.5 followed by a steep decrease at pH 8 (**Fig. 1d**). pKa-based calculations of the pI of the whole mAb, its Fab, and its Fc fragment, yielded, respectively, 7.9, 6.2, and 8.3. This supports the hypothesis of a different local surface charge behavior of PPI-1. For example, a strong self-association is detected from pH 6 (**Figs. 1, 2, 4,** and **6**), which is close to the Fab pI, a milder self-association is detected at pH 8 (**Fig. 1c**) as the Fab is negatively charged, and no self-association at pH 5 were the Fab is positively charged (**Figs. 1** and **6**). Therefore, we digested PPI-1 to study the behavior of its fragments. We demonstrate that the native reversible self-association of PPI-1 is driven by Fab-Fab interaction (**Fig. 6**), by exploiting PPI-1 temperature-dependent behavior (**Figs. 4** and **6**). It has been shown that high mAb viscosity can be mediated by Fab-Fab self-association[11] and that IgG1 self-association can be driven by either the Fab or the Fc region.[11,12] The hydrophobic aggregation-prone regions, identified by Chennmsetty et al,[50,51] are more frequently found on the Fc region than on the Fab regions. However, we observed that PPI-1 Fab fragment presents a higher degree of hydrophobicity compared to the Fc fragment by RP-UPLC (**Fig. 8**). As the primary sequence of molecules studied in literature is unfortunately not available, it is impossible to compare IgG1 molecules showing pronounced self-association and identify molecular moieties on the Fab or Fc that could drive the association. Herein, we provide the primary sequence of PPI-1 to possibly increase the molecular understanding of IgG1 self-association (SI 9). We suggest that PPI-1 native reversible association is (1) highly dominated by the Fab fragments interaction, (2) that the locus of the interaction is located on the Fab fragment, and (3) that the association is due to hydrophobic interactions. The short-range nature of the hydrophobic interaction can explain the pH-dependent behavior (**Fig. 1**). Higher charge at low pH values prevents the short-range interaction from being accessible. If only the Fab is considered, with net neutral charge species at pH values close to 6, the salting-in effect (**Figs. 1** and **2**) can be rationalized. For charge-neutral species, many theoretical considerations were developed to explain initial salting-in of proteins. [52-54] It has been shown that mAbs close to their pI reveal a general salting-in effect by all anions.[55] The Debye-Huckel theory, in combination with the Kirkwood's theory expression of the dipole moments, actually predicts a salting-in effect which is consistent with the observation of protein behavior near their respective pIs.[56] This would explain the strong effect of Na^+ and Cl^- to dampen the PPI-1 native self-association (**Figs. 1** and **2**). We expect Na^+ to weakly interact with the fragment surface, and Cl^- could specifically bind to the protein surface. The idea of attractive electrostatic interaction is supported by salting-in behavior of carboxyhemoglobin close to its pI by potassium fluoride.[57] In addition, this is in agreement with the observation that a chaotropic monovalent ion binds more strongly to a net-charge neutral molecule.[58,59] Therefore, attractive electrostatic interactions may dominate at the Fab-Fab interaction site(s) at pHs near the Fab pI, where the cation and anion binding strengths with the protein surface determine the decreasing protein native self-association as the salt is increased. Further, a hydrophobic interaction supports the oligomers

association by decreasing the temperature of PPI-1 and its Fab fragments (**Figs. 4** and **6**). Hydrophobic association occurs as a result of a more ordered water structure in the proximity of nonpolar hydrocarbon groups.[60] It has been estimated that the change of entropy for protein-protein hydrophobic associations is positive and it becomes less positive at higher temperatures.[61] This would thermodynamically support an association at lower temperature as entropically driven. As expected, at increased PPI-1 concentrations, an increased degree of self-association was observed (**Fig. 1b**). In fact, higher mAb concentration increases the entropy of mixing which will tend to decrease the activation energy required to aggregate by increasing the potential energy baseline.[49] Interestingly, PPI-1 shows the formation of large particles once a lower concentration threshold is reached and exceeded (**Figs. 1b** and **5**), and this is only observed under conditions that prompt PPI-1 native reversible self-association. The mAbs self-association is driven by the minimization of solvent exposed hydrophobic surfaces on the Fab fragment. We hypothesize that under a critical value, the mAb concentration is not enough to self-stabilize, and therefore a phase separation occurs. However, further studies are necessary to properly characterize PPI-1 behavior at very low concentration.

4.2 Lessons learned: pitfalls to study PPI-1 reversible native self-association

Batch DLS, batch SLS, SEC-MALS, AF4-MALS, CG-MALS, DLSMALS, AUC, nanoDSF, and SAXS were applied to investigate PPI-1. Owing to the ubiquitous native reversible self-association of PPI- 1, only the techniques capable to measure the naïve sample (batch SLS and DLS, AUC, SAXS) allowed proper assessment of the size and, or, amount of the reversible oligomers. On the other hand, care in the interpretation of the results is necessary if the technique applied involves the modification of either pH, ionic strength, temperature, or protein concentration, as the equilibrium of the system will be shifted. Owing to the unusual behavior of PPI-1 as a function of its concentration (**Figs. 1b** and **5**), pH (**Fig. 1**), and salt concentration (**Figs. 1c** and **2**), the SEC-MALS (**Fig. 3**) cannot be applied to investigate PPI-1 reversible self-association. In fact, the buffer exchange and dilution over the column impacts the mass recovery even without stress (**Fig. SI 6**). However, SEC-MALS remains a valuable tool to characterize the formation of irreversible nonnative aggregation. Other fractionation methods had similar issues; for example, AF4 does not allow to properly characterize the sample due to the intense dilution over the channel (**Fig. SI 5**). Further, the uncommon behavior of PPI-1 as a function of the concentration **(Figs**. **1c** and **5**) does not allow the assessment of the stoichiometry and constants of dissociation with limited amount of material (e.g. CG-MALS, AUC). This could be a limiting factor for mAbs in early stage of development, such as PPI-1. Thus, we suggest the use of nanoDSF, DLS, and SLS as high-throughput technologies and AUC as a gold-standard to characterize native reversible self-association.

5 Aknowledgements

This study was funded by a project part of the EU Horizon 2020 Research and Innovation program under the Marie Skłodowska-Curie grant agreement No 675074. The first author would like to thank the whole PIPPI consortium (http://www.pippi.kemi.dtu.dk) for the continuous support offered. The first author thanks Wyatt Technology staff members for their many contributions, with a special mention to Felix Gloge for the passionate discussion related to this work. The ESRF synchrotron, BM29 bioSAXS beamline at Grenoble is acknowledged for providing beamtime for the project. PH and SM acknowledge DanSCATT for funding.

6 References

1. Wang W, Singh S, Zeng DL, King K, Nema S. Antibody structure, instability, and formulation. J Pharm Sci. 2007;96(1):1-26.

2. Pavlou AK, Belsey MJ. The therapeutic antibodies market to 2008. Eur J Pharm Biopharm. 2005;59(3):389-396.

3. Roque AC, Lowe CR, Taipa MA. Antibodies and genetically engineered related molecules: production and purification. Biotechnol Prog. 2004;20(3):639-654.

4. Jain T, Sun T, Durand S, et al. Biophysical properties of the clinical-stage antibody landscape. Proc Natl Acad Sci U S A. 2017;114(5):944-949.

5. Wang W. Protein aggregation and its inhibition in biopharmaceutics. Int J Pharm. 2005;289(1-2):1-30.

6. Hermeling S, Crommelin DJ, Schellekens H, Jiskoot W. Structure-immunogenicity relationships of therapeutic proteins. Pharm Res. 2004;21(6):897-903.

7. Braun A, Kwee L, Labow MA, Alsenz J. Protein aggregates seem to play a key role among the parameters influencing the antigenicity of interferon alpha (IFN-alpha) in normal and transgenic mice. Pharm Res. 1997;14(10):1472-1478.

8. Telikepalli S, Shinogle HE, Thapa PS, et al. Physical characterization and in vitro biological impact of highly aggregated antibodies separated into size-enriched populations by fluorescence-activated cell sorting. J Pharm Sci. 2015;104(5): 1575-1591.

9. Stefani M, Dobson CM. Protein aggregation and aggregate toxicity: new insights into protein folding, misfolding diseases and biological evolution. J Mol Med (Berl). 2003;81(11):678-699.

10. Roberts CJ. Non-native protein aggregation kinetics. Biotechnol Bioeng. 2007;98(5):927-938.

11. Kanai S, Liu J, Patapoff TW, Shire SJ. Reversible self-association of a concentrated monoclonal antibody solution mediated by fabefab interaction that impacts solution viscosity. J Pharm Sci. 2008;97(10):4219-4227.

12. Nishi H, Miyajima M, Wakiyama N, et al. Fc domain mediated self-association of an IgG1 monoclonal antibody under a low ionic strength condition. J Biosci Bioeng. 2011;112(4):326-332.

13. Yadav S, Sreedhara A, Kanai S, et al. Establishing a link between amino acid sequences and self-associating and viscoelastic behavior of two closely related monoclonal antibodies. Pharm Res. 2011;28(7):1750-1764.

14. Shire SJ, Shahrokh Z, Liu J. Challenges in the development of high protein concentration formulations. J Pharm Sci. 2004;93(6):1390-1402.

15. Liu J, Nguyen MD, Andya JD, Shire SJ. Reversible self-association increases the viscosity of a concentrated monoclonal antibody in aqueous solution. J Pharm Sci. 2005;94(9):1928-1940.

16. Hall CG, Abraham GN. Reversible self-association of a human myeloma protein. Thermodynamics and relevance to viscosity effects and solubility. Biochemistry. 1984;23(22):5123-5129.

17. Hall CG, Abraham GN. Size, shape, and hydration of a self-associating human IgG myeloma protein: axial asymmetry as a contributing factor in serum hyperviscosity. Arch Biochem Biophys. 1984;233:330-337.

18. Lindsley H, Teller D, Noonan B, Peterson M, Mannik M. Hyperviscosity syndrome in multiple myeloma. A reversible, concentration-dependent aggregation of the myeloma protein. Arch Biochem Biophys. 1984;233(2):330-337.

19. Pope RM, Fletcher MA, Mamby A, Shapiro CM. Rheumatoid arthritis associated with hyperviscosity syndrome and intermediate complex formation. Arch Intern Med. 1975;135(2):281-285.

20. Liu J, Andya JD, Shire SJ. A critical review of analytical ultracentrifugation and field flow fractionation methods for measuring protein aggregation. AAPS J. 2006;8(3):E580-E589.

21. Saluja A, Badkar AV, Zeng DL, Nema S, Kalonia DS. Ultrasonic storage modulus as a novel parameter for analyzing protein-protein interactions in high protein concentration solutions: correlation with static and dynamic light scattering measurements. Biophys J. 2007;92(1):234-244.

22. Saluja A, Badkar AV, Zeng DL, Nema S, Kalonia DS. Application of high frequency rheology measurements for analyzing protein-protein interactions in high protein concentration solutions using a model monoclonal antibody (IgG2). J Pharm Sci. 2006;95(9):1967-1983.

23. Philo JS, Arakawa T. Mechanisms of protein aggregation. Curr Pharm Biotechnol. 2009;10(4):348-351.

24. Schuck P. Size-distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and Lamm equation modeling. Biophys J. 2000;78(3): 1606-1619.

25. Pernot P, Round A, Barrett R, et al. N Upgraded ESRF BM29 beamline for SAXS on macromolecules in solution. J Synchrotron Radiat. 2013;20(Pt 4):660-664.

26. Konarev PV, Volkov VV, Sokolova AV, Koch MHJ, Svergun DI. PRIMUS: a Windows PC-based system for small-angle scattering data analysis. J Appl Crystallogr. 2003;36:1277-1282.

27. Franke D, Petoukhov MV, Konarev PV, et al. Atsas 2.8: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. J Appl Crystallogr. 2017;50(Pt 4):1212-1225.

28. Semenyuk AV, Svergun DI. GNOMea program package for small-angle scattering data processing. J Appl Crystallogr. 1991;24:537-540.

29. Mylonas E, Svergun DI. Accuracy of molecular mass determination of proteins in solution by small-angle X-ray scattering. J Appl Crystallogr. 2007;40:245-249.

30. Gentiluomo L, Roessner D, Augustijn D, et al. Application of interpretable artificial neural networks to early monoclonal antibodies development. Eur J Pharm Biopharm. 2019;141:81-89.

31. Schuck P. On the analysis of protein self-association by sedimentation velocity analytical ultracentrifugation. Anal Biochem. 2003;320(1):104-124.

32. Minton AP. Recent applications of light scattering measurement in the biological and biopharmaceutical sciences. Anal Biochem. 2016;501:4-22.

33. Cole JL, Lary JW, P Moody T, Laue $_{TM}$. Analytical ultracentrifugation: sedimentation velocity and sedimentation equilibrium. Methods Cell Biol. 2008;84:143-179.

34. Pekar AH, Frank BH. Conformation of Proinsulin. A comparison of insulin and Proinsulin self-association at neutral pH. Biochemistry. 1972;11(22):4013-4016.

35. 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. 2003;20(9):1325-1336.

36. Dobson CM. Principles of protein folding, misfolding and aggregation. Semin Cell Dev Biol. 2004;15(1):3-16.

37. Fink AL. Protein aggregation: folding aggregates, inclusion bodies and amyloid. Fold Des. 1998;3(1):R9-R23.

38. Goers J, Permyakov SE, Permyakov EA, Uversky VN, Fink AL. Conformational prerequisites for alpha-lactalbumin fibrillation. Biochemistry. 2002;41(41): 12546-12551..

39. Grillo AO, Edwards KL, Kashi RS, et al. Conformational origin of the aggregation of recombinant human factor VIII. Biochemistry. 2001;40(2):586-595.

40. Khurana R, Gillespie JR, Talapatra A, et al. Partially folded intermediates as critical precursors of light chain amyloid fibrils and amorphous aggregates. Biochemistry. 2001;40(12):3525-3535.

41. Linding R, Schymkowitz J, Rousseau F, Diella F, Serrano L. A comparative study of the relationship between protein structure and b-aggregation in globular and intrinsically disordered proteins. J Mol Biol. 2004;342(1):345-353.

42. Uversky VN, Fink AL. Conformational constraints for amyloid fibrillation: the importance of being unfolded. Biochim Biophys Acta. 2004;1698(2):131-153.

43. Sahin E, Grillo AO, Perkins MD, Roberts CJ. Comparative effects of pH and ionic strength on protein-protein interactions, unfolding, and aggregation for IgG1 antibodies. J Pharm Sci. 2010;99(12):4830-4848.

44. Jion AI, Goh LT, Oh SK. Crystallization of IgG1 by mapping its liquid-liquid phase separation curves. Biotechnol Bioeng. 2006;95(5):911-918.

45. Ahamed T, Esteban BN, Ottens M, et al. Phase behavior of an intact monoclonal antibody. Biophys J. 2007;93(2):610-619.

46. Dumetz AC, Chockla AM, Kaler EW, Lenhoff AM. Protein phase behavior in aqueous solutions: crystallization, liquid-liquid phase separation, gels, and aggregates. Biophys J. 2008;94(2):570-583.

47. Nishi H, Miyajima M, Nakagami H, Noda M, Uchiyama S, Fukui K. Phase separation of an IgG1 antibody solution under a low ionic strength condition. Pharm Res. 2010;27(7):1348-1360.

48. Israelachvili JN. Intermolecular and Surface Forces: With Applications to Colloidal and Biological Systems. 2nd ed. London: Academic Press; 1992.

49. Laue T. Proximity energies: a framework for understanding concentrated solutions. J Mol Recognit. 2012;25(3):165-173.

50. Chennamsetty N, Helk B, Voynov V, Kayser V, Trout BL. Aggregation-prone motifs in human immunoglobulin G. J Mol Biol. 2009;391(2):404-413.

51. Chennamsetty N, Voynov V, Kayser V, Helk B, Trout BL. Design of therapeutic proteins with enhanced stability. Proc Natl Acad Sci U S A. 2009;106(29):11937- 11942.

52. Baldwin RL. How Hofmeister ion interactions affect protein stability. Biophys J. 1996;71(4):2056-2063.

53. Arakawa T, Timasheff SN. Theory of protein solubility. Methods Enzymol. 1985;114:49-77.

54. Melander W, Horv_ath C. Chromatography on hydrophobic interactions of Proteins: an interpretation in precipitation and of the lyotropic series. Arch Biochem Biophys. 1977;183(1):200-215.

55. Zhang L, Tan H, Fesinmeyer RM, et al. Antibody solubility behavior in monovalent salt solutions reveals specific anion effects at low ionic strength. J Pharm Sci. 2012;101(3):965-977.

56. Arakawa T, Timasheff SN. The stabilization of proteins by osmolytes. Biophys J. 1985;47(3):411-414.

57. Green AA. Studies in the physical chemistry of the proteins X. The solubility of hemoglobin in solutions of chlorides and sulfates of varying concentration. Journal of Biological Chemistry. 1932;95:47-66.

58. Chen X, Flores SC, Lim SM, et al. Specific anion effects on water structure adjacent to protein monolayers. Langmuir. 2010;26(21):16447-16454.

59. Chen X, Yang T, Kataoka S, Cremer PS. Specific ion effects on interfacial water structure near macromolecules. J Am Chem Soc. 2007;129(40):12272-12279.

60. Frank HS, Evans MW. Free volume and entropy in condensed systems III. Entropy in binary liquid mixtures; partial molal entropy in dilute solutions; structure and thermodynamics in aqueous electrolytes. J Chem Phys. 1945;13: 507-532.

60. Ross PD, Subramanian S. Thermodynamics of protein association reactions: forces contributing to stability. Biochemistry. 1981;20(11):3096-3102

.

7 Supplementary information

List of supplementary information

SI 1. SAXS supplementary information

SI 2. PPI-1 DLS and nanoDSF results.

SI 3. The effect of NaCl on the sedimentation coefficient of PPI-1 by sedimentation velocity. AUC.

SI 4. SAXS results (Rg)

SI 5. AF4-MALS chromatograms

SI 6. Size exclusion chromatograms.

SI 7. Temperature ramp curve measured by DLS

SI 8. Temperature cycles investigated by DLS of the undigested mAb, Fab and Fc.

SI 9. PPI-1 primary sequence

SI 1. SAXS supplementary information

Table SI 1.1: Experimental set up of SAXS measurements

Instrument	ESRF, Grenoble (France), bioSAXS beamline (BM29)
Wavelength (Å)	0.99
q-range (Å^{-1})	0.004 – 0.49
Sample-to-detector distance (m)	2.864
Detector	Pilatus 1M
Flux (photons/s)	2 x 10^{12}
Beam size (μm^2)	700 x 700
Sample configuration	1.8 mm quartz glass capillary
Absolute scaling method	Comparison to water in sample capillary
Normalization	To transmitted intensity by beam-stop counter
Monitoring for radiation damage	Control of un-subtracted and scaled subtracted data for systematic changes typical for radiation damage

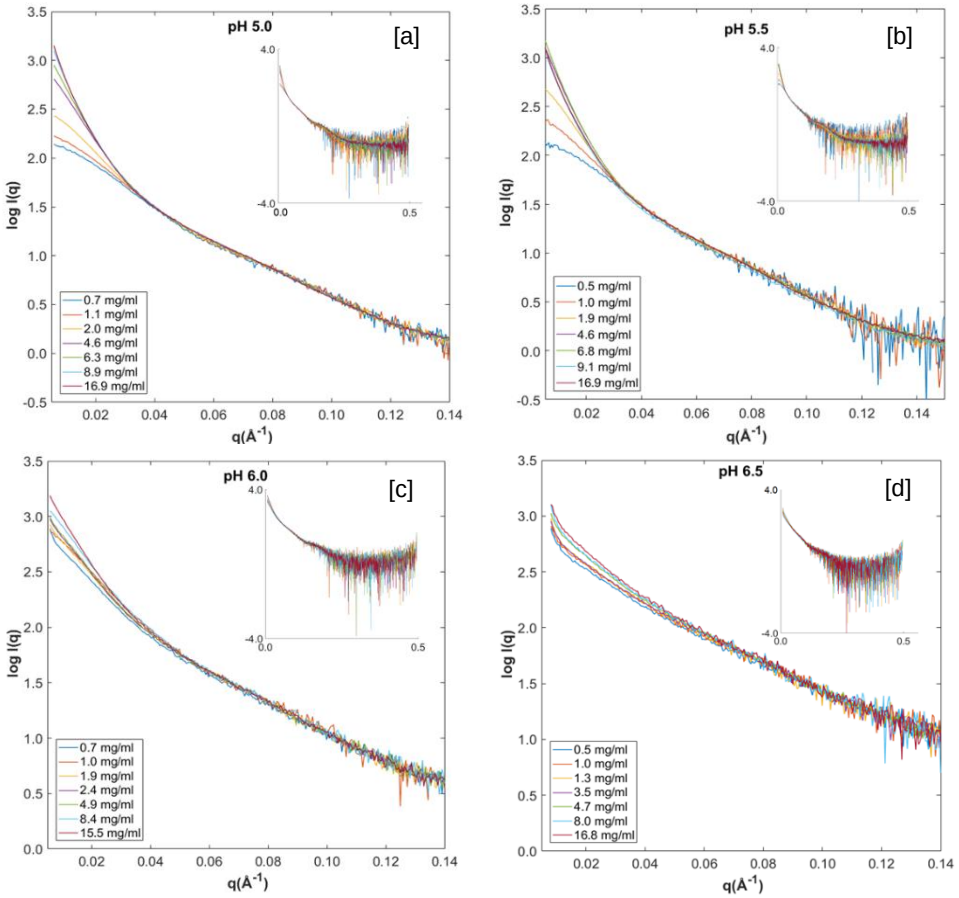


Figure SI 1.2 SAXS scattering curves*: a) 10mM_Histidine_pH 5.0, b) 10mM_Histidine_pH 5.5, c) 10mM_Histidine_pH 6.0, d) 10mM_Histidine_pH 6.5. Data are shown for different PPI-1 formulation conditions with increasing concentrations.*

Table SI 1.3 An overview of the samples measured by SAXS and data treatment parameters:

a) 10mM histidine pH 5.0

Protein concentration (mg/ml)	R_G (Gnom) (nm)	$I(0)/c$ (Gnom)	M_w (kDa)
0.74	6.01	0.11	156
1.10	6.79	0.14	188
2.01	9.80	0.24	333
4.56	15.34	0.66	916
6.26	19.32	1.05	1451
8.94	23.24	1.74	2412
16.89	25.82	2.24	3109

b) 10mM histidine pH 5.5

Protein concentration (mg/ml)	R_G (Gnom) (nm)	$I(0)/c$ (Gnom)	M_w (kDa)
0.47	6.32	0.12	163
1.00	8.06	0.18	249
1.93	13.55	0.44	610
4.65	19.71	1.32	1931
6.82	20.64	1.39	1828
9.11	21.51	1.42	1966
16.93	23.09	1.77	2456

c) 10mM histidine pH 6.0

Protein concentration (mg/ml)	R_G (Gnom) (nm)	$I(0)/c$ (Gnom)	M_w (kDa)
0.70	12.70	0.48	663
1.04	15.34	0.56	774
1.88	15.98	0.69	949
2.41	19.32	1.05	1451
4.92	20.24	1.58	2190
8.43	24.42	1.70	2351
15.50	26.52	2.16	2998

d) 10m histidine pH 6.5

Protein concentration (mg/ml)	R_G (Gnom) (nm)	$I(0)/c$ (Gnom)	M_w (kDa)
0.48	17.48	0.81	1119
1.01	19.74	1.26	1747
1.34	22.09	1.54	2129
3.52	23.38	1.84	2555
4.74	23.94	2.38	3298
8.03	25.48	2.50	3459
16.82	29.36	2.96	4106

Table SI 2. PPI-1 DLS and nanoDSF results.

Protein	pH	[NaCl]	k_D (mg/mL)	T_{agg} (°C)	$T_{m,on}$ (°C)	T_{m1} (°C)
PPI-1	5	0	-2.46E-02	52.69	52.12	57.44
PPI-1	5.5	0	-1.90E-02	54.28	53.43	58.84
PPI-1	6	0	-1.72E-02	55.38	54.48	60.73
PPI-1	6.5	0	-2.94E-02	56.96	54.90	62.55
PPI-1	7	0	-2.45E-02	56.8	55.29	64.62
PPI-1	7.5	0	-2.34E-02	50.82	53.78	64.31
PPI-1	8	0	-1.81E-02	49.47	56.77	69.97
PPI-1	9	0	-1.87E-02	56.5	58.98	69.99
PPI-1	5	70	-2.39E-02	45.97	49.98	54.91
PPI-1	5.5	70	-2.01E-02	50.7	52.01	57.44
PPI-1	6	70	-1.89E-02	51.98	54.61	60.30
PPI-1	6.5	70	-4.05E-02	54.41	55.64	63.14
PPI-1	7	70	-4.44E-02	51.56	56.27	63.73
PPI-1	7.5	70	-3.62E-02	55.76	56.57	64.00
PPI-1	8	70	-3.37E-02	55.48	53.60	69.93
PPI-1	9	70	2.11E-02	56.48	59.72	70.69
PPI-1	5	140	-2.01E-02	50.24	49.19	54.30
PPI-1	5.5	140	-1.74E-02	47.11	52.17	57.12
PPI-1	6	140	-2.46E-02	52.63	54.10	59.85
PPI-1	6.5	140	-1.90E-02	55.98	56.00	62.65
PPI-1	7	140	-1.72E-02	55.78	56.59	63.40
PPI-1	7.5	140	-2.94E-02	55.84	56.65	63.94
PPI-1	8	140	-2.45E-02	56.81	55.17	70.59
PPI-1	9	140	-2.34E-02	56.09	58.75	70.67

*T_m fit formula: 37.102+3.875pH – (0.005[NaCl])**

Table SI 3. The effect of NaCl on the sedimentation coefficient of PPI-1 by sedimentation velocity. AUC. *The table lists the S_w from the NaCl tritation experiment showed in Fig. 2.*

System (in 10 mM Histidine)			Monomer			TOTAL AVERAGE		
pH	NaCl (mM)	PPI-01 (mg/mL)	S_w	Std. Dev.	%	S_w	Std. Dev.	%
5	0	1	**7.31**	0.277	99	**7.31**	0.277	99
5	0	0.5	**7.35**	0.621	99	**7.35**	0.621	99
5	0	0.1	**7.36**	0.343	99	**7.36**	0.343	99
5.5	0	1	**7.69**	0.293	83	**7.82**	0.528	100
5.5	0	0.5	**7.52**	0.385	92	**7.75**	1.02	99
5.5	0	0.1	**7.42**	0.394	94	**10.7**	3.1	99
6	0	1	**7.8**	0.734	39	**10.77**	3.5	100
6	0	0.5	**7.85**	0.521	69	**9.28**	2.9	100
6	0	0.1	**7.4**	0.444	81	**7.89**	1.578	95
6.5	0	1	**7.78**	0.864	24	**16.82**	6.9	100
6.5	0	0.5	**7.55**	0.392	30	**11.67**	4.05	99
6.5	0	0.1	*	*	*	*	*	*
7	0	1	-	-	-	**30.1**	15.3	100
7	0	0.5	*	*	*	*	*	*
7	0	0.1	*	*	*	*	*	*
7.5	0	1	*	*	*	*	*	*
7.5	0	0.5	*	*	*	*	*	*
7.5	0	0.1	*	*	*	*	*	*
6	10	1	**7.8**	1.2	92	**8.1**	1.683	99
6	20	1	**7.5**	0.8	92	**7.73**	1.21	99
6	30	1	**7.33**	0.433	97	**7.5**	0.85	100
6	40	1	**7.3**	0.429	95	**7.3**	0.429	99
6	50	1	**7.14**	0.356	99	**7.14**	0.356	99
6	60	1	**7.23**	0.48	100	**7.23**	0.48	100
6	140	1	**6.9**	0.48	99	**6.9**	0.48	99
6.5	10	1	**8.15**	1.384	99	**8.15**	1.384	99
6.5	20	1	**7.6**	0.751	90	**7.98**	1.6	99
6.5	30	1	**7.47**	0.634	97	**7.54**	0.766	99
6.5	40	1	**7.33**	0.64	95	**7.49**	1.2	99
6.5	50	1	**7.3**	0.65	93	**7.52**	1.1	99
6.5	60	1	**7.21**	0.62	93	**7.41**	1.2	99
6.5	140	1	**7**	0.53	93	**7**	1	99
7	10	1	**7.62**	0.41	60	**9.51**	3.35	98
7	20	1	**7.43**	0.4	77	**8.23**	1.93	98
7	30	1	**7.27**	0.231	95	**7.43**	0.87	100
7	40	1	**7.3**	0.37	99	**7.3**	0.37	99
7	50	1	**7.22**	0.6	97	**7.22**	0.6	97
7	60	1	**7.1**	0.57	99	**7.1**	0.57	99
7	140	1	**6.8**	0.7	99	**6.8**	0.7	99

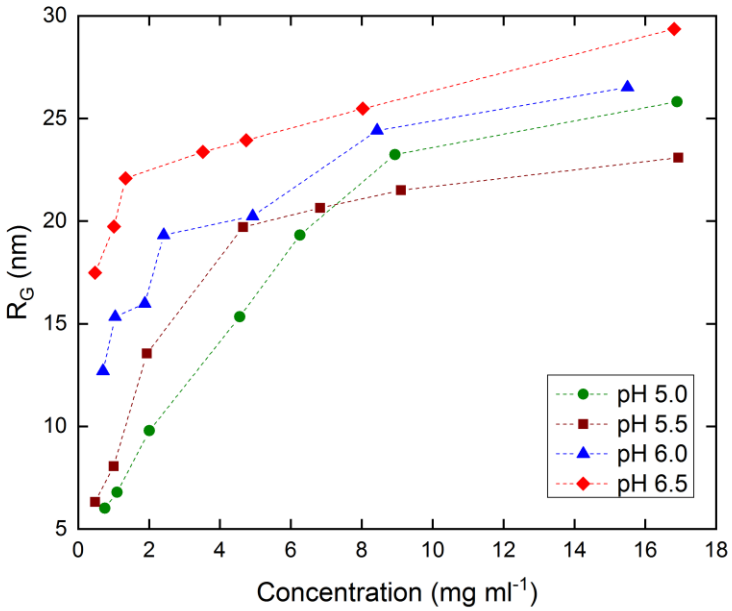


Figure SI 4 SAXS results. *Four formulations (without salt) were formulated at pH 5, 5.5, 6.0 and 6.5 depicted respectively in green, brown, blue and red.*

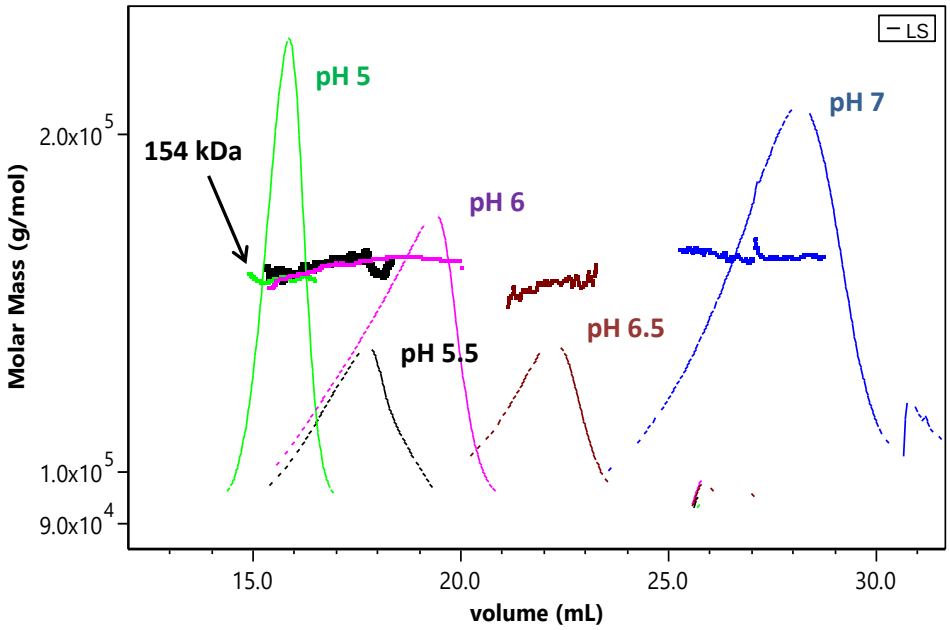


Figure SI 5. AF4-MALS chromatograms. *The light scattering signal is showed for PPI-1 in different formulations. The mobile phase always matches the formulation (His 10 mM for all the pHs investigated).*

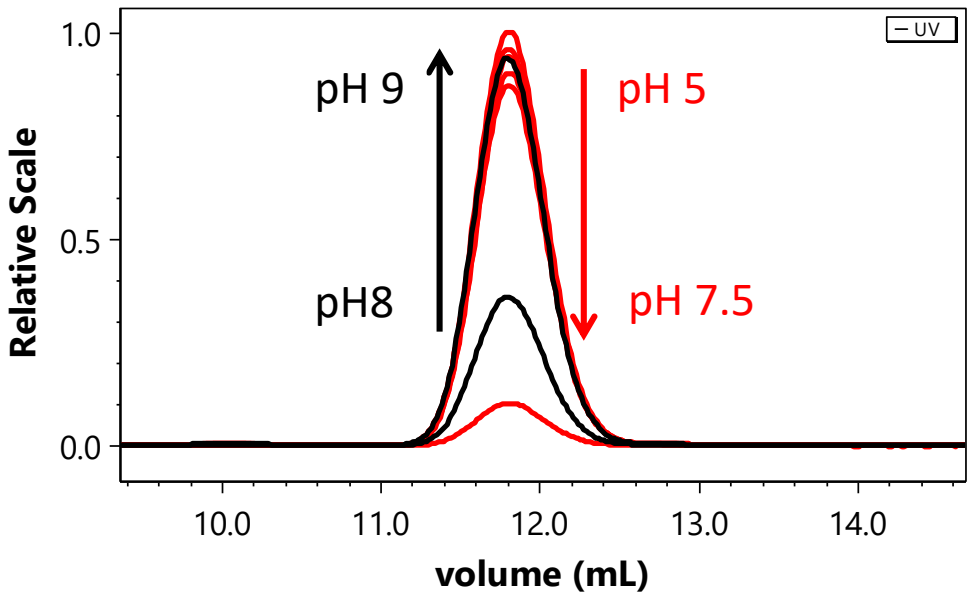


Figure SI 6. Size exclusion chromatograms. *The figure includes PPI-1 formulations with no salt from pH 5 to pH 9. The red and black chromatograms represent respectively a decrease of column recovery with the pH and an increase of column recovery with the pH.*

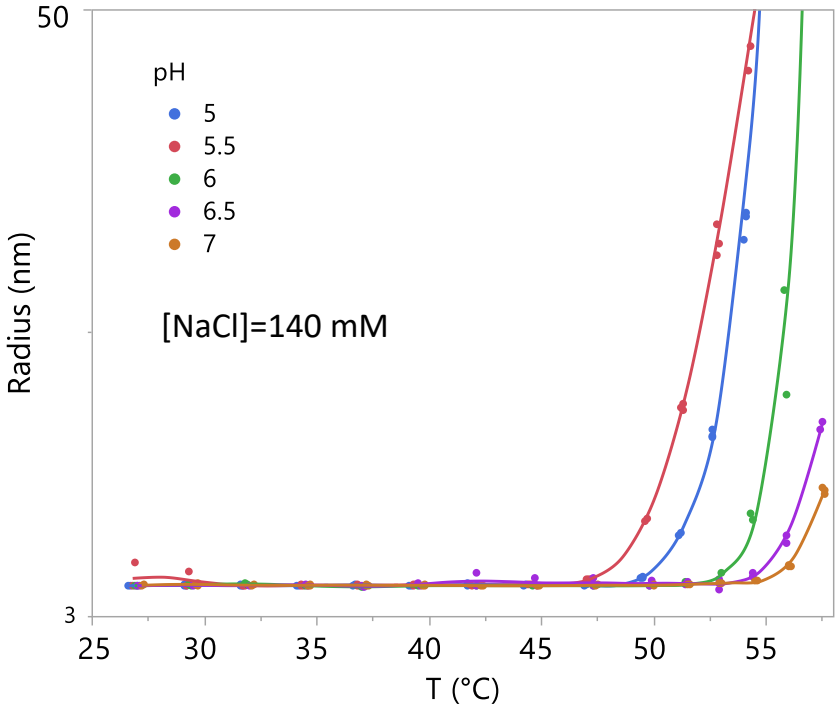


Figure SI 7. Temperature ramp curve measured by DLS. *All formulations where investigated in 10 mM His, 140 mM NaCl from pH 5 to 7 and the data points were depicted as shown in the legend.*

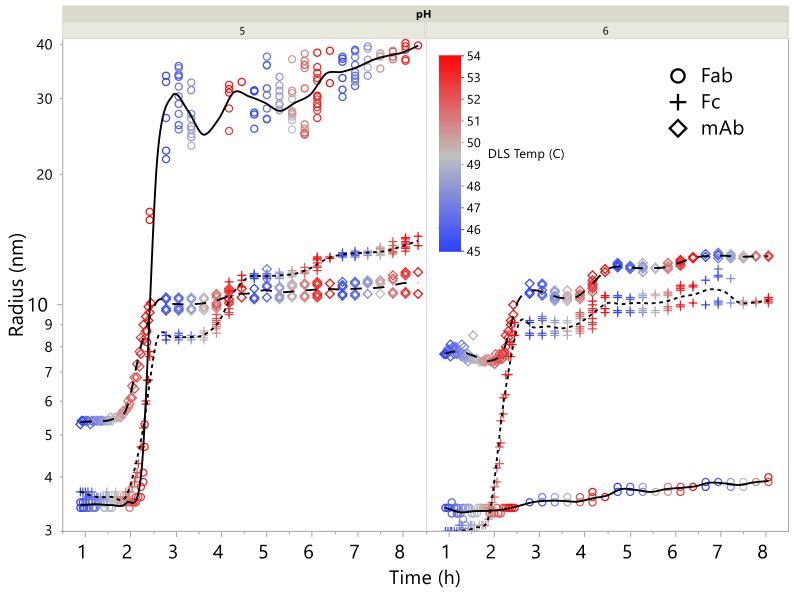


Figure SI 8. Temperature cycles investigated by DLS of the undigested mAb, Fab and Fc. *Temperature cycles from 45 °C to 54 °C are shown for all samples at pH 5 and 6 (10 mM His) respectively on the left and right of the graph. On the ordinates the apparent Rh is shown. mAb, Fab and Fc are respectively depicted as shown in the legend.*

SI 9: PP-1 (IgG1) primary sequence.

Heavy chain
EVQLVQSGAEVKKPGATVKISCKVYGYIFTDYNIYWVRQAPGKGLEWMGLIDPDNGETFYAEKFQGRAT
MTADTSSDRAYMELSSLRFEDTAVYYCATVMGKWIKGGYDYWGRGTLVTVSSASTKGPSVFPLAPSSK
STSGGTAALGCLVKDYFPEPVTVSWNSGALTSGVHTFPAVLQSSGLYSLSSVVTVPSSSLGTQTYICNVN
HKPSNTKVDKKVEPKSCDKTHTCPPCPAPELLGGPSVFLFPPKPKDTLMISRTPEVTCVVVDVSHEDPEV
KFNWYVDGVEVHNAKTKPREEQY**N**STYRVVSVLTVLHQDWLNGKEYKCKVSNKALPAPIEKTISKAKGQ
PREPQVYTLPPSRDELTKNQVSLTCLVKGFYPSDIAVEWESNGQPENNYKTTPPVLDSDGSFFLYSKLTV
DKSRWQQGNVFSCSVMHEALHNHYTQKSLSLSPGK

N-glycosylation site

Light chain
QSVLTQPPSVSGAPGQRVTISCTGSSSNIGAGYDVHWYQQLPGTAPKLLIYDNFNRPSGVPPRFSGSKS
GTSASLAITGLQAEDEADYYCQSYDSPTLTSPFGTGT
LTVLGQPKAAPSVTLFPPSSEELQANKATLVCLISDFYPGAVTVAWKADSSPVKAGVETTTPSKQSNNKY
AASSYLSLTPEQWKSHRSYSCQVTHEGSTVEKTVAPTECS

CHAPTER V: Application of machine learning to predict monomer retention of therapeutic proteins after long term storage

Lorenzo Gentiluomo[1,2,3*], Dierk Roessner[1], Wolfgang Frieß[2]

[1] Wyatt Technology Europe GmbH, Hochstrasse 18, 56307 Dernbach, Germany

[2] Department of Pharmacy: Pharmaceutical Technology and Biopharmaceutics; Ludwig-Maximilians-Universitaet Muenchen, Butenandtstrasse 5, 81377 Munich, Germany

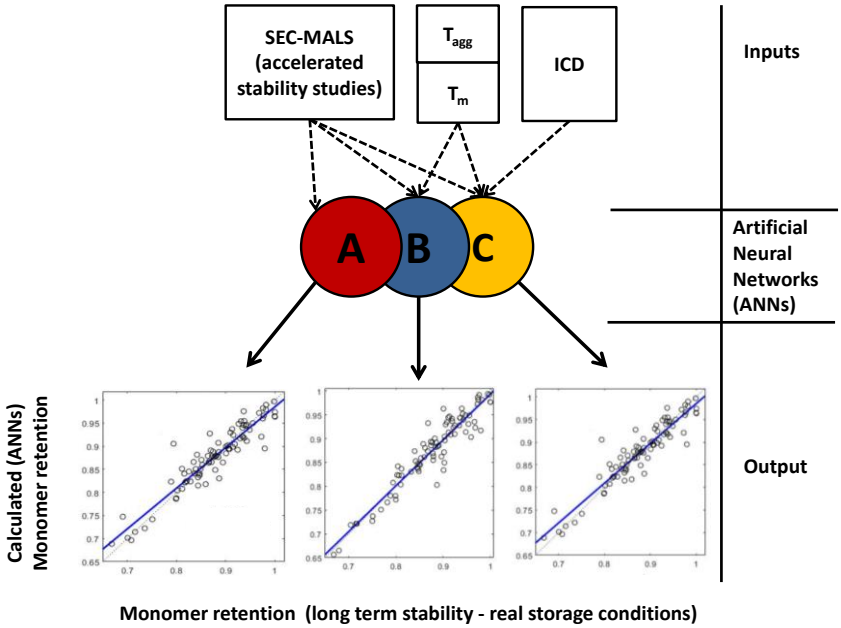
[3]Present address: Coriolis Pharma, Fraunhoferstraße 18B, 82152, Planegg, Germany

*denotes corresponding author.

CONTRIBUTION: **Lorenzo Gentiluomo**, Wolfgang Frieß and Dierk Roessner conceived and designed the study. **Lorenzo Gentiluomo** conducted and analyzed the SEC-MALS experiments. **Lorenzo Gentiluomo** performed the calculations. **Lorenzo Gentiluomo** and Wolfgang Frieß contributed to the writing of the manuscript.

Abstract

An important aspect of initial developability assessments as well formulation development and selection of therapeutic proteins is the evaluation of data obtained under accelerated stress condition, i.e. at elevated temperatures. We propose the application of artificial neural networks (ANNs) to predict long term stability in real storage condition from accelerated stability studies and other high-throughput biophysical properties e.g. the first apparent temperature of unfolding (T_m). Our models have been trained on therapeutic relevant proteins, including monoclonal antibodies, in various pharmaceutically relevant formulations. Further, we developed network architectures with good prediction power using the least amount of input features, i.e. experimental effort to train the network. This provides an empiric means to highlight the most important parameters in the prediction of real-time protein stability. Further, several models were developed by a different validation means (i.e. leave-one-protein-out cross-validation) to test the robustness and the limitations of our approach. Finally, we apply surrogate machine learning algorithms (e.g. linear regression) to build trust in the ANNs decision making procedure and to highlight the connection between the leading inputs and the outputs.



Graphical abstract. *Application of machine learning to predict monomer retention of therapeutic proteins after long term storage*

List of abbreviations

ANNs – artificial neural networks

IgG - immunoglobulin

mAb – monoclonal antibody

T_m – first apparent temperature of unfolding

$T_{agg,on}$ – onset temperature of aggregation

C_{m1} - denaturant concentration at half transition of the first apparent unfolding

m1 - slope of the first unfolding transition curve of chemical denaturation

$T_{m,on}$ - onset temperature of unfolding

R_h – hydrodynamic radius

SEC – size-exclusion chromatography

MALS - multi angle (laser) light scattering

R – Pearson´s correlation coefficient

Mw – molecular weight

RMSE – root means square error

DLS – dynamic light scattering

ICD – isothermal chemical denaturation

a_n^x – monomer retention after 6 months of storage at temperature X

m_n^x - monomer retention after 2 weeks of storage at temperature X

LS_m^X - light scattering area ratio after 2 weeks of storage at the temperature X

1 Introduction

Protein therapeutics are used against multiple severe diseases.[1,2] Their success lies in the specificity for therapeutic targets, which is rooted in the diversity and the complexity of protein structures. Promising candidates have to exhibit desirable biophysical properties that allow for sufficient stability during manufacturing, shipping, storage, handling and administration. Even immunoglobulins (IgGs) of the same subclass that share common structural features often behave quite differently in solution.[3] An important aspect of initial developability assessments of therapeutic proteins is the evaluation of data obtained by stressing the candidates under several stressing conditions, i.e. accelerated stability studies.[4,5] Aim of these studies is to gain an understanding of the modes of instability that could impact the drug product during its lifecycle, which is often difficult to predict. Afterwards this information is used to design formulation strategies that mitigate protein instability. Amongst these instabilities protein drugs aggregation presents one of the most common and troubling challenges.[6] It has been shown that aggregates have reduced activity and greater immunogenicity.[7,8] Arrhenius kinetics could be applied to extrapolate approximate aggregation rates at intended storage condition from data obtained at higher temperature in different formulations. Nevertheless, with increasing temperature, other reactions could become rate-determining that cannot be described by Arrhenius kinetics. Ultimately, only real-time stability testing permits the establishment of recommended storage conditions and shelf life. On the other hand, for developability assessment and formulation development and selection, higher temperature studies are indispensable.

Various tools for the prediction of aggregation rates have been reported, some of which have been validated with therapeutically relevant proteins.[9–14] Notably, the spatial aggregation potency tool was validated using long term stability data, and was found beneficial to rank therapeutically relevant monoclonal antibodies (mAbs).[15,16] These tools are usually related to a protein class, e.g. mAbs, and require the protein structure. Additionally, they render an intrinsic aggregation propensity, even though aggregation depends on both the protein and its surrounding environment, the formulation. In this study we applied machine learning to predict the real-time stability, in our case the six months stability at 4°C, 25°C and 40°C, from accelerated stability studies. All the proteins presented in this work have recently been extensively characterized and their primary sequence is publicly available.[17] We focused our

investigation on the prediction of monomer retention in size exclusion chromatography (SEC). We coupled SEC with multi angle laser light scattering (MALS) which is useful for studying and characterizing aggregation due to its high sensitivity towards high molecular weight species.[18,19] SEC-MALS results provide an absolute means for qualitative and quantitative analysis of protein aggregation kinetics.[20–23] For the prediction of long term stability from accelerated stability studies we propose the use of artificial neural networks (ANNs). ANNs have been applied in the pharmaceutical research for many different purposes.[24–33] However, a significant limitation of ANNs is their interpretability. Algorithms, such as TREPANs, have been used to extract a comprehensible concept description from a trained network to solve classification problems.[34] For certain regression problems another approach is to use surrogate models.[35] Therefore, we investigated several machine learning algorithms (e.g. Support vector machines) as surrogate models to understand and build trust in the ANNs decision making process. Amongst the one investigated we selected "white box" algorithms (i.e. human friendly), namely linear regression and a decision trees, as no significant benefit was observed in applying more complex models. Fourteen therapeutically relevant proteins were investigated in twenty-four pharmaceutically relevant formulation conditions. Our hypothesis was that the output of suitably designed short term accelerated stability assays could be used to generate formulation dependent protein stability "finger-prints". These features are used as input for the ANNs to predict the real-time stability of proteins. Furthermore, we developed multiple algorithms which included other additional high-throughput biophysical parameter typically used to investigate protein stability, e.g. the temperature of unfolding (T_m). We investigate ANN architectures with the highest prediction power and the least amount of input features, i.e. experimental effort to train the network. This provides an empiric means to highlight the most important parameters in the prediction of the monomer retention after six months of stress from the data collected after only two weeks of stress.

2 Material and methods

2.1 Sample preparation

Five antibodies IgG1s (PPI-02, PPI-03, PPI-04, PPI-10, PPI-13), one bispecific antibody (PPI-08), one IgG2 (named PPI-17), and one HSA-fusion protein (named PPI-18) were provided by AstraZeneca (UK).

Interferon alpha-2a (named PPI30) was provided from Roche (Switzerland). Two lipases (PPI-45 and PPI-46) were provided by Novozymes (Denmark). Recombinant human transferring and human serum albumin (respectively PPI-44 and PPI-49) were provided by Albumedix (UK). A summary of the protein properties is listed in **Table 1**. The proteins were dialyzed overnight using Slide-A-Lyzer™ cassettes (Thermo Fisher Scientific, USA) with suitable membrane cut-off against excess of 10 mM of buffer at several pHs 5.0, 5.5, 6.5, 7.5. The excipients (e.g. Sucrose) stock solutions were prepared in the respective buffers. The 24 investigated formulations are listed in **Table 1**. Protein concentration was measured on a Nanodrop 2000 (Thermo Fisher Scientific, USA) using the protein extinction coefficient. The extinction coefficient was double-checked experimentally by means of a refractive index detector (Optilab T-ReX, Wyatt technology, USA) through the software ASTRA V7.2 (Wyatt technology, USA). All conditions were prepared in 1.5 mL non-coated PP Eppendorf tubes. Finally, the formulations were sterile filtered with 0.22 µm cellulose acetate filters from VWR International (Germany). The purity of the proteins before formulation was studied by SEC and cEIF.[17] All the proteins presented in this work have recently been extensively characterized and their primary sequence is publicly available.[17]

2.2 Size exclusion chromatography combined with multi angle light scattering (SEC–MALS)

SEC-MALS was conducted on Agilent 1260 Bio-Inert system (Agilent Technologies, Germany) with a variable wavelength Ultraviolet (UV) detector operated at 280 nm (Thermo Fischer Scientific, USA), followed by a TREOS II detector (Wyatt Technology, USA) and an Optilab T-rEX (Wyatt Technology, USA). The temperature controlled autosampler was kept at 4 °C. Separation was performed with a Superdex 200 increased 10/30 GL column. Data collection and processing were performed using the ASTRA® software V7.2 (Wyatt Technology, USA). The aqueous mobile phase consisted of 38 mM NaH2PO4, 12 mM Na2HPO4, 150 mM NaCl and 200 ppm NaN3 at pH 7.4 dissolved in HPLC-grade water. The mobile phase was filtered through Durapore VVPP 0.1 µm membrane filters (Millipore, USA). The samples were centrifuged and were injected at a volume of 25 µl.

2.3 Stress assays

0.2 mL of each protein solution was aliquoted at a concentration of 1 mg/mL and sterile filtered in 0.5 mL sterile non-coated PP Eppendorf tubes. The samples were incubated at 4 °C, 25 °C, and 40°C for 2 weeks (i.e. accelerated assay) and 6 months (i.e. real time-storage). The samples were routinely analyzed, i.e. after 1 and 3 months of stress. After storage, the samples were left at 4 °C and measured within two weeks. Due to the high number of samples (i.e. 1008 per time point) only one replicate was collected by SEC-MALS. Chromatograms were routinely controlled and additional injections were conducted when deemed necessary.

2.4 Dynamic light scattering (DLS)

DynaPro® III plate reader (Wyatt Technology, USA) was used for the dynamic light scattering experiments. The measurements were performed in 1536 LoBase Assay Plates (Aurora Microplates, USA) in triplicates using 4 µL of sample sealed with a few µL of silicone oil. The plate was centrifuged for 1 min at 2000 rpm before placed in the plate reader. Data was collected and processed with the DYNAMICS® software V7.8 (Wyatt Technology, USA). The coefficient of self-diffusion, D, and the polydispersity index (PDI) were calculated from the obtained autocorrelation functions using cumulant analysis. The Stokes-Einstein equation was used to calculate R_h from D. The aggregation onset temperature (T_{agg}) was determined using protein concentration of 1 mg/mL. A temperature ramp of 0.1 °C/min was applied from 25 °C to 80 °C. One measurement included 3 acquisitions of 3 s. T_{agg} was calculated by the DYNAMICS® software V7.8 from the increase in R_h during heating.

2.5 Differential scanning fluorimetry with intrinsic protein fluorescence detection (nanoDSF)

Samples containing 1 mg/mL protein in the respective formulations were filled in standard nanoDSF capillaries (NanoTemper Technologies, Germany). Measurements were performed using the Prometheus

NT.48 (NanoTemper Technologies, Germany) system that measures the intrinsic protein fluorescence intensity at 330 nm and 350 nm after excitation at 280 nm. A temperature ramp of 1°C/min was applied from 20 to 95°C. The fluorescence intensity ratio (F350/F330) was plotted against the temperature, the inflection points of the unfolding transitions were determined from the first derivative of each measurement using the PR Control software V1.12 (NanoTemper Technologies, Germany).

2.6 Isothermal chemical denaturation (ICD)

ICD studies were performed on Unchained Labs HUNK system (Unchained Labs, USA). The excitation wavelength was 285 nm and emission intensities were recorded from 300 nm to 450 nm. The incubation time was varied depending on the protein studied. 48-point linear gradient was automatically generated for each condition. Guanidine hydrochloride (GnHCl) was used as denaturants. 6 M GnHCl stock solutions were prepared in each condition. Protein stock solutions were prepared at 1 mg/ml and diluted 12.5 times to the final condition. The data analysis was performed using the software Formulator V3.02 (Unchained Labs, USA). For the native protein, the fluorescence emission maximum λmax (native) was selected from the spectrum of the sample containing no denaturant. For the unfolded state, the fluorescence emission maximum λmax (den) was chosen from the fluorescence emission spectrum of the sample containing 5.5 M GnHCl. C_{m1} and m_1 were calculated for the first transitions.[36,37] Different state models were applied for different proteins.

2.7 Learning algorithms

All the algorithms presented in this work have been coded in MATLAB R2018a (MathWorks, USA)[38], unless differently stated, and the trained models may be available upon request. The functions used in the algorithms are part of the Statistic and Machine Learning Toolbox™ and of the Deep Learning Toolbox™. Data preprocessing involved the scaling of the features before training.

Table 1. *List of the proteins and formulations investigated.*

Protein					
Protein ID	Type	Batch	M_W (kDa)	pI	Notes
PPI-01	IgG1λ	SP11-255	144.8	7.96	-
PPI-02	Human IgG1κ	SP12-394	148.2	8.53	-
PPI-03	Human IgG1κ	SP12-423	144.8	8.44	Wild-type IgG
PPI-04	IgG1λ YTE	SP08-128/e	146.2	8.99	-
PPI-08	IgG1κ + scFv	72635-132	204.4	8.9-9.2	Bispecific
PPI-10	Human IgG1	SP07-212/1	144.2	8.95	-
PPI-11	IgG4λ-P YTE	SP09-384	146.3	9	-
PPI-13	Human IgG1κ TM	SP10-016	148.9	9.04	BFI, no VRF
PPI-17	IgG2κ	72635-056	145.1	7.78	-
PPI-18	HSA-NEP	SP11-285	146.7	5.8	Conjugate
PPI-30	Ifnα-2A (ROC)	N/A	19.2	5.97	Interferon alpha-2A
PPI-44	Transferrin	UK08E0103A	74.9	6.58	-
PPI-45	Lipozyme RM	U8BJ6	29.5	4.7	pdb code: 5TGL
PPI-46	Lipolase	U8D3D	29	4.8	pdb code: 1GT6
PPI-49	HAS	RF20-001	66.4	5.67	

Formulation					
Rank**	Buffer	pH	NaCl (mM)	Excipient type	Excipient concentration (mM)
5	His	5 / 5.5*	0	Proline	280
10	His	5 / 5.5*	0	Arg.HCl	140
9	His	5 / 5.5*	0	Sucrose	280
7	His	5 / 5.5*	0	-	-
11	His	5 / 5.5*	140	Proline	280
9	His	5 / 5.5*	140	Arg.HCl	140
9	His	5 / 5.5*	140	Sucrose	280
8	His	5 / 5.5*	140	-	-
6	Acetate	5 / 5.5*	0	Proline	280
10	Acetate	5 / 5.5*	0	Arg.HCl	140
8	Acetate	5 / 5.5*	0	Sucrose	280
5	Acetate	5 / 5.5*	0	-	-
6	His	6.5 / 7.5*	0	Proline	280
6	His	6.5 / 7.5*	0	Arg.HCl	140
8	His	6.5 / 7.5*	0	Sucrose	280
5	His	6.5 / 7.5*	0	-	-
5	His	6.5 / 7.5*	140	Proline	280
6	His	6.5 / 7.5*	140	Arg.HCl	140
5	His	6.5 / 7.5*	140	Sucrose	280
7	His	6.5 / 7.5*	140	-	-
9	Phosphate	6.5 / 7.5*	0	Proline	280
10	Phosphate	6.5 / 7.5*	0	Arg.HCl	140
14	Phosphate	6.5 / 7.5*	0	Sucrose	280
10	Phosphate	6.5 / 7.5*	0	-	-

**PPI-30, PPI-45, PPI-46; **see section 3.1*

This step was necessary to investigate features with different range of values (e.g. T_m and LS_m). 5 fold cross validation was preferred for the decision trees and linear regressions algorithms, while for ANNs different approaches were applied (see next section). ANNs and linear regression were applied to fit the data while the decision tree algorithms were applied to build classifiers. The analyzed data used in this study will be available via a specially-designed publicly-available database for protein formulation (https://pippi-data.kemi.dtu.dk/).

2.7.1 Application of machine learning to predict long term storage stability

Our target is to predict the monomer retention at the temperature X (a_n^X) after 6 months of incubation, which is calculated from:

Eq. 1
$$a_n^X = (\frac{U_{mon}^X}{U_{tot}^X}) \cdot R_X$$

Where U_{mon}^X is the monomer UV area after storage at the temperature X, U_{tot}^X is the UV area of all the peaks after storage at the temperature X, and R_X is the calculated recovered monomer mass ratio at the temperature X. R_x is calculated as the total peaks mass divided by the total injected mass. Therefore, a_n^X is "normalized" by taking into account the mass loss due to soluble and insoluble aggregates (which are filtered through the column or before injection). Protein column recovery was tested right before formulation to confirm complete recovery.

The primary features selected to predict a_n^X were the reduced monomer retention m_n^X, which is calculated similarly to a_n^X in Eq.1 using data after two weeks of storage, and the light scattering area ratio at the temperature X after two weeks of stress, calculated from:

Eq.2
$$LS_m^X = \frac{LS_{tot}^X}{LS_{mon}^X}$$

Where LS_{tot}^X is the sum of the light scattering peaks area at the temperature X, and LS_{mon}^X is the light scattering monomer area at the temperature X. Furthermore, in order to tell the "prediction power" of

typical high-throughput biophysical assays we developed a series of models which took additional features into account, namely: the melting temperature of the first transition (T_m), the onset temperature of unfolding ($T_{m,\,on}$), the onset temperature of aggregation ($T_{agg,\,on}$), the denaturant concentrations for 50% of the first chemical unfolding process (C_{m1}), the slope of the first chemical unfolding process (m_1). Then features were optimized as appropriate for each algorithm type to individuate the minimum amount necessary for an accurate prediction. Multiple models were developed when necessary including different features.

2.7.2 Artificial neural networks

The architecture of our feed-forward back-propagation ANNs is shown in **Fig. 1**. This presents one hidden layer, which is usually sufficient to provide adequate predictions even when continuous variables are adopted as units in the output layer.[24,39]

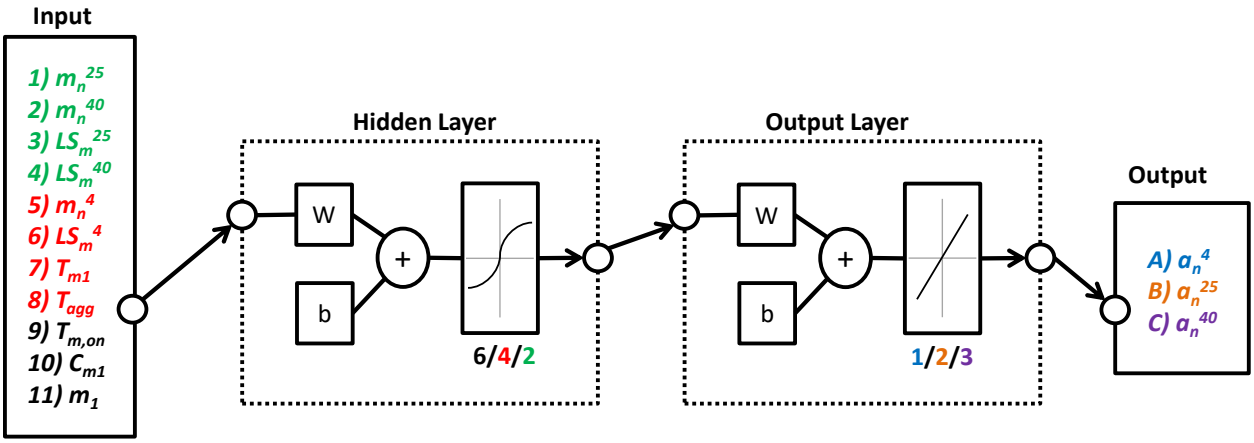


Figure 1. ***Neural network architectures.*** *The input layer is fully connected to the neurons in the hidden layer by the weights, w, which multiplies the corresponding signal. All the weighted signals are summed by a summing function and an external bias, b, is applied. Finally, the signal is connected to a sigmoidal activation function, which limits the amplitude of the signal to the output layer. The neurons used are 2, 4 and 6 respectively for models using the inputs 1-4, 1-8 and 1-11. Similarly, the output layer differs based on the prediction output.*

The initial number of neurons to optimize was selected as previously described.[24] Then, due to the low computational time required to train our shallow ANNs, usually less than a minute, the optimal number of neurons was optimized based on grind search, which was evaluated on the test set. Similarly, combinations of input and output parameters were optimized by investigating all the possible combinations. We selected a Bayesian regularization backpropagation according to Levenberg-Marquardt optimization[40] to minimize a linear combination of squared errors and weights so that at the end of training the resulting network has good generalization qualities.[41,42]

This algorithm typically requires more time (yet within a minute of computation in our case), but can result in good generalization for difficult, small or noisy datasets. As the training stops according to adaptive weight minimization, no validation was necessary and we randomly divided the dataset for these ANNs into training and test sets (respectively 70% and 30%). The features and outputs of each ANN validated in such way are enlisted in **Table 2**. Similarly to the features selection a grind search of output combinations has been explored, which resulted in multi-output architecture for the prediction of both a_n^{25} and a_n^{40}. Other ANNs models were developed by a different cross-validation method, i.e. "leave-one-protein-out". These experiments were executed to test the robustness of our approach. In this case the data of only one protein at the time was used to test the models, which in turn were trained with the remaining data (**Table 3**).

2.7.3 Linear regression

We assessed the full model using all the available features, and then reduced the model to only the terms that were statistically deemed relevant. A curvature response was allowed by assessing the quadratic term considering also two-way interactions. The reduced model was obtained using a backward stepwise elimination. The F-statistic approach was used to perform the effect test considering a value of 0.05 or less as statistically significant. This yielded the SEC-MALS parameters, namely m_n^{25}, m_n^{40}, LS_m^{25},LS_m^{40}, as the statistically relevant features to be used. These calculations were performed by the statistical

software JMP® v 14.0 (SAS Institute Inc., Cary, USA). As no significant interaction was observed we trained the linear model assuming no interactions of the selected features.

Table 2. *List of the features and outputs for models A1-3, B1-3, C1-3, D1-3 and E1-3. "Y" stands for "yes", i.e. included as feature, while "N" stands for "no", i.e. not included as feature. Notice that models D and E include only data from mAbs and non-mAbs respectively, while the other models include the whole dataset.*

ID	Outp.	R (Training)	R (Test)	RMSE (Test)	m_n^{25}	m_n^{40}	LS_m^{25}	LS_m^{40}	m_n^{4}	LS_m^{4}	T_m	$T_{agg,on}$	$T_{m,on}$	C_{m1}	m_1
A1	a_n^4	0.94	0.91	$5.8 \cdot 10^{-3}$	Y	Y	Y	Y	N	N	N	N	N	N	N
A2	a_n^4 a_n^{25}	0.93	0.84	$3.8 \cdot 10^{-3}$	Y	Y	Y	Y	N	N	N	N	N	N	N
A3	a_n^4 a_n^{25} a_n^{40}	0.91	0.77	$1.7 \cdot 10^{-2}$	Y	Y	Y	Y	N	N	N	N	N	N	N
B1	a_n^4	0.97	0.95	$1.8 \cdot 10^{-3}$	Y	Y	Y	Y	Y	Y	Y	Y	N	N	N
B2	a_n^4 a_n^{25}	0.96	0.89	$5.0 \cdot 10^{-3}$	Y	Y	Y	Y	Y	Y	Y	Y	N	N	N
B3	a_n^4 a_n^{25} a_n^{40}	0.96	0.84	$6.0 \cdot 10^{-2}$	Y	Y	Y	Y	Y	Y	Y	Y	N	N	N
C1	a_n^4	0.98	0.94	$0.1 \cdot 10^{-3}$	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
C2	a_n^4 a_n^{25}	0.95	0.87	$1.1 \cdot 10^{-3}$	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
C3	a_n^4 a_n^{25} a_n^{40}	0.96	0.84	$1.3 \cdot 10^{-2}$	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
D1	a_n^4	0.95	0.91	$2.0 \cdot 10^{-3}$	Y	Y	Y	Y	Y	Y	Y	Y	N	N	N
D2	a_n^4 a_n^{25}	0.94	0.83	$6.2 \cdot 10^{-3}$	Y	Y	Y	Y	Y	Y	Y	Y	N	N	N
D3	a_n^4 a_n^{25} a_n^{40}	0.94	0.81	$1.5 \cdot 10^{-2}$	Y	Y	Y	Y	Y	Y	Y	Y	N	N	N
E1	a_n^4	0.99	0.93	$1.1 \cdot 10^{-3}$	Y	Y	Y	Y	Y	Y	Y	Y	N	N	N
E2	a_n^4 a_n^{25}	0.98	0.95	$3.3 \cdot 10^{-3}$	Y	Y	Y	Y	Y	Y	Y	Y	N	N	N
E3	a_n^4 a_n^{25} a_n^{40}	0.99	0.93	$1.7 \cdot 10^{-2}$	Y	Y	Y	Y	Y	Y	Y	Y	N	N	N

2.7.4 Decision tree

Our top-down induction decision tree used the Gini's diversity index as splitting criterion.[43] The maximum number of splits was initially set to 10 and then gradually reduced until a balance between model performance and generality was achieved. This search was based on the average R^2 and RMSE values from the 5-fold cross validation by reducing the split until a significant drop of the prediction power was observed. Further, the true and false positive rates were investigated by means of confusion matrixes and receiver operator characteristic (ROC) curves. Similarly to the linear regression models, the optimized

selection of splits included only rules based on the SEC-MALS parameters, namely m_n^{25}, m_n^{40}, LS_m^{25},LS_m^{40} . Samples were classified as high stability (H), medium stability (M) and low stability (L) respectively when they presented $a_n^X > 0.8$, $0.8 \leq a_n^X \leq 0.5$ and $a_n^X < 0.5$. This classification was arbitrarily selected, as previously described,[17] in order to investigate the prediction power of the decision tree algorithm and to easily visualize the data as described in section 2.1.

3 Results

3.1 The target: protein monomer retention after long term storage

Although accelerated stability testing is an attractive practice it is beset by many pitfalls that arise from complex deterioration profiles.[4] To develop an accurate empirical model based on data acquirable in a timely manner we investigated samples by SEC-MALS stored at 4°C, 25°C and 40°C for 2 and 24 weeks. The samples were analyzed also after 4 and 12 weeks as a control. We used the SEC-MALS data after two weeks of storage as feature to predict the monomer retention after six months of storage. The light scattering detector can often identify high molecular weight species which are not clearly detectable by UV. The M_w of those species cannot be calculated as the concentration of the sample is unknown.[44] Therefore, in most cases these small populations of species are disregarded during the calculations. Nevertheless, these small populations of sample after two weeks of stress can be well correlated with the long term monomer retention (**Fig. S1**). However, the relationship between the formation of small population of high molecular weight species detected by MALS and the monomer retention is not linear, it can involve multiple peaks and it is formulation and protein dependent. Our hypothesis was that the light scattering signal could be used to refine our models, especially at higher temperatures. Moreover, we processed the monomer column retention to take into account also the formation of insoluble particles, which do not elute in SEC (see material and methods).

A total of 336 samples were investigated. The samples included 14 protein formulated in 24 different conditions (**Table 1**) stored at 4°C, 25°C and 40°C, for a total of 1008 data points at each time point. As a brief overview, we clustered the result by protein (**Fig. 2**) and by formulation (**Fig. 3**).

Most proteins show a rather high stability at 4°C. This is understandable as the protein drug candidates are engineered to be sufficiently stable. Overall PPI-03, PPI-13, PPI-2 and PPI-49 are the most stable, while PPI-18 and PPI-45 show the least thermal stability. We noticed formation of big insoluble aggregates in some cases. The formulation cluster (**Fig. 3**) indicates a rather consistent distribution of low, medium and high stability samples for each formulation. After 6 month storage at 4°C the phosphate buffered formulations comprised the lowest number of highly stabile samples and the highest number of lowly stable samples. The formulations could be ranked by the sum of sample with low stability at the three temperatures (**Table 1**). A slight linear correlation was observed between the monomer retention after 6 months at 4 and 40°C (R=0.23) and at 4 and 25 °C (R=0.48).

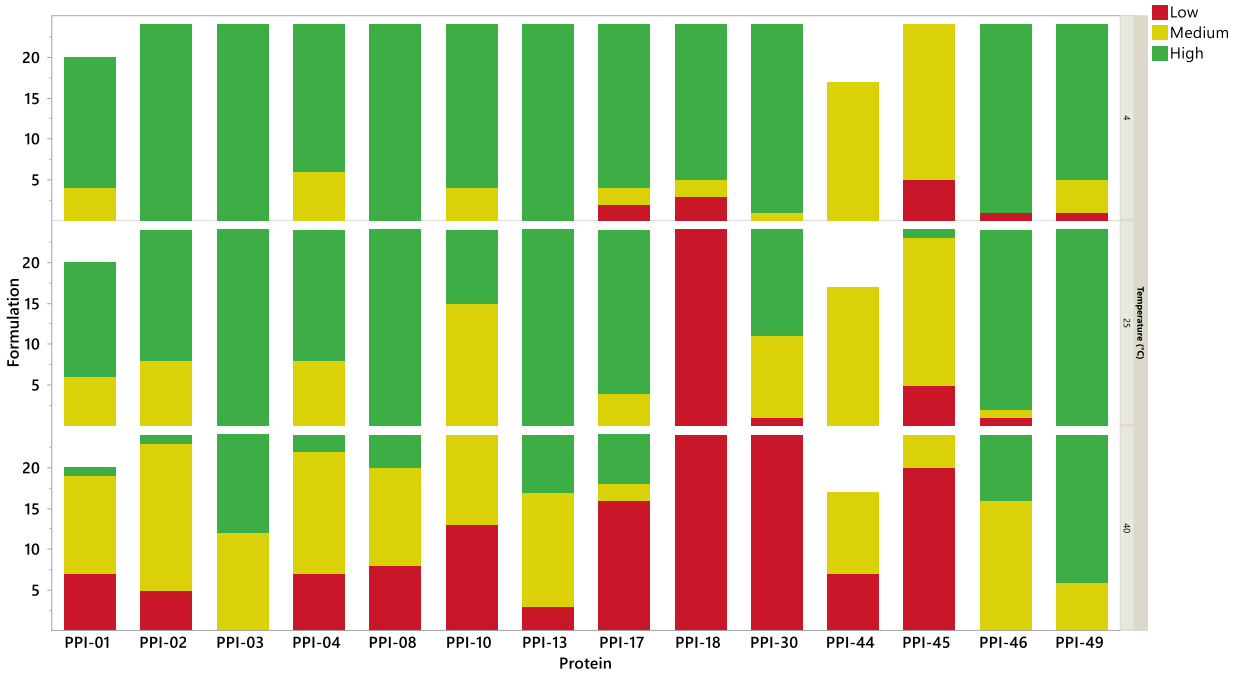


Figure 2. *The samples were classified as high stability (Green-High), medium stability (Yellow-medium) and low stability (red-low) respectively when they presented* a_n^X *>0.8, 0.8≤* a_n^X *≤0.5 and* a_n^X*<0.5. The total number of resulting formulations is clustered by protein.*

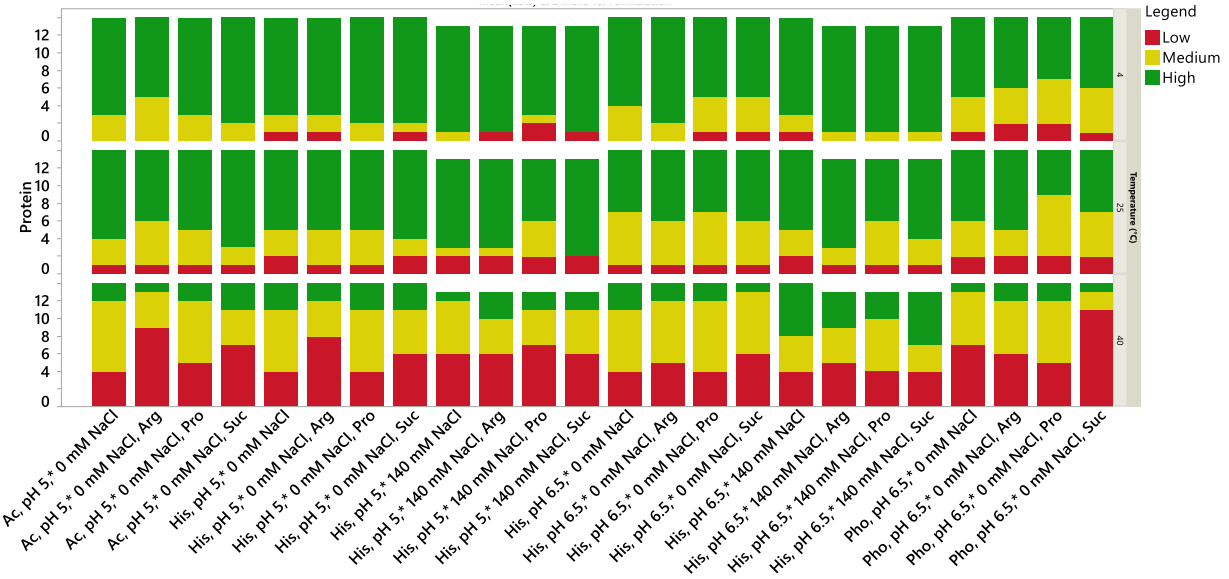


Figure 3. *The samples were classified as high stability (Green-High), medium stability (Yellow-medium) and low stability (red-low) respectively when they presented* a_n^X *>0.8, 0.8≤* a_n^X *≤0.5 and* a_n^X*<0.5. The number of proteins is clustered by formulation.*

3.2 Artificial neural networks

ANNs were applied to predict a_n^4, a_n^{25} and a_n^{40}. We developed a total of 11 models validated by random holdout to train and test ANNs for the whole dataset (**Fig. 4, Table 2**). In **Fig. 1** the models architecture is shown, including the list of features, amount of neurons in the hidden layer and the output layer. Multi-outputs algorithms were selected to predict a_n^{25} and a_n^{40}, e.g. model A3. This approach was preferred to a single output, e.g. a_n^{40}, as the fitting and prediction of these parameters was significantly improved with the former approach. This is possibly due to the increased amount of training data to predict the more complex and less linear behavior at higher temperature. In fact, the number of data point available increased for each output included in the prediction, i.e. 336 runs for a_n^4 , 672 for a_n^4 and a_n^{25},1008 for a_n^4, a_n^{25} and a_n^{40}. Models A1, A2 and A3 (**Fig. 4, Table 2**) used only SEC-MALS data to predict a_n^4, a_n^4-a_n^{25}, *or* a_n^4-a_n^{25}-a_n^{40}. Similarly, models B1, B2, B3 (**Fig. 4, Table 2**) utilized the same features as models A1, A2 and A3 plus $T_{agg,on}$, T_{m1}, m_n^4 and LS_m^4. Finally, models C1, C2, C3 (**Fig. 4, Table 2**) use the same features as B1, B2 and B3 plus $T_{m,on}$, C_{m1}, m_1. We observed a good prediction power for A1, B1 and C1

(i.e. to predict a_n^4) with a R values between 0.91 to 0.94, and low RMSEs (**Fig. 4, Table 2**). Thus SEC-MALS data after 2 weeks of stress in combination with a simple network are sufficient to predict 6 month monomer retention at intended storage temperature. In addition ANNs showed a good accuracy in predicting monomer retention after storage at elevated temperature (a_n^{25} and a_n^{40}). This in turn allows to better comprehend the sample stability especially for highly stable samples (e.g. PPI-3 and PPI-46). This is possibly due to the rather good accuracy of the models A2-3, B2-3 and C2-3 (**Fig. 4, Table 2**). We noticed that including more biophysical parameters (e.g. C3) yielded a better linear correlation between models output and actual data. This is due to a more accurate prediction of samples with low stability. The models B1, B2 and B3 offer a good high-throughput approach to predict monomer retention after long term storage at refrigerated and elevated temperature as all the data necessary for the models can be collected in a timely manner and with low material consumption. Differently, C1, C2 and C3 necessitate isothermal chemical denaturation (ICD) data which requires more time and material.[45] Therefore, we selected the models B1, B2 and B3 to further investigate the robustness of our approach for the different classes of proteins investigated. We divided our dataset into mAbs (models D1, D2 and D3, **Fig. 5, Table 2**) and non-mAb proteins (models E1, E2 and E3, **Fig. 5, Table 2**). Good prediction was achieved for both datasets (**Fig. 5, Table 2**). Furthermore we noticed that D1 presented the lowest RMSE compared to the other models. This is probably due to the high stability of the investigated mAbs.

Overall, ANNs offer higher of accuracy compared to other approaches (e.g. linear regression). In fact, ANNs are known to have robust performance in dealing with noisy or incomplete datasets and the ability to generalize highly non-linear problem with a high fault tolerance.[46] Further, we investigated the robustness of the approach when dealing with unknown proteins as described in section 3.3.

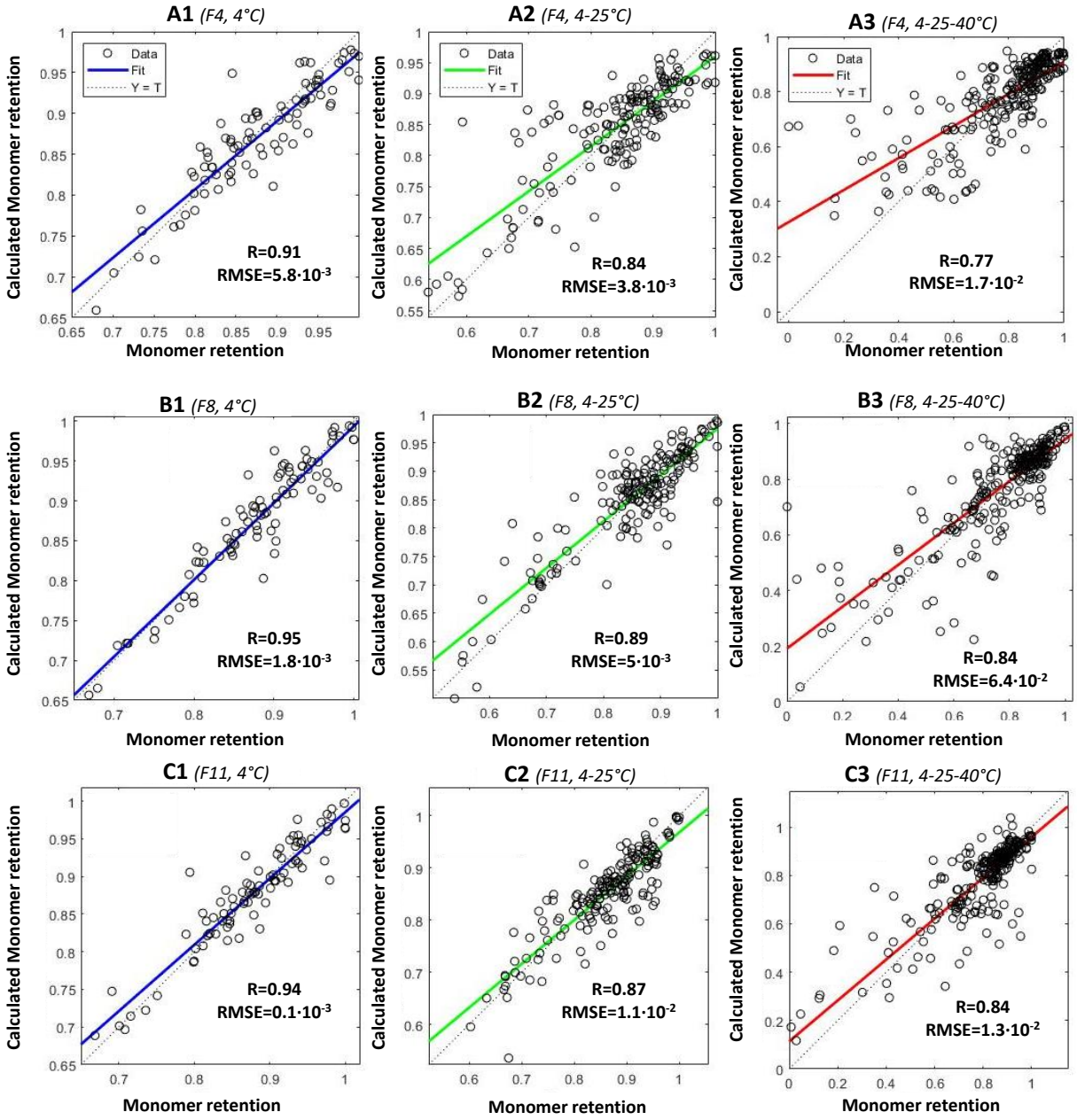


Figure 4. Neural network models to predict long term stability of therapeutic protein at different storage temperature. *he models A1-3, B1-3 and C1-3 represent the linear relationship between the targets and the model outputs. The features used for the model are either 4, 8 or 11 respectively for the A, B and C model. The list of the features is shown in Fig. 1. In brackets information on each model is provided, e.g. "(F11, 4-25°C)" stands for 11 inputs to predict the monomer retention after 6 month of storage at 4 and 25°C.*

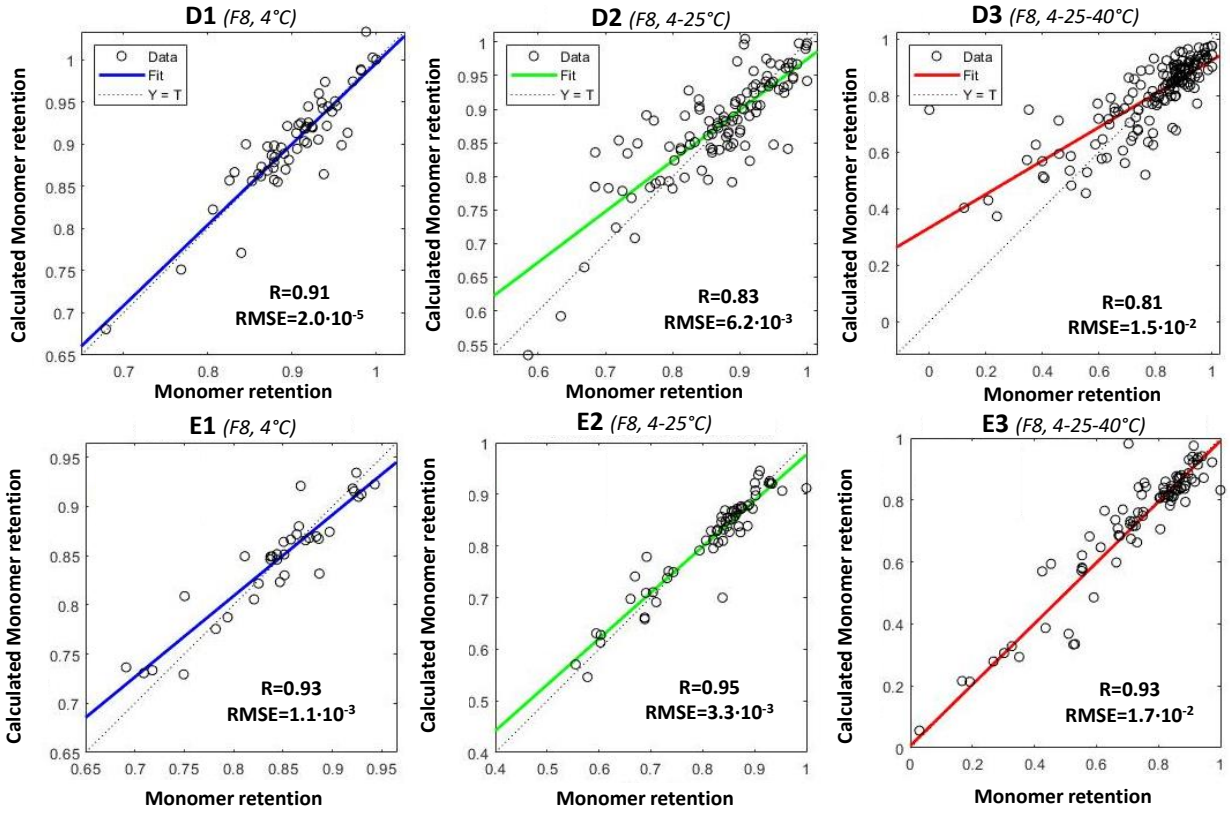


Fig. 5 Neural network models to predict long term stability of mAb and non-mAb. *The figure is organized similarly to Fig. 4. D-models are relative to mAbs only, while E-models are relative to non-mAbs molecules.*

3.3 Cross-validation: leave-one-protein-out

Although formulation ranking is of paramount importance in product development, we expect our approach to be of use especially to predict the monomer retention after long term storage of new proteins from short term data (i.e. proteins inexperienced by the models during the training phase). Therefore new models were trained and cross-validated leaving the data of one protein out for each model. The models investigated presented the same architecture as the previous models. The results for the computed 378 models are summarized in **Table 3**. We observed a drop of R for the testing of PPI-13 and 46 after 6 months storage at 4°C as both proteins vary only minimally in monomer retention between the formulations and exhibit a very high stability. In turn the models for testing PPI-13 and 46 predict high

stability and cluster all the data points in the right region, yielding low Rs and low RMSE (**Table 3**). This means that the models are accurate, and cluster all the data points properly (i.e. all the values predicted are clustered around the value of 1, which means no difference in formulation is observed). At higher storage temperature the data points are more distributed and the R values of these two sets increase. This case demonstrates that access to the prediction at different temperatures can be beneficial.

Most non-mAb proteins presented lower R values and usually higher RMSE values. In all cases the formulations are properly ranked, but the fitting line presented a slightly different slope. The prediction of monomer retention decreased in accuracy at high temperature (with the exception for the model tested by PPI-13 and PPI-46 as aforementioned). The prediction of PPI-30 stability at 40°C presented negative R for models of type A and B. However, the addition of ICD features in C models allows a proper positive correlation. Differently, the prediction of PPI-49 at high temperature is impaired for model type C.

In **Fig. 6** the averaged R values and standard deviations calculated from **Table 3** for the train and validation sets are showed. The standard deviations are significantly reduced if only mAbs are included. As expected, the models predicting a_n^4 present the highest prediction accuracy. A significant drop in prediction power was observed for multi-output models. Further, we observed that the implementation of more input parameter to predict a_n^4 has only a limited effect on the prediction accuracy, while slightly increasing the fitting capabilities of the network. On the other hand, architecture B and C presented higher R values and lower standard deviation for the multi-output models.

3.4 Linear regression and decision tree classifier

ANNs have a series of known limitations, namely overfitting, chance effect, overtraining and difficult interpretability.[47–50] The interpretation of ANNs is still an open field of research and ANNs are often named as black-box models. One approach for the interpretation of ANNs is to develop human-friendly surrogate models. In order to provide such surrogate model we developed linear regression and decision tree models.

Table 3. Leave-one-protein-out cross-validation (Part 1 of 2). *Summary of the results for the 378 models generated by the leave-one-protein-out cross-testing procedure. The features used for the model are either 4, 8 or 11 respectively for the A, B and C model, similarly to models showed in Fig. 5. The list of the features is shown in Fig. 7. Information for each model is summarized for each model type, e.g. "(F11, 4-25°C)" stands for 11 inputs to predict the monomer retention after 6 month of storage at 4 and 25°C.*

Models C	***F11 4°C***			***F11 4-25°C***			***F11 4-25-40°C***		
Protein	**R (Train)**	**R (Test)**	**RMSE (Test)**	**R (Train)**	**R (Test)**	**RMSE (Test)**	**R (Train)**	**R (Test)**	**RMSE (Test)**
PPI-01	0.98	0.95	0.024	0.93	0.84	0.031	0.94	0.85	0.036
PPI-02	0.98	0.95	0.0358	0.94	0.81	0.023	0.94	0.84	0.021
PPI-03	0.97	0.96	0.008	0.93	0.8	0.001	0.93	0.89	0.041
PPI-04	0.98	0.91	0.077	0.94	0.44	0.1824	0.94	0.68	0.1936
PPI-08	0.98	0.86	0.0177	0.94	0.6	0.0296	0.95	0.76	0.056
PPI-10	0.98	0.9	0.006	0.93	0.7	0.042	0.94	0.75	0.094
PPI-13	0.98	0.46	0.009	0.94	0.44	0.033	0.95	0.77	0.0291
PPI-17	0.97	0.55	0.0246	0.93	0.41	0.02	0.93	0.73	0.097
PPI-18	0.98	0.96	0.0185	0.94	0.83	0.0185	0.94	0.78	0.0185
PPI-30	0.98	0.73	0.1452	0.94	0.72	0.16	0.94	-0.6	0.3514
PPI-44	0.96	0.7	0.047	0.93	0.57	0.065	0.93	0.44	0.154
PPI-45	0.97	0.56	0.001	0.92	0.5	0.029	0.95	0.87	0.06
PPI-46	0.98	0.17	0.003	0.93	0.21	0.0354	0.94	0.71	0.193
PPI-49	0.98	0.93	0.049	0.94	0.49	0.031	0.94	0.53	0.0408
Models B	***F8 4°C***			***F8 4-25°C***			***F8 4-25-40°C***		
Protein	**R (Train)**	**R (Test)**	**RMSE (Test)**	**R (Train)**	**R (Test)**	**RMSE (Test)**	**R (Train)**	**R (Test)**	**RMSE (Test)**
PPI-01	0.96	0.93	0.036	0.93	0.83	0.025	0.89	0.84	0.053
PPI-02	0.96	0.93	0.017	0.93	0.84	0.018	0.89	0.85	0.021
PPI-03	0.95	0.94	0.004	0.92	0.83	0.009	0.89	0.88	0.022
PPI-04	0.95	0.88	0.026	0.93	0.6	0.118	0.9	0.68	0.195
PPI-08	0.96	0.86	0.005	0.93	0.57	0.0465	0.87	0.77	0.059
PPI-10	0.95	0.93	0.009	0.94	0.68	0.048	0.9	0.7	0.109
PPI-13	0.96	0.14	0.0348	0.93	0.56	0.019	0.89	0.77	0.0146
PPI-17	0.96	0.61	0.0349	0.93	0.36	0.0239	0.89	0.78	0.071
PPI-18	0.96	0.84	0.0259	0.92	0.89	0.0259	0.91	0.9	0.0259
PPI-30	0.95	0.53	0.1956	0.93	0.64	0.187	0.87	-0.77	0.4164
PPI-44	0.96	0.61	0.051	0.9	0.53	0.073	0.89	0.55	0.176
PPI-45	0.93	0.62	0.066	0.91	0.54	0.023	0.86	0.83	0.075
PPI-46	0.97	0.46	0.0194	0.93	0.2	0.0213	0.9	0.75	0.171
PPI-49	0.96	0.94	0.046	0.93	0.31	0.0432	0.9	0.35	0.059

Table 3. Leave-one-protein-out cross-validation (Part 2 of 2).

Models A	*F4 4°C*			*F4 4-25°C*			*F4 4-25-40°C*		
Protein	**R (Train)**	**R (Test)**	**RMSE (Test)**	**R (Train)**	**R (Test)**	**RMSE (Test)**	**R (Train)**	**R (Test)**	**RMSE (Test)**
PPI-01	0.94	0.91	0.038	0.85	0.81	0.026	0.86	0.86	0.04
PPI-02	0.93	0.92	0.021	0.86	0.67	0.032	0.81	0.8	0.012
PPI-03	0.93	0.92	0.0143	0.84	0.83	0.024	0.82	0.83	0.104
PPI-04	0.92	0.91	0.0294	0.87	0.3	0.194	0.83	0.77	0.599
PPI-08	0.92	0.83	0.0121	0.85	0.69	0.01	0.78	0.68	0.054
PPI-10	0.94	0.94	0.005	0.87	0.66	0.046	0.81	0.79	0.081
PPI-13	0.92	0.46	0.004	0.84	0.4	0.012	0.85	0.84	0.027
PPI-17	0.94	0.35	0.023	0.86	0.2	0.034	0.79	0.36	0.04
PPI-18	0.91	0.84	0.01	0.85	0.72	0.01	0.77	0.55	0.01
PPI-30	0.93	0.83	0.0334	0.87	0.63	0.0562	0.77	0.87	0.182
PPI-44	0.91	0.6	0.052	0.79	0.31	0.077	0.76	0.27	0.145
PPI-45	0.88	0.76	0.0518	0.86	0.46	0.041	0.77	0.83	0.113
PPI-46	0.92	0	0.005	0.85	0	0.033	0.78	0.68	0.134
PPI-49	0.93	0.9	0.0371	0.86	0.29	0.034	0.79	0.26	0.116

Linear regression (**Fig.7**) enabled to predict a_n^4 *and* a_n^{25} with R^2 values of 0.79 and 0.62 respectively, by means of the following equations:

Eq. 3 $$a_n^4 = 0.1149 + 0.8856 \cdot m_n^{25} + 0.0569 \cdot LSm^{25} - 0.0487 \cdot m_n^{40}$$

Eq. 4 $$a_n^{25} = 0.1091 + 0.8723 \cdot m_n^{25} + 0.0658 \cdot LSm^{40} - 0.0817 \cdot m_n^{40}$$

The features in Eq. 3 and 4 are listed in order of relevance evaluated by F-statistic approach. The F-statistic approach applied for the linear regression to select statistically significant features confirmed that other investigated parameters (e.g. T_m) were found to be statistically insignificant to linearly fit a_n^X. Differently, ANNs fitting capability was slightly improved by including the biophysical parameter especially to predict the monomer retention at 40°C. m_n^{25} was confirmed as the most important factor to predict a_n^4 *and* a_n^{25}. No fit was achieved for samples stored at 40°C. LS_m^{25} and LS_m^{40} have been found statistically significant respectively to predict a_n^4 and a_n^{25}. The root means square error (RMSE) averaged over the cross-validation was 3.3e-2 for a_n^4 and 6.5e-2 for a_n^{25}, which are an order of degree higher than the RMSE yielded by ANNs.

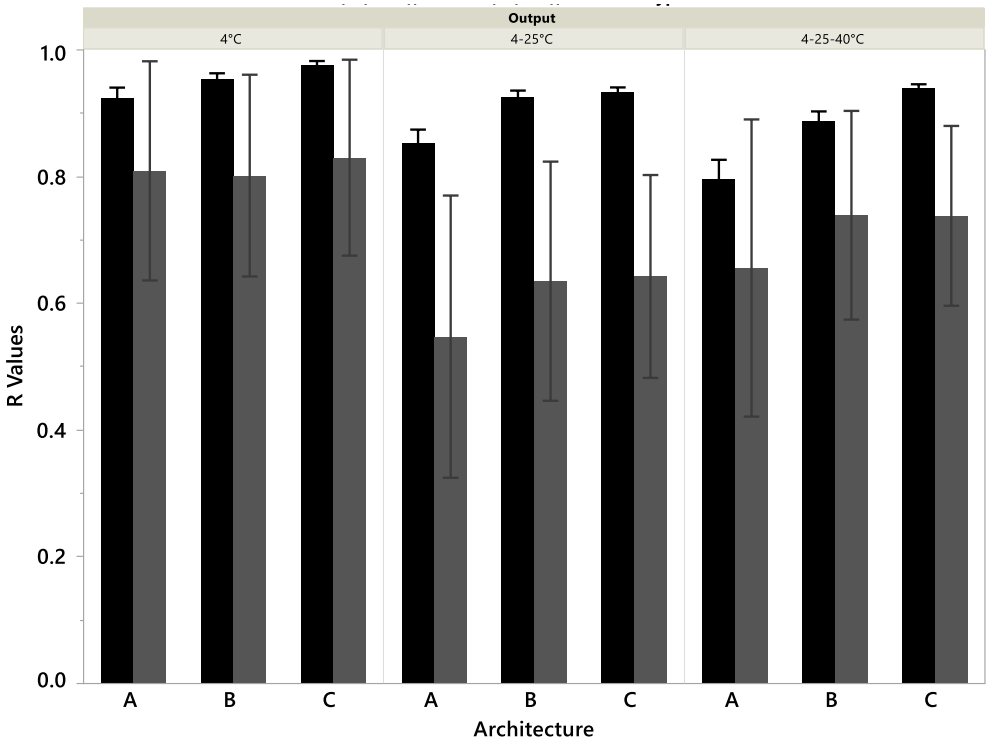


Fig. 6 Leave-one-protein-out cross-validation. *The mean R values for the train and validation set are depicted respectively in black and grey. The error bars represent the standard deviation. The values were averaged from the data listed in Table 3. PPI-13 and PPI-46 were not included in the calculation as the R values are not representative, as mentioned in the text, of the prediction accuracy. Similarly, PPI-30 resulted in negative values and was considered as an outliner and not included.*

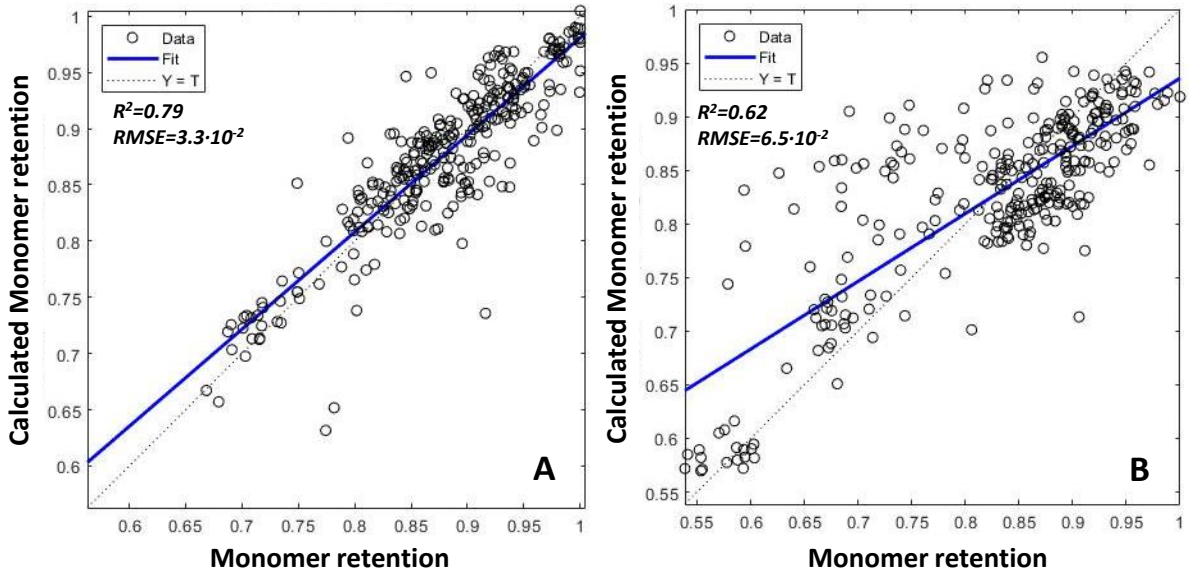


Figure 7. Linear regression prediction at 4°C (A) and 4°-25°C (B).

We also included in this work the evaluation of decision tree models. In **Fig. 8** the calculated rules of the selected model are shown. The averaged cross-validation accuracy is of 90.5% and 80.7% to predict the classes (calculated as described in material and method) for the samples stored at 4 °C and 25°C respectively. No reasonable prediction was achieved if samples stored at 40°C were included. Further pruning did not significantly improve the performance of the model. The individuated rules suggest that if a sample after two weeks of stress yields a $m_n^{25} > 0.77$ and *a* $m_n^{40} > 0.32$ it has very high chances to be stable after six months of stress. However, the model fails to classify poorly stable samples (i.e. a_n^X<0.5) and the accuracy of cross-validation is due to the high number of highly stable sample (i.e. a_n^X>0.8). Therefore, we concluded that decision tree models were not suitable to predict a_n^X. Nevertheless, the decision tree algorithm intuitively highlighted the relative importance of the parameters relative to two weeks accelerated stability studies calculated by SEC-MALS, which were also selected from the linear models and the ANNs.

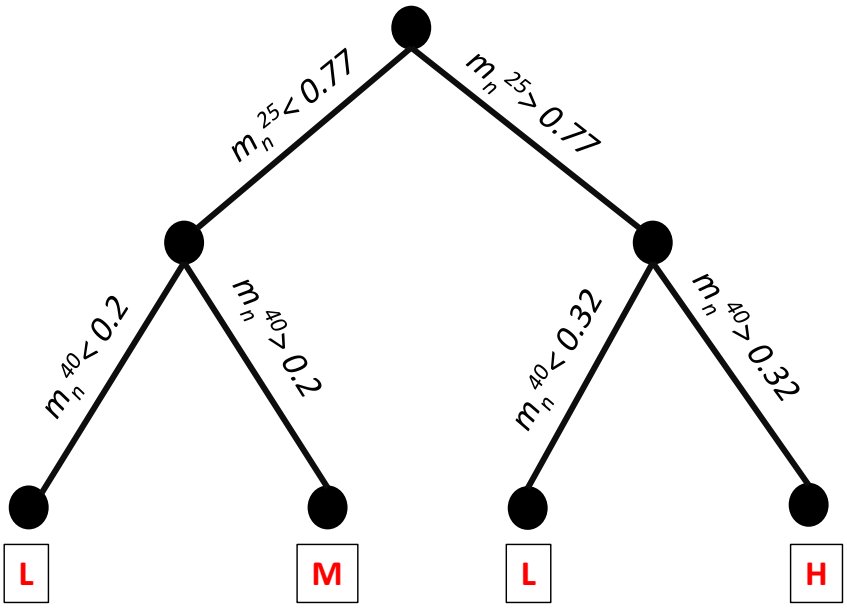


Figure 8. Decision Tree. *Each node (black dots) corresponds to one input variables, while each leaf (squares) represent the target value given the values of the input variables. The connection between the dots represents a conditional statement control. The samples were classified as high stability (H), medium stability (M) and low stability (L) respectively when they presented* a_n^X *>0.8, 0.8≤* a_n^X *≤0.5 and* a_n^X*<0.5.*

4 Discussion

Proteins tend to aggregate under a variety of conditions. The extent of aggregation depends on several factors that can be generally distinguished as intrinsic (e.g. primary structure) and extrinsic (e.g. the formulation). In our study we developed empirical ANNs capable to accurately predict the real-time aggregation of therapeutically relevant proteins in pharmaceutically relevant formulations based on accelerated stability studies and typically used biophysical parameters. Other *in silico* tools have been developed which predict the intrinsic aggregation propensity of certain classes of proteins[9–13]. However, our ANNs have a series of advantages: i) the networks output provide an accurate prediction of the sample (meant as the combination of a certain protein in a certain formulation) monomer retention after long term storage at refrigerated condition, room temperature and elevated temperature, allowing a better understanding of the sample thermal stability ii) the networks output is a value calculated by SEC and could be validated in every lab, iii) the networks can be retrained in different laboratories to further refine their performance iv) no primary sequence or protein structure is necessary, v) the ANNs have been trained and tested over different classes of therapeutically relevant protein in several formulations, including different excipients, providing a means to predict protein stability as a function of the formulation. We propose 11 models which necessitate different levels of analytical effort. These have been cross-validated by randomly dividing the dataset in test and validation set. SEC-MALS analysis of samples stored for two weeks at 25 °C and 40°C is the key to predict the monomer retention of proteins after 6 months storage at refrigerated conditions. We have recently shown that such accelerated stability studies do not correlate with other biophysical parameters when several proteins and formulations are considered.[17] The models present a reduced, but still acceptable accuracy in prediction of monomer retention after long term storage at elevated temperatures i.e. 25 °C and 40°C. To increase the accuracy the high throughput biophysical parameters $T_{agg,on}$ and T_m can be included. We showed in a previous study how those two can be accurately computed ab-priori by means of different ANNs as a function of basic formulation conditions.[24] Finally, to further refine the prediction of samples stored at 40°C ICD data can be included. A different cross-validation approach was also investigated to test the robustness of our approach when dealing with unknown proteins. 378 models were developed and cross-validated to predict new proteins. We demonstrate a rather high robustness of the models and individuate that the

stability of new mAbs have the highest chances of being accurately predicted. This highlights the robustness and validity of our approach to predict the stability of candidates at long term refrigerated storage. Monomer retention of the monoclonal antibodies, but for PPI-17, was accurately predicted. PPI-17 presents the formation of insoluble aggregates which has been shown to be difficult to predict also with other approaches.[24] Interestingly, high prediction power has been observed for PPI-01, which presents a complex reversible native self-association mediated by the Fab-Fab fragments.[51] The prediction of the monomer retention for non-mAbs was less accurate since only mAbs were extensively represented in the investigated data set. For instance, models which do not include PPI-30 in the training set cannot gain experience on the aggregation pathway of such protein class.

The two validation approaches, i.e. random holdout and "leave-one-protein-out", served to two different scopes. The former yielded intrinsically more accurate models because some data points of each protein were randomly included in the training phase. Therefore, this validation approach provides models which might be more general when tested with data from different groups in the future. On the other hand, the latter validation process provides a means of testing the robustness of our approach by using our dataset only. Further, it provides room to discuss the limitations of the models when the aim is the prediction of protein classes which has not been experienced by the network. Possibly, this validation method would provide better results with larger dataset, where multiple proteins of each class are included. In order to overcome this limitation, we developed a protein formulation database, which could serve to this scope in the future (https://pippi-data.kemi.dtu.dk/).[17]

In recent years, many models used to support several kinds of decision making processes have been constructed as black boxes, meaning that their internal logic is unknown to the user.[35] Therefore, we developed two types of human-friendly surrogate global machine learning algorithms: decision tree and linear regression. The first is considered as a transparency machine learning model with the ability of nonlinear relationship modeling, while the second is the most common linear model applied in science. We evaluated also other surrogate algorithms (e.g. support vector machines) achieving similar or slightly better results. Nevertheless, we selected the aforementioned algorithms as their scope is to support the connection between the features and the output of the ANNs algorithm. Thanks to these two "white box"

algorithms we build trust in the hidden relationships of the ANNs. In fact, we could confirm that accelerated stability studies with SEC-MALS analysis are the key to predict the monomer retention after long term storage at refrigerated and elevated temperatures. The linear regression shows indeed a simple linear relationship at low temperature which gets lost with increasing temperature. As therapeutic proteins are usually stored at refrigerated conditions the linear model provides a rather accurate algorithm to differentiate samples with high and medium stability. Similarly, the tree decision could simply spot the most important parameters for the prediction of real-time stability. Further, the rules from such decision trees could be used to individuate cut-off values to generate Stability Risk Scores.[17] On the other hand, ANNs are the most accurate approach and manage to provide a more complete characterization of protein thermal stability.

5 Acknowledgment

This study was funded by a project part of the EU Horizon 2020 Research and Innovation program under the Marie Skłodowska-Curie grant agreement No 675074. We thank the whole PIPPI consortium (http://www.pippi.kemi.dtu.dk) for the continuous support offered and Wyatt Technology staff members for their many contributions. Special thanks to Dillen Augustijn and Åsmund Rinnan for the discussion related to the machine algorithms. We thank Hristo Svilenov, Andreas Tosstorff, Marcello Morales and Maria Laura Greco for formulating the samples presented in this work. We thank Hristo Svilenov, Andreas Tosstorff, Sujata Mahapatra , Alina Kulakova, Christin Pohl, for collecting and processing DLS, nanoDSF and ICD data.

6 References

1. D. S. Dimitrov, Therapeutic proteins, Methods Mol. Biol. 899 (2012) 1-26.

2. J. G. Elvin, R. G. Couston, C. F. Van Der Walle, Therapeutic antibodies: Market considerations, disease targets and bioprocessing, Int. J. Pharm. 440 (2013) 83–98.

3. Y. Lu, S.E. Harding, T.E. Michaelsen, E. Longman, K.G. Davis, A. Ortega, J.G. Grossmann, I. Sandlie, J. García de la Torre, Solution conformation of wild-type and mutant IgG3 and IgG4 immunoglobulins using crystallohydrodynamics: Possible implications for complement activation, Biophys. J. 93 (2007) 3733-44.

4. F. Franks, Bioproducts : Attractions and Pitfalls, Trends Biotechnol. , 12 (1994) 114–117.

5. M. S. Tydeman, T. B. L. Kirkwood, Design and analysis of accelerated degradation tests for the stability of biological standards I. Properties of maximum likelihood estimators, J. Biol. Stand., 12 (1984) 195–206.

6. W. Wang, Protein aggregation and its inhibition in biopharmaceutics, Int. J. Pharm. 289 (2005) 1–30.

7. S. Hermeling, D. J. A. Crommelin, H. Schellekens, W. Jiskoot, Structure-immunogenicity relationships of therapeutic proteins, Pharm. Res., 21 (2004) 897-903.

8. A. Braun, L. Kwee, M. A. Labow, J. Alsenz, Protein aggregates seem to play a key role among the parameters influencing the antigenicity of interferon alpha (IFN-α) in normal and transgenic mice, Pharm. Res. 14 (1997) 1472-8.

9. A. M. Fernandez-Escamilla, F. Rousseau, J. Schymkowitz, L. Serrano, Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins, Nat. Biotechnol. 22 (2004) 1302-6.

10. G. G. Tartaglia, A. Cavalli, R. Pellarin, A. Caflisch, Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences, Protein Sci. 14 (2005) 2723-34.

11. M. Belli, M. Ramazzotti, F. Chiti, Prediction of amyloid aggregation in vivo, EMBO Rep., 12 (2011) 657-63.

12. N. J. Agrawal, S. Kumar, X. Wang, B . Helk, S.K. Singh, B. L. Trout, Aggregation in protein-based biotherapeutics: Computational studies and tools to identify aggregation-prone regions, J. Pharm. Sci. 100 (2011) 5081-95.

13. O. Obrezanova, A. Arnell, R. G. de la Cuesta, M. E. Berthelot, T. R. Gallagher, J. Zurdo, Y. Stallwood, Aggregation risk prediction for antibodies and its application to biotherapeutic development, MAbs, 7 (2015) 352–363.

14. X. Wang, T. K. Das, S. K. Singh, S. Kumar, Potential aggregation prone regions in biotherapeutics: A survey of commercial monoclonal antibodies. MAbs, 1 (2009) 254–267.

15. T. M. Lauer, N. J. Agrawal, N. Chennamsetty, K. Egodage, B. Helk, B.L. Trout. Developability index: A rapid in silico tool for the screening of antibody aggregation propensity, J. Pharm. Sci., 101 (2012) 102-15.

16. N. Chennamsetty, B. Helk, V. Voynov, V. Kayser, B. L. Trout, Aggregation-Prone Motifs in Human Immunoglobulin G, J. Mol. Biol. 14 (2009) 404-13.

17. L. Gentiluomo, et al. Advancing therapeutic protein discovery and development through comprehensive computational and biophysical characterization. Mol. Pharmaceutics (2019) Just Accepted Manuscript.

18. E. Sahin, C. J. Roberts, Size-exclusion chromatography with multi-angle light scattering for elucidating protein aggregation mechanisms, In Therapeutic proteins (2012) 403-423, Humana Press, Totowa, NJ.

19. H. Ye, Simultaneous determination of protein aggregation, degradation, and absolute molecular weight by size exclusion chromatography-multiangle laser light scattering, Anal. Biochem, 356 (2006) 76-85.

20. Y. Li, B. A. Ogunnaike, C. J. Roberts, Multi-variate approach to global protein aggregation behavior and kinetics: Effects of pH, NaCl, and temperature for α-chymotrypsinogen A, J. Pharm. Sci., 99 (2010) 645-662.

21. R. K. Brummitt, D. P. Nesta, L. Chang, A. M. Kroetsch, C. J. Roberts, Nonnative aggregation of an IgG1 antibody in acidic conditions, part 2: Nucleation and growth kinetics with competing growth mechanisms, J. Pharm. Sci. 100 (2011) 2104-2119.

22. Y. Li, C. J. Roberts, Lumry-eyring nucleated-polymerization model of protein aggregation kinetics. 2. Competing growth via condensation and chain polymerization, J. Phys. Chem. B, 113 (2009) 7020-32.

23. E. Sahin, A. O. Grillo, M. D. Perkins, C. J. Roberts, Comparative effects of pH and ionic strength on protein-protein interactions, unfolding, and aggregation for IgG1 antibodies. J. Pharm. Sci. 99 (2010) 4830-48.

24. L. Gentiluomo, D. Roessner, D. Augustijn, H. Svilenov, A. Kulakova, S. Mahapatra, G. Winter, W. Streicher, Å . Rinnan, G. H. J. Peters, P. Harris, W. Frieß, Application of interpretable artificial neural networks to early monoclonal antibodies development, Eur. J. Pharm. Biopharm. 141 (2019) 81–89.

25. A. S. Hussain, X. Yu, R. D. Johnson, Application of Neural Computing in Pharmaceutical Product Development. Pharm. Res., 8 (1991) 1248-1252.

26. K. Takayama, M. Fujikawa, T. Nagai, Artificial neural network as a novel method to optimize pharmaceutical formulations, Pharm. Res., 16 (1999) 1-6.

27. A. S. Achanta, J. G. Kowalski, C. T. Rhodes, Artificial neural networks: Implications for pharmaceutical sciences, Drug Dev. Ind. Pharm. 21 (1995) 119-55.

28. A. C. King, M. Woods, W. Liu, Z. Lu, D. Gill, M. R. Krebs, High-throughput measurement, correlation analysis, and machine-learning predictions for pH and thermal stabilities of Pfizer-generated antibodies, Protein Sci. 20, (2011) 1546–57.

29. Y. Yang, Z. Ye, Y. Su, Q. Zhao, X. Li, D. Ouyang. Deep learning for in vitro prediction of pharmaceutical formulations, Acta Pharm. Sin. B, 9 (2019) 177–185.

30. Z. Ye, Y. Yang, X. Li, D. Cao, D. Ouyang, An Integrated Transfer Learning and Multitask Learning Approach for Pharmacokinetic Parameter Prediction, Mol. Pharm., 16 (2019) 533–541.

31. J. Bourquin, H. Schmidli, P. Van Hoogevest, H. Leuenberge, Application of artificial neural networks (ANN) in the development of solid dosage forms, 2 (1997) 111-21.

32. A. P. Plumb, R. C. Rowe, P. York, C. Doherty, The effect of experimental design on the modeling of a tablet coating formulation using artificial neural networks, Eur. J. Pharm. Sci. 16 (2002) 281-288.

33. B. Aksu, A. Paradkar, M. de Matas, O. Ozer, T. Güneri, P. York. Quality by Design Approach: Application of Artificial Intelligence Techniques of Tablets Manufactured by Direct Compression, AAPS PharmSciTech, 13 (2012) 1138-46.

34. M. Craven, S. W. Jude, Extracting tree-structured representations of trained networks, (1996) 24:30

35. R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A Survey Of Methods For Explaining Black Box Models, ACM computing surveys 51 (2018) 93.

36. J. K. Myers, C. N. Pace, J. M. Scholtz, Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding, Protein Sci., 4 (1995) 2138–48.

37. L. Wafer, M. Kloczewiak, S. M. Polleck, Y. Luo, Isothermal chemical denaturation of large proteins: Path-dependence and irreversibility, Anal. Biochem. 539 (2017) 60:69.

38. W. Y. Yang, W. Cao, T. S. Chung, J. Morris, Applied Numerical Methods Using MATLAB®. Applied Numerical Methods Using MATLAB® (2005), John Wiley & Sons.

39. R. P. Lippmann, An Introduction to Computing with Neural Nets, *IEEE ASSP* 4 (1987) 4:22.

40. J. J. More, The Levenberg-Marquard Algorithm; Implementation and Theory. Numerical analysis, (1978) 105-116, Springer, Berlin, Heidelberg.

41. D. J. C. MacKay, Baysian Interpolation, Neural interpolation, 447 (1992) 415–447.

42. F. D. Foresee, M. T. Hagan, Gauss-Newton approximation to Bayesian learning, Proceedings of International Conference on Neural Networks, 3 (1997) 1930-1935.

43. J. R. Quinlan, Induction of Decision Trees. Mach. Learn. (1986) 81-106.

44. B. H. Zimm, The scattering of light and the radial distribution function of high polymer solutions, J. Chem. Phys., 16 (1948) 1093–99.

45. H. Svilenov, L. Gentiluomo, W. Friess, D. Roessner, G. Winter, A New Approach to Study the Physical Stability of Monoclonal Antibody Formulations—Dilution From a Denaturant, J. Pharm. Sci., 107 (2018) 3007–13.

46. D.W. Patterson, Artificial Neural Networks: Theory and Applications, Prentice Hall Asia, 1998.

47. D. J. Livingstone, D. T. Manallack, I. V. Tetko, Data modelling with neural networks: Advantages and limitations, J. Comput. Aided. Mol. Des., 11 (1997) 135-142.

48. D. J. Livingstone, D. T. Manallack, Statistics Using Neural Networks: Chance Effects, Journal of Medicinal Chemistry, 36 (1993) 1295-97.

49. D. T. Manallack, D. D. Ellis, D. J. Livingstone, Analysis of Linear and Nonlinear QSAR Data Using Neural Networks, J. Med. Chem. 37 (1994) 3758-67.

50. D. T. Manallack, D. J. Livingstone, Artificial neural networks: application and chance effects for QSAR data analysis, Med. Chem. Res. 2 (1992) 181-90.

51. L. Gentiluomo, D. Roessner, W. Streicher, S. Mahapatra, P. Harris, W. Frieß, (2019). Characterization of native reversible self-association of a monoclonal antibody mediated by Fab-Fab interaction. J. Pharm. Sci. (2019) In Press.

8 Supplementary information

List of supplementary information

SI 1. Purity data (i.e. cIEF and SEC) of the protein bulk

SI 2. SEC-MALS exemplary chromatogram.

Table SI 1 Purity data (i.e. cIEF and SEC) of the protein bulk

Method	cIEF	SEC
Protein	Isoelectric point Ip	Monomer mass fraction (%)
PPI-01	7.2	99.7
PPI-02	9.2 – 9.3	98.3
PPI-03	9.3 – 9.4	99.8
PPI-04	8.8 – 9.0	99.1
PPI-08	9.0 – 9.2	99.7
PPI-10	8.9 – 9.2	96.3
PPI-13	8.8 – 8.9	99.4
PPI-17	8.9 – 9.1	98.5
PPI-18	5.2 – 5.6	98.3
PPI-30	6.2	100
PPI-44	5.2 – 5.5	85.1
PPI-45	4.7 – 4.8	100
PPI-46	4.3 – 4.4	100
PPI-49	4.7 – 4.9	98.1

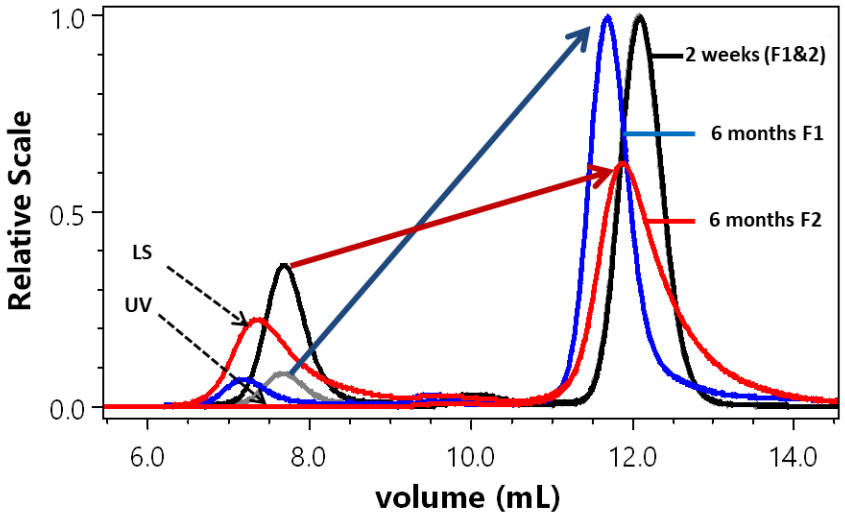


Figure SI 2. SEC-MALS exemplary chromatogram. The SEC-MALS chromatogram of two samples, stored for two weeks at 40°C, are colored in black and gray, representing respectively PPI-1 formulated in 10 mM Histidine at pH 5 (F1) and with 280 mM Sucrose as excipient (F2). Similarly, F1 and F2 after 6 months of storage at 25°C are colored, respectively, in blue and red. The light scattering (LS) and UV signal are superimposed. The higher molecular weight species are visible only by LS. F1 presents a lower amount of HWM compared to F2 after two weeks of stress, which correlated with a lower monomer retention after 6 months of stress.

CHAPTER VI: Coupling multi-angle light scattering to ultra-high-pressure reverse-phase chromatography (UPLC-RP-MALS) for monoclonal antibodies characterization

Lorenzo Gentiluomo[1,2,3,]*, Vanessa Schneider,[1] Dierk Roessner[1], Wolfgang Frieß[2]

[1] Wyatt Technology Europe GmbH, Hochstrasse 18, 56307 Dernbach, Germany

[2] Department of Pharmacy: Pharmaceutical Technology and Biopharmaceutics; Ludwig-Maximilians-Universitaet Muenchen, Butenandtstrasse 5, 81377 Munich, Germany

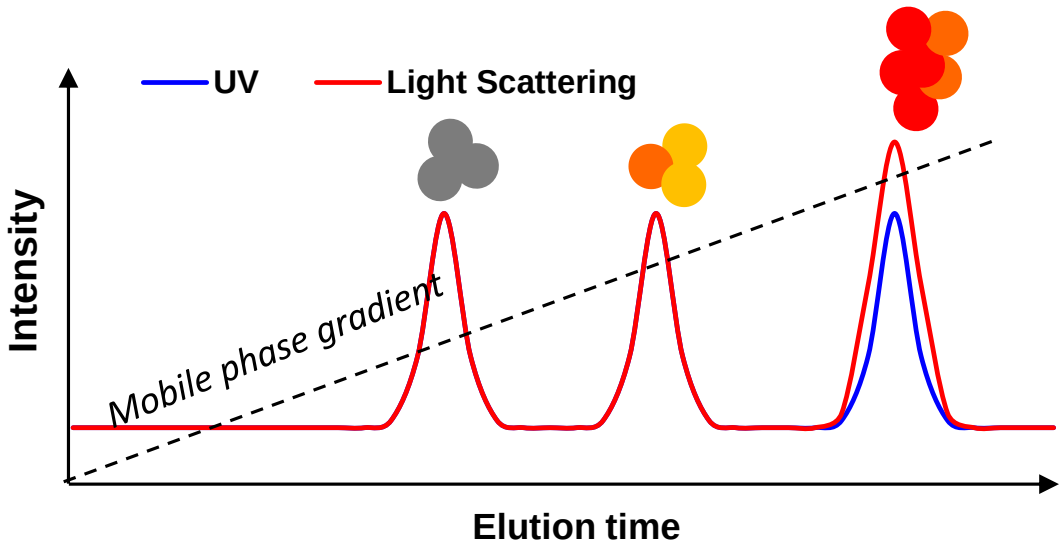
[3]Present address: Coriolis Pharma, Fraunhoferstraße 18B, 82152, Planegg, Germany

* denotes corresponding author.

CONTRIBUTION: **Lorenzo Gentiluomo**, Wolfgang Frieß and Dierk Roessner conceived and designed the study. **Lorenzo Gentiluomo** conducted and analyzed the SEC-MALS experiments. Vanessa Schneider conducted and analyzed with the supervision of **Lorenzo Gentiluomo** the RP-MALS measurements. **Lorenzo Gentiluomo** performed the calculations. **Lorenzo Gentiluomo** and Wolfgang Frieß contributed to the writing of the manuscript.

Abstract

Multi-angle light scattering coupled with size-exclusion chromatography (SEC-MALS) is a standard approach for protein characterization. Recently MALS detection has been coupled with ion-exchange chromatography (IEX) which demonstrated the feasibility and high value of MALS in combination with non-sized-based fractionation methods. In this study we coupled reverse-phase ultra-high pressure liquid chromatography (RP-UPLC) with a low-dispersion MALS detector for the characterization of intact monoclonal antibody (mAbs) and their fragments. We confirmed a constant refractive index increment value for mAbs in RP gradients, in good agreement with the values in literature for other classes of proteins. We showed that the impurities eluting from a RP column can often be related to aggregated species and we confirmed that in most cases those oligomers are present also in SEC-MALS. Yet, in few cases small aggregates fractions in RP-UPLC are an artifact. In fact, proteins presenting thermal and physical stability not suitable for the harsh condition applied during the RP separation of mAbs (i.e. organic solvents at high temperature) can aggregate. Further, we applied RP-UPLC-MALS during a long term stability studies. The different principle of separation used in RP-UPLC- MALS provides an additional critical level of protein characterization compared to SEC-MALS and IEX-MALS.



Graphical abstract. *Coupling multi-Angle light scattering to ultra-high-pressure reverse-phase chromatography (UPLC-RP-MALS) for monoclonal antibodies characterization*

1 Introduction

Light scattering is one of the widely-used techniques for the characterization of macromolecules and particles in solution in biological and biopharmaceutical sciences[1]. By far the most common application of light scattering in this field is the determination of mass and size of proteins by means of multi-angle light scattering coupled to size-exclusion chromatography (SEC-MALS)[2] or field flow fractionation (FFF-MALS)[3]. Other important applications include the characterization of protein conformational and colloidal stability and the characterization of both specific and non-specific protein-protein interaction[1]. The use of MALS with fractionated samples yields a calculation of the absolute molecular weight (Mw) at each point of the chromatogram. As the Mw estimated by the retention time is often inaccurate,[4,5] SEC-MALS provides a useful tool for determination of accurate monomer and fragment Mw, oligomeric state and hydrodynamic radius (R_h)[1,2,6]. Recently the advantages of coupling MALS with ion exchange chromatography (IEX) have been demonstrated[7]. IEX separates proteins according to surface charge based on differences in ionic interaction with the support matrix[8]. The different principle used in the separation of IEX-MALS provides additional critical information and can resolve SEC-MALS shortcomings[7]. In this study, we coupled MALS with another type of liquid chromatography, reversed-phase (RPLC). RPLC is a highly promising technique to study chemical changes[9-11] and to quantify[12,13] peptides and proteins, including monoclonal antibodies (mAbs). Historically, the use of RP to monitor intact mAb was limited because the complex hydrophobic and hydrophilic nature of these large proteins caused poor recovery and limited resolution. More recently, the use of columns with large pores (300 Å) at high temperatures (60–75 °C) in combination with non-traditional solvent system containing ion pairing agents has been consolidated as standard procedure for the analysis of mAbs, overcoming previous difficulties[14,15]. Small chemical differences cannot be separated by standard RP-HPLC[16], as they are often insufficient to yield significant changes in polarity[17]. Here, we took advantage of ultra-high pressure LC (UPLC) instrumentation to further refine the separation of mAb species and their derivatives. We investigated RP-UPLP-MALS for mAb characterization, focusing on two common applications: (i) analysis and characterization of mAb fragments, which are typically studied by mass spectrometry, (ii) analysis of mAbs after long term storage. The former is a real-time stability testing which permits the establishment of recommended storage condition and shelf life of the bio-therapeutic products. The addition of MALS allows the Mw assignment for each individual peak in the chromatogram enabling differentiation between chemical variants of the monomeric form and other impurities or degradation products as aggregates and fragments.

2 Result and discussion

2.1 RP-MALS technique.

The principle of RP-HPLC-MALS is the combination of RP chromatography with an online MALS detector. As shown in **Fig. 1**, multiple hydrophobic areas of protein molecules interact with the alkyl silane-derived surface of the stationary phase.

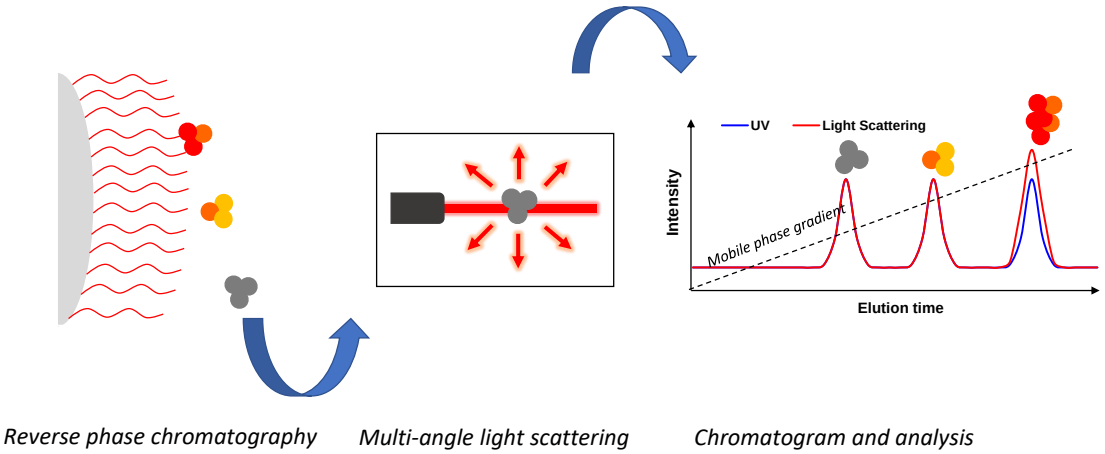


Figure 1. ***Schematic illustration of the RP-UPLC-MALS method.*** *A protein sample is injected in the RP chromatography column in-line with a MALS detector. The protein interacts with the hydrophobic matrix.*

The separation is achieved by decreasing the water concentration in the mobile phase increasing the organic solvent fraction (e.g. acetonitrile). This in turn weakens the hydrophobic attraction of the protein to the column. During elution from the column the molecules are then introduced into a concentration detector (i.e. UV) and subsequently in a MALS detector. Using these detectors to measure the Mw of eluting molecules is especially important as no column calibration procedure, analogous to that of analytical SEC, can be applied to relate the size of a molecule to its hydrophobic interaction with a column matrix.

2.2 Development of UPLC-RP-MALS

Good RP-HPLC conditions for intact protein analysis are typically achieved with a UPLC, a stationary phase with short alkyl chain length and large pore size, a strong ion-pairing agent and an adequate gradient decreasing the water content of the mobile phase at high temperature[9]. We coupled a low-volume, low-dispersion MALS detector to our UPLC system allowing for small peak width and high resolution. Six different IgG1s (PPI01, PPI02, PPI03, PPI04, PPI10, PPI13), one IgG2 (PPI17), one bispecific (PPI08), and one protein-drug conjugate (PPI18) were used to develop and assess our RP-

UPLC-MALS method. IFNα2a served as a reference, as RPLC is a well-established technique to detect its chemically-changed species[18-21]. During the development of the RP-UPLC-MALS method column type, temperature, flow rate, injection volume, mobile phase and gradient were evaluated[14,15]. Some proteins presented better resolution with the BEH-300 C4 column compared with the Zorbax 300SB-C8 column. However, we noticed a fast decrease of efficiency with the BEH-300 C4 after just 400 injections, while the Zorbax 300SB-C8 showed good robustness. This is possibly due to the fact that the C4 phase chemistry is less resistant to hydrolysis in acidic media than the C8 phase chemistry. As screens of proteins in multiple formulations and across many time points involve thousands of injections, we selected the Zorbax 300SB-C8 as workhorse.

In order to determine Mw correctly, it is necessary to know the refractive index increment of solute in solution value dn/dc and the concentration for each slice of a peak. It has been shown that MALS is compatible with RP elution gradients[22]. Different classes of proteins have been investigated in literature with various mobile phase compositions containing aqueous buffer and acetonitrile yielding a dn/dc values close to 0.175 ml/g[22-25]. It has been shown that assuming a constant dn/dc in the narrow interval of an eluting peak only induces an error at most 3–4%[22]. This is due to the fact that the solvent refractive index changes only very slightly within the time frame of peak elution[26]. We first calculated the protein Mw using the dn/dc of proteins in water at 660 nm of 0.185 mL/g[27]. The obtained Mw was approx. 25% below the Mw calculated based on the primary sequence. Consequently, we fixed the Mw of the monomer as calculated from the primary sequence and confirmed by SEC-MALS to obtain a dn/dc in the RP-MALS eluent. This yielded a dn/dc value of 0.1742 +/− 0.0017 mL/g for the proteins, which is in very good agreement with the literature[22-25], and was used for calculating the Mw of the investigated proteins.

2.3 Analysis of intact monoclonal antibodies using UPLC-RP-MALS.

Proteins with similar size cannot be separated by SEC, but if they have a different hydrophobicity they can be separated by RP-UPLC. In our study we encountered three cases: (i) The Mw of all peaks reflects monomeric variants (e.g. PPI01 and PPI10), (ii) The main peak represents a monomeric form while other impurity peaks are identified as aggregates (e.g. PPI04), (iii) The main peak represents a monomeric form while other impurities peaks are either identified as aggregates, fragments or close to, but not equal within the experimental error, to the monomer Mw (e.g. PPI02) (**Fig. 2**). Dimers detected in SEC-MALS (**Fig. 3**) were not found in RP-UPLC-MALS (**Fig. 2**).

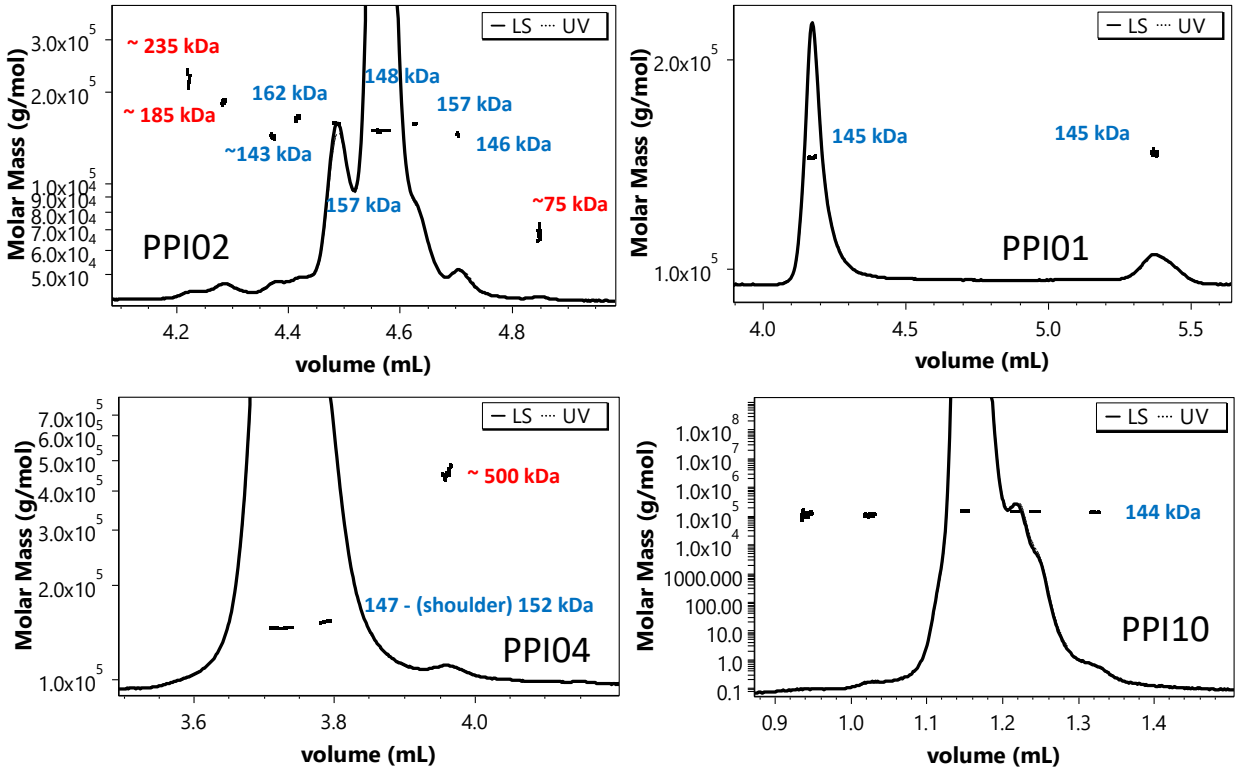


Figure 2. RP-UPLC-MALS of mAbs. *Typical chromatograms showing the UV and the MALS signals of PPI02, PPI01, PPI04 and PPI10 analyzed by RP-UPLC-MALS. The MW of the monomer, aggregates/fragments, and dimers are highlighted in blue and respectively. (*) denotes aggregates.*

As the RP-UPLC recovery was often close or exactly 100% (**Table 1**) we hypothesize that (i) the monomer-dimer equilibrium is completely shifted towards the monomeric form in the RPLC eluent, (ii) the dimers are prompted to further aggregation, (iii) the dimers are lost over the column. Both RP-UPLC-MALS and SEC-MALS confirmed the absence of oligomers beyond the dimers visible in SEC for PPI01 and PPI10 (**Fig. 2**). Similar conclusions were reached for PPI13, PPI08 and PPI17 (Supplementary information - **SI 1**). Differently, PPI04 (**Fig. 2**) and PPI18 (**SI 2**) showed a very small fraction of oligomers by RP-UPLC-MALS, which were not detected in SEC-MALS (**Fig. 3**).These oligomers may have been induced by the high temperature of 75 °C applied during the RP separation. The first temperature of unfolding (T_{m1}), the temperature of aggregation (T_{agg}), and the diffusion interaction parameter (k_D) for PPI01, PPI02, PPI03, PP10 and PP17 are 66, 61 °C and 5.6 mg/L (data averaged from 24 formulation conditions, Gentiluomo L, et al.)[28] as compared to 54 °C, 47 °C and 4.7 mg/L resp. for PPI18 and 64 °C, 55 °C and −1.9 mg/L for PPI04. This lower thermal and/or colloidal stability of PPI18 and PP4 could explain their susceptibility to aggregation under the RP conditions. Finally, PPI02 showed aggregates and fragments (highlighted in red in **Fig. 2**) that were also detected in SEC-MALS (**Fig. 3**). The averaged Mw of the PPI02 aggregates from SEC-MALS and RP-UPLC-MALS are respectively of 250 kDa and 235 kDa. This difference is probably due to the high error in the Mw calculations, which is in turn due to the small

concentration of such aggregates. Further, the 235 kDa aggregate in RP-UPLC-MALS is not baseline separated. Moreover, PPI02 presented a series of peaks and shoulders with 5 to 15 kDa difference to the monomer Mw, which were not visible by SEC-MALS. The Mw difference may be possibly due to post-translational modifications of the IgG. These typically include methionine oxidation, asparagine and glutamine deamidation, N-terminal acetylation or cyclization, glycation of lysine and variable glycosylation[29]. Physically, the refractive index increment is insensitive to the long-range structure of macromolecules[27] and is nearly independent on its amino acid composition[30]. However, carbohydrate moieties do affect the refractive index value[31]. This would suggest that PPI02 comes with a high degree of variation in glycosylation.

2.4 Characterization of Fab and Fc fragments.

Complete proteolytic digestion of mAb (peptide mapping) followed by RP-UPLC coupled with mass spectrometry (MS) is a well-established method for the identification and quantification of chemical modification of mAbs[32,33]. Alternatively, the analysis by MALS of large fragments, such as Fab and Fc, requires little sample preparation and can provide a high-throughput alternative. The preparation and purification of the fragments was performed as described in material and methods. Subsequently, we investigated the Fab and Fc fragments of PPI01 by RP-UPLC-MALS. The Fc fragment eluted before the intact mAb which in turn eluted before the Fab fragment (**Fig. 4**). The latter exhibited two shoulders on the left and right of the 47 kDa monomer with a Mw close to that of a Fab dimer (~90 kDa). The Fc fragment elutes with a series of peaks after the main peak of ~110, ~700, ~170 kDa with longer elution time. SEC-MALS measurements on the purified fragments confirmed the presence of Fab dimer and of Fc dimer and trimer (fragments showed in **SI 3**, intact mAb showed in **Fig. 3**). However, the 700 kDa Fc aggregate was not detected in SEC-MALS. As previously mentioned, the formation of small fraction of high molecular-weight oligomers due to the RP conditions can affect proteins with insufficient thermal and/or colloidal stability.

PP01 shows averaged T_{m1}, typically reflecting unfolding of the CH2 domain and T_{m2}, typically reflecting unfolding of the CH3 and Fab fragment, of 64 °C and 77 °C[34-36]. This would explain the higher susceptibility of the Fc fragment to unfolding and aggregation. Thus, it could be useful to couple MALS with RP-UPLC-MS to differentiate between monomer and aggregates peak before analyzing the MS spectra.

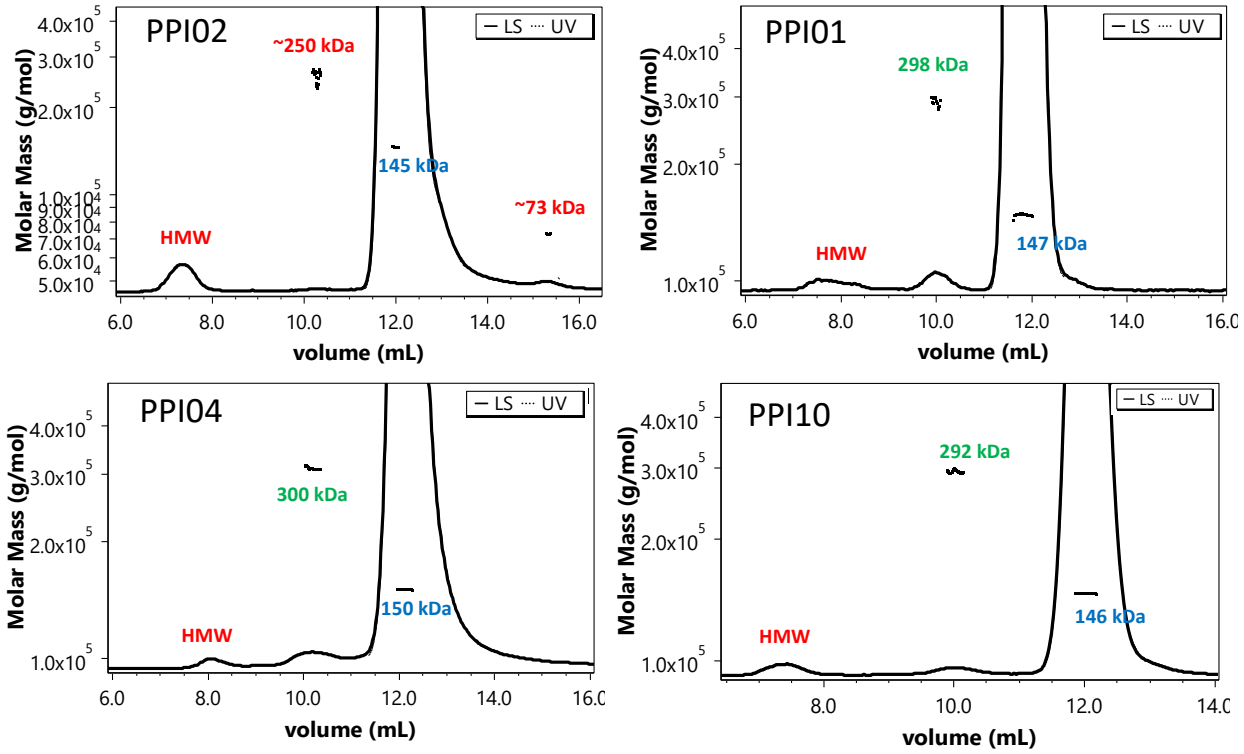


Figure 3. SEC-MALS of mAbs. *Typical chromatograms of the proteins investigated by SEC-MALS showing UV and LS signals along superimposed with calculated molar mass. The MW of the monomer, aggregates/fragments and dimers are highlighted in blue, red and green, respectively. HMW stands for high-molecular weight species, which are usually not separated, and in all our investigated cases presented no UV detectable signal. (*) denotes aggregates; (**) denotes dimers.*

2.5 Long term stability studies.

Finally, we performed a long term stability study and analyzed samples with the RP-UPLC-MALS method developed herein to learn whether we can gain additional insights from the MALS information on the chemical stability of our proteins. PP02, PP03, PP04, PP08, PP10, PP13 were tested in 8 different formulations for six month at 4 °C and 25 °C (see **SI 4** for the formulations list). We observed an overall high chemical stability. Significant changes upon storage stress occurred only in a few conditions. PP10, formulated in 10 mM His at pH 6.5 stored at 25 °C, exhibited an increased hydrophobicity of the shoulder, presenting the same Mw of the monomer (**Fig. 5**). Chemical changes can perturb the local conformation backbone of proteins, such in the case of deamidation, the most common hydrolytic reaction for protein, and Asp isomerization. Conformational variants of proteins often present increased hydrophobicity and are more prone to aggregate[37].

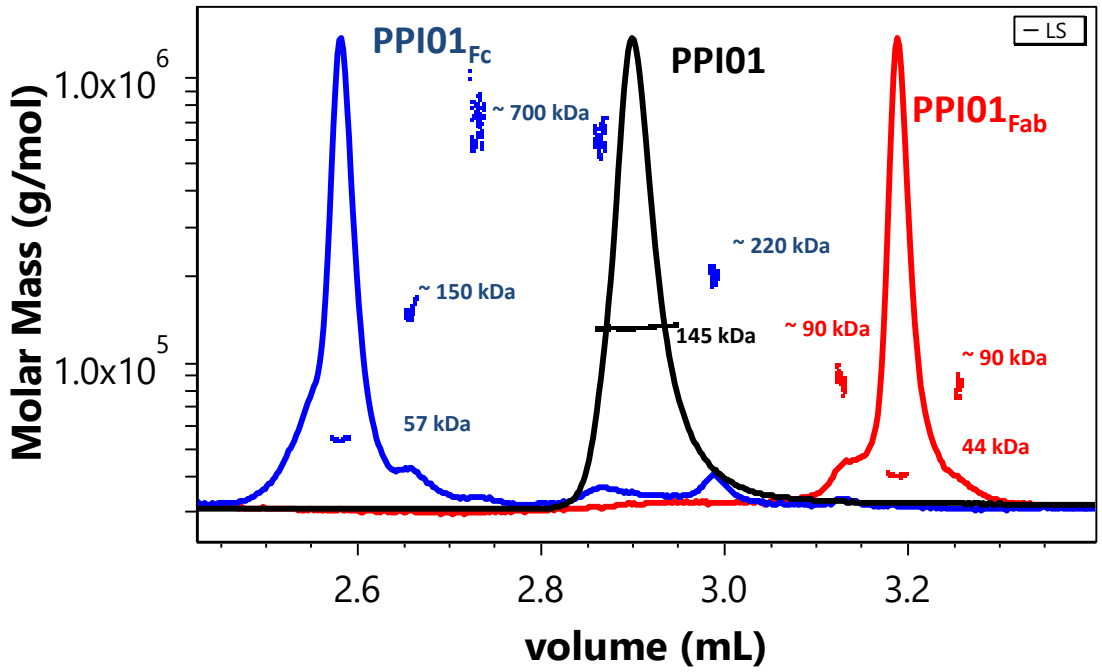


Figure 4. UPLC-RP-MALS of PPI-1 and its fragments. *PPI1 Fc fragment, PPI1 (whole mAb) and PPI1 Fc fragment are plotted in blue, black and red lines, respectively.*

Other chemical reactions, such Met oxidation, could on the other side decrease the hydrophobicity of proteins[38]. However, RP-UPLC-MALS cannot provide mechanistic insight behind an increased hydrophobicity after isothermal stress. For such purpose mass spectroscopy, which could be coupled with RP-UPLC-MALS, could provide quantitation of degradation products, such as in the case of deamidation products[39]. PPI08 stored at 25 °C in 10 mM Histidine at pH 5 showed a new peak with an Mw of 225 kDa, which was not observed in any other formulation and was not noticeable in SEC- MALS (**Fig. 5**). This aggregate is probably made of a mixture of fragments formed during the stress e.g. Fab, Fc, Heavy chain or by a complex formed by monomer and light chain. Comparison with SEC-MALS confirmed the presence of fragments (**Fig. 5**). As baseline separation was not obtained between the monomer and the dimer, we could not tell whether the small complex is present in the formulation or formed during the RP separation. Regardless, MALS provided the exact Mw of the peaks eluting upon RP-UPLC, which allowed differentiation between chemical variants of the monomer (i.e. in cases of PPI10) and aggregates (i.e. in case of PPI08) formed during long term storage.

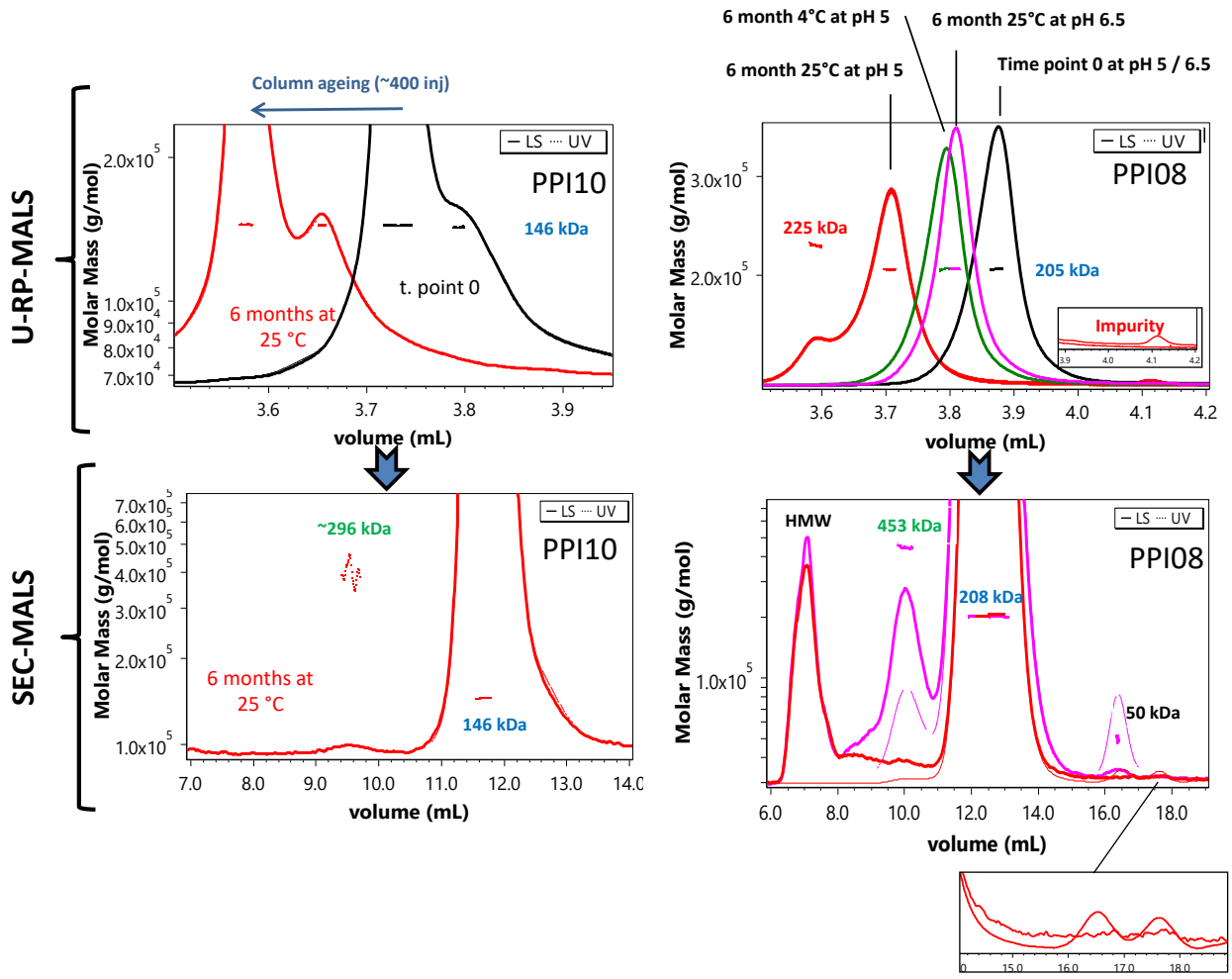


Figure 5. RP-UPLC-MALS and SEC-MALS of mAbs for long term stability studies. *Typical chromatograms of the long-term-storage study, showing the regions of eluting sample. Top: RP-UPLC-MALS chromatograms; bottom: SEC-MALS chromatograms. Left: PPI10; right: PPI08. The Mw of the monomer, aggregates/fragments, and dimers are highlighted in blue, red and green, respectively. HMW stands for high-molecular-weight species, which are usually not separated, and in all our investigated cases presented no detectable UV signal. A magnified section shows the impurities for PP08. The shifting of the chromatograms at different time points is due to column ageing. PPI10 is shown in one formulation only (His 10 mM at pH 6.5), where the chromatograms before and after 6 months at 25 °C are depicted in black and red, respectively. PPI08 is shown formulated at pH 6.5 (His 10 mM) before stress, in black, and after 6 months at 25 °C, in magenta, and formulated at pH 5 (His 10 mM) before, in black, and after 6 months at 4 °C, in green, and 6 months at 25 °C, in red. PPI08 fragments are zoomed.*

3 Conclusion

We successfully coupled RP-UPLC with MALS to calculate the Mw of each eluting peak of intact mAbs and of Fc and Fab fragments. The different principle of separation used in RP-UPLC-MALS provides an additional critical level of protein characterization compared to SEC-MALS and IEX-MALS. RP is one of the most promising analytical techniques to analyze proteins[11,12,40]. Yet, peaks eluting from the column can often be related to aggregated species. Thanks to MALS, it is possible to tell whether an impurity is indeed a chemical variant of the monomer, an aggregate or a fragment. Furthermore, we highlight that the organic solvent and the temperature applied during the RP separation of mAbs could artificially induce aggregates which may lead to false interpretation of protein purity. Nonetheless, MALS could not be enough to describe detailed mechanisms and further coupling with MS (i.e. RP-UPLC-MALS-MS) could prove in the future natural development to characterize RP chromatograms.

4 Material and methods

4.1 Sample preparation.

Five antibodies IgG1s (PPI02, PPI03, PPI04, PPI10, PPI13), one bispecific antibody (PPI08), one IgG2 (PPI17), and one HSA-fusion protein (PPI18) were provided by AstraZeneca (Cambridge, UK). Interferon alpha-2a (PPI30) was provided from Roche Diagnostics GmbH. A summary of the protein's physical properties is listed in **Table 1**. The proteins were dialyzed overnight using Slide-A-Lyzer™ cassettes (Thermo Fisher Scientific, USA) with suitable membrane cut-off against excess of 10 mM of histidine HCl buffer with pH 5.0, 5.5, 6.0, 6.5, 7.0, 7.5. The excipient (e.g. NaCl) stock solutions were prepared in the respective buffers. Protein concentration was measured on a Nanodrop 2000 (Thermo Fisher Scientific, Waltham, USA) using the protein extinction coefficient calculated from the primary sequence. All conditions were prepared in 1.5 mL non-coated PP Eppendorf tubes. Finally, the formulations were sterile-filtered with 0.22 µm cellulose acetate filters from VWR International (Germany). The purity of the proteins was studied by SEC and cEIF (**SI 5**).

4.2 Ultra-high-pressure reverse-phase chromatography combined with multi-angle light scattering (UPLC-RP-MALS)

RP-UPLC-MALS was conducted on an ACQUITY UPLC H-Class system (Waters, USA) equipped with a quaternary pump, an autosampler, UV detector and a µDAWN detector (Wyatt Technology, USA). The separation was performed with both an Acquity BEH-300 C4 (Waters, USA) and a Zorbax 300SB-C8

column (Agilent Technologies, Germany). The samples were diluted to 1 mg/mL before injection. For monoclonal antibodies a pilot gradient of 20 to 40% of eluent B in A over 20 minutes was used. Eluent A consisted of 10% w/v acetonitrile and 0.1% w/v trifluoracetic acid in ultrapure water. Eluent B consisted of 0.1% w/v trifluoracetic acid in acetonitrile. The flow rate was 0.2 mL/min. The column oven temperature was set at 75 °C. A preheater was included before the column. Subsequently, depending on the protein and the column used the gradient was fine-tuned.

Table 1. Information on the investigated protein. *The theoretical Mw is calculated from the primary sequence. Mass recovery is calculated over all the visible UV peaks as described in material and method.*

Type	Provider	Ɛ at 280 nm (mg/ml/cm)	Theoretical M_W (kDa)	MALS M_w (kDa)	Mass recovery	pI	Notes	ID
IgG1λ	AstraZeneca	1.56	144.8	144.1 ± 0.2%	99.9%	7.96	-	PPI01
Human IgG1κ	AstraZeneca	1.47	148.2	148.1 ± 0.1%	100%	8.53	-	PPI02
Human IgG1κ	AstraZeneca	1.435	144.8	144.6 ± 0.2%	100%	8.44	WT IgG	PPI03
IgG1λ YTE	AstraZeneca	1.755	146.2	146.5 ± 0.1%	97.2%	8.99	-	PPI04
IgG1κ + scFv	AstraZeneca	1.57	204.4	204.4 ± 0.1%	98%	9.2	Bispecific	PPI08
Human IgG1	AstraZeneca	1.533	144.2	144.6 ± 0.2%	96.5%	8.95	-	PPI10
Human IgG1κ	AstraZeneca	1.66	148.9	148.7 ± 0.2%	100%	9.04	-	PPI13
IgG2κ	AstraZeneca	1.31	145.1	145.6 ± 0.3%	99.9%	7.78		PPI17
HSA-NEP	AstraZeneca	1.04	146.7	146.3 ± 0.1%	100%	5.8	Conjugate	PPI18
Intα-2A	Roche	0.972	19.2	20.1 ± 7.5%	100%	5.97	-	PPI30

All methods were based on a gradient from 20–25 to 40%. On-column adsorption of the mAbs was evaluated systematically and almost complete mass recovery was reached for all the protein (**Table 1**). All the calculations were performed with ASTRA V7.1 software (Wyatt Technology, USA). Mass recovery is calculated from the injected mass versus the calculated mass from the concentration detector (i.e. UV). Therefore, to achieve an accurate determination of the mass recovery the sample concentration needs to be accurately measured. Thus, the concentration was measured again before injection in real triplicates by a Nanodrop One (Thermo Fisher Scientific, USA). The theoretical extinction coefficients were double-checked re-calculating the values from the RI monomeric peaks during the SEC-MALS experiments. PPI30 (int-2alpha) was used as a standard. Finally, to achieve a flat baseline, we collected and subtracted the blanks by the algorithm included in the ASTRA V7.1 software.

4.3 Size-exclusion chromatography combined with multi angle light scattering (SEC–MALS)

SEC-MALS was conducted on Agilent 1260 Bio-Inert system with a variable wavelength UV detector operated at 280 nm (Thermo Fischer Scientific, USA), followed by a TREOS II detector (Wyatt

Technology, USA) and an Optilab T-rEX (Wyatt Technology, USA). The temperature controlled-autosampler was kept at 4 °C. Separation was performed with a Superdex 200 increased 10/30 GL column. Data was collected and processed using the ASTRA® software V7.2 (Wyatt Technology, USA). The aqueous mobile phase consisted of 38 mM NaH2PO4, 12 mM Na2HPO4, 150 mM NaCl and 200 ppm NaN3 at pH 7.4 dissolved in HPLC-grade water, filtered through Durapore VVPP 0.1 m membrane filters (Millipore, USA). The samples were centrifuged and injected in duplicates of 25 µl.

4.4 Stress assay

0.2 mL of each protein solution was aliquoted at a concentration of 1 mg/mL and filtered in 0.5 mL sterile non-coated PP Eppendorf tubes. The samples were incubated at 4 °C and 25 °C, for 6 months. After storage, the samples were quenched in an ice bath, left at 4 °C and measured within two weeks. Sample concentration was measured after the stress in real triplicates by a Nanodrop One (Thermo Fisher Scientific, USA). Similarly, the pH was measured after the stress showing no changes within the experimental error (i.e. ±0.1).

4.5 Preparation and purification of Fab and Fc fragments

Immobilized Papain (Thermo Fisher Scientific, USA) was used to digest PPI01 into its Fab and Fc fragments. PPI01 at 20 mg/mL was pipetted into 15 mL glass vial, the vial capped with the resin separator provided with the kit to remove all the air-liquid interfaces. The vial was gently rotated by a Sunlab rotator SU1100 for 5 h at 37 °C. An ÄKTA purifier 10 (GE Healthcare, Uppsala, Sweden) equipped with a Pierce Protein A chromatography cartridge (Thermo Fisher Scientific, USA) (column volume, CV = 5 ml) was used to separate Fc (and undigested mAb) from the Fab fragments. The binding buffer was made of 100 mM sodium phosphate with 150 mM NaCl at pH 7.2. The column was equilibrated with 2 CV of binding buffer with a flow of 2 ml/min. Fractions were collected in 15-ml PP tubes using a Frac 920 fraction collector (GE Healthcare, Sweden) capturing any unbound species e.g. Fab. The elution buffer (100 mM sodium phosphate at pH 3) was kept at 100% over 7 CV. The eluting protein was collected in 15-ml PP tubes using the fraction collector, and was immediately neutralized with a 1 M sodium phosphate buffer at pH 8.5. Ultrafiltration was performed using Vivaspin® tubes with a 10 kDa MWCO PES membrane (Sartorius Stedim Biotech, Germany). Success of the purification was monitored by HP-SEC (see 3.4).

5 Acknowledgments

This study was funded by a project part of the EU Horizon 2020 Research and Innovation program under the Marie Skłodowska-Curie grant agreement No 675074. The first author would like to thank the whole PIPPI consortium (http://www.pippi.kemi.dtu.dk) for the continuous support offered. The first author thanks Wyatt Technology staff members for their many contributions, with a special mention to Dan Some for the passionate discussion related to this work and for reviewing the manuscript. The first author would like to thank Hristo Svinelov, Andreas Tosstorff, Marcello Morales and Maria Laura Greco for formulating some of the samples presented in this work.

6 References

1. Minton, A. P. Recent applications of light scattering measurement in the biological and biopharmaceutical sciences. Anal. Biochem. 501, 4–22 (2016).

2. Wyatt, P. J. Light scattering and the absolute characterization of macromolecules. Anal. Chim. Acta 272, 1–40 (1993).

3. Wittgren, B. & Wahlund, K. G. Fast molecular mass and size characterization of polysaccharides using asymmetrical fow feld-fow fractionation-multiangle light scattering. J. Chromatogr. A 760, 205–218 (1997).

4. Sahin, E. & Roberts, C. J. Size-exclusion chromatography with multi-angle light scattering for elucidating protein aggregation mechanisms. Methods Mol. Biol. 899, 403–423 (2012).

5. Goyon, A. et al. Evaluation of size exclusion chromatography columns packed with sub-3 μm particles for the analysis of biopharmaceutical proteins. J. Chromatogr. A 1498, 80–89 (2017).

6. Nettleship, J. E., Brown, J., Groves, M. R. & Geerlof, A. Methods for Protein Characterization by Mass Spectrometry, Termal Shif (TermoFluor) Assay, and Multiangle or Static Light Scattering. Methods Mol. Biol. 426, 299–318 (2008).

7. Amartely, H., Avraham, O., Friedler, A., Livnah, O. & Lebendiker, M. Coupling Multi Angle Light Scattering to Ion Exchange chromatography (IEX-MALS) for protein characterization. Sci. Rep. 8, 1–9 (2018).

8. Fekete, S., Beck, A., Veuthey, J. L. & Guillarme, D. Ion-exchange chromatography for the characterization of biopharmaceuticals. J. Pharm. Biomed. Anal. 113, 43–55 (2015).

9. Staub, A., Guillarme, D., Schappler, J., Veuthey, J. L. & Rudaz, S. Intact protein analysis in the biopharmaceutical feld. J. Pharm. Biomed. Anal. 55, 810–822 (2011).

10. Grotefend, S. et al. Protein quantitation using various modes of high performance liquid chromatography. J. Pharm. Biomed. Anal. 71, 127–138 (2012).

11. Fekete, S., Rudaz, S., Fekete, J. & Guillarme, D. Analysis of recombinant monoclonal antibodies by RPLC: Toward a generic method development approach. J. Pharm. Biomed. Anal. 70, 158–168 (2012).

12. Molnár, I. & Horváth, C. Separation of amino acids and peptides on non-polar stationary phases by high-performance liquid chromatography. J. Chromatogr. A 142, 623–640 (1977).

13. Navas, N. et al. Quantifcation of an intact monoclonal antibody, rituximab, by (RP)HPLC/DAD in compliance with ICH guidelines. Anal. Bioanal. Chem. 405, 9351–9363 (2013).

14. Dillon, T. M., Bondarenko, P. V. & Speed Ricci, M. Development of an analytical reversed-phase high-performance liquid chromatography-electrospray ionization mass spectrometry method for characterization of recombinant antibodies. J. Chromatogr. A 1053, 299–305 (2004).

15. Dillon, T. M. et al. Optimization of a reversed-phase high-performance liquid chromatography/mass spectrometry method for characterizing recombinant antibody heterogeneity and stability. J. Chromatogr. A 1120, 112–120 (2006).

16. Yang, J., Wang, S., Liu, J. & Raghani, A. Determination of tryptophan oxidation of monoclonal antibody by reversed phase high performance liquid chromatography. J. Chromatogr. A 1156, 174–182 (2007).

17. Martínez-Ortega, A. et al. Study and ICH validation of a reverse-phase liquid chromatographic method for the quantifcation of the intact monoclonal antibody cetuximab. J. Pharm. Anal. 6, 117–124 (2016).

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

19. Sharma, V. K. & Kalonia, D. S. Polyethylene glycol-induced precipitation of interferon alpha-2a followed by vacuum drying: Development of a novel process for obtaining a dry, stable powder. AAPS PharmSci 6 (2004).

20. Mohl, S. & Winter, G. Continuous release of rh-interferon α-2a from triglyceride implants: Storage stability of the dosage forms. Pharm. Dev. Technol. 11, 103–110 (2006).

21. Hermeling, S. et al. Structural characterization and immunogenicity in wild-type and immune tolerant mice of degraded recombinant human interferon alpha2b. Pharm. Res. 22, 1997–2002 (2005).

22. Astafeva, I. V., Eberlein, G. A. & Wang, Y. J. Absolute on-line molecular mass analysis of basic fbroblast growth factor and its multimers by reversed-phase liquid chromatography with multi-angle laser Light scattering detection. J. Chromatogr. A 740, 215–229 (1996).

23. Mhatre, R., Krull, I. S. & Stuting, H. H. Determination of biopolymer (protein) molecular weights by gradient elution, reversedphase high-performance liquid chromatography with low-angle laser light scattering detection. J. Chromatogr. A 502, 21–46 (1990).

24. Mhatre, R. & Krull, I. S. Determination of On-Line Differential Refractive Index and Molecular Weight via Gradient HPLC Interfaced with Low-Angle Laser Light Scattering, Ultraviolet, and Refractive Index Detection. Anal. Chem. 65, 283–286 (1993).

25. Stuting, H. H. & Krull, I. S. Determination of pituitary and recombinant human growth hormone molecular weights by modern high-performance liquid chromatography with low angle laser light scattering detection. J. Chromatogr. A 539, 91–109 (1991

26. An, N., Zhuang, B., Li, M., Lu, Y. & Wang, Z. G. Combined Teoretical and Experimental Study of Refractive Indices of WaterAcetonitrile-Salt Systems. J. Phys. Chem. B 119, 10701–10709 (2015).

27. Zhao, H., Brown, P. H. & Schuck, P. On the distribution of protein refractive index increments. Biophys. J. 100, 2309–2317 (2011).

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

29. Walsh, C. T., Garneau-Tsodikova, S. & Gatto, G. J. Protein posttranslational modifications: The chemistry of proteome diversifcations. Angew. Chem. Int. Ed. Engl. 44, 7342–7372 (2005).

30. Wen, J., Arakawa, T. & Philo, J. S. Size-exclusion chromatography with on-line light-scattering, absorbance, and refractive index detectors for studying proteins and their interactions. Anal. Biochem. 240, 155–166 (1996).

31. Arakawa, T. & Wen, J. Determination of carbohydrate contents from excess light scattering. Anal. Biochem. 299, 158–161 (2001).

32. Lundell, N. & Schreitmüller, T. Sample preparation for peptide mapping - A pharmaceutical quality-control perspective. Anal. Biochem. 266, 31–47 (1999).

33. Williams, K. R. & Stone, K. L. Identifying Sites of Posttranslational Modifcations in Proteins Via HPLC Peptide Mapping. Methods Mol Biol. 40, 157–175 (1995).

34. Menzen, T. & Friess, W. High-throughput melting-temperature analysis of a monoclonal antibody by differential scanning fuorimetry in the presence of surfactants. J. Pharm. Sci. 102, 415–428 (2013).

35. Ionescu, R. M., Vlasak, J., Price, C. & Kirchmeier, M. Contribution of variable domains to the stability of humanized IgG1 monoclonal antibodies. J. Pharm. Sci. 97, 1414–1426 (2008).

36. Garber, E. & Demarest, S. J. A broad range of Fab stabilities within a host of therapeutic IgGs. Biochem. Biophys. Res. Commun. 355, 751–757 (2007)..

37. Wang, W. & Roberts, C. J. Aggregation of therapeutic protein. John Wiley & Sons (2010).

38. Uversky, V. N., Yamin, G., Souillac, P. O., Goers, J. & Glaser, C. B. Methionine oxidation inhibits fibrillation of human alphasynuclein in vitro. FEBS Lett. 517, 239–44 (2002).

39. Barnes, C. A., Lim, A. Applications of mass spectrometry for the structural characterization of recombinant protein pharmaceuticals. Mass Spectrom Rev. 26, 370–388.

40. Molnar, I., Boysen, R. I. & Erdmann, V. A. High-performance-liquid chromatography of Thermus aquaticus 50S and 30S ribosomal proteins. Chromatographia 28, 39–44 (1989).

8 Supplementary information

List of supplementary information:

SI 1. PPI-8, PPI-13 & PPI-17 UPLC-RPLC-MALS Chromatograms

SI 2. PPI-18 UPLC-RPLC-MALS Chromatogram

SI 3. Fab & Fc SEC-MALS Chromatograms

SI 4. Formulations list for long term stability studies

SI 5. Purity data (i.e. cIEF and SEC) of the protein bulk

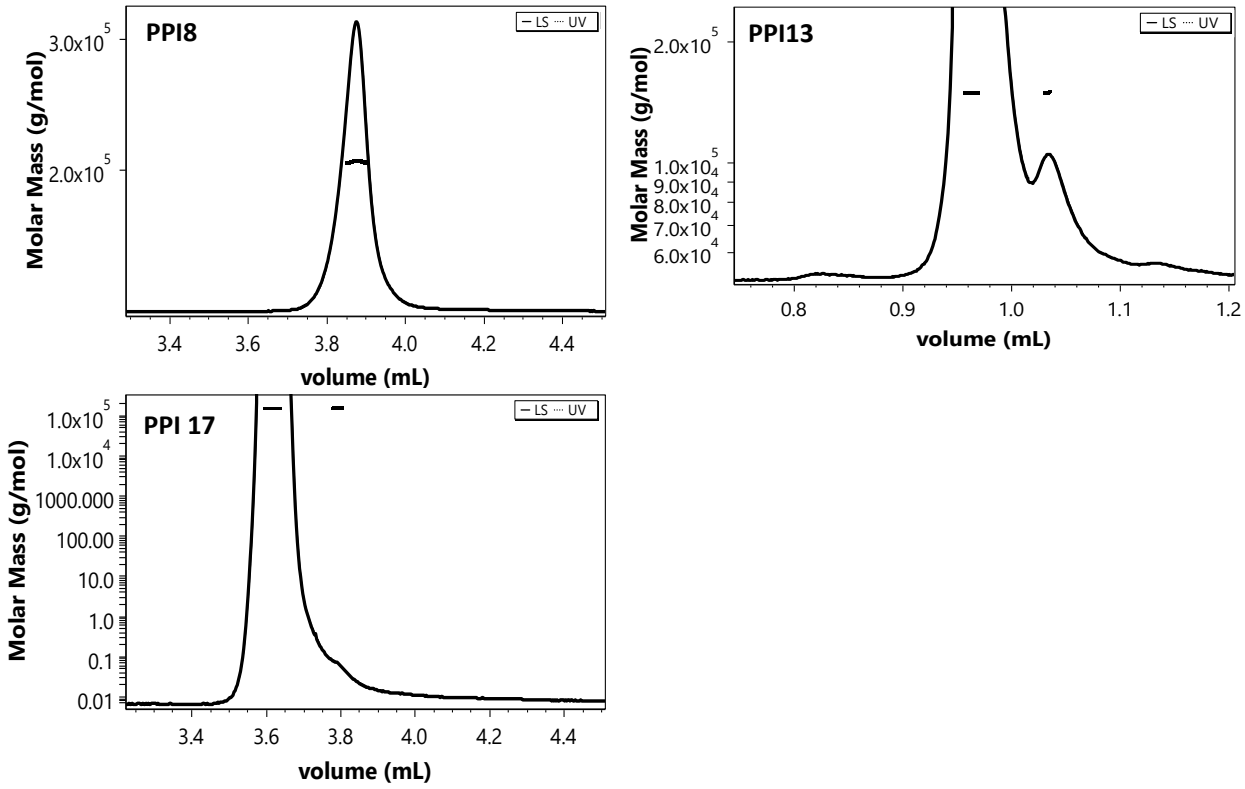


Figure SI 1. UPLC-RPLC-MALS of PPI-8, PPI-13 & PPI-17. *MALS results confirmed the absence of oligomers for PPI-8, PPI-13 and PPI-17.*

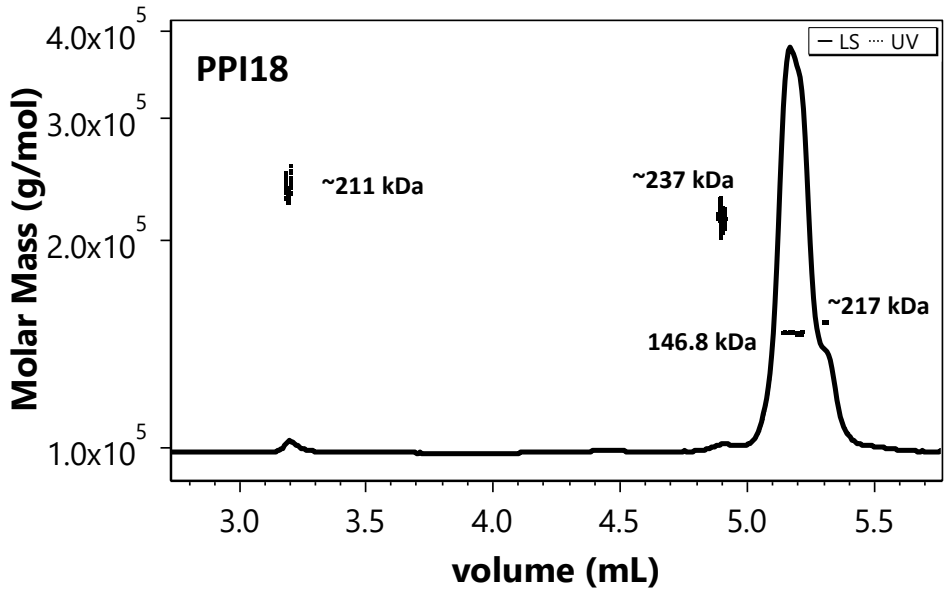


Figure SI 2. UPLC-RPLC-MALS of PPI-18. *MALS results confirmed the presence of oligomers for PPI18.*

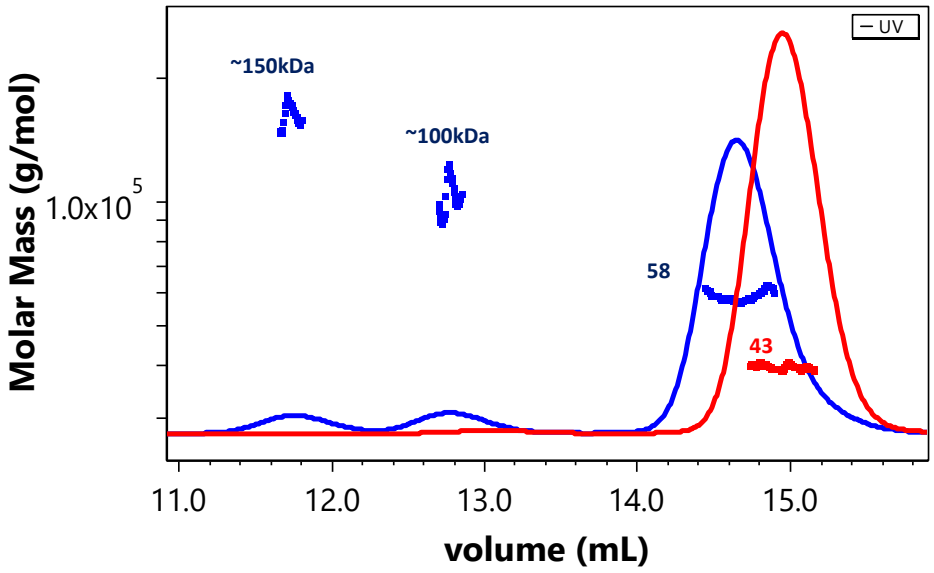


Figure SI 3. SEC-MALS of PPI-1 Fc and Fab fragments. *PPI1 Fc fragment and PPI1 Fc fragment are plotted in blue, and red lines, respectively.*

Table SI 4. Long-term formulations list. *Formulations list for long term stability studies.*

Formulation	Buffer	pH	Excipient	Salt
A	10 mM Histidine	5	-	-
B	10 mM Acetate	5	-	-
C	10 mM Histidine	6.5	280 mM Proline	-
D	10 mM Histidine	6.5	140 mM ArgHCl	-
E	10 mM Histidine	6.5	280 mM Sucroce	-
F	10 mM Histidine	6.5	-	-
G	10 mM Histidine	6.5	-	140 mM NaCl
H	10 mM Phosphate	6.5	-	-

Table SI 5 Purity data. *cIEF and SEC results of the protein bulk (i.e. before formulating).*

Method	cIEF	SEC
Protein	Isoelectric point Ip	Monomer mass fraction (%)
PPI-01	7.2	99.7
PPI-02	9.2 – 9.3	98.3
PPI-03	9.3 – 9.4	99.8
PPI-04	8.8 – 9.0	99.1
PPI-08	9.0 – 9.2	99.7
PPI-10	8.9 – 9.2	96.3
PPI-13	8.8 – 8.9	99.4
PPI-17	8.9 – 9.1	98.5
PPI-18	5.2 – 5.6	98.3
PPI-30	6.2	100

SUMMARY

The control of protein stability is an important component to ensure safety and efficacy of protein drugs. In **Chapter II** we argue that to understand what exactly indicates intrinsic stability of a protein molecule requires more information on various therapeutically-relevant proteins, including their primary sequences, purity data, and computational and biophysical characterization in different solution conditions. Therefore, we made publicly available a comprehensive dataset, which leads the foundation for a protein formulation database. Moreover, we could demonstrate that the relationships between some biophysical parameters cannot be generalized for a heterogeneous population of proteins in a diverse set of formulation conditions (**Chapter II**). These findings highlight that "protein stability" cannot be well described by using a single biophysical parameter, nor by studying a protein in a single solution condition. Therefore, design of experiments (DoE) approaches and response surface methodology (RMS) remain pivotal for the optimization of robust protein formulations. We presented a simple parallel approach which combine multiple parameters into stability risk scores (SRS).

The application of empirical models to predict protein stability and aggregation can be a powerful practical method to support the selection of lead products. Among these models, shallow ANNs showed the best performance as they can solve problems that are complex, ill-defined, highly nonlinear, of many and different variables, and/or stochastic (**Chapter III** and **V**). Thanks to our trained model it is possible to achieve a better understanding of protein stability even before expression (**Chapter III**) or predict monomer retention in pharmaceutically relevant formulation after long term storage (**Chapter V**). Furthermore, these models can be used to highlight the most important biophysical assays to predict aggregation. Importantly, machine learning models can be designed in a way that allows continuous validation and improvement. They could be even more efficacious in industrial environments as the large amount of data usually available allows for the use of deep ANNs. However, the interpretation of "black-box" models is an open field of research. To address this problem we designed surrogate "white-box" models. While the quality of the prediction of simpler model might be lower, they can highly valuable for a better understanding of the process.

The major drawback of empirical algorithms is that their output does not provide any causations (i.e. understanding of the process) but only correlations. Even "white-box" models such as a simple linear model do "only" highlight useful data patterns. Therefore, the characterization of case studies for protein aggregation is yet extremely important. Following an extensive screening of several therapeutic proteins (**Chapter II**) we individuated the intense native reversible self-association of one IgG1, namely PPI-1 (**Chapter IV**). This process has been observed in literature at high mAb concentration, nonetheless PPI-1 exerts its self-association at low concentration. The nature of the self-association of the full-length IgG1 as well as the corresponding Fab and Fc fragments was investigated by several orthogonal methods. We

rationalized the self-association as a combination of hydrophobic and electrostatic interactions driven by the Fab fragments. A long term storage study demonstrated that PPI-1 is a perfect example of a protein having multiple non-exclusive aggregation pathways.

Finally, the development of new information rich techniques to investigate protein chemical and physical stability is of primary importance. In this direction we coupled UPLC-RP with MALS (**Chapter VI**), which is a natural means to characterize protein aggregates. The different principle of separation used in UPLC-RP-MALS provides an additional level of protein characterization compared to SEC-MALS and IEX-MALS. The MALS detection allows telling whether an impurity detected in UPLC-RP is indeed a chemical variant of the monomer or an aggregate or fragment.

APPENDIX

1 List of publications

1.1 Publications related to this thesis

1. Gentiluomo, L., Svilenov, H., Augustijn, D., El Bialy, I., Greco, M. L., Kulakova, A., Indrakumar. S., Mahapatra, S., Morales, M. M., Pohl, C., Roche, A., Tosstorff, A., Curtis, R., Derrick, J. P., Noergaard, A., Khan, T. A., Peters, G. H. J., Pluen, A., Rinnan, Å, Streicher, W. W., van der Walle, C. F., Uddin, S., Winter, G., Roessner, D., Harris, P., Frieß, W. Advancing therapeutic protein discovery and development through comprehensive computational and biophysical characterization. Molecular Pharmaceutics (2019).
2. Gentiluomo, L., Roessner, D., Augustijn, D., Svilenov, H., Kulakova, A., Mahapatra, S., Winter, G., Streicher, W., Rinnan, Å, Peters, G. H. J., Harris, P., Frieß, W. Application of interpretable artificial neural networks to early monoclonal antibodies development. European Journal of Pharmaceutics and Biopharmaceutics 141 (2019): 81-89.
3. 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 – Journal of pharmaceutical science 109 (2019): 443-451.
4. Gentiluomo, L., Roessner, D., Frieß, W. Application of machine learning to predict monomer retention of therapeutic proteins after long term storage. International Journal of Pharmaceutics (2020). Just accepted.
5. Gentiluomo, L., Schneider, V., Roessner, D., Frieß, W. Coupling Multi-Angle Light Scattering to Reverse-Phase Ultra-High-Pressure Chromatography (RP-UPLC-MALS) for the characterization monoclonal antibodies. Scientific reports 9 (2019): 1-8.

1.2 Publication not included into this thesis

1. 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. Journal of Pharmaceutical Sciences, 107 (2018): 3007-13.
2. Kulakova, A., Indrakumar, S., Sänderby, P., Gentiluomo, L., Streicher, W., Roessner, D., Frieß, W., Peters, G. H. J., Harris, P. Small angle X-ray scattering and molecular dynamic simulation provide molecular insight for stability of recombinant human transferring. Journal of Structural Biology X (2019).

Publications currently in preparation:

3. Gentiluomo L., Roessner, D., Frieß, W. Applications of online differential viscosimetry for protein characterization - In preparation
4. ElBialy, I., Gentiluomo, L., Roessner, D. and Frieß, W. Detection of Weak Protein-Surfactant Interactions Using Composition Gradient Multi-Angle Light Scattering - In preparation
5. Roche, A., Gentiluomo, L., Hebditch, M., Morales, M. M., Greco, M. L., Van Der Walle, C. F., Roessner, D., Frieß, W., Warwicker, J. and Curtis R. Using pH-Dependence of Monoclonal Antibody Electrophoretic Mobility to Determine Effective Isoelectric Points, Folded State Stability and Shape Effects on Net Charge Estimations - In preparation
6. Roche A., Gentiluomo, L., Roessner, D., Frieß, W., Trainoff, S. and Curtis, R. A Novel Technique for Measuring the Huggins' Coefficient in Protein and its Correlation to Concentrated Solution Behaviour - In preparation.
7. Mahapatra, S., Curtis, R., Gentiluomo, L., Peters, G. H. J., Nørgaard, A., Harris, P., Streicher, W. W. Self-Interactions of Monoclonal Antibodies at High Protein Concentrations - In preparation.
8. Kulakova, A., Rinnan, Å., Mahapatra, S., Pohl, C., ElBialy, I., Gentiluomo, L., Greco, M. L., Morales, M. M., Svilenov, H. L., Tosstorf, A., Curtis, R., Nørgaard, A., Khan, T., van der Walle, C. F., Uddin, S., Winter, G., Roessner, D., Frieß, W., Streicher, W. W., Peters, G. H. J. and Harris, P. Effect of Different Buffers and Excipients on Protein Stability - In preparation.
9. Frederiksen, T. M., Gentiluomo, L. et al. Oligomerization of a Glucagon-like Peptide 1 characterized by Field Flow fractionation and computational studies - In preparation
10. Augustijn D., Gentiluomo L. et al. The PIPPI protein formulation database - In preparation

1.3 Patent applications not included into this thesis

1. Lorenzo Gentiluomo. U.S. Provisional Patent Application at the U.S. Patent & Trademark Office, Owned by Wyatt Technology, 2019

1.4 Oral Presentations

1. PEGS, 2017, Lisbon (PT) : *„Application of Artificial Neural Networks in the development of protein formulation"*
2. 18th symposium on Field- and Flow-Based Separation, 2017, Columbia (US) : *„AF4-MALS in protein formulation development"*
3. Controlled Released Society, 2017, Marburg (DE) : *„Use of AF4-MALS to prove the mechanism of action of silver-nanolipid and proteins-drugs complexes"*
4. Columbia University, 2017, Columbia (US).: *„AF4 Workshop: Basic theory"*

5. Wyatt Technology, 2017, Santa Barbara (USA): *„The Pharmaceutical industry in Europe“*
6. Wyatt Technology, 2017, Dernbach (DE): *„Data mining workshop“*
7. CLS Behring, 2017, Malburg (DE): *“Application of interpretable artificial neural networks to early monoclonal antibodies development”*
8. Wyatt Protein Summit, 2017, Dernbach (DE): „*A New Approach to Study the Physical Stability of Monoclonal Antibody Formulations—Dilution From a Denaturant*“
9. Wyatt Technology, 2018, Dernbach (DE): “CG-MALS workshop”
10. Roche, 2018, Basel (CH): *„Advancing therapeutic protein discovery and development through comprehensive computational and biophysical characterization”*
11. Webinar, 2018: *„Predicting and evaluating the stability of therapeutic protein formulations by dynamic light scattering and machine learning“*
12. Wyatt Protein Summit, 2018, Dernbach (DE): *„Advancing therapeutic protein discovery and development through comprehensive computational and biophysical characterization”*
13. Max Planck Institute, 2019, Cologne (DE): “CG-MALS workshop”
14. PIPPI protein-protein interaction, 2019, Heidelberg (DE): *„Boosting therapeutic protein development by publicly available datasets including comprehensive biophysical and computational characterization“*
15. Biological and Pharmaceutical Complex Fluids 2019, Vienna (AT): *Application of machine learning to predict monomer retention of therapeutic proteins after long term storage at refrigerated and elevated temperatures*
16. Wyatt Protein Tour, 2019, Munich (DE): “*Applications of online differential viscosimetry for protein characterization”*
17. Wyatt Technology, 2019, Dernbach (DE): „*Protein formulation workshop“*
18. Nanotemper, 2019, Munich (DE): *„Application of machine learning to predict monomer retention of therapeutic proteins after long term storage”*
19. Coriolis Pharma, 2020, Munich (DE): *„Application of machine learning to predict monomer retention of therapeutic proteins after long term storage”*

1.5 Poster presentations

1. Protein aggregation, 2016, Manchester (UK): “*Application of UPLC-MALS to characterize protein aggregation”*
2. PEGS, 2017, Lisbon (PT): *„Application of Artificial Neural Networks in the development of protein formulation“*
3. PEGS, 2018, Lisbon (PT): *„Advancing therapeutic protein discovery and development through comprehensive computational and biophysical characterization”*
4. 11th Pharmaceutical world meeting, 2018, Granada (ES): *„Application of Artificial Neural Networks*

in the development of protein formulation“

5. Biotherapeutics and vaccines, 2019, Houston (USA): *„Advancing therapeutic protein discovery and development through comprehensive computational and biophysical characterization”*
6. PIPPI protein-protein interaction, 2019, Heidelberg: *„Characterization of native reversible self-association of a monoclonal antibody mediated by Fab-Fab interaction”*

1.6 Oral presentations and poster presentations from the international training network

1. PIPPI meeting, 2016, Cambridge (UK)
2. PIPPI meeting, 2017, Manchester (UK)
3. PIPPI meeting, 2017, Dernbach (DE)
4. PIPPI meeting, 2017, Copenhagen (DK)
5. PIPPI meeting, 2018, Munich (DE)
6. PIPPI meeting, 2018, Copenhagen (DK)

2 Biophysical parameter tables

(Part 1 of 22)

Parts of the data discussed in Chapter II (i.e. SI 13) are listed in the following.

Legend:

**Not detected.*
***Could not be determined accurately.*
****Not included for these conditions.*
#Not detected in the tested PEG concentration range.
##Not injected due to precipitation
PPI-04 unfolds already at low GuHCl concentrations, no pre-unfolding baseline can be obtained and no model can be fit to the data.
^ Value could not be detected with this experimental setup. PPI-18 is very hydrophobic and the dye already binds to the native protein.
^^ PPI-18 unfolds already at low GuHCl concentrations, no pre-unfolding baseline can be obtained and no model can be fit to the data.
^^^ Not included due to solubility problems.
~ Not measured due to limited sample amount.
~~ Value could not be detected with this experimental setup.
~~~ Probable binding of the dye to the native state

Biophysical parameter tables (Part 2 of 22)

Protein	Buffer	pH	NaCl (mM)	$T_{on,int}$ (°C)	$T_{m1,int}$ (°C)	$T_{m2,int}$ (°C)	$T_{on,ex}$ (°C)	$T_{m1,ex}$ (°C)	T_{agg} (°C)	k_D	ζ (mV)	R_h (nm)	PD (%)
PPI-01	10 mM histidine	5	0	52.12	57.44	*	46.89	55.17	52.69	-2.01E-02	9.81	5.55	13
PPI-01	10 mM histidine	5.5	0	53.43	58.84	*	49.61	56.86	54.28	-3.62E-02	7.50	5.83	7
PPI-01	10 mM histidine	6	0	54.48	60.73	*	48.42	58.35	55.38	-4.05E-02	5.83	7.96	15
PPI-01	10 mM histidine	6.5	0	54.90	62.55	*	51.42	59.96	56.96	-3.37E-02	2.94	12.92	18
PPI-01	10 mM histidine	7	0	55.29	64.62	*	51.52	62.20	56.8	-4.44E-02	-0.44	19.65	16
PPI-01	10 mM histidine	7.5	0	53.78	64.31	*	**	62.03	50.82	2.11E-02	-2.44	6.67	30
PPI-01	10 mM tris	8	0	56.77	69.97	76.03	**	63.54	49.47	***	-4.19	6.86	11
PPI-01	10 mM tris	9	0	58.98	69.99	75.88	51.68	62.99	**	***	-10.46	7.60	51
PPI-01	10 mM histidine	5	70	49.98	54.91	*	44.66	52.67	45.97	-1.90E-02	-	5.49	4
PPI-01	10 mM histidine	5.5	70	52.01	57.44	*	47.76	55.43	50.7	-1.72E-02	-	5.48	4
PPI-01	10 mM histidine	6	70	54.61	60.30	*	46.78	58.02	51.98	-2.46E-02	-	5.59	5
PPI-01	10 mM histidine	6.5	70	55.64	63.14	*	48.79	60.88	54.41	-2.45E-02	-	5.62	5
PPI-01	10 mM histidine	7	70	56.27	63.73	*	51.49	62.55	51.56	-2.94E-02	-	5.60	7
PPI-01	10 mM histidine	7.5	70	56.57	64.00	*	51.87	63.17	55.76	-2.34E-02	-	3.83	17
PPI-01	10 mM tris	8	70	53.60	69.93	77.11	50.94	64.17	55.48	***	-	5.52	5
PPI-01	10 mM tris	9	70	59.72	70.69	76.43	**	61.64	56.48	***	-	5.48	5
PPI-01	10 mM histidine	5	140	49.19	54.30	*	43.18	51.11	50.24	-1.74E-02	-	5.48	5
PPI-01	10 mM histidine	5.5	140	52.17	57.12	*	44.88	54.48	47.11	-1.87E-02	-	5.46	5
PPI-01	10 mM histidine	6	140	54.10	59.85	*	45.35	58.15	52.63	-1.81E-02	-	5.52	4
PPI-01	10 mM histidine	6.5	140	56.00	62.65	*	49.53	59.98	55.98	-2.39E-02	-	5.49	5
PPI-01	10 mM histidine	7	140	56.59	63.41	*	49.43	60.50	55.78	-2.01E-02	-	5.44	5
PPI-01	10 mM histidine	7.5	140	56.65	63.94	*	48.46	62.02	55.84	-1.89E-02	-	5.51	7
PPI-01	10 mM tris	8	140	55.17	70.59	77.41	48.83	61.89	56.81	***	-	5.49	5
PPI-01	10 mM tris	9	140	58.75	70.67	76.83	50.28	63.10	56.09	***	-	5.42	3
PPI-02	10 mM histidine	5	0	59.09	63.38	79.01	52.79	59.78	80	4.76E-02	12.00	5.38	8
PPI-02	10 mM histidine	5.5	0	60.48	65.94	80.73	57.99	64.69	80	5.09E-02	11.12	5.32	8
PPI-02	10 mM histidine	6	0	64.30	68.81	81.85	60.33	67.50	80	7.30E-02	10.83	5.20	9
PPI-02	10 mM histidine	6.5	0	66.86	71.14	83.00	62.02	69.04	80	6.97E-02	9.74	4.89	7
PPI-02	10 mM histidine	7	0	67.31	71.36	82.81	63.11	69.72	76.19	2.33E-03	4.69	4.56	7
PPI-02	10 mM histidine	7.5	0	67.33	71.37	82.76	63.07	69.59	76.24	2.55E-02	4.08	4.54	10
PPI-02	10 mM tris	8	0	64.70	71.10	82.77	63.00	69.34	70	***	0.53	5.29	3
PPI-02	10 mM tris	9	0	65.38	70.42	83.71	59.00	69.06	70	***	-3.83	5.27	2
PPI-02	10 mM histidine	5	70	55.83	60.45	77.17	47.05	55.56	68.1	5.13E-03	-	5.46	7
PPI-02	10 mM histidine	5.5	70	59.29	63.62	80.49	53.82	60.85	73	3.13E-03	-	5.42	4
PPI-02	10 mM histidine	6	70	62.38	66.48	82.85	56.49	64.21	73.69	3.90E-03	-	5.44	5
PPI-02	10 mM histidine	6.5	70	64.38	69.29	82.30	59.83	67.38	71.67	-3.03E-03	-	3.66	15
PPI-02	10 mM histidine	7	70	65.29	70.08	81.81	56.75	63.60	76.48	2.51E-02	-	5.38	4
PPI-02	10 mM histidine	7.5	70	65.69	70.35	81.68	61.39	68.99	74.05	-3.84E-03	-	5.57	10
PPI-02	10 mM tris	8	70	65.65	70.47	82.12	61.78	68.78	**	***	-	5.44	4
PPI-02	10 mM tris	9	70	65.60	70.09	85.39	59.81	68.59	**	***	-	5.47	4
PPI-02	10 mM histidine	5	140	55.58	59.62	75.82	46.29	55.29	67.06	1.23E-02	-	5.68	9
PPI-02	10 mM histidine	5.5	140	58.34	63.03	79.70	52.60	60.57	73.26	7.89E-03	-	5.49	7
PPI-02	10 mM histidine	6	140	61.92	66.15	82.38	57.38	64.13	75.8	-1.35E-03	-	5.47	7
PPI-02	10 mM histidine	6.5	140	63.78	68.72	82.14	58.47	67.08	74.34	-3.01E-03	-	5.46	6
PPI-02	10 mM histidine	7	140	64.85	69.67	81.01	60.41	68.05	65	1.06E-03	-	5.43	5
PPI-02	10 mM histidine	7.5	140	64.99	69.96	81.22	60.58	68.33	75.38	-3.58E-03	-	5.46	5
PPI-02	10 mM tris	8	140	65.55	70.01	81.52	61.39	68.00	**	***	-	5.45	3
PPI-02	10 mM tris	9	140	65.17	69.63	85.12	59.43	68.28	**	***	-	5.47	5

Biophysical parameter tables (Part 3 of 22)

Protein	Buffer	pH	NaCl (mM)	$T_{on,int}$ (ºC)	$T_{m1,int}$ (ºC)	$T_{m2,int}$ (ºC)	$T_{on,ex}$ (ºC)	$T_{m1,ex}$ (ºC)	T_{agg} (ºC)	k_D	ζ (mV)	R_h (nm)	PD (%)
PPI-03	10 mM histidine	5	0	55.08	61.05	75.66	**	58.97	65.98	3.91E-02	15.85	5.36	7
PPI-03	10 mM histidine	5.5	0	59.97	65.72	76.88	**	58.90	73.32	4.27E-02	12.66	5.21	5
PPI-03	10 mM histidine	6	0	59.75	69.30	76.98	**	68.42	75.89	8.63E-02	13.39	5.08	5
PPI-03	10 mM histidine	6.5	0	57.26	70.32	76.59	**	57.62	74.67	8.41E-02	10.85	5.45	7
PPI-03	10 mM histidine	7	0	56.43	70.53	76.25	**	56.86	73.56	9.35E-03	8.55	5.44	5
PPI-03	10 mM histidine	7.5	0	59.16	70.33	75.78	**	56.17	72.16	1.26E-02	7.01	5.55	8
PPI-03	10 mM tris	8	0	60.61	70.61	75.83	**	56.54	69.7	***	5.14	**	**
PPI-03	10 mM tris	9	0	59.58	69.93	75.55	**	56.84	67.94	***	-1.55	5.27	3
PPI-03	10 mM histidine	5	70	52.48	58.03	73.83	**	57.41	63.34	-3.36E-03	-	5.27	9
PPI-03	10 mM histidine	5.5	70	56.79	62.41	75.71	**	59.50	70.08	-1.44E-03	-	5.40	6
PPI-03	10 mM histidine	6	70	60.15	66.56	76.97	**	65.48	70.25	-2.88E-04	-	5.45	9
PPI-03	10 mM histidine	6.5	70	62.78	69.02	77.19	**	67.71	71.07	1.70E-04	-	5.34	4
PPI-03	10 mM histidine	7	70	54.22	70.37	77.17	**	57.40	70.45	-6.28E-03	-	5.38	6
PPI-03	10 mM histidine	7.5	70	58.01	70.74	77.06	**	56.49	60.87	1.66E-03	-	5.38	5
PPI-03	10 mM tris	8	70	59.03	70.90	76.94	**	57.21	**	***	-	5.48	14
PPI-03	10 mM tris	9	70	60.42	70.84	76.32	**	56.20	**	***	-	5.42	10
PPI-03	10 mM histidine	5	140	52.53	58.09	74.16	**	56.82	65.89	-5.39E-03	-	5.40	4
PPI-03	10 mM histidine	5.5	140	56.62	62.08	75.94	**	61.01	70.47	-3.44E-03	-	5.42	5
PPI-03	10 mM histidine	6	140	60.48	66.08	77.24	**	64.76	69.15	2.47E-03	-	5.44	8
PPI-03	10 mM histidine	6.5	140	63.23	69.17	77.53	**	67.61	70.82	-5.96E-03	-	5.41	5
PPI-03	10 mM histidine	7	140	55.36	70.22	77.49	**	68.49	59.86	-1.02E-03	-	5.41	5
PPI-03	10 mM histidine	7.5	140	54.75	70.69	77.29	**	68.59	70.96	-6.00E-03	-	5.46	9
PPI-03	10 mM tris	8	140	55.49	70.68	77.23	**	68.84	70.3	***	-	5.40	4
PPI-03	10 mM tris	9	140	59.60	70.92	76.80	**	68.36	69.38	***	-	5.62	10
PPI-04	10 mM histidine	5	0	52.77	61.64	*	50.52	61.75	65	3.19E-02	13.97	5.48	14
PPI-04	10 mM histidine	5.5	0	54.00	63.15	*	54.51	64.47	63.99	1.04E-02	8.40	5.71	24
PPI-04	10 mM histidine	6	0	54.75	64.24	74.90	56.50	65.30	63.49	3.80E-02	8.67	5.48	11
PPI-04	10 mM histidine	6.5	0	52.65	64.98	75.31	56.68	65.32	59.83	2.68E-02	6.94	5.43	9
PPI-04	10 mM histidine	7	0	54.41	65.19	*	55.83	65.38	58.73	-1.02E-02	2.70	5.83	27
PPI-04	10 mM histidine	7.5	0	53.31	65.59	*	56.81	65.43	55.7	-9.64E-03	**	6.17	38
PPI-04	10 mM tris	8	0	53.41	66.53	*	56.40	65.52	51.71	***	2.46	7.15	9
PPI-04	10 mM tris	9	0	55.27	66.95	*	52.70	66.08	44.88	***	-5.11	6.77	7
PPI-04	10 mM histidine	5	70	52.49	58.06	68.42	44.61	57.07	59.65	-3.97E-03	-	5.78	12
PPI-04	10 mM histidine	5.5	70	53.31	61.17	69.70	50.11	61.42	55.94	-5.45E-03	-	5.60	8
PPI-04	10 mM histidine	6	70	54.67	63.75	*	53.65	64.65	55.6	-5.64E-03	-	3.89	21
PPI-04	10 mM histidine	6.5	70	55.25	65.27	*	54.54	65.60	55.62	-2.76E-03	-	5.47	5
PPI-04	10 mM histidine	7	70	55.75	66.24	*	54.89	66.33	54.19	-2.44E-03	-	5.65	5
PPI-04	10 mM histidine	7.5	70	56.11	66.55	*	56.35	66.45	54.62	-9.06E-03	-	5.70	11
PPI-04	10 mM tris	8	70	56.56	66.98	*	53.96	66.67	55	***	-	5.66	4
PPI-04	10 mM tris	9	70	55.50	67.08	*	55.39	66.63	51.66	***	-	5.79	14
PPI-04	10 mM histidine	5	140	52.88	57.31	67.91	43.09	56.28	50.63	-4.91E-03	-	5.77	11
PPI-04	10 mM histidine	5.5	140	53.08	60.52	69.02	46.99	59.54	53.44	-6.17E-03	-	5.62	14
PPI-04	10 mM histidine	6	140	55.98	63.35	*	51.39	63.44	53.67	-8.87E-03	-	5.57	7
PPI-04	10 mM histidine	6.5	140	56.17	65.03	*	53.43	65.01	54.28	-1.15E-02	-	5.74	10
PPI-04	10 mM histidine	7	140	57.01	60.67	*	57.28	65.63	54.43	-1.39E-02	-	5.62	5
PPI-04	10 mM histidine	7.5	140	56.34	66.07	*	54.92	66.35	53.52	-1.40E-02	-	5.63	6
PPI-04	10 mM tris	8	140	54.83	66.59	*	55.90	66.39	53.61	***	-	5.62	4
PPI-04	10 mM tris	9	140	55.03	66.50	*	55.59	66.49	52.73	***	-	5.94	19

Biophysical parameter tables (Part 4 of 22)

Protein	Buffer	pH	NaCl (mM)	$T_{on,int}$ (ºC)	$T_{m1,int}$ (ºC)	$T_{m2,int}$ (ºC)	$T_{on,ex}$ (ºC)	$T_{m1,ex}$ (ºC)	T_{agg} (ºC)	k_D	ζ (mV)	R_h (nm)	PD (%)
PPI-08	10 mM histidine	5	0	49.65	67.96	*	49.23	61.15	57.99	2.74E-02	10.03	6.84	31
PPI-08	10 mM histidine	5.5	0	58.51	68.41	*	49.81	63.36	58.73	3.56E-02	9.46	6.37	9
PPI-08	10 mM histidine	6	0	64.15	75.71	*	52.85	64.97	58.17	5.11E-02	9.38	6.96	11
PPI-08	10 mM histidine	6.5	0	64.61	75.51	*	53.56	65.37	60.27	4.87E-02	8.47	6.02	16
PPI-08	10 mM histidine	7	0	64.42	74.71	*	54.07	66.05	56.03	6.64E-03	6.45	5.64	10
PPI-08	10 mM histidine	7.5	0	64.20	74.31	*	54.11	65.64	59.16	5.06E-02	4.41	6.17	45
PPI-08	10 mM tris	8	0	63.24	74.10	*	54.13	65.34	57.09	***	4.14	6.20	1
PPI-08	10 mM tris	9	0	61.47	73.55	*	53.40	65.07	50.51	***	-0.38	6.44	7
PPI-08	10 mM histidine	5	70	50.72	64.58	*	45.84	58.58	52.16	-5.59E-03	-	6.70	16
PPI-08	10 mM histidine	5.5	70	55.42	65.90	*	49.94	61.49	53.23	-6.84E-03	-	6.57	13
PPI-08	10 mM histidine	6	70	62.21	74.74	*	52.96	64.39	54.18	-8.39E-03	-	6.29	4
PPI-08	10 mM histidine	6.5	70	62.53	74.77	*	51.54	63.79	54.54	-9.14E-03	-	6.37	10
PPI-08	10 mM histidine	7	70	63.03	74.84	*	53.54	65.40	54.77	-2.99E-03	-	6.31	4
PPI-08	10 mM histidine	7.5	70	63.16	74.85	*	53.28	65.64	56.71	-7.03E-03	-	6.31	4
PPI-08	10 mM tris	8	70	62.65	74.75	*	52.37	66.45	56.47	***	-	6.31	3
PPI-08	10 mM tris	9	70	61.48	72.23	*	54.35	67.91	55.98	***	-	6.44	8
PPI-08	10 mM histidine	5	140	50.53	63.77	*	41.32	56.44	**	-5.65E-03	-	6.40	5
PPI-08	10 mM histidine	5.5	140	54.37	65.36	*	47.18	60.32	52.14	-6.89E-03	-	6.41	4
PPI-08	10 mM histidine	6	140	61.60	74.16	*	50.78	62.99	53.66	-9.97E-03	-	6.43	5
PPI-08	10 mM histidine	6.5	140	61.73	74.56	*	52.07	65.22	54.18	-1.04E-02	-	6.38	4
PPI-08	10 mM histidine	7	140	62.84	75.20	*	51.76	65.42	55.43	-9.40E-03	-	6.38	4
PPI-08	10 mM histidine	7.5	140	62.61	74.93	*	52.33	65.99	55.3	-1.29E-02	-	6.42	6
PPI-08	10 mM tris	8	140	62.97	74.00	*	53.07	67.15	54.92	***	-	7.54	8
PPI-08	10 mM tris	9	140	60.95	72.81	*	53.20	68.54	54.55	***	-	6.44	3
PPI-10	10 mM histidine	5	0	55.57	62.97	82.15	49.34	58.67	33.41	3.00E-02	12.73	6.00	31
PPI-10	10 mM histidine	5.5	0	59.88	65.91	82.65	54.56	63.46	50	4.74E-02	9.86	5.83	24
PPI-10	10 mM histidine	6	0	61.96	69.11	82.88	58.23	67.68	70	1.44E-01	6.77	5.23	8
PPI-10	10 mM histidine	6.5	0	63.44	71.44	82.57	60.56	69.72	70	7.52E-02	7.15	5.21	15
PPI-10	10 mM histidine	7	0	62.68	71.82	82.22	59.60	70.13	74.91	7.25E-04	6.42	4.90	16
PPI-10	10 mM histidine	7.5	0	63.39	72.01	82.10	58.56	70.36	66.48	2.47E-02	4.31	5.15	23
PPI-10	10 mM tris	8	0	63.47	71.73	82.66	56.90	69.85	71.45	***	3.88	5.47	4
PPI-10	10 mM tris	9	0	63.36	71.16	80.77	52.92	69.89	70	***	-0.54	5.60	5
PPI-10	10 mM histidine	5	70	52.98	59.10	80.37	42.76	53.39	35	-2.97E-03	-	6.07	22
PPI-10	10 mM histidine	5.5	70	56.52	62.74	80.86	53.49	61.66	65	-4.52E-03	-	5.75	17
PPI-10	10 mM histidine	6	70	60.52	66.48	81.38	55.70	65.55	68.91	-1.04E-02	-	5.57	7
PPI-10	10 mM histidine	6.5	70	62.47	69.39	82.56	58.36	67.79	65	-3.23E-03	-	5.60	9
PPI-10	10 mM histidine	7	70	62.91	70.35	82.40	56.26	68.82	66.05	-1.90E-04	-	5.53	5
PPI-10	10 mM histidine	7.5	70	62.80	70.71	82.06	55.04	69.03	42.32	-1.57E-03	-	5.58	6
PPI-10	10 mM tris	8	70	64.02	70.92	82.10	55.53	68.78	**	***	-	5.71	13
PPI-10	10 mM tris	9	70	63.76	70.47	80.73	52.64	68.90	70.56	***	-	5.57	7
PPI-10	10 mM histidine	5	140	53.27	58.20	79.16	43.28	53.17	40	-2.32E-03	-	5.77	16
PPI-10	10 mM histidine	5.5	140	57.01	62.16	80.38	50.81	59.81	66.64	-1.14E-02	-	**	**
PPI-10	10 mM histidine	6	140	60.44	65.79	81.68	54.32	64.52	63.44	-8.23E-03	-	6.87	25
PPI-10	10 mM histidine	6.5	140	62.66	68.78	82.16	57.85	67.12	71.57	-5.58E-02	-	5.71	10
PPI-10	10 mM histidine	7	140	63.71	69.85	82.06	**	68.57	68.91	-8.53E-03	-	5.60	7
PPI-10	10 mM histidine	7.5	140	63.27	70.35	81.82	54.55	69.37	**	-1.22E-02	-	5.60	6
PPI-10	10 mM tris	8	140	63.98	70.57	81.53	53.25	68.98	51.91	***	-	5.63	10
PPI-10	10 mM tris	9	140	63.43	69.94	80.15	53.85	69.00	64.99	***	-	5.67	9

Biophysical parameter tables (Part 5 of 22)

Protein	Buffer	pH	NaCl (mM)	$T_{on,int}$ (ºC)	$T_{m1,int}$ (ºC)	$T_{m2,int}$ (ºC)	$T_{on,ex}$ (ºC)	$T_{m1,ex}$ (ºC)	T_{agg} (ºC)	k_D	ζ (mV)	R_h (nm)	PD (%)
PPI-13	10 mM histidine	5	0	52.54	56.28	81.44	42.86	54.85	75	3.91E-02	10.46	5.39	10
PPI-13	10 mM histidine	5.5	0	55.99	60.01	83.00	50.40	58.37	78.69	4.27E-02	8.78	5.37	7
PPI-13	10 mM histidine	6	0	58.17	62.82	83.28	52.60	61.00	75.88	8.63E-02	6.26	5.44	18
PPI-13	10 mM histidine	6.5	0	59.12	64.13	82.60	50.66	61.40	65	8.41E-02	5.49	5.23	17
PPI-13	10 mM histidine	7	0	59.14	64.57	82.18	47.42	56.16	65	9.35E-03	3.21	5.31	22
PPI-13	10 mM histidine	7.5	0	58.85	64.46	81.70	50.92	56.60	73.63	4.50E-02	1.78	5.16	34
PPI-13	10 mM tris	8	0	58.64	64.14	81.64	50.88	58.63	42.89	***	0.60	5.76	12
PPI-13	10 mM tris	9	0	57.57	63.44	81.14	49.99	57.12	69.95	***	-3.36	5.78	6
PPI-13	10 mM histidine	5	70	50.01	53.64	79.15	41.60	52.60	59.3	-5.65E-03	-	5.58	6
PPI-13	10 mM histidine	5.5	70	53.88	57.65	81.21	47.24	57.14	68.11	-1.59E-03	-	5.71	7
PPI-13	10 mM histidine	6	70	56.89	61.08	82.23	51.30	59.31	60	-2.88E-04	-	5.78	8
PPI-13	10 mM histidine	6.5	70	58.66	63.25	82.48	53.18	60.54	55	1.70E-04	-	5.72	6
PPI-13	10 mM histidine	7	70	59.11	63.97	82.51	51.79	62.04	45.21	-6.28E-03	-	5.79	11
PPI-13	10 mM histidine	7.5	70	59.08	64.32	82.04	47.10	57.19	68.13	1.66E-03	-	5.95	13
PPI-13	10 mM tris	8	70	59.33	64.37	81.95	48.67	63.23	73.3	***	-	5.95	10
PPI-13	10 mM tris	9	70	58.53	63.96	80.38	45.93	55.99	70.05	***	-	6.14	24
PPI-13	10 mM histidine	5	140	49.44	52.90	77.84	42.86	51.62	62.13	-5.39E-03	-	5.56	5
PPI-13	10 mM histidine	5.5	140	53.72	57.20	80.40	47.65	56.08	64.03	-3.44E-03	-	5.97	28
PPI-13	10 mM histidine	6	140	56.74	60.61	81.97	51.16	59.52	64.49	2.20E-03	-	5.81	7
PPI-13	10 mM histidine	6.5	140	58.50	62.93	82.46	53.83	62.02	65.64	-5.96E-03	-	5.77	7
PPI-13	10 mM histidine	7	140	59.53	64.22	82.19	55.59	63.26	55.44	-1.02E-03	-	5.88	13
PPI-13	10 mM histidine	7.5	140	59.95	64.57	81.55	56.13	63.24	65	-6.00E-03	-	5.76	6
PPI-13	10 mM tris	8	140	59.58	64.67	81.49	54.84	63.11	66.06	***	-	6.48	14
PPI-13	10 mM tris	9	140	59.70	64.51	79.91	55.29	63.35	72.77	***	-	5.82	7
PPI-17	10 mM histidine	5	0	56.78	62.71	76.83		59.96	60.00	2.49E-02	4.98	5.90	44
PPI-17	10 mM histidine	5.5	0	59.43	66.09	77.32	47.98	64.68	67.68	2.87E-02	5.97	6.01	54
PPI-17	10 mM histidine	6	0	55.25	69.55	77.92	54.86	67.96	62.33	2.30E-02	4.24	6.58	48
PPI-17	10 mM histidine	6.5	0	61.09	77.93	*	59.33	69.52	68.90	2.05E-02	3.37	5.78	34
PPI-17	10 mM histidine	7	0	62.15	77.35	*	58.41	69.88	63.10	5.76E-03	1.19	5.51	30
PPI-17	10 mM histidine	7.5	0	62.16	76.94	*	59.51	70.04	62.47	-8.28E-03	0.89	6.16	42
PPI-17	10 mM tris	8	0	***	***	***	58.29	70.31	63.79	***	0.68	6.10	39
PPI-17	10 mM tris	9	0	***	***	***	58.91	72.04	55.00	***	-3.14	6.31	43
PPI-17	10 mM histidine	5	70	52.22	58.02	70.55	38.75	54.90	62.86	-1.76E-03	-	6.23	41
PPI-17	10 mM histidine	5.5	70	53.72	61.88	75.78	44.98	61.53	61.21	-2.14E-03	-	8.02	50
PPI-17	10 mM histidine	6	70	59.64	66.57	76.96	51.41	65.45	65.43	-3.76E-03	-	6.19	40
PPI-17	10 mM histidine	6.5	70	57.26	70.24	77.65	57.01	68.02	63.72	-3.23E-03	-	7.82	48
PPI-17	10 mM histidine	7	70	60.06	77.37	*	57.23	68.59	62.62	1.04E-02	-	6.88	48
PPI-17	10 mM histidine	7.5	70	61.17	77.00	*	58.27	69.44	65.67	-3.47E-03	-	6.66	48
PPI-17	10 mM tris	8	70	***	***	***	58.48	69.43	65.34	***	-	6.26	30
PPI-17	10 mM tris	9	70	***	***	***	57.38	69.25	61.97	***	-	6.56	50
PPI-17	10 mM histidine	5	140	50.70	56.24	69.27	39.78	53.94	54.37	-4.25E-03	-	6.55	50
PPI-17	10 mM histidine	5.5	140	54.90	61.01	75.26	41.88	59.00	58.01	-3.10E-03	-	6.49	50
PPI-17	10 mM histidine	6	140	58.51	65.49	76.28	49.68	63.62	65.94	-5.93E-03	-	6.50	45
PPI-17	10 mM histidine	6.5	140	59.72	69.44	77.09	54.37	67.48	66.13	-5.99E-03	-	6.30	47
PPI-17	10 mM histidine	7	140	59.21	77.00	*	54.37	68.36	63.53	-5.96E-03	-	6.24	52
PPI-17	10 mM histidine	7.5	140	60.88	76.81	*	55.66	68.55	65.36	-1.00E-02	-	6.29	34
PPI-17	10 mM tris	8	140	***	***	***	57.78	68.71	65.81	***	-	8.68	50
PPI-17	10 mM tris	9	140	***	***	***	56.48	68.75	64.33	***	-	7.08	50

Biophysical parameter tables (Part 6 of 22)

Protein	Buffer	pH	NaCl (mM)	$T_{on,int}$ (ºC)	$T_{m1,int}$ (ºC)	$T_{m2,int}$ (ºC)	$T_{on,ex}$ (ºC)	$T_{m1,ex}$ (ºC)	T_{agg} (ºC)	k_D	ζ (mV)	R_h (nm)	PD (%)
PPI-18	10 mM histidine	5	0	39.59	50.20	55.50	^	^	39.28	-9.78E-03	-0.46	5.76	19
PPI-18	10 mM histidine	5.5	0	44.40	51.11	55.46	^	^	40.77	-4.38E-03	-2.96	5.67	16
PPI-18	10 mM histidine	6	0	38.07	44.99	*	^	^	43.76	1.29E-02	-6.41	5.74	17
PPI-18	10 mM histidine	6.5	0	41.26	48.21	60.07	^	^	48.1	2.35E-02	-8.09	5.50	18
PPI-18	10 mM histidine	7	0	42.75	50.62	63.43	^	^	45.79	2.03E-02	-12.20	5.12	24
PPI-18	10 mM histidine	7.5	0	40.55	52.68	65.79	^	^	52.74	4.50E-02	-12.68	4.75	40
PPI-18	10 mM tris	8	0	39.14	47.85	53.82	^	^	51.18	***	-14.85	4.97	9
PPI-18	10 mM tris	9	0	37.22	46.27	54.20	^	^	56.88	***	-14.64	3.62	37
PPI-18	10 mM histidine	5	70	46.02	53.43	*	^	^	36.16	-5.76E-03	-	7.97	34
PPI-18	10 mM histidine	5.5	70	46.96	53.47	*	^	^	44	-4.93E-03	-	6.53	30
PPI-18	10 mM histidine	6	70	37.81	44.94	56.65	^	^	47.93	2.57E-03	-	7.63	42
PPI-18	10 mM histidine	6.5	70	41.55	48.45	60.22	^	^	47.61	-1.86E-03	-	6.47	39
PPI-18	10 mM histidine	7	70	44.20	50.88	65.00	^	^	44.93	2.86E-03	-	8.55	46
PPI-18	10 mM histidine	7.5	70	41.77	52.34	67.27	^	^	49.32	5.53E-03	-	6.17	38
PPI-18	10 mM tris	8	70	38.70	53.34	*	^	^	47.35	***	-	5.64	14
PPI-18	10 mM tris	9	70	37.73	46.65	53.85	^	^	49.21	***	-	5.74	24
PPI-18	10 mM histidine	5	140	46.24	53.27	*	^	^	36.59	-1.35E-02	-	9.55	41
PPI-18	10 mM histidine	5.5	140	47.29	53.35	*	^	^	40.98	3.01E-04	-	7.35	25
PPI-18	10 mM histidine	6	140	59.05	65.86	*	^	^	44.96	-1.01E-04	-	5.68	36
PPI-18	10 mM histidine	6.5	140	59.05	65.86	*	^	^	48.64	-2.75E-03	-	5.74	15
PPI-18	10 mM histidine	7	140	62.75	65.62	*	^	^	49.1	1.19E-03	-	6.52	46
PPI-18	10 mM histidine	7.5	140	62.75	65.62	*	^	^	49.17	-5.26E-04	-	5.98	24
PPI-18	10 mM tris	8	140	62.32	65.75	*	^	^	49.13	***	-	5.65	9
PPI-18	10 mM tris	9	140	62.32	65.75	*	^	^	48.63	***	-	5.91	25
PPI-30	10 mM histidine	5	0	59.05	65.86	*	~~~	~~~	55.59	~	8.36	3.03	21.73
PPI-30	10 mM histidine	5.5	0	59.77	65.80	*	~~~	~~~	^^^	~	5.53	^^^	^^^
PPI-30	10 mM histidine	6	0	^^^	^^^	*	~~~	~~~	^^^	~	^^^	^^^	^^^
PPI-30	10 mM histidine	6.5	0	^^^	^^^	*	~~~	~~~	56	~	^^^	6.07	33.20
PPI-30	10 mM histidine	7	0	62.75	65.61	*	~~~	~~~	57.49	~	-9.70	4.80	34.30
PPI-30	10 mM histidine	7.5	0	62.32	65.75	*	~~~	~~~	58.4	~	-11.49	**	**
PPI-30	10 mM tris	8	0	62.62	66.14	*	~~~	~~~	58.33	~	-18.45	4.20	30.07
PPI-30	10 mM tris	9	0	61.45	64.88	*	~~~	~~~	48.6	~	-	2.67	31.57
PPI-30	10 mM histidine	5	70	56.34	63.75	*	~~~	~~~	**	~	-	**	**
PPI-30	10 mM histidine	5.5	70	58.53	64.43	*	~~~	~~~	^^^	~	-	^^^	^^^
PPI-30	10 mM histidine	6	70	^^^	^^^	*	~~~	~~~	^^^	~	-	^^^	^^^
PPI-30	10 mM histidine	6.5	70	^^^	^^^	*	~~~	~~~	54.29	~	-	5.03	16.93
PPI-30	10 mM histidine	7	70	63.32	66.88	*	~~~	~~~	54.87	~	-	5.37	19.00
PPI-30	10 mM histidine	7.5	70	63.33	66.98	*	~~~	~~~	55.62	~	-	5.77	15.20
PPI-30	10 mM tris	8	70	63.20	66.96	*	~~~	~~~	55.83	~	-	5.70	19.43
PPI-30	10 mM tris	9	70	62.03	66.16	*	~~~	~~~	48.4	~	-	2.47	13.10
PPI-30	10 mM histidine	5	140	56.62	62.60	*	~~~	~~~	**	~	-	**	**
PPI-30	10 mM histidine	5.5	140	58.40	63.80	*	~~~	~~~	^^^	~	-	^^^	^^^
PPI-30	10 mM histidine	6	140	^^^	^^^	*	~~~	~~~	^^^	~	-	^^^	^^^
PPI-30	10 mM histidine	6.5	140	^^^	^^^	*	~~~	~~~	52.88	~	-	4.40	23.23
PPI-30	10 mM histidine	7	140	63.06	67.08	*	~~~	~~~	52.91	~	-	4.50	19.03
PPI-30	10 mM histidine	7.5	140	63.07	67.15	*	~~~	~~~	54.3	~	-	5.13	21.70
PPI-30	10 mM tris	8	140	63.05	67.09	*	~~~	~~~	54.19	~	-	5.20	21.10
PPI-30	10 mM tris	9	140	61.86	66.35	*	~~~	~~~	**	~	-	**	**

Biophysical parameter tables (Part 7 of 22)

Protein	Buffer	pH	NaCl (mM)	$T_{on,int}$ (ºC)	$T_{m1,int}$ (ºC)	$T_{m2,int}$ (ºC)	$T_{on,ex}$ (ºC)	$T_{m1,ex}$ (ºC)	T_{agg} (ºC)	k_D	ζ (mV)	R_h (nm)	PD (%)
PPI-44	10 mM histidine	5	0	51.92	61.90	67.77	51.69	60.09	69.32	1.46E-03	11.54	4.23	18
PPI-44	10 mM histidine	5.5	0	57.42	66.06	73.23	54.55	64.25	63.16	-3.26E-04	4.05	4.88	39
PPI-44	10 mM histidine	6	0	57.99	69.05	75.43	56.49	67.42	56.48	-9.84E-04	-3.01	4.14	41
PPI-44	10 mM histidine	6.5	0	53.68	80.32	*	~~	~~	55.00	-8.85E-03	-4.49	4.59	41
PPI-44	10 mM histidine	7	0	59.03	78.95	*	~~	~~	59.94	-4.08E-03	-10.21	4.84	39
PPI-44	10 mM histidine	7.5	0	73.41	85.54	*	~~	~~	65.34	2.02E-02	-12.20	4.57	43
PPI-44	10 mM tris	8	0	77.48	87.19	*	~~	~~	61.21	***	-	4.70	49
PPI-44	10 mM tris	9	0	76.03	88.20	*	~~	~~	**	***	-	4.13	44
PPI-44	10 mM histidine	5	70	36.85	40.80	59.82	47.69	58.08	58.51	-1.45E-02	-	4.73	27
PPI-44	10 mM histidine	5.5	70	56.60	65.90	*	~~	63.76	61.13	-1.42E-02	-	4.22	17
PPI-44	10 mM histidine	6	70	61.80	69.82	*	55.87	67.93	60.00	2.03E-03	-	4.10	12
PPI-44	10 mM histidine	6.5	70	65.36	73.12	77.06	~~	~~	64.59	-4.23E-03	-	4.13	13
PPI-44	10 mM histidine	7	70	65.34	82.60	*	~~	~~	65.45	-1.61E-03	-	5.22	27
PPI-44	10 mM histidine	7.5	70	70.98	84.57	*	~~	~~	62.00	-3.44E-03	-	5.54	40
PPI-44	10 mM tris	8	70	75.56	85.96	*	~~	~~	61.74	***	-	4.62	44
PPI-44	10 mM tris	9	70	57.79	85.88	*	~~	~~	57.53	***	-	4.84	50
PPI-44	10 mM histidine	5	140	28.79	39.84	58.44	39.11	57.22	57.40	-1.42E-02	-	4.49	27
PPI-44	10 mM histidine	5.5	140	56.20	65.30	*	51.83	63.50	60.50	-1.16E-02	-	4.29	17
PPI-44	10 mM histidine	6	140	61.32	69.63	*	56.25	67.80	60.00	-4.14E-03	-	4.30	30
PPI-44	10 mM histidine	6.5	140	66.28	73.22	76.43	~~	~~	64.70	-7.51E-03	-	4.45	35
PPI-44	10 mM histidine	7	140	61.41	72.95	80.37	~~	~~	66.84	-2.34E-03	-	4.28	28
PPI-44	10 mM histidine	7.5	140	66.63	83.73	*	~~	~~	62.75	-4.19E-03	-	6.40	31
PPI-44	10 mM tris	8	140	72.52	85.06	*	~~	~~	61.13	***	-	4.69	33
PPI-44	10 mM tris	9	140	72.48	85.18	*	~~	~~	57.85	***	-	4.77	41
PPI-45	10 mM histidine	5	0	34.61	57.77	*	45.95	54.38	37.56	~	~~	6.82	50
PPI-45	10 mM histidine	5.5	0	37.36	58.46	*	45.75	56.16	39.37	~	~~	6.77	50
PPI-45	10 mM histidine	6	0	39.43	58.75	*	47.81	56.37	42.52	~	~~	17.24	50
PPI-45	10 mM histidine	6.5	0	39.45	58.92	*	47.56	57.04	46.50	~	~~	20.41	50
PPI-45	10 mM histidine	7	0	43.78	58.88	*	49.71	56.95	46.36	~	~~	19.68	50
PPI-45	10 mM histidine	7.5	0	52.20	59.33	*	50.66	57.66	46.62	~	~~	22.26	50
PPI-45	10 mM tris	8	0	54.58	59.39	*	48.34	56.12	48.27	~	~~	4.29	50
PPI-45	10 mM tris	9	0	47.14	53.95	*	46.20	54.35	50.58	~	~~	16.48	50
PPI-45	10 mM histidine	5	70	35.38	57.70	*	43.80	54.47	39.62	~	-	5.31	50
PPI-45	10 mM histidine	5.5	70	40.50	58.51	*	45.65	55.78	42.10	~	-	4.99	50
PPI-45	10 mM histidine	6	70	40.46	58.65	*	46.22	57.44	44.63	~	-	11.61	50
PPI-45	10 mM histidine	6.5	70	39.41	58.79	*	47.85	56.94	46.44	~	-	10.82	50
PPI-45	10 mM histidine	7	70	38.80	58.38	*	48.54	56.67	47.07	~	-	5.57	50
PPI-45	10 mM histidine	7.5	70	39.48	57.89	*	48.34	56.21	47.16	~	-	13.63	50
PPI-45	10 mM tris	8	70	47.00	56.77	*	47.73	55.78	45.62	~	-	5.82	50
PPI-45	10 mM tris	9	70	46.25	54.90	*	45.30	53.95	46.42	~	-	5.78	50
PPI-45	10 mM histidine	5	140	36.76	57.76	*	44.63	54.02	39.63	~	-	5.02	50
PPI-45	10 mM histidine	5.5	140	38.49	58.61	*	45.54	56.96	42.32	~	-	6.09	50
PPI-45	10 mM histidine	6	140	40.38	58.88	*	47.14	57.28	44.97	~	-	6.60	50
PPI-45	10 mM histidine	6.5	140	37.86	59.04	*	46.82	57.68	46.86	~	-	12.36	50
PPI-45	10 mM histidine	7	140	39.33	58.82	*	49.29	57.01	47.35	~	-	4.72	50
PPI-45	10 mM histidine	7.5	140	39.74	58.30	*	48.70	56.53	46.28	~	-	5.34	50
PPI-45	10 mM tris	8	140	47.15	57.41	*	47.88	56.00	45.72	~	-	4.35	50
PPI-45	10 mM tris	9	140	42.65	55.92	*	45.58	54.52	46.50	~	-	5.85	50

Biophysical parameter tables (Part 8 of 22)

Protein	Buffer	pH	NaCl (mM)	$T_{on,int}$ (ºC)	$T_{m1,int}$ (ºC)	$T_{m2,int}$ (ºC)	$T_{on,ex}$ (ºC)	$T_{m1,ex}$ (ºC)	T_{agg} (ºC)	k_D	ζ (mV)	R_h (nm)	PD (%)
PPI-46	10 mM histidine	5	0	64.13	69.28	*	44.63	61.17	42.67	4.38E-02	~~	2.77	12
PPI-46	10 mM histidine	5.5	0	67.04	71.72	*	45.50	64.86	44.48	2.20E-02	~~	2.70	11
PPI-46	10 mM histidine	6	0	69.81	73.27	*	49.49	64.94	50.94	2.42E-02	~~	2.60	13
PPI-46	10 mM histidine	6.5	0	71.07	74.09	*	50.80	68.61	57.08	4.22E-02	~~	2.53	12
PPI-46	10 mM histidine	7	0	71.51	76.21	*	50.51	64.02	58.26	6.69E-02	~~	2.40	10
PPI-46	10 mM histidine	7.5	0	65.62	72.80	*	50.42	68.20	59.33	6.11E-02	~~	2.30	7
PPI-46	10 mM tris	8	0	64.84	72.83	*	51.10	64.26	62.13	***	~~	3.07	48
PPI-46	10 mM tris	9	0	62.04	71.25	*	50.71	66.53	61.58	***	~~	2.40	6
PPI-46	10 mM histidine	5	70	65.54	69.23	*	43.15	61.23	45.15	1.08E-01	-	2.77	10
PPI-46	10 mM histidine	5.5	70	68.40	72.20	*	48.32	61.32	46.17	3.21E-03	-	2.87	18
PPI-46	10 mM histidine	6	70	71.33	74.67	*	51.11	65.98	51.92	1.39E-03	-	2.70	9
PPI-46	10 mM histidine	6.5	70	71.80	79.22	*	51.89	66.64	56.73	2.77E-03	-	3.30	26
PPI-46	10 mM histidine	7	70	72.02	77.01	*	53.26	70.96	57.09	1.76E-02	-	2.83	17
PPI-46	10 mM histidine	7.5	70	66.04	72.24	*	53.43	70.45	57.35	1.34E-01	-	2.70	9
PPI-46	10 mM tris	8	70	64.67	71.49	*	51.12	69.68	57.70	***	-	2.90	17
PPI-46	10 mM tris	9	70	63.84	70.94	*	48.35	63.26	58.34	***	-	2.70	7
PPI-46	10 mM histidine	5	140	64.81	68.84	*	44.17	59.59	45.10	1.92E-01	-	2.83	12
PPI-46	10 mM histidine	5.5	140	68.42	72.18	*	47.61	61.42	46.94	-2.28E-03	-	2.73	9
PPI-46	10 mM histidine	6	140	71.49	74.63	*	51.82	66.43	52.48	8.20E-04	-	3.17	38
PPI-46	10 mM histidine	6.5	140	71.03	78.15	*	53.62	68.47	56.49	3.49E-03	-	2.77	11
PPI-46	10 mM histidine	7	140	72.56	77.14	*	53.86	71.17	56.71	5.39E-03	-	2.70	10
PPI-46	10 mM histidine	7.5	140	62.42	71.21	*	52.17	70.37	57.24	1.89E-03	-	2.80	10
PPI-46	10 mM tris	8	140	61.05	70.81	*	51.03	69.56	56.62	***	-	2.90	21
PPI-46	10 mM tris	9	140	60.87	68.81	*	46.86	67.66	56.70	***	-	2.80	14
PPI-49	10 mM histidine	5	0	38.62	47.07	*	~~~	~~~	66.26	1.78E-02	~~	3.66	12
PPI-49	10 mM histidine	5.5	0	44.83	52.75	*	~~~	~~~	62.06	2.35E-02	~~	3.73	14
PPI-49	10 mM histidine	6	0	48.44	56.14	*	~~~	~~~	52.86	1.98E-02	~~	4.58	35
PPI-49	10 mM histidine	6.5	0	49.53	57.73	*	~~~	~~~	48.38	3.61E-02	~~	3.44	26
PPI-49	10 mM histidine	7	0	50.04	58.63	*	~~~	~~~	51.34	5.12E-02	~~	2.70	15
PPI-49	10 mM histidine	7.5	0	50.92	59.61	*	~~~	~~~	48	4.44E-02	~~	2.77	26
PPI-49	10 mM tris	8	0	52.17	59.77	*	~~~	~~~	52.37	***	~~	3.47	9
PPI-49	10 mM tris	9	0	44.08	54.41	*	~~~	~~~	65	***	~~	3.33	24
PPI-49	10 mM histidine	5	70	40.57	48.76	*	~~~	~~~	70	5.78E-03	-	3.93	30
PPI-49	10 mM histidine	5.5	70	51.07	57.73	*	~~~	~~~	53.07	5.52E-03	-	3.72	25
PPI-49	10 mM histidine	6	70	56.49	62.90	*	~~~	~~~	51.74	5.10E-03	-	3.66	20
PPI-49	10 mM histidine	6.5	70	59.05	65.26	*	~~~	~~~	50.56	4.38E-03	-	4.14	26
PPI-49	10 mM histidine	7	70	59.84	65.57	*	~~~	~~~	54.67	7.89E-03	-	3.66	11
PPI-49	10 mM histidine	7.5	70	58.99	64.31	*	~~~	~~~	56.27	8.05E-03	-	3.64	13
PPI-49	10 mM tris	8	70	59.10	64.26	*	~~~	~~~	54.25	***	-	3.83	20
PPI-49	10 mM tris	9	70	51.36	59.40	*	~~~	~~~	55.4	***	-	3.87	21
PPI-49	10 mM histidine	5	140	41.90	50.00	*	~~~	~~~	74.51	1.77E-03	-	4.32	23
PPI-49	10 mM histidine	5.5	140	51.15	57.86	*	~~~	~~~	53.26	2.33E-02	-	3.79	21
PPI-49	10 mM histidine	6	140	57.47	63.62	*	~~~	~~~	50.24	9.13E-04	-	3.82	19
PPI-49	10 mM histidine	6.5	140	60.83	66.84	*	~~~	~~~	51.51	1.55E-03	-	3.72	14
PPI-49	10 mM histidine	7	140	62.01	67.42	*	~~~	~~~	50	6.38E-03	-	3.57	18
PPI-49	10 mM histidine	7.5	140	61.06	65.99	*	~~~	~~~	50	8.56E-03	-	3.58	10
PPI-49	10 mM tris	8	140	60.74	65.54	*	~~~	~~~	54.86	***	-	3.71	19
PPI-49	10 mM tris	9	140	55.20	61.53	*	~~~	~~~	56.04	***	-	3.96	24

Biophysical parameter tables (Part 9 of 22)

Protein	Buffer	pH	dG_1 (kcal/mol)	m_1	C_{m1} (M)	dG_2 (kcal/mol)	m_2	C_{m2} (M)	D_0 (cm2/s)	PEG_{TMP} (%)
PPI-01	10 mM histidine	5	6.82	4.79	1.42	7.45	2.88	2.58	4.35E-07	#
PPI-01	10 mM histidine	5.5	8.63	5.12	1.69	6.05	2.23	2.71	3.47E-07	***
PPI-01	10 mM histidine	6	8.71	5.26	1.65	4.21	1.82	2.31	2.33E-07	2.8
PPI-01	10 mM histidine	6.5	8.83	5.23	1.69	3.93	1.66	2.37	1.69E-07	***
PPI-01	10 mM histidine	7	9.34	5.02	1.86	4.6	1.74	2.65	4.58E-07	1.62
PPI-01	10 mM histidine	7.5	8.54	4.63	1.84	6.25	2.06	3.04	3.43E-07	***
PPI-01	10 mM tris	8	7.86	4.07	1.93	7.26	2.62	2.77	***	0
PPI-01	10 mM tris	9	5.22	2.4	2.17	12.77	4.48	2.85	***	1.7
PPI-01	10 mM histidine	5	5.94	4.05	1.47	7.94	2.86	2.78	4.44E-07	3.52
PPI-01	10 mM histidine	5.5	7.14	4.58	1.56	5.54	2.19	2.54	4.52E-07	***
PPI-01	10 mM histidine	6	8.89	4.88	1.82	5.02	1.79	2.8	4.59E-07	3.62
PPI-01	10 mM histidine	6.5	9.21	4.94	1.86	4.71	1.67	2.82	4.48E-07	***
PPI-01	10 mM histidine	7	8.61	4.77	1.81	5.23	1.83	2.85	3.41E-07	3.8
PPI-01	10 mM histidine	7.5	8.40	4.36	1.93	7.53	2.27	3.31	4.44E-07	***
PPI-01	10 mM tris	8	7.60	3.72	2.04	8.91	2.99	2.98	***	0
PPI-01	10 mM tris	9	4.06	1.73	2.34	15.8	5.26	3	***	3.72
PPI-01	10 mM histidine	5	6.40	4.27	1.5	7.34	2.75	2.66	4.52E-07	3.52
PPI-01	10 mM histidine	5.5	8.65	5.36	1.61	5.81	2.12	2.74	4.46E-07	***
PPI-01	10 mM histidine	6	10.31	6.06	1.7	4.41	1.73	2.54	4.52E-07	3.6
PPI-01	10 mM histidine	6.5	11.09	6.37	1.74	4.09	1.59	2.57	4.52E-07	***
PPI-01	10 mM histidine	7	11.73	6.31	1.86	4.45	1.69	2.63	4.46E-07	6.65
PPI-01	10 mM histidine	7.5	10.54	5.86	1.8	5.73	2.03	2.82	4.11E-07	***
PPI-01	10 mM tris	8	8.88	5.02	1.77	7.16	2.62	2.73	***	0
PPI-01	10 mM tris	9	5.20	2.21	2.36	13.22	4.52	2.92	***	2.82
PPI-02	10 mM histidine	5	4.71	2.44	1.93	11.24	3.13	3.59	4.29E-07	#
PPI-02	10 mM histidine	5.5	4.67	2.06	2.27	13.91	3.48	4	4.21E-07	***
PPI-02	10 mM histidine	6	4.02	1.77	2.27	14.56	3.7	3.93	4.16E-07	#
PPI-02	10 mM histidine	6.5	3.77	1.6	2.36	14.55	3.81	3.82	4.56E-07	***
PPI-02	10 mM histidine	7	3.66	1.54	2.38	15.06	3.79	3.97	4.33E-07	#
PPI-02	10 mM histidine	7.5	3.78	1.58	2.39	13.96	3.66	3.81	4.88E-07	***
PPI-02	10 mM tris	8	3.92	1.73	2.27	10.98	3.41	3.22	***	3.77
PPI-02	10 mM tris	9	4.91	2.36	2.08	8.84	2.55	3.47	***	***
PPI-02	10 mM histidine	5	2.87	1.23	2.34	19.37	4.97	3.9	4.23E-07	12.04
PPI-02	10 mM histidine	5.5	3.28	1.48	2.22	15.88	4.07	3.9	4.23E-07	***
PPI-02	10 mM histidine	6	3.93	1.69	2.32	13.76	3.36	4.1	4.16E-07	7.65
PPI-02	10 mM histidine	6.5	4.27	1.87	2.28	11.38	2.81	4.05	4.37E-07	***
PPI-02	10 mM histidine	7	4.39	2.01	2.18	8.96	2.44	3.67	3.49E-07	**
PPI-02	10 mM histidine	7.5	4.64	2.11	2.2	8.24	2.25	3.66	4.33E-07	***
PPI-02	10 mM tris	8	4.89	2.18	2.25	7.98	2.24	3.56	***	4.64
PPI-02	10 mM tris	9	4.70	2.19	2.14	9.47	2.74	3.46	***	**
PPI-02	10 mM histidine	5	2.82	1.24	2.27	15.62	4.09	3.82	3.94E-07	3.62
PPI-02	10 mM histidine	5.5	3.46	1.44	2.4	16.04	4	4.02	3.96E-07	***
PPI-02	10 mM histidine	6	3.78	1.61	2.35	15.35	3.87	3.97	4.36E-07	3.43
PPI-02	10 mM histidine	6.5	4.06	1.75	2.32	14.16	3.7	3.82	4.30E-07	***
PPI-02	10 mM histidine	7	4.18	1.85	2.25	13.35	3.5	3.81	4.19E-07	11.04
PPI-02	10 mM histidine	7.5	4.54	1.93	2.35	12.15	3.26	3.73	4.40E-07	***
PPI-02	10 mM tris	8	5.08	1.97	2.58	11.25	2.98	3.77	***	6.1
PPI-02	10 mM tris	9	4.05	1.96	2.06	7.77	2.32	3.35	***	**

Biophysical parameter tables (Part 10 of 22)

Protein	Buffer	pH	dG_1 (kcal/mol)	m_1	C_{m1} (M)	dG_2 (kcal/mol)	m_2	C_{m2} (M)	D_0 (cm2/s)	PEG_{TMP} (%)
PPI-03	10 mM histidine	5	8.14	3.4	2.21	2.49	1.13	2.4	4.61E-07	#
PPI-03	10 mM histidine	5.5	6.80	3.8	1.79	6.21	2.2	2.83	4.64E-07	***
PPI-03	10 mM histidine	6	7.16	4.1	1.75	7.85	2.98	2.64	4.43E-07	#
PPI-03	10 mM histidine	6.5	8.97	4.29	2.09	10.08	3.47	2.9	4.97E-07	***
PPI-03	10 mM histidine	7	9.08	4.39	2.07	11.24	3.68	3.05	4.47E-07	#
PPI-03	10 mM histidine	7.5	8.75	4.38	2	10.49	3.61	2.91	6.63E-07	***
PPI-03	10 mM tris	8	11.36	4.27	2.66	11.39	3.24	3.51	***	6.04
PPI-03	10 mM tris	9	10.11	3.74	2.7	5.37	1.65	3.25	***	2.23
PPI-03	10 mM histidine	5	6.15	4.41	1.4	5.32	2.07	2.57	4.71E-07	13.38
PPI-03	10 mM histidine	5.5	6.25	3.58	1.75	7.47	2.61	2.86	4.67E-07	***
PPI-03	10 mM histidine	6	6.12	3.21	1.91	9.32	2.99	3.12	4.64E-07	14.13
PPI-03	10 mM histidine	6.5	6.48	3.28	1.98	9.64	3.2	3.02	4.57E-07	***
PPI-03	10 mM histidine	7	7.43	3.81	1.95	9.21	3.24	2.84	4.73E-07	10.92
PPI-03	10 mM histidine	7.5	8.52	4.79	1.78	9.1	3.12	2.92	4.42E-07	***
PPI-03	10 mM tris	8	15.43	6.23	2.48	9.04	2.83	3.19	***	4.18
PPI-03	10 mM tris	9	27.71	10.45	2.65	5.15	1.76	2.92	***	3.67
PPI-03	10 mM histidine	5	8.58	5.51	1.56	4.91	1.89	2.59	4.72E-07	10.19
PPI-03	10 mM histidine	5.5	8.84	5.14	1.72	7.08	2.49	2.84	4.73E-07	***
PPI-03	10 mM histidine	6	8.65	4.94	1.75	8.25	2.91	2.83	4.55E-07	1.12
PPI-03	10 mM histidine	6.5	8.99	4.89	1.84	9.05	3.15	2.88	4.74E-07	***
PPI-03	10 mM histidine	7	10.30	5.01	2.06	9.47	3.2	2.96	4.61E-07	11.3
PPI-03	10 mM histidine	7.5	10.05	5.29	1.9	8.54	3.07	2.79	4.64E-07	***
PPI-03	10 mM tris	8	15.32	5.74	2.67	9.56	2.75	3.47	***	5.82
PPI-03	10 mM tris	9	19.73	7.12	2.77	4.83	1.58	3.06	***	4.02
PPI-04	10 mM histidine	5	###	###	###	###	###	###	4.47E-07	#
PPI-04	10 mM histidine	5.5	###	###	###	###	###	###	4.41E-07	***
PPI-04	10 mM histidine	6	###	###	###	###	###	###	4.25E-07	#
PPI-04	10 mM histidine	6.5	###	###	###	###	###	###	4.51E-07	***
PPI-04	10 mM histidine	7	###	###	###	###	###	###	4.08E-07	#
PPI-04	10 mM histidine	7.5	###	###	###	###	###	###	3.97E-07	***
PPI-04	10 mM tris	8	###	###	###	###	###	###	***	2.01
PPI-04	10 mM tris	9	###	###	###	###	###	###	***	0
PPI-04	10 mM histidine	5	###	###	###	###	###	###	4.31E-07	4.43
PPI-04	10 mM histidine	5.5	###	###	###	###	###	###	4.46E-07	***
PPI-04	10 mM histidine	6	###	###	###	###	###	###	4.42E-07	3.01
PPI-04	10 mM histidine	6.5	###	###	###	###	###	###	4.25E-07	***
PPI-04	10 mM histidine	7	###	###	###	###	###	###	4.00E-07	4.27
PPI-04	10 mM histidine	7.5	###	###	###	###	###	###	4.20E-07	***
PPI-04	10 mM tris	8	###	###	###	###	###	###	***	2.78
PPI-04	10 mM tris	9	###	###	###	###	###	###	***	0
PPI-04	10 mM histidine	5	###	###	###	###	###	###	4.49E-07	3.5
PPI-04	10 mM histidine	5.5	###	###	###	###	###	###	4.24E-07	***
PPI-04	10 mM histidine	6	###	###	###	###	###	###	4.48E-07	3.2
PPI-04	10 mM histidine	6.5	###	###	###	###	###	###	4.49E-07	***
PPI-04	10 mM histidine	7	###	###	###	###	###	###	4.49E-07	6.67
PPI-04	10 mM histidine	7.5	###	###	###	###	###	###	4.44E-07	***
PPI-04	10 mM tris	8	###	###	###	###	###	###	***	2.92
PPI-04	10 mM tris	9	###	###	###	###	###	###	***	0

Biophysical parameter tables (Part 11 of 22)

Protein	Buffer	pH	dG_1 (kcal/mol)	m_1	C_{m1} (M)	dG_2 (kcal/mol)	m_2	C_{m2} (M)	D_0 (cm2/s)	PEG_{TMP} (%)
PPI-08	10 mM histidine	5	5.45	3.85	1.41	6.93	3.18	2.18	3.90E-07	#
PPI-08	10 mM histidine	5.5	5.30	3.46	1.53	7.7	3.28	2.35	3.85E-07	***
PPI-08	10 mM histidine	6	5.16	3.16	1.63	7.55	3.3	2.29	3.74E-07	#
PPI-08	10 mM histidine	6.5	4.45	2.96	1.51	7.47	3.24	2.3	4.06E-07	***
PPI-08	10 mM histidine	7	5.06	2.84	1.78	7.91	3.11	2.54	3.71E-07	#
PPI-08	10 mM histidine	7.5	5.20	2.81	1.85	6.88	2.89	2.38	4.16E-07	***
PPI-08	10 mM tris	8	3.49	2.87	1.22	5.77	2.6	2.22	***	2.85
PPI-08	10 mM tris	9	7.08	3.25	2.17	3.91	1.78	2.2	***	2.78
PPI-08	10 mM histidine	5	5.07	3.46	1.47	6.98	3.04	2.3	3.88E-07	7.68
PPI-08	10 mM histidine	5.5	4.64	3.18	1.46	6.85	3.1	2.21	3.90E-07	***
PPI-08	10 mM histidine	6	4.62	2.93	1.58	7.57	3.12	2.43	3.97E-07	7.53
PPI-08	10 mM histidine	6.5	4.80	2.71	1.77	7.6	3.1	2.45	4.00E-07	***
PPI-08	10 mM histidine	7	4.50	2.51	1.79	7.13	3.04	2.34	3.82E-07	4.44
PPI-08	10 mM histidine	7.5	4.24	2.35	1.81	7.22	2.95	2.44	3.92E-07	***
PPI-08	10 mM tris	8	3.63	2.21	1.64	6.56	2.83	2.32	***	2.88
PPI-08	10 mM tris	9	3.62	2.01	1.8	5.92	2.46	2.4	***	2.82
PPI-08	10 mM histidine	5	5.35	3.82	1.4	9.08	4.08	2.23	3.79E-07	4.87
PPI-08	10 mM histidine	5.5	5.75	3.56	1.61	8.89	3.62	2.45	3.91E-07	***
PPI-08	10 mM histidine	6	5.71	3.3	1.73	7.8	3.23	2.42	3.99E-07	3.94
PPI-08	10 mM histidine	6.5	4.42	3.02	1.46	6.62	2.89	2.29	4.00E-07	***
PPI-08	10 mM histidine	7	4.48	2.74	1.64	6.17	2.61	2.36	3.98E-07	3.06
PPI-08	10 mM histidine	7.5	3.59	2.44	1.47	5.33	2.39	2.23	4.01E-07	***
PPI-08	10 mM tris	8	4.20	2.13	1.97	4.71	2.23	2.11	***	2.84
PPI-08	10 mM tris	9	2.47	1.49	1.66	4.58	2.08	2.21	***	2.85
PPI-10	10 mM histidine	5	2.78	1.31	2.13	8.39	2.31	3.64	4.75E-07	#
PPI-10	10 mM histidine	5.5	3.44	1.58	2.18	9.55	2.46	3.89	4.51E-07	***
PPI-10	10 mM histidine	6	3.79	1.82	2.08	9.97	2.59	3.85	3.45E-07	#
PPI-10	10 mM histidine	6.5	4.60	2.04	2.25	10.57	2.7	3.91	5.00E-07	***
PPI-10	10 mM histidine	7	4.97	2.24	2.22	10.89	2.8	3.89	4.37E-07	#
PPI-10	10 mM histidine	7.5	4.99	2.41	2.07	10.78	2.88	3.75	5.42E-07	***
PPI-10	10 mM tris	8	5.66	2.56	2.21	10.79	2.94	3.67	***	3.73
PPI-10	10 mM tris	9	5.38	2.77	1.94	10.47	3.01	3.48	***	1.79
PPI-10	10 mM histidine	5	3.51	1.7	2.06	9.1	2.47	3.68	4.47E-07	10.38
PPI-10	10 mM histidine	5.5	4.01	1.89	2.12	9.67	2.61	3.71	4.66E-07	***
PPI-10	10 mM histidine	6	4.66	2.05	2.28	10.94	2.72	4.02	4.57E-07	10.48
PPI-10	10 mM histidine	6.5	4.55	2.16	2.1	11.27	2.81	4.01	4.56E-07	***
PPI-10	10 mM histidine	7	4.90	2.25	2.18	10.91	2.88	3.79	4.40E-07	9.83
PPI-10	10 mM histidine	7.5	4.72	2.29	2.06	11.31	2.93	3.86	4.34E-07	***
PPI-10	10 mM tris	8	4.94	2.3	2.15	10.84	2.95	3.67	***	4.13
PPI-10	10 mM tris	9	4.65	2.2	2.11	10.78	2.93	3.68	***	4.14
PPI-10	10 mM histidine	5	3.51	1.55	2.26	9.45	2.56	3.7	4.35E-07	8.85
PPI-10	10 mM histidine	5.5	3.56	1.83	1.94	9.54	2.51	3.8	4.58E-07	***
PPI-10	10 mM histidine	6	4.09	2.07	1.98	9.4	2.49	3.77	4.51E-07	9.40
PPI-10	10 mM histidine	6.5	4.78	2.26	2.12	9.68	2.51	3.86	5.23E-07	***
PPI-10	10 mM histidine	7	4.90	2.4	2.04	10.05	2.56	3.93	4.44E-07	9.83
PPI-10	10 mM histidine	7.5	5.17	2.49	2.07	9.87	2.64	3.73	4.48E-07	***
PPI-10	10 mM tris	8	5.28	2.54	2.07	9.68	2.76	3.51	***	4.35
PPI-10	10 mM tris	9	5.53	2.5	2.21	10.85	3.09	3.51	***	3.73

Biophysical parameter tables (Part 12 of 22)

Protein	Buffer	pH	dG_1 (kcal/mol)	m_1	C_{m1} (M)	dG_2 (kcal/mol)	m_2	C_{m2} (M)	D_0 (cm2/s)	PEG_{TMP} (%)
PPI-13	10 mM histidine	5	4.83	3.63	1.33	5.36	1.83	2.93	4.61E-07	#
PPI-13	10 mM histidine	5.5	4.75	3.25	1.46	6.61	1.96	3.37	4.64E-07	***
PPI-13	10 mM histidine	6	4.79	2.93	1.63	7	2.11	3.32	4.43E-07	#
PPI-13	10 mM histidine	6.5	5.28	2.68	1.97	7.85	2.28	3.44	4.97E-07	***
PPI-13	10 mM histidine	7	4.58	2.48	1.85	8.78	2.47	3.55	4.47E-07	9.58
PPI-13	10 mM histidine	7.5	4.32	2.35	1.84	8.88	2.69	3.3	6.07E-07	***
PPI-13	10 mM tris	8	4.47	2.27	1.97	9.31	2.92	3.18	***	2.62
PPI-13	10 mM tris	9	4.07	2.31	1.76	10.66	3.46	3.08	***	2.39
PPI-13	10 mM histidine	5	3.65	2.41	1.52	6.06	1.92	3.16	4.74E-07	0.87
PPI-13	10 mM histidine	5.5	3.73	2.51	1.48	6.56	2.1	3.13	4.67E-07	***
PPI-13	10 mM histidine	6	4.44	2.64	1.68	7.82	2.26	3.46	4.64E-07	1.66
PPI-13	10 mM histidine	6.5	5.46	2.78	1.97	8.62	2.41	3.58	4.57E-07	***
PPI-13	10 mM histidine	7	5.31	2.93	1.81	8.22	2.53	3.25	4.73E-07	2.87
PPI-13	10 mM histidine	7.5	5.57	3.11	1.79	9.02	2.64	3.41	4.42E-07	***
PPI-13	10 mM tris	8	6.30	3.3	1.91	8.89	2.73	3.25	***	2.56
PPI-13	10 mM tris	9	6.67	3.75	1.78	9.04	2.86	3.16	***	2.86
PPI-13	10 mM histidine	5	4.01	2.71	1.48	6.84	2.25	3.04	4.72E-07	0.20
PPI-13	10 mM histidine	5.5	3.70	2.19	1.69	7.67	2.28	3.36	4.73E-07	***
PPI-13	10 mM histidine	6	3.53	1.96	1.8	7.68	2.31	3.32	4.58E-07	0.87
PPI-13	10 mM histidine	6.5	4.00	2.04	1.96	8.1	2.34	3.46	4.74E-07	***
PPI-13	10 mM histidine	7	4.46	2.42	1.85	8.11	2.38	3.41	4.61E-07	3.63
PPI-13	10 mM histidine	7.5	5.54	3.09	1.79	7.88	2.41	3.27	4.64E-07	***
PPI-13	10 mM tris	8	7.19	4.07	1.77	7.43	2.45	3.04	***	2.23
PPI-13	10 mM tris	9	12.06	6.92	1.74	7.63	2.52	3.02	***	2.44
PPI-17	10 mM histidine	5	5.72	3.85	1.49	10.01	4.35	2.3	4.26E-07	#
PPI-17	10 mM histidine	5.5	5.56	3.64	1.53	7.71	3.25	2.37	4.16E-07	***
PPI-17	10 mM histidine	6	5.42	3.44	1.58	7.99	3.42	2.34	4.30E-07	#
PPI-17	10 mM histidine	6.5	5.59	3.24	1.73	11.05	4.85	2.28	4.41E-07	***
PPI-17	10 mM histidine	7	5.73	3.04	1.89	18.03	7.54	2.39	4.13E-07	3.47
PPI-17	10 mM histidine	7.5	5.69	2.83	2.01	26.49	11.5	2.3	4.51E-07	***
PPI-17	10 mM tris	8	***	***	***	***	***	***	***	0
PPI-17	10 mM tris	9	***	***	***	***	***	***	***	0
PPI-17	10 mM histidine	5	6.97	4.42	1.58	9.62	4.1	2.35	3.83E-07	3.37
PPI-17	10 mM histidine	5.5	6.42	4.26	1.51	9.32	4.02	2.32	4.18E-07	***
PPI-17	10 mM histidine	6	6.59	4	1.65	10.32	4.26	2.42	4.17E-07	2.77
PPI-17	10 mM histidine	6.5	6.28	3.64	1.73	11.42	4.82	2.37	4.18E-07	
PPI-17	10 mM histidine	7	5.42	3.17	1.71	12.95	5.7	2.27	3.84E-07	2.34
PPI-17	10 mM histidine	7.5	5.18	2.6	1.99	16.54	6.9	2.4	4.11E-07	***
PPI-17	10 mM tris	8	***	***	***	***	***	***	***	0
PPI-17	10 mM tris	9	***	***	***	***	***	***	***	0
PPI-17	10 mM histidine	5	5.26	3.59	1.46	7.62	3.37	2.26	4.21E-07	2.1
PPI-17	10 mM histidine	5.5	5.44	3.24	1.68	11.34	4.63	2.45	3.97E-07	***
PPI-17	10 mM histidine	6	5.46	3.11	1.75	12.87	5.37	2.4	4.19E-07	14.19
PPI-17	10 mM histidine	6.5	5.51	3.22	1.71	12.7	5.58	2.28	4.19E-07	***
PPI-17	10 mM histidine	7	6.49	3.55	1.83	12.7	5.27	2.41	4.15E-07	2.86
PPI-17	10 mM histidine	7.5	6.69	4.11	1.63	9.91	2.23	2.23	4.24E-07	***
PPI-17	10 mM tris	8	***	***	***	***	***	***	***	0
PPI-17	10 mM tris	9	***	***	***	***	***	***	***	0

Biophysical parameter tables (Part 13 of 22)

Protein	Buffer	pH	dG_1 (kcal/mol)	m_1	C_{m1} (M)	dG_2 (kcal/mol)	m_2	C_{m2} (M)	D_0 (cm2/s)	PEG_{TMP} (%)
PPI-18	10 mM histidine	5	^^	^^	^^	^^	^^	^^	3.97E-07	3.48
PPI-18	10 mM histidine	5.5	^^	^^	^^	^^	^^	^^	4.53E-07	***
PPI-18	10 mM histidine	6	^^	^^	^^	^^	^^	^^	4.69E-07	8.78
PPI-18	10 mM histidine	6.5	^^	^^	^^	^^	^^	^^	4.86E-07	***
PPI-18	10 mM histidine	7	^^	^^	^^	^^	^^	^^	4.45E-07	#
PPI-18	10 mM histidine	7.5	^^	^^	^^	^^	^^	^^	5.86E-07	***
PPI-18	10 mM tris	8	^^	^^	^^	^^	^^	^^	***	#
PPI-18	10 mM tris	9	^^	^^	^^	^^	^^	^^	***	#
PPI-18	10 mM histidine	5	^^	^^	^^	^^	^^	^^	2.63E-07	6.88
PPI-18	10 mM histidine	5.5	^^	^^	^^	^^	^^	^^	3.46E-07	***
PPI-18	10 mM histidine	6	^^	^^	^^	^^	^^	^^	4.50E-07	#
PPI-18	10 mM histidine	6.5	^^	^^	^^	^^	^^	^^	4.62E-07	***
PPI-18	10 mM histidine	7	^^	^^	^^	^^	^^	^^	4.63E-07	12.93
PPI-18	10 mM histidine	7.5	^^	^^	^^	^^	^^	^^	4.46E-07	***
PPI-18	10 mM tris	8	^^	^^	^^	^^	^^	^^	***	#
PPI-18	10 mM tris	9	^^	^^	^^	^^	^^	^^	***	#
PPI-18	10 mM histidine	5	^^	^^	^^	^^	^^	^^	1.28E-07	6.7
PPI-18	10 mM histidine	5.5	^^	^^	^^	^^	^^	^^	4.03E-07	***
PPI-18	10 mM histidine	6	^^	^^	^^	^^	^^	^^	4.50E-07	#
PPI-18	10 mM histidine	6.5	^^	^^	^^	^^	^^	^^	4.71E-07	***
PPI-18	10 mM histidine	7	^^	^^	^^	^^	^^	^^	4.58E-07	#
PPI-18	10 mM histidine	7.5	^^	^^	^^	^^	^^	^^	4.62E-07	***
PPI-18	10 mM tris	8	^^	^^	^^	^^	^^	^^	***	#
PPI-18	10 mM tris	9	^^	^^	^^	^^	^^	^^	***	#
PPI-30	10 mM histidine	5	7.44	2.12	3.51	*	*	*	~	~
PPI-30	10 mM histidine	5.5	7.39	2.1	3.52	*	*	*	~	~
PPI-30	10 mM histidine	6	^^^	^^^	^^^	*	*	*	~	~
PPI-30	10 mM histidine	6.5	^^^	^^^	^^^	*	*	*	~	~
PPI-30	10 mM histidine	7	6.44	1.83	3.52	*	*	*	~	~
PPI-30	10 mM histidine	7.5	8.08	2.27	3.56	*	*	*	~	~
PPI-30	10 mM tris	8	8.98	2.48	3.62	*	*	*	~	~
PPI-30	10 mM tris	9	8.58	2.27	3.79	*	*	*	~	~
PPI-30	10 mM histidine	5	6.07	1.72	3.53	*	*	*	~	~
PPI-30	10 mM histidine	5.5	7.50	2.1	3.57	*	*	*	~	~
PPI-30	10 mM histidine	6	^^^	^^^	^^^	*	*	*	~	~
PPI-30	10 mM histidine	6.5	^^^	^^^	^^^	*	*	*	~	~
PPI-30	10 mM histidine	7	8.24	2.38	3.46	*	*	*	~	~
PPI-30	10 mM histidine	7.5	7.84	2.22	3.53	*	*	*	~	~
PPI-30	10 mM tris	8	9.08	2.45	3.71	*	*	*	~	~
PPI-30	10 mM tris	9	9.03	2.46	3.68	*	*	*	~	~
PPI-30	10 mM histidine	5	5.67	1.59	3.56	*	*	*	~	~
PPI-30	10 mM histidine	5.5	6.66	1.87	3.57	*	*	*	~	~
PPI-30	10 mM histidine	6	^^^	^^^	^^^	*	*	*	~	~
PPI-30	10 mM histidine	6.5	^^^	^^^	^^^	*	*	*	~	~
PPI-30	10 mM histidine	7	6.95	2	3.48	*	*	*	~	~
PPI-30	10 mM histidine	7.5	7.34	2.08	3.54	*	*	*	~	~
PPI-30	10 mM tris	8	8.52	2.31	3.69	*	*	*	~	~
PPI-30	10 mM tris	9	8.55	2.29	3.73	*	*	*	~	~

Biophysical parameter tables (Part 14 of 22)

Protein	Buffer	pH	dG_1 (kcal/mol)	m_1	C_{m1} (M)	dG_2 (kcal/mol)	m_2	C_{m2} (M)	D_0 (cm2/s)	PEG_{TMP} (%)
PPI-44	10 mM histidine	5	6.56	6.43	1.02	11.06	6.69	1.65	6.19E-07	#
PPI-44	10 mM histidine	5.5	6.41	6.45	0.99	9.07	5.28	1.72	6.29E-07	***
PPI-44	10 mM histidine	6	7.28	6.29	1.16	7.23	4.14	1.75	6.10E-07	3.27
PPI-44	10 mM histidine	6.5	8.93	5.97	1.5	7.59	3.24	2.34	6.35E-07	***
PPI-44	10 mM histidine	7	9.16	5.47	1.67	6.96	2.61	2.67	5.93E-07	3.22
PPI-44	10 mM histidine	7.5	8.08	4.8	1.68	6.17	2.24	2.76	6.09E-07	***
PPI-44	10 mM tris	8	7.96	3.97	2.01	6.3	2.12	2.97	***	#
PPI-44	10 mM tris	9	3.77	1.78	2.12	7.58	2.67	2.84	***	#
PPI-44	10 mM histidine	5	3.45	3.97	0.87	10.7	6.51	1.64	5.79E-07	6.74
PPI-44	10 mM histidine	5.5	5.52	5.73	0.96	8.51	5.14	1.66	6.02E-07	***
PPI-44	10 mM histidine	6	7.74	6.9	1.12	7.1	4.03	1.76	5.99E-07	9.96
PPI-44	10 mM histidine	6.5	10.39	7.49	1.39	7.02	3.17	2.22	6.13E-07	***
PPI-44	10 mM histidine	7	12.09	7.49	1.61	6.61	2.56	2.59	6.12E-07	11.94
PPI-44	10 mM histidine	7.5	10.89	6.91	1.58	5.79	2.2	2.63	5.87E-07	***
PPI-44	10 mM tris	8	10.47	5.74	1.83	6.06	2.1	2.89	***	#
PPI-44	10 mM tris	9	3.76	1.64	2.3	6.82	2.65	2.58	***	#
PPI-44	10 mM histidine	5	5.43	6.83	0.8	9.95	6.25	1.59	5.53E-07	#
PPI-44	10 mM histidine	5.5	6.57	6.84	0.96	8.6	5.07	1.7	5.93E-07	***
PPI-44	10 mM histidine	6	7.65	6.68	1.15	7.27	4.09	1.78	5.99E-07	#
PPI-44	10 mM histidine	6.5	9.20	6.36	1.45	7.39	3.33	2.22	6.27E-07	***
PPI-44	10 mM histidine	7	9.00	5.88	1.53	7.28	2.77	2.62	6.09E-07	13.64
PPI-44	10 mM histidine	7.5	8.75	5.23	1.67	6.76	2.43	2.79	5.91E-07	***
PPI-44	10 mM tris	8	8.21	4.42	1.86	6.68	2.29	2.92	***	#
PPI-44	10 mM tris	9	4.64	2.32	2	7.5	2.64	2.84	***	#
PPI-45	10 mM histidine	5	5.70	2.45	2.33	*	*	*	~	~
PPI-45	10 mM histidine	5.5	5.10	2.27	2.25	*	*	*	~	~
PPI-45	10 mM histidine	6	4.79	2.13	2.25	*	*	*	~	~
PPI-45	10 mM histidine	6.5	4.66	2.1	2.28	*	*	*	~	~
PPI-45	10 mM histidine	7	4.47	1.98	2.26	*	*	*	~	~
PPI-45	10 mM histidine	7.5	4.58	1.96	2.33	*	*	*	~	~
PPI-45	10 mM tris	8	4.42	1.99	2.23	*	*	*	~	~
PPI-45	10 mM tris	9	4.85	2.16	2.25	*	*	*	~	~
PPI-45	10 mM histidine	5	5.08	2.18	2.33	*	*	*	~	~
PPI-45	10 mM histidine	5.5	4.85	2.15	2.25	*	*	*	~	~
PPI-45	10 mM histidine	6	5.02	2.13	2.36	*	*	*	~	~
PPI-45	10 mM histidine	6.5	5.06	2.09	2.38	*	*	*	~	~
PPI-45	10 mM histidine	7	4.70	2.07	2.27	*	*	*	~	~
PPI-45	10 mM histidine	7.5	4.92	2.04	2.41	*	*	*	~	~
PPI-45	10 mM tris	8	4.58	2.02	2.27	*	*	*	~	~
PPI-45	10 mM tris	9	4.44	1.96	2.27	*	*	*	~	~
PPI-45	10 mM histidine	5	5.16	2.16	2.39	*	*	*	~	~
PPI-45	10 mM histidine	5.5	5.04	2.22	2.27	*	*	*	~	~
PPI-45	10 mM histidine	6	5.12	2.24	2.33	*	*	*	~	~
PPI-45	10 mM histidine	6.5	5.14	2.22	2.31	*	*	*	~	~
PPI-45	10 mM histidine	7	4.94	2.18	2.27	*	*	*	~	~
PPI-45	10 mM histidine	7.5	4.66	2.1	2.22	*	*	*	~	~
PPI-45	10 mM tris	8	4.41	1.98	2.22	*	*	*	~	~
PPI-45	10 mM tris	9	4.32	1.89	2.28	*	*	*	~	~

Biophysical parameter tables (Part 15 of 22)

Protein	Buffer	pH	dG_1 (kcal/mol)	m_1	C_{m1} (M)	dG_2 (kcal/mol)	m_2	C_{m2} (M)	D_0 (cm2/s)	PEG_{TMP} (%)
PPI-46	10 mM histidine	5	8.29	2.66	3.12	*	*	*	2.87E-08	~
PPI-46	10 mM histidine	5.5	9.39	2.66	3.53	*	*	*	6.99E-07	~
PPI-46	10 mM histidine	6	11.39	2.87	3.97	*	*	*	8.46E-07	~
PPI-46	10 mM histidine	6.5	7.86	2.26	3.47	*	*	*	8.79E-07	~
PPI-46	10 mM histidine	7	6.77	1.91	3.54	*	*	*	9.17E-07	~
PPI-46	10 mM histidine	7.5	8.08	2.26	3.57	*	*	*	9.18E-07	~
PPI-46	10 mM tris	8	8.99	2.48	3.62	*	*	*	***	~
PPI-46	10 mM tris	9	8.67	2.29	3.79	*	*	*	***	~
PPI-46	10 mM histidine	5	7.86	2.66	2.95	*	*	*	2.38E-08	~
PPI-46	10 mM histidine	5.5	9.65	2.7	3.57	*	*	*	7.49E-07	~
PPI-46	10 mM histidine	6	10.87	2.84	3.82	*	*	*	8.45E-07	~
PPI-46	10 mM histidine	6.5	7.32	2.15	3.41	*	*	*	8.87E-07	~
PPI-46	10 mM histidine	7	8.23	2.38	3.45	*	*	*	8.36E-07	~
PPI-46	10 mM histidine	7.5	7.43	2.12	3.5	*	*	*	4.62E-02	~
PPI-46	10 mM tris	8	9.43	2.54	3.71	*	*	*	***	~
PPI-46	10 mM tris	9	9.35	2.53	3.7	*	*	*	***	~
PPI-46	10 mM histidine	5	8.55	2.83	3.02	*	*	*	1.81E-08	~
PPI-46	10 mM histidine	5.5	8.48	2.41	3.51	*	*	*	7.40E-07	~
PPI-46	10 mM histidine	6	10.10	2.62	3.86	*	*	*	8.01E-07	~
PPI-46	10 mM histidine	6.5	7.72	2.24	3.44	*	*	*	8.21E-07	~
PPI-46	10 mM histidine	7	7.18	2.06	3.48	*	*	*	8.44E-07	~
PPI-46	10 mM histidine	7.5	8.41	2.35	3.58	*	*	*	8.62E-07	~
PPI-46	10 mM tris	8	8.68	2.35	3.7	*	*	*	***	~
PPI-46	10 mM tris	9	8.48	2.27	3.74	*	*	*	***	~
PPI-49	10 mM histidine	5	5.65	2.4	2.35	*	*	*	5.62E-07	***
PPI-49	10 mM histidine	5.5	5.43	2.47	2.2	*	*	*	5.76E-07	***
PPI-49	10 mM histidine	6	5.46	2.5	2.19	*	*	*	6.46E-07	***
PPI-49	10 mM histidine	6.5	5.13	2.47	2.08	*	*	*	6.48E-07	***
PPI-49	10 mM histidine	7	4.83	2.39	2.02	*	*	*	5.98E-07	***
PPI-49	10 mM histidine	7.5	4.58	2.27	2.02	*	*	*	7.89E-07	***
PPI-49	10 mM tris	8	4.07	2.1	1.94	*	*	*	***	***
PPI-49	10 mM tris	9	3.19	1.61	1.98	*	*	*	***	***
PPI-49	10 mM histidine	5	5.80	2.38	2.44	*	*	*	6.24E-07	***
PPI-49	10 mM histidine	5.5	5.43	2.43	2.23	*	*	*	6.31E-07	***
PPI-49	10 mM histidine	6	5.65	2.43	2.32	*	*	*	6.37E-07	***
PPI-49	10 mM histidine	6.5	4.89	2.38	2.06	*	*	*	6.43E-07	***
PPI-49	10 mM histidine	7	4.37	2.27	1.93	*	*	*	6.32E-07	***
PPI-49	10 mM histidine	7.5	4.08	2.1	1.94	*	*	*	6.31E-07	***
PPI-49	10 mM tris	8	3.65	1.88	1.94	*	*	*	***	***
PPI-49	10 mM tris	9	2.95	1.27	2.33	*	*	*	***	***
PPI-49	10 mM histidine	5	6.29	2.56	2.46	*	*	*	6.37E-07	***
PPI-49	10 mM histidine	5.5	5.96	2.4	2.49	*	*	*	5.33E-07	***
PPI-49	10 mM histidine	6	5.02	2.24	2.24	*	*	*	6.56E-07	***
PPI-49	10 mM histidine	6.5	4.30	2.09	2.06	*	*	*	6.60E-07	***
PPI-49	10 mM histidine	7	3.87	1.95	1.99	*	*	*	6.54E-07	***
PPI-49	10 mM histidine	7.5	3.65	1.81	2.01	*	*	*	6.38E-07	***
PPI-49	10 mM tris	8	4.02	1.69	2.38	*	*	*	***	***
PPI-49	10 mM tris	9	2.53	1.45	1.75	*	*	*	***	***

Biophysical parameter tables (Part 16 of 22)

Protein	Buffer	pH	$m_{25,rec}$ (%)	$m_{40,rec}$ (%)	$m_{50,rec}$ (%)	M_{25} (%)	M_{40} (%)	M_{50} (%)	LSA_{25}	LSA_{40}	LSA_{50}	$R_{R,25}$	$R_{R,40}$	$R_{R,50}$
PPI-01	10 mM histidine	5	90.4	85.4	5.8	99.70	99.10	11.10	0.96	0.97	0.14	0.99	1.03	1.82
PPI-01	10 mM histidine	5.5	86.3	85.2	25.4	99.60	98.90	52.80	0.97	0.94	0.47	0.98	1.05	2.01
PPI-01	10 mM histidine	6	97.8	94.4	46.4	99.50	98.90	76.20	0.97	0.94	0.58	0.98	1.03	1.04
PPI-01	10 mM histidine	6.5	88.2	89.7	44.9	99.30	98.10	86.20	0.97	0.92	0.46	0.95	1.05	1.55
PPI-01	10 mM histidine	7	88.6	81.1	36.3	99.20	98.60	97.40	0.97	0.95	0.94	0.93	0.99	0.83
PPI-01	10 mM histidine	7.5	91.7	77.1	37.6	99.00	97.90	88.80	0.90	0.97	0.82	1.09	0.90	1.04
PPI-01	10 mM tris	8	87.1	89.7	51.5	99.60	99.30	97.60	0.95	0.99	0.93	0.94	0.97	0.98
PPI-01	10 mM tris	9	97.3	95.9	44.2	99.40	98.20	77.80	0.95	0.91	0.34	0.96	1.27	2.60
PPI-01	10 mM histidine	5	81.4	69.8	0.8	99.60	97.10	1.90	0.98	0.92	0.04	0.99	1.95	6.97
PPI-01	10 mM histidine	5.5	90.4	89.9	2.1	99.70	99.20	4.50	0.98	0.98	0.13	1.01	1.05	23.10
PPI-01	10 mM histidine	6	89.9	88.6	42.6	99.70	99.70	98.90	0.98	0.99	0.67	0.98	0.98	**
PPI-01	10 mM histidine	6.5	94.7	94.8	55.0	99.70	99.60	99.30	0.98	1.00	1.00	0.99	0.98	10.53
PPI-01	10 mM histidine	7	93.4	96.7	51.8	99.70	99.60	99.10	0.98	0.99	0.96	0.99	0.98	1.00
PPI-01	10 mM histidine	7.5	81.4	81.7	47.3	99.60	99.50	99.00	0.98	0.99	0.98	1.45	1.42	1.49
PPI-01	10 mM tris	8	102.1	96.9	52.9	99.50	99.40	98.50	0.99	0.97	0.90	1.00	0.99	1.10
PPI-01	10 mM tris	9	82.8	81.4	40.9	99.50	98.70	86.20	0.98	0.97	0.39	0.99	1.07	3.78
PPI-01	10 mM histidine	5	79.6	50.4	##	99.60	88.20	##	0.99	0.27	##	0.99	5.82	103.35
PPI-01	10 mM histidine	5.5	92.2	93.2	##	99.60	98.70	##	0.99	0.98	##	1.00	1.18	84.32
PPI-01	10 mM histidine	6	75.0	85.0	38.2	99.60	99.50	97.50	0.98	0.98	0.40	0.98	0.99	13.45
PPI-01	10 mM histidine	6.5	81.3	90.2	64.3	99.50	99.60	99.20	0.98	0.99	0.81	0.99	0.98	1.73
PPI-01	10 mM histidine	7	92.5	93.4	53.6	99.70	99.60	98.60	0.99	0.99	0.93	0.99	0.99	1.23
PPI-01	10 mM histidine	7.5	80.3	80.6	46.0	99.70	99.40	97.80	0.99	0.99	0.85	0.99	0.98	1.57
PPI-01	10 mM tris	8	93.5	94.4	47.4	99.60	99.40	97.20	0.99	0.98	0.03	1.00	0.98	2.50
PPI-01	10 mM tris	9	89.5	92.5	35.9	99.50	98.50	79.90	0.98	0.96	0.37	1.00	1.16	3.13
PPI-02	10 mM histidine	5	77.7	77.7	67.1	99.80	98.50	98.50	0.99	0.96	0.98	1.05	1.02	1.07
PPI-02	10 mM histidine	5.5	89.1	78.4	73.3	100.00	98.70	98.60	0.99	0.97	0.98	0.98	1.00	1.04
PPI-02	10 mM histidine	6	113.5	84.6	68.7	99.90	98.80	98.70	0.99	0.97	0.98	0.97	1.00	1.02
PPI-02	10 mM histidine	6.5	89.3	77.0	52.2	99.90	98.80	98.70	0.99	0.97	0.42	0.70	1.01	1.05
PPI-02	10 mM histidine	7	85.7	75.1	65.1	99.90	98.70	98.90	0.99	0.97	0.91	1.02	1.05	1.06
PPI-02	10 mM histidine	7.5	84.1	76.0	55.8	99.90	99.20	98.70	0.99	0.97	0.49	1.06	1.13	1.09
PPI-02	10 mM tris	8	82.3	78.1	68.5	99.70	98.60	98.80	0.99	0.97	0.98	1.01	1.04	1.01
PPI-02	10 mM tris	9	83.8	80.1	68.6	99.80	98.40	97.90	0.99	0.96	0.97	1.00	0.99	1.03
PPI-02	10 mM histidine	5	79.0	74.2	63.7	99.90	98.30	97.70	0.99	0.96	0.97	1.26	1.01	1.07
PPI-02	10 mM histidine	5.5	76.8	71.0	65.6	99.80	98.60	98.40	0.99	0.97	0.94	1.00	0.98	1.06
PPI-02	10 mM histidine	6	87.4	83.4	74.5	99.90	98.50	98.70	0.99	0.97	0.99	0.99	0.99	1.04
PPI-02	10 mM histidine	6.5	72.8	69.7	63.6	99.90	99.00	98.80	0.99	0.97	0.98	1.47	1.49	1.51
PPI-02	10 mM histidine	7	69.7	69.0	60.5	99.90	98.80	98.80	0.99	0.97	0.98	1.00	1.00	1.00
PPI-02	10 mM histidine	7.5	84.6	79.8	72.8	99.80	98.50	98.80	0.99	0.97	0.98	0.97	0.99	0.99
PPI-02	10 mM tris	8	33.3	30.9	26.4	99.80	98.70	98.80	0.99	0.97	0.99	0.99	0.99	1.02
PPI-02	10 mM tris	9	78.0	73.3	63.2	99.90	98.40	97.80	0.99	0.97	0.97	0.99	1.00	0.99
PPI-02	10 mM histidine	5	72.5	66.4	65.0	99.80	97.60	94.60	0.99	0.96	0.91	0.96	0.95	1.03
PPI-02	10 mM histidine	5.5	88.9	74.5	68.6	100.00	98.40	98.40	0.99	0.97	0.98	0.98	1.01	1.00
PPI-02	10 mM histidine	6	80.8	74.0	68.2	99.90	98.40	98.70	0.99	0.97	0.99	0.99	0.99	1.01
PPI-02	10 mM histidine	6.5	86.9	81.1	68.9	99.90	98.80	98.80	0.99	0.97	0.99	1.00	1.04	1.01
PPI-02	10 mM histidine	7	81.9	79.3	70.9	100.00	98.80	98.80	0.99	0.97	0.99	0.99	0.98	1.01
PPI-02	10 mM histidine	7.5	78.9	75.2	69.0	99.90	98.60	98.80	0.99	0.97	0.99	1.00	1.00	1.00
PPI-02	10 mM tris	8	71.5	64.1	57.7	99.90	99.00	98.80	0.99	0.97	0.98	0.99	1.01	1.01
PPI-02	10 mM tris	9	81.1	77.1	67.8	99.50	98.50	98.20	0.99	0.97	0.98	0.99	0.99	1.01

Biophysical parameter tables (Part 17 of 22)

Protein	Buffer	pH	$m_{25,rec}$ (%)	$m_{40,rec}$ (%)	$m_{50,rec}$ (%)	M_{25} (%)	M_{40} (%)	M_{50} (%)	LSA_{25}	LSA_{40}	LSA_{50}	$R_{R,25}$	$R_{R,40}$	$R_{R,50}$
PPI-03	10 mM histidine	5	98.4	97.0	85.4	99.20	99.60	98.90	0.99	0.99	0.98	0.98	1.00	1.01
PPI-03	10 mM histidine	5.5	101.2	97.7	86.0	99.10	99.60	99.00	0.99	0.99	0.99	1.00	0.99	1.02
PPI-03	10 mM histidine	6	106.0	98.2	88.9	99.10	99.70	99.20	0.99	0.99	0.98	1.00	1.00	1.02
PPI-03	10 mM histidine	6.5	101.9	98.9	83.7	99.00	99.80	99.20	0.99	0.99	0.98	0.87	0.90	0.90
PPI-03	10 mM histidine	7	103.0	100.3	94.4	98.90	99.70	99.20	0.99	0.99	0.99	0.77	0.78	0.81
PPI-03	10 mM histidine	7.5	83.4	97.1	##	98.40	99.60	##	0.99	0.99	##	1.07	0.74	0.80
PPI-03	10 mM tris	8	85.9	82.8	68.2	98.00	99.50	98.80	0.99	0.98	0.97	**	**	**
PPI-03	10 mM tris	9	100.6	97.2	78.6	98.10	98.50	97.00	0.99	0.95	0.93	0.97	1.01	1.04
PPI-03	10 mM histidine	5	100.4	92.6	80.0	99.20	99.20	97.90	0.99	0.99	0.99	1.02	1.06	1.10
PPI-03	10 mM histidine	5.5	94.1	103.0	85.0	99.20	99.50	98.70	0.99	1.00	0.99	1.03	1.00	1.05
PPI-03	10 mM histidine	6	99.8	90.9	79.1	99.30	99.60	99.20	0.99	0.99	1.00	0.99	0.97	1.02
PPI-03	10 mM histidine	6.5	98.7	102.1	86.7	99.10	99.70	99.20	0.99	1.00	1.00	1.00	1.00	1.02
PPI-03	10 mM histidine	7	84.6	91.6	83.3	99.20	99.60	99.30	0.99	1.00	0.99	1.00	1.00	1.00
PPI-03	10 mM histidine	7.5	92.9	88.7	74.6	99.00	99.60	99.00	0.99	1.00	1.00	1.02	1.03	1.00
PPI-03	10 mM tris	8	96.3	103.5	88.4	98.90	99.50	99.10	0.99	0.99	0.98	0.99	0.99	0.99
PPI-03	10 mM tris	9	96.6	67.2	69.3	98.70	99.10	98.20	0.99	0.97	0.97	1.00	1.00	1.03
PPI-03	10 mM histidine	5	110.3	**	82.9	99.50	**	95.20	0.99	**	0.95	1.00	**	1.12
PPI-03	10 mM histidine	5.5	87.0	98.8	80.2	99.20	99.40	98.40	0.99	0.99	0.99	1.00	1.00	1.02
PPI-03	10 mM histidine	6	96.4	75.9	74.0	99.20	99.60	99.10	0.99	0.99	1.00	0.99	0.99	1.03
PPI-03	10 mM histidine	6.5	103.6	107.5	93.9	99.30	99.70	99.30	0.99	1.00	1.00	1.00	0.99	1.00
PPI-03	10 mM histidine	7	84.5	100.0	82.2	99.00	99.60	99.20	0.99	1.00	1.00	1.00	1.00	1.00
PPI-03	10 mM histidine	7.5	83.3	85.6	76.3	99.20	99.60	99.10	0.99	0.99	0.99	0.99	0.98	1.01
PPI-03	10 mM tris	8	99.3	101.0	86.0	98.90	99.50	99.10	0.99	0.99	0.98	1.00	1.01	1.01
PPI-03	10 mM tris	9	77.6	95.2	81.6	98.80	99.00	97.60	0.96	0.98	0.97	0.96	0.98	0.98
PPI-04	10 mM histidine	5	80.8	60.1	42.2	99.20	99.10	94.80	0.96	0.98	0.95	0.99	1.05	1.19
PPI-04	10 mM histidine	5.5	102.3	79.4	68.7	99.10	99.50	97.20	0.96	1.00	0.98	0.93	0.95	1.17
PPI-04	10 mM histidine	6	88.3	57.7	64.2	99.10	99.00	97.50	0.96	0.95	0.94	0.97	1.10	1.30
PPI-04	10 mM histidine	6.5	87.8	66.8	68.3	99.00	98.60	97.10	0.95	0.98	0.92	0.71	1.11	1.74
PPI-04	10 mM histidine	7	88.9	75.0	57.5	98.90	97.00	94.40	0.95	0.97	0.86	0.92	1.13	1.72
PPI-04	10 mM histidine	7.5	90.8	68.5	53.8	98.40	96.50	93.70	0.94	0.97	0.97	0.96	1.48	**
PPI-04	10 mM tris	8	70.9	47.1	32.9	98.00	94.20	92.80	0.93	0.94	0.82	1.00	1.03	0.97
PPI-04	10 mM tris	9	41.5	16.2	6.9	98.10	95.20	88.30	0.93	0.94	0.76	0.95	0.89	0.85
PPI-04	10 mM histidine	5	113.5	65.1	7.4	99.20	99.00	92.10	0.96	0.99	0.97	0.99	1.45	21.06
PPI-04	10 mM histidine	5.5	100.0	71.4	55.6	99.20	98.90	98.60	0.96	0.98	1.00	0.99	1.29	3.02
PPI-04	10 mM histidine	6	97.2	78.6	65.8	99.30	98.90	98.50	0.96	0.98	1.00	1.42	1.73	2.14
PPI-04	10 mM histidine	6.5	83.5	67.3	48.3	99.10	98.70	98.50	0.96	0.99	0.94	1.04	1.16	1.12
PPI-04	10 mM histidine	7	90.1	66.3	65.4	99.20	98.60	98.50	0.95	0.98	1.00	0.99	1.08	0.70
PPI-04	10 mM histidine	7.5	104.6	82.0	64.5	99.00	98.10	97.40	0.95	0.98	0.99	1.01	1.05	0.98
PPI-04	10 mM tris	8	84.2	62.2	52.4	98.90	97.70	97.40	0.94	0.91	0.92	1.03	1.10	1.02
PPI-04	10 mM tris	9	88.3	66.2	37.2	98.70	93.50	88.80	0.93	0.86	0.74	0.99	1.84	3.06
PPI-04	10 mM histidine	5	80.1	48.9	6.0	99.50	99.20	95.00	0.97	1.00	1.00	1.00	1.91	307.51
PPI-04	10 mM histidine	5.5	80.5	57.1	35.4	99.20	99.30	98.70	0.96	1.00	1.00	1.00	1.54	4.48
PPI-04	10 mM histidine	6	93.3	86.2	64.4	99.20	98.90	98.80	0.96	0.99	1.00	0.99	1.29	1.58
PPI-04	10 mM histidine	6.5	82.2	66.0	59.0	99.30	98.60	98.60	0.96	0.98	1.00	0.97	1.23	1.25
PPI-04	10 mM histidine	7	80.0	64.3	55.3	99.00	98.40	98.40	0.95	0.99	1.00	1.00	1.25	1.15
PPI-04	10 mM histidine	7.5	93.0	70.7	69.1	99.20	97.50	97.50	0.95	0.98	0.95	1.00	1.20	1.15
PPI-04	10 mM tris	8	87.9	71.8	64.1	98.90	97.30	97.70	0.94	0.93	0.93	1.02	1.30	1.21
PPI-04	10 mM tris	9	86.7	91.8	40.3	98.80	93.20	88.60	0.94	0.72	0.47	0.96	1.86	4.22

Biophysical parameter tables (Part 18 of 22)

Protein	Buffer	pH	$m_{25,rec}$ (%)	$m_{40,rec}$ (%)	$m_{50,rec}$ (%)	M_{25} (%)	M_{40} (%)	M_{50} (%)	LSA_{25}	LSA_{40}	LSA_{50}	$R_{R,25}$	$R_{R,40}$	$R_{R,50}$
PPI-08	10 mM histidine	5	72.6	62.8	68.1	100.00	99.50	98.70	0.99	0.99	0.90	0.92	0.93	1.23
PPI-08	10 mM histidine	5.5	48.8	39.3	54.0	100.00	99.60	99.10	1.00	0.99	0.99	0.98	1.03	1.19
PPI-08	10 mM histidine	6	71.4	57.7	77.7	100.00	99.40	99.20	1.00	0.98	0.99	0.89	0.89	0.94
PPI-08	10 mM histidine	6.5	59.6	48.9	67.2	99.90	99.30	99.10	1.00	0.98	0.97	1.00	1.00	1.10
PPI-08	10 mM histidine	7	64.6	48.0	71.4	100.00	99.30	98.90	1.00	0.99	0.98	1.01	1.11	1.15
PPI-08	10 mM histidine	7.5	68.4	62.0	75.9	100.00	99.20	98.70	1.00	0.98	0.95	1.12	1.08	2.17
PPI-08	10 mM tris	8	66.1	66.9	65.3	100.00	99.00	98.40	1.00	0.99	0.91	1.04	1.01	1.63
PPI-08	10 mM tris	9	66.8	55.5	39.6	92.50	94.60	93.50	0.98	0.95	0.91	0.95	1.05	0.96
PPI-08	10 mM histidine	5	76.8	53.9	2.1	100.00	99.00	77.20	1.00	0.99	0.82	0.95	1.51	4.81
PPI-08	10 mM histidine	5.5	71.4	57.2	44.6	100.00	99.20	98.30	1.00	0.99	0.99	0.98	**	5.77
PPI-08	10 mM histidine	6	52.9	43.1	56.6	100.00	99.30	99.30	1.00	0.99	1.00	1.01	1.03	2.66
PPI-08	10 mM histidine	6.5	62.2	49.1	66.2	100.00	99.40	99.30	1.00	0.99	0.99	1.00	1.02	3.71
PPI-08	10 mM histidine	7	62.3	48.7	67.9	100.00	99.20	99.10	1.00	0.99	1.00	1.00	0.67	1.08
PPI-08	10 mM histidine	7.5	61.8	43.8	64.3	100.00	98.90	99.10	1.00	0.99	0.99	1.00	1.00	1.01
PPI-08	10 mM tris	8	68.8	39.9	69.3	100.00	98.90	98.80	1.00	0.96	0.83	1.01	1.01	1.23
PPI-08	10 mM tris	9	56.6	59.9	46.2	100.00	98.30	97.80	1.00	0.98	1.23	0.99	0.99	0.98
PPI-08	10 mM histidine	5	54.4	28.0	##	100.00	98.80	##	1.00	1.00	##	1.00	5.98	6.29
PPI-08	10 mM histidine	5.5	65.7	53.9	##	99.80	99.20	##	1.00	1.00	##	1.00	1.27	43.60
PPI-08	10 mM histidine	6	62.8	49.0	55.0	100.00	99.30	99.20	1.00	0.99	0.99	1.00	1.02	1.17
PPI-08	10 mM histidine	6.5	66.7	51.8	71.1	100.00	99.30	99.20	1.00	1.01	1.00	0.99	1.00	1.04
PPI-08	10 mM histidine	7	60.8	44.7	63.7	100.00	99.00	99.10	1.00	1.01	0.99	1.00	1.01	1.00
PPI-08	10 mM histidine	7.5	66.2	44.9	68.9	100.00	99.00	99.10	1.00	0.97	1.02	0.99	0.99	1.01
PPI-08	10 mM tris	8	38.6	63.6	63.2	100.00	98.90	99.00	0.82	0.98	0.98	0.85	0.85	0.87
PPI-08	10 mM tris	9	61.1	63.5	39.5	100.00	98.40	97.30	1.00	0.99	1.05	0.99	1.02	1.00
PPI-10	10 mM histidine	5	84.7	83.3	67.4	99.10	99.10	97.80	0.92	0.96	0.90	0.97	0.76	**
PPI-10	10 mM histidine	5.5	93.2	77.7	67.4	99.00	98.90	97.80	0.92	0.95	0.87	1.01	**	**
PPI-10	10 mM histidine	6	92.2	85.0	74.7	98.90	99.20	98.30	0.90	0.96	0.91	1.08	1.11	1.11
PPI-10	10 mM histidine	6.5	86.3	80.1	72.2	98.70	98.80	98.30	0.87	0.95	0.92	1.09	1.08	1.28
PPI-10	10 mM histidine	7	84.8	75.1	67.6	98.50	98.90	98.20	0.91	0.95	0.94	1.09	1.19	1.18
PPI-10	10 mM histidine	7.5	84.5	64.9	64.2	98.50	98.00	97.00	0.87	0.93	0.85	1.11	1.48	**
PPI-10	10 mM tris	8	78.2	71.8	61.2	98.30	97.70	97.70	0.89	0.92	0.91	1.02	1.07	0.99
PPI-10	10 mM tris	9	86.0	75.1	64.4	98.10	97.00	95.10	0.89	0.90	0.90	0.99	1.04	1.03
PPI-10	10 mM histidine	5	76.7	70.2	54.9	99.30	98.10	93.20	0.85	0.94	0.78	1.19	0.94	**
PPI-10	10 mM histidine	5.5	84.3	80.2	70.5	99.10	99.00	97.90	0.83	0.88	0.84	1.27	1.29	**
PPI-10	10 mM histidine	6	90.8	84.9	76.1	98.80	99.00	98.30	0.88	0.96	0.86	1.08	1.53	**
PPI-10	10 mM histidine	6.5	85.4	76.5	63.6	98.70	99.00	98.30	0.89	0.95	0.85	1.01	1.36	1.16
PPI-10	10 mM histidine	7	44.4	82.1	75.2	98.50	98.50	98.30	0.80	0.94	0.85	1.05	1.39	**
PPI-10	10 mM histidine	7.5	85.1	79.3	72.8	98.60	98.70	98.00	0.87	0.94	0.86	1.11	0.82	**
PPI-10	10 mM tris	8	88.3	85.4	73.8	98.30	98.60	98.00	0.89	0.93	0.91	0.98	0.96	0.98
PPI-10	10 mM tris	9	69.8	68.6	57.9	98.30	98.10	96.50	0.90	0.93	0.90	0.99	1.01	0.99
PPI-10	10 mM histidine	5	94.5	91.2	70.4	99.20	98.40	92.80	0.88	0.93	0.82	1.00	0.97	1.09
PPI-10	10 mM histidine	5.5	80.5	82.0	71.3	99.10	98.90	97.90	0.88	0.94	0.88	**	**	**
PPI-10	10 mM histidine	6	82.7	85.4	75.3	99.10	99.00	98.30	0.89	0.96	0.89	0.83	1.11	**
PPI-10	10 mM histidine	6.5	83.6	81.5	76.1	98.80	99.10	98.50	0.90	0.95	0.94	1.01	1.44	1.01
PPI-10	10 mM histidine	7	81.1	79.7	71.7	98.60	98.80	98.40	0.90	0.94	0.94	1.07	1.30	1.02
PPI-10	10 mM histidine	7.5	87.8	83.1	77.6	98.60	98.50	98.30	0.90	0.92	0.99	1.02	0.99	0.99
PPI-10	10 mM tris	8	83.2	76.0	70.1	98.60	98.60	97.70	0.92	0.93	0.98	0.98	0.98	0.98
PPI-10	10 mM tris	9	79.8	76.1	65.5	98.20	97.50	96.50	0.89	0.91	0.89	0.99	0.98	0.98

Biophysical parameter tables (Part 19 of 22)

Protein	Buffer	pH	$m_{25,rec}$ (%)	$m_{40,rec}$ (%)	$m_{50,rec}$ (%)	M_{25} (%)	M_{40} (%)	M_{50} (%)	LSA_{25}	LSA_{40}	LSA_{50}	$R_{R,25}$	$R_{R,40}$	$R_{R,50}$
PPI-13	10 mM histidine	5	73.8	102.8	87.2	98.60	98.60	98.40	0.92	0.95	0.96	1.00	1.04	1.00
PPI-13	10 mM histidine	5.5	80.7	100.5	85.0	98.80	98.80	98.60	0.94	0.96	0.98	1.03	1.06	1.03
PPI-13	10 mM histidine	6	65.6	86.8	86.1	98.70	98.70	98.50	0.94	0.96	0.98	0.99	0.96	0.99
PPI-13	10 mM histidine	6.5	68.5	95.2	89.6	98.70	98.80	98.90	0.95	0.96	0.98	0.99	0.96	0.99
PPI-13	10 mM histidine	7	76.1	86.6	94.1	98.80	98.90	98.70	0.95	0.96	0.98	1.17	0.89	0.92
PPI-13	10 mM histidine	7.5	65.6	87.7	103.1	98.80	98.90	98.80	0.94	0.96	0.98	1.05	0.95	0.97
PPI-13	10 mM tris	8	78.1	94.9	99.4	98.60	98.50	98.30	0.95	0.95	0.97	0.98	0.97	0.98
PPI-13	10 mM tris	9	60.4	78.4	74.5	98.80	98.10	97.60	0.95	0.94	0.96	0.97	0.98	0.97
PPI-13	10 mM histidine	5	98.9	115.0	87.8	98.60	97.70	98.00	0.96	0.94	0.95	0.68	0.98	1.07
PPI-13	10 mM histidine	5.5	66.0	74.7	79.2	98.60	98.70	98.70	0.96	0.95	0.97	0.96	0.99	1.00
PPI-13	10 mM histidine	6	61.2	83.2	89.4	98.80	98.80	98.80	0.96	0.96	0.98	0.99	0.98	1.00
PPI-13	10 mM histidine	6.5	68.1	83.4	86.8	98.70	98.90	98.80	0.96	0.96	0.98	1.04	0.98	1.00
PPI-13	10 mM histidine	7	94.0	83.0	98.8	98.70	98.80	98.80	0.96	0.96	0.89	1.01	0.99	1.01
PPI-13	10 mM histidine	7.5	63.5	76.6	79.7	98.70	98.80	98.80	0.96	0.96	0.99	0.97	0.95	0.97
PPI-13	10 mM tris	8	71.0	83.0	90.0	98.60	98.60	97.00	0.94	0.95	0.96	0.97	0.96	0.96
PPI-13	10 mM tris	9	77.2	85.6	##	98.70	98.00	##	0.95	0.94	##	0.93	0.91	1.02
PPI-13	10 mM histidine	5	74.8	82.4	81.2	98.70	98.70	92.50	0.96	0.96	0.67	1.22	0.99	1.07
PPI-13	10 mM histidine	5.5	69.8	87.0	##	98.70	98.80	##	0.96	0.96	##	8.20	0.93	2.82
PPI-13	10 mM histidine	6	79.1	96.1	103.0	98.80	98.90	98.80	0.96	0.96	0.98	0.99	0.99	1.00
PPI-13	10 mM histidine	6.5	70.4	84.3	100.4	98.70	98.80	98.90	0.96	0.96	0.98	1.10	1.00	1.02
PPI-13	10 mM histidine	7	71.9	85.0	89.3	98.60	98.90	98.70	0.96	0.96	0.97	1.16	0.96	0.97
PPI-13	10 mM histidine	7.5	80.4	97.6	97.7	98.60	98.90	98.90	0.96	0.96	0.98	0.86	0.98	0.99
PPI-13	10 mM tris	8	75.3	91.9	96.4	98.60	98.70	98.60	0.95	0.95	0.98	0.78	0.87	0.88
PPI-13	10 mM tris	9	70.2	84.4	90.9	98.60	98.20	97.60	0.95	0.95	0.97	0.79	0.96	0.98
PPI-17	10 mM histidine	5	73.3	78.2	82.8	98.00	98.10	98.30	0.79	0.80	0.80	0.96	0.97	0.99
PPI-17	10 mM histidine	5.5	88.8	97.0	97.2	98.10	98.20	98.40	0.79	0.80	0.80	0.93	0.95	0.97
PPI-17	10 mM histidine	6	79.9	89.0	90.3	98.10	98.10	98.40	0.79	0.81	0.82	0.92	0.87	0.90
PPI-17	10 mM histidine	6.5	82.1	88.1	87.3	98.10	98.10	98.50	0.81	0.82	0.82	0.95	0.98	0.97
PPI-17	10 mM histidine	7	78.9	89.7	91.0	98.10	97.90	98.20	0.82	0.83	0.83	1.06	0.99	1.05
PPI-17	10 mM histidine	7.5	79.3	84.4	82.1	98.40	97.80	97.70	0.79	0.82	0.82	2.79	0.98	0.94
PPI-17	10 mM tris	8	76.7	81.0	79.3	98.00	97.60	97.70	0.82	0.81	0.80	0.96	0.98	1.00
PPI-17	10 mM tris	9	79.4	87.7	##	98.30	97.20	##	0.81	0.82	##	1.36	0.95	1.57
PPI-17	10 mM histidine	5	85.9	90.4	78.8	98.00	97.80	80.40	0.80	0.78	0.44	0.93	0.96	1.29
PPI-17	10 mM histidine	5.5	77.2	85.5	83.8	98.10	97.90	97.70	0.80	0.80	0.75	0.75	0.74	0.77
PPI-17	10 mM histidine	6	80.4	91.5	78.4	98.10	97.90	98.00	0.80	0.80	0.77	0.95	0.97	0.97
PPI-17	10 mM histidine	6.5	85.4	92.2	94.0	98.20	97.90	98.00	0.80	0.80	0.73	0.82	0.75	0.77
PPI-17	10 mM histidine	7	82.6	91.1	91.3	98.10	97.80	97.60	0.80	0.80	0.78	0.85	0.86	0.87
PPI-17	10 mM histidine	7.5	81.0	89.5	91.5	98.10	97.70	97.50	0.80	0.80	0.71	1.18	0.89	0.92
PPI-17	10 mM tris	8	81.7	90.0	92.0	98.00	97.50	97.50	0.79	0.80	0.78	0.96	0.96	0.99
PPI-17	10 mM tris	9	80.3	85.3	84.5	97.80	97.00	95.00	0.81	0.80	0.77	0.90	0.91	1.24
PPI-17	10 mM histidine	5	81.6	86.8	53.3	98.00	97.50	63.50	0.80	0.78	0.16	0.90	0.94	1.69
PPI-17	10 mM histidine	5.5	80.3	84.3	83.7	98.10	97.80	96.50	0.80	0.81	0.87	0.90	0.93	0.98
PPI-17	10 mM histidine	6	76.6	88.3	86.5	98.10	97.90	97.80	0.80	0.80	0.78	0.91	0.91	0.94
PPI-17	10 mM histidine	6.5	85.6	96.3	88.6	98.10	97.90	97.90	0.80	0.80	0.79	1.00	0.94	0.96
PPI-17	10 mM histidine	7	82.9	94.7	92.7	98.10	97.90	97.60	0.80	0.80	0.69	0.95	0.96	0.99
PPI-17	10 mM histidine	7.5	73.4	83.3	83.7	98.10	97.70	97.20	0.79	0.80	0.74	0.93	1.00	0.99
PPI-17	10 mM tris	8	84.2	92.4	86.7	98.00	97.40	96.90	0.79	0.80	0.78	0.83	0.71	0.75
PPI-17	10 mM tris	9	85.7	90.4	85.2	97.90	96.90	92.30	0.80	0.79	0.77	2.10	0.62	1.37

Biophysical parameter tables (Part 20 of 22)

Protein	Buffer	pH	$m_{25,rec}$ (%)	$m_{40,rec}$ (%)	$m_{50,rec}$ (%)	M_{25} (%)	M_{40} (%)	M_{50} (%)	LSA_{25}	LSA_{40}	LSA_{50}	$R_{R,25}$	$R_{R,40}$	$R_{R,50}$
PPI-18	10 mM histidine	5	90.0	0.0	##	88.5	7	##	0.80	0.28	##	2.96	2.09	4.08
PPI-18	10 mM histidine	5.5	66.3	0.5	##	96.9	6.3	##	0.96	0.14	##	1.21	140.53	97.36
PPI-18	10 mM histidine	6	63.6	0.5	##	97.7	0.4	##	0.96	0.11	##	1.01	7.85	12.88
PPI-18	10 mM histidine	6.5	63.6	0.4	##	97.8	0.4	##	0.96	0.03	##	1.11	3.10	13.46
PPI-18	10 mM histidine	7	62.8	0.7	##	98.1	1	##	0.97	0.02	##	1.52	1.51	3.80
PPI-18	10 mM histidine	7.5	78.3	1.4	##	98	1.2	##	0.96	0.01	##	1.29	0.91	1.66
PPI-18	10 mM tris	8	137.5	0.0	##	97.8	0	##	0.96	0.00	##	1.17	1.21	2.43
PPI-18	10 mM tris	9	59.7	0.0	##	97.6	0	##	0.96	0.00	##	1.62	73.08	1.92
PPI-18	10 mM histidine	5	23.0	0.7	##	75.3	9.6	##	0.90	0.15	##	2.94	44.33	155.51
PPI-18	10 mM histidine	5.5	79.9	0.0	##	93.6	0	##	0.86	0.00	##	1.74	46.11	0.06
PPI-18	10 mM histidine	6	52.2	0.0	##	98.1	0	##	0.96	0.00	##	0.83	3.44	43.36
PPI-18	10 mM histidine	6.5	71.1	0.0	##	98.1	0	##	0.96	0.00	##	0.94	5.99	3.38
PPI-18	10 mM histidine	7	71.9	0.6	##	98.2	0.7	##	0.96	0.01	##	0.71	1.37	2.92
PPI-18	10 mM histidine	7.5	79.8	0.0	##	98	0	##	0.96	0.00	##	0.97	1.74	4.50
PPI-18	10 mM tris	8	105.3	0.0	##	97.6	0	##	0.96	0.00	##	1.06	9.82	5.79
PPI-18	10 mM tris	9	68.2	12.5	##	96.8	15.7	##	0.95	0.08	##	1.03	2.28	3.65
PPI-18	10 mM histidine	5	19.1	0.7	##	72.5	44.2	##	0.89	0.02	##	0.42	0.07	192.70
PPI-18	10 mM histidine	5.5	80.6	0.3	##	88.3	2.6	##	0.68	0.00	##	0.54	54.10	176.22
PPI-18	10 mM histidine	6	76.2	0.0	##	98.1	0	##	0.96	0.00	##	0.72	5.45	50.48
PPI-18	10 mM histidine	6.5	71.8	1.0	##	98.1	0.9	##	0.96	0.02	##	0.73	2.32	8.83
PPI-18	10 mM histidine	7	62.3	67.3	##	98.2	71.8	##	0.97	0.24	##	0.64	1.19	5.60
PPI-18	10 mM histidine	7.5	65.6	41.3	##	98	54.4	##	0.97	0.24	##	0.68	1.98	4.88
PPI-18	10 mM tris	8	99.0	0.0	##	97.8	0	##	0.96	0.00	##	0.77	2.79	5.75
PPI-18	10 mM tris	9	60	0.0	##	96.8	0	##	0.95	0.00	##	0.69	2.15	4.40
PPI-30	10 mM histidine	5	9.8	34.6	##	100.00	99.70	##	0.44	1.02	##	1.73	1.04	**
PPI-30	10 mM histidine	5.5	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^
PPI-30	10 mM histidine	6	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^
PPI-30	10 mM histidine	6.5	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^
PPI-30	10 mM histidine	7	60.1	80.9	39.3	99.80	99.00	99.80	0.56	0.43	0.31	1.21	1.96	41.79
PPI-30	10 mM histidine	7.5	107.9	89.5	36.2	99.20	99.40	94.50	0.69	0.42	0.40	0.92	1.39	106.97
PPI-30	10 mM tris	8	75.8	81.2	33.5	100.00	99.40	95.20	0.86	0.80	0.40	0.00	**	**
PPI-30	10 mM tris	9	95.5	105.5	30.4	100.00	94.30	71.90	0.89	0.74	0.00	1.02	1.10	9.67
PPI-30	10 mM histidine	5	91.1	90.4	##	100.00	99.40	##	1.01	0.99	##	1.00	1.04	290.40
PPI-30	10 mM histidine	5.5	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^
PPI-30	10 mM histidine	6	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^
PPI-30	10 mM histidine	6.5	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^
PPI-30	10 mM histidine	7	78.6	41.6	11.5	99.40	98.10	100.00	1.00	0.55	0.00	0.99	1.00	2.76
PPI-30	10 mM histidine	7.5	80.5	78.9	4.8	100.00	100.20	15.80	1.00	0.88	0.10	0.90	1.19	9.00
PPI-30	10 mM tris	8	83.1	81.4	52.1	100.00	93.60	93.10	1.00	0.88	0.53	0.77	1.92	5.07
PPI-30	10 mM tris	9	48.2	75.8	15.9	100.80	95.60	100.00	0.70	0.78	0.23	0.98	1.76	77.03
PPI-30	10 mM histidine	5	117.4	74.6	##	99.80	100.60	##	1.01	0.99	##	1.01	1.04	**
PPI-30	10 mM histidine	5.5	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^
PPI-30	10 mM histidine	6	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^
PPI-30	10 mM histidine	6.5	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^	^^^
PPI-30	10 mM histidine	7	31.3	78.2	19.4	94.50	98.80	100.00	0.44	0.98	0.19	1.01	1.07	14.70
PPI-30	10 mM histidine	7.5	102.3	105.7	65.4	99.70	98.50	100.00	1.00	0.96	0.72	1.03	1.07	6.57
PPI-30	10 mM tris	8	114.5	90.0	65.0	99.20	96.80	100.00	0.79	0.84	NaN	1.08	3.92	15.65
PPI-30	10 mM tris	9	63.6	80.2	21.9	100.00	96.80	100.00	0.79	0.80	0.31	1.06	3.47	38.51

Biophysical parameter tables (Part 21 of 22)

Protein	Buffer	pH	$m_{25,rec}$ (%)	$m_{40,rec}$ (%)	$m_{50,rec}$ (%)	M_{25} (%)	M_{40} (%)	M_{50} (%)	LSA_{25}	LSA_{40}	LSA_{50}	$R_{R,25}$	$R_{R,40}$	$R_{R,50}$
PPI-44	10 mM histidine	5	60.4	70.3	64.7	84.30	83.70	87.60	0.68	0.77	0.81	1.01	0.99	1.02
PPI-44	10 mM histidine	5.5	94.6	95.3	90.0	84.90	83.20	86.90	0.69	0.75	0.62	0.87	0.86	1.17
PPI-44	10 mM histidine	6	82.4	87.2	90.0	85.10	83.00	87.40	0.69	0.74	0.61	**	1.00	1.60
PPI-44	10 mM histidine	6.5	80.5	84.7	85.0	85.00	83.10	87.80	0.69	0.66	0.66	0.90	1.36	1.06
PPI-44	10 mM histidine	7	92.4	89.5	127.1	84.90	82.90	88.10	0.68	0.66	0.46	0.88	0.93	21.50
PPI-44	10 mM histidine	7.5	115.5	30.0	30.0	84.40	83.30	88.20	0.67	0.70	0.68	**	0.87	1.38
PPI-44	10 mM tris	8	142.1	30.0	30.0	85.00	83.80	88.60	0.69	0.69	0.66	0.91	0.99	1.40
PPI-44	10 mM tris	9	88.3	30.0	30.0	86.00	86.80	91.90	0.71	0.79	0.82	1.03	1.02	0.94
PPI-44	10 mM histidine	5	36.4	36.4	32.1	83.80	82.20	86.20	0.69	0.77	0.78	**	0.68	**
PPI-44	10 mM histidine	5.5	66.3	75.6	88.9	84.30	82.50	87.40	0.69	0.67	0.72	1.05	1.42	1.08
PPI-44	10 mM histidine	6	93.6	91.5	90.0	85.00	83.20	87.60	0.69	0.67	0.73	1.15	1.10	1.00
PPI-44	10 mM histidine	6.5	91.6	100.0	90.0	84.70	83.30	88.00	0.68	0.72	0.72	**	1.06	0.99
PPI-44	10 mM histidine	7	80.9	85.6	80.0	85.20	83.50	88.50	0.69	0.74	0.68	0.89	0.84	**
PPI-44	10 mM histidine	7.5	84.2	89.6	80.0	85.30	84.70	89.90	0.70	0.75	0.74	0.78	0.76	1.09
PPI-44	10 mM tris	8	110.5	30.0	30.0	85.30	84.90	89.80	0.70	0.75	0.69	**	1.05	1.82
PPI-44	10 mM tris	9	85.7	30.0	30.0	86.50	88.40	91.70	0.72	0.79	0.79	0.91	0.95	1.10
PPI-44	10 mM histidine	5	30.2	31.6	21.2	86.30	84.80	90.40	0.72	0.83	0.88	0.95	1.20	1.20
PPI-44	10 mM histidine	5.5	100.2	95.7	103.2	85.10	83.70	87.60	0.70	0.76	0.75	**	1.09	1.06
PPI-44	10 mM histidine	6	86.0	104.9	125.3	85.30	83.40	88.20	0.69	0.68	0.73	1.00	1.10	0.96
PPI-44	10 mM histidine	6.5	74.6	100.4	121.9	85.20	83.50	88.20	0.70	0.74	0.73	0.97	0.92	0.97
PPI-44	10 mM histidine	7	86.8	97.0	90.0	85.20	83.90	89.10	0.69	0.74	0.71	**	0.98	**
PPI-44	10 mM histidine	7.5	84.0	96.0	90.0	84.90	85.00	90.10	0.69	0.76	0.75	0.70	0.64	0.94
PPI-44	10 mM tris	8	102.5	30.0	20.0	85.50	85.30	90.20	0.70	0.74	0.72	0.97	0.93	1.66
PPI-44	10 mM tris	9	81.5	30.0	20.0	87.30	89.10	91.70	0.74	0.82	0.80	**	0.90	0.90
PPI-45	10 mM histidine	5	52.1	30.0	##	100.00	**	##	0.81	**	##	0.99	**	20.36
PPI-45	10 mM histidine	5.5	99.0	73.9	##	100.00	100.00	##	0.66	0.81	##	0.93	**	28.21
PPI-45	10 mM histidine	6	99.4	88.5	##	100.00	100.00	##	0.76	0.55	##	0.34	0.14	14.07
PPI-45	10 mM histidine	6.5	86.6	93.9	##	100.00	100.00	##	0.76	0.58	##	0.19	0.32	40.44
PPI-45	10 mM histidine	7	99.0	95.7	6.2	100.00	100.00	50.90	0.80	0.55	0.02	0.18	0.18	46.56
PPI-45	10 mM histidine	7.5	98.6	95.3	24.8	100.00	100.50	57.50	0.81	0.55	0.02	0.19	0.62	1.76
PPI-45	10 mM tris	8	98.1	75.6	1.7	100.00	99.50	16.70	0.56	0.10	0.02	1.07	2.70	200.77
PPI-45	10 mM tris	9	96.2	91.4	12.1	100.00	100.00	16.70	0.39	0.08	0.00	0.52	0.24	0.81
PPI-45	10 mM histidine	5	98.8	##	##	100.00	##	##	0.88	##	##	0.77	0.48	40.80
PPI-45	10 mM histidine	5.5	99.7	96.0	##	100.00	100.00	##	0.72	0.96	##	0.85	0.51	42.20
PPI-45	10 mM histidine	6	101.1	90.2	##	100.00	100.00	##	0.81	0.89	##	0.40	0.26	14.11
PPI-45	10 mM histidine	6.5	99.9	93.7	##	100.00	100.00	##	0.74	0.90	##	0.37	0.29	24.19
PPI-45	10 mM histidine	7	99.4	##	##	100.00	##	##	0.73	##	##	1.08	0.56	44.73
PPI-45	10 mM histidine	7.5	94.6	88.4	1.5	100.00	100.00	100.00	0.65	0.71	0.21	0.27	0.37	70.81
PPI-45	10 mM tris	8	97.1	85.6	0.2	100.00	99.30	100.00	0.79	0.06	1.00	1.68	6.48	174.90
PPI-45	10 mM tris	9	96.6	71.3	##	100.00	88.90	##	0.32	0.01	##	0.70	6.34	15.99
PPI-45	10 mM histidine	5	100.8	##	##	100.00	##	##	0.67	NaN	##	0.74	0.48	47.55
PPI-45	10 mM histidine	5.5	92.6	82.6	##	100.00	100.00	##	0.77	0.91	##	0.74	0.43	28.07
PPI-45	10 mM histidine	6	98.6	90.9	##	100.00	100.00	##	0.76	0.84	##	0.65	0.39	37.94
PPI-45	10 mM histidine	6.5	102.7	101.3	##	100.00	100.00	##	0.73	0.86	##	0.29	0.27	21.23
PPI-45	10 mM histidine	7	100.1	94.5	##	100.00	100.00	##	0.67	0.87	##	0.91	0.66	38.72
PPI-45	10 mM histidine	7.5	98.8	89.5	##	100.00	100.00	##	0.63	0.82	##	1.00	0.59	42.43
PPI-45	10 mM tris	8	98.5	90.0	##	100.00	100.00	##	0.45	0.42	##	0.93	**	272.16
PPI-45	10 mM tris	9	96.8	74.8	##	100.00	99.20	##	0.40	0.04	##	0.56	27.97	157.13

Biophysical parameter tables (Part 22 of 22)

Protein	Buffer	pH	$m_{25,rec}$ (%)	$m_{40,rec}$ (%)	$m_{50,rec}$ (%)	M_{25} (%)	M_{40} (%)	M_{50} (%)	LSA_{25}	LSA_{40}	LSA_{50}	$R_{R,25}$	$R_{R,40}$	$R_{R,50}$
PPI-46	10 mM histidine	5	97.4	86.3	##	97.70	98.80	##	0.60	0.97	##	1.10	1.07	95.61
PPI-46	10 mM histidine	5.5	99.7	93.7	##	97.50	98.00	##	0.67	0.72	##	1.12	1.05	151.16
PPI-46	10 mM histidine	6	98.4	101.7	87.5	97.50	97.90	98.80	0.72	0.48	0.69	1.03	22.26	337.40
PPI-46	10 mM histidine	6.5	0.6	97.2	90.0	97.50	97.20	95.40	0.01	0.72	0.06	1.04	1.03	9.07
PPI-46	10 mM histidine	7	101.7	102.7	98.4	97.50	97.60	91.50	0.72	0.81	0.37	1.10	1.10	3.46
PPI-46	10 mM histidine	7.5	99.1	103.1	97.3	97.50	97.60	95.60	0.74	0.81	0.60	1.12	1.04	1.48
PPI-46	10 mM tris	8	98.1	99.8	93.9	97.50	97.80	97.70	0.75	0.81	0.78	0.99	0.85	1.07
PPI-46	10 mM tris	9	101.7	97.5	92.8	97.60	97.60	99.30	0.76	0.95	0.79	1.06	1.04	**
PPI-46	10 mM histidine	5	103.4	86.9	##	97.60	98.30	##	0.77	0.97	##	1.01	1.02	109.64
PPI-46	10 mM histidine	5.5	99.8	96.9	##	97.50	97.70	##	0.72	0.95	##	0.95	0.94	64.14
PPI-46	10 mM histidine	6	100.1	99.8	97.0	97.50	97.70	98.10	0.73	0.79	0.83	2.01	1.04	37.42
PPI-46	10 mM histidine	6.5	94.3	102.5	104.0	97.50	97.40	97.90	0.70	0.73	0.75	0.82	0.85	56.81
PPI-46	10 mM histidine	7	101.4	101.6	100.0	97.50	97.40	94.80	0.74	0.81	0.08	0.95	1.14	7.88
PPI-46	10 mM histidine	7.5	99.8	100.0	99.7	97.50	97.50	93.40	0.72	0.79	0.18	1.00	1.04	4.83
PPI-46	10 mM tris	8	95.4	100.6	103.3	97.40	96.70	91.90	0.75	0.87	0.25	0.95	9.84	1.67
PPI-46	10 mM tris	9	145.5	100.0	100.0	97.50	96.10	91.20	0.74	0.77	0.38	1.01	1.17	**
PPI-46	10 mM histidine	5	102.5	99.9	##	97.50	97.70	##	0.49	0.53	##	0.95	1.12	75.89
PPI-46	10 mM histidine	5.5	101.8	97.0	##	97.50	97.70	##	0.73	0.79	##	0.96	1.01	53.59
PPI-46	10 mM histidine	6	104.5	100.0	100.1	97.50	97.60	98.00	0.73	0.78	0.81	1.17	0.89	3.94
PPI-46	10 mM histidine	6.5	102.5	100.3	113.2	97.50	97.50	98.10	0.75	0.78	0.78	0.88	1.05	1.14
PPI-46	10 mM histidine	7	100.8	107.2	102.3	97.50	97.50	96.50	0.74	0.84	0.14	0.90	1.04	14.77
PPI-46	10 mM histidine	7.5	114.2	100.0	100.0	97.50	97.50	93.80	0.74	0.85	0.11	1.17	1.04	7.38
PPI-46	10 mM tris	8	100.4	98.2	103.9	97.50	96.70	92.40	0.75	0.79	0.20	0.90	1.03	**
PPI-46	10 mM tris	9	100.3	97.0	105.4	97.60	95.30	90.80	0.76	0.68	0.31	**	1.14	**
PPI-49	10 mM histidine	5	74.8	81.4	61.0	98.80	97.70	76.10	0.91	0.79	0.36	0.98	1.03	1.64
PPI-49	10 mM histidine	5.5	79.7	84.5	84.2	98.20	98.70	96.40	0.93	0.94	0.79	0.96	0.97	1.17
PPI-49	10 mM histidine	6	75.3	79.5	84.6	98.40	98.60	98.40	0.92	0.92	0.98	0.76	0.81	0.78
PPI-49	10 mM histidine	6.5	41.1	76.4	79.4	98.50	98.60	98.60	0.73	0.90	0.98	0.96	0.87	0.99
PPI-49	10 mM histidine	7	72.0	79.0	80.1	98.80	98.60	98.70	0.91	0.90	0.98	1.01	0.99	0.98
PPI-49	10 mM histidine	7.5	72.9	80.3	81.5	98.60	98.70	98.80	0.91	0.91	0.99	1.01	1.05	0.99
PPI-49	10 mM tris	8	86.1	82.9	88.7	98.60	98.80	98.70	0.92	0.84	0.92	0.98	0.72	1.03
PPI-49	10 mM tris	9	75.3	71.3	78.1	98.60	98.90	98.80	0.91	0.86	0.94	0.90	0.99	0.98
PPI-49	10 mM histidine	5	59.0	62.6	39.3	97.20	96.80	53.70	0.94	0.89	0.05	0.92	0.95	5.82
PPI-49	10 mM histidine	5.5	71.7	77.5	76.2	98.20	98.10	97.40	0.92	0.94	0.89	0.97	1.05	1.03
PPI-49	10 mM histidine	6	77.7	84.8	88.1	98.80	98.60	98.70	0.93	0.93	0.98	0.97	0.99	10.39
PPI-49	10 mM histidine	6.5	72.2	78.9	81.7	98.50	98.60	98.70	0.96	0.93	0.92	0.86	0.89	0.87
PPI-49	10 mM histidine	7	85.1	86.3	90.0	98.30	98.60	98.80	0.92	0.94	0.98	0.98	1.03	0.98
PPI-49	10 mM histidine	7.5	75.2	82.7	85.8	98.60	98.70	98.80	0.92	0.92	0.99	1.30	0.72	1.01
PPI-49	10 mM tris	8	81.2	88.3	88.8	98.10	98.30	98.40	0.91	0.85	0.92	0.95	0.78	0.96
PPI-49	10 mM tris	9	79.1	85.9	90.2	98.20	98.10	97.90	0.92	0.91	0.92	0.95	0.96	0.99
PPI-49	10 mM histidine	5	74.4	80.4	46.6	98.30	98.40	76.60	0.93	0.91	0.03	0.83	8.03	25.23
PPI-49	10 mM histidine	5.5	69.9	74.3	75.0	98.60	98.60	98.40	0.93	0.94	0.90	0.95	0.99	0.98
PPI-49	10 mM histidine	6	81.5	90.1	89.3	98.30	98.70	98.70	0.92	0.94	0.98	0.94	14.97	0.94
PPI-49	10 mM histidine	6.5	77.3	83.4	85.3	98.40	98.50	98.70	0.93	0.93	0.98	0.97	8.78	0.97
PPI-49	10 mM histidine	7	74.8	81.4	85.8	98.60	98.60	98.80	0.93	0.93	0.98	1.01	0.72	1.01
PPI-49	10 mM histidine	7.5	81.7	87.3	89.6	98.30	98.60	98.70	0.93	0.89	0.98	1.00	1.07	1.02
PPI-49	10 mM tris	8	77.9	85.6	94.4	98.60	98.60	98.70	0.92	0.89	0.92	0.97	0.97	1.01
PPI-49	10 mM tris	9	81.3	62.5	90.2	98.50	98.70	98.80	0.92	0.75	0.93	0.92	**	1.00

3 Neural network constants to predict monomer retention of therapeutic proteins after long term storage

The constants of the trained networks, cross-validated by random holdout and presented in Chapter V, are listed in the following. The algorithms were coded in MaTLab2018a.

3.1 Network constants of model A1 (F4, 4°C):

```
% Input 1
x1_step1.xoffset = [0;0;6.26045452316477e-05;0.528539388524551];
x1_step1.gain = [2;2;2.00012521692961;4.349472238051 72];
x1_step1.ymin = -1;

% Layer 1
b1 = [0.39607276174183170259;0.069832958543102174875];
IW1_1 = [0.054787589085189290772 -0.12042662105899816505 -
0.86412089177281237173 -0.023444560190131236849;-0.15709513045521628083
0.1791570368552709569 1.9379742155004295867 -1.2217577740399869768];

% Layer 2
b2 = 0.67426946728449099755;
LW2_1 = [-2.5360709220595802904 -1.4723612317324712429];

% Output 1
y1_step1.ymin = -1;
y1_step1.gain = 6.03314122048985;
y1_step1.xoffset = 0.668497731628166;
```

3.2 Network constants of model A2 (F4, 4-25°C):

```
% Input 1
x1_step1.xoffset = [0;0;6.26045452316477e-05;0.528539388524551];
x1_step1.gain = [2;2;2.00012521692961;4.349472238051 72];
x1_step1.ymin = -1;

% Layer 1
b1 = [0.85152986843293176911;-0.25524368507062633604];
IW1_1 = [1.4743080093768090588 -0.2940277278565914254 0.031206545797845947759
-0.11861344619187222382;0.40555147492113685637 -0.28700432140874782005
0.20456067146874123974 -1.1068203174392445831];

% Layer 2
b2 = [-0.53354832751902425692;-0.80071332697122044042];
LW2_1 = [0.49319977233629452851 -
0.98297374637185896962;0.47019185209484098964 -1.5393358670505212782];

% Output 1
y1_step1.ymin = -1;
y1_step1.gain = [4.33570182775714;6.03314122048985];
y1_step1.xoffset = [0.538713666332862;0.668497731628166];
```

3.3 Network constants of model A3 (F4, 4-25-40°C):

```
% Input 1
x1_step1.xoffset = [0;0;6.26045452316477e-05;0.528539388524551];
x1_step1.gain = [2;2;2.00012521692961;4.34947223805172];
x1_step1.ymin = -1;

% Layer 1
b1 = [-0.31029856834989372194;0.76982633197127103841];
IW1_1 = [0.54404922496028662593 -0.86845005290765930894 -
0.49642368338054460031 2.5081753874866263843;-0.033641901691432676835
0.27720939466867877687 0.44063306735669627301 -1.3468487764394425632];

% Layer 2
b2 = [-0.41984844539812471442;-0.032993606706157134478;-
1.2737377318650751068];
LW2_1 = [0.9638667187140489423 0.93912771017501128767;0.75365431795556214745
0.21322517036493496279;1.266486106656306454 2.5013054792766702406];

% Output 1
y1_step1.ymin = -1;
y1_step1.gain = [4.33570182775714;6.03314122048985;2.000765738785O7];
y1_step1.xoffset =
[0.538713666332862;0.668497731628166;0.000382722859664607];
```

3.4 Network constants of model B1 (F8, 4°C):

```
% Input 1
x1_step1.xoffset = [0;0;-0.055930255733897;0;0;0;0;0];
x1_step1.gain =
[2;1.97773680791308;0.837709383383849;2;2;2;2;2.02354333654555];
x1_step1.ymin = -1;

% Layer 1
b1 =
[0.8779997622410470326;0.10753112403923095752;0.72929536021262109102;0.158050
91849032293361];
IW1_1 = [-0.17326065554008712732 0.66468983461611630403
0.32595155596861163927 -0.36137342811901923545 -0.042703439202685322673 -
0.12374793018432769798 -0.3854442434007692708Z -
1.376151006477973171;0.48814174494277512428 -0.628145959329965528l4
0.30482271140650085695 0.16425750645652179016 0.57771024463333131571 -
0.60250586389755833583 1.1863527189703910736 -
0.67995786710895378402;0.072566515494120728547 -0.005268751368928266745G -
0.93631930905823168576 1.1539023858665489186 -0.22651903446358676031
0.48696296959063539722 0.44694277884513966548 -2.4646930827591799762;-
0.34360307669374717809 -1.2893058020283305609 0.35791055369090751803 -
0.7338328532590642217 -0.16031149640635158904 0.062142584158095870439 -
0.57656272646510597912 3.0102907296633789613];

% Layer 2
b2 = 0.75360231365168361961;
LW2_1 = [-1.7342886448572534253 -1.0839900956883172789 -1.403105782608680796
-1.7783634984817375813];
% Output 1
```

```
y1_step1.ymin = -1;
y1_step1.gain = 6.03314122048985;
y1_step1.xoffset = 0.668497731628166;
```

3.5 Network constants of model B2 (F8, 4-25°C):

```
% Input 1
x1_step1.xoffset = [0;0;-0.055930255733897;0;0;0;0;0];
x1_step1.gain =
[2;1.97773680791308;0.837709383383849;2;2;2;2;2.02354333654555];
x1_step1.ymin = -1;

% Layer 1
b1 =
[1.5009845240882822015;1.2862442384726000544;0.013942400782822867919;0.190654
39568694014794];
IW1_1 = [-0.33170518133754711698 -1.3216132016790063108 1.0196929560781884394
-0.35848364752385786236 -0.40396642151839945445 0.19955983755454656525
2.2901855611214654296 -0.79212838080443814803;-0.047373128779114447029 -
0.058451875819191485839 -0.2211035611823531366 0.093867836371450671 94 -
0.024347942117175266907 -0.080163488893689946457 -0.35635980392357297086 -
1.5065724845955823508;0.2333158921606397 7621 -0.42968165287315113199 -
1.3438292988631241887 0.078075177955313296296 -0.70808704455796267485
0.41732452241238165991 -0.25877976696381332911 -0.43184035435404971581;-
0.076827353061130546075 -0.31886648785949528451 -0.46766852837849298696 -
0.082861684731440593543 -0.59230282950113366525 0.31979437785681585904
1.2691109715603916541 -0.71410263252733252592];

% Layer 2
b2 = [0.5755932112000898826;0.58129103813642424292];
LW2_1 = [-2.0467370355969158879 -0.11896333240695314282 -
1.6564400304863364788 2.6716730662630170379;-1.0346482905317837542 -
1.4234301168013716588 -0.25416988415433788839 0.81132432669165677197];

% Output 1
y1_step1.ymin = -1;
y1_step1.gain = [4.33570182775714;6.03314122048985];
y1_step1.xoffset = [0.538713666332862;0.668497731628166];
```

3.6 Network constants of model B3 (F8, 4-25-40°C):

```
% Input 1
x1_step1.xoffset = [0;0;-0.055930255733897;0;0;0;0;0];
x1_step1.gain =
[2;1.97773680791308;0.837709383383849;2;2;2;2;2.02354333654555];
x1_step1.ymin = -1;

% Layer 1
b1 = [-1.6678886944407154136;1.1599726637959426601;-0.013493001919060821497;-
0.030017944816860285 54];
IW1_1 = [-0.17550700259775048329 0.12437900796796157177
0.11079239025341047764 -0.034075895754986283703 0.29483175937824362967 -
0.057713815834810146266 -1.1193397324834686302
```

```
3.3403573184268307017;0.10382108796636763115 -0.26423074546211317015 -
0.07040560218118190353 0.14880835326845129329 0.24531852798260075921 -
0.14005801785651375568 -0.86108523618232735153 -0.49447826497060376472;-
1.5375680323729121746 0.10744233312355987164 1.6972164684725878914
0.023475570635027292327 0.18654826701683385748 -0.3040185727588216591 -
1.2667740544547705372 2.5996354505903949139;1.6759218404130238333 -
0.99220912659864490823 -1.7639973679471612211 0.24409267318162125582
0.6703697837820276062 0.057992176831529004821 -3.6546840051134310556
1.9322862201354018552];

% Layer 2
b2 = [-0.20064561634415761793;0.2126716541953492523;-1.207669621029798579];
LW2_1 = [1.4903682160365423481 1.2342221870312257703 -0.74689452973439007799
-0.84407765909843834873;0.93997435989942523804 -0.5797040945375588894 -
0.2791793199562800365 -0.061694321259383247402;2.5958358576357181136
4.5098690559625476126 -2.2368002370158452941 -2.2187453127702148059];

% Output 1
y1_step1.ymin = -1;
y1_step1.gain = [4.33570182775714;6.03314122048985;2.00076573878507];
y1_step1.xoffset =
[0.538713666332862;0.668497731628166;0.000382722859664607];
```

3.7 Network constants of model C1 (F11, 4°C):

```
% Input 1
x1_step1.xoffset = [0;0;0;0;0;-0.055930255733897;0;0;0;0;0];
x1_step1.gain =
[2;2;2;2;1.97773680791308;0.837709383383849;2;2;2;2;2.023543336545555];
x1_step1.ymin = -1;

% Layer 1
b1 = [-
0.40370089747207732467;0.17871799244745403823;0.47721400744757097589;0.520897
95999478671895;0.35457508070124665878;-0.94286052713759127553];
IW1_1 = [0.42155318445361472657 -0.51524473126784586796 -
0.20416798404505290265 0.20768760396018018377 0.91916844755465487715 -
0.12411818824628013946 0.062744834704219876187 0.067716436169300994741
0.45633171680676221094 1.1312218874864743423 -0.99368871304347861084;-
0.075393579991349463088 0.16334887328936875672 -0.59583923552529727274 -
0.41668703852470406757 0.044162310996705117505 -0.5351140441134089043 -
0.36265903159808504652 -0.1946690382647009443 -0.34075115001898026446 -
0.48514089676642330184 -0.99573410884504554019;-1.1018482189815230754
0.39109425926540880569 0.5022508519833684959 -0.18066345514459647426 -
0.73383218576374342756 0.23175528400746367441 0.49056863989596116715
0.68421672228251062542 -0.091714531470580146211 0.16413163147728729307
0.18544990585682508133;-0.32220459213361984929 -0.44611626501836698644
0.5951783055127251032 0.081814668854340133475 -0.35094977437538699139
0.03667013504512791483 -0.15108827441790620427 0.80839056868347070051
0.018060888995733902218 -0.1729977215935263879 -0.29692770056889195018;-
1.678856968015912221 -0.49662208635976368143 0.58022662747001263472
0.96352965387047762036 0.56931444262515684684 -0.59179359526713404271 -
0.14693476133743244016 0.51195120845292796385 0.15873225673975835393
0.30521451541199917656 -1.057796546570762386;0.11424739908715322856
```

```
0.26971448480340631804 0.316969740906699049 -1.1481567825044523001 -
0.0074052056914349273162 -0.052248831785534813332 0.23327881834227992552
0.84472418487630918094 -0.00072994081188172102621 0.34210671310969131831
0.60528776425713382014];

% Layer 2
b2 = 0.25004823586103314126;
LW2_1 = [-1.4222597480419465477 -1.5149632866517936947 -1.4028569325369002474
-1.113295912051026626 1.2217769190397147483 1.2258728538589300783];

% Output 1
y1_step1.ymin = -1;
y1_step1.gain = 6.03314122048985;
y1_step1.xoffset = 0.668497731628166;
```

3.8 Network constants of model C2 (F11, 4-25°C):

```
% Input 1
x1_step1.xoffset = [0;0;0;0;0;-0.055930255733897;0;0;0;0;0];
x1_step1.gain =
[2;2;2;2;1.97773680791308;0.837709383383849;2;2;2;2;2.02354333654555];
x1_step1.ymin = -1;

% Layer 1
b1 = [-0.52651855988476814474;0.69062321846358665223;-
0.17167419635148478685;-0.15499703436247458388;0.52209410001840095283;-
0.42767933782322692382];
IW1_1 = [0.10022963166839572724 -0.84081543522351886999 -
0.29543577802368514718 0.72116697155261100249 1.7127660871314180913 -
0.28516433906903887463 -0.72880523384219852989 -0.17856100516108489029
0.15093786032072054049 0.95412463959634619126 -0.10710561056643286282;-
0.00039829718886674617476 0.22050275586323619947 -0.39226632266018590656 -
0.54392347313409739051 -0.34878852547786320937 0.17925726534135802415
0.67261144648981685545 -0.31054145611810962668 0.30790081361015109174 -
0.46277435567193164756 -1.4642725542925394411;0.13625355556901347365
0.00246314086345979926 8 -0.52802226153241338391 0.28977916859303337604 -
0.96768573666869950323 1.210825525780141243 0.57813464266175240702
0.30559296977505984172 -0.14514216772172061631 -0.35949722848328380387
1.3094910292552288666;0.26425597946439760566 -0.073738425137265431464 -
1.5029833750217902466 -0.076281370054950892179 0.16865405701609323397 -
0.22884309627202747972 0.93123848734340863853 -0.38073160815216555708
0.35296788766705933238 -0.30270105573622169137 -0.60867937149247641759;-
0.13786229440281241865 -0.23317168465003168532 -0.30418458333108283620 -
0.17396401655378310180 -1.0132457062355946231 -1.6272761142785407973 -
0.57072078290912131582 0.021820235472612806854 -0.30091898545501122308 -
0.12427108770706132845 -0.78640154956120456031;0.14631461314929727902 -
0.08723448337497741193 43 -0.23492433614969499001 -0.47138504446961276839
0.58554162201984205627 0.064136046786829034594 0.53819358487453516915 -
0.50161955465587315661 0.48357451118477440266 -0.15300712300686425404 -
0.25504133871203688022];

% Layer 2
b2 = [-0.39575365639336035350;-0.0027204676170238142711];
```

```
LW2_1 = [-1.0283769261819599716 -1.175600542265540005 -1.8198451666350585754
1.2973431152349981232 -1.390529074198237458 -0.73936386857524893923;-
1.0551096428032396002 -1.7581362654837433546 -0.43157668385251063414
0.37343596504681847437 -0.37581698320677126013 1.1150136839987450088];

% Output 1
y1_step1.ymin = -1;
y1_step1.gain = [4.33570182775714;6.03314122048985];
y1_step1.xoffset = [0.538713666332862;0.668497731628166];
```

3.9 Network constants of model C3 (F11, 4-25-40°C):

```
% Input 1
x1_step1.xoffset = [0;0;0;0;0;-0.055930255733897;0;0;0;0;0];
x1_step1.gain =
[2;2;2;2;1.97773680791308;0.837709383383849;2;2;2;2;2.023543336545555];
x1_step1.ymin = -1;

% Layer 1
b1 = [-1.2392226155832075651;1.2966158332307400336;0.26850488476994649467;-
1.3247219500176135032;-0.88022712900115407386;-1.2810062537233188884];
IW1_1 = [0.71608821579649795108 0.3409317793897489457 -
0.052174061111371225485 0.9538860365435065658 1.2854636201136866447
0.94644461581302075093 -1.0273497022101258924 0.58676415695967398811 -
0.10807895724704920382 -0.64795847557246744053 1.8686243371798876289;-
0.11270553452187728749 0.22679310657614148994 -0.10237681129151732229 -
0.068361740010054655725 -0.42314162205500049918 -0.036313823835854804056
0.003172797347704695535 -0.23063687649130920621 0.016941200328373838546 -
0.62152320192993015802 -1.281050571462963416 6;-0.0074507819657740363348 -
0.23125036136245119556 0.049380463780601693835 1.5293577813736947135 -
1.0525655494245482746 -2.3552070267476157284 -1.4673893145569010255
0.31312216229156331071 -0.17754766724191686778 -2.1577111566370517082
1.1076873834512623862;0.029634417142328804695 1.353012543655713884
0.040160093153926372522 0.46566651949555820655 1.1707714247437135402
1.0221934464931057285 -0.82324845039265837077 0.059661774486995243494 -
0.14242170493113240348 -1.2880827257448252965
0.61536103684240961442;0.25106860345269838009 0.22842046826466586018
0.020308366311532363713 1.2462430374475896855 0.56942799937270838218 -
2.3034564935030985566 -1.1559582095849907368 0.16244021157367546926
0.11342535548220293395 0.45229451592467639909 -0.020907643342617097948;-
0.033749069795194333587 1.1331913951069494395 0.044510223619863037436
0.84194905452217483344 1.07052264342128578 -2.2858954648655465114 -
0.70836733338416313366 -1.1301562327285641363 0.70502761775772204711
0.47647127560363750076 -0.10546778216772731418];

% Layer 2
b2 = [0.75074727525263573735;0.63407654340164654183;1.0970393464380399084];
LW2_1 = [-0.36967834613234545893 -1.3950388413634842166
0.045354031686088633335 0.81608230579108209657 -0.039415582913781693764
0.16730421296430131317;-0.23016961987202153028 -1.8128752736744009333
0.23354928990336648598 0.53929324135714873378 -0.71210704965445881243
0.38990595802020555638;-1.6266627863393188445 0.14695715082251606387 -
1.8434275379145830254 2.5533514120532867508 2.6418053462848876123 -
1.3196798194018877215];
```

```
% Output 1
y1_step1.ymin = -1;
y1_step1.gain = [4.33570182775714;6.03314122048985;2.00076573878507];
y1_step1.xoffset =
[0.538713666332862;0.668497731628166;0.000382722859664607];
```

3.10 Network constants of model D1 (F8, 4°C):

```
% Input 1
x1_step1.xoffset =
[0;0.578048457148629;0.504630685370859;0.262552631822902;0.558329429597553;0.
121445902425566;0.320586542346591;0.568642574177782];
x1_step1.gain =
[2;4.61671515771224;3.29535895127815;3.07013207334982;4.5282618630841;2.27646
767059846;2.94371560861879;4.76505063157985];
x1_step1.ymin = -1;

% Layer 1
b1 = [0.46226033058223148231;-
0.039604028491902795728;0.15149371605456501966;-0.65341946296139263861];
IW1_1 = [-1.7324195799026460563 0.39858007103103948454 -
0.35916257909204302168 0.28681331343778226906 0.3898331574879249084 -
0.73046225506045325826 -0.71760775935311360563 -
0.35928426760370385518;0.17348226008647976704 -0.75171379011693073569
0.26841949147381283236 0.155759909769252608 -0.62917756745993347334
0.58356411232913307785 0.81309283502202722715 -
0.68249943503489951535;1.1979955287650545515 0.21395335052703146062
0.064102074526654867492 0.49362767499456261522 1.1551988574146536681 -
0.53126238742993037167 -0.10958601862512362679 0.39500070889342187019;-
0.026582418915605761367 -0.39777051900722937283 0.055903900628677893891 -
1.074746352346635625 -0.71204013805111610047 0.014875178202156688276
0.25143122081670804535 -0.645956195480287598271;

% Layer 2
b2 = -0.42250793271618813929;
LW2_1 = [-1.1463831735757361585 -0.94600724021728355151 -
1.2483796052899607254 -1.20816704317177081581;

% Output 1
y1_step1.ymin = -1;
y1_step1.gain = 6.03314122048985;
y1_step1.xoffset = 0.668497731628166;
```

3.11 Network constants of model D2 (F8, 4-25°C):

```
% Input 1
x1_step1.xoffset =
[0;0.578048457148629;0.504630685370859;0.262552631822902;0.558329429597553;0.
121445902425566;0.320586542346591;0.568642574177782];
x1_step1.gain =
[2;4.61671515771224;3.29535895127815;3.07013207334982;4.5282618630841;2.27646
767059846;2.94371560861879;4.76505063157985];
```

```
x1_step1.ymin = -1;

% Layer 1
b1 = [-0.35617905549234829543;-
0.41892582824143997922;0.37924793923566973985;-0.86286447773348495804];
IW1_1 = [0.33925165458226053117 -0.078989479449508237519
0.1567430586683562499 0.131269349818998321 -0.30242796330672022664
0.67047375572527356802 0.34308796891194076606 0.50782648386046003175;-
0.89456258860463566673 -0.51942793654267949144 0.95420138459822190224 -
1.2765454490159577894 -0.27523691667659827553 0.030725564925658463422
1.4870587829083758979 -0.64106319981723003387;-0.45314488733215435579
0.43453231756131183161 0.49401812675466100222 -0.96302257691206805124 -
0.0227142942216037938 -0.97778109368067067653 -0.5973241123265220498
0.95429919414000108357;-0.50324774746997547048 0.40648668289087752648
0.30289911830334803611 -0.34008757912086523545 -0.26366064002048639203
0.3037322423339041122 1.5117781997273185279 -1.3163712418748116928];

% Layer 2
b2 = [0.60287279178289909787;0.14744378905376265276];
LW2_1 = [0.3916673119489742505 -1.1129497118103595099 0.8139650978018614369
1.4800344186853791051;0.91197984162503187289 -0.58451683891640904278
0.71861410265830516497 0.59524747655955612125];

% Output 1
y1_step1.ymin = -1;
y1_step1.gain = [4.95521477085903;6.03314122048985];
y1_step1.xoffset = [0.584773656687707;0.668497731628166];
```

3.12 Network constants of model D3 (F8, 4-25-40°C):

```
% Input 1
x1_step1.xoffset =
[0;0.578048457148629;0.504630685370859;0.262552631822902;0.558329429597553;0.
121445902425566;0.320586542346591;0.568642574177782];
x1_step1.gain =
[2;4.61671515771224;3.29535895127815;3.07013207334982;4.5282618630841;2.27646
767059846;2.94371560861879;4.76505063157985];
x1_step1.ymin = -1;

% Layer 1
b1 = [0.12306361225536781301;-0.53966669081909068861;0.36656064188294035233;-
0.69299295051835085868];
IW1_1 = [-2.0074569595974565139 -1.2204017982957546806 0.2394366041436190029
-2.2865540946300626501 1.7889668248195083056 1.8471183335868692765 -
2.2035606648373744498 -2.4642331554415477513;-0.026883572229301603423 -
0.045566037425901810554 -1.2473170442658954027 0.6036941653605939484
0.84349832461384210980 0.69587596367047099832 0.13328751860562118581 -
0.52425761264123627026;-1.0210932003967743142 -1.4392998980433400202
0.65963496468599791545 -2.0737626651702525393 0.77284705469432279923
0.55003979801526159399 -1.1656031342103754245 -
2.3066866344921295173;0.89647078775850619348 -0.4402694322699585161 -
1.6606750170070705419 1.5774424716933317381 0.95254372678695786725
1.0524785096407160978 0.46410107597776611987 -1.9739027732084308830];
```

```
% Layer 2
b2 = [0.36210794046796063173;0.30607573585256131476;0.69879923580211900092];
LW2_1 = [-0.30817073546736140077 0.84117526057646840876
0.17481725739552503307 -0.70673235285474189649;-0.20518617236931485648
0.70398339089493167098 -0.053575454849440345972 -0.71988351691421448297;-
2.0277897846080130684 1.9551992706789758092 2.1209403038154373 -
0.99650157056928245769];

% Output 1
y1_step1.ymin = -1;
y1_step1.gain = [4.95521477085903;6.03314122048985;2.09609572180438];
y1_step1.xoffset =
[0.584773656687707;0.668497731628166;0.000382722859664607];
```

3.13 Network constants of model E1 (F8, 4°C):

```
% Input 1
x1_step1.xoffset = [0.399577464788732;0;-0.055930255733897;0;0;0;0;0];
x1_step1.gain =
[3.88261381937096;2.17131722032741;0.837709383383849;2;2;2.13304186684969;2.0
6416932772396;2.0700417231272];
x1_step1.ymin = -1;

% Layer 1
b1 = [0.45868763237835985525;-0.1995172999306056294;1.0939337440584191441;-
0.72761038376368381009];
IW1_1 = [0.10555030380314357415 0.51202744134167166301 -
0.022483404811268613288 0.057463524659501719061 1.1783596839634353781
0.18380019133974687118 -0.90184015028218167398 0.0045616387928068208649;-
0.22326814743801826069 0.059150798279541470792 -0.15282454681264190377
0.35902897499569036421 -0.65059939937768007656 -0.53169983899447104125 -
0.4424668821019595577 1.3218425385318728349;-0.16447769642528858869
0.15535708961337046929 -0.22702765953263073784 0.5133975598846085564 -
0.3719939804139695827 -0.38641070519496484303 -0.48387943503280800295 -
1.2101368354462489485;-0.22932119423103722311 -0.23730571938086264527
0.11860738977645129966 0.82641174546414386537 -1.2033985442193688709 -
0.38987750380609359491 -0.01893426825632058344 -0.10079708225399422461];

% Layer 2
b2 = 0.008940583180508129487 7;
LW2_1 = [0.94700689655140113388 1.4963083025905123957 -1.7229745092008086349
1.1208835042803368953];

% Output 1
y1_step1.ymin = -1;
y1_step1.gain = 7.00496287781719;
y1_step1.xoffset = 0.687464394929014;
```

3.14 Network constants of model E2 (F8, 4-25°C):

```
% Input 1
x1_step1.xoffset = [0.399577464788732;0;-0.055930255733897;0;0;0;0;0];
x1_step1.gain =
[3.88261381937096;2.17131722032741;0.837709383383849;2;2;2.13304186684969;2.0
6416932772396;2.0700417231272];
x1_step1.ymin = -1;

% Layer 1
b1 = [-0.42710442531050368808;1.2638378257505151758;-0.10969403914571353165;-
0.038794447932146439162];
IW1_1 = [0.18178196914529881623 1.3135120099494961288 -0.52035306520478963321
-1.379498215899536584 -0.15344586940627280569 -0.24848675240437048672 -
1.882272769956559344 0.52908063364162105646;0.12015626854582947547
0.24229591420878365815 -0.54126195653649777917 0.38153979770303148022 -
0.029537862278413165584 -0.093929092588251536511 -0.74333789945365635887 -
1.506832516907202546;-0.06522503739089847985 0.35499834665457374872 -
0.18797076112611943444 1.5338148262230968388 0.071905588981503912316 -
0.044348224000649534571 0.82938131836907591499 -
0.67041499335623988554;0.092649307683572823535 0.27662443836092337479 -
0.30769087853402427024 0.23460425128056169863 -0.073662407454700901122 -
0.082591988291961002111 -0.93259083801826270221 0.34143976345947585349];

% Layer 2
b2 = [-0.067444105885872587813;0.21408896718345235555];
LW2_1 = [1.3273582649521054044 -0.80680244946838763909 1.7674672902300629929
-1.5796403416202389192;0.083006575384792044559 -1.929784396263211077
0.30116897971609174478 1.7874686426362318414];

% Output 1
y1_step1.ymin = -1;
y1_step1.gain = [4.33570182775714;7.00496287781719];
y1_step1.xoffset = [0.538713666332862;0.687464394929014];
```

3.15 Network constants of model E3 (F8, 4-25-40°C):

```
% Input 1
x1_step1.xoffset = [0.399577464788732;0;-0.055930255733897;0;0;0;0;0];
x1_step1.gain =
[3.88261381937096;2.17131722032741;0.837709383383849;2;2;2.13304186684969;2.0
6416932772396;2.0700417231272];
x1_step1.ymin = -1;

% Layer 1
b1 = [-1.2410411149750029125;0.69914969161205176018;0.47151978599762162103;-
1.9627228748189600172];
IW1_1 = [1.0897581664238331633 -0.60734205231348847853 -
0.64191966422586410612 0.85524763529426728326 -1.440017180225172666 -
0.16230116294272059485 -0.77410814485782908356 -1.4772771508868229873;-
0.20166922399038414437 -1.4753746185813385594 -0.013286099395561063938
0.30205714692312435332 -0.029752259352854026897 0.013367550331830988647
0.30121733026169528191 -1.486869252450622847;-0.53711453215301041553 -
2.1825440178469355246 -0.010153265258977521579 0.20587348584335951607
```

```
0.01397506571221173062 0.14222300140092034026 -1.2405066176131749334
1.4973333114983937708;-0.117222951459633698 0.89731970163599861223
0.085010789539900016853 -0.20544179548860450057 0.00019588949115233099895
0.058259981703963750754 0.71495147213064758063 1.8396645955968276009];

% Layer 2
b2 = [0.44041004544225931250;0.81721162716614814681;-0.11842909309622452751];
LW2_1 = [0.40505951576539822589 0.29286339004681977150 -0.03867368802576221648
1.11158448141700105790;1.36473707033482716260 -0.44704939357110989873
0.63958233224248617343 1.29518174580938527640;-1.11684226733356561920
2.37940379147984026890 -1.37474671312366480970 1.19792172465310531320];

% Output 1
y1_step1.ymin = -1;
y1_step1.gain = [4.33570182775714;7.00496287781719;2.01612903225806];
y1_step1.xoffset = [0.538713666332862;0.687464394929014;0.008];
```