Decoding Mechanisms of PTEN Missense Mutations in Cancer and Autism Spectrum Disorder using Interpretable Machine Learning Approaches

Miao Yang[1#], Jingran Wang[1#], Ziyun Zhou[1], Wentian Li[1], Gennady Verkhivker[2,3], Fei Xiao[1], Guang Hu[1,4,5,6]**

[1] MOE Key Laboratory of Geriatric Diseases and Immunology, Suzhou Key Laboratory of Pathogen Bioscience and Anti-infective Medicine, Department of Bioinformatics and Computational Biology, School of Life Sciences, Suzhou Medical College of Soochow University, Suzhou, 215213, China

[2]Keck Center for Science and Engineering, Schmid College of Science and Technology, Chapman University, One University Drive, Orange 92866, California, United States.

[3]Department of Biomedical and Pharmaceutical Sciences, Chapman University School of Pharmacy, Irvine 92618, California, United States.

[4]Jiangsu Province Engineering Research Center of Precision Diagnostics and Therapeutics Development, Soochow University, Suzhou 215123, China

[5]Key Laboratory of Alkene-carbon Fibres-based Technology & Application for Detection of Major Infectious Diseases, Soochow University, Suzhou 215123, China

[6]Jiangsu Key Laboratory of Infection and Immunity, Soochow University, Suzhou 215123, China

# Equal contribution

*To whom correspondence should be addressed: Fei Xiao, xiaofei@suda.edu.cn; Guang Hu, huguang@suda.edu.cn

ABSTRACT

Missense mutations in oncogenic proteins that are concurrently associated with neurodevelopmental disorders have garnered significant attention. Phosphatase and tensin homolog (PTEN) serves as a paradigmatic model for mapping its mutational landscape and identifying genotypic predictors of distinct phenotypic outcomes, including cancer and autism spectrum disorder (ASD). Despite extensive research into the genotype-phenotype correlations of PTEN mutations, the mechanisms underlying the dual association of specific PTEN mutations with both cancer and ASD (PTEN-cancer/ASD mutations) remain elusive. This study introduces an integrative approach that combines machine learning (ML) with structural dynamics to elucidate the molecular effects of PTEN-cancer/ASD mutations. Analysis of biophysical and network biology-based signatures reveals a complex energetic and functional landscape. Subsequently, an ML model and corresponding integrated score were developed to classify and predict PTEN-cancer/ASD mutations, underscoring the significance of protein dynamics in predicting cellular phenotypes. Further molecular dynamics simulations demonstrated that PTEN-cancer/ASD mutations induce dynamic alterations characterized by open conformational changes restricted to the P loop and coupled with inter-domain allosteric regulation. This research aims to enhance the genotypic and phenotypic understanding of PTEN-cancer/ASD mutations through an interpretable ML model integrated with structural dynamics analysis. By identifying shared mechanisms between cancer and ASD, the findings pave the way for the development of novel therapeutic strategies.

1. INTRODUCTION

Epidemiological studies have demonstrated that certain cancers occur more frequently in individuals with neurodevelopmental disorders (NDDs), suggesting a significant correlation between these conditions[1]. Cancer originates from the dysregulation of cellular growth, proliferation, and differentiation[2], whereas NDDs arise from disruptions in nervous system development, affecting cognitive, motor, social, and behavioral functions[3]. Despite their distinct clinical manifestations, emerging evidence has uncovered shared cellular pathways, proteins and genetic mutations underlying both conditions[4-6]. For instance, both diseases involve dysfunctions in key biological processes, including chromatin remodeling and signaling through the PI3K/mTOR and MAPK pathways[7]. Notably, over one-third of genes causally linked to cancer have also been implicated in the onset of NDDs[8], with phosphatase and tensin homolog (PTEN) protein exemplifying this intersection[9].

PTEN comprises several functional domains, including N-terminal PIP2-binding domain (PBD), catalytic phosphatase domain (PD), membrane-binding C2 domain (C2D), and C-terminal tail (CTT), each essential for its diverse roles in cellular signaling[10]. Missense mutations in the PTEN gene are associated with a broad spectrum of diseases[11], encompassing various cancers and Autism Spectrum Disorder (ASD)[12]. These mutations can be categorized into three types based on clinical outcomes: (1) PTEN-caner mutations, present in patients with cancer only; (2) PTEN-ASD mutations, identified in patients with ASD only; and (3) PTEN-cancer/ASD mutations, observed in patients with both cancer and ASD[12-14]. Compared to the first two categories, PTEN-cancer/ASD mutations warrant significant attention due to their dual phenotypic associations. For example, the PTEN-cancer/ASD mutation R173C has been

identified in conditions such as PTEN hamartoma tumor syndrome, Cowden syndrome, ASD, and various cancers, involving colorectal cancer and brain neoplasms, disrupting the catalytic site regulation and altering its phospho-regulation[14]and increasing domain flexibility and fluctuation in the CBR3 loop, interdomain regions, and C-terminal tail[15], which resulted in the loss of phosphatase function and enhanced cellular transformation[16]. However, the dual structural and functional impacts of PTEN-cancer/ASD mutations remain underexplored.

Various methodologies have been employed to analyze PTEN mutations and their clinical relevance. Experimental approaches, including biophysical analyses[17], structural and statistical assessments[18], deep mutational scanning[19], and phenotypic studies in model organisms[20], have been used to investigate the functional impacts of these mutations. Nevertheless, these methods often lack scalability and fail to elucidate detailed molecular mechanisms, thereby limiting the comprehensive understanding of PTEN mutations. In contrast, computational approaches, such as molecular dynamics (MD) simulations[21], elastic network models (ENMs)[22], and protein structure network (PSN)-based analyses[21] have uncovered conformational changes, long-range allosteric effects, and disruptions in PTEN function and stability. These studies have been pivotal in distinguishing PTEN-cancer from PTEN-ASD mutations, where PTEN-ASD mutations primarily induce localized dynamic increases, particularly at the active site, leading to local instability that may affect substrate binding and catalytic activity, thereby contributing to ASD pathology. Conversely, PTEN-cancer mutations result in significant global structural destabilization, including reduced active site stability and increased interdomain dynamics[23], potentially compromising PTEN's tumor-suppressive functions and promoting oncogenesis. Furthermore, a machine learning (ML) model[24] has been developed to classify PTEN mutations by analyzing their molecular effects on protein structure and function. Although effective in

distinguishing PTEN-cancer and PTEN-ASD mutations, the mechanistic basis of PTEN-cancer/ASD mutations and their role in the genotype-phenotype relationship remains poorly understood, necessitating further investigation.

Biophysical insights into missense mutations are critical for evaluating protein functions[25-29]. This study aims to bridge the gap between PTEN mutations and their disease phenotypes by leveraging PTEN structural dynamics, including three steps (Figure 1). First, three types of mutations were curated from gene mutation databases, and molecular and network signatures were assigned to each mutation type through multi-level analyses. Next, an ML model incorporating an integrated scoring function bases on the above features was developed to distinguish phenotypic similarities, enabling the classification of PTEN-cancer/ASD mutations. Lastly, the molecular mechanisms underlying PTEN-cancer/ASD mutations were investigated by exploring the interpretable natures of the ML model and quantifying their conformational dynamics and long-range perturbation dynamics. Taken together, the methodological advances presented here elucidate molecular mechanisms underlying pathogenic mutations involving co-occurring diseases, and thus assist in assigning novel therapeutic strategies.

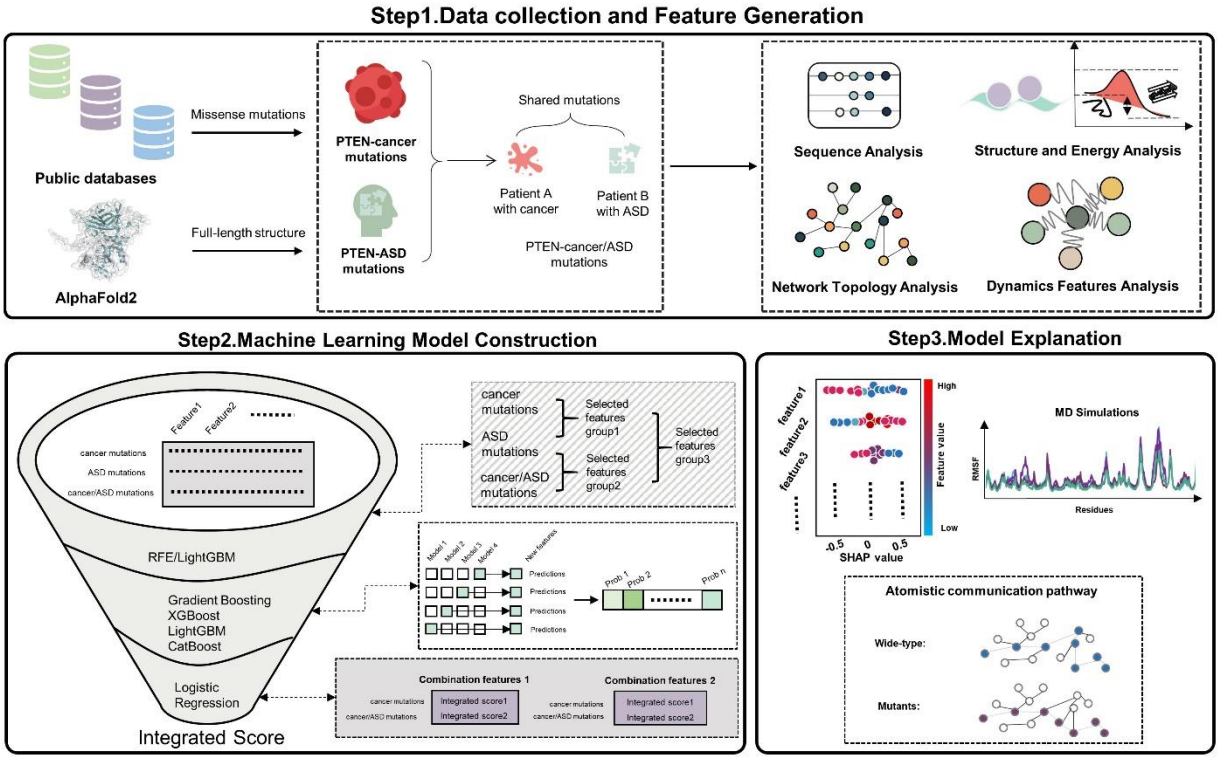


Figure 1. Overview of the integrative workflow for PTEN mutation classification and analysis. It consists of data collection and feature generation, machine learning model construction and model explanation. Model interpretability is achieved via SHAP-values to highlight influential features, while MD simulations and communication pathway analysis provide deeper insights into local and long-range structural effects distinguishing wild-type (WT) from mutants.

2. MATERIALS AND METHODS

Detailed information of structural modeling, molecular feature calculations, and MD simulations are provided in the Text S1.

Data collection and data preparation. Three types of PTEN missense mutations including PTEN-cancer, PTEN-ASD, and PTEN-cancer/ASD mutations, as well as their phenotypic information, were systematically retrieved from the ClinVar database[30], the Genome

Aggregation database[31], and the Cosmic database[32]using the search parameters "Missense", "Pathogenic" and "SNVs". To further ensure comprehensive data collection, additional PTEN mutations with corresponding phenotypic annotations from previously published studies were incorporated [14, 17] (Table S1).

Molecular features calculation. A total of molecular features at sequence, structure and dynamic levels were calculated to classify PTEN missense mutations, encompassing relative Accessible Surface Area ($RASA$), conservation score, Shannon information entropy $S(i)$, mutual information (MI), folding Gibbs free energy ($\Delta\Delta G$), Degree Centrality (DC), Clustering Coefficient (C), Betweenness Centrality (BC), Closeness Centrality (CC), Eigenvector Centrality (EC), Mean-Square Fluctuations (MSF), stiffness, Dynamic Flexibility Index (DFI), sensitivity, effectiveness.

Machine learning framework. To construct the predictive model (Figure 2), multiple interpretable ML algorithms were employed. The framework commenced with Recursive Feature Elimination (RFE)[33] in conjunction with the LightGBM classifier[34] to filter features, thereby identifying those most critical for disease classification with mean precision obtained via 5-fold cross validation. RFE systematically reduces the number of features under consideration by evaluating each feature's impact on model performance to select the optimal subset. Utilizing the feature importance rankings from LightGBM, the most effective feature combinations were identified and subsequently input into various classification models to construct a stacking model. The classification models included Gradient Boosting[35], XGBoost[36], LightGBM[34], and CatBoost[37]. This approach enhanced model' robustness and accuracy in classifying the effects of PTEN missense mutations. Stratified 10-fold cross-validation was employed to evaluate the classification performance of the models. The predicted probabilities from these models were

consolidated into a probability matrix and subsequently input into a logistic regression model for final integration.

Additionally, our model develops an IS to distinguish mutations causing similar phenotypes. The score is derived from the probabilities predicted by the model and facilitates an initial assessment of the potential phenotypes associated with relevant missense mutations. For binary classification distinguishing different phenotypes, the IS ranges from 0 to 1, with mutations with IS close to 0 or 1 being more likely to correspond to their associated phenotypes. This scoring system provides an initial evaluation of unknown missense mutations, aiding their differentiation based on phenotypic impact.

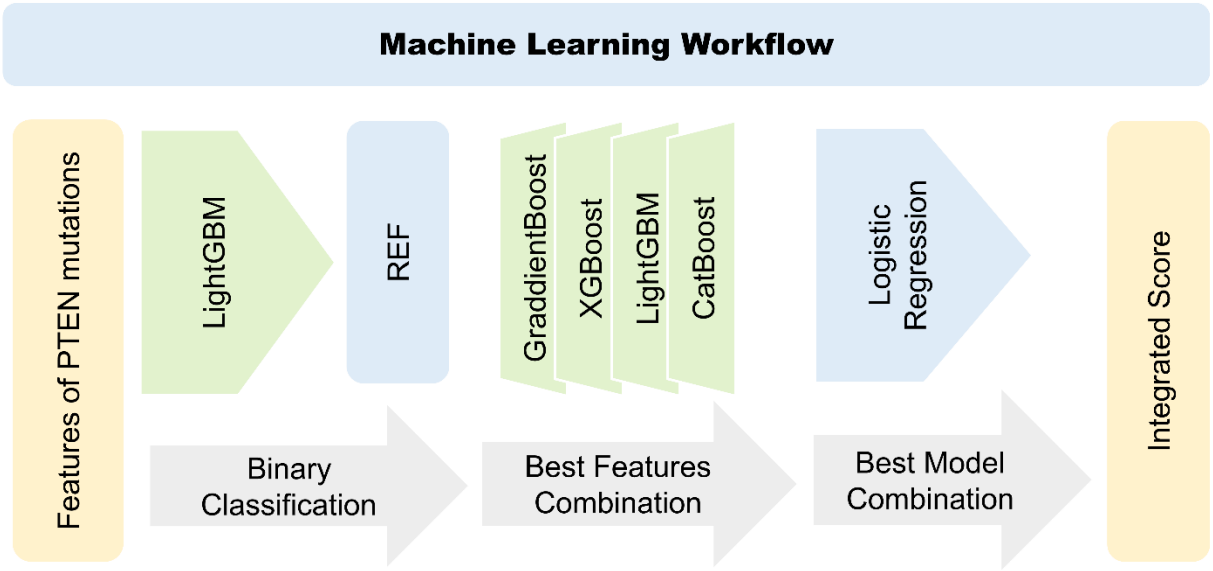


Figure 2. Overview of multi-feature integration and ML workflow. PTEN mutations are first extracted and then binary classified by LightGBM. The feature set is then optimised using RFE. The optimal combination of features is determined using ensemble models (GradientBoost, XGBoost, LightGBM and CatBoost). The optimal combination of models is then determined using logistic regression, and an Integrated Score is generated for each mutation with different phenotypes.

Performance analysis and model interpretability. The model's performance was evaluated using Receiver Operating Characteristic (ROC) curves and the corresponding Area Under the Curve (AUC) values. The AUC, ranging from 0 to 1, quantifies the model's ability to distinguish between positive and negative samples, with higher values indicating superior performance across all classification thresholds. The AUC was instrumental in determining the optimal model and parameter combinations within the stacking model. For model interpretability, Tree SHapley Additive exPlanations (SHAP)[38] were employed to enhance both global and individual interpretability within the LightGBM framework. SHAP, grounded in cooperative game theory and inspired by Shapley values, provides a robust methodology for attributing the contribution of individual features to model predictions. On a global scale, SHAP quantifies feature importance, elucidating their aggregate impact on model behavior. At the individual level, SHAP reveals the contribution of specific features to each prediction, offering a granular view of the decision-making process. By leveraging SHAP, comprehensive insights into the interplay of features were attained, enriching both the global and individual interpretability of the model.

Communication pathway analysis. Based on MD simulations (see Text S1), two methodologies were employed to explore the allosteric influences of mutations. Dynamic residue network (DRN) analysis was performed using the MD-TASK software[39]. In this approach, network nodes were defined by the $C\alpha$ and $C\beta$ atoms extracted from the MD trajectories, and edges were established when the interatomic distance $\leq$ 6.5 Å. The optimal allosteric pathways were identified by evaluating the product of edge occurrence frequencies along the path and site betweenness derived from the set of shortest paths. Additionally, the Floyd-Warshall algorithm was utilized to delineate the shortest edges connecting mutational sites to allosteric sites, thereby revealing the allosteric communication pathways[40].

3. RESULTS AND DISSCUSSION

PTEN-cancer/ASD mutations are preferentially buried, located in the Phosphatase Domain, and exhibit higher conservation. A total of 541 missense mutations were collected and categorized based on their phenotypic associations into three groups: PTEN-cancer (409 mutations), PTEN-ASD (79 mutations), and PTEN-cancer/ASD (53 mutations) (Figure 3a, Figure S1). These mutations were mapped onto the PTEN protein structure (Figure 3b, Figure S1), revealing distinct localization patterns across the groups. Specifically, PTEN-cancer mutations were predominantly enriched in the ATP-binding Type A region, situated between the P loop and the WPD loop. In contrast, PTEN-ASD mutations were distributed across the ATP-binding Type A and surrounding regions, with a subset extending into the ATP-binding Type B region. Notably, the PTEN-cancer, PTEN-ASD, and PTEN-cancer/ASD mutations predominantly cluster within the PD, whereas relatively fewer mutations occur in the PBD, C2D, and CTT (Figures 3c-e). Further analysis of the *RASA* (Figures 3f-h and Table S2) indicated differential solvent accessibility among the mutation groups. Specifically, 63% of PTEN-cancer mutations, 72% of PTEN-ASD mutations, and 81% of PTEN-cancer/ASD mutations were situated in buried or partially exposed regions. This trend underscores a higher propensity for PTEN-cancer/ASD mutations to occur in less solvent-accessible areas of the protein.

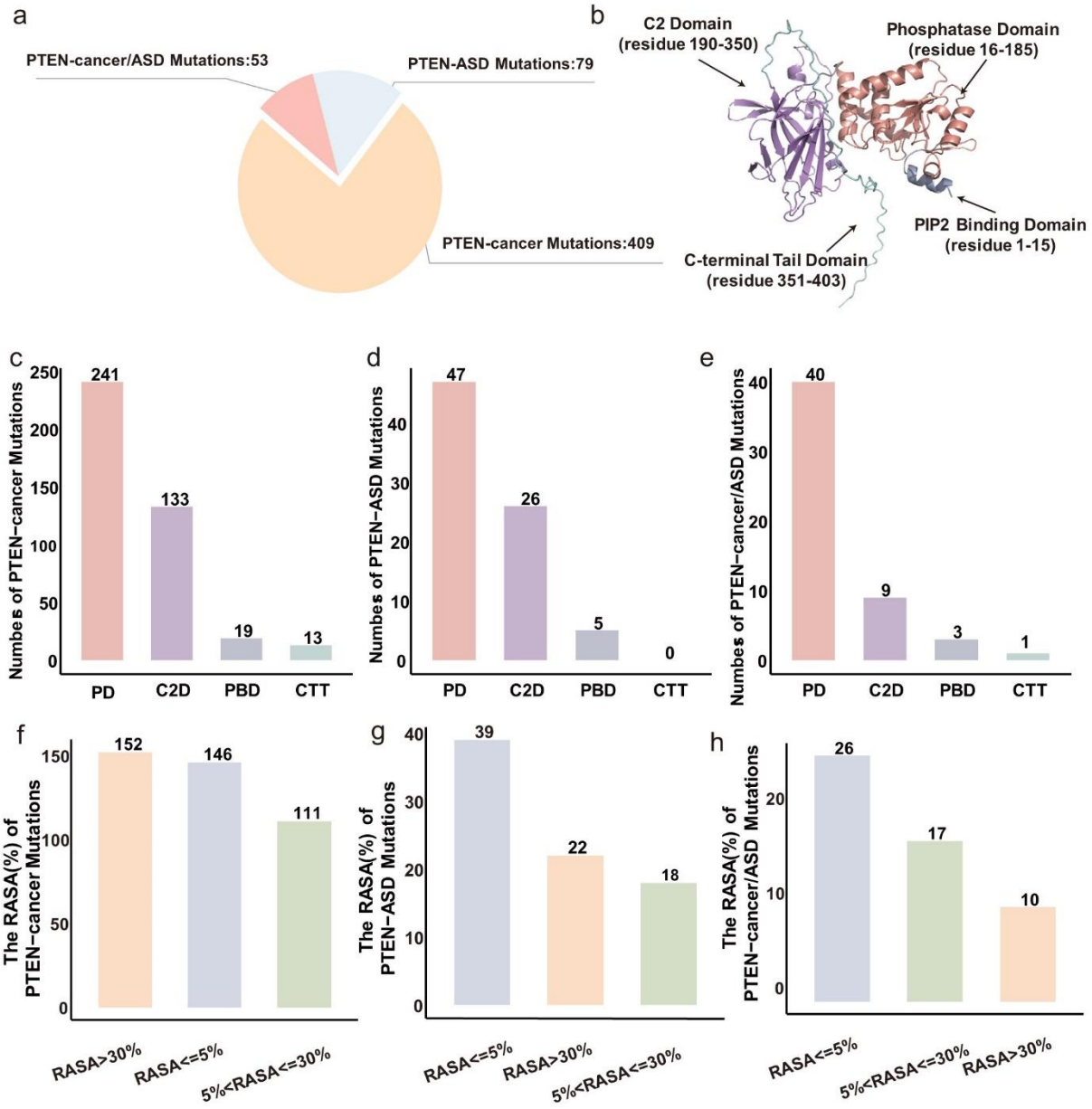


Figure 3. Landscape of PTEN mutation data. (a) Statistical distribution of PTEN missense mutations categorized into three phenotypic groups: PTEN-cancer (orange), PTEN-ASD (blue), and PTEN-cancer/ASD (red). (b) Full-length AlphaFold structure of PTEN highlighting its functional domains: PBD, PD, C2D, and CTT domain. (c) PTEN-cancer, (d) PTEN-ASD, (e) PTEN-cancer/ASD groups' distribution across the PBD, PD, C2D, and CTT. (f) PTEN-cancer, (g) PTEN-ASD, (h) PTEN-cancer/ASD groups' *RASA* calculation results.

To elucidate the sequence characteristics underlying the three types of PTEN mutations, three sequence-based signature including conservation score, $S(i)$, and MI were calculated. Using the conservation score threshold of ≥ 5 to define conserved residues, 215 out of 403 residues (53.34%) in the PTEN sequence were identified as conserved (Figure S2). As depicted in Figure 4a, conserved residues were mutated in 68% of the PTEN-cancer group (183/242), 81% of the PTEN-ASD group (60/67), and 83% of the PTEN-cancer/ASD group (37/43) (Table S3). This distribution indicated a higher propensity for PTEN-cancer/ASD mutations to occur at conserved sites. According to $S(i)$, 63% of conserved residues (259/409) were mutated in the PTEN-cancer group, 73% (58/79) in the PTEN-ASD group, and 81% (43/53) in the PTEN-cancer/ASD group (Figures 4b, Table S3). Co-evolutionary analysis revealed that PTEN-cancer/ASD mutations exhibit lower co-evolutionary scores compared to PTEN-cancer and PTEN-ASD mutations (Figure 4c, Table S3). This suggests that PTEN-cancer/ASD mutations may have a more significant impact on protein function due to their occurrence in residues that are less evolutionarily constrained. Statistical comparisons of conservation score, $S(i)$, and MI among the mutation groups revealed significant differences between the PTEN-cancer and PTEN-ASD groups, as well as between the PTEN-cancer and PTEN-cancer/ASD groups (Figures 4d-f).

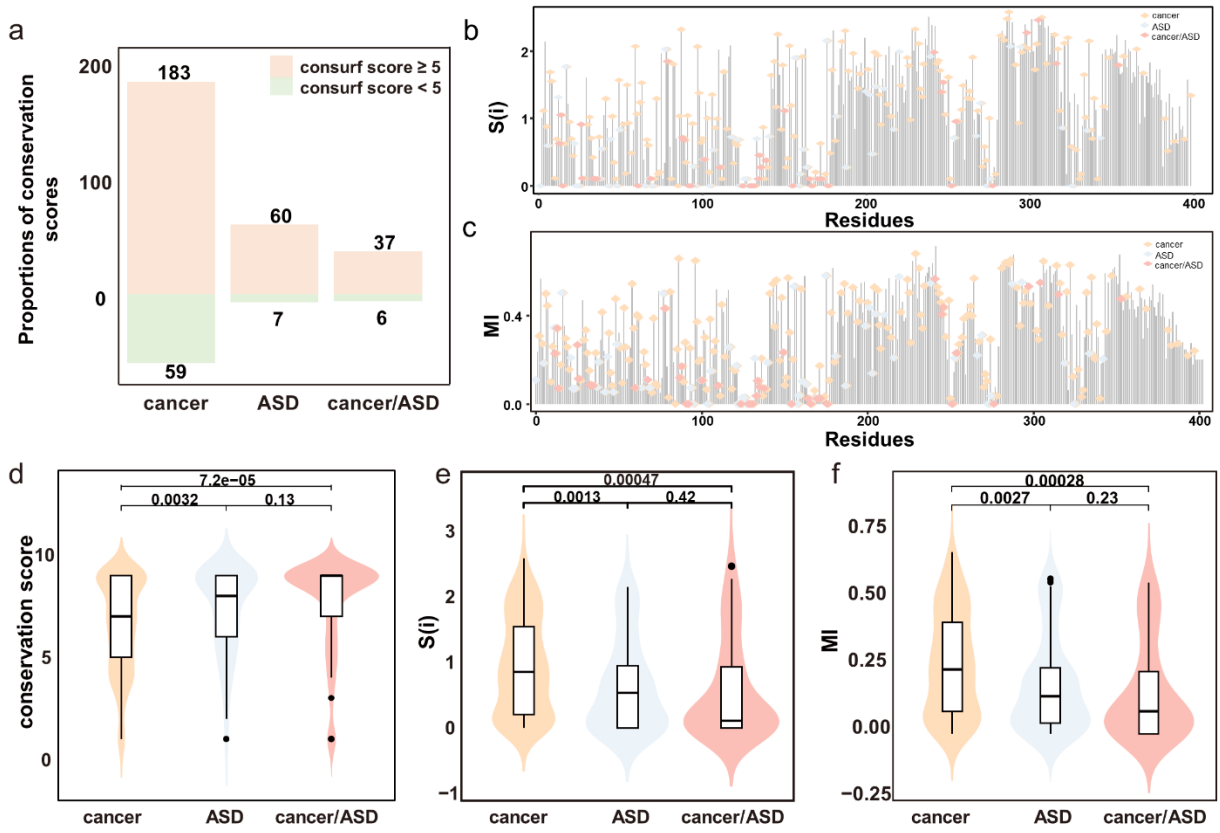


Figure 4. Sequence landscape of PTEN mutations. (a) Proportions of conservation scores for PTEN missense mutations. (b) $S(i)$, and (c) MI values for three types of PTEN mutations. (d-f) Significant differences (P-values<0.05) were observed in all three sequence features between the PTEN-cancer group and the other two groups.

Distinct Energetic Landscapes of PTEN-cancer/ASD mutations. To elucidate the thermodynamic determinants of mutation hotspots and their functional relevance, FoldX[41] was employed to construct mutants for each mutation group and calculate the local free energy changes induced by these mutations (Table S4). As depicted in Figures 5a and 5b, compared with the mutations related to ASD, significant energy changes could be observed between the PTEN- cancer/ASD and the cancer mutations. Specifically, the most pronounced energy changes were observed in mutations localized to the PD and the C2D (Figure 5c) ,which were particularly

enriched in critical functional regions, such as the P loop and TI loop[13]. The top 10 PTEN-cancer/ASD mutations with the largest increase in free energy include C136R ($\Delta\Delta G$=12.456 kcal/mol), I135R($\Delta\Delta G$=12.252 kcal/mol), G132D($\Delta\Delta G$=11.151 kcal/mol), T26P($\Delta\Delta G$=9.521 kcal/mol), G165V($\Delta\Delta G$=8.964 kcal/mol), F241S($\Delta\Delta G$=6.012 kcal/mol), D252G($\Delta\Delta G$=5.086 kcal/mol), M35V($\Delta\Delta G$=4.846 kcal/mol), I135K($\Delta\Delta G$=4.720 kcal/mol), Y68N($\Delta\Delta G$=4.5985 kcal/mol), which may induce severe disturbances in intramolecular dynamics or impair the interactions[23, 42, 43] Among them, I135R and C136R, located in the PD, are classified as severe pathogenic mutations implicated in both cancer and ASD[23]. Similarly, the G165V mutation, situated in the TI loop, is identified as a loss-of-function variant, further highlighting its detrimental effect on PTEN function[43]. These findings underscore the substantial thermodynamic and functional impact of PTEN-cancer/ASD mutations, particularly those affecting key structural regions essential for PTEN activity and regulation.

However, based on our energy analysis, other key PTEN-cancer/ASD mutations, such as R130Q($\Delta\Delta G$=1.827 kcal/mol), R173C($\Delta\Delta G$=3.115 kcal/mol), Y155C($\Delta\Delta G$=4.054 kcal/mol) did not exhibit significant energy increases with only C136R($\Delta\Delta G$=12.456 kcal/mol) showing a substantial energy increase as shown in Figure 5c and 5d. R130Q, and Y155C are located within the ATP-binding type A region of the active site, whereas the R173C is located at the interdomain interface of Motif 1[23]. Additionally, some mutations were found to enhance rather than destabilize protein stability. For instance, the H93R mutation ($\Delta\Delta G$=-1.430 kcal/mol), located in WPD loop, is strongly associated with both cancer and ASD, and its importance has been highlighted in previous studies[42, 44]. Similarly, the Y167N mutation ($\Delta\Delta G$=-0.009 kcal/mol), located near Motif 1, exhibited increased protein stability in our energy analysis, and has been reported to induce long-range conformational changes, significantly perturbing intramolecular

dynamics[23]. Moreover, the A79T mutation ($\Delta\Delta G$=0.132 kcal/mol) located near the P loop, has been linked to potential functional disruption of PTEN[45].

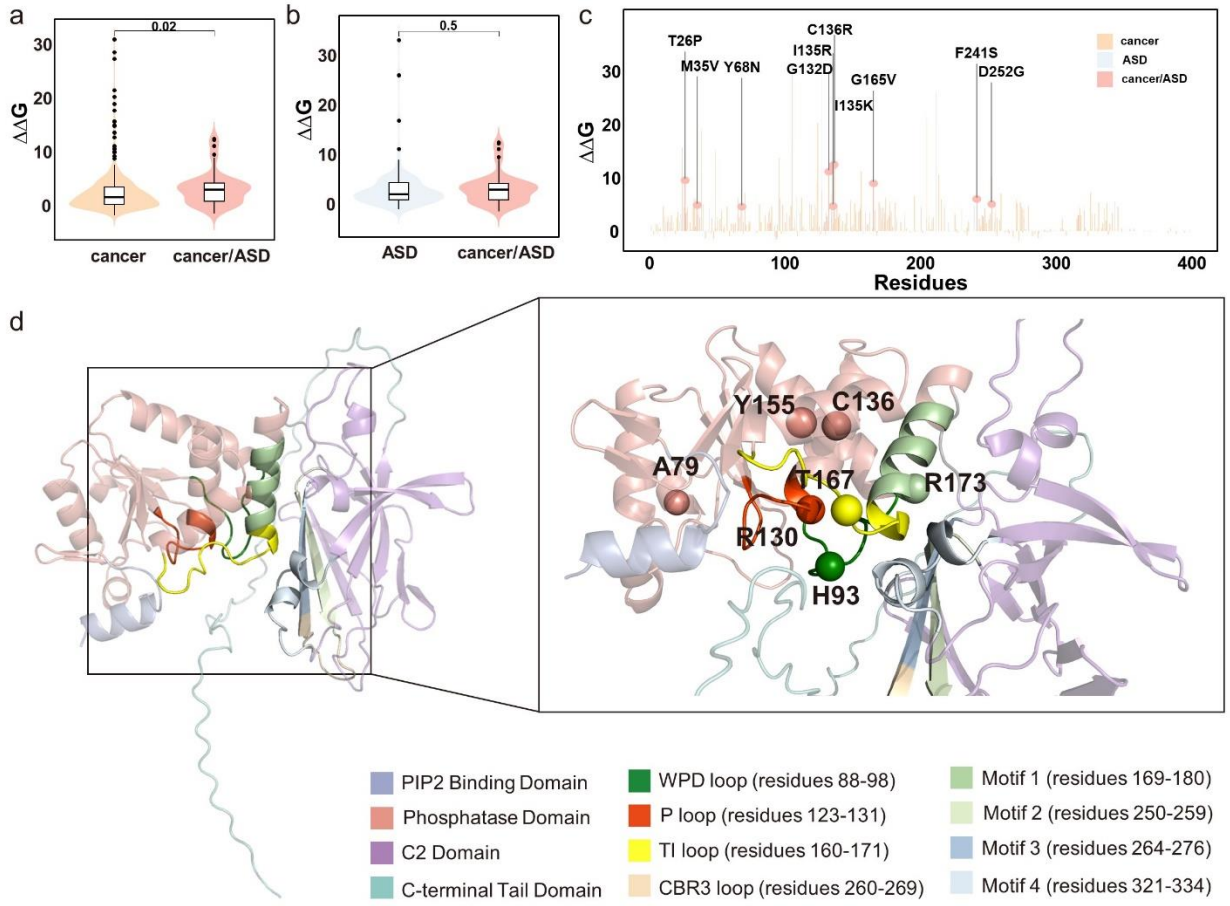


Figure 5. Energetic landscapes of PTEN-cancer/ASD mutations. Significance test analysis of free energy change between PTEN-cancer/ASD (red) and (a) PTEN-cancer mutations (yellow) as well as (b) PTEN-ASD mutations (blue) respectively. (c) Distribution of free energy changes in the PTEN protein sequence for PTEN-cancer, PTEN-ASD mutations, and PTEN-cancer/ASD mutations. The top 10 PTEN-cancer/ASD mutations with the highest free energy changes are marked in red font. (d) Detailed visualization of specific mutations and their corresponding free energy changes in the PTEN sequence.

Our free energy profiles support previous findings that strong hotspot mutations lead to a cancer phenotype, while weak or moderate mutations are associated with NDDs[46]. However, challenges remain in revealing the complex landscape of energy changes induced by PTEN-cancer/ASD mutations, as their heterogeneous consequences may depend on the cellular context and dynamics.

The network and dynamic landscape of PTEN-cancer/ASD mutations. To systematically assess the structural and dynamics impacts of PTEN mutations, we employed an amino acid contact energy network and computed five topological features for each mutation. The differences in network centrality metrics between mutant and WT proteins were calculated to characterize the topological changes induced by PTEN mutations (Figure 6, Table S5).

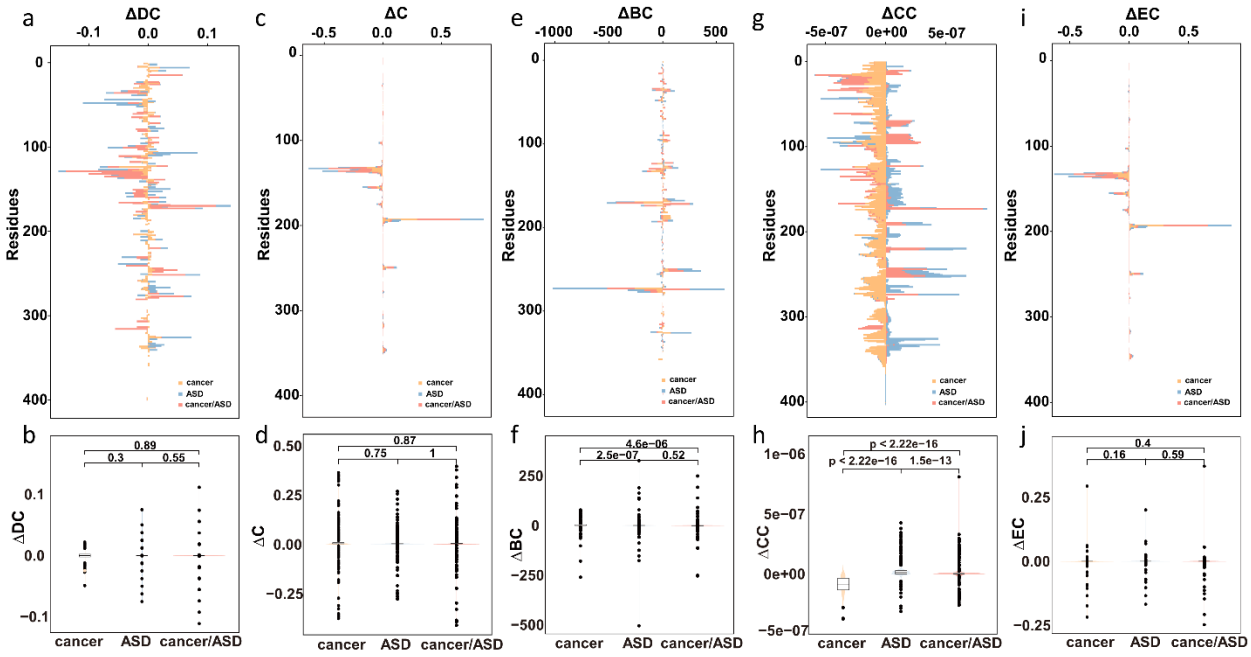


Figure 6. Network analysis of PTEN mutations. Distribution of the difference in (a) ΔDC, (c) ΔC, (e) ΔBC, (g) ΔCC, (i) ΔEC, respectively, between the mutated and the WT. Statistical analysis of (b) ΔDC, (d) ΔC, (f) ΔBC, (h) ΔCC, (j) ΔEC, respectively, across groups.

Examination of ΔDC revealed varied topological alterations among the groups. The PTEN-cancer/ASD cohort exhibited a broader distribution of ΔDC compared to the PTEN-cancer and PTEN-ASD groups, indicating differential degrees of network centrality perturbation (Figure 6a). However, no significant differences in ΔDC were observed across the groups (Figure 6b). Similarly, the assessment of the ΔC demonstrated no statistically significant variations among the PTEN-cancer, PTEN-ASD, and PTEN-cancer/ASD groups (Figures 6c and 6d), which suggests that clustering behavior within the network remains relatively unaffected across the different PTEN mutation groups. In contrast to ΔDC and ΔC, ΔBC analysis highlighted distinct network disruptions (Figures 6e and 6f). The PTEN-cancer/ASD group displayed a dispersed ΔBC distribution, indicating intermediate topological changes compared to the pronounced alterations in critical residues observed in the PTEN-cancer group and the relatively stable profile with fewer outliers in the PTEN-ASD group. Moreover, statistical network analysis revealed significant differences in ΔBC between the PTEN-cancer and PTEN-ASD groups (P-value=$2.5e^{-7}$) as well as between the PTEN-cancer and PTEN-cancer/ASD groups (P-value=$4.6e^{-6}$). Notably, the PTEN-cancer/ASD group exhibited the highest ΔBC among all cohorts, whereas no significant difference was observed between the PTEN-ASD and PTEN-cancer/ASD groups. Furthermore, the evaluation of ΔCC indicated substantial dispersion within the PTEN-cancer/ASD group compared to the PTEN-cancer and PTEN-ASD groups (Figures 6g and 6h). Statistical comparisons confirmed significant differences in ΔCC among all group pairings: PTEN-cancer vs. PTEN-ASD (P-value<$2.22e^{-16}$), PTEN-cancer vs. PTEN-cancer/ASD (P-valuee<$2.22e^{-16}$), and PTEN-ASD vs. PTEN-cancer/ASD (P-value=$1.5e^{-13}$). Specifically, the PTEN-cancer/ASD mutations exhibited a reduction in ΔCC, highlighting notable alterations in network closeness centrality within this group. In contrast, the analysis of ΔEC did not reveal

significant differences among the three groups (Figures 6i and 6j), suggesting that the influence of individual residues within the overall network remains consistent across the different PTEN mutation contexts.

Given the significant distinctions observed in ΔBC and ΔCC between the PTEN-cancer/ASD group and the PTEN-cancer or PTEN-ASD groups, further analysis was conducted to elucidate the underlying structural implications. Residues exhibiting elevated ΔBC changes in the PTEN-cancer/ASD group were primarily located at interdomain interfaces and in the vicinity of the P loop. In contrast, mutations within the PTEN-cancer group were predominantly observed around the TI loop and at the junction between the PD and C2D, while those in the PTEN-ASD group were concentrated at interdomain interfaces. Additionally, the mutations associated with the most substantial decreases in ΔCC (A34V, I135R, I135K, H61R, H61Y) were all localized within the PD. Notably, H61R and H61Y are situated in the ATP Binding Type A motif, whereas I135R and I135K are located in the WPD loop, a region critical for substrate catalysis. These findings underscore the differential impact of PTEN mutations on network centrality metrics and further corroborate our previous results regarding the structural and functional significance of PTEN-cancer/ASD mutations[23].

We performed ENM modeling to investigate the dynamic effects of PTEN-cancer/ASD mutations (Table S6). Distinct trends were observed in the effectiveness, DFI, stiffness, MSF, and sensitivity metrics for the three mutation types, particularly in their distribution within critical functional regions of the PD (Figure 7a-7e). Our analysis revealed that PTEN-cancer/ASD mutations are characterized by higher effectiveness, DFI, and stiffness values, coupled with lower MSF and sensitivity values. These mutations predominantly cluster within

the ATP Binding Motif Type A, especially around the P loop, within the PD. In contrast, PTEN-ASD and PTEN-cancer mutations exhibit broader distributions, including regions such as the WPD loop.

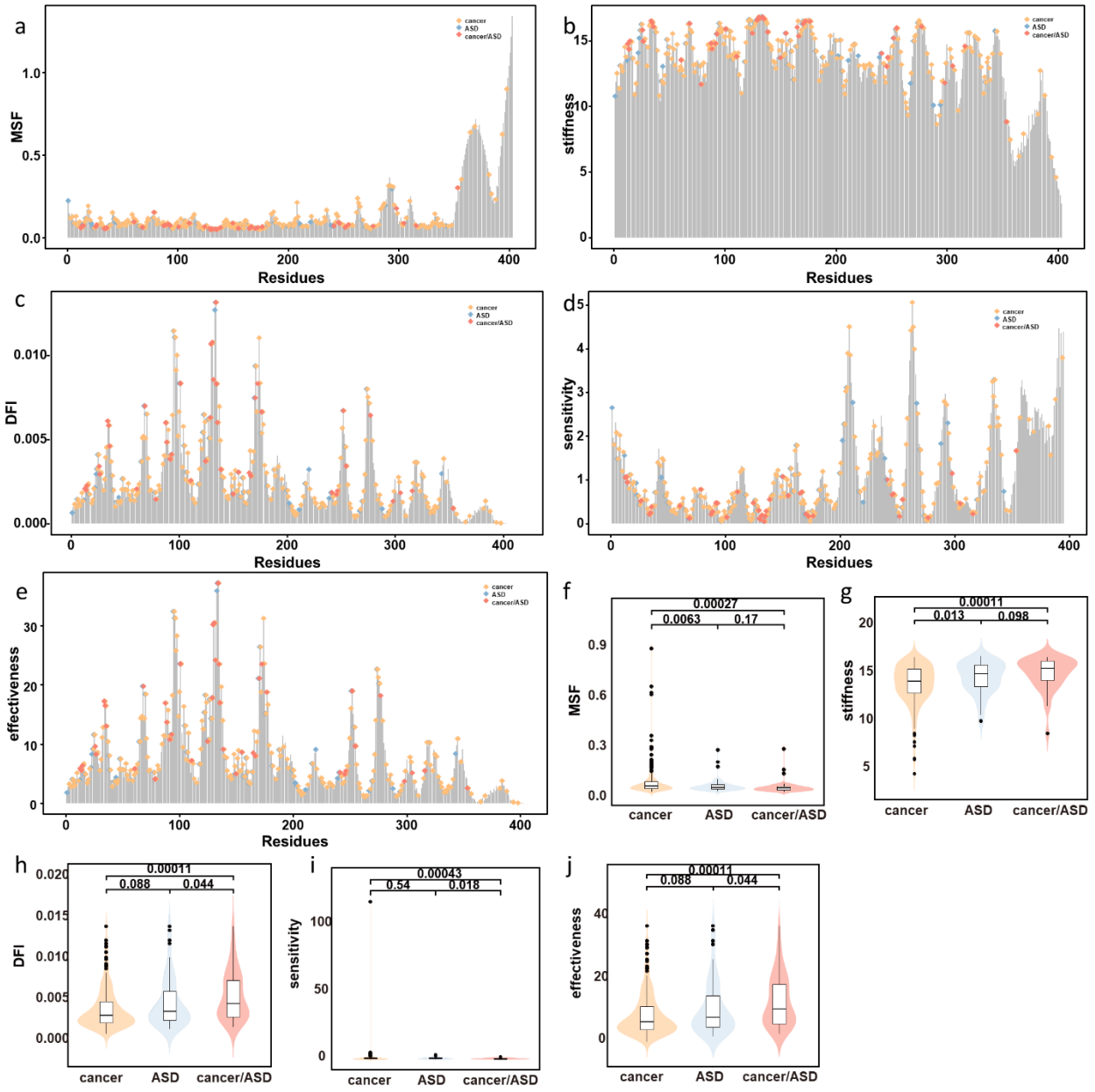


Figure 7. Elastic network model-based features of PTEN mutations. (a) *MSF*, (b) stiffness, (c) *DFI*, (d) sensitivity, (e) effectiveness values, respectively, across PTEN residues (b) Statistical

comparison of (f) MSF, (g) stiffness, (h) DFI, (i) sensitivity, (j) effectiveness, respectively, values among PTEN-cancer, PTEN-ASD, and PTEN-cancer/ASD groups.

Statistical analysis revealed significant differences in MSF only between the PTEN-cancer and PTEN-ASD groups (P-value=0.0063) and between PTEN-cancer and PTEN-cancer/ASD groups (P-value=0.00027) (Figure 7f). Similarly, stiffness values showed significant differences only between the PTEN-cancer and PTEN-ASD groups (P-value=0.013), and between PTEN-cancer and PTEN-cancer/ASD groups (P-value=0.00011) (Figures 7g). The DFI analysis demonstrated significant differences across all three groups (Figures 7h), highlighting unique flexibility profiles associated with PTEN-cancer/ASD mutations. Sensitivity values differed significantly between the PTEN-ASD and PTEN-cancer/ASD groups (P-value=0.018) and between PTEN-cancer and PTEN-cancer/ASD groups (P-value=0.00043), with no significant difference observed between PTEN-cancer and PTEN-ASD groups (Figures 7i). Effectiveness values also exhibited significant differences across all three groups (Figures 7j).

These findings suggest that PTEN-cancer/ASD mutations confer distinct dynamic properties to the protein, particularly in regions critical for its catalytic and regulatory functions. The observed increases effectiveness, DFI, and stiffness may indicate altered allosteric communication and reduced flexibility, potentially impacting the interactions of PTEN with substrates and regulatory proteins. Conversely, the decreases in MSF and sensitivity reflect a more rigid structural environment, potentially compromising PTEN's functional dynamics. Overall, the unique ENM-derived features associated with PTEN-cancer/ASD mutations highlight their substantial impact on PTEN's structural and functional integrity, providing insight into the molecular mechanisms underlying their dual roles in cancer and NDDs.

Machine Learning Classification and integrated score for PTEN-cancer/ASD mutations. To effectively distinguish PTEN-cancer/ASD mutations from PTEN-cancer and PTEN-ASD mutations, ROC analysis was conducted on 15 molecular features. The contribution of each feature to phenotype differentiation was evaluated. Based on these selected features, classification models were developed to differentiate among the PTEN-cancer, PTEN-ASD, and PTEN-cancer/ASD groups. Comparison of the AUC values for the 15 features reveled that ΔCC demonstrated the highest discriminatory power across the three groups (Figures 8a-c, Figure S3). Specifically, for distinguishing between the PTEN-cancer and PTEN-ASD groups,ΔCC achieved an AUC of 0.9311, indicating high reliability. However, when distinguishing PTEN-cancer/ASD from PTEN-cancer groups and PTEN-ASD groups, the AUC values for ΔCC were only 0.7747 and 0.6883, respectively, highlighting challenges in achieving precise differentiation.

Machine learning methods were employed to iteratively select multiple features until the AUC exceeded 0.8, demonstrating the model's ability to effectively distinguish the PTEN-cancer/ASD group from the other two groups (Figures 8d and 8f). The final feature set revealed that distinguishing PTEN-cancer from PTEN-cancer/ASD mutations could be achieved using a combination of ΔCC, ΔBC, ΔEC, sensitivity, conservation score, effectiveness, stiffness, *MI*, $RASA$, and $\Delta\Delta G$ (Figure 8e). And, distinguishing PTEN-ASD from PTEN-cancer/ASD mutations could be accomplished using a combination of ΔCC, ΔEC, sensitivity, ΔBC, $RASA$, and $\Delta\Delta G$ (Figure 8g). Notably, the selected feature sets consistently included at least one feature from sequence-based, structural, and dynamical properties. These highlights that the model comprehensively accounts for all potential contributors to the distinct allosteric properties associated with these different phenotypes. The top four features critical for distinguishing the

PTEN-cancer/ASD group were ΔCC, ΔBC, ΔEC, and sensitivity, which are crucial for revealing the allosteric properties of the PTEN-cancer/ASD mutations.

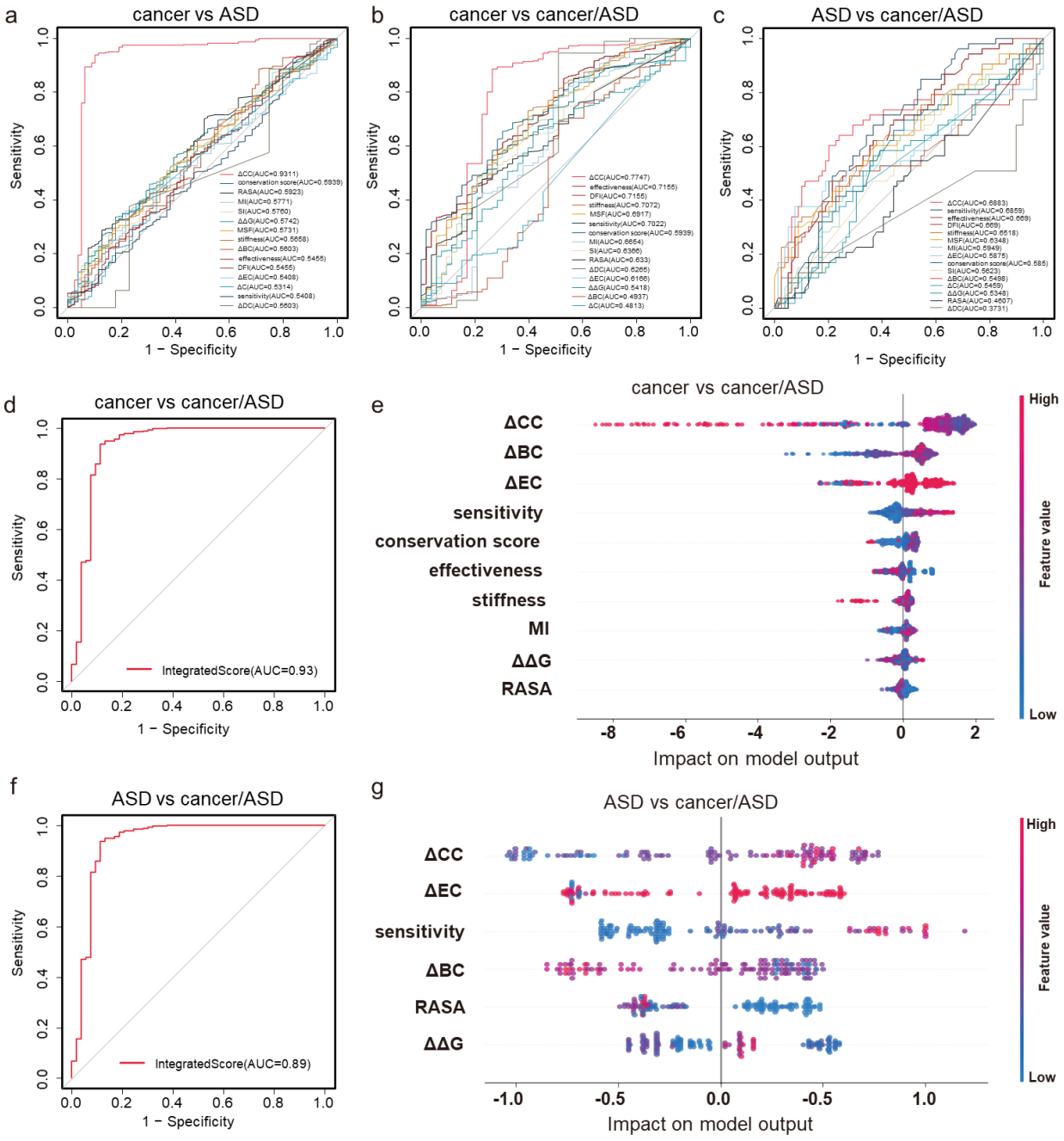


Figure 8. Feature selection to distinguish different phenotypes. The ROC curve and model interpretation plot for single feature selection for distinguishing the PTEN-cancer group from the (a) PTEN-ASD group and (b) PTEN-cancer/ASD group. (c) The ROC curve and model

interpretation plot for single feature selection for distinguishing PTEN-ASD group from PTEN-cancer/ASD group. (d) The ROC curve and (e) model interpretation plot for the ML model used to distinguish the PTEN-cancer group from the PTEN-cancer/ASD group. This model was generated by the optimal feature combination composed of ΔCC, ΔBC, ΔEC, sensitivity, conservation score, effectiveness, stiffness, MI, $RASA$, and $\Delta\Delta G$. (f) The ROC curve and (g) model interpretation plot for the ML model used to distinguish the PTEN-ASD group from the PTEN-cancer/ASD group. This model included multiple features consisting of ΔCC, ΔEC, sensitivity, ΔBC, $RASA$, and $\Delta\Delta G$.

Additionally, the developed ML model can generate an IS for multi-type variants, facilitating a more granular classification of similar phenotypes. The IS confirmed that PTEN-cancer/ASD mutations, including R130Q, R173C[13, 14, 23], Y155C[13, 43], and C136R[13, 23], all exhibited scores greater than 0.5, validating their classification as PTEN-cancer/ASD mutations and further supporting the model's effectiveness and accuracy.

Previous studies have suggested that PTEN-ASD mutations predominantly affect local regions, especially within the PD, causing local allosteric effects that disrupt communication and functionality within the domain[13]. In contrast, PTEN-cancer mutations typically induce global allosteric effects, influencing the conformation and function of the entire protein via long-range structural communication pathways. However, the specific nature of communication (local, global, or both) for PTEN-cancer/ASD mutations remains unclear. Additionally, we also identified several PTEN-cancer/ASD mutations with exhibited high IS compared to PTEN-cancer and PTEN-ASD mutations (Figure S4, Table S7), including G132A, R130P, T131I and Y68N. These mutations are believed to exhibit significant allosteric properties. To investigate

further, MD simulations were performed on these four PTEN-cancer/ASD mutations to explore their dynamic effects in greater detail.

Dynamics of PTEN-cancer/ASD mutations: local open conformation coupled with long-range allosteric communications. Through MD simulations (see Text S2), we observed that the four PTEN-cancer/ASD mutations — Y68N, R130P, T131I, and G132A (Figures S5 and S6) — differentially affected the flexibility of both local and distal amino acid residues within the PTEN protein. All four mutations altered the flexibility of key functional loops (WPD, P, and TI loops), transitioning them from a closed to an open conformation. Specifically, the R130P, T131I, and G132A mutations in the P loop notably increased the distance between residues R130 and D92 (Figure 9a). This disruption of the compact WT conformation was especially evident in the TI loop, which exhibited increased flexibility and a marked deviation from its native closed conformation, suggesting structural destabilization that could affect distal regions (Figure S7). MD simulations were conducted for 500 ns for each of the four mutations and the WT PTEN, eventually reaching relative stability (Figure 9b). The Y68N mutation not only affected Region 1 and Region 2 in the PD but also influenced the flexibility of residue A79, transmitting the allosteric signal to a more distal region within the same domain. This change was reflected in the reduced stability of residue A79 and its surrounding area, indicating a structural alteration in the distal region of the PD (Figures 9c and 9d). Similarly, the R130P mutation significantly impacted the stability of residues H93 and T167, located on the WPD loop and TI loop, respectively. The mutation at position 130 residue within the P loop, further confirmed the functional synergy between the WPD loop, TI loop, and P loop (Figures 9e and 9f). The T131I mutation notably affected the stability of Region0 (residue: L42-V45) centered around residue E43 in the same

domain (Figures 9g and 9h). In contrast, the G132A mutation transmitted the signal to the distal residue G143, leading to changes in the stability of the surrounding region (Figures 9i and 9j).

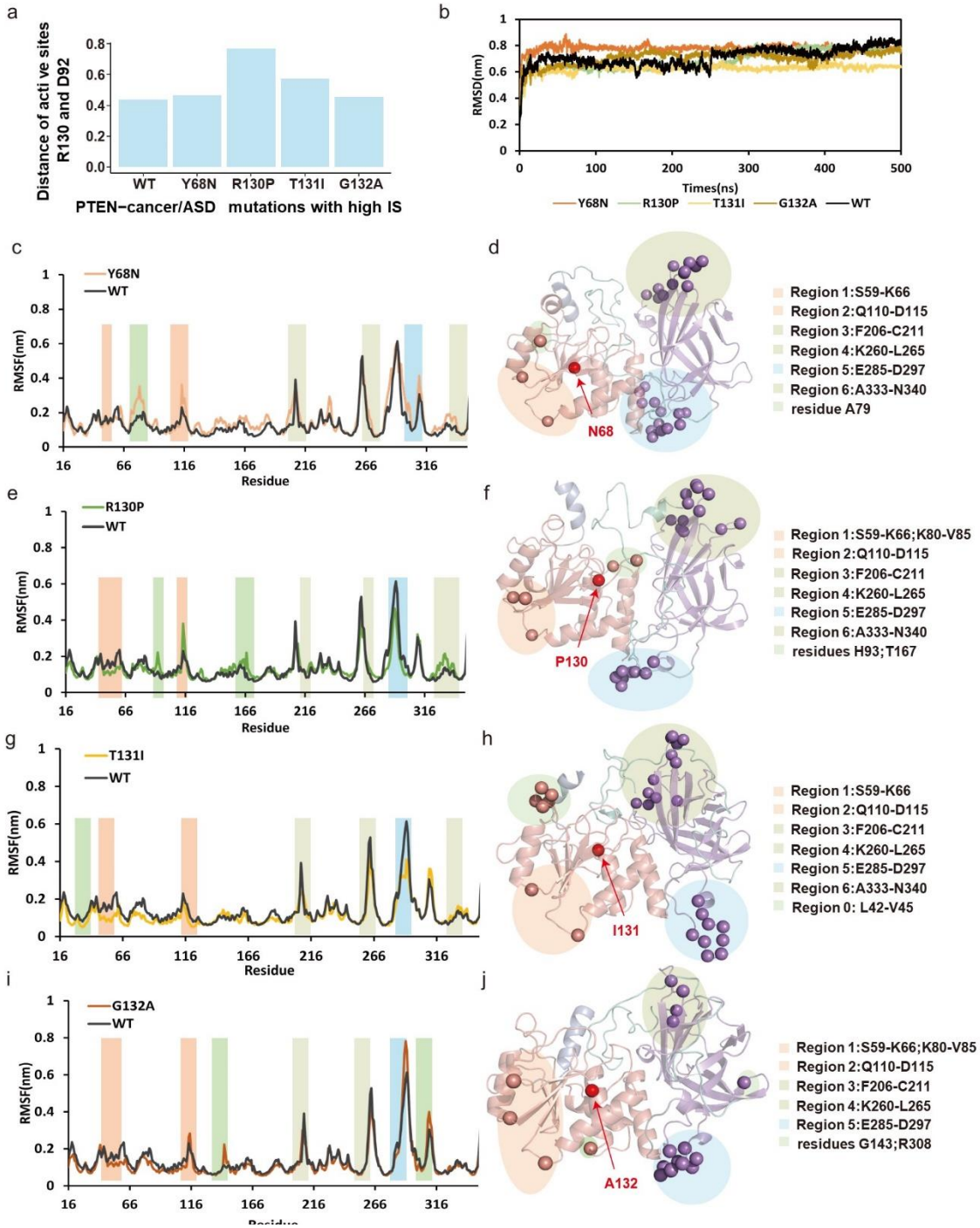


Figure 9. Conformational effects in WT PTEN system and high-scoring PTEN-cancer/ASD systems. (a) Distance of active sites D92 and R130 with PTEN-cancer/ASD mutations. (b)

RMSF results for residues in the WT PTEN and mutants Y68N, R130P, T131I, and G132A. RMSF for (c) Y68N, (e) R130P, (g)T131I and (i) G132A and Spatial distribution Spatial distribution of the residues with significant flexibility changes due to the (d) Y68N, (f) R130P, (h)T131I and (j) G132A mutations. Detailed information about these regions is provided in the Figure S1.

These results suggest that PTEN-cancer/ASD mutations with high scores in the P loop region not only affect the stability of the active site but also influence the stability of key functional loop regions. Furthermore, these mutations influence the stability of residues surrounding the mutation site and potentially transmit their effects to more distal regions, suggesting that such high-scoring mutations exhibit stronger allosteric effects. The residues with the highest flexibility changes in the Y68N, R130P, T131I, and G132A mutants were L295, T286, T286, and K289, respectively, indicating potential allosteric communication between the mutation sites and these regions. Utilizing a dynamic network model-based shortest path algorithm, our results demonstrate that mutants exhibit shorter allosteric pathways, further supporting our hypothesis that PTEN-cancer/ASD mutations induce a clear allosteric tendency (Figure 10).

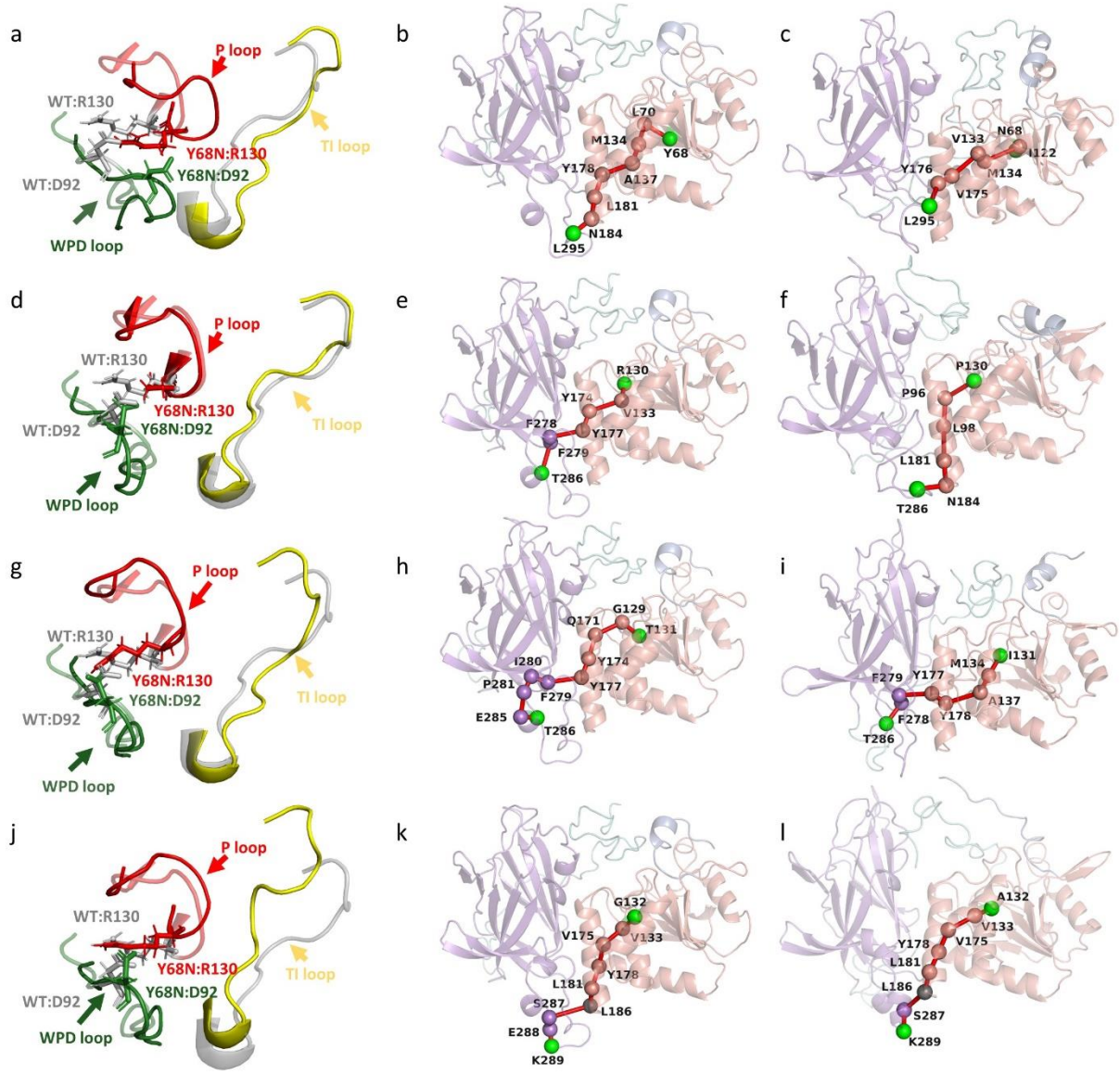


Figure 10. Conformational pathways of high-IS PTEN-cancer/ASD mutations selected by the model. (a) Local conformational changes in the active site of WT-Y68 and mutant Y68N. Conformational pathway of (b) WT-Y68, (c) mutant Y68N, with the starting point at residue 68 and the endpoint at residue L295. (d) Local conformational changes in the active site of WT-R130 and mutant R130P. Conformational pathway of (e) WT-R130, (f) mutant R130P, with the starting point at residue 130 and the endpoint at residue T286. (g) Local conformational changes in the active site of WT-T131 and mutant T131I. Conformational pathway of (h) WT-T131, (i)

mutant T131I, with the starting point at residue 131 and the endpoint at residue T286. (j) Local conformational changes in the active site of WT-G132 and mutant G132A. Conformational pathway of (k) WT-G132, (l) mutant G132A, with the starting point at residue G132 and the endpoint at residue K289.

In the WT Y68, the shortest path from residue Y68 to residue L295 required passing through the L181 and N184 regions to transmit the signal to the distal region. However, following the Y68N mutation, the active site adopts an open conformation (Figure 10a), and the shortest path bypasses the L181 and N184 regions, instead choosing a shorter, more direct route to transmit the signal to the distal region (Figures 10b and 10c). This alteration is likely due to the opening of the active site region. Furthermore, after the Y68N mutation, the shortest path no longer traverses residue L70, but shifts to a path passes through the P loop. Similarly, the R130P mutation also resulted a shorter, more direct pathways following the opening of the region (Figure 10d). Before the R130 mutation, the optimal shortest path predominantly passed through Motif 1, reaching the C2D to transmit the allosteric signal to residue T286. After the R130P mutation, the shortest path navigated through the WPD loop and interdomain residues L181 and N184 to transmitting the signal to T286. While the path before and after the mutation was similar, the post-mutation path short, involving fewer residues passing through the WPD loop. This alteration is likely related to the open states of the WPD loop, P loop, and TI loop (Figures 10e and 10f). The T131I and G132A mutants also exhibited shortened paths, with signals passing through residues Y177 and L186 (Figures 10g-l). These findings suggest that PTEN-cancer/ASD mutations with high IS not only perturb the stability of the active site but also propagate this disturbance to distal regions, inducing conformational changes in those areas.

In these pathways, we observed that, in the wild-type PTEN, signal propagation largely depends on interdomain interfaces and key surrounding residues such as L181, N183, and N184. However, PTEN-cancer/ASD mutations tend bypass these interdomain connection points, transmitting the signal to distal regions. Molecular feature analysis further reveals that residues V133, M134, Y177, and Y178 exhibit relatively high effectiveness and low sensitivity, serving as key sites for structural signal propagation between the mutated domain and interdomain regions.

4. Conclusions

Predicting the specific phenotype of missense mutations is critical for advancing precision medicine. PTEN, a well-studied protein, is associated with diverse mutations linked to multiple phenotypes, including cancer and ASD. While numerous studies have focused on distinguishing PTEN-cancer and PTEN-ASD mutations, PTEN-cancer/ASD mutations present distinct molecular characteristics, integrating both oncogenic and neurodevelopmental effects within a shared structural framework. This dual functionality complicates the understanding of their specific pathways and raises two key questions: (1) Why do certain PTEN mutations promote both cancerous and neurodevelopmental phenotypes in different individuals or contexts? (2) How do these dual-faceted mutations structurally modulate PTEN's activity, resulting in variability in disease outcomes?

In this study, we developed an integrative ML model combined with structural dynamics and network-based analyses to systematically evaluate PTEN-cancer/ASD mutations. Previous ML approaches have focused on classifying PTEN mutations based on their molecular consequences on protein structure and function, distinguishing PTEN-cancer, ASD, and non-pathogenic

mutations [17]. Our model builds on this foundation by incorporating four network and dynamic-based features[47], introduce four network/dynamics-based features—ΔBC, ΔCC, ΔEC, and sensitivity—which complement the conventional set of molecular descriptors (e.g., $\Delta\Delta G$). Compared to the commonly used import features, ΔCC and $\Delta\Delta G$, these new features collectively form an "intrinsically accessible spectrum of modes of motions" that enables adaptation to adapt to different environments and interactions while maintaining its structural fold and functional integrity. Among these dynamics features, allosteric signaling effectiveness emerged as a key factor. Overall, our model underscores the critical role of structural dynamics in predicting the phenotypic effects of missense mutations. This approach led to the development of a novel integrative score, which not only classifies PTEN mutations but also ranks the likelihood of mutations belonging to the PTEN-cancer/ASD group.

To enhance the interpretability of ML predictions and the underlying molecular mechanism, we performed MD simulations combined with network analysis on selected PTEN-cancer/ASD mutations. Several comprehensive computational studies have previously leveraged MD simulations for analyze PTEN mutations. For example, previous investigations have examined the conformational alterations of the catalytic core in the PD[24], while other studies have demonstrated the allosteric effects of PTEN missense mutations[48]. By integrating network analysis, MD simulations have revealed distinct differences in allosteric regulation between PTEN-ASD and PTEN-cancer mutations[21]. These differences arise from the coupled interplay of CTT phosphorylation dynamics. Notably, two key studies employing MD-based computational approaches provided valuable insights into the genotype-phenotype relationships of PTEN-cancer and PTEN-ASD mutations[13, 14]. Their findings reveal that 1) PTEN-cancer mutations induce long-range communication pathways that span the inter-domain interface, while

maintaining a closed conformation at the active site, 2) PTEN-ASD mutations cause localized destabilization restricted to the PD, leading to partial opening of the active site. Building on these studies, we applied a similar MD-network-based method to investigate the dynamic effects of PTEN-cancer/ASD mutations. Our results demonstrate that PTEN-cancer/ASD mutations exhibit coupled dynamics characterized by local conformational changes and long-range allosteric communications. Specifically, these mutations induce a partially open conformation at the P loop, altering PTEN's structural dynamics. Accordingly, we hope that our work will expand the genotype-phenotype map of PTEN mutations, shedding light on the unique pathways associated with PTEN-cancer/ASD mutations.

To enhance the of our ML models, the structural dynamics analysis provided valuable insights into the molecular mechanisms underlying PTEN-cancer/ASD mutations. Among the most significant features, ΔCC and ΔBC capture signatures of local conformational changes, while effectiveness and sensitivity elucidate how mutations act as molecular drives of allosteric communication pathways. Furthermore, our results highlight the potential of ML models in precision medicine, particularly in the context of PTEN-cancer/ASD mutations. One promising strategy involves targeting mutations located at active sites or orthosteric sites by designing effective modulators to control local conformational change. Another viable approach is the development of allosteric drugs that modulate the long-range dynamics caused by these mutations. Interestingly, mutations in other drug targets[49, 50] show similar dynamics profiles, with both localized effects and long-range allosteric changes. This observation suggests the combining orthosteric and allosteric drugs could form a transferable therapeutic paradigm for treating multi-phenotypic mutations[51].

ASSOCIATED CONTENT

Data Availability Statement

The codes for ML model and the data for MD simulations are available at https://zenodo.org/records/14636196.

Supporting Information.

Text S1. Detailed information of structural modeling, molecular feature calculations, and MD simulations.

Text S2. The results of the MD simulations.

Figure S1. PTEN structure and mutation mapping.

Figure S2. Sequence conservation scoring results of PTEN.

Figure S3. Feature correlation analysis.

Figure S4. Distribution of IS for PTEN-cancer/ASD mutations.

Figure S5. MD trajectory convergence and reproducibility.

Figure S6. RMSF comparison among independent MD replicates

Figure S7. Conformational effects in wild-type PTEN and high-IS PTEN-cancer/ASD mutants.

Table S1. Overview of collected PTEN missense mutation data. These data were curated from public databases, ensuring a broad and representative set of pathogenic variants for subsequent analyses.

Table S2. RASA values computed by PASIS for PTEN. These data were used to evaluate whether mutations were buried, partially exposed, or exposed, providing important structural context for PTEN-cancer, PTEN-ASD, and PTEN-cancer/ASD mutations.

Table S3. PTEN sequence-based features computation results. These data were used to examine how sequence conservation and co-evolutionary differ across the three PTEN mutation groups.

Table S4. Folding free energy (ΔΔG) computation results. These data reveal the thermodynamic impacts of mutations and highlight notable hotspots in PTEN-cancer/ASD variants.

Table S5. PTEN amino acid contact energy network features. These network-based features are pivotal in distinguishing local vs. global structural perturbations among PTEN variants.

Table S6. Elastic network model-based dynamic features. These data elucidate the dynamic consequences of PTEN-cancer, PTEN-ASD, and PTEN-cancer/ASD mutations.

Table S7. Machine learning integrated scores for PTEN-cancer/ASD classification. This set of scores corroborates the ML-based findings discussed in the main text and reveals high-IS mutations that merit further structural and functional analysis.

AUTHOR INFORMATION

Corresponding Authors

Guang Hu - MOE Key Laboratory of Geriatric Diseases and Immunology, Suzhou Key Laboratory of Pathogen Bioscience and Anti-infective Medicine, Department of Bioinformatics and Computational Biology, School of Life Sciences, Suzhou Medical College of Soochow University, Suzhou 215123, China, orcid: 0000000287541541; Email: huguang@suda.edu.cn

Fei Xiao - Department of Bioinformatics and Computational Biology, School of Life Sciences, Suzhou Medical College of Soochow University, Suzhou 215123, China, orcid: 0000000152622475; Email: xiaofei@suda.edu.cn

Authors

Miao Yang - Department of Bioinformatics and Computational Biology, School of Life Sciences, Suzhou Medical College of Soochow University, Suzhou, 215213, China

Jingran Wang - Department of Bioinformatics and Computational Biology, School of Life Sciences, Suzhou Medical College of Soochow University, Suzhou, 215123, China

Ziyun Zhou- Department of Bioinformatics and Computational Biology, School of Life Sciences, Suzhou Medical College of Soochow University, Suzhou, 215123, China

Wentian Li - Department of Bioinformatics and Computational Biology, School of Life Sciences, Suzhou Medical College of Soochow University, Suzhou, 215123, China

Gennady Verkhivker - Department of Biomedical and Pharmaceutical Sciences, Chapman University School of Pharmacy, Irvine 92618, California, United States

Author Contributions

[#]M. Y. and J. W. contributed equally. The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (32271292), the Project of MOE Key Laboratory of Geriatric Diseases and Immunology (JYN202404), and a Project Funded by the Priority Academic Program Development (PAPD) of Jiangsu Higher Education Institutions. This research was also funded by the National Institutes of Health under Award 1R01AI181600-01 and Subaward 6069-SC24-11 to Gennady Verkhivker.

REFERENCES

(1) Nussinov, R.; Yavuz, B. R.; Arici, M. K.; Demirel, H. C.; Zhang, M.; Liu, Y.; Tsai, C.-J.; Jang, H.; Tuncbag, N. Neurodevelopmental disorders, like cancer, are connected to impaired chromatin remodelers, PI3K/mTOR, and PAK1-regulated MAPK. *Biophysical reviews* **2023**, *15* (2), 163-181. DOI: 10.1007/s12551-023-01054-9.

(2) Crespi, B. Autism and cancer risk. *Autism research : official journal of the International Society for Autism Research* **2011**, *4* (4), 302-310. DOI: 10.1002/aur.208.

(3) Moreno-De-Luca, A.; Myers, S. M.; Challman, T. D.; Moreno-De-Luca, D.; Evans, D. W.; Ledbetter, D. H. Developmental brain dysfunction: revival and expansion of old concepts based on new genetic evidence. *Lancet Neurology* **2013**, *12* (4), 406-414. DOI: 10.1016/s1474-4422(13)70011-5.

(4) Yavuz, B. R.; Arici, M. K.; Demirel, H. C.; Tsai, C.-J.; Jang, H.; Nussinov, R.; Tuncbag, N. Neurodevelopmental disorders and cancer networks share pathways, but differ in mechanisms, signaling strength, and outcome. *Npj Genomic Medicine* **2023**, *8* (1). DOI: 10.1038/s41525-023-00377-6.

(5) Nussinov, R.; Tsai, C.-J.; Jang, H. How can same-gene mutations promote both cancer and developmental disorders? *Science Advances* **2022**, *8* (2). DOI: 10.1126/sciadv.abm2059.

(6) Qi, H.; Dong, C.; Chung, W. K.; Wang, K.; Shen, Y. Deep Genetic Connection Between Cancer and Developmental Disorders. *Human Mutation* **2016**, *37* (10), 1042-1050. DOI: 10.1002/humu.23040.

(7) Glaviano, A.; Foo, A. S. C.; Lam, H. Y.; Yap, K. C. H.; Jacot, W.; Jones, R. H.; Eng, H.; Nair, M. G.; Makvandi, P.; Geoerger, B.; et al. PI3K/AKT/mTOR signaling transduction pathway and targeted therapies in cancer. *Molecular Cancer* **2023**, *22* (1). DOI: 10.1186/s12943-023-01827-6.

(8) Nussinov, R.; Yavuz, B. R.; Demirel, H. C.; Arici, M. K.; Jang, H.; Tuncbag, N. Review: Cancer and neurodevelopmental disorders: multi-scale reasoning and computational guide. *Frontiers in Cell and Developmental Biology* **2024**, *12*. DOI: 10.3389/fcell.2024.1376639.

(9) Jang, H.; Smith, I. N.; Eng, C.; Nussinov, R. The mechanism of full activation of tumor suppressor PTEN at the phosphoinositide-enriched membrane. *Iscience* **2021**, *24* (5). DOI: 10.1016/j.isci.2021.102438.

(10) Smith, I. N.; Briggs, J. M. Structural mutation analysis of PTEN and its genotype-phenotype correlations in endometriosis and cancer. *Proteins-Structure Function and Bioinformatics* **2016**, *84* (11), 1625-1643. DOI: 10.1002/prot.25105.

(11) Rademacher, S.; Eickholt, B. J. PTEN in Autism and Neurodevelopmental Disorders. *Cold Spring Harbor Perspectives in Medicine* **2019**, *9* (11). DOI: 10.1101/cshperspect.a036780.

(12) Yehia, L.; Keel, E.; Eng, C. The Clinical Spectrum of PTEN Mutations. *Annual Review of Medicine, Vol* **2020**, *71*, 103-116. DOI: 10.1146/annurev-med-052218-125823.

(13) Smith, I. N.; Thacker, S.; Jaini, R.; Eng, C. Dynamics and structural stability effects of germline PTEN mutations associated with cancer versus autism phenotypes. *Journal of Biomolecular Structure & Dynamics* **2019**, *37* (7), 1766-1782. DOI: 10.1080/07391102.2018.1465854.

(14) Smith, I. N.; Thacker, S.; Seyfi, M.; Cheng, F.; Eng, C. Conformational Dynamics and Allosteric Regulation Landscapes of Germline PTEN Mutations Associated with Autism Compared to Those Associated with Cancer. *American Journal of Human Genetics* **2019**, *104* (5), 861-878. DOI: 10.1016/j.ajhg.2019.03.009.

(15) Shan, L.; Yu, J.; He, Z.; Chen, S.; Liu, M.; Ding, H.; Xu, L.; Zhao, J.; Yang, A.; Jiang, H. Defining relative mutational difficulty to understand cancer formation. *Cell Discovery* **2020**, *6* (1). DOI: 10.1038/s41421-020-0177-8.

(16) Ng, P. K.-S.; Li, J.; Jeong, K. J.; Shao, S.; Chen, H.; Tsang, Y. H.; Sengupta, S.; Wang, Z.; Bhavana, V. H.; Tran, R.; et al. Systematic Functional Annotation of Somatic Mutations in Cancer. *Cancer Cell* **2018**, *33* (3), 450-+. DOI: 10.1016/j.ccell.2018.01.021.

(17) Portelli, S.; Barr, L.; de Sa, A. G. C.; Pires, D. E. V.; Ascher, D. B. Distinguishing between PTEN clinical phenotypes through mutation analysis. *Computational and Structural Biotechnology Journal* **2021**, *19*, 3097-3109. DOI: 10.1016/j.csbj.2021.05.028.

(18) Mighell, T. L.; Evans-Dutson, S.; O'Roak, B. J. A Saturation Mutagenesis Approach to Understanding PTEN Lipid Phosphatase Activity and Genotype-Phenotype Relationships. *American Journal of Human Genetics* **2018**, *102* (5), 943-955. DOI: 10.1016/j.ajhg.2018.03.018.

(19) Mighell, T. L.; Thacker, S.; Fombonne, E.; Eng, C.; O'Roak, B. J. An Integrated Deep-Mutational-Scanning Approach Provides Clinical Insights on PTEN Genotype-Phenotype Relationships. *American Journal of Human Genetics* **2020**, *106* (6), 818-829. DOI: 10.1016/j.ajhg.2020.04.014.

(20) Post, K. L.; Belmadani, M.; Ganguly, P.; Meili, F.; Dingwall, R.; McDiarmid, T. A.; Meyers, W. M.; Herrington, C.; Young, B. P.; Callaghan, D. B.; et al. Multi-model functionalization of disease-associated PTEN missense mutations identifies multiple molecular mechanisms underlying protein dysfunction. *Nature Communications* **2020**, *11* (1). DOI: 10.1038/s41467-020-15943-0.

(21) Smith, I. N.; Dawson, J. E.; Eng, C. Comparative Protein Structural Network Analysis Reveals C-Terminal Tail Phosphorylation Structural Communication Fingerprint in PTEN-Associated Mutations in Autism and Cancer. *Journal of Physical Chemistry B* **2023**, *127* (3), 634-647. DOI: 10.1021/acs.jpcb.2c06776.

(22) Smith, I. N.; Dawson, J. E.; Krieger, J.; Thacker, S.; Bahar, I.; Eng, C. Structural and Dynamic Effects of PTEN C-Terminal Tail Phosphorylation. *Journal of Chemical Information and Modeling* **2022**, *62* (17), 4175-4190. DOI: 10.1021/acs.jcim.2c00441.

(23) Murthy, A. S. N.; Suresh, R. V.; Nallur, B. R. Comprehensive in silico mutational-sensitivity analysis of PTEN establishes signature regions implicated in pathogenesis of Autism Spectrum Disorders. *Genomics* **2021**, *113* (1), 999-1017. DOI: 10.1016/j.ygeno.2020.10.035.

(24) Sinha, S.; Li, J.; Tam, B.; Wang, S. M. Classification of PTEN missense VUS through exascale simulations. *Briefings in Bioinformatics* **2023**, *24* (6). DOI: 10.1093/bib/bbad361.

(25) Ponzoni, L.; Bahar, I. Structural dynamics is a determinant of the functional significance of missense variants. *Proceedings of the National Academy of Sciences of the United States of America* **2018**, *115* (16), 4164-4169. DOI: 10.1073/pnas.1715896115.

(26) Xiao, F.; Song, X.; Tian, P.; Gan, M.; Verkhivker, G. M.; Hu, G. Comparative Dynamics and Functional Mechanisms of the CYP17A1 Tunnels Regulated by Ligand Binding. *Journal of Chemical Information and Modeling* **2020**, *60* (7), 3632-3647. DOI: 10.1021/acs.jcim.0c00447.

(27) Di Paola, L.; Hadi-Alijanvand, H.; Song, X.; Hu, G.; Giuliani, A. The Discovery of a Putative Allosteric Site in the SARS-CoV-2 Spike Protein Using an Integrated Structural/Dynamic Approach. *Journal of Proteome Research* **2020**, *19* (11), 4576-4586. DOI: 10.1021/acs.jproteome.0c00273.

(28) Xiao, F.; Zhou, Z.; Song, X.; Gan, M.; Long, J.; Verkhivker, G.; Hu, G. Dissecting mutational allosteric effects in alkaline phosphatases associated with different Hypophosphatasia phenotypes: An integrative computational investigation. *Plos Computational Biology* **2022**, *18* (3). DOI: 10.1371/journal.pcbi.1010009.

(29) Liang, Z.; Verkhivker, G. M.; Hu, G. Integration of network models and evolutionary analysis into high-throughput modeling of protein dynamics and allosteric regulation: theory, tools and applications. *Briefings in Bioinformatics* **2020**, *21* (3), 815-835. DOI: 10.1093/bib/bbz029.

(30) Landrum, M. J.; Chitipiralla, S.; Brown, G. R.; Chen, C.; Gu, B.; Hart, J.; Hoffman, D.; Jang, W.; Kaur, K.; Liu, C.; et al. ClinVar: improvements to accessing data. *Nucleic Acids Research* **2020**, *48* (D1), D835-D844. DOI: 10.1093/nar/gkz972.

(31) Chen, S.; Francioli, L. C.; Goodrich, J. K.; Collins, R. L.; Kanai, M.; Wang, Q.; Alfoldi, J.; Watts, N. A.; Vittal, C.; Gauthier, L. D.; et al. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **2024**, *625* (7993). DOI: 10.1038/s41586-023-06045-0.

(32) Sondka, Z.; Dhir, N. B.; Carvalho-Silva, D.; Jupe, S.; McLaren, K.; Starkey, M.; Ward, S.; Wilding, J.; Ahmed, M.; Argasinska, J.; et al. COSMIC: a curated database of somatic variants and clinical data for cancer. *Nucleic Acids Research* **2023**. DOI: 10.1093/nar/gkad986.

(33) Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Machine Learning* **2002**, *46* (1-3), 389-422. DOI: 10.1023/a:1012487302797.

(34) Hajihosseinlou, M.; Maghsoudi, A.; Ghezelbash, R. A Novel Scheme for Mapping of MVT-Type Pb-Zn Prospectivity: LightGBM, a Highly Efficient Gradient Boosting Decision Tree Machine Learning Algorithm. *Natural Resources Research* **2023**, *32* (6), 2417-2438. DOI: 10.1007/s11053-023-10249-6.

(35) Natekin, A.; Knoll, A. Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics* **2013**, *7*. DOI: 10.3389/fnbot.2013.00021.

(36) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Arxiv* **2016**. DOI: arXiv:1603.02754.

(37) Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Veronika Dorogush, A.; Gulin, A. CatBoost: unbiased boosting with categorical features. *Arxiv* **2019**. DOI: arXiv:1706.09516.

(38) Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Arxiv* **2017**. DOI: arXiv:1705.07874.

(39) Brown, D. K.; Penkler, D. L.; Amamuddy, O. S.; Ross, C.; Atilgan, A. R.; Atilgan, C.; Bishop, O. T. MD-TASK: a software suite for analyzing molecular dynamics trajectories. *Bioinformatics* **2017**, *33* (17), 2768-2771. DOI: 10.1093/bioinformatics/btx349.

(40) Floyd, R. W. ALGORITHM-97 - SHORTEST PATH. *Communications of the Acm* **1962**, *5* (6), 345-345. DOI: 10.1145/367766.368168.

(41) Schymkowitz, J.; Borg, J.; Stricher, F.; Nys, R.; Rousseau, F.; Serrano, L. The FoldX web server: an online force field. *Nucleic Acids Research* **2005**, *33*, W382-W388. DOI: 10.1093/nar/gki387.

(42) Tilot, A. K.; Frazier, T. W., II; Eng, C. Balancing Proliferation and Connectivity in PTEN-associated Autism Spectrum Disorder. *Neurotherapeutics* **2015**, *12* (3), 609-619. DOI: 10.1007/s13311-015-0356-8.

(43) Chow, J. T.-S.; Salmena, L. Recent advances in PTEN signalling axes in cancer. *Faculty reviews* **2020**, *9*, 31-31. DOI: 10.12703/r/9-31.

(44) Rodriguez-Escudero, I.; Oliver, M. D.; Andres-Pons, A.; Molina, M.; Cid, V. J.; Pulido, R. A comprehensive functional analysis of PTEN mutations: implications in tumor- and autism-related syndromes. *Human Molecular Genetics* **2011**, *20* (21), 4132-4142. DOI: 10.1093/hmg/ddr337.

(45) Kaymakcalan, H.; Kaya, I.; Binici, N. C.; Nikerel, E.; Ozbaran, B.; Aksoy, M. G.; Erbilgin, S.; Ozyurt, G.; Jahan, N.; Celik, D.; et al. Prevalence and clinical/molecular characteristics of PTEN mutations in Turkish children with autism spectrum disorders and macrocephaly. *Molecular Genetics & Genomic Medicine* **2021**, *9* (8). DOI: 10.1002/mgg3.1739.

(46) Nussinov, R.; Liu, Y.; Zhang, W.; Jang, H. Cell phenotypes can be predicted from propensities of protein conformations. *Current Opinion in Structural Biology* **2023**, *83*. DOI: 10.1016/j.sbi.2023.102722.

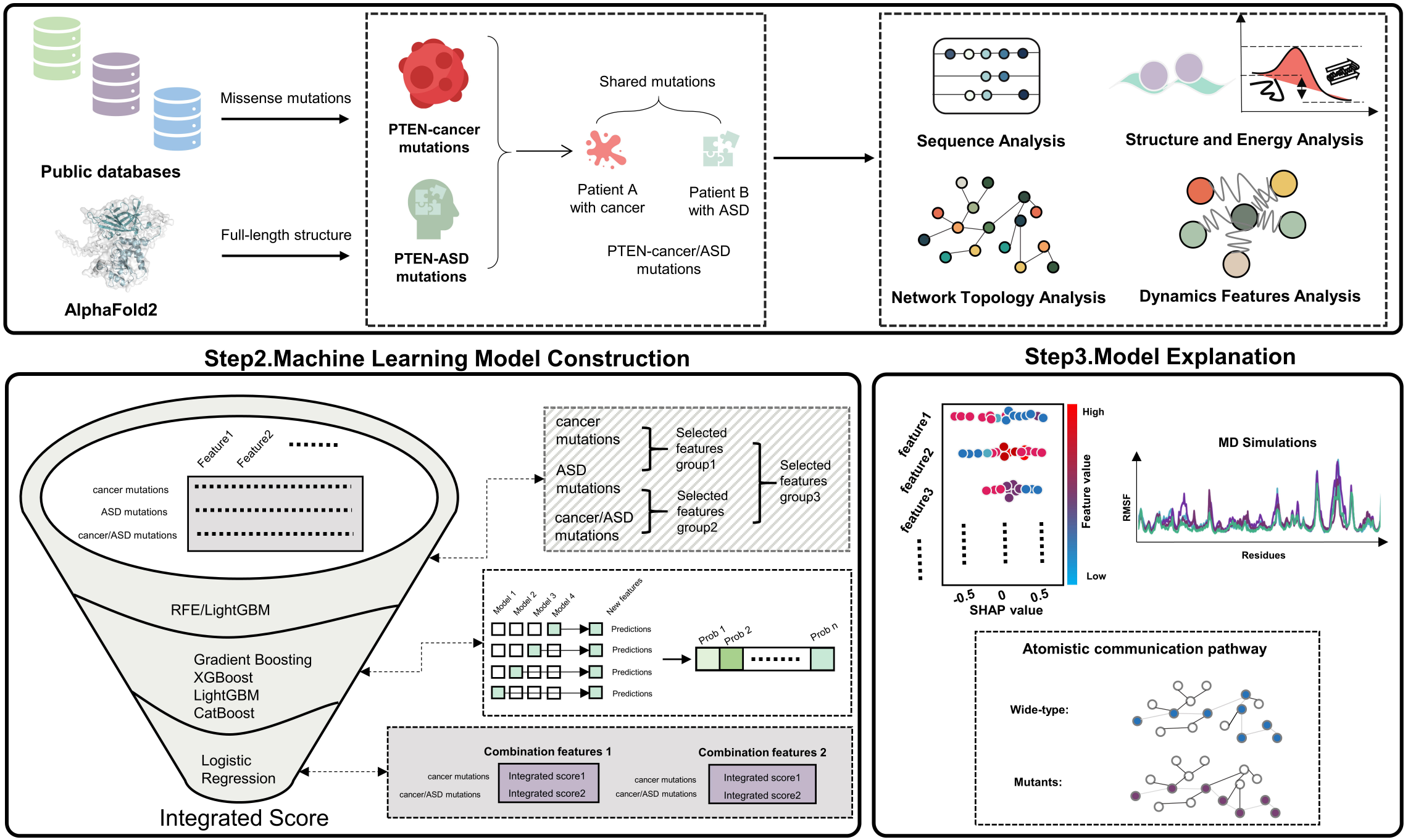
(47) Banerjee, A.; Saha, S.; Tvedt, N. C.; Yang, L.-W.; Bahar, I. Mutually beneficial confluence of structure-based modeling of protein dynamics and machine learning methods. *Current Opinion in Structural Biology* **2023**, *78*. DOI: 10.1016/j.sbi.2022.102517.

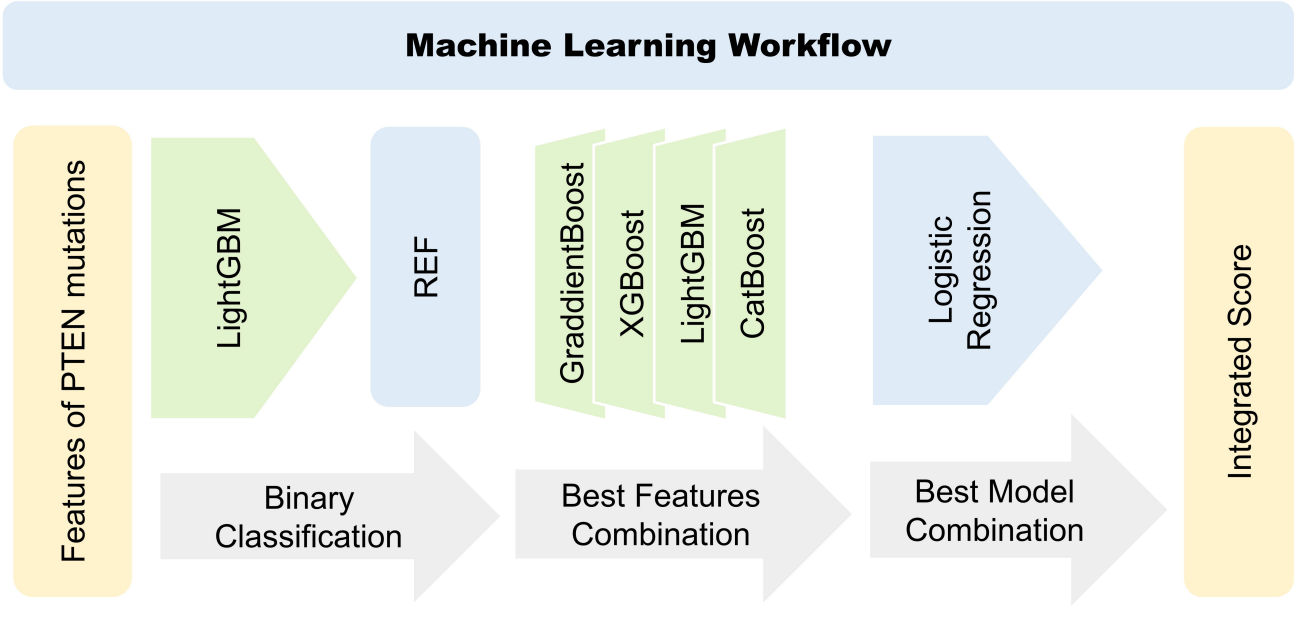
(48) Jang, H.; Chen, J.; Iakoucheva, L. M.; Nussinov, R. Cancer and Autism: How PTEN Mutations Degrade Function at the Membrane and Isoform Expression in the Human Brain. *Journal of Molecular Biology* **2023**, *435* (24). DOI: 10.1016/j.jmb.2023.168354.

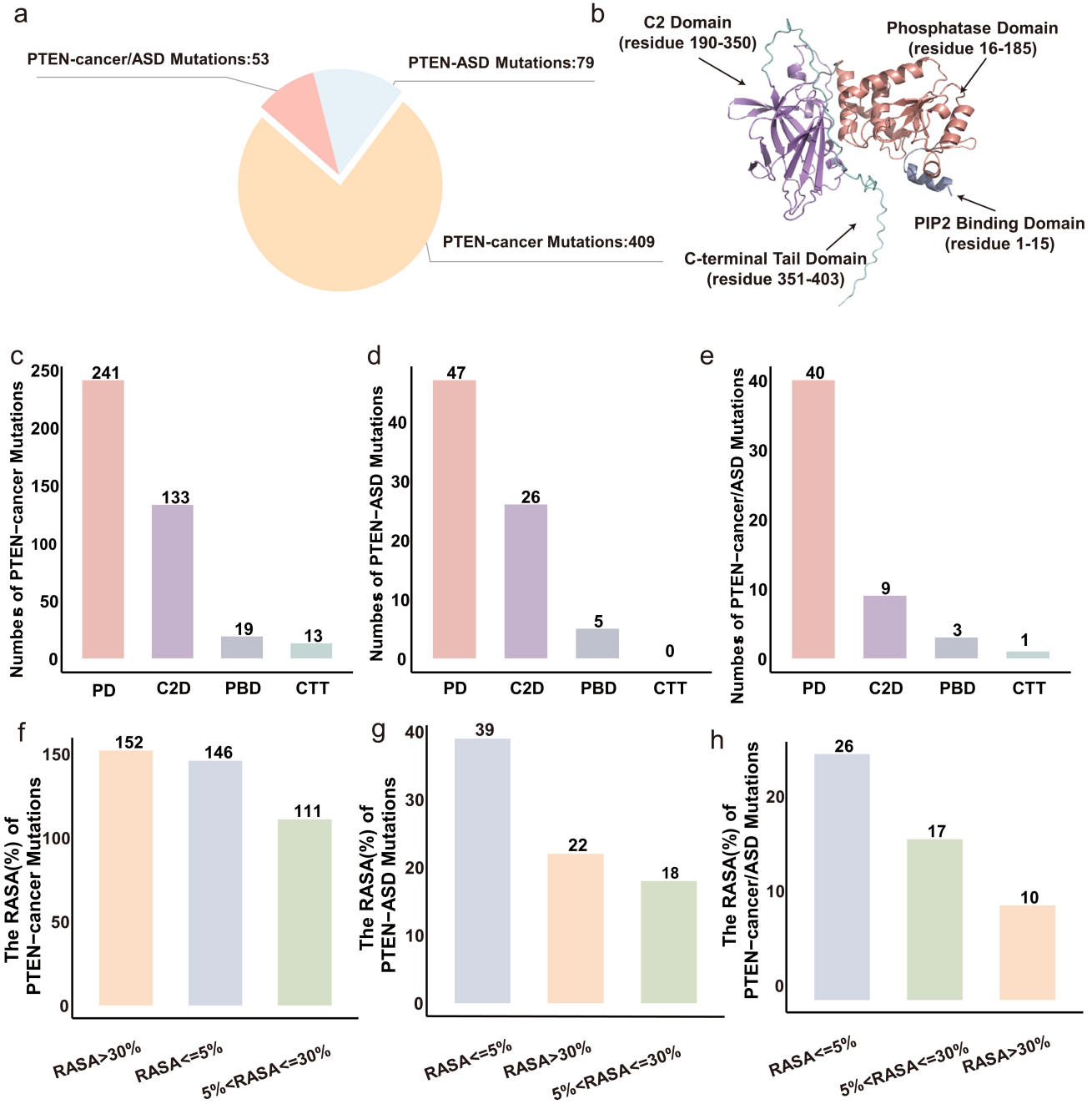
(49) Liang, Z.; Zhu, Y.; Long, J.; Ye, F.; Hu, G. Both intra and inter-domain interactions define the intrinsic dynamics and allosteric mechanism in DNMT1s. *Computational and Structural Biotechnology Journal* **2020**, *18*, 749-764. DOI: 10.1016/j.csbj.2020.03.016.

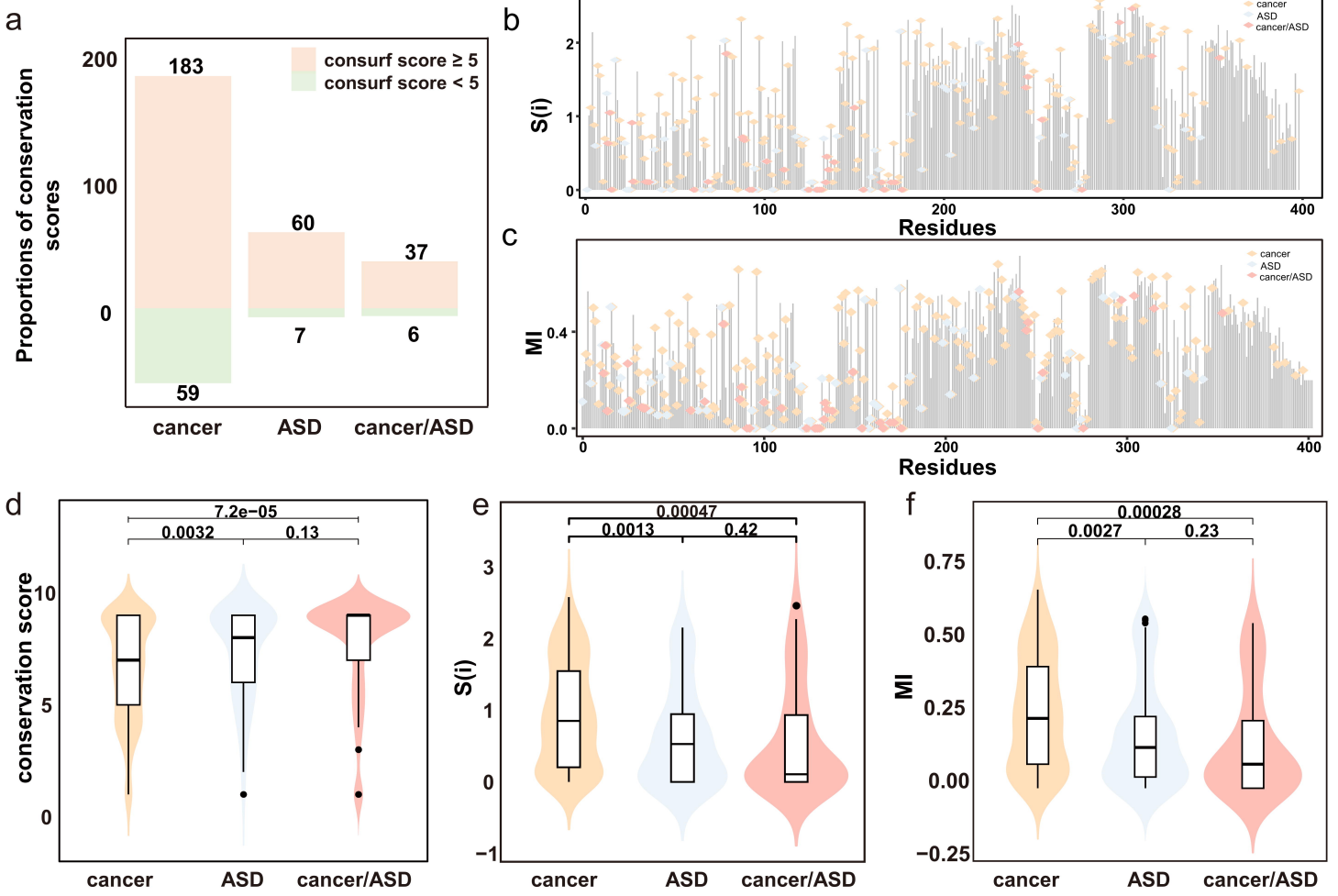
(50) Degn, K.; Beltrame, L.; Hede, F. D.; Sora, V.; Nicolaci, V.; Vabistsevits, M.; Schmiegelow, K.; Wadt, K.; Tiberti, M.; Lambrughi, M.; Papaleo, E. Cancer-related Mutations with Local or Long-range Effects on an Allosteric Loop of p53. *Journal of Molecular Biology* **2022**, *434* (17). DOI: 10.1016/j.jmb.2022.167663.

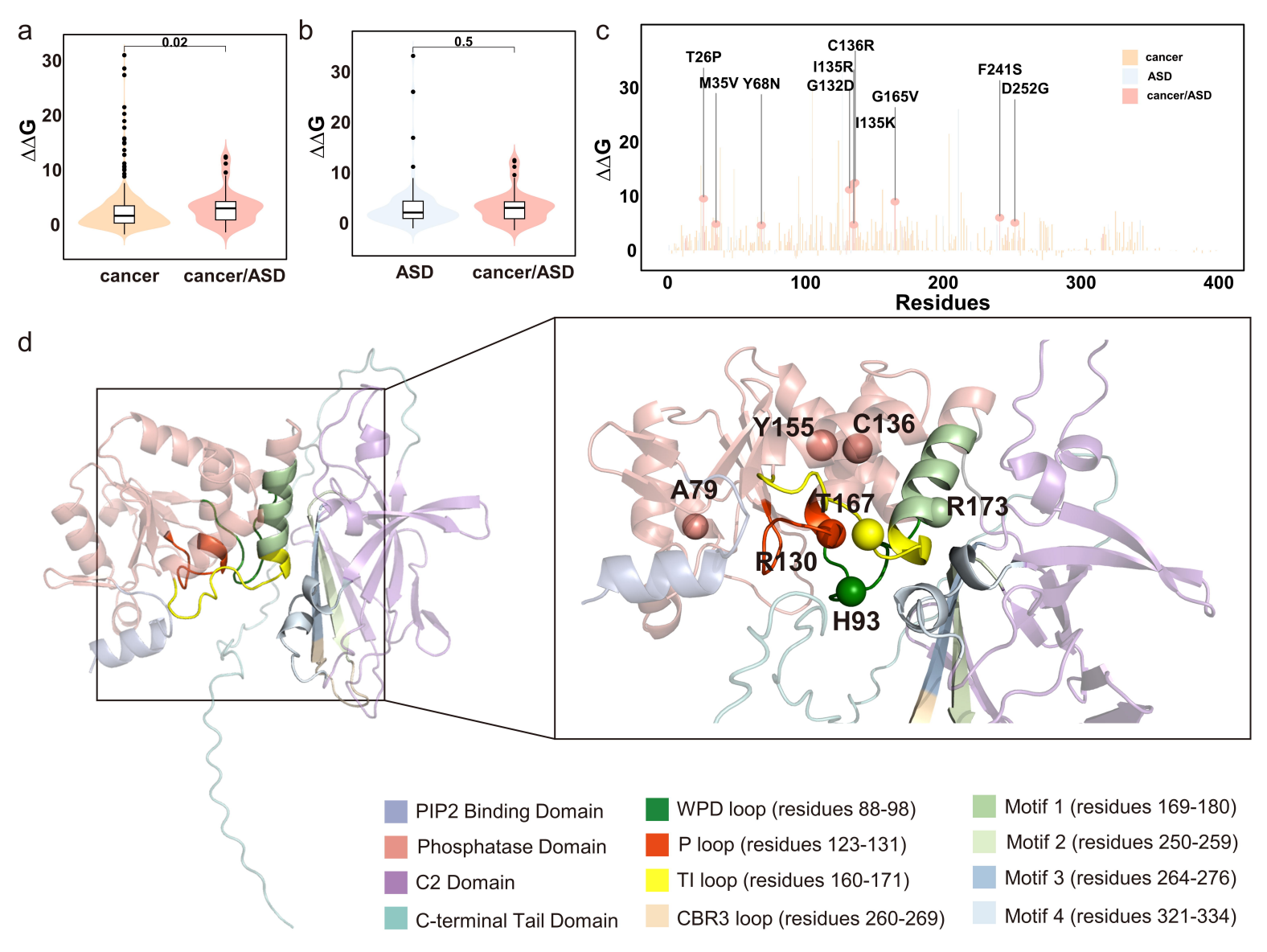
(51) Zhang, H.; Gur, M.; Bahar, I. Global hinge sites of proteins as target sites for drug binding. *Proceedings of the National Academy of Sciences of the United States of America* **2024**, *121* (49), e2414333121-e2414333121. DOI: 10.1073/pnas.2414333121.

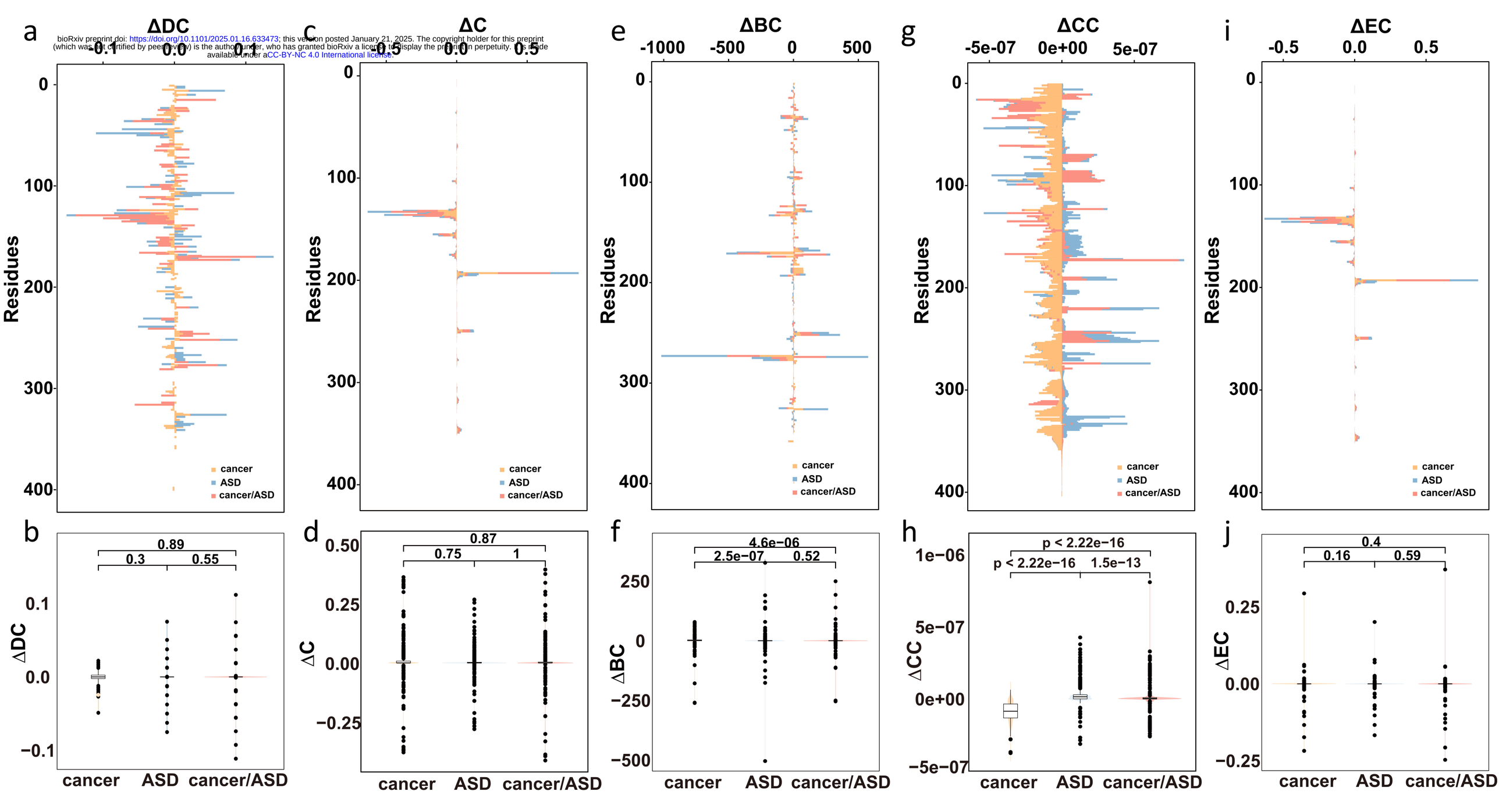

Step1.Data collection and Feature Generation
Public databases
Missense mutations
AlphaFold2
Full-length structure
PTEN-cancer mutations
PTEN-ASD mutations
Shared mutations
Patient A with cancer
Patient B with ASD
PTEN-cancer/ASD mutations
Sequence Analysis
Structure and Energy Analysis
Network Topology Analysis
Dynamics Features Analysis
Step2.Machine Learning Model Construction
Feature1
Feature2
cancer mutations
ASD mutations
cancer/ASD mutations
RFE/LightGBM
Gradient Boosting
XGBoost
LightGBM
CatBoost
Logistic Regression
Integrated Score
Selected features group1
Selected features group2
Selected features group3
Model 1
Model 2
Model 3
Model 4
New features
Predictions
Prob 1
Prob 2
Prob n
Combination features 1
Combination features 2
Integrated score1
Integrated score2
Step3.Model Explanation
feature1
feature2
feature3
High
Feature value
Low
-0.5
0
0.5
SHAP value
MD Simulations
RMSF
Residues
Atomistic communication pathway
Wide-type:
Mutants:

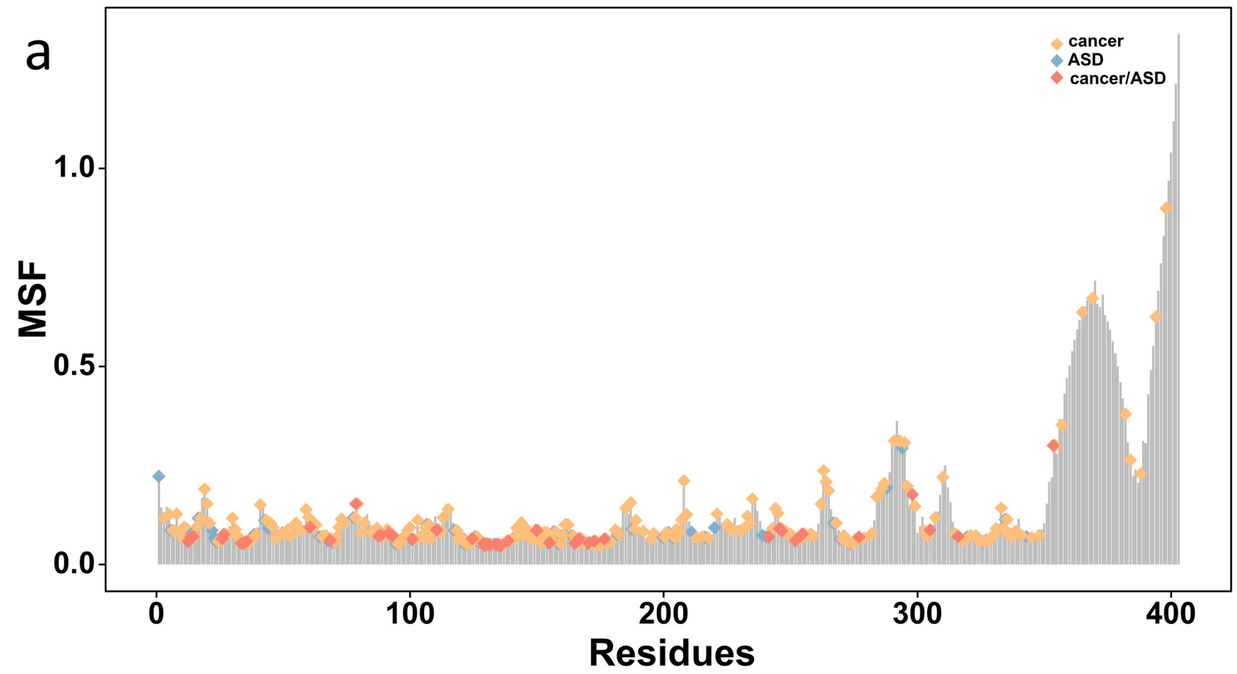

Machine Learning Workflow
Features of PTEN mutations
LightGBM
REF
GraddientBoost
XGBoost
LightGBM
CatBoost
Logistic Regression
Integrated Score
Binary Classification
Best Features Combination
Best Model Combination

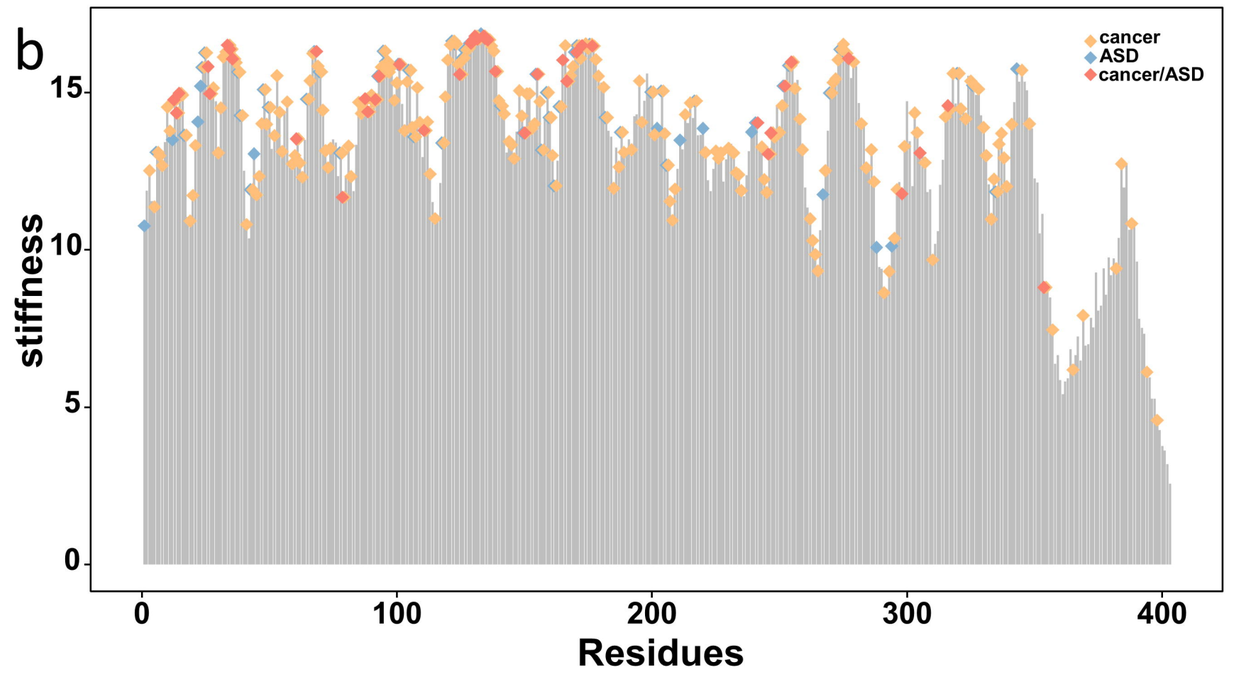

a
PTEN-cancer/ASD Mutations:53
PTEN-ASD Mutations:79
PTEN-cancer Mutations:409
b
C2 Domain
(residue 190-350)
Phosphatase Domain
(residue 16-185)
PIP2 Binding Domain
(residue 1-15)
C-terminal Tail Domain
(residue 351-403)
c
Numbes of PTEN−cancer Mutations
241
133
19
13
PD
C2D
PBD
CTT
d
Numbes of PTEN−ASD Mutations
47
26
5
0
PD
C2D
PBD
CTT
e
Numbes of PTEN−cancer/ASD Mutations
40
9
3
1
PD
C2D
PBD
CTT
f
The RASA(%) of
PTEN−cancer Mutations
152
146
111
RASA>30%
RASA<=5%
5%<RASA<=30%
g
The RASA(%) of
PTEN−ASD Mutations
39
22
18
RASA<=5%
RASA>30%
5%<RASA<=30%
h
The RASA(%) of
PTEN−cancer/ASD Mutations
26
17
10
RASA<=5%
5%<RASA<=30%
RASA>30%

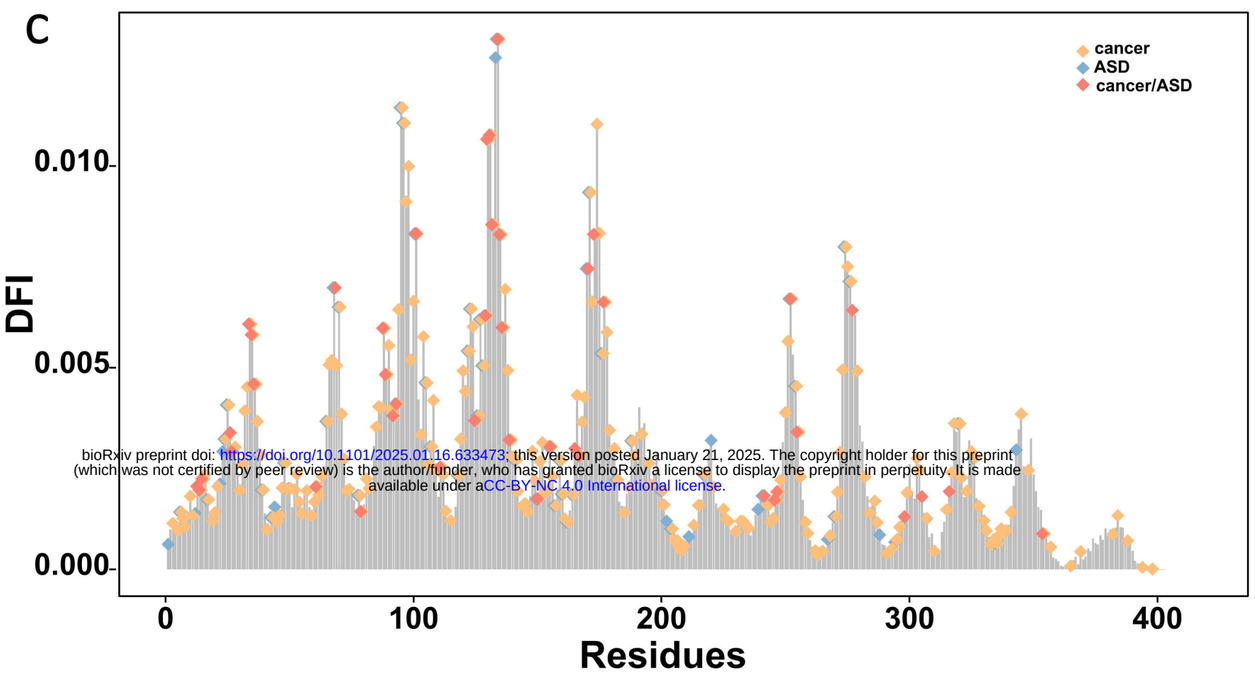

a
Proportions of conservation scores
200
100
0
consurf score ≥ 5
consurf score < 5
183
59
60
7
37
6
cancer
ASD
cancer/ASD
b
S(i)
Residues
cancer
ASD
cancer/ASD
c
MI
Residues
d
conservation score
7.2e−05
0.0032
0.13
e
S(i)
0.00047
0.0013
0.42
f
MI
0.00028
0.0027
0.23

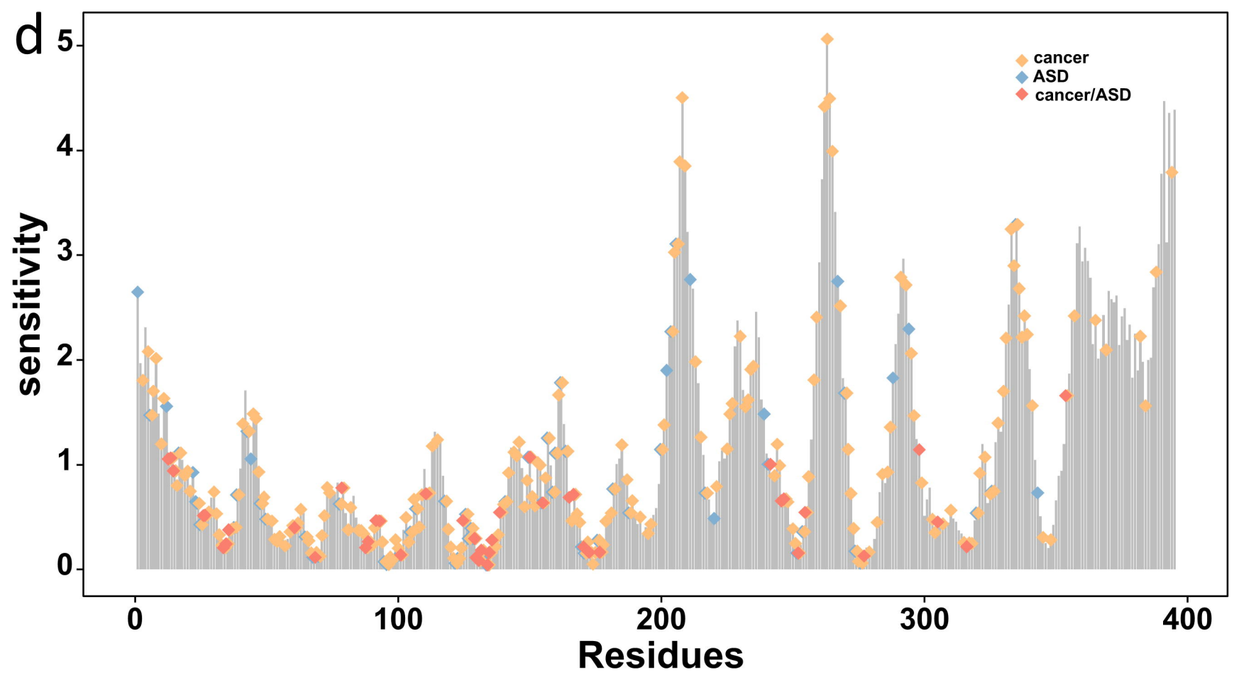

a
0.02
ΔΔG
cancer
cancer/ASD
b
0.5
ASD
cancer/ASD
c
T26P
M35V Y68N
C136R
I135R
G132D
G165V
I135K
F241S
D252G
cancer
ASD
cancer/ASD
Residues
d
Y155
C136
A79
T167
R173
R130
H93
PIP2 Binding Domain
Phosphatase Domain
C2 Domain
C-terminal Tail Domain
WPD loop (residues 88-98)
P loop (residues 123-131)
TI loop (residues 160-171)
CBR3 loop (residues 260-269)
Motif 1 (residues 169-180)
Motif 2 (residues 250-259)
Motif 3 (residues 264-276)
Motif 4 (residues 321-334)

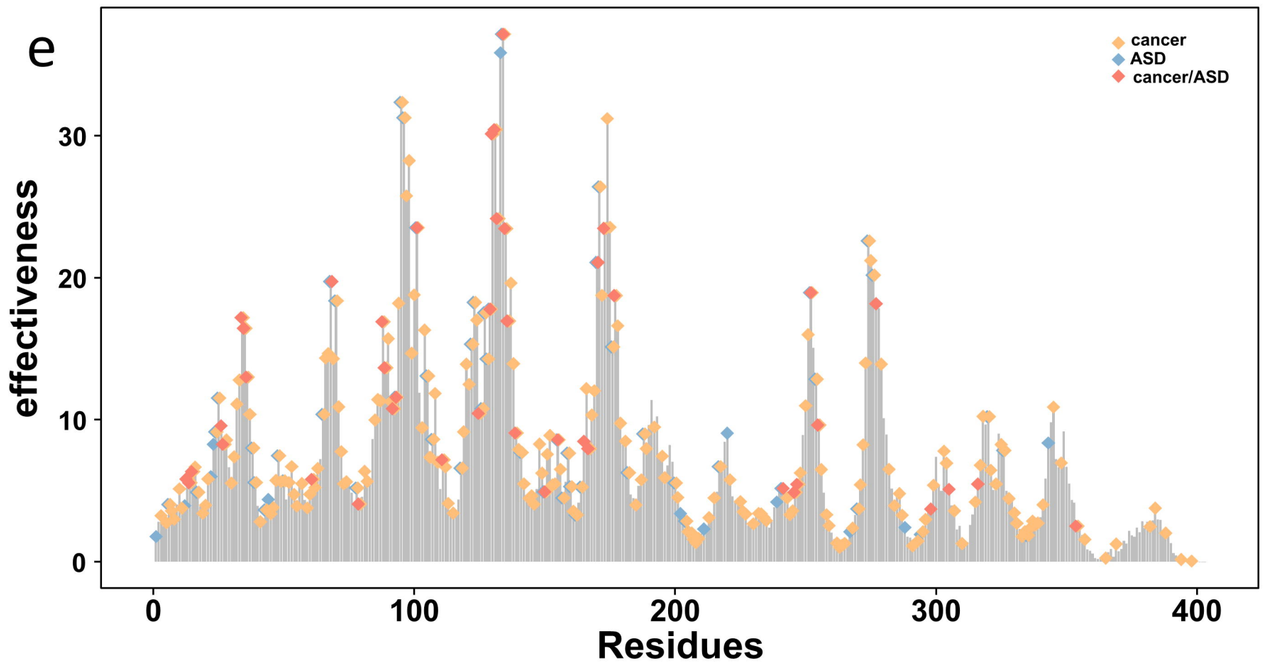


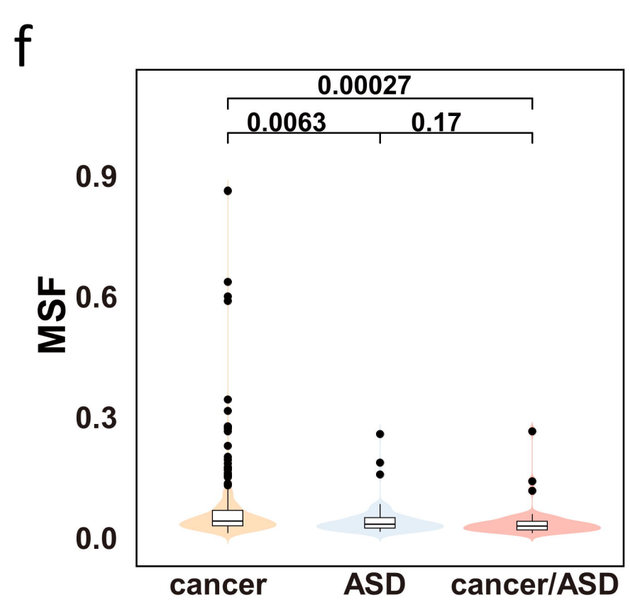


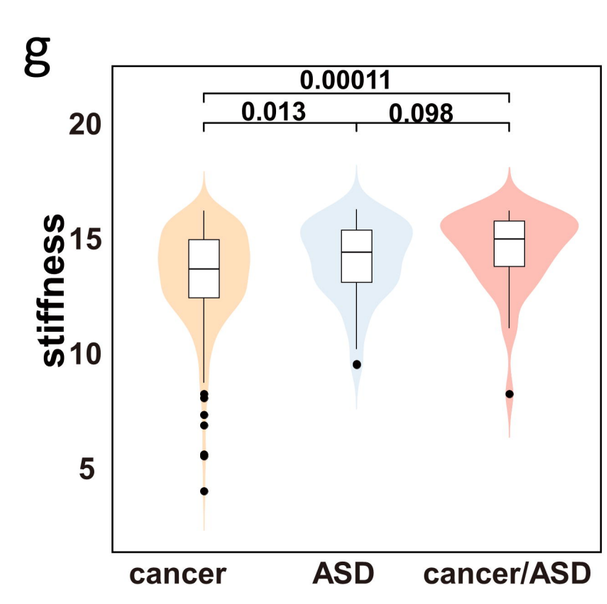


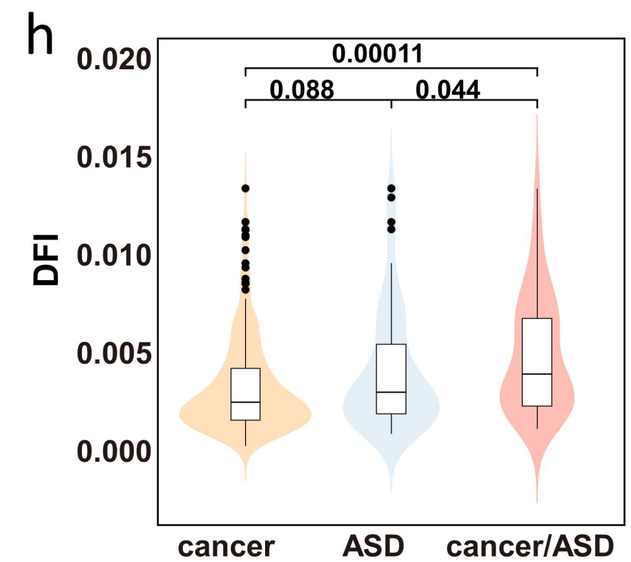


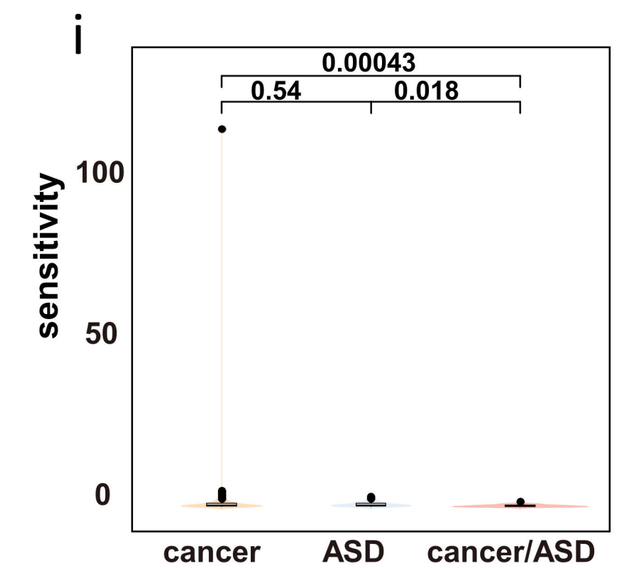


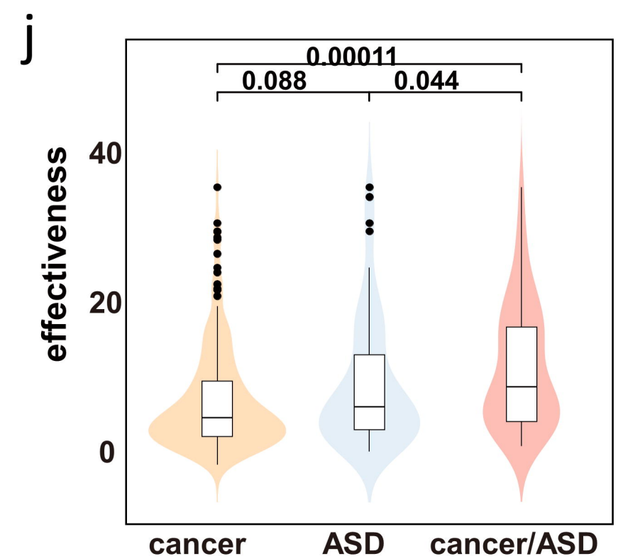


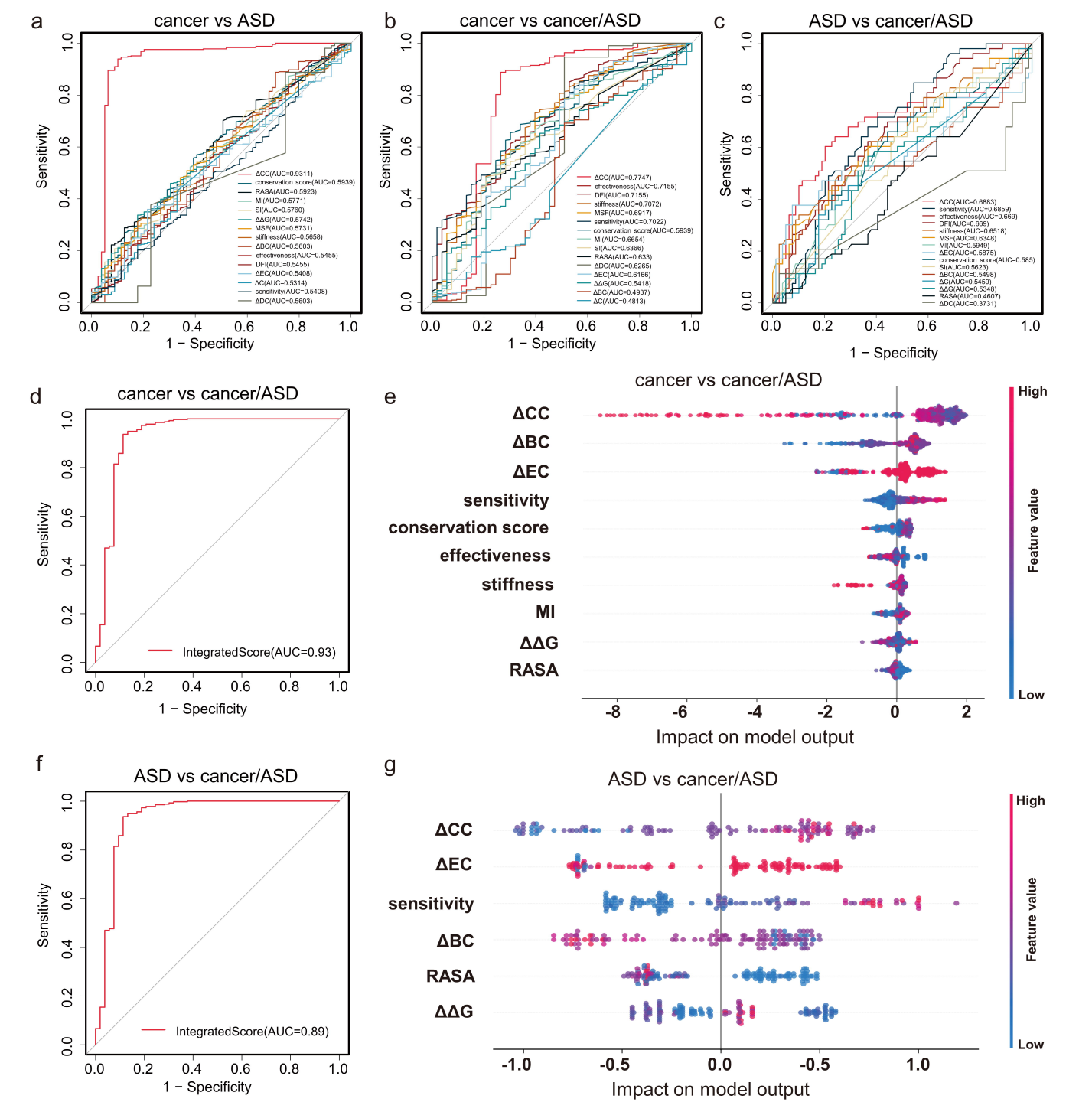


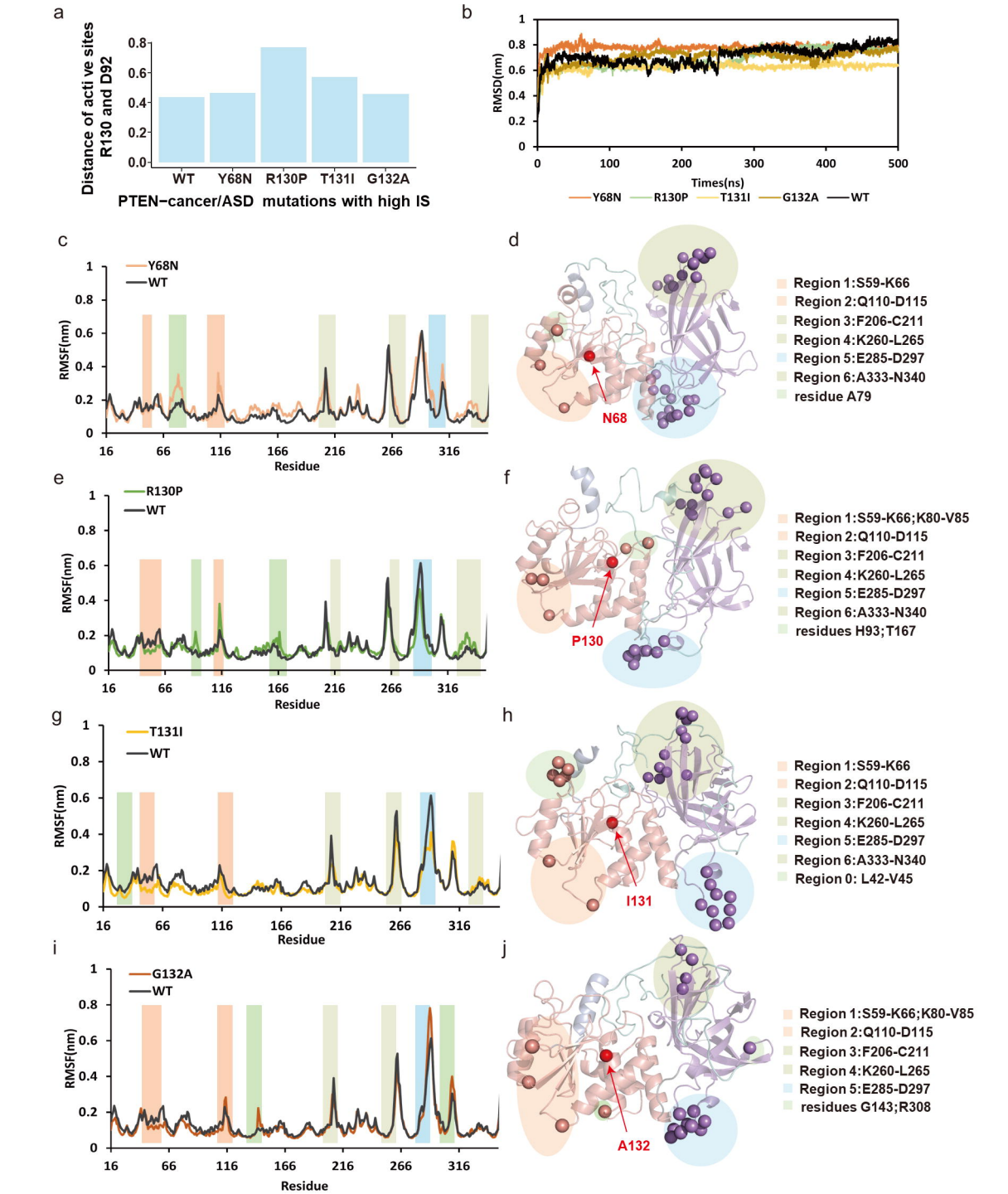


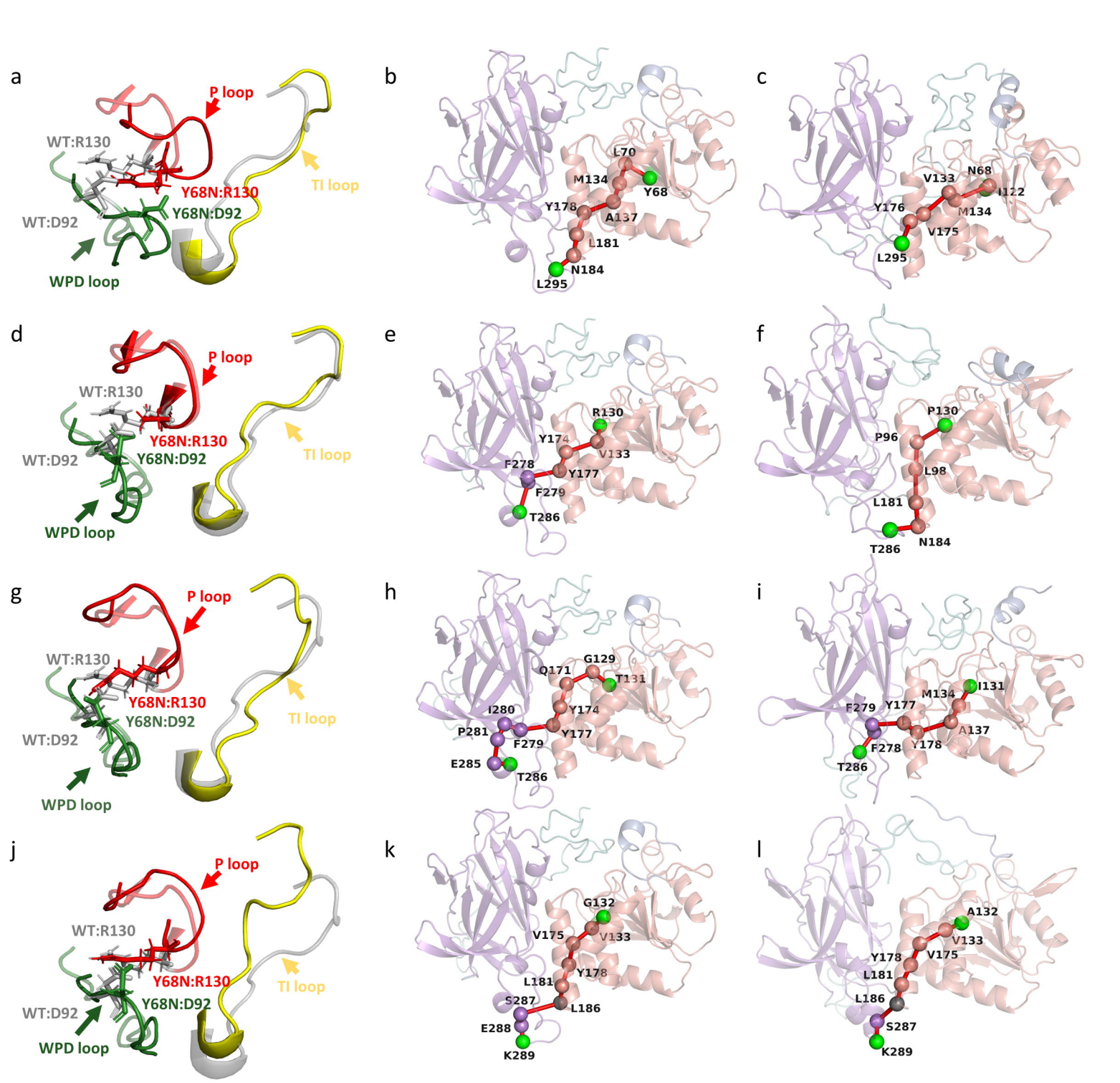







a
cancer vs ASD
Sensitivity
1 − Specificity
ΔCC(AUC=0.9311)
conservation score(AUC=0.5939)
RASA(AUC=0.5923)
MI(AUC=0.5771)
SI(AUC=0.5760)
ΔΔG(AUC=0.5742)
MSF(AUC=0.5731)
stiffness(AUC=0.5658)
ΔBC(AUC=0.5603)
effectiveness(AUC=0.5455)
DFI(AUC=0.5455)
ΔEC(AUC=0.5408)
ΔC(AUC=0.5314)
sensitivity(AUC=0.5408)
ΔDC(AUC=0.5603)
b
cancer vs cancer/ASD
Sensitivity
1 − Specificity
ΔCC(AUC=0.7747)
effectiveness(AUC=0.7155)
DFI(AUC=0.7155)
stiffness(AUC=0.7072)
MSF(AUC=0.6917)
sensitivity(AUC=0.7022)
conservation score(AUC=0.5939)
MI(AUC=0.6654)
SI(AUC=0.6366)
RASA(AUC=0.633)
ΔDC(AUC=0.6285)
ΔEC(AUC=0.6166)
ΔΔG(AUC=0.5418)
ΔBC(AUC=0.4937)
ΔC(AUC=0.4813)
c
ASD vs cancer/ASD
Sensitivity
1 − Specificity
ΔCC(AUC=0.6883)
sensitivity(AUC=0.6859)
effectiveness(AUC=0.669)
DFI(AUC=0.669)
stiffness(AUC=0.6518)
MSF(AUC=0.6348)
MI(AUC=0.5949)
ΔEC(AUC=0.5875)
conservation score(AUC=0.585)
SI(AUC=0.5623)
ΔBC(AUC=0.5498)
ΔC(AUC=0.5459)
ΔΔG(AUC=0.5349)
RASA(AUC=0.4607)
ΔDC(AUC=0.3731)
d
cancer vs cancer/ASD
Sensitivity
1 − Specificity
IntegratedScore(AUC=0.93)
e
cancer vs cancer/ASD
ΔCC
ΔBC
ΔEC
sensitivity
conservation score
effectiveness
stiffness
MI
ΔΔG
RASA
-8
-6
-4
-2
0
2
Impact on model output
High
Feature value
Low
f
ASD vs cancer/ASD
Sensitivity
1 − Specificity
IntegratedScore(AUC=0.89)
g
ASD vs cancer/ASD
ΔCC
ΔEC
sensitivity
ΔBC
RASA
ΔΔG
-1.0
-0.5
0.0
-0.5
1.0
Impact on model output
High
Feature value
Low


a
Distance of acti ve sites R130 and D92
0.8
0.6
0.4
0.2
0.0
WT
Y68N
R130P
T131I
G132A
PTEN−cancer/ASD mutations with high IS
b
RMSD(nm)
1
0.8
0.6
0.4
0.2
0
0
100
200
300
400
500
Times(ns)
Y68N
R130P
T131I
G132A
WT
c
Y68N
WT
RMSF(nm)
1
0.8
0.6
0.4
0.2
0
16
66
116
166
216
266
316
Residue
d
N68
Region 1:S59-K66
Region 2:Q110-D115
Region 3:F206-C211
Region 4:K260-L265
Region 5:E285-D297
Region 6:A333-N340
residue A79
e
R130P
WT
RMSF(nm)
1
0.8
0.6
0.4
0.2
0
16
66
116
166
216
266
316
Residue
f
P130
Region 1:S59-K66;K80-V85
Region 2:Q110-D115
Region 3:F206-C211
Region 4:K260-L265
Region 5:E285-D297
Region 6:A333-N340
residues H93;T167
g
T131I
WT
RMSF(nm)
1
0.8
0.6
0.4
0.2
0
16
66
116
166
216
266
316
Residue
h
I131
Region 1:S59-K66
Region 2:Q110-D115
Region 3:F206-C211
Region 4:K260-L265
Region 5:E285-D297
Region 6:A333-N340
Region 0: L42-V45
i
G132A
WT
RMSF(nm)
1
0.8
0.6
0.4
0.2
0
16
66
116
166
216
266
316
Residue
j
A132
Region 1:S59-K66;K80-V85
Region 2:Q110-D115
Region 3:F206-C211
Region 4:K260-L265
Region 5:E285-D297
residues G143;R308


a
P loop
WT:R130
Y68N:R130
TI loop
WT:D92
Y68N:D92
WPD loop
b
L70
M134
Y68
Y178
A137
L181
N184
L295
c
V133
N68
I122
Y176
M134
V175
L295
d
P loop
WT:R130
Y68N:R130
TI loop
WT:D92
Y68N:D92
WPD loop
e
R130
Y174
V133
F278
Y177
F279
T286
f
P130
P96
L98
L181
T286
N184
g
P loop
WT:R130
Y68N:R130
TI loop
WT:D92
Y68N:D92
WPD loop
h
G129
Q171
T131
Y174
I280
P281
F279
Y177
E285
T286
i
I131
M134
F279
Y177
A137
F278
Y178
T286
j
P loop
WT:R130
Y68N:R130
TI loop
WT:D92
Y68N:D92
WPD loop
k
G132
V175
V133
Y178
L181
S287
L186
E288
K289
l
A132
V133
Y178
V175
L181
L186
S287
K289