CamSol design report for input 5vnw2

General remarks. This CamSol design procedure is aimed at optimising the solubility of the target protein while retaining or improving its native state stability. This task is achieved by combining four approaches. (1) The CamSol structurally-corrected prediction is carried out to identify potential aggregation hotspots on the surface of the target protein, whose presence may elicit aggregation from the native state. Those residues contributing most to the hotspots aggregation propensity are flagged as candidate mutation sites. (2) The CamSol intrinsic profile is used to identify further mutation sites that contribute strongly to the poor solubility of the unfolded state. In fact, thermal fluctuations may lead to the transient exposure of otherwise buried aggregation-promoting regions leading to aggregation from partially or fully unfolded states. After (1) & (2) the CamSol intrinsic algorithm is used to quickly screen all possible point mutations at these identified sites to create a long-list of candidate mutations that would in principle increase solubility. (3) Sequences homologous to the target protein are automatically identified, and a multiple-sequence alignment (MSA) is performed between them. A position-specific scoring-matrix (PSSM) is calculated from this alignment to identify those residues that are most likely to be found at each position. Mutations to residues more conserved (i.e. with higher PSSM frequency) in the alignments of homologous sequences have been shown to correlate with increased native-state stability (albeit the correlation is not perfect), and more generally mutations to such residues are much better tolerated, and highly unlikely to be deleterious for the native fold. This PSSM is used to filter the long-list of candidate solubilising mutations by selecting only those mutations that increase the frequency at the position under scrutiny. (5) A structure-based atomistic prediction of the stability change upon mutation is carried out with the energy function Fold-X for each shortlisted mutations. Mutations with calculated DDG<0 are predicted to be stabilising. The correlation between measured stability and Fold-X predicted DDG is statistically significant but far from perfect. However, the atomistic calculations of Fold-X are fully independent from, and therefore highly complementary to, the evolutionary frequency difference calculated from the PSSM. Therefore mutations with increased evolutionary frequency and with a negative calculated DDG should be stabilising, or at least should not negatively impact stability. Conversely, the CamSol methods has been shown to be highly quantitative in recapitulating the effect of mutations on measured solubility in different contexts (R~0.9). Consequently mutations that (i) increase the CamSol solubility score, (ii) have a FoldX DDG smaller than 0 and (iii) increase the PSSM frequency are expected to increase protein solubility, while not affecting or even improving native fold stability.

CamSol analysis of input pdb file: 5vnw2_clean.pdb

The following sequence positions are excluded from the design (but their presence is considered in solubility calculations - these may be excluded because given as input or becase of e.g. missing PSSM information):
Chain B: all
Chain C: L119, E120

These amino acids are excluded from the list of potential substitution targets: C, M

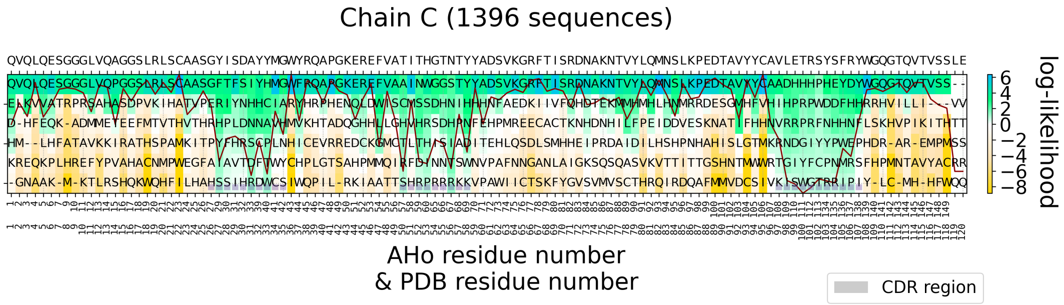
Sequences extracted from input pdb and used for analysis:
lower case residues (if any) are residues of missing coordinates but present in the SEQRES field of the input pdb file. Such residues are considered for the solubility calculation, but are never targeted for mutations and their potential impact on stability is not considered.

```
> 5vnw2_clean:B
dahKSEVAHRFKDLGEENFKALVLIAFAQYLQQCPFEDHVKLVNEVTEFAKTCVADESAENCDKSLHTLFGDKLCTVATLRETYGEMADCCAKQEPERNECFLQHKDDNPNLPRLVRPEVDVMCTAFHDNEETFLKKYLYEIARRHPYFYAPE

> 5vnw2_clean:C
QVQLQESGGGLVQAGGSLRLSCAASGYISDAYYMGWYRQAPGKEREFVATITHGTNTYYADSVKGRFTISRDNAKNTVYLQMNSLKPEDTAVYYCAVLETRSYSFRYWGQGTQVTVSSLE
```

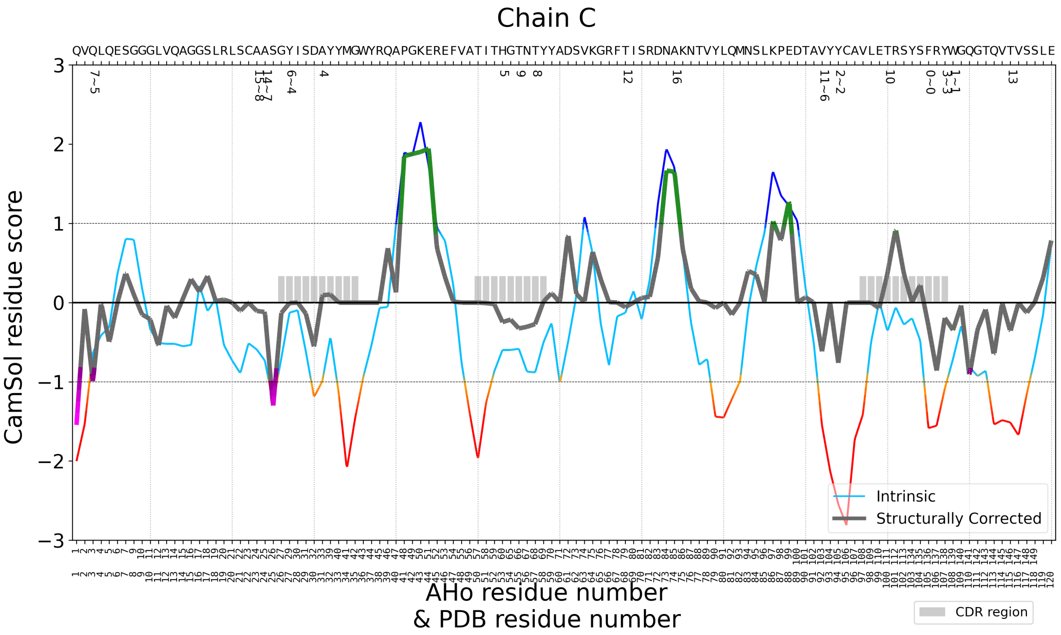
Using log-likelihood pssm. Considering only candidate mutations with positive enrichment (log-likelihood > 0), and further restricting the space of candidate substitutions at each position to those residues that are more likely than the WT one

PSSM used to pick candidate mutations



Position-specific scoring matrix (PSSM), as calculated from a multiple-sequence alignment (MSA) of similar sequences. The observed residue frequency (color-bar) is used to select candidate amino acid substitutions. The sequence above the panels is the wild-type (input) sequence as read from the alignment. The red line (if present) is the conservation index of each position (high means position highly conserved). PSSM obtained from MSA of chain C containing 1396 Fv sequences (Fv region only).

CamSol intrinsic and structurally corrected profiles

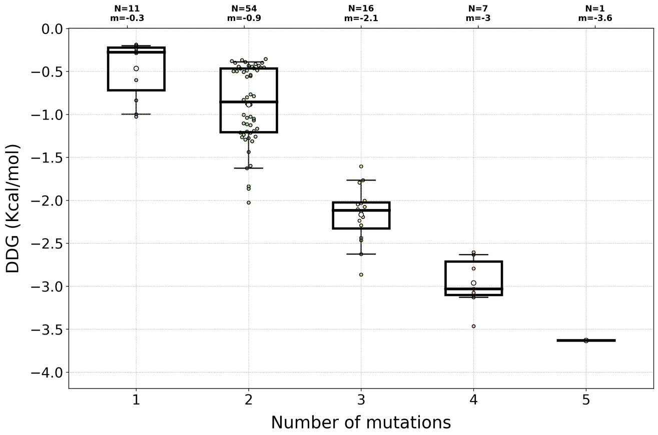


The CamSol intrinsic profile is colour-coded red to blue, where red means aggregation-prone and blue aggregation-resistant. It is common for folded proteins to have large aggregation-prone regions in their intrinsic profile that typically drive the hydrophobic collapse during folding. The structurally corrected profile is color-coded in gray/green/magenta, regions of low negative scores (magenta) are potential aggregation hotspots, regions of high score (green) are solubility promoting. Numbers below the amino acid sequence at the top denote potential mutation sites, identified according to their contribution to the solubility, as well as their accessibility to the solvent.

Solubility and Stability enhancing single-point mutants combinations

In the second part of the pipeline the single-point mutaions are combined to produce combination of mutations that are predicted to enhance both solubility and stability of the input protein. The combination process does not rely on running all the computational methods above mentioned for each possible combination since such a process would be too computationally costly. The single point mutants are combined by means of the sum of two of their main characteristics, namely DDG and Delta frequency. The CamSol score is then computed for each combination. Owing to its high computation speed, re-running the CamSol method for each combination does not slow down the algorithm. Once the three metrics are collected the mutation score for the combinations is calculated. Combinations are being produced until the limit of simultaneously occurring mutations set by the user is reached. Of all combination groups only some are selected as "best" groups. The best combination groups are those ones whose maximum mutation score marks a change in the behavior of the increment of the mutation score thorughout the different combination grous. Such a change in behavior is interpreted as a point in the combination proces from which the contribution of an additional mutant is not as favorable as it has been until that point. For all those groups label as "best" up to 3 combinations are modeled, while for the other groups just one model is returned. In generating the protein model bearing the mutation combination the DDG path of its mutations is checked. For DDG path we mean the contribution that each single mutant in the combination gives, in terms of stability, to the protein. If in applying one mutation after the other we register a non negative DDG value the algorithm tries to correct for it by swapping the single point mutation under scrutiny with another taken from the list of the other mutation combinations of the same combination group and that happens at the same point in the DDG path. If no viable alternative is found after three retries the mutation is simply skipped.

Best mutant combination groups identified for 1,4 simultaneous single-point mutants



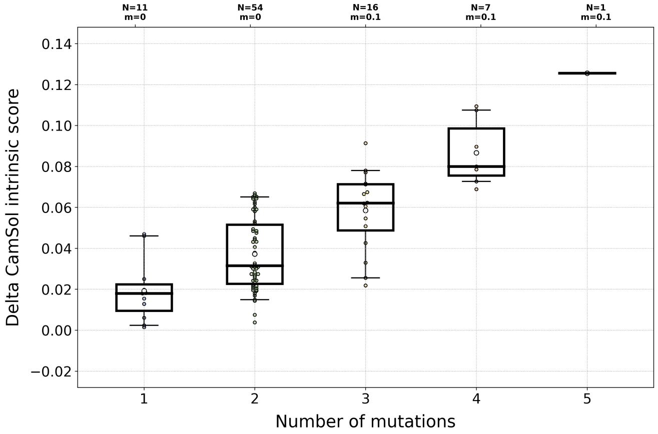


Table with identified best single-point mutant combinations

Design Name	Number of Mutations	Design Number	Mutations in Combination	Stability Bin	Solubility Bin	Mutation Score
model_2	4	2	AC31D,TC55G,EC99R,TC100R	0	0	0.544
model_1	1	1	AC31D	0	1	0.184
model_3	1	3	TC55G	1	1	0.157
model_4	1	4	EC99R	0	1	0.113
model_6	3	6	AC31D,TC55G,EC99R	1	0	0.454
model_5	2	5	AC31D,TC55G	0	1	0.341

This table contains information on the identified mutation combinations, selected on the basis of their contribution to the overall solubility and stability of the protein. The column 'design name' identifies the pdb file bearing the mutations listed in the column 'mutations in combination'. Combinations appertaining to the best combination groups are listed first and sorted according to their mutation score. The remaining combinations are liste afterwards and are sorted as well. Mutation combinations are grouped in solubility and stability bins according to their delta CamSol score and DDG value. The combinations listed are expected to increase both solubility and stability and contain single point mutants that are more conserved than their WT counterpart according to the log2(enrichment ratio) score from the input PSSM

Best Models

4 Simultaneous Mutants:

Model Name: Model_2

Mutations by chain

Chain C :AC31D,TC55G,EC99R,TC100R

Solubility Bin: 0 | Delta CamSol score: 0.09

Stability Bin: 0 | DDG: -3.464 kcal/mol

Mutation Score: 0.544

> Model_2 chain B pages/camsolcombination/data/output

KSEVAHRFKDLGEENFKALVLIAFAQYLQQCPFEDHVKLVNEVTEFAKTCVADESAENCDKSLHTLFGDKLCTVATLRETYGEMADCCAKQEPERNECFLQHKDDNPNLPRLVRPEVDVMC

TAFHDNEETFLKKYLYEIARRHPYFYAPELLFFAKRYKAAFTECCQAADKAACLLPKLDELRDEGKASSAKQRLKCASLQKFGERAFKAWAVARLSQRFPKAEFAEVSKLVTDLTKVHTE

CCHGDLLECADDRADLAKYICENQDSISSKLKECCEKPLLEKSHCIAEVENDEMPADLPSLAADFVESKDVCKNYAEAKDVFLGMFLYEYARRHPDYSVVLLLRLAKTYETTLEKCCAAA

DPHECYAKVFDEFKPLVEEPQNLIKQNCELFEQLGEYKFQNALLVRYTKKVPQVSTPTLVEVSRNLGKVGSKCCKHPEAKRMPCAEDYLSVVLNQLCVLHEKTPVSDRVTKCCTESLVNR

RPCFSALEVDETYVPKEFNAETFTFHADICTLSEKERQIKKQTALVELVKHKPKATKEQLKAVMDDFAAFVEKCCKADDKETCFAEEGKKLVAASQAALG

> Model_2 chain C pages/camsolcombination/data/output | AC31D TC55G EC99R TC100R

QVQLQESGGGLVQAGGSLRLSCAASGYISD**A**YYMGWYRQAPGKEREFVATITHG**T**NTYYADSVKGRFTISRDNAKNTVYLQMNSLKPEDTAVYYCAVL**ET**RSYSFRYWGQGTQVTVSSLE

1 Simultaneous Mutants:

Model Name: Model_1

Mutations by chain

Chain C :AC31D

Solubility Bin: 1 | Delta CamSol score: 0.047

Stability Bin: 0 | DDG: -1.027 kcal/mol

Mutation Score: 0.184

> Model_1 chain B pages/camsolcombination/data/output

KSEVAHRFKDLGEENFKALVLIAFAQYLQQCPFEDHVKLVNEVTEFAKTCVADESAENCDKSLHTLFGDKLCTVATLRETYGEMADCCAKQEPERNECFLQHKDDNPNLPRLVRPEVDVMC

TAFHDNEETFLKKYLYEIARRHPYFYAPELLFFAKRYKAAFTECCQAADKAACLLPKLDELRDEGKASSAKQRLKCASLQKFGERAFKAWAVARLSQRFPKAEFAEVSKLVTDLTKVHTE

CCHGDLLECADDRADLAKYICENQDSISSKLKECCEKPLLEKSHCIAEVENDEMPADLPSLAADFVESKDVCKNYAEAKDVFLGMFLYEYARRHPDYSVVLLLRLAKTYETTLEKCCAAA

DPHECYAKVFDEFKPLVEEPQNLIKQNCELFEQLGEYKFQNALLVRYTKKVPQVSTPTLVEVSRNLGKVGSKCCKHPEAKRMPCAEDYLSVVLNQLCVLHEKTPVSDRVTKCCTESLVNR

RPCFSALEVDETYVPKEFNAETFTFHADICTLSEKERQIKKQTALVELVKHKPKATKEQLKAVMDDFAAFVEKCCKADDKETCFAEEGKKLVAASQAALG

> Model_1 chain C pages/camsolcombination/data/output | AC31D

QVQLQESGGGLVQAGGSLRLSCAASGYISD**A**YYMGWYRQAPGKEREFVATITHGTNTYYADSVKGRFTISRDNAKNTVYLQMNSLKPEDTAVYYCAVLETRSYSFRYWGQGTQVTVSSLE

Model Name: Model_3

Mutations by chain

Chain C :TC55G

Solubility Bin: 1 | Delta CamSol score: 0.018

Stability Bin: 1 | DDG: -0.999 kcal/mol

Mutation Score: 0.157

> Model_3 chain B pages/camsolcombination/data/output

KSEVAHRFKDLGEENFKALVLIAFAQYLQQCPFEDHVKLVNEVTEFAKTCVADESAENCDKSLHTLFGDKLCTVATLRETYGEMADCCAKQEPERNECFLQHKDDNPNLPRLVRPEVDVMC

TAFHDNEETFLKKYLYEIARRHPYFYAPELLFFAKRYKAAFTECCQAADKAACLLPKLDELRDEGKASSAKQRLKCASLQKFGERAFKAWAVARLSQRFPKAEFAEVSKLVTDLTKVHTE

CCHGDLLECADDRADLAKYICENQDSISSKLKECCEKPLLEKSHCIAEVENDEMPADLPSLAADFVESKDVCKNYAEAKDVFLGMFLYEYARRHPDYSVVLLLRLAKTYETTLEKCCAAA

DPHECYAKVFDEFKPLVEEPQNLIKQNCELFEQLGEYKFQNALLVRYTKKVPQVSTPTLVEVSRNLGKVGSKCCKHPEAKRMPCAEDYLSVVLNQLCVLHEKTPVSDRVTKCCTESLVNR

RPCFSALEVDETYVPKEFNAETFTFHADICTLSEKERQIKKQTALVELVKHKPKATKEQLKAVMDDFAAFVEKCCKADDKETCFAEEGKKLVAASQAALG

> Model_3 chain C pages/camsolcombination/data/output | TC55G

QVQLQESGGGLVQAGGSLRLSCAASGYISDAYYMGWYRQAPGKEREFVATITHG**T**NTYYADSVKGRFTISRDNAKNTVYLQMNSLKPEDTAVYYCAVLETRSYSFRYWGQGTQVTVSSLE

Model Name: Model_4

Mutations by chain

Chain C :EC99R

Solubility Bin: 1 | Delta CamSol score: 0.001

Stability Bin: 0 | DDG: -0.838 kcal/mol

Mutation Score: 0.113

> Model_4 chain B pages/camsolcombination/data/output

KSEVAHRFKDLGEENFKALVLIAFAQYLQQCPFEDHVKLVNEVTEFAKTCVADESAENCDKSLHTLFGDKLCTVATLRETYGEMADCCAKQEPERNECFLQHKDDNPNLPRLVRPEVDVMC

TAFHDNEETFLKKYLYEIARRHPYFYAPELLFFAKRYKAAFTECCQAADKAACLLPKLDELRDEGKASSAKQRLKCASLQKFGERAFKAWAVARLSQRFPKAEFAEVSKLVTDLTKVHTE

CCHGDLLECADDRADLAKYICENQDSISSKLKECCEKPLLEKSHCIAEVENDEMPADLPSLAADFVESKDVCKNYAEAKDVFLGMFLYEYARRHPDYSVVLLLRLAKTYETTLEKCCAAA

DPHECYAKVFDEFKPLVEEPQNLIKQNCELFEQLGEYKFQNALLVRYTKKVPQVSTPTLVEVSRNLGKVGSKCCKHPEAKRMPCAEDYLSVVLNQLCVLHEKTPVSDRVTKCCTESLVNR

RPCFSALEVDETYVPKEFNAETFTFHADICTLSEKERQIKKQTALVELVKHKPKATKEQLKAVMDDFAAFVEKCCKADDKETCFAEEGKKLVAASQAALG

> Model_4 chain C pages/camsolcombination/data/output | EC99R

QVQLQESGGGLVQAGGSLRLSCAASGYISDAYYMGWYRQAPGKEREFVATITHGTNTYYADSVKGRFTISRDNAKNTVYLQMNSLKPEDTAVYYCAVL**E**TRSYSFRYWGQGTQVTVSSLE

Table with identified candidate mutation sites (17 sites)

Mutation	Mutation	pssm-	site	solvent	solubilization	Problematic	region	residue	residue	wt	identified

site (seq index)	site (pdb number)	allowed mutations	conservation	exposure	potential	region+site score	size	stru. corr. score	intrinsic score	residue frequency score	from
F104.C	FC105	Y H W I P N	0.435	0.306	0.06	4.27	4	-0.292	-1.585	2.563	Solub. Seq.
W107.C	WC108	R	0.89	0.482	0.059	3.394	4	-0.348	-0.712	5.936	Solub. Seq.
Y93.C	YC94	F H	0.83	0.275	0.054	7.705	6	-0.764	-2.546	4.524	Solub. Seq.
Y106.C	YC107	H F P S	0.652	0.188	0.047	3.742	4	-0.193	-1.085	4.202	Solub. Seq.
Y26.C	YC27	F R H I S N	0.333	1.0	0.034	1.883	3	-0.007	-0.128	1.423	Solub. Seq.
Q2.C	QC3	K	0.683	1.0	0.031	3.295	2	-1.005	-0.631	4.154	Solub. Seq.
V91.C	VC92	I	0.665	0.396	0.009	6.723	6	-0.623	-1.546	3.897	Solub. Seq.
A23.C	AC24	V	0.768	0.143	0.0	2.417	3	-0.12	-0.734	3.988	Solub. Seq.
A22.C	AC23	T V	0.698	0.616	0.0	2.274	3	-0.106	-0.592	3.855	Solub. Seq.
A30.C	AC31	I H N S D R T F	0.225	0.576	0.047	2.658	2	0.092	-0.996	-1.37	Solub. Seq.
H52.C	HC53	W S R I P T N	0.317	0.873	0.037	3.511	5	-0.246	-0.598	1.309	Solub. Seq.
T56.C	TC57	I P	0.699	0.434	0.02	2.253	2	-0.271	-0.881	4.042	Solub. Seq.
T54.C	TC55	G H D S N	0.472	1.0	0.02	1.957	2	-0.327	-0.584	-0.117	Solub. Seq. & Conservation
T99.C	TC100	H R P G Y W F L D N A I S K Q	0.087	0.445	0.017	5.467	6	0.368	-0.365	0.806	Solub. Seq.
T67.C	TC68	I	0.879	0.681	0.001	1.118	1	-0.058	-0.128	4.353	Solub. Seq.
T114.C	TC115		0.945	0.425	0.0	5.239	8	-0.367	-1.514	4.449	Solub. Seq.
A73.C	AC74		0.789	1.0	0.0	1.107	6	1.65	1.703	4.026	Solub. Seq.

This table contains information on the identified mutation sites, selected on the basis of their contribution to the overall solubility and their solvent exposure, or on the basis of their conservation in the PSSM. The column 'Problematic region+site score' is an indicator (the higher the more aggregation-promoting) of how much a site is expected to contribute to the aggregation propensity accounting for both its own CamSol score and that of the broader sequence region that contains it. Sites identified from the structurally corrected profile are listed first, then those identified from the intrinsic solubility profile, which are also solvent-exposed but don't have a particularly negative structurally-corrected score, and finally those identified from the Conservation (see column 'identified from'). Mutations sites identified from solubility are ranked according to the column 'solubilization potential', which indicates how much the solubility can be improved by mutating that site according to those mutations allowed by the PSSM. As some problematic sites may be highly conserved, the 'solubilization potential' does not necessarily correlate with 'Problematic region+site score'. The 'sovlent exposure' is a score that ranges from 0 (non-exposed) to 1 (as exposed as in the context of a Gly-AminoAcidUnderScrutiny-Gly 3-peptide in an extended conformation), note that a solvent exposure of 0.5 would already offer the largest area to a potential aggregation partner. Sites identifide from 'Conservation' are those where the WT residue has negative log2(enrichment ratio) score from the input PSSM, and the site is relatively well conserved (conservation index > 0.25). The conservation index ranges in 0 to 1 and indicates how conserved a position in the alignment is (red line in the plot of the PSSM).

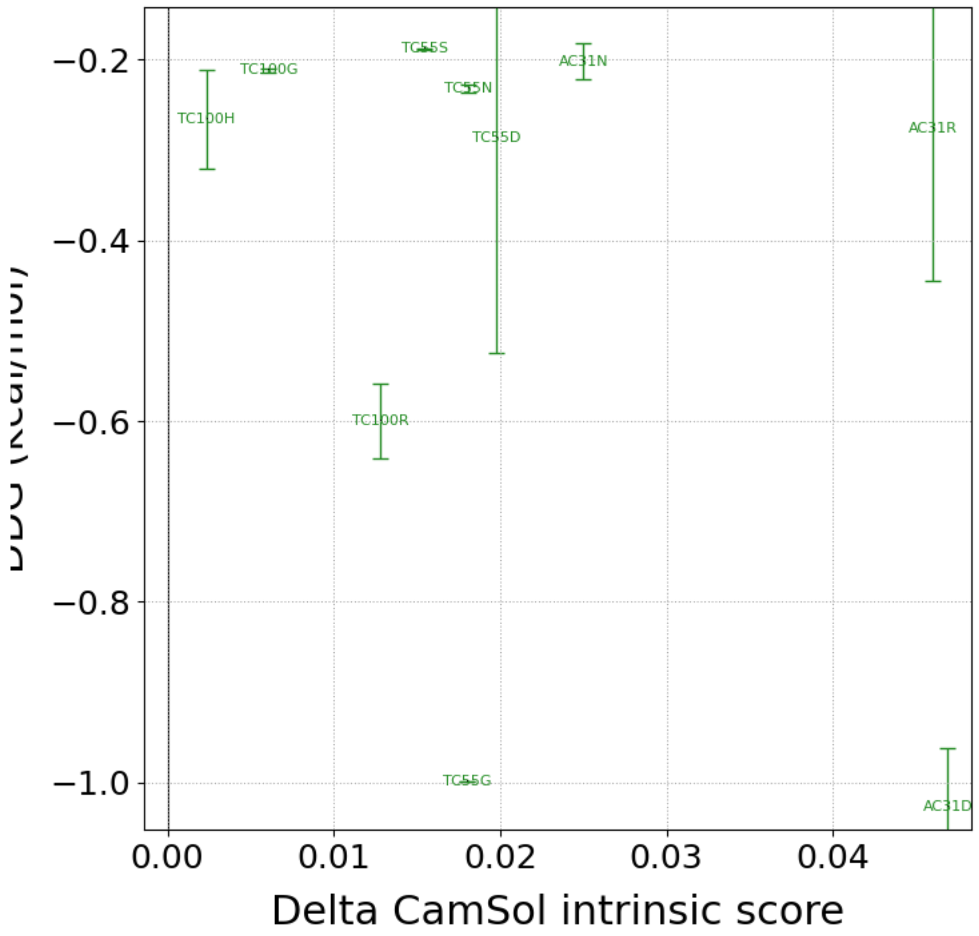
Table with the result of a single-mutation scanning at all suitable sites

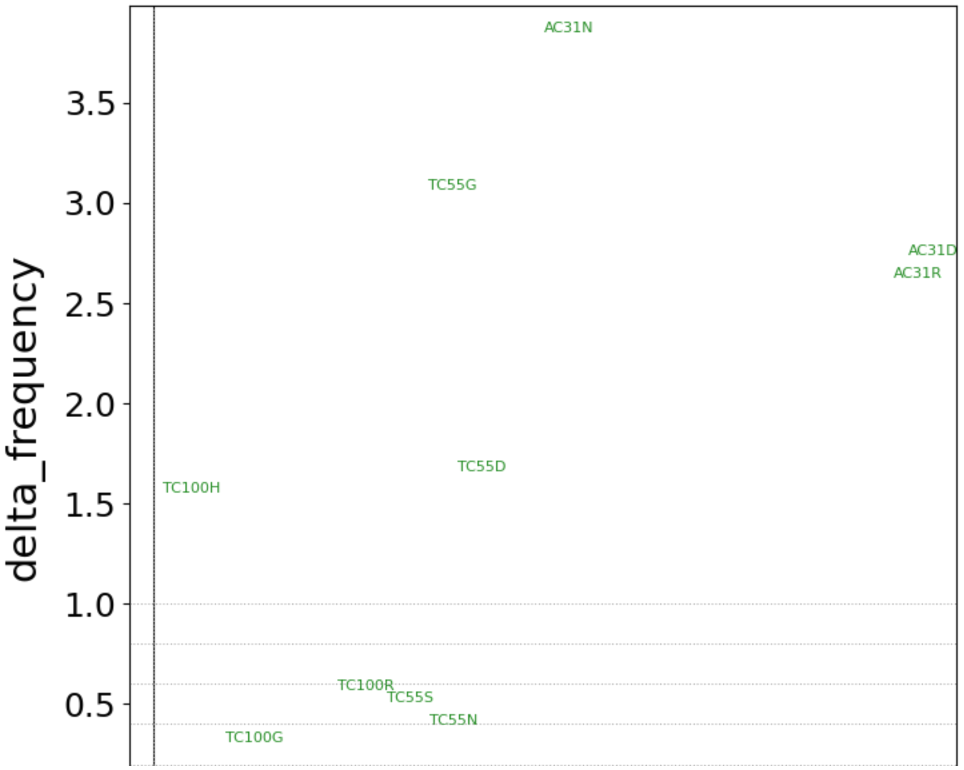
mut_id_seqIndex	mut_id_pdb	Mutation Score	Delta CamSol intrinsic score	delta_frequency	DDG (kcal/mol)	Mutation type	CamSol intrinsic score	Mutation frequency	PDB
WT	WT	0	0.0	0.0		n/a	0.468	n/a	
AC30N	AC31N	0.277	0.025	3.869	-0.202	Solub. Seq.	0.493	2.499	Optimized_5vnw2_clean_
AC30D	AC31D	0.315	0.047	2.757	-1.027	Solub. Seq.	0.515	1.387	Optimized_5vnw2_clean_
AC30R	AC31R	0.232	0.046	2.644	-0.276	Solub. Seq.	0.514	1.274	Optimized_5vnw2_clean_
TC54G	TC55G	0.303	0.018	3.085	-0.999	Solub.	0.486	2.968	Optimized_5vnw2_clean_

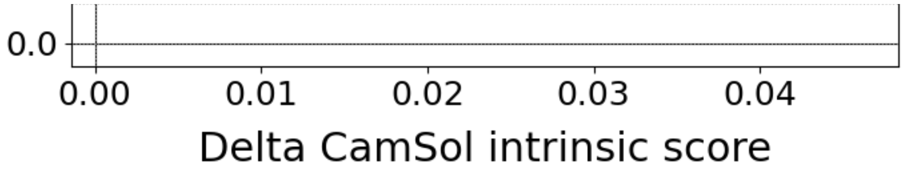
						Seq.			
TC54D	TC55D	0.149	0.02	1.681	-0.287	Solub. Seq.	0.488	1.564	Optimized_5vnw2_clean_
TC54S	TC55S	0.066	0.015	0.528	-0.188	Solub. Seq.	0.484	0.412	Optimized_5vnw2_clean_
TC54N	TC55N	0.066	0.018	0.418	-0.232	Solub. Seq.	0.486	0.302	Optimized_5vnw2_clean_
EC98R	EC99R	0.216	0.001	2.187	-0.838	Exposed Solub.	0.47	2.251	Optimized_5vnw2_clean_
TC99H	TC100H	0.124	0.002	1.577	-0.266	Solub. Seq.	0.471	2.383	Optimized_5vnw2_clean_
TC99R	TC100R	0.108	0.013	0.587	-0.6	Solub. Seq.	0.481	1.394	Optimized_5vnw2_clean_
TC99G	TC100G	0.047	0.006	0.329	-0.211	Solub. Seq.	0.474	1.135	Optimized_5vnw2_clean_
SC101R	SC102R	0.038	0.012	0.162	-0.167	Exposed Solub.	0.48	0.607	Optimized_5vnw2_clean_

This table contains information on all possible single mutations at sites identified on the basis of their contribution to the overall solubility and their solvent exposure (see previous table), as well as to their conservation in the MSA. The column 'Mutation type' describes how a site has been identified: Solub. Stru. denotes that it was a site contributing to poor local solubility in the structurally corrected profile; Solub. seq. is the same but for the sequence-based intrinsic profile; Conservation indicates that the frequency of the wild-type amino acid is lower than that of other amino acids at this site in the PSSM (these are like in the previous table); Exposed Solub. are additional solvent-exposed sites where the WT residue is not strongly conserved that may be mutated to further increase solubility (albeit unlike Solub. Seq. and Solub. Stru. these are typically not close to or within candidate aggregation hotspots). The column 'mut_id_seqIndex' contains candidate mutations numbered according to the index of the mutation site along the input sequence (the first amino acid has index 0), while in that 'mut_id_pdb' mutations are numbered according to the residue number in the input pdb file. This table has been filtered to contain only point-mutations predicted to increase both frequency and solubility, and - if FoldX calculations are carried out - also to increase predicted stability (DDG<0). While these is the safest approach, mutations that instead decrease the frequency from the PSSM may still be solubilising and stabilising, and are therefore may be considered in the general design pipeline.

Result of single-mutation scanning at all suitable sites

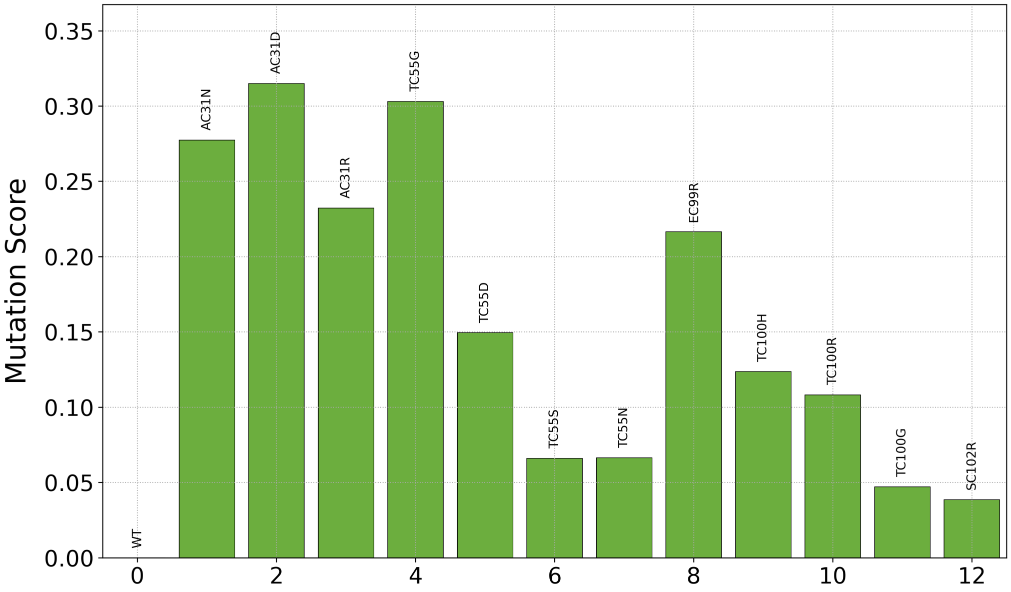

DDG (kcal/mol)
−0.2
−0.4
−0.6
−0.8
−1.0
0.00
0.01
0.02
0.03
0.04
Delta CamSol intrinsic score
TC55S
TC100G
AC31N
TC55N
TC100H
TC55D
AC31R
TC100R
TC55G
AC31D


delta_frequency
3.5
3.0
2.5
2.0
1.5
1.0
0.5
AC31N
TC55G
AC31D
AC31R
TC55D
TC100H
TC100R
TC55S
TC55N
TC100G



Plot of the data in the previous table, points are candidate mutations numbered according to the residue number within the input pdb file. The best mutations are those with large Delta CamSol score and large negative predicted DDG (if FoldX was run) and/or large positive Delta Frequency in MSA. Points in black correspond to mutation sites selected according to their conservation (see previous table), in green according to the sequence-based intrinsic solubility prediction, and in blue according to the structurally corrected one. Mutations at sites selected according to the solubility profiles (blue and green) are expected to have more impact on solubility than mutations at sites selected from conservation (black).

Mutation score of results of single-mutation scanning at all suitable sites



The plotted score is a rather arbitrary combination of delta solubility score, delta frequency, and predicted DDG (if FoldX was run). The highest this score the best the mutation. This score provides a nice visual ranking of mutations, but in practice one should refer to actual delta solubility score, delta frequency, and predicted DDG values to choose suitable mutations.