Computational and Structural Biotechnology Journal 20 (2022) 1439–1455

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Investigation of the pH-dependent aggregation mechanisms of GCSF using low resolution protein characterization techniques and advanced molecular dynamics simulations

Suk Kyu Ko [a,*], Carolin Berner [b], Alina Kulakova [c], Markus Schneider [d], Iris Antes [d,1], Gerhard Winter [b], Pernille Harris [c], Günther H.J. Peters [a,*]

[a] Technical University of Denmark, Department of Chemistry, 2800 Kongens Lyngby, Denmark
[b] Ludwig Maximilian University of Munich, Department of Pharmacy, 81377 Munich, Germany
[c] University of Copenhagen, Department of Chemistry, 2100 Copenhagen, Denmark
[d] Technical University of Munich, TUM School of Life Sciences, 85354 Freising, Germany

ARTICLE INFO

Article history:
Received 15 December 2021
Received in revised form 13 March 2022
Accepted 14 March 2022
Available online 17 March 2022

Keywords:
Aggregation
Granulocyte stimulating factor
Molecular dynamics
Coarse grained simulations
Small angle X-ray scattering

ABSTRACT

Granulocyte-colony stimulating factor (GCSF) is a widely used therapeutic protein to treat neutropenia. GCSF has an increased propensity to aggregate if the pH is increased above 5.0. Although GCSF is very well experimentally characterized, the exact pH-dependent aggregation mechanism of GCSF is still under debate. This study aimed to model the complex pH-dependent aggregation behavior of GCSF using state-of-the-art simulation techniques. The conformational stability of GCSF was investigated by performing metadynamics simulations, while the protein-protein interactions were investigated using coarse-grained (CG) simulations of multiple GCSF monomers. The CG simulations were directly compared with small-angle X-ray (SAXS) data. The metadynamics simulations demonstrated that the orientations of Trp residues in GCSF are dependent on pH. The conformational change of Trp residues is due to the loss of Trp-His interactions at the physiological pH, which in turn may increase protein flexibility. The helical structure of GCSF was not affected by the pH conditions of the simulations. Our CG simulations indicate that at pH 4.0, the colloidal stability may be more important than the conformational stability of GCSF. The electrostatic potential surface and CG simulations suggested that the basic residues are mainly responsible for colloidal stability as deprotonation of these residues causes a reduction of the highly positively charged electrostatic barrier close to the aggregation-prone long loop regions.

1. Introduction

Proteins are widely applied as medicines due to their high specificity compared to small chemicals [1,2]. However, protein drugs exhibit additional challenges when it comes to the development of formulations that can preserve their stability [3,4]. Protein aggregation is a commonly encountered problem in the development of biopharmaceuticals that can affect the efficacy of the product and cause undesired immune reactions in patients [5]. Both protein colloidal and conformational stability have been related to protein aggregation [5]. The colloidal stability of a protein is related to weak net interactions between the protein molecules in solution, which can either be attractive or repulsive. The conformational stability is defined by the equilibrium between folded and unfolded states of a protein, and a slight deviation from these

Abbreviations: A3D, aggrescan3d; CD, circular dichroism; CG, coarse-grained; cMD, conventional molecular dynamics; COM, center of mass; CV, collective variable; D_0, infinite dilution diffusion coefficient; DLS, dynamic light scattering; DSF, differential scanning fluorimetry; ESRF, European Synchrotron Radiation Facility; FES, free energy surface; FF, force field; GCSF, granulocyte-colony stimulating factor; HDX, hydrogen deuterium exchange; MD, molecular dynamics; NMR, nuclear magnetic resonance; PDB, Protein Data Bank; PME, particle-mesh-Ewald; PPI, protein-protein interaction; $r(H)_0$, infinite dilution values for the hydrodynamic radius; SAP, spatial aggregation propensity; SAXS, small-angle X-ray scattering; SIRAH, South American Initiative for a Rapid and Accurate Hamiltonian.

* Corresponding authors.
E-mail addresses: sukkyk@kemi.dtu.dk (S.K. Ko), ghp@kemi.dtu.dk (G.H.J. Peters).

[1] Passed away Aug. 4, 2021.

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

optimal conditions may shift the equilibrium towards unfolded protein species, which are often prone to form aggregates [6,7]. Various factors including the protein's amino acid sequence and environmental factors such as pH, buffers, protein concentration, ionic strength, and storage conditions have an impact on protein aggregation. Therefore, the prevention of protein aggregation is a major challenge in the formulation development process in the biopharmaceutical industry. Due to the lack of a complete molecular understanding and predictability of protein aggregation, formulation development is still done in a trial and error approach [8,9].

In this study, we investigated the aggregation mechanism of the therapeutic model protein granulocyte-colony stimulating factor (GCSF). Native GCSF is a 19.6 kDa glycoprotein with 174 amino acid residues [10], which mediates the proliferation of granulocytes through receptor binding. Filgrastim, the non-glycosylated, recombinant form of GCSF with an additional N-terminal methionine group is a licensed drug to treat neutropenia [11]. Filgrastim, hereafter referred to as GCSF, is a hydrophobic cytokine with a molecular weight of 18.7 kDa as a result of the removal of the glycosylation [12]. The structure of GCSF is characterized by a four-helix-bundle fold with two long loops connecting helices αA and αB as well as αC and αD. An additional short 3_{10}-helix is located within the loopAB and is perpendicular to the four antiparallel α-helices (Fig. 1).

The stability of GCSF is highly pH-dependent with maximum stability at pH 4.0 and low stability and fast aggregation at physiological pH [15–17]. With an isoelectric point of around 6.1 [16], GCSF is highly positively charged at acidic pH, resulting in electrostatic repulsion between the protein molecules. Furthermore, the addition of salt at pH 3.5 causes aggregation [16] indicating a strong impact of electrostatic interactions on the aggregation of GCSF. Nevertheless, the pH-dependent behavior of GCSF is still discussed in the literature. Narhi et al. reported an increase of α-helical content of GCSF at pH 4.0 compared to neutral pH using circular dichroism (CD) measurements [18]. Another study used hydrogen deuterium exchange (HDX) mass spectrometry to compare the local changes of relative uptake difference between pH 4.0 and 7.0 and could not observe a change in the α-helical content [19].

Narhi et al. observed quenching of Trp residue(s) at pH 4.0 in fluorescence intensity measurements [18]. Similarly, an Nuclear Magnetic Resonance (NMR) study performed by Aubin et al. showed that Trp configuration is pH-dependent [20].

To provide a detailed molecular understanding of the pH-dependent aggregation mechanisms of GCSF, we performed a multi-scale modeling approach using full atomic and coarse-grained (CG) molecular dynamics (MD) simulations. The aggregation mechanism was explored by CG simulations of GCSF using the SIRAH force field (FF) [21,22] developed to simulate proteins in explicit solvent conditions. WT4 models describe the CG water molecules, where one WT4 model consists of four beads that are connected in a tetrahedral form. The protein backbone in the SIRAH FF is defined by 3 beads representing nitrogen, alpha carbon, and oxygen atoms and thereby allowing for movement of the secondary structure since no constraint is applied to fix the protein backbone. Each side chain was modeled specifically based on a combination of physicochemical characteristics. The SIRAH FF is a relatively new force field that was recently used to study the process of seeding peptide aggregation [23]. SIRAH was chosen as an alternative to MARTINI since it has been shown that the MARTINI FF overestimates protein-protein interaction (PPI) for membrane proteins [24]. The conformational stability of GCSF at varying pH values was studied by carrying out full atomic MD simulations in the pH range of 4.0 to 7.5. We could observe that the conformational state of GCSF is very similar at varying pH values in unbiased

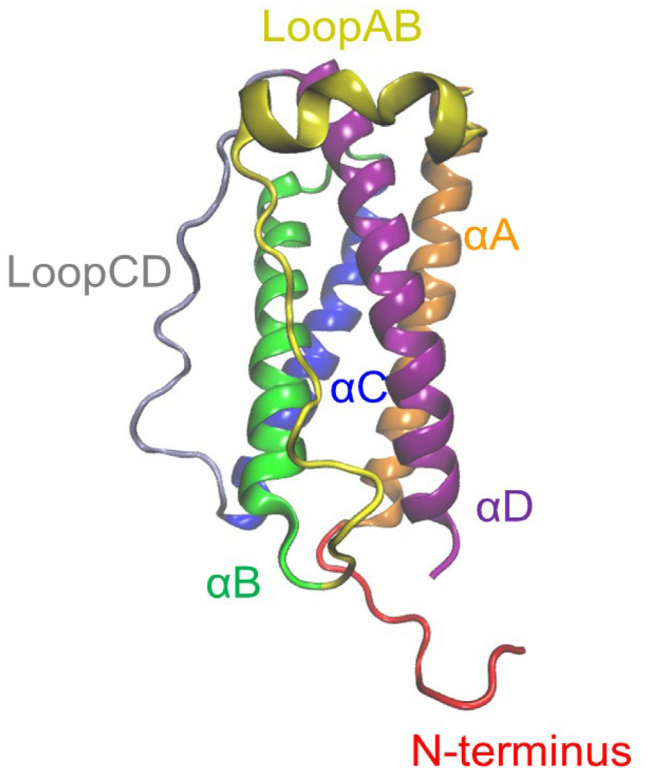


Fig. 1. The structure of GCSF was obtained from the Protein Data Bank (PDB) (PDB code: 1CD9 [13]). MODELLER was used to generate the first five disordered residues [14]. The secondary structure of GCSF is shown with different color schemes: N-terminus (Met1-Pro11), αA (Gln12-Tyr40), loopAB (Lys41-Gln71), αB (Leu72-Leu93), αC (Leu100-Leu125), loopCD (Gly126-Ser143), and αD (Ala144-Pro175).

systems. To ensure that the system is not trapped in a local minimum, we carried out metadynamics simulations. We compared our *in silico* results with experimental data obtained from fluorescence intensity, CD spectroscopy, nanoDSF, and DLS measurements as well as modeling based on small-angle X-ray scattering (SAXS).

2. Methods

2.1. Conventional molecular dynamics simulations

The structure of GCSF is available from X-ray (1CD9 [13], 1RHG [10], and 2D9Q [25]) and NMR (1GNC [26]) studies, of which 1CD9 has been widely used as the GCSF model structure in various MD simulation studies [27–29]. The initial structure of GCSF for the conventional molecular dynamics simulation (cMD) study was prepared using PDB entry 1CD9 (solved at pH 7.5) [13]. The missing five residues were added using Modeller software 9.21 [14]. The *PDB2PQR* server was used to protonate the titratable residues [30] at pH 4.0, 5.0, and 7.5. The full atomic cMD simulations were carried out using the *AMBER* software 20 [31] and GCSF was parametrized using the force field FF14sb [32]. The protein was inserted into a cubic periodic boundary box, where the minimum distance between the protein and the edge of the box was set to 15 Å. The TIP4P Ewald water model [33] was used to solvate the system. The system was neutralized by adding either sodium or chloride ions. The initial structures were minimized using 10,000 cycles. The first 5000 cycles were computed using the steepest descend algorithm. The remaining 5000 cycles were carried out using the conjugate gradient algorithm. The cut-off distance of the non-bonding interactions was set to 12 Å. The electrostatic long-range interactions were evaluated using the particle-mesh-Ewald (PME)

method [34]. The SHAKE algorithm was applied to fix the bonds involving hydrogen [35,36]. The system was heated to 300 K in the *NVT* ensemble (constant N = Number of atoms, V = Volume, T = Temperature) for 0.3 ns, using the Langevin thermostat [37] with a collision frequency of 5 ps^{-1}. The system was then subjected to a short equilibration run for 2.2 ns in the *NPT* (constant N = Number of atoms, P = Pressure, T = temperature) ensemble, while the pressure was kept at 1 bar using a Monte Carlo barostat [38]. The final production run was carried out using the *NPT* ensemble for 400 ns.

2.2. Metadynamics

To ensure that GCSF conformation is not trapped in the local minima during the simulations, we have carried out metadynamics simulations using *AMBER* software 20 [31] and *PLUMED 2* [39]. The initial structures for the metadynamics simulations were obtained from the final frame of the cMD simulations. All metadynamics simulations were carried out in the *NVT* ensemble for 400 ns using the Langevin thermostat [37] with a collision frequency of 5 ps^{-1}. The well-tempered metadynamics scheme [40] was used to ensure a smooth convergence of the free energy landscape. The collective variables (CVs) were chosen based on the experimental observations [18,19,41,19], and included the center of the mass distance (COM) between Trp and His residues to monitor the interactions between, Trp59-His157 (*d1*) and His80-Trp119 (*d2*), and the α-helical content (α) (Table 1).

2.3. CG simulations

The CG simulations were carried out using the *Gromacs* software 2018 [42] with the SIRAH 2 force field [21,21,43]. The CG model of GCSF at pH 4.0, 5.0, and 7.5 was obtained by coarse-graining the full atomic GCSF models that were obtained from the *PDB2PQR* [30] web server using the SIRAH toolbox [43]. For each simulation, 8 GCSF monomers were added to the system. The initial GCSF monomer was translated and duplicated along the *x*-, *y*-, and *z*-axes where the center of the mass distance between replicates was set to 7.5 nm. An alternative approach could have been to sample the initial structures from a population-density of structures determined from single monomer metadynamics simulations. However, without any input from experimental results, this will give rise to a large number of combinations, and we decided therefore to use the initial structure of cMD simulations.

The distance between solute and box was set to 0.75 nm resulting in a concentration of ≈30 mg/mL. Note that a too small simulation box will cause an immediate aggregation of the proteins while too large box sizes will increase the simulation time. The optimal protein-protein and protein-box distances were chosen empirically to reduce the computational burden for sampling the aggregation. The system was solvated by adding SIRAH-based WT4 [44] molecules. After the solvation, the system was neutralized by adding either sodium or chloride ions. In addition to the pH study, the effect of salt was monitored by adding 150 mM of NaCl (in CG mode) to the systems at different pH conditions. The initial minimization was conducted using the steepest descent algorithm, followed by the conjugate gradient algorithm. The maximum number of each minimization scheme was set to 50,000. The heating was performed for 2 ns where the system was coupled to the Berendsen thermostat and barostat [45]. After the heating, the system was equilibrated for 500 ns using a time step of 10 fs. To accurately sample the *NPT* ensemble, the system was coupled to the stochastic velocity rescaling thermostat [46] and the Parrinello-Rahman barostat [47]. The production run was performed for 3 μs. For each condition, 5 replicate simulations were carried out amounting to 15 μs per condition.

2.4. Materials

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

2.5. Sample dialysis and preparation

The buffer was exchanged by extensive dialysis to the respective buffer at the given pH (10 mM sodium acetate at pH 4.0 and pH 5.0, 10 mM potassium phosphate at pH 7.5) for 24 h at 2–8 °C using a Spectra/Por® dialysis membrane (cutoff 6–8 kDa, Spectrum Laboratories, Rancho Dominguez, CA, USA) or a Slide-A-Lyzer™ MINI Dialysis Device (cutoff 3.5 kDa, Thermo Fisher Scientific, Germany). The samples were collected in microcentrifuge tubes and centrifuged at 10.000*g* for 10 min and subsequently filtered with 0.02 μm Anotop® membrane filters (Whatman, FP 30/0.2 CA-S, GE Healthcare, Buckinghamshire, UK). Stock solutions of sodium chloride were prepared in the respective buffer and spiked into the dialysed protein stock to prepare samples containing 100 mM of sodium chloride. For measurements that required higher protein concentrations, the protein solutions were upconcentrated using Vivaspin 20 5 MWCO PES centrifugal concentrators (Sartorius Lab Instruments, Goettingen, Germany). The concentration was measured again, and the solutions were sterile filtered with 0.02 μm Anotop® membrane filters.

2.6. Intrinsic fluorescence spectroscopy

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

2.7. Circular Dichroism (CD) spectroscopy

Near- and far-UV circular dichroic spectra were collected at 25 °C with a Jasco J-810 spectropolarimeter (JASCO Deutschland GmbH, Pfungstadt, Germany). All samples contained 1 g/L of protein. Quartz cuvettes (Hellma GmbH, Muellheim, Germany) with 10 mm and 0.1 mm wavelength path were used for the measurements, respectively. 5 accumulations of each sample were taken at a speed of 20 nm/min. The spectrum of the respective buffer was subtracted for each sample and smoothing of the spectra was performed using the Savitzky-Golay algorithm with 9 smoothing points. The mean residue ellipticity (MRE) of the protein at each wavelength was calculated as described elsewhere [48].

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

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

2.8. Differential Scanning Fluorimetry (nanoDSF)

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

2.9. Dynamic Light Scattering (DLS)

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

2.10. Small Angle X-ray Scattering (SAXS)

For SAXS measurements, samples with initial protein concentrations of 2, 5, and 7 mg/mL were prepared and shipped to the ESRF (The European Synchrotron Radiation Facility, Grenoble, France) on dry ice. Before measurements, the samples were thawed at room temperature and centrifuged at 10,000 rpm for 10 min. Data collection was performed at the ID02 beamline. Data collection is summarized in Table S1. The DOI for the data is https://doi.org/10.15151/ESRF-ES-404440738.

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

3. Results and discussion

3.1. pH-dependent structural differences of GCSF

We have investigated the effect of the pH on the secondary and tertiary structure of GCSF using a combination of modeling and biophysical techniques. The characteristic far-UV CD spectra with two minima at 209 and 222 nm confirm the presence of α-helical protein structure at all conditions and showed no difference between pH 4.0 and 5.0 and only a slight decrease of helical content when increasing the pH to 7.5 (Fig. 2a). In agreement with our findings, the GCSF structure solved at pH 3.5 (PDB code: 1GNC) has a similar helix content to the GCSF structure solved at pH 7.5 (PDB code: 1CD9). A similar trend could be observed from HDX-measurement performed by Wood et al., who could not find any clear evidence for a change of helical contents between pH 4.25 and 7.4 [19]. In contrast, Narhi et al. used CD spectroscopy and showed that the helical content is noticeably increased at low pH (pH 4.5: helical content 75% vs. pH 7.5: helical contents 66%) [18].

The near-UV CD spectra of GCSF at pH 5.0 and 7.5 are very similar in the wavelength region from 250 to 295 nm (Fig. 2b). Surprisingly, GCSF at pH 7.5 shows a negative CD signal at wavelengths from 300 to 340 nm which is very unusual for a protein in inorganic buffer but has been previously observed for filgrastim [55]. It is presumably caused by aggregates in the sample. The near-UV CD spectrum at pH 4.0 slightly deviates from the other spectra determined at pH 5.0 and 7.5 in the wavelength region 250 to 295 nm, but the characteristic features of the spectra remain the same. Therefore, GCSF has a well-defined tertiary structure with only a little difference between the three tested pH values.

The tryptophan fluorescence of GCSF is significantly quenched when the pH is decreased from pH 7.5 to pH 4.0 indicating that the Trp residues are in different conformational states at the different pH values (Fig. 2c) This observation is in accordance with the findings of Narhi et al. [18]. GCSF contains two Trp residues: Trp59 and Trp119 which are located close to His157 and His80, respectively (Fig. 2d). The change in pH causes a conformational change of Trp that promotes interactions between Trp and positively charged His leading to the quenching of Trp. Furthermore, the pH-dependent change of the Trp residues is observed in the available PDB structures. The NMR structure of GCSF at pH 3.5 (PDB code: 1GNC) revealed that the Trp residues can interact with the neighboring His residues [26]. On the other hand, the X-ray structure obtained at pH 7.5 (PDB code: 1CD9), shows that Trp59 points away from His157. In contrast to 1CD9, Trp59 is pointing upwards in 1GNC (Fig. 3), indicating that the conformation of the

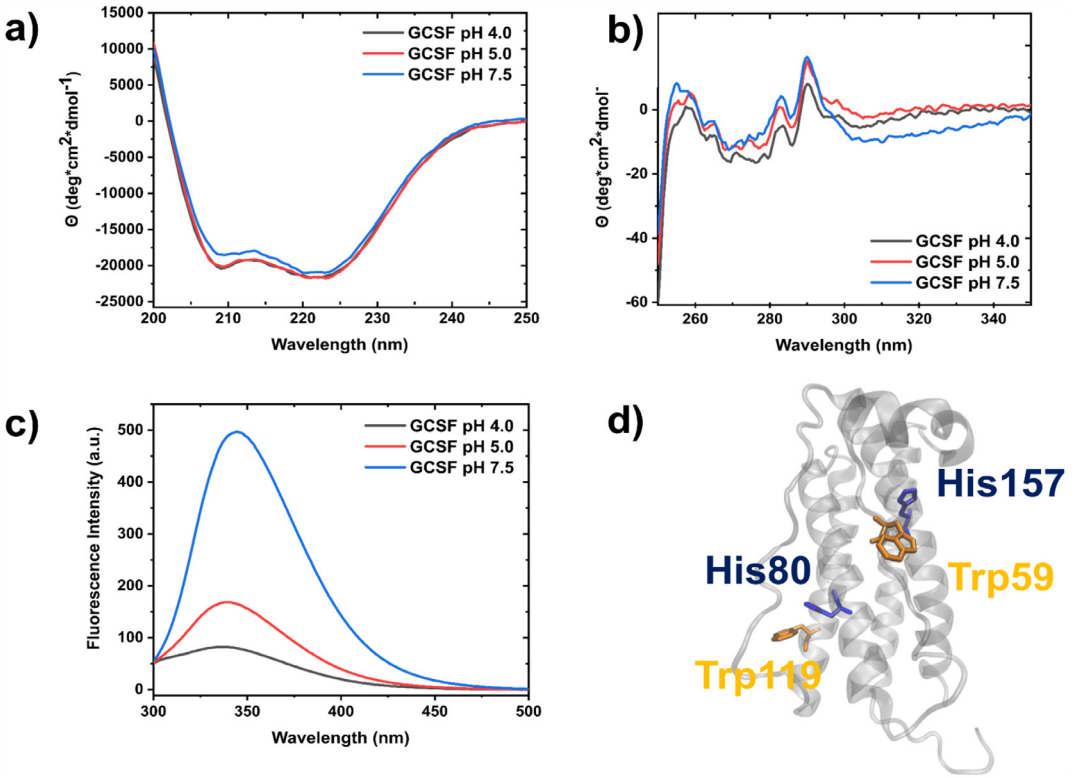


Fig. 2. Effect of pH on the GCSF secondary structure studied with a) far-UV circular dichroism; and on the GCSF tertiary structure studied with b) near-UV circular dichroism. c) fluorescence intensity measurements which indicate that the Trp residues in GCSF are quenched at pH 4.0. d) The location of Trp and His residues in GCSF (PDB code: 1CD9). The protein is shown in a transparent cartoon structure. The investigated Trp and His residues are shown as sticks and colored in orange and blue, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Trp residues is dependent on pH. Based on these findings, we hypothesized that the Trp configuration is an important factor influencing the conformational stability of GCSF.

To further investigate the effect of Trp configurations on the GCSF structural integrity, we performed cMD simulations with the crystal structure 1CD9 as the starting structure. The simulations were carried out for 400 ns. During the simulations, no significant conformational changes of the Trp residues could be observed. Presumably, 400 ns cMD simulations were not sufficient to induce noticeable structural changes. Therefore, we continued with well-tempered metadynamics simulations where bias potentials are added as a function of the center of mass (COM) distances between Trp and His side chains. To check the overall conformational stability, the α-helical content was chosen as the third CV. The 2D and 3D free energy surfaces (FES) of the CVs are shown in Fig. 4, and Fig. S3 respectively. The time evolution of the FES is provided in Figure S1-S2.

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

Contrarily, it is difficult to separate His80-Trp119 more than 1 nm, and only one local minimum could be found from the

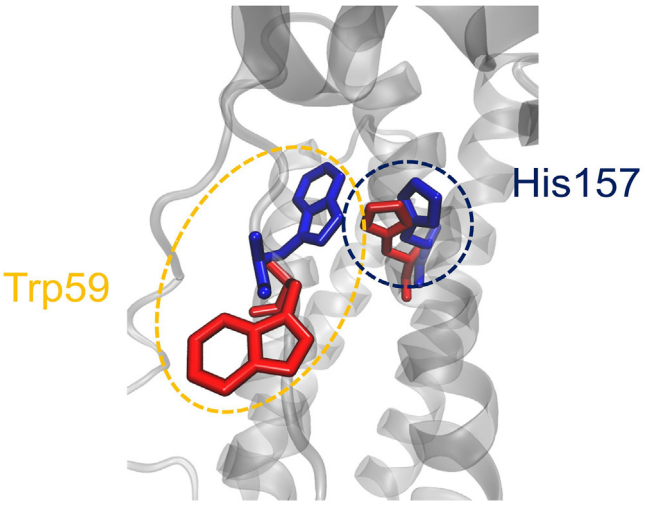


Fig. 3. The conformational change of Trp59. The sidechain structure of GCSF at pH 4.0 (PDB code: 1GNC) is colored blue. The sidechain structure at pH 7.0 (PDB code: 1CD9) is colored red. Note, Trp59 forms an upward configuration at pH 4.0. The protein is shown using in transparent cartoon structure. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

COM distance. At pH 7.5, the COM distance of His80-Trp119 remained around 0.7 nm. At lower pH, the COM distance of His80-Trp119 remained around 0.5 nm. In addition, relatively larger fluctuation could be observed at pH 7.5 compared to lower pH, indicating that even though His80-Trp119 are located close to each other, they are not able to form a strong cation-π-interaction, since there is no cation at pH 7.5 (i.e., neutral His).

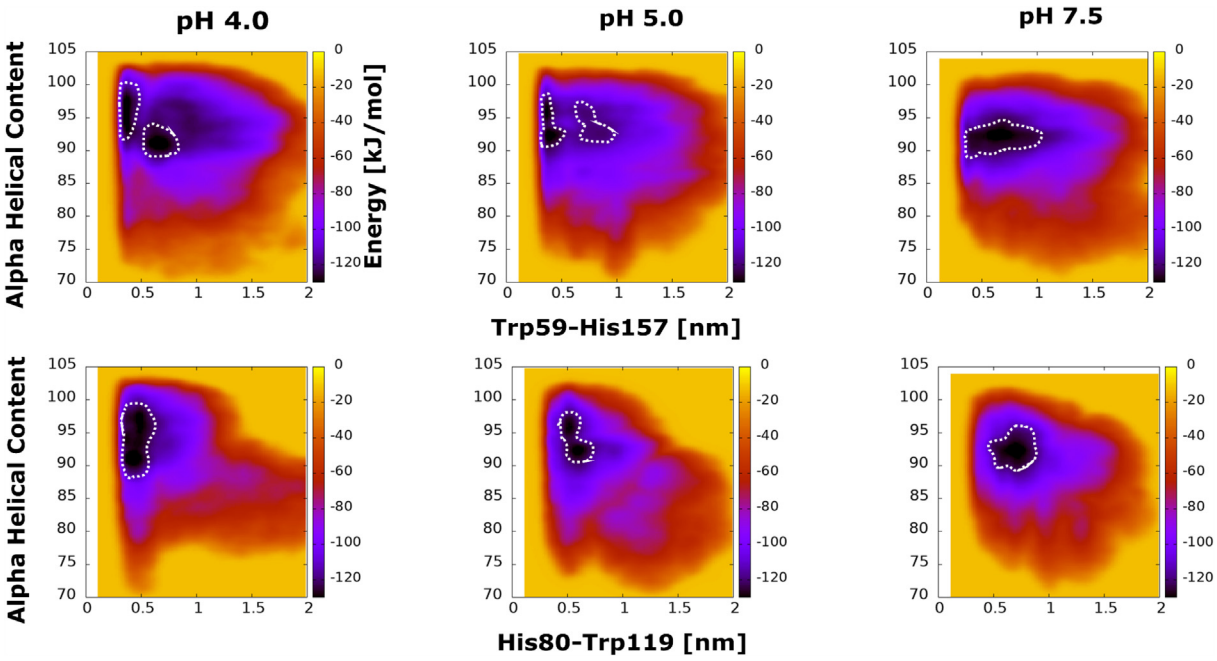


Fig. 4. Estimate of the FES of GCSF at different pH values. Each energy surface is obtained as a function of the α-helical content and the distance between His and Trp residues. The local minima are highlighted with dashed white circles. Top panel: The COM distance between Trp59 and His157 is on the x-axis. Bottom panel: The COM distance between His80 and Trp119 is on the x-axis.

The histidine residues located closely to the Trp residues in GCSF will be protonated at low pH. The FES has shown that the interaction between Trp and its neighboring His residue is much more favorable at pH 4.0. This interaction between Trp and His residues may stabilize GCSF at pH 4.0 compared to pH 7.5 by clamping loopAB to helix αD and helix αB to helix αC, therefore making the structure locally less flexible. Aubin et al. investigated the interactions between Trp and His residues at pH 4.3, 5.0, and 6.4 using NMR [20]. Based on chemical shift analysis, the authors could show that changes in Trp-His interactions affect the conformational stability of GCSF [20]. In addition, Ghasriani et al. have determined the relaxation parameters of GCSF using NMR spectroscopy and assessed the protein flexibility from the calculated order parameters [41]. The authors found that the main difference between pH 4.0 and 6.0 was due to the change in the loop and helical flexibility. The authors observed that the flexibility of loopCD was increased at pH 4.0, whereas a very slight increase of flexibility was seen for loopAB at pH 4.0. The π-cation interaction between Trp59 and His157 can prevent an increase of loopAB mobility. In contrast, a decrease of flexibility could be observed for the helical packing at pH 4.0, and the authors suggested that the His80-Trp119 interaction can be the factor that is involved in reducing the flexibility. On the other hand, Wood et al. have reported an increase of the loopCD deuterium uptake at pH 7.40 [19]. One of the challenging parts of the experimental characterization of GCSF above pH 6.2 is that an extensive aggregation can occur in the sample[19]. Compared to experiments, the metadynamics simulations (performed on a single GCSF molecule) provide an option to study protein conformation in highly aggregation-prone physicochemical conditions without the interference of protein-protein interactions. The current FES study focused on the CVs that can be directly observed in the fluorescence intensity and CD measurements (Fig. 2), and since the FES are based on a few local CVs, the magnitude of flexibility may be dependent on the choice of CVs.

The observation made from HDX [19] and NMR [20] experiments is in good agreement with our metadynamics simulations, where the overall α-helical content is not significantly affected by adjusting the pH. The MD study indicates that the interactions between Trp and His residues may affect the local structural conformation and loop mobility. The interactions between loopAB-αD (Trp59-His157) and αB-αC (His80-Trp119) are lost at pH 7.5, suggesting that GCSF will be more flexible at pH 7.5.

3.2. Effect of pH and sodium chloride on the thermal unfolding and aggregation of GCSF

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

The colloidal stability of GCSF in all tested conditions was assessed by means of the interaction parameter k_D which describes the interaction of proteins in solution [56] (Fig. 5, Table 3). k_D is commonly used as a surrogate parameter for the osmotic second virial coefficient B_{22}, which is directly related to PPIs, whereas k_D provides a less direct relationship. In general, positive k_D values indicate net repulsive PPIs, and negative values correspond to net attractive interactions. However, the reversal does not occur exactly at zero. The excluded volume contribution to k_D is smaller

Table 2
T_m and T_{agg} of GCSF were determined with the PR. ThermControl software from the thermal unfolding curves and the increase in the signal from the backreflection of the nanoDSF measurements. T_m and T_{agg} are mean of triplicates with standard deviations. (NA- no detection of aggregates. At pH 4, no aggregates were detected.).

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

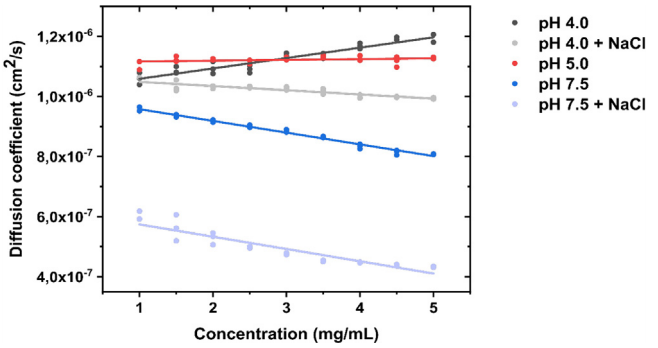


Fig. 5. Diffusion coefficients at increasing protein concentrations assessed with DLS at pH 4.0, 5.0, and 7.5 with and without the addition of 100 mM NaCl.

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

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

than for $B_{22,}$ and therefore values of k_D can be negative when B_{22} values are still positive.

There is considerable variation in the *y*-intercept, i.e. the diffusion coefficient at infinite dilution, for the measured conditions which could be due to the formation of irreversible species which do not dissociate upon dilution or due to protein conformational changes. Since we could not observe large conformational changes but a tendency to form aggregates in the other methods, we assume that the samples contained irreversible aggregates. To confirm this hypothesis, we used the Stokes-Einstein relation to calculate the infinite dilution values for the hydrodynamic radius $r(H)_0$ from the infinite dilution diffusion coefficients (D_0) for each condition (Table 3). The $r(H)_0$ values range from 2.1 up to 4.0 nm, whereas the reported value is 2.0 nm [57], which confirms the presence of larger species in our samples. This in turn impedes the correct determination of the diffusion interaction parameter k_D. Additionally, the partial specific volume of the protein is expected to be a function of pH and could significantly contribute to differences in k_D. However, the partial specific volume should only change upon unfolding which could neither be observed in CD measurements nor MD simulations. Therefore, we do not expect the partial specific volume to have a drastic effect on the k_D values. To support this hypothesis, we submitted the last frames of the conventional all-atom MD simulations at the respective pH values to the HullRad webserver, which calculates the partial specific volume of a protein from a PDB structure. A partial specific volume of 0.75 mL/g was calculated for all three pH values.

GCSF shows a positive k_D and repulsion at pH 4.0 which is in agreement with the proposed highly positive electrostatic surface at low pH (Fig. 6). The addition of salt screens the surface charges of the protein resulting in a negative k_D. These observations correlate well with the strongly decreased thermal stability at low pH upon addition of salt. A k_D of almost zero could be observed at pH 5.0, which indicates no strong attractive nor repulsive forces between the protein monomers. This behavior is expected at a pH close to the isoelectric point where the protein has (almost) no net charge. Due to the very high level of the aggregation, the k_D could not be measured when salt was added to the pH 5.0 formulation. This result is in accordance with the observations from Chi et al. [16]. The authors used static light scattering experiments to obtain the osmotic second virial coefficient (B_{22}) value. A positive and negative B_{22} value could be determined at pH 3.5 and pH 6.1, respectively [16]. Aggregation of GCSF occurred when 150 mM of NaCl was added to the formulation, and B_{22} value could not be determined due to the precipitation [16]. The negative k_D value at pH 7.5 suggests that the GCSF monomers attract each other.

This is in accordance with the surface potential of GCSF which is highly pH-dependent (Fig. 6) as GCSF contains a relatively large number of charged residues. The net charge of GCSF at pH 4.0 is estimated to +13 e using *PDB2PQR* [30]. The electrostatic surface of the helical bundle is highly positively charged. Therefore, it is expected that GCSF will be repulsive at pH 4.0. At pH 5.0, the net charge of GCSF is decreased to +1 e, and it becomes −4 e at pH 7.5. Hence, electrostatic interactions play a substantial role in the aggregation process of GCSF.

In order to simulate the aggregation behavior of GCSF at the different pH values, we performed CG simulations with eight monomers in a pH series. The snapshots of the first 500 ns simulation before the aggregation are shown in Fig. S4. The aggregation behavior of GCSF at different conditions was estimated by tracking the number of protein-protein interactions (PPIs) during the simulations (Fig. 7). The number of PPIs during the 3 μs of the production run was defined as the number of observed intermolecular residue pairs with pair distance less than 4 Å.

Since the total number of the PPIs is highly dependent on the simulation time and the size of the simulation box, the number of interactions is normalized by the number of interactions obtained at pH 4.0. Addition of NaCl or increasing the pH value to 5.0 resulted in a 2-fold increase of the sampled PPIs compared to pH 4.0. A 2.5-fold increase of the PPIs could be observed at pH 7.5 (Fig. 7a). The 2-fold increase is following the trend that was observed for the k_D data (Table 3), and the SIRAH FF model could reproduce the increase of PPIs at the aggregating conditions. Fig. 7b shows the time evolution of the relative number of PPIs during the simulations. It is interesting to note that the total number of PPIs increases with simulation time, indicating that the overall tendency is an aggregation (irreversible oligomer formation) rather than an association (reversible oligomer formation). An increase in the relative number of PPIs can also be observed at pH 4.0. However, this is not completely surprising since the CG simulations were performed at relatively high protein concentrations (approximately 30 mg/mL) to reduce the computational time for sampling PPIs. Interestingly, the slope of the time evolution of the PPIs is different at each pH. At pH 7.5, a much faster increase of the relative number of PPIs is observed when compared to pH 5.0. Since the number of monomers is limited to 8, the relative number of PPIs at 3 μs is very similar for pH 5.0 and 7.5, indicating that the difference in the observed relative number of PPIs will also be dependent on the simulation time. Accuracy and performance will

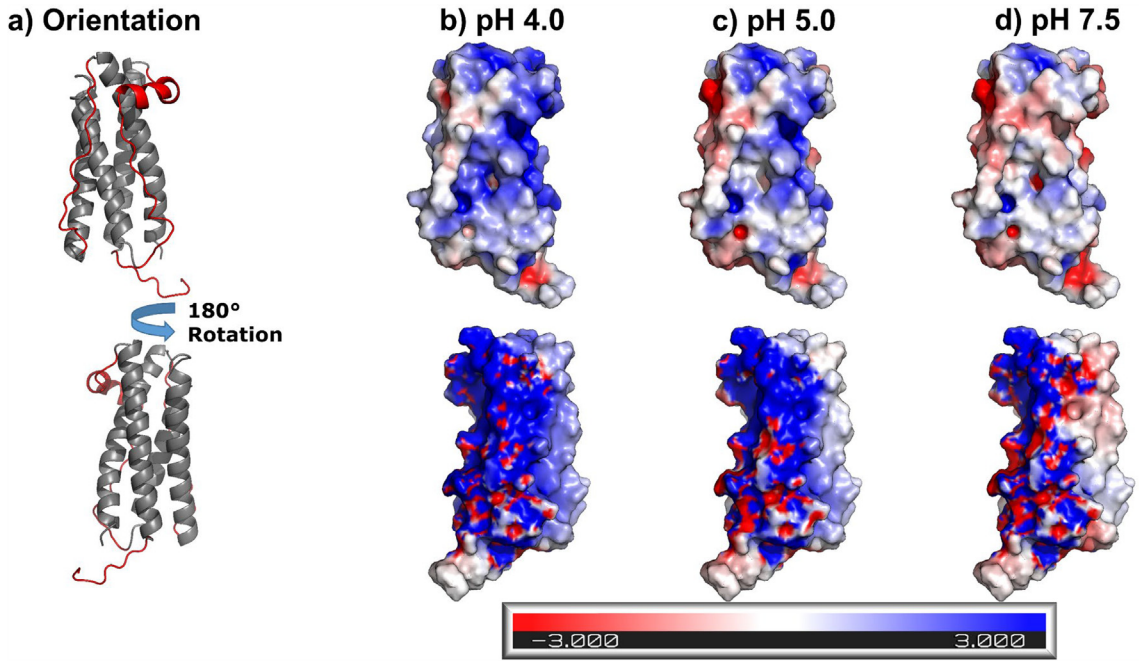


Fig. 6. Electrostatic surface properties of GCSF at different pH values. a) Orientation of the structure corresponding to the orientation of the electrostatic surfaces in b)-d). The flexible N-terminus, loopAB, and loopCD are colored in red. Top: The region containing loopAB and loopCD is on the front view. Bottom: The helical bundle without any long loop structures is on the front view. b)-d) The electrostatic potential surface at different pH values was calculated using the *APBS electrostatics plugins* [58] in PyMOL [59] and *PDB2PQR* [30]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

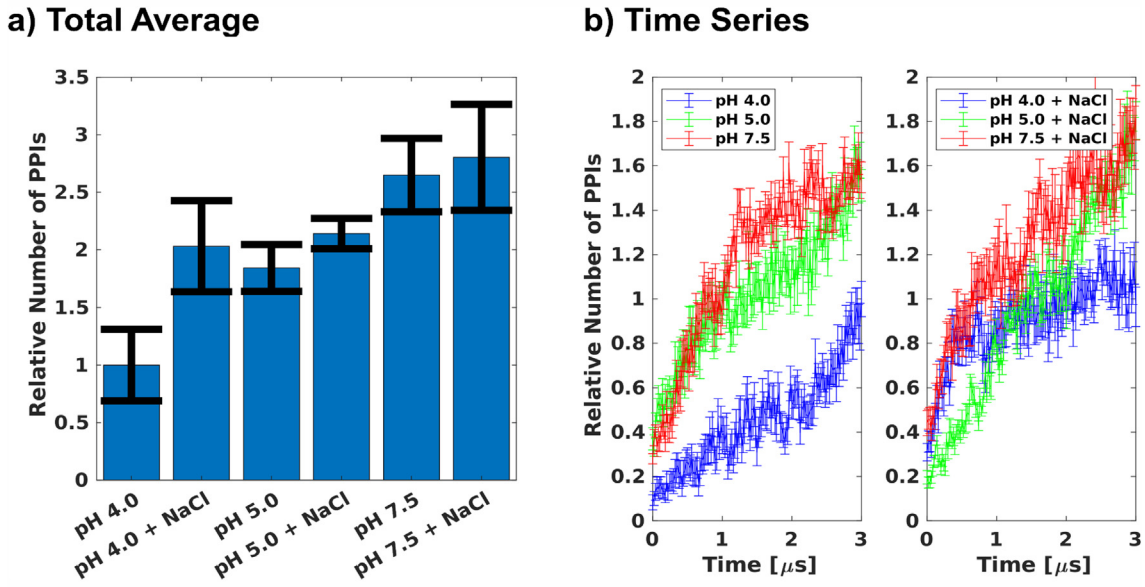


Fig. 7. The relative number of protein-protein interactions (PPIs) obtained from the CG simulations. The number of PPIs during the 3 μs of the production run was defined as the number of observed intermolecular residue pairs with pair distance less than 4 Å. The number of PPIs is normalized to the number of interactions observed in the pH 4.0 simulations. a) The total relative number of PPIs during the simulations. Each error bar represents the mean and the standard error of the mean of the five trajectories performed at each condition. b) Time series of the relative number of PPIs. Each error bar represents the mean and the standard error of the mean of the five trajectories that are observed in the current MD frame.

always be a trade-off when sampling PPIs between diffusing protein monomers. In an ideal case, very long CG sampling with a relatively large box with many protein monomers will give more accurate sampling at the expense of computational time, but it is expected that the results will show a similar tendency observed here.

In the CG simulations, no significant increase of aggregation propensity could be sampled between pH 4.0 + NaCl, pH 5.0, and pH 5.0 + NaCl. The highest aggregation behavior could be observed at pH 7.5. Note that the degree of increase in the PPIs may be dependent on the size of the simulation box and the number of the GCSF monomers. This implies that the relative number of PPIs might change when the simulation conditions are changed. However, the overall trend is expected to be the same.

The results from the CG simulations suggest three different aggregation states: 1) weak aggregation at pH 4.0, 2) moderate

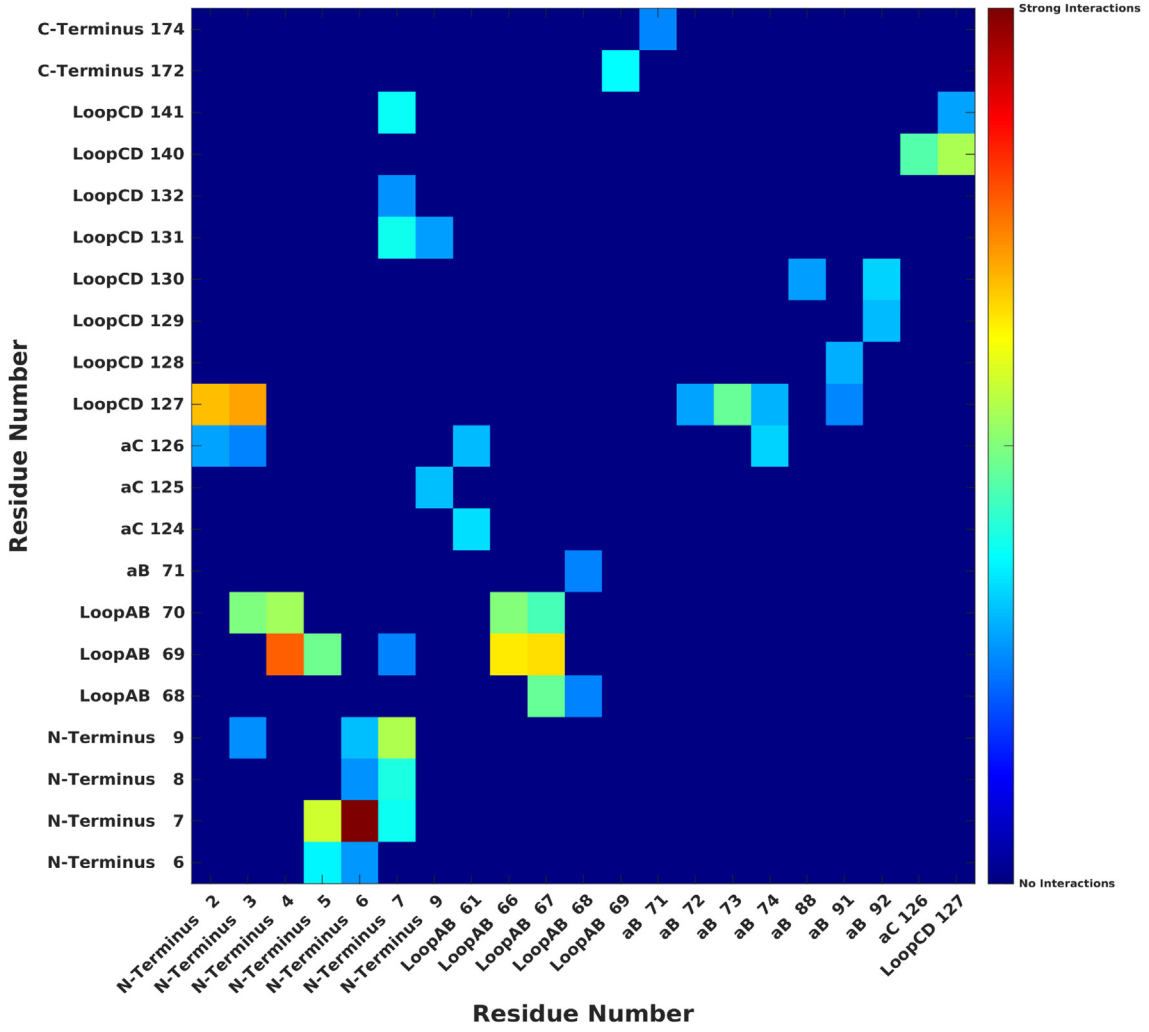


Fig. 8. Protein-protein interaction (PPI) heatmap at pH 4.0. The *x*- and *y*-axes describe the residue number and their secondary structural localization of the interacting residue pair. The interacting residue pair between different monomers from all five simulations are collected into one data set. The color scale indicates the occurrence of the interactions between specific residue pairs. The color bar is scaled to the strongest interaction that occurred in the pH 4.0 simulations; here N-Terminus 6 - N-Terminus 7 interaction.

aggregation at pH 4.0 + NaCl, pH 5.0, and pH 5.0 + NaCl, and 3) strong aggregation at pH 7.5 and pH 7.5 + NaCl.

To characterize the region of the aggregation-prone residues, interacting residue pairs in the CG simulations were depicted in heatmaps. The pixels of the heatmap are assigned by the number of frames with the pair-distance less than 4 Å that was registered from all five trajectories. The color bar is scaled to the strongest interactions that could be observed in the pH 4.0 simulations. Residue pairs were only included in the heatmap if the interactions occurred for more than 25% of the strongest interaction observed in the pH 4.0 simulations. An example of the heatmap is shown in Fig. 8. To visualize the residues that are involved in the PPIs, the residues in the GCSF structure were colored in a similar color scheme as in the heatmaps (Fig. 9).

At pH 4.0, the N-terminal part of GCSF is the main region participating in aggregation (Figs. 8 and 9a). Since GCSF is highly charged at pH 4.0, it is expected that the GCSF monomers will repel each other. However, since the N-terminal part of GCSF does not contain any charged residues, is very flexible and exposed to the solvent, it can still interact with other GCSF monomers. Therefore, it appears that N- to N-terminus interactions may be one of the dominant PPIs at pH 4.0 (Figs. 8 and 9a). Shibuya et al. studied the colloidal stability of the backbone circularized GCSF, i.e. the N- and C-termini of GCSF are connected. Their study revealed that backbone circularization of GCSF at pH 4.0 leads to a more aggregation-resistant GCSF when a protein denaturant is added [60]. When NaCl is added to the simulations at pH 4.0, both loopAB and loopCD are participating in the PPI (Fig. 9d), which indicates that the electrostatic repulsion between the GCSF monomers is the main limiting factor of the intermolecular long loop interactions. The aggregation-prone residues at pH 5.0 and pH 5.0 + NaCl are following a similar pattern as seen for pH 4.0 + NaCl (Fig. 9b, d, and e). The relative number of interactions is very similar in these conditions (Fig. 7). This suggests that the aggregation behavior at these conditions mainly originated from the loss of the repulsion between the GCSF molecules, i.e. that colloidal stability plays a larger role than conformational stability.

At pH 7.5, the short helix in loopAB and the bottom part of the helix bundle located close to the N- and C-termini become more prone to aggregation. Addition of NaCl at pH 7.5 has a minimal effect on aggregation which indicates that the screening of electrostatic interactions does not have a noticeable effect on GCSF aggregation at this pH which is in accordance with our experimental data.

We also determined the aggrescan3d (A3D) score and the spatial aggregation propensity (SAP) [61] using PBD structure 1CD9

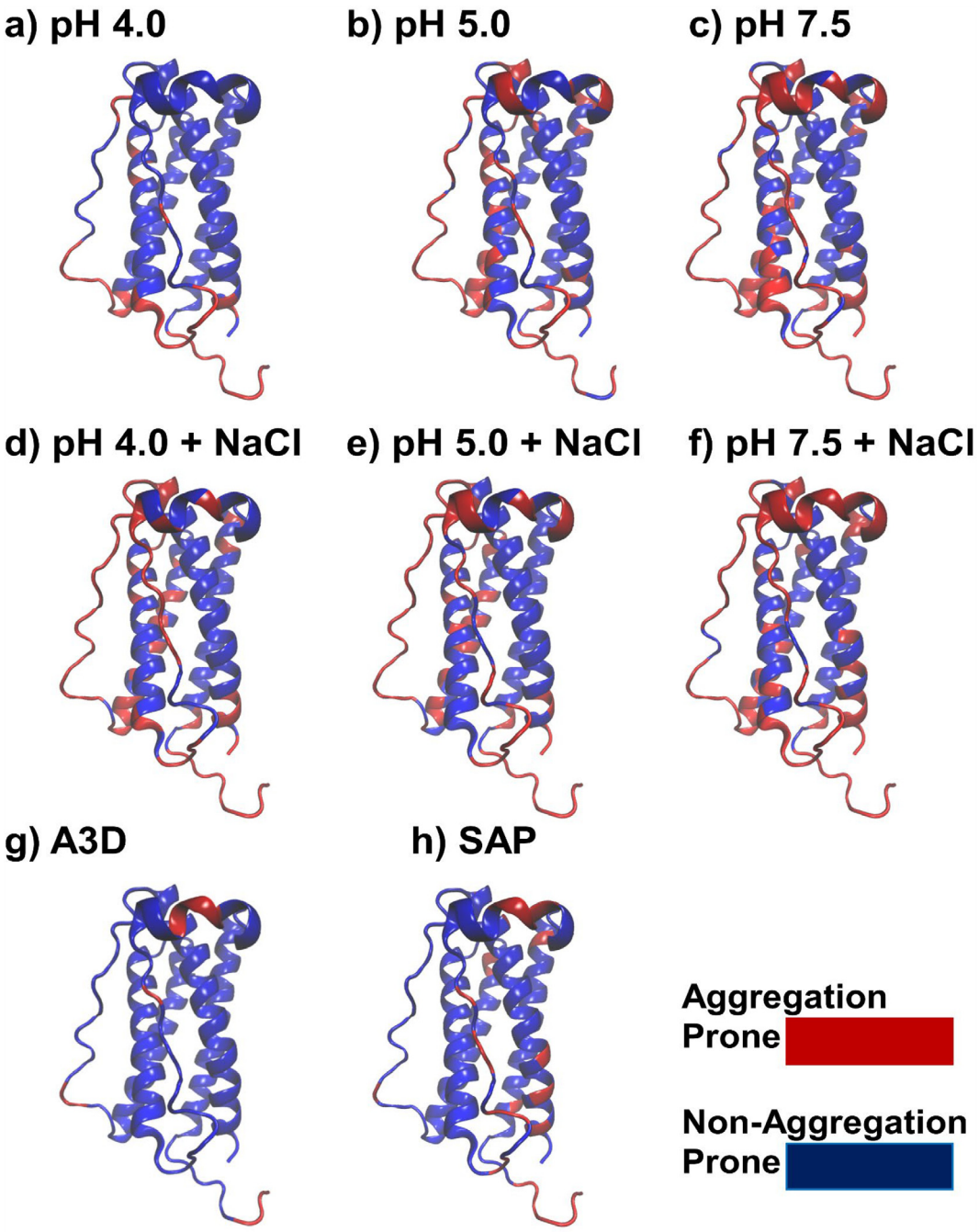


Fig. 9. The aggregation-prone residues determined from the CG simulations and prediction algorithms based on the PDB structure. a)-f) Interacting residues determined from the CG simulations. Aggregation-prone residues that were involved in the PPI in the CG simulations are colored red. Residues not prone to participate in aggregation (blue) interacted 25% or less compared to the strongest interaction at pH 4.0. g-h) Aggregation-prone residues are predicted from the initial PDB structure (PDB code: 1CD9) using aggrescan3d (A3D) and spatial aggregation propensity (SAP). Red residues represent aggregation-prone residues. Blue residues represent non-aggregation prone residues based on A3D-score or SAP lower than 25% of the strongest A3D/SAP score from 1CD9. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(Fig. 9g) to compare the results obtained from relatively fast prediction algorithms with results from computational demanding CG simulations. Interestingly, the A3D/SAP calculations were able to predict the N-terminus and large area of the loopAB region as aggregation-prone regions (Fig. 9g–h). Those aggregation-prone residues follow a similar pattern as seen from the CG simulations. However, CG simulations have the advantage that aggregation-prone regions can be determined in a pH-dependent manner revealing additional aggregation-prone regions.

Observing the overall pattern of the aggregation-prone residues from the CG simulations, it becomes clear that the aggregation mechanism of GCSF is non-specific, e.g. more than one aggregation site exists in GCSF. Previously, Meric et al. used multiple aggregation prediction algorithms and suggested that Leu83 located at αB is the most aggregation-prone residue [62]. However, the authors found that the point mutation Leu83 to Ala did not improve the aggregation propensity of GCSF [62], which is in line with our results that indicate a non-specific aggregation mechanism.

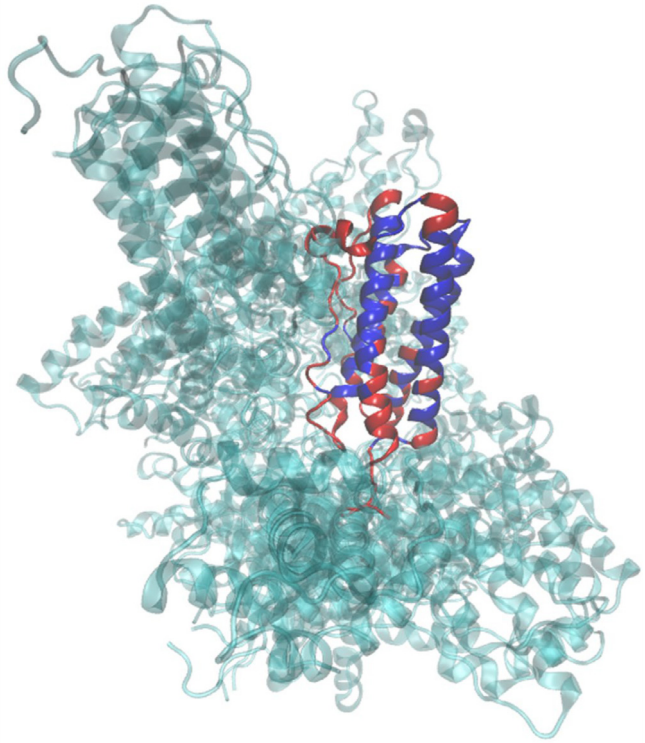


Fig. 10. GCSF aggregation ensemble. For each condition, multiple dimer structures containing several different strong interaction clusters were manually extracted from the CG simulations. Two to four dimers were extracted from each condition. In total 18 dimers were extracted from the CG simulations. The first chain of the extracted dimer was aligned to the reference PDB structure (1CD9). The second chains of the ensemble are shown in transparent structures and colored cyan'. The reference structure is colored according to the scheme used for aggregation-prone residues of pH 7.5 + NaCl (see Fig. 9). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

To extract the important PPIs of the GCSF oligomers, we further analyzed the heatmap (Fig. 8). For each condition, 2 to 4 dimers that contain several different strong interaction clusters (marked in Figs. S6–S11) were manually extracted from the CG simulations. The extracted GCSF dimers are used to provide an aggregation ensemble of GCSF oligomers. The first chain of the GCSF dimers was aligned to the reference PDB structure (1CD9). After the alignments, only the second chain is kept together with the reference structure, mapping the different protein-protein interfaces in GCSF aggregates (Fig. 10).

Various types of dimers can be observed from the aggregation ensemble. Both aggregation ensemble and the simulated aggregation-prone residues suggest that the exposed long loop regions are highly prone to aggregation. Interestingly, the exposed helical structures are not prone to aggregate, suggesting that a combination of electrostatic repulsion and compactness of the helical bundle prevents aggregation of helices. At pH 4.0, the long loop regions show a positively charged electrostatic surface (Fig. 6 top). Since loopAB and loopCD become aggregation-prone when sodium chloride is added or the pH is increased to 5.0, one may argue that electrostatic repulsion of the long loop region is one of the most important factors to avoid the aggregation of GCSF. Our metadynamics simulations suggest an increase of flexibility at pH 7.5 due to the loss of the Trp-His interactions. Since the probability to obtain unfolding of an α-helix at standard conditions is low without adding protein denaturants or heating the system, it suggests that the increase of loop flexibility initiates the aggregation of GCSF.

Since the CG model has a limited atomic resolution, careful consideration is required when interpreting CG simulation results. In an attempt to validate and inspect the aggregation mechanism of GCSF, we compared the CG simulation results to SAXS measurements of GCSF at different pH and NaCl concentrations (Fig. 11). SAXS can be applied to investigate the inter-particle interactions of therapeutic protein [63]. However, it is extremely challenging to model the PPIs in irreversible aggregating conditions using SAXS data [64]. On other hand, SAXS data still provides valuable information when it is combined with the CG simulations since the combination of SAXS and CG simulations enables the direct comparison between computational and experimentally determined aggregation behavior.

The SAXS data indicate that the only non-aggregating condition of GCSF is at pH 4.0, where repulsion between GCSF molecules is observed (Fig. 11a). Since the data measured at the highest concentration is less noisy, it was used for the modeling process.

The aggregation of GCSF was initially investigated by inspecting the low-*q* region of the SAXS data. According to the data shown in Fig. 12, increasing pH and the addition of NaCl lead to an increase in aggregation, which is in agreement with the aggregation profile deduced from the CG simulations (Fig. 7).

In order to investigate the fraction of higher order species of the SAXS data, the dimer fraction of the considered samples was calculated. The obtained molecular weight of GCSF in the aggregating conditions was between the molecular weight of monomer and dimer (supplementary data Table S2). Therefore, we decided that in the modeling step, only monomer and/or dimer will be included (i.e., no larger oligomers). The following modeling approaches were applied: i) fitting a dimer structure that was obtained from rigid-body modeling using *SASREFMX* [53] with two high-resolution monomer structures (1CD9) as an input (see Fig. 13a) and ii) fitting of the monomer (ICD9) and the dimer structures that were extracted from the CG simulations (see Fig. 10) using *OLIGOMER* [54] (see Fig. 13b). Both models assume that the scattering data are from the mixture, meaning that both monomer and dimer structures will be fitted to the SAXS data simultaneously. Furthermore, *CRYSOL* [52] was used to fit the stand-alone monomer structure that is obtained from 1CD9. To validate the dimer structure, the χ^2 value of the monomer fitting was compared to the outcomes of the dimer fitting. The obtained dimer fraction is shown in Fig. 13.

Overall, monomer + dimer has a better fit on the experimental data than the monomer only fit, meaning that both are present at all studied conditions. (Fig. 14). The rigid-body modeling approach had higher freedom to create the dimer structure to obtain an optimal fitting. Therefore, the result obtained from *SASREFMX* usually gave a better fit (Fig. 14b). However, one must note that the dimer structure that is generated from the rigid body modeling may not have a physically realistic protein-protein interface. The structures obtained from the rigid body modeling can be found in the supplementary (Fig. S18).

One interesting outcome is that the dimer fraction from the SAXS modeling (Fig. 13) follows a similar trend as obtained from the CG simulations (Fig. 7). Two entirely different modeling approaches could separate PPIs at 3 different levels: 1) at pH 4.0, 2) pH 4.0 + NaCl, pH 5, pH 5.0 + NaCl, and 3) pH 7.5, pH 7.5 + NaCl.

However, the modeling based on SAXS data (Fig. 13) provided a more pronounced increase of dimer fraction at pH 7.5 compared to our results from the CG simulations (Fig. 7). Note that the increase of dimer fraction can also indicate bigger aggregation species. One must note that the CG simulation results will be dependent on the size of the system, the number of included monomers, and the simulation time. Thus, the CG simulations alone do not provide an accurate description of the level of

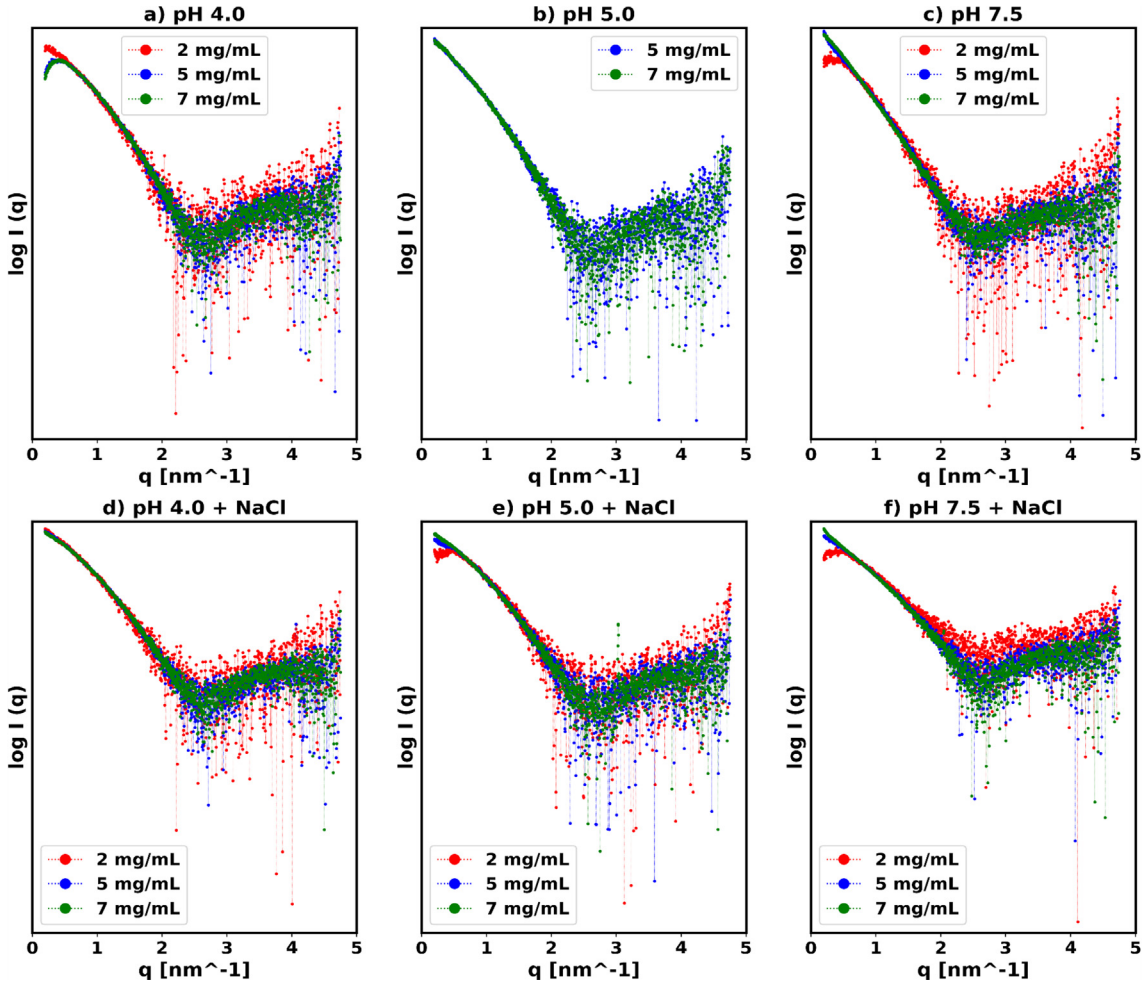


Fig. 11. SAXS scattering curves of GCSF at a) pH 4.0, b) pH 5.0, c) pH 7.5, d) pH 4.0 + 100 mM NaCl, e) pH 5.0 + 100 mM NaCl, and f) pH 7.5 + 100 mM NaCl. The protein concentration range: 2–7 mg/mL. The data set for pH 5.0 2 mg/mL was not included due to technical problems occuring during the measurements.

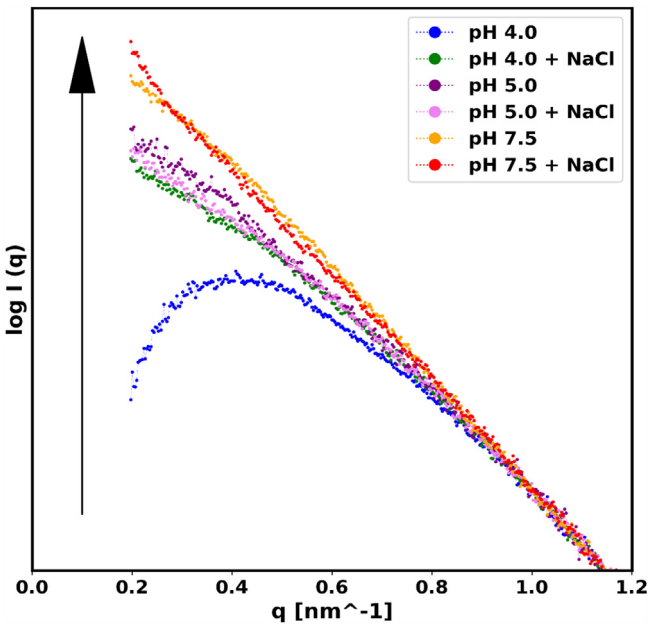


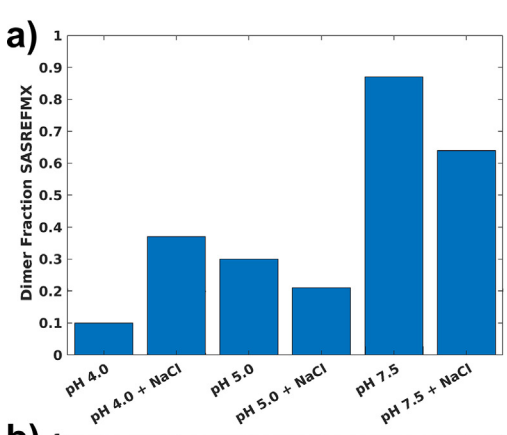
Fig. 12. SAXS scattering profile plotted at the low-*q* region. The arrow illustrates the aggregation of GCSF with increasing pH. The protein concentrations of the samples are 7 mg/mL.

aggregation compared to experimental data. Therefore, the SAXS data was included to predict the level of aggregation in a more physically correct manner. The SAXS data could validate the trend that was obtained from CG simulations. Furthermore, it was possible to propose a possible GCSF dimer structure by combining both CG simulations and SAXS measurements (Fig. 15). The proposed GCSF oligomer structure contains realistic PPIs, and the structure could be directly related to the experimental data.

It is worthwhile to mention that the *ab-initio* models in Fig. 15 have different shapes, which indicates that the amount of aggregating species are different at the different pH conditions, since the *ab-initio* model describes an averaged protein shape in solution. The selected dimer structure suggests that the N- to N-terminus interactions are dominant at pH 4.0, and that the long loops are involved in the aggregation at pH 5.0 and pH 7.5. We propose that the SAXS models can serve as an extension to the CG aggregation model of GCSF, where the SAXS models can be used to provide a bridge between the CG modeled GCSF and the real system.

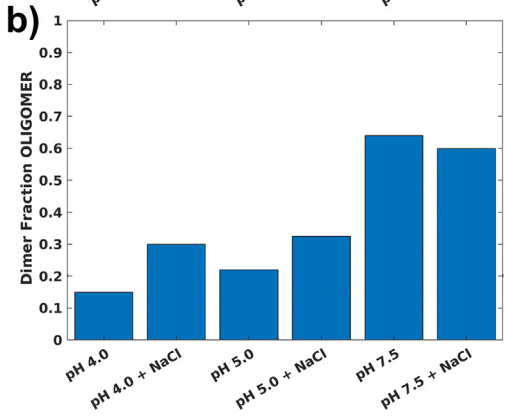
3.3. Future perspectives and potential challenges

Our study on GCSF demonstrates that the application of orthogonal techniques can provide a molecular understanding of the driving forces for PPIs. Since soluble aggregates are usually transient,



SASREFMX	χ^2
pH 4.0	0.65
pH 4.0 + NaCl	0.80
pH 5.0	0.54
pH 5.0 + NaCl	1.7
pH 7.5	0.88
pH 7.5 + NaCl	0.64

CRYSOL	χ^2
pH 4.0	1.14
pH 4.0 + NaCl	2.16
pH 5.0	1.11
pH 5.0 + NaCl	2.58
pH 7.5	7.60
pH 7.5 + NaCl	4.56



OLIGOMER	χ^2
pH 4.0	0.74
pH 4.0 + NaCl	1.13
pH 5.0	0.65
pH 5.0 + NaCl	1.58
pH 7.5	0.99
pH 7.5 + NaCl	0.94

Fig. 13. Analysis of the dimer fractions of GCSF at different pH conditions. a) Dimer fitting using *SASREFMX* [53], and b) Dimer fitting using *OLIGOMER* [54]. The χ^2 of *CRYSOL* [52] is generated from the fitting of the GCSF monomer structure. The modeling was performed with SAXS data obtained for protein concentration 7 mg/mL.

heterogeneous and present at very low concentrations, an ideal experimental technique would be able to simultaneously quantify the size and conformation of each species in a sample without immobilization or labeling. However, the applied experimental techniques in this work are measuring averages of all species present so that species with a low population are being neglected. In some cases, the presence of (irreversible) aggregates even impedes the data analysis, for example in the determination of k_D and the modeling of the SAXS data. The MD simulations allowed us to determine pH and ionic strength-dependent changes on conformation and PPIs at an atomistic level, yet needed experimental validation. Applying multiple techniques, the shortcomings of each technique can be compensated.

Due to its distinct pH-dependent behavior, GCSF is particularly suited as a model protein for this study. However, not all proteins show such dramatic pH-dependent differences and it has to be seen if the computational approaches are sensitive enough to distinguish the aggregation behavior. Nevertheless, we propose that our approach could be extended to other proteins/systems. The challenge will be as the protein size and complexity of the system increase (e.g., glycosylation or by including excipients in the simulations), the computational cost will increase. However, with the increasing processor and network technology performance, it will become feasible to simulate systems with higher complexity. Including excipients in simulations may also require additional force field development of these molecules in the coarse-grained presentation. CG simulations have been applied to investigate self-interactions of antibodies [66,67], where the antibody is usually coarse-grained to much smaller beads (6–12 beads) [66], or the self-association is monitored by simulating only two antibodies [67].

4. Conclusions

We have demonstrated that multiple approaches are required to shed light on the aggregation process of GCSF. Full atomic simulations have shown that it was very difficult to capture the conformations of GCSF in an unbiased system. The metadynamics study revealed that one of the most distinct conformational changes of GCSF at the different pH values occurs due to the loss of cation-π interactions between Trp and the neighboring His residues. The α-helix stability is not noticeably affected by pH, indicating that helical destabilization is not the main issue in standard formulations. However, the results show that the loss of Trp59-His157 and His80-Trp119 interactions will cause a local perturbation that may contribute to the increased flexibility of GCSF at higher pH values. The CG simulations could provide the pH-dependent aggregation-prone regions of GCSF, which were in accordance with the SAP results. The predicted aggregation-prone regions are the N-terminal region and the two long loops parts of GCSF. We have inspected the electrostatic surface to explain the pH-dependent change in the aggregation promoting regions and found that the long loop regions are repulsive at pH 4.0 due to the positively charged surface potential. The addition of salt or increase in pH will make GCSF more aggregation-prone since it will reduce the electrostatic charge located closely to the highly aggregation-prone loop regions. Inspecting the dimer struc-

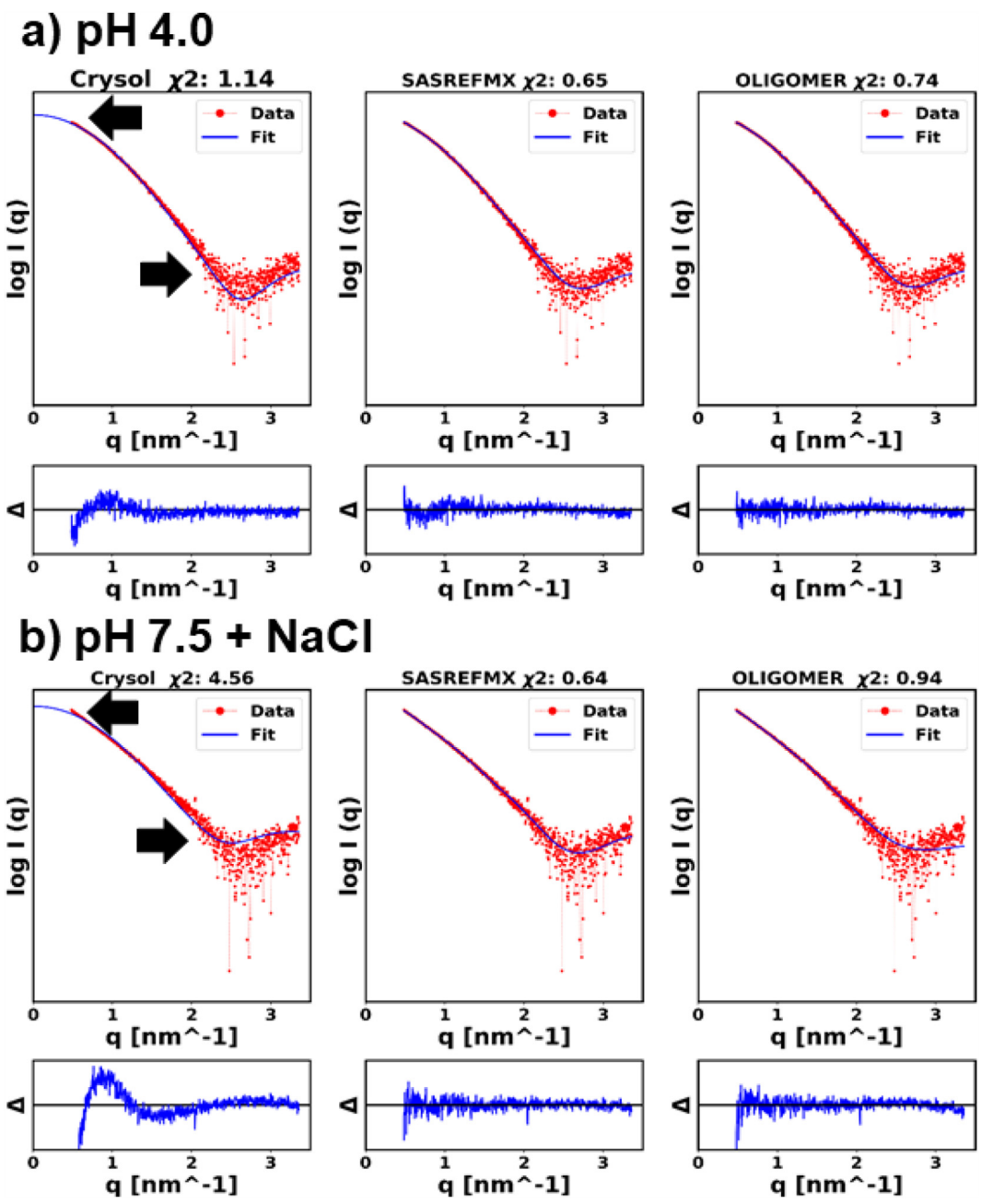


Fig. 14. The results of the different modeling techniques. Two conditions are chosen as examples: a) non-aggregating pH 4.0, and b) highly aggregating pH 7.5 + NaCl. The arrows indicate a misfit at the low-*q* region. The difference between the fit and scattering profile (Δ) is plotted below the fittings. The horizontal line indicates Δ = 0. All *y*-axes are scaled to the same arbitrary chosen range. The modeling was performed with SAXS data obtained for protein concentration 7 mg/mL.

ture from the CG simulations, we observe that the α-helical structures are not participating in the aggregation and that the aggregation of GCSF is highly unspecific where multiple forms of GCSF dimer can exist. The CG simulations lack atomic resolution, and it is not feasible to simulate the GCSF aggregation in a physically realistic size scale. To overcome this problem, we have included SAXS data for validation and interpretation of the simulation outcomes. We could obtain a reasonable fitting by including the dimer structures extracted from CG simulations during the modeling based on the SAXS data. The dimer fraction from SAXS data and the number of interactions from CG simulations followed a similar trend. Since both modeling and experiments of the protein aggregation process are extremely challenging, it requires a combination of multiple approaches to compensate for the weakness of each. We found that the combination of various modeling approaches could shed light on the complex pH-dependent aggregation process of GCSF.

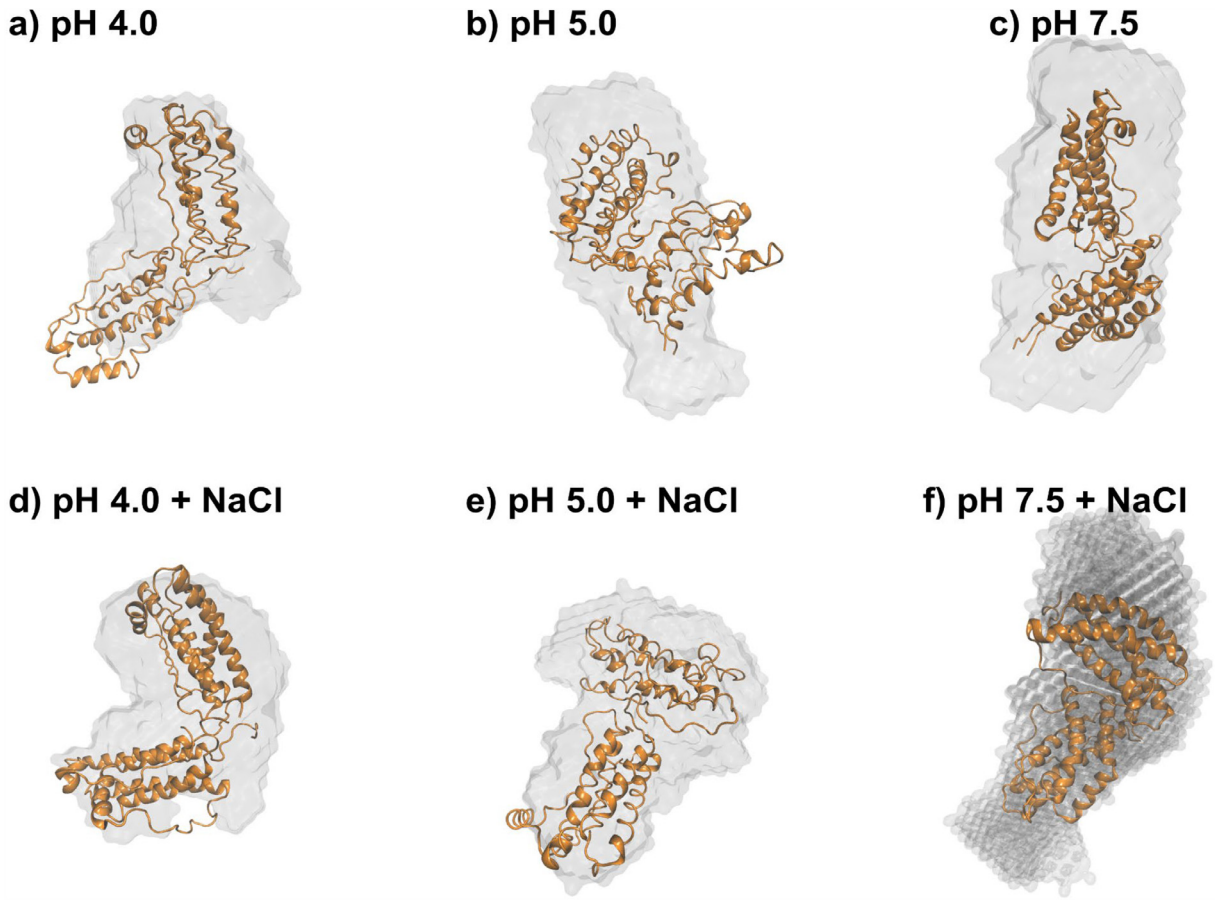


Fig. 15. Dimer models of GCSF at different conditions combining dimers extracted from CG simulations and SAXS measurements. Note, the fitting has been performed using monomer and dimer mixtures. The simulated CG dimer structures were back-mapped and then fitted to the SAXS data at the corresponding condition. The dimer with the χ^2 value closest to 1 was selected. All dimer models selected from the CG simulations are aligned to the *ab-initio* model of the corresponding SAXS data (gray envelope) using *DAMMIF* [65]. However, the GCSF SAXS data showed a certain fraction of higher order species, and therefore, the interpretation of the particle shape derived from the SAXS data must be assessed with caution.

CRediT authorship contribution statement

Suk Kyu Ko: Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Carolin Berner:** Conceptualization, Methodology, Investigation, Writing – review & editing, Visualization. **Alina Kulakova:** Methodology, Writing – review & editing. **Markus Schneider:** Methodology, Writing – review & editing. **Iris Antes:** Methodology, Resources, Writing – review & editing, Supervision. **Gerhard Winter:** Conceptualization, Methodology, Resources, Writing – review & editing, Supervision. **Pernille Harris:** Methodology, Resources, Writing – review & editing, Supervision. **Günther H.J. Peters:** Conceptualization, Methodology, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of interest

The authors declare that there are no known conflicts of interest. The financial support of the research or personal relationships do not have any significant impact on the outcomes reported in this works.

Acknowledgments

We thank the European Synchrotron Radiation Facility for providing beam time for performing the SAXS experiments and Beamline Operator Manager Michael Sztucki for the measurements at beamline ID02. We thank Sergio Pantano for providing the SIRAH parameters for the protonated histidine. The simulations were carried at the High-Performance Computing (HPC) cluster at DTU. SKK received a Ph.D. scholarship funded through the DTU Alliance program. MS was supported by the TUM International Graduate School of Science and Engineering (IGSSE). In memory of Prof. Dr. Iris Antes who unfortunately passed away during the duration of this work on Aug. 4, 2021.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.03.012.

References

[1] de la Torre BG, Albericio F. The pharmaceutical industry in 2019. An analysis of FDA drug approvals from the perspective of molecules. Molecules 2020;25:745.
[2] H Tobin P, H Richards D, A Callender R, J Wilson C. Protein engineering: a new frontier for biological therapeutics. Curr Drug Metab 2014;15:743–56.
[3] Manning MC, Patel K, Borchardt RT. Stability of protein pharmaceuticals. Pharm Res 1989;6:903–18.
[4] Manning MC, Chou DK, Murphy BM, Payne RW, Katayama DS. Stability of protein pharmaceuticals: an update. Pharm Res 2010;27:544–75.
[5] Roberts CJ. Therapeutic protein aggregation: mechanisms, design, and control. Trends Biotechnol 2014;32:372–80.

[6] Chi EY, Krishnan S, Randolph TW, Carpenter JF. Physical stability of proteins in aqueous solution: mechanism and driving forces in nonnative protein aggregation. Pharm Res 2003;20:1325–36.
[7] Clarkson BR, Schön A, Freire E. Conformational stability and self-association equilibrium in biologics. Drug Discov Today 2016;21:342–7.
[8] Wang W. Protein aggregation and its inhibition in biopharmaceutics. Int J Pharm 2005;289:1–30.
[9] Zalar M, Svilenov HL, Golovanov AP. Binding of excipients is a poor predictor for aggregation kinetics of biopharmaceutical proteins. Eur J Pharm Biopharm 2020;151:127–36.
[10] Hill CP, Osslund TD, Eisenberg D. The structure of granulocyte-colony-stimulating factor and its relationship to other growth factors. Proc Natl Acad Sci U S A 1993. https://doi.org/10.1073/pnas.90.11.5167.
[11] Weiss M, Voglic S, Harms-Schirra B, Lorenz I, Lasch B, Dumon K, et al. Effects of exogenous recombinant human granulocyte colony-stimulating factor (filgrastim, rhG-CSF) on neutrophils of critically ill patients with systemic inflammatory response syndrome depend on endogenous G-CSF plasma concentrations on admission. Intensive Care Med 2003. https://doi.org/10.1007/s00134-003-1734-y.
[12] Welte K, Gabrilove J, Bronchud MH, Platzer E, Morstyn G. Filgrastim (r-metHuG-CSF): The first 10 years. Blood 1996. https://doi.org/10.1182/blood.v88.6.1907.bloodjournal8861907.
[13] Aritomi M, Kunishima N, Okamoto T, Kuroki R, Ota Y, Morikawa K. Atomic structure of the GCSF-receptor complex showing a new cytokine- receptor recognition scheme. Nature 1999. https://doi.org/10.1038/44394.
[14] Šali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234:779–815.
[15] Krishnan S, Chi EY, Webb JN, Chang BS, Shan D, Goldenberg M, et al. Aggregation of granulocyte colony stimulating factor under physiological conditions: Characterization and thermodynamic inhibition. Biochemistry 2002. https://doi.org/10.1021/bi012006m.
[16] Chi EY, Krishnan S, Kendrick BS, Chang BS, Carpenter JF, Randolph TW. Roles of conformational stability and colloidal stability in the aggregation of recombinant human granulocyte colony-stimulating factor. Protein Sci 2003. https://doi.org/10.1110/ps.0235703.
[17] Robinson MJ, Matejtschuk P, Bristow AF, Dalby PA. T m-values and unfolded fraction can predict aggregation rates for granulocyte colony stimulating factor variant formulations but not under predominantly native conditions. Mol Pharm 2018;15:256–67.
[18] Narhi LO, Kenney WC, Arakawa T. Conformational changes of recombinant human granulocyte-colony stimulating factor induced by pH and guanidine hydrochloride. J Protein Chem 1991. https://doi.org/10.1007/BF01025250.
[19] Wood VE, Groves K, Cryar A, Quaglia M, Matejtschuk P, Dalby PA. HDX and in silico docking reveal that excipients stabilize G-CSF via a combination of preferential exclusion and specific hotspot interactions. Mol Pharm 2020;17:4637–51.
[20] Aubin Y, Hodgson DJ, Thach WB, Gingras G, Sauvé S. Monitoring effects of excipients, formulation parameters and mutations on the high order structure of filgrastim by NMR. Pharm Res 2015;32:3365–75.
[21] Darré L, Machado MR, Brandner AF, González HC, Ferreira S, Pantano S. SIRAH: a structurally unbiased coarse-grained force field for proteins with aqueous solvation and long-range electrostatics. J Chem Theory Comput 2015;11:723–39.
[22] Machado MR, Barrera EE, Klein F, Sónora M, Silva S, Pantano S. The SIRAH 2.0 Force Field: Altius, Fortius. Citius J Chem Theory Comput 2019. https://doi.org/10.1021/acs.jctc.9b00006.
[23] Barrera EE, Zonta F, Pantano S. Dissecting the role of glutamine in seeding peptide aggregation. Comput Struct Biotechnol J 2021;19:1595–602.
[24] Javanainen M, Martinez-Seara H, Vattulainen I. Excessive aggregation of membrane proteins in the Martini model. PLoS ONE 2017;12:e0187936.
[25] Tamada T, Honjo E, Maeda Y, Okamoto T, Ishibashi M, Tokunaga M, et al. Homodimeric cross-over structure of the human granulocyte colony-stimulating factor (GCSF) receptor signaling complex. Proc Natl Acad Sci 2006;103:3135–40.
[26] Zink T, Ross A, Lüers K, Cieslar C, Holak TA, Rudolph R. Structure and dynamics of the human granulocyte colony-stimulating factor determined by NMR spectroscopy. Loop mobility in a four-helix-bundle protein. Biochemistry 1994. https://doi.org/10.1021/bi00194a009.
[27] Chu J-W, Yin J, Wang DIC, Trout BL. Molecular dynamics simulations and oxidation rates of methionine residues of granulocyte colony-stimulating factor at different pH values. Biochemistry 2004. https://doi.org/10.1021/bi0356000.
[28] Singh SK, Mishra A, Goel G, Chirmule N, Rathore AS. Modulation of granulocyte colony stimulating factor conformation and receptor binding by methionine oxidation. Proteins Struct Funct Bioinforma 2021;89:68–80.
[29] Rospiccio M, Arsiccio A, Winter G, Pisano R. The role of cyclodextrins against interface-induced denaturation in pharmaceutical formulations: A molecular dynamics approach. Mol Pharm 2021.
[30] Dolinsky TJ, Nielsen JE, McCammon JA, Baker NA. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. Nucleic Acids Res 2004;32:W665–7. https://doi.org/10.1093/nar/gkh381.
[31] Case DA, Belfon K, Ben-Shalom I, Brozell SR, Cerutti D, Cheatham T, et al. Amber 2020:2020.
[32] Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. J Chem Theory Comput 2015;11:3696–713.
[33] Horn HW, Swope WC, Pitera JW, Madura JD, Dick TJ, Hura GL, et al. Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. J Chem Phys 2004;120:9665–78. https://doi.org/10.1063/1.1683075.
[34] Darden T, York D, Pedersen L. Particle mesh Ewald: An N log(N) method for Ewald sums in large systems. J Chem Phys 1993;98:10089–92. https://doi.org/10.1063/1.464397.
[35] Ryckaert J-P, Ciccotti G, Berendsen HJC. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J Comput Phys 1977;23:327–41. https://doi.org/10.1016/0021-9991(77)90098-5.
[36] Miyamoto S, Kollman PA. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. J Comput Chem 1992;13:952–62. https://doi.org/10.1002/jcc.540130805.
[37] Pastor RW, Brooks BR, Szabo A. An analysis of the accuracy of Langevin and molecular dynamics algorithms. Mol Phys 1988;65:1409–19.
[38] Åqvist J, Wennerström P, Nervall M, Bjelic S, Brandsdal BO. Molecular dynamics simulations of water and biomolecules with a Monte Carlo constant pressure algorithm. Chem Phys Lett 2004;384:288–94.
[39] Tribello GA, Bonomi M, Branduardi D, Camilloni C, Bussi G. PLUMED 2: New feathers for an old bird. Comput Phys Commun 2014;185:604–13.
[40] Barducci A, Bussi G, Parrinello M. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. Phys Rev Lett 2008;100:20603.
[41] Ghasriani H, Frahm GE, Johnston MJW, Aubin Y. Effects of excipients on the structure and dynamics of filgrastim monitored by thermal unfolding studies by CD and NMR spectroscopy. ACS Omega 2020;5:31845–57.
[42] Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX 2015;1:19–25.
[43] Machado MR, Pantano S. SIRAH tools: mapping, backmapping and visualization of coarse-grained models. Bioinformatics 2016;32:1568–70.
[44] Darré L, Machado MR, Dans PD, Herrera FE, Pantano S. Another coarse grain model for aqueous solvation: WAT FOUR? J Chem Theory Comput 2010;6:3793–807.
[45] Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. J Chem Phys 1984;81:3684–90. https://doi.org/10.1063/1.448118.
[46] Bussi G, Donadio D, Parrinello M. Canonical sampling through velocity rescaling. J Chem Phys 2007;126:14101.
[47] Parrinello M, Rahman A. Polymorphic transitions in single crystals: A new molecular dynamics method. J Appl Phys 1981;52:7182–90.
[48] Kelly SM, Price NC. The use of circular dichroism in the investigation of protein structure and function. Curr Protein Pept Sci 2000;1:349–84.
[49] Harding SE, Johnson P. The concentration-dependence of macromolecular parameters. Biochem J 1985;231:543–7.
[50] Connolly BD, Petry C, Yadav S, Demeule B, Ciaccio N, Moore JMR, et al. Weak interactions govern the viscosity of concentrated antibody solutions: high-throughput analysis using the diffusion interaction parameter. Biophys J 2012;103:69–78.
[51] Franke D, Petoukhov MV, Konarev PV, Panjkovich A, Tuukkanen A, Mertens HDT, et al. ATSAS 2.8: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. J Appl Crystallogr 2017;50:1212–25.
[52] Svergun D, Barberato C, Koch MH. CRYSOL - A program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. J Appl Crystallogr 1995;28:768–73. https://doi.org/10.1107/S0021889895007047.
[53] Petoukhov MV, Franke D, Shkumatov AV, Tria G, Kikhney AG, Gajda M, et al. New developments in the ATSAS program package for small-angle scattering data analysis. J Appl Crystallogr 2012;45:342–50. https://doi.org/10.1107/S0021889812007662.
[54] 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;36:1277–82. https://doi.org/10.1107/S0021889803012779.
[55] Nikravesh FY, Shirkhani S, Bayat E, Talebkhan Y, Mirabzadeh E, Sabzalinejad M, et al. Extension of human GCSF serum half-life by the fusion of albumin binding domain. Sci Rep 2022;12:1–13.
[56] Zhang J, Liu XY. Effect of protein–protein interactions on protein aggregation kinetics. J Chem Phys 2003;119:10972–6.
[57] Jing W, Roberts JW, Green DE, Almond A, DeAngelis PL. Synthesis and characterization of heparosan-granulocyte-colony stimulating factor conjugates: a natural sugar-based drug delivery system to treat neutropenia. Glycobiology 2017;27:1052–61.
[58] Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA. Electrostatics of nanosystems: Application to microtubules and the ribosome. Proc Natl Acad Sci 2001;98:10037–41. https://doi.org/10.1073/pnas.181342398.
[59] Pymol DWL. An open-source molecular graphics tool. CCP4 Newsl Protein Crystallogr 2002;40:82–92.
[60] Shibuya R, Miyafusa T, Imamura H, Ooishi A, Honda S. Effect of backbone circularization on colloidal stability: Compaction of unfolded structures improves aggregation resistance of granulocyte colony-stimulating factor. Int J Pharm 2021;605:120774.
[61] Chennamsetty N, Voynov V, Kayser V, Helk B, Trout BL. Prediction of aggregation prone regions of therapeutic proteins. J Phys Chem B 2010;114:6614–24.

[62] Meric G, Naik S, Hunter AK, Robinson AS, Roberts CJ. Challenges for design of aggregation-resistant variants of granulocyte colony-stimulating factor. Biophys Chem 2021:106630.
[63] Pohl C, Mahapatra S, Kulakova A, Streicher W, Peters GHJ, Nørgaard A, et al. Combination of high throughput and structural screening to assess protein stability–A screening perspective. Eur J Pharm Biopharm 2022;171:1–10.
[64] Kikhney AG, Svergun DI. A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. FEBS Lett 2015;589:2570–7.
[65] Franke D, Svergun DI. DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering. J Appl Crystallogr 2009;42:342–6. https://doi.org/10.1107/S0021889809000338.
[66] Izadi S, Patapoff TW, Walters BT. Multiscale coarse-grained approach to investigate self-association of antibodies. Biophys J 2020;118:2741–54.
[67] Mahapatra S, Polimeni M, Gentiluomo L, Roessner D, Frieß W, Peters GHJ, et al. Self-interactions of two monoclonal antibodies: small-angle X-ray scattering, light scattering, and coarse-grained modeling. Mol Pharm 2021.