Advancing Therapeutic Protein Discovery and Development through Comprehensive Computational and Biophysical Characterization

Gentiluomo, Lorenzo; Svilenov, Hristo L; Augustijn, Dillen; El Bialy, Inas; Greco, Maria Laura; Vitaliyivna Kulakova, Alina; Indrakumar, Sowmya; Mahapatra, Sujata; Morales, Marcello Martinez; Pohl, Christin

Total number of authors:
26

Published in:
Molecular Pharmaceutics

Link to article, DOI:
10.1021/acs.molpharmaceut.9b00852

Publication date:
2020

Document Version
Peer reviewed version

Advancing therapeutic protein discovery and development through comprehensive computational and biophysical characterization

Lorenzo Gentiluomo[1,2,10], Hristo L. Svilenov[2], Dillen Augustijn[3], Inas El Bialy[2], Maria Laura Greco[4], Alina Kulakova[5], Sowmya Indrakumar[5], Sujata Mahapatra[6], Marcello Martinez 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 W. Streicher[6], Christopher F. van der Walle[4], Shahid Uddin[4,11], Gerhard Winter[2], Dierk Roessner[1], Pernille Harris[5]*, Wolfgang Frieß[2]

* Corresponding author. E-mail: ph@kemi.dtu.dk

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

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

[5] Department of Chemistry, Technical University of Denmark, 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

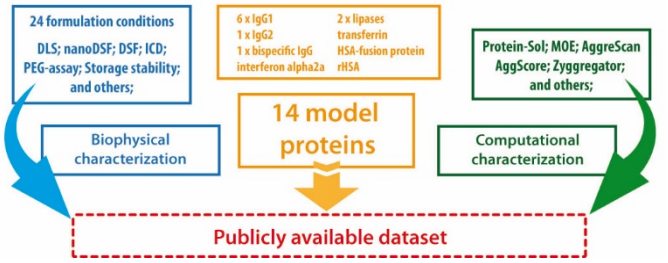
[8] School of Biological Sciences, Faculty of Biology, Medicine and Health, Manchester Academic Health Science Centre, The University of Manchester, Oxford Road, Manchester M13 9PT U.K.

[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

KEYWORDS: Therapeutic proteins; Protein formulation; Developability assessment; Protein characterization;



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 research 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.

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 prerequisites 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 unpublished 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. MATERIALS 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 - SI1. The proteins in the bulk solutions were dialyzed overnight using Slide-A-Lyzer™ cassettes (Thermo Fisher Scientific, Waltham, 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 (Darmstadt, 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 generated in the consortium. 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, Zscore was calculated using the standalone version of Prosa2003[32]. The generated models have an overall negative Zscore indicating a good quality of built structures (Supplementary information - SI2).

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., New York, New York, 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, Santa Barbara, 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 were collected and processed with the DYNAMICS® software V7.8 (Wyatt Technology, Santa Barbara, 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 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 $R_{h,o}$ 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 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 (Nano-Temper 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 SI3.

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-point 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 (Ward Hill, 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_o 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 run 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, Santa Barbara, 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, Santa Barbara, 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, 40 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 percent as M_{25}, M_{40} and M_{50} in the Supplementary Table.

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

2.11. Stress study

Protein samples with concentration 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, 25, 40 °C and 50 °C for two weeks, and in a separate study at 4 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 analysis (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 – SI4. All the results were calculated using the statistical software JMP® v 14.0 (SAS Institute Inc., Cary, 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 param-

eters 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. 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 sample. The data points of the whole dataset are also provided in Supplementary information – SI5.

2.14. Principal Component Analysis (PCA)

In order to get a quick overview of all the data stored, 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 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 - SI1. 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 was 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 – SI6. 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 neighbouring 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.

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 parameter.

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 used often in published work on the characterization of therapeutic proteins. Although this selection might be subjective, it is based on our experience and the availability of the techniques in the consortium.

The type of molecular descriptors calculated with MOE and ProDCal are summarized in Supplementary information SI7. The parameters from AggScore, Zyggregator and Aggrescan are presented in Supplementary information SI8.

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, 70 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. 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. Comparison 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/).

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. 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 SI9. Also, in SI10 the correlations in other subsets are shown. In general, the strength of the correlations observed in Figure 1a can change slightly when only a subset of the proteins like the one in SI9 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 SI7 (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.

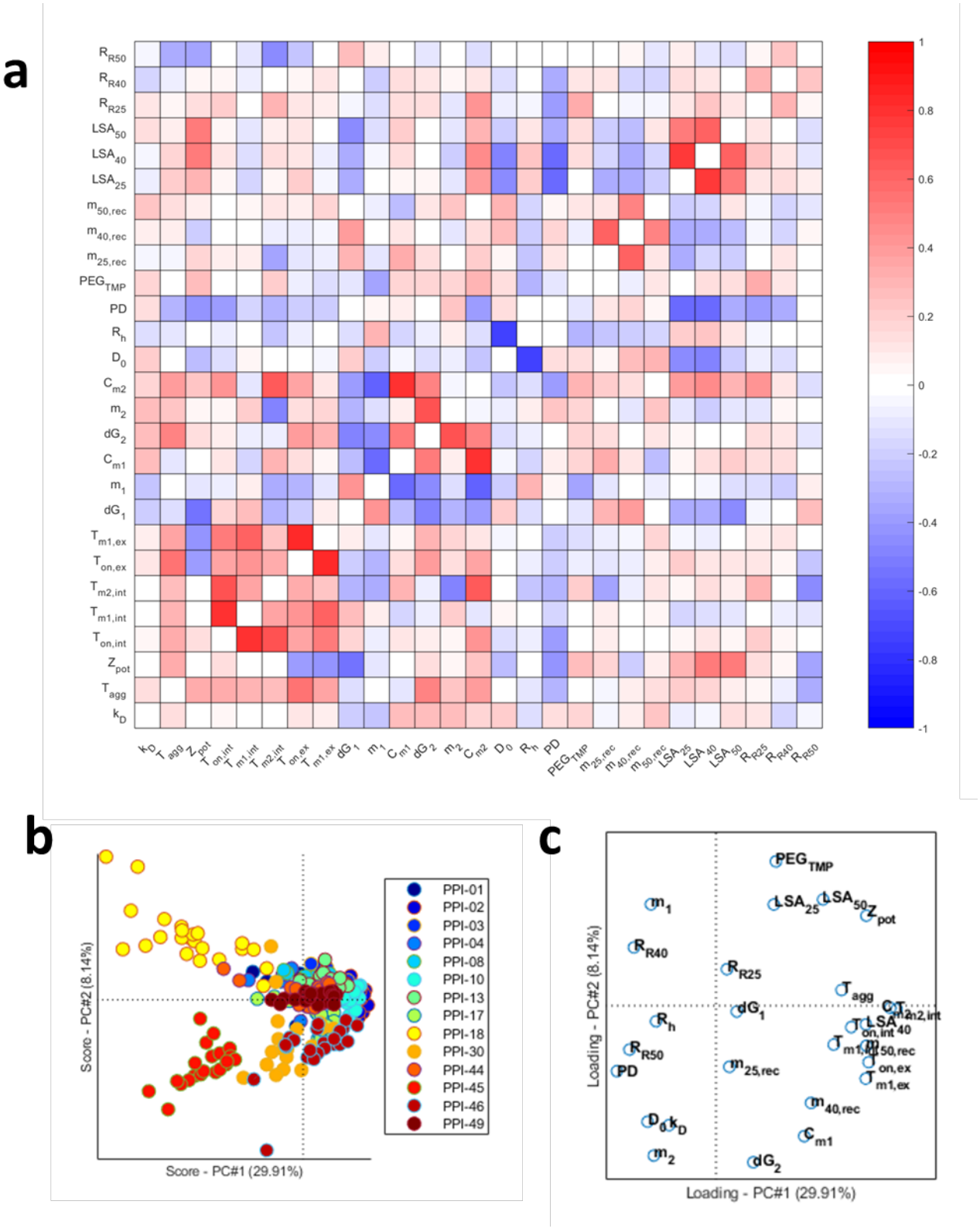


Figure 1. *(**a**) - Pairwise linear correlations between measured biophysical parameters in the entire dataset including 14 proteins and 24 different formulation conditions. The strength of these correlations was assessed using t-tests. p-values <0.05 are statistically significant at a 95 % confidence 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.)*

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 °C 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 %. 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 the selected cut-off 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 present 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.

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.

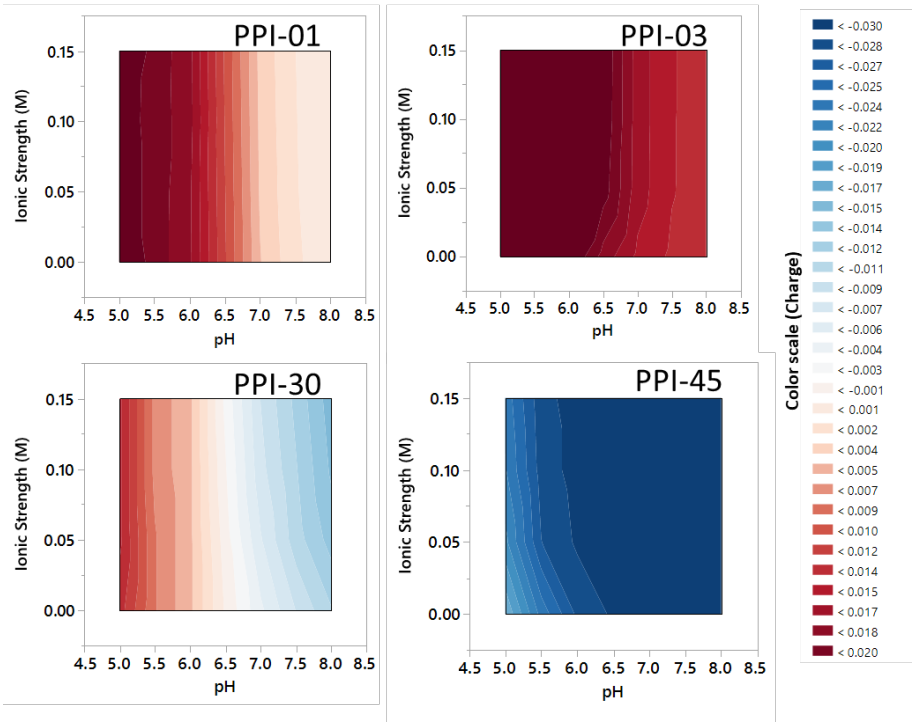


Figure 2. *Calculated charge per amino acid as a function of pH value and ionic strength for two antibodies (PPI-01, PPI-03), interferon α2a (PPI-30) and one lipase (PPI-45).*

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.

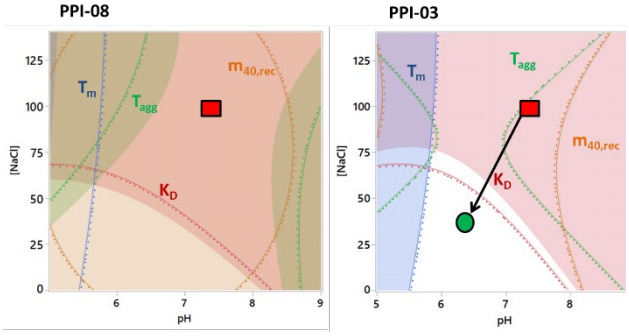
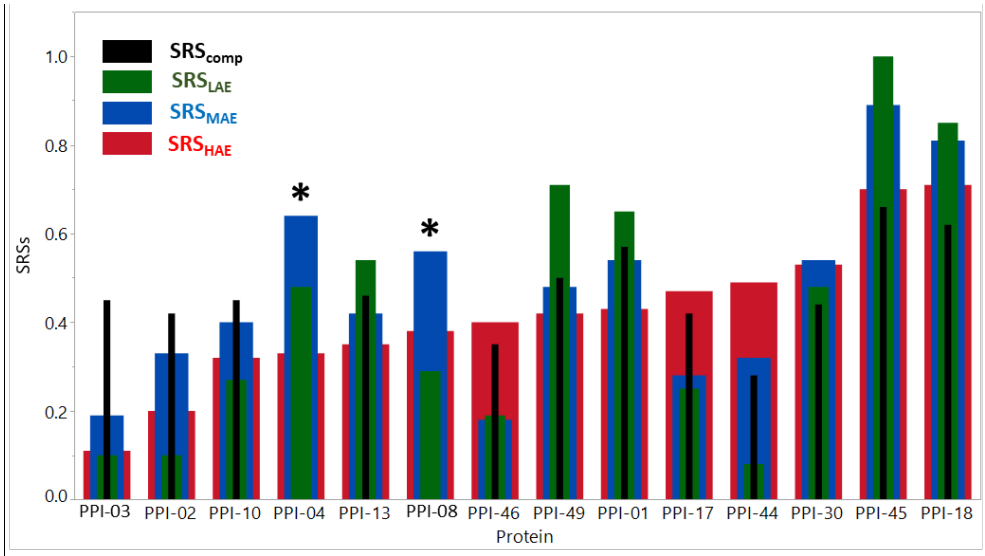


Figure 3. *Contour plot representing the change of several critical biophysical parameters with pH and salt concentration for PPI-08 and PPI-03 - white areas present regions where all four parameters are above the defined cut-off value. Areas colored in red, green, blue and yellow represent areas where 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.*

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 between 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.



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. SRS$_{comp}$, SRS$_{LAE}$, SRS$_{MAE}$, SRS$_{HAE}$ are represented in black, green, blue and red, respectively. The proteins are ordered by SRS$_{HAE}$ value in ascending order. A higher SRS indicates an increased risk of stability issues. The asterisk (*) indicates proteins with no primary sequence available and therefore without a SRS$_{comp}$ in this study.*

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 the proteins. Shifting the SRSs cut-off to an upper or lower limit, thus forcing the SRSs to 0 or 1 for all the 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 – SI12. 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 (Figure 4, blue line). 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 effort. Interestingly, the SRS_{comp} correlates well with the SRS_{LAE} (Figure 5). 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 between 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.

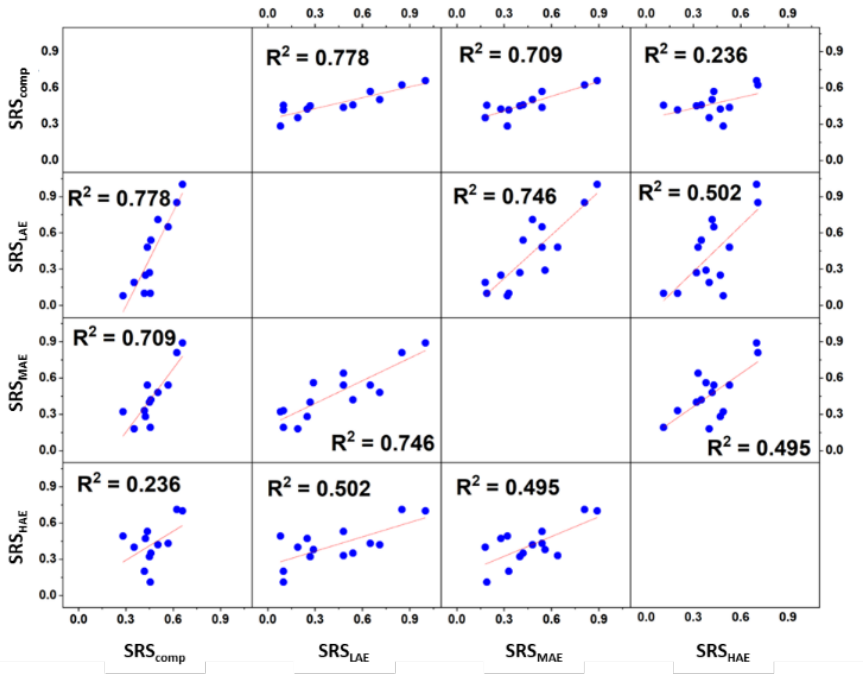


Figure 5. *Linear correlation between the protein rankings based on four stability risk scores. For more information regarding the calculation and definition of the different stability risk scores refer to the main manuscript text and Figure 4.*

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 ^{0}C. 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 of the experimental SRSs for physical stability risk assessment and 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 the proteins, formulations and temperatures of stress studied. SRS_{comp} present the lowest mean correlation and highest variability. As expected, with increasing 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. Interestingly SRS_{LAE} and SRS_{MAE} present similar prediction power which confirms that an early rough ranking by use of few high throughput biophysical parameters, namely T_{agg} and $T_{m,int}$, 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).

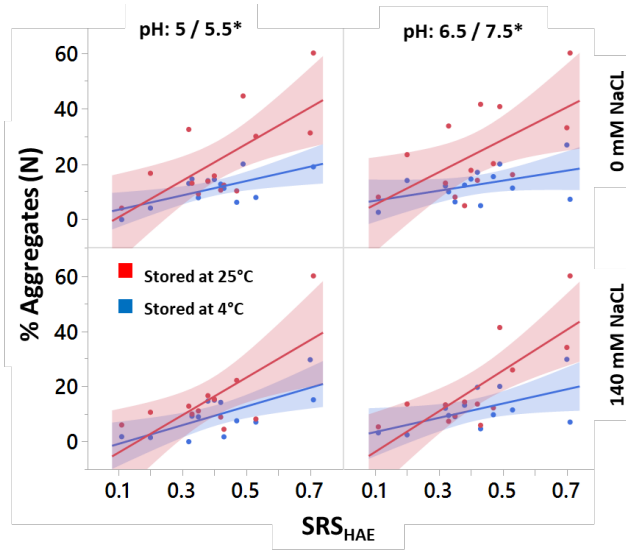


Figure 6. *Linear correlation between SRS_{HAE} and the percentage of aggregates after 6 months of storage at 4 °C (in blue) and 25 °C (in red). A total of four formulation were studied i) 10 mM His at pH 5 ii) 10 mM His and 140 mM NaCl at pH 5, iii) 10 mM His at pH 6, iv) 10 mM His and 140 mM NaCl at 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.*

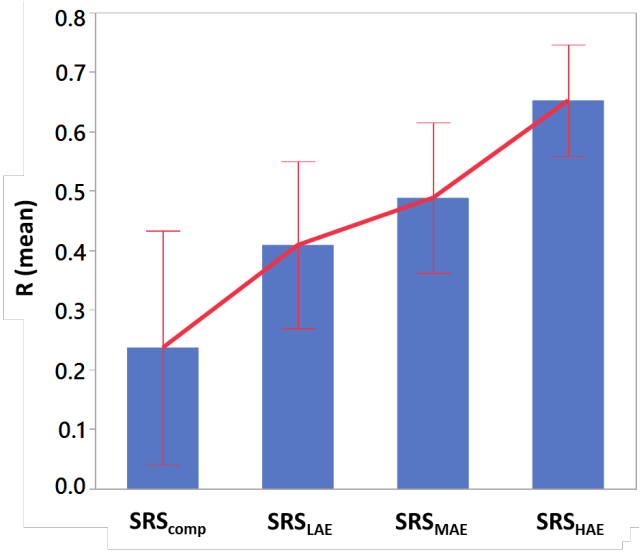


Figure 7. *Averaged Pearson's correlation coefficient R between SRSs and the percentage of aggregates after 6 months of storage at 4 and 25 °C. The mean and error bars are calculated as a standard variation of the R value between SRSs and the amount of aggregates after 6 months of storage over all the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

variety of biophysical parameters that are considered to be indicative for protein stability. We find linear correlations between some biophysical parameters, but not between 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 between 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 correlates better 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 and 25 ^{0}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.

ASSOCIATED CONTENT

This supplementary information included is mentioned below and available free of charge via the internet at **http://pubs.acs.org**.

SI1. Primary sequences of the studied proteins, SI2. Zscore values for the homology model structures, SI3. Details on the calculation used for extrinsic DSF, SI4. Fitting from the response surface methodology (RSM), SI5. Multivariate matrix including all datapoints, SI6. Separations obtained with SEC-MALS and cIEF for the proteins in the dataset, SI7. List of the molecular descriptor calculated by MOE and ProDCal, SI8. Parameters from AggScore, Zyggregator and Aggrescan, SI9. Pairwise correlations between biophysical parameters in a subset including only mAbs, SI10. Pairwise correlations between biophysical parameters in subsets including different proteins, S11. Surface profiles of studied protein relative to a reduced subset of several biophysical parameter, SI12. Cut off values for used for the calculation of the different stability risk scores (SRSs) and S13 a table containing all the fitted biophysical parameters.

AUTHOR INFORMATION

Corresponding Author

Pernille Harris

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

E-mail: ph@kemi.dtu.dk

Author Contributions

Lorenzo Gentiluomo, Hristo L. Svilenov, Gerhard Winter, Wolfgang Frieß wrote the manuscript. 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 L. Svilenov, performed and analyzed nanoDSF on 20 % of the protein library. Inas El Bialy, performed and analyzed DSF on the 100 % of the protein library. Maria Laura Greco, performed PEG-assay on 20 % of the protein library. Alina Kulakova and Sujata Mahapatra, performed and analyzed nanoDSF and ICD on 90 % of the protein library. Sujata Mahapatra and Alina Kulakova performed the purification of 15% of the tprotein library. Marcello Morales, performed and analyzed PEG-assay on 80 % of the protein library. Christin Pohl, performed and analyzed nanoDSF and ICD 10 % of the protein library. Christin Pohl and Sujata Mahapatra performed DLS and T_{agg} 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. All authors have given approval to the final version of the manuscript.

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. The first author thanks Wyatt Technology staff members for their many contributions.

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 analysis; 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;

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.