Automated optimisation of solubility and conformational stability of antibodies and proteins

Angelo Rosace[a,b,1,†], Anja Bennett[a,c,d,†], Marc Oeller[a], Mie M. Mortensen[e,f], Laila Sakhnini[a], and Pietro Sormanni[a,*]

[†] These authors contributed equally to this work
[*]Correspondence to ps589@cam.ac.uk

a. Centre for Misfolding Diseases, Yusuf Hamied Department of Chemistry, University of Cambridge, Lensfield road CB2 1EW Cambridge, UK
b. Master in Bioinformatics for Health Sciences, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain
c. Department of Mammalian Expression, Global Research Technologies, Novo Nordisk A/S, Novo Nordisk Park 1, 2760 Måløv, Denmark
d. BRIC, Faculty of Health and Medical Sciences, University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen N, Denmark
e. Department of Purification Technologies, Global Research Technologies, Novo Nordisk A/S, Novo Nordisk Park 1, 2760 Måløv, Denmark
f. Faculty of Engineering and Science, Department of Biotechnology, Chemistry and Environmental Engineering, University of Aalborg, Fredrik Bajers Vej 7H, 9220 Aalborg, Denmark
1. Current address: Institute for Research in Biomedicine (IRB), The Barcelona Institute of Science and Technology, Barcelona, Catalonia, Spain

Abstract

Developability issues, such as aggregation, low expression yield, or instability over long-term storage often hamper the development of biologics. Developability is underpinned by a complex interplay of biophysical properties, including conformational stability and solubility. Advances have been made in the optimisation of individual properties, however, multi-trait optimisation remains highly challenging, as mutations that improve one property often negatively impact the others. Here, we introduce a fully automated computational strategy for the rational design of conformationally stable and soluble protein and antibody variants harbouring multiple mutations. This pipeline, called CamSol Combination, leverages phylogenetic information to reduce false positive predictions, and combines a rapid method of predicting solubility changes with an empirical energy function. We experimentally validate the method's predictions on a nanobody isolated with yeast-display from a synthetic library, by producing 12 designs ranging from single-point mutants to a quadruple mutant. All 12 designed variants retained antigen binding, had improved thermal stability, and decreased propensity to precipitate (increased relative solubility), and 8 also had reduced cross-reactivity. The melting temperature was improved by 13.6 ºC with 4 rationally designed mutations. We make the method available as a webserver at www-cohsoftware.ch.cam.ac.uk

Significance

Protein-based biologics, such as antibodies and enzymes, are crucial reagents in research, industrial biotechnology, and diagnostics, and are increasingly used to treat a wide range of diseases as biotherapeutics. Often, biologics with suitable functionality are discovered, but their development into practically useful molecules is impeded by developability issues. Conformational stability and solubility are arguably the most important biophysical properties underpinning developability potential, as they determine colloidal stability and aggregation, and correlate with expression yield and poly-specificity. Here we introduce and experimentally validate a fully automated computational method and associated webserver for the rational design of proteins and antibodies with improved stability and solubility. Computational methods are rapid, inexpensive, and have no material requirements, which makes their implementation into biologic development pipelines particularly attractive.

Introduction

Over the past few decades, protein-based biologics have risen to become a key class of therapeutic molecules (1, 2). These biologics offer a range of favourable characteristics, such as high specificity, low toxicity and immunogenicity, and the possibility of replacing or supplementing endogenous proteins and hormones, which make them very valuable for drug development purposes (3). Furthermore, proteins and antibodies are also key reagents in biomedical research and diagnostics, and enzymes find crucial applications in industrial biotechnology.

However, proteins and antibodies destined to research, diagnostic, biotechnology, and especially therapeutic applications are required to endure a wide range of stresses related to manufacturing, development, shipping, storage, and administration. As these stresses are not present in vivo, natural proteins and antibodies have not evolved to withstand them (4). Besides, therapeutic applications, in particular, have stringent requirements, which include strong biological activity at relatively low administration dosage and frequency, high concentration formulations, and long product shelf-life (5).

Taken together, the stresses that biologics must endure, and the requirements they need to meet imply that in most cases all biophysical properties must be optimised beyond their typical natural levels (5). While methods of in vitro directed evolution are routinely employed to optimise binding affinity, the simultaneous optimisation of multiple biophysical traits remains highly challenging (6). Functional proteins and antibodies exhibit a balance of different biophysical traits, including conformational stability, solubility, interaction affinity, and selectivity. As the fundamental forces that drive protein folding, aggregation, and binding are ultimately the same, such traits are often conflicting, in the sense that mutations that improve one of them tend to worsen the others (5, 7–9). This process of multi-traits optimization can be compared to solving a Rubik's cube, where each face represents one biophysical property. Changing one face will affect other faces, perhaps detrimentally, and solving a single face is much simpler than completing the puzzle. In protein engineering, it remains highly challenging to select mutations that selectively improve properties of interest while leaving the others unaffected, and there is an unmet need to develop technologies that enable the simultaneous optimisation of different traits.

Computational approaches offer a promising avenue to generate such technologies, as at least in principle they allow for highly controlled parallel screenings of multiple biophysical properties. Moreover, computer calculations are rapid, inexpensive, and have no material requirements, which makes their implementation in antibody and protein development pipelines particularly attractive.

In this work, we contribute to addressing this challenge by introducing a fully automated computational strategy for the simultaneous optimisation of solubility and conformational stability, which have often been reported to be conflicting (5, 10–12). Our approach leverages

phylogenetic information to reduce false positive predictions and to prevent the modification of functionally relevant sites. Furthermore, the pipeline relies on the CamSol method (13) to carry out predictions of solubility changes upon mutation, and on the energy function Fold-X (14) for the predictions of the associated stability changes. Both algorithms have separately been extensively validated on experimental data for a wide range of proteins and antibodies (CamSol Refs (4, 13, 15–21) and FoldX Refs (10, 22–26)). Our pipeline also includes an ad-hoc recipe to obtain and exploit suitable phylogenetic information for immunoglobulin variable domains, as these are a key class of biologics that, because of their modular nature, cannot be handled with standard tools for searching homologs. In the case of antibodies, our method exploits custom-built multiple sequence alignments (MSAs) selectable depending on the user's requirements and representing the mutational space observed in human antibodies, mouse antibodies, nanobodies, and post-phase-I clinical-stage antibodies.

Solubility and conformational stability are among the most important properties underpinning the developability potential of biologics, which is defined as the likelihood of a drug candidate with suitable functionality to be developed into a manufacturable, stable, safe, and effective drug that can be formulated to high concentrations while retaining a long shelf-life (27–29). Poor solubility is a major bottleneck for manufacturing (28, 30), quality, and safety of pharmaceutical formulations (27, 30, 31), and it has also been linked to diminished binding specificity (32, 33). Similarly, high conformational stability is essential to ensure efficacy and safety during manufacturing, formulation, storage, shipping, and administration (28, 34). Various studies have reported a correlation between conformational stability and recombinant expression yield (24, 35, 36), and high stability translates into enhanced ability to withstand the physical or chemical stresses encountered during manufacturing and formulation (37).

Fully or partially unfolded proteins and antibodies not only lose their activity, but can readily aggregate, as the solubility of the unfolded state is typically much lower than that of the native state (15). Aggregation from partially unfolded states can be irreversible, and aggregated biologics have been reported to induce immunogenicity upon injection (38, 39), which can be fatal in some cases (40). Although it is still not clear which types of aggregates mediate immunogenicity, regulatory agencies require formulations with minimal amount of aggregates at the end of the formulation shelf-life to grant market approval (41, 42).

Solubility and conformational stability ultimately depend on the amino acid sequence of the protein under scrutiny, as well as on some extrinsic properties of the buffer such as pH, temperature, and the presence of excipients (4). Together, they influence other essential properties that underpin developability potential, including viscosity and colloidal stability, which also strongly depend on the protein concentration of the formulation. Colloidal stability is a kinetic property that depends on the height of the free energy barrier that separates the native state from the aggregated states (4). While solubility is defined as the critical concentration observed when soluble and insoluble phases are in equilibrium, colloidal stability is defined by the long-term integrity of a formulation, and hence by the time it takes for

aggregation to occur (4). Solubility and conformational stability determine colloidal stability through their link with aggregation. Self-association can be triggered via two main pathways. In one, aggregation hotspots on molecular surfaces drive the assembly of two or more molecules, forming aggregates that then may act as seeds to drive further aggregation and may also increase viscosity (43, 44). In the other, the presence of partially or fully unfolded states lead to the transient exposure of hydrophobic patches that can elicit the formation of misfolded aggregates (15). Therefore, maximising solubility and conformational stability can be expected to translate into increased colloidal stability and hence lower aggregation rates.

As aggregation is the most common of the various degradation reactions that a protein can experience during its biotechnological or biotherapeutic development (45), effective computational methods to rationally design mutations that reduce aggregation without hampering other properties can be expected to substantially streamline biologics development pipelines. The automated rational design pipeline that we introduce here works by removing surface-exposed aggregation hotspots leading to poor solubility, and by proposing mutations expected to increase conformational stability and solubility, thus decreasing the population of partially or fully unfolded states in solution. We validate this pipeline experimentally by using it to design mutational variants of an anti-human serum albumin (HSA) nanobody derived from a yeast-display naïve synthetic library (46).

Results

The goal of the computational pipeline is to design protein or antibody variants by predicting combinations of mutations that increase both stability and solubility, or one of the two without affecting the other. The method relies on the knowledge of the native structure of the target protein, or on the availability of a good structural model, and on a multiple sequence alignment (MSA) of homologous sequences. A position specific scoring matrix (PSSM) is then extracted from the MSA, to provide information on the frequency of the amino acids observed among homologous proteins at each position of the input structure.

The algorithm is built to handle input proteins consisting of multiple polypeptide chains, as well as bound structures where the binding partner can be excluded from the design. Furthermore, in the case of antibody variable domains (Fvs), a custom pipeline is provided to automatically generate PSSMs from the input sequences, based on user-selectable databases of species-specific antibodies or therapeutic antibodies. Users are also allowed to provide additional input parameters as explained below in the *Algorithm Pipeline* section.

Phylogenetic information reduces false positive predictions of stability change

The phylogenetic information is used in the pipeline to enable the identification of potential mutations based on their observed frequencies in natural variants of the protein or antibody under scrutiny. Therefore, these mutations are more likely to be well-tolerated, and thus less likely to disrupt stability (47). We hypothesized that restricting the mutational space to those mutations that are enriched in natural variants will greatly decrease the number of False

Positive (FP) predictions, which can be quantified through the False Discovery Rate (FDR, see Supplementary Methods).

For stability predictions, we define False Positives those mutations predicted to be stabilizing, while being destabilizing in practice, and False Negatives (FN) as the opposite (see **Fig. 1A**). In the context of computational design, and in particular of algorithms aimed at yielding mutational variants that improve one or more biophysical properties, FNs can be regarded as missed opportunities. These mutations will not be suggested by the algorithm, even if they would be beneficial. However, provided that at least some beneficial mutations are identified, FNs are not the main concern. FP mutations on the other hand will be suggested by the algorithm, leading to a potential waste of time and resources in the experimental production and characterization of designs containing such prediction mistakes.

It is therefore of paramount importance that a method for the automated design of mutational variants has the lowest possible FDR (48). We decided to assess the performance using a recently published database of experimentally characterized mutations (49), which is a curated subset of the ProTherm database where inaccuracies and biases have been removed (49–51). This database contains thermodynamic information on the stability changes caused by 755 point mutations within 81 different proteins, and it was developed for the specific purpose of benchmarking computational methods of predicting stability changes (49).

The results of our analysis show that the FDR of the FoldX energy function can be decreased in a statistically significant way by incorporating filters based on phylogenetic information (**Fig. 1B**). More specifically, we have screened different parameters for the search of homologous sequences and different implementations of the PSSM (see Supplementary Methods and **Fig. S1**). Our results show that the FDR of the stability predictions can be decreased from ~26% to ~21% ($p<0.0001$) by restricting the search space to mutations with positive log-likelihood, that is mutations that are observed at that position more often than their background probability (i.e., more often than expected by random chance). The FDR can then be further improved to ~15% ($p<0.00001$) by only considering mutations that both have a positive log-likelihood and increase the frequency over that of the WT residue (i.e., positive Δlog-likelihood, **Fig. 1B**). These results are in line with strategies of consensus designs (52, 53), as well as with previously reported findings for different forcefields (24, 36). All p-values were calculated explicitly through random resampling (see Supplementary Methods), which demonstrates that it is the specific choice of phylogenetic filtering that is behind this performance improvement, and not a generic restriction of the database size. We also verified that the observed performance improvement is not a consequence of simply removing many of those mutations with predicted ΔΔG close to 0, which would be within the expected error of FoldX (see Supplementary Methods and **Fig. S3**).

We further verified that similar results can be obtained when using a PSSM containing only raw frequency counts, as opposed to log-likelihood scores (**Fig. S1**). Such PSSM of raw

frequencies is often referred to as Position Probability Matrix (PPM) or Position Weight Matrix (PWM). While log-likelihood scores are generally preferable, as they correct the observed frequencies for the expected background probability of observing each amino acid at a given position by random chance, their calculations are often unreliable for alignments containing small numbers of sequences (e.g., < 50). Therefore, in cases where the input protein only has a few homologs, the employment of a PWM provides more reliable information, and our algorithm automatically switches to it when less than 50 sequences are contained in the input MSA.

We note that a similar analysis was not possible for the solubility prediction, as a dataset of experimentally measured solubility changes upon mutation, which is large enough to draw any statistically significant conclusions on changes in FDR, is not currently available. Moreover, though all globular proteins must retain good stability to function, one may expect to find an evolutionary pressure towards maintaining high solubility only for those proteins that are expressed to high concentrations, suggesting that the phylogenetic filtering we implemented may not be as beneficial for solubility predictions (15, 54). However, the typical performance of the CamSol method in ranking protein and antibody mutational variants is high (Pearson $R \gtrsim 0.9$) (4, 13, 16, 18, 21, 33), thus indicating that the FDR of CamSol predictions is already low.

In summary, implementing phylogenetic filtering can reduce the FDR of stability predictions by about 10%, at the expense of being left with a smaller – but typically large enough – mutational space to sample. A reduced mutational space means that sometimes potentially beneficial mutations will be left out, but also that the overall pipeline will run much faster, as it only needs to sample a sub-region of the mutational space that is evolutionarily grounded.

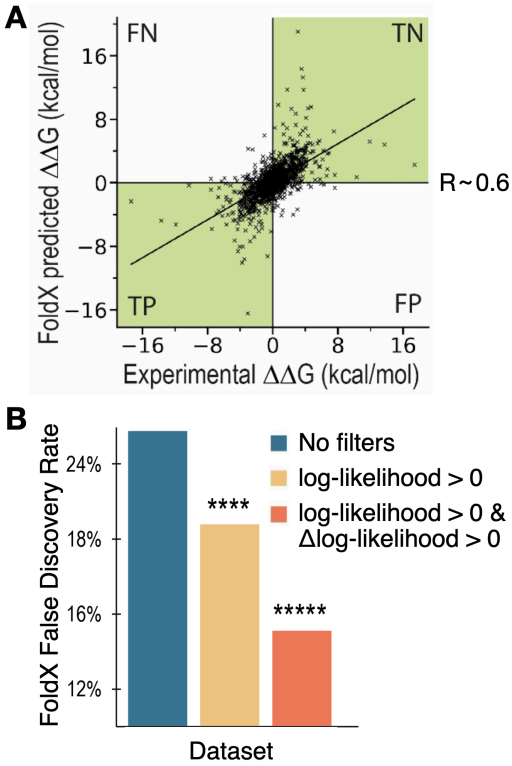

Figure 1. (**A**) Comparison of the experimental ΔΔG against the FoldX-predicted ΔΔG values. Labels are applied to the four quadrants of the graph. (FN: False Negatives, TN: True Negatives, TP: True Positives, FP: False Positives). We define positive mutations those that result in a negative ΔΔG value, and hence are stabilizing. The Pearson correlation coefficient (R) for the dataset is also shown. The green quadrants identify mutations whose overall effect (stabilizing or destabilizing) is correctly predicted. (**B**) Bar plot of the FDR of the FoldX prediction obtained by applying different filtering based on phylogenetic information (see legend). The PSSMs employed were obtained with HHblits(55) using a coverage parameter of 60 and identity of 95. Statistical significance was calculated explicitly with random resampling (see Supplementary Methods, **** $p < 0.0001$ and ***** $p < 0.00001$).

Algorithm pipeline

(i) Input processing

The method is implemented as a webserver (*www-cohsoftware.ch.cam.ac.uk/index.php/camsolcombination*). Therefore, the user interacts with it through an input form. The required input consists of a protein structure in pdb format, its type (antibody Fv/nanobody or other protein) and the alignment file(s) to use in the pipeline. If the input protein is not an antibody or nanobody, the user must provide .hhm or .a3m files obtained with HHblits for each chain in the input protein that is not manually excluded from the design. Such files may be obtained by running a HHblits search with a local installation (55), or more straightforwardly from the MPI Bioinformatics Toolkit webserver (56) by following the instructions linked on the CamSol combination webserver homepage.

If the input protein is an antibody or nanobody, the user can choose between a set of pre-compiled alignments of species-specific antibodies (human or mouse), or therapeutic antibodies (post phase-I clinical trial) (57, 58), or single-domain antibodies obtained from multiple camelid species (59) (see Supplementary Methods). The need for special alignments for antibodies is due to their modularity, and high degree of conservation in the framework regions coupled with high variability in the binding loops. These features make standard alignment tools unsuitable for antibodies, as an excessive number of gaps is typically introduced in loop regions. Furthermore, especially when antibodies are destined to therapeutic applications, it is important to be able to select candidate substitutions that are observed in the relevant species (e.g., from sequences of human origins) to reduce the chances of introducing immunogenic motifs. Similarly, for single-domain antibodies, candidate substitutions must be selected based on alignments of VH sequences known to be able to autonomously fold and remain soluble in the absence of a VL partner.

The user can also provide optional parameters to the design calculations. These include residue positions or whole chains to exclude from the design, the maximum number of mutations to try simultaneously, and residues to exclude from the list of potential substitution targets. For example, the default procedure is to exclude cysteines and methionines, as these are known oxidation sites, and solvent-exposed cysteines may trigger covalent dimerization. Users may also consider excluding asparagines, as these are often prone to deamidation or glycosylation.

The algorithm automatically identifies groups of identical chains that may be present in the input structure. In this way, different mutations are never carried out on chains that are identical in sequence, as in practice such chains would be encoded by the same gene. Identical chains are identified from the *seqres* field in the header of the input PDB file, and from the *atom* sequence only if the *seqres* field is not present. In the PDB format, the *seqres* sequence corresponds to the sequence that was used in the structural determination experiment, while the *atom* sequence is the one for which 3D coordinates have been obtained. The two sequences may differ for example when there are regions of missing electron density. An option is also provided to manually input sequence identity specifications, as groups of chain IDs of identical polypeptide chains, so to offer maximum flexibility to accommodate 'non-canonical' PDB structures.

Upon submission, a unique identifier is assigned to the web server job to track its calculation and correctly handle its output files. The job is then added to a job queue. When started, the algorithm first calculates a log-likelihood Position Specific Scoring Matrix (PSSM) from the MSA (**Fig. 2**). If the number of aligned sequences is below 50, a simpler raw frequency PWM is used instead. The PSSM provides information on which residues are most conserved at each position in the protein sequence, and which mutations are observed in natural variants at each site.

As an example, **Supplementary File 1** is the final report of a run on bacillus licheniformis alpha-amylase (PDB ID 1bli). The PSSM plotted therein, calculated from an alignment of homologs obtained from the HHblits web server by simply uploading the sequence (56), reveals that the active site of the enzyme (Asp 231, Glu 261 and Asp 328) is highly conserved, and so are known ion-binding sites on its surface. Given the high conservation in the PSSM, no mutation would be tried at these positions in the automated pipeline. This example shows that incorporating phylogenetic information can preserve functionally relevant residues in automated computational design pipelines, without the need to manually exclude them from the calculation, which, depending on the case, may require high domain expertise.

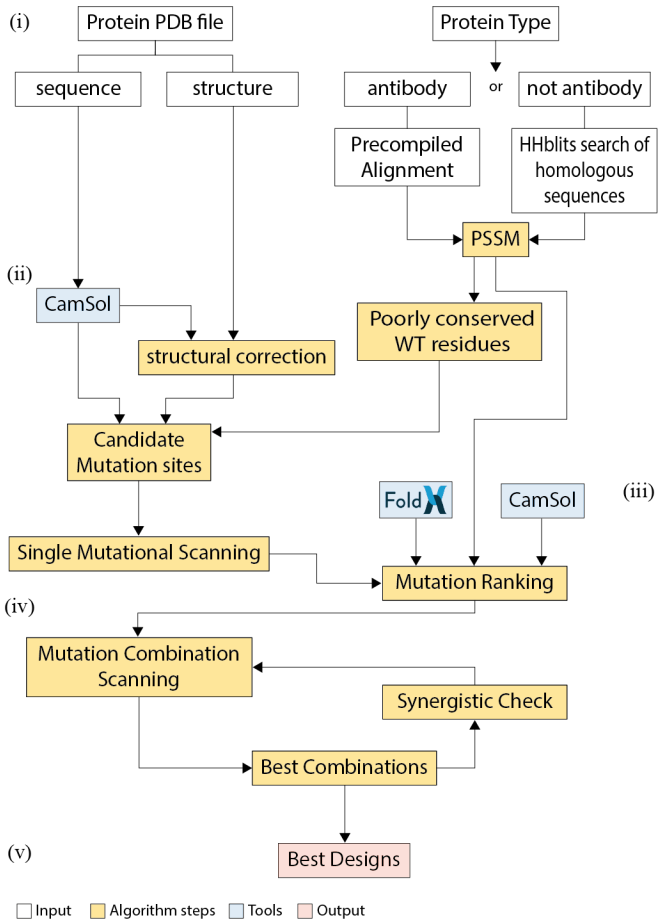

Figure 2. Schematic representation of the algorithm pipeline. Roman numbers refer to the subsections of the algorithm pipeline section in the main text. Coloured in white are the input processing steps, in blue the tools included in the algorithm, in yellow the core algorithm steps, and in red the output.

(ii) Selection of candidate mutation sites

After processing the input, the algorithm automatically identifies candidate mutation sites. Candidate sites for mutation are identified based on their contribution to solubility, as predicted with the CamSol method (13), and their level of conservation, as obtained from the PSSM (**Fig. 2**).

First, CamSol is used to calculate intrinsic solubility profiles for each chain of the input structure. These profiles associate each residue with a number that reflects its impact on the overall solubility (13, 15). This calculation relies solely on the knowledge of the amino acid sequence, which is extracted from the *seqres* field of the input pdb file (if present, otherwise from the *atom* sequence) to account for the contributions to the solubility of possible regions with unassigned 3D coordinates. CamSol predictions rely on a combination of physicochemical properties of amino acids, including charge, hydrophobicity, and propensity to form secondary structure elements (13). These properties are first considered at the individual residue level, then averaged locally across sequence regions to account for the influence of neighbouring residues, and finally considered globally to yield a solubility score (15, 16). This score, which is one number for the whole sequence or multiple sequences in the case of a complex (16), will later be used to rank different mutational variants. Intrinsic profiles are used to identify candidate mutation sites (i.e., residues in the sequence) that contribute strongly to poor solubility of the unfolded state of the protein. At this step, only those sites that are solvent-exposed on the protein surface are flagged as possible candidate mutation sites, as buried poorly soluble residues are expected within the hydrophobic core of globular proteins. More specifically, residues are considered solvent-exposed if they have a solvent exposure of at least 10%, calculated by dividing the observed solvent accessible surface area of each residue by that of the amino acid type under scrutiny in the context of a fully extended Gly-Xxx-Gly three-peptide (13). Typically, a relatively low solvent exposure results in structurally corrected solubility scores close to zero, because of the effect of the exposure weight in the CamSol structurally corrected calculations (see Ref. (13)). However, thermal fluctuations may transiently expose such poorly soluble motifs, thus leading to aggregation. It is therefore important to identify poorly soluble sites that are close to the surface and may therefore become exposed through thermal fluctuations. We refer to this group of candidate mutation sites as *identified from the solubility of the sequence*.

Second, the algorithm focuses on predicting potential aggregation hotspots on the protein surface itself, which are defined as groups of residues on the surface that create patches of poor solubility. The prediction is based on the CamSol structurally corrected calculation as described in detail in Ref. (13). In this calculation, the intrinsic profiles are modified to account for the proximity of amino acids in the 3D structure and for their solvent exposure. The corrected profiles are used to identify those mutation sites that contribute most to the overall aggregation propensity of possible surface hotspots. We refer to this group of candidate mutation sites as *identified from the solubility of the structure*.

Third, also those sites that are relatively well conserved (with conservation index greater than 0.25, see Supplementary Methods), and where the frequency of the WT residue is low (log-likelihood ≤ 0, or in case of a PWM frequency < 0.05) are added to the list of candidate mutation sites, since they may be stability liabilities and represent good candidate sites to further optimize solubility and stability. The lower bound on the conservation index is necessary to avoid flagging positions that are intrinsically poorly conserved (e.g. within the

CDR3 of an antibody). We refer to this group of candidate mutation sites as *identified from conservation*.

Fourth, those residues that are exposed on the surface, and that have PSSM-permitted mutations predicted to increase the solubility are also added to the list of candidate mutation sites, unless the position has a conservation index greater than 0.7 and the wild-type residue is already the consensus residue. This group of candidate mutation sites is referred to as *identified from exposed solubility*, as mutations here can be exploited to increase the solubility even if the sites are not predicted as liabilities. If the input protein has already got many candidate mutation sites in the previous three classes (maximum total number currently set to 100), then sites in this latest class are discarded from the list of candidates to safeguard computational efficiency.

Finally, an option is provided for the user to input *custom* mutation sites, which, if given, are considered in the next steps of the pipeline alongside those identified automatically.

(iii) Single mutational scanning

Once the candidate mutation sites are identified, all possible amino acid substitutions allowed by the PSSM are tested as candidate mutations at each site. The user can decide which criterion to use for allowed substitutions from the filtering in **Fig. 1B**. The two options are either all substitutions with log-likelihood > 0 (i.e. observed more than random, positively enriched), or only those that also increase the frequency from the WT residues (log-likelihood > 0 & Δlog-likelihood > 0), which is the default. Based on the results in **Fig. 1B**, the first option should be used in those cases where too few or no mutations are suggested by the algorithm at the end of the pipeline when running with the default log-likelihood > 0 & Δlog-likelihood > 0.

Allowed substitutions at each site are then singly ranked using the CamSol intrinsic solubility score, as this calculation is extremely fast (4). Except for mutations happening at non-solvent-exposed sites identified from conservation, all mutations predicted to decrease solubility are discarded at this step. The output of this first part of the algorithm thus consists of a longlist of all PSSM-permitted mutations at all identified sites, comprising only solubility-improving mutations at those sites identified based on their low solubility.

This longlist is then used as the starting point for the calculation of stability change upon mutation with the energy-function Fold-X (14). The energy-function is used to calculate for each longlisted mutation a ΔΔG score. This is the predicted difference in stability (ΔG) between the wild type and the mutant (60). Only mutations with calculated ΔΔG < 0 are predicted to be stabilizing (or at least not de-stabilizing), and consequently further considered in the method pipeline. Therefore, the shortlist of candidate mutations contains single-point mutants characterized by the difference in CamSol solubility score between mutant and WT (Δ CamSol score), the difference in calculated stability (ΔΔG), and the difference in frequency (Δlog-

likelihood). We are thus left with a list of mutations that are singly predicted to increase protein solubility and/or conformational stability.

If identical chains were identified in the input processing, the effect of the shortlisted point mutations is propagated to all chains identical to the chains on which they had been modelled. This means that the CamSol score is re-calculated after mutating the sequence of identical chains, and that the ΔΔG, initially calculated on only one chain for computational efficiency, is multiplied by the number of chains in the same group.

Finally, these scores reflecting the changes in solubility, stability, and conservation are combined into the Mutation Score (see Supplementary Methods), which is used to provide a preliminary ranking of the single mutations. An intermediate output table is produced, containing the results of the single-mutational scanning. This table lists all the "beneficial" single-point mutations identified and all their aforementioned scores. Mutations are named with a single string concatenating WT amino acid, Chain ID, pdb residue number, and mutated amino acid (e.g. LA24D to denote a leucine to aspartic acid substitution at residue number 24 of chain A).

(iv) Combining multiple mutations

The shortlisted mutations are then combined to create designs harbouring multiple mutations. Before beginning the combination, the single-point mutants with a negative Mutation Score (if any) are discarded, and after that, all Δlog-likelihood listed in the single mutational scanning output table are normalized to make them comparable across different chains. This normalization consists in dividing the Δlog-likelihood by the standard deviation of the log-likelihood from the PSSM of the chain in which the mutation was carried out. Afterwards, the Mutation Score is recalculated with the updated Δlog-likelihood. To increase computational efficiency, the mutation combination process does not re-run all calculations for each possible combination of mutations, as in particular FoldX takes a while to run. In the first instance, point mutations are combined together by summing their ΔΔGs and Δlog-likelihoods, while the intrinsic solubility score is re-calculated from the sequences harbouring all mutations, as the CamSol intrinsic calculation is adequately fast (~200 sequences/seconds on a single core (15)). A combination of mutations is flagged as "potentially interacting" if at least two of its mutation sites in proximity in the input structure, as assessed from the contact map that was calculated at the beginning in the CamSol structurally corrected solubility calculation (13) (see Supplementary Methods). The underlying assumption for this procedure is that mutations that are distant in the structure may be expected to yield an addictive contribution to the stability, or in other words that the overall stability change can be calculated by summing the stability changes of the individual mutations. However, this assumption breaks down when the mutations are close to each other in the structure. Therefore, combinations containing at least two mutations in proximity are flagged so that, if they occur in the final shortlist, their predicted stability can be recalculated by explicitly modelling these combinations (see step (v)). Once the three metrics (ΔCamSol score, FoldX ΔΔG, and Δlog-likelihood) are computed for all

combinations, the Mutation Score is calculated. If multiple identical chains are present in the input structure, the CamSol and ΔΔG scores are re-calculated as explained in section (iii). Point mutations are combined recursively in this way until the maximum number of simultaneous mutations decided by the user is reached, or until all identified suitable mutation sites have been combined together, whichever happens first.

In the first step of combining the mutations, the single-point mutants are combined to generate double mutants. At each recursive step, the single-point mutants are combined with the mutant combinations generated in the preceding step. Care is taken to ensure that each mutation site is represented only once in a combination, and that identical combinations are never re-assessed (e.g., HA52P, TA54G and TA54G, HA52P). To preserve computational efficiency, the combinatorial space is gradually reduced during the combination process. Starting from the triple combinations onward, the algorithm considers only the top-ranking substitution for each candidate mutation site from the single-mutational scanning, as ranked by their Mutation Score.

We define a group of combinations as the ensemble of all combinations with the same total number of mutations (e.g. double mutants, triple mutants, etc.). Within each group, combinations are ranked according to their Mutation Score. The predicted top-ranking combination across all groups is almost certainly in the group with the maximum number of mutations attempted. However, this may not be beneficial in practice, as with each new mutation the chances of introducing a false positive in the combination increase. Therefore, a procedure is implemented to select those groups that embody the best balance between the number of mutations and predicted gain in solubility and/or stability.

The best groups are identified by those points in the recursive combination process, where the gain of carrying out an additional mutation becomes less favourable than it has been for the preceding mutations, as assessed by the growth of the Mutation Score as a function of the number of mutations (see Supplementary Methods). This procedure thus identifies one or more groups that embody the maximum gain with the minimum number of mutations (**Fig. S4**). Then, the algorithm creates a final shortlist of designed variants, which contains the three top-ranking combinations for each of these best groups, as well as the top-ranking combination from all other groups, so that at least one combination per total number of mutations is proposed in the final output for the user to consider.

(v) Check for potentially interacting mutations

If any design flagged as "potentially interacting" ends up in the final shortlist, then its mutations are explicitly modelled one by one to make sure that the stability of the variant is not compromised by unfavourable interactions between different mutations. This process enables to test the assumption that the ΔΔG for a combination can be calculated as the sum of the ΔΔGs of its single-point mutations. Operationally, this is achieved by carrying out each mutation in the combination under scrutiny subsequently, using the output model of the previous step as input, so that the final model contains all mutations.

All mutations considered up to now where singly predicted to be stabilising ($\Delta\Delta G < 0$). Therefore, if this test discovers a mutation predicted to be destabilizing ($\Delta\Delta G \geq 0$), it indicates unfavourable interactions with other mutations that have already been modelled at nearby positions. If such mutation is found, the algorithm tries to replace it with another substitution among those that were shortlisted for the same position by the single mutational scanning step. If applying this replacement results in a $\Delta\Delta G < 0$, then the alternative mutation is kept. If this is not the case, the process is repeated to explore up to three alternative substitutions per position. If no suitable alternative is found, the disruptive mutation site is removed from the combination under scrutiny.

After this check, the Mutation Score of the combination is updated with the new $\Delta\Delta G$ that has been explicitly calculated. If the new score is lower than that initially predicted by summing the $\Delta\Delta G$ of the point mutations, a comparison is carried out with all other designs in the same group, as now a different combination with a higher Mutation Score might exist. If this is the case, the new top-ranking design is shortlisted, and the mutation-interaction check is repeated on it. If needed, this procedure is repeated up to three times for each group. Ultimately, the best ranking design among those combinations that have been checked (or one without the "potentially interacting" flag) is returned as the best for its group. This process therefore updates the final shortlist and the predicted best designs.

The final output consists of an html report with the description of the top-ranking designs, and the key results of each step of the pipeline, including graphs and descriptions (see for example **Supplementary files 1** and **2**). A zip folder is also provided containing detailed csv tables with the results of all calculations, and the modelled structure of all top-ranking designs in PDB format.

Experimental validation on a nanobody

After developing the pipeline, we applied it to a nanobody that was isolated from a recently introduced naïve yeast-display library (46). This nanobody, called Nb.b201 in the original work (46), binds to human serum albumin with a KD in the high nanomolar range. The automated design procedure was run by allowing a maximum of 6 simultaneous mutations, with the phylogenetic filtering of log-likelihood>0 only, and by using the pre-compiled MSA of single-domain antibody sequences (see Supplementary Methods). As input, we used the structure of Nb.b201 (chain C of PDB ID 5vnw), and paratope residues were excluded from the list of candidate mutation sites (PDB residue numbers: 33,50,52,58,98,102-105).

The first step of the method is to compare the input sequence with the PSSM of the single-domain antibody MSA (**Fig. 3A**). Nb.b201 originated from a library based on the consensus sequence of nanobodies observed in the PDB (46). Therefore, we find that the framework regions are mostly identical to the consensus sequence (top row of the matrix), which focuses the candidate mutation sites to the CDR regions. The CamSol structurally corrected profile

(**Fig. 3B**) reveals some small aggregation hotspots that cluster together on the surface (**Fig. 3C**), the larger of which comes from a paratope region in the CDR3, around residue F105. However, because the paratope regions were excluded from the design, the algorithm did not directly flag these sites for mutation. Conversely, using the criteria described in section "(ii) Selection of candidate mutation sites", the algorithm identified a total of 30 possible sites for mutation (**Fig. 3C** and **Supplementary File 2**), some of which are relatively close to the hotspots and may thus provide compensatory mutations. Of these 30 sites, 7 were shortlisted at the end of the single mutational scanning step, yielding a total of 17 different point mutations (**Fig. 3D**). The other sites were discarded because either no mutation was allowed by the PSSM at that position, or none of the allowed mutations was predicted to be solubilising by CamSol, or none had a negative FoldX ΔΔG. All mutations at these seven sites were then combined into multiple designs harbouring up to 6 simultaneous mutations and, following the check for potentially interacting mutations, the best designs were returned (**Fig. 3E**).

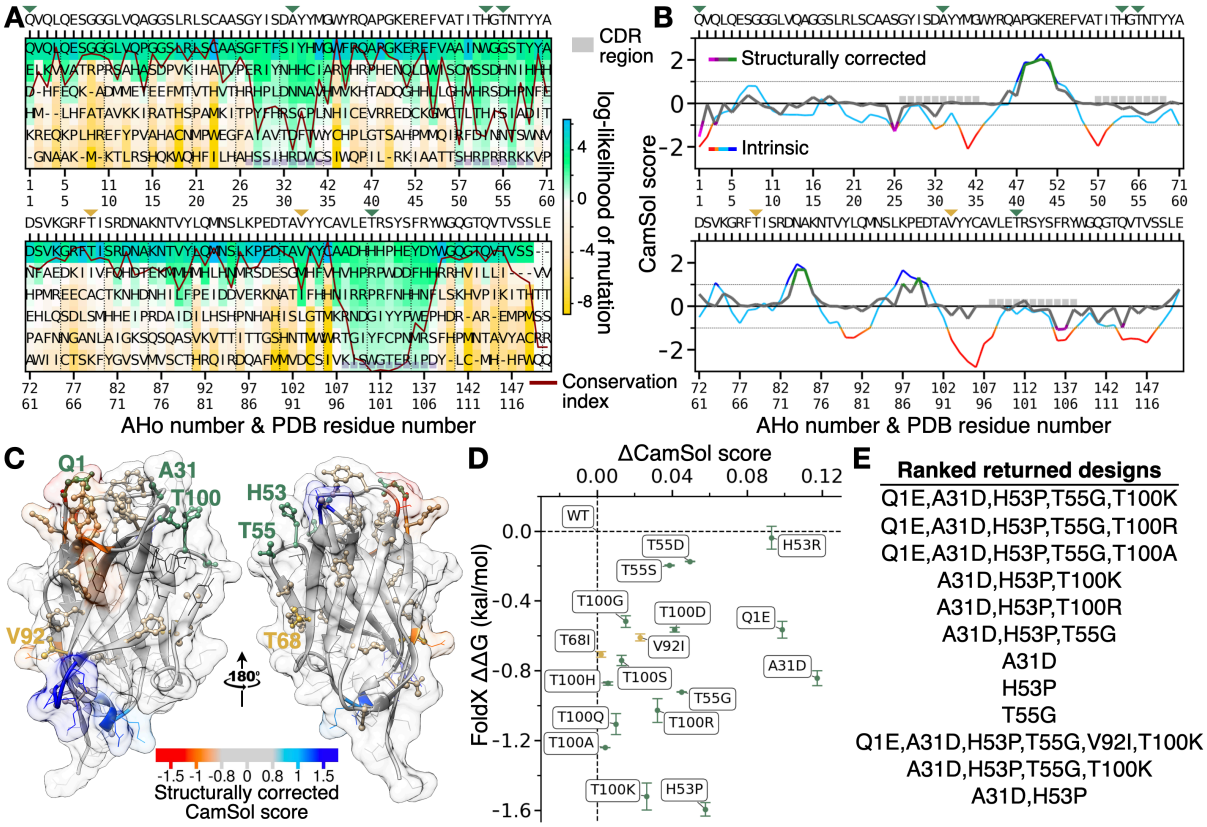

Figure 3. Main pipeline steps and returned designs for the nanobody Nb.b201. (**A**) Representation of the PSSM calculated from an alignment of 1396 single-domain antibody sequences. The observed residue log-likelihood at each position (colour-bar) is used to select candidate amino acid substitutions (allowed only log-likelihood > 0). The sequence above the panels is the wild-type Nb.b201 sequence. The red line is the conservation index of each position (high means position more conserved). X-axis labels are residue numbers from the PDB file, and AHo scheme immunoglobulin numbers, which are used to align the single domain antibody sequences; grey boxes denote CDR positions. Green or gold triangles point to mutation sites where at least one candidate mutation was identified by the single-mutational scanning step (panel D). The last two residues (LE) don't have PSSM information as these are not part of the Fv region, but rather come from a cloning restriction site present in the input structure. (**B**) CamSol profiles of Nb.b201. The CamSol intrinsic profile is coloured red to blue, where red means aggregation-prone and blue aggregation-resistant. It is common for globular proteins to have large aggregation-prone regions in their intrinsic profile that typically drive the hydrophobic collapse during folding.

The structurally corrected profile is coloured in grey/green/magenta, regions of low negative scores (magenta) are potential aggregation hotspots, regions of high score (green) are solubility promoting. (**C**) The CamSol structurally corrected profile is colour-coded (see colour-bar) on the surface of Nb.B201 (PDB ID 5vnw chain C, which is given as input to the pipeline). Sidechains shown as "ball & stick" are identified candidate mutation sites, those labelled in gold or green are sites where at least one mutation was predicted to increase both stability and solubility at the single mutational scanning step, those labelled in green are also found among the top-ranking designs returned at the end. Figure made with UCSF Chimera. (**D**) Results of the single mutational scanning step as a scatter plot of ΔCamSol score (x-axis) against FoldX ΔΔG (y-axis) for all mutations predicted to improve both properties (ΔCamSol>0 & ΔΔG<0). (**E**) Final designs returned by the pipeline, ranked as described in the main text.

Based on these predictions, we set out to produce some Nb.b201 variants and measure their stability and solubility (see Supplementary Methods). More specifically, we decided to test seven different single mutational variants, so to singly cover more of the mutations predicted to be beneficial, three double mutants and the best predicted triple and quadruple mutants (**Table 1**). Although the top-ranking designs contained five mutations (**Fig. 3E**), we excluded Q1E from experimental testing, as the nanobody were produced in HEK293 cells, which yields N-terminal pyroglutaminated glutamine (see **Table S2** and Supplementary Methods). Therefore, had we tested this mutation, we would have measured the difference between pyroglutamine and glutamic acid, instead of the predicted difference between glutamine and glutamic acid. We note, however, that the mutation Q1E is well-known in the nanobody field, and it was previously shown to be stabilising or neutral in a highly diverse set of eight nanobodies (mean sequence identity = 0.67) (61), thus suggesting that this prediction is unlikely to be wrong.

All nanobody variants were obtained at high purity (**Table S2**). The circular dichroism (CD) spectra of triple and quadruple mutants, as well as those of the single mutants not contained in the quadruple mutant, were undistinguishable from that of the WT and fully compatible with a well-folded VHH domain (**Fig. S5**). A biolayer interferometry (BLI) experiment confirmed that all variants bound their antigen (HSA) with K_D values in the high nanomolar range in agreement with previous reports for Nb.b201 (46) (**Fig. S6** and **Table S2**).

Conformational stability was measured with heat denaturation using nano differential scanning fluorimetry (nanoDSF, see Supplementary Methods). Strikingly, all variants had an apparent melting temperature greater than that of the WT (**Fig. 4** and **Table 1**). In particular, the most stable variant was the quadruple mutant, with an apparent melting temperature measured to be 13.6 ºC above that of the WT.

We then attempted to measure relative solubility with polyethylene glycol (PEG) precipitation, using a recently introduced method that only requires roughly 270 µg of purified protein material (62). However, a first experiment carried out with WT and quadruple mutant revealed that these nanobodies do not precipitate in PEG-6K at PEG concentrations up to 30%

(weight/volume). A mild drop in soluble concentration was observed only for the WT at 33% PEG (**Fig. S7**), a concentration at which the pipetting of the automated robot starts to become less accurate because of the very high viscosity of the PEG solution (62). While this result hints at a higher relative solubility of the quadruple mutant, it shows that PEG precipitation cannot be used to measure the relative solubility of all designed variants. Therefore, we resorted to using ammonium sulphate (AMS) instead of PEG as a precipitant in the assay. While the two precipitants work through different principles, good correlations between these two types of protein precipitation measurements are reported in the literature (63), and AMS has been used to assess the relative solubility of monoclonal antibodies (33). Our results reveal that all designed variants had a midpoint of AMS precipitation greater than that of the WT (**Fig. 4A** and **S8**). The assay, however, was not accurate enough to determine the rank-order of all variants with certainty, and the confidence intervals of the fitted $AMS_{50\%}$ are typically very broad (**Fig. 4A**, **S8** and **Table 1**).

To get more accurate estimates, we also carried out measurements of cross interaction chromatography (CIC). In CIC, bulk IgGs purified from human serum are chemically coupled to an NHS-activated chromatography resin (64, 65) (see Supplementary Methods). Higher retention times in this chromatography technique indicate a higher degree of non-specific interactions with this IgG mixture or increased 'stickiness'. Owing to the strong correlation observed with solubility measurement, CIC was originally proposed as an assay to identify highly soluble antibody candidates (64). A similar correlation was also reported more recently for a library of monoclonal antibodies (17). We find that, unlike AMS-precipitation measurements, CIC measurements were highly accurate. For example, the measured retention times from duplicate experiments carried out on subsequent days were virtually identical (**Fig. 4B**).

Our results reveal that 8 designs, including triple and quadruple mutants, had better CIC performance than the WT, showing up to 49 seconds decrease in retention time (**Fig. 4B**). Two variants, H53P and H53P T55G, showed similar performance to the WT, and two others, H53R and T100R, showed worse performance by 10.1 and 4.1 seconds, respectively. Notably, these two variants with higher CIC retention times than that of the WT are both mutations to arginine. Arginine residues in antibody CDRs have been associated with reduced specificity by several investigations (58, 66–68). Therefore, it is perhaps not surprising that these arginine variants perform worse in cross-interaction assays than they do in relative solubility assays such as the AMS precipitation.

Taken together, the results of this validation show that all designed variants tested experimentally had improved stability and relative solubility, and most of these designs also reduced cross interactions. We also observed statistically significant correlations between the in-silico predictions underpinning variant selection in our pipeline (FoldX and CamSol) and the corresponding experimental measurements (**Fig. S9**).

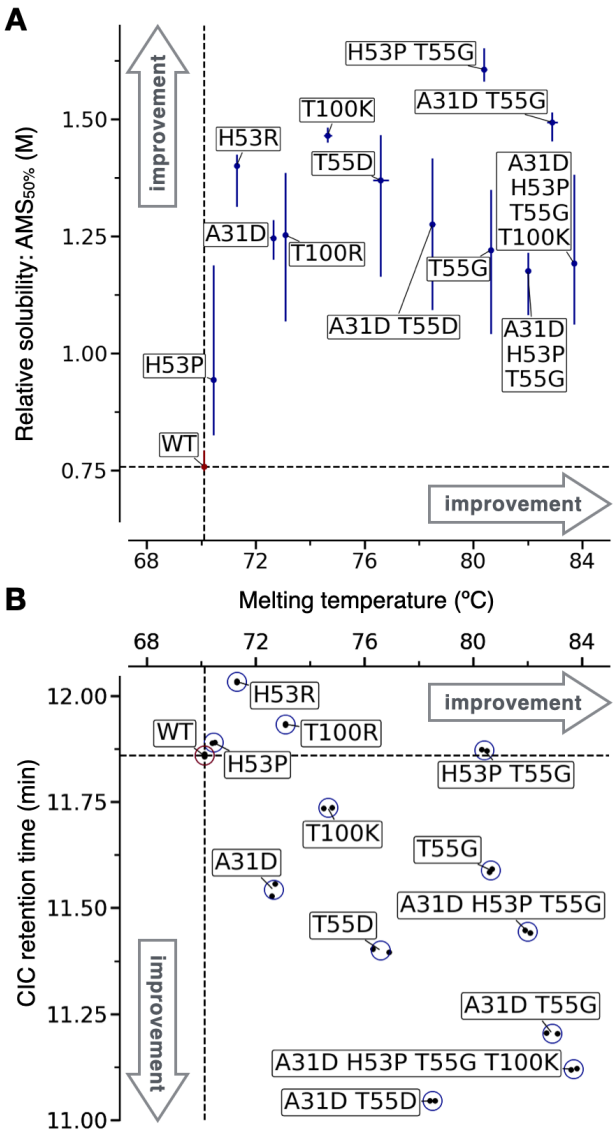

Figure 4. Experimental characterisation of designed nanobody variants. (**A**) Scatter plot of the melting temperature (T_m) against the ammonium sulphate midpoint of precipitation ($AMS_{50\%}$, used as a proxy for relative solubility). Vertical error bars are the 95% confidence interval on the fitted $AMS_{50\%}$, horizontal error bars, which are often smaller than the data point, are standard deviation over two independent melting experiments (shown individually in B). (**B**) Scatter plot of the measured T_m against the cross-interaction chromatography (CIC) retention time. Data from two independent experiments (black points) are reported. In both panels the dashed lines are drawn across the mean measurements of the WT Nb.b201 (red marker).

Nanobody variant	Apparent T_m (ºC)	$AMS_{50\%}$ (M)	CIC RT (min)
WT	70.1 ± 0.05	0.76 in (0.75,0.79)	11.859 ± 0.003
A31D	72.7 ± 0.1	1.25 in (1.20,1.28)	11.543 ± 0.014
H53P	70.5 ± 0.05	0.94 in (0.82,1.19)	11.889 ± 0.001
H53R	71.3 ± 0.05	1.40 in (1.31,1.42)	12.033 ± 0.002
T55D	76.6 ± 0.3	1.37 in (1.16,1.47)	11.399 ± 0.003
T55G	80.7 ± 0.1	1.22 in (1.04,1.35)	11.588 ± 0.003

T100K	74.7 ± 0.1	1.47 in (1.45,1.48)	11.736 ± 0.001
T100R	73.1 ± 0.05	1.25 in (1.07,1.39)	11.932 ± 0.002
A31D T55D	78.5 ± 0.1	1.27 in (1.09,1.42)	11.045 ± 0.001
A31D T55G	82.9 ± 0.2	1.49 in (1.45,1.51)	11.204 ± 0.001
H53P T55G	80.4 ± 0.1	1.61 in (1.58,1.65)	11.872 ± 0.002
A31D H53P T55G	82.0 ± 0.1	1.18 in (1.08,1.22)	11.444 ± 0.003
A31D H53P T55G T100K	83.7 ± 0.1	1.19 in (1.06,1.38)	11.120 ± 0.001

Table 1. Experimentally characterised designed variants. WT and rationally designed mutational variants with their measured conformational stability (apparent melting temperature, T_m; apparent because heat denaturation is not reversible), relative solubility in ammonium sulphate precipitation ($AMS_{50\%}$) and cross-reactivity as assessed with cross interaction chromatography retention time (CIC RT). Values are mean ± standard deviation, except for the relative solubility column, which reports fitted $AMS_{50\%}$ and lower and upper 95% confidence interval on this fitting parameter (see **Fig. S8**). T_m was measured in duplicates and reported at a 0.1 ºC resolution, when the two measurements were identical a standard deviation of 0.05 was arbitrarily assigned.

Discussion

In this work we have presented a fully automated computational pipeline and associated webserver for the design of conformationally stable and soluble protein variants.

The simultaneous improvement of stability and solubility is of high relevance, as these two properties underpin many potential liabilities that can hamper the development of biologics, such as aggregation, low expression yield, or instability upon long-term storage.

The novelty of our approach is in the simultaneous optimisation of these two biophysical properties in combination with the inclusion of phylogenetic information. We have shown that including phylogenetic information significantly reduces the false discovery rate of stability predictions by analysing a large database of experimental data (**Fig. 1**). Furthermore, all 12 designs we tested experimentally had a melting temperature greater than that of the WT, confirming a low FDR (**Fig. 4** and **S9**). Our algorithm can handle input proteins comprising multiple polypeptide chains, takes into consideration residues with missing coordinates in its solubility predictions (using the *seqres* sequence), and handles immunoglobulin variable domains using custom-built precompiled MSAs.

The pipeline is slightly different depending on whether the input is an immunoglobulin variable region or a generic protein. For a generic protein like an enzyme, a MSA is constructed from homologs found with the HHblits algorithm. In this case, the inclusion of phylogenetic information also safeguards against predicting mutations at functionally relevant sites, as these are typically conserved by evolution and as such would not be touched by our algorithm. Conversely, different pre-compiled MSAs are provided for immunoglobulin variable domains, representing the mutational space observed in human antibodies, mouse antibodies, nanobodies, and post-phase-I clinical-stage antibodies. Functionally relevant antibody residues (i.e., the paratope) are found within hypervariable regions and therefore cannot be inferred

from phylogenetic information. However, when bound structures or mutational studies are not available to determine paratope residues, paratope predictors can be used. These algorithms predict those residues most likely to be directly involved in antigen binding from the antibody sequence. Many such predictors are available and their accuracy has been steadily increasing (69–73), and one, called Parapred (69), can be accessed directly on our webserver. Whether predicted or experimentally determined, functionally relevant positions can be excluded from the design, as we have done for Nb.b201 in this work.

Our pipeline is fully automated and accessible via a user-friendly webserver. The input page enables the customisation of various parameters and settings, thus making the design highly tuneable to accommodate user-specific needs. Users can decide on target residues to exclude, sites or whole chains that should not be touched, can input extra "custom" mutation sites to test, choose between different PSSM thresholding and pre-compiled antibody MSA, and even upload user-built MSAs. The web server implementation of the method is complemented by an easy-to-use graphical interface that guides the user, including a simple guide on how to generate and download suitable alignments of homolog sequences from the HHblits webserver (56).

The required input is the structure of the protein to be optimised. While an experimentally determined structure may not always be available, great advances have been made in the de novo modelling of proteins (74, 75) and unbound antibody structures (76–79). Models generated with such software, or simpler homology models, can readily be used as input for our algorithm, albeit it remains to be seen whether they provide sufficient accuracy for atomistic stability-design calculations.

We experimentally validated the predictions of the algorithm in the case of a yeast-display-derived nanobody. We tested 12 designs consisting of seven different single mutants, three double, one triple, and one quadruple mutants. Remarkably, all variants tested had increased stability and relative solubility, assessed respectively through apparent melting temperatures and midpoints of AMS precipitation. Furthermore, most variants also had reduced retention times in CIC columns, which shows they also have diminished cross-reactivity.

We also noted that the two variants with worsened CIC retention times were both mutations to arginine, which are well-known to contribute to poor specificity (58, 66–68). This type of liabilities are easy to spot, as most software for molecular viewing can highlight patches of positive charges, and these can also be identified directly from the CDR sequences, and are tabulated in various publications (66, 68). The assessment of sequence-based liabilities will be the first area of improvement for future versions of the method, including known drivers of cross-reactivity (e.g. number of CDR arginines & tryptophans) as well as chemical liabilities (e.g. deamidation sites or post-translational-modification sites).

In conclusion, we have introduced and experimentally validated a fully automated computational method that provides a time- and cost-effective way to improve the stability and solubility of proteins and antibodies, through the rational design of mutations. We anticipate that this algorithm will find broad applicability in the optimisation of the developability potential of lead proteins and antibodies destined to applications in research, biotechnology, diagnostics, and therapeutics.

Acknowledgments
P.S. is a Royal Society University Research Fellow (URF\R1\201461). This work was partly funded by a Research Grant (RGS\R1\211126) from the Royal Society. Biomolecular production and some of the characterisation were funded by Novo Nordisk. M.O. is a PhD student funded by AstraZeneca. A.B. and M.M.M are PhD students within the Novo Nordisk R&D STAR Fellowship programme and are partially funded by Innovation Fund Denmark. We are grateful to Christian Poulsen, Nikolai Lorenzen (Novo Nordisk) and Michele Vendruscolo (University of Cambridge) for constructive feedback and helpful discussions.

Author contributions
A.R. and P.S. designed and coded the algorithm. A.B., M.O., L.S. and P.S. designed experiments. A.B., M.O., M.M.M. and L.S. carried out experiments. P.S. conceived and supervised the project. A.R., A.B. and P.S. wrote the first draft of the paper. All authors contributed new reagents/analytic tools, analysed data, and edited the paper.

Disclosure statement
A.B. and M.M.M are industrial PhD students at Novo Nordisk. P.S. is one of the developers of the original CamSol method, which is available as a free webserver, but also through commercial licenses.

References
1. H. Kaplon, A. Chenoweth, S. Crescioli, J. M. Reichert, Antibodies to watch in 2022. *mAbs* **14**, 2014296 (2022).
2. A. Beck, T. Wurch, C. Bailly, N. Corvaia, Strategies and challenges for the next generation of therapeutic antibodies. *Nat. Rev. Immunol.* **10**, 345–352 (2010).
3. P. J. Carter, Introduction to current and future protein therapeutics: a protein engineering perspective. *Exp. Cell Res.* **317**, 1261–1269 (2011).
4. A.-M. Wolf Perez, N. Lorenzen, M. Vendruscolo, P. Sormanni, Assessment of Therapeutic Antibody Developability by Combinations of In Vitro and In Silico Methods. *Methods Mol. Biol. Clifton NJ* **Therapeutic Antibodies: Methods and Protocols** (2021).
5. P. Sormanni, F. A. Aprile, M. Vendruscolo, Third generation antibody discovery methods: in silico rational design. *Chem. Soc. Rev.* **47**, 9137–9157 (2018).
6. L. A. Rabia, A. A. Desai, H. S. Jhajj, P. M. Tessier, Understanding and overcoming trade-offs between antibody affinity, specificity, stability and solubility. *Biochem. Eng. J.* **137**, 365–374 (2018).

7. I. Pallarès, S. Ventura, Understanding and predicting protein misfolding and aggregation: Insights from proteomics. *Proteomics* **16**, 2570–2581 (2016).
8. D. U. Ferreiro, E. A. Komives, P. G. Wolynes, Frustration in biomolecules. *Q. Rev. Biophys.* **47**, 285–363 (2014).
9. S. Gianni, *et al.*, Understanding the frustration arising from the competition between function, misfolding, and aggregation in a globular protein. *Proc. Natl. Acad. Sci.* **111**, 14141–14146 (2014).
10. A. Broom, Z. Jacobi, K. Trainor, E. M. Meiering, Computational tools help improve protein stability but with a solubility tradeoff. *J. Biol. Chem.* **292**, 14349–14361 (2017).
11. J. Van Durme, *et al.*, Solubis: a webserver to reduce protein aggregation through mutation. *Protein Eng. Des. Sel. PEDS*, gzw019 (2016).
12. M. Gil-Garcia, *et al.*, Combining structural aggregation propensity and stability predictions to redesign protein solubility. *Mol. Pharm.* **15**, 3846–3859 (2018).
13. P. Sormanni, F. A. Aprile, M. Vendruscolo, The CamSol method of rational design of protein mutants with enhanced solubility. *J. Mol. Biol.* **427**, 478–490 (2015).
14. J. Schymkowitz, *et al.*, The FoldX web server: an online force field. *Nucleic Acids Res.* **33**, W382-8 (2005).
15. P. Sormanni, M. Vendruscolo, Protein solubility predictions using the CamSol method in the study of protein homeostasis. *Cold Spring Harb. Perspect. Biol.* (2019) https:/doi.org/10.1101/cshperspect.a033845.
16. P. Sormanni, L. Amery, S. Ekizoglou, M. Vendruscolo, B. Popovic, Rapid and accurate in silico solubility screening of a monoclonal antibody library. *Sci. Rep.* **7**, 8200 (2017).
17. A.-M. Wolf Pérez, *et al.*, In vitro and in silico assessment of the developability of a designed monoclonal antibody library. *mAbs* **11**, 388–400 (2019).
18. C. Camilloni, *et al.*, Rational design of mutations that change the aggregation rate of a protein while maintaining its native structure and stability. *Sci. Rep.* **6**, 25559 (2016).
19. A. Achour, *et al.*, Biochemical and biophysical comparison of human and mouse beta-2 microglobulin reveals the molecular determinants of low amyloid propensity. *FEBS J.* (2019) https:/doi.org/10.1111/febs.15046.
20. M. Stenvang, *et al.*, Corneal dystrophy mutations drive pathogenesis by targeting TGFBIp stability and solubility in a latent amyloid-forming domain. *J. Mol. Biol.* **430**, 1116–1140 (2018).
21. L. Shan, *et al.*, Developability assessment of engineered monoclonal antibody variants with a complex self-association behavior using complementary analytical and in silico tools. *Mol. Pharm.* (2018) https:/doi.org/10.1021/acs.molpharmaceut.8b00867.
22. S. Sirin, J. R. Apgar, E. M. Bennett, A. E. Keating, AB-Bind: Antibody binding mutational database for computational affinity predictions: Antibody-Antigen Affinity Database and Computational Benchmarks. *Protein Sci.* **25**, 393–409 (2016).
23. Y. Myung, C. H. M. Rodrigues, D. B. Ascher, D. E. V. Pires, mCSM-AB2: guiding rational antibody design using graph-based signatures. *Bioinformatics* **12**, 553 (2019).
24. D. Bednar, *et al.*, FireProt: Energy- and evolution-based computational design of thermostable multiple-point mutants. *PLoS Comput. Biol.* **11**, e1004556 (2015).
25. B. Li, Y. T. Yang, J. A. Capra, M. B. Gerstein, Predicting changes in protein thermodynamic stability upon point mutation with deep 3D convolutional neural networks. *PLOS Comput. Biol.* **16**, e1008291 (2020).

26. A. Broom, K. Trainor, Z. Jacobi, E. M. Meiering, Computational modeling of protein stability: Quantitative analysis reveals solutions to pervasive problems. *Structure* (2020) https:/doi.org/10.1016/j.str.2020.04.003.
27. S. J. Shire, Z. Shahrokh, J. Liu, Challenges in the development of high protein concentration formulations. *J. Pharm. Sci.* **93**, 1390–1402 (2004).
28. M. Vázquez-Rey, D. A. Lang, Aggregates in monoclonal antibody manufacturing processes. *Biotechnol. Bioeng.* **108**, 1494–1508 (2011).
29. M. C. Manning, D. K. Chou, B. M. Murphy, R. W. Payne, D. S. Katayama, Stability of protein pharmaceuticals: an update. *Pharm. Res.* **27**, 544–575 (2010).
30. C. J. Roberts, Protein aggregation and its impact on product quality. *Curr. Opin. Biotechnol.* **30**, 211–217 (2014).
31. C. J. Roberts, Therapeutic protein aggregation: mechanisms, design, and control. *Trends Biotechnol.* **32**, 372–380 (2014).
32. Y. Xu, *et al.*, Addressing polyspecificity of antibodies selected from an in vitro yeast presentation system: a FACS-based, high-throughput selection and analytical tool. *Protein Eng. Des. Sel.* **26**, 663–670 (2013).
33. A.-M. W. Pérez, *et al.*, In vitro and in silico assessment of the developability of a designed monoclonal antibody library. *mAbs* **11**, 388–400 (2019).
34. V. Kayser, N. Chennamsetty, V. Voynov, B. Helk, B. L. Trout, Conformational stability and aggregation of therapeutic monoclonal antibodies studied with ANS and Thioflavin T binding. *mAbs* **3**, 408–411 (2011).
35. T. Jain, *et al.*, Biophysical properties of the clinical-stage antibody landscape. *Proc. Natl. Acad. Sci.* **114**, 944–949 (2017).
36. A. Goldenzweig, *et al.*, Automated structure- and sequence-based design of proteins for high bacterial expression and stability. *Mol. Cell* **63**, 337–346 (2016).
37. S. J. Shire, Formulation and manufacturability of biologics. *Curr. Opin. Biotechnol.* **20**, 708–714 (2009).
38. M. L. E. Lundahl, S. Fogli, P. E. Colavita, E. M. Scanlan, Aggregation of protein therapeutics enhances their immunogenicity: causes and mitigation strategies. *RSC Chem. Biol.*, 10.1039.D1CB00067E (2021).
39. K. D. Ratanji, J. P. Derrick, R. J. Dearman, I. Kimber, Immunogenicity of therapeutic proteins: influence of aggregation. *J. Immunotoxicol.* **11**, 99–109 (2014).
40. T. T. Hansel, H. Kropshofer, T. Singer, J. A. Mitchell, A. J. T. George, The safety and side effects of monoclonal antibodies. *Nat. Rev. Drug Discov.* **9**, 325–338 (2010).
41. J. F. Carpenter, *et al.*, Overlooking Subvisible Particles in Therapeutic Protein Products: Gaps That May Compromise Product Quality. *J. Pharm. Sci.* **98**, 1201–1205 (2009).
42. &Na; &Na;, Points to Consider in the Manufacture and Testing of Monoclonal Antibody Products for Human Use (1997): *J. Immunother.* **20**, 214–215 (1997).
43. J. Arora, *et al.*, Charge-mediated Fab-Fc interactions in an IgG1 antibody induce reversible self-association, cluster formation, and elevated viscosity. *mAbs* **8**, 1561–1574 (2016).
44. V. Kumar, N. Dixit, L. (Lisa) Zhou, W. Fraunhofer, Impact of short range hydrophobic interactions and long range electrostatic forces on the aggregation kinetics of a monoclonal antibody and a dual-variable domain immunoglobulin at low and high concentrations. *Int. J. Pharm.* **421**, 82–93 (2011).
45. Z. Hamrang, N. J. W. Rattray, A. Pluen, Proteins behaving badly: emerging technologies in profiling biopharmaceutical aggregation. *Trends Biotechnol.* **31**, 448–458 (2013).

46. C. McMahon, *et al.*, Yeast surface display platform for rapid discovery of conformationally selective nanobodies. *Nat. Struct. Biol.* **25**, 289–296 (2018).
47. B. Steipe, B. Schiller, A. Plückthun, S. Steinbacher, Sequence Statistics Reliably Predict Stabilizing Mutations in a Protein Domain. *J. Mol. Biol.* **240**, 188–192 (1994).
48. J. Weinstein, O. Khersonsky, S. J. Fleishman, Practically useful protein-design methods combining phylogenetic and atomistic calculations. *Curr. Opin. Struct. Biol.* **63**, 58–64 (2020).
49. B. Frenz, *et al.*, Prediction of Protein Mutational Free Energy: Benchmark and Sampling Improvements Increase Classification Accuracy. *Front. Bioeng. Biotechnol.* **8**, 558247 (2020).
50. M. M. Gromiha, *et al.*, ProTherm: Thermodynamic Database for Proteins and Mutants. 3.
51. R. Nikam, A. Kulandaisamy, K. Harini, D. Sharma, M. M. Gromiha, ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years. *Nucleic Acids Res.* **49**, D420–D424 (2021).
52. B. T. Porebski, A. M. Buckle, Consensus protein design. *Protein Eng. Des. Sel.* **29**, 245–251 (2016).
53. M. Sternke, K. W. Tripp, D. Barrick, Consensus sequence design as a general strategy to create hyperstable, biologically active proteins. *Proc. Natl. Acad. Sci.* **116**, 11275–11284 (2019).
54. G. Vecchi, *et al.*, Proteome-wide observation of the phenomenon of life on the edge of solubility. *Proc. Natl. Acad. Sci.* **14**, 201910444 (2019).
55. M. Remmert, A. Biegert, A. Hauser, J. Söding, HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173–175 (2011).
56. L. Zimmermann, *et al.*, A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *J. Mol. Biol.* **430**, 2237–2243 (2018).
57. M. I. J. Raybould, *et al.*, Thera-SAbDab: the Therapeutic Structural Antibody Database. *Nucleic Acids Res.* **48**, D383–D388 (2020).
58. M. I. J. Raybould, *et al.*, Five computational developability guidelines for therapeutic antibody profiling. *Proc. Natl. Acad. Sci.* **33**, 201810576 (2019).
59. E. E. Wilton, M. P. Opyr, S. Kailasam, R. F. Kothe, H.-J. Wieden, sdAb-DB: The Single Domain Antibody Database. *ACS Synth. Biol.* **7**, 2480–2484 (2018).
60. R. Kazlauskas, Engineering more stable proteins. *Chem. Soc. Rev.* (2018) https:/doi.org/10.1039/c8cs00014j.
61. P. Kunz, *et al.*, Exploiting sequence and stability information for directing nanobody stability engineering. *Biochim. Biophys. Acta BBA - Gen. Subj.* **1861**, 2196–2205 (2017).
62. M. Oeller, P. Sormanni, M. Vendruscolo, An open-source automated PEG precipitation assay to measure the relative solubility of proteins with low material requirement. *Sci. Rep.* **11**, 21932 (2021).
63. R. M. Kramer, V. R. Shende, N. Motl, C. N. Pace, J. M. Scholtz, Toward a molecular understanding of protein solubility: increased negative surface charge correlates with increased solubility. *Biophys. J.* **102**, 1907–1915 (2012).
64. S. A. Jacobs, S.-J. Wu, Y. Feng, D. Bethea, K. T. O'Neil, Cross-Interaction Chromatography: A Rapid Method to Identify Highly Soluble Monoclonal Antibody Candidates. *Pharm. Res.* **27**, 65–71 (2010).
65. L. I. Sakhnini, *et al.*, Improving the Developability of an Antigen Binding Fragment by Aspartate Substitutions. *Biochemistry* **58**, 2750–2759 (2019).

66. R. L. Kelly, D. Le, J. Zhao, K. D. Wittrup, Reduction of nonspecificity motifs in synthetic antibody libraries. *J. Mol. Biol.* **430**, 119–130 (2018).
67. Y. Zhang, *et al.*, Physicochemical rules for identifying monoclonal antibodies with drug-like specificity. *Mol. Pharm.*, acs.molpharmaceut.0c00257 (2020).
68. A. Azevedo Reis Teixeira, *et al.*, Drug-like antibodies with high affinity, diversity and developability directly from next-generation antibody libraries. *mAbs* **13**, 1980942 (2021).
69. E. Liberis, P. Velickovic, P. Sormanni, M. Vendruscolo, P. Lio, Parapred: Antibody paratope prediction using convolutional and recurrent neural networks. *Bioinformatics* **273**, 927 (2018).
70. A. Deac, P. Velickovic, P. Sormanni, Attentive cross-modal paratope prediction. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* **26**, 536–545 (2019).
71. S. Lu, Y. Li, X. Nan, S. Zhang, "A Sequence-Based Antibody Paratope Prediction Model Through Combing Local-Global Information and Partner Features" in *Bioinformatics Research and Applications*, Lecture Notes in Computer Science., Y. Wei, M. Li, P. Skums, Z. Cai, Eds. (Springer International Publishing, 2021), pp. 179–190.
72. A. Del Vecchio, A. Deac, P. Liò, P. Veličković, Neural message passing for joint paratope-epitope prediction. *ArXiv210600757 Cs Q-Bio* (2021) (April 8, 2022).
73. S. Daberdaku, C. Ferrari, Antibody interface prediction with 3D Zernike descriptors and SVM. *Bioinformatics* (2018) https:/doi.org/10.1093/bioinformatics/bty918.
74. M. Baek, *et al.*, Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, eabj8754 (2021).
75. J. Jumper, D. Hassabis, Protein structure predictions to atomic accuracy with AlphaFold. *Nat. Methods* **19**, 11–12 (2022).
76. J. Dunbar, *et al.*, SAbPred: a structure-based antibody prediction server. *Nucleic Acids Res.* **44**, W474-8 (2016).
77. B. Abanades, G. Georges, A. Bujotzek, C. M. Deane, ABlooper: fast accurate antibody CDR loop structure prediction with accuracy estimation. *Bioinformatics* **38**, 1877–1880 (2022).
78. G. Lapidoth, J. Parker, J. Prilusky, S. J. Fleishman, AbPredict 2: a server for accurate and unstrained structure prediction of antibody variable domains. *Bioinformatics* **35**, 1591–1593 (2019).
79. J. A. Ruffolo, J. Sulam, J. J. Gray, Antibody structure prediction using interpretable deep learning. *Patterns* **3**, 100406 (2022).