MultiAlloDriver: a multi-model method to predict and identify cancer driver mutations

Abstract

A minority of driver mutations in cancer significantly alter protein structure and key functionalities, thereby driving cancer progression. Consequently, the prediction and identification of driver mutations hold critical implications for targeted cancer therapy. This study introduces MultiAlloDriver, a novel multi-modal machine learning model based on an attention mechanism, which for the first time, incorporates protein surface information. By integrating information from three dimensions-protein sequence, structure, and surface-the model achieves high-accuracy driver target prediction with an accuracy exceeding 93%. Notably, we utilized this tool to predict and identify the driver effect of the F90S mutation in the PTEN tumor suppressor gene, uncovering mutations associated with cancer signaling mechanisms. Overall, MultiAlloDriver contributes to elucidating the underlying mechanisms of cancer development and progression, providing a robust framework for the identification of driver targets.

Introduction

Cancer is a disease caused by the cumulative effects of mutations[1–3]. During its development, a small subset of mutations that significantly alter protein structure, phenotype, or function, thereby driving carcinogenesis, are termed driver mutations[4], while the majority of mutations with minimal impact on protein structure or function are classified as passenger mutations. Numerous studies have demonstrated that driver mutations are closely associated with critical processes in cancer progression, including signal transduction, aberrant proliferation, metastasis, and drug resistance[5–7]. Consequently, a pivotal scientific challenge in cancer-targeted therapy lies in the accurate prediction and identification of driver mutations. High-precision detection of mutation types not only elucidates potential mechanisms underlying cancer progression but also facilitates the discovery of key therapeutic targets and guides the development of targeted anticancer drugs.

Mutations in protein-coding regions can alter amino acid sequences, leading to modified local folding patterns, structural perturbations (local or global), and ultimately functional changes that disrupt interactions with upstream/downstream signaling molecules[8]. Based on their topological and spatial relationships with functional domains, mutation sites can be categorized as orthosteric or allosteric[9–11]. Orthosteric sites reside within functional regions, where mutations directly interfere with domain structure and function, thereby impacting molecular interactions. In contrast, allosteric sites are distal to functional regions. Emerging evidence suggests that such mutations may indirectly modulate functional domains through long-range structural rearrangements (e.g., altered hydrogen bonding networks or conformational dynamics)[12–14]. However, due to mechanistic complexity and limitations of traditional methodologies, allosteric driver mutations have not been as extensively studied as their orthosteric counterparts. Thus, the prediction and characterization of allosteric driver mutations hold significant potential for uncovering novel therapeutic targets in cancer.

Protein surface information, a high-level representation derived from boundary atoms forming a mesh-like structure, encodes critical physicochemical properties such as geometric configuration, charge distribution, hydrophobicity, and hydrogen-bonding capacity. Studies indicate that proteins with highly similar sequences or structures may exhibit divergent molecular binding affinities due to surface variations, whereas surface similarity often correlates with conserved biological activity[15,16]. Compared to sequence or structural data, surface features

more directly govern protein interaction capabilities, particularly for functional domains[17]. For many driver mutations, structural destabilization caused by residue substitutions frequently coincides with localized surface alterations (e.g., hydrophobicity shifts or charge redistribution), which may directly influence folding pathways and secondary structure formation. KRAS, a GTPase belonging to the RAS family, serves as a critical molecular switch in cellular signal transduction, particularly regulating pathways governing cell growth, proliferation, and survival[18]. In its GTP-bound state, KRAS adopts an active conformation to propagate downstream signals, whereas hydrolysis of GTP to GDP transitions it to an inactive state—a cycle tightly regulated under physiological conditions[18,19]. KRAS mutations are strongly associated with malignancies such as pancreatic, colorectal, and non-small cell lung cancers[20,21]. Among these, the G12V variant represents one of the most oncogenic mutations. This substitution introduces a hydrophobic residue that impairs GTPase-mediated hydrolysis, locking KRAS in a constitutively active GTP-bound state[22]. Clinically, G12V has been linked to aggressive tumor phenotypes and poor prognosis[23]. Structural analyses of wild-type versus mutant KRAS (Fig. 1) reveal that the G12V missense mutation induces significant local backbone rearrangements, reshaping the surface topology of the GTP-binding pocket[24]. These alterations correlate with diminished GTPase activity, underscoring the interplay between mutational effects and surface feature remodeling. Anaplastic lymphoma kinase (ALK) is a member of the insulin receptor protein-tyrosine kinase superfamily. Under physiological conditions, ALK is activated by phosphorylation at Y1258 and regulates cell proliferation, differentiation, and survival through downstream signaling pathways (e.g., PI3K/AKT, MAPK/ERK, JAK/STAT), primarily functioning during central nervous system development. Aberrant ALK activation has been implicated in diverse malignancies, including anaplastic large-cell lymphoma, lung cancer, and neuroblastoma[25–27]. The F1174L mutation is a well-characterized pathogenic allosteric driver mutation in ALK. Structural studies demonstrate that substitution of phenylalanine with leucine at position 1174 promotes tighter interactions between the αC-helix and proximal A-loop helix, while inducing distal A-loop distortion[28]. This conformational rearrangement exposes the P+1 pocket, leading to ligand-independent constitutive kinase activation even in the absence of Y1258 phosphorylation. Further analysis of the protein structure revealed that the F1174L mutation induces significant alterations in the geometry of adjacent surface pockets. The DEAD-box RNA helicase DDX3X belongs to the

Superfamily 2 (SF2) of RNA helicases and contains two tandem RecA-like domains (D1 and D2)[29]. It catalyzes ATP hydrolysis and remodels RNA secondary structure through several conserved helicase motifs. DDX3X plays critical roles in various cellular processes, including the regulation of translation of mRNAs with complex 5′ untranslated regions (UTRs). Loss of DDX3X function has been implicated in several cancers, including medulloblastoma[30]. Among the disease-associated mutations, D354V occurs at the interface between the D1 and D2 domains. This mutation replaces a hydrophilic, negatively charged aspartic acid (D) with a more hydrophobic valine (V), thereby disrupting electrostatic interactions and increasing hydrophobicity at the domain interface. Crystallographic analysis indicates that while the D354V mutation does not significantly alter the individual structures of D1 or D2, it does induce a substantial rearrangement of their relative orientation, with D2 rotated approximately 180° compared to the wild type. This conformational shift results in an "open but dynamic" structure, which likely contributes to crystallization stability[31]. More importantly, although D354V does not directly impair canonical ATP-binding motifs, it affects the N-terminal ATP-binding loop (ABL, residues 135–168), thereby indirectly reducing RNA-stimulated ATPase activity. The mutant retains catalytic ATP hydrolysis capacity; however, it exhibits only mild defects in dsRNA binding and ATPase stimulation in the presence of double-stranded RNA. These findings suggest that D354V compromises the protein's responsiveness to specific RNA structures, which may underlie its pathogenic effects. The protein tyrosine kinase JAK2 is a key signaling molecule downstream of a wide range of cytokine receptors, including those for erythropoietin, growth hormone, interferon-γ, and interleukins such as IL-3 and IL-5. It functions via the JAK-STAT signaling pathway to regulate hematopoietic cell proliferation, differentiation, and immune responses[32]. JAK2 is composed of several conserved domains: an N-terminal FERM domain responsible for receptor binding, a SH2-like domain with unclear function, a pseudokinase domain (JH2), and a C-terminal catalytic tyrosine kinase domain (JH1)[33]. Mutations in the JH2 domain of JAK2 are strongly associated with myeloproliferative neoplasms (MPNs), which are clonal hematopoietic disorders characterized by excessive proliferation of myeloid lineages. The V617F mutation in JH2 is the most common pathogenic alteration and may lead to the development of polycythemia vera, essential thrombocythemia, and primary myelofibrosis. Structurally, V617F replaces a small, nonpolar valine with a bulky aromatic phenylalanine, enhancing hydrophobic interactions and rigidifying the adjacent αC-helix within

the N-lobe of the JH2 domain. This conformational stabilization promotes the stimulatory interaction of JH2 with JH1, facilitating constitutive trans-phosphorylation of the activation loop in JH1 independent of cytokine stimulation[34]. Computational prediction using FPocket confirmed substantial differences in both pocket volume and druggability scores, underscoring a direct correlation between the mutation and structural remodeling of surface pockets.

Correlatively, James W. McCormick et al. (2021) revealed that the allosteric H124Q mutation in Dihydrofolate Reductase markedly increased E. coli photoresponse, with synergistic mutations distributed across the protein surface amplifying allosteric effects[35]. Similarly, Nikita R. Dsouza et al. (2024) demonstrated that mutations in Granzyme H and IGF1R statistically altered surface charge distributions, explaining crystallographic disparities between wild-type and mutant proteins[16]. These findings underscore the critical role of surface properties in mediating mutation-induced functional changes, highlighting their importance for allosteric driver site identification.

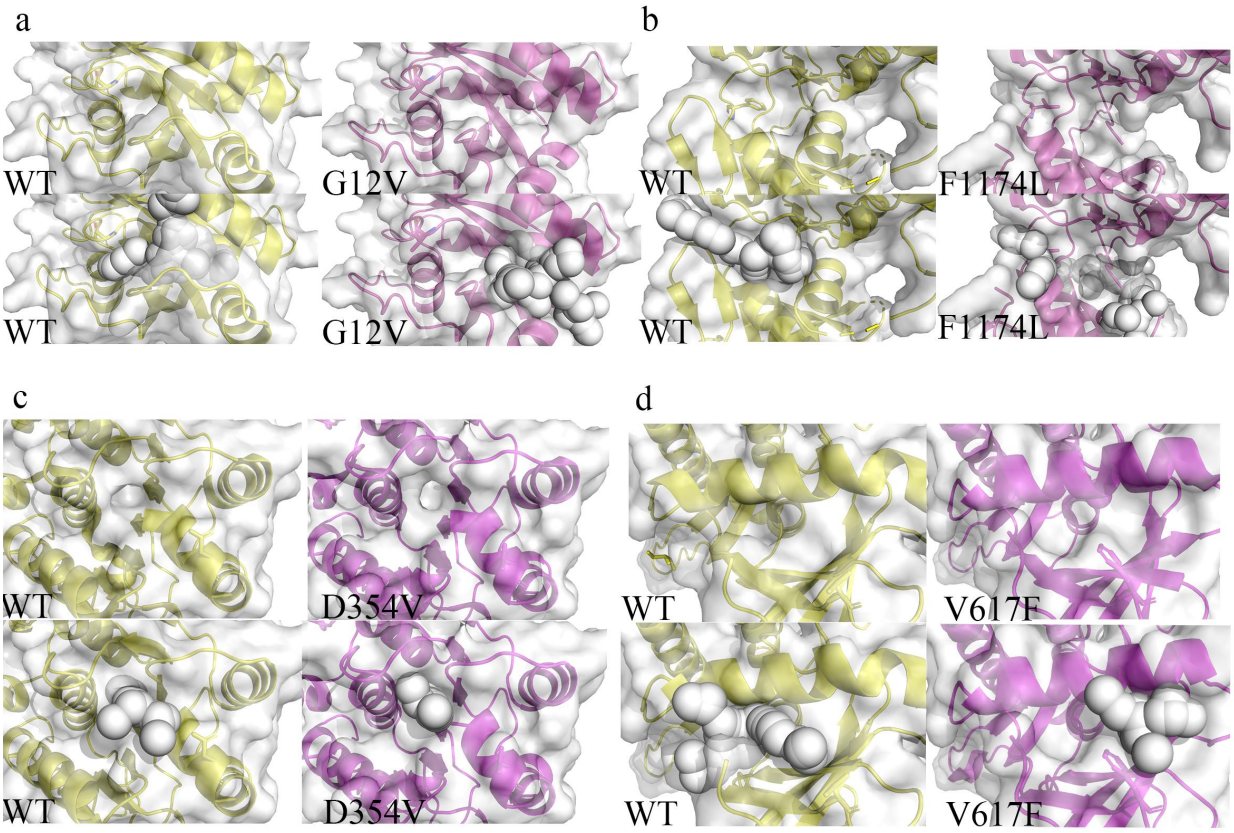

Figure 1：Protein surface change before and after mutation, the original one and the one after using FPocket to represent the structure. a. G12V mutation in KRAS. **b.**F1174L in ALK. **c.**D354V in DDX3X. **d.**V617F in JAK2.

While surface information has been widely applied in drug design and molecular docking[36], its inherent complexity, high-dimensional representation, and technical challenges in feature extraction historically limited its utilization. Recent advances in machine learning, however, have enabled deep mining of surface features. The Masif model (2020)[37], employing geodesic convolutional neural networks to process surface meshes, demonstrated robust performance in molecular docking but suffered from non-end-to-end processing and high memory demands. Subsequently, the dMasif framework addressed these limitations by representing surfaces as atomic point clouds, achieving enhanced computational efficiency and performance[38].Despite advances facilitated by AlphaFold and machine learning, current models for predicting allosteric driver sites predominantly rely on sequence and structural features[39–42] while neglecting surface information — a key determinant of molecular interactions, especially for surface-localized allosteric mutations. To address this gap, we propose MultiAlloDriver, a multimodal deep learning framework integrating sequence, structure, and surface features based on our past research[43,44]. Our approach employs: (1) ESM-2 for extracting long-range sequence dependencies, (2) GVP-GNNs for structural graph representations, and (3) dMasif-based atomic point clouds for surface feature encoding. A transformer-based attention mechanism aligns and fuses these multimodal features. MultiAlloDriver achieved 94% accuracy and precision on test datasets. Furthermore, we validated its utility by identifying the F90S allosteric driver mutation in the tumor suppressor PTEN, elucidating its impact on dephosphorylation activity. As the first model incorporating surface information for mutation effect prediction, MultiAlloDriver unveils novel cancer mechanisms and provides a powerful tool for guiding targeted therapy development.Materials and Methods.

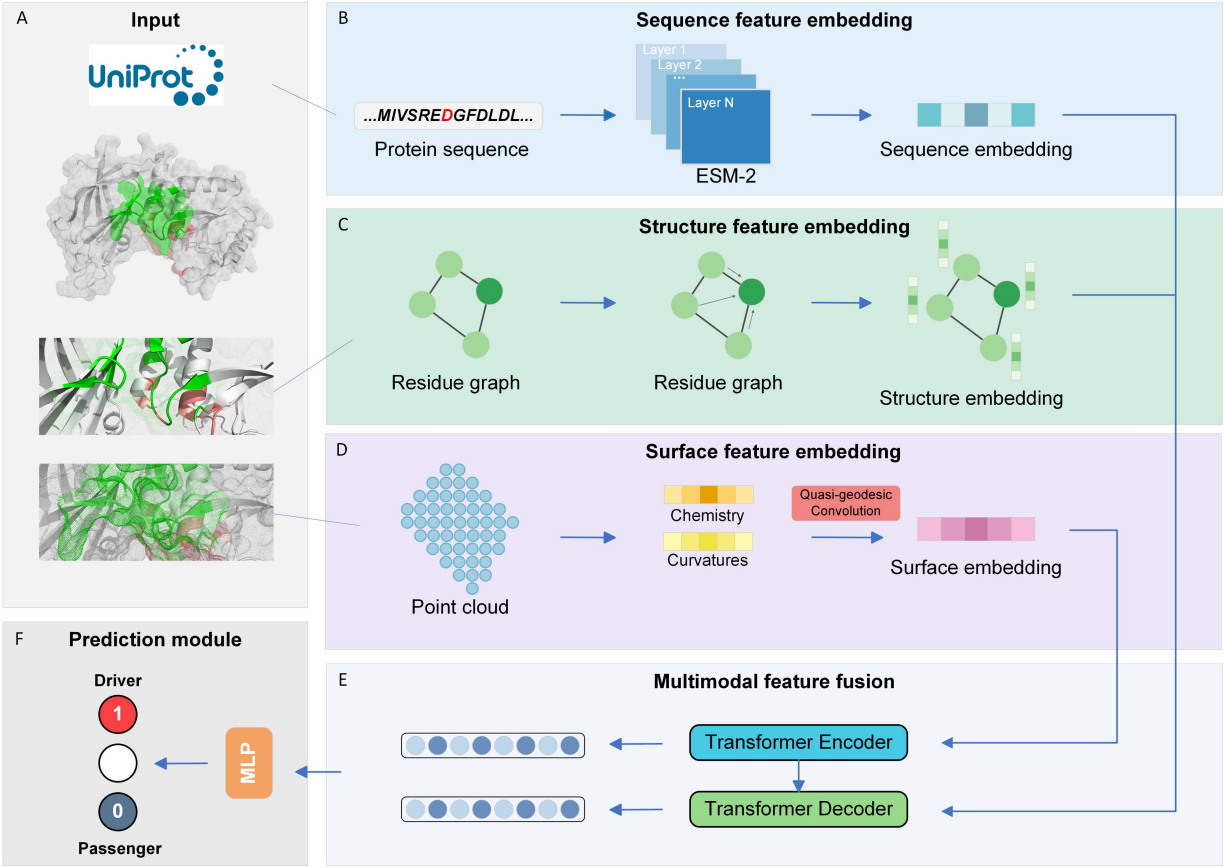

Figure 2: The workflow of MultiAlloDriver

Performance

We randomly divided the 17,130 allosteric driver and passenger mutations into training, validation, and test sets in an 8:1:1 ratio (13,704 for training, 1,713 for validation, and 1,713 for testing).

Using these data, we built multimodal input features to train and assess a driver-mutation recognition model. The resulting deep network fuses primary protein sequence, 3-D backbone geometry, and surface chemo-geometric properties to discriminate drivers from passengers. After training, the model outputs continuous probabilities on previously unseen data; following the literature, scores above 0.50 are labeled as drivers, otherwise passengers.

On the validation set the model achieved an accuracy of 0.9348, and the overall accuracy reached 0.9365. The confusion-matrix metrics show balanced precision (0.9314) and recall (0.9371), yielding an F_1 -score of 0.9342—indicating a high detection rate for driver mutations while controlling false positives. Crucially, the threshold-independent ROC-AUC reached 0.97,

far exceeding the 0.80 benchmark for "excellent" performance and demonstrating an ideal trade-off between true- and false-positive rates across all thresholds. Together, these results place our framework at the forefront of current methods in both accuracy and robustness, offering a reliable tool for large-scale functional annotation of cancer variants and for clinical precision-medicine screening.

Surface Feature Ablation Analysis

To evaluate the contribution of surface features to model performance and validate their importance in protein feature extraction, we conducted ablation experiments (Figure n) comparing the complete MultiAlloDriver model with a degraded version lacking surface information inputs. Performance metrics were systematically compared between the full model (integrating sequence, structure, and surface features) and the ablated model (using only sequence and structural data). Results demonstrated significant declines across all metrics upon surface feature removal: accuracy decreased by 21%, AUC dropped by 12%, and recall exhibited the most pronounced reduction (41%). These findings confirm that surface information effectively enhances protein feature representation and substantially improves predictive capability for cancer-related allosteric driver mutations.

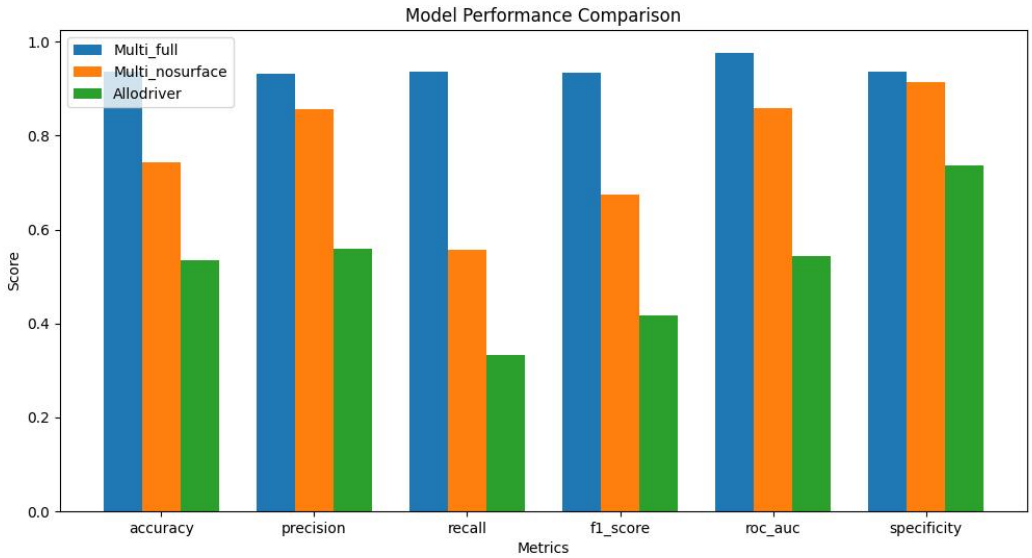

Figure 3. Bar plots comparing performance metrics between the full MultiAlloDriver model (blue) and the surface-ablated model (yellow). Key observations: accuracy (-21%), recall (-41%), and AUC (-12%) exhibited substantial reductions upon surface feature removal.

Case Studies

1. PI3Kα (Phosphoinositide 3-Kinase Alpha)

As a critical member of the class I PI3K family, PI3Kα functions as a heterodimer comprising a catalytic subunit (p110α) and a regulatory subunit (p85α/β). The p110α subunit catalyzes the phosphorylation of phosphatidylinositol 4,5-bisphosphate (PIP2) to phosphatidylinositol 3,4,5-trisphosphate (PIP3), while the p85 subunit modulates enzymatic activity and mediates interactions with upstream signaling molecules (e.g., receptor tyrosine kinases). PI3Kα plays a central role in the PI3K/AKT/mTOR pathway, regulating cell growth, proliferation, survival, and metabolism. Dysregulated PI3Kα activation is strongly associated with malignancies such as breast, colorectal, and endometrial cancers. MultiAlloDriver identified the E542K and E545K mutations within the helical domain of p110α as high-confidence allosteric driver sites (prediction probabilities: 99.74% and 99.56%, respectively). These mutations localize to a key allosteric pocket, where they disrupt interfacial salt bridges, relieve autoinhibition, and induce active-site remodeling—mechanisms corroborated by prior studies linking these variants to constitutive PI3Kα activation[14,45].

2. PTEN (Phosphatase and Tensin Homolog)

PTEN, a pivotal tumor suppressor, antagonizes PI3K signaling by dephosphorylating PIP3 to PIP2, thereby inhibiting the PI3K/AKT/mTOR pathway and regulating apoptosis, proliferation, and migration. Its functional domains include the phosphatase domain, C2 domain (mediating membrane binding), and PDZ-binding motif. MultiAlloDriver predicted the F90S mutation within the phosphatase domain as a potent allosteric driver (probability: 99.86%). Structural mapping revealed that F90S resides in an allosteric network involving Phe90, Glu91, Asp92, His93, Asn94, Pro95, Pro96, Gln97, Gln219, Leu220, Lys221, and Asp252. This mutation induces long-range allosteric effects that impair C2 domain-membrane interactions (Reference), significantly reducing PTEN's dephosphorylation activity[46]. These results highlight the therapeutic potential of targeting F90S for novel anticancer drug development.

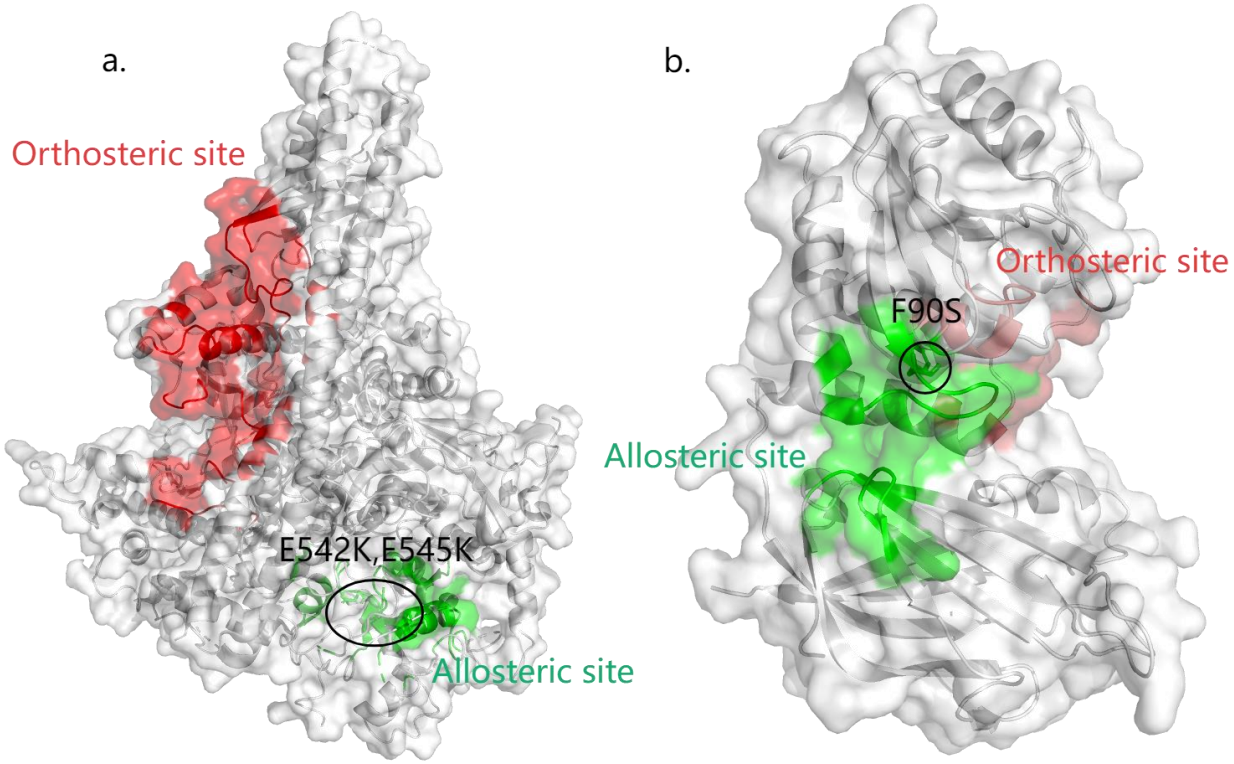

Figure 4. Examples of MultiAlloDriver. a. The driver mutation E542K, E545K is found at the PI3Kα allosteric site (PDB ID: 4OVV). The protein structure is exhibited in both cartoon and transparent surface modes with allosteric site coloured in green and orthosteric site in red. **b.** The driver mutation F90S is found at the PTEN allosteric site (PDB ID:1D5R). The protein structure is exhibited in both cartoon and transparent surface modes with allosteric site coloured in green and orthosteric site in red.

Methods

Dataset

The dataset used in this study was compiled by our group in earlier work and was originally collected and constructed systematically by [Author Name] et al. (see Reference [XX])[47]. It integrates driver-mutation information from multiple public databases, including the Cancer Genome Interpreter (CGI)[48], Integrative OncoGenomics (IntOGen)[49], Clinical Interpretation of Variants in Cancer (CIViC)[50], the Database of Curated Mutations (DoCM)[51], MSK's OncoKB[52], and the Precision Medicine Knowledgebase (PMKB)[53]. After rigorous standardization and

filtering, we obtained 715 driver genes with a total of 17,181 mutation records, among which 8,565 are confirmed driver mutations.

To create a balanced training set, we randomly selected an equal number of passenger mutations from other sources, yielding a balanced dataset of 17,130 samples comprising both driver and passenger mutations.

Mutimodal structure

Surface feature embedding

Protein surfaces are rich in chemical and geometric features that directly reflect interactions between a mutation site and nearby residues[54]. Unlike the costly, non-differentiable pre-computations required by MaSIF, dMaSIF employs a fully end-to-end differentiable network that takes only atom types and 3-D coordinates as input, efficiently reconstructing and encoding the local geometry and chemical environment of the surface[55]. Concretely, for each mutation site we extract an allosteric site by cropping the 512 atoms lying within a ~7 Å radius. From this region we derive surface descriptors—such as curvature distribution, charge density, and the spatial arrangement of polar groups—and finally generate a fixed-dimensional embedding vector that represents the allosteric site.

dMaSIF treats a protein as an unordered point cloud of atoms. For each atom type, it first builds a 22-dimensional one-hot vector to encode its chemical identity. It then assigns a characteristic influence radius to every atom type and, on that basis, defines a smooth distance field that gives the shortest distance from any point in space to the molecular surface. By solving the iso-distance condition on this field, the method uses numerical optimization to uniformly sample surface points at the target distance, achieving a resolution of about 1 Å and typically producing several thousand to tens of thousands of points—enough to cover the entire exterior and capture fine concave-convex details.

At each sampled point, the algorithm computes the gradient of the distance field to estimate the surface normal. It then constructs two tangent vectors in the plane orthogonal to this normal. Together, the normal and the two orthogonal tangent vectors define a local 3-D coordinate frame. This frame provides invariance to rigid-body motions and local rotations in the downstream network, enabling a robust encoding of protein shape.

For chemical feature extraction, we query the 16 nearest atomic centers around each surface point. The one-hot encodings of these atoms are passed through a lightweight multilayer perceptron (MLP), whose nonlinear mapping yields a fixed-dimensional chemical embedding. For geometric feature extraction, we combine the curvature information at that point with the directions of the local tangent vectors. Using either local surface fitting or spectral techniques, we derive a geometric vector that captures the degree of concavity/convexity and bending of the surface patch.

The chemical embedding and geometric vector are then concatenated to form a multimodal feature vector for every surface point. This vector simultaneously encodes the chemical environment of nearby atomic clusters and the fine-grained morphology of the molecular surface, supplying downstream deep networks with a comprehensive, compact, and fully differentiable representation—thereby enhancing the prediction of allosteric sites and the functional impact of driver-mutation sites.

To further aggregate the multimodal features of surface points and generate the final embedding, we employ quasi-geodesic convolution. Designed specifically for protein surface–oriented point clouds, this algorithm propagates information along approximate geodesic paths within local neighborhoods, allowing the network to adaptively learn feature representations pertinent to driver-mutation prediction. Unlike conventional Euclidean convolutions, quasi-geodesic convolution is naturally equivariant or invariant to 3-D rotations and translations—meaning that, regardless of how the protein is oriented in space, the resulting convolved features remain consistent.

Concretely, for each sampled surface point the method first constructs an approximate geodesic neighborhood using the local tangent vector field. Within this neighborhood it defines a set of learnable filter kernels, which aggregate features through weighted summation. Stacking multiple such convolutional layers enables the network to capture progressively larger surface-context information. This process preserves fine-scale geometric curvature and chemical-distribution details while distilling global conformational patterns layer by layer, providing the downstream multimodal network with a highly robust and biologically meaningful representation that markedly enhances the identification and classification of driver-mutation sites.

Structure feature embedding

In deep learning, a protein's three-dimensional structure encodes abundant physicochemical information, clearly revealing how each residue interacts with its surrounding amino acids and how signals propagate through the molecule. To mine these structural cues efficiently, a protein is typically abstracted as a residue graph—nodes represent residues (or atoms), while edges link nodes according to spatial proximity or chemical bonds—and a graph neural network (GNN) serves as the core feature extractor.

Through an iterative message-passing scheme, a GNN captures not only local inter-residue dependencies but, when multiple layers are stacked, the protein's global topology as well. At each layer, a node receives messages from its neighbors; these messages are generated by learnable functions that combine the neighbors' features with relative spatial encodings. An aggregation operation (such as a weighted sum or normalized mean) pools the messages into a neighborhood representation, which is then fused with the node's own features to update its embedding. Modern GNN designs commonly incorporate residual connections and layer normalization between layers to ensure stable information flow in deep networks and to reduce over-smoothing.

By jointly leveraging residue sequences, chemical attributes, and geometric coordinates, GNN-based models have delivered strong results across a variety of tasks, including functional prediction, active-site identification, protein–protein interaction classification, and drug-binding affinity estimation[56,57,58].

To achieve simultaneous invariance to 3-D rotations and translations, we embed a Geometric Vector Perceptron (GVP) into the standard message-passing framework[59]. Designed for biomolecular data, GVP separates each node's representation into scalar and vector channels. In the message and update functions, scalar and vector features are processed through independent linear projections followed by nonlinear activations. Before producing the new scalar output, the L_2 norm of the incoming vector channel is concatenated to the projected scalar features, preserving rotational invariance. The new vector output is obtained by applying a linear transform directly to the input vector, ensuring equivariance with respect to 3-D rotations. Through this mechanism, each round of message passing updates chemical and sequence information while precisely capturing local geometric directions and deformation trends, yielding node embeddings that remain biologically meaningful and robust to rigid-body motions.

322 Our network also adopts a multi-scale aggregation strategy in which skip connections fuse
323 GVP outputs from different depths. By integrating signals that span from atomic to whole-
324 protein scales, the model can detect subtle geometric changes in local allosteric pockets as well
325 as large-scale backbone rearrangements. This comprehensive structural awareness significantly
326 enhances driver-mutation classification performance and supports more insightful biological
327 interpretation.
328
329 **Sequence feature embedding**
330 ESM-2 (Evolutionary Scale Modeling 2) is currently the best-performing general-purpose
331 protein language model[60]. Built on a Transformer architecture, it has been self-supervised on
332 large sequence repositories such as UniProt via a masked-language-modeling objective, enabling
333 it to capture evolutionary signals and latent structural information directly from sequences.
334 Beyond predicting structure and function for single proteins, the model supports a wide array of
335 downstream tasks, including active-site identification and discrimination of protein–protein
336 interaction interfaces[61,62].
337 In this study we benchmarked four differently sized ESM-2 variants: a 6-layer model with 8
338 million parameters, a 12-layer model with 35 million parameters, a 30-layer model with 150
339 million parameters, and a 33-layer model with 650 million parameters. Although performance
340 generally increases with model size, larger models impose greater demands on computation and
341 GPU memory. Balancing accuracy gains against hardware constraints, we ultimately selected the
342 largest variant—ESM2_t33_650M_UR50D, containing 33 layers and roughly 650 million
343 parameters—for our experiments.
344 In practice, every layer of the model assigns a 1,280-dimensional hidden vector to each
345 amino-acid residue in the sequence. These vectors fuse local sequence context with information
346 about evolutionary conservation and structural propensities. We first retrieve the full amino-acid
347 sequence of the target protein via its UniProt ID and replace the wild-type residue at the mutation
348 site to obtain the mutated sequence. Both sequences are then passed through the pre-loaded
349 ESM2_t33_650M_UR50D model, from which we extract the 1,280-dimensional residue
350 embeddings produced by the deepest—33rd—layer; prior studies indicate that representations
351 from the final layer capture higher-level abstractions capable of distinguishing diverse functional
352 sites.

Finally, we perform global average pooling over the entire sequence of residue vectors to produce a fixed-length sequence feature. This vector not only quantifies the impact of a single-point mutation in the sequence-representation space but also integrates seamlessly with structural and surface-level modalities, providing a rich and efficient descriptor for accurately classifying driver versus passenger mutations.

Multimodal feature fusion

In practical settings, combining multiple modalities poses a number of challenges. The data distributions and feature scales differ markedly across modalities: sequence embeddings encode linguistic information as high-dimensional dense vectors, whereas surface point-cloud features are sparse geometric–chemical descriptors. Simple concatenation or naïve blending can let one modality dominate while drowning out the others. Moreover, sequence, structural, and surface data are neither spatially nor sequentially synchronized, and without a clear alignment mechanism it is difficult to establish precise correspondences among them.

Multimodal fusion is still hampered by heterogeneous data distributions, mismatched feature scales, and the lack of an explicit alignment mechanism—issues that can cause one modality to dominate and make it difficult to establish precise correspondences across sequence, structure, and surface representations. Against this backdrop, the Transformer architecture offers natural advantages. It can model serialized features and capture long-range dependencies through self-attention, and—by means of cross-attention—build direct links between different feature spaces, enabling deep alignment and complementary integration of information.

We first feed the surface-point features extracted by dMaSIF into a dedicated Transformer encoder. Multi-head self-attention lets the model evaluate similarity and interaction between any two points globally, automatically focusing on critical regions such as surface pockets and charge patterns. Subsequent feed-forward layers, together with residual connections and layer normalization, yield a high-dimensional representation that optimally summarizes global surface geometry. In parallel, we project the ESM-2 sequence embeddings and the GVP-processed graph embeddings through separate linear layers and add positional encodings to preserve both backbone order and spatial relationships. Each is then refined by an additional self-attention layer, which amplifies long-range co-evolutionary signals in the sequence and global coupling among sub-domains in the graph.

During fusion, we supply the concatenated sequence + structure embeddings as the queries and the surface embeddings as keys and values to a Transformer decoder. Cross-attention directly gauges the similarity and complementarity between sequence–backbone features and surface descriptors, aligning their shared patterns at driver-mutation sites. After several layers of cross-attention and feed-forward refinement, the decoder outputs a unified fusion vector that retains the unique information of each modality while achieving complementary optimization in global context—furnishing the downstream classifier with a highly discriminative multimodal representation.

Model Training

Loss Function

The model produces a probability score indicating whether a single–point mutation is a driver. We compute the discrepancy between predictions and ground-truth labels with *Binary Cross-Entropy with Logits* (BCEWithLogitsLoss). Because this formulation fuses the Sigmoid activation with cross-entropy in a numerically stable way, it alleviates issues arising from extreme probability values and promotes smooth gradient flow, helping the network converge reliably.

Optimizer

Training employs the Adam optimizer, whose adaptive learning-rate scheme blends first- and second-moment estimates of the gradients. This accelerates convergence while dampening the impact of gradient noise throughout training.

Training and Validation Procedure

An epoch consists of a full pass over the training set followed by a single validation run. During training, batch losses are calculated with BCEWithLogitsLoss and back-propagated through Adam to update all learnable parameters. In validation, only forward inference is executed, and intermediate metrics—such as accuracy, precision/recall, and ROC-AUC—are recorded to monitor generalization and guide hyper-parameter tuning.

Discussion

This study presents a novel end-to-end differentiable multimodal framework, MultiAlloDriver, which integrates protein surface information with sequence and structural features to achieve precise identification of cancer driver mutations. First, the ESM-2 protein language model was employed to capture long-range sequence dependencies, while GVP-GNNs enabled equivariant encoding of three-dimensional structural features. Concurrently, atomic-level surface geometry and physicochemical properties were reconstructed using the dMaSIF algorithm. A transformer-based architecture, leveraging self-attention and cross-attention mechanisms, facilitated global alignment and synergistic fusion of these multimodal features, ensuring complementary information flow across sequence, structure, and surface representations. Evaluated on a large-scale driver/passenger mutation dataset previously curated by our group, the model demonstrated robust performance, achieving an accuracy and precision of approximately 94% on the test set. Its biological relevance was further validated through the identification of the F90S allosteric driver mutation in PTEN, where MultiAlloDriver successfully predicted its impact on dephosphorylation activity, underscoring the framework's generalizability and mechanistic interpretability.

By pioneering the integration of protein surface information—a high-dimensional functional descriptor—into driver mutation prediction, MultiAlloDriver not only enhances the detection of complex allosteric sites but also provides new avenues to elucidate protein-protein and protein-ligand interaction mechanisms. Future efforts will focus on extending this framework to additional cancer-related genes and ligand-binding systems, refining model architectures and feature representation strategies. Such advancements are anticipated to yield computational tools with enhanced predictive power for targeted drug discovery and precision oncology.

FUNDING

This study was supported by grants from the National Key R&D Program of China (NO.2023YFC3404700)，the National Natural ScienceFoundation of China (No. 81925034).

REFERENCES

(1) Kiri, S.; Ryba, T. Cancer, Metastasis, and the Epigenome. *Mol Cancer* **2024**, *23* (1), 154. https://doi.org/10.1186/s12943-024-02069-w.

(2) Australian Pancreatic Cancer Genome Initiative; ICGC Breast Cancer Consortium; ICGC MMML-Seq Consortium; ICGC PedBrain; Alexandrov, L. B.; Nik-Zainal, S.; Wedge, D. C.; Aparicio, S. A. J. R.; Behjati, S.; Biankin, A. V.; Bignell, G. R.; Bolli, N.; Borg, A.; Børresen-Dale, A.-L.; Boyault, S.; Burkhardt, B.; Butler, A. P.; Caldas, C.; Davies, H. R.; Desmedt, C.; Eils, R.; Eyfjörd, J. E.; Foekens, J. A.; Greaves, M.; Hosoda, F.; Hutter, B.; Ilicic, T.; Imbeaud, S.; Imielinski, M.; Jäger, N.; Jones, D. T. W.; Jones, D.; Knappskog, S.; Kool, M.; Lakhani, S. R.; López-Otín, C.; Martin, S.; Munshi, N. C.; Nakamura, H.; Northcott, P. A.; Pajic, M.; Papaemmanuil, E.; Paradiso, A.; Pearson, J. V.; Puente, X. S.; Raine, K.; Ramakrishna, M.; Richardson, A. L.; Richter, J.; Rosenstiel, P.; Schlesner, M.; Schumacher, T. N.; Span, P. N.; Teague, J. W.; Totoki, Y.; Tutt, A. N. J.; Valdés-Mas, R.; Van Buuren, M. M.; Van 'T Veer, L.; Vincent-Salomon, A.; Waddell, N.; Yates, L. R.; Zucman-Rossi, J.; Andrew Futreal, P.; McDermott, U.; Lichter, P.; Meyerson, M.; Grimmond, S. M.; Siebert, R.; Campo, E.; Shibata, T.; Pfister, S. M.; Campbell, P. J.; Stratton, M. R. Signatures of Mutational Processes in Human Cancer. *Nature* **2013**, *500* (7463), 415–421. https://doi.org/10.1038/nature12477.

(3) Kandoth, C.; McLellan, M. D.; Vandin, F.; Ye, K.; Niu, B.; Lu, C.; Xie, M.; Zhang, Q.; McMichael, J. F.; Wyczalkowski, M. A.; Leiserson, M. D. M.; Miller, C. A.; Welch, J. S.; Walter, M. J.; Wendl, M. C.; Ley, T. J.; Wilson, R. K.; Raphael, B. J.; Ding, L. Mutational Landscape and Significance across 12 Major Cancer Types. *Nature* **2013**, *502* (7471), 333–339. https://doi.org/10.1038/nature12634.

(4) Martínez-Jiménez, F.; Muiños, F.; Sentís, I.; Deu-Pons, J.; Reyes-Salazar, I.; Arnedo-Pac, C.; Mularoni, L.; Pich, O.; Bonet, J.; Kranas, H.; Gonzalez-Perez, A.; Lopez-Bigas, N. A Compendium of Mutational Cancer Driver Genes. *Nat Rev Cancer* **2020**, *20* (10), 555–572. https://doi.org/10.1038/s41568-020-0290-x.

(5) Vasan, N.; Razavi, P.; Johnson, J. L.; Shao, H.; Shah, H.; Antoine, A.; Ladewig, E.; Gorelick, A.; Lin, T.-Y.; Toska, E.; Xu, G.; Kazmi, A.; Chang, M. T.; Taylor, B. S.; Dickler, M. N.; Jhaveri, K.; Chandarlapaty, S.; Rabadan, R.; Reznik, E.; Smith, M. L.; Sebra, R.; Schimmoller, F.; Wilson, T. R.; Friedman, L. S.; Cantley, L. C.; Scaltriti, M.; Baselga, J. Double *PIK3CA* Mutations in Cis Increase Oncogenicity and Sensitivity to PI3Kα Inhibitors. *Science* **2019**, *366* (6466), 714–723. https://doi.org/10.1126/science.aaw9032.

(6) Bozic, I.; Antal, T.; Ohtsuki, H.; Carter, H.; Kim, D.; Chen, S.; Karchin, R.; Kinzler, K. W.; Vogelstein, B.; Nowak, M. A. Accumulation of Driver and Passenger Mutations during Tumor Progression. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107* (43), 18545–18550. https://doi.org/10.1073/pnas.1010978107.

(7) Daniel, A. H.; Jeff, S. Cancer: Drivers and Passengers.

(8) Vila, J. A. Analysis of Proteins in the Light of Mutations. *Eur Biophys J* **2024**, *53* (5–6), 255–265. https://doi.org/10.1007/s00249-024-01714-y.

(9) Ostroverkhova, D.; Przytycka, T. M.; Panchenko, A. R. Cancer Driver Mutations: Predictions and Reality. *Trends in Molecular Medicine* **2023**, *29* (7), 554–566. https://doi.org/10.1016/j.molmed.2023.03.007.

(10) Changeux, J.-P.; Christopoulos, A. Allosteric Modulation as a Unifying Mechanism for Receptor Function and Regulation. *Cell* **2016**, *166* (5), 1084–1102. https://doi.org/10.1016/j.cell.2016.08.015.

(11) Van Westen, G. J. P.; Gaulton, A.; Overington, J. P. Chemical, Target, and Bioactive Properties of Allosteric Modulation. *PLoS Comput Biol* **2014**, *10* (4), e1003559. https://doi.org/10.1371/journal.pcbi.1003559.

(12) Liu, J.; Nussinov, R. Allostery: An Overview of Its History, Concepts, Methods, and Applications. *PLoS Comput Biol* **2016**, *12* (6), e1004966. https://doi.org/10.1371/journal.pcbi.1004966.

(13) Lu, S.; Qiu, Y.; Ni, D.; He, X.; Pu, J.; Zhang, J. Emergence of Allosteric Drug-Resistance Mutations: New Challenges for Allosteric Drug Discovery. *Drug Discovery Today* **2020**, *25* (1), 177–184. https://doi.org/10.1016/j.drudis.2019.10.006.

(14) Nussinov, R.; Zhang, M.; Maloney, R.; Liu, Y.; Tsai, C.-J.; Jang, H. Allostery: Allosteric Cancer Drivers and Innovative Allosteric Drugs. *Journal of Molecular Biology* **2022**, *434* (17), 167569. https://doi.org/10.1016/j.jmb.2022.167569.

(15) Tseng, Y. Y.; Li, W.-H. Classification of Protein Functional Surfaces Using Structural Characteristics. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109* (4), 1170–1175. https://doi.org/10.1073/pnas.1119684109.

(16) Dsouza, N. R.; Haque, N.; Tripathi, S.; Zimmermann, M. T. Assessing Protein Surface-Based Scoring for Interpreting Genomic Variants. *IJMS* **2024**, *25* (22), 12018. https://doi.org/10.3390/ijms252212018.

(17) Riahi, S.; Lee, J. H.; Sorenson, T.; Wei, S.; Jager, S.; Olfati-Saber, R.; Zhou, Y.; Park, A.; Wendt, M.; Minoux, H.; Qiu, Y. Surface ID: A Geometry-Aware System for Protein Molecular Surface Comparison. *Bioinformatics* **2023**, *39* (4), btad196. https://doi.org/10.1093/bioinformatics/btad196.

(18) Luo, J. KRAS Mutation in Pancreatic Cancer. *Seminars in Oncology* **2021**, *48* (1), 10–18. https://doi.org/10.1053/j.seminoncol.2021.02.003.

(19) Punekar, S. R.; Velcheti, V.; Neel, B. G.; Wong, K.-K. The Current State of the Art and Future Trends in RAS-Targeted Cancer Therapies. *Nat Rev Clin Oncol* **2022**, *19* (10), 637–655. https://doi.org/10.1038/s41571-022-00671-9.

(20) Yang, Y.; Zhang, H.; Huang, S.; Chu, Q. KRAS Mutations in Solid Tumors: Characteristics, Current Therapeutic Strategy, and Potential Treatment Exploration. *JCM* **2023**, *12* (2), 709. https://doi.org/10.3390/jcm12020709.

(21) Wu, X.; Song, W.; Cheng, C.; Liu, Z.; Li, X.; Cui, Y.; Gao, Y.; Li, D. Small Molecular Inhibitors for KRAS-Mutant Cancers. *Front. Immunol.* **2023**, *14*, 1223433. https://doi.org/10.3389/fimmu.2023.1223433.

521 (22) Pantsar, T. The Current Understanding of KRAS Protein Structure and Dynamics.
522 *Computational and Structural Biotechnology Journal* **2020**, *18*, 189–198.
523 https://doi.org/10.1016/j.csbj.2019.12.004.

524 (23) Pan, L.; Ma, Y.; Li, Z.; Hu, J.; Xu, Z. KRAS G12V Mutation Upregulates PD-L1
525 Expression via TGF-β/EMT Signaling Pathway in Human Non-small-cell Lung Cancer. *Cell*
526 *Biology International* **2021**, *45* (4), 795–803. https://doi.org/10.1002/cbin.11524.

527 (24) Saetang, J.; Yamabhai, M.; Rangnoi, K.; Prompat, N.; Haewphet, T.; Sangkhathat, S.;
528 Tipmanee, V.; Benjakul, S. Identification and Characterization of Oncogenic KRAS G12V
529 Inhibitory Peptides by Phage Display, Molecular Docking and Molecular Dynamic Simulation.
530 *Computers in Biology and Medicine* **2025**, *192*, 110272.
531 https://doi.org/10.1016/j.compbiomed.2025.110272.

532 (25) Salaverria, I.; Beà, S.; Lopez-Guillermo, A.; Lespinet, V.; Pinyol, M.; Burkhardt, B.;
533 Lamant, L.; Zettl, A.; Horsman, D.; Gascoyne, R.; Ott, G.; Siebert, R.; Delsol, G.; Campo, E.
534 Genomic Profiling Reveals Different Genetic Aberrations in Systemic ALK-positive and
535 ALK-negative Anaplastic Large Cell Lymphomas. *Br J Haematol* **2008**, *140* (5), 516–526.
536 https://doi.org/10.1111/j.1365-2141.2007.06924.x.

537 (26) Katayama, R. Drug Resistance in Anaplastic Lymphoma Kinase-rearranged Lung Cancer.
538 *Cancer Science* **2018**, *109* (3), 572–580. https://doi.org/10.1111/cas.13504.

539 (27) Chen, Y.; Takita, J.; Choi, Y. L.; Kato, M.; Ohira, M.; Sanada, M.; Wang, L.; Soda, M.;
540 Kikuchi, A.; Igarashi, T.; Nakagawara, A.; Hayashi, Y.; Mano, H.; Ogawa, S. Oncogenic
541 Mutations of ALK Kinase in Neuroblastoma. *Nature* **2008**, *455* (7215), 971–974.
542 https://doi.org/10.1038/nature07399.

543 (28) Jiang, C.-H.; Huang, C.-X.; Chen, Y.-J.; Chuang, Y.-C.; Huang, B.-Y.; Yang, C.-N.
544 Molecular Modeling for Structural Insights Concerning the Activation Mechanisms of F1174L
545 and R1275Q Mutations on Anaplastic Lymphoma Kinase. *Molecules* **2018**, *23* (7), 1610.
546 https://doi.org/10.3390/molecules23071610.

547 (29) Mallam, A. L.; Campo, M. D.; Gilman, B.; Sidote, D. J.; Lambowitz, A. M. Structural
548 Basis for RNA Duplex Recognition and Unwinding by the DEAD-Box Helicase Mss116p.
549 *Nature* **2012**, *490* (7418), 121–125. https://doi.org/10.1038/nature11402.

550 (30) Robinson, G.; Parker, M.; Kranenburg, T. A.; Lu, C.; Chen, X.; Ding, L.; Phoenix, T. N.;
551 Hedlund, E.; Wei, L.; Zhu, X.; Chalhoub, N.; Baker, S. J.; Huether, R.; Kriwacki, R.; Curley, N.;
552 Thiruvenkatam, R.; Wang, J.; Wu, G.; Rusch, M.; Hong, X.; Beckford, J.; Gupta, P.; Ma, J.;
553 Easton, J.; Vadodaria, B.; Onar-Thomas, A.; Lin, T.; Li, S.; Pounds, S.; Paugh, S.; Zhao, D.;
554 Kawauchi, D.; Roussel, M. F.; Finkelstein, D.; Ellison, D. W.; Lau, C. C.; Bouffet, E.; Hassall,
555 T.; Gururangan, S.; Cohn, R.; Fulton, R. S.; Fulton, L. L.; Dooling, D. J.; Ochoa, K.; Gajjar, A.;
556 Mardis, E. R.; Wilson, R. K.; Downing, J. R.; Zhang, J.; Gilbertson, R. J. Novel Mutations
557 Target Distinct Subgroups of Medulloblastoma. *Nature* **2012**, *488* (7409), 43–48.
558 https://doi.org/10.1038/nature11213.

559 (31) Epling, L. B.; Grace, C. R.; Lowe, B. R.; Partridge, J. F.; Enemark, E. J. Cancer-
560 Associated Mutants of RNA Helicase DDX3X Are Defective in RNA-Stimulated ATP

HydrolysisRNA 解旋酶 DDX3X 的癌症相关突变体在 RNA 刺激的 ATP 水解中存在缺陷. *Journal of Molecular Biology* **2015**, *427* (9), 1779–1796. https://doi.org/10.1016/j.jmb.2015.02.015.

(32) Levy, D. E.; Darnell, J. E. STATs: Transcriptional Control and Biological Impact. *Nat Rev Mol Cell Biol* **2002**, *3* (9), 651–662. https://doi.org/10.1038/nrm909.

(33) Ghoreschi, K.; Laurence, A.; O'Shea, J. J. Janus Kinases in Immune Cell Signaling. *Immunol Rev* **2009**, *228* (1), 273–287. https://doi.org/10.1111/j.1600-065X.2008.00754.x.

(34) Bandaranayake, R. M.; Ungureanu, D.; Shan, Y.; Shaw, D. E.; Silvennoinen, O.; Hubbard, S. R. Crystal Structures of the JAK2 Pseudokinase Domain and the Pathogenic Mutant V617F. *Nat Struct Mol Biol* **2012**, *19* (8), 754–759. https://doi.org/10.1038/nsmb.2348.

(35) McCormick, J. W.; Russo, M. A.; Thompson, S.; Blevins, A.; Reynolds, K. A. Structurally Distributed Surface Sites Tune Allosteric Regulation. *eLife* **2021**, *10*, e68346. https://doi.org/10.7554/eLife.68346.

(36) *Molecular Modeling of Proteins*; Kukol, A., Ed.; Walker, J. M., Series Ed.; Methods in Molecular Biology; Humana Press: Totowa, NJ, 2008; Vol. 443. https://doi.org/10.1007/978-1-59745-177-2.

(37) Sverrisson, F.; Feydy, J.; Correia, B. E.; Bronstein, M. M. Fast End-to-End Learning on Protein Surfaces.

(38) Zheng, Z.; Jiao, Y.; You, H.; An, J.; Sun, Y. Fast End-to-End Surface Interpretation of SARS-CoV-2 Variants by Differentiable Molecular Surface Interaction Fingerprinting Method. *Computational and Structural Biotechnology Journal* **2023**, *21*, 4816–4824. https://doi.org/10.1016/j.csbj.2023.09.033.

(39) Cheng, J.; Novati, G.; Pan, J.; Bycroft, C.; Žemgulytė, A.; Applebaum, T.; Pritzel, A.; Wong, L. H.; Zielinski, M.; Sargeant, T.; Schneider, R. G.; Senior, A. W.; Jumper, J.; Hassabis, D.; Kohli, P.; Avsec, Ž. Accurate Proteome-Wide Missense Variant Effect Prediction with AlphaMissense. *Science* **2023**, *381* (6664), eadg7492. https://doi.org/10.1126/science.adg7492.

(40) Cheng, P.; Mao, C.; Tang, J.; Yang, S.; Cheng, Y.; Wang, W.; Gu, Q.; Han, W.; Chen, H.; Li, S.; Chen, Y.; Zhou, J.; Li, W.; Pan, A.; Zhao, S.; Huang, X.; Zhu, S.; Zhang, J.; Shu, W.; Wang, S. Zero-Shot Prediction of Mutation Effects with Multimodal Deep Representation Learning Guides Protein Engineering. *Cell Res* **2024**, *34* (9), 630–647. https://doi.org/10.1038/s41422-024-00989-2.

(41) Tzavella, K.; Diaz, A.; Olsen, C.; Vranken, W. Combining Evolution and Protein Language Models for an Interpretable Cancer Driver Mutation Prediction with D2Deep. *Briefings in Bioinformatics* **2024**, *26* (1), bbae664. https://doi.org/10.1093/bib/bbae664.

(42) Dieckhaus, H.; Brocidiacono, M.; Randolph, N.; Kuhlman, B. Transfer Learning to Leverage Larger Datasets for Improved Prediction of Protein Stability Changes.

(43) Song, K.; Li, Q.; Gao, W.; Lu, S.; Shen, Q.; Liu, X.; Wu, Y.; Wang, B.; Lin, H.; Chen, G.; Zhang, J. AlloDriver: A Method for the Identification and Analysis of Cancer Driver Targets. *Nucleic Acids Research* **2019**, *47* (W1), W315–W321. https://doi.org/10.1093/nar/gkz350.

(44) Song, Q.; Li, M.; Li, Q.; Lu, X.; Song, K.; Zhang, Z.; Wei, J.; Zhang, L.; Wei, J.; Ye, Y.; Zha, J.; Zhang, Q.; Gao, Q.; Long, J.; Liu, X.; Lu, X.; Zhang, J. DeepAlloDriver: A Deep Learning-Based Strategy to Predict Cancer Driver Mutations. *Nucleic Acids Research* **2023**, *51* (W1), W129–W133. https://doi.org/10.1093/nar/gkad295.

(45) Nussinov, R.; Zhang, M.; Tsai, C.-J.; Jang, H. Phosphorylation and Driver Mutations in PI3Kα and PTEN Autoinhibition. *Molecular Cancer Research* **2021**, *19* (4), 543–548. https://doi.org/10.1158/1541-7786.MCR-20-0818.

(46) Masson, G. R.; Williams, R. L. Structural Mechanisms of PTEN Regulation. *Cold Spring Harb Perspect Med* **2020**, *10* (3), a036152. https://doi.org/10.1101/cshperspect.a036152.

(47) Song, Q.; Li, M.; Li, Q.; Lu, X.; Song, K.; Zhang, Z.; Wei, J.; Zhang, L.; Wei, J.; Ye, Y.; Zha, J.; Zhang, Q.; Gao, Q.; Long, J.; Liu, X.; Lu, X.; Zhang, J. DeepAlloDriver: A Deep Learning-Based Strategy to Predict Cancer Driver Mutations. *Nucleic Acids Research* **2023**, *51* (W1), W129–W133. https://doi.org/10.1093/nar/gkad295.

(48) Tamborero, D.; Rubio-Perez, C.; Deu-Pons, J.; Schroeder, M. P.; Vivancos, A.; Rovira, A.; Tusquets, I.; Albanell, J.; Rodon, J.; Tabernero, J.; de Torres, C.; Dienstmann, R.; Gonzalez-Perez, A.; Lopez-Bigas, N. Cancer Genome Interpreter Annotates the Biological and Clinical Relevance of Tumor Alterations. *Genome Medicine* **2018**, *10* (1), 25. https://doi.org/10.1186/s13073-018-0531-8.

(49) Zhang, J.; Lee, D.; Dhiman, V.; Jiang, P.; Xu, J.; McGillivray, P.; Yang, H.; Liu, J.; Meyerson, W.; Clarke, D.; Gu, M.; Li, S.; Lou, S.; Xu, J.; Lochovsky, L.; Ung, M.; Ma, L.; Yu, S.; Cao, Q.; Harmanci, A.; Yan, K.-K.; Sethi, A.; Gürsoy, G.; Schoenberg, M. R.; Rozowsky, J.; Warrell, J.; Emani, P.; Yang, Y. T.; Galeev, T.; Kong, X.; Liu, S.; Li, X.; Krishnan, J.; Feng, Y.; Rivera-Mulia, J. C.; Adrian, J.; Broach, J. R.; Bolt, M.; Moran, J.; Fitzgerald, D.; Dileep, V.; Liu, T.; Mei, S.; Sasaki, T.; Trevilla-Garcia, C.; Wang, S.; Wang, Y.; Zang, C.; Wang, D.; Klein, R. J.; Snyder, M.; Gilbert, D. M.; Yip, K.; Cheng, C.; Yue, F.; Liu, X. S.; White, K. P.; Gerstein, M. An Integrative ENCODE Resource for Cancer Genomics. *Nat Commun* **2020**, *11* (1), 3696. https://doi.org/10.1038/s41467-020-14743-w.

(50) Griffith, M.; Spies, N. C.; Krysiak, K.; McMichael, J. F.; Coffman, A. C.; Danos, A. M.; Ainscough, B. J.; Ramirez, C. A.; Rieke, D. T.; Kujan, L.; Barnell, E. K.; Wagner, A. H.; Skidmore, Z. L.; Wollam, A.; Liu, C. J.; Jones, M. R.; Bilski, R. L.; Lesurf, R.; Feng, Y.-Y.; Shah, N. M.; Bonakdar, M.; Trani, L.; Matlock, M.; Ramu, A.; Campbell, K. M.; Spies, G. C.; Graubert, A. P.; Gangavarapu, K.; Eldred, J. M.; Larson, D. E.; Walker, J. R.; Good, B. M.; Wu, C.; Su, A. I.; Dienstmann, R.; Margolin, A. A.; Tamborero, D.; Lopez-Bigas, N.; Jones, S. J. M.; Bose, R.; Spencer, D. H.; Wartman, L. D.; Wilson, R. K.; Mardis, E. R.; Griffith, O. L. CIViC Is a Community Knowledgebase for Expert Crowdsourcing the Clinical Interpretation of Variants in Cancer. *Nat Genet* **2017**, *49* (2), 170–174. https://doi.org/10.1038/ng.3774.

(51) Ainscough, B. J.; Griffith, M.; Coffman, A. C.; Wagner, A. H.; Kunisaki, J.; Choudhary, M. N.; McMichael, J. F.; Fulton, R. S.; Wilson, R. K.; Griffith, O. L.; Mardis, E. R. DoCM: A

Database of Curated Mutations in Cancer. *Nat Methods* **2016**, *13* (10), 806–807. https://doi.org/10.1038/nmeth.4000.

(52) Chakravarty, D.; Gao, J.; Phillips, S.; Kundra, R.; Zhang, H.; Wang, J.; Rudolph, J. E.; Yaeger, R.; Soumerai, T.; Nissan, M. H.; Chang, M. T.; Chandarlapaty, S.; Traina, T. A.; Paik, P. K.; Ho, A. L.; Hantash, F. M.; Grupe, A.; Baxi, S. S.; Callahan, M. K.; Snyder, A.; Chi, P.; Danila, D. C.; Gounder, M.; Harding, J. J.; Hellmann, M. D.; Iyer, G.; Janjigian, Y. Y.; Kaley, T.; Levine, D. A.; Lowery, M.; Omuro, A.; Postow, M. A.; Rathkopf, D.; Shoushtari, A. N.; Shukla, N.; Voss, M. H.; Paraiso, E.; Zehir, A.; Berger, M. F.; Taylor, B. S.; Saltz, L. B.; Riely, G. J.; Ladanyi, M.; Hyman, D. M.; Baselga, J.; Sabbatini, P.; Solit, D. B.; Schultz, N. OncoKB: A Precision Oncology Knowledge Base. *JCO Precis Oncol* **2017**, *1*, PO.17.00011. https://doi.org/10.1200/PO.17.00011.

(53) Huang, L.; Fernandes, H.; Zia, H.; Tavassoli, P.; Rennert, H.; Pisapia, D.; Imielinski, M.; Sboner, A.; Rubin, M. A.; Kluk, M.; Elemento, O. The Cancer Precision Medicine Knowledge Base for Structured Clinical-Grade Mutations and Interpretations. *Journal of the American Medical Informatics Association* **2017**, *24* (3), 513–519. https://doi.org/10.1093/jamia/ocw148.

(54) Gainza, P.; Sverrisson, F.; Monti, F.; Rodolà, E.; Boscaini, D.; Bronstein, M. M.; Correia, B. E. Deciphering Interaction Fingerprints from Protein Molecular Surfaces Using Geometric Deep Learning. *Nat Methods* **2020**, *17* (2), 184–192. https://doi.org/10.1038/s41592-019-0666-6.

(55) Sverrisson, F.; Feydy, J.; Correia, B. E.; Bronstein, M. M. Fast End-to-End Learning on Protein Surfaces. bioRxiv December 29, 2020, p 2020.12.28.424589. https://doi.org/10.1101/2020.12.28.424589.

(56) Jang, Y. J.; Qin, Q.-Q.; Huang, S.-Y.; Peter, A. T. J.; Ding, X.-M.; Kornmann, B. Accurate Prediction of Protein Function Using Statistics-Informed Graph Networks. *Nat Commun* **2024**, *15* (1), 6601. https://doi.org/10.1038/s41467-024-50955-0.

(57) Knutson, C.; Bontha, M.; Bilbrey, J. A.; Kumar, N. Decoding the Protein–Ligand Interactions Using Parallel Graph Neural Networks. *Sci Rep* **2022**, *12* (1), 7624. https://doi.org/10.1038/s41598-022-10418-2.

(58) Dandibhotla, S.; Samudrala, M.; Kaneriya, A.; Dakshanamurthy, S. GNNSeq: A Sequence-Based Graph Neural Network for Predicting Protein–Ligand Binding Affinity. *Pharmaceuticals* **2025**, *18* (3), 329. https://doi.org/10.3390/ph18030329.

(59) Jing, B.; Eismann, S.; Suriana, P.; Townshend, R. J. L.; Dror, R. Learning from Protein Structure with Geometric Vector Perceptrons. arXiv May 16, 2021. https://doi.org/10.48550/arXiv.2009.01411.

(60) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; Rives, A. Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model. *Science* **2023**, *379* (6637), 1123–1130. https://doi.org/10.1126/science.ade2574.

(61) Zeng, W.; Dou, Y.; Pan, L.; Xu, L.; Peng, S. Improving Prediction Performance of General Protein Language Model by Domain-Adaptive Pretraining on DNA-Binding Protein. *Nat Commun* **2024**, *15* (1), 7838. https://doi.org/10.1038/s41467-024-52293-7.

(62) Cordoves-Delgado, G.; García-Jacas, C. R. Predicting Antimicrobial Peptides Using ESMFold-Predicted Structures and ESM-2-Based Amino Acid Features with Graph Deep Learning. *Journal of Chemical Information and Modeling* **2024**. https://doi.org/10.1021/acs.jcim.3c02061.