Predicting Mutation-Induced Relative Protein-Ligand Binding Affinity Changes via Conformational Sampling and Diversity Integration with Subsampled Alphafold2 in Few-Shot Learning

Liangxu Xie[1,#], Xiaohua Lu[1,#], Dawei Zhang[1], Lei Xu[1], Mingjun Yang[2], Young-Seob Jeong[3], Shan Chang[1,*], Xiaojun Xu[1,*]

[1] Institute of Bioinformatics and Medical Engineering, Jiangsu University of Technology, Changzhou, 213001, China

[2] Shenzhen Jingtai Technology Co., Ltd. (XtalPi), Shenzhen 518045, China

[3] Department of Computer Engineering, Chungbuk National University, Cheongju, 28644, South Korea

[#]Equivalent authors

Correspondence: S.C.: schang@jsut.edu.cn; X.X.: xuxiaojun@jsut.edu.cn

Abstract

Predicting the impact of mutations on protein-ligand binding affinity is crucial in drug discovery, particularly in addressing drug resistance and repurposing existing drugs. Current structure-based methods, including deep learning and physics-based computational techniques, are constrained by their dependence on known protein-ligand complex structures. The lack of structural data for mutated proteins, coupled with the potential for mutations to alter protein conformational states, poses challenges for reliable predictions. To address the challenge, we propose a few-shot learning approach that integrates AlphaFold2 subsampling, ensemble docking, and Siamese learning to predict the relative binding affinity changes induced by mutations. We benchmark our approach for predicting binding affinity changes in 31 variants of the mutated Abelson tyrosine kinase (ABL). The conformation ensembles for each ABL variant are generated using AlphaFold2 subsampling. Subsequently, the most likely conformational states are selected through Gaussian fitting to construct the pairwise structural inputs and compute relative binding energies. To leverage the structural information, we utilize the multimodal inputs including contact maps and molecular graphs as the molecular representation in the Siamese learning network. The relative binding affinities are derived by averaging the values predicted by the Siamese learning model across the constructed conformational pairs. Benchmarking on the Tyrosine kinase inhibitors (TKI) dataset and the direct binding affinity using the refined set of PDBbind database, our proposed approach achieves increased prediction accuracy in estimating relative binding affinities by using the conformations generated from AlphaFold2 subsampling, which includes structural diversity in the prediction and avoids reliance on crystal structures. We anticipate the approach will be

valuable for predicting binding energy changes induced by mutations in proteins or ligands, facilitating drug discovery efforts.

Introduction

Proteins are responsible for life-essential specific functions such as cellular structural support, immune protection, enzyme catalysis, cell signaling to transcription and translation functions[1, 2, 3, 4, 5]. These functions are closely related to structural changes in the proteins. While protein sequence information is abundant in the UniProtKB database, the experimentally determined three-dimensional structures remain limited due to the complexity and expense of techniques like X-ray crystallography, nuclear magnetic resonance, and electron cryo-microscopy (cryo-EM). Consequently, prediction of protein structures through computational methods has become crucial[6, 7, 8]. Deep learning techniques, notably AlphaFold2 (AF2), have revolutionized the field of protein structure prediction. AF2, the second version of AlphaFold, provides prediction accuracies close to experimentally determined methods when it first appeared in the 14th Critical Assessment of Protein Structure Prediction (CASP14)[9]. Following the success of AF2, many predictors have been generated such as RoseTTAFold[10], ColabFold, OpenFold[11]; ESMfold[12], RGN2[13], and OmegaFold[14]. The success of AlphaFold2 brings new possibilities for many downstream explorations, such as analyzing missense variants in genome sequences[15], predicting protein-ligand molecule interaction in drug discovery[16, 17], protein engineering in synthetic biology[18, 19], and metallic pathway analysis for biological systems[20].

While it effectively predicts static protein structures, the default AF2 has limited capability in capturing the dynamic behavior and conformational changes of proteins. Proteins are not static and undergo conformational changes upon binding of a specific ligand, whereby the side chains of residues in the pocket adjust their conformation to optimize their interaction with the ligand[21]. Thermodynamic fluctuations in proteins can trigger dynamic structural changes[22]. These suitable conformations may become stabilized upon ligand binding, providing important clues for the design of highly selective drugs. The dynamics of proteins have been realized by recent work on protein design to capture the excited states besides ground states[23]. Many efforts have been devoted to optimizing AF2 to capture different conformations. The AF2 model constructs protein 3D structures in two stages by using Multiple Sequence Alignment (MSA) and initial pairwise distance measurements as inputs[24]. Transformer-based Evoformer network layer is used to process the MSA and distance information to extract the co-evolutionary features. MSA encodes coevolutionary information of proteins, therefore manipulation of MSA becomes a reasonable way to alter the conformations by controlling coevolutionary information. Successful approaches have been reported for predicting conformational ensembles of proteins, for example, AF-Cluster via sequence clustering[25], subsampling MSA[26], energetic frustration analysis of MSA[27], SPEECH_AF via in silico mutagenesis of MSA[28], randomized alanine sequence scanning[29], etc. Leveraging sampling techniques, the AF2 model has been adopted to predict membrane proteins and G protein-coupled receptors (GPCRs) in different conformations by taking subsampling MSA to drive AF2 to sample multiple conformations[30]. A recent successful case has applied AF2 to directly predict the relative populations of different protein conformations by subsampling the comparison of multiple sequences[26]. The analysis of MSA has also been used to analyze patterns of relevant mutations within protein families[31, 32], predict interacting residues based on sequences[33, 34], and construct structure prediction

models guided by energy-based processes[35, 36].

Predicting conformational ensembles can significantly enhance drug discovery, particularly when protein targets exhibit multiple possible conformational states. After predicting the structures of proteins, binding affinity can be predicted for assessing the strength and stability of small molecules binding to protein targets. With advances in databases and computational power, predictions can be made using physics-based methods and deep learning methods[37, 38, 39, 40]. However, the binding sites of proteins in different conformational states can be significantly different, which can affect the binding mode of the drug. Therefore, the challenge has changed from "Is there a protein structure available?" to "How to appropriately select a proper structure?". To avoid biasing towards single protein conformation and implicitly including protein flexibility, the ensemble of representative structures is used to dock candidate ligands[41]. Accounting for protein flexibility is crucial for accurate binding affinity prediction, as exemplified by Abelson tyrosine kinase (ABL). ABL kinase undergoes significant structural changes in its active and inactive states, which can be selectively bounded by different types of inhibitors. ABL kinase plays a key role in multiple cell signaling pathways that regulate cell proliferation, differentiation, and apoptosis.[42, 43, 44]. While Tyrosine kinase inhibitors (TKIs) like imatinib have revolutionized chronic myeloid leukemia (CML) treatment [45], drug resistance emerges through mutations that alter the conformation, and consequently affect the drug binding mode and binding affinity[46, 47]. The scarcity of mutated types of kinase limits the accuracy of physics-based methods and machine/deep learning methods[48].

The mutation-induced structural changes bring two limitations for the traditional computational methods: (1) the determination of the mutation-induced structures of proteins, and (2) the selection of the most rational structure from conformation ensembles. A real-world scenario in drug development involves understanding the resistance profiles of certain drugs against various mutated structures[49]. Take ABL kinase for example, mutations in ABL alter the relative population of the active state and inactivate state, and both states can be selectively targeted by the different types of inhibitors[50]. Several states of ABL are revealed through molecular dynamics simulation, which are structurally similar to known crystal structures capable of binding a variety of inhibitors[51]. Consequently, the conformation ensemble of kinase proteins should be incorporated into the prediction of binding free energy changes induced by mutations. The accuracy of AF2 can extend to the resolution of individual amino acids, enabling precise predictions of protein structures at an atomic level[15, 52]. Here, we propose an integrated approach to solve the two above-mentioned limitations by streamlining AF2 subsampling, ensemble docking, and Siamese learning. AF2 has learned an approximate biophysical energy function that can be used to assess the quality of mutated protein structures with high accuracy[53]. Subsampled AF2 predicts the conformational landscape and the relative populations of conformations at single mutation resolution[26]. To include structural diversity, we adopt data augmentation through subsampled AF2 and ensemble docking. We benchmark our proposal by investigating drug resistance in ABL kinase. The conformation ensembles of ABL kinase were

generated using subsampled AF2, and this generated data is paired with experimental values for data augmentation. Conformation selection is achieved through ensemble docking and Gaussian fitting filtering strategy. Siamese network has been used to predict relative binding affinity for the static structures[54]. We include three structural features in Siamese learning: the contact map from 3D structures of proteins, and molecular graph from both ligand and protein-ligand interactions. To efficiently process the structural features, we propose an STGNet (Structure and graph-aware Siamese network) model. Besides, we compare the statistical relative value prediction method using our proposed method and the direct absolute value prediction method. The results demonstrate that our method improves the accuracy and stability of prediction. Through RMSD calculations and clustering analysis of the data augmentation methods, we observe that different TKIs exhibit different preferences for protein conformations, thereby validating the rationale for selection based on Gaussian fitting. In few-shot learning, our approach demonstrates higher ranking accuracy in predicting binding free energy changes induced by mutations, without relying on crystal structures.

Results

Construction of augmented pairwise dataset and deep learning architecture

Predicting relative properties has significant applications in drug discovery, for example, changes in absorption, distribution, metabolism, excretion, and toxicity (ADMET) during drug lead optimization[55, 56], as well as in assessing relative binding affinities between congeneric ligands[57]. Various neural network models have been proposed for this purpose as illustrated in **Fig. 1**. The relative properties can be computed using a standard neural network, where the output is derived from the difference between two predicted values. The delta model is a specialized architecture that focuses on learning these output differences[56]. The Siamese architecture is designed to compare two inputs by generating embeddings, with the output defined as the absolute difference between the learned representations, thus capturing subtle distinctions between similar inputs[54]. The pairwise Siamese network further expands on this concept by integrating pairs of representations and calculating average values across various input aspects[55]. These models are particularly useful for computing relative binding affinities in protein-ligand complexes, which are influenced by ligand structure, protein mutations, and conformational variations during interactions. A commonly employed method for predicting relative binding affinity is the free energy perturbation method[58], which illustrates energy changes between a reference state and a query state. In our study, we expand the scope of perturbation analysis by shifting from comparing absolute values to computing differences between relative values, as illustrated in **Fig. 1F**. Specifically, we aim to analyze the relative changes in binding energy induced by mutations, comparing experimental data with predicted values. To achieve this, we will employ deep learning techniques to establish a mapping between relative scores and the corresponding relative binding energies associated with mutations. To enhance the reliability of our predictions, we also implement a pairing and averaging process as shown in **Fig. 1G**. This method integrates the contributions from both

multiple conformational states and ligand variations to derive a consensus binding score, thereby improving the accuracy of our predictions.

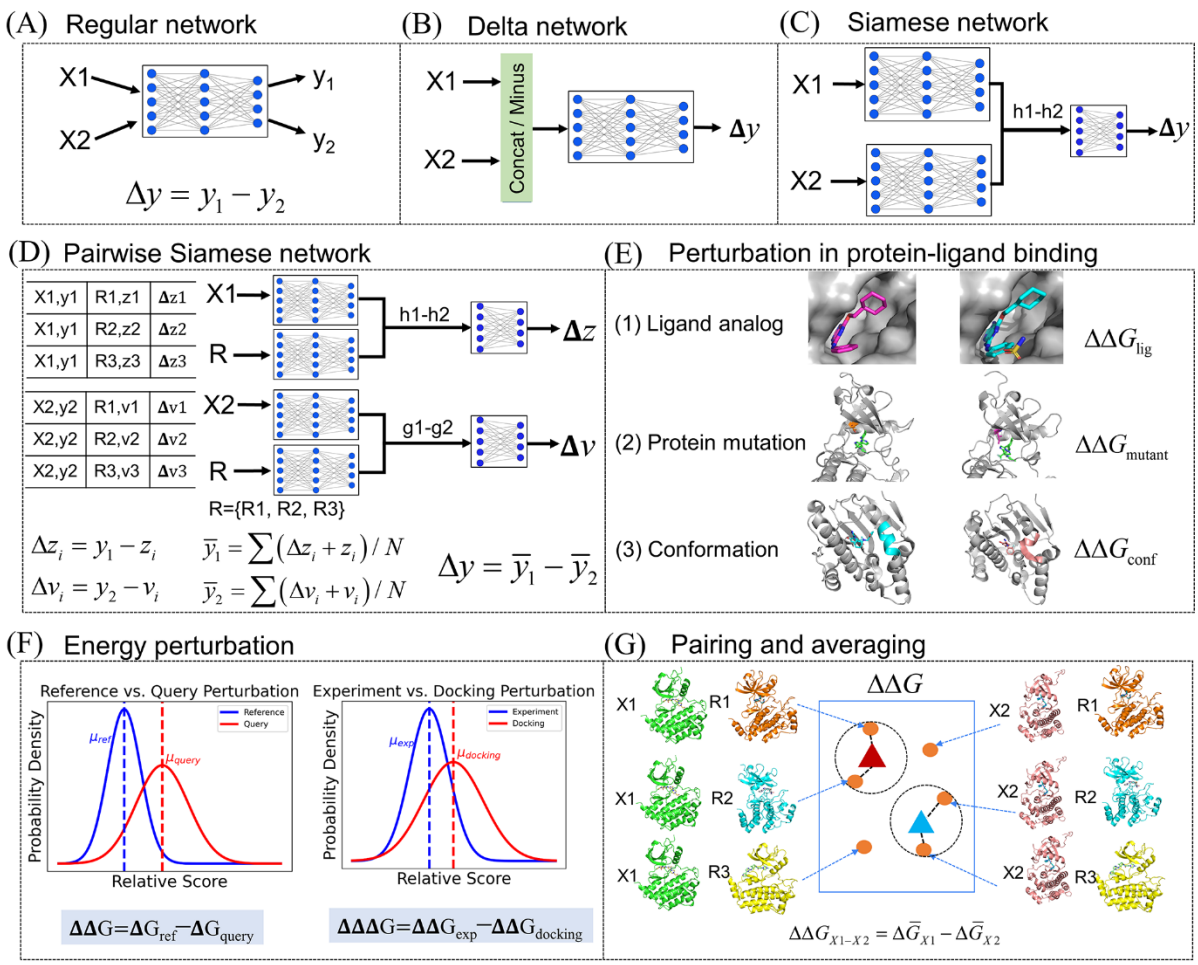

Fig. 1 Models for relative property prediction. (A) Regular neural network. (B) Delta neural network. (C) Siamese network. (D) Pairwise Siamese network. (E) Perturbation in protein-ligand binding includes variations of ligand analogs, protein mutation, and conformations. (F) Two types of energy perturbation. (G) Pairing and averaging with the most populated states when estimating the relative binding free energy. X1 and X2 represent the input, and y_1 and y_2 are the output. R1, R2, and R3 represent the reference states.

The augmented dataset is constructed by using AF2 subsampling and ensemble docking (refer to the method section for details). In the workflow illustrated in **Fig. 2**, we selected 31 mutants of ABL kinase and the 6 TKI molecules to construct our dataset. The aim is to construct reference-generated conformation pairs for the referred mutant. The reference value is from the TKI dataset[59]. The generated value is predicted from subsampling AF2 for 31 mutants and wild-type of ABL kinase. The conformations from each type of protein were docked with six TKI molecules using AutoDock Vina following the procedure of ensemble docking, thus obtaining docking score values for all mutants and wild-type conformations. The relative docking score $\Delta\Delta G_{dock}$ is computed from the docking score of mutants and the wild type. We selected the most likely conformations based on the distance between the relative docking

scores and the Gaussian mean value. The selected relative docking score will be paired with the experimental value. The relative binding affinity for each experimental value will be augmented to a specified number of values. Then, we can obtain the relative binding free energy difference (referred to as $\Delta\Delta\Delta G_{cal}$) by subtracting the experimental value from the relative docking score. $\Delta\Delta\Delta G_{cal}$ represents the binding energy change induced by conformational variations rather than mutations. By specifying the type of mutation, the pairs do not account for other mutational effects; instead, they focus on the differences in conformation and the variations in bound ligands. The advantage of using the relative binding affinity difference is that it allows Siamese learning to focus on the difference between the reference state and the generated conformational states.

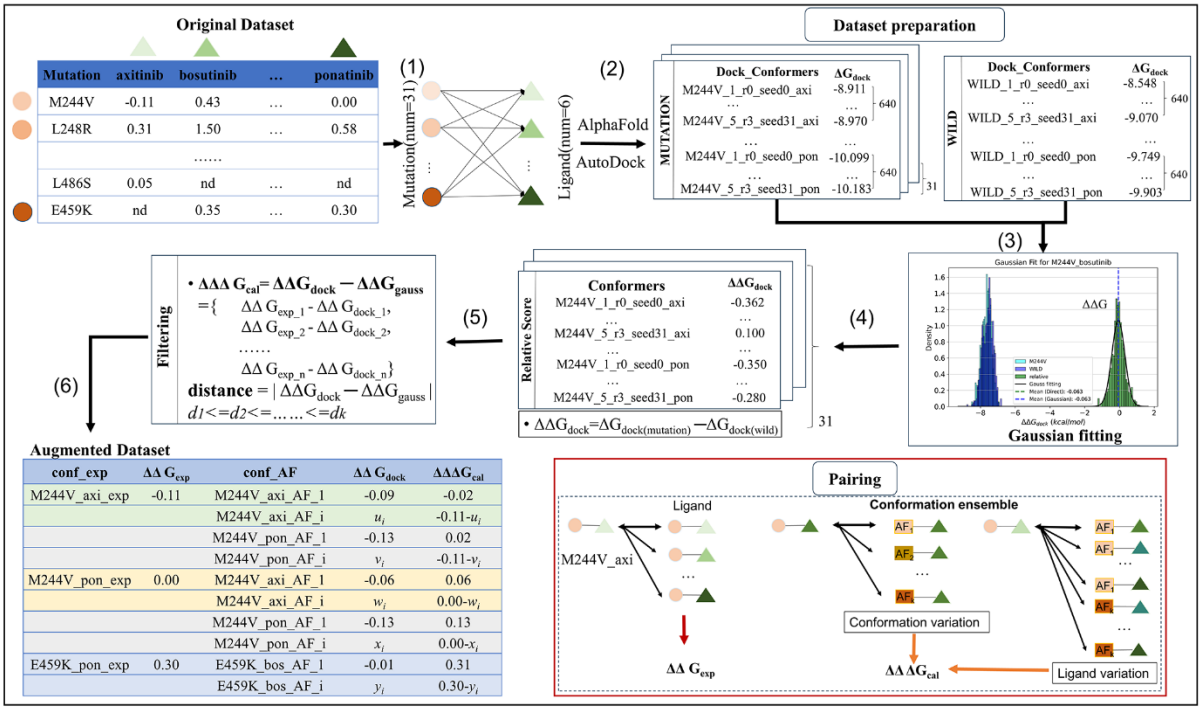

Fig. 2 Pipeline of construction of the augmented pairwise dataset. The workflow includes the creation of initial mutant-ligand data pairs, conformational sampling, ensemble molecular docking, relative binding energy calculation, data filtering, and construction of the final dataset. The red line highlights the pairing between the reference state and the query states, including the conformation variation and ligand variation.

Data augmentation has been demonstrated to enhance the prediction of binding affinities, offering improved accuracy and robustness[39]. The effect of single-conformation and multi-conformation data on the predictive ability of the model is analyzed by controlling the number of conformations by taking the number of conformations with k=1, 5, and 50, respectively. In the original dataset, there is one relative binding affinity for each protein-TKI pair. Through this workflow, we augmented the dataset by k-fold by pairing the experimental data with the generated data. As a result, we include the structural diversity for each protein-ligand pair by conformation variation and ligand variation. After obtaining the augmented dataset, the dataset

is partitioned into training and test sets based on the names of different TKIs for cross-validation. The splitting method reflects real-world scenarios in which we have established knowledge of protein mutations and their associated changes in binding free energy. Leveraging this prior knowledge, we can predict the relative binding free energy of a new drug molecule across all mutations. More details are described in Methods and Fig. S1 in supporting information. The investigated TKI molecules display different similarities and binding types as shown in Fig. S2. All molecules exhibit low similarity, which ensures that pairing with the experimental values does not lead to information leakage.

Similar pairwise approaches have been employed to predict relative binding free energies[54, 57, 60]. Our approach differs from existing work in two key aspects. First, previous work, for example, pairwise binding comparison network (PBCNet), primarily focuses on the structural analogs of ligands while conserving identical pockets of proteins. Their models predict the binding energy changes induced by variations in ligands. In contrast, our approach emphasizes mutations in proteins. Our method can be considered as the extension of PBCNet, specifically aimed at predicting binding energy changes resulting from mutations in proteins. Second, we propose using the relative binding free energy difference as the prediction target rather than relative binding free energies. This allows us to train the Siamese learning model to more effectively capture the differences among various conformations.

To harness the structural information, the input for our Siamese learning includes contact maps of each conformation, molecular graph of ligand, and molecular graph of protein-ligand interaction as shown in **Fig. 3**. The protein contact map with BLOSUM62 matrix scores[61] can reflect the spatial contacts between CA atoms of each residue in the three-dimensional (3D) structure of a protein. The ligand graph contains the chemical properties within the molecule. The protein-ligand interaction graph considers covalent interaction between intramolecular and non-covalent atomic-level interaction information of intermolecular. Three types of features capture both physical and spatial interactions between proteins and ligands. To learn the structural features, we constructed the STGNet (Structure and graph-aware Siamese network) model to learn the structural information using the Message Passing Neural Networks (MPNN) network, and to learn interaction information using the Geometric Interaction Graph Neural Network (GINN) network. MPNN has graph isomorphism invariance for better handling of graph structure data and can extract local as well as global information of nodes and edges through multi-layer message passing and aggregation. The protein contact map and ligand graph are learned using MPNN networks in the Siamese architecture, which helps to capture the 3D folding structures of the protein as well as the internal chemical properties of the ligand. GIGN can learn the binding patterns between proteins and ligands, by combining the properties of the nodes and the 3D coordinate information. The edge features in the protein-ligand interaction graph include not only covalent bonds between proteins and ligands but also non-covalent bonds across molecules. By using MPNN and GIGN together, the model can focus on intramolecular and intermolecular features, respectively. The experimental value and the paired AF2-generated data will be input into the shared Siamese Network. Therefore, the network can better learn the difference between the reference state and the other conformational state, and provide the prediction of relative binding affinity difference.

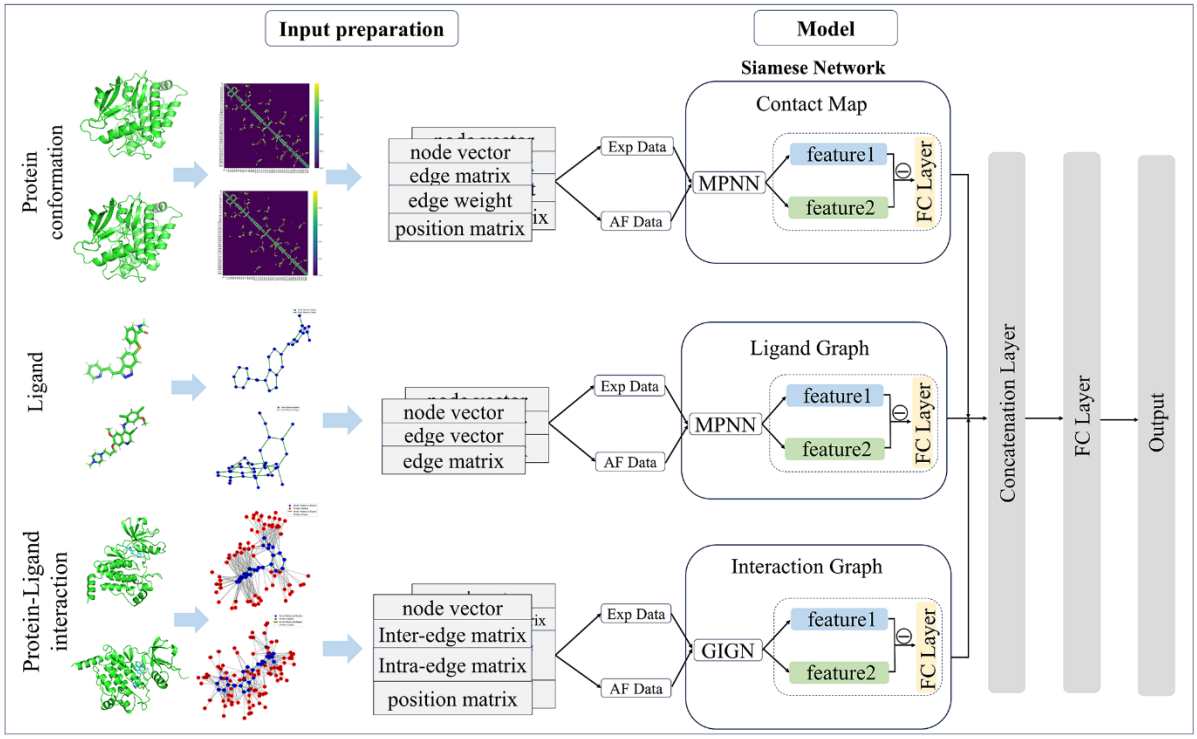

Fig. 3 Architecture of the STGNet model. The reference data and the AF2-generated data are processed by the Siamese network to capture the feature difference. Protein contact maps and ligand graphs are processed by MPNN, and protein-ligand interaction graphs are processed by GIGN. Ligands show as blue nodes and edges, and proteins show as red nodes and edges in molecular graph.

Conformation sampling for ABL kinase using subsampled AF2

Conformational changes in ABL kinase have important implications for their function and drug resistance[62]. Mutations induce obvious structural changes in the P-loop and A-loop region, thereby resulting in the change of binding affinity of the drug to ABL kinase by interfering with the stability of specific conformations (e.g., DFG-in and DFG-out)[63]. The mutation effect on drug resistance has been investigated by integrating structural properties with evolutionary data[48, 64]. Structural difference analyses for type I (DFG-in) and type II (DFG-out) inhibitors have shown that these two classes of drugs have significantly different sensitivities to conformational changes in ABL kinase, with type II inhibitors generally being more resistant to the effects of mutations. These studies indicate it requires conformational changes when investigating drug resistance. **Fig. 4** demonstrates three conformational states of ABL kinase, including the active state, inactive state I_1, and inactive state I_2. In the active state, P-Loop and A-Loop are in the open state and αC-Helix is in the inward state. Inactive state I_2 shows obvious structural changes with activate state. The αC-Helix is shifted from the in position to the out position, with the P-Loop becoming closed, and the A-Loop is partially folded. The inactive state I_1 shows fewer conformational changes to the active state and αC-Helix and P-Loop change to a small extent. This suggests that the conformational changes of ABL kinase between the active and inactive states are characterized by a remarkable dynamic equilibrium that is

critical for its functional state and drug binding[65]. Mutations may indirectly reduce the availability of drug-binding sites by stabilizing a certain conformation state and altering the percentage of the active state of the kinase.

AF2 has been demonstrated to predict the structures at the resolution of individual amino acids[15, 52]. The previous study also presents a conformational dynamic equilibrium between the active and two inactive states (I_1 and I_2) of ABL kinase achieved by subsampling MSA[26]. Subsampled AF2 can efficiently identify key conformations of ABL kinase, including active state and inactive state, which are critical for mechanistic studies of drug binding and resistance. We adopted the same procedure of subsampling MSA and obtained the conformation ensemble for all 31 mutants. The deviation of the different conformation distributions of the mutant (M244V) and the wild type (WILD) relative to the active and inactive states I_2 are quantified by computing root mean square deviation (RMSD) distribution, which has been used to characterize the ground state and inactive I_2-like state in the previous study[26]. **Fig. 4B** depicts the RMSD distribution densities of the loop regions of 31 mutants and the wild-type (WILD) in different conformations. It can be seen that the main peaks of the RMSD densities of the wild-type are concentrated at RMSD < 3.5 Å, consistent with previous study[26]. A different distribution suggests that the mutation alters the conformational balance of the ABL kinase. The percentage of active state is 83%, which is close to the experimental value of 88%. The subsampling can reproduce the relative population of ABL kinase and other populations of mutants, and the relative population is summarized in Table S1.

To include structural diversity in estimating binding affinity, we conduct the ensemble docking for all conformations. The distribution of docking score for mutant-ligand (M244V_axitinib) and the wild-type (WILD_axitinib) are shown in **Fig. 4C-D**. The Gaussian fitting indicates the binding free energy distribution follows the principle of maximum likelihood. Other mutations also demonstrate the Gaussian distribution as shown in Figs. S3-S8. **Fig. 4E** confirms the Gaussian distribution of relative binding free energy distributions $\Delta\Delta G_{dock}$ of the M244V mutant and the wild-type of ABL kinase, which is validated by the Sharipo-Wilk test. The docking distribution can be conceptualized as a result of a perturbation applied to the experimental distribution as shown in **Fig. 1F**. In this context, we aim to determine the relative energy changes induced by mutation through the use of conformation ensembles. Based on the principle of maximum likelihood, we select the k conformations and their relative binding affinities by minimizing the deviation of these affinities from the Gaussian mean value. This approach utilizes statistical principles to replace the manual selection of conformations. The proposed filtering approach offers a statistical framework to select the most populated states, rather than merely prioritizing the best conformation for scoring.

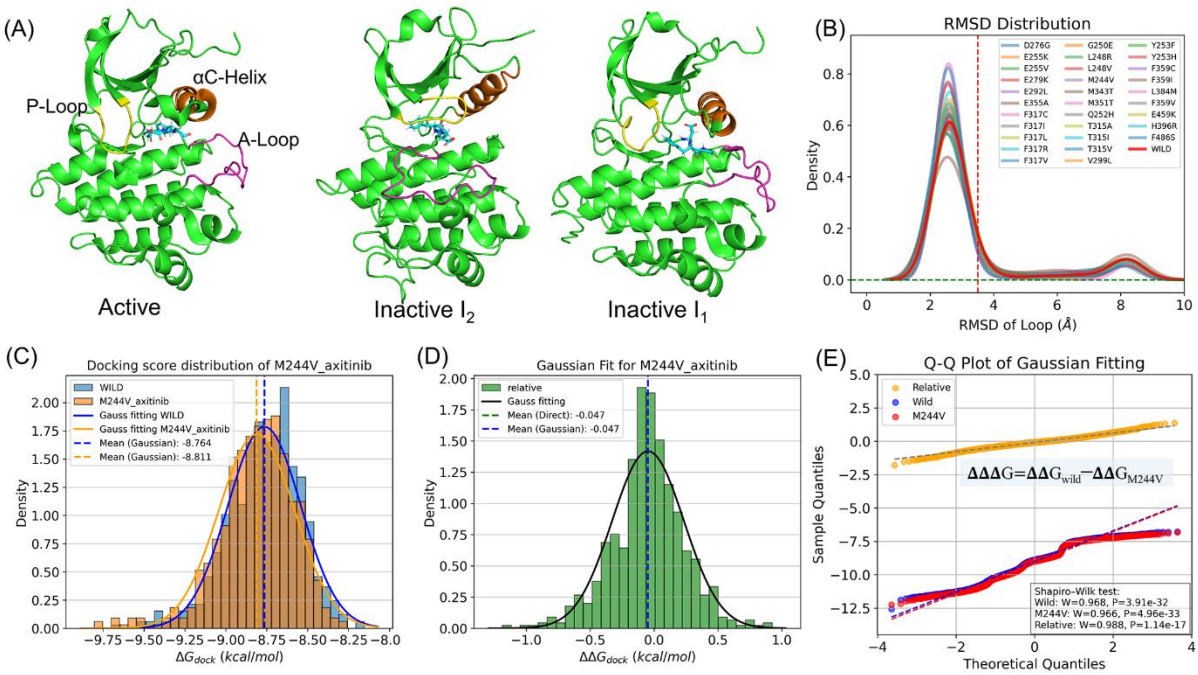

Fig. 4 Mutation effect on the conformation of ABL kinase and the drug binding affinity. (A) ABL kinase conformation in active state (PDB ID: 6XR6) and two inactive conformational states I_1 and I_2 (PDB ID:6XR7, 6XRG). (B) RMSD distribution of the loop regions (A-loop, P-loop, and C-helix) for wild type and 31 mutants. (C) Distribution of AutoDock Vina docking score of the wild and mutant (M244V) of ABL kinase. (D) Density distribution plot based on the relative binding energy obtained from the mutant (M244V) and the wild type. (E) The Quantile-Quantile Plot for the docking score and the relative docking score for the mutant (M244V) and the wild type.

Improved Predictive performance over physics-based methods (docking and FEP)

Prediction of relative binding free energy delivers import value in drug discovery applications. However, accurate prediction is limited due to technical and practical challenges in physics-based modeling and statistical mechanics methods[66]. Two physics-based methods (molecular docking and free energy perturbation (FEP+)) were compared to our SGTNet model. The predicted relative binding affinity and the experimental value are shown in **Fig. 5**. Molecular docking can provide a quick ranking, but it fails to reproduce the relative binding free energy between them. FEP+ can reproduce the relative binding affinity at the chemical accuracy, but it requires high computational resources. For our method, we provide the relative binding affinity for all experiment-generated pairs, then we compute the statistical average for the required protein-ligand pair, which enhances the stability and reduces the impact of the outlier.

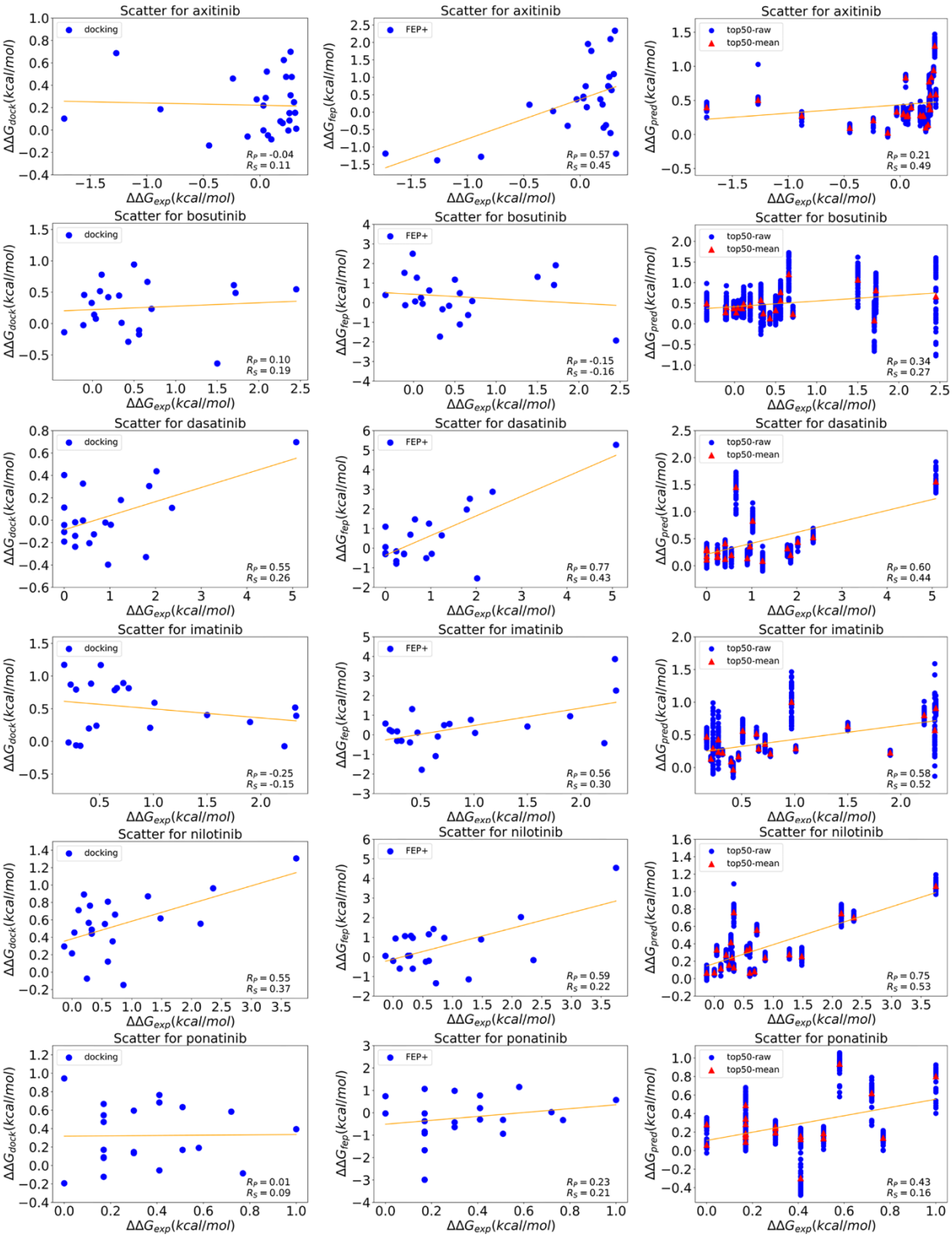

Fig. 5 Comparison of experimental and predicted relative binding energies for six TKI molecules. The predictions were generated using molecular docking, FEP+, and STGNet (denoted as top50-raw for the raw prediction and top50-mean for the average across the 50 data points derived from pairs sharing the same reference state). Each blue sphere represents one prediction and the red triangle represents the average value.

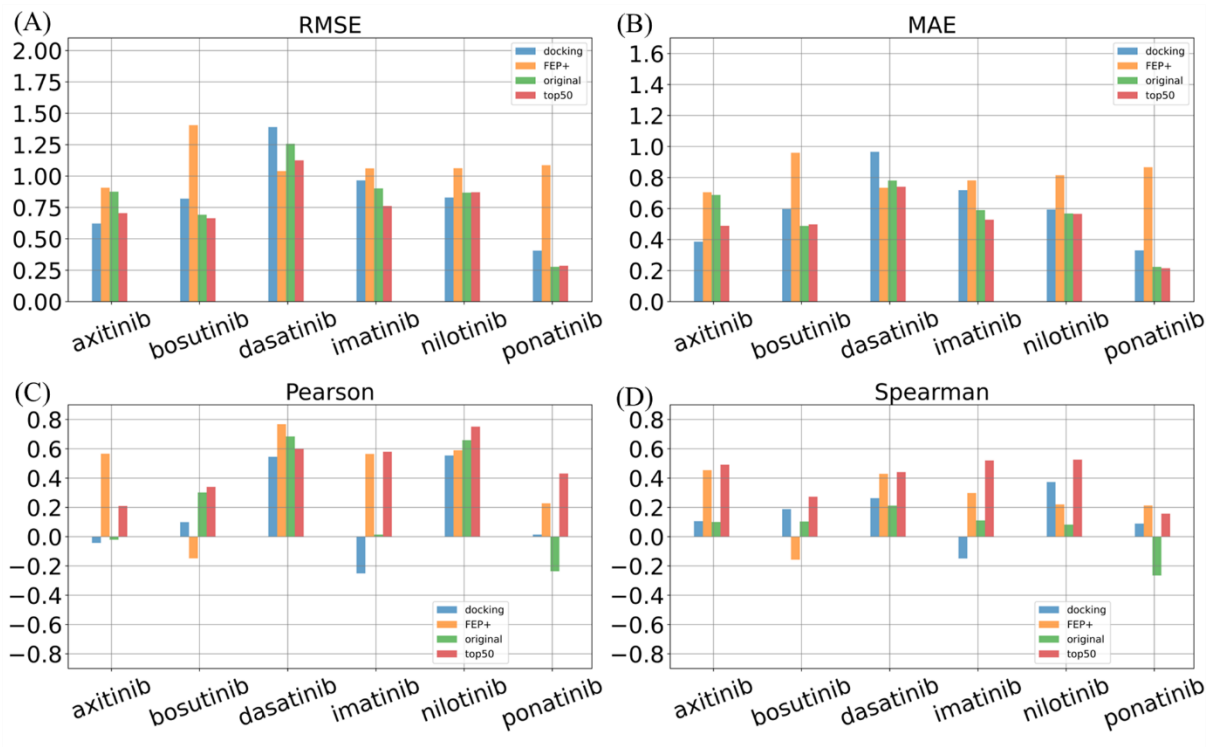

Fig. 6 Evaluation metrics of investigated methods. Performance is evaluated by RMSE, MAE, Pearson coefficients, and Spearman coefficients for docking, FEP+, STGNet using the original TKI dataset (denoted as original) and using top50 pairwise samples (denoted as top50).

The performance of the models was evaluated by calculating RMSE, MAE, Pearson, and Spearman coefficients, which are summarized in **Fig. 6** and Table S2-S5. It can be seen that the docking method has limited performance for protein-ligand relative binding free energy calculation. The docking-based method shows a low correlation between the axitinib, bosutinib, and ponatinib datasets, and the Pearson and Spearman coefficients for these datasets. The FEP+ method calculates the free energy based on molecular dynamics simulation. The FEP+ results obtained from Hauser et al[59], are superior to that of molecular docking, with the Pearson and Spearman coefficients for the axitinib, dasatinib, imatinib, and ponatinib significantly higher than those of molecular docking. The STGNet, using top50 data-enhanced dataset, achieves better performance on RMSE, MAE, and Pearson coefficients, obtaining the best values for four out of six TKIs: bosutinib, imatinib, nilotinib, and ponatinib. The Spearman correlation coefficient is more suitable than Pearson for ranking data points with multiple identical values. Our method outperforms both molecular docking and FEP+ methods, as evidenced by the highest Spearman correlation coefficients obtained for axitinib, bosutinib, dasatinib, imatinib, and nilotinib, highlighting its enhanced ranking performance. By introducing a multi-conformational free energy distribution filtering strategy and applying Siamese learning, our method significantly improves the accuracy and generalization ability in predicting the relative binding free energy changes induced by mutation. The relative docking score is utilized to describe the effects of mutations and is paired with the experimental reference. The relative binding energy is predicted based on structural data and is subsequently recalibrated to achieve true experimental accuracy.

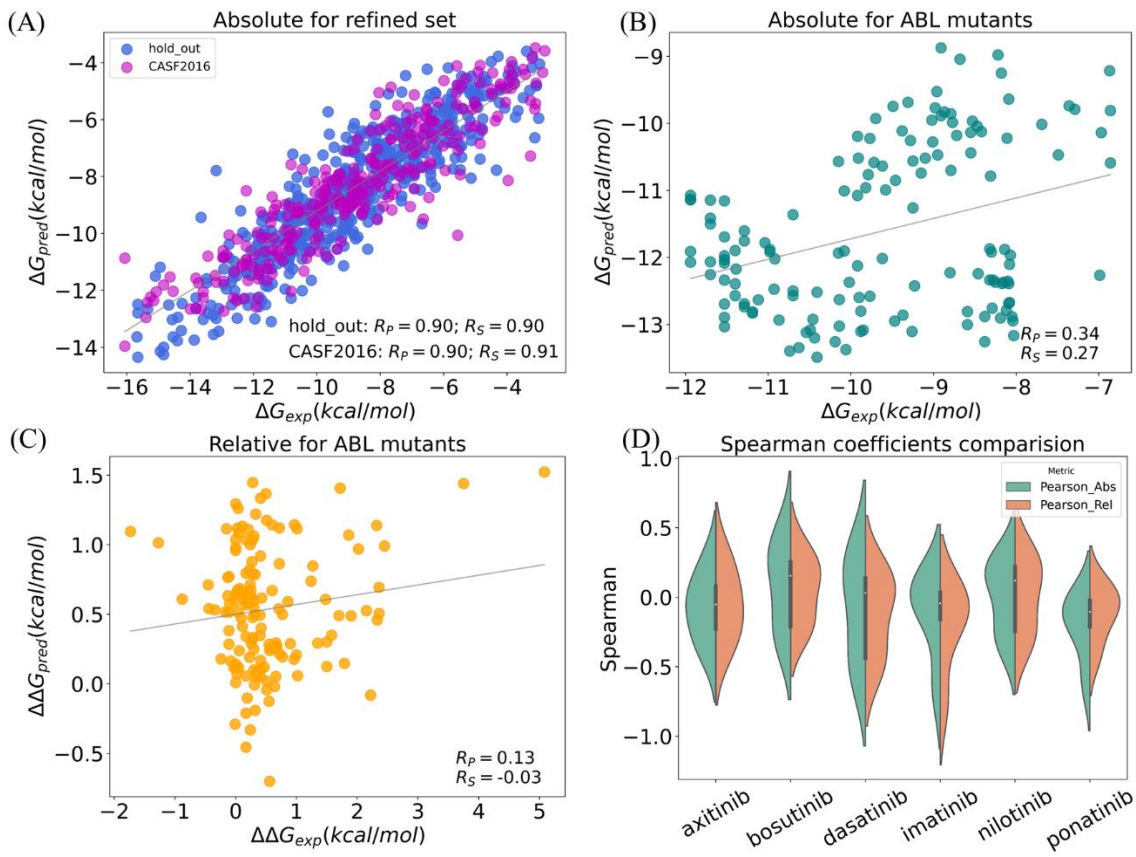

Fig. 7 Performance of regular network on the refined set and TKI set in the absolute binding energy prediction. (A) Correlation between the absolute values of the prediction and experimental results in the CASF-2016 set. (B) Correlation between the absolute values of prediction and experimental results in the TKI set. (C) Correlation between the relative values of prediction and experimental results in the TKI set. (D) Spearman correlation coefficients for both absolute values and the relative values of predictions compared to experimental results.

Relative Value Prediction in Siamese Learning Outperforms Absolute Value Prediction

We compare the data augmentation method with the direct absolute prediction of binding energy using the refined set of PDBbind. The refined set of PDBbind provides more diverse structures and binding affinity than the TKI dataset. The model for the refined dataset is shown in Fig. S9, which was trained using a similar network, except for removing the Simase network. The model predicts the absolute value of the binding free energy. We first benchmarked the molecular representation using the regular network. The results, as shown in **Fig. 7A**, demonstrate that our model achieves satisfactory performance in predicting absolute binding affinities, with strong correlations (Pearson and Spearman coefficients are 0.90, and 0.90 for the hold-out set, and for the CASF-2016 set). However, the model exhibits limitations in predicting both absolute and relative binding energies for the ABL mutants, as illustrated in **Fig. 7B and 7C**. The relative binding energies are computed by subtraction for the same pairs. The model cannot accurately capture the effects of single mutations in proteins, as evidenced by the low correlation coefficients. The Spearman coefficients in **Fig. 7D** support this

observation, highlighting the regular model's inability to reliably predict relative binding energy changes induced by single mutations.

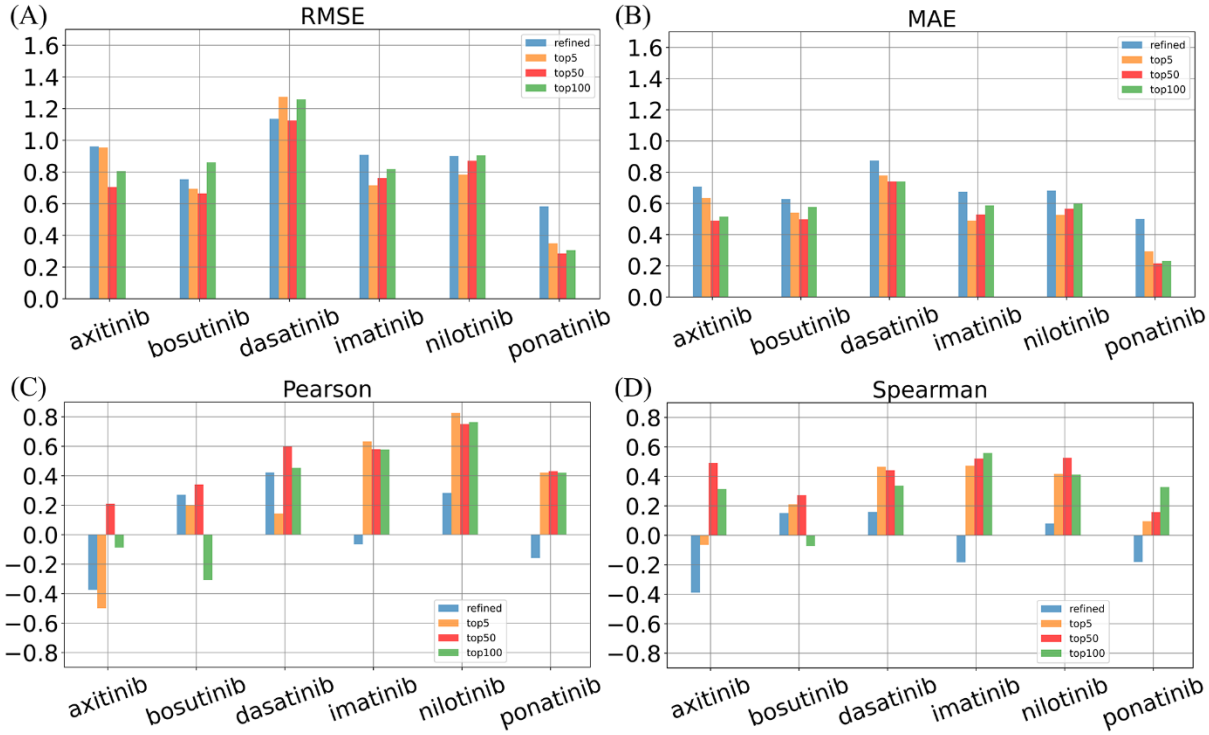

Fig. 8 Evaluation metrics of the investigated methods. Performance is evaluated using RMSE, MAE, Pearson coefficients, and Spearman coefficients for a refined set of PDBbind, and our method is evaluated by using 5, 50, and 100 samples.

The predictive performance of the model using a refined set was limited by the lack of comprehensive consideration of conformational dynamics and diversity. We compare the regular model with our STGNet model. In contrast, the data-enhanced method provides more comprehensive input data for each mutation-ligand pair by introducing multi-conformation information generated by AF2 subsampling. As can be seen in **Fig. 8**, the RMSE and MAE of STGNet using the different numbers of samples (5, 50, and 100 are denoted as top5, top50, and top100) were lower than those of the refined dataset for five out of six molecules, including axitinib, bosutinib, dasatinib, nilotinib, and ponatinib. STGNet presents higher Pearson coefficients than that of the refined set. Combining experimental and AF2 data, the STGNet method is better able to capture the dynamic changes of proteins in different conformations and the effects of mutations on these changes. Overall, our method comprehensively improves the prediction performance of mutation effects by integrating the information of multiple conformations, introducing a reasonable filtering strategy, and combining the multidimensional feature extraction of graph data with the STGNet model. By combining the results from **Fig. 7** and **Fig. 8**, we demonstrate that our approach significantly enhances the ability of the STGNet model to recognize and interpret structural and energetic perturbations induced by single amino acid substitutions.

Impact of Data Augmentation Ratios on Prediction Accuracy

We evaluate the impact of sample size on prediction performance by varying the number of conformations to 5, 50, and 100 samples, designated as top5, top50, and top100, for each experimental protein-ligand pair. The filtering conformation process involves minimizing the deviation from the mean value in the distribution of relative binding energy. Through the data augmentation workflow, the dataset is expanded by factors of 5, 50, and 100 compared to the original dataset. The original dataset only includes the crystal structures and fails to reflect the global impact of mutations and conformational changes on free energy due to the singularity and limitations of the sample. As can be seen from **Fig. 8**, the prediction based on the original dataset results in higher errors and lower correlations. The best performance is achieved when the number of data is expanded by 50-fold (top50) on the original dataset. When the dataset is expanded by 5-fold (top5) or 100-fold (top100) compared to the original dataset, the RMSE and MAE have larger errors for axitinib, bosutinib, and ponatinib compared to the top50 dataset. The top50 dataset reduced the RMSE and MAE by the data enhancement strategy, especially in the case of bosutinib, dasatinib, and ponatinib. In contrast, top5 and top100 performed slightly lower than top50. This indicates that the proper choice in the number of conformational filtering can further improve the accuracy of the model prediction. From Pearson and Spearman coefficients, the correlation of the original dataset was lower than that of top5, top50, and top100, especially on axitinib and ponatinib, which even showed a negative correlation. In contrast, both top5 and top50 showed higher correlations on these datasets, especially top50 showed the best correlation on almost all drugs. The top100 achieved higher Spearman coefficients in imatinib and ponatinib, while the top50 achieved higher values in the other four TKIs. The top50 also shows less variance over other datasets as shown in Fig. S10. The top50 may better achieve the balance between capturing the structural diversity and minimizing noise, while the top100 dataset may encompass large-scale conformational changes, which also introduces variability in prediction.

The Pearson and Spearman values were higher than those of the physics-based method and direct binding energy prediction method. The increase in data is not proportional to model performance. Due to the inconsistent conformational distribution among TKIs, both excessive and insufficient data can hinder improvements in model performance. The key lies in the relevance and correlation of the data. This indicates that data augmentation strategies play an important role in improving model performance, which requires consideration of the quality of the dataset and the correlation of the data. Further optimizing the ratio of the number of samples used can improve the prediction performance. By integrating multimodal dynamic information, introducing a rational filtering strategy, and STGNet modeling for feature extraction, the proposed method has significant potential to enhance prediction accuracy in producing the relative binding energy.

Conformations distribution after filtering

By calculating the RMSD of the backbone (backbone) of each conformation of the data-enhanced top50, it is possible to quantify the extent to which the overall conformation of the protein is shifted relative to the reference structure (active and inactive states). The RMSD distribution characterizes the relative distribution of proteins between the active and inactive states after mutation. The filtered conformations of mutant ligands were clustered based on RMSD with the reference crystal structures. The clustering result in **Fig. 9A** reflects the preference of TKIs for kinase conformations, and the distribution of active and inactive states affects the predictive performance of the model under different TKIs.

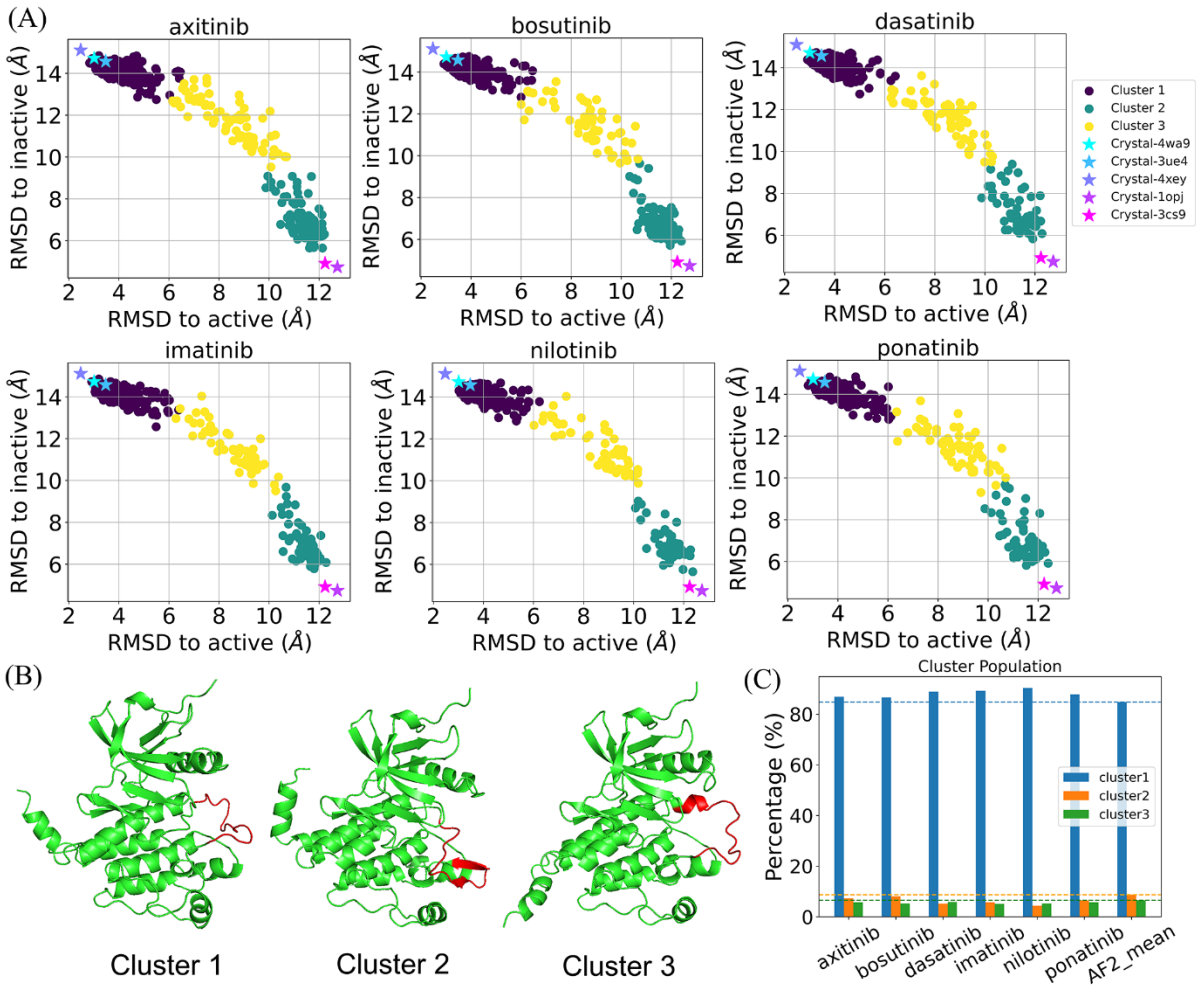

Fig. 9 RMSD clustering of top50 dataset and the representative conformations. (A) RMSD distributions of conformations docked for six TKIs. (B) Representative conformations in each cluster center. (C) Percentage changes in each cluster after filtering. Crystal structures are labeled as the stars. The A-Loop region is highlighted in red in the structures.

As can be seen in **Fig. 9**, cluster 1 distributes close to the active state region, cluster 2 is close to the inactive state region, and cluster 3 is located in the middle transition region. The structures nearest to the center of the clusters were extracted out respectively as the representative structures as shown in **Fig. 9B**. The A-Loop stretched away from the core region of the protein in cluster1, which is characteristic of the active state. The representative structure

cluster 2 is more folded and tightly closed, consistent with the inactive state. The representative structure of cluster 3 shows a partially folded A-Loop. Similar binding modes are observed for six TKI molecules as shown in Fig. S11. We also observe the slight changes in conformational distributions in the three clusters as shown in **Fig. 9C**. The selection based on the most probable distribution in the data filtering answers the question "How to appropriately select a proper structure?". This suggests that kinase mutations follow an intrinsic link between conformations and functional states in different TKIs. Even though molecular docking cannot predict the accurate binding affinity, we can obtain a rational distribution for different conformational ensembles from our approach.

We demonstrate the application of subsampled AF2 in the downstream prediction of drug resistance. By integrating with ensemble docking and Siamese learning, we achieved better accuracy in ranking relative binding affinities for different inhibitors compared to physics-based methods. This approach offers new insights into studying the dynamic properties of protein-ligand interactions in drug discovery. Leveraging AF2 predicted structures for protein-ligand binding affinity prediction is poised to become a major focus in computational biology, offering transformative potential for downstream applications in drug design and molecular interactions.

Conclusion

We present a method for predicting relative free energy in protein mutations by generating the relative distribution of conformations through AF2 subsampling. The data enhancement approach significantly improves the prediction performance of mutation effects by integrating the structural diversity of conformational ensembles, implementing a robust filtering strategy, and employing multi-dimensional feature extraction of graph data using the STGNet model. Using the TKI dataset and a refined set of PDBbind, we verified the improved prediction performance of our proposed method over the physics-based method and the direct absolute value prediction method. The results show that the dataset augmented by 50-fold (top50), created using AF2 subsampling and Gaussian fitting, not only increases the structural diversity of mutation-ligand pairs but also effectively captures potential interaction patterns between different mutations and ligands. STGNet model proposed in this study not only effectively captures the key features of protein-ligand interactions, but also significantly improves the accuracy and stability of mutation effect prediction. Additionally, RMSD calculation and clustering analysis reveal the dynamic distribution of kinase mutant conformations among active, inactive, and intermediate states, reflecting the binding preferences of different TKIs on kinase conformations. In summary, we extend the application of integrating subsampled AF2 and Siamese learning in the realm of prediction of drug resistance. This approach offers a promising tool for predicting the relative binding energy of mutation-ligand interactions related to drug resistance without relying on crystal structures.

Methods

Mathematical formulation

The energy difference between two states can be analyzed using perturbation theory. The perturbation depends on the selection of the reference state. In cases where the two states differ in the bound ligands, the perturbation corresponds to the energy difference in binding affinity induced by variations in the ligand. The energy perturbation obeys $\Delta G_{i,j} = \Delta G_{ref} + \Delta\Delta G_{lig}$. If two states differ in the conformational states of the protein, the perturbation reflects the energy difference in binding affinity resulting from conformational changes, which can be expressed as follows.

$$\Delta G_{i,j} = \Delta G_{ref} + \Delta\Delta G_{conf} \tag{1}$$

Transitioning from absolute binding energy $\Delta G_{i,j}$ to relative binding energy $\Delta\Delta G_{i,j}$, we can define the reference state as the relative binding energy obtained from experimental values, and the perturbed state is represented by the relative binding energy obtained from molecular docking. Thus, in our case, we can interpret the distribution of relative docking scores as the perturbations applied to the experimental distribution. The energy difference is computed from the reference states plus the energy variation as follows.

$$\Delta\Delta G_{i,j} = \Delta\Delta G_{ref} + \Delta\Delta\Delta G_{conf_lig} \tag{2}$$

Therefore, the purpose of employing deep learning is to compute the differences in relative binding energy $\Delta\Delta\Delta G_{conf_lig}$.

After generating the conformation ensemble, we conduct the molecular docking. We obtain the relative docking score difference between mutants and the wild-type protein.

$$\Delta\Delta G_{\text{dock}} = \Delta G_{\text{dock(mutation)}} - \Delta G_{\text{dock(wild)}} \tag{3}$$

The relative docking score is used to select the most probable distribution based on the deviation from the Gaussian fitting mean value. The selected docking score and their corresponding conformation will then be paired with the experimental values. In the current study, we select the experimental value as the reference state. The difference in relative binding energy is computed from the following equation:

$$\Delta\Delta\Delta G_{\text{cal}} = \Delta\Delta G_{\text{exp}} - \Delta\Delta G_{\text{dock}} \tag{4}$$

$\Delta\Delta\Delta G_{\text{cal}}$ represents the energy difference caused by the conformation change and ligand variety. We aim to include conformational diversity in the prediction. Siamese networks are particularly well-suited for learning the differences between the two so we constructed a relative difference based on the experimental values and generated values, which serves as the regression target for our network learning.

The molecular representations and the difference binding energies are input into STGNet. After training the model, we predicted the binding energy difference $\Delta\Delta\Delta G_{\text{pred}}$ for the test dataset. Siamese network was employed to learn the energy differences resulting from both conformation changes and mutations in protein sequences. The predicted energy differences $\Delta\Delta\Delta G_{\text{pred}}$ are recovered to the predicted relative binding energy $\Delta\Delta G_{\text{pred}}$ from the following equation:

$$\Delta\Delta G_{\text{pred}} = \Delta\Delta\Delta G_{\text{pred}} + \Delta\Delta G_{\text{dock}} \tag{5}$$

Each experimental value has been paired with N-generated samples, thereby we obtained N predicted relative binding energy. The final predicted relative binding energy is the average of the predicted relative binding energy:

$$\overline{\Delta\Delta G_{\text{pred}}} = \frac{\Delta\Delta G_{\text{pred}}}{N} \tag{6}$$

From the above workflow, we introduce the conformation ensemble in the prediction of relative binding energy. The relative binding energy can be recalibrated to the experimental level $\overline{\Delta\Delta G_{\text{pred}}}$ by computing the statistical average of STGNet predictions for the specified ligand molecule.

Workflow for constructing the augmented dataset

As shown in **Fig. 2**, the workflow outlines a systematic approach to enhancing drug docking studies through the integration of reference data, conformational generation, and statistical analysis. By utilizing AlphaFold2 for conformational modeling and implementing ensemble docking procedures, we obtain relative docking scores that are subsequently analyzed using Gaussian fitting. The workflow emphasizes ranking these scores, filtering conformations based on probability proximity to the average, and ultimately constructing an augmented dataset.

Step 1: Construct reference data. Compile the original dataset, which includes 31 types of mutations and 6 associated compounds. Prepare mutation-ligand pairs and the corresponding relative binding energies.

Step 2: Generate conformations and conduct the docking. Utilize subsampled AlphaFold2 to generate protein conformations for wild and 31 types of ABL mutants. We generate 640 conformations for each type of mutation. Conduct molecular docking simulations on these generated conformations to assess binding affinities.

Step 3: Obtain docking scores and perform Gaussian fitting. Calculate the docking scores from the docking using AutoDock Vina. Check Gaussian fitting on the distribution of these docking scores to identify statistical averages.

Step 4: Prepare relative docking scores. Obtain relative docking scores between the wild and mutants to evaluate the binding efficacy of each conformation.

Step 5: Filter conformations. Filter the conformations and their corresponding relative docking scores to identify those that are close to the average score with the highest probability density.

Step 6: Construct an augmented dataset. Create an augmented dataset that includes the filtered conformations, their associated relative docking scores, and the difference in the relative docking scores. Ensure that the augmented dataset captures variability and distinct conformational states for further analysis.

Datasets and preprocessing

TKI dataset. We validated the proposed approach on the TKI dataset collected by Hauser et al., which includes binding affinity data for mutation-ligand pairs ($\Delta\Delta G$ in Kcal/mol)[59]. The 'nd' values were excluded based on the original dataset, resulting in the selection of 144 valid docking energy data for the mutants. In total, 31 mutant and 1 wild-type proteins were selected. The conformations from each protein were then adopted as the target to dock with different TKI molecules using AutoDock Vina, thus obtaining docking energy data for both mutant and wild-type conformations. The investigated TKI molecules include axitinib, bosutinib, dasatinib, imatinib, nilotinib, and ponatinib. The data is partitioned into training and test sets for cross-validation based on the name of TKI.

Conformational ensemble prediction from AF2 subsampling

Monteiro et al. developed subsampled AlphaFold2 to predict the relative populations of protein conformations by subsampling MSA[26]. Their results show that the subsampling method can accurately predict the relative conformational populations of proteins when compared to the data from NMR experiments. They also obtained the accurate relative state population changes for the eight mutants for ABL kinase[26]. We employed the same implementation to sample conformation ensembles of 31 mutants of ABL kinase. The residue length indices from 248 to 534 according to the crystal structure (PDB ID: 6XR6), following the same prediction procedure using ColabFold. ColabFold utilizes the fast homology search of MMseqs2 with AF2 to facilitate rapid prediction of protein structures and complexes [67]. The value of max_seq:extra_seq was set to 256:512, which would result in the most diverse results concerning the activation loop conformation. We utilized 5 models of AF2 per prediction, performed 4 rounds of optimization, and set the number of seeds as 32. In total, we generated 640 conformations for each mutation.

Molecular docking and preprocessing of the relative binding energy

Different conformations of the proteins were obtained through subsampling, resulting in a relative conformational distribution. Molecular docking for each sampled conformation was performed using AutoDock Vina to determine the binding energy of the ligand across different mutations as shown in Step 2 in **Fig. 2**. The docking center is set as the coordinate center of four residues with residue index 267, 288, 378, and 399. The exhaustiveness of AutoDock Vina is set as 32. The docking energies for both the mutant and wild type were derived from ensemble docking. We computed the difference in docking score between the mutant and the wild type in Step 3 and Step 4. Then we filtered the conformational data with the smallest

deviation from the Gaussian mean of the binding energy distribution of each mutant-ligand combination in Step 5. We selected the top k (eg. 5, 50, and 100) values with the smallest absolute differences value from the Gaussian mean value for pairing. We select the k values near the average because the Gaussian fitting indicates that values close to the average have the highest probability distribution. We calculate $\Delta\Delta\Delta G_{\text{cal}}$ different TKIs by calculating the differences between the experimental value and the generated value in Step 6. Finally, all composite confirmation data are grouped to construct the required data-augmented dataset. The total number of raw predictions is equal to the number of conformations multiplied by the number of ligands, represented as $N=k\times 6$. As shown in **Fig. 2**, conf_exp consists of a single conformation, while conf_AF comprises k conformations generated from subsampled AF2 corresponding to the same mutation type.

Refined set of PDBbind for direct absolute value prediction

The refined set of PDBbind version 2020 was selected to compare with our approach. After the removal of ABL samples, the refined set was divided into the training and validation sets in a 9:1 ratio. The CASF-2016 is used to benchmark the performance of the regular network model. The test set data comprises the 144 mutation data in the original TKI dataset alongside the six TKIs for wild type. The training was conducted using a similar network architecture, except for removing the Siamese network. After training the model, we predict the absolute binding free energy separately for both the wild type and mutations. Subsequently, we computed the relative binding affinity by subtracting the absolute binding energy of mutation from that of the wild type, respectively.

Input features of the Siamese learning

We adopt both the structural information and molecular graph as the molecular representation to capture the contribution from the conformation and the interaction between proteins and ligands. The molecular representations are summarized in Table S6.

Protein contact maps. Predicting geometrical relationships between residues (including distances and orientations) has demonstrated its importance in structure predictions[68]. Protein contact maps use the adjacency matrix $A_{i,j}$ to represent the graph structure of a protein, which contains information about whether or not amino acids are found in the protein chain, specifically, a protein map $G=\left(N_p, M_p\right)$ is a contact map generated by taking an input protein containing L_p residues. A pair of residues is considered to be in contact when the Euclidean distance $d_{i,j}$ between the CA atoms of the two pairs of amino acids is less than a threshold d_c. The residues in a pair are considered to be in contact with each other. These

connections are determined from the three-dimensional structure of the proteins. We set a threshold of 10 Å.

$$A_{i,j} = \begin{cases} d_{i,j} & d_{i,j} \le d_c \\ 0 & d_{i,j} > d_c \end{cases} \tag{7}$$

The protein contact map of the MPNN structure can be represented as $G_{contactMap} = (V, E)$, where V denotes a set of n atomic nodes, and each node feature vector denotes (x_v), including the one-hot coding of the protein sequence, the 12 amino acid nature features, and the BLOSUM62 matrix features[61]. E denotes the relationship of the edges of the neighbor matrix in the molecular graph, and the eigenvector of each edge is denoted (e_{vw}) as the Euclidean distance between each pair of amino acids.

Ligand graph. We choose the molecular graph as the representation of the ligands. The molecular graph of the ligand for the MPNN structure can be represented as $G_{ligand} = (V, E)$. where V denotes a set of n atomic nodes, each represented by an eigenvector of 75 values (x_v), including 44 atomic types, 11 atomic degrees, 7 implicit valence states, charge, number of free electrons, 5 hybridization types, aromaticity, and 5 hydrogen atoms. E represents the relationship of the edges of the neighbor matrix in the molecular graph. Each edge is represented by 6 values called eigenvectors (e_{vw}), including 4 bond types, conjugacy, and intra-ring bonds. The molecular graph has been validated in our previous work[69].

Protein-ligand interaction graph. We adopt the 3D interaction graph structure data used by GIGN[70]. The complex is represented as a 3D interaction graph $G_{inter} = (V, \varepsilon, R) = (V_l \cup V_p, \varepsilon_l \cup \varepsilon_{lp}, R_l \cup R_p)$, where V_l and V_p are sets of ligand and protein nodes (atoms), ε_l and ε_p are sets of ligand and protein covalent edges (covalent bonds/covalent interactions). R_l and R_p are sets of ligand and protein 3D coordinates. Each node $v_i \in V$ has initial features $x_i \in \mathrm{R}^n$ and 3D coordinates $r_i \in \mathrm{R}^3$. When atoms v_i and v_j are connected by a covalent edge, the covalent edge $e_{ij} \in \varepsilon_l \cup \varepsilon_p$. A noncovalent edge $e_{ik} \in \varepsilon_{lp}$ exists when $d_{ik} < 15$ Å, where $d_{ik} = \|r_i - r_k\|^2$ is the distance between the ligand atom $v_i \in V_l$ and the protein atom $v_k \in V_p$.

Siamese learning

Message Passing Neural Networks. We used message-passing networks to process contact maps for proteins and graph data for ligands. Gilmer et al. designed an MPNN framework for processing chemical graphs[71]. Forward propagation consists of two phases, the message-passing phase and the readout phase. The message passing phase runs for T time steps and is defined according to the message function M_t and vertex update function U_t. In the message passing phase, the hidden state h_v^t of each node is updated according to the message m_v^{t+1} as follows.

$$m_v^{t+1} = \sum_{w \in N(v)} M_t\left(h_v^t, h_w^t, e_{vw}\right) \tag{8}$$

$$h_v^{t+1} = U_t\left(h_v^t, m_v^{t+1}\right) \tag{9}$$

Where, $N(v)$ denotes the neighbor of node v in graph G, M_t denotes the message function, U_t denotes the vertex update function. The readout phase uses the readout function R to compute the feature vectors of the entire graph according to the following equation.

$$y = R\left(\left\{h_v^T \mid v \in G\right\}\right) \tag{10}$$

where R denotes the readout function.

Geometric interaction graph neural network. Yang et al. designed an ensemble interaction graph neural network, a Geometric interaction graph neural network (GIGN) that combines 3D structural and physical interactions to learn node representations more efficiently through a heterogeneous interaction layer that unifies covalent and non-covalent interactions into a message-passing phase[70].

In the Geometric Interaction Graph Neural Network (GIGN), the process of integrating covalent and noncovalent interactions into node representations is handled through a message-passing mechanism that accounts for the heterogeneity of these interactions. Each node in the 3D interaction graph, representing an atom, updates its representation by aggregating messages from its neighboring nodes, which are categorized as either covalent or noncovalent interactions. The messages from covalent neighbors are calculated using a function $M^{(\mathrm{cov})}$ that incorporates the node features x and their 3D coordinates r, while messages from noncovalent neighbors are computed using a similar function $M^{(\mathrm{ncov})}$. These messages are then aggregated separately, respecting their distinct nature.

The aggregated messages, which differ in scale due to the significant difference in the number of noncovalent versus covalent interactions, are normalized to the same scale by multilayer perceptrons $U^{(\mathrm{cov})}$ and $U^{(\mathrm{ncov})}$ before integration. This integration step ensures that the

information from both interaction types is effectively combined to refine the node's representation. The final representation update for each node is given as follows.

$$h_i^{(t+1)} = U^{(\mathrm{cov})}\left(h_i^{(t)}, m_i^{(\mathrm{cov})}\right) + U^{(\mathrm{ncov})}\left(h_i^{(t)}, m_i^{(\mathrm{ncov})}\right) \tag{11}$$

Where, m_i^{cov} and m_i^{ncov} are the aggregated messages from covalent and noncovalent neighbors, respectively, and $h_i^{(t)}$ is the current representation of node v_i at layer t. This approach maintains the invariance property concerning input translations and rotations, avoiding the need for computationally expensive data augmentation and leading to a more efficient and biologically meaningful prediction of protein-ligand binding affinities.

STGNet model. We propose the SGTNet model for predicting the relative binding energy for protein-ligand complexes. The SGTNet model uses dual inputs of experimental data and generated data as shown in **Fig. 3**. The pairings are made between the reference conformation and the generated conformations. This enables the fusion of ligand-binding features with information on the structural changes triggered by the mutation. For the learning with protein contact maps and small molecules, we take the graph information as inputs and perform feature extraction through the MPNN model with shared parameters to get the feature differences between the two inputs, followed by further extraction of the deep features through the fully connected layer. For the learning of protein-ligand interactions, we take the graph information of the docked protein-ligand structure as inputs perform feature extraction via the GIGN model with shared parameters, and further extract deep features via the fully-connected layer after calculating the feature differences. Then, all the learned graph features are concatenated together to form a joint feature, which generates the final prediction $\Delta\Delta\Delta G_{pred}$ through a fully connected layer.

To obtain the target prediction value $\overline{\Delta\Delta G_{pred}}$, the calculation method is shown in Fig. S1. The results of the raw predictions for a given test set include specific ligands, and each mutation corresponds to the predicted binding energies for these ligands. Consequently, the target prediction value needs to be averaged from the raw predicted results. We evaluated the STGNet model using the original dataset, and three pairwise augmented datasets with *k* chosen as 5, 50, and 100, respectively.

Data availability

AlphaFold2 subsampling prediction data generated in this study have been deposited in the Zenodo: https://zenodo.org/records/14607966.

Code availability

The code used to generate the results shown in this study is available under an MIT License

via GitHub at https://github.com/AIMedDrug/STGNet.

References

1. Tunyasuvunakool K, *et al.* Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590 (2021).
2. Watson JL, *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089-1100 (2023).
3. Abramson J, *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493-500 (2024).
4. Pillai A, *et al.* De novo design of allosterically switchable protein assemblies. *Nature* **632**, 911-920 (2024).
5. Krishna R, *et al.* Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science*, eadl2528 (2024).
6. Senior AW, *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706-710 (2020).
7. AlQuraishi M. Protein-structure prediction revolutionized. *Nature* **596**, 487-488 (2021).
8. Marx V. Method of the Year: protein structure prediction. *Nat Methods* **19**, 5-10 (2022).
9. Jumper J, *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).
10. Humphreys IR, *et al.* Computed structures of core eukaryotic protein complexes. *Science* **374**, eabm4805 (2021).
11. Ahdritz G, *et al.* OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *Nat. Methods* **21**, 1514-1524 (2024).
12. Lin Z, *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123-1130 (2023).
13. Chowdhury R, *et al.* Single-sequence protein structure prediction using a language model and deep learning. *Nat. Biotechnol.* **40**, 1617-1623 (2022).
14. Wu R, *et al.* High-resolution de novo structure prediction from primary sequence. *BioRxiv*, (2022).
15. Cheng J, *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492.
16. Bryant P, Pozzati G, Elofsson A. Improved prediction of protein-protein interactions using AlphaFold2. *Nat. Commun.* **13**, 1265 (2022).
17. Yang Z, Zeng X, Zhao Y, Chen R. AlphaFold2 and its applications in the fields of biology and medicine. *Signal Transduction and Targeted Therapy* **8**, 115 (2023).
18. Ma W, *et al.* Enhancing protein function prediction performance by utilizing AlphaFold-predicted protein structures. *Journal of Chemical Information Modeling* **62**, 4008-4017 (2022).
19. Kortemme T. De novo protein design-From new structures to programmable functions. *Cell* **187**, 526-544 (2024).
20. Grinter R, *et al.* Structural basis for bacterial energy extraction from atmospheric hydrogen. *Nature* **615**, 541-547 (2023).
21. Amaro RE. Will the real cryptic pocket please stand out? *Biophys. J* **116**, 753-754

(2019).
22. Davis MI, *et al.* Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* **29**, 1046-1051 (2011).
23. Broerman AJ, *et al.* Design of facilitated dissociation enables control over cytokine signaling duration. *bioRxiv*, 2024.2011.2015.623900 (2024).
24. Jänes J, Beltrao P. Deep learning for protein structure prediction and design—progress and applications. *Mol. Syst. Biol.* **20**, 162-169 (2024).
25. Wayment-Steele HK, *et al.* Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature* **625**, 832-839 (2024).
26. Monteiro da Silva G, Cui JY, Dalgarno DC, Lisi GP, Rubenstein BM. High-throughput prediction of protein conformational distributions with subsampled AlphaFold2. *Nat Commun* **15**, 2464 (2024).
27. Guan X, *et al.* Predicting protein conformational motions using energetic frustration analysis and AlphaFold2. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2410662121 (2024).
28. Stein RA, McHaourab HS. SPEACH_AF: Sampling protein ensembles and conformational heterogeneity with Alphafold2. *PLoS Comput. Biol.* **18**, e1010483 (2022).
29. Raisinghani N, *et al.* Integration of a Randomized Sequence Scanning Approach in AlphaFold2 and Local Frustration Profiling of Conformational States Enable Interpretable Atomistic Characterization of Conformational Ensembles and Detection of Hidden Allosteric States in the ABL1 Protein Kinase. *J. Chem. Theory Comput.* **20**, 5317-5336 (2024).
30. Del Alamo D, Sala D, Mchaourab HS, Meiler J. Sampling alternative conformational states of transporters and receptors with AlphaFold2. *Elife* **11**, e75751 (2022).
31. Freiberger MI, *et al.* Local energetic frustration conservation in protein families and superfamilies. *Nat. Commun.* **14**, 8379 (2023).
32. Tessa A, *et al.* Identification of mutations in AP4S1/SPG52 through next generation sequencing in three families. *Eur. J. Neurol.* **23**, 1580-1587 (2016).
33. Jones DT, Singh T, Kosciolek T, Tetchner S. MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics* **31**, 999-1006 (2015).
34. Wang S, Sun S, Li Z, Zhang R, Xu J. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput. Biol.* **13**, e1005324 (2017).
35. Heo L, Feig M. Experimental accuracy in protein structure refinement via molecular dynamics simulations. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 13276-13281 (2018).
36. Park H, Ovchinnikov S, Kim DE, DiMaio F, Baker D. Protein homology model refinement by large-scale energy optimization. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 3054-3059 (2018).
37. Steinbrecher TB, *et al.* Accurate binding free energy predictions in fragment optimization. *J. Chem. Inf. Model.* **55**, 2411-2420 (2015).
38. Moore JH, Margreitter C, Janet JP, Engkvist O, de Groot BL, Gapsys V. Automated relative binding free energy calculations from SMILES to ΔΔG. *Chem. Commun.* **6**, 82 (2023).
39. Cao D, *et al.* Generic protein–ligand interaction scoring by integrating physical prior

knowledge and data augmentation modelling. *Nat. Mach. Intell.* **6**, 688-700 (2024).
40. Gilson MK, Zhou H-X. Calculation of Protein-Ligand Binding Affinities. *Annu. Rev. Biophys.* **36**, 21-42 (2007).
41. Amaro RE*, et al.* Ensemble Docking in Drug Discovery. *Biophys. J* **114**, 2271-2278 (2018).
42. Kharbanda S, Yuan Z-M, Weichselbaum R, Kufe D. Determination of cell fate by c-Abl activation in the response to DNA damage. *Oncogene* **17**, 3309-3318 (1998).
43. Van Etten RA. Cycling, stressed-out and nervous: cellular functions of c-Abl. *Trends Cell Biol.* **9**, 179-186 (1999).
44. Wang JY. The capable ABL: what is its biological function? *Mol. Cell. Biol.* **34**, 1188-1197 (2014).
45. O'Brien SG*, et al.* Imatinib compared with interferon and low-dose cytarabine for newly diagnosed chronic-phase chronic myeloid leukemia. *N. Engl. J. Med.* **348**, 994-1004 (2003).
46. Azam M, Latek RR, Daley GQ. Mechanisms of autoinhibition and STI-571/imatinib resistance revealed by mutagenesis of BCR-ABL. *Cell* **112**, 831-843 (2003).
47. Shah NP*, et al.* Multiple BCR-ABL kinase domain mutations confer polyclonal resistance to the tyrosine kinase inhibitor imatinib (STI571) in chronic phase and blast crisis chronic myeloid leukemia. *Cancer Cell* **2**, 117-125 (2002).
48. Zhou Y*, et al.* Structure-guided machine learning prediction of drug resistance mutations in Abelson 1 kinase. *Comput. Struct. Biotechnol. J.* **19**, 5381-5391 (2021).
49. Friedman R. Computational studies of protein–drug binding affinity changes upon mutations in the drug target. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **12**, e1563 (2022).
50. Roskoski R. Classification of small molecule protein kinase inhibitors based upon the structures of their drug-enzyme complexes. *Pharmacol. Res.* **103**, 26-48 (2016).
51. Paul F, Meng Y, Roux B. Identification of Druggable Kinase Target Conformations Using Markov Model Metastable States Analysis of apo-Abl. *J. Chem. Theory Comput.* **16**, 1896-1912 (2020).
52. McBride JM, Polev K, Abdirasulov A, Reinharz V, Grzybowski BA, Tlusty T. AlphaFold2 Can Predict Single-Mutation Effects. *Phys. Rev. Lett.* **131**, 218401 (2023).
53. Roney JP, Ovchinnikov S. State-of-the-art estimation of protein model accuracy using AlphaFold. *Phys Rev Lett* **129**, 238101 (2022).
54. McNutt AT, Koes DR. Improving ΔΔG Predictions with a Multitask Convolutional Siamese Network. *J. Chem. Inf. Model.* **62**, 1819-1829 (2022).
55. Feng B*, et al.* A bioactivity foundation model using pairwise meta-learning. *Nat. Mach. Intell.* **6**, 962-974 (2024).
56. Fralish Z, Chen A, Skaluba P, Reker D. DeepDelta: predicting ADMET improvements of molecular derivatives with deep learning. *J. Cheminf.* **15**, 101 (2023).
57. Yu J*, et al.* Computing the relative binding affinity of ligands based on a pairwise binding comparison network. *Nat. Comput. Sci.* **3**, 860-872 (2023).
58. Jorgensen WL, Thomas LL. Perspective on Free-Energy Perturbation Calculations for Chemical Equilibria. *J. Chem. Theory Comput.* **4**, 869-876 (2008).
59. Hauser K*, et al.* Predicting resistance of clinical Abl mutations to targeted kinase inhibitors using alchemical free-energy calculations. *Communications Biology* **1**, 70

(2018).
60. Jiménez-Luna J*, et al.* DeltaDelta neural networks for lead optimization of small molecule potency. *Chem. Sci.* **10**, 10911-10918 (2019).
61. Eddy SR. Where did the BLOSUM62 alignment score matrix come from? *Nat. Biotechnol.* **22**, 1035-1036 (2004).
62. Lyczek A*, et al.* Mutation in Abl kinase with altered drug-binding kinetics indicates a novel mechanism of imatinib resistance. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2111451118 (2021).
63. O'Hare T, Eide CA, Deininger MWN. Bcr-Abl kinase domain mutations, drug resistance, and the road to a cure for chronic myeloid leukemia. *Blood* **110**, 2242-2249 (2007).
64. Fernandes PO, Martins DM, de Souza Bozzi A, Martins JPA, de Moraes AH, Maltarollo VG. Molecular insights on ABL kinase activation using tree-based machine learning models and molecular docking. *Mol. Divers.* **25**, 1301-1314 (2021).
65. Xie T, Saleh T, Rossi P, Kalodimos CG. Conformational states dynamically populated by a kinase determine its function. *Science* **370**, eabc2754 (2020).
66. Cournia Z, Allen B, Sherman W. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *J. Chem. Inf. Model.* **57**, 2911-2937 (2017).
67. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger MJNm. ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679-682 (2022).
68. Du Z*, et al.* The trRosetta server for fast and accurate protein structure prediction. *Nat. Protoc.* **16**, 5634-5651 (2021).
69. Lu X, Xie L, Xu L, Mao R, Xu X, Chang S. Multimodal fused deep learning for drug property prediction: Integrating chemical language and molecular graph. *Comput. Struct. Biotechnol. J.* **23**, 1666-1679 (2024).
70. Yang Z, Zhong W, Lv Q, Dong T, Yu-Chian Chen C. Geometric interaction graph neural network for predicting protein–ligand binding affinities from 3d structures (GIGN). *J Phys Chem Lett* **14**, 2020-2033 (2023).
71. Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. Neural message passing for quantum chemistry. In: *Proceedings of the 34th International Conference on Machine Learning*). JMLR.org (2017).

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62373172, 22003020 and 12074151), the Natural Science Foundation of Jiangsu Province (BK20191032), and Changzhou Sci. & Tech. Program (CJ20200045).

Author contributions

L.Xie.: Initiated the project and proposed the method. X.L.: Performed modeling. L.Xie. and X.L.: Conducted data analysis. D.Z. and L.Xu.: Contributed the data analysis. S.C. and X.X.:

Supervised the work.

Competing interests

The authors declare no competing interests.

Additional information

The online version contains supplementary material.