molecular pharmaceutics

Article

Combining structural aggregation propensity and stability predictions to re-design protein solubility

Marcos Gil-Garcia, Manuel Baño-Polo, Nathalia Varejao, Michal Jamroz, Aleksander Kuriata, Marta Diaz Caballero, Jara Lascorz, Bertrand Morel, Susanna Navarro, David Reverter, Sebastian Kmiecik, and Salvador Ventura

Mol. Pharmaceutics, **Just Accepted Manuscript** • DOI: 10.1021/acs.molpharmaceut.8b00341 • Publication Date (Web): 23 Jul 2018

ACS Publications

is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

Combining structural aggregation propensity and stability predictions to re-design protein solubility

Marcos Gil-Garcia[1], Manuel Bañó-Polo[1], Nathalia Varejão[1], Michal Jamroz[2], Aleksander Kuriata[2], Marta Diaz-Caballero[1], Jara Lascorz[1], Bertrand Morel[3], Susanna Navarro[1], David Reverter[1], Sebastian Kmiecik[2] and Salvador Ventura[1*]

[1]Institut de Biotecnologia i de Biomedicina and Departament de Bioquímica i Biologia Molecular; Universitat Autònoma de Barcelona; Bellaterra (Barcelona) 08193, Spain.

[2]Biological and Chemical Research Centre, Faculty of Chemistry; University of Warsaw; Warsaw, Poland.

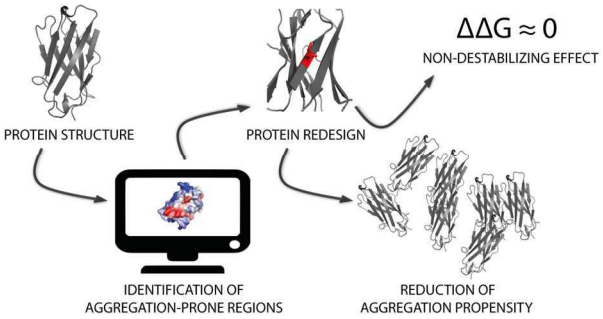
[3]Departamento de Química Física e Instituto de Biotecnología, Facultad de Ciencias, Universidad de Granada, 18071 Granada, Spain.

Author information

Corresponding author: Salvador Ventura

Tel: 93 58689 56; Fax: 93 5812011, E- mail: salvador.ventura@uab.cat

Table Of Contents



Abstract

The aggregation propensity of each particular protein seems to be shaped by evolution according to its natural abundance in the cell. The production and downstream processing of recombinant polypeptides implies attaining concentrations that are orders of magnitude above their natural levels, often resulting in their aggregation; a phenomenon that precludes the marketing of many globular proteins for biomedical or biotechnological applications. Therefore, there is a huge interest in methods aimed to rise proteins solubility above their natural limits. Here, we demonstrate that an updated version of our AGGRESCAN 3D structural aggregation predictor, that now takes into account protein stability, allows designing mutations at specific positions in the structure that improve the solubility of proteins without compromising their conformation. Using this approach, we have designed a highly soluble variant of the Green Fluorescent Protein (GFP) and a human single-domain VH antibody displaying significantly reduced aggregation propensity. Overall, our data indicate that the solubility of unrelated proteins can be easily tuned by *in silico*-designed non-destabilizing amino acid changes at their surfaces.

Keywords

Protein Aggregation; Protein Stability; Protein Structure; Green Fluorescent Protein; Aβ peptide; Single-domain Antibodies.

Introduction

Proteins are molecules involved in a wide range of molecular functions. To perform their functions, globular proteins have to acquire a defined 3D structure, skipping misfolding and side aggregation reactions[1]. Protein aggregation involves the establishment of intermolecular interactions and therefore, it is exquisitely sensitive to protein concentration[2]. Indeed, it appears that the aggregation propensities of natural proteins correlate with their abundances in the cell[3]. This suggests that proteins have been shaped to remain soluble at those concentrations required to sustain function under physiological conditions, according to the "living on the edge" hypothesis[4]. Thus, many proteins might be only marginally soluble above their natural levels[5].

Despite their evident economic interest, there are challenging barriers in the production and manufacture of therapeutic protein-based drugs that preclude the arrival of many of these molecules to the market[6]. Among the various degradation reactions that a protein might experiment during its biotechnological production, aggregation is the most common[7]. This is not surprising, if we think that therapeutic applications require protein concentrations that are several orders of magnitude beyond the cellular levels. Aggregation can occur at every phase of protein-based drugs development, from recombinant expression to packaging and storage of the final product and administration[8]. Aggregation leads to a decrease in protein activity and might elicit an immunological response, having also regulatory implications[9]. Thus, pharmaceutical and biotechnological industries spend large efforts to prevent the eventual aggregation of their protein products[10], in many cases using trial and error assays with different buffers, salts or additives to increase solubility. These approaches are costly

and time-consuming and the adequate conditions for increased solubility may never be found.

In principle, one can modify the amino acid sequence of a globular protein in order to increase its solubility. However, this is not a straightforward task, mainly because the physicochemical properties that allow a globular protein to fold into a stable structure overlap with those triggering aggregation[11]. Indeed, the most aggregation-prone regions in a globular protein are usually found at the hydrophobic core, whose correct packing is a crucial determinant of native structures[12]. This makes difficult to decrease a protein's aggregation propensity without impacting at the same time its stability and conformation[11,13].

On the one hand, mutations that decrease the thermodynamic and/or kinetic stability of a globular protein often facilitate aggregation, favoring the population of partially folded states, which expose to the solvent hydrophobic regions that are otherwise protected in the native folded conformation[11,13,14]. On the other hand, proteins can also aggregate from native or native-like conformations without the need of extensive destabilization and unfolding[15,16], this is especially true when these polypeptides are very concentrated, as in it occurs in most protein biopharmaceutics formulations[10]. In these cases, the key to re-design protein solubility consists in identifying specific positions in protein structures for which mutations that uncouple aggregation from thermodynamic stability can be introduced. Novel algorithms able to analyze the aggregation propensity of folded proteins structures constitute privileged tools to identify such potential amino acids substitutions in a cost- and time-effective manner[9,17,18,19].

In general, mutations at the protein surface are best tolerated that residue changes in the protein interior[20] and, accordingly, surface residues evolve faster than core ones[21]. Surface mutations are more frequent because of their weaker destabilizing effects[22]. In addition, in folded globular proteins, the solubility is strongly influenced by the identity of surface residues[23]. Thus, the identification of point mutations that decrease surface aggregation propensity without destabilizing the overall protein structure appears as a promising avenue to optimize protein solubility. We have recently developed AGGRESCAN 3D (A3D) a structure-based aggregation prediction tool that seems optimal to implement such an approach, since it already predicts the aggregation properties of globular protein surfaces[18]. Here, we coupled these predictions with calculations of protein stability changes upon mutation by incorporating stability predictions using the FoldX empirical force field[24] in the A3D pipeline. The value of this combined prediction scheme is experimentally demonstrated by re-designing the solubility of two unrelated protein folds of biotechnological interest.

Experimental Section

In silico analysis

A3D predictions and virtual mutations were made with the algorithm default parameters: static mode and a distance of aggregation analysis of 10 Å. The "Total Score" value results from the sum of the A3D scores of all individual residues; The "Average score", results from normalizing the "Total Score" relative to the number of total amino acids in the protein[25]. For both scores, the more negative the value, the higher the solubility. We used as inputs the folding reporter GFP (fr-GFP) X-ray structure (PDB: 2B3Q) and a 3D model generated with SwissModel[26] (based on PDB: 5I19) in the case of the DP47 antibody. The impact of mutations on protein stability was initially analyzed using the FoldX force field[24]. Protein structures and their surfaces were represented with Pymol (DeLano Scientific LLC). Prediction of protein stability was implemented in A3D web server[18] and it is automatically invoked when user introduces mutation(s). A3D calculates stability for the input and mutated structure (using FoldX[24]) and report the energy difference between the input and mutated protein in kcal/mol. Therefore, negative and positive values indicate stabilization and destabilization respectively, with respect to the input structure. The report appears under the 'Project details" tab of the A3D web server[18] (see example output in **Figure S9**).

Protein engineering, expression and purification

fr-GFP[27] and Aβ42-GFP[28] were cloned in pET21a and pET28a vectors (Addgene, USA), respectively. Mutations were introduced using the Quickchange mutagenesis kit (Stratagene,USA). *E. coli* BL21 (DE3) competent cells were transformed with the correspondent plasmids.

DP47 VH antibody[29] was cloned in pET12a (Novagen, USA), mutagenesis performed as above, and *E. coli* BL21-Gold competent cells (Stratagene, USA) were transformed with correspondent plasmids.

For the production of fr-GFP and Aβ42-GFP and their variants, transformed *E. coli* BL21 (DE3) cells were grown aerobically in LB medium supplemented with 50 μg/mL kanamycin and 100 μg/mL ampicillin, respectively. Protein expression was induced with 1 mM IPTG and cells grown at 25 ^{o}C for 12 h. For protein purification, cells expressing fr-GFP, GFP/KKK and GFP/DDD were harvested by centrifugation (13,000 rpm for 15 min). After cell lysis by sonication and fractionation, the soluble fraction was collected and injected in a *Histrap^{TM} FF 5 ml* column (GE Healthcare, USA) using an ÄKTA (GE Healthcare, USA). The proteins were re-purified by size-exclusion chromatography in a *HiLoad^{TM} 26/600 Superdex^{TM} 75 pg* column (GE Healthcare, USA) and Native Buffer (50 mM Tris, 100 mM NaCl pH 8.4) was used as a mobile phase. The purity of the proteins was checked by 15 % SDS-PAGE.

Antibody fragments were expressed using a periplasmic expression system in *E. coli* BL21-Gold at 30 ^{o}C in 2XTY medium supplemented with 100 μg/mL ampicillin and tetracycline. Protein expression was induced with 1 mM IPTG for 48 h at 30 ^{o}C. Filtered supernatant was added to protein A resin column (GE Healthcare, USA). The proteins were eluted with PBS pH 3.2 and neutralized with PBS pH 9.0 to a pH of 7.4. The purity of the proteins was checked by 18 % SDS-PAGE.

GFP fluorescence

Excitation and emission spectra of fr-GFP, GFP/KKK and GFP/DDD were analyzed in a Jasco FP-8200 fluorescence spectrofluorometer (Jasco Corporation, Japan) using 0.5

mg/mL of protein in Native Buffer at pH 8. Three spectra were accumulated at 25 ^{o}C, with slit widths of 5 nm, a 0.5 nm interval and 1000 nm/min scan rate. Emission spectra was obtained recording emission between 500-600 nm. Excitation spectra was obtained exciting in a 450-600 nm range. Fluorescence of fr-GFP and GFP/KKK purified proteins was recorded using excitation and emission wavelengths of 485 nm and 515 nm, respectively.

For thermal denaturation, GFP fluorescence was recorded in a range of 25 ^{o}C to 90 ^{o}C with an increasing heat rate of 1 ^{o}C/min, slit widths of 5 nm, 0.5 nm interval and 1000 nm/min scan rate. Proteins at 10 μM were excited at 485 nm and emission recorded at 515 nm each 0.5 ^{o}C.

For chemical denaturation, fr-GFP, GFP/KKK and GFP/DDD proteins were mixed with increasing concentrations of guanidinium hydrochloride, from 0 to 6 M, keeping a constant protein concentration of 10 μM. Samples were left to stand for 48 h at 25 ^{o}C in order to equilibrate. Then, fluorescence was analyzed exciting at 485 nm and recording emission in a 500-600 nm range, with a bandwidth of 1 nm. Experiments were performed in duplicates.

For refolding kinetics, protein samples at 0.5 mg/mL in Native Buffer pH 8 were unfolded at 95 ^{o}C for 5 min and allowed to refold at 25 ^{o}C for 30 min. The recovery of GFP fluorescence was monitored exciting at 485 nm and recording the emission fluorescence at 515 nm every 5 seconds during 30 min.

For analyzing the dependence of the fluorescence emission on the pH, fr-GFP, GFP/KKK and GFP/DDD proteins were diluted in different buffers ranging from pH 2 to pH 13, at a final protein concentration of 10 μM and incubated for 2 h. Experiments were performed in duplicate.

In vivo GFP fluorescence

Induced *E. coli* cells expressing the Aβ42-GFP variants were harvested by centrifugation, washed with phosphate-buffered saline (PBS buffer pH 7.0) and adjusted to an OD of 0.2. GFP fluorescence of intact cells was measured on a Jasco FP-8200 fluorescence spectrofluorometer (Jasco Corporation, Japan), with slit widths of 5 nm, a 0.5 nm interval and 1000 nm/min scan rate, exciting the sample at 485 nm and recording the average of three accumulated emission spectra from 500 to 600 nm. Experiments were performed in triplicates.

Fluorescence microscopy

Induced Aβ42-GFP variants cultures were centrifuged, washed in PBS buffer, and 10 μL of cells were deposited on top of microscopy glass slides. Images were obtained in a Leica fluorescence microscope (Leica Microsystems, Germany) under UV light with a filter for GFP excitation (450-500 nm) and an emission filter (515-560 nm).

Immunoblotting

10 μL of fractioned cell extracts of Aβ42-GFP were loaded on a 15 % SDS-PAGE gel containing 0.1 % SDS. Proteins were transferred to a polyvinilidene difluoride membrane (PVDF) (MerckMillipore, Germany) with a Mini Trans-Blot Cell (Bio-Rad, USA) using Transfer Buffer (0.5 M Tris, 0.04 M glycine, 0.04 % SDS, 20 % methanol) at 100 V for 1 h. Membranes were blocked in blocking solution (0.05 M Tris-HCl, pH 7.6, 0.15 NaCl, 0.1 % Tween 20, 5 % milk) ON and incubated with 1:1000 anti-GFP antibody (Roche, Switzerland). Blot was then incubated with 1:2000 dilution of goat anti-mouse IgG

conjugated with peroxidase for 1 h at room temperature. Detection was carried out using an enhanced chemiluminescence (ECL) SuperSignal kit (Pierce Protein Biology, USA).

Flow cytometry analysis

Flow cytometry was performed using a BD FACSCanto flow cytometer (BD Biosciences,USA) software equipped with 488 nm and 635 nm lasers. Cells were first gated (P1) for forward scatter (FSC) and side scatter (SCC), both set on logarithmic amplification. A threshold was adjusted on FSC to reduce electronic background noise. Cells in P1 were then re-gated (P2) for FITC-channel to acquire GFP emission with a 530/30 nm band pass filter. A total of 20,000 events were acquired. Non-induced cells were used as a GFP signal controls. Data were collected with the FACSDiva Software (BD Biosciences, USA) and data analysis was performed with the FlowJo software. Experiments were made in triplicates.

Protein crystallization and X-ray structure determination

GFP/KKK was crystallized in MRC2 commercial crystallization plates at 18 °C, using the sitting drop vapour diffusion method; mixing 300 nL of protein (22.5 mg/mL) with 300 nL of precipitant (0.1 M MMT buffer, 25 % PEG 1500, pH 4.0) (PACT permier HT-96, Molecular Dimensions), which was allowed to equilibrate against a well containing 50 µl of precipitant. Single crystals appeared after two weeks. Crystals were cryo-protected in reservoir buffer containing 10% glycerol and flash-frozen in liquid nitrogen prior to diffraction analysis.

Diffraction data were collected from cryo-cooled crystals (100 K) on ALBA synchrotron (BL13-XALOC beamline), Barcelona, Spain[30].

Data were integrated and merged using XDS[31] and scaled, reduced and further analyzed using CCP4[32]. The structure of GFP/KKK was determined from the X-ray data by molecular replacement using the wtGFP crystal structure (PDB: 1GFL) as a starting model with phaser.phenix. Refinements and model rebuilding of GFP/KKK crystal structure were performed using PHENIX[33] and Coot[34]. All structure figures were generated using PyMOL (DeLano Scientific LLC). The data-collection and refinement statistics are summarized in the **Table S1**. The atomic coordinates and structure factors for GFP/KKK structure have been deposited in the Protein Data Bank (PDB entry 6FWW).

Circular dichroism spectroscopy

Thermal denaturation experiments, were performed using a JASCO J-715 spectropolarimeter (Jasco Corporation, Japan) equipped with a Peltier system for temperature control. The protein concentration was 10 μM in PBS pH 7.4. The CD signal at at 202 nm was recorded from 25 ^{o}C to 90 ^{o}C, with a heating rate of 1 ^{o}C/min. Far-UV spectra were recorded before and after heating. Five accumulations were averaged for each sample.

Light scattering

Aggregation kinetics of GFP and DP47 variants were followed on a Jasco FP-8200 fluorescence spectrofluorometer (Jasco Corporation, Japan) by exciting at 330 nm and recording orthogonal light scattering signal in a 320-340 nm range, with slit widths of 5 nm, a 0.5 nm interval and 1000 nm/min scan rate.

For temperature induced aggregation, 0.5 mg/mL GFP and GFP/KKK in Native Buffer, pH 8 were incubated at 80 ^{o}C. Measurements were made every 10 min during 70 min.

For chemical induced aggregation, 0.5 mg/mL GFP and GFP/KKK were incubated in the presence of 40 % TFE at 25 ^{o}C. The signal was acquired every 10 min for a total of 90 min in the same conditions as above.

DP47 and DP47/KKK aggregation kinetics were analyzed at 37 ^{o}C with an agitation of 300 rpm for 72 h. The protein concentration was 10 μM in PBS at pH 7.4.

Intrinsic tryptophan fluorescence

Tryptophan intrinsic fluorescence of DP47 and DP47/KKK was analyzed at 25 ^{o}C in a Jasco FP8200 spectrofluorometer (Jasco Corporation, Japan). Three averaged spectra were accumulated using an excitation wavelength of 280 nm and recording the emission from 300 to 400 nm with slit widths of 5 nm, 0.5 nm interval and 1000 nm/min scan rate.

Electrophoretic mobility assay

The EMSA reaction with (FITC)-labeled 50b random ssDNA was prepared with a constant DNA concentration of 0.2 μM, and various concentrations of purified protein in binding buffer (10 mM HEPES–KOH pH 7.5, 2 mM 2-ME). After 30 min incubation at 30 °C and addition of 10% (v/v) glycerol, free DNA and DNA–protein mixtures were resolved by electrophoresis for 90 min at 10 V, on 0.5% (w/v) TAE-agarose gels at 4°C. Gels with FITC labeled short ssDNA were detected directly on a Versadoc system (BioRad) with excitation at 488 nm.

Size-Exclusion Chromatography

500 µl of protein solution of fr-GFP and GFP/KKK at 10 µM in Native Buffer pH 8.0 were loaded onto a *Superdex^{TM} 75 10/300 GL* (GE Healthcare, USA) column coupled to an *ÄKTA-Fast Protein Liquid Chromatography* device (GE Healthcare, USA), previously equilibrated with Native Buffer pH 8. Protein elution was monitored by following absorbance at 280 nm.

Differential Scanning Calorimetry (DSC)

Samples were dialyzed in 20 mM phosphate buffer, 100 mM NaCl pH 7.95. Samples concentrations were adjusted to 15 µM. Experiments were carried out on a VP-DSC capillary-cell microcalorimeter from MicroCal (Malvern Panalytical Ltd, UK) at a scan rate of 120 °C/h from 5 ºC to 105 ºC. Calorimetric cells (operating volume 0.134 ml) were kept under pressure (60 psi) to prevent sample degassing. Several buffer-buffer baselines were obtained before each protein run to ascertain proper equilibration of the instrument. Reheating showed that the unfolding process is irreversible.

Results

A3D description

The A3D algorithm uses the protein 3D structure as an input, which is subsequently energetically minimized using the FoldX force field[24]. Then, an aggregation propensity score is calculated for all the spheres with a 10 Å radius in the protein structure. The variables that contribute to the A3D score are (i) the experimentally-derived individual amino acid aggregation propensities[35], (ii) the surface area exposure of the amino acids in the sphere and (iii) the effective distance between adjacent residues and the central amino acid in the sphere. Therefore, the A3D score is structurally corrected and, in contrast to sequence-based aggregation predictors, provides information on the structural aggregation propensity (STAP) of globular proteins in their functional folded states[18].

Rational design of an aggregation-resistant GFP

We initiated our re-design exercise trying to generate a more soluble variant of an already folding-optimized GFP version: the 'folding reporter' GFP (fr-GFP)[27], a folding-enhanced GFP containing the 'cycle-3' mutations[36] and the 'enhanced GFP' (EGFP) mutations[37] (**Figure 1A and 1B**). The 'cycle-3' mutations F99S, M153T, and V163A reduced the hydrophobicity of the protein, rendering it significantly more resistant to aggregation than the wild type GFP[38], whereas the EGFP mutations F64L and S65T modify the chromophore for simple absorption spectrum and strong fluorescence[39]. fr-GFP was the variant on top of which superfolder-GFP was evolved upon four rounds of DNA shuffling and paneling[40].

The structure of fr-GFP (PDB: 2B3Q) was analyzed using the A3D algorithm. Only three aggregation-prone residues were identified in the fr-GFP A3D profile: V11, Y39 and L221

(**Figure 1C**). In order, to obtain a more soluble version of fr-GFP, we exploited the so-called aggregation gatekeeper residues. They correspond to either charged amino acids (Asp, Glu, Arg or Lys) or Pro, all shown to contribute to modulate the aggregation properties of natural proteins[3,41,42]. The three fr-GFP aggregation-prone residues were virtually and individually mutated by gatekeepers and the theoretical most soluble variants selected according to their Average and A3D Total scores. Mutations to Asp or Lys were top ranked. Accordingly, variants containing one, two or the three aggregation-prone residues mutated either to Asp or to Lys were analyzed with A3D. Predictions indicated that the successive introduction of charged residues at the fr-GFP surface would potentially abrogate the presence of aggregation-prone exposed patches and accordingly decrease the STAP of the re-designed proteins (**Table 1 and Figure 2**).

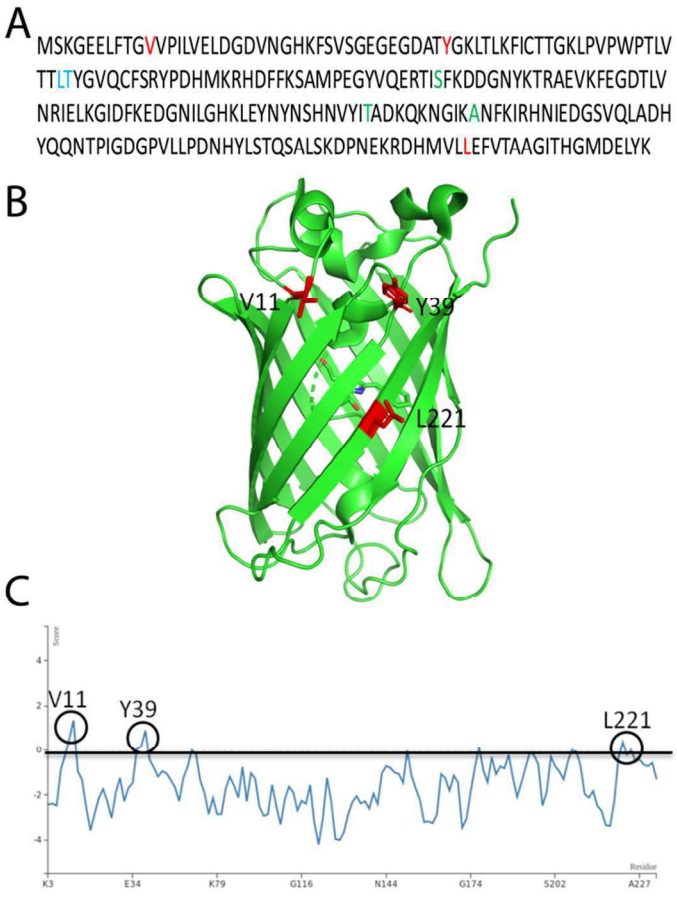


Figure 1. Sequence, structure and aggregation profile of GFP. A) Amino acid sequence of fr-GFP. Residues shared with Cycle-3 and EGFP variants are indicated in green and blue, respectively. **B)** 3D structure of fr-GFP (PDB: 2B3Q). **C)** Aggregation profile of fr-GFP according to A3D. In A) and B) predicted aggregation-prone residues are indicated in red, in C) they are circled.

Even when a particular residue substitution at the protein surface is predicted to increase the protein solubility, this positive effect might be canceled in case the mutation impacts the thermodynamic stability, since it would facilitate protein fluctuations and the exposure of aggregation-prone regions previously protected in the structure[43]. This might render the mutant variant even more aggregation-prone than the original protein. We calculated the impact of the designed mutations on fr-GFP stability using the FoldX force field[24]. FoldX is among the most popular stability predictors and was already implemented in the original version of A3D to minimize the input PDB files. Despite computational predictions of stability changes upon mutation are still far from being perfect, FoldX has been described as an accurate stability predictor[44] with a low prediction bias[45]. Moreover, FoldX has been already used to explore the connection between protein stability and aggregation[3,46]. Mutations to Asp where predicted to be destabilizing ($\Delta\Delta G > 1$ kcal/mol), whereas mutations to Lys where predicted to be neutral ($0.5 > \Delta\Delta G > -0.5$ kcal/mol). This provided us a unique opportunity to evaluate the interplay between predicted STAP and stability in determining fr-GFP solubility and, more generally, on the eventual benefit of incorporating stability predictions in our A3D pipeline.

Table 1. A3D and FoldX analysis of fr-GFP and its variants. More negative average and total A3D scores indicate higher solubility. Negative and positive ΔΔG values correspond to over- and de-stabilizing mutations, respectively.

GFP MUTATIONS	NAME	AVERAGE SCORE	TOTAL SCORE	ΔΔG (kcal/mol)
-	fr-GFP	-1.026	-231.814	-
V11K	GFP/K	-1.078	-243.631	-0.065
V11K/L221K	GFP/KK	-1.102	-248.993	-0.026
V11K/Y39K/L221K	GFP/KKK	-1.132	-255.922	0.017
V11D	GFP/D	-1.09	-246.239	1.08
V11D/L221D	GFP/DD	-1.107	-250.147	2.555
V11D/Y39D/L221D	GFP/DDD	-1.140	-257.73	3.545

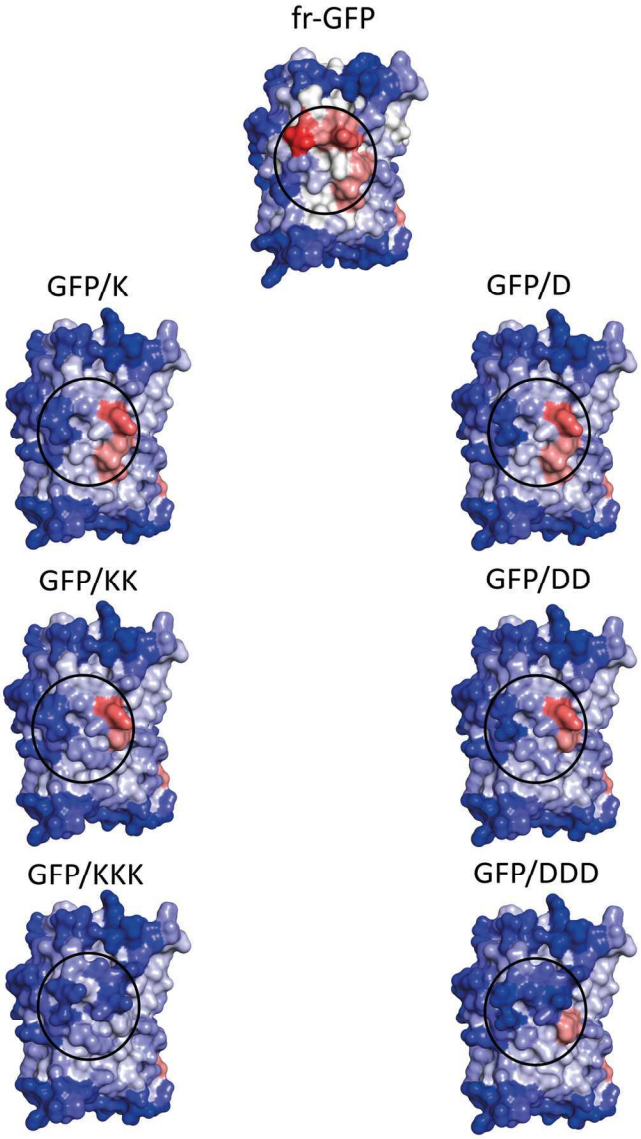


Figure 2. A3D structural predictions for fr-GFP and its variants. 3D Structures of fr-GFP and its mutational variants structures. The protein surface is colored according to A3D score in a gradient from blue (high-predicted solubility) to white (negligible impact on protein aggregation) to red (high-predicted aggregation propensity).

Designed lysine mutations increase Aβ42-GFP solubility upon overexpression

The already optimized folding and solubility properties of fr-GFP[27] make the selection of a more robust variant a challenging exercise. We reasoned that this can be achieved by fusing the above six designed mutants with an N-terminal aggregation-prone bait that would favor the aggregation of the fusion protein. To this aim we used a fusion of the amyloidogenic Aβ Alzheimer's peptide (Aβ42) with fr-GFP (Aβ42-GFP) or its variants. We have used this kind of fusion extensively to analyze the impact of Aβ42 mutations on aggregation-propensity[28], toxicity[47] or cellular aging[48], by mutating specific residues in the Aβ42 moiety, while leaving intact the GFP domain. Independently of the Aβ42 mutation, Aβ42-GFP was shown to accumulate always in the insoluble fraction as inclusion bodies[28]. However, the presence of active GFP in such aggregates differed significantly, displaying a high correlation with the intrinsic aggregation propensity of the Aβ42 mutants[49]. This is due to the existence of an *in vivo* kinetic competition between the attainment of the GFP functional structure and the aggregation of the fusion protein, directed by the Aβ moiety[50]; highly aggregation-prone Aβ variants resulting in low fluorescence and the other way around.

In this work, we proceeded in the opposite way, testing if mutations in the GFP domain can somehow counteract the high aggregation-propensity of wild-type Aβ42. In this system, it is expected that the more soluble fr-GFP variants would result in higher levels of active

GFP upon intracellular over-expression, by decreasing the overall aggregation rate of the protein fusion. Therefore, we measured the GFP fluorescence of intact cells over-expressing Aβ42 fused to fr-GFP (Aβ42-GFP) as well as to our six fr-GFP variants using spectrofluorimetry. As shown in **Figure 3A**, a single Lys mutation at position 11 (Aβ42-GFP/K) did not suffice to increase significantly GFP emission, relative to cells expressing the original fusion, despite it was predicted to be the individual Lys mutation with the largest impact on protein solubility. However, the introduction of two and three Lys point mutations in fr-GFP (Aβ42-GFP/KK and Aβ42-GFP/KKK) promoted a significant increase in the fluorescence emitted by the respective cells (**Figure 3A**). The need for several simultaneous mutations to attain a significant solubilizing effect has been evidenced in different computer-assisted protein solubility re-design exercises [3,17,42,51]. Cells expressing Aβ42-GFP/KKK were three-fold more fluorescent than those expressing Aβ42-GFP. We also assayed the GFP fluorescence of the different cell populations using flow cytometry (FC) (**Figure 3C**). The ability of FC to analyze a large number of cells (20.000 cells/sample) averages experimental variability and provides more quantitative results[47]. The analysis showed that, independently of the considered variant, a large majority of the cells were somehow fluorescent, indicating that they were expressing the correspondent protein fusions (data not shown). The mean fluorescence values obtained for the different variants were in excellent agreement with those obtained by spectrofluorimetry, confirming that cells expressing Aβ42-GFP/KKK were three times more fluorescent than the ones expressing the original fusion.

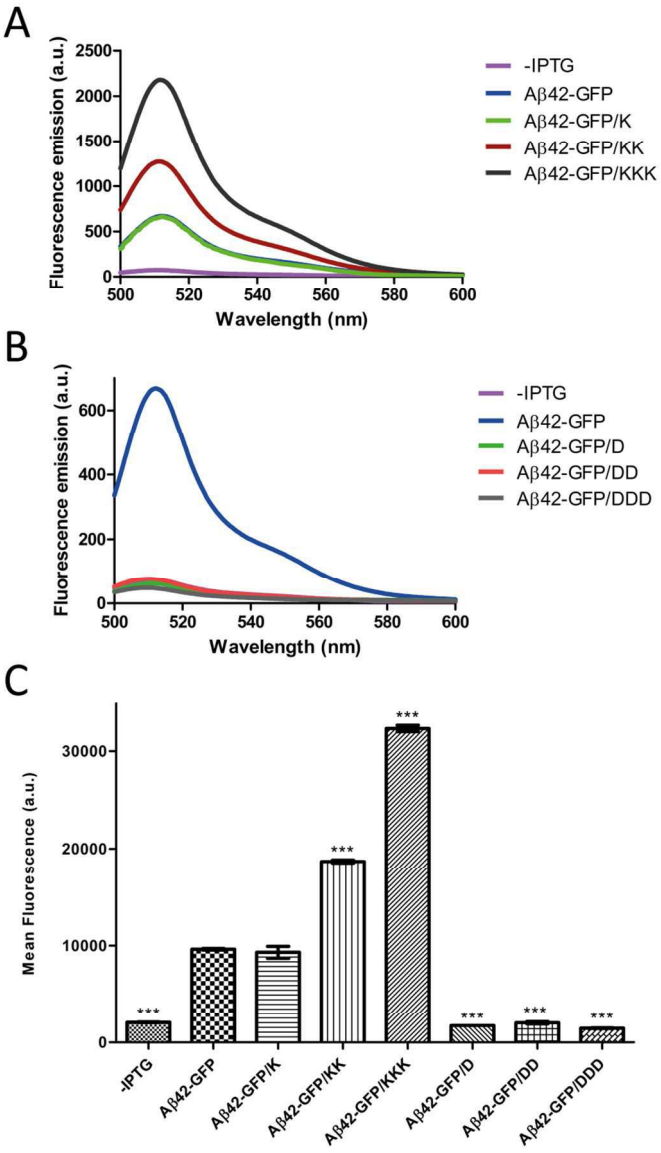


Figure 3. GFP fluorescence of intact cells expressing Aβ42-GFP variants. A) Quantification of the fluorescence emission of different lysine variants in *E. coli* cells measured by fluorescence spectroscopy. **B)** Quantification of the fluorescence emission of different aspartic acid variants in *E. coli* cells measured by fluorescence spectroscopy. **C)** Histogram of Aβ42-GFP variants showing the average value of fluorescence intensity as detected by flow cytometry. All experiments correspond to triplicates and the statistical values were obtained using an Unpaired *T-student test.* Statistically significant values as compared with Aβ42-GFP are indicated: *p*-value < 0,001 (***). Error bars correspond to SEM.

The distribution of the recombinant protein fusions between the soluble and insoluble cellular fractions was analyzed by western blotting (**Figure 4A**). As previously reported[48,49] all Aβ42-GFP resided in the insoluble fraction, as it was the case for Aβ42-GFP/K. In contrast, despite a majority of Aβ42-GFP/KK and Aβ42-GFP/KKK fusions were still located in the insoluble fraction, increasing amounts of soluble protein were detected for these two variants. It is worth to emphasize here that, in our previous studies, we only observed the Aβ42-GFP fusion in the soluble fraction when at least two highly disruptive mutations where introduced in the Aβ42 moiety[47].

Finally, we used fluorescence microscopy to identify the cellular location of the detected GFP emission in *E.coli* cells. As expected, the cells become more fluorescent as the number of Lys in fr-GFP increased (**Figure 4B)**. In good agreement with fractionation analysis, the fluorescence was confined mainly in inclusion bodies at the poles of the cells. The high fluorescence of Aβ42-GFP/KKK inclusions indicates that the higher solubility of this fr-GFP variant competes the aggregation of the Aβ42 peptide, allowing a higher amount of fluorophore formation before the intracellular aggregation occurs.

Interestingly enough, despite A3D STAP predictions for Asp variants were, on the overall, equivalent to those obtained for the Lys counterparts (**Table 1 and Figure 2**), both spectrofluorimetry and FC data indicated that bacterial cells expressing Aβ42-GFP/D, Aβ42-GFP/DD and Aβ42-GFP/DDD fusion proteins displayed fluorescence intensities equivalent to that of non-induced cultures and thus, significantly lower than induced Aβ42-GFP (**Figure 3B and 3C**), even if they were expressed at similar levels than the original fusion (**Figure S1**). This highlighted the important role played by protein stability on

effective protein solubility and advised us on the convenience of incorporating stability in our A3D prediction scheme.

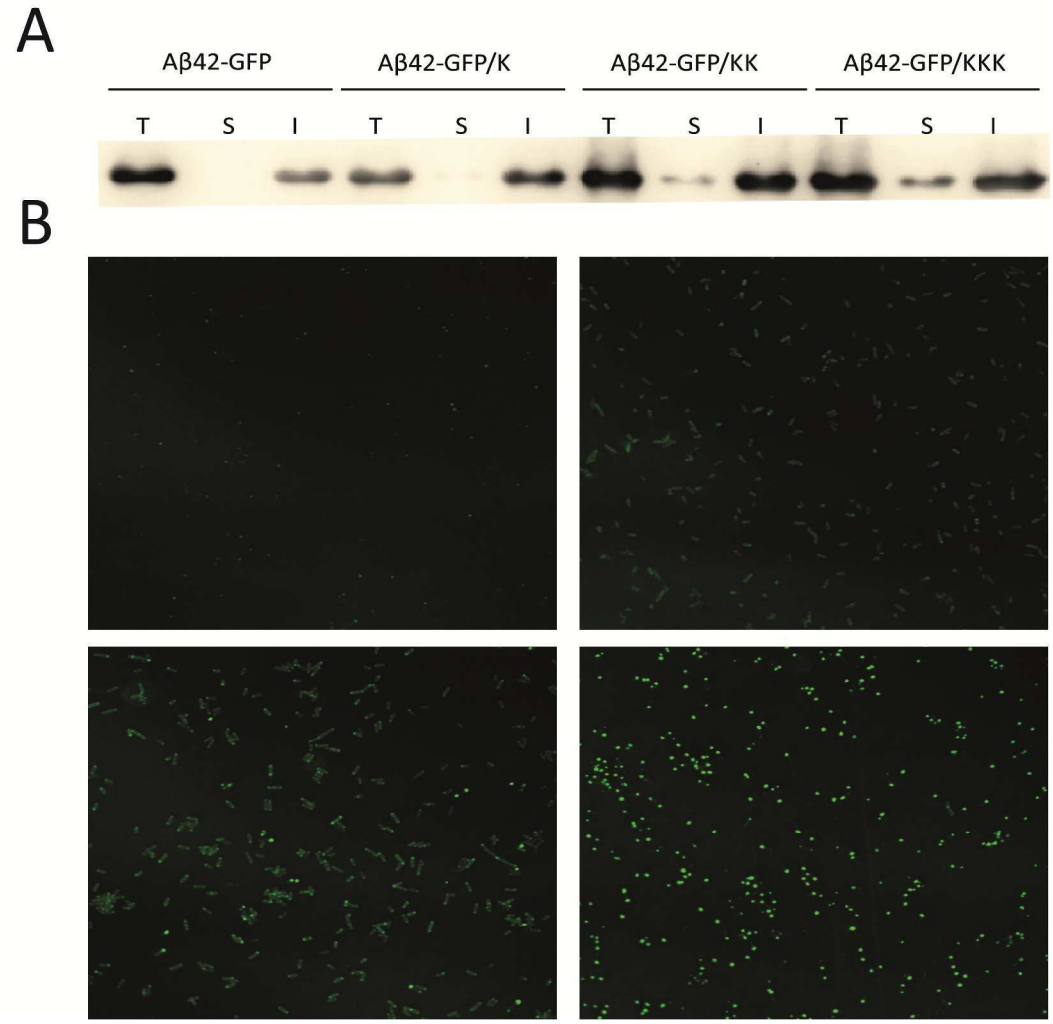


Figure 4. Cellular distribution of Aβ42-GFP fusions. A) Western blot of Aβ42-GFP in the total (T), soluble (S) and insoluble (I) fractions of *E. coli* cells. **B**) Visualization of GFP fluorescence in intact *E. coli* cells expressing Aβ42-GFP (Top, left), Aβ42-GFP/K (Top, right), Aβ42-GFP/KK (Bottom, left) and Aβ42-GFP/KKK (Bottom, right).

GFP/KKK is folded and stable

Once *in cell* analysis suggested GFP/KKK being the more soluble among the generated variants, we wanted to asses first that, once purified, it was well-folded and stable. To this aim the mutant GFP domain alone was His-tagged, recombinantly produced and purified

using affinity and size exclusion chromatography. The same procedure was used for the original fr-GFP protein. We also produced and purified GFP/DDD in order to assess if, as predicted by FoldX, this variant was destabilized relative to both GFP/KKK and fr-GFP. Both GFP/DDD and Aβ42-GFP/DDD exhibited anomalous electrophoretic mobility (**Figures S1 and S2)**, a property already described for other proteins in which mutations to acidic residues have been introduced[52].

The fluorescence emission spectra of fr-GFP, GFP/KKK and GFP/DDD were superimposable (**Figure 5A**). The fluorescence excitation of GFP/KKK and GFP/DDD were slightly blue-shifted relative to fr-GFP, an effect that was also observed for superfolder GFP[40]. At pH 8.0, the emission fluorescence intensities of GFP/KKK and GFP/DDD at 515 nm were 91 % and 87 % of that exhibited by fr-GFP, respectively.

In order to confirm that the introduced mutations did not impact significantly protein stability, as predicted, fr-GFP and GFP/KKK were submitted to thermal and chemical denaturation. The thermal stability of the proteins was analyzed at pH 8.0 by monitoring the changes in GFP fluorescence emission. We obtained cooperative, superimposable denaturation curves with Tm = 81.72 ± 0.17 ^{o}C and Tm = 81.47± 0.16 ^{o}C, for fr-GFP and GFP/KKK, respectively (**Figure 5B**). As predicted by FoldX, the two variants displayed a higher Tm than GFP/DDD, with a Tm = 76.06 ± 0.10 ^{o}C (**Figure 5B**). The heat denaturation of all the variants in these conditions was irreversible, which precluded calculation of the thermodynamic constants. To confirm that fr-GFP and GFP/KKK differ from GFP/DDD in terms of thermal stability, we performed thermal denaturation assays for the three variants in the presence of increasing concentrations of guanidinium hydrochloride (GuHCl) (**Figure S3**). As it can be deduced from the calculated Tms in the

different conditions (**Figure 5D**) fr-GFP and GFP/KKK display similar stability, both being more stable than GFP/DDD, for which the denaturation curve in the presence of 2.4 M GuHCl was not longer cooperative. To validate that fr-GFP and GFP/KKK have similar thermal stability we used differential scanning calorimetry (DSC) (**Figure S4**). We calculated Tm = 84.6 ± 0.1 ^{o}C and Tm = 84.0 ± 0.1 ^{o}C for fr-GFP and GFP/KKK, respectively. For GFPs the calculated Tms depend significantly on the scan rates[53] this would explain the difference between the Tms obtained for both variants in fluorimetry and DSC analysis. In any case, all the results converge to indicate that, in contrast to GFP/DDD, GFP/KKK is very similar to fr-GFP in terms of thermal stability.

The resistance against chemical denaturation with guanidinium hydrochloride (GuHCl) was also monitored by recording the changes in GFP fluorescence in the presence of increasing concentrations of denaturant after equilibration for 48 h. The obtained curves were cooperative and displayed a midpoint of denaturation of 3.19 ± 0.06 M, 3.08 ± 0.08 M and 2.56 ± 0.10 M (**Figure 5C**). It is worth to note that the conditions of the assay are of pseudo-equilibrium, since the equilibrium of GFP is not attained even after several weeks[54] and that the reaction is not reversible due to the hysteresis caused by the denaturation of the chromophore during the unfolding reaction[55], both factors impeding the calculation of thermodynamic parameters.

Overall, as predicted, in GFP/KKK the introduced mutations did not compromise GFP function or stability, whereas the GFP/DDD variant is significantly less stable.

Finally, we calculated the dependence of the chromophore fluorescence on the pH for the three variants (**Figure S5**). The three proteins remained inactive below pH 5.0. fr-GFP and

GFP/DDD exhibited a very similar profile, displaying maximum fluorescence emission at pH 9.0. GFP/KKK is more fluorescent at alkaline pH, reaching a maximum at pH 11.0. Upon incubation a pH 12.0 and above all the proteins become inactive.

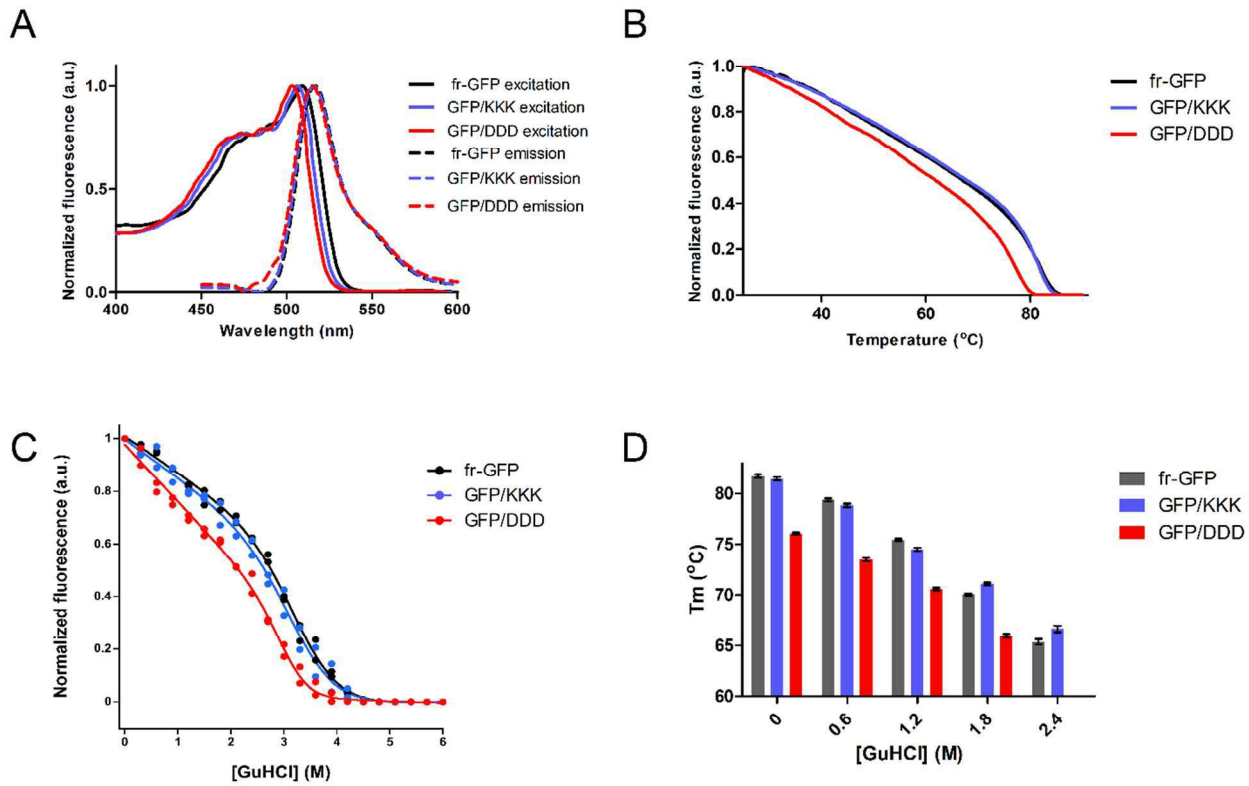


Figure 5. Spectral properties and stability of fr-GFP, GFP/KKK and GFP/DDD. A) Normalized excitation and emission spectra of different GFP variants. **B)** Thermal unfolding curves followed by GFP fluorescence from 25 °C to 90 °C with an increasing rate of 1 ° C/min. **C)** Chemical denaturation of variants at 10 µM with increasing concentrations of GuHCl. Samples were equilibrated during 48 h. **D)** Melting Temperature values in the presence of increasing GuHCl concentrations. Experiments were performed in duplicates. Error bars correspond to SEM.

GFP/KKK is highly resistant to aggregation

In order to confirm that the increased solubility of the Aβ42-GFP/KKK inside cells, relative to Aβ42-GFP, can be assigned to a more soluble GFP moiety, we compared the aggregation properties of purified GFP/KKK and fr-GFP.

First, we monitored the kinetics of aggregation for the two variants by following the increase in light scattering when the proteins were heated at 80 ^{o}C. The aggregation rate was significantly lower for the re-designed protein (k_1= 0.017 ± 0.005 min^{-1}) than for fr-GFP (k_1= 0.030 ± 0.009 min^{-1}), as well as the scattering at the end of the reaction (**Figure 6A**). Next, we tested the proteins resistance to aggregation in the presence of 40 % of TFE, a strongly denaturing condition, known to promote the aggregation of β-sheet rich proteins[23,56]. We recorded the light scattering signal at different time points after addition of the co-solvent. As shown in **Figure 6B**, fr-GFP already started to aggregate after 10 min, whereas for GFP/KKK the detection of the initial aggregates occurred only after 50 min, exhibiting lower scattering intensity than the original protein at any analyzed time point.

Finally, we analyzed the ability of both proteins to refold into the native state upon thermal denaturation. fr-GFP and GFP/KKK were incubated at 95 ^{o}C for 5 min and allowed to refold at 25 ^{o}C, the recovery of GFP fluorescence along time was monitored (**Figure 6C**). The refolding reactions were adjusted to double-exponential functions ($R \geq 0.99$). GFP/KKK folds more than 3-fold faster (k_1= 0.175 ± 0.004 min^{-1}) than fr-GFP (k_1= 0.045 ± 0.001 min^{-1}). In addition, GFP/KKK recovers two times more fluorescence than fr-GFP at the end of the reaction.

Overall, the data in this section demonstrates how the combined prediction of STAP and stability allowed us to design a GFP variant with significantly improved solubility and folding properties by mutating only three residues at the protein surface. This result is remarkable if we take into account that fr-GFP was already designed to have optimal properties as a reporter of protein folding.

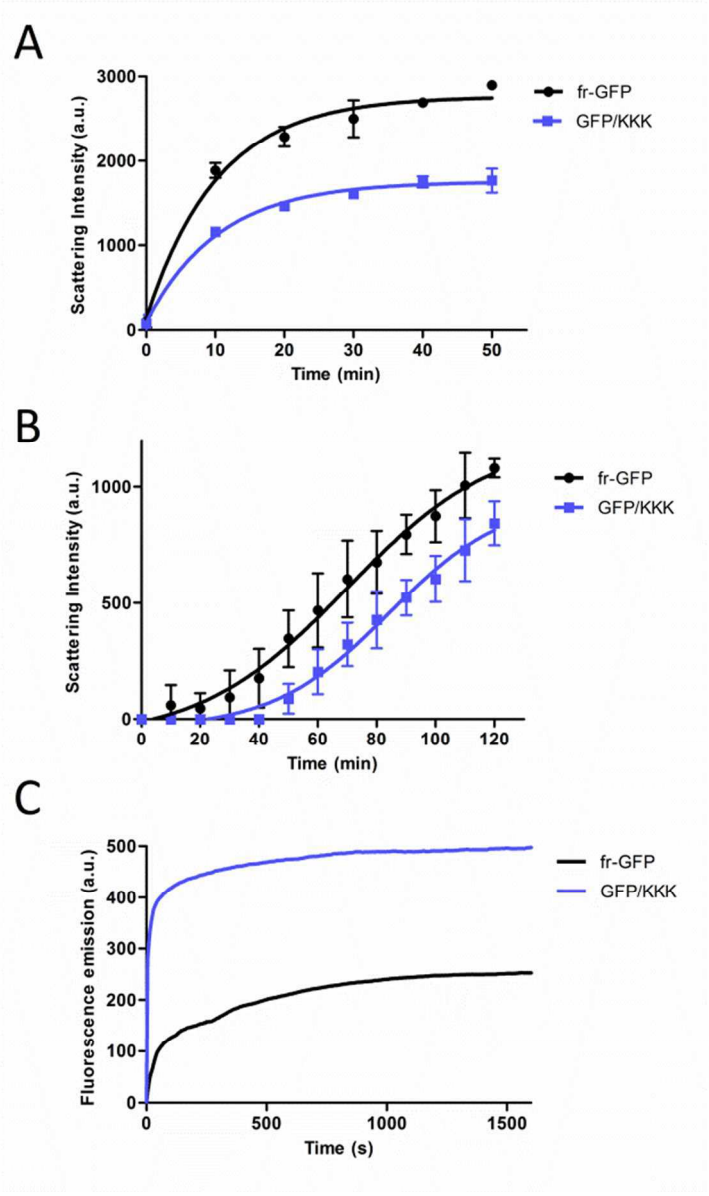


Figure 6. Comparison of the aggregation properties of fr-GFP and GFP/KKK. A) Evolution of the light scattering signal upon incubation at 80 ^{o}C. **B)** Evolution of the light scattering signal upon incubation in 40 % TFE. Both experiments were performed in duplicates. **C)** Refolding kinetics of both proteins at 25 ^{o}C after thermal unfolding at 95

°C, as measured by monitoring the recovery of GFP fluorescence emission. All the experiments were performed at 0.5 mg/mL concentration in Native Buffer pH 8.0 and in duplicates. Error bars correspond to SEM.

X-ray structure of GFP/KKK

To investigate if the inserted mutations in GFP/KKK impact the 3D structure of GFP, we solved the crystal structure of this mutant at 1.13 Å resolution (PDB: 6FWW). The side chains of the newly introduced Lys residues are well-defined in the electron density maps, all being exposed to the solvent (**Figure 7A**). The chromophore formed from residues T65-Y66-G67 (annotated as CRO in the PDB file) is correctly formed and buried in the middle of the central helix (**Figure 7A and 7B**). The Cα backbone trace of wild-type GFP (1GFL) and GFP/KKK could be superimposed with a root mean square deviation (r.m.s.d.) of 0.52 Å (**Figure 7C**), which is comparable to the r.m.s.d. of 0.50 Å between the fr-GFP and GFP/KKK. Thus, as expected, the three-surface mutations did not change the overall GFP fold, indeed no evident structural rearrangements were observed around the mutated regions (**Figure 7D**). An analysis of the GFP/KKK structure with A3D confirms that, in contrast to fr-GFP, this variant lacks any surface exposed aggregation-prone region (**Figures 7E and 7F**).

Although purified fr-GFP and GFP/KKK proteins were monomeric (**Figure S6**), it has been reported that fr-GFP crystallizes as an antiparallel dimer (PDB:2B3Q) similar to other reported GFP crystal structures, likely because GFP has certain tendency to dimerize in solution[40]. In contrast GFP/KKK crystallizes as a monomer. Using the dimeric structure of fr-GFP, we generated a model of a GFP/KKK dimer. The L221K residue of each monomer protrudes into the symmetry-related monomer facing the equivalent L221K residue,

potentially generating repulsion and hindering dimerization (**Figure S7**). Calculation of the impact of this single mutation on the stability of the dimeric interface with FoldX[24] confirms that it is highly disruptive ($\Delta\Delta G$ = 4.96 kcal/mol). We find out that, Lippincott-Schwartz and co-workers introduced exactly the same single mutation in GFP to generate a variant that could not dimerize[57,58]. This effect might contribute to the higher solubility of this re-engineered variant. In addition, in the crystal structure of GFP/KKK the side chains of the other two mutated residues, Lys 11 and Lys 39, are at less than 5 Å distance (**Table S2**), resulting in a local clustering of positive charges, which will likely cause electrostatic repulsion between GFP/KKK molecules.

It has been shown that a positively supercharged version of GFP binds strongly to nucleic acids, which compromises its use in a cellular context[59]. Electrophoretic mobility shift assays showed that increasing concentrations of fr-GFP or GFP/KKK did not reduce the mobility of a random FITC-labeled 50 base long ssDNA in the native agarose gel even when the protein concentration exceeds 50 molar times over DNA (**Figure S8**). This result indicates that the insertion of the three mutated Lys is not sufficient to promote the binding of GFP/KKK to DNA.

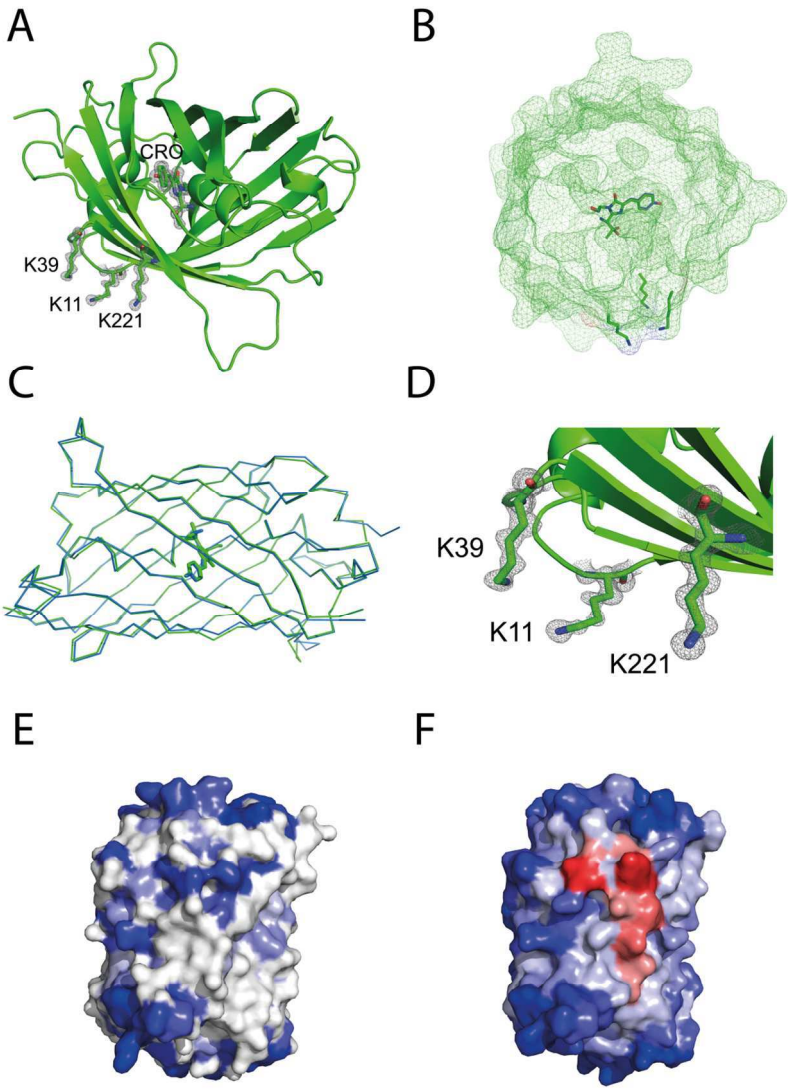


Figure 7. 3D structure and A3D structural prediction of GFP/KKK. A) Crystal structure of GFP/KKK (PDB: 6FWW) mutant showing the differential map (2mFo-Fc contoured at 1.2σ) surrounding the mutated lysine residues 11, 39 and 221 and the fluorescent cromophore which are depicted in mesh and stick. **B)** Surface representation of the superposed GFP structures showing the perfect overlay of the fluorescent chromophore (blue and green sticks for wtGFP and GFP/KKK mutant, respectively). **C)** Superposition of wtGFP (blue, PDB: 1GFL) and GFP/KKK (green) monomers. The overall rmsd for the 226 superposed Cα atoms is 0.52 Å with 96.9 % identity. **D)** Zoom of GFP/KKK structure around the mutated lysines. **E-F)** A3D of GFP/KKK and fr-GFP crystal structures, respectively. The protein surface is colored according to A3D score in a gradient from blue (high-predicted solubility) to white (negligible impact on protein aggregation) to red (high-predicted aggregation propensity).

Incorporating stability predictions in A3D

The above-described data clearly indicated the convenience of skipping the experimental validation of mutations that despite optimizing STAP impact negatively protein stability. To this aim, we implemented the FoldX empirical force field[24] in A3D web server for automated protein stability calculations if a user introduces any mutation(s) (see "In silico analysis" in Experimental Section). An example of the output is presented in **Figure S9B**. We wanted to prevent A3D users from producing re-designed proteins that ultimately will aggregate at the same or higher levels than the original protein only because they are destabilized. Therefore, we took a conservative criterion for structural integrity, in which a mutation resulting in a $\Delta\Delta G \geq 1$ kcal/mol is considered disruptive. Indeed, all fr-GFP Asp mutants overpass this destabilization limit. Accordingly, a warning message is now displayed in the "Project details" screen of A3D when sequence changes are predicted to destabilize the protein ≥ 1 kcal/mol (**Figure S9C).**

Rational design of a VH antibody with increased aggregation resistance

We decided to test the novel version of A3D in the re-design of a therapeutically relevant scaffold. There is a trend in the industry toward the use of small antibody formats for imaging and tumor targeting purposes. They include human single chain fragments (scFv) and human single domain antibodies[29]. However, the biophysical properties of these human proteins are poor, and their aggregation propensity is even higher than the one of complete antibodies, due to the lack of inter-domain stabilization. Here we tried to improve the solubility of a Variable Heavy (VH) segment of the human antibody germline (DP47)[29]. DP47 VH is a monomer consisting of 3 α-helices, 10 β-strands, 1 disulfide bond and 3

complementarity-determining regions (CDRs) essential for antigen binding (**Figure 8A and 8B**). DP47 VH was selected because the evolution of more soluble variants of this specific antibody has been previously addressed using phage display[29]. In this work, CDR3 was excluded from mutation to limit effects on antigen binding. Despite 6 mutations that increased the solubility of the molecule were identified, all them were clustered at CDR1, while mutations at other positions had little effect[60]. This was expected, because CDRs tend to be sticky, in order to recognize antigens. Nevertheless, mutations at CDRs can compromise significantly binding and, accordingly, they are not a preferred strategy to generate more soluble antibody variants.

Here, we aimed at generating a DP47 variant with increased aggregation resistance without changing any of its CDRs residues. No structure is available for wild-type DP47 and the version with mutated CDRs was crystallized as a domain-swapped dimer (PDB: 3UPC)[29]. Therefore, no suitable DP47 monomeric structure was available for direct A3D analysis. Accordingly, we created a homology model on top of sequentially related available structures (PDB:5I19) using Swiss-Model[25]. This model has a GMQE (Global Model Quality Estimation) of 0.99 and a QMEAN Z-score of 0.6 (**Figure 8B**). As expected, A3D predicted a high aggregation propensity for the three DP47 CDRs. Apart from them, only three other residues where predicted as aggregation-prone, Leu3 and Leu9 at the N-terminus and Leu109 at the C-terminus (**Figure 8C**). These residues were virtually mutated by the different gatekeeper residues using the new A3D algorithm and again a triple Lys mutant (DP47/KKK) appeared as the version that best combined predicted solubility and stability (**Table 2**, **Figure 9**), whereas an equivalent DP47/DDD version was again predicted to be destabilized (**Table S3**).

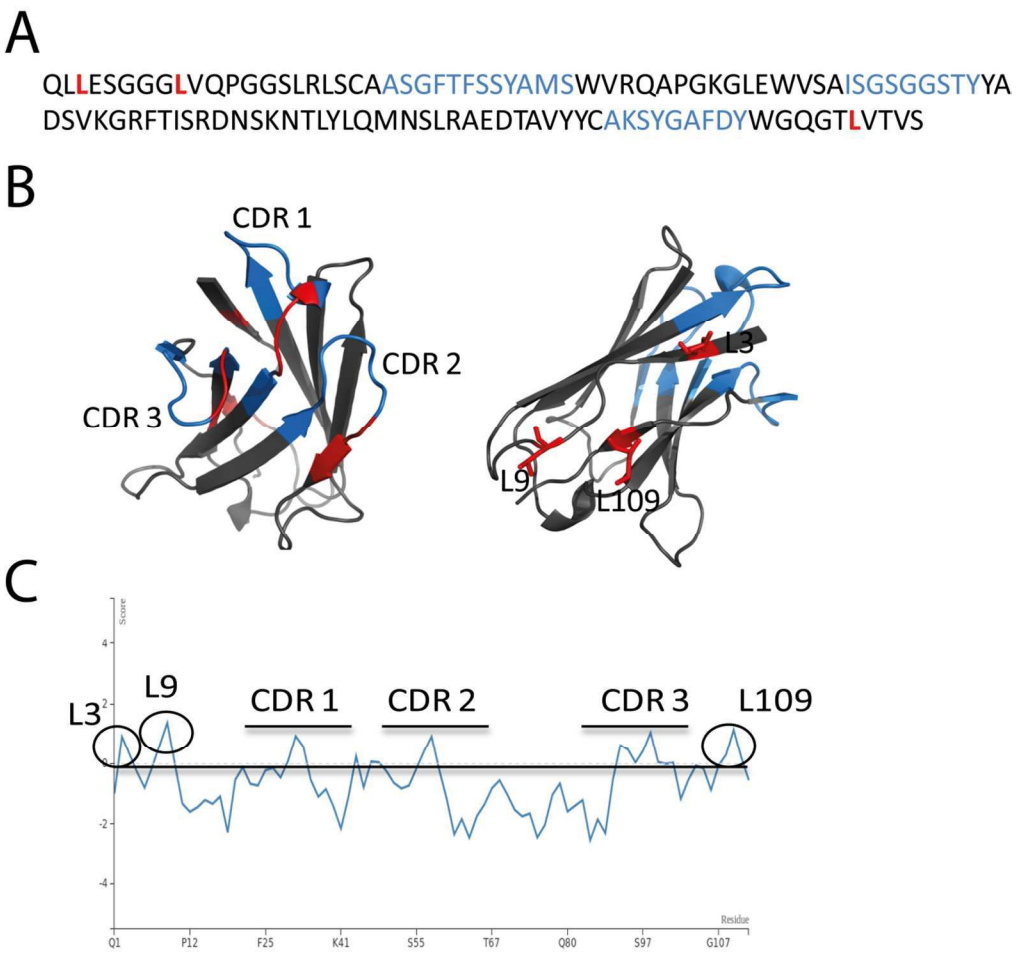


Figure 8. Sequence, structure and aggregation profile of DP47. A) Amino acid sequence of DP47. **B)** 3D structure of DP47 model obtained from SwissModel. **C)** Aggregation profile of DP47 model according to A3D. In A) and B) predicted aggregation-prone residues are indicated in red and CDR residues in blue, in C) they are circled and underlined, respectively.

Table 2. A3D analysis of DP47 and its variants. More negative average and total A3D scores indicate higher solubility. Negative ΔΔG values correspond to over-stabilizing mutations.

DP47 PROTEIN	AVERAGE SCORE	TOTAL SCORE	ΔΔG (kcal/mol)
DP47	-0.48	-54.248	-
DP47/K	-0.6021	-68.0402	-0.5196
DP47/KK	-0.6993	-79.0154	-0.014
DP47/KKK	-0.758	-85.681	-0.427

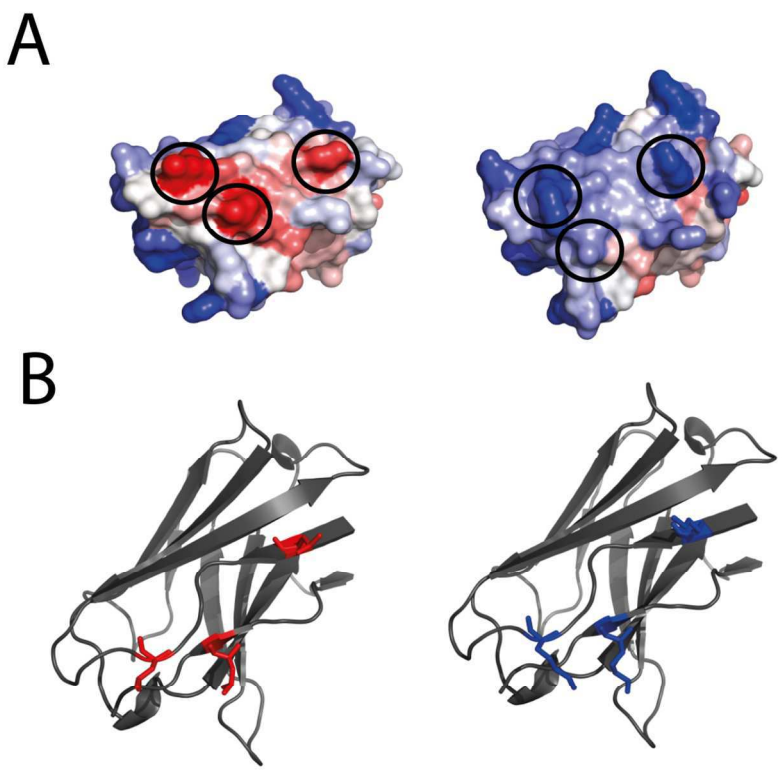


Figure 9. A3D structural predictions for DP47 and its variant. 3D Structures of DP47 (left) and its mutational variant structure (right). **A**) The protein surface is colored according to A3D score in a gradient from blue (high-predicted solubility) to white (negligible impact on protein aggregation) to red (high-predicted aggregation propensity). **B**) Ribbon representation of the two models indicating the wild-type Leu residues in red and the mutated Lys residues in blue.

DP47/KKK is folded and stable

Wild-type DP47 and DP47/KKK were recombinantly expressed in *E. coli*, and purified to homogeneity from the extracellular medium by affinity on a protein A column. Then we compared the biophysical properties of both single-domain antibodies.

The circular dichroism (CD) spectra of both proteins showed the characteristic β-sheet minimum at 218 nm and overlapped, indicating that they essentially shared the same secondary structure content (**Figure 10A**). DP47 contains three Trp residues at positions 34, 45 and 104 being W34 inside the structure and W45 and W104 exposed to the solvent.

Its intrinsic fluorescence spectra in the folded state exhibits a maximum at 345.5 nm. The Trp intrinsic fluorescence spectra of DP47/KKK shows the same intensity than that of DP47, although the maximum is slightly red-shifted (maximum at 346.5 nm) (**Figure 10B**), suggesting that despite the two antibodies share a similar globular tertiary structure, they can exhibit certain local conformational differences.

Finally, in order to confirm that, as predicted, the introduced mutations did not impact significantly the antibody stability (predicted $\Delta\Delta G$= -0.43 kcal/mol), DP47 and DP47/KKK were submitted to thermal denaturation. The thermal stability of the proteins was analyzed by monitoring the changes in CD signal at 202 nm. Both proteins unfolded cooperatively, displaying a very similar Tm of 60.0 ± 0.1 ^{o}C and 60.4 ± 0.1 ^{o}C for DP47 and DP47/KKK, respectively (**Figure 10C**). The reaction was not reversible and thermodynamic constants could not be calculated.

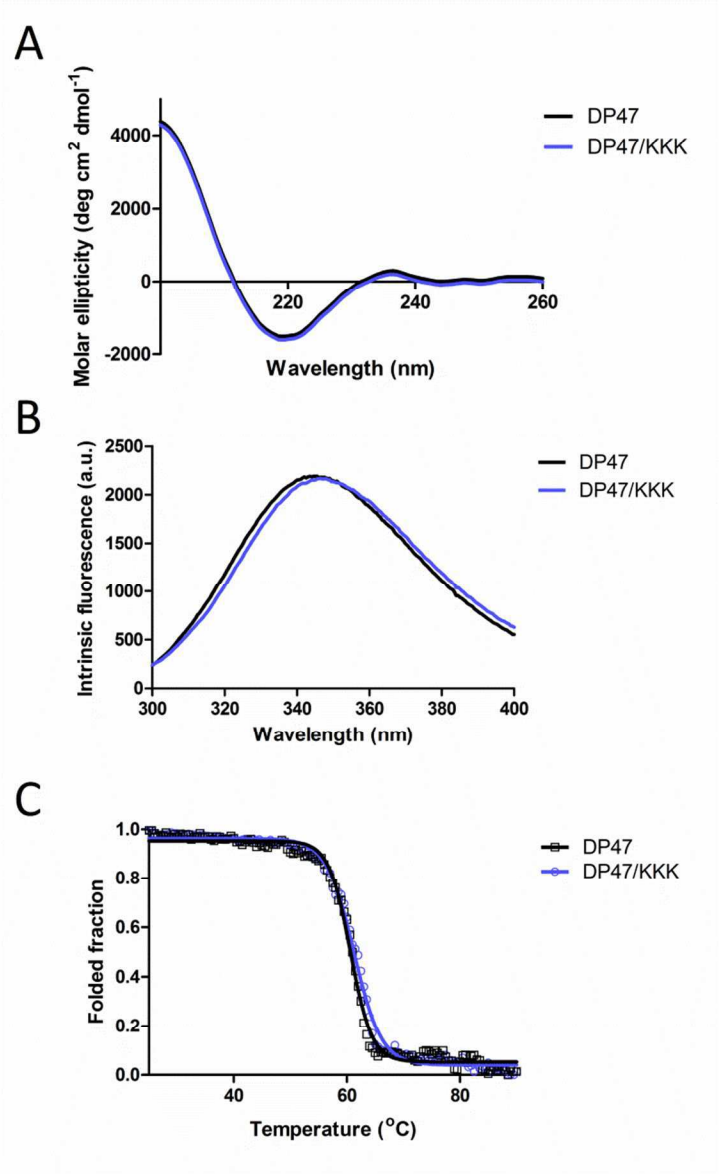


Figure 10. Biophysical characterization of DP47 variants. A) Far-UV CD spectra with a minimum value at 218 nm. **B)** Tryptophan intrinsic fluorescence. **C)** Thermal unfolding curves followed by Far-UV from 25 °C to 90 °C with a heating rate of 1 °C/min. The experiments were performed at 10 µM protein in PBS pH 7.4.

DP47/KKK displays increased aggregation resistance

To address if, as intended, the three Leu to Lys mutations increased the resistance of the antibody against aggregation in close to physiological conditions, DP47 and DP47/KKK

were incubated in PBS at pH 7.4. and 37 °C for 3 days and we monitored the changes in light scattering. **Figure 11** illustrates how the scattering of DP47 solution steadily increases with time reaching a maximum after two days. In contrast, the scattering of the DP47/KKK solution remains constant and is lower than that of the wild-type protein at any time point. Despite the final scattering signal attained in these conditions is low, compared with the one obtained after heat denaturation, the differences between proteins are statistically significant, indicating that the re-design has succeeded in reducing the inherent high aggregation propensity of this particular antibody.

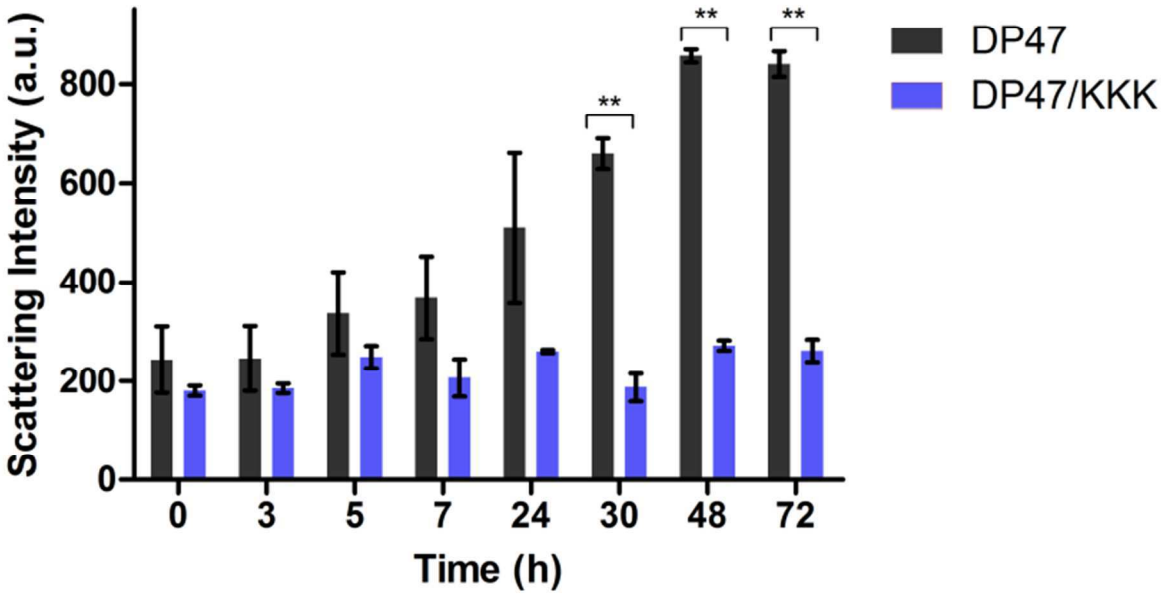


Figure 11. Comparison of the aggregation properties of DP47 variants. The two antibodies were incubated at 37 °C with an agitation of 300 rpm for 72 h. The protein concentration was 10 µM in PBS at pH 7.4. The light scattering signal was recorded along time. Statistical values were obtained using an Unpaired *T-student test*. Statistically significant values are indicated: *p*-value < 0,01 (**). Error bars correspond to SEM.

Discussion

Protein aggregation remains a major factor limiting the biotechnological and therapeutic use of many proteins[10]. When globular proteins are employed for such applications, they should remain soluble and active in artificial conditions for which they have not been optimized by evolutionary selection. Therefore, it is not surprising that they tend to aggregate, since they are expressed and manufactured at concentrations that are far beyond their natural limits and stored for periods of time that are much longer than their cellular lifetimes, both factors favoring the establishment of intermolecular interactions that can nucleate the aggregation reaction, especially in the absence of the protective mechanisms provided *in vivo* by the protein quality surveillance machinery. This implies that the use of proteins above their cellular concentrations would require certain sequential adaptations; however, it is not known which the margin of improvement is, if any, for such an approach[61].

Protein aggregation not only affects protein production, but also downstream processes like purification, storage, administration and activity[10]. In this context, effective computational approaches able to rank protein therapeutics according to their solubility and improve this particular property upstream in the industrial pipeline hold the promise to result in significant savings in time and money, since, in principle, they can sift the properties of hundreds to thousands protein variants before performing any experiment.

Re-designing globular proteins to reduce their natural aggregation propensity is, however, not an easy task, since the more potent aggregation-promoting regions lie at the hydrophobic core and it turns difficult to mutate them without disrupting the protein

structure and function[3]. Because mutations at the surface are, in general, less disruptive, an alternative is to mutate those exposed residues that contribute significantly to the aggregation properties of globular structures. The identification of such amino acid required a novel generation of aggregation-prediction algorithms that work on top of protein structures and not on protein sequences. SAP, CamSol and A3D are examples of these kind of programs[9,17,18]. An important feature of these computational tools is that they allow to detect aggregating regions that are relevant only in the native state. These regions are generally formed by residues that, despite being scattered in the sequence, come together in the structure upon folding to form an aggregation-prone surface. Indeed, the three aggregation-prone residues predicted by A3D and mutated here in both fr-GFP and DP47 lie in the same face of the protein, despite some of them are N- and others C-terminal in the respective sequences. Nevertheless, in their present versions, none of these programs takes into account protein stability and thus they do not allow assessing the impact of the designed mutations in the protein structural integrity. This is an important issue, because mutations that impact negatively the thermodynamic and/or kinetic stability would likely facilitate aggregation, increasing the exposure of hydrophobic regions to the solvent, irrespective of whether they are predicted to increase or decrease the protein surface aggregation propensity in the native state. In fact, all these programs predict the fr-GFP Lys and Asp mutational series to be similarly solubilizing, whereas, in contrast, our results indicate that this is not the case. This stresses the advantage of integrating protein stability predictions in A3D, since it allows to discard from further experimental validation potentially destabilized or not-folding protein variants.

Solubis is a recent program that combines aggregation and stability predictions with remarkable success[3,19,42]. However, in contrast to A3D it looks for contiguous aggregation-prone residues in the sequence and not in the structure, implementing the TANGO sequential prediction algorithm[61]. It aims to introduce solubilizing mutations in the middle of these sequence stretches without disturbing the protein structure. V11 and Y39 residues in fr-GFP do not lie inside an aggregation-prone region (APR) according to TANGO. L221 maps in a C-terminal APR, but TANGO predicts that its mutation to Lys would increase the aggregation potential of the sequence (**Figure S10**). Therefore, the gain in solubility observed for GFP/KKK cannot be predicted by Solubis. In the case of DP47 only Leu109 lies in a TANGO-predicted APR, comprising residues 107-GT**L**VTV-112 (**Figure S10**), illustrating again how, despite in certain cases sequential and structural aggregation-prone regions might overlap, this is not generally the case. Thus, the new version of A3D incorporates features that are not combined in any pre-existent algorithm.

Our data demonstrate that there is room to increase the solubility of globular proteins by introducing a small number of mutations that reduce the STAP while preserving the structure. GFP/KKK is clearly more-aggregation resistant than the folding reporter version of the fluorescent protein[27], being able to provide intracellular solubility to the highly aggregation-prone Aβ42 peptide. GFP/KKK only shares with superfolder-GFP a mutation at position 39 (Y39N), indicating that experimental direct evolution and computational re-design can arrive to different solutions in order to optimize protein solubility. This is best illustrated in the case of antibodies, whereas the overlap between the regions responsible for epitope binding and those accounting for their aggregation potential critically compromise their redesign. In fact, for DP47, phage display screening could only identify

solubilizing mutations mapping in the CDRs, without any other surface exposed residue impacting aggregation[29]. In contrast, A3D detected three of these residues, whose mutation to Lys effectively increased the antibody aggregation-resistance, allowing to avoid the modification of amino acids involved in antigen recognition.

As in our designs, the aggregation of several proteins has been shown to be reduced by introducing charged residues[3,42,60]. It has been argued that this effect results from an increase in colloidal stability by increasing the net charge of the protein and therefore that it is rather unspecific[62]. However, the introduction of three Lys residues in fr-GFP reduces the net charge of the protein from -7 to -4, whereas in the non-functional GFP/DDD variant the net charge increases to -10. This suggests that, in our study, the observed impact in solubility result from structural gatekeeping and preservation of the protein conformation, and not from the mere change in the physicochemical characteristics of the protein sequence.

Overall, we developed an improved prediction tool (accessible at http://biocomp.chem.uw.edu.pl/A3D/) that should allow to re-design solubility through protein engineering without impacting the native structure, stability of the target protein. Despite the functionality of the re-designed human antibody should be still confirmed, with an increasing number of protein-based therapeutics under development[63], this algorithm might become a powerful tool to assist the design of well folded and soluble versions of these drugs and might compete or at least complement experimental screening efforts.

ASSOCIATED CONTENT

Supporting Information.

X-ray data collection and refinement statistics spectra, GFP Asp mutants' expression gels, distances between residues in crystal structures, model of GFP/KKK dimer, EMSA assay, screenshots of A3D predictions, table of predicted A3D scores for Asp mutants of DP47, and TANGO predictions for fr-GFP, DP47 and KKK mutants. This material is available free of charge via the Internet at http://pubs.acs.org.

Notes

The authors declare no competing financial interest.

Acknowledgments

This work was funded by the Spanish Ministry of Economy and Competitiveness (BFU2013-44763-P and BIO2016-783-78310-R to SV). SV has been granted an ICREA ACADEMIA award. NV acknowledges support from Science Without Borders (CNPq, Brazil). MGG was supported by the Spanish Ministry of Science and Innovation via a doctoral grant (FPU16/02465). X-ray experiments were performed at the BL-13 beamline at the ALBA synchrotron in collaboration with the ALBA staff. We thank Daniel Christ for kindly providing us the wild type DP47 plasmid.

References

(1) Chiti, F.; Dobson, C. M. Protein Misfolding, Amyloid Formation, and Human Disease: A Summary of Progress Over the Last Decade. *Annu. Rev. Biochem.* **2017**, *86* (1), 27–68.

(2) Ellis, R. J. Macromolecular Crowding: Obvious but Underappreciated. *Trends Biochem. Sci.* **2001**, *26* (10), 597–604.

(3) Ganesan, A.; Siekierska, A.; Beerten, J.; Brams, M.; Van Durme, J.; De Baets, G.; Van der Kant, R.; Gallardo, R.; Ramakers, M.; Langenberg, T.; Wilkinson, H.; De Smet, F.; Ulens, C.; Rousseau, F.; Schymkowitz, J. Structural Hot Spots for the Solubility of Globular Proteins. *Nat. Commun.* **2016**, *7*, 10816.

(4) Tartaglia, GG.; Pechmann, S.; Dobson, C. M.; Vendruscolo, M. Life on the edge: A link between gene expression levels and aggregation rates of human proteins. *Trends in Biochem. Sci.* **2007**, *32 (5)*, 204-206.

(5) Ciryam, P.; Kundra, R.; Morimoto, R. I.; Dobson, C. M.; Vendruscolo, M. Supersaturation Is a Major Driving Force for Protein Aggregation in Neurodegenerative Diseases. *Trends Pharmacol. Sci.* **2015**, *36* (2), 72–77.

(6) Roberts, C.J. Therapeutic Protein Aggregation: Mechanisms, Design, and Control. *Trends Biotechnol.* **2014**, *32* (7), 372–380.

(7) Hamrang, Z.; Rattray, N. J. W.; Pluen, A. Proteins Behaving Badly: Emerging Technologies in Profiling Biopharmaceutical Aggregation. *Trends Biotechnol.* **2013**, *31* (8), 448–458.

(8) Wang, W.; Nema, S.; Teagarden, D. Protein Aggregation-Pathways and Influencing Factors. *Int. J. Pharm.* **2010**, *390* (2), 89–99.

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

(10) Roberts, C. J. Protein Aggregation and Its Impact on Product Quality. *Curr. Opin. Biotechnol.* **2014**, *30*, 211–217.

(11) Espargaró, A.; Castillo, V.; de Groot, N. S.; Ventura, S. The in Vivo and in Vitro Aggregation Properties of Globular Proteins Correlate With Their Conformational Stability: The SH3 Case. *J. Mol. Biol.* **2008**, *378* (5), 1116–1131.

(12) Ventura, S.; Vega, M. C.; Lacroix, E.; Angrand, I.; Spagnolo, L.; Serrano, L. Conformational Strain in the Hydrophobic Core and Its Implications for Protein Folding and Design. *Nat. Struct. Biol.* **2002**, *9* (6), 485–493.

(13) Castillo, V.; Espargaró, A.; Gordo, V.; Vendrell, J.; Ventura, S. Deciphering the Role of the Thermodynamic and Kinetic Stabilities of SH3 Domains on Their Aggregation inside Bacteria. *Proteomics* **2010**, *10* (23), 4172–4185.

(14) Jahn, T. R.; Radford, S. E. The Yin and Yang of Protein Folding. *FEBS J.* **2005**, *272* (23), 5962–5970.

(15) Bemporad, F.; Chiti, F. "Native-like Aggregation" of the Acylphosphatase from Sulfolobus Solfataricus and Its Biological Implications. *FEBS Lett.* **2009**, *583* (16), 2630–2638.

(16) Garcia-Pardo, J.; Graña-Montes, R.; Fernandez-Mendez, M.; Ruyra, A.; Roher, N.; Aviles, F. X.; Lorenzo, J.; Ventura, S. Amyloid Formation by Human Carboxypeptidase D Transthyretin-like Domain under Physiological Conditions. *J. Biol. Chem.* **2014**, *289* (49), 33783–33796.

(17) Sormanni, P.; Aprile, F. a.; Vendruscolo, M. The CamSol Method of Rational Design of Protein Mutants with Enhanced Solubility. *J. Mol. Biol.* **2015**, *427* (2), 478–490.

(18) Zambrano, R.; Jamroz, M.; Szczasiuk, A.; Pujols, J.; Kmiecik, S.; Ventura, S. AGGRESCAN3D (A3D): Server for Prediction of Aggregation Properties of Protein Structures. *Nucleic Acids Res.* **2015**, *43* (W1), W306–W313.

(19) De Baets, G.; Van Durme, J.; Van Der Kant, R.; Schymkowitz, J.; Rousseau, F. Solubis: Optimize Your Protein. *Bioinformatics* **2015**, *31* (15), 2580–2582.

(20) Franzosa, E. a.; Xia, Y. Structural Determinants of Protein Evolution Are Context-Sensitive at the Residue Level. *Mol. Biol. Evol.* **2009**, *26* (10), 2387–2395.

(21) Sasidharan, R.; Chothia, C. The Selection of Acceptable Protein Mutations. *Proc. Natl. Acad. Sci.* **2007**, *104* (24), 10080–10085.

(22) Tokuriki, N.; Stricher, F.; Schymkowitz, J.; Serrano, L.; Tawfik, D. S. The Stability Effects of Protein Mutations Appear to Be Universally Distributed. *J. Mol. Biol.* **2007**, *369* (5), 1318–1332.

(23) Lawrence, M. S.; Phillips, K. J.; Liu, D. R. Supercharging Proteins Can Impart Unusual Resilience. *J. Am. Chem. Soc.* **2007**, *129* (33), 10110–10112.

(24) Schymkowitz, J.; Borg, J.; Stricher, F.; Nys, R.; Rousseau, F.; Serrano, L. The FoldX Web Server: An Online Force Field. *Nucleic Acids Res.* **2005**, *33*, 382–388.

(25) Pujols, J; et al. 2018. AGGRESCAN3D: Toward the Prediction of the Aggregation Propensities of Protein Structures. Methods in Molecular Biology. *Springer. Computational Drug Discovery and Design Book*. **2018,** 4427-443.

(26) Schwede, T.; Kopp, J.; Guex, N.; Peitsch, M. C. SWISS-MODEL: An Automated Protein Homology-Modeling Server. *Nucleic Acids Res.* **2003**, *31* (13), 3381–3385.

(27) Waldo, G. S.; Standish, B. M.; Berendzen, J.; Terwilliger, T. C. Rapid Protein-Folding Assay Using Green Fluorescent Protein. *Nat. Biotechnol.* **1999**, *17* (7), 691–695.

(28) De Groot, N. S.; Aviles, F. X.; Vendrell, J.; Ventura, S. Mutagenesis of the Central Hydrophobic Cluster in Aβ42 Alzheimer's Peptide: Side-Chain Properties Correlate with Aggregation Propensities. *FEBS J.* **2006**, *273* (3), 658–668.

(29) Dudgeon, K.; Rouet, R.; Kokmeijer, I.; Schofield, P.; Stolp, J.; Langley, D.; Stock, D.; Christ, D. General Strategy for the Generation of Human Antibody Variable Domains with Increased Aggregation Resistance. *Proc. Natl. Acad. Sci.* **2012**, *109* (27), 10879–10884.

(30) Juanhuix, J.; Gil-Ortiz, F.; Cuní, G.; Colldelram, C.; Nicolás, J.; Lidón, J.; Boter, E.; Ruget, C.; Ferrer, S.; Benach, J. Developments in Optics and Performance at BL13-XALOC, the Macromolecular Crystallography Beamline at the Alba Synchrotron. *J. Synchrotron Radiat.* **2014**, *21* (4), 679–689.

(31) Kabsch, W. Xds. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2010**, *66* (2), 125–132.

(32) Winn, M. D.; Ballard, C. C.; Cowtan, K. D.; Dodson, E. J.; Emsley, P.; Evans, P. R.; Keegan, R. M.; Krissinel, E. B.; Leslie, A. G. W.; McCoy, A.; McNicholas, S. J.; Murshudov, G. N.; Pannu, N. S.; Potterton, E.A.; Powell, H. R.; Read, R. J.; Vagin, A.; Wilson, K. S. Overview of the CCP4 Suite and Current Developments. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2011**, *67* (4), 235–242.

(33) Adams, P. D.; Afonine, P. V.; Bunkóczi, G.; Chen, V. B.; Davis, I. W.; Echols, N.; Headd, J. J.; Hung, L. W.; Kapral, G. J.; Grosse-Kunstleve, R. W.; McCoy, A. J.; Moriarty, N.W.; Oeffner, R.; Read, R. J.; Richardson, D. C.; Richardson, J. S.; Terwilliger, T. C.; Zwart, P. H. PHENIX: A Comprehensive Python-Based System for Macromolecular Structure Solution. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2010**, *66* (2), 213–221.

(34) Emsley, P.; Lohkamp, B.; Scott, W. G.; Cowtan, K. Features and Development of Coot. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2010**, *66* (4), 486–501.

(35) 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* (1), 65.

(36) Crameri, A.; Whitehorn, E. a.; Tate, E.; Stemmer, W. P. C. Improved Green Fluorescent Protein by Molecular Evolution Using DNA Shuffling. *Nat. Biotechnol.* **1996**, *14* (3), 315–319.

(37) Patterson, G. H.; Knobel, S. M.; Sharif, W. D.; Kain, S. R.; Piston, D. W. Use of the Green Fluorescent Protein and Its Mutants in Quantitative Fluorescence Microscopy. *Biophys. J.* **1997**, *73* (5), 2782–2790.

(38) Fukuda, H.; Arai, M.; Kuwajima, K. Folding of Green Fluorescent Protein and the Cycle3 Mutant. *Biochemistry* **2000**, *39* (39), 12025–12032.

(39) Yang, T. T.; Cheng, L.; Kain, S. R. Optimized Codon Usage and Chromophore Mutations Provide Enhanced Sensitivity with the Green Fluorescent Protein. *Nucleic Acids Res.* **1996**, *24* (22), 4592–4593.

(40) Pédelacq, J.-D.; Cabantous, S.; Tran, T.; Terwilliger, T. C.; Waldo, G. S. Engineering and Characterization of a Superfolder Green Fluorescent Protein. *Nat. Biotechnol.* **2006**, *24* (1), 79–88.

(41) Beerten, J.; Schymkowitz, J.; Rousseau, F. Aggregation Prone Regions and Gatekeeping Residues in Protein Sequences. *Curr. Top. Med. Chem.* **2013**, *12* (22), 2470–2478.

(42) Van der Kant, R.; Karow-Zwick, A. R.; Van Durme, J.; Blech, M.; Gallardo, R.; Seeliger, D.; Aβfalg, K.; Baatsen, P.; Compernolle, G.; Gils, A.; Studts, J. M.; Schulz, P.;

943 Garidel, P.; Schymkowitz, J.; Rousseau, F. Prediction and Reduction of the Aggregation of
944 Monoclonal Antibodies. *J. Mol. Biol.* **2017**, *429* (8), 1244–1261.
945 (43) Chiti, F.; Dobson, C. M. Amyloid Formation by Globular Proteins under Native
946 Conditions. *Nat. Chem. Biol.* **2009**, *5* (1), 15–22.
947 (44) Buß, O.; Rudat, J.; Ochsenreither, K. FoldX as Protein Engineering Tool: Better
948 Than Random Based Approaches? *Comput. Struct. Biotechnol. J.* **2018**, *16*, 25–33.
949 (45) Usmanova, D. R.; Bogatyreva, N. S.; Ariño Bernad, J.; Eremina, A. a; Gorshkova,
950 A. a; Kanevskiy, G. M.; Lonishin, L. R.; Meister, A. V; Yakupova, A. G.; Kondrashov, F.
951 a; et al. Self-Consistency Test Reveals Systematic Bias in Programs for Prediction Change
952 of Stability upon Mutation. *Bioinformatics* **2018**, 1–6.
953 (46) Kumar, V.; Rahman, S.; Choudhry, H.; Zamzami, M. a.; Sarwar Jamal, M.; Islam,
954 A.; Ahmad, F.; Hassan, M. I. Computing Disease-Linked SOD1 Mutations: Deciphering
955 Protein Stability and Patient-Phenotype Relations Article. *Sci. Rep.* **2017**, *7* (1), 1–13.
956 (47) Navarro, S.; Villar-Piqué, A.; Ventura, S. Selection against Toxic Aggregation-
957 Prone Protein Sequences in Bacteria. *Biochim. Biophys. Acta - Mol. Cell Res.* **2014**, *1843*
958 (5), 866–874.
959 (48) Villar-Pique, A.; De Groot, N. S.; Sabaté, R.; Acebrón, S. P.; Celaya, G.;
960 Fernàndez-Busquets, X.; Muga, A.; Ventura, S. The Effect of Amyloidogenic Peptides on
961 Bacterial Aging Correlates with Their Intrinsic Aggregation Propensity. *J. Mol. Biol.* **2012**,
962 *421* (2-3), 270–281.
963 (49) De Groot, N. S.; Pallarés, I.; Avilés, F. X.; Vendrell, J.; Ventura, S. Prediction of
964 "Hot Spots" of Aggregation in Disease-Linked Polypeptides. *BMC Struct. Biol.* **2005**, *5*, 1–
965 15.

(50) De Groot, N. S.; Ventura, S. Protein Activity in Bacterial Inclusion Bodies Correlates with Predicted Aggregation Rates. *J. Biotechnol.* **2006**, *125* (1), 110–113.

(51) Camilloni, C.; Sala, B. M.; Sormanni, P.; Porcari, R.; Corazza, A.; Rosa, M. De; Zanini, S.; Barbiroli, A.; Esposito, G.; Bolognesi, M.; et al. Rational Design of Mutations That Change the Aggregation Rate of a Protein While Maintaining Its Native Structure and Stability. *Sci. Rep.* **2016**, 6:25559.

(52) Varejão, N.; Ibars, E.; Lascorz, J.; Colomina, N.; Torres-rosell, J.; Reverter, D. DNA Activates the Nse 2 / Mms 21 SUMO E 3 Ligase in the Smc 5 / 6 Complex. *EMBO J.* **2018**, *37*: e98306.

(53) Melnik, T.; Povarnitsyna, T.; Solonenko, H.; Melnik, B. Studies of Irreversible Heat Denaturation of Green Fluorescent Protein by Differential Scanning Microcalorimetry. *Thermochim. Acta* **2011**, *512* (1-2), 71–75.

(54) Jackson, S. E.; Craggs, T. D.; Huang, J. R. Understanding the Folding of GFP Using Biophysical Techniques. *Expert Rev. Proteomics* **2006**, *3* (5), 545–559.

(55) Hsu, S. T. D.; Blaser, G.; Jackson, S. E. The Folding, Stability and Conformational Dynamics of B-Barrel Fluorescent Proteins. *Chem. Soc. Rev.* **2009**, *38* (10), 2951–2965.

(56) Pallarès, I.; Vendrell, J.; Avilés, F. X.; Ventura, S. Amyloid Fibril Formation by a Partially Structured Intermediate State of A-Chymotrypsin. *J. Mol. Biol.* **2004**, *342* (1), 321–331.

(57) Patterson, G. H. A Photoactivatable GFP for Selective Photolabeling of Proteins and Cells. *Science (80).* **2002**, *297* (5588), 1873–1877.

(58) Zacharias, D. a; Violin, J. D.; Newton, A. C. Partitioning of Lipid-Modified Monomeric CFPs into Membrane Microdomains of Live Cells. *Science* **2002**, *296*, 913–916.

(59) Snapp, E. L.; Hegde, R. S.; Francolini, M.; Lombardo, F.; Colombo, S.; Pedrazzini, E.; Borgese, N.; Lippincott-Schwartz, J. Formation of Stacked ER Cisternae by Low Affinity Protein Interactions. *J. Cell Biol.* **2003**, *163* (2), 257–269.

(60) Perchiacca, J. M.; Bhattacharya, M.; Tessier, P. M. Mutational Analysis of Domain Antibodies Reveals Aggregation Hotspots within and near the Complementarity Determining Regions. *Proteins Struct. Funct. Bioinforma.* **2011**, *79* (9), 2637–2647.

(61) Fernandez-Escamilla, A.-M.; Rousseau, F.; Schymkowitz, J.; Serrano, L. Prediction of Sequence-Dependent and Mutational Effects on the Aggregation of Peptides and Proteins. *Nat. Biotechnol.* **2004**, *22* (10), 1302–1306.

(62) Kramer, R. M.; Shende, V. R.; Motl, N.; Pace, C. N.; Scholtz, J. M. Toward a Molecular Understanding of Protein Solubility: Increased Negative Surface Charge Correlates with Increased Solubility. *Biophys. J.* **2012**, *102* (8), 1907–1915.

(63) Elgundi, Z.; Reslan, M.; Cruz, E.; Sifniotis, V.; Kayser, V. The State-of-Play and Future of Antibody Therapeutics. *Adv. Drug Deliv. Rev.* **2016,** *122*, 2-19.