Scientific Reports

https://doi.org/10.1038/s41598-026-38805-z

Article in Press

Computational optimization of DEK1 calpain domain solubility through integrated structural modelling and data-driven targeted mutagenesis

Received: 8 September 2025

Accepted: 31 January 2026

Published online: 08 February 2026

Cite this article as: Dabiri M., Levarski Z., Struhárňanská E. *et al.* Computational optimization of DEK1 calpain domain solubility through integrated structural modelling and data-driven targeted mutagenesis. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-38805-z

Mohammad Dabiri, Zdenko Levarski, Eva Struhárňanská, Viktor Demko, Vladimir Beneš, Jan Turňa & Stanislav Stuchlík

Computational optimization of DEK1 calpain domain solubility through integrated structural modelling and data-driven targeted mutagenesis

Mohammad Dabiri[1, *], Zdenko Levarski[1, 2, *], Eva Struharnanska[1], Viktor Demko[3,4], Vladimir Benes[5], Jan Turna[1, 2], Stanislav Stuchlík[1, 2]

1. Department of Molecular Biology, Faculty of Natural Sciences, Comenius University in Bratislava, 842 15 Bratislava, Slovak Republic. Correspondence author: dabiri2@uniba.sk
2. Science Park, Comenius University in Bratislava, 841 04 Bratislava, Slovak Republic.
   Correspondence author: zdenko.levarski@uniba.sk
3. Department of Plant Physiology, Faculty of Natural Sciences, Comenius University in Bratislava, 842 15 Bratislava, Slovak Republic.
4. Plant Science and Biodiversity Centre, Slovak Academy of Sciences, 84523 Bratislava, Slovak Republic.
5. Genomics Core Facility, EMBL Heidelberg, 69117 Heidelberg, Germany.

Abstract

The DEFECTIVE KERNEL 1 (DEK1) protein plays essential functions throughout plant development. DEK1 is a multidomain 240 kDa protein with yet unsolved 3D structure. To facilitate structural and functional studies of DEK1, here we investigate its calpain protease core domain (CysPc) from *Physcomitrium patens*. Using integrated structural modelling we propose targeted mutagenesis of CysPc to enhance its solubility during recombinant protein production. We created a pipeline to predict the topology of the CysPc domain with improved precision, providing a robust framework for further exploration. We evaluated the native and mutant structures by MD simulations, concentrating on several solubility-related parameters. Following these features, we implemented specific single, double, and triple amino acid mutagenesis to select variants with improved solubility. Our method preserves overall structural integrity while reducing aggregation-prone traits. We advocate

for the utilization of optimized data driven method that can effectively traverse the extensive combinatorial space and prioritize mutation sets with the greatest potential for enhancing solubility. This framework provides a logical, data-driven approach to improving protein solubility, particularly beneficial in situations lacking high-resolution structural data.

Keywords

Calpain, Protein solubility, Mutagenesis, Molecular dynamics, Structure prediction, Data driven

Introduction

The high solubility and native structure of proteins is a desired outcome of their production in heterologous hosts, particularly *Escherichia coli* with its well-known drawbacks such as restricted post-translational modification capability, recurrent development of inclusion bodies, and the propensity to generate insoluble or misfolded proteins under overexpression conditions. Low solubility can generate inclusion bodies, which seriously impede structural studies, functional tests, and protein recovery [1]. Clearly showing this issue for eukaryotic and membrane-associated proteins produced in bacterial systems are variations in codon usage, lack of post-translational modifications, and predisposition for misfolding in complex protein domains [2]. In both academic and industrial biotechnology, protein insolubility thus presents a major obstacle that calls for the development of methodical and successful solutions to improve solubility.

Structure determination of eukaryotic calpain proteases has been challenging due to insolubility and aggregation issues. Currently, the structure of mammalian calpains has been solved [3–5]. Calpains however represent diverse family with neither the 3D structure nor the biological role determined for majority of them [6]. The solubility issue is particularly relevant to the study of DEFECTIVE KERNEL1 (DEK1), a large, multi-domain membrane protein. DEK1 is a 240 kDa protein with a 23-

spanning transmembrane domain, a cytosolic linker segment, and a C-terminal calpain protease [7] CysPc. Genetic analyses in cereals and diverse model plant species showed that DEK1 plays essential functions throughout plant development via regulation of meristematic cell divisions and cell fate control [8,9]. In addition, point mutagenesis and genetic complementation studies indicated that the calpain moiety of DEK1 is essential for the protein function and represents the protein's effector [10]. Despite its essential function, the 3D structure of DEK1 has not been solved yet. Consequently, performing mutagenesis studies on CycPc's structure is particularly challenging due to the absence of crystallographic data or any experimentally determined structures in public databases, as this limits the accuracy of structural predictions and the rational design of beneficial mutations.

Traditional methods for addressing this difficulty encompass empirical mutagenesis [11], co-expression with molecular chaperones [2,12], codon optimization [13], and fusion with solubility-enhancing tags [14–16]. Nevertheless, these processes are frequently protracted, laborious, and contingent upon circumstances. In recent years, computational approaches have emerged as valuable tools for targeted mutagenesis aimed at enhancing protein solubility [17]. Still, these processes are often time-consuming, labour-intensive, dependent on circumstance. Recent years have seen computational methods become useful tools for targeted mutagenesis to increase protein solubility [18–21]. By use of structural modelling, molecular dynamics (MD) simulations, and machine learning-generated solubility predictions, these approaches identify alterations enhancing folding stability and reducing aggregation tendency [22–24]. For eukaryotic proteins produced in prokaryotic systems like *E. coli*, computational optimization provides a systematic and effective approach to enhance solubility by targeted mutagenesis which related studies revealed surface patch analysis was used to design rHuEPO variants in *E. coli*, where reducing positively charged patches (e.g., F48D, R150D) enhanced solubility by up to 60%, while increasing them (e.g., E13K) led to significant solubility loss compared to the wild type [25]. Also, it has

been reported that these methodologies applied integrate structural modelling, MD simulations, and solubility prediction methods to systematically diminish aggregation-prone areas [26–28].

Based on recent research regarding the solubility engineering of aggregation-prone eukaryotic proteins and the *de novo* structural modelling framework and MD; we introduce a computational framework designed to enhance the solubility of the DEK1 calpain-like (CysPc) domain from *P. patens* by rational mutagenesis. The *P. patens* has been chosen as this model plant enables to use effective targeted mutagenesis to link protein structure with function [29,30]. Through this manner, we identify surface-exposed residues exhibiting significant aggregation propensity or detrimental solvation patterns. We suggest specific changes that will improve solubility by making the surface less hydrophobic and increasing local flexibility while keeping the structure stable. This method uses tried-and-true methods for improving the solubility of eukaryotic proteins that tend to clump together and provides a way to study the structure and function of DEK1 domains that have been hard to study experimentally because they don't dissolve well enough [31–35]. Here. First we developed a workflow procedure to increase the accuracy of prediction the structure and based on provided reconstructed structure we come up with a computational method using focused mutagenesis to maximize the solubility of DEK1's CysPc domain. We find residues causing low solubility using structural modelling and MD simulations, then we offer mutation candidates expected to increase solubility while preserving domain stability. This work opens the route for functional and structural studies of DEK1 and offers a paradigm for solubility-oriented protein engineering of challenging eukaryotic domains.

Results

Modelling and validation of wild-type and mutated protein structures

We employed various structural assessment techniques to meticulously evaluate all modelled structures, encompassing both native and mutant protein variations (Fig. 1). We additionally contrasted our integrated homology modelling with direct threading modelling. The projected models' TM-score, US-score, QMEANDisCo, ProQ3D, and Z-score suggest that the structure generated by AlphaFold2 provides high-confidence predictions among the structures examined, indicating reliable representation of the target protein’s structure. (Table. 1).

The SAVES v6.1 server's assessment of stereochemical quality revealed that most residues in all structures resided within the favoured and allowed sections of the Ramachandran plot, signifying appropriate backbone geometry. The QMEAN scoring and analysis conducted using the SWISS-MODEL Assessment tools verified that both the normal and mutants’ models exhibited satisfactory overall quality and reliability ratings. Result revealed that average score of QMEAN for all analysed structure with integrated homology modelling was 1.02 which considered as an acceptable score while this score was 1.53 in direct threading method of structure prediction of CysPc domain. Also, the ERRAT analysis corroborated these findings, revealing that all models predicted good overall quality factors, indicating a lack of substantial non-bonded interaction defects. The amalgamation of these validation techniques verifies that all produced protein structures both wild-type and mutated exhibit good structural integrity and are appropriate for further analysis. All tests which carried out with SAVES v6.1 server for integrated homology modelling indicate that the average of overall quality factor of predicted structures was 97.32 and for verified 3D structure reached to 86.53% while these metrics for direct method of structure prediction was 91.33 and 74.26% respectively which significantly increased. Ramachandran plot analyses comparing the wild-type CysPc structure with the predicted model indicate that the applied consensus-based method produced a higher proportion of residues in favoured and allowed regions, reflecting improved stereochemical quality and backbone geometry in the validated models (Fig. 2).

MD simulations were performed on the wild CysPc domain and each individual mutant to thoroughly evaluate the structural and dynamic effects of sequence alterations. For each structure, we extracted a collection of pertinent features: SASA, H-bond, RDF, Rg, Diffusion coefficient, RMSD, RMSF, minDist. We utilized a weighted scoring approach to statistically compare and prioritize individual mutants based on these criteria. The weights for each characteristic were established using statistical study of the link between feature alterations (from wild to single mutant) and their correlation with favourable enhancements in protein solubility. The resultant set of experimentally derived weights is as follows: SASA (1.2), H-bond (1.2), RDF (1.0), Rg (1.1), Diffusion (0.8), RMSD (-1.4), RMSF (-1.4), and minDist (1.0). The weighted differences were normalized to facilitate direct comparison among features with varying scales, and the resultant Weight(x) score (Eq. 1) permitted efficient ranking of all individual mutants. Mutants with the highest composite scores were recognized as the most promising candidates for the subsequent phase of combinatorial mutant type and optimization.

Evaluation of single mutants; structural and dynamic changes

Single-point mutations in the CysPc sequence were predicted via AlphaFold2 and underwent 200 ns MD simulations to assess their structural and dynamic characteristics. The backbone RMSD values of all mutants stabilized post-equilibration and remained analogous to the wild type, indicating that the overall conformation of the CysPc domain was maintained over the simulation time. Analyses of RMSF indicated that mutations in surface-exposed or loop regions generally affected local flexibility, whereas substitutions within core secondary structures typically decreased RMSF. Alterations in SASA were observed according to the properties of the substituted residues; hydrophilic or charged alterations at the surface were associated with increased solvent exposure, which may correlate with enhanced solubility according to computational metrics. The total count of intra-protein hydrogen bonds remained comparable to the wild-type; however, certain mutants displayed novel

surface-exposed hydrogen bonds that could potentially improve solubility which may contribute to local structural reinforcement or influence solubility. The radius of gyration values for all mutants exhibited negligible variation, affirming that domain compactness was preserved. The results as a whole show that most of the individual mutations in the CysPc domain, as shown by AlphaFold2 and suggested by long-timescale MD simulations, do not change the overall structure of the domain. However, they may change how flexible it is and how much solvent it is exposed to, which are important for improving stability and solubility.

Iterative generation and assessment of double and triple mutants

We meticulously developed and evaluated a library of double mutants in the CysPc domain of DEK1 by extensive *in silico* screening. Subsequently, we selected the 25 out of 45 most promising candidates from double mutants based on various MD features associated with solubility. We meticulously selected each double mutant by amalgamating two single-point mutations that had previously suggested enhanced solubility metrics independently. MD simulations of the double mutants suggested a notable decrease in RMSD, indicating reduced backbone fluctuations relative to the wild-type and single mutants relative to the wild-type and single mutants. Concurrently, RMSF analyses indicated less residue-level fluctuations in critical locations, indicating reduced residue-level fluctuations in critical locations, suggesting increased local rigidity. These data aligned with signs of increased solubility, since the mutants seemed more solvent-accessible and exhibited improved hydration. To corroborate these findings, we additionally assessed other structural and solubility-related characteristics.

Certain double mutants exhibited enhancements in various solubility descriptors concurrently. These mutants are excellent subjects for further mutational investigation or experimental validation. In summary, these 25 double mutants exhibited several enhanced solubility indicators and are promising candidates for subsequent triple mutants' selection.

Table 2 presents their ranking and fitness of MD-derived profiles, demonstrating the efficacy of iterative, data-driven mutagenesis techniques in systematically enhancing protein solubility *in silico*.

Identification of top mutant candidates for improved solubility

The identification and creation of triple mutants were executed via a complex computational pipeline that utilized both empirical MD data and sophisticated data driven method, specifically employing a Proximal Policy Optimization (PPO) framework. Initially, a comprehensive dataset was created through *in silico* mutagenesis of the CysPc domain, producing single and double mutants and simulating the dynamics of each variant to derive critical solubility-related descriptors: RMSD, RMSF, SASA, H-bond, Rg, radial distribution RDF, minimum residue distance, and diffusion coefficient. The descriptors were quantitatively assessed for each mutation and normalized against the wild-type protein to create a comprehensive feature collection.

The PPO methodology was subsequently employed to systematically explore the extensive combinatorial space of potential triple mutants. The optimization framework employed a data-driven scoring function that integrated changes across all solubility-relevant descriptors, rewarding additive or synergistic improvements while penalizing destabilizing or non-informative substitutions. The agent's action space was dynamically limited: each time, it only looked at positions and substitutions that had been shown to improve solubility in single or double mutants. At the same time, it looked at the wild-type conformation to make sure that no mutations were introduced that would break important native contacts or structural motifs.

For every prospective triple mutant suggested by the PPO agent, an *in silico* MD simulation was conducted, and the resultant structural and solubility metrics were incorporated into the algorithm to enhance its exploration approach. The algorithm's policy was enhanced by integrating knowledge of epistatic effects situations where one mutation's influence is altered by another's presence achieved through a

systematic comparison of the triple mutant's characteristics with those of its individual single and double mutants. This method facilitated the identification of combinations in which the third mutation yielded a genuinely advantageous effect just when associated with double mutant backgrounds, rather than independently (Fig. 3). Furthermore, the wild-type sequence was used at each evaluation stage as a reference to indicate that the proposed triple mutants would not deviate significantly from the protein's original conformation or stability characteristics which all resulted in MUT347 as rank 1 of triple mutant selection. We ran extra MD analyses on the triple mutants that obtained the best scores to make sure they did what they were meant to do and to see how they would change the structure and flexibility of proteins. This process made sure that only the best and most lasting alternatives made it to the final phase of testing. It combined the benefits of data-driven prediction and biological plausibility.

Stability and flexibility of wild-type and mutated protein structures

Over a 200 ns trajectory MD models were run to see how the surface mutations affected the stability and dynamic behaviour of both the wild-type and mutant proteins. RMSD plots (Fig. 4, B) show that the mutant protein (MUT347) equilibrated faster and had consistently lower RMSD values than the wild-type protein. The wild-type structure went up steadily and stopped at about 0.70 Å, but the mutant form stopped moving earlier and stayed at 0.49 Å throughout the exercise. The results indicating reduced backbone fluctuations over the simulation period, suggesting increased structural rigidity, most likely by reducing the changes in shape that come from surface-solvent interactions. We also applied RMSF analysis to check the flexibility of the area (Fig. 4, A). As expected, both protein versions showed similar fluctuation patterns, with most of the flexible areas being found at the N and C-termini. The mutant structure had slightly lower RMSF values in several loop areas, especially between residues 0:20, 50:57, 190:209, and 322:354, which meant that the area was stiffer. These reduced fluctuations may reflect

effective surface remodelling, potentially increasing solvent engagement and better local compactness. Also, persistent fluctuation in active site of both structures still high, suggests that applied mutants have negligible impact on the structure integrity on catalytic area. Additionally, Rg analysis between wild-type and mutated structure supports this matter that mutated structure has slightly enhanced compactness and reduced conformational fluctuations over time of simulation and lower fluctuation (Fig. 4, C). The RMSD and RMSF results support the idea that surface-targeted mutations may contribute to reduced conformational fluctuations in solution, especially when there are a lot of them and depletion-induced aggregation can happen because of short interactions between molecules.

Implications of secondary structure stability and solubility

DSSP was used for secondary structure analysis during the exercise to find out if the changes seen in the dynamic behaviour led to changes in the structure. The time-resolved secondary structure maps (Fig. 5) showed that the overall secondary structure makeup of both the native and mutant forms stayed the same over the 200 ns trajectory. When looked at more closely, it was found that the mutant protein kept its α-helix and β-sheet structure more consistently, while the wild-type protein changed between structured and unstructured states more often, especially in loop-rich areas. This was suggested by the quantitative study of the average secondary structure makeup. The mutant had a slightly higher content of helixes and turns compared to the normal variant, but a significantly lower content of coils and bends. There were also fewer standard deviations across simulation blocks. This means that the structure exhibits slightly more persistent secondary structure elements and compactness, which computationally suggests a potential correlation with solubility. These results are important because they match the decrease in RMSF seen in the mutant. This means that the designed surface mutations made the area less flexible while also making the native secondary structure motifs more persistent. From a

physiological point of view, this drop in surface-exposed dynamic disorder may limit the ability for depletion to cause aggregation, especially when there is a lot going on inside cells or in formulations. The combined study of structural and secondary structures shows that changes made to the surface have a good effect on solubility while keeping the structure's integrity. In addition, minimum distance matrix (mdmat) of last 20 ns for residue-residue contact pattern has a higher proof on this matter that applied mutants has a preserved global contact map and no major rearrangement and misfolding occurred and there isn’t any disrupt domain architecture on internal topology (Fig. 5, B).

Analyses of hydrogen bonding and solvent accessibility area

We examined protein-solvent hydrogen bonds (between structure and surrounding solvent) and intra-protein hydrogen (between residues) bonds and SASA during the simulation (dynamic mode) and in static mode to investigate the structural basis for the enhanced solubility of the wild-type and mutated proteins. The overall count of intra-protein hydrogen bonds was persistently elevated in the mutant structure throughout the trajectory, averaging 274.98 ± 9.6 H-bonds, in contrast to 229.17 ± 7.6 in the wild-type (Fig. 6, A). Also, slight reduction of hydrogen bond between structures and solvent was not statistically significant but higher formation of hydrogen bond between residues (intra-molecular) detected which this matter which may contribute to increased local structural rigidity in mutated structure. Furthermore, static mode of SASA profiles showed that the two protein variants had very different amounts of surface contact which statistically significant. The mutated structure had a higher average SASA (21108 $Å^2$), but the wild-type protein had a lower surface area (17821 $Å^2$), which supports the successful surface design and presents a greater surface area available for interaction with water molecules. These changes may be associated with local stabilization and compactness and potentially relate to the observed RMSD and RMSF trends.

There aren’t any significant differences between SASA dynamic mode of wild-type and mutant in all over trajectories, which where 185.59 ± 1.97 and 183.19 ± 3.7 nm/S^2/N respectively. Notwithstanding, better stable fluctuation and higher accessible area were detected on mutated structure at finals frames. On the other hand, analyses of free energy surface (FES) in basis of RMSD and SASA (protein-surface) a slight better reduction of energy resulted for mutant structure in comparison with wild-type (Fig. 6, B).

A bigger solvent accessible surface area on static mode and stronger intra-molecular hydrogen bonds supports the idea that the mutant has changed into a better shape that makes it easier to dissolve. Higher SASA means that more polar and hydrophilic surface residues are exposed, which lets them interact with water molecules around them in a wider area. Furthermore, analyses of H-bond among wild type and mutated structure revealed no significant difference but along with overall solvation, per-residue H-bond analysis showed that some residues close to the mutation sites were more regularly involved in water interactions in the mutant structure (Fig. 6, A). This localized rise in hydration may work as a buffer against nonspecific inter-protein contacts, making it less likely that proteins will clump together when they are crowded. These results add to the SASA data, which showed a moderate drop in global solvent exposure but a better distribution of polar residues that could be accessed by solvents. In opposition, a significant difference of H-bond discovered in intra residues level which could improve with potential of reduce in aggregation and supports solubility via structure better compactness which Rg analyses also supports this matter. Also, studies of hydrogen bond lifetimes showed no difference between wild-type and mutant. This suggests that the applied mutant has no impact on structure destabilization and flexibility.

Figure 7, Panel A illustrates that alterations in pH result in negligible, limited variations, particularly at mutated sites, with the traces for both the wild-type and Mut347 predominantly overlapping. The principal

flexibility is restricted to the N- and C-terminal peaks, and no new hotspots arise throughout the pH series. In contrast, Panels B and C (NaCl and $(NH_4)_2SO_4$, within a concentration range of 0.2–1.2 mol/L) suggested distinct ionic-strength effects: as salt concentration increases, the overall RMSF envelope diminishes, terminal peaks reduce in size, and loop areas become more uniform, with these enhancements consistently more evident in Mut347. Significantly, at the mutation site and its proximal neighbours in Mut347, the coloured traces at elevated salt concentrations are positioned beneath the wild-type curves, suggesting RMSF profiles under different salt conditions show reduced flexibility at terminal and loop regions for MUT347. These trends are consistent with potential stabilization effects predicted by the simulations. The reduction of terminal mobility (both N and C) is more pronounced in Mut347, resulting in cleaner, lower-amplitude endpoints compared to the wild-type. Although NaCl and $(NH_4)_2SO_4$ exhibit a similar qualitative trend, the reduction in RMSF at the altered residue, in adjacent loops, and at the termini is somewhat more significant in $(NH_4)_2SO_4$ at equivalent doses.

Discussion

The integrated modelling and validation technique employed merging template searches (Dali/COFACTOR, RCSB PDB/UniProt) with various predictors and rigorous quality assessments yielded a dependable structural foundation for molecular dynamics. Dali and COFACTOR are recognized for structure/function transfer from homologs, whilst RCSB PDB and UniProt function as the principal, curated repositories for 3D structures and sequence annotations, respectively [36,37]. Across independent quality metrics (e.g., QMEANDisCo, ProQ3D, Ramachandran/PROCHECK, ERRAT), the AlphaFold2 model consistently scored best, aligning with broad evidence that AlphaFold2 attains near-experimental accuracy for many single-domain proteins. Specifically, QMEANDisCo and ProQ3D are cutting-edge model-quality evaluators, while the SAVES/ERRAT toolbox continues to serve as a benchmark for

stereochemical assessments and the identification of non-bonded interaction anomalies [38,39]. This consensus *de novo* approach is supported by recent benchmarks, which predicted that the combination of models from several predictors can perform better than any one approach, especially for challenging targeted structure especially when homology modelling with single template is insufficient [40,41]. Our findings corroborate this pattern: the consensus-based modelling workflow yielded an estimated 86% confidence for the CysPc domain, whereas the single-model AlphaFold2 prediction reached 74%. This difference reflects the increased robustness gained from integrating multiple independent structural models rather than relying on a single predictive source [42]. Also, in support by previous research, consensus-based modelling approaches often outperform single-model predictions by integrating information from multiple independent structures, thereby increasing robustness and predictive accuracy [43,44].

The molecular dynamics investigations (RMSD, RMSF, Rg, SASA) indicate that the designed surface replacements enhanced global dynamics while maintaining the structural conformation. RMSD indicates the overall deviation from the reference conformation, whereas RMSF identifies flexibility along the sequence; the observed first: accelerated equilibration/lower RMSD and second: attenuated RMSF at termini and loops are characteristic indicators of reduced conformational fluctuations and increased structural rigidity. The minor decrease and refinement of Rg indicate a more compact ensemble, aligning with conventional interpretations of protein compactness in simulations and statistical evaluations of PDB structures [45,46]. Additionally, alterations in SASA and hydrogen-bond counts indicate enhanced internal packing and a restructured hydration shell; increased intra-protein hydrogen bonding, together with slight modifications in protein–solvent hydrogen bonds, is a prevalent mechanism for preserving the native state without altering the fold [47].

At the sequence level, the Arg-containing substitutions near the surface likely enhance both stability and solubility by augmenting local hydrogen-bonding opportunities (including weak yet frequent C–H···O interactions) and redistributing surface charge to promote hydration and diminish aggregation-prone positive regions. The stabilizing function of canonical and weak (C–H···O) hydrogen bonds at interfaces is well established, and Arg-rich motifs can engage in both canonical hydrogen bonding and advantageous CH···O interactions [48–51]. Furthermore, surface electrostatics significantly affect solubility; enhanced negative/polar exposure and a reduction in extensive positive regions are associated with improved soluble expression aligning with SASA/H-bond patterns and the diminished variations in exposed loops [52].

The mutant indicates an elevated number of intramolecular hydrogen bonds without an increase in hydrogen bond lifetimes, indicating a denser yet not necessarily slower network, which may contribute to incremental local reinforcement of structure without excessively restricting dynamics. Comprehensive mutational thermodynamics indicate that backbone and side-chain hydrogen bonding positively influence stability in a context-dependent manner, aligning with our finding that an increased number of internal hydrogen bonds correlates with decreased stability [47].

The MD data indicate a consistent effect of reduced flexibility for the surface-engineered mutant. The accelerated equilibration and consistently reduced RMSD of MUT347 compared to the wild-type suggest a more restricted conformational ensemble, a result anticipated when surface electrostatics are positively altered by mutation. Previous research indicates that targeting solvent-exposed regions and improving surface charge can enhance thermodynamic stability without compromising the fold, aligning with our observations [53–55]. In molecular dynamics, RMSD and RMSF serve as conventional metrics for flexibility and structural fluctuations; a reduction in global RMSD accompanied by a decrease in local RMSF often indicates a more rigid, compact ensemble

[56,57].

Furthermore, RMSF profiles related to implemented surface design indicate that the most significant reductions occur in loops and at the termini, whereas the structured core maintains rigidity specifically where stabilization typically initially emerges. The DSSP analysis verifies that secondary-structure content is maintained or somewhat enhanced (increased helix/turn, decreased coil/bend), suggesting that the mutant's diminished fluctuations reflect better retention of native motifs rather than a modification of fold [58,59].

The subdued pH dependence and pronounced salt-dependent attenuation of RMSF most evident at the N/C-terminal peaks, solvent-exposed loops, and the mutated site align with traditional ionic-strength screening: elevated salt concentrations diminish the Debye length, mitigate long-range charge–charge repulsion, and promote short-range interactions, consequently reducing backbone fluctuations in the most unstable regions [60].In accordance with ion specificity, the marginally enhanced damping observed with $(NH_4)_2SO_4$ compared to NaCl is anticipated for a more kosmotropic salt positioned higher in the Hofmeister series, which facilitates compaction and salting-out at similar ionic strengths [61,62].Notably, Mut347 exhibits superior enhancement (reduced RMSF) with elevated salt concentrations compared to the wild-type particularly at and near the mutated site and suggested a slightly improved global profile; such sequence-dependent advancements are expected when local charge distributions are optimized by screened electrostatics and when substitutions improve interfacial interactions [63,64]. In this context, mutations at residue 21 and residue 342 may enhance stabilization by facilitating stronger hydrogen bonds with adjacent chains, a recognized factor in complex/interface stability, including weak yet frequent CH-O and canonical H-bonds at residue-residue interfaces [65,66].

Although, RDF, Diffusion coefficient and minDist had no substantial changes in all over mutants since the primary aim of mutagenesis process was to enhance solubility while conserving total structure

integrity Consequently, significant alterations in these parameters were not expected. Our integrated computational pipeline, which includes iterative mutagenesis, extensive MD analyses, and data-driven through proximal policy optimization, is much better than traditional methods for structure-based protein engineering and increasing solubility. In most cases, mutagenesis and improving protein solubility use single-point mutations or random library screening. These methods are time-consuming and don't work well for finding complex epistatic interactions between many residues [67,68].

On the other hand, our method uses MD simulations in a systematic way to get high-dimensional structural and solubility descriptors from different single and double mutants' library. This creates a strong data base for the future use of any data-driven techniques. Our implemented input information to find its way through the complex world of triple mutants and focuses on mutation combinations that are most likely to improve solubility in an additive or synergistic way, while also avoiding substitutions that are harmful to structure.

Taken together, the computational analyses support that the change in surface physicochemical properties, along with increased internal hydrogen bonding, may be associated with enhanced solvation and hydration shell formation, potentially contributing to improved solubility and reduced aggregation propensity. These findings indicate that the surface-engineered triple mutation results in a more compact and internally reinforced structure, with a more accessible surface that may diminish the tendency for depletion-induced aggregation. However, the suggested alterations and developed method were discerned via computational modelling; nonetheless, experimental validation is crucial to ascertain their influence on solubility and to extrapolate these approaches to divers' proteins, particularly those with structural solubility issue

Methods

Structure prediction, preparation and validation

The amino acid sequence of the DEK1 calpain-type cysteine protease domain was retrieved from the NCBI protein database (https://www.ncbi.nlm.nih.gov/), specifically from *P. patens* (accession number: XP_024400265.1). This sequence corresponds to the predicted CysPc domain of DEK1, which plays a critical role in the protein's proteolytic activity. The sequence was used as the basis for homology modelling and subsequent structural analyses. Prior to structure prediction, the sequence was analysed to identify domain boundaries, surface residues, and potential disordered regions using standard bioinformatics tools. For finding reference structure to carry out homology modelling, threading (fold recognition) model of CysPc domain which was predicted by C-I-ITASSER server (https://zhanggroup.org/C-I-TASSER) [69], and then applied in different blast engine on uniport, (against targeted data base of UniProtKB with 3D structure) (https://www.uniprot.org/blast), RCSB PDB structure similarity search (https://www.rcsb.org/search/advanced), Dali server (http://ekhidna2.biocenter.helsinki.fi) [37] and COFACTOR (https://zhanggroup.org/COFACTOR) [36,70]. All search engine resulted in structure of mu-like calpain (PDB: 1QXP) and related structure retrieved from protein data bank (RCSB). Because experimental structure of 1QXP has 1571 modelled residue count and deposited structure had 1800 residues, closest structure of 1QXP retrieved from RCSB protein data bank according to the result of structural search engines (1KEU, 2P0R, and 6BDT)) and Modeller 10.7 with multimode prediction strategy applied to predict complete structure with reference structures to reaching to the best model possible as reference structure [71,72]. The 3D structure of CysPc domain was modelled using AlphaFold2 [42,73], SWISS-MODEL (http://swissmodel.expasy.org) [74], and I-TASSER (https://zhanggroup.org/I-TASSER/) [75] based on homology modelling to the experimental crystal structure of mu-like calpain (PDB: 1QXP) , ensuring high coverage and confidence in the predicted conformation.

The resulting model was then refined by Galaxy Refine for structure refinement (https://galaxy.seoklab.org) [76] and ModLoop (https://modbase.compbio.ucsf.edu/) for loops refinements [77] . Furthermore, the achieved 3D protein structure minimized any analysis about the structural integrity of the protein structure by SAVES v6.1 server (https://saves.mbi.ucla.edu/) for 3D verification, ERRAT analyses [78] and SWISS-MODEL assess (https://swissmodel.expasy.org/assess) for QMEAN per residues errors [74], QMEANDisCo for distance limitation on model quality estimation [39], TM-score (with superimposed predicted model against each other) and US-Score (predicted model against 1QXP) to assess the similarity structure of predicted model [79,80], Z-Score for overall model quality with ProSA-web [81], ProQ3D for improved model quality assessment (https://proq3.bioinfo.se/pred/) [82], and also, Molprobity server (http://molprobity.biochem.duke.edu/) applied for Ramachandran analyses [83–85]. All process related to the prediction of structures and preparing structural library are presented in Fig. 2.

MD simulation and trajectory condition design

The predicted 3D conformation of the protein was generated for MD simulation utilizing the CHARMM-GUI server (https://www.charmm-gui.org) [86]. The system was parameterized using the CHARMM36m force field and transformed into GROMACS-compatible input files [87]. Solvation was conducted utilizing the TIP3P water model [88], and the system was neutralized with suitable counterions. Each system was solubilized in a cubic simulation box maintaining a minimum solute box distance of 1.0 nm, thereafter, neutralized with counterions and adjusted to physiological ionic strength. Periodic boundary conditions (PBC) were applied in all three dimensions to mimic an infinite aqueous environment and prevent artefacts arising from finite box boundaries [89,90]. The Particle Mesh Ewald (PME) approach was employed for long-range electrostatics [91], while van der Waals and Coulomb interactions utilized a conventional cutoff scheme, generally ranging from 1.0 to 1.2 nm. The Verlet system was employed to update neighbour listings [92]. The temperature was regulated with a V-rescale thermostat at 300 K, while

the pressure was managed with a Parrinello–Rahman barostat at 1 bar [93]. Energy minimization was followed by NVT and NPT equilibration phases, after which a 200 ns production MD simulation was performed for each variant. The configured system was subsequently employed for MD simulations in GROMACS 2022.1 [94], with production runs conducted for 200 ns, and 100M steps according to the duration of selected simulation and ensuring stabilization of soluble-related parameters and dependable time averaged data across root means square deviation (RMSD) and root means square fluctuation (RMSF) [94,95]. The simulations were conducted to examine the solubility behaviour and structural dynamics of the protein under physiological settings. In addition, MD simulation carried out in ionic conditions, sodium chloride (NaCl) as physiological ionic strength agent and ammonium sulfate ($(NH_4)_2SO_4$) as protein participant agent in concentration 0.2, 0.6, 0.9 and 1.2 mol/L by manual setup of input files [96–99]. Also, for different pH condition behaviour, PDB files were adjusted with APBS-PDB2PQR server (https://pdb2pqr.readthedocs.io) [100] in range of 4, 5.5, 7 and 9. All trajectory-based studies, encompassing RMSD, RMSF, SASA, radius of gyration, hydrogen-bond quantification, and diffusion coefficient estimates, were executed utilizing standard GROMACS analytic tools (gmx rms, gmx rmsf, gmx sasa, gmx gyrate, gmx hbond, and gmx msd, respectively).

Assessment and scoring of single-point mutants for solubility improvement

Following the initial 200 ns MD simulation, residue-level analyses were conducted to identify structurally unstable or flexible regions that could serve as potential mutant targets. Important metrics included RMSF to detect residues with high positional variability, solvent accessible surface area (SASA) to identify exposed areas that may affect solubility, and hydrogen bond (H-bond) occupancy and secondary structure stability analyses to determine structural integrity throughout the simulation. Also, ensemble-based technique was utilized for RMSF evaluations across several trajectories or conformations, identifying residues with

consistently heightened fluctuations, increased solvent exposure, or unstable secondary structural configurations as mutation hotspots [101]. The substitution of residues aimed to generate a novel mutated sequence by replacing residues characterized by high flexibility and hydrophobicity with those of the same category that exhibit greater similarity and enhanced hydrophilicity [54]. In table 1, all applied mutations with replaced residues are presented. To preserve functional activity, all selected residues were visually examined using UCSF ChimeraX version 1.8 software, and only those situated on the protein's surface and distanced from the active site were selected for mutation. After 28 potential mutants were chosen based on RMSF, SASA, H-bond analyses, and the Ramachandran plot, they were all run through CamSol (https://www-cohsoftware.ch.cam.ac.uk/) [102], a server that calculates an intrinsic solubility profile and solubility gain upon mutagenesis, and Protein-Sol (https://protein-sol.manchester.ac.uk/) [103], a tool that estimates overall solubility propensity based on sequence and structural features. For re-prioritization structure and MD simulation, the top 10 highest scoring candidate models were chosen from all that were analysed and subsequent to the identification of the 10 premier mutant candidates, an exhaustive MD study was conducted on all associated protein structures. The evaluation encompassed RMSD, RMSF, SASA, numbers of H-bond, diffusion coefficient, radial distribution function (RDF), radius of gyration (Rg), and minimum distance (minDist) measurements. Finally assigning secondary structure to the residues (DSSP) [94], structural alignment with RMSD score [104], and minimum distance matrix (mdmat) [105] carried out in order to final check and comparison of mutated structures against wild structure has not any structural modification. Also, A 2D free energy surface (FES) was generated using RMSD and SASA as collective variables. The probability density was calculated from the MD trajectory and converted to free energy using gmx sham toolbox. The resulting RMSD–SASA landscape was used to identify dominant conformational states and their relative stabilities. on each cycle of mutant selection to library, structures with more than 5% changes in residue counts removed

from selection and only structures with higher confidence than 95% with standard deviation of SE: ±0.05, and RMSD score less than 1.5 Å (highly similar structure against wild-type) advanced to the final round of selection and sorted in library according to the DSSP analyses results. In addition, all selected structure of final library, computationally sequence alignment scored and visually matched against the reference structure (wild-type). Ultimately, top scored triple mutants analysed in aspect of H-bond [106] (protein-solvent and intra-protein state), RMSD, RMSF, Rg and SASA (static mode and dynamic mode) [107].

Synergistic mutagenesis design; identifying optimal mutant's combinational state

We utilized a systematic combinatorial design to uncover optimal synergistic mutation combinations that enhance protein solubility, focusing on the top 10 single-point mutants previously selected by solubility-related MD characteristics. All conceivable double and triple combinations were selected and generated as described in follow, and for each, short MD simulations in 200 ns were conducted to derive solubility-relevant descriptors, including SASA, number of hydrogen bonds, radial distribution function (RDF), radius of gyration (Rg), diffusion coefficient, RMSD, RMSF, and minimum distance to solvent (minDist). in next, top 25 promising combination according to the complementary effects on solubility selected and finally a library consists of wild-type, single and combined (double and triple combination) mutants generated for next phase of analyses to find the best possible combination.

Each combination was assessed utilizing a weighted solubility scoring algorithm, with weights allocated to each attribute according to their established or presumed impact on solubility. For each mutation, we calculated eight MD-derived descriptors □□.

RMSD, RMSF, SASA, Rg, hydrogen bond count, radial distribution function, diffusion coefficient, minimum distance. Descriptor values were normalized using z-scores across the entire mutant cohort. We delineated the weighted deviation from native as:

$$(1)\ \text{Weight}(x) = \sum_{i=1}^{8} w_i.(f_{i,\ \text{mutant}} - f_{i,\text{wild}})$$

where f_i represents the normalized value of feature i and w_i denotes the associated feature weight. Positive weights were allocated to qualities linked to enhanced solubility (e.g., SASA, H-bond, RDF, Rg, diffusion), whilst negative weights were designated to features related to structural instability (e.g., RMSD, RMSF, minDist). This synergistic strategy aimed to uncover mutation combinations that yield improved solubility effects surpassing those of single-point variants, therefore informing the design of optimal multi-point mutants with superior biophysical features. The ultimate solubility score was standardized to a 0–1 scale utilizing:

$$(2)\ \text{Solubility Score} = \sum_{i=1}^{8} w_i.f_i$$

where f_i represents the normalized value of feature i and w_i denotes the associated feature weight. Positive weights were allocated to qualities linked to enhanced solubility (e.g., SASA, H-bond, RDF, Rg, diffusion), whilst negative weights were designated to features related to structural instability (e.g., RMSD, RMSF, minDist). This synergistic strategy aimed to uncover mutation combinations that yield improved solubility effects surpassing those of single-point variants, therefore informing the design of optimal multi-point mutants with superior biophysical features. The ultimate solubility score was standardized to a 0–1 scale utilizing:

$$(3)\ \text{Scaled score} = 100 \times \frac{\sum_{i=1}^{8} w_i.f_i - \text{min score}}{\text{max score} - \text{min score}}$$

The solubility score (Eq. 1–2) was designed to quantify the relative improvement in solubility-related properties for each mutant compared to the natural type. It aggregates eight descriptors and minimum distance to solvent, using a weighted sum of their normalized differences. The weighting coefficients (w_i) were determined during the validation step by analyzing the correlation between each descriptor and solubility trends observed in the single-point mutant dataset. Features with stronger predictive influence were assigned higher weights, while less impactful descriptors received proportionally lower weights. All

weights were normalized so that $\sum w_i = 1$, ensuring comparability and preventing scale bias.

A optimization technique called Proximal Policy Optimization (PPO)[108] used to quickly move through the combinatorial space and find the best mutation sets. To get the highest solubility score, PPO agent was taught to pick mutation combinations that work well together based on input from simulated MD-derived attributes.

The PPO implementation was designed as an actor-critic framework with a feed-forward neural network comprising two hidden layers (128 neurons each, ReLU activation) for both policy and value networks. The input features were normalized MD-derived descriptors for each candidate mutant set and minimum distance to solvent. The reward function was defined as the scaled solubility score (Eq. 3), encouraging mutation combinations that maximize solubility while maintaining structural stability. PPO hyperparameters included a learning rate of 3×10^{-4}, discount factor $\gamma = 0.99$, clipping parameter $\varepsilon = 0.2$, batch size of 64, and entropy coefficient of 0.01. Training was conducted for 10,000 episodes, and convergence was assessed by monitoring reward stabilization and policy entropy reduction. Epistatic interactions were captured implicitly by representing mutation sets as joint states and optimizing the policy over their combined effects on solubility descriptors. Performance was evaluated by comparing PPO-selected combinations against exhaustive search and random sampling, demonstrating superior solubility scores and structural integrity.

This integrated design approach makes it easier to choose multi-point mutants that are logical, based on data, and have solubility properties that are noticeably better than those possible with single mutations alone. Schematic diagram of applied method is presented in Fig. 3.

References

1. Villaverde, A. & Mar Carrió, M. Protein aggregation in recombinant bacteria: biological role of inclusion bodies. *Biotechnology Letters* **25**, 1385–1395 (2003).
2. Baneyx, F. & Mujacic, M. Recombinant protein folding and misfolding in Escherichia coli. *Nat Biotechnol* **22**, 1399–1408 (2004).
3. Nemova, N. N., Lysenko, L. A. & Kantserova, N. P. Proteases of the calpain family: Structure and functions. *Russ J Dev Biol* **41**, 318–325 (2010).
4. Melloni, E., Salamino, F. & Sparatore, B. The calpain-calpastatin system in mammalian cells: properties and possible functions. *Biochimie* **74**, 217–223 (1992).
5. Suzuki, K., Hata, S., Kawabata, Y. & Sorimachi, H. Structure, Activation, and Biology of Calpain. *Diabetes* **53**, S12–S18 (2004).
6. Zhao, S. *et al.* Massive expansion of the calpain gene family in unicellular eukaryotes. *BMC Evol Biol* **12**, 193 (2012).
7. Johansen, W. *et al.* The DEK1 calpain Linker functions in three-dimensional body patterning in Physcomitrella patens. *Plant Physiol.* pp.00925.2016 (2016) doi:10.1104/pp.16.00925.
8. Johnson, K. L., Faulkner, C., Jeffree, C. E. & Ingram, G. C. The Phytocalpain Defective Kernel 1 Is a Novel *Arabidopsis* Growth Regulator Whose Activity Is Regulated by Proteolytic Processing. *The Plant Cell* **20**, 2619–2630 (2008).
9. Lid, S. E. *et al.* The *defective kernel 1* (*dek1*) gene required for aleurone cell development in the endosperm of maize grains encodes

a membrane protein of the calpain gene superfamily. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 5460–5465 (2002).

10. Demko, V. *et al.* Regulation of developmental gatekeeping and cell fate transition by the calpain protease DEK1 in Physcomitrium patens. *Commun Biol* **7**, 261 (2024).
11. Pantophlet, R., Wilson, I. A. & Burton, D. R. Improved design of an antigen with enhanced specificity for the broadly HIV-neutralizing antibody b12. *Protein Engineering Design and Selection* **17**, 749–758 (2004).
12. De Marco, A., Deuerling, E., Mogk, A., Tomoyasu, T. & Bukau, B. Chaperone-based procedure to increase yields of soluble recombinant proteins produced in E. coli. *BMC Biotechnol* **7**, (2007).
13. Gustafsson, C., Govindarajan, S. & Minshull, J. Codon bias and heterologous protein expression. *Trends in Biotechnology* **22**, 346–353 (2004).
14. Chatterjee, D. K. & Esposito, D. Enhanced soluble protein expression using two new fusion tags. *Protein Expression and Purification* **46**, 122–129 (2006).
15. Esposito, D. & Chatterjee, D. K. Enhancement of soluble protein expression through the use of fusion tags. *Current Opinion in Biotechnology* **17**, 353–358 (2006).
16. Sachdev, D. & Chirgwin, J. M. Properties of Soluble Fusions Between Mammalian Aspartic Proteinases and Bacterial Maltose-Binding Protein. *J Protein Chem* **18**, 127–136 (1999).

17. Agostini, F., Cirillo, D., Bolognesi, B. & Tartaglia, G. G. X-inactivation: quantitative predictions of protein interactions in the Xist network. *Nucleic Acids Research* **41**, e31–e31 (2013).
18. Kulshreshtha, S., Chaudhary, V., Goswami, G. K. & Mathur, N. Computational approaches for predicting mutant protein stability. *J Comput Aided Mol Des* **30**, 401–412 (2016).
19. Damborsky, J. & Brezovsky, J. Computational tools for designing and engineering enzymes. *Current Opinion in Chemical Biology* **19**, 8–16 (2014).
20. Ebert, M. C. & Pelletier, J. N. Computational tools for enzyme improvement: why everyone can – and should – use them. *Current Opinion in Chemical Biology* **37**, 89–96 (2017).
21. Broom, A., Jacobi, Z., Trainor, K. & Meiering, E. M. Computational tools help improve protein stability but with a solubility tradeoff. *Journal of Biological Chemistry* **292**, 14349–14361 (2017).
22. Childers, M. C. & Daggett, V. Insights from molecular dynamics simulations for computational protein design. *Mol. Syst. Des. Eng.* **2**, 9–33 (2017).
23. Rouhani, M., Khodabakhsh, F., Norouzian, D., Cohan, R. A. & Valizadeh, V. Molecular dynamics simulation for rational protein engineering: Present and future prospectus. *Journal of Molecular Graphics and Modelling* **84**, 43–53 (2018).
24. Pikkemaat, M. G., Linssen, A. B. M., Berendsen, H. J. C. & Janssen, D. B. Molecular dynamics simulations as a tool for improving protein

stability. *Protein Engineering, Design and Selection* **15**, 185–192 (2002).

25. Carballo-Amador, M. A., McKenzie, E. A., Dickson, A. J. & Warwicker, J. Surface patches on recombinant erythropoietin predict protein solubility: engineering proteins to minimise aggregation. *BMC Biotechnol* **19**, (2019).
26. Kumar, S., Kumar Bhardwaj, V., Singh, R. & Purohit, R. Explicit-solvent molecular dynamics simulations revealed conformational regain and aggregation inhibition of I113T SOD1 by Himalayan bioactive molecules. *Journal of Molecular Liquids* **339**, 116798 (2021).
27. Chennamsetty, N., Voynov, V., Kayser, V., Helk, B. & Trout, B. L. Prediction of Aggregation Prone Regions of Therapeutic Proteins. *J. Phys. Chem. B* **114**, 6614–6624 (2010).
28. Agrawal, N. J. *et al.* Aggregation in Protein-Based Biotherapeutics: Computational Studies and Tools to Identify Aggregation-Prone Regions. *Journal of Pharmaceutical Sciences* **100**, 5081–5095 (2011).
29. Ako, A. E. *et al.* An intragenic mutagenesis strategy in Physcomitrella patens to preserve intron splicing. *Sci Rep* **7**, 5111 (2017).
30. Perroud, P. *et al.* Defective Kernel 1 (DEK 1) is required for three-dimensional growth in *P hyscomitrella patens* . *New Phytologist* **203**, 794–804 (2014).
31. Navarro, S. & Ventura, S. Computational re-design of protein structures to improve solubility. *Expert Opinion on Drug Discovery* **14**, 1077–1088 (2019).

32. Trainor, K., Broom, A. & Meiering, E. M. Exploring the relationships between protein sequence, structure and solubility. *Current Opinion in Structural Biology* **42**, 136–146 (2017).
33. Gupta, J., Nunes, C., Vyas, S. & Jonnalagadda, S. Prediction of Solubility Parameters and Miscibility of Pharmaceutical Compounds by Molecular Dynamics Simulations. *J. Phys. Chem. B* **115**, 2014–2023 (2011).
34. Ganugapati, J. & Akash, S. Multi-template homology based structure prediction and molecular docking studies of protein 'L' of Zaire ebolavirus (EBOV). *Informatics in Medicine Unlocked* **9**, 68–75 (2017).
35. Lu, H., Cheng, Z., Hu, Y. & Tang, L. V. What Can De Novo Protein Design Bring to the Treatment of Hematological Disorders? *Biology* **12**, 166 (2023).
36. Roy, A., Yang, J. & Zhang, Y. COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Research* **40**, W471–W477 (2012).
37. Holm, L., Laiho, A., Törönen, P. & Salgado, M. DALI shines a light on remote homologs: One hundred discoveries. *Protein Science* **32**, e4519 (2023).
38. Chan, P., Curtis, R. A. & Warwicker, J. Soluble expression of proteins correlates with a lack of positively-charged surface. *Sci Rep* **3**, 3333 (2013).
39. Studer, G. *et al.* QMEANDisCo—distance constraints applied on model quality estimation. *Bioinformatics* **36**, 1765–1771 (2020).

40. Van Den Bedem, H. & Fraser, J. S. Integrative, dynamic structural biology at atomic resolution—it's about time. *Nat Methods* **12**, 307–318 (2015).
41. Heo, L. & Feig, M. High-accuracy protein structures by combining machine-learning with physics-based refinement. *Proteins* **88**, 637–642 (2020).
42. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
43. Alden, K., Veretnik, S. & Bourne, P. E. dConsensus: a tool for displaying domain assignments by multiple structure-based algorithms and for construction of a consensus assignment. *BMC Bioinformatics* **11**, 310 (2010).
44. Batista, P. R. *et al.* Consensus modes, a robust description of protein collective motions from multiple-minima normal mode analysis—application to the HIV-1 protease. *Phys. Chem. Chem. Phys.* **12**, 2850 (2010).
45. Lobanov, M. Yu., Bogatyreva, N. S. & Galzitskaya, O. V. Radius of gyration as an indicator of protein structure compactness. *Mol Biol* **42**, 623–628 (2008).
46. Abouzied, A. S. *et al.* Structural and free energy landscape analysis for the discovery of antiviral compounds targeting the cap-binding domain of influenza polymerase PB2. *Sci Rep* **14**, 25441 (2024).
47. Pace, C. N. *et al.* Contribution of hydrogen bonds to protein stability. *Protein Science* **23**, 652–661 (2014).

48. Jiang, L. & Lai, L. CH···O Hydrogen Bonds at Protein-Protein Interfaces. *Journal of Biological Chemistry* **277**, 37732–37740 (2002).
49. Tsumoto, K. *et al.* Role of Arginine in Protein Refolding, Solubilization, and Purification. *Biotechnol. Prog.* **20**, 1301–1308 (2004).
50. Strub, C. *et al.* Mutation of exposed hydrophobic amino acids to arginine to increase protein stability. *BMC Biochem* **5**, 9 (2004).
51. Warwicker, J., Charonis, S. & Curtis, R. A. Lysine and Arginine Content of Proteins: Computational Analysis Suggests a New Tool for Solubility Design. *Mol. Pharmaceutics* **11**, 294–303 (2014).
52. Kramer, R. M., Shende, V. R., Motl, N., Pace, C. N. & Scholtz, J. M. Toward a Molecular Understanding of Protein Solubility: Increased Negative Surface Charge Correlates with Increased Solubility. *Biophysical Journal* **102**, 1907–1915 (2012).
53. Mills, B. J. & Laurence Chadwick, J. S. Effects of localized interactions and surface properties on stability of protein-based therapeutics. *Journal of Pharmacy and Pharmacology* **70**, 609–624 (2018).
54. Trevino, S. R., Scholtz, J. M. & Pace, C. N. Measuring and Increasing Protein Solubility. *Journal of Pharmaceutical Sciences* **97**, 4155–4166 (2008).
55. Kuhn, A. B. *et al.* Improved Solution-State Properties of Monoclonal Antibodies by Targeted Mutations. *J. Phys. Chem. B* **121**, 10818–10827 (2017).
56. Ghahremanian, S., Rashidi, M. M., Raeisi, K. & Toghraie, D. Molecular dynamics simulation approach for discovering potential inhibitors

against SARS-CoV-2: A structural review. *Journal of Molecular Liquids* **354**, 118901 (2022).

57. Xiao, S. *et al.* Rational modification of protein stability by targeting surface sites leads to complicated results. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 11337–11342 (2013).
58. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
59. Mokmak, W., Chunsrivirot, S., Assawamakin, A., Choowongkomon, K. & Tongsima, S. Molecular dynamics simulations reveal structural instability of human trypsin inhibitor upon D50E and Y54H mutations. *J Mol Model* **19**, 521–528 (2013).
60. Zhou, H.-X. & Pang, X. Electrostatic Interactions in Protein Structure, Folding, Binding, and Condensation. *Chem. Rev.* **118**, 1691–1741 (2018).
61. Gregory, K. P. *et al.* Understanding specific ion effects and the Hofmeister series. *Phys. Chem. Chem. Phys.* **24**, 12682–12718 (2022).
62. Hyde, A. M. *et al.* General Principles and Strategies for Salting-Out Informed by the Hofmeister Series. *Org. Process Res. Dev.* **21**, 1355–1370 (2017).
63. Tadeo, X., López-Méndez, B., Castaño, D., Trigueros, T. & Millet, O. Protein Stabilization and the Hofmeister Effect: The Role of Hydrophobic Solvation. *Biophysical Journal* **97**, 2595–2603 (2009).

64. Tadeo, X., Pons, M. & Millet, O. Influence of the Hofmeister Anions on Protein Stability As Studied by Thermal Denaturation and Chemical Shift Perturbation. *Biochemistry* **46**, 917–923 (2007).
65. Sammond, D. W. *et al.* Structure-based Protocol for Identifying Mutations that Enhance Protein–Protein Binding Affinities. *Journal of Molecular Biology* **371**, 1392–1404 (2007).
66. Xu, D., Tsai, C. J. & Nussinov, R. Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein Engineering Design and Selection* **10**, 999–1012 (1997).
67. Goldenzweig, A. & Fleishman, S. J. Principles of Protein Stability and Their Application in Computational Design. *Annu. Rev. Biochem.* **87**, 105–129 (2018).
68. Rosano, G. L. & Ceccarelli, E. A. Recombinant protein expression in Escherichia coli: advances and challenges. *Front. Microbiol.* **5**, (2014).
69. Zheng, W. *et al.* Folding non-homologous proteins by coupling deep-learning contact maps with I-TASSER assembly simulations. *Cell Reports Methods* **1**, 100014 (2021).
70. Zhang, C., Freddolino, L. & Zhang, Y. COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information. *Nucleic Acids Research* **45**, W291–W299 (2017).
71. Šali, A. & Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *Journal of Molecular Biology* **234**, 779–815 (1993).

72. Fiser, A. & Šali, A. Modeller: Generation and Refinement of Homology-Based Protein Structure Models. in *Methods in Enzymology* vol. 374 461–491 (Elsevier, 2003).
73. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat Methods* **19**, 679–682 (2022).
74. Waterhouse, A. M. *et al.* The structure assessment web server: for proteins, complexes and more. *Nucleic Acids Research* **52**, W318–W323 (2024).
75. Zheng, W. *et al.* Deep-learning-based single-domain and multidomain protein structure prediction with D-I-TASSER. *Nat Biotechnol* https://doi.org/10.1038/s41587-025-02654-4 (2025) doi:10.1038/s41587-025-02654-4.
76. Ko, J., Park, H., Heo, L. & Seok, C. GalaxyWEB server for protein structure prediction and refinement. *Nucleic Acids Research* **40**, W294–W297 (2012).
77. Fiser, A. & Sali, A. ModLoop: automated modeling of loops in protein structures. *Bioinformatics* **19**, 2500–2501 (2003).
78. Lüthy, R., Bowie, J. U. & Eisenberg, D. Assessment of protein models with three-dimensional profiles. *Nature* **356**, 83–85 (1992).
79. Zhang, C., Shine, M., Pyle, A. M. & Zhang, Y. US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nat Methods* **19**, 1109–1115 (2022).
80. Xu, J. & Zhang, Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **26**, 889–895 (2010).

81. Wiederstein, M. & Sippl, M. J. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Research* **35**, W407–W410 (2007).
82. Uziela, K., Menéndez Hurtado, D., Shu, N., Wallner, B. & Elofsson, A. ProQ3D: improved model quality assessments using deep learning. *Bioinformatics* **33**, 1578–1580 (2017).
83. Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology* **7**, 95–99 (1963).
84. Lovell, S. C. *et al.* Structure validation by Cα geometry: ϕ,ψ and Cβ deviation. *Proteins* **50**, 437–450 (2003).
85. Williams, C. J. *et al.* MolProbity: More and better reference data for improved all-atom structure validation. *Protein Science* **27**, 293–315 (2018).
86. Brooks, B. R. *et al.* CHARMM: The biomolecular simulation program. *J Comput Chem* **30**, 1545–1614 (2009).
87. Huang, J. *et al.* CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat Methods* **14**, 71–73 (2017).
88. MacKerell, A. D. *et al.* All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **102**, 3586–3616 (1998).
89. Berendsen, H. J. C., Postma, J. P. M., Van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics* **81**, 3684–3690 (1984).

90. Allen, M. P. & Tildesley, D. J. *Computer Simulation of Liquids*. (Oxford University PressOxford, 2017). doi:10.1093/oso/9780198803195.001.0001.

91. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems. *The Journal of Chemical Physics* **98**, 10089–10092 (1993).

92. Chialvo, A. A. & Debenedetti, P. G. On the use of the Verlet neighbor list in molecular dynamics. *Computer Physics Communications* **60**, 215–224 (1990).

93. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics* **52**, 7182–7190 (1981).

94. Bauer, P., Hess, B. & Lindahl, E. GROMACS 2022 Source code. Zenodo https://doi.org/10.5281/ZENODO.6103835 (2022).

95. Lee, J. *et al.* CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. *J. Chem. Theory Comput.* **12**, 405–413 (2016).

96. Galm, L., Amrhein, S. & Hubbuch, J. Predictive approach for protein aggregation: Correlation of protein surface characteristics and conformational flexibility to protein aggregation propensity. *Biotech & Bioengineering* **114**, 1170–1183 (2017).

97. Soares, C. M., Teixeira, V. H. & Baptista, A. M. Protein Structure and Dynamics in Nonaqueous Solvents: Insights from Molecular Dynamics Simulation Studies. *Biophysical Journal* **84**, 1628–1641 (2003).

98. Friedman, R., Nachliel, E. & Gutman, M. Molecular Dynamics of a Protein Surface: Ion-Residues Interactions. *Biophysical Journal* **89**, 768–781 (2005).

99. Zhang, Y. & Cremer, P. Interactions between macromolecules and ions: the Hofmeister series. *Current Opinion in Chemical Biology* **10**, 658–663 (2006).

100. Jurrus, E. *et al.* Improvements to the APBS biomolecular solvation software suite. *Protein Science* **27**, 112–128 (2018).

101. Arantes, P. R., Ligabue-Braun, R. & Pedebos, C. eRMSF: A Python Package for Ensemble-Based RMSF Analysis of Biomolecular Systems. *J. Chem. Inf. Model.* acs.jcim.5c02413 (2025) doi:10.1021/acs.jcim.5c02413.

102. Sormanni, P., Aprile, F. A. & Vendruscolo, M. The CamSol Method of Rational Design of Protein Mutants with Enhanced Solubility. *Journal of Molecular Biology* **427**, 478–490 (2015).

103. Hebditch, M., Carballo-Amador, M. A., Charonis, S., Curtis, R. & Warwicker, J. Protein–Sol: a web tool for predicting protein solubility from sequence. *Bioinformatics* **33**, 3098–3100 (2017).

104. Meng, E. C., Pettersen, E. F., Couch, G. S., Huang, C. C. & Ferrin, T. E. Tools for integrated sequence-structure analysis with UCSF Chimera. *BMC Bioinformatics* **7**, 339 (2006).

105. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).

106. Van Der Spoel, D., Van Maaren, P. J., Larsson, P. & Tîmneanu, N. Thermodynamics of Hydrogen Bonding in Hydrophilic and Hydrophobic Media. *J. Phys. Chem. B* **110**, 4393–4398 (2006).

107. Bondi, A. van der Waals Volumes and Radii. *J. Phys. Chem.* **68**, 441–451 (1964).

108. Schulman, J., Wolski, F., Dhariwal, P., Radford, A. & Klimov, O. Proximal Policy Optimization Algorithms. Preprint at https://doi.org/10.48550/ARXIV.1707.06347 (2017).

Funding

This work is the result of implementation of Slovak Research and Development Agency grants APVV-21-0227, APVV-21-0215, APVV-22-0161, UK/1088/2025 and by implementation of the project 101160008 “Fostering Excellence in Advanced Genomics and Proteomics Research at Comenius University in Bratislava – FORGENOM II” funded by the Horizon Europe program.

Acknowledgments

MD is thankful to Daniel Kráľ and Milan Melicherčík, Faculty of Mathematics, Physics, and Informatics (FMPI) for their kind recommendation and guidance.

Authur Contributions

MD and ZL: conceived and designed the study, MD: developed the structural prediction and mutagenesis workflow, and performed all molecular dynamics simulations and solubility analyses. Data processing, interpretation, and manuscript writing and editing were carried out by MD and ZL. VD and ES: Provided recommendations and scientific guidance of research. All aspects of the research were conducted under

the academic supervision of JT, VD, VB and SS who provided critical feedback on the study design and manuscript.

Competing interests

The authors declare no competing interests.

Data Availability

The data generated and analysed during this study, including molecular dynamics simulation outputs, structural models, and solubility feature datasets and code resources will be made available from the corresponding authors upon reasonable request. Due to the size and computational nature of the datasets, they are not hosted in a public repository. Full mutagenesis dataset and related materials are provided in supplementary information.

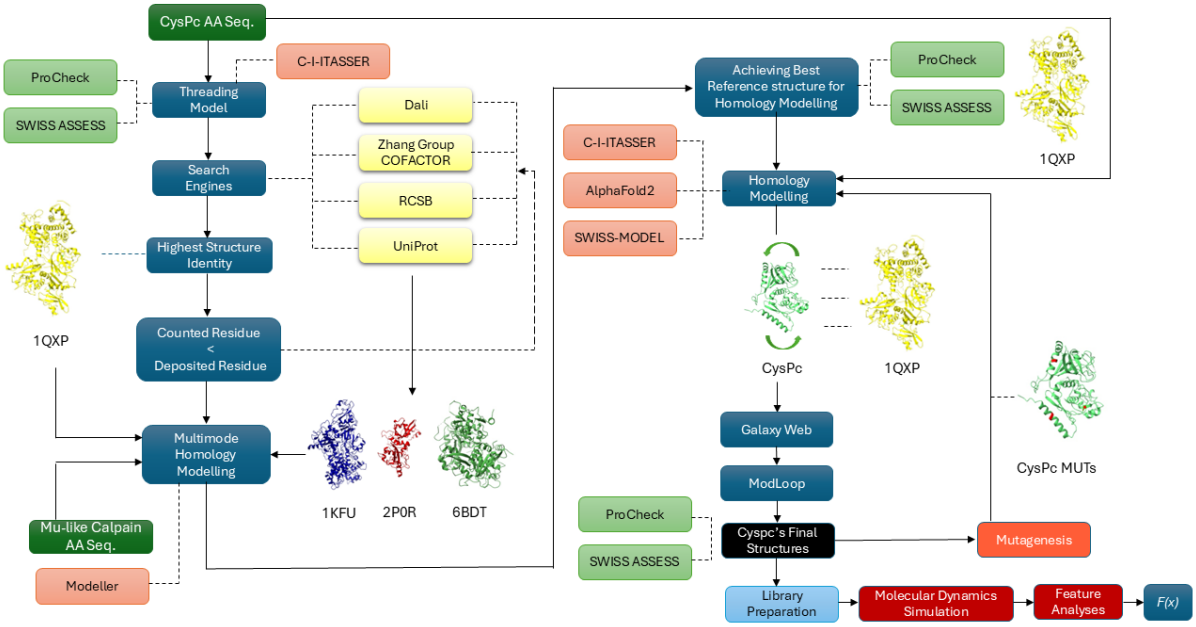


Fig. 1: process related to the finding best reference sequence and integrated homology modelling of CysPc domain and library preparation.

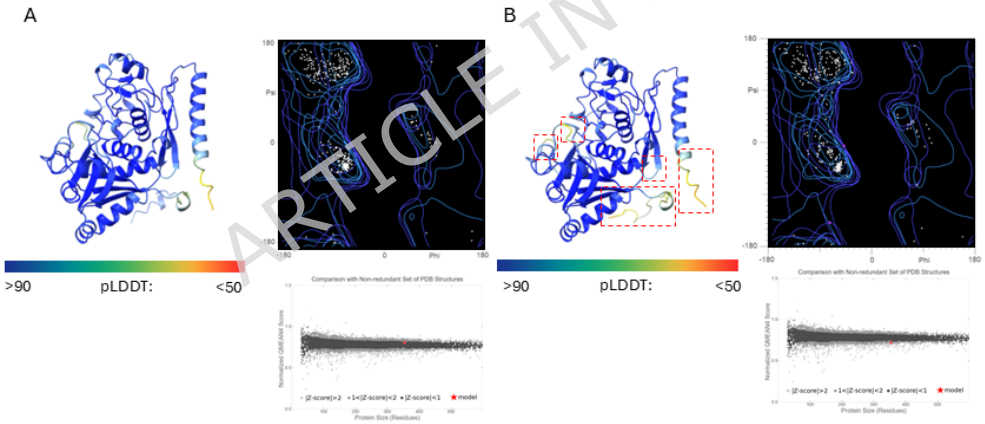


Fig. 2: A: consensus-based integrated modelling approach. B: single-model prediction. These analyses highlight the improved stereochemical quality and model confidence achieved with the consensus-based approach.

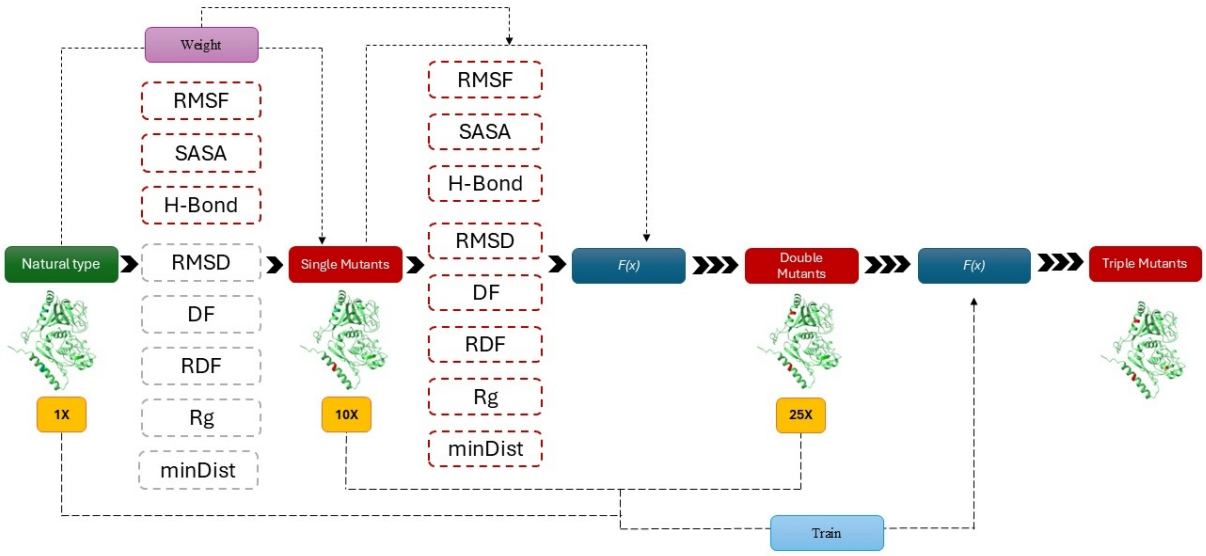


Fig. 3: Workflow for multi-point mutation protein optimization using MD features and iterative training. This diagram illustrates a stepwise approach for optimizing protein solubility (or other properties) via single, double, and triple mutants, guided by MD simulation features and iterative learning.

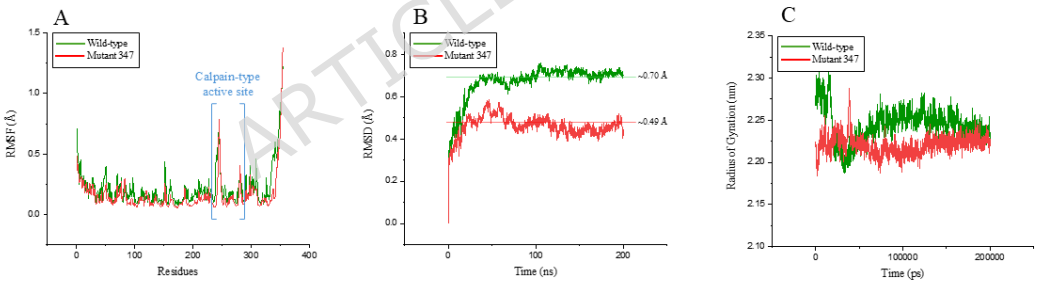


Fig. 4: A: MD simulation RMSF plot per residue for the wild-type (green) and mutant (red) proteins. Most residues show comparable fluctuation patterns in both profiles, indicating local flexibility. Both variants have slightly greater RMSF values at the C-terminal region, with the wild-type showing slightly more flexibility. Some loop areas of the mutated protein have much less variation, which suggests that changes to the surface made the protein more stable in those areas. B: RMSD of backbone atoms over time (ns) for proteins that are normally formed (green) and proteins that have been changed (red). The RMSD of the changed structure stays lower

than that of the wild-type structure after it is stabilized quickly. The smaller RMSD of the changed protein supports the idea that certain surface changes make structures more compact and less likely to change. C: Rg analysis between natural-type and mutated structures reveals better compactness and integrity all over trajectory.

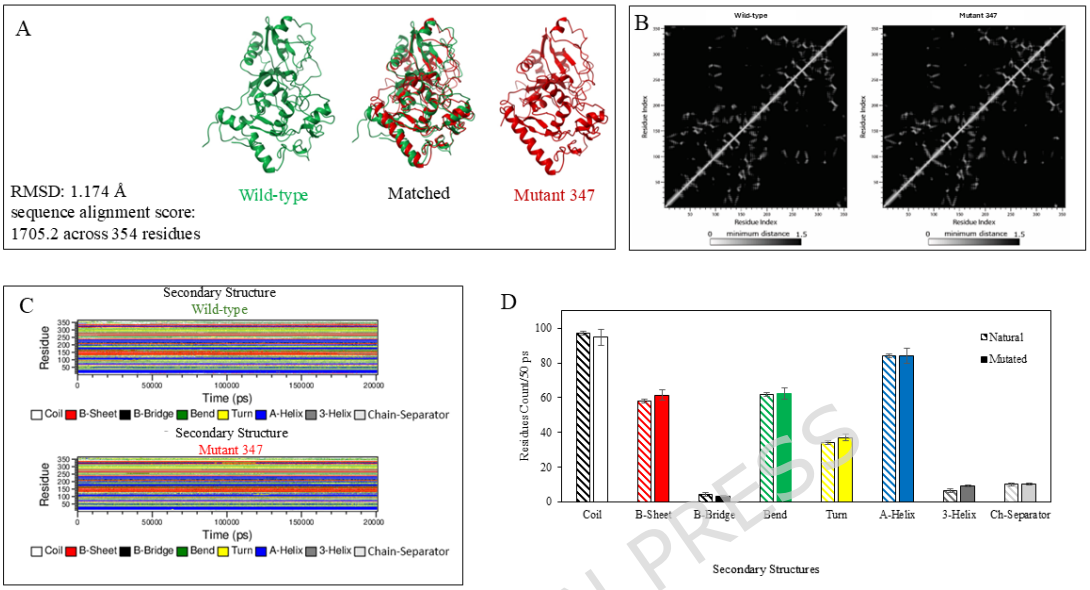


Fig. 5: Comparison of the global folded structures and secondary structures of the native protein and the triple-mutant variant. A: Structure alignment of natural type (green) and triple-mutated (red) type and superimposed structure (middle) of CysPc domain with RMSD of 1.174 Å. B: mdmat analyses of wild-type and MUT347 in residue-residue level. C: time resolved DSSP secondary structures of natural and triple-mutated type in 200 ns time-laps. D: Quantitative analysis of secondary structure composition over blocks of 50 ps; block averaging with SE ± 0.05. Colour coded as follows: coil (white), β-sheet (red), β-bridge (black), bend (green), turn (yellow), α-helix (blue) 3_{10}-helix (dark gray), chain separator (light gray); Natural type (patterned bars), mutated type (solid bars).

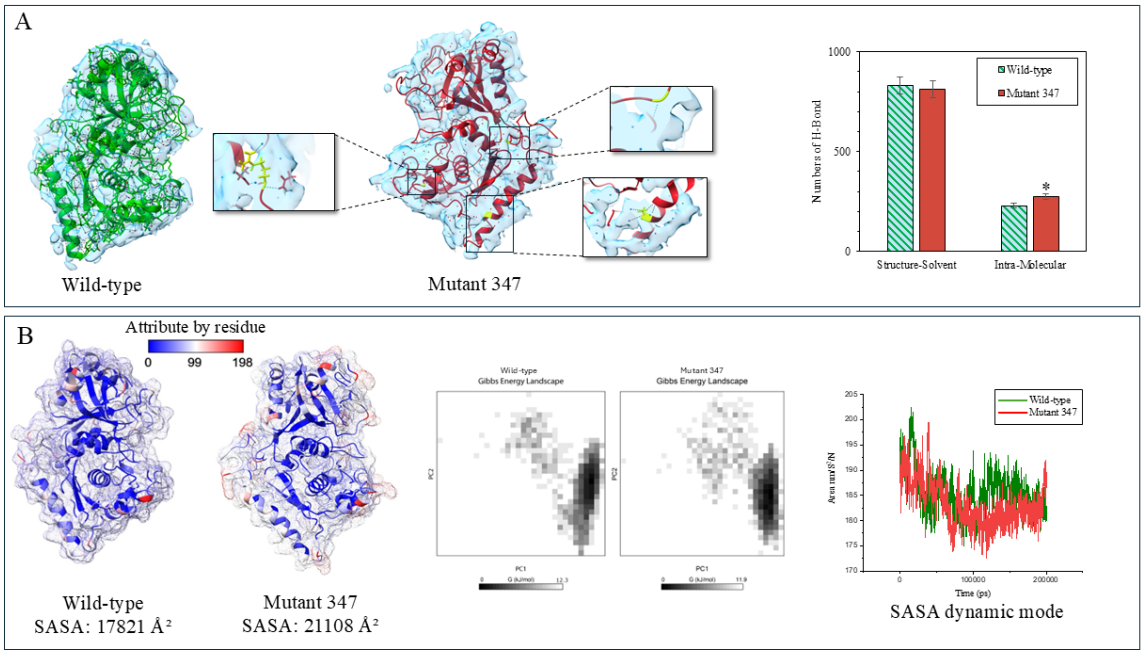
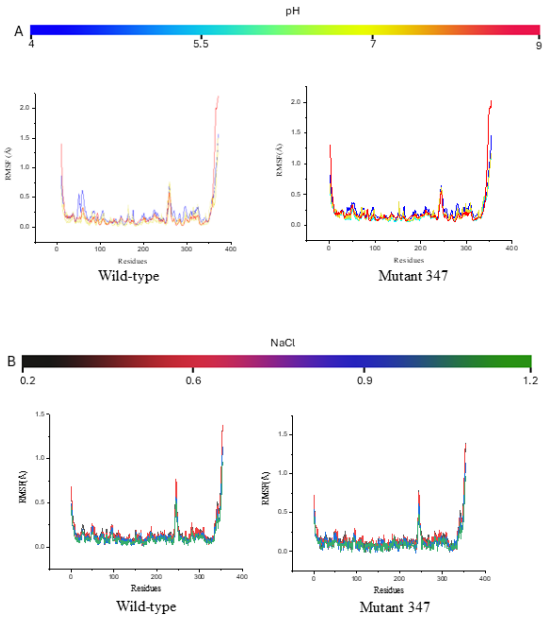


Fig. 6: Structural comparison and surface study relates to solubility of the wild-type and mutant protein. A: Surface and water molecules distribution representations illustrate mutation-induced structural alterations (represent in yellow), with the mutant variant displaying enhanced intra-molecular hydrogen bonding, as assessed in the accompanying bar chart. B: SASA mapping by residue and area indicates that the mutant has a higher static surface exposure (21108 Å²) than the wild-type (17821 Å²). Dynamic SASA analysis, however, reveals that the mutant sustains a slightly more compact conformation but better stability in finals steps. Free energy surface (FES) calculated using RMSD-SASA as reaction coordinates for wild-type and mutant 347.



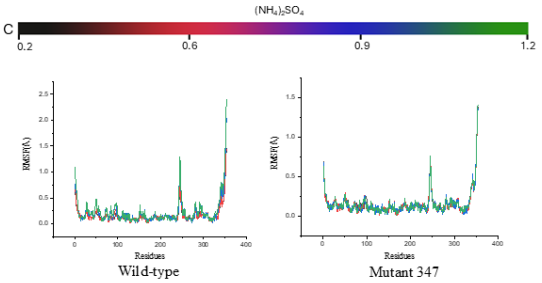


Fig. 7: Effect of pH and ionic strength on backbone flexibility of the wild-type protein and the mutant variant. A: pH series. Per-residue Cα RMSF profiles for wild-type (left) and mutant (right) across four pH conditions spanning acidic to basic. Line colours follow the top colour bar (blue to red = low high pH). Residue index is on the x-axis; RMSF (Å) is on the y-axis. B: NaCl series. RMSF profiles at increasing NaCl concentrations (0.2–1.2 M). Line colours map to the middle colour bar (black to green = low to high salt). Both variants show broadly similar shapes, with only modest salt-dependent dampening of fluctuations in solvent-exposed regions. C: $(NH_4)_2SO_4$ series. RMSF profiles at increasing $(NH_4)_2SO_4$ concentrations (0.2–1.2 M). Colores follow the bottom colour bar (black to green).

Table 1: Benchmarking C-I-TASSER, AlphaFold2, and SWISS-MODEL based on structural accuracy scores.

Modelling software	QMEANDisCo Global	TM-Score	US-Score	Z-Score	ProQ3D
C-I-ITASSER	0.59±0.05	0.70	0.70	-9.64	0.701
Alphafold2	0.69±0.05	0.72	0.73	-6.57	0.713
SWISS-MODEL	0.64±0.05	0.66	0.68	-.8.51	0.688

Table. 2: Candidate residues and position number and replaced residues for single-pointed mutants.

Mutant	Original residue	position	mutated residue	Structural type
MUT1	Valine	17	Arginine	α-Helix
MUT2	Leucine	19	Serine	α-Helix
MUT3	Isoleucine	21	Serine	α-Helix
MUT4	Valine	47	Arginine	Loop
MUT5	Phenylalanine	291	Tyrosine	Loop
MUT6	Leucine	323	Serine	α-Helix
MUT7	Valine	342	Asparagine	Loop
MUT8	Lysine	7	Arginine	α-Helix
MUT9	Glutamine	54	Arginine	α-Helix
MUT10	Lysine	339	Arginine	Loop

Table 3: Fitness-based ranking of protein mutants across mutation levels (description: mutant identifiers follow the format: MUTxyz, where each digit (x, y, z) represents the point of a mutation in the protein sequence. For example, MUT347 indicates mutations includes single mutants of 3, 4, and 7 (as triple mutant).

Rank	Single Mutant	Fitness	Double Mutant	Fitness	Triple Mutant	Fitness
1	MUT3	0.1622	MUT37	0.1432	MUT347	01368
2	MUT7	0.1719	MUT34	0.1443	MUT246	0.1475
3	MUT2	0.1759	MUT67	0.1498	MUT367	0.1498
4	MUT5	0.1775	MUT35	0.1519	MUT345	0.1519
5	MUT4	0.1803	MUT24	0.1696	MUT245	0.1705
6	MUT6	0.1850	MUT25	0.1705	MUT136	0.1776
7	MUT1	0.2163	MUT26	0.1712	MUT247	0.1909
8	MUT8	0.2184	MUT36	0.1741	MUT236	0.1950
9	MUT10	0.2209	MUT45	0.1775	MUT457	0.2101
10	MUT9	0.2226	MUT13	0.1776	MUT146	0.2551