Calcium Binding Affinity in the Mutational Landscape of Troponin-C: Free Energy Calculation, Co-evolution modeling and Machine Learning

Pooja and Pradipta Bandyopadhyay*

School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi 110067, INDIA

E-mail: praban07@gmail.com

Abstract

Computational protein science has made substantial headway, but accurately predicting the functional effects of mutation in Calcium-binding proteins (CBPs) on Ca^{2+} binding affinity proves obscure. The complexity lies in the fact that only sequence features or structural information individually offer an incomplete picture on their own. To triumph over this adversity, we introduce a pioneering framework that effortlessly integrates protein sequence evolution information, structural characteristics, and Ca^{2+} binding interaction properties into a machine learning algorithm. This synthesis has been carried out poised to significantly enhance accuracy and precision in the prediction of the Ca^{2+} binding affinity towards CBP variants. In our study, we have developed a Ca^{2+} binding affinity prediction model for various mutants of cardiac Troponin-C protein, to uncover the molecular determinants that contribute binding affinity across protein variants. Our method combines state-of-the-art

practices, including a physics-based approach that uses relative binding free-energy (BFE) calculations to assess mutations with implicit polarization. Additionally, it incorporates the impact of evolutionary factors on protein mutations through a theoretical deep mutational scan using a statistical probability model. Support Vector Regression (SVR) algorithms have been used to predict Ca^{2+} binding affinity based on sequence information, structural properties, and interactions of water molecules with Ca^{2+} in the EF-hand loop. Our model demonstrates high accuracy and can potentially be generalized for other calcium-binding proteins to predict the effects of point mutations on Ca^{2+} binding affinity for CBPs.

Introduction

Calcium-binding proteins (CBPs) are a diverse group of intracellular signaling molecules characterized by their ability to bind calcium ions through specific domain. These proteins act as calcium sensor inside the cell and mediator of various signaling pathways that regulate cellular processes related to muscle contraction, neuronal signal transduction, and cell proliferation. When calcium ion bind to CBPs such as troponin-C, calmodulin, and paralbumin etc., they undergo conformational changes resulting in the exposure of a hydrophobic patch, which interacts with target protein and cellular process occur.[1] The binding of Ca^{2+} in the conserved helix-loop-helix motif of these proteins, is facilitated by the presence of specific coordinating residues in a pentagonal bipyramidal geometry.[2] A single point mutation in the critical amino acid residue of these proteins disrupts Ca^{2+} binding affinity, affecting its functionality, and may lead to severe diseases including cardiomyopathies, neurodegenerative disorder and cancer.[3] Due to the profound significance of these proteins in regulation of cellular pathways, it has been the subject of thorough investigation, encompassing both experimental studies using Isothermal Titration Calorimetry (ITC), Fluorescence and Surface Plasmon Resonance[4–9] and theoretical[10–16] approaches. But a complete understanding of the mechanism that govern the Ca^{2+} binding properties of CBPs still eludes us. Accurate estimation of Ca^{2+} binding affinity to the different variants (wildtype and mutants) of CBPs is extremely important as alteration of the binding affinity for mutants can have potentially severe consequences. For instance, mutation in cardiac troponin-C (cTn-C) protein changes the Ca^{2+} binding affinity towards protein variant, can lead to cardiac diseases such as hypercardiomyopathy and dilated or restricted cardiomyopathy.[5] Likewise, mutation in calmodulin (CaM) protein affects Ca^{2+} binding affinity and thus affects the regulatory pathways that may cause cardiac arrhythmia, long QT syndrome (LQTS), Alzheimer's disease.[17] However, considering the presence of hundreds of CBPs in mammalian systems and their very large number of mutations, computational methods can be extremely valuable.

Thus, the need of the hour is to develop a significantly accurate and efficient prediction

model to calculate the Ca^{2+} binding affinity using simple features of these protein variants. Understanding the reasons for the change in binding affinities can give avenues to modulation of Ca^{2+} binding and eventually to CBP design. For the calculation of Ca^{2+} binding affinity to CBPs, various structure and sequence-based approach have been used sporadically, but still, there are several technical challenges.[16,18–20] Studies carried out on eleven distinct mutants with substitutions in the hydrophobic core observed that despite the absence of direct interactions between these residues and Ca^{2+}, the hydrophobic residues played a significant role in influencing the Ca^{2+}-binding properties of calbindin.[21] Consistent observations with their research demonstrate that a single mutation in the N-terminal domain disrupts the hydrophobic interactions in calmodulin, and causes conformational shift.[22] Some empirical studies detail how replacing the polar side chain residue glutamine and lysine only at specific positions with non-polar side chains decreases Ca^{2+} binding affinity.[23] Earlier studies carried out by our research group led to the development of a regression model for Ca^{2+} binding free energy calculation, which provides accurate predictions for calmodulin.[24,25] However till now, no generalized Ca^{2+} binding affinity prediction model tailored specifically for CBP variants has incorporated sequence, structural along with some specific features that affect these interactions.

In our current study, we have developed an accurate and streamlined machine learning (ML) prediction model for the calculation of Ca^{2+} binding affinity using essential structure and sequence characteristics in conjunction with Ca^{2+} interaction properties of protein variants. Our model has been built in the following manner. First, an accurate molecular dynamics-based protocol has been used to calculate the binding free energy (BFE) of mutations using the MM-PBSA (Molecular Mechanics Poisson-Boltzmann Surface Area) method. The protocol has been optimized using the experimental binding affinity data for the observed mutants of cardiac troponin-C (cTn-C). These mutants are primarily disease mutations (DM) connected to various heart disorders. In the ensuing step, we have augmented our mutation dataset by generating additional plausible mutations (PM) from the

perspective of sequence evolution of cTn-C protein. The evolutionary mutation prediction method "EVmutation" has been used for this purpose developed by Hopf et al.(2017).[26] Next, we calculated the BFE for those plausible mutations using the same protocol we had used earlier, and Ca^{2+} BFE dataset of all the mutations was collected. Both DM and PM features have been extracted and the relative BFE with wild-type protein of all the 75 mutation dataset was calculated to build the Ca^{2+} binding affinity prediction model. Following the refinement and tailoring process, the most effective ML model was selected based on a rigorous evaluation of the performance metrics of various algorithms. Finally, the Support Vector Machine (SVM) algorithm was used to build the Ca^{2+} binding affinity prediction model using discernible sequence, structure, and Ca^{2+} interaction features extracted from the mutation dataset.

The challenges associated with each step of our investigation have been circumvented in the following way. Empirical evidence supports that non-polarizable force fields in atomistic simulations are not good enough to calculate accurate BFE for Ca^{2+} binding to CBPs.[27,28] Some studies including our lab have found that charge scaling of Ca^{2+} and its coordinating atoms with a scaled factor of 0.75 incorporate polarization in classical MD simulations, increases accuracy in BFE calculations.[25,28,29] To fine-tune and elevate the precision of our calculations, we calibrated the internal dielectric constant of protein in the MM-PBSA BFE calculations to establish a good concordance with the experiment calculations.

Support Vector Regression (SVR) algorithm has been deployed to predict the relative binding free energy ($\Delta\Delta G_{bind}$) of the cTn-C mutation using nine features of the mutation dataset. To determine the best-fitting hyperplane through hyperparameter fine-tuning, we tune the regularization parameter (C), type of kernel, and margin of tolerance (ϵ). Our optimized model with poly kernel gives R^2 of 0.77, Mean Squared Error (MSE) of 0.16 along with pearson's correlation coefficient (R_p) of 0.92. The favorable results of these evaluation metrics highlight the effectiveness of the model. To further explore the contribution of each feature in our model, we conducted feature analysis using three methods, including the

correlation coefficient of each feature, permutation feature importance and SHap(the shapley additive explanations) value. Top three best features were selected that showed the highest contribution to the model, including the number of water molecules coordinating within $3\mathring{A}$ of Ca^{2+}, type of amino acid substitution, and distance of mutation to the Ca^{2+}. Next, for the classification of the mutation dataset, a Support Vector Classifier (SVC) with a linear kernel was utilized to establish a decision boundary between high and low Ca^{2+} binding affinity mutation concerning wild-type cTn-C protein. We gauged the performance of the SVC model gives an accuracy of 0.76 and a receiver operating characteristic (ROC) curve of 0.86, which effectively showcases the predictive power of our model. Our combined ML approach shows a high accuracy.

The present study outlines an integrative computational approach utilizing a physics-based strategy, an evolutionary approach using a statistical probability method, and a support vector machine algorithm, utilizing sequence, structure and Ca^{2+} interaction features, for the prediction of Ca^{2+} binding affinity towards cTn-C variants. As the methodology is general, it can be used for any other CBPs and it is expected to lay the path for an integrative approach to predict and design CBPs with tailor-made Ca^{2+} binding affinity.

Methodology

The ensuing sections succinctly outline the methodology deployed in this study along with its theoretical foundations, and illuminate the protocols employed for the calculations presented herein.

1. Implementation of charge scaling for Ca^{2+} and its coordinating ions in classical Force Field

In the scrutiny of biological interactions in aqueous environment, electronic polarization effects play an important role in the interactions of charged ions such as Ca^{2+} and salts. However, the majority of biomolecular simulations opt for non-polarizable force fields due to the complexities and computational constraints of explicitly including electronic polarization. But in the presence of polyvalent ions, non-polarizable force fields often lead to a pronounced overestimation of ion pairing, potentially distorting simulation outcomes.[27,30]

A little more than one decade back Stuchebrukhov, Leontyev, Marshall, and co-workers came up with a model, where the electronic part of the dielectric constant of the medium is modeled by a simple charge scaling procedure.[27,31] This method, within its approximations, has been used for both isolated ions in water and protein in water. The amount of scaling depends on the electronic part of the dielectric constant, which is not possible to determine exactly for a protein system. Following previous works, we scaled charges of Ca^{2+}, salt ions (K^+,Cl^-) and Ca^{2+} coordinating specific oxygen atoms by a scaling factor of 0.75 systematically. Our main aim is to establish a computational procedure that gives novel good agreement with experimental binding free energy.

2. Molecular Dynamic Simulation and Parameters

All simulations in this study were performed using the AMBER20 package.[32] To set up for molecular dynamics simulation of cTn-C protein, we started with the wild-type (WT)

protein structure of the calcium-saturated N-terminal regulatory domain of human cardiac troponin C obtained from the Protein Data Bank (PDB ID:1AP4) shown in Figure 1. The initial structure served as the template for constructing mutant proteins by changing specific residues in the wild-type protein structure. Missing atoms and missing hydrogen in the mutant protein were added using the tleap module of the Amber20 package. The protonation state of residue was determined at a neutral pH using the pdb2pqr server (https://server.poissonboltzmann.org/pdb2pqr).[33]

The all-atom ff14SB force field[34] was used to describe protein parameters and system was solvated in TIP3P water model containing 0.15 M of KCl, where atomic charges of Ca^{2+} and its coordinating residue oxygen atoms were scaled by a factor of 0.75 systematically (details in supplementary Table S1 and S2). The minimum distance between the protein and the cubic box edge was set to 10Å. Periodic boundary conditions were applied throughout all simulations. For the monovalent and divalent ions, we have used the Joung-Cheatham force field parameters and Li-Merz force field parameters respectively, ensuring that the specific interactions of these ions are well-represented. Once the appropriate parameters set were established, we commenced by minimizing the system through 5000 steps using the steepest descent method, followed by an additional 5000 steps of the conjugate gradient method. During this phase, the protein was restrained with a force constant of 2 kcal/mol/Å^2 on the heavy atoms of protein and Ca^{2+}. Subsequently, we proceeded to relax the entire system for another 10,000 steps. van der Waals and short-range electrostatic interactions were truncated at a spherical cutoff distance of 10Å and long-range electrostatic interactions were accounted by the particle mesh Ewald (PME) summation algorithm. SHAKE algorithm was used to constrain the bond distances involving hydrogen atoms with a time step of 2 fs for integrating the equation of motion. Following the minimization process, we gradually increased the temperature of the system to reach 300 K and kept it constant using a Langevin thermostat with a collision frequency of $2ps^{-1}$ and the Berendsen barostat with a pressure relaxation time of 1ps, was utilized for constant pressure. After 10 ns of heating and subse-

quent 10 ns of density equilibration, the system underwent a 10 ns of equilibrium using the canonical ensemble, allowing the system to adjust while maintaining a constant temperature and pressure. After the completion of the equilibration phase, the system was simulated for 100 ns to achieve stable root mean square fluctuation (RMSF), root mean square deviation (RMSD), and some other parameters. Next, for BFE calculations, ten parallel trajectories of 10ns were conducted to avoid false conclusions from a single simulation trajectory. Statistical significant value (details in supplementary) was calculated for our data to avoid errors in saving uncorrelated configurations for Ca^{2+} binding free energy calculation. All simulations were performed using the GPU-enabled cuda using the AMBER20 software package and pmemd.MPI program was used to perform steps upto equilibration of the system. Furthermore, the cpptraj program was utilized to analyze system properties such as RMSD, RMSF, number of hydrogen bond interactions between protein, and protein-water. VMD visualization tool[35] tcl script was used to calculate an average number of water molecules coordinating within $3\mathring{A}$ of Ca^{2+} ion during simulation.

3. BFE Calculation using MM-PBSA(Molecular Mechanics Poisson-Boltzmann Surface Area) method

Total binding free energy (ΔG_{bind}) of protein-ion complex in MM-PBSA is approximated as:

$$\Delta G_{bind} = \Delta E_{MM} + \Delta G_{solv} - T\Delta S \tag{1}$$

where ΔE_{MM} represents a change in the gas-phase interaction and ΔG_{solv} is the change in solvation-free energy upon ion binding. ΔE_{MM} is calculated as the sum of electrostatic changes and van der Waals interactions upon ion binding. The decomposition of energy term ΔE_{MM} in MM-PBSA is given as:

$$\Delta E_{MM} = \Delta E_{VdW} + \Delta E_{electrostatics} \tag{2}$$

The solvation free energy (ΔG_{solv}) consists of two components related to ion binding as:

$$\Delta G_{solv} = \Delta G_{polar} + \Delta G_{non-polar} \tag{3}$$

The calculation of ΔG_{polar} using Poisson-Boltzmann equation, involves estimating the electrostatic solvation energy of polar groups in the protein-ion complex upon ion binding. The decomposition of non-polar energy terms into cavitation and dispersion is given as:

$$\Delta G_{non-polar} = \gamma SASA + c + \Delta G_{dispersion} \tag{4}$$

here γ is the proportionality constant representing the change in free energy per unit molecular surface area and SASA is Molecular surface area of the protein-ion complex. Here constant terms are c = -0.569 kcal/mol and γ is 0.0378 kcal/mol A^2 used for our calculations as implemented in AMBER20 package.[36] The entropy component (-TΔS) is not considered in our calculations.

MMPBSA.py program was conducted with the ionic strength of 0.15 M and dielectric constant of 12 and 78.35 for the protein and the water respectively providing the most accurate agreement of calculated relative binding free energy with experimental observations. Dielectric constant of the protein was optimized because it is instrumental in computing the polar component of the solvation-free energy. For this $\Delta\Delta G_{bind}$ of all mutations has been calculated at different dielectric constant of protein ranging from 8 to 14 and found that 12 gives the lowest RMSD between calculated and experiment $\Delta\Delta G_{bind}$ as shown in Figure 2.

4. Evolutionary approach using a statistical probability method

A theoretical deep mutational scan along the sequence of cTn-C protein was done using a statistical probability method "EVmutation" developed by Hopf et al.[26] This method includes both site-specific constraint and pairwise constraints between two amino acid positions i and j, represented in Figure 3.[37]

So, total energy of specific sequence E(σ) is:

$$E(\sigma) = \sum_{i} h_i(\sigma_i) + \sum_{i<j} J_{ij}(\sigma_i, \sigma_j) \tag{5}$$

Here, E(σ) is an energy function used to capture different types of constraint in the sequence. h_i represents amino acid positions i as site-specific constraints to determine the likelihood of each amino acid residue within sequence σ of length N. $J_{ij}(\sigma_i, \sigma_j)$ represents the coupling strength between residues i and j within sequence σ. This coupling strength denotes the degree to which substituting an amino acid at the ith position influences the fitness of an amino acid at the jth position, and vice versa.

Using these parameters, the effect of each possible substitution has been quantified by evaluating the log-odds ratio of sequence probabilities between the wild-type and mutant sequence as evolutionary statistical energy change (ΔE). Equation is given below:

$$\Delta E(\sigma^{mut}, \sigma^{WT}) = log\frac{P(\sigma^{mut})}{P(\sigma^{WT})} = E(\sigma^{mut}) - E(\sigma^{WT}) \tag{6}$$

This EVmutation model incorporates the sequence context into the calculation of mutation effects.

4.1 Selection of plausible mutations using EVmutation.py

In our study, we have used the N-terminal domain of cardiac troponin-C sequence (Pfam:PF13833) to access the quantitative effect of mutations, facilitated by a Python framework EVmutation.py. Multiple sequence alignment of the protein was done using the Jackhammer tool using the UniRef100 database and a length-normalized bit score threshold of 0.3 was applied. We have identified 76,931 homologous sequences of the target protein, with 1221 effective sequences (Meff/L) between proteins of different lengths. To enhance its quality, a post-processing step was conducted, filtering out positions with over 30% gaps and sequences aligning less than 50% of the target sequence length, and model complexity was controlled

by applying l_2-regularization. EVmutation model has been run and results were analyzed, where evolutionary statistical energy change upon mutation (ΔE) was extracted from the mutational landscape. For the selection of beneficial substitutions, mutations having positive ΔE were selected out of all possible mutations in the protein sequence. ΔE value greater than zero indicates mutant sequences are more probable and potentially beneficial. Conversely, values below zero suggest mutant sequences are less probable and possibly deleterious. Hence, we chose 61 beneficial mutations that are presumed to be constituting our plausible mutation (PM) dataset.

5. Machine Learning algorithm

The machine learning (ML) models utilized for the prediction of Ca^{2+} relative binding free energy, denoted as:

$$\Delta\Delta G_{\text{pred},i} = F(x_1, x_2, \ldots, x_9) \tag{7}$$

where F is the ML model adopting a nonlinear function, and $x_1, x_2, ...x_9$ are the nine input features forming a feature vector.

5.1 Benchmarking Input Feature Selection

The gist of machine learning lies in generating distinctive input features. Our model includes simple sequence, structure, and Ca^{2+} interaction properties and it is comprised of nine features that are systematically categorized based on their intrinsic characteristics.

Sequence-based features: We have included properties such as conservation score of each residue along the protein sequence, and the evolutionary statistical energy difference upon mutation (ΔE) was extracted from the evolutionary approach. The conservation score represents the importance of conserved amino acid residue during sequence evolution, which may be critical for the stability of protein structure and thus function. Primary sequence properties include the position of the mutation and change in charge of the system upon

mutation which usually changes where charged amino acids are involved.

Structure-based features: We have included properties such as type of amino acid substitution, secondary structure of the residue at the mutation site, distance of mutation from Ca^{2+} in three-dimensional space, and solvent-accessible surface area (SASA) change upon mutation. Amino acid substitution describes amino acid properties, whether it is charged, hydrophobic, aromatic or polar. Secondary structure of the residue at the mutation site indicating whether the mutation took place in helix, sheet, coil, or turn. SASA was calculated to quantify the exposure of a residue to the solvent and thus determine its accessibility to the molecular surface. The values taken for the accessible surface area used in this reference paper[38] were used, presented in Table S3.

Ca^{2+} ***interaction properties:*** It include number of water molecules coordinating within $3\AA$ radius of Ca^{2+}, extracted from simulation. The physical basis for choosing this is because the number of water molecules present in the vicinity of Ca^{2+} disturb the coordination shell of Ca^{2+} binding loop and thus change in Ca^{2+} binding affinity.

5.2 Machine learning algorithms

Prediction model: Several ML models were leveraged for the prediction of Ca^{2+} relative binding free energy of cTn-C protein variants. Support vector regression, decision tree, and random forest algorithms were used to build the prediction model and assess their performance. This comparative analysis aims to identify the best prediction algorithm customized for our mutation dataset. Out of these three algorithms, Support Vector Regression exhibits superior performance and thus most effective model for our dataset. SVR conducts regression tasks by seeking an optimized hyperplane that minimizes the sum of distances between data points and the hyperplane. In our study, we employed SVR using scikit-learn[39] library, where input features were standardized using the StandardScaler to ensure uniform scaling. Mutation dataset was split randomly into training and test sets using an 80%-20% split

ratio respectively. Hyperparameter tuning was conducted using GridSearchCV, exploring various combinations of hyperparameters, such as the regularization parameter (C), margin of tolerance (ϵ), and type of kernel used. Following tuning of optimized hyperparameters, the model was trained on the training set and the performance of the model was carried out on the test set using multiple evaluation metrices including Mean Absolute Error (MAE), Mean Squared Error (MSE), R-squared (R^2) coefficient, Pearson correlation coefficient (R_p).

Assess the contribution of each feature: In the succeeding steps, we plan to carry out an analysis to identify the contribution of each input feature used in our ML model. For this purpose, we applied three different approaches to discern the features that contribute most significantly to the model. First, we calculate the correlation coefficients of each feature with $\Delta\Delta G_{bind}$ using linear, polynomial, and radial basis function kernel. Next, the Permutation feature importance method was used which assesses the impact of each feature on model performance by measuring the change in accuracy when the values of that feature are randomly shuffled. To further investigate the role of each feature, we have used the SHAP (SHapley Additive exPlanations) method to provide a unified measure by assigning a contribution score to each feature for the prediction of model, based on SHapley values. The use of these three approaches for calculating the contribution of each feature enhances the interpretability and transparency of the model.

Classification model: Support Vector Classifier (SVC) algorithm was used with a linear kernel for the classification of high and low Ca^{2+} binding affinity cTn-C mutations. The SV classifier performance was evaluated on the test dataset with 10-fold cross-validation and the mean value of accuracy scores for each fold were computed to assess the overall performance of the classification model.

Results and Discussion

1. Binding Free Energy Calculations

The charge scaling protocol with MM-PBSA method was principally used to add polarization and accurately calculate the Ca^{2+} BFE of cTn-C mutants. First, we collected 24 cTn-C diseased mutations Ca^{2+} binding affinity (mutation sites are represented in Figure S2) associated with different types of cardiomyopathies, with experimental measurements available in the literature represented in Table 1. For the theoretical ΔG_{bind} calculations, we have selected 500 statistically significant configurations. The mean values and the mean of the standard error of the mean (SEM) of ΔG_{bind} and its different components such as ΔE_{vdW},ΔE_{elec}, ΔE_{PB},$\Delta E_{cavitation}$ and $\Delta E_{dispersion}$ were calculated for wild type (WT) and diseased mutations are summarized in Table 2. Next, we calculated the relative BFE (where $\Delta\Delta G_{bind}$ = ΔG_{mutant} - ΔG_{WT}) for diseased mutations represented in Figure 4 shows that some mutations such as Y5H, S37G, E59D, I61Q, D87A, D88A shows very low $\Delta\Delta G_{bind}$ values, indicating that they are neutral in terms of their mutation effect on binding. Our observation reveals a pattern where negatively charged residue substitution to other amino acids reduces Ca^{2+} binding affinity. For example, mutations such as E40A, D67A, D73A, and D75Y present in the loop show a significant reduction in Ca^{2+} binding affinity, whereas D87A and D88A mutations, located outside the loop, have a less noticeable effect. This indicates that, beyond the type of mutation, the position of the mutation within the three-dimensional structure is also crucial. In the pursuit of this observation, hydrophobic to glutamine residue substitution was seen in mutations such as F20Q, A23Q, L29Q, V44Q, L48Q, I61Q, V79Q, M81Q showing a consistent increase in Ca^{2+} binding affinity qualitatively irrespective of the position of the mutation except L57Q. Additionally, in some mutations such as E59D and I61Q, it appears that there is an opposite trend between calculated and experiment $\Delta\Delta G_{bind}$, however difference is not significant. An additional observation highlighted by the figure is that a few mutations such as A8V, F20Q, I61Q, D75Y, V79Q, M81Q show large quanti-

tative differences between the experiment and calculations. To access the potential reason for this difference, we checked the contributions of each coordinating and non-coordinating residue present in Ca^{2+} binding loop II. Residue-wise decomposition of loop II residues from 65 to 76 has been done and a comparison of each residue contribution to BFE in the WT and mutant systems are shown in Figure 5. From this observation, we can conclude that negatively charged amino acids such as aspartic acid present at 1,3,9th position and glutamic acid present at 12th position show a significantly high contribution to BFE in the WT system. Therefore, we compare these residue contributions in some mutations and found in the case of A8V, aspartic acid present at the 9th position shows more contribution than WT. Concerning F20Q, aspartic acid present at the 3rd position shows a higher contribution than WT, and for V79Q and M81Q, aspartic acid present at the 3rd and 9th position shows higher contribution than WT, hence responsible for higher $\Delta\Delta G_{bind}$ than WT. As for D75Y mutation, the contribution of all the negative charge residues decreases resulting in abrupt decrease in $\Delta\Delta G_{bind}$.

Role of water in coordination sphere: By scrutinizing the simulation outputs of WT and mutations, highlights the contribution of one coordinating water molecule with Ca^{2+} in the Ca^{2+} coordination sphere of the WT system. Few mutations such as D75Y show the presence of three coordinating water molecules with Ca^{2+} disrupting the coordination sphere and causing low binding affinity. M45Q has one coordinating water molecule whereas V79Q mutation has no water molecules present in Ca^{2+} coordination sphere and shows increased binding affinity. Diagrammatic representation is shown in Figure 6.

Regression model: Moving forward, our next step involves quantifying the accuracy of our calculated $\Delta\Delta G_{cal}$ with experimental $\Delta\Delta G_{exp}$ represented in Table 3. Figure 7 represents a linear regression model using the Ordinary Least Square module of sklearn in Python based on BFE components as descriptors for the prediction of the BFE. Evaluation metrics such as Pearson's correlation coefficient (R_p), spearman's correlation coefficient (R_s)

and R^2 have been used to check the accuracy of the model. R_p value of 0.87, R_s value of 0.88 and R^2 of 0.75, reflect highly favorable outcomes.

2. Generation of plausible mutations using evolutionary information

In this section, we augmented the mutation dataset by generating additional plausible mutations (PM) using the "EVmutation" described in the method section. Proceed with analyzing results obtained from the theoretical deep mutational scan using equation 7 and the effect of each mutation was analyzed in terms of evolutionary statistical energy difference (ΔE). Figure 8a represents the evolutionary landscape of each mutation in the protein and demonstrates the evolutionary fitness of particular mutation in terms of ΔE. Beneficial mutations show positive ΔE ranging from 0 to +4 were selected for our PM dataset. Hence, we selected a total of 61 mutations are shown in Figure 8b and BFE calculations of these PM have been done using the same charge scaling with MM-PBSA protocol and the values of ΔG_{bind} are represented in Table S4 and S5. To develop our machine learning-based BFE prediction model for cTN-C mutations, we calculated the relative BFE ($\Delta\Delta G_{cal}$) of PM and values are presented in Table 4 and 5. Next, the conservation score of each residue in the protein sequence aiding in the identification of critical regions that are essential for protein function and Ca^{2+} binding activity was extracted from the results obtained through EVmutation. Focusing on the conservation pattern reveals residues having a high conservation score are mostly present in the binding loop region while residues present outside the binding loop show a low conservation score, illustrated in Figure 9. Residues present in the protein sequence from 29 to 40 (loop I) and 65-76 (loop II) are conserved except few.

3. BFE Prediction Model using Machine Learning

We have used two advanced machine learning models for the prediction and classification of Ca^{2+} binding affinity of cTn-C protein variants. Support Vector Regression (SVR) was utilized for prediction and Support Vector Classification (SVC) was applied to categorize the mutation dataset into two distinct classes.

3.1 Support Vector Regression

As detailed in the Methods section, we have used nine different feature for our ML model represented in Table 6. Throughout the model training process, the mean $\Delta\Delta G_{cal}$ values served as reference benchmarks for 75 mutations (DM+PM). Employing hyperparameter tunning on the mutation dataset, found that a margin of tolerance (ϵ) value of 0.5 and regularization parameter of 10, with a polynomial kernel gives the best statistical parameters. Pearson's correlation coefficient of 0.92 and R^2 of 0.77, along with MSE of 0.16 kcal/mol, are represented in Figure 10. Also, we check the SVR model on linear and rbf kernel with same values of ϵ and C, and statistical parameters obtained using various kernels are shown in Figure 11 and summarized in Table S6.

Assess the contribution of each feature: Next, to evaluate the contribution of each feature used in the model, we calculated the correlation coefficients of each feature with $\Delta\Delta G_{bind}$ using all three kernels as shown in Figure 12. The number of water molecules coordinating within $3\mathring{A}$ of Ca^{2+}, type of amino acid substitution, and distance of mutation with Ca^{2+} features show the best correlation coefficients in poly and rbf kernel. For deeper analysis, the contribution of each feature in the prediction model was calculated using permutation feature importance and obtained results represented in terms of ranking of each feature are shown in Figure 13.

Next in continuation of this analysis, SHAP values were calculated and used to interpret the impact of individual features on the model. The SHAP plot is shown in Figure 14, in

which each point represents a single prediction. Position on the x-axis indicating the SHAP value for a specific feature. For instance, the positive SHAP value of the feature number of water molecules coordinating within $3\mathring{A}$ of Ca^{2+} generally increases the prediction of the model. The SHAP heatmap shown in Figure 15, reveal a consistent observation of high contributions of number of water molecules coordinating within $3\mathring{A}$ of Ca^{2+} feature. Following the analysis of feature importance using three approaches, we found the three most prominent features includes the number of water molecules coordinating within $3\mathring{A}$ of Ca^{2+}, type of amino acid substitution, and distance of mutation with Ca^{2+}.

Subsequently, we closely examined how effectively these features affect the evaluation metrics of the SVR model. so we calculate statistical parameters obtained through different combinations of feature types used to evaluate the model such as structure-based features in combination with sequence-based features. etc. Evaluation metrics of SVR model are presented in Table 7. Data shows that only structure-based descriptor gives R^2 of 0.57 and R_p of 0.79 and MSE of 0.31 kcal/mol, hence the most contributing class of features.

3.2 Support Vector Classification

Support Vector Classification (SVC) was employed to classify all the mutations into two categories, one with high Ca^{2+} binding affinity and the other with low Ca^{2+} binding affinity concerning WT protein. The SVC model achieved a mean accuracy of 0.76 and MSE of 0.29 kcal/mol, and ROC curve shows an AUC of 0.86 indicating good in classification of mutations.

From these ML models, we have found that the SVR model with polynomial kernel gives the best prediction of the BFE for our mutation dataset, compared with other models used in our research. Our analysis also concludes that the number of water molecules coordinating within $3\mathring{A}$ of Ca^{2+} is the most important descriptor for the model, followed by, type of amino acid substitution and the distance of the mutation from the Ca^{2+} individually. Our classification model, with AUC of 0.86 , exhibits considerable robustness. Overall, We have

developed models that are capable of predicting binding free energy of given mutation in cTn-C protein based on some physical descriptors, with high accuracy. In the scope of the future studies, our model can be used to design mutations in CPBs to acheive desirable Ca^{2+} binding affinity.

Figures and Tables

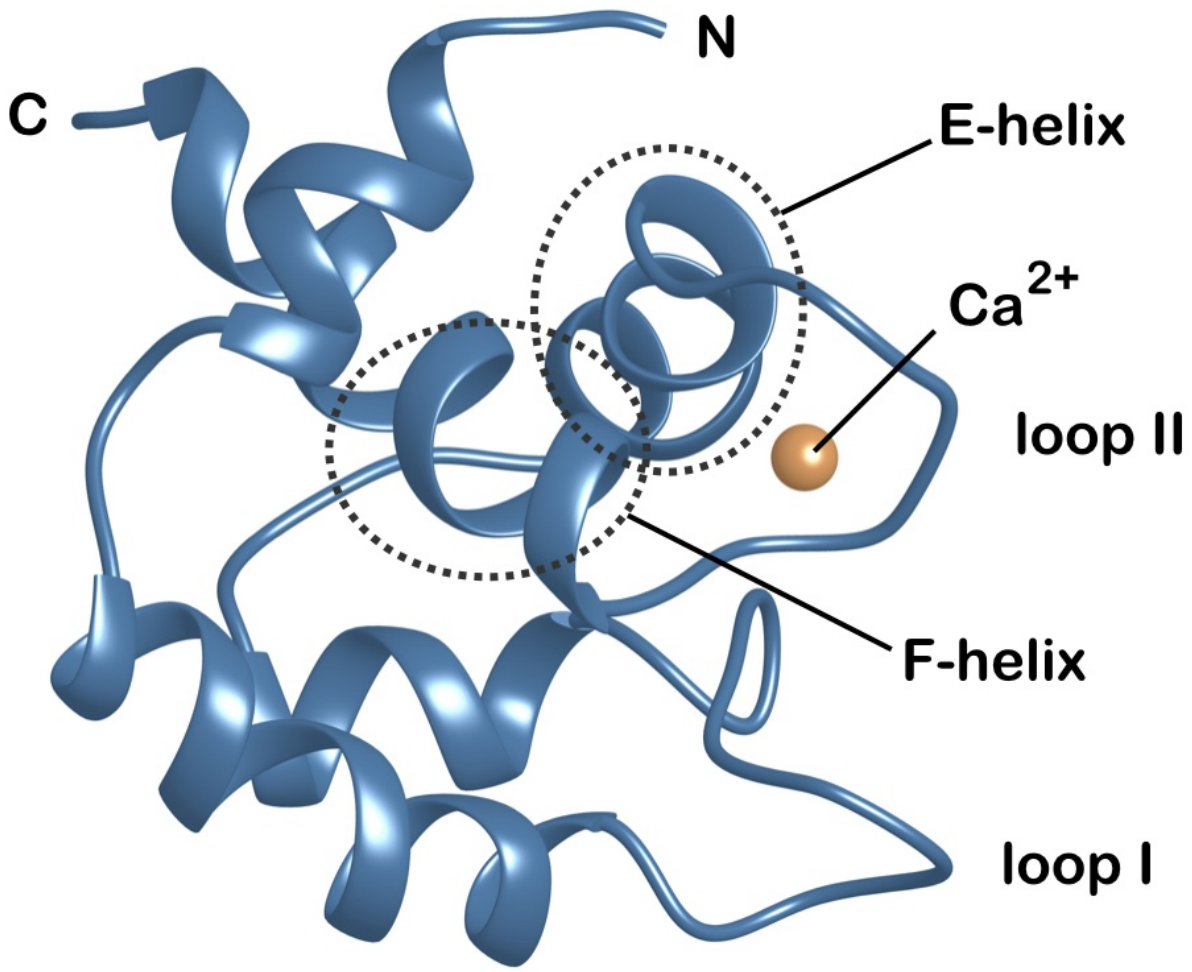


Figure 1: **Schematic representation of N-terminal of cTn-C in presence of Ca^{2+} (PDB id:1AP4). Figure illustrates the E-helix and F-helix conformations in Ca^{2+} bound state.**

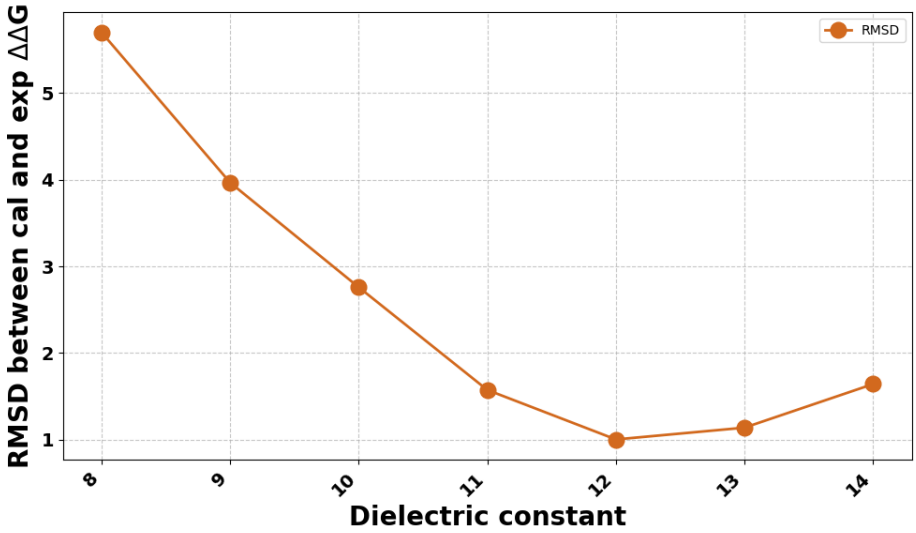


Figure 2: **Graphical representation of the RMSD between Calculated $\Delta\Delta G_{cal}$ and experimental $\Delta\Delta G_{exp}$ for diseased mutations at various dielectric constant of the protein in MM-PBSA calculations.**

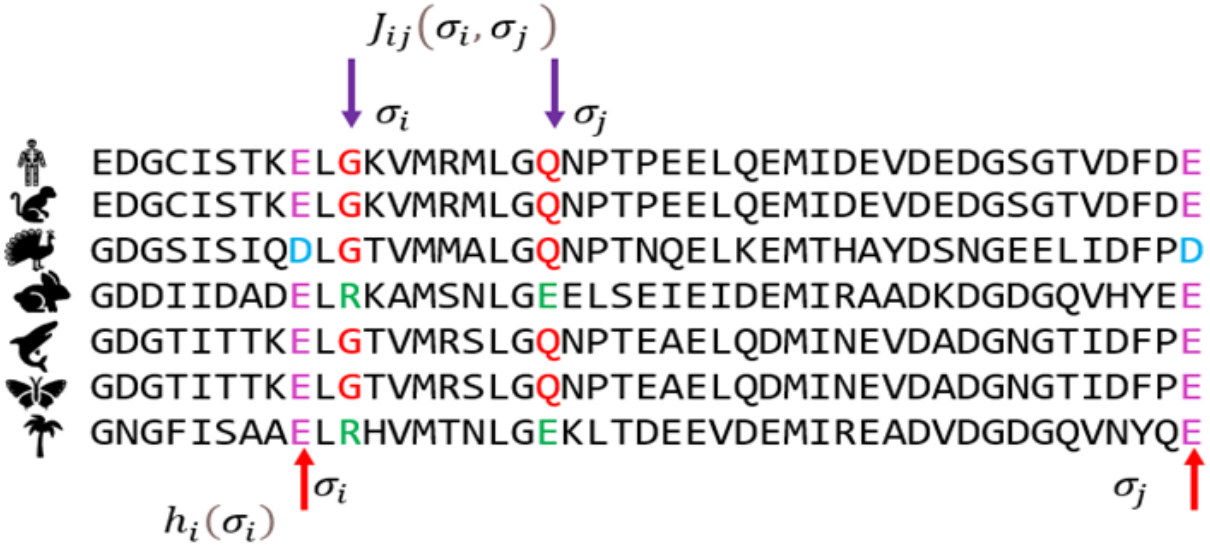


Figure 3: **Schematic representation of multiple sequence alignment of a cTn-C protein sequence with site-specific constraints (h_i) and coupling constraints (J_{ij}) between residues i and j within sequence σ.**

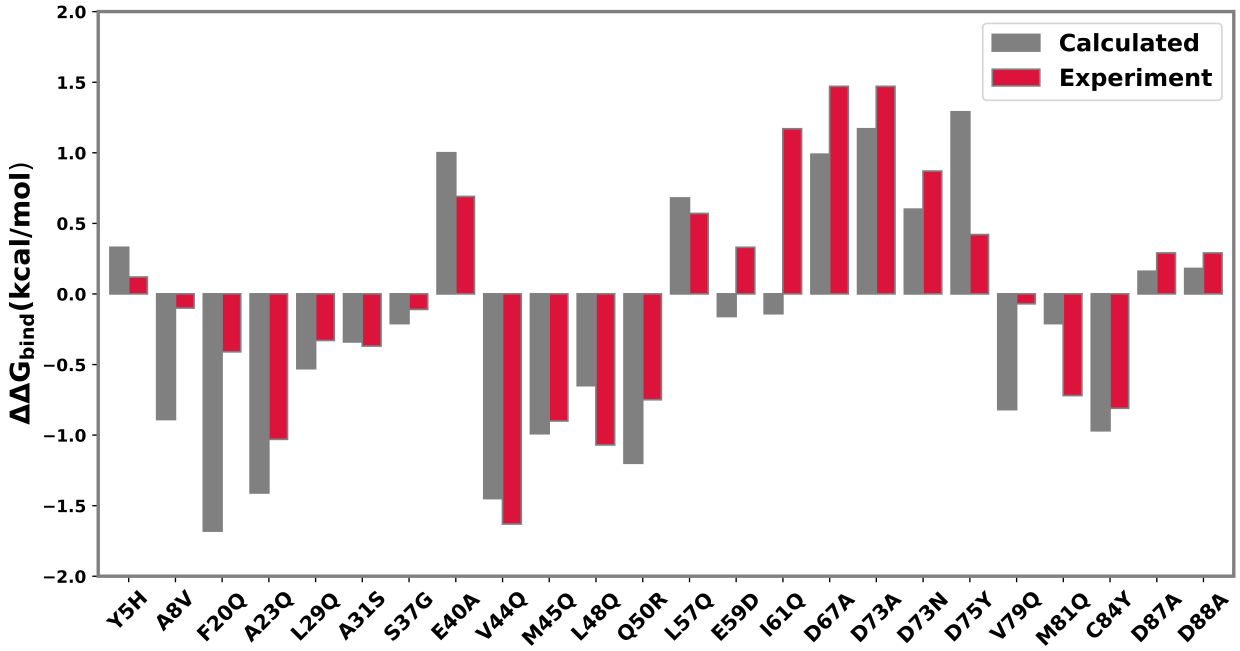


Figure 4: **Barplot representation of the calculated $\Delta\Delta G_{cal}$ with their corresponding experimental $\Delta\Delta G_{exp}$ values (in kcal/mol) for diseased mutations of cTn-C protein.**

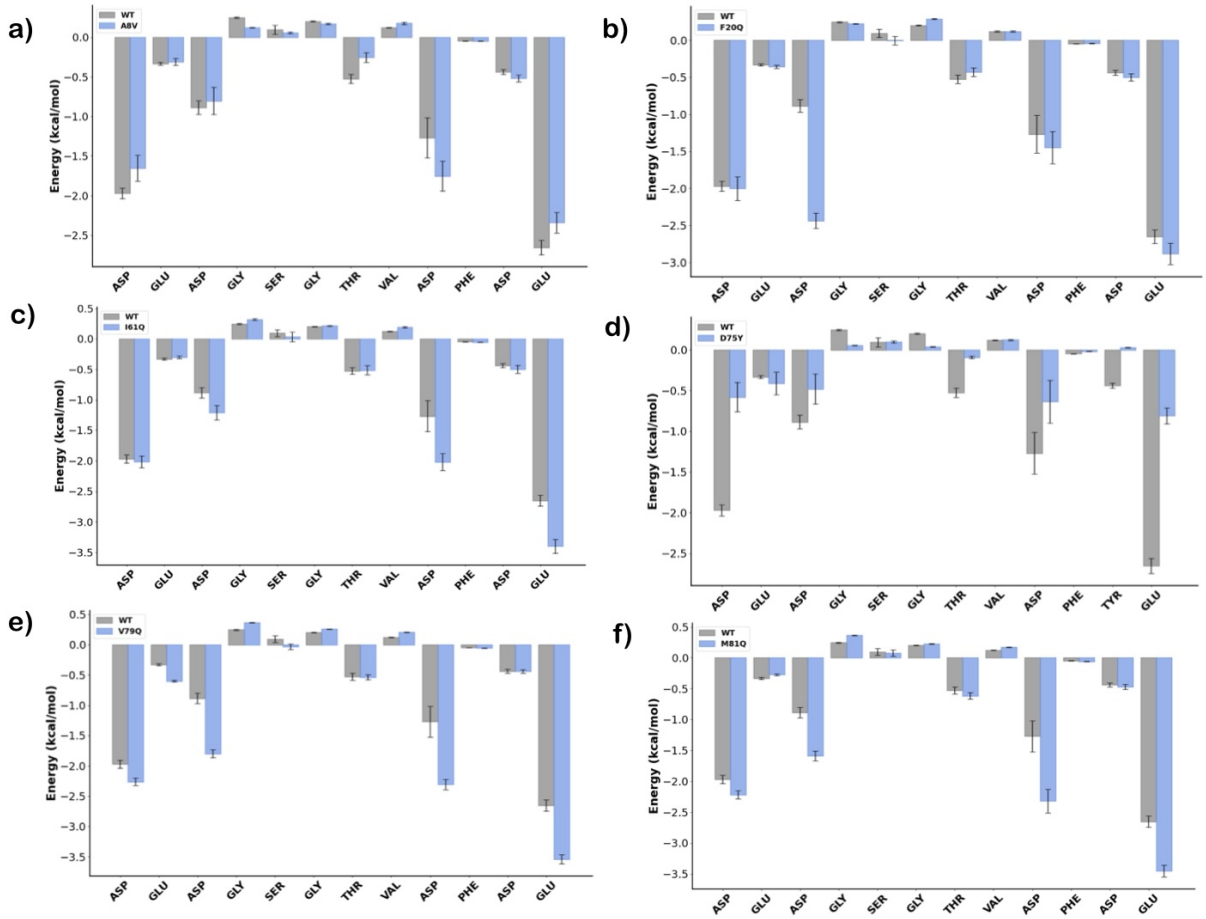


Figure 5: **Barplot representation of residue wise binding free energy decomposition of EF-hand loop II residue. The figure illustrates the contributions of the wild-type protein compared to several key mutations.**

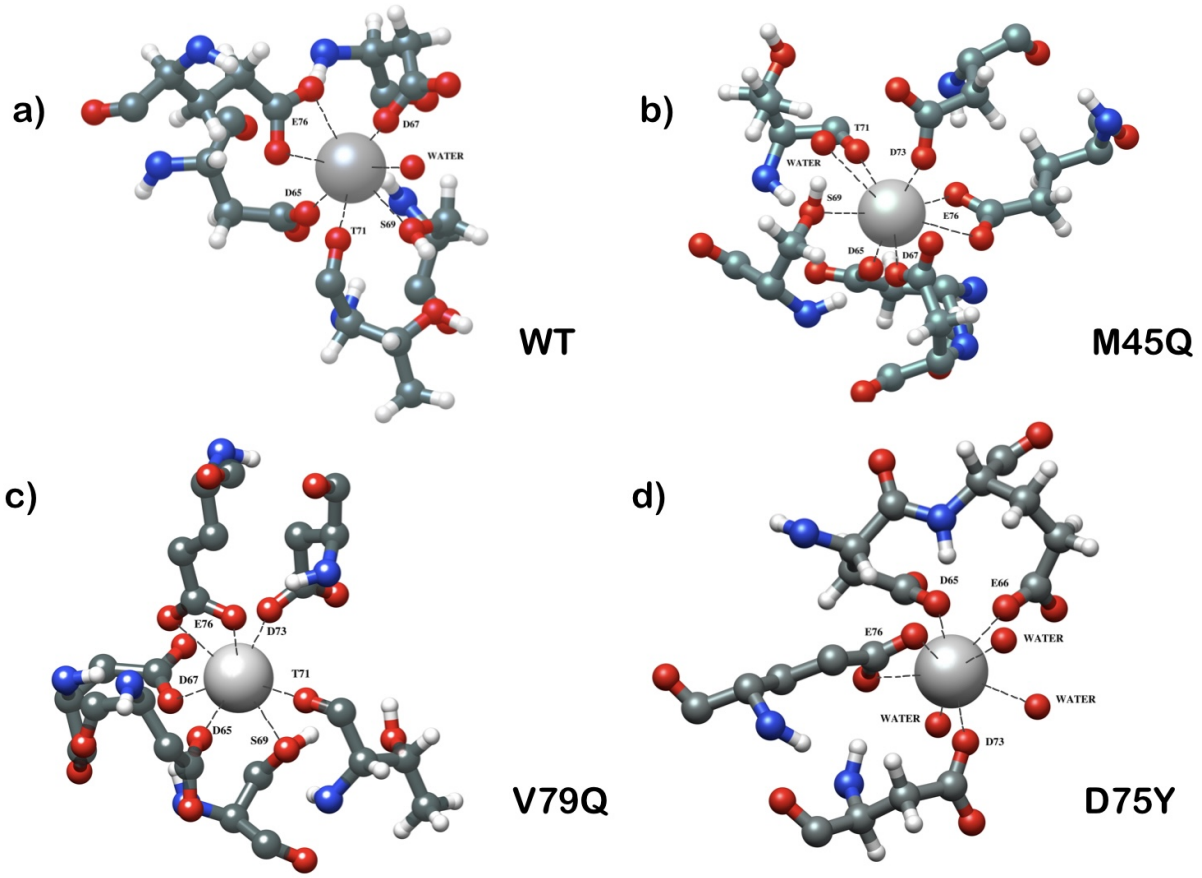


Figure 6: **Schematic representation of oxygen atoms of coordinating residues in EF-hand loop II and oxygen atoms of water molecule with Ca^{2+} in wild type and few mutations.**

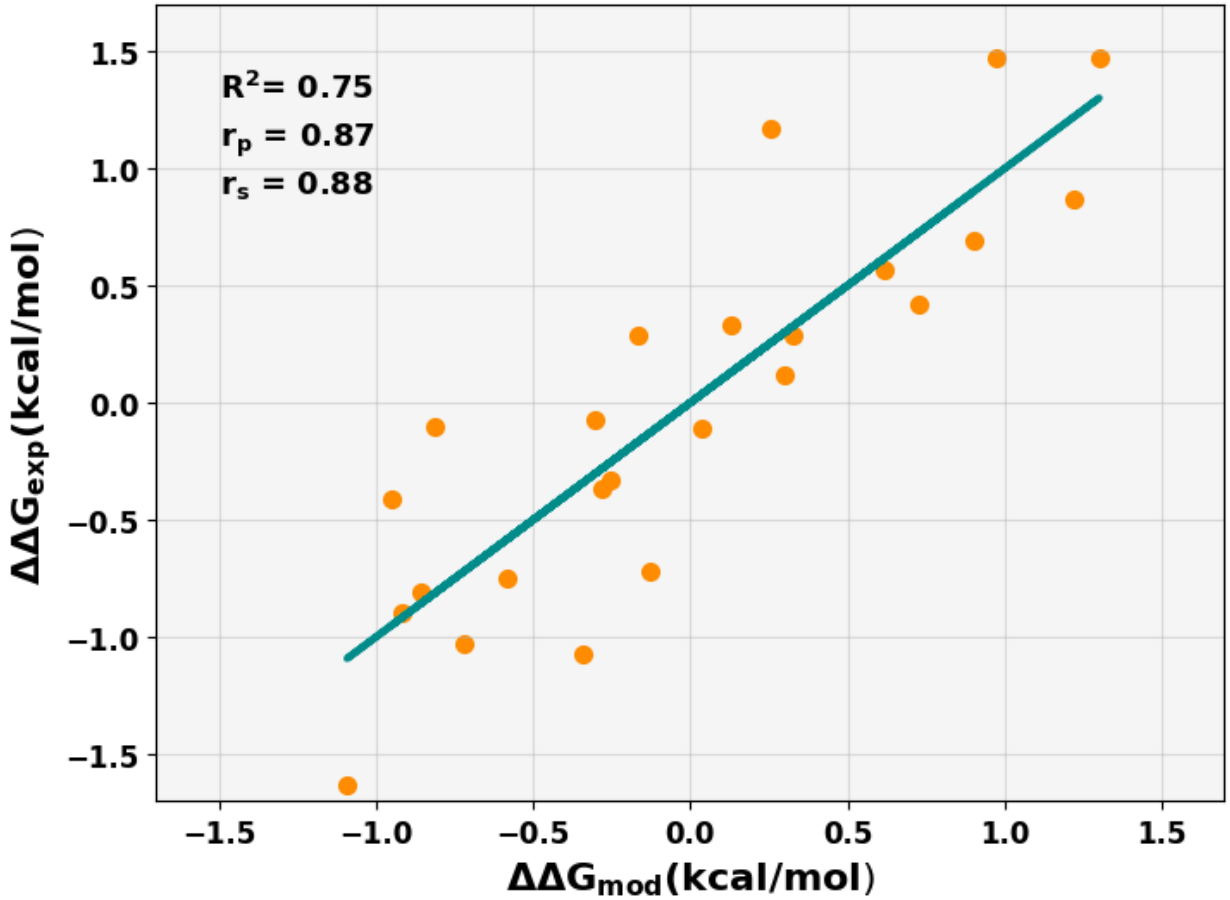
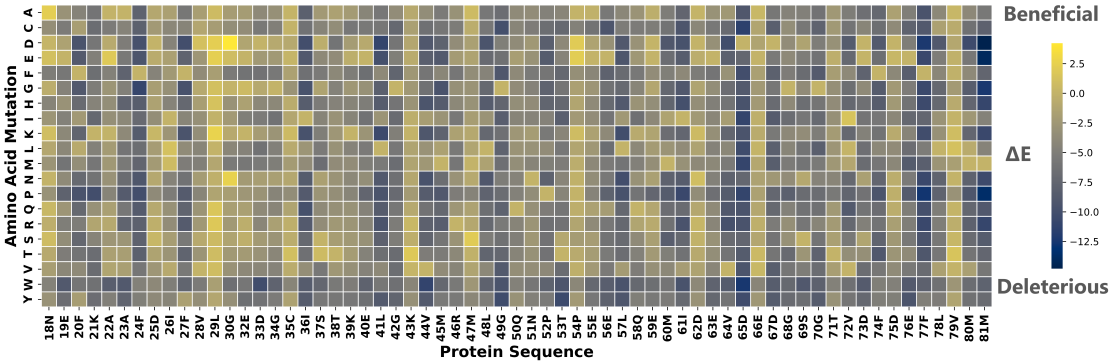
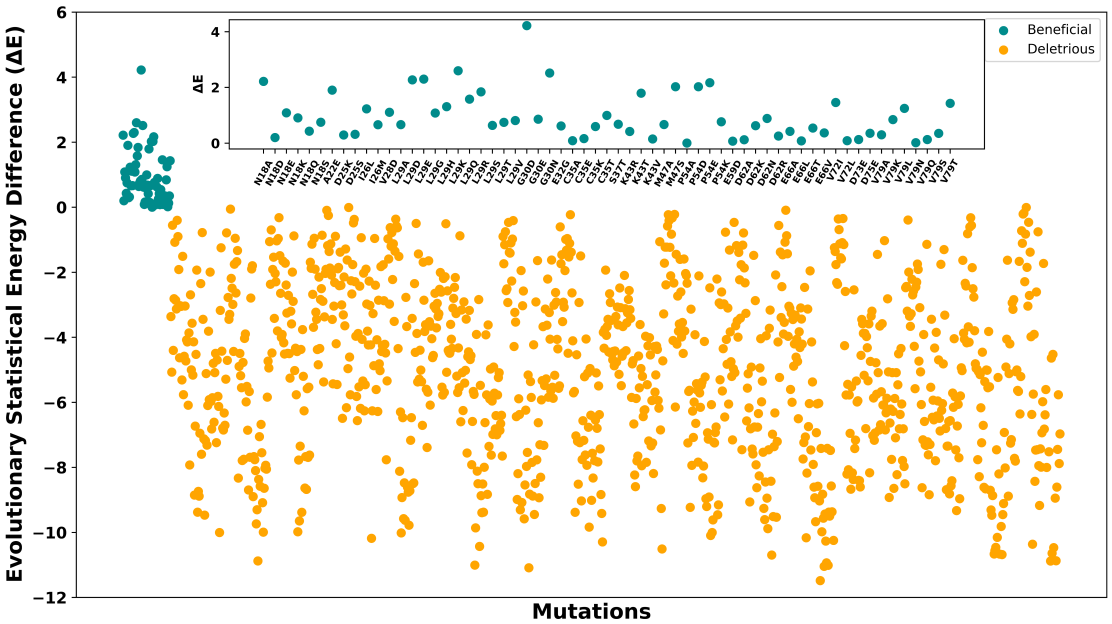


Figure 7: **This scatter plot illustrates the relationship between modellled $\Delta\Delta G_{mod}$ and experimental $\Delta\Delta G_{exp}$ values (in kcal/mol) for diseased mutations of cTn-C protein. Mutations data represented with orange dots and a solid line is linear fit for the dataset. $\Delta\Delta G_{mod}$ values are derived through a multivariate approach that encompasses components of BFE.**



(a) Mutational landscape of cTn-C protein sequence.



(b) Evolutionary statistical energy difference of mutations.

Figure 8: **(a)The mutational landscape of cTn-C protein sequence as illustrated, with the x-axis representing the residue position and the y-axis representing substituted residue. The color gradient from yellow to blue reflects the ΔE values of the each mutations in landscape in terms of benefical and deletrious mutation respectively. (b) This scatter plot represents ΔE for all mutations categorized as beneficial (positive values) and deleterious (negative values) shown in cyan and yellow respectively.**

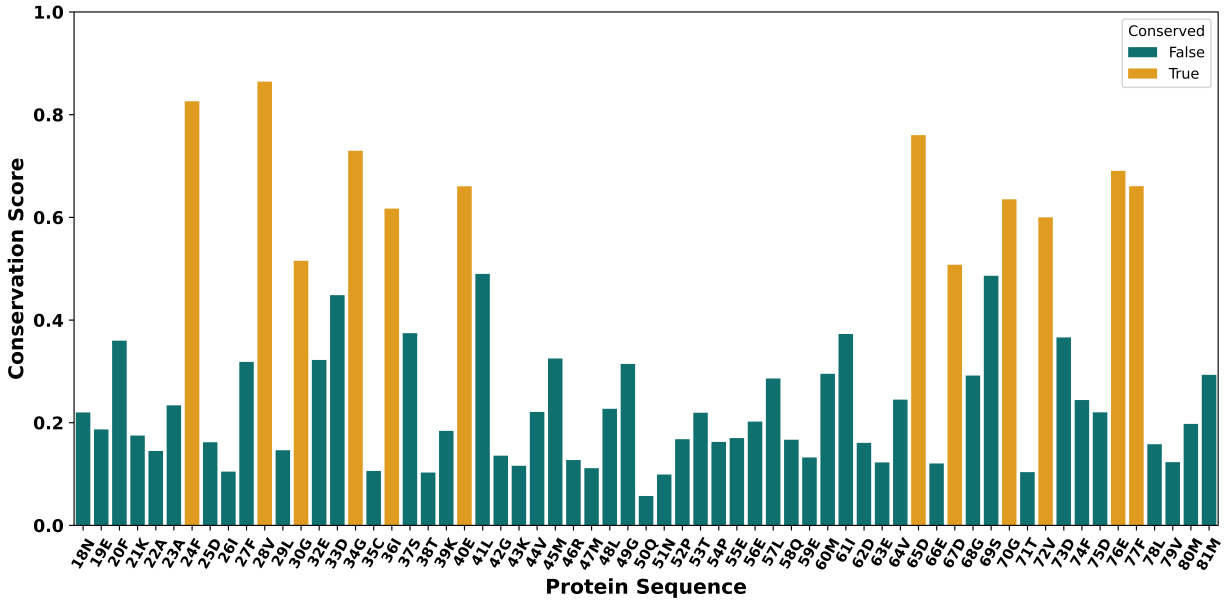


Figure 9: **This barplot represents the conservation score of each residue along the protein sequence. Residues with the highest conservation scores, indicated by yellow bars, while green bars represent residues with lower conservation score.**

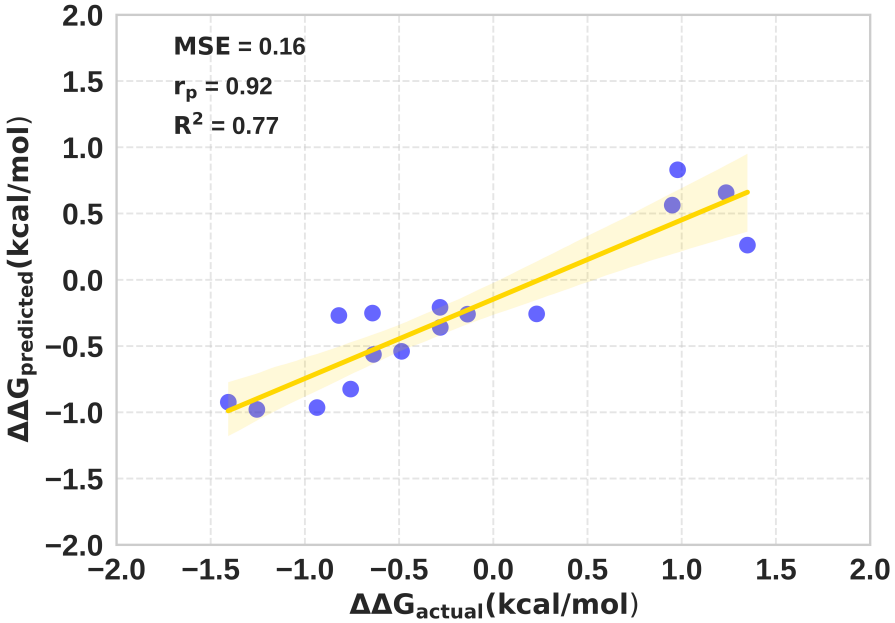


Figure 10: **This figure depicts a fitted curve using a polynomial SVR model. The x-axis represents the calculated BFE while the y-axis represents predicted BFE and evaluation metrics are shown on upper left column of the graph.**

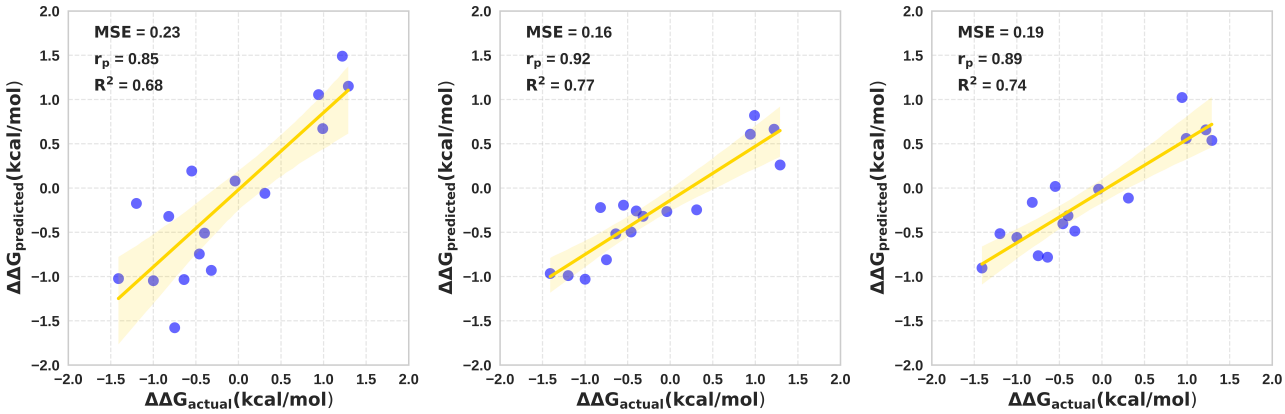


Figure 11: **This figure depicts a comparative SVR plot using linear, poly and rbf kernel to compare various evaluation metrices.**

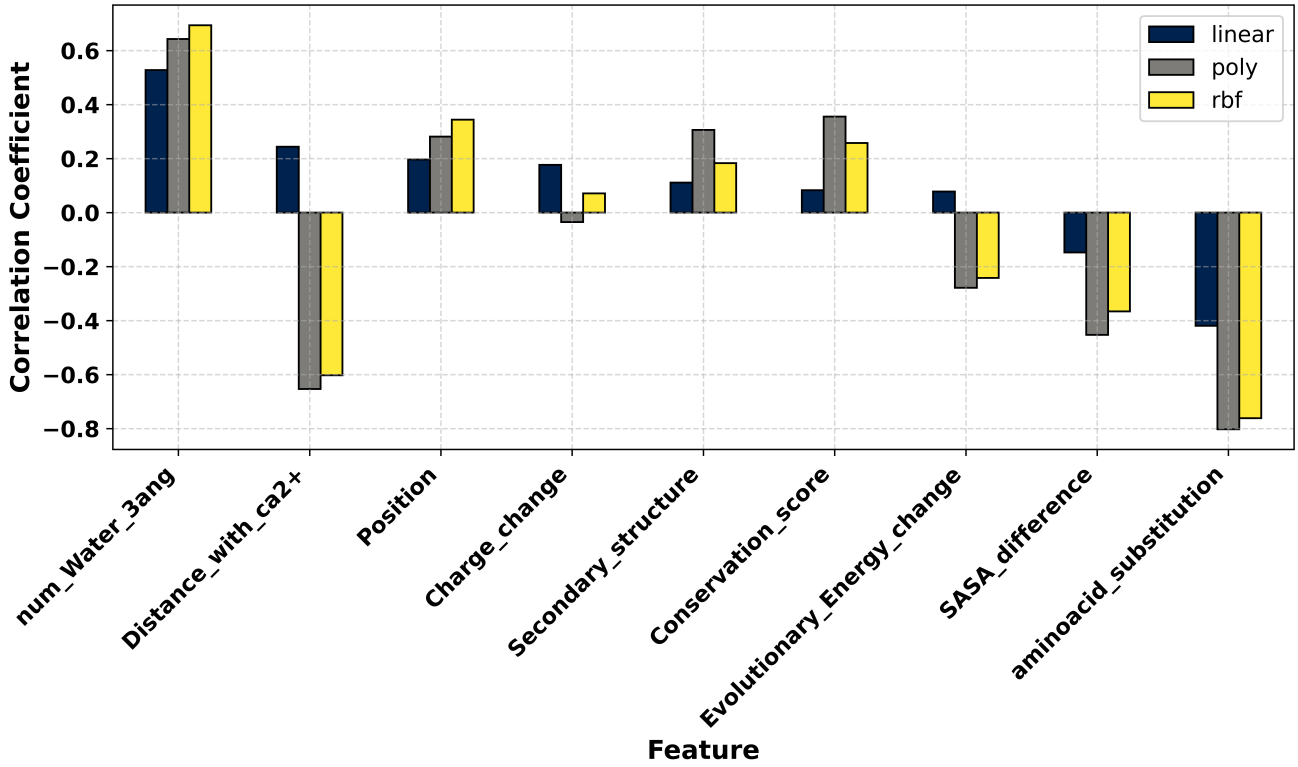


Figure 12: **Barplot graph represents correlation coefficients of each feature with $\Delta\Delta G_{bind}$ using various kernels in SVR model.**

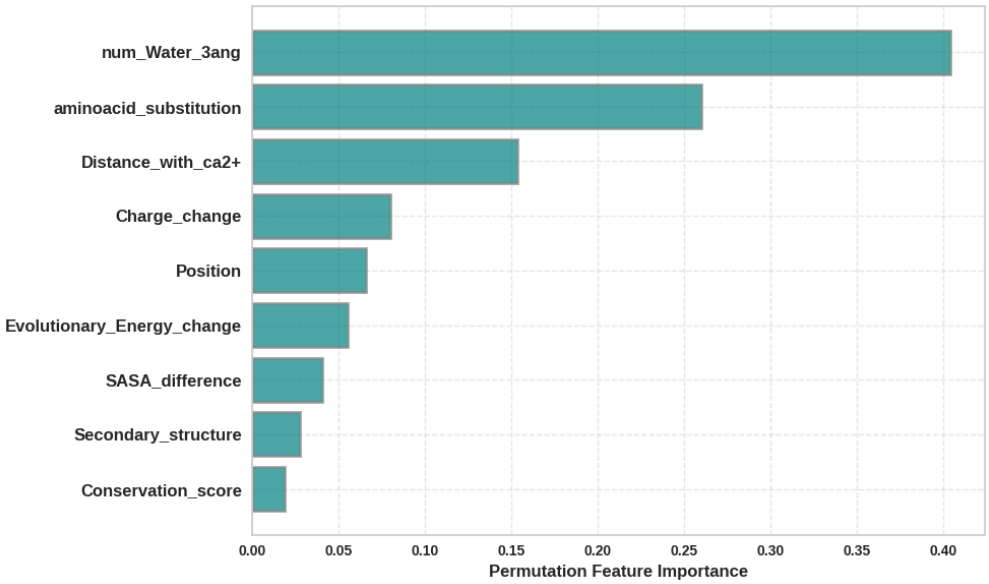


Figure 13: **Figure represents the ranking of individual feature using 'permutation feature importance' method.**

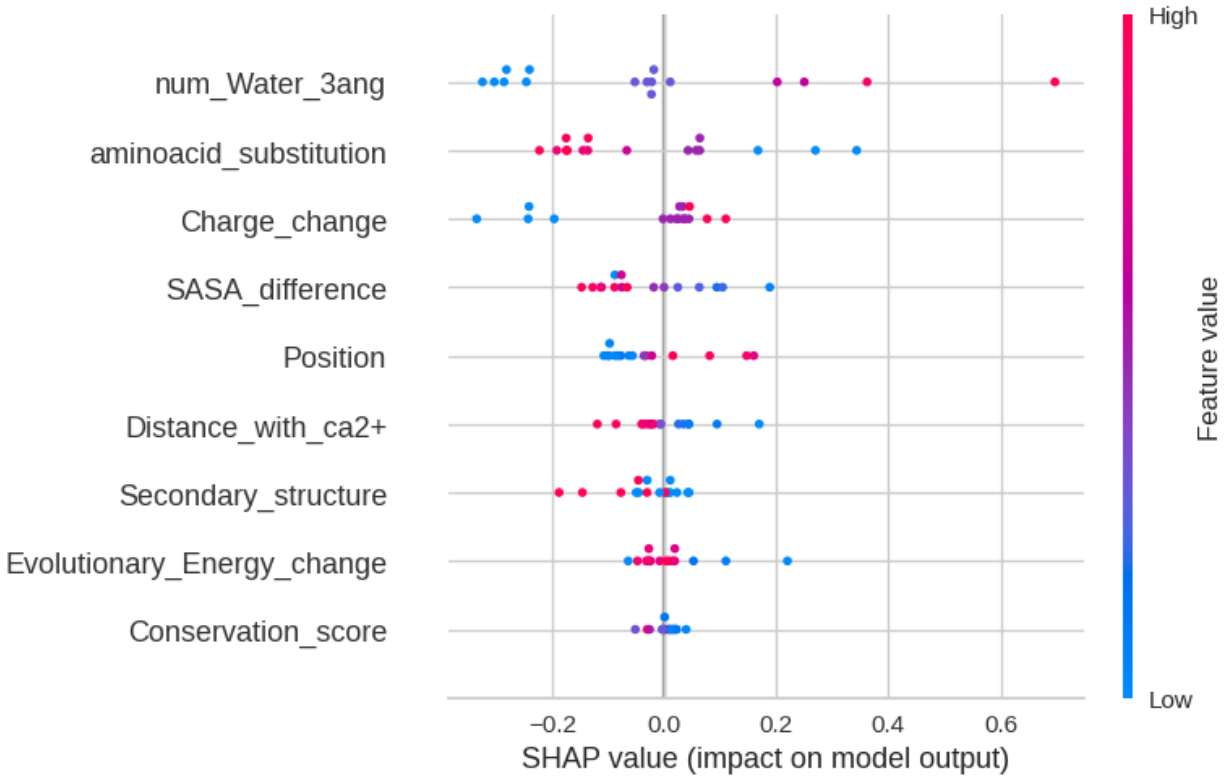


Figure 14: **Feature importance bee swarm plot ranking feature according to their impact on model output measured using SHap(thee shapley additive explanations). Each dot represents the value of a datapoint and coloured according to the feature value.**

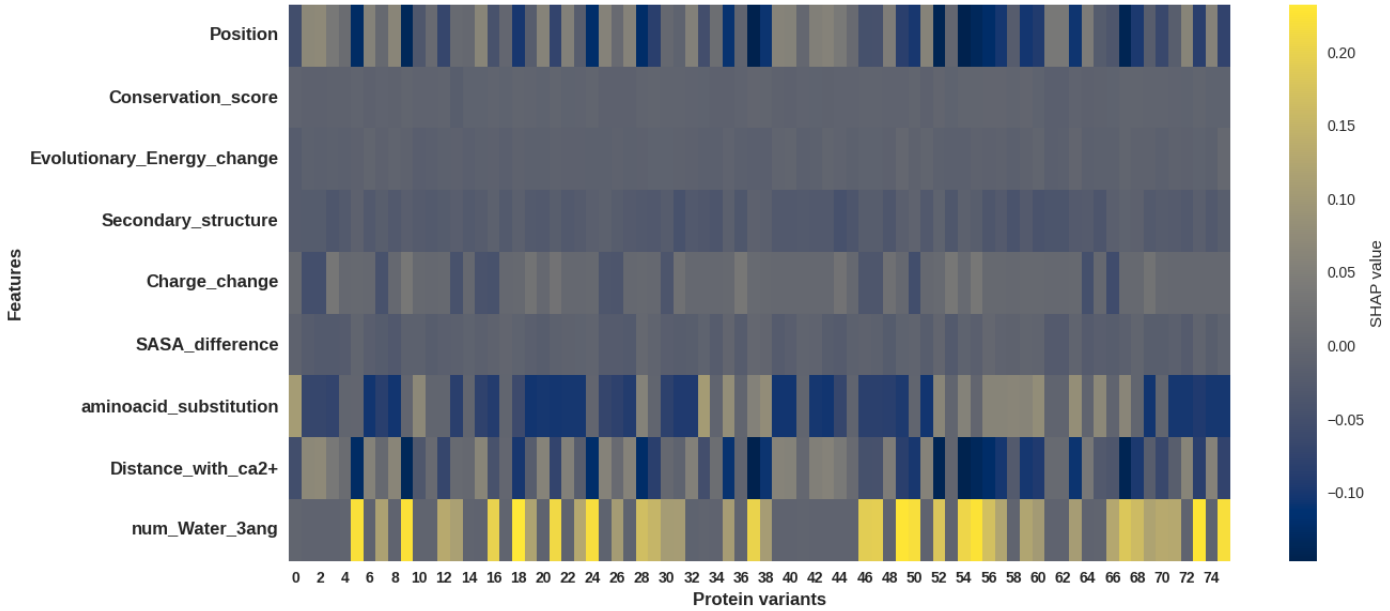


Figure 15: **Figure represents the heatmap of the each feature's SHAP value across the 75 mutation dataset.**

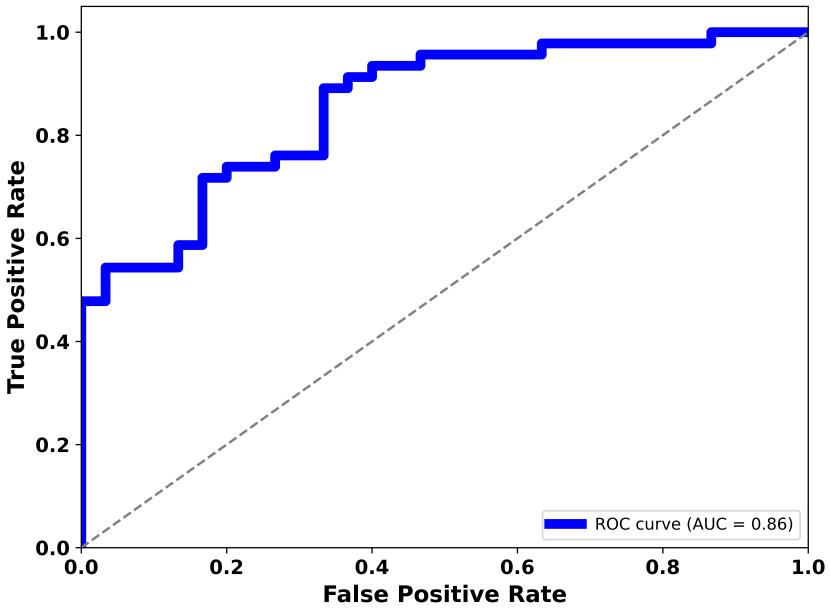


Figure 16: **The ROC curve with AUC is shown alongside the diagonal reference line. X-axis represents False Positive Rate (FPR) and y-axis denotes True Positive Rate (TPR) or sensitivity.**

Table 1: **Experimental Binding Free Energy values (ΔG_{exp}) with standard error(in Kcal/mol) of cTn-C WT protein and diseased mutations.**

System	$\Delta G_{exp}(kcal/mol)$
WT	-6.56 (0.03)[4]
Y5H	-6.44 (0.02)[5]
A8V	-6.66 (0.01)[4]
F20Q	-6.97 (0.06)[14]
A23Q	-7.59 (0.13)[6]
L29Q	-6.89 (0.03)[4]
A31S	-6.93 (0.06)[4]
S37G	-6.67 (0.03)[6]
E40A	-5.87 (0.13)[6]
V44Q	-8.19 (0.05)[14]
M45Q	-7.46 (0.02)[14]
L48Q	-7.63 (0.01)[4]
Q50R	-7.31 (0.06)[4]
L57Q	-5.99 (0.02)[7]
E59D	-6.23 (0.01)[40]
I61Q	-5.39 (0.02)[7]
D67A	-5.09 (0.06)[15]
D73A	-5.09 (0.06)[15]
D73N	-5.69 (0.08)[8]
D75Y	-6.14 (0.06)[40]
V79Q	-6.63 (0.05)
M81Q	-7.28 (0.03)[14]
C84Y	-7.37 (0.12)[4]
D87A	-6.27 (0.01)[3]
D88A	-6.27 (0.01)[3]

Table 2: **Calculated binding free energy(ΔG_{bind}) and its components with mean of standard error of mean for cTnC WT protein and diseased mutations. All values are expressed in kcal/mol, with mean of standard error of the mean provided in parentheses.**

System	ΔE_{vdW}	ΔE_{el}	ΔG_{PB}	ΔG_{cav}	ΔG_{disp}	ΔG_{bind}
WT	16.89(0.54)	-74.03(0.32)	52.61(0.14)	-3.48(0.00)	1.59(0.02)	-6.42(0.44)
Y5H	15.23(0.28)	-74.10(0.20)	54.54(0.10)	-3.46(0.00)	1.70(0.01)	-6.09(0.22)
A8V	17.74(0.32)	-77.61(0.20)	54.61(0.12)	-3.48(0.00)	1.43(0.01)	-7.31(0.27)
F20Q	21.48(0.18)	-82.47(0.07)	55.03(0.04)	-3.49(0.00)	1.34(0.00)	-8.10(0.15)
A23Q	21.29(0.52)	-81.65(0.12)	54.70(0.12)	-3.49(0.00)	1.33(0.00)	-7.83(0.44)
L29Q	20.12(0.18)	-79.27(0.07)	54.40(0.04)	-3.49(0.00)	1.29(0.00)	-6.95(0.15)
A31S	15.43(0.57)	-74.07(0.34)	53.70(0.15)	-3.47(0.01)	1.65(0.02)	-6.76(0.45)
S37G	20.33(0.18)	-79.81(0.07)	55.04(0.04)	-3.49(0.00)	1.30(0.00)	-6.63(0.16)
E40A	16.19(0.56)	-71.84(0.36)	52.18(0.17)	-3.46(0.01)	1.50(0.03)	-5.42(0.45)
V44Q	19.21(0.58)	-78.35(0.29)	53.43(0.14)	-3.49(0.00)	1.33(0.01)	-7.87(0.48)
M45Q	18.80(0.54)	-79.11(0.25)	55.06(0.13)	-3.48(0.00)	1.33(0.01)	-7.41(0.45)
L48Q	20.15(0.53)	-79.39(0.20)	54.38(0.12)	-3.49(0.00)	1.29(0.00)	-7.07(0.46)
Q50R	20.02(0.55)	-77.74(0.22)	52.26(0.14)	-3.49(0.00)	1.33(0.01)	-7.62(0.47)
L57Q	10.39(0.47)	-65.03(0.26)	50.15(0.16)	-3.22(0.02)	1.98(0.02)	-5.74(0.38)
E59D	14.09(0.48)	-71.52(0.23)	52.48(0.13)	-3.47(0.01)	1.85(0.02)	-6.58(0.41)
I61Q	19.91(0.58)	-76.91(0.24)	52.62(0.13)	-3.49(0.00)	1.31(0.01)	-6.56(0.49)
D67A	14.19(0.29)	-73.05(0.37)	55.38(0.26)	-3.19(0.02)	1.43(0.01)	-5.25(0.21)
D73A	14.55(0.56)	-64.52(0.26)	46.64(0.14)	-3.45(0.01)	1.53(0.02)	-5.25(0.43)
D73N	19.01(0.56)	-69.72(0.23)	47.07(0.13)	-3.49(0.00)	1.30(0.01)	-5.82(0.46)
D75Y	14.10(0.18)	-72.17(0.15)	54.69(0.07)	-3.36(0.01)	1.61(0.01)	-5.13(0.14)
V79Q	19.33(0.17)	-75.25(0.08)	50.86(0.04)	-3.49(0.00)	1.31(0.00)	-7.24(0.15)
M81Q	14.62(0.25)	-71.44(0.14)	51.97(0.07)	-3.45(0.00)	1.67(0.01)	-6.63(0.19)
C84Y	18.65(0.19)	-78.28(0.09)	54.39(0.06)	-3.49(0.00)	1.34(0.00)	-7.39(0.15)
D87A	16.32(0.29)	-71.81(0.16)	51.69(0.11)	-3.47(0.00)	1.52(0.01)	-6.26(0.22)
D88A	14.37(0.33)	-74.49(0.21)	55.65(0.09)	-3.45(0.00)	1.68(0.02)	-6.24(0.27)

Table 3: **Calculated relative binding free energy ($\Delta\Delta G_{cal}$) of Ca^{2+} towards cTn-C diseased mutations are compared with experimental values given in kcal/mol.**

System	$\Delta\Delta G_{cal}$	$\Delta\Delta G_{exp}$
Y5H	0.33	0.12
A8V	-0.89	-0.1
F20Q	-1.68	-0.41
A23Q	-1.41	-1.03
L29Q	-0.53	-0.33
A31S	-0.34	-0.37
S37G	-0.21	-0.11
E40A	1.00	0.69
V44Q	-1.45	-1.63
M45Q	-0.99	-0.9
L48Q	-0.65	-1.07
Q50R	-1.2	-0.75
L57Q	0.68	0.57
E59D	-0.16	0.33
I61Q	-0.14	1.17
D67A	0.99	1.47
D73A	1.17	1.47
D73N	0.6	0.87
D75Y	1.29	0.42
V79Q	-0.82	-0.07
M81Q	-0.21	-0.72
C84Y	-0.97	-0.81
D87A	0.16	0.29
D88A	0.18	0.29

Table 4: **Calculated relative binding free energy ($\Delta\Delta G_{cal}$) value (in kcal/mol) of cTn-C plausible mutations.**

System	$\Delta\Delta G_{cal}$
N18A	-0.16
N18D	-1.21
N18E	-1.57
N18K	-0.47
N18Q	-0.61
N18S	0.27
A22E	-1.02
D25K	1.87
D25S	0.24
I26L	-0.44
I26M	0.45
V28D	-0.77
L29A	0.36
L29D	-0.75
L29E	-0.46
L29G	-0.12
L29H	0.38
L29K	-0.44
L29Q	-0.53
L29R	1.54
L29S	-0.32
L29T	-0.40
L29V	-0.12
G30D	-1.14
G30E	-1.0
G30N	-0.69
E32G	1.22
C35A	0.94
C35E	-2.02
C35K	-0.89
C35T	-0.64

Table 5: **continued...**

System	$\Delta\Delta G_{cal}$
S37T	-0.78
K43R	-0.23
K43T	1.95
K43V	0.05
M47A	0.20
M47S	-0.29
P54A	-0.79
P54D	0.77
P54E	0.34
P54K	-1.90
E59D	-0.16
D62A	0.25
D62K	-0.04
D62N	0.18
D62R	0.31
E66A	0.51
E66L	2.81
E66T	0.21
E66V	2.88
V72I	-1.18
V72L	-0.56
D73E	-0.24
D75E	-0.82
V79A	-0.05
V79K	0.31
V79L	-0.55
V79N	0.26
V79Q	-0.82
V79S	0.62
V79T	-1.20

Table 6: **Table represented sequence, structure and Ca^{2+} interaction based descriptors used for ML model.**

Feature type	**Features**
Sequence	Residue conservation
	Evolutionary statistical energy
	Charge of the system
	Position of mutation in protein sequence
Structure	Secondary structure of mutant site
	Distance of mutation from Ca^{2+}
	SASA
	Type of amino acid substitution
Ca^{2+} interactions	Number of water within $3\mathring{A}$ of Ca^{2+}

Table 7: **Table shows a combination of different types of features with their evaluation metrics such as R^2, RMSE, MAE, MSE, R_P and kernel used in SVR model.**

Descriptors combination	R^2	MSE	MAE	RMSE	pearson's	kernel
Sequence+Structure+Ca^{2+} interaction	0.77	0.16	0.30	0.41	0.92	poly
Sequence+Structure	0.51	0.35	0.51	0.59	0.80	rbf
Sequence+Ca^{2+} interaction	0.66	0.25	0.44	0.50	0.81	rbf
Structure+Ca^{2+} interaction	0.72	0.20	0.36	0.45	0.85	rbf
Sequence	0.34	0.47	0.57	0.69	0.59	rbf
Structure	0.57	0.31	0.44	0.56	0.79	linear
Ca^{2+} interaction	0.47	0.38	0.54	0.62	0.69	linear

Acknowledgement

...

Supporting Information Available

Detail of the parameters and additional plots and data are given in Supporting Information.

References

(1) Carafoli, E.; Krebs, J. Why calcium? How calcium became the best communicator. *Journal of Biological Chemistry* **2016**, *291*, 20849–20857.

(2) Gifford, J. L.; Walsh, M. P.; Vogel, H. J. Structures and metal-ion-binding properties of the Ca2+-binding helix–loop–helix EF-hand motifs. *Biochemical Journal* **2007**, *405*, 199–221.

(3) Swindle, N.; Albury, A. N.; Baroud, B.; Burney, M.; Tikunova, S. B. Molecular and functional consequences of mutations in the central helix of cardiac troponin C. *Archives of biochemistry and biophysics* **2014**, *548*, 46–53.

(4) Rayani, K.; Hantz, E. R.; Haji-Ghassemi, O.; Li, A. Y.; Spuches, A. M.; Van Petegem, F.; Solaro, R. J.; Lindert, S.; Tibbits, G. F. The effect of Mg2+ on Ca2+ binding to cardiac troponin C in hypertrophic cardiomyopathy associated TNNC1 variants. *The FEBS Journal* **2022**, *289*, 7446–7465.

(5) Pinto, J. R.; Siegfried, J. D.; Parvatiyar, M. S.; Li, D.; Norton, N.; Jones, M. A.; Liang, J.; Potter, J. D.; Hershberger, R. E. Functional characterization of TNNC1 rare variants identified in dilated cardiomyopathy. *Journal of Biological Chemistry* **2011**, *286*, 34404–34412.

(6) Parvatiyar, M. S.; Pinto, J. R.; Liang, J.; Potter, J. D. Predicting cardiomyopathic phenotypes by altering Ca2+ affinity of cardiac troponin C. *Journal of Biological Chemistry* **2010**, *285*, 27785–27797.

(7) Wang, D.; McCully, M. E.; Luo, Z.; McMichael, J.; Tu, A.-Y.; Daggett, V.; Regnier, M. Structural and functional consequences of cardiac troponin C L57Q and I61Q Ca2+-desensitizing variants. *Archives of biochemistry and biophysics* **2013**, *535*, 68–75.

(8) McConnell, B. K.; Singh, S.; Fan, Q.; Hernandez, A.; Portillo, J. P.; Reiser, P. J.; Tikunova, S. B. Knock-in mice harboring a Ca2+ desensitizing mutation in cardiac troponin C develop early onset dilated cardiomyopathy. *Frontiers in Physiology* **2015**, *6*, 242.

(9) Kreutziger, K. L.; Piroddi, N.; McMichael, J. T.; Tesi, C.; Poggesi, C.; Regnier, M. Calcium binding kinetics of troponin C strongly modulate cooperative activation and tension kinetics in cardiac muscle. *Journal of molecular and cellular cardiology* **2011**, *50*, 165–174.

(10) Smith, M.; Li, P. Molecular Insights into the Calcium Binding in Troponin C through a Molecular Dynamics Study. *Journal of Chemical Information and Modeling* **2022**, *63*, 354–361.

(11) Hantz, E. R.; Lindert, S. Adaptative steered molecular dynamics study of mutagenesis effects on calcium affinity in the regulatory domain of cardiac troponin C. *Journal of chemical information and modeling* **2021**, *61*, 3052–3057.

(12) Bowman, J. D.; Coldren, W. H.; Lindert, S. Mechanism of cardiac troponin C calcium sensitivity modulation by small molecules illuminated by umbrella sampling simulations. *Journal of chemical information and modeling* **2019**, *59*, 2964–2972.

(13) Tikunova, S. B.; Rall, J. A.; Davis, J. P. Effect of hydrophobic residue substitutions with glutamine on Ca2+ binding and exchange with the N-domain of troponin C. *Biochemistry* **2002**, *41*, 6697–6705.

(14) Tikunova, S. B.; Davis, J. P. Designing calcium-sensitizing mutations in the regulatory domain of cardiac troponin C. *Journal of Biological Chemistry* **2004**, *279*, 35341–35352.

(15) Rayani, K.; Seffernick, J.; Li, A. Y.; Davis, J. P.; Spuches, A. M.; Van Petegem, F.; Solaro, R. J.; Lindert, S.; Tibbits, G. F. Binding of calcium and magnesium to human cardiac troponin C. *Journal of Biological Chemistry* **2021**, *296*.

(16) Mazumder, M.; Kumar, S.; Kumar, D.; Bhattacharya, A.; Gourinath, S. Machine learning-based modulation of Ca2+-binding affinity in EF-hand proteins and comparative structural insights into site-specific cooperative binding. *International Journal of Biological Macromolecules* **2023**, *248*, 125866.

(17) Berchtold, M. W.; Zacharias, T.; Kulej, K.; Wang, K.; Torggler, R.; Jespersen, T.; Chen, J.-N.; Larsen, M. R.; La Cour, J. M. The arrhythmogenic calmodulin mutation D129G dysregulates cell growth, calmodulin-dependent kinase II activity, and cardiac function in zebrafish. *Journal of Biological Chemistry* **2016**, *291*, 26636–26646.

(18) Mazumder, M.; Padhan, N.; Bhattacharya, A.; Gourinath, S. Prediction and analysis of canonical EF hand loop and qualitative estimation of Ca2+ binding affinity. *PLoS One* **2014**, *9*, e96202.

(19) Franke, S.; Herfurth, J.; Hoffmann, D. Estimating affinities of calcium ions to proteins. *Advances and Applications in Bioinformatics and Chemistry* **2010**, 1–6.

(20) Boguta, G.; Stepkowski, D.; Bierzyński, A. Theoretical estimation of the calcium-binding constants for proteins from the troponin C superfamily based on a secondary structure prediction method: II. Applications. *Journal of theoretical biology* **1988**, *135*, 63–73.

(21) Kragelund, B. B.; Jönsson, M.; Bifulco, G.; Chazin, W. J.; Nilsson, H.; Finn, B. E.; Linse, S. Hydrophobic core substitutions in calbindin D9k: effects on Ca2+ binding and dissociation. *Biochemistry* **1998**, *37*, 8926–8937.

(22) Halder, R.; Warshel, A. Energetic and structural insights behind calcium induced conformational transition in calmodulin. *Proteins: Structure, Function, and Bioinformatics* **2024**, *92*, 384–394.

(23) Ababou, A.; Desjarlais, J. R. Solvation energetics and conformational change in EF-hand proteins. *Protein Science* **2001**, *10*, 301–312.

(24) Basit, A.; Mishra, R. K.; Bandyopadhyay, P. Calcium ion binding to calmodulin: binding free energy calculation using the molecular mechanics Poisson-Boltzmann surface area (MM-PBSA) method by incorporating implicit polarization. *Journal of Biomolecular Structure and Dynamics* **2021**, *39*, 7213–7222.

(25) Basit, A.; Yadav, A. K.; Bandyopadhyay, P. Calcium ion binding to the mutants of calmodulin: a structure-based computational predictive model of binding affinity using a charge scaling approach in molecular dynamics simulation. *Journal of Chemical Information and Modeling* **2022**, *62*, 2821–2834.

(26) Hopf, T. A.; Ingraham, J. B.; Poelwijk, F. J.; Schärfe, C. P.; Springer, M.; Sander, C.; Marks, D. S. Mutation effects predicted from sequence co-variation. *Nature biotechnology* **2017**, *35*, 128–135.

(27) Leontyev, I.; Stuchebrukhov, A. Accounting for electronic polarization in non-polarizable force fields. *Physical Chemistry Chemical Physics* **2011**, *13*, 2613–2626.

(28) Kohagen, M.; Lepsik, M.; Jungwirth, P. Calcium binding to calmodulin by molecular dynamics with effective polarization. *The journal of physical chemistry letters* **2014**, *5*, 3964–3969.

(29) Duboue-Dijon, E.; Javanainen, M.; Delcroix, P.; Jungwirth, P.; Martinez-Seara, H. A practical guide to biologically relevant molecular simulations with charge scaling for electronic polarization. *The Journal of Chemical Physics* **2020**, *153*.

(30) Kirby, B. J.; Jungwirth, P. Charge scaling manifesto: A way of reconciling the inherently macroscopic and microscopic natures of molecular simulations. *The journal of physical chemistry letters* **2019**, *10*, 7531–7536.

(31) Marshall, G. R. Limiting assumptions in molecular modeling: electrostatics. *Journal of computer-aided molecular design* **2013**, *27*, 107–114.

(32) Case, D. A.; Aktulga, H. M.; Belfon, K.; Ben-Shalom, I.; Brozell, S. R.; Cerutti, D. S.; Cheatham III, T. E.; Cruzeiro, V. W. D.; Darden, T. A.; Duke, R. E., et al. *Amber 2021*; University of California, San Francisco, 2021.

(33) Dolinsky, T. J.; Nielsen, J. E.; McCammon, J. A.; Baker, N. A. PDB2PQR: an automated pipeline for the setup of Poisson–Boltzmann electrostatics calculations. *Nucleic acids research* **2004**, *32*, W665–W667.

(34) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *Journal of chemical theory and computation* **2015**, *11*, 3696–3713.

(35) Humphrey, W.; Dalke, A.; Schulten, K. VMD: visual molecular dynamics. *Journal of molecular graphics* **1996**, *14*, 33–38.

(36) Tan, C.; Tan, Y.-H.; Luo, R. Implicit nonpolar solvent models. *The Journal of Physical Chemistry B* **2007**, *111*, 12263–12274.

(37) Kamisetty, H.; Ovchinnikov, S.; Baker, D. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence-and structure-rich era. *Proceedings of the National Academy of Sciences* **2013**, *110*, 15674–15679.

(38) Ooi, T.; Oobatake, M.; Nemethy, G.; Scheraga, H. A. Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proceedings of the National Academy of Sciences* **1987**, *84*, 3086–3090.

(39) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V., et al. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **2011**, *12*, 2825–2830.

(40) Lim, C. C.; Yang, H.; Yang, M.; Wang, C.-K.; Shi, J.; Berg, E. A.; Pimentel, D. R.; Gwathmey, J. K.; Hajjar, R. J.; Helmes, M., et al. A novel mutant cardiac troponin

C disrupts molecular motions critical for calcium binding affinity and cardiomyocyte contractility. *Biophysical journal* **2008**, *94*, 3577–3589.