Computational and Structural Biotechnology Journal 20 (2022) 2415–2433

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

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

Protposer: The web server that readily proposes protein stabilizing mutations with high PPV

Helena García-Cebollada [a,b,c], Alfonso López [a,b,c], Javier Sancho [a,b,c,*]

[a] Department of Biochemistry, Molecular and Cell Biology, Faculty of Science, University of Zaragoza, 50009 Zaragoza, Spain
[b] Biocomputation and Complex Systems Physics Institute (BIFI), Unit and GBs-CSIC, University of Zaragoza, 50018 Zaragoza, Spain
[c] Aragon Health Research Institute (IIS Aragón), 50009 Zaragoza, Spain

ARTICLE INFO

Article history:
Received 16 March 2022
Received in revised form 5 May 2022
Accepted 5 May 2022
Available online 10 May 2022

Keywords:
Protein stabilization
Protein thermostabilization
Stability predictor
Protein engineering
Protein biotechnology
Protein expression

ABSTRACT

Protein stability is a requisite for most biotechnological and medical applications of proteins. As natural proteins tend to suffer from a low conformational stability *ex vivo*, great efforts have been devoted toward increasing their stability through rational design and engineering of appropriate mutations. Unfortunately, even the best currently used predictors fail to compute the stability of protein variants with sufficient accuracy and their usefulness as tools to guide the rational stabilisation of proteins is limited. We present here ***Protposer***, a protein stabilising tool based on a different approach. Instead of quantifying changes in stability, ***Protposer*** uses structure- and sequence-based screening modules to nominate candidate mutations for subsequent evaluation by a logistic regression model, carefully trained to avoid overfitting. Thus, ***Protposer*** analyses PDB files in search for stabilization opportunities and provides a ranked list of promising mutations with their estimated success rates (eSR), their probabilities of being stabilising by at least 0.5 kcal/mol. The agreement between eSRs and actual positive predictive values (PPV) on external datasets of mutations is excellent. When ***Protposer*** is used with its Optimal kappa selection threshold, its PPV is above 0.7. Even with less stringent thresholds, ***Protposer*** largely outperforms FoldX, Rosetta and PoPMusiC. Indicating the PDB file of the protein suffices to obtain a ranked list of mutations, their eSRs and hints on the likely source of the stabilization expected. ***Protposer*** is a distinct, straightforward and highly successful tool to design protein stabilising mutations, and it is freely available for academic use at http://webapps.bifi.es/the-protposer.

1. Introduction

The beginning of protein design was marked by the application of site-directed mutagenesis to the modification of protein active sites, which paved the way for many subsequent biotechnological and biomedical advances [1–6]. One of the main goals in protein design is to gain a deep understanding of protein stability and how it can be modulated by single amino acid variations, as this is the key to efficient protein production. If we manage to design protein stabilising mutations in an easy, quick and accurate manner (i.e. with a high positive predictive value (PPV)), the biotechnological and biomedical use of proteins as analytic, synthetic, and therapeutic tools will be tremendously boosted. Protein-based biosensors [7–11] are widely used in medicine by practitioners and patients. Their usefulness in developing countries can greatly benefit from the availability of more stable proteins, permitting easier storage conditions and longer shelf lives. The generalization of biological catalysis in food, fuel, pharmaceutical or more conventional chemical production industries is slowed down by the low stability of proteins in organic solvents, high temperatures or extreme pH conditions sometimes required. Availability of more stable enzymes will translate into higher production rates and reduced replacement costs [12–16]. Moreover, protein-based biological products [17–21], such as those used in cancer immunotherapy [22–23], provide new treatments that are revolutionizing and personalizing medicine. As with other protein products, their production, transport, long-term storage and ease of administration can substantially benefit from protein stabilization.

The biopharmaceutical market (vaccines excluded) is estimated at U.S.$208 billion by the end of 2020. [24]. While the economic and social impact of being able to design protein-stabilizing mutations is clear and many different approaches have been suggested

* Corresponding author at: Department of Biochemistry, Molecular and Cell Biology, Faculty of Science, University of Zaragoza, 50009 Zaragoza, Spain.
E-mail address: jsancho@unizar.es (J. Sancho).

https://doi.org/10.1016/j.csbj.2022.05.008
2001-0370/

for this purpose, the present performance of software implementing those approaches provides considerable room for improvement [25]. Modarres et al. reviewed 22 standalone calculation tools described as capable of predicting stabilizing mutations [16]. Those predictors were based on analysis of the protein sequence and/or structure and typically used machine learning or potential energy functions, although a few ones used fuzzy queries (FQ-STAB [26]), graphs (mCSM [27]), or Normal Mode Analysis (ENCoM [28]). It seems that even structure-based predictors, which are usually more accurate than sequence-based ones [16], present serious limitations for calculating the precise change in stability a given mutation will bring about (change in unfolding free energy: ΔΔG). Even for the simpler task of classifying mutations as either stabilizing or destabilizing, stability predictors suffer from a very limited accuracy. A comparative study by Khan and Vihinen [29] indicated that the accuracy of the best three predictors (I-Mutant [30], D-mutant [31] and FoldX [32]) among those compared was only around 60%. In another systematic evaluation of mutation stability predictors, Potapov et al. [33] showed that combining different methods could not significantly enhance the accuracy.

The ΔΔG calculated by structure-based predictors should be very useful for deciding whether to implement a particular mutation in order to stabilise a biotechnologically relevant protein. Unfortunately, even the best predictors calculate ΔΔG values with average unsigned errors over 1 kcal/mol [33] (i.e., in average, the predicted ΔΔG value is over 1 kcal/mol away from the experimental value). Furthermore, the correlation coefficients between experimental and predicted ΔΔG values are below 0.6 [33], so the experimental and predicted data do not correlate well and, therefore, the average unsigned errors cannot be due to the predicted values being a multiple of the experimental ones. Additionally, the self-consistency biases reported are over 0.7 kcal/mol [34], showing that the thermodynamic assumption that $\Delta\Delta G_{A\rightarrow B} = -\Delta\Delta G_{B\rightarrow A}$ is not being fulfilled by predictors. A recent review by Pucci et al. [35] covering from very complex deep learning algorithms (e.g., ThermoNet [36]) to very simple three-parameter predictions (e.g., SimBa [37]) confirms that the average unsigned errors of the predictors have been stagnated at around that value of 1 kcal/mol for over 15 years. The fact that simple models may perform as well as very complex ones has been discussed by Semenova et al. [38], whose recommendations for improving predictors stagnated around a certain predictive value align with finding simple rational models that, at least, may provide a better understanding of the problem.

Additional limitations of structure-based predictors that calculate ΔΔG values are that they usually evaluate mutations previously defined by the user and do not provide much insight on the physical cause of the predicted stabilization. For researchers unfamiliar with structural computational biophysics, assuming the task of conceiving potentially stabilizing mutations so that a specialized software analyses them and eventually confirms their usefulness could be troublesome. On the other hand, trying to circumvent the problem by asking the software to compute all possible mutations (19 times the length of the sequence) will set them to struggle with a large amount of data. Some programs, such as PoPMuSiC [39–41], tackle this problem by making a map of mutation hot spots, based on the ΔΔG values calculated for all possible mutations of each residue. However, interpretation of a hot spot map may not be straightforward, as not all mutations in a hot spot will be equally stabilizing.

Fortunately, thermodynamic analysis of thousands of point mutations generated by protein engineers over decades has provided a wealth of data allowing biotechnologists to identify protein-stabilizing strategies. While they are far from infallible, the chances of stabilizing a protein by taking them into account are much greater than by making random mutations. Well-known examples of successful stabilizing strategies for which a biophysical explanation can be offered include sequence optimization of α-helical segments [42–46], filling of internal protein cavities [47–48], optimization of electrostatic interactions between charged residues [49], replacement of hyperexposed hydrophobic residues [50] or underexposed polar ones [51], or replacement of hydrogen-bonded acidic residues by neutral isosters [52]. Other successful stabilizing strategies that find justification on the grounds of evolutionary reasoning rely on mutating residues to return to a previously calculated consensus sequence of the protein family studied [53] or to the calculated ancestral sequence of that family [54]. Once stabilizing mutations have been introduced and tested in a protein, the more stabilizing ones can be combined, sometimes in a close to additive manner, to obtain highly stabilized variants of the initial unstable wild type protein [55–56]. Drawing from this accumulated knowledge on protein stabilization strategies, we develop here Protposer, a straightforward web server that proposes protein stabilizing point mutations for any protein of known structure. To do that, Protposer analyses the PDB file indicated by the user in search for stabilization opportunities. Then it internally assesses mutations that could increase protein stability by amending the structural weaknesses found, and finally it returns a simple ranked list of potentially stabilizing mutations along with their estimated success rate (i.e., the probability a proposed mutation has of increasing the stability of the protein by>0.5 kcal/mol).

2. Materials and methods

2.1. Algorithm

Protposer operates through four main steps, displayed in Fig. 1: 1) PDB file retrieval and preparation; 2) internal nomination of candidate mutations (including filtering); 3) evaluation of candidates; and 4) scoring. In the initial step, if the user provides a PDB ID code, the corresponding PDB file is downloaded from the wwPDB database (www.wwpdb.org [57–59]) and a new PDB file containing only the indicated target chain is prepared. If the user provides its own PDB file, a target chain must be also indicated and a PDB file for the specified chain is similarly prepared. In step 2, the PDB file is scanned in search for specified features of the protein (Fig. 1) that allow the program to nominate potential mutations for further evaluation, if some criteria are met (see the subsection "Modules" below, for details). The mutations nominated in this step are filtered to discard those likely to produce steric clashes. Additionally, mutations proposed by the "cavities" module are filtered by SASA in order to prevent artifacts caused by the opening of internal cavities to solvent upon mutation. In the candidates evaluation step (step 3), each candidate mutation that has passed the filters is evaluated to obtain numerical scores for each of several properties (Fig. 1). The scoring metrics for each property have been defined considering recommendations from Pucci et al. [60] for getting unbiased scoring models, as well as recommendations from Fang [61] for selecting informative features in predictive models. Details on each specific metric used are given in the "Modules" subsection below. In the scoring step (step 4), the scores for all the metrics calculated for each mutation are standardized and fed to a supervised logistic regression model, previously trained, in order to calculate each mutation probability of being stabilizing, which is defined as the probability of having $\Delta\Delta G_{unf}^{wt\rightarrow mut} > 0.5$ kcal/mol. As these probabilities underestimate the actual positive predictive value (PPV) of the mutations, they are transformed, as explained in the results section, into estimated success rates (eSR), which are then used for ordering the mutations

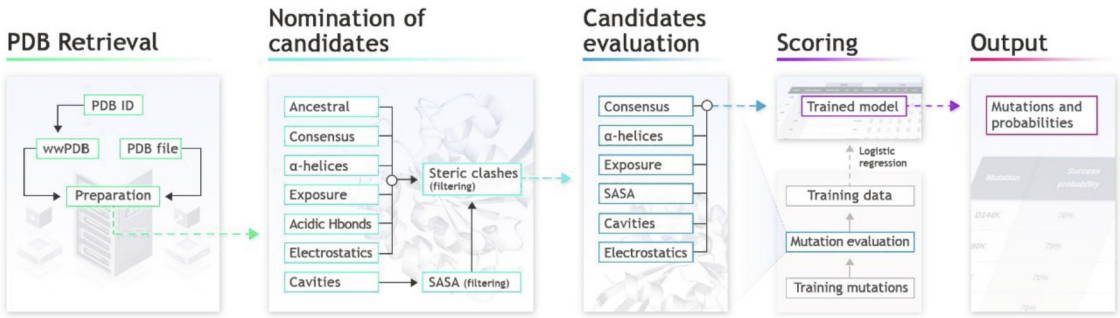

Fig. 1. General workflow of *Protposer*. Each step of the general workflow is represented with one colour in the underline of the step title, and the boxes for each module or process inside that step. Boxes in "Nomination of candidates" and "Candidates evaluation" steps correspond to the different nominating and evaluating modules. The user supplies a PDB file or its ID code in the worldwide PDB database and indicates the chain to be analysed. The structure is then prepared for the nomination step (first column in "Nomination of candidates") and then, some mutations are filtered out to avoid steric clashes and other artifacts. The remaining candidate mutations have some of their features evaluated, and the values obtained are passed to a trained logistic regression model to get the probability of each mutation of being stabilizing, which is then converted to an estimated success rate indicated in the results report emailed back to the user. The model is trained using a set of mutations obtained after filtering the database ProTherm [87–88].

from most to least potentially stabilizing, as reported in the results page returned to the user.

Protposer has been implemented as a web server using Joomla, a content management system, with PHP as a connector between the input form and the main Python script performing the calculations behind Protposer, either directly or calling additional software. The web page and the email sent to the user is also returned by PHP. All the programs and Python modules used by Protposer and their respective versions are shown in Supplementary Table S1.

2.2. PDB preparation

To ensure that the PDB files are correctly analysed by all modules in Protposer, a routine check is performed at the beginning of the workflow. If any correction is needed, the PDB file is automatically edited to solve the issue detected. As part of this process, all ligands, heteroatoms and residues with an incomplete backbone (usually near the ends of the structure) are removed by editing the PDB file with Python, the modified residues are substituted by their precursor using SCWRL [62], and the numbering of residues is checked in order to detect if different residues have been numbered as the same position. If more than one residue has been given the same number, an error message will be returned to the user indicating the problematic residues. A PDB file including only a correct version of the target chain is generated for further steps.

2.3. Modules

Protposer has a modular structure. Nine different modules are involved in the nomination of candidates and evaluation steps. Seven of them ("ancestral", "consensus", "alpha helices", "exposure", "acidic H-bonds", "cavities" and "electrostatics") search in the protein structure for the presence of defined features and propose potentially stabilizing mutations accordingly. Two additional modules ("SASA" and "steric clashes") are used to quickly filter out nominated mutations that might easily turn out to be destabilizing due to their specific location in the structure.

The "ancestral" module searches for closely related proteins using blastp [63] against the non-redundant protein database of the NCBI server, including all non-redundant GenBank CDS translations [64], PDB [57–59], SwissProt [65–66], PIR [66–67] and PRF, excluding environmental samples from WGS projects (PRF/SEQDB database, Protein Research Foundation, Osaka, Japan). Then, the retrieved sequences are grouped in clusters of 90 % sequence similarity using CD-HIT [68]. A phylogenetic tree is constructed using PhyML with default parameters [69] and the common ancestor sequence is calculated using PAML [70] with the JTT amino acid substitution model [71]. Mutations are proposed for those residues differing in the original and ancestral sequences, which consist of replacement of the wild type residue for the ancestral one.

The "consensus" module generates, from the sequences with a higher similarity score, as previously found and calculated using blastp [63], a multiple alignment using MUSCLE [72]. Then, BioPython alignment tools [73] are used to calculate the consensus sequence. Mutations are then proposed for those residues differing in the original and consensus sequences consisting of replacement of the wild type residue by that in the consensus sequence. The score given by this module to a given mutation (either proposed by this module or otherwise) is calculated using the number of occurrences at that position in the multiple alignment of the wild type residue (n_{WT}) and of the residue that is proposed to replace it (n_{mut}), and the total number of sequences (N), according to:

$$Score = (n_{mut} - n_{WT})/N$$

The "alpha helices" module identifies the helices and their ends using DSSP [74] for the prediction of secondary structure and criteria defined by Leader et al. [75] Hydrogen bonds are calculated using HBPlus [76]. Mutations are proposed in order to improve the stability of the helix. As suggested in the literature [42–46], three different helical positions are considered: N-cap, C-cap, and inner helix. Specifically, the theoretically most stabilizing residues for the N- and C-caps and for the inner residues of the helices, according to the scale defined by Muñoz et al. [43–45], are proposed to replace any non-buried helical residue whose side chain does not establish hydrogen bonds with protein atoms in the wild type structure. Values of ΔG_{hel} calculated by Muñoz et al. [43–45] as the difference in free energy between the random coil and helix states are used to determine the best options of residues for each position: Ala for inner residues, Asp for N-cap or Gly for C-cap. The estimation of $\Delta\Delta G_{unf}^{wt\rightarrow mut}$ according to their work is used as score.

The "exposure" module calculates the relative exposure of a residue (100 × folded exposure/unfolded exposure) using data from DSSP and the average exposure for each residue in the unfolded state, as calculated by Estrada et al. [77–78] Buried polar residues with relative exposure under 15 % are proposed for mutations, following recommendations by Ayuso-Tejedor et al. [51] Mutations for overexposed apolar residues with relative exposure over 100 % are proposed to be replaced by more polar residues of similar size and structure (e.g. Val to Thr, Phe to Tyr or Leu to Gln). For scoring, an empirical equation from Ayuso-Tejedor et al.

[51] has been adapted to full residue Accessible Solvent Areas (ASA, the only exposure data returned by DSSP) by using the following assumptions: a) all backbone is considered to be apolar, b) side chains are considered as either completely polar or completely apolar, c) the size of the backbone is approximately that of an alanine (per residue), and d) backbone and side chains are equally buried, being the burial percentage the complementary of the relative exposure (1 - RE). The final score is calculated with the following equation:

$$\begin{cases} [0.0276(ASA_{Ala} - ASA_{WT}) + 0.0072(ASA_{Mut} - ASA_{Ala})](1 - RE) if\ polar \rightarrow apolar \\ [0.0276(ASA_{Mut} - ASA_{Ala}) + 0.0072(ASA_{Ala} - ASA_{WT})](1 - RE) if\ apolar \rightarrow polar \end{cases}$$

The "acidic hydrogen bonds" module analyses the relative exposures and hydrogen bonds, as calculated before, to identify acidic residues (Asp or Glu) with a relative exposure over 85% and, at least, one hydrogen bond formed by their side chain, and proposes their mutation to isosteric neutral residues (Asn or Gln, respectively), as suggested by Irún et al. [52].

The "cavities" module finds residues at the surface of internal cavities of the protein identified using Voronoia [79] and proposes mutations to bigger, apolar residues in order to fill the cavities (e.g. from Phe to Trp or from Val to Ile, Leu, Phe, Tyr or Trp). For their evaluation, the mutations are generated using SCWRL [62], keeping the backbone and side chains of the protein static while allowing movement for the side chain of the mutated residue in order to minimize an energy function. The difference in size (in Å^3) of the volume of internal cavities between the wild type and mutant proteins is used as score.

The "electrostatics" module is based on the work by Estrada et al. [49], using MODELLER [80–82] and REDUCE [83] for the preparation of the structure and Delphi [84] for the calculation of the potential map of the protein. The stabilization profile is calculated using the Native only model, and only the residues with a calculated ionization energy under −2.5 kJ/mol are proposed for mutations. Basic aliphatic residues (Lys or Arg) are mutated to Ala (neutralization) or Glu (inversion), while acidic residues (Asp or Glu) are mutated to their isosteric neutral residues, Asn or Gln, (neutralization) or to Lys (inversion). The score of a mutation is the calculated ionization energy multiplied by −1 (neutralization) or −2 (inversion).

The "steric clashes" module filters out mutations that are likely to generate steric clashes. For that purpose, the difference in energy between the wild type and mutant protein is calculated by SCWRL [62]. If the difference is more destabilizing than 50 (in SCWRL energy units), the mutation is discarded. This value has been estimated using several mutations to bigger residues giving rise or not to steric clashes (data not shown).

The "SASA" module calculates the difference in Solvent Accessible Surface Area (SASA) between the SCWRL-generated mutant and the wild type protein, using the Gromacs sasa module [85–86]. For cavity filling mutations, if SASA grows by >0.275 nm^2 upon mutation, the mutation is discarded, as such growth may be revealing the opening of an internal cavity to the solvent.

Whether to consider or not for scoring in step 4 any of the protein features identified by the modules described above was decided in the model training process, using the training set (see subsection "Logistic regression model training and testing").

2.4. Mutation Databases used and names of models

Several databases and logistic regression models are considered along this work. They have been named as summarized in Table 1. Two datasets (PT^{ori} and PT^{dup}), jointly referred to as PT, have been obtained from ProTherm [87–88]. PT^{ori} is a filtered, therefore reduced, version of ProTherm, while PT^{dup} is a duplicated database containing all mutations in PT^{ori} plus their corresponding reverse mutations. Either version of PT has been used for training and testing a logistic regression model (*Lr*) that calculates the probability of a given mutation of being stabilizing. The *Lr* models obtained for PT^{ori} and PT^{dup} are termed *Lr*^{ori} and *Lr*^{dup}, respectively. The training and testing process for either of these *Lr* models has been done after splitting the corresponding PT in two subgroups, respectively used for training and for testing (e.g., for training and testing *Lr*^{ori}, the PT^{ori} dataset has been split into train-PT^{ori} and test-PT^{ori}).

Table 1
Nomenclature of datasets and predictive models.

Nomenclature		Description
Datasets		
	PT^{ori}	Original PT dataset filtered from Protherm
	PT^{dup}	Duplicated dataset, including PT^{ori} and its reverse mutations
	train-PT^{ori}	Training subset extracted from PT^{ori} in a stratified manner
	test-PT^{ori}	Test subset extracted from PT^{ori} in a stratified manner
	train-PT^{dup}	Training subset extracted from PT^{dup} in a stratified manner
	test-PT^{dup}	Test subset extracted from PT^{dup} in a stratified manner
	ED	External dataset obtained from ThermoMutDB excluding mutations in PT^{ori}
	ED^{+}	ED with the addition of the experimental data from 1PGA and 1FTG
Lr models		
	Lr^{ori}	Logistic regression model trained on PT^{ori}
	Lr^{dup}	Logistic regression model trained on PT^{dup}
Protposer versions		
	Protposer^{ori}	Full ***Protposer*** workflow including the nominating algorithm, *Lr*^{ori} and the sigmoidal model estimating eSR, trained on PT^{ori}
	Protposer^{dup}	Full ***Protposer*** workflow including the nominating algorithm, *Lr*^{dup} and the sigmoidal model estimating eSR, trained on PT^{ori}
Results subsets		
	Protposer$^{ori}_{HM}$	Results of ***Protposer***^{ori} selected according to the half of mutations ($_{HM}$) criterion
	Protposer$^{ori}_{classic}$	Results of ***Protposer***^{ori} selected according to the classic ($_{classic}$) criterion
	Protposer$^{ori}_{O\kappa}$	Results of ***Protposer***^{ori} selected according to the Optimal kappa ($_{O\kappa}$) criterion
Specification of dataset predicted		
	Lr^{dup} → PT^{ori} [a]	Predictions of *Lr*^{dup} on PT^{ori}

[a] The same convention is used throughout the text to indicate the results of any specified model on any specified dataset.

For each dataset derived from Protherm, and for external datasets that will be described below, partitions have been made based on different properties of the protein or the mutation itself: relative exposure, change of size of the mutated residue, protein fold and protein length. Relative exposure has been calculated with the "exposure" module and each mutation has been classified as exposed if the relative exposure is over 30% and as buried otherwise, as previously defined by Caldararu et al [89]. For the change of size of the mutated residue, a mutation is considered volume-changing if the change of volume in the mutated residue is, in absolute value, higher than 30 Å^3, as defined by Caldararu et al

[89]. According to this, there are large-to-small (L2S) and small-to large (S2L) volume-changing mutations, and equal-to-equal (E2E) size ones. The fold of the protein has been obtained from CATH [90] and the length has been taken from the structure as in the PDB database, the protein being classified as long if the length is over 150 residues and short otherwise [89]. These partitions are useful to assess potential biases in the final model.

The role of the alternative *Lr* models (Lr^{ori} and Lr^{dup}) is limited to calculating the probability of individual mutations of being stabilizing. ***Protposer*** is the conjunction of an algorithm designed to nominate potentially stabilizing mutations, an *Lr* model used for evaluating them and a sigmoidal model used to calculate eSR from *Lr* probabilities. The nominating algorithm strongly reduces the number of mutations that are subsequently evaluated by the *Lr* model and improves the interpretability and quality of the results. Depending on the *Lr* model used for evaluating the mutations nominated by the algorithm, two versions of ***Protposer,*** termed ***Protposer***^{ori} and ***Protposer***^{dup}, have been built. Either version of ***Protposer*** has been made to operate always on original, publicly available PDB files (i.e., not usually modelled to include mutations). Therefore, the ***Protposer*** performance has been tested always on PT^{ori}, irrespective of the *Lr* model (Lr^{ori} or Lr^{dup}) implemented in it. To avoid confusion, results are presented through this work indicating both the *Lr* model or ***Protposer*** version used and the database on which it is being evaluated, connected by an arrow (e.g., $Lr^{dup} \rightarrow PT^{ori}$).

In some cases, additional specifications have been applied to ***Protposer*** results to influence the size of the output (a ranked list of likely stabilizing mutations) and its predictive quality. This is done by setting specific minimum eSR values only above which a mutation evaluated is actually considered to have been proposed. Those specifications to ***Protposer*** results are explained in detail in the "***Protposer*** performance assessment" section below. They will be indicated throughout this work using subindices (e.g., $Protposer^{dup}_{Ok}$ for the Optimal kappa specification).

Once ***Protposer*** has been fully trained and tested using the PT datasets, further tests of the final version (i.e., ***Protposer***^{dup}) have been performed on proteins or mutations that were not present in PT. For this purpose, two external datasets, ED and ED^{+}, have been built. ED contains mutations, not present in PT^{ori}, corresponding to the 9 proteins with more newly reported mutations in ThermoMutDB [91] that were not present in ProTherm. ED^{+} is a larger database, additionally including mutations from proteins with PDB IDs 1PGA and 1FTG, whose effects on protein stability have been characterized by Nisthal et al. [92] and Sancho and coworkers [47–48,51–52,55,93–100], respectively.

2.5. Training and testing of the logistic regression model

Experimental mutation stability data have been extracted from the ProTherm database (last release, February 2013) [87–88], filtering to obtain single amino acid point mutations corresponding to proteins of known three dimensional structure, analysed in the 6.0–8.0 pH and 5–45 °C temperature ranges. Data has been consistently checked for misannotations in order to get a high quality database, as suggested by Yang et al. [101] The resulting filtered ProTherm database is termed here PT^{ori}. On the other hand, to get a better balance of stabilizing and destabilising mutations, a larger dataset has been generated (PT^{dup} dataset) under the assumption that $\Delta\Delta G$ for the reverse mutation equals $-\Delta\Delta G$ for the direct mutation (i.e. $\Delta\Delta G^{mut\rightarrow wt}_{unf} = -\Delta\Delta G^{wt\rightarrow mut}_{unf}$) [60,102]. In either dataset, $\Delta\Delta G$ is calculated as $\Delta G^{mut}_{unf} - \Delta G^{wt}_{unf}$, and mutations are considered to be stabilizing if their $\Delta\Delta G$ value is >0.5 kcal/mol. Under this criterion, PT^{ori} and PT^{dup} contain approximately 11 % and 37 % of stabilizing mutations, respectively.

The so filtered experimental stability data deriving from 91 proteins (1692 single mutations for PT^{ori} and 3384 for PT^{dup}) have been further filtered, leaving out the mutations for which the standard deviation of the available stability determinations was higher than the difference between its mean and the 0.5 kcal/mol threshold used to define a mutation as stabilizing. The resulting datasets (1641 mutations for PT^{ori} and 3236 for PT^{dup}) were divided into two groups in a stratified manner, in order to preserve the ratio of positives to negatives in each group. The training groups (train-PT^{ori} and train-PT^{dup}) contain 80 % of the corresponding original data, and the test groups (test-PT^{ori} and test-PT^{dup}), the remaining 20%.

To select, among those provided by the different modules, the optimal features to combine in the mutation evaluation step, the weight coefficient versus regularization strength (C-value) plots were represented. As the scores of all the features considered have been designed to be positive for presumed stabilizing mutations, those features exhibiting abnormal behaviour, such as alternating from positive to negative weights or their weights being always negative or 0, have been eliminated. (Supplementary figure S1). The discarded metrics were those of the "ancestral", "acidic hydrogen bonds" and "steric clashes" modules. Thus, in its current version, the program computes the stabilizing probabilities of the mutations by weighing the scores provided by the "consensus", "alpha helices", "exposure", "cavities", "SASA" and "electrostatics" modules. To select the hyperparameters (L1 or L2 regularization and C value) which best fit each logistic regression model, 10-fold stratified cross-validation has been performed on the training groups, using learning and validation curves (Supplementary figure S2) for both PT^{ori} and PT^{dup}. Once the best hyperparameters have been selected, each logistic regression model has been trained with the corresponding full training set, and it has been evaluated on both the training and test sets, in a holdout manner, so obtaining the weights for the Lr^{ori} and Lr^{dup} models.

2.6. Predictive quality assessment of the Lr models and of Protposer

For comparing the performance of the *Lr* models, and their implementations in ***Protposer***, with that of other currently available classifiers, we have used the same measures as Yang et al. [101] (i.e. accuracy, positive predictive value (PPV), negative predictive value (NPV), sensitivity or true positive rate (TPR), specificity or true negative rate (TNR) and Matthews correlation coefficient (MCC)) on both test-PT^{ori} and test-PT^{dup}. For external validation, only the PPV for different levels of stabilization has been calculated on the external databases ED and ED^{+}.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$PPV = \frac{TP}{TP + FP}$$

$$NPV = \frac{TN}{TN + FN}$$

$$TPR = \frac{TP}{TP + FN}$$

$$TNR = \frac{TN}{TN + FP}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FN) \times (TN + FP)}}$$

where TP, TN, FP and FN are the numbers of true positives, true negatives, false positives and false negatives, respectively. Also, Recei-

ver Operating Characteristic (ROC) curves have been calculated, where the False Positive Rate (FPR) is defined as 1-TNR. As both PT^{ori} and PT^{dup}, as well as subsets thereof used for training and testing, are imbalanced datasets, normalized values for each measure [101] have also been obtained by multiplying the number of negative samples (either TN or FP) by the factor $\frac{TP+FN}{FP+TN}$, thus forcing the number of positive samples (TP + FN) to be the same as that of negative ones (TN + FP).

2.7. Specific quality assessments of Protposer predictive performance

Using the optimal weights for each version of *Lr* obtained as described above, ***Protposer*** (Fig. 1) has been run on the 91 protein structures present in PT^{ori} (***Protposer*** → PT^{ori}). The probabilities calculated by either *Lr* model for each of the mutations nominated by the algorithm that were present in PT^{ori} have been used to obtain the actual PPV of ***Protposer*** as a function of the calculated probability threshold. For that, increasingly smaller subsets of mutations have been selected by using increasingly larger *Lr* probability threshold values. The PPVs calculated for those mutations with a *Lr* probability over that threshold have been plotted as a function of such threshold (Supplementary figure S3). As it is clear in the figure that the probabilities reported by each *Lr* model underestimate the actual PPV of ***Protposer***, realistic estimations of the actual PPV have been obtained from a fit of the plotted data to the following sigmoidal equation:

$$estimated\ Success\ Rate(\%) = \left(b + \frac{h}{1 + e^{-k\times(P-P_0)}}\right) \times 100$$

where b is the minimum value (baseline), h is the difference between the minimum and maximum values (height), k is the steepness of the curve, P is the probability calculated by the *Lr* model and P_0 is the midpoint: the value in which the curvature of the function changes. The newly estimated PPVs obtained with this formula are referred to as "estimated success rates" (eSR) in the output report provided by the ***Protposer*** server and from now on along this work. The eSR reported for a given mutation should not be interpreted as the probability of the individual mutation of being stabilizing, but as the average probability of being stabilising of the different mutations exhibiting that eSR value or higher.

For further evaluating the predictive quality of the ***Protposer*** → PT^{ori} proposal and scoring and to provide general interpretation advice to users, three different mutation selection eSR thresholds have been considered to define shorter lists of mutations from the full proposal. The so-termed "Classic" eSR threshold (eSR > 50%) selects the mutations whose chances of being stabilizing (i.e. of having $\Delta\Delta G$ > 0.5 kcal/mol) are higher than those of not being stabilising. The "Half of Mutations" (HM) eSR threshold is simply designed to select the half of the proposed mutations with higher eSRs for all the mutations nominated for all proteins in PT^{ori}. The Optimal Kappa (Oκ) eSR threshold is set at the maximum Cohen's kappa value, which is calculated using and algorithm similar to the one developed in GHOST [103]. Cohen's kappa [104], a statistic value measuring the agreement between two classifications having into account the chance of agreement at random, is determined using the following formula: $\kappa = 1 - \frac{1-p_o}{1-p_e}$, where p_o is the relative observed agreement and p_e is the randomly expected one. Perfect agreement renders a κ of 1, and worse than random agreement would give negative values of κ. Using the Cohen's kappa value as a selection threshold ensures maximal agreement between proposals and empirical data.

2.8. Validation of Protposer on external databases

Additional testing of ***Protposer*** has been performed by running the program on proteins or mutations not included in the training dataset. Nine proteins have been selected from the relatively new protein stability database ThermoMutDB [91] in order to build the ED external database. Those nine proteins are particularly useful to test **Protposer** for two reasons: they are the proteins with more single point mutations described in ThermoMutDB that are absent in PT^{ori}, and the percentage of their mutations present in ThermoMutDB that was already present in Protherm is low (from 0 to 12 % overlap; Supplementary Table S2). These proteins thus have contributed very little to the training of ***Pirepred***, and the specific mutations selected to conform the ED database have not contributed at all, as mutations in PT^{ori} have been filtered out of ED.

On the other hand, two extensively characterised proteins, for which thermodynamic data on point mutations abound, have been selected for a detailed study of the performance of ***Protposer*** (Supplementary table S3). One of them is the 56-residue β1 immunoglobulin-binding domain of streptococcal protein G (structure with PDB code 1PGA). ***Protposer*** predictions on this protein have been compared with the experimental data obtained by Nisthal et al. [105], who have constructed almost every single mutant of this domain and measured their stabilities using liquid-handling automation and deep mutational scanning techniques. The other protein is the 168-residue *Nostoc* sp. apoflavodoxin (PDB code 1FTG) for with abundant stability data have been reported in previous experimental studies published by our group [47–48,51–52,55,93–100].

The set of novel mutations conforming ED plus those from β1 immunoglobulin-binding domain and apoflavodoxin jointly constitute the external dataset ED^+. ***Protposer*** predictions on ED (***Protposer*** → ED) and on the two additional proteins selected have been compared with the corresponding experimental data in order to find out whether mutations with experimentally determined $\Delta\Delta G$ > 0, 0.25, 0.4 or 0.5 kcal/mol are predicted as stabilizing, and whether the estimated eSR shown in the results report of the ***Protposer*** server agrees with that calculated from the ED^+ data.

2.9. Influence of structure resolution on Protposer performance. PDB file coverage analysis

The proposal dependence on the resolution of the specific 3D structure used has been checked by running ***Protposer*** on structures of three widely characterised proteins: barnase (1A2P, 1BRS), flavodoxin (1FLV, 1RCF) and CheY (1EHC, 3CHY, 5CHY) obtained at different resolution. In the case of CheY, two of the PDB structures contain a mutation (i.e., D13K in 1EHC and Y106W in 5CHY), so any proposed mutation in the position of the experimentally mutated residue (e.g., D13K and D13Q, proposed for 3CHY and 5CHY, K13D for 1EHY or W106Y for 5CHY) has been excluded from the comparison.

For improving and assessing the coverage of structures on which ***Protposer*** can work, two rounds of testing and problem solving with 100 protein structures each have been performed. The 200 protein structures used have been randomly selected from the subset of representative structures at less than 90% sequence identity present in the Protein Data Bank which contain at least one protein chain with a length between 50 and 400 residues and are not part of a big protein complex. For the latter purpose, the query has been filtered for protein structures with less than 4 entities in the PDB file, omitting large structures and not containing "ribosome", "ribosomal", "ribosomic" or derivatives in the title. In the structures with more than one chain, ***Protposer*** has been run on the first one appearing in the PDB and fulfilling the length criterion.

2.10. Comparison of the final version of Protposer with other protein stability servers

The performance of ***Protposer*** (as implemented in the server) on the external database ED^{+} has been compared to that of three representative stability predictive servers (FoldX [32], Rosetta [106–107] and PoPMuSiC [39–41]) selected based on their popularity and accuracy. The mutants used for FoldX and Rosetta assessment have been built using SCWRL4 [62], as described previously for the "cavities" module. For direct comparison with the mutants, the corresponding WT structures have been processed using the same parameters. Calculations with FoldX have been performed with the Stability command with default parameters. In Rosetta, the Relax protocol with 3 separate relaxation trajectories over 5 cycles of sidechain repacking and minimization has been used to calculate, for each mutant, the minimum score of the three trajectories, with the rest of parameters as default. In PoPMuSiC, the systematic procedure available on the server has been used.

To simulate a real-case scenario of optimization of a protein for which no previous quantitative stability determinations are available, only the 10 best mutations predicted by each program for each of the proteins in ED^{+} have been selected for analysis. To simulate a case in which the user is more advanced, an additional criterion for mutation selection has been adopted. In FoldX, only mutations at least 0.5 kcal/mol more stable than the WT protein are considered. In Rosetta, only mutations more stable than the WT protein (in Rosetta Energy Units) are selected, and in PoPMuSiC, only mutations with a negative score (i.e., predicted as stabilizing) are used. The criteria used to select mutations from ***Protposer*** are indicated in the ***Protposer*** performance assessment subsection and they are detailed in the results section.

3. Results

3.1. Dataset properties

Single point mutations have been extracted from ProTherm [87–88] and filtered as indicated in the Methods section. Misannotations in units of temperature and differences in Gibbs energy have been corrected, as suggested by Yang et al. [101] The original resulting dataset, PT^{ori}, is composed of all the mutations that have passed the filters. If more than one $\Delta\Delta G$ value is available for a given mutation, the average is taken as its $\Delta\Delta G$ value if the whole range of average ± SD is completely above or completely below the free energy threshold used for classification. If this condition is not met, the data is discarded. PT^{ori} contains 1641 mutations from 91 proteins, 179 being stabilizing (according to $\Delta\Delta G > 0.5$ kcal/mol) and 1462 are non-stabilizing ($\Delta\Delta G \leq 0.5$ kcal/mol). To be able to assess the impact of the high size imbalance of the stabilizing and non-stabilizing classes present in PT^{ori} (only $\sim$ 11 % of positives, i.e. stabilizing mutations), the approach from Capriotti et al [102] has been additionally implemented. Considering that $\Delta\Delta G_{A\rightarrow B} = -\Delta\Delta G_{B\rightarrow A}$, a duplicated dataset ($PT^{dup}$) has been formed encompassing 3236 mutations from the same 91 proteins as in PT^{ori}, 1208 of which are stabilizing and 2028 are non-stabilizing mutations ($\sim$37 % of positive data, i.e., stabilising mutations). With a symmetrized dataset as PT^{dup}, reaching a 50% of positives is not expected as the dataset is symmetric around 0 kcal/mol but the threshold between positives and negatives (i.e., stabilising and non-stabilising mutations) is 0.5 kcal/mol.

Additional interesting features of a mutation dataset other than containing a balanced number of stabilising and destabilising mutations have been recently reviewed [89,108]. Of the 380 pair-wise substitutions that can be made using the 20 protein amino acids, PT^{dup} contains 329 of them, covering 87 % of all possible substitutions (see the PT^{dup} mutation matrix in Fig. 2 and related train-PT^{dup} and test-PT^{dup} matrices in Supplementary figure S4). As is often the case in mutation datasets, replacements to alanine are overrepresented in PT^{dup}. This type of mutation is popular because it is easier to interpret in the absence of structure for the mutant protein. To counterbalance a predominance of alanine mutations, special datasets have been proposed [89] that limit the occurrence of any type of replacement to be below a certain number. This seems a useful strategy when the goal is advancing in the computing of accurate $\Delta\Delta G$ values for mutations. However, this strategy significantly reduces dataset size and does not seem appropriate for our distinct approach focused on proposing mutations that are likely stabilizing. In contrast with alanine replacements, those involving Cys, Pro or Trp residues are scarce in PT^{dup}. As these residues are either infrequent in proteins or difficult to engineer obtaining a positive effect on protein stability, their low representation in PT^{dup} is of little concern for our approach.

On the other hand, it has been proposed that, in addition to having a balanced number of stabilising and destabilising mutations, datasets should display a balanced distribution of $\Delta\Delta G$ values. As expected of a duplicated dataset, PT^{dup} is highly balanced in this respect and it displays a smooth and rather symmetric distribution of free energies (Fig. 2). The same is true for its train-PT^{dup} and test-PT^{dup} subsets (Supplementary figure S4). It is also appropriate that datasets contain approximately similar fractions of buried and exposed mutations, as this may help to obtain models not limited to issue good predictions for either one type of mutation or the other. To assess the content of buried and exposed mutations in PT^{dup}, the relative solvent exposure of all mutated residues, as calculated using the ProtSA server [77–78], has been retrieved from the "exposure" module. Although PT^{dup} contains more mutations (83 %) from short proteins ($\leq$150 residues) than from longer ones (17 %), it includes a similar percentage of mutations involving exposed (49.6 %) and buried residues (50.4 %) using 30 % relative solvent exposure as cutoff [89] (see Table 2 for PT^{dup} and Supplementary Table S4 for related train-PT^{dup} and test-PT^{dup} datasets). In addition, PT^{dup} is also reasonably balanced in terms of mutations belonging to α proteins (14.4 %), β proteins (49.3 %) and αβ proteins (34.1 %). Finally, PT^{dup} is also balanced in terms of volume change upon mutation. In 31 % of cases the volume of the new residue exceeds that of the wild type one by >30 $Å^3$ [89] (small-to-large substitutions), in 37.0 % of cases the volume is similar (equal-to equal size), and in 32 % of cases the new residue is smaller by >30 $Å^3$ (large-to-small substitutions).

As of the external dataset ED, the distribution of $\Delta\Delta G$ values is also smooth (Fig. 2). However, as ED is not a duplicated dataset, it contains more destabilising than stabilizing mutations and the distribution is left-skewed as expected. The smooth, skewed distribution of ED is well retained by the ED partitions into exposed/buried residues, α/β/αβ fold, short/long length, or small-to-large/equal-to-equal/large-to-small volume (not shown). ED also displays (Table 2) an equilibrated content of mutations of different exposures (45 % exposed and 56 % buried), volume changes (21% small-to-large, 42 % equal-to-equal, 37 % large-to-small) and occurrence of mutations in particular folds (50% α, 14% β, 28 % αβ) or chain lengths (42 % short, 57 % long). Interestingly, the fact that these mutation distributions, particularly those related to protein fold and protein length, differ from those in PT^{dup}, does not apparently have a significant effect in the predictive performance of ***Protposer*** (Table S5), which seems to be robust against dataset composition variation in these terms. The composition of the datasets used to train and test the optimized method are provided, together with the experimental and computed output, as supplementary tabulated txt files.

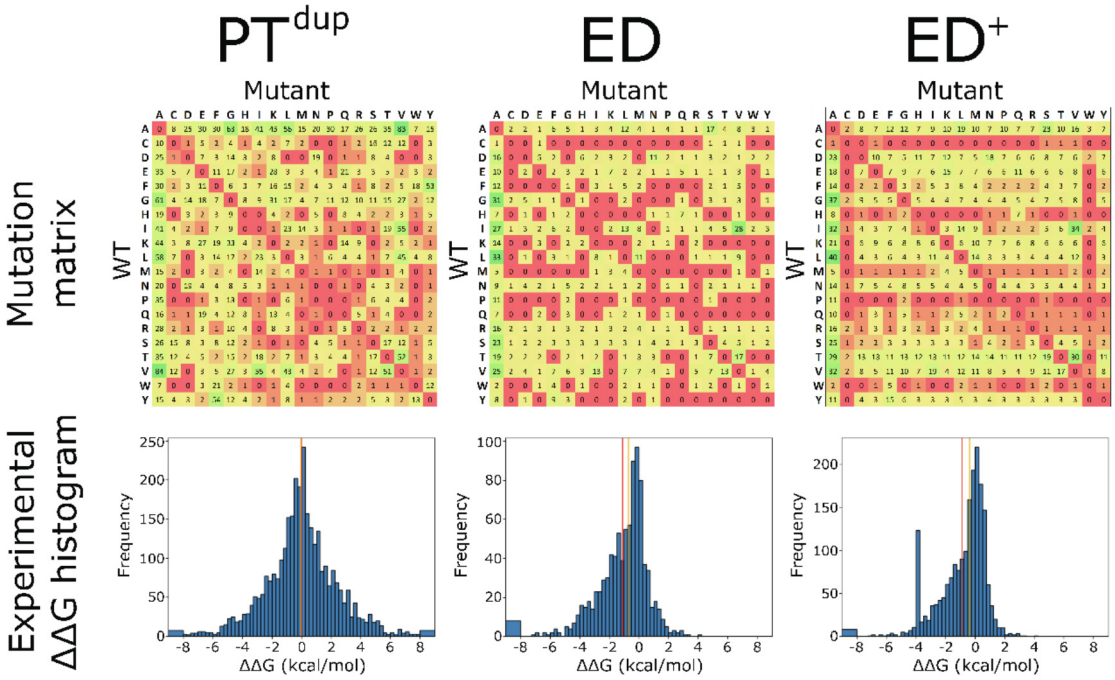

Fig. 2. Dataset properties. In each column, the mutation substitution matrix (first row) and the experimental ΔΔG histogram (second row) is shown for a given dataset: PT^{dup}, ED or ED^{+}. In the mutation matrices, a gradient from red (no mutations) to green (maximum number of mutations for a given type in the dataset) going through yellow (50% percentile) is shown to represent the number of mutations of each kind, also displayed as a number in the centre of its respective square. For the histograms, bins of 0.25 kcal/mol were made, including two bins for mutations over 8 kcal/mol and under −8 kcal/mol. The vertical red lines represent the mean ΔΔG value for each dataset (−0.017 kcal/mol for PT^{dup}, −1.104 kcal/mol for ED and −0.868 kcal/mol for ED^{+}), while the orange ones represent the median (−0.03 kcal/mol for PT^{dup}, −0.7 kcal/mol for ED and −0.36 for ED^{+}). The histogram of PT^{dup} is not perfectly symmetrical due to the asymmetric definition of the bins intervals and the filtering of mutations whose ΔΔG standard deviation was bigger than the difference between its mean ΔΔG value and 0.5 kcal/mol. The outlier bar in the ED^{+} dataset is due to a technical limitation from the data derived of the work of Nisthal et al. [105], in which clearly destabilizing mutations too unstable to be measured or with ΔΔG values under –4 kcal/mol were represented as –4 kcal/mol. However, these mutations are not proposed by ***Protposer*** and do not affect to the binary classification of the mutations according to their experimental ΔΔG. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Mutation type distribution in the datasets.

Type	PT^{dup}	ED	ED^{+}
Buried[a]	50.36%	54.99%	38.78%
Exposed[a]	49.64%	45.01%	61.22%
α[b]	14.43%	50.44%	24.11%
β[b]	49.26%	13.65%	6.52%
α + β[b]	34.09%	27.62%	65.40%
Other fold[b]	2.22%	8.30%	3.97%
Short[c]	82.63%	43.12%	69.15%
Long[c]	17.37%	56.88%	30.85%
S2L[d]	31.24%	21.07%	30.90%
E2E[d]	36.99%	42.14%	38.62%
L2S[d]	31.77%	36.79%	30.48%
Total mutations[e]	3236	916	1916

[a] The exposure classification is performed according to the relative exposure calculated by the exposure module, being buried those mutations with a relative exposure under 30% and exposed those over 30%.
[b] The fold classification is performed according to CATH.
[c] The length classification is performed based on the residue length of the structure in the PDB file, being long over 150 residues and short under 150 residues.
[d] Volume-changing mutations, as opposed to equal-to-equal size mutations (E2E) are those in which the residue volume change upon mutation is >30 $Å^3$. Those are classified as small-to-large (S2L) o large-to-small (L2S) depending on whether the mutated residue is bigger than the WT.
[e] The total number of mutations present in each dataset is directly retrieved from each dataset.

3.2. Performance of the logistic regression models Lr^{ori} and Lr^{dup}

After the nomination of candidates step (Fig. 1) ***Protposer*** runs a candidates evaluation step using a logistic regression model (*Lr*). Alternative logistic regression models for ***Protposer*** have been trained using either PT^{ori} or PT^{dup} and have been assessed using a holdout approach. In each alternative training, a 20% of the corresponding dataset has been selected in a stratified manner as a test group (test-PT datasets) and the remaining 80% has been used as training set (train-PT datasets). The best hyperparameters for the training of each model have been identified using a 10-fold stratified cross-validation, being L2 regularization with a C-value of 0.1 for both Lr^{ori} and Lr^{dup}. Once trained, each model has been separately tested on both its training and test set to check for possible overfitting.

The performance analysis of the alternative models shows (Fig. 3) that Lr^{dup}, the one trained on PT^{dup}, outperforms Lr^{ori} (trained on PT^{ori}), in both the PPV and the ROC curves. The closeness between the Lr^{dup} → train-PT^{dup} and Lr^{dup} → test-PT^{dup} PPV and ROC curves (Fig. 3B and 3D) suggests there is no overfitting, which is confirmed by their learning curves (Supplementary S2D). As for Lr^{ori}, although a gap between Lr^{ori} → train-PT^{ori} and Lr^{ori} → test-PT^{ori} is seen in their ROC curves, their learning curves (Supplementary S2C) show no signs of overfitting either. The apparently lower performance of Lr^{ori} in their respective datasets may be due to the imbalance in positive (36) and negative (287) mutations in test-PT^{ori}, making the PPV and other performance metrics very sensitive to outliers.

The performance of Lr^{ori} and Lr^{dup} on their respective test-PT datasets is compared in Table 3 with that of other classifiers, previously analysed by Yang et al. [101] Lr^{dup} outperforms the rest of classifiers at the metrics that are more relevant for a stabilizing mutation proposing objective. Of special interest for that is the PPV, as the main goal of the program is to propose the user a small set of mutations highly enriched in truly stabilizing ones. Our program outstands in this metric as, out of the four other methods

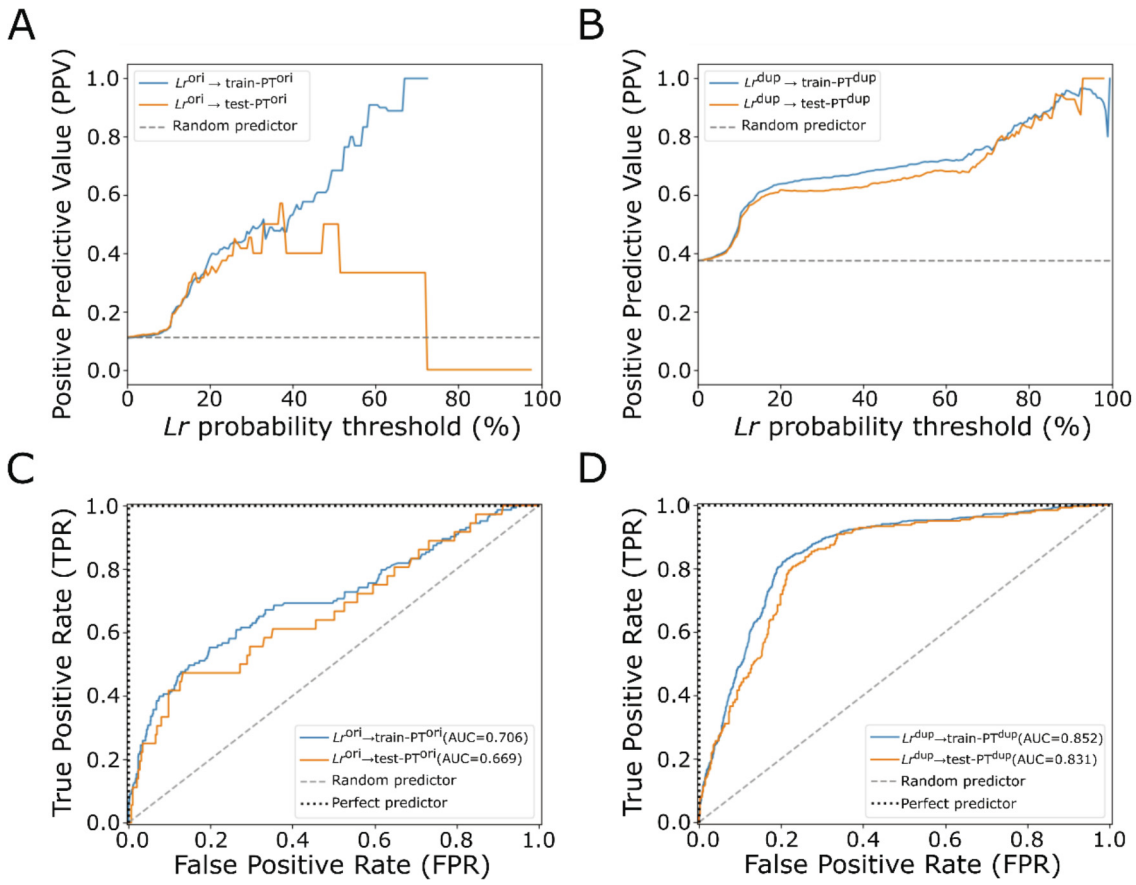

Fig. 3. Evaluation of the alternative logistic regresion models (*Lr*) on their respective training PT datasets. Blue lines are used to represent the quality of the prediction issued by the indicated *Lr* model on the training set, while orange lines represent the evaluation of the predictions on the test set. Dashed grey lines show the performance of a random predictor. A) and B) PPV for different *Lr* probability threshold values for Lr^{ori} (A) and Lr^{dup} (B). C) and D) Receiver Operator Curves (ROC) for Lr^{ori} (C) and Lr^{dup} (D). The area under the curve (AUC) for each ROC curve is indicated. A higher value of AUC corresponds to a better performing model, being 1 for a perfect predictor (dotted black line) and 0.5 for a random predictor (dashed grey line). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Comparison of the predictive performance of *Lr* with several predictors.

Performance measures [a]	Predictors					
	Lr^{ori} [b]	Lr^{dup} [b]	EASE_MM[c]	I-Mutant[c]	INPS[c]	PON-tstab[c]
Predicted mutations	323	636	40	40	15	165
TP	2/2	131/131	0	0	0	3/3
TN	35.75	59/32.9	34/6	33/5.8	15	126/20.4
FP	2/0.25	246/137.1	0	1/0.2	0	16/2.6
FN	34/34	39/39	6/6	6/6	0	20/20
Accuracy	0.89/0.52	0.78/0.79	0.85/0.5	0.83/0.49	1	0.78/0.51
Sensitivity	0.06/0.06	0.84/0.84	0/0	0/0	NA[d]	0.13/0.13
Specificity	0.99/0.99	0.74/0.74	1/1	0.97/0.97	1	0.89/0.89
PPV	0.50/0.89	0.66/0.76	NA[d]	0/0	NA[d]	0.16/0.54
NPV	0,89/0,51	0.88/0.82	0.85/0.50	0.85/0.49	1	0.86/0.51
MCC	0.14/0.14	0.56/0.58	NA[d]	−0.07/-0.12	NA[d]	0.02/0.03

[a] Results before and after normalization of positive and negative cases (stabilising and non-stabilising mutations, respectively) are separated by a slash (see main text for further explanation on normalization of these data).
[b] Performance of *Lr* models evaluated on their respective test subset (test-PT^{ori} and test-PT^{dup}).
[c] Data extracted from Yang et al. [101].
[d] NA: Not Available, division by 0.

compared, only PON-tstab gets a PPV value different from 0 which is, nevertheless, much lower than that obtained for our method. Lr^{ori} predicts few positives when the standard probability cut-off of 50% is used, which hampers drawing significant conclusions on some of the performance metrics, such as PPV. The restrictiveness of Lr^{ori} for predicting mutations as positive is most probably due to the fact that the cost function of the machine learning algorithm is based on accuracy. As there are very few positive cases in PT^{ori}, not predicting them as positive is less costly than predicting some negatives as positive, therefore making the model less prone to predicting positives. Importantly, if a lower probability threshold is used (i.e., when mutations over 33–38 % probability are predicted as positive), Lr^{ori} achieves a PPV of 50% or higher (Fig. 3A), which still outperforms the other methods (Table 3).

3.3. Performance of Protposer on PT^{ori}

Protposer has been run on the 91 proteins present in PT^{ori} in order to evaluate its performance (Fig. 1), fit the sigmoidal model that calculates the eSRs (Supplementary figure S3), and select the final evaluating model (Lr^{ori} or Lr^{dup}) which will be implemented in the version available to users. To first assess whether the ***Protposer*** combination of a nominating algorithm with an *Lr* model represents an improvement versus using an *Lr* model alone, the performance of ***Protposer***^{ori} and ***Protposer***^{dup} (illustrated in Fig. 4) is directly compared to that of Lr^{ori}, Lr^{dup}, on PT^{ori} in Supplementary figure S5. The decline of the relative number of proposals from the *Lr* models operating on their own occurs in a more abrupt way and with a lower eSR than when incorporated to the corresponding ***Protposer*** version. As each ***Protposer*** version uses for mutation scoring the same parameters as its corresponding *Lr* model, the nominating algorithm incorporated in ***Protposer*** appears to effectively select a subset of mutations enriched in those exhibiting higher eSRs. Comparison of the PPV and average experimental ΔΔG values as a function of eSR threshold obtained by either version of ***Protposer*** with those of their respective *Lr* models (Supplementary figures S5B and S5C), reveals that improvements of around 25% in PPV and of 0.2 to 0.4 kcal/mol in ΔΔ*G* are obtained using ***Protposer***. This implies that the rational rules used for nominating mutations efficiently retrieve mutations sets that are enriched in stabilizing mutations.

To select the best performing version of ***Protposer***, we have considered the fact that the nominating modules only analyse WT structures. Therefore, only PT^{ori} has been used for the comparison of the two ***Protposer*** versions. For either version, 7318 mutations in the 91 proteins have been nominated by the modules of the nominating algorithm. Of those mutations, 114 are present in PT^{ori}. The ***Protposer***^{ori} → PT^{ori} and ***Protposer***^{dup} → PT^{ori} results are compared in Fig. 4. The dependence of the relative amount of proposed mutations on the eSR threshold imposed (Fig. 4A) shows that, as the threshold increases, fewer mutations remain. Although the number of mutations proposed by ***Protposer***^{ori} and ***Protposer***^{dup} is similar, ***Protposer***^{dup} may be slightly preferred as it appears to propose more mutations in the higher eSR range. Along the entire PPV and average experimental ΔΔG curves (Fig. 4B and C), ***Protposer***^{ori} and ***Protposer***^{dup} perform similarly well, except for a slight improvement of ***Protposer***^{dup} over ***Protposer***^{ori} in the average measured ΔΔG values corresponding to eSR thresholds between 55% and 70%. Based on this and on the slightly higher number of proposals issued by ***Protposer***^{dup} (Fig. 4A) we have selected ***Protposer***^{dup} as the version implemented in the server. The maximum PPV for ***Protposer***^{dup} (71.4%) is reached using an estimated eSR threshold of 67.0 % (Fig. 4B). Actually, a PPV of

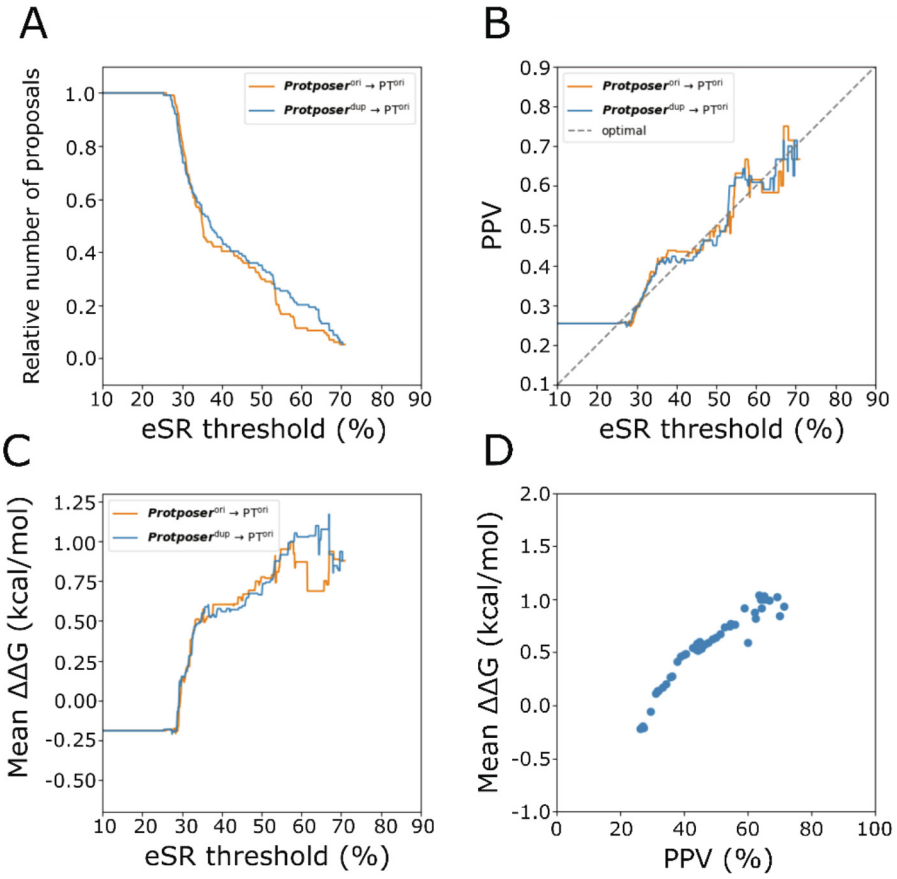

Fig. 4. Predictive performance of the alternative versions of *Protposer* on PT^{ori}. Orange lines describe the performance of ***Protposer***^{ori}, while blue lines refer to the performance of ***Protposer***^{dup}. The grey line in panel B indicates the expected behavior for an optimal sigmoidal model for the calculation of the estimated success rate (eSR), where the eSR given coincide with the actual positive predictive values (PPV). A) Relative number of proposals for each model with an eSR over a certain threshold. The relative number of 1 is obtained using 0% estimated eSR as threshold, which corresponds to 7318 proposed mutations. B) PPV for each subset of proposed mutations in PT^{ori} all having eSRs above the indicated threshold. C) Average of the experimental ΔΔG values determined for the proposed mutations in PT^{ori} for which the eSRs are above the indicated threshold. D) Average of the experimental ΔΔG values determined for the proposed mutations in PT^{ori} vs Positive Predictive Value. The plot is obtained by calculating the PPV and mean ΔΔG for eSR thresholds between 0 and 100%, separated by 0.5%. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

approximately 70 % is maintained all over the 65–70% range of estimated eSR threshold values. A decay of PPV is observed at estimated eSR thresholds above 70%. As it is most likely due to the scarceness of positive results, it is not shown in the figures. ***Protposer*** shows a fine performance as it reaches a PPV above 70%. In addition, the eSR given by ***Protposer*** in its output corresponds well to the actual PPV (diagonal line in Fig. 4B) and therefore eSR constitutes a reliable and easy-to-interpret metric for users. While the eSR in ***Protposer*** is not directly related to the ΔΔG values of the mutations used for training, higher eSR threshold values lead to higher PPVs, and there is a clear trend indicating that, as the PPV increases, the mean experimental ΔΔG value increases as well (Fig. 4D). This trend has been observed not only in the subset of mutations proposed by ***Protposer*** but also in the whole datasets (Supplementary Figure S6).

When disaggregating the proposed mutations on PT^{ori} by different properties (Supplementary table S5), no clear effects on PPV are observed depending on exposure or protein fold, other than the high PPV derived from the low number of mutations with α fold. Differences between different protein length and volume change upon mutation are found in PT^{ori}, but they are clearly reduced or even inverted in ED^{+}, so they are likely due to the low number of proposed mutations in both datasets and not to those properties being confounders.

To simplify comparison with other protein stability software and to provide interpretation insight that may be useful to users, an additional evaluation of ***Protposer*** has been carried out using three different eSR thresholds to define mutation subsets from the total proposal done by either ***Protposer***^{ori} or ***Protposer***^{dup}. The Classic threshold selects the mutations with eSR > 50%; the Half of Mutations threshold selects the mutations with an eSR higher than the median eSR value for all proposed mutations for all proteins in PT^{ori}; and the Optimal κ threshold selects mutations with eSR higher than the eSR value at which Cohen's kappa [104] is maximized. The values of the thresholds obtained by each version of ***Protposer***, as well as the PPVs when using them on ED^{+} (see below) are shown in Table 4.

3.4. Performance of Protposer on external datasets

There is evidence that the reported performance of a classifying algorithm on the data set used for its training tends to be higher than when it is tested on external datasets [35,60–61,109]. Throughout this work, we have carefully minimized the possibilities of overfitting in ***Protposer***, and we have shown that its performance in the subset of mutations left aside from the training (test-PT) is as good as its performance on the training set (train-PT). Still, we deem it necessary to further test the ***Protposer*** performance by evaluating the predictions it makes on mutations that had not been known at the time the final server was ready (i.e., mutations that were not present in either train-PT or test-PT. To that end, we have performed an additional validation test on an external mutation dataset of 9 proteins derived from ThermoMutDB [91] (ED), and on two individual proteins: the rather small, 56-residue β1 immunoglobulin-binding domain of streptococcal protein G (PDB structure 1PGA), and the larger, but still one-domain, 168-residue *Nostoc* sp. apoflavodoxin (PDB structure 1FTG).

First, data recently obtained by Nisthal et al. [105] have allowed us to test the performance on a new set of mutations available for PDB structure 1PGA (Fig. 5, supplementary table S3). Although this protein was present in the training database (5 mutations in 1PGA were present in PT^{ori}) none of the mutations proposed by ***Protposer*** for 1PGA had been included in ProTherm and, therefore, in PT datasets. Out of the 7 mutations proposed by ***Protposer*** for 1PGA, 3 are stabilizing according to our general criterion ($\Delta\Delta G \geq 0.5$ kcal/mol), one additional mutation is very close to it ($\Delta\Delta G \geq 0.4$ kcal/mol), and only one is destabilizing. The average ΔΔG for the 7 proposed mutations is 0.13 kcal/mol. To consider alternative scenarios of ***Protposer*** use in real cases, two additional indicators can be calculated. Accumulated ΔΔG is the sum of the experimental ΔΔG of all proposed mutations (if more than one mutation is proposed at a position, the ΔΔG of the first one indicated by ***Protposer*** is selected). It represents the user expectation when all the mutations proposed are engineered simultaneously and their effects are accumulative. The accumulated ΔΔG for 1PGA is 0.92 kcal/mol. On the other hand, maximal ΔΔG is the sum of the experimental ΔΔG of all proposed mutations that are non-destabilising. It represents the user expectation when the mutations are engineered individually and, afterwards, the non-destabilising ones are jointly engineered. The maximal ΔΔG for 1PGA is 2.69 kcal/mol. When using the Optimal Kappa criterion, only one stabilizing mutation is proposed, with a ΔΔG of 0.97 kcal/mol, being this the average, the accumulated and the maximal ΔΔG.

Table 4
Predictive performance of different versions of *Protposer* on ED^{+}.

Model [a]	SR decision threshold [b]	# mutations	With data in ED^{+}	PPV 0 [c]	PPV 0.25[c]	PPV 0.4 [c]	PPV 0.5 [c]
Random[d]	NA	24,472	2006	34.4%	24.0%	14.4%	10.4%
Protposer[e]	25%	640	83	51.8%	39.8%	32.5%	26.5%
Protposer$^{ori}_{HM}$ [f]	35%	323	49	73.5%	57.1%	46.9%	38.8%
Protposer$^{dup}_{HM}$ [f]	38%	308	55	61.8%	47.3%	40.0%	34.5%
Protposer$^{ori}_{classic}$ [g]	50%	133	21	76.2%	66.7%	66.7%	57.1%
Protposer$^{dup}_{classic}$ [g]	50%	180	26	73.1%	61.5%	57.7%	50.0%
Protposer$^{ori}_{Oκ}$ [h]	43%	185	27	74.1%	59.3%	55.6%	48.1%
Protposer$^{dup}_{Oκ}$ [h]	64.3%	93	14	78.6%	78.6%	78.6%	78.6%

[a] Model used for the selection of the mutations predicted as stabilizing. It indicates the training dataset, being the original dataset (^{ori}) or the duplicated symmetrized dataset (^{dup}), and the criteria for the selection of the estimated eSR decision threshold ($_{HM}$, $_{Oκ}$ or $_{classic}$).
[b] Minimum value of the estimated eSR calculated by ***Protposer*** for a mutation in order to be considered as predicted positive. Not available for a random predictor (NA).
[c] Positive Predictive Value considering as actually positive (stabilising) mutations those with an experimental ΔΔG value higher than 0, 0.25, 0.4 or 0.5 kcal/mol, respectively, as indicated.
[d] Calculated as the statistics for the whole dataset.
[e] All mutations proposed by Protposer. The decision threshold here is the minimum value achievable using the reported equation for the calculation of estimated eSR (see methods).
[f] The decision threshold has been selected in order to discard half of the proposed mutations in PT^{ori}.
[g] The decision threshold has been selected with the classic criterion of being positive only when the probability of being a true positive (as given by the calculated eSR) is higher than the probability of being a false positive.
[h] The decision threshold has been selected in order to get the maximal value of Cohen's kappa coefficient of agreement between the classification of the predicted mutations and the experimental values.

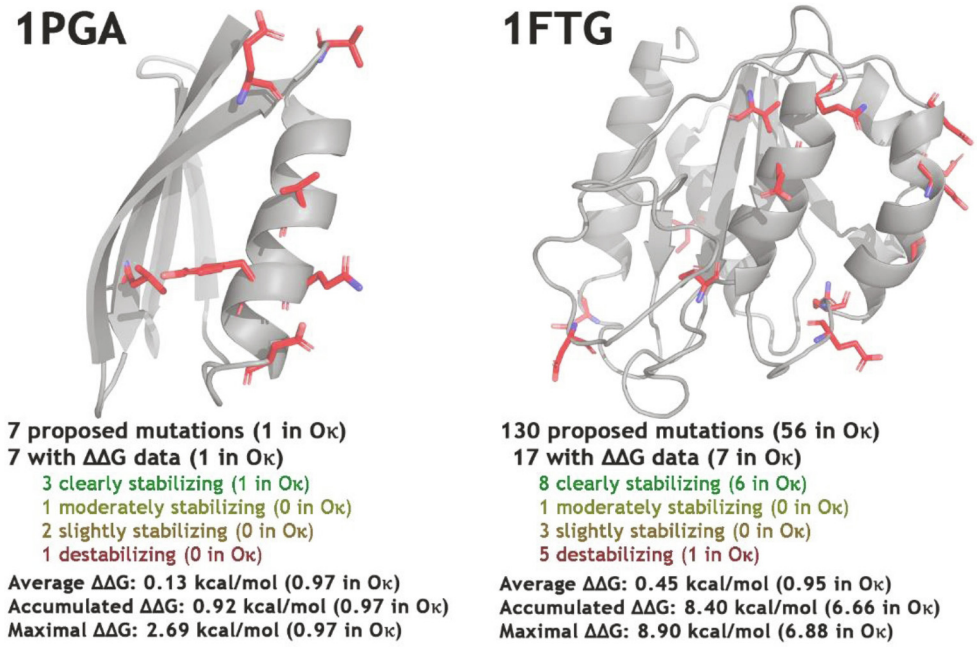

Fig. 5. Predictive performance of *Protposer* on 1PGA and 1FTG. A comparison of ***Protposer*** predictions on the structures of the 56-residue β1 immunoglobulin-binding domain of streptococcal protein G (1PGA) and of the 168-residue apoflavodoxin (1FTG) with empirical data obtained by Nisthal et al [92] and Sancho and coworkers [47–48,51–52,55,93–100], respectively, is summarized in the lower part of the figure. Cartoon representations of the three-dimensional structures of each protein are shown with the residues proposed for mutation displayed as red sticks. Colours used in the summary for each type of mutations according to their ΔΔG experimental values follow the same code as in supplementary table S2. The average ΔΔG is the average of the ΔΔG values for all proposed mutations for the protein. The accumulated ΔΔG estimates the ΔΔG for a mutant protein including all proposed mutations by summing up their individual ΔΔG values. If several mutations are proposed for a given position, the first in the proposed list returned by Protposer is used, as they are ordered from higher to lower eSR. The maximal ΔΔG estimates a scenario where all mutations are tested and only those with positive ΔΔG are engineered in the final protein, so only positive values of ΔΔG are summed. If several stabilizing mutations are proposed for a given position, the most stabilizing one is selected. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

A second protein-specific test can be done from the data obtained in our group on the conformational stability of apoflavodoxin [47–48,51–52,55,93–100] using the PDB structure 1FTG (Fig. 5, supplementary table S3). This protein is not a totally naïve example, as 26 of the 71 mutations that have been experimentally analysed were already present in PT^{ori}. ***Protposer***^{dup} did 130 proposals for apoflavodoxin, of which 17 had been experimentally characterized (7 were present in PT^{ori}). Out of the 17 proposed mutations, 5 have a negative experimental $\Delta\Delta G$ value, while 8 are clearly stabilizing ($\Delta\Delta G \geq 0.5$ kcal/mol). The obtained average $\Delta\Delta G$ is 0.45 kcal/mol, with an accumulated $\Delta\Delta G$ of 8.4 kcal/mol and a maximal $\Delta\Delta G$ of 8.9 kcal/mol. Importantly, the mutations that are proposed with a high estimated eSR are very successful as 6 out of the 7 mutations fulfilling the Optimal Kappa criterion display $\Delta\Delta G \geq 0.5$ kcal/mol and only 1 is destabilising. For these mutations, the average $\Delta\Delta G$ is 0.95 kcal/mol, the accumulated $\Delta\Delta G$ is 6.66 kcal/mol and the maximal $\Delta\Delta G$ is 6.88 kcal/mol. These results illustrate the usefulness of ***Protposer*** in proposing mutation lists highly enriched in stabilizing mutations, and indicate that the eSR provided by the server is a good indicator of the success expectation for proteins that are analysed for the first time. They also illustrate that, while selecting only the best mutations greatly improves the average stabilization per mutation, even bigger absolute stabilizations can be obtained, at a higher experimental cost, by individually determining the $\Delta\Delta G$ of the proposed mutations and combining the non-destabilising ones.

Finally, to carry out a wider and even blinder analysis of ***Pirepred***^{dup} performance, a completely external mutation dataset (ED) has been constructed from ThermoMutDB [91], a fairly new database containing thermodynamic data for protein mutations. First, any mutation present in PT^{ori} was filtered out from ThermoMutDB. Then, the 9 proteins with >50 mutations left in ThermoMutDB (1BPI: Bovine pancreatic trypsin inhibitor; 1BVC: Apomyoglobin from *Physeter macrocephalus*; 1LZ1: Human lysozyme; 1RGG: Guanyl-specific ribonuclease Sa from *Kitasatospora aureofaciens*; 1RX4: Dihydrofolate reductase from *E. coli*; 1SHF: SH3 domain of human Fyn; 1TEN: Fibronectin type III domain from human tenascin; 2LZM: Bacteriophage T4 lysozime and 2RN2: Ribonuclease H from *E. coli*) were selected to conform the external database named ED. Those 9 proteins were analysed using ***Protposer***^{dup}, obtaining with the optimal kappa criterion a PPVs of 66.7%, slightly higher than the corresponding eSR threshold of 64.3 %.

For the larger ED^{+} external database encompassing both ED and the two proteins (1PGA and 1FTG) individually analysed, the good correlation between the estimated eSR and the real PPV determined from the published stability data (Table 4, Fig. 6B) is maintained. Besides, the similar behaviour exhibited by ***Protposer***^{dup} when tested on either PT^{ori} or ED^{+} (supplementary figure S7) indicates there is no sign of overfitting in the trained *Lr* model (Lr^{dup}) used for evaluating the mutations. A strong indication of ***Protposer*** usefulness for pinpointing protein stabilising mutations in ED^{+} is the high PPV obtained (73.3 %) and the high mean experimental $\Delta\Delta G$ (0.8 kcal/mol) for the 15 mutations proposed with the optimal kappa criterion (Table 4, Fig. 6C).

3.5. Protposer coverage, sensitivity to X-ray resolution, and number of mutation proposals to expect

200 PDB structures were randomly selected to perform 2 cycles of testing and debugging, using 100 structures in each cycle. After this process, the percentage of structures for which ***Protposer*** successfully returned results was of 95%. The 5 % of structures for which ***Protposer*** could not finalize the predictions encompassed

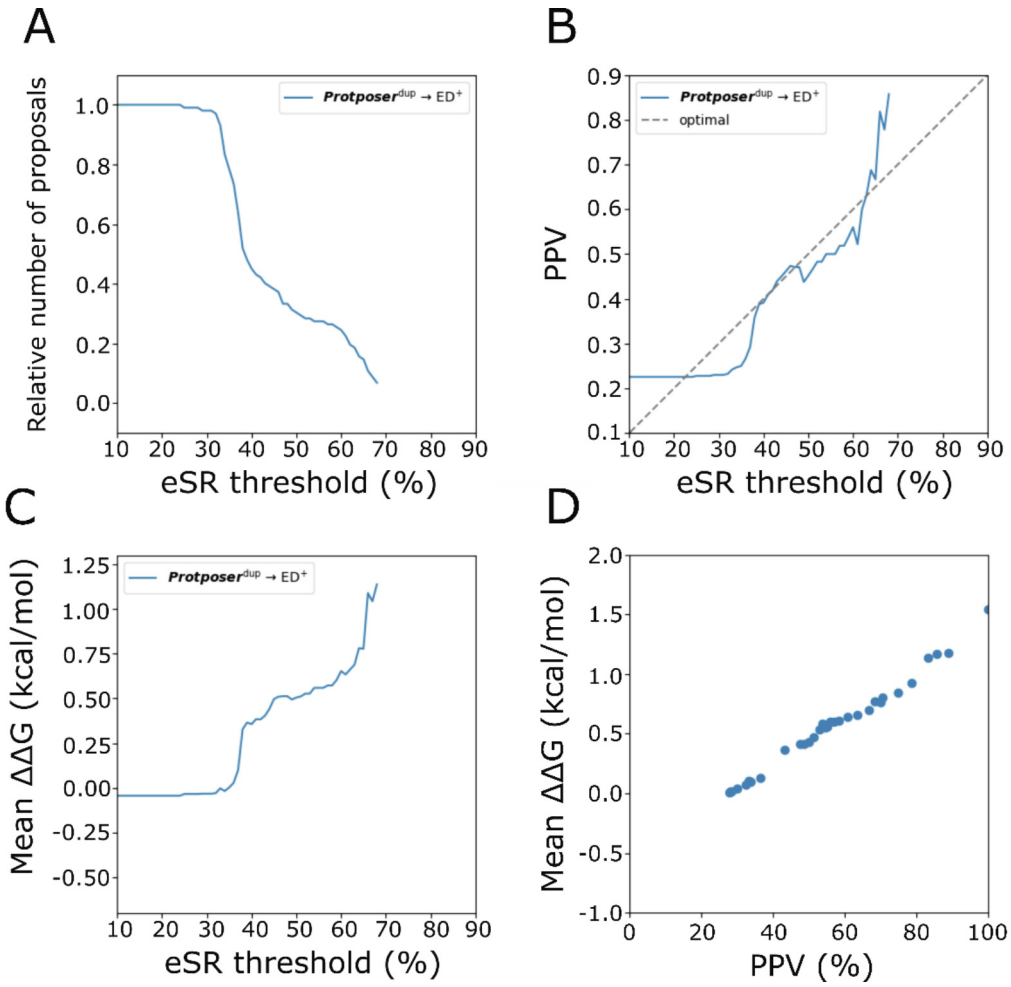

Fig. 6. Predictive performance of the final version of *Protposer* on ED⁺. Blue lines describe the performance of ***Protposer***^{dup}, the version of ***Protposer*** selected for the server. The grey line in panel B indicates the behavior expected for an optimal sigmoidal model for the calculation of estimated success rates (eSR), where the estimated success rates given coincide with the actual positive predictive values. A) Relative number of proposals for each model with eSR over a certain threshold. The relative number of 1 is obtained using 0% eSR as threshold, which corresponds to 672 proposed mutations. B) Positive Predictive Values (PPV) for each subset of proposed mutations in ED⁺ all having estimated eSRs above the indicated threshold. C) Average of the experimental ΔΔG values determined for the proposed mutations in ED⁺ for which the estimated eSRs are above the indicated threshold. D) Average of the experimental ΔΔG values determined for the proposed mutations in ED⁺ vs Positive Predictive Value. The plot is obtained by calculating the PPV and mean ΔΔG for eSR thresholds between 0 and 100%, separated by 0.5%. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

a 2% of structures with incorrect numbering of the PDB file and a 3% of structures rendering a wide variety of rare errors. For the 2 % of structures with incorrect numbering, an informative email is returned to the user to facilitate the renumbering needed for resubmitting.

On the other hand, the sensitivity of the program to the X-ray resolution of the protein structure analysed has been investigated running the program on several structures of the same proteins (barnase, flavodoxin and CheY) solved at different resolutions. The predictions are similar for structures differing in resolution (from 1.08 to 2.26 Å) (Supplementary Table S6). For these proteins, the structure with less proposals contained at least 77 % of the total of all proposed mutations in any of the analyzed structures. The estimated eSR calculated for the mutations proposed in common in structures of different resolution correlate well, with Pearson correlation coefficients always over 0.90, with an average of 0.95.

A rough estimation of the number of mutations that ***Protposer*** will propose for a PDB file of a given length has been obtained from analysis of predictions on the external database ED⁺. For those proteins (lengths ranging from 56 to 168 residues), 672 proposals were made. A linear fit of number of proposals and sequence length (supplementary figure S8) rendered the following equation with an R^2 value of 0.749.

$$\#proposals = 0.82 \times length - 42.85$$

Having into account that the median size of proteins coded in the genome of *Homo sapiens* is 414 residues [110], and extrapolating the obtained fit, the average expected number of proposals for a human protein is 299, of which an approximate 14% (43 proposals) will fulfill the most restrictive selection criterion (optimal κ) yielding a PPV of 69%. Thus, around 29 mutations, each increasing the stability by > 0.5 kcal/mol, should be expected. However, this kind of analysis may be more illustrative if done on folding domains rather than entire multidomain proteins. The reason is that proteins containing several structural domains [111], and even some single domain proteins [99], may exhibit a non-fully cooperative stability behaviour. In those proteins, an individual domain should be the specific target for stabilization [55]. As the average length of protein domains is around 150 residues [112–113], the expected number of proposals per domain is of 80, with 11 of them expected to fulfill the optimal κ criterion, meaning that 7–8 of them will each stabilise the domain by at least 0.5 kcal/mol.

3.6. Comparison with current methods

To compare ***Protposer*** performance with that of other software currently used for calculation of protein stability changes upon mutation (Table 5), the ED^{+} dataset has been analysed with FoldX [32], Rosetta [106–107] and PoPMuSiC [39–41]. These servers have been selected for comparison with the implemented version of ***Protposer*** available to users, due to their representativeness and accuracy. One comparison has been done selecting the 10 best ranked predictions of each server regardless of their predicted energies. Out of the three tested predictors, the best performance, evaluated by their PPV, is obtained by PoPMuSiC (28.2 %), followed by Rossetta (25.5 %) and FoldX (20.9 %). As the best predicted mutations of the servers may not be stabilising in all cases, a more stringent and meaningful comparison has been done by focusing on the best mutations that are specifically predicted as stabilising, thus discarding any destabilising one. In this selection scenario, which would be likely favoured by an experienced user, the performances of the three servers increase a bit, the best still being that of PoPMuSiC (PPV of 32.3%), followed by Rosetta (PPV of 29.0 %) and FoldX (22.1 %). As ***Protposer*** estimates the success rate of the mutations it proposes, its PPV depends on the eSR threshold used to define the short list of mutations to be implemented. If no threshold is used, ***Protposer*** PPV is of 36.1 %, which increases to 42.9 % if the Half of Mutations threshold is used, to 56.0 % if the Classic threshold is used, and to 78.6 % using the Optimal kappa eSR threshold. The much higher PPV of ***Protposer*** with the Oκ threshold (78.6%) compared to that of PoPMuSiC (32.3 %) is statistically significant, having a p-value of 0.0006 in one-sided T test. If the 7 mutations proposed for 1FTG that are present in ED^{+} but also in PT^{ori} are removed from the analysis, the PPV of ***Protposer*** with the Oκ threshold is of 71.4 %, still much higher than the 32.2 % calculated in that case for PoPMuSiC (Supplementary table S7), with a p-value in one-sided T test of 0.0193, still significant. Furthermore, if not only the 10 best mutations for each protein but all mutations fulfilling the selection criteria for each program are selected (Supplementary table S8), the predictive performance gap between PoPMuSiC and ***Protposer*** with the Oκ threshold remains of a similar magnitude but, as more mutations are available, the significance of these results increases, with a p-value of 0.0005.

3.7. Interface

As previously indicated, ***Protposer***^{dup} has been implemented in the final version of ***Protposer*** (Fig. 1). A user-friendly interface has been designed to allow both experienced users from protein-related fields as well as novices an easy access to the server. Few, simple parameters are requested to launch the calculations, without any special knowledge of bioinformatics required (Fig. 7). The input screen (Fig. 7A) enables users to upload their own PDB files or to name one for automated retrieval from the Protein Data Bank (PDB). In either case, the user must indicate the PDB file chain on which the calculations will be performed. Besides, a name for the project and an email address are requested in order to send the results to the user.

The results page (Fig. 7B) consists of a self-explanatory table displaying the mutations proposed as rows ordered from higher to lower estimated eSR (i.e. higher to lower chance of increasing the stability of the protein by >0.5 kcal/mol). The columns are organized as follows. The first column corresponds to the mutation, expressed in O#M format, where O and M are, respectively, the original and mutated residues in one letter amino acid code, and # is a number specifying the position of the mutation (e.g., D144K represents the replacement of an aspartic acid residue by lysine at position 144). The second column shows the estimated eSR calculated for the mutation, which reflects the actual probability of the mutation being stabilizing by at least 0.5 kcal/mol, as calculated by ***Protposer***. The third to fifth columns give warnings if the mutation affects a catalytic or binding site of the protein, as defined in the Catalytic Site Atlas (CSA) or annotated in the PDB file (SITE) or, for oligomeric proteins, if it takes place at the interface formed by the specific polypeptide chain that has been analysed and another chain present in the original PDB file. The sixth column indicates the ***Protposer*** module that initially nominated the proposed mutation. The remaining six columns show qualitatively the individual evaluations done by each of the evaluating modules, color-coded and using signs, so that it is accessible for colour-blind users. Each column contains a brief explanatory help that can be visualized by hovering over the heading of the column. A legend at the end of the table provides the full names of the evaluation modules and their colour and sign codes. Additionally, in the web version of the report, a protein structure viewer is available for locating the mutation site in the structure upon clicking its name.

4. Discussion

4.1. Need of a simple protein stabilising program with high PPV

Biotechnological and medical use of proteins often requires previously increasing their conformational stability so that they can remain active in harsh solution conditions (presence of organic solvents, extreme values of pH or temperature) or be transported and stored in non-demanding manners. Purely experimental approaches for protein stabilization have been developed based on generating protein variants and selecting or identifying the more stable ones [114–116]. Advantages and disadvantages of these methods have been discussed elsewhere [117]. On the other hand, protein biophysicists and bioinformaticians have struggled to develop automated computational methods that were able to calculate the effect of specified mutations on protein stability. Some of those computational methods [32,118–126] have specialised for their use in genetic interpretation [110] by combining an evaluation of stability determinants with evolutionary data reporting in both stability and function. Such methods are not of use in protein engineering, as they do not provide for the mutations tested neither an estimation of the stability change nor a classification of mutations purely based on stability considerations.

Many other computational methods have focused on calculating the change in folding free energy brought about by point mutations ($\Delta\Delta G_{fol}^{wt-mut}$) [27,30,32,60,106,40–41]. As a direct calculation of that change is still not possible, except for conservative mutations and using computationally demanding methods [127–128], a variety of empirical energy functions have been derived as proxies for the changes in folding free energy. Several critical evaluations of those methods [25,60] indicate that the correlations they get between calculated and experimentally determined $\Delta\Delta G$ values display Pearson correlation coefficients under 0.6, average unsigned errors over 1 kcal/mol [33], and self-consistency biases over 0.7 kcal/mol [34]. Those unsigned errors are in the range of the stability effects commonly observed for point mutations in proteins. Thus, the calculated stability differences often miss even the right sign of $\Delta\Delta G$, predicting as stabilizing mutations that turn out to be destabilising, or the other way around. In general, those predictors are much more accurate at calculating destabilizing mutations (average success rate of 69%) than stabilizing ones (average success rate of 29.4%) [25,60]. Both the known high imbalance between abundant destabilizing and scarce stabilizing mutations in the datasets used for training these methods and suboptimal selection of scoring features may contribute to their poor performance [60,101]. It has been suggested that, for reaching higher accuracies, the training datasets have to be improved

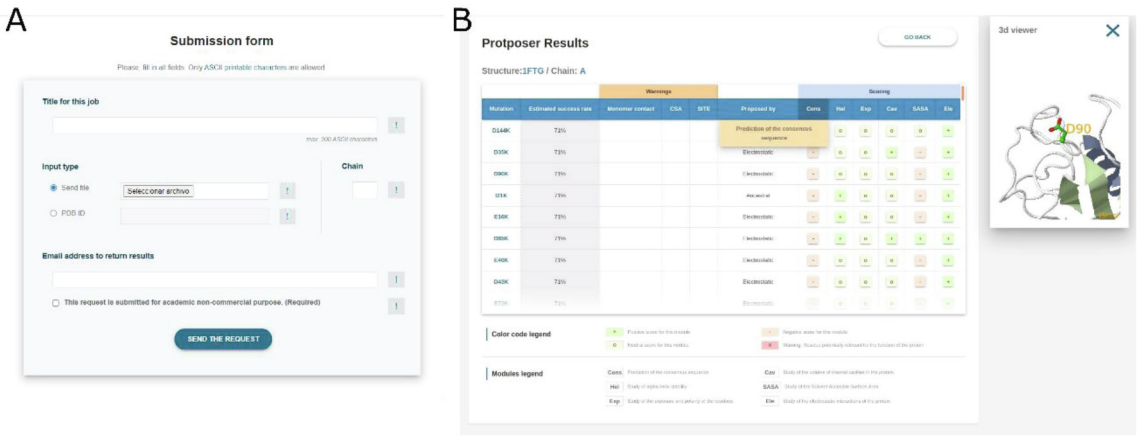

Fig. 7. ***Protposer*** **input and output interfaces.** A) Input screen of the server, with the required fields: the title for the job, the input field to select from uploading a PDB file or indicating a PDB ID, the target chain, the email address to which the results must be returned, and a checkbox to assert the use of ***Protposer*** is non-commercial. B) Output report. It consists of a self-explanatory result table where, by hovering over the titles of the columns, their contents are briefly explained, as shown for the "Cons" column. The scoring results are shown both with a color code and a symbol code to allow for color-blind accessibility. On the right, the 3D visualizer that pops up when a proposed mutation is clicked on shows mutation D90K in its structural context in PDB 1FTG. Legends at the bottom explain the color code used and the modules that appear in the table.

[25,60,101]. Approaches suggested for that include intensive review of existing databases [101] and their extension using results from novel techniques such as deep mutational scanning [60,115]. In this line, in addition to a filtered version of ProTherm, we have used ThermoMutDB [91], a manually curated database for protein mutation thermodynamic data. ThermoMutDB includes some mutations from ProTherm, but it also provides tools for building external datasets. These tools have been useful to derive here a naïve external dataset (ED) that does not overlap with Protherm and has been used to carry out the final test of ***Protposer*** performance.

Two main considerations have made us try in this work a totally different computational approach for protein stabilization. One is, indeed, the current difficulty in accurately calculating the change in ΔΔG associated to a given mutation, which hinders the use of the quantitative results provided by the programs discussed above as a guidance to select mutations for engineering more stable proteins. The other is the realization that a biotechnologically oriented user may have the need to stabilize a protein, but not necessarily the knowledge for even anticipating which mutations should be uploaded to those programs for evaluation. Having this in mind, we have implemented a different approach that aims at being useful for both experienced and non-experienced users, as it does not require previous expertise in the protein stabilization field to obtain the results needed. The method is plainly based in analysing the protein structure of interest in search for opportunities of stabilization by combining and exploiting some of the more successful protein stabilizing strategies discovered by protein engineers over many years. Some of those strategies are rooted in Biophysics (e.g. electrostatic interactions optimization or cavity filling) [42–52] and others are based on sequence analyses and evolutionary interpretation (e.g. return to consensus or ancestral sequences) [53–54]. The method integrates the different strategies considered in a modular nominating algorithm and a logistic regression, machine learning evaluating model (Lr^{dup}) that performs the analysis of the protein structure in a fully automated way. Then, it provides the user with a ranked list of probably stabilizing mutations ordered by their estimated success rates. It additionally provides accompanying information that helps to understand why each of the mutations has been proposed and what makes it to be likely stabilizing, as well as warnings for mutations that may compromise protein functional sites. Selecting the high ranked mutations for engineering into the protein of interest renders a superior performance (higher PPV) to that of current software commonly used for similar purposes.

4.2. Protposer displays an excellent PPV on an external dataset of mutations from 11 proteins

Adhering to concerns in the field about the training problem posed by the low number of stabilizing mutations present in protein stability mutation databases [60,102,115] we have trained our model in two ways. On one hand we have trained it using PT^{ori}, a filtered version of Protherm containing 1641 mutations (179 stabilizing: ΔΔG > 0.5 kcal/mol, and 1462 non-stabilizing: ΔΔG ≤ 0.5 kcal/mol). On the other, we have trained it using a duplicated [60,102], better balanced dataset (PT^{dup}) containing 1208 stabilizing and 2028 non-stabilizing mutations. Besides, our scoring features have been selected conforming to the recommendation [101] that they should analyse the type of residues involved in the mutation and the changes introduced in the surrounding space in order to reduce overfitting issues. As the imbalance between positive and negative cases (i.e., stabilising and non-stabilising mutations) could cause defects and bias in the training and in the evaluation of the models, we have evaluated the model forcing the number of positives and negatives to be the same (see methods). The classification of mutations based on either trained *Lr* model is better than that offered by other mutation evaluating software (Table 3). Among the two models trained, Lr^{dup} performed better than Lr^{ori} towards its training dataset (PT^{dup}) but similarly well when it was tested on PT^{ori} (supplementary figure S5). Additional analyses performed on Lr^{dup} (Supplementary figure S2) indicate it does not show signs of overfitting. ***Protposer***^{dup} has been finally selected for implementation on the server available to users because, when tested on PT^{ori} (Fig. 4, supplementary figure S5), it reaches a slightly higher PPV as well as higher mean experimental ΔΔG values in the proposed mutations than ***Protposer***^{ori}. The final version of ***Protposer*** has been used to propose mutations for the 91 proteins in PT^{ori}. Out of the 7318 mutations proposed, 114 are present in PT^{ori} and can thus be used to estimate the actual PPV of the method. The PPV achieved by ***Protposer*** on PT^{ori} for proposed mutations with eSR above the optimal kappa value is of 61.9 % (Fig. 4, Supplementary Table S5), remarkably high and far above the reported [25,32] PPV of 29 % for FoldX when that server is used

for the search of stabilizing mutations. The PPV values achieved by ***Protposer*** on PT^{ori} (Supplementary Table S5) do not seem to be greatly affected by the exposed or buried position of the mutation, by its occurrence in β or αβ proteins or by the mutation replacing a residue by another of similar or of larger volume. For α proteins, short length proteins and large to small substitutions the number of mutations involved is too low to draw conclusions.

Most importantly, the ***Protposer*** predictive power has also been tested on ED^+, an external dataset consisting of 9 structures selected from ThermoMutDB plus structures 1PGA and 1FTG (Table 4, Fig. 6). The eSRs calculated by ***Protposer*** for each proposed mutation in ED^+ (which, in the output of the server appear reported as "estimated success rate") have been compared with the actual PPVs calculated using the experimental thermodynamic stability data available for those 11 proteins. The agreement between estimated eSR and actual PPV is excellent (Fig. 6B), which suggests the training of the evaluating *Lr* model in ***Protposer*** has not caused overfitting. Therefore, the proposals of mutations in proteins outside the training set will be as successful as described by the estimated eSRs provided by ***Pirepred***. Interestingly, the observed agreement between estimated eSRs and actual PPVs seems independent of the crystallographic resolution of the structure used to propose mutations for a given protein (supplementary table S6). Moreover, as seen for PT^{ori} partitions, the PPV values achieved by ***Protposer*** on ED^+ partitions are similar to those obtained for the entire ED^+ dataset (Supplementary Table S5). Thus, no bias has been found due to structural resolution, residue size change, residue exposure, protein fold or protein size.

Finally, the usefulness of ***Protposer*** as a program readily proposing protein stabilising mutations has been compared to that of current commonly used software such as FoldX [32], Rosetta [106–107] and PoPMuSiC [39–41] using the external database ED^+. Compared with those programs, ***Protposer*** offers a clear improvement (Table 5) in the predictions of stabilizing mutations (defined as ΔΔG > 0.5 kcal/mol). Out of the 11 proteins for which the 4 predictors offer their predictions, ***Protposer*** exhibits the higher individual PPV in 5 proteins. Globally, the PPV for the whole ED^+ offered by the 4 predictors are: 22% (FoldX); 29% (Rossetta); 32% (PoPMuSiC); and 79% (***Protposer***). The superiority of ***Protposer*** is even larger if a mutation is considered stabilising in a non restrictive manner (e.g. ΔΔG > 0.0 kcal/mol) (supplementary table S9) but we believe the biotechnologist should be primarily interested on the prediction of significantly stabilising mutations (i.e. ΔΔG > 0.5 kcal/mol), as summarised in Table 5.

4.3. Reflections for future development

Currently, ***Protposer*** only proposes single point mutations, not being able to propose multiple point mutants exhibiting additive or synergic effects. Some work in the field is being done in the development of predictors (e.g., Fireprot [56]) that evaluate the compatibility and synergies of combined mutations. Testing the usefulness of these approaches is hampered by the lack of large databases reporting thermodynamic data on single mutations as well as their combinations. Such databases would be very useful and they would allow training ***Protposer***-like algorithms to learn and propose combined mutations. From an experimental perspective, the combination of individually stabilising mutations has proved to be very efficient for stabilising proteins by as much as 32 °C [55,116]. A collection of successful application of rational protein engineering for the stabilization of biocatalysts can be found in Bommarius et al. [129] Additionally, the development of databases describing combinations of single point mutations would enable an in depth study of disulfide bonds, which have also proved useful for protein stabilisation [130–131].

At this point, we would like to bring attention to an intrinsic limitation of any method that aspires to design protein stabilizing mutations. Protein engineers tend to assume a two-state equilibrium behaviour for their proteins of interest. Where that is true, the local stabilization impact introduced by a given point mutation will certainly increase the overall stability of the protein, thus having the expected beneficial biotechnological effect. However, as it has been illustrated by theory and experiment [98–99,132], this is not necessarily true for non-two state proteins. In fact, when

Table 5
Comparison of predictive performance in a real case approach with ED^+ between *Protposer* and currently used similar purpose software[a].

Dataset or PDB file	Predictor											
	Random[b]	FoldX	FoldX sel[c]	Rosetta	Rosetta sel[d]	PoPMuSiC	PoPMuSiC sel[e]	*Protposer*^{dup} [f]	*Protposer*$^{dup}_{HM}$ [f]	*Protposer*$^{dup}_{classic}$ [f]	*Protposer*$^{dup}_{Oκ}$ [f]
ED^+ [g]	**14.6%**	**20.9%**	**22.1%**	**25.5%**	**29.0%**	**28.2%**	**32.3%**	**36.1%**	**42.9%**	**56.0%**	**78.6%**
1BPI	**1.3%**	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	NA
1BVC	12.2%	10.0%	10.0%	40.0%	40.0%	20.0%	20.0%	50.0%	50.0%	**66.7%**	NA
1FTG	22.2%	30.0%	30.0%	30.0%	33.3%	30.0%	30.0%	60.0%	60.0%	60.0%	**85.7%**
1LZ1	10.0%	10.0%	10.0%	10.0%	**100.0%**	20.0%	20.0%	33.3%	50.0%	NA	NA
1PGA	18.4%	10.0%	10.0%	30.0%	30.0%	50.0%	50.0%	42.8%	50.0%	50.0%	**100.0%**
1RGG	20.5%	**50.0%**	**50.0%**	20.0%	20.0%	30.0%	33.3%	20.0%	25.0%	25.0%	33.3%
1RX4	5.3%	10.0%	10.0%	0.0%	0.0%	**20.0%**	**20.0%**	0.0%	0.0%	0.0%	NA
1SHF	1.2%	0.0%	0.0%	**20.0%**	**20.0%**	0.0%	0.0%	0.0%	0.0%	0.0%	NA
1TEN	5.3%	10.0%	**20.0%**	10.0%	11.1%	10.0%	0.0%	0.0%	NA	NA	NA
2LZM	7.9%	30.0%	30.0%	30.0%	30.0%	30.0%	30.0%	20.0%	33.3%	100.0%	**100.0%**
2RN2	40.8%	70.0%	70.0%	**100.0%**	**100.0%**	**100.0%**	**100.0%**	60.0%	50.0%	**100.0%**	**100.0%**

[a] The predictive performance is shown as the PPV, considering as actually positive mutations those with a ΔΔG higher than 0.5 kcal/mol for the best 10 mutations of each predictor for each PDB, if >10 proposed mutations available. If there are less than 10 proposed mutations, all proposed mutations are considered. The best result for the whole ED^+ dataset (in bold) or for each individual PDB is underlined and in bold. NA means there were no mutations fulfilling criteria for selection. The number of proposed mutations for each PDB or the total for the ED^+ dataset can be seen in Table S9.
[b] Calculated as the statistics for the whole ED^+ dataset.
[c] Mutations analyzed in FoldX with a predicted stabilizing ΔΔG (at least 0.5 kcal/mol lower ΔG_{fol} for the mutant than for the WT structure after SCWRL processing).
[d] Mutations analyzed in Rosetta with a predicted stabilizing Rosetta Energy Units (more negative score for the mutant than for the WT structure after SCWRL processing).
[e] Mutations analyzed in PoPMuSiC with a predicted stabilizing score (negative score).
[f] Mutations analyzed in ***Protposer***^{dup}. The subscript indicates the selection criterion used: the $_{HM}$ criterion uses an eSR threshold that leaves out half of the proposed mutations (38%), the $_{classic}$ uses an eSR threshold of 50% and the $_{Oκ}$ uses a value to optimize Cohen's kappa coefficient (64,3%).
[g] Calculated as the weighted mean of the PPV for all PDBs, with the number of mutations evaluated as weights.

the goal is to stabilise a three-state (or more-state) protein, the least stable domain should be specifically acted upon [55], because acting on any of the other domains will only stabilise partly unfolded intermediates of the protein likely deprived of biotechnological interest [98]. It is important to tackle and overcome this limitation, and further work to intertwine algorithms for protein stabilization, such as ***Protposer,*** with algorithms predicting unstable subdomains [133] may result in further advances in the field of protein stabilization.

In its final web implementation, **Protposer** constitutes a very efficient mutation classifier that provides an unusual and very convenient functionality lacking in other protein stability predicting software. From a PDB file, ***Protposer*** identifies highly likely stabilizing mutations for biotechnologists and other users who do not need to be trained in structural biophysics to run a query in the server or interpret the resulting predictions. By doing so, ***Protposer*** allows users to focus their efforts on expressing and testing the potentially improved variants. For an average protein domain of 150 residues, ***Protposer*** users will be typically offered 11 mutations for testing, of which 7–8 will increase the stability of the protein by>0.5 kcal/mol each, which exceed the performance currently offered by programs commonly used for the same purpose. The ***Protposer*** server is freely available for academic use at http://webapps.bifi.es/the-protposer.

5. Conclusions

We have designed ***Protposer***, a user-friendly protein mutation proposal and evaluation software of free academic use, destined to facilitate protein stabilisation for biotechnological applications. ***Protposer*** overcomes some limitations of previous stability predictors providing a higher positive predictive value, which allows users to focus on the engineering and testing of just a few mutations exhibiting a high probability of improving the stability of their proteins of interest.

Funding

We acknowledge financial support from MICINN, Spain, (PID2019-107293GB-I00 and PDC2021-121341-I00 grants) and Gobierno de Aragón, Spain, (E45_20R). H. G.-C. is the recipient of an FPU16/04232 doctoral contract from MCINN.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.05.008.

References

[1] Winter G, Fersht AR, Wilkinson AJ, Zoller M, Smith M. Redesigning enzyme structure by site-directed mutagenesis: tyrosyl tRNA synthetase and ATP binding. Nature 1982;299:756–8.
[2] Dalbadie-McFarland G et al. Oligonucleotide-directed mutagenesis as a general and powerful method for studies of protein function. Proc Natl Acad Sci 1982;79.
[3] Sigal IS, Harwood BG, Arentzen R. Thiol-beta-lactamase: replacement of the active-site serine of RTEM beta-lactamase by a cysteine residue. Proc Natl Acad Sci 1982;79. 7157 LP – 7160.
[4] Leisola M, Turunen O. Protein engineering: opportunities and challenges. Appl Microbiol Biotechnol 2007;75:1225–32.
[5] Brannigan JA, Wilkinson AJ. Protein engineering 20 years on. Nat Rev Mol Cell Biol 2002;3:964–70.
[6] Bornscheuer UT et al. Engineering the third wave of biocatalysis. Nature 2012;485:185–94.
[7] Yoo E-H, Lee S-Y. Glucose biosensors: an overview of use in clinical practice. Sensors (Basel) 2010;10:4558–76.
[8] Fan J, Wang M, Wang C, Cao Y. Advances in human chorionic gonadotropin detection technologies: a review. Bioanalysis 2017;9:1509–29.
[9] Fenollar F et al. Evaluation of the panbio COVID-19 rapid antigen detection test device for the screening of patients with COVID-19. J Clin Microbiol 2021;59.
[10] Han X, Li S, Peng Z, Othman AM, Leblanc R. Recent Development of Cardiac Troponin I Detection. ACS Sensors 2016;1:106–14.
[11] Ertürk G, Hedström M, Tümer MA, Denizli A, Mattiasson B. Real-time prostate-specific antigen detection with prostate-specific antigen imprinted capacitive biosensors. Anal Chim Acta 2015;891:120–9.
[12] Klein-Marcuschamer D, Oleskowicz-Popiel P, Simmons BA, Blanch HW. The challenge of enzyme cost in the production of lignocellulosic biofuels. Biotechnol Bioeng 2012;109:1083–7.
[13] Sheldon RA, van Pelt S. Enzyme immobilisation in biocatalysis: why, what and how. Chem Soc Rev 2013;42:6223–35.
[14] Yazbeck DR, Martinez CA, Hu S, Tao J. Challenges in the development of an efficient enzymatic process in the pharmaceutical industry. Tetrahedron Asymmetry 2004;15:2757–63.
[15] Woodley JM. Protein engineering of enzymes for process applications. Curr Opin Chem Biol 2013;17:310–6.
[16] Modarres HP, Mofrad MR, Sanati-Nezhad A. Protein thermostability engineering. RSC Adv 2016;6:115252–70.
[17] Chirino AJ, Mire-Sluis A. Characterizing biological products and assessing comparability following manufacturing changes. Nat Biotechnol 2004;2211 (22):1383–91.
[18] Ding S et al. Protein-based nanomaterials and nanosystems for biomedical applications: A review. Mater Today 2021;43:166–84.
[19] Dong QY et al. Alginate-based and protein-based materials for probiotics encapsulation: a review. Int J Food Sci Technol 2013;48:1339–51.
[20] Ansari SA, Husain Q. Potential applications of enzymes immobilized on/in nano materials: A review. Biotechnol Adv 2012;30:512–23.
[21] Borrebaeck CAK. Antibodies in diagnostics – from immunoassays to protein chips. Immunol Today 2000;21:379–82.
[22] Iqbal H et al. Serum protein-based nanoparticles for cancer diagnosis and treatment. J Control Release 2021;329:997–1022.
[23] Schirrmacher V. From chemotherapy to biological therapy: A review of novel concepts to reduce the side effects of systemic cancer treatment (Review). Int J Oncol 2019;54:407–19.
[24] Kesik-Brodacka M. Progress in biopharmaceutical development. Biotechnol Appl Biochem 2018;65:306–22.
[25] Buß O, Rudat J, Ochsenreither K. FoldX as Protein Engineering Tool: Better Than Random Based Approaches? Comput Struct Biotechnol J 2018;16:25–33.
[26] Huang L-T, Lai L-F, Wu C-C. A Fuzzy Query Method Based on Human-Readable Rules for Predicting Protein Stability Changes. Open Struct Biol J 2009;3:143–8.
[27] Pires DEV, Ascher DB, Blundell TL. mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics 2013;30:335–42.
[28] Frappier V, Chartier M, Najmanovich RJ. ENCoM server: exploring protein conformational space and the effect of mutations on protein function and stability. Nucleic Acids Res 2015;43:W395–400.
[29] Khan S, Vihinen M. Performance of protein stability predictors. Hum Mutat 2010;31:675–84.
[30] Capriotti E, Fariselli P, Casadio R. I-Mutant2. 0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res 2005;33:W306–10.
[31] Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 2002;11:2714–26.
[32] Schymkowitz J et al. The FoldX web server: an online force field. Nucleic Acids Res 2005;33:W382–8.
[33] Potapov V, Cohen M, Schreiber G. Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. Protein Eng Des Sel 2009;22:553–60.
[34] Usmanova DR et al. Self-consistency test reveals systematic bias in programs for prediction change of stability upon mutation. Bioinformatics 2018;34:3653–8.
[35] Pucci F, Schwersensky M, Rooman M. Artificial intelligence challenges for predicting the impact of mutations on protein stability. Curr Opin Struct Biol 2021;72:161–8.
[36] Li B, Yang YT, Capra JA, Gerstein MB. Predicting changes in protein thermodynamic stability upon point mutation with deep 3D convolutional neural networks. PLoS Comput Biol 2020;16:e1008291.
[37] Caldararu O, Blundell TL, Kepp KP. Three Simple Properties Explain Protein Stability Change upon Mutation. J Chem Inf Model 2021;61: 1981–8.
[38] Semenova L, Rudin C, Parr R. A study in Rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning. arXiv Prepr 2019, 1908,:01755.

[39] Dehouck Y et al. Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. Bioinformatics 2009;25:2537–43.
[40] Dehouck Y, Kwasigroch JM, Gilis D, Rooman M. PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. BMC Bioinf 2011;12:151.
[41] Gonnelli G, Rooman M, Dehouck Y. Structure-based mutant stability predictions on proteins of unknown structure. J Biotechnol 2012;161:287–93.
[42] Serrano L, Sancho J, Hirshberg M, Fersht AR. Alpha-Helix stability in proteins. I. Empirical correlations concerning substitution of side-chains at the N and C-caps and the replacement of alanine by glycine or serine at solvent-exposed surfaces. J Mol Biol 1992;227:544–59.
[43] Munoz V, Serrano L. Elucidating the Folding Problem of Helical Peptides Using Empirical Parameters. Nat Struct Biol 1994;1:399–409.
[44] Muñoz V, Serrano L. Elucidating the folding problem of helical peptides using empirical parameters. II. Helix macrodipole effects and rational modification of the helical content of natural peptides. J Mol Biol 1995;245:275–96.
[45] Muñoz V, Serrano L. Elucidating the folding problem of helical peptides using empirical parameters. III. Temperature and pH dependence. J Mol Biol 1995;245:297–308.
[46] Fernández-Recio J, Sancho J. Intrahelical side chain interactions in alpha-helices: Poor correlation between energetics and frequency. FEBS Lett 1998;429:99–103.
[47] Bueno M, Campos L, a, Estrada, J. & Sancho, J.. Energetics of aliphatic deletions in protein cores. Protein Sci 2006;15:1858–72.
[48] Bueno M, Cremades N, Neira JL, Sancho J. Filling Small, Empty Protein Cavities: Structural and Energetic Consequences. J Mol Biol 2006;358:701–12.
[49] Estrada J, Echenique P, Sancho J. Predicting stabilizing mutations in proteins using Poisson-Boltzmann based models: study of unfolded state ensemble models and development of a successful binary classifier based on residue interaction energies. PCCP 2015;17:31044–54.
[50] Strub C et al. Mutation of exposed hydrophobic amino acids to arginine to increase protein stability. BMC Biochem 2004;5:9.
[51] Ayuso-Tejedor S, Abián O, Sancho J. Underexposed polar residues and protein stabilization. Protein Eng Des Sel 2011;24:171–7.
[52] Irun MP, Maldonado S, Sancho J. Stabilization of apoflavodoxin by replacing hydrogen-bonded charged Asp or Glu residues by the neutral isosteric Asn or Gln. Protein Eng 2001;14:173–81.
[53] Sternke M, Tripp KW, Barrick D. Consensus sequence design as a general strategy to create hyperstable, biologically active proteins. Proc Natl Acad Sci U S A 2019;166:11275–84.
[54] Merkl R, Sterner R. Ancestral protein reconstruction: Techniques and applications. Biol Chem 2016;397:1–21.
[55] Lamazares E, Clemente I, Bueno M, Velázquez-Campoy A, Sancho J. Rational stabilization of complex proteins: a divide and combine approach. Sci Rep 2015;5:9129.
[56] Musil M et al. FireProt: web server for automated design of thermostable proteins. Nucleic Acids Res 2017;45:W393–9.
[57] Berman H, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. Nat Struct Mol Biol 2003;10:980.
[58] Berman H, Henrick K, Nakamura H, Markley JL. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res 2006;35:D301–3.
[59] wwPDB consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res 2018;47:D520–D528.
[60] Pucci F, Bernaerts K, Kwasigroch JM, Rooman M. Quantification of biases in predictions of protein stability changes upon mutations. Bioinformatics 2018. bty348–bty348.
[61] Fang J. A critical review of five machine learning-based algorithms for predicting protein stability changes upon mutation. Brief Bioinform 2019. https://doi.org/10.1093/bib/bbz071.
[62] Krivov GG, Shapovalov MV, Dunbrack RL. Improved prediction of protein side-chain conformations with SCWRL4. Proteins Struct Funct Bioinforma 2009;77:778–95.
[63] Madden TL, Tatusov RL, Zhang J. Applications of network BLAST server. Methods Enzymol 1996;266:131–41.
[64] Benson DA et al. GenBank. Nucleic Acids Res 2013;41.
[65] Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL. Nucleic Acids Res 1997;25:31–6.
[66] Bateman A. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 2019;47:D506–15.
[67] Wu CH et al. The Protein Information Resource. Nucleic Acids Res 2003;31:345–7.
[68] Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006;22:1658–9.
[69] Guindon S et al. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Syst Biol 2010;59:307–21.
[70] Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. Bioinformatics 1997;13:555–6.
[71] Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 1992;8:275–82.
[72] Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinf 2004;5:113.
[73] Cock PJA et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 2009;25:1422–3.
[74] Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–637.
[75] Leader DP, Milner-White EJ. The structure of the ends of ??-helices in globular proteins: Effect of additional hydrogen bonds and implications for helix formation. Proteins Struct Funct Bioinforma 2011;79:1010–9.
[76] McDonald IK, Thornton JM. Satisfying Hydrogen Bonding Potential in Proteins. J Mol Biol 1994;238:777–93.
[77] Estrada J, Bernadó P, Blackledge M, Sancho J. ProtSA: a web application for calculating sequence specific protein solvent accessibilities in the unfolded ensemble. BMC Bioinf 2009;10:104.
[78] Bernadó P, Blackledge M, Sancho J. Sequence-specific solvent accessibilities of protein residues in unfolded protein ensembles. Biophys J 2006;91:4536–43.
[79] Rother K, Hildebrand PW, Goede A, Gruening B, Preissner R. Voronoia: analyzing packing in protein structures. Nucleic Acids Res 2009;37.
[80] Webb B, Sali A. Protein Structure Modeling with MODELLER. Methods Mol Biol 2017;1654:39–54.
[81] Šali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234:779–815.
[82] Fiser A, Kinh R, Do G, Ali AS. ˙ Modeling of loops in protein structures. Protein Sci 2000;9:1753–73.
[83] Word JM, Lovell SC, Richardson JS, Richardson DC. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. J Mol Biol 1999;285:1735–47.
[84] Li L et al. DelPhi: a comprehensive suite for DelPhi software and associated resources. BMC Biophys 2012;5.
[85] Lindahl E, Hess B, van der Spoel D. GROMACS 3.0: a package for molecular simulation and trajectory analysis. Mol Model Annu 2001, 2001,;78 (7):306–17.
[86] Abraham MJ et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX 2015;1–2:19–25.
[87] Bava KA, Gromiha MM, Uedaira H, Kitajima K, Sarai A. ProTherm, version 4.0: Thermodynamic database for proteins and mutants. Nucleic Acids Res 2004;32.
[88] Kumar MDS et al. ProTherm and ProNIT: thermodynamic databases for proteins and protein–nucleic acid interactions. Nucleic Acids Res 2006;34:204–6.
[89] Caldararu O, Mehra R, Blundell TL, Kepp KP. Systematic investigation of the data set dependency of protein stability predictors. J Chem Inf Model 2020;60:4772–84.
[90] Orengo CA et al. CATH – a hierarchic classification of protein domain structures. Structure 1997;5:1093–109.
[91] Xavier JS et al. ThermoMutDB: a thermodynamic database for missense mutations. Nucleic Acids Res 2021;49:D475–9.
[92] Nisthal A, Wang CY, Ary ML, Mayo SL. Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis. Proc Natl Acad Sci 2019;116. 16367 LP – 16377.
[93] Fernández-Recio J, Romero A, Sancho J. Energetics of a hydrogen bond (charged and neutral) and of a cation-π interaction in apoflavodoxin. J Mol Biol 1999;290:319–30.
[94] Irún MP, Garcia-Mira MM, Sanchez-Ruiz JM, Sancho J. Native hydrogen bonds in a molten globule: The apoflavodoxin thermal intermediate. J Mol Biol 2001;306:877–88.
[95] López-Llano J, Campos LA, Sancho J. Alpha-helix stabilization by alanine relative to glycine: roles of polar and apolar solvent exposures and of backbone entropy. Proteins 2006;64:769–78.
[96] Ayuso-Tejedor S, Nishikori S, Okuno T, Ogura T, Sancho J. FtsH cleavage of non-native conformations of proteins. J Struct Biol 2010;171:117–24.
[97] Bueno M, Ayuso-Tejedor S, Sancho J. Do Proteins with Similar Folds Have Similar Transition State Structures? A Diffuse Transition State of the 169 Residue Apoflavodoxin. J Mol Biol 2006;359:813–24.
[98] Campos LA, Garcia-Mira MM, Godoy-Ruiz R, Sanchez-Ruiz JM, Sancho J. Do proteins always benefit from a stability increase? Relevant and residual stabilisation in a three-state protein by charge optimisation. J Mol Biol 2004;344:223–37.
[99] Campos LA, Bueno M, Lopez-Llano J, Jiménez MÁ, Sancho J. Structure of stable protein folding intermediates by equilibrium φ-analysis: The apoflavodoxin thermal intermediate. J Mol Biol 2004;344:239–55.
[100] Campos LA, Cuesta-López S, López-Llano J, Falo F, Sancho J. A double-deletion method to quantifying incremental binding energies in proteins from experiment: Example of a destabilizing hydrogen bonding pair. Biophys J 2005;88:1311–21.
[101] Yang Y et al. PON-tstab: protein variant stability predictor. Importance of training data quality. Int J Mol Sci 2018;19:1009.
[102] Capriotti E, Fariselli P, Rossi I, Casadio R. A three-state prediction of single point mutations on protein stability changes. BMC Bioinf 2008;9:S6.
[103] Esposito C, Landrum GA, Schneider N, Stiefl N, Riniker S. GHOST: Adjusting the Decision Threshold to Handle Imbalanced Data in Machine Learning. J Chem Inf Model 2021;61:2623–40.
[104] Cohen J. A Coefficient of Agreement for Nominal Scales. Educ Psychol Meas 1960;20:37–46.

[105] Nisthal A, Wang CY, Ary ML, Mayo SL. Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis. Proc Natl Acad Sci U S A 2019;116:16367–77.
[106] Alford RF et al. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. J Chem Theory Comput 2017;13:3031–48.
[107] Park H et al. Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. J Chem Theory Comput 2016;12:6201–12.
[108] Bæk KT, Kepp KP. Data set and fitting dependencies when estimating protein mutant stability: Toward simple, balanced, and interpretable models. J Comput Chem 2022;43:504–18.
[109] Roelofs R et al. A Meta-analysis of overfitting in machine learning. Advances in Neural Information Processing Systems (eds. Wallach, H. et al.) vol. 32. Curran Associates Inc; 2019.
[110] Galano-Frutos JJ, Garciá-Cebollada H, Sancho J. Molecular dynamics simulations for genetic interpretation in protein coding regions: Where we are, where to go and when. Briefings Bioinf 2021;22:3–19.
[111] Pedroso I, Irún MP, Machicado C, Sancho J. Four-State Equilibrium Unfolding of an scFv Antibody Fragment†. Biochemistry 2002;41:9873–84.
[112] Shen MY, Davis FP, Sali A. The optimal size of a globular protein domain: A simple sphere-packing model. Chem Phys Lett 2005;405:224–8.
[113] Sandhya S et al. Length variations amongst protein domain superfamilies and consequences on structure and function. PLoS ONE 2009;4.
[114] Packer MS, Liu DR. Methods for the directed evolution of proteins. Nat Rev Genet 2015;16:379–94.
[115] Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. Nat Methods 2014;11:801–7.
[116] Ahmed S, Manjunath K, Chattopadhyay G, Varadarajan R. Identification of stabilizing point mutations through mutagenesis of destabilized protein libraries. J Biol Chem 2022:101785.
[117] Lane MD, Seelig B. Advances in the directed evolution of proteins. Curr Opin Chem Biol 2014;22:129–36.
[118] Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 2009, 2009,;47(4):1073–81.
[119] Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res 2001;11:863–74.
[120] Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 2003;31:3812–4.
[121] Ng PC, Henikoff S. Predicting the Effects of Amino Acid Substitutions on Protein Function. 2006;7:61–80. https://doi.org/10.1146/annurev.genom.7.080505.115630.
[122] Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics 2015;31:2745–7.
[123] Choi, Y., Craig, J. and Institute, V. A Fast Computation of Pairwise Sequence Alignment Scores Between a Protein and a Set of Single-Locus Variants of Another Protein General Terms. *Proc. ACM Conf. Bioinformatics, Comput. Biol. Biomed. - BCB '12*. doi:10.1145/2382936.
[124] Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS ONE 2012;7.
[125] Adzhubei IA et al. A method and server for predicting damaging missense mutations. Nat Methods 2010;7:248–9.
[126] Kircher M et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 2014;46:310.
[127] Udhaya Kumar S, Bithia R, DT.K., Doss CGP, Zayed H. Mutational landscape of K-Ras substitutions at 12th position-a systematic molecular dynamics approach. J Biomol Struct Dyn 2022;40:1571–85.
[128] Galano-Frutos JJ, Sancho J. Accurate calculation of barnase and SNase folding energetics using short molecular dynamics simulations and an atomistic model of the unfolded ensemble: evaluation of force fields and water models. J Chem Inf Model 2019. https://doi.org/10.1021/ACS.JCIM.9B00430.
[129] Bommarius AS, Paye MF. Stabilizing biocatalysts. Chem Soc Rev 2013;42:6534–65.
[130] Eijsink VGH et al. Rational engineering of enzyme stability. J Biotechnol 2004;113:105–20.
[131] Liu T et al. Enhancing protein stability with extended disulfide bonds. Proc Natl Acad Sci U S A 2016;113:5910–5.
[132] Sancho J et al. The 'relevant' stability of proteins with equilibrium intermediates. Sci World J 2002;2:1209–15.
[133] Angarica VE, Sancho J. Protein dynamics governed by interfaces of high polarity and low packing density. PLoS ONE 2012;7.