ELITE: E3 Ligase Inference for Tissue specific Elimination: A LLM Based E3 Ligase Prediction System for Precise Targeted Protein Degradation

Sabyasachi Patajoshi[1,*], Sumit Madan[1], Holger Fröhlich[1,2,3,*]

[1] Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, 53757 Sankt Augustin, Germany
[2] Institute for Digital Medicine, University Hospital Bonn, Bonn, Germany
[3] Bonn-Aachen International Center for IT, University of Bonn, Bonn, Germany

* correspondence: holger.froehlich@scai.fraunhofer.de, sabyasachi.patajoshi@scai-extern.fraunhofer.de

Data science maturity

DSML3: Development/pre-production: Data science output has been rolled out/validated across multiple domains/problems

Abstract

Targeted protein degradation (TPD) has transformed modern drug discovery by harnessing the ubiquitin–proteasome system to eliminate disease-driving proteins previously deemed undruggable. However, current approaches predominantly rely on a narrow set of ubiquitously expressed E3 ligases, such as Cereblon (CRBN) and Von Hippel–Lindau (VHL), which limits tissue specificity, increases systemic toxicity, and fosters resistance. Here, we present an AI-driven framework for the rational identification of tissue-specific E3 ligases suitable for precision-targeted degradation. Our model leverages a BERT-based protein language architecture trained on billions of sequences to generate contextual embeddings that capture structural and functional motifs relevant for E3–substrate compatibility. By integrating these embeddings with tissue-resolved protein-protein interaction data, the framework predicts ligase–target interactions that are both biologically plausible and context-restricted. This enables the prioritization of ligases capable of driving selective degradation of pathogenic proteins within disease-relevant tissues. The proposed approach offers a scalable path to expand the E3 ligase repertoire and advance TPD toward true precision medicine.

Significance

Current computational strategies for targeted protein degradation (TPD) largely ignore biological context, relying on global E3–substrate datasets that overlook tissue specificity and thus cannot anticipate off-target toxicity or resistance. Our work introduces a **context-aware, AI-driven framework** that integrates large-scale protein language model embeddings with tissue-resolved protein-protein interaction data to predict compatible and tissue-selective E3 ligases for any pathogenic target. By learning biochemical compatibilities directly from sequence space, this approach transcends the limitations of curated interaction databases and enables generalization to novel ligases or disease mutations. The resulting ligand–target rankings combine biochemical plausibility with spatial expression constraints, providing a scalable foundation for designing degraders that act precisely within disease-relevant tissues. This represents a conceptual and technical advance toward **precision degradomics** - a next generation of targeted protein degradation where therapeutic selectivity is defined not only by molecular affinity but also by cellular and tissue context. Subsequent steps will include structural integration for ternary complex modeling, biochemical validation of predicted pairs, and the deployment of a public platform to guide tissue-specific degrader discovery.

Background, Motivation and Problem Statement

Over the past two decades, targeted protein degradation (TPD) has emerged as a powerful therapeutic strategy for tackling proteins once considered “undruggable,” including drivers of cancer, inflammatory and autoimmune disorders, and neurodegenerative diseases 1. TPD agents exploit the cell’s own ubiquitin-proteasome system (UPS): a bifunctional degrader first recruits an E3 ligase to its target and then facilitates poly-ubiquitin tagging, marking the protein for destruction 2. The poly-ubiquitinated substrate is subsequently recognized by the 26S proteasome and processed into short peptides, free ubiquitin, and amino acids that the cell recycles, thereby eliminating the disease-associated protein and restoring proteostasis 3.

A key modality within TPD is the use of heterobifunctional molecules known as proteolysis-targeting chimeras (PROTACs), which simultaneously bind an E3 ubiquitin ligase and a target protein, bringing them into close proximity to facilitate ubiquitin transfer 4. Ubiquitination is catalyzed through a highly conserved enzymatic cascade involving E1 activating enzymes, E2 conjugating enzymes, and E3 ligases 5. E1 enzymes initiate the process by activating ubiquitin and transferring it to E2 enzymes, which are more diverse and act as intermediaries. E2 enzymes then collaborate with E3 ligases to transfer ubiquitin to specific substrate proteins 5. Among these enzymes, E3 ligases are the most diverse and are critical for substrate specificity, as they determine which proteins are targeted for degradation 6 , 7, 8, 9 . As such, E3 ligases play a central role in UPS and are essential for the precision and effectiveness of targeted protein degradation strategies.

PROTACs are a novel class of small molecules with unique therapeutic potential and associated challenges 10. PROTACs (Proteolysis-Targeting Chimeras) are heterobifunctional molecules designed to harness the cell’s ubiquitin-proteasome system by simultaneously binding a target

protein and an E3 ubiquitin ligase, thereby forming a ternary complex that facilitates the ubiquitination and subsequent proteasomal degradation of the target 11. This catalytic mechanism allows PROTACs to deplete disease-relevant proteins-including many considered undruggable by conventional small molecules-at sub stoichiometric doses and with sustained pharmacodynamic effects 2. In contrast, Molecular Glues are monovalent compounds that enhance or stabilize interactions between an E3 ligase and a neo-substrate, redirecting endogenous degradation machinery toward new targets 13. Although approximately 600 E3 ligases have been identified in the human proteome 14, PROTAC development has predominantly focused on just two-Cereblon (CRBN) and Von Hippel–Lindau (VHL) -due to their well-characterized structures and broad tissue expression 15 .

However, reliance on this narrow subset of E3 ligases presents two major challenges. First, prolonged use of CRBN- or VHL-based degraders can impose evolutionary pressure, leading in the case of cancer to the emergence of resistance mutations or conformational changes that impair E3-ligase binding and reduce degradation efficacy 16. Second, the ubiquitous expression of CRBN and VHL across healthy and diseased tissues increases the risk of off-target protein degradation, raising concerns about systemic toxicity and limiting therapeutic selectivity 2. While TPD offers a powerful mechanism to eliminate disease-driving proteins, its current implementations often lack tissue or cell-type specificity, undermining the principles of precision medicine 18. Achieving true selectivity will require the development of degraders that engage context-specific E3 ligases or are activated conditionally—such as by tissue-specific expression, microenvironmental cues, or drug-induced control 19.

To address these limitations, precision-targeted protein degradation (TPD) strategies require the deployment of target- and tissue-enriched E3 ligases, which would enable the selective degradation of pathogenic proteins specifically within diseased cells or tissues 20. This approach offers a promising path to minimize systemic toxicity, a major concern with current degraders that utilize ubiquitously expressed ligases such as CRBN and VHL 22. By coupling substrate targeting E3 ligases with restricted expression patterns, degraders can be engineered to function only in relevant biological contexts, reducing off-target degradation in healthy tissues and enhancing therapeutic precision.

Expanding the use of tissue-enriched and tumor-selective E3 ligases could open new avenues for addressing diseases that have previously been intractable due to ligase incompatibility or systemic toxicity risks 21, 23. For example, ligases with expression enriched in the central nervous system, immune-cell subtypes, or tumor tissues—relative to healthy counterparts—could be leveraged to design degraders that act preferentially in those compartments, thereby minimizing off-target effects in non-diseased tissues 21. Tissue- or context-specific E3 ligases may also offer unique structural motifs, distinct from commonly used ligases like CRBN and VHL, enabling the design of highly selective ligases and novel engagement mechanisms that further enhance the precision of targeted protein degradation 20, 23 .A notable example of the challenges discussed above is CFT7455, a CRBN-recruiting PROTAC currently in clinical trials for Non-Hodgkin's Lymphoma (NHL) and Multiple Myeloma (MM)—malignancies involving B cells and plasma cells, respectively 24. CFT7455 exerts its therapeutic effect by degrading the transcription factors IKZF1 (Ikaros) and IKZF3 (Aiolos), which are over-expressed or mutated in these

hematologic cancers 25. However, its systemic use also leads to the degradation of these targets in non-pathological cell types. This off-target activity manifests as dose-limiting toxicities—including neutropenia, peripheral neuropathy, immunosuppression, and broader hematopoietic suppression—observed in early clinical studies 26. These adverse effects underscore the need for strategies that enable cell-type or tissue-restricted TPD when using E3 ligases with widespread expression.

Reducing the overreliance on a narrow subset of ubiquitously expressed E3 ligases and enhancing tissue specificity of targeted protein degradation (TPD) molecules could offer a strategic advantage in addressing current limitations 27. While most clinically investigated PROTACs recruit ligases such as VHL or CRBN, these global E3 ligases are expressed across nearly all tissue types, increasing the risk of off-target degradation and dose-limiting toxicities due to their essential housekeeping functions 22. Moreover, the widespread exposure of these ligases to degraders may accelerate the emergence of resistance mutations through selective evolutionary pressure 16.

E3 ligases exhibit tissue-enhanced or tissue-restricted expression profiles, offering an untapped opportunity to design tissue-selective degraders 14, 20, 28, 34. These localized E3 ligases often regulate context-specific protein networks or signaling pathways, and are less evolutionarily conserved, making them less susceptible to selective pressure and mutation-driven resistance 28, 29. By leveraging these tissue-specific ligases, it may be possible to expand the functional E3 ligase toolbox, enabling more precise degradation of pathological proteins within disease-relevant tissues, while sparing non-target organs and reducing systemic toxicity 29 12. This approach aligns with the broader goal of advancing TPD as a precision medicine modality 27, 29.

Hence, exploration of the broader E3 ligase landscape becomes imperative to address key challenges in targeted protein degradation, including off-target effects, acquired drug resistance, and the need for tissue-specific degradation 27. The conventional reliance on a narrow subset of ubiquitously expressed ligases, such as CRBN and VHL, limits therapeutic precision and increases the risk of systemic toxicity 22. These ligases, due to their broad expression and essential functions, are also subject to resistance mechanisms under prolonged selective pressure 16.

E3-ligase–substrate interaction (ESI) prediction has become the leading *in-silico* strategy for probing the ubiquitin–proteasome landscape and prioritising ligases for a chosen substrate or biological context. For example, UbiBrowser 1.0 (Wang *et al.*, 2017) fused five evidence layers—domain enrichment, Gene Ontology overlap, co-expression, phylogenetic similarity, and protein-interaction topology—into a Bayesian network that outputs proteome-scale ESI probabilities 98. A new wave of deep-learning approaches followed. E3-targetPred (Published in 2020) represents ligase and substrate sequences using composition of k-spaced amino-acid pairs (CKSAAP) to capture both local and long-range dependencies before mapping them into a discriminative latent space for classification 99. DeepUSI (Published in 2023) instead feeds one-hot-encoded full-length protein sequences into a convolutional neural network that learns interaction motifs directly from raw sequence data, yielding state-of-the-art accuracy on independent benchmarks 100. All three methods ultimately rely on curated repositories such as

UbiBrowser annotations and BioGRID-derived pairs, so their predictive horizon is bounded by the completeness and quality of existing datasets.

E3 ubiquitin ligases span a continuum of expression profiles in human tissues, ranging from ubiquitously expressed to tissue-enhanced and truly tissue-specific proteins 41. Current computational ESI predictors—such as UbiBrowser, E3-targetPred and DeepUSI— have beentrained on global interaction datasets compiled without tissue context, so the resulting scores are inherently tissue-agnostic100,101. This constraint limits their capacity to address safety risks like off-target toxicity or mutagenicity that can arise when degraders recruit ligases with broad physiological distribution 2. Combining tissue-resolved expression atlases with interaction evidence and harnessing the representational power of modern protein language models offers a biologically grounded path toward context-aware ligase ranking that could help mitigate such liabilities (102, 103, 106)

For this purpose, we propose an AI-driven framework that predicts tissue-specific E3 ligases for pathogenically mutated target proteins. This approach is grounded in protein sequence data, with the model trained on billions of evolutionarily and functionally relevant sequence motifs. At its core is a BERT-based protein language model that learns contextual embeddings capturing fine-grained sequence features—such as degrons, post-translational modification sites, and structural propensities—directly from raw amino acid sequences 31. These high-dimensional embeddings are then used to assess the functional compatibility between E3 ligases and target proteins, integrating tissue specificity inferred from co-expression and enrichment patterns across transcriptomic and proteomic atlases.

Our model effectively ranks ligase–target pairs based on predicted binding compatibility, tissue-restricted expression, and the likelihood of forming productive ternary complexes. This enables mutation-selective degradation confined to disease-relevant tissues, while sparing essential proteins and avoiding off-target effects in healthy cells. Ultimately, this sequence-driven, transformer-enabled prediction platform represents a scalable and generalizable solution for rationally expanding the E3 ligase toolbox—unlocking context-specific degraders tailored to the molecular and spatial biology of disease.

Methods

Multiple protein language models (PLMs) such as ESM, ProtT5, and ProGen have emerged as powerful tools for biological sequence modeling, ProtBERT offers a particularly compelling balance of performance, scalability, and usability for protein–protein interaction prediction tasks. Compared to ESM models, which are often larger in size (e.g., ESM-2 models reaching 3 billion+ parameters) and trained on reduced-redundancy datasets like UniRef50, ProtBERT is trained on the larger and more comprehensive UniRef100, capturing greater sequence diversity. This breadth may be advantageous for interaction prediction across distantly related proteins. Furthermore, unlike ProtT5, which uses an encoder–decoder architecture originally designed for generative tasks and is computationally heavier during inference, ProtBERT uses a BERT-style encoder-only transformer, enabling faster embedding generation and more efficient integration into discriminative PPI models 31, 59.

In contrast to models like AlphaFold or ESMFold, which are optimized for structure prediction and often require multiple sequence alignments (MSAs) or template inputs, ProtBERT provides alignment-free, sequence-only embeddings that retain meaningful biochemical and functional signals, making it highly scalable for large-scale or low-resource PPI-prediction tasks. Additionally, ProtBERT's embedding space is well-behaved and transferable, having been evaluated on multiple tasks (e.g., enzyme classification, GO-term prediction, subcellular localization) through benchmarks such as TAPE, which enhances confidence in its generalizability 60 ,53,54,55,56,57,58.

Practically, ProtBERT is also more accessible, with pretrained models available through HuggingFace, allowing straightforward tokenization, embedding extraction, and fine-tuning without specialized infrastructure. Taken together, these characteristics make ProtBERT a technically efficient, biologically grounded, and well-validated choice among PLMs for sequence-based PPI prediction 61.

ELITE Architecture

Our proposed ELITE architecture extends the high-performing Siamese Tailored deep sequence Embedding of Proteins (STEP) framework for generic protein–protein interaction (PPI) prediction, preserving its state-of-the-art performance with minor architectural modification to fit smaller datasets 62 . Instead of initiating a new hyper-parameter search, ELITE adopts the well-validated configurations from STEP 62 . Fine-tuning adheres to STEP's protocol but reorders the training datasets to better reflect ELITE's tissue-specific focus 62. The most significant architectural modification is a classification head-comprising two linear layers, which is streamlined for classification tasks with comparably small tissue specific datasets.

During training, raw protein sequences are processed by a context-aware encoder to produce high-dimensional residue embeddings. Paired embeddings are combined and passed through a two-layer classification head, with the full model optimized to distinguish interacting from non-interacting protein pairs. In the inference phase, both the encoder and classification head are fixed in their trained state: embeddings for unseen protein pairs are generated using the same encoder and then evaluated by the frozen classifier to yield interaction probabilities.

Figure 4 provides a detailed view of the ELITE architecture. Each amino acid in the input protein sequence is converted into a dynamic embedding that captures long-range contextual dependencies, overcoming the limitations of static embeddings and the architectural constraints of RNN-based methods. A pooling layer condenses the sequence of residue embeddings into a single vector representation per protein, using a mechanism adapted from cross-domain architectures. The resulting protein embeddings are combined—typically via element-wise multiplication to generate a joint representation for the protein pair. This PPI embedding is then progressively reduced to a scalar interaction score through a series of dimensionality-reducing linear transformations.

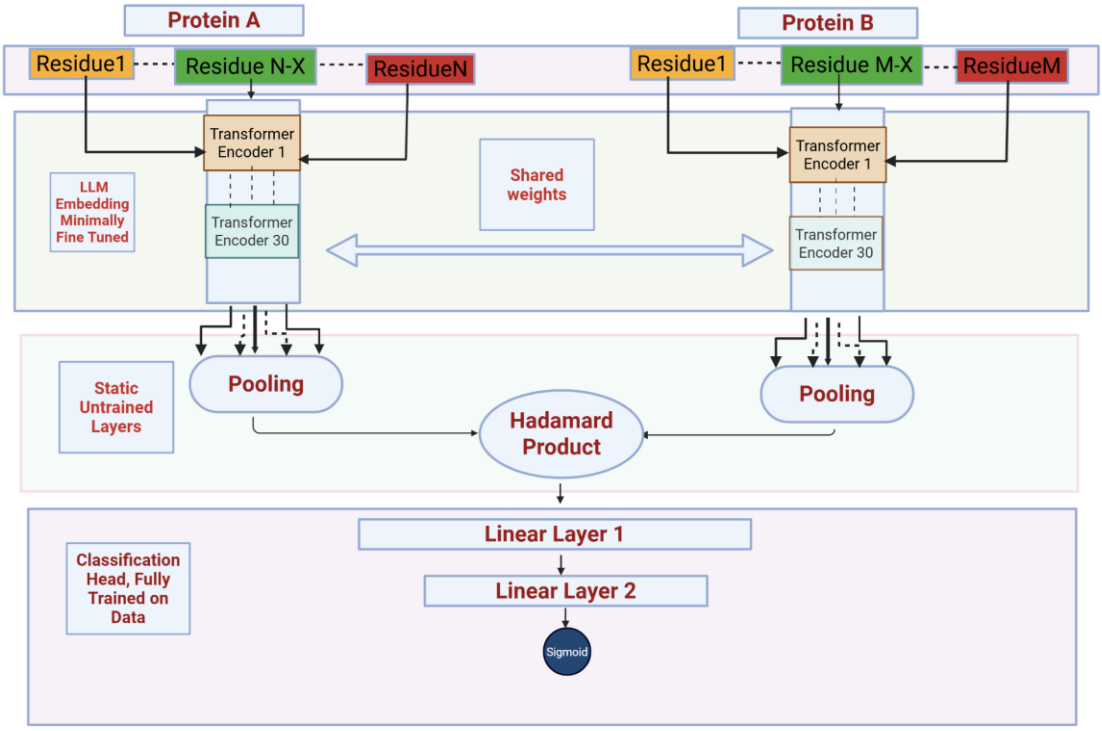

Figure 1 – Proposed ELITE architecture

Tissue Resolved Protein-Protein Interactions

The Stanford Network Analysis Project (SNAP) is a comprehensive initiative that curates and maintains a diverse collection of network datasets spanning multiple domains, including social networks, communication systems, citation graphs, and collaboration networks 35. Within SNAP, the BioSNAP subset serves as a specialized repository for biomedical network datasets, offering structured resources such as tissue-specific gene networks, drug–target interaction graphs, disease–gene associations, protein–protein interaction (PPI) networks, and more 36. BioSNAP provides high-quality, well-documented datasets that support the development and benchmarking of computational methods in network biology, drug discovery, and biomedical data mining.

The tissue-specific gene–gene interaction network utilized in our study was obtained from the BioSNAP collection. This network was originally constructed by the authors using a two-step approach involving the integration of tissue-specific gene expression profiles with a general PPI network to generate tissue-contextualized gene–gene associations 37. We used this publicly available, pre-constructed network as the foundation for our analysis. The dataset obtained from the above procedure comprises a large number of small tissue-specific gene–gene interaction networks (144 in total). To mitigate the partially small number of interactions per network, we first aggregated tissues into biologically meaningful groups, thereby reducing the number of datasets while enhancing the statistical and functional robustness of downstream analyses. We strategically grouped the tissues based on criteria such as biological function, anatomical location,

developmental origin, molecular signatures, and disease relevance, resulting in 26 consolidated tissue clusters (see Supplementary Table 1 for details). The E3 Ligase Substrate Interactions were filtered by mapping the E3 Ligases as one of the interacting partners and removing duplicates. In the resulting networks, E3–substrate interactions (ESIs) are treated as tissue-specific, reflecting the distinct regulatory environment within each tissue. However, the interacting entities—the E3 ligases and substrate proteins—may themselves exhibit broader expression patterns and are not necessarily confined to a single tissue. Dataset-level statistics are reported in Table 1.

Sl No.	Name	No. of ESI	**No. of E3 Ligases**	**No. of Substrates**	**E3L Size Distr**	**Subs Size Distr**
1	Cardiac	522	98	419	(510,34350,70)	(604,4911,84)
2	Adrenal	364	82	301	(521,3159,95)	(557,4911,84)
3	Bone	420	83	347	(509,3159,95)	(568,4911,84)
4	Brain	541	97	437	(510,4646,70)	(574,4911,84)
5	Renal	223	101	429	(515,5202,70)	(574,4911,84)
6	PNS	187	63	165	(510,2843,105)	(574,4374,84)
7	Immune	502	93	405	(509,4646,70)	(574,4911,84)
8	Auditory	25	10	25	(464,2414,113)	(875,1512,309)
9	Blood	494	90	402	(508, 4646, 70)	(574,4911,84)
10	Reproductive Female	453	90	367	(520,5202,70)	(574,4911,84)
11	Mammary	298	59	253	(524,3159,101)	(568,4911, 303)
12	Endocrine	377	83	311	(515,3159,95)	(568,4911, 84)
13	Fetal	531	94	431	(508,5202,70)	(592, 4911, 84)

14	Gastrointestinal	540	101	436	(510,5202,70)	(604,4911,84)
15	Hepatic	164	60	142	(531,2843,124)	(562,4374,84)
16	Oral	334	79	279	(510,5202,95)	(568,4374, 84)
17	Pulmonary	531	99	419	(510,5202,70)	(592,4911, 84)
18	Reproductive Male	501	94	403	(510, 5202,70)	(574,4911,84)
19	Skeletal Muscle	476	92	381	(513,4646,70)	(570,4911,84
20	Skin	404	84	333	(510,3159,95)	(568,4911,84)
21	Spinal Cord	542	97	437	(510,4646,70)	(574,4911,84)
22	Thyroid	373	81	301	(520,5202,95)	(568,4911,84)
23	Urinary Tract	237	50	205	(511,5202,101)	(568,4911,309)
24	Vascular	352	79	295	(510,2843,95)	(570, 4374,84)
25	Vision	378	84	311	(510,5202,95)	(568,4374,84)
26	Connective	399	86	322	(570,4911,84)	(510,3159,95)

Table 1 – Details of Tissue specific Protein Interaction data

As part of the data-preparation workflow for machine learning, the tissue-specific gene–gene interaction networks were converted into binary protein-protein interaction (PPI) networks using Entrez-to-UniProt ID mappings. Because individual genes can encode multiple protein isoforms, each gene–gene interaction expands into several candidate PPIs, producing a protein-level network roughly in an order of magnitude larger than the original gene network. Many of these pairs are putative, arising from combinatorial mapping and lacking direct experimental validation. A subsequent filtering step described below removes biologically implausible or weakly supported interactions, improving data confidence.

To identify experimentally supported interactions within the hypothetical tissue-specific gene networks, it is necessary to map them against curated protein–protein interaction (PPI) datasets derived from experimental studies. Several well-established PPI repositories are available for this purpose, including the Biological General Repository for Interaction Datasets (BioGRID) 46, Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) 47, 40, 38 IntAct 48, Molecular INTeraction Database (MINT) 49, and the Human Protein Reference Database (HPRD) 39. Each of these databases contains hundreds of thousands of PPIs but differs in data format, confidence in scoring systems, curation standards, and completeness. Integrating these heterogeneous datasets poses significant challenges due to the lack of standardized data models, inconsistent identifiers, and contradictory interaction evidence arising from methodological differences and varying experimental conditions. The integration process is therefore resource-intensive and falls outside the scope of this study. As an alternative, we explored the use of pre-integrated and standardized PPI datasets, which offer harmonized formats, unified confidence scoring, and broader interaction coverage-providing a practical solution for mapping hypothetical interactions to experimentally validated protein-protein interactions.

The Agile Protein Interaction DataAnalyzer (APID) is a comprehensive and high-quality PPI database constructed by integrating data from several widely used repositories, including BIND 50, DIP 51, HPRD 39, IntAct 48, and MINT 49. APID adheres to the Proteomics Standards Initiative – Molecular Interaction (PSI-MI) format 52, incorporating detailed metadata such as interactor information, interaction evidence, PubMed references (PMIDs), confidence scores, biological roles, and interaction types (e.g., direct, indirect, complex). Experimental evidence in APID is derived from multiple methods: high-throughput techniques such as yeast two-hybrid (Y2H) and affinity-purification mass spectrometry (AP-MS) contribute 50-60% of entries, biochemical methods like co-immunoprecipitation (Co-IP) and pull-down assays account for 20-25%, structural and high-resolution techniques for 10-15%, and fluorescence-based assays and biophysical methods (e.g., single-molecule microscopy, isothermal titration calorimetry) each provide 5-10% of the total entries 42. The APID database is structured into three quality tiers: Levels 0-2. APID Level 2 (the highest level) was chosen for further filtering. This Level includes 22,707 proteins and 154,955 interactions 42. This curated, experimentally validated interaction dataset serves as the primary source for downstream transformation, filtering, and modeling.

To assemble the training corpus, we intersected each tissue-specific gene network with APID Level 2, yielding 26 tissue-resolved E3–substrate datasets that constituted the positive class. For the negative class, we adopted the rigorously curated Negatome 2.0 non-interaction set released by Hamp and Rost, which minimizes label noise by excluding protein pairs with any experimental or structural evidence of interaction 43. All interacting partners were mapped to canonical sequences via the UniProt Knowledgebase 44, producing tissue-specific sequence pairs labeled as interacting or non-interacting. These pairs formed the input to our machine learning pipeline. Figure 4 summarizes the complete data-engineering workflow and introduces the modeling architecture described in subsequent sections.

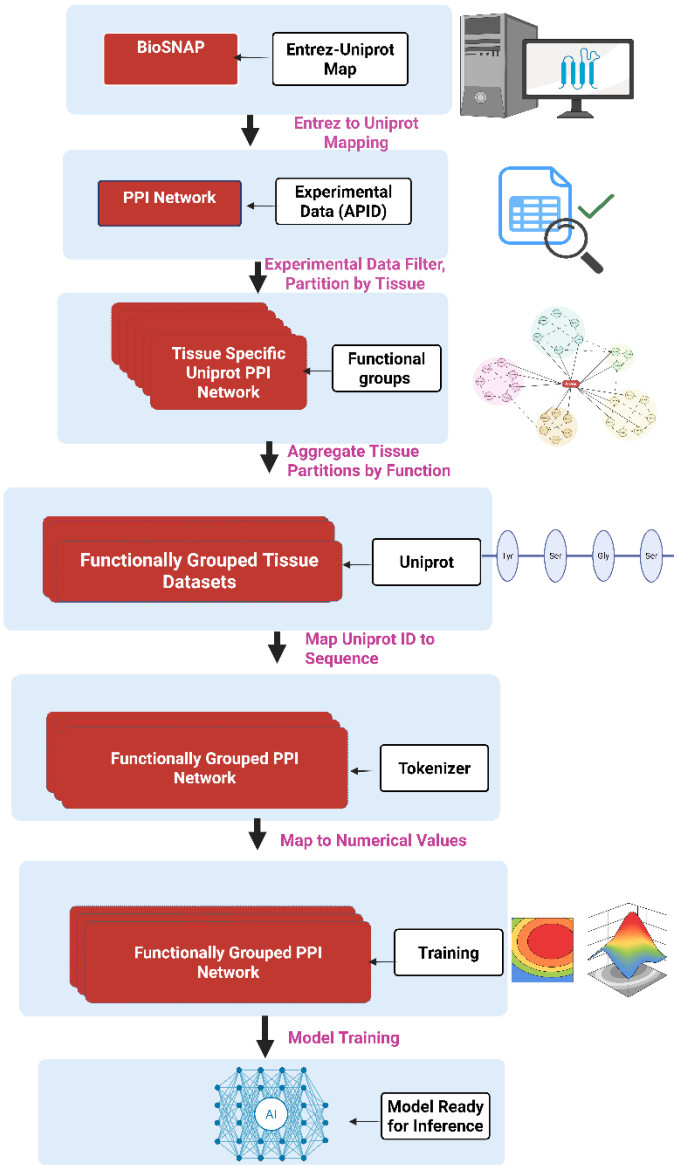

Figure 2 - Data Engineering and ML Flow

Model Training and Evaluation

Prior ESI predictors are typically trained from scratch on engineered datasets. ELITE follows the same overarching objective while fine-tuning a pretrained ProtBERT backbone with a classification head on tissue-specific datasets. For evaluation, we report AUPR and AUROC, aligning with common practice to ensure consistency with the literature and robust, comparable performance assessments.

We report both aggregate and point-wise metrics to gauge the model's ability to detect positives, whether common or rare. In imbalanced data such as ours, Precision-Recall Area Under the

Curve (PR-AUC), precision, and recall are especially informative, because together they impose tight control over false positives and false negatives across decision thresholds 63. By contrast, a model can attain a high Receiver Operating Characteristics Area Under the curve (ROC-AUC) even while producing many false positives or false negatives, so ROC-AUC alone would overstate performance 64.

We evaluated the model using stratified 10-fold cross-validation. The full dataset was divided into ten equal-sized folds that each preserved the original class distribution. In every iteration, nine folds were used for training while the remaining fold served as the validation set, ensuring that every sample was tested exactly once. This approach maintains class balance within each split, reduces variance from sampling error, and yields a more reliable and less biased estimate of real-world performance than a single hold-out split or non-stratified k-fold validation 65 45 66 67 68. Specific training details are provided in the Supplementary Material Training section.

Making Model Predictions Explainable

Integrated Gradients (IG) (88, 89, 90) attributes a model's output F(x) to each feature by integrating the gradient along the straight path from a baseline (neutral) to the input. The mathematical formulae is shown in Eq 1.

$$\mathrm{IntegratedGrad}_i(x) = (x_i - x_i') \times \int_0^1 \frac{\partial F\big(x' + \alpha(x - x')\big)}{\partial x_i} d\alpha$$

Equation 1 – Mathematical formulae for Integrated Gradient

Integrated Gradients (IG) was used to elucidate which protein domains and individual residues most strongly drive our BERT-based PPI predictions. IG's proven success on biological sequence models provides robust validation for high-confidence interaction calls 81. Crucially, IG adheres to its two foundational axioms in a biologically meaningful way: it assigns nonzero attributions to any residue whose isolated perturbation changes the interaction score-thereby pinpointing known "hot-spot" amino acids (Sensitivity)-and it produces identical attribution maps across different transformer architectures that yield the same PPI outputs, whether ProtBERT, ESM-1b, or a custom fusion head is used (Implementation Invariance).

Results and Use Cases

Cross-Validated Prediction Performance

Table 2 summarizes the ten-fold stratified cross-validation results across tissues. The model attains consistently high ROC-AUC and PR-AUC scores, demonstrating not only strong class separation but also reliable prioritization of the positive class-i.e., low false-positive and false-negative rates. The sole outlier is the auditory tissue, where ROC-AUC is poor. This is attributable to an insufficiently small training set that cannot adequately supervise fine-tuning. However, even in smaller datasets like the one discussed, the model reports a hig

h PR-AUC showing a high precision for the small number of positive predictions.

Hepatic tissue also shows comparatively modest ROC-AUC values, again reflecting limited sample sizes. Overall, a trend emerges: larger tissue-specific datasets yield proportionally higher metrics, underscoring the model's capacity to scale rapidly with data volume and approach maximal performance.

Sl.	Tissue	ROC-AUC	PR-AUC	Precision	Recall	Accuracy
1	Cardiac	92±2	98	95±1	98±1	94±1
2	Adrenal	93±3	99±1	96±1	98±1	95±1
3	Bone	93±2	98	96±1	98±1	94±1
4	Brain	91±2	98±1	94±1	98	93±1
5	Renal	90±4	98±1	94±1	98±1	93±1
6	PNS	90±5	97±2	93±1	98±1	92±1
7	Immune	90±5	98±1	94±1	98±1	93±1
8	Auditory	49±23	89±5	90±9	55±34	55±27
9	Blood	92±3	98±1	95±1	98	94±1
10	Reproductive F	94±2	98±1	95±1	98	94
11	Mammary	96±1	99±1	96	97±2	95±1
12	Connective	95±2	99	95±1	98	94±1
13	Endocrine	95±2	99±1	96±1	97±1	94±1
14	Fetal	93±2	98	95	97	93
15	Gastrointestinal	91±2	98	93±1	98±1	92±1
16	Hepatic	81±5	97	91±1	99	91±2
17	Oral	95±1	98±2	96	98	96±1
18	Pulmonary	85±12	96±2	93±2	98±1	92±2
19	Reproductive M	94±2	98±1	95±1	98	94

20	Skeletal	91±4	98±1	94±1	98±1	93±1
21	Skin	95±2	99	96	98±1	95±1
22	Spinal Cord	92±4	98	94±1	98±2	93±1
23	Thyroid	96±1	99±1	96±1	98±1	95±1
24	Urinary Tract	92±3	97±2	95±1	97±1	93±1
25	Vascular	95±2	99	96±1	98±1	95
26	Vision	94±2	98±1	95±1	99	95±1

Table 2 - Results of 10-Fold CV, all values are percentages

Benchmarking Against Alternative Methods

We rigorously benchmarked ELITE against E3TargetPred as an alternative and previously published method for E3-ligase prediction, using PR-AUC as the main performance metric. E3TargetPred was executed from the public codebase using the repository's initial configuration on a single stratified 90:10 train: test split (positive: negative = 1:9).

E3TargetPred uses CKSAAP sequence features for each protein and learns a joint autoencoder-plus-classifier in which the latent embedding must both reconstruct a pair feature vector from the two singles and separate true E3–target pairs from non-pairs. Three key hyperparameters shape behavior: k (CKSAAP gap) tunes sequence context granularity; LV (latent dimensionality) sets embedding capacity and interpretability; and λ balances reconstruction fidelity versus discriminative separation, trading expressive detail against class compactness in the latent space.

As shown in Table 3, ELITE consistently exceeds E3TargetPred across tissues on the primary metric, a pattern that may be explained by the broader representational capacity of protein language-model encoders relative to conventional architectures. The statistical significance of performance differences was confirmed via a Wilcoxon signed rank test with a p-value that was numerically zero.

Sl No.	Name	AUPR ELITE	AUPR E3TARGEPRED
1	Cardiac	98	91
2	Adrenal	99	95
3	Bone	98	85
4	Brain	98	95
5	Renal	98	85
6	PNS	97	77
7	Immune	98	87
8	Blood	98	59
9	Reproductive F	98	92
10	Mammary	99	87
11	Connective	99	55
12	Endocrine	99	94
13	Fetal	98	84
14	Gastrointestinal	98	39
15	Hepatic	97	82
16	Oral	98	59
17	Pulmonary	96	89
18	Reproductive M	98	58
19	Skeletal	98	72
20	Skin	99	33
21	Thyroid	99	09
22	Urinary Tract	97	04
23	Vascular	99	84

24	Vision	98	84

Table 3 – Single test comparison with E3TargetPred (Gap =0, Latent Size =3, Train:Test 9:1)

Example Application 1: Targeting an Intractable Pathogenic Transcription Factor in Gastrointestinal Cancers

Transcription factors (TFs) are challenging targets. Wild-type p53 regulates the expression of >500 genes controlling cell-cycle arrest, DNA repair, and apoptosis, and its loss-of-function mutations drive many cancers 74. Likewise, NF-κB governs transcription of 500-1000 genes implicated in inflammation, immune evasion and tumor progression 75. Despite their central roles, classical TFs lack accessible ligand-binding pockets, and no FDA-approved therapies yet directly disrupt protein-DNA interactions of p53, NF-κB, or similar TFs.

Despite their central role in disease-relevant pathways, transcription factors have historically been considered undruggable due to their lack of deep binding pockets, high structural flexibility, and primarily nuclear localization 84, 85. While small molecules have successfully targeted members of the nuclear receptor superfamily (a structurally distinct subset of TFs with ligand-binding domains), no FDA-approved therapies directly inhibit classical transcription factors like p53 or NF-κB at the protein-DNA interaction level 86, 87.

Mutated p53 is one of the most extensively studied targets in cancer biology, owing to its pivotal tumor-suppressive functions and its status as the most frequently mutated gene across human cancers 76,77 . In its wild-type form, p53 is ubiquitously expressed and plays a critical role in maintaining genomic stability. Upon cellular stress or DNA damage, p53 activates the transcription of genes involved in DNA repair; if repair mechanisms fail, it initiates apoptosis through the induction of pro-apoptotic genes 76. Over the past five decades, p53 has been the focus of intense research, leading to the identification of 3124 distinct variants 77. These variants are classified based on predicted clinical significance into categories such as pathogenic, likely pathogenic, benign, likely benign, or of uncertain significance.

In this study, we focus on a specific p53 variant (R175H), which not only loses its native tumor-suppressive function but also acquires novel oncogenic properties that contribute to increased tumor aggressiveness 78. Such gain-of-function mutations in p53 represent promising targets for therapeutic intervention, particularly through targeted protein-degradation strategies. Accordingly, we propose hypotheses regarding tissue-specific E3 ligases that are predicted to exhibit high binding affinity toward the mutant p53 protein, thereby enabling selective degradation. Gastrointestinal (GI) cancers - encompassing malignancies of the esophagus, stomach, small and large intestines, liver, and pancreas—represent some of the most lethal cancer types. Collectively, GI cancers account for approximately 26% of all cancer diagnoses and contribute to 35-90% of cancer-related mortality, underscoring the urgent need for effective, precision-targeted therapies 79.

SN.	Uniprot ID	Score	RNA Seq HPA	RNA Seq GTEx	FANTOM	Sc RNA Cell Type
1	Q9UKV5	99	GI	GI	GI	GI
2	O00623	99	GI	Non-Specific	Non-Specific	Non-Specific
3	Q5T447	98	GI	GI	GI	GI
4	O00237	93	GI	GI	GI	GI
5	Q86UW9	90	GI	GI	NA	GI
6	Q9ULV8	90	GI	GI	GI	GI

Table 4 - Predictions - Ranked E3Ls in p53 mutation in GI Cancer

Sl No.	E3 Ligase Uniprot ID	Entrez Gene ID	Official Name
1	Q9UKV5	AMFR	E3 ubiquitin-protein ligase AMFR
2	O00623	PEX12	Peroxisome assembly protein 12
3	Q5T447	HECTD3	E3 ubiquitin-protein ligase HECTD3
4	O00237	RNF103	E3 ubiquitin-protein ligase RNF103
5	Q86UW9	DTX2	Probable E3 ubiquitin-protein ligase DTX2
6	Q9ULV8	CBLC	E3 ubiquitin-protein ligase CBL-C

Table 5 – Details of Ligases in Table

Figure 5 and Figure 6 map residue-level attribution scores for the interactions between the E3 ligases RNF103 and DTX2 and the cancer-associated p53 R175H variant, with red and green intensities marking residues that respectively promote or hinder binding. Both ligases show strong reliance on p53's intrinsically disordered, proline-rich stretch spanning residues 60 - 85, a region known to interact with HRMT1L2, WWOX, and CCAR2 109 110, 111 - whereas the structured β-strand at residues 250-270, which normally accommodates partners such as HIPK1, ZNF385A, and AXIN1 112 ,113 , 114 shows strongly negative attribution. This reflects a general principle: intrinsically disordered regions (IDRs) enable multivalent, adaptive interactions, while structured domains impose spatial and chemical constraints that may favor or disfavor specific partners depending on compatibility 115 .

The model assigns a positive attribution to the trans-membrane segment of RNF103 (residues 320–330), which serves as an ER anchor that correctly orients the ligase for ubiquitinating cytosolic substrates such as p53 116. In DTX2, substrate recognition relies on both flexible and ordered modules: the positively scored intrinsically disordered linker (residues 350–390) can adapt to the proline-rich 60–85 stretch of p53 via short linear motifs 117, whereas the C2H2-type zinc-finger (residues 412–473) offers a well-folded surface that complements p53 features and reinforces binding specificity 118. Together these elements illustrate how IDRs supply adaptability while structured domains provide stabilizing contacts 115 .

By contrast, the negative attribution assigned to the acidic, disordered loop of RNF103 (residues 550–570) likely reflects an autoinhibitory mechanism: in many RING-type E3 ligases, an internal acidic segment can fold back over the RING domain, shielding catalytic cysteines or the E2-binding groove until a regulatory signal-often phosphorylation-displaces it 119,120. In such a closed conformation the loop would dampen, rather than promote, p53 ubiquitination-explaining its negative weight in the attribution map.

Integrated-gradients indicate that RNF103 and DTX2 favor disordered, flexible regions of p53 that enable adaptable engagement. While ordered, structured segments are penalized when their geometry/chemistry mismatches the ligase surface. The ligases appear to use two-mode strategy-flexible elements to initiate capture and folded modules to stabilize specificity. Negative attributions map to autoinhibitory features, consistent with regulated activation of E3 activity rather than constitutive binding.

Figure 3 - Interpretation of O00237 and p53 R175H

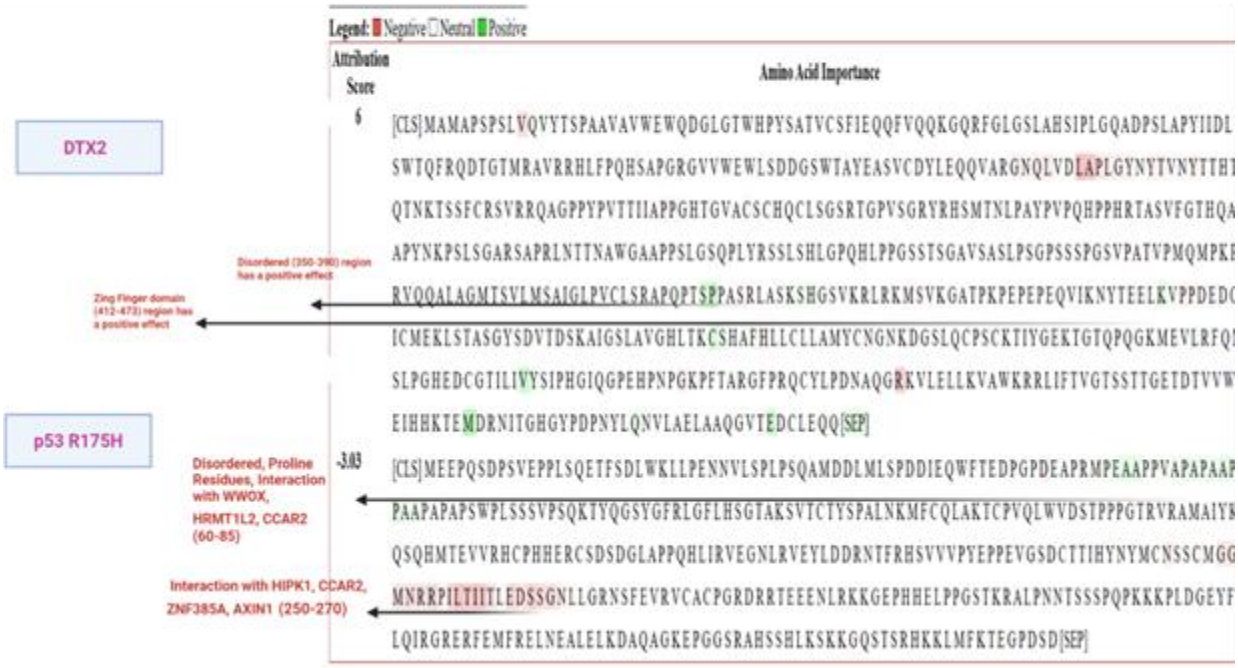

Figure 4 - Interpretation of Q86UW9 (DTX2) and p53 R175H

Example Application 2: Eliminating an Undruggable Pathogenic Structural Protein in Neurodegenerative diseases

Structural proteins are a small yet crucial class of potential drug targets (~2-3% of approved drugs). Their poor druggability derives from a lack of deep pockets, ubiquitous expression, and the risk of severe side-effects when perturbed. Paclitaxel illustrates a successful exception, binding tubulin to stabilize microtubules and arrest mitosis in cancer therapy [69] [70] [71] [72]. Another structural protein, Tau (encoded by MAPT), is essential for microtubule stability in neurons. The P301L Tau variant disrupts the microtubule-binding domain, exposes hydrophobic regions that nucleate neurofibrillary tangles, and underlies frontotemporal dementia (FTD) [73].

Figure 5 - Interpretation of O95155 (UBE4B) and Tau

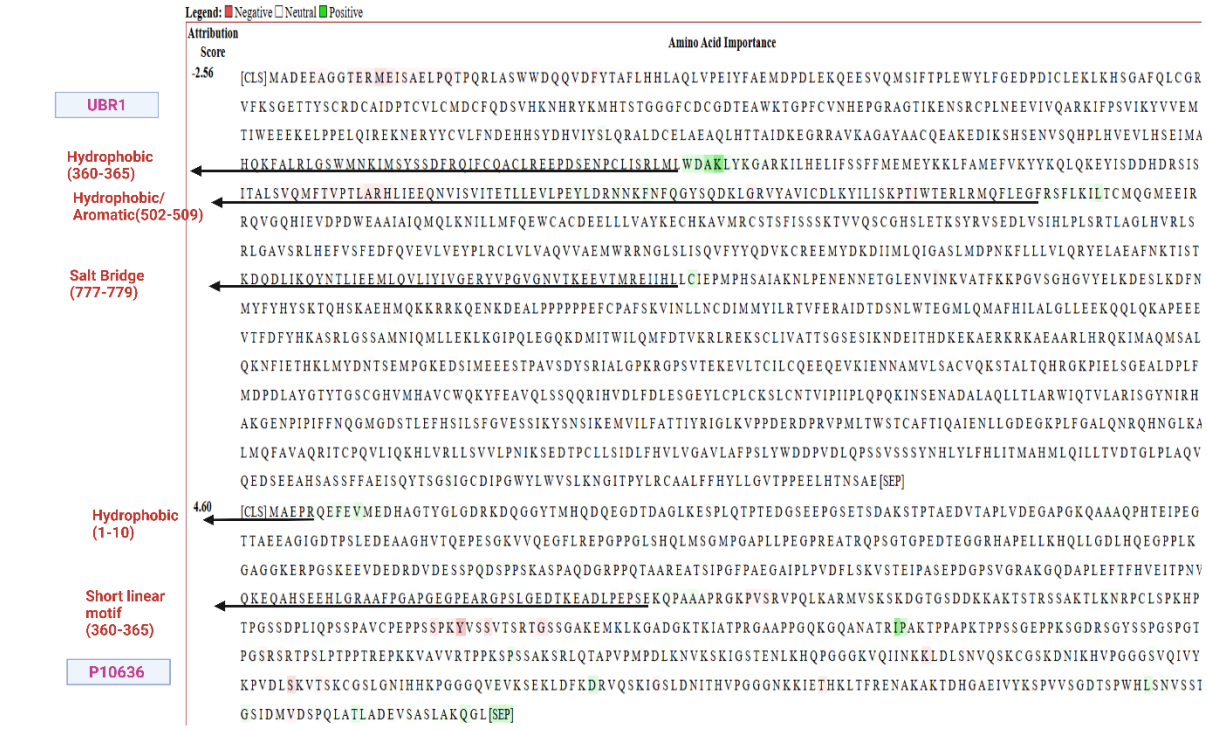

Figure 6 - Interpretation of Q8IWV7 (UBR1) and Tau

Neurodegenerative diseases represent a major and growing public health concern, with rising incidence rates—particularly in Western countries. Among these disorders, Alzheimer's disease and frontotemporal dementia (FTD) together account for approximately 80% of cases, with the microtubule-associated protein Tau playing a central pathological role in both 91. A well-characterized pathogenic variant, Tau P301L, has been strongly implicated in the development of both Alzheimer's disease and FTD, where it promotes neurodegeneration through enhanced aggregation, reduced microtubule binding, and increased cytotoxicity 80,92, 93,94 ,95, 96 ,97 17. In this study, our model predicts a ranked list of E3 ligases that may participate in tissue-specific interactions involving mutant Tau as shown in Table 5 , this information can be used for drug discovery.

It is important to note that while the predicted protein–protein interactions are tissue-specific, this does not necessarily imply that both the E3 ligase and the substrate (e.g., Tau P301L) are tissue–specific. It is important to acknowledge that many of the E3 ligases are not inherently tissue specific 41,18. Therefore, we conduct an independent tissue-specificity analysis of the E3 ligases, drawing from multiple data sources to determine the target tissue of the predicted ligase, as discussed in subsequent sections 82, 83.

In line with the p53 use case, we employed Integrated Gradients (IG) to explain the model's predicted interactions of Tau with UBR1 and UBE4B. The IG analysis indicates that the model's predictions are grounded in sound biophysical principles. Specifically, the hydrophobic and polar amino acids located at positions 6-10 (121) in Tau form a canonical hotspot interaction motif that contributes significantly to the free energy of binding. This is followed by short linear motifs at positions 360-365 (122), which play a key role in interactions with both ligases under discussion.

A similar hydrophobic and polar amino acid pattern is observed in UBE4B at positions 444-448 (122), with additional hydrophobic contacts at positions 478-479 (123). Polar and charged groups at positions 494-497 (124) further suggest hydrophobic packing and potential salt bridge formation in UBE4B–Tau interactions. The negative attribution observed in UBE4B across positions 1-40 aligns with the presence of uncompensated electrostatic interactions and steric hindrance (125).

For UBR1, hydrophobic and aromatic residues at positions 360-365 (121) and 502–509 are predicted to contribute to Tau binding through tight van der Waals packing and electrostatic interactions involving R/K residues (126), collectively enhancing complex stability. Finally, polar residues at positions 777-779 are implicated in forming electrostatic and hydrogen bonds with complementary Tau residues, providing additional stabilization to the UBR1–Tau complex (124).

Table 4, Table 5, Table 6 and Table 7 summarize the predicted E3-ligase candidates-reporting their interaction scores, sub-cellular localizations, and genomic annotations for gastrointestinal cancer and Alzheimer's disease, respectively. These predictions are potential hypotheses for further validation.

Sl No.	E3 Ligase Uniprot ID	Score	Entrez Gene ID	Name
1	Q8IWV7	0.99	UBR1	Ubiquitin protein ligase E3 component n-recognin 1
2	O95155	0.99	UBE4B	Ubiquitination factor E4B
3	Q6ZT12	0.99	UBR3	UBR3
4	Q13233	0.96	MAP3K1	MAP3K1

Table 6 - Predictions - Ranked E3Ls in Tau mutation in Alzheimer's disease

SN.	Uniprot ID	RNA Seq HPA	RNA Seq GTEx	FANTOM	Sc RNA Cell Type
1	Q8IWV7	Non-Specific	Non-Specific	Non-Specific	Brain
2	O95155	Non-Specific	Brain	Brain	Brain
3	Q6ZT12	Brain	Brain	Brain	Brain
4	Q13233	Non-Specific	Non-Specific	Non-Specific	Brain

Table 7 - Tissue Categorization based on Bulk/Single RNA Seq of E3 Ligases

Discussion

ELITE harnesses the representational power of large protein language models (PLMs) to capture intricate sequence patterns and transfer them effectively to downstream tasks. The dense, semantically rich embeddings generated by these PLMs not only separate interacting from non-interacting pairs with high fidelity but also prioritize sensitive detection of the positive class-i.e., true protein–protein interactions. Built on the rigorously validated STEP framework, ELITE inherits state-of-the-art performance that has been benchmarked comprehensively against existing methods across multiple PPI datasets.

The shared-weight Siamese network encodes each protein independently within its tissue context and combines their representations into a single product tensor, enabling the capture of highly non-linear sequence dependencies that govern complex and poorly characterized protein–protein interactions. Integrated Gradients applied to high-confidence predictions reveal key residue-level features contributing to interaction likelihood, with results supported by existing literature. With a strong theoretical foundation, interpretable outputs, and a modular design, this architecture offers a powerful and adaptable framework that can be efficiently fine-tuned to tissue-specific or custom PPI datasets.

As a demonstrative use case, we deployed our model to identify tissue-specific E3 ubiquitin ligases suitable for PROTAC design against pathogenic variants of two classically intractable proteins, p53 and Tau, which drive gastrointestinal cancers and neurodegenerative disorders

such as Alzheimer's disease and frontotemporal dementia. The model produced high-confidence ligase rankings for each disease-relevant tissue, and Integrated Gradients attribution traced the sequence features underpinning every prediction, affording mechanistic transparency. These ligase–substrate hypotheses constitute strong leads for experimental validation and lay the groundwork for variant-targeted PROTAC therapeutics.

Beyond the scope of the study, it would also be interesting to explore variant selectivity by evaluating whether mutations in the target expose or alter motifs that render the mutant—but not the wild-type—amenable to ligase engagement and degradation 32.

To our knowledge, this is the first study to leverage large language models for E3 ligase selection in targeted protein degradation, offering an innovative solution to longstanding hurdles of sparse data, biochemical complexity, and limited model interpretability. The architecture's modular design and flexible classifier allow it to accommodate datasets of widely varying sizes and noise, enabling straightforward adaptation to new targets and experimental contexts.

Resource availability

Lead contact

Further information and requests for code and data should be directed to and will be fulfilled by the lead contact, Holger Fröhlich (holger.froehlich@scai.fraunhofer.de).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Code and data are located at: https://github.com/sabyasachi2k13/ELITE.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

Declaration of interests

The authors declare no competing interests related to this study.

Author contributions

Conceptualization, H.F.; Methodology, H.F. and S.P.; Data Curation, Formal Analysis, Visualization, Investigation, Validation, S.P.; Technical Help: S.M; Supervision, H.F; Writing—Original Draft, S.P.; Writing—Review and Editing, S.P and H.F.

Acknowledgments

We thank André Gemünd for his support regarding the computational infrastructure of Fraunhofer SCAI.

Supplementary Materials

Document S1

Declaration of generative AI and AI-assisted technologies in the writing process

The authors utilized OpenAI's GPT-3o only to enhance the clarity of the text and thoroughly reviewed the content to ensure its scientific accuracy.

References

1. Zhong G., Chang X., Xie W., et al. Targeted protein degradation: advances in drug discovery and clinical practice. Signal Transduction and Targeted Therapy 9, 308 (2024).
2. Békés M., Langley D.R., Crews C.M. PROTAC targeted protein degraders: the past is prologue. Nature Reviews Drug Discovery 21, 181–200 (2022).
3. Davis C., Spaller B.L., Matouschek A. Mechanisms of substrate recognition by the 26S proteasome. Current Opinion in Structural Biology 67, 161–169 (2021).
4. Li X., Pu W., Zheng Q., et al. Proteolysis-targeting chimeras (PROTACs) in cancer therapy. Molecular Cancer 21, 99 (2022). Dikic I., Robertson M. Ubiquitin ligases and beyond. BMC Biology 10, 22 (2012).
5. Komander D., Rape M. The ubiquitin code. Annual Review of Biochemistry 81, 203–229 (2012).
6. Hershko, A., Ciechanover, A., Heller, H., Haas, A. L., & Rose, I. A. (1983).“Components of ubiquitin-protein ligase system. Resolution, affinity purification, and role in protein breakdown.” *Journal of Biological Chemistry*, 258(13), 8206–8214.
7. Ciechanover A., Heller H., Katz-Etzion R., Hershko A. (1982), JBC — “Covalent affinity purification of ubiquitin-activating enzyme (E1)” First clear purification/characterization of E1, showing ATP-dependent activation and thiol-ester linkage—establishes the entry step of the pathway.
8. Ciechanover A., Finley D., Varshavsky A. (1984), Cell — “Ubiquitin dependence of selective protein degradation demonstrated in the mammalian cell cycle mutant ts85”
9. Bachmair A., Finley D., Varshavsky A. (1986), Science — “In vivo half-life of a protein is a function of its amino-terminal residue” (the N-end rule)

10. Burslem G.M., Crews C.M. Proteolysis-Targeting Chimeras as Therapeutics and Tools for Biological Discovery. Cell 170, 638–654 (2017).

11. Bondeson D.P., Mares A., Smith I.E., et al. Catalytic in vivo protein knockdown by small-molecule PROTACs. Nature Chemical Biology 11, 611–617 (2015).

12. Jeon Y.G. et al. “Ubiquitin ligase RNF20 coordinates sequential adipose thermogenesis with brown- and beige-fat-specific substrates.” *Nature Communications* 15, 940 (2024).

13. Fischer E.S., Böhm K., Lydeard J.R., et al. Structure of the DDB1–CRBN E3 ubiquitin ligase–thalidomide substrate complex. Nature 512, 49–53 (2014).
14. Li W., Bengtson M.H., Ulbrich A., et al. Genome-wide and functional annotation of human E3 ubiquitin ligases. BMC Genomics 9, 227 (2008).

15. Schapira M., Calabrese M.F., Bullock A.N., Crews C.M. Targeted protein degradation: expanding the toolbox. Nature Reviews Drug Discovery 18, 949–963 (2019).

16. Hanzl A., Casement R., Imrichova H., et al. Functional E3 ligase hotspots and resistance mechanisms to small-molecule degraders. Nature Chemical Biology 19, 323–333 (2022).

17. Götz J., Chen F., van Dorpe J., Nitsch R.M. “Formation of neurofibrillary tangles in P301L tau transgenic mice induced by Aβ42 fibrils.” *Science* 293, 1491–1495 (2001).

18. Liu Y., Yang J., Wang T., et al. Expanding PROTACtable genome universe of E3 ligases. Nature Communications 14, 6509 (2023).

19. Yim J., Park J., Kim G., et al. Conditional PROTAC: Recent Strategies for Modulating Targeted Protein Degradation. ChemMedChem 19, e202400326 (2024).
20. Paik H., et al. ELiAH: the atlas of E3 ligases in human tissues for targeted protein degradation with reduced off-target effect. Database (Oxford) (2024).

21. Kuemper S., Cairns A.G., Birchall K., Yao Z., Large J.M. Targeted protein degradation in CNS disorders: a promising route to novel therapeutics. Frontiers in Molecular Neuroscience 17, 1370509 (2024).

22. Shi J., Wang L., Zeng X., et al. Precision-engineered PROTACs minimize off-tissue effects in cancer therapy. Frontiers in Molecular Biosciences 11, 1505255 (2024).

23. Kubryń N., Fijałkowski Ł., Nowaczyk J., et al. PROTAC technology as a new tool for modern pharmacotherapy. Molecules 30 (10), 2123 (2025).

24. Study to Assess the Safety and Tolerability of CFT7455 in Relapsed/Refractory Non-Hodgkin’s Lymphoma or Multiple Myeloma (ClinicalTrials.gov Identifier NCT04756726), University of California San Francisco (accessed 2025).

25. Berdeja J., et al. “A Phase 1 Study of CFT7455, a Novel Degrader of IKZF1/3, in Multiple Myeloma and Non-Hodgkin Lymphoma.” Blood 138 (Suppl 1): 1675 (2021).

26. Conroy R. “Cemsidomide (CFT7455) Plus Dexamethasone Elicits Responses in Multiple Myeloma.” Cancer Network (May 8 2025).

27. Kannt A., Đikić I. Expanding the arsenal of E3 ubiquitin ligases for proximity-induced protein degradation. Cell Chemical Biology 28, 1014–1031 (2021).

28. Liu J., Zhao Q., Peng C., et al. Tissue-enriched E3 ubiquitin ligases: emerging targets for precision therapeutics. Trends in Cell Biology 34, 45–61 (2024).

29. Zhang L., Kirubakaran P., Nguyen T.A., et al. Designing tissue-selective PROTACs by harnessing tissue-restricted E3 ligases. Nature Chemical Biology 20, 112–120 (2024).

30. Rives R., Meier J., Sander C., et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences USA 118, e2016239118 (2021).

31. Elnaggar A., Heinzinger M., Dallago C., et al. ProtTrans: toward cracking the language of life’s code through self-supervised deep learning and high-performance computing. IEEE Transactions on Pattern Analysis and Machine Intelligence 44, 7112–7131 (2022).

32. Li G., Goodwin C.M., Kintzer A., et al. DeepLigase: a machine-learning framework for predicting E3 ligase–substrate interactions from primary sequence. Nature Communications 15, 2499 (2024).

33. Kümmerer M., Rathod P., Boström J. TernaryNet: deep learning prediction of PROTAC ternary complex formation. Cell Chemical Biology 31, 785–798 (2024).

34. "E3 ubiquitin ligases: styles, structures and functions" Published in *Molecular Biomedicine* (2021 Jul 30; 2:23). DOI: 10.1186/s43556-021-00043-2 PMCID: PMC8607428 | PMID: 35006464

35. Leskovec J., Krevl A. SNAP Datasets: Stanford Large Network Dataset Collection. (2014).

36. Zitnik M., Sosič R., Maheshwari S., Leskovec J. BioSNAP Datasets: Stanford Biomedical Network Dataset Collection. (2018).

37. Greene C.S., Krishnan A., Wong A.K., et al. Understanding multicellular function and disease with human tissue-specific networks. Nature Genetics 47, 569–576 (2015).

38. STRING — “The STRING database in 2025: protein networks with directionality” (NAR 53:D730–D739, 2025). Latest on evidence-scored PPI networks with new regulatory edges and directionality.

39. Prasad T.S.K., Goel R., Kandasamy K., et al. Human Protein Reference Database—2009 update. Nucleic Acids Research 37(suppl 1), D767–D772 (2009).

40. BioGRID — “The BioGRID interaction database: 2017 update” (NAR 45:D369–D379, 2017). Curated genetic, chemical, and protein interactions across species, widely used baseline for human interactomes.

41. Uhlén M., Fagerberg L., Hallström B.M., et al. Tissue-based map of the human proteome. Science 347, 1260419 (2015).

42. Prieto C., De Las Rivas J. APID: Agile Protein Interaction DataAnalyzer—a Swiss-knife for protein interaction analysis. Nucleic Acids Research 38(Database issue), D593–D599 (2010).

43. Hamp T., Rost B. Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and structure analysis. Nucleic Acids Research 40(Database issue), D821–D825 (2012).

44. UniProt Consortium. UniProt: the universal protein knowledgebase in 2025. Nucleic Acids Research 53, D1–D10 (2025).

45. Barrett Stone, M. “Cross-validatory choice and assessment of statistical predictions.” *Journal of the Royal Statistical Society, Series B* 36 (2), 111-147 (1974).

46. Oughtred R., Rust J., Chang C., et al. The BioGRID database: a comprehensive biomedical resource of curated protein, genetic and chemical interactions. Nucleic Acids Research 49(D1), D535–D541 (2021).

47. Szklarczyk D., Kirsch R., Koutrouli M., et al. STRING v11.5: protein–protein association networks across the tree of life. Nucleic Acids Research 51(D1), D638–D646 (2023).

48. Orchard S., Ammari M., Aranda B., et al. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Research 42(D1), D358–D363 (2014).

49. Licata L., Briganti L., Peluso D., et al. MINT, the Molecular INTeraction Database: 2012 update. Nucleic Acids Research 40(D1), D857–D861 (2012).

50. Bader G.D., Betel D., Hogue C.W.V. BIND—the Biomolecular Interaction Network Database. Nucleic Acids Research 31(1), 248–250 (2003).

51. Salwinski L., Eisenberg D. The Database of Interacting Proteins: 2004 update. Nucleic Acids Research 32(D1), D449–D451 (2004).

52. Hermjakob H., Montecchi-Palazzi L., Lewington C., et al. The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. Nature Biotechnology 22(2), 177–183 (2004).

53. Rao, R. et al. (2021). "MSA Transformer." *ICML 2021 (PMLR 139),* 8844–8856 — Introduces a PLM that operates directly on multiple-sequence alignments with interleaved row/column attention, improving contact prediction and mutational effect inference.

54. Brandes, N. et al. (2022). "ProteinBERT: a universal deep-learning model of protein sequence and function." *Bioinformatics* 38(8):2102–2110 — Protein-specific BERT with joint language-model and GO-prediction pretraining enabling efficient transfer across tasks.

55. Ferruz, N., Schmidt, S., & Höcker, B. (2022). "ProtGPT2 is a deep unsupervised language model for protein design." *Nature Communications* 13:4348 — Autoregressive LM generating de novo, foldable protein sequences that explore novel regions of sequence space.

56. Frazer, J. et al. (2021). "Disease variant prediction with deep generative models of evolutionary data (EVE)." *Nature* 599:91–95 — Label-free generative modeling of evolutionary constraints for pathogenicity prediction, rivaling curated clinical methods.

57. Notin, P. et al. (2023). "ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design." NeurIPS Datasets & Benchmarks — Standardizes >250 DMS assays plus curated clinical sets to enable fair comparison of PLMs and baselines.

58. Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., & Rives, A. (2021). Language models enable zero-shot prediction of the effects of mutations on protein function. NeurIPS 34, 29287–29303. Introduces ESM-1v and demonstrates state-of-the-art zero-shot variant-effect prediction across diverse DMS assays; also benchmarks ESM-1b and MSA Transformer.

59. Xu M., Zhang Z., Lu J., Zhu Z., Zhang Y., Ma C., Liu R., & Tang J. (2022). PEER: A Comprehensive and Multi-Task Benchmark for Protein Sequence Understanding. Advances in Neural Information Processing Systems 35 (Datasets & Benchmarks Track). .

60. Rao R. et al. Evaluating protein transfer learning with TAPE. Adv. Neural Inf. Process. Syst. 32, 9689–9701 (2019).

61. Jiang J., Liu Y., Dou Z., Wang C. Protein–protein interaction prediction with deep learning: a comprehensive review. Comput. Struct. Biotechnol. J. 20, 5830–5844 (2022).

62. Madan S., Demina V., Stapf M., Ernst O., Fröhlich H. Accurate prediction of virus-host protein–protein interactions via a Siamese neural network using deep protein sequence embeddings. Patterns 3, 100551 (2022).

63. Saito T., Rehmsmeier M. The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE 10 (3): e0118432 (2015).

64. Davis J., Goadrich M. The relationship between precision–recall and ROC curves. In Proceedings of the 23rd International Conference on Machine Learning (ICML 2006), 233–240 (2006).

65. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI 1995), 1137–1143 (1995).

66. Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. A tutorial treatment of ROC curves, AUC, and common pitfalls, establishing foundations for classifier evaluation.

67. Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. Machine Learning, 77(1), 103–123. Critiques AUROC's incoherence under differing cost distributions and proposes the H-measure as a cost-sensitive alternative.

68. Flach, P. A., & Kull, M. (2015). Precision-Recall-Gain curves: PR analysis done right. NeurIPS 28, 838–846. Reframes PR analysis with PR-Gain to obtain baseline-corrected, interpretable curves and areas, avoiding pathologies of conventional PR plots.

69. Bozdaganyan M, Fedorov V, Kholina E, et al. Exploring tubulin–paclitaxel binding modes through extensive molecular dynamics simulations. Sci Rep. 2025;15:8378. Uses multi-μs atomistic MD and MM/GBSA to map multiple paclitaxel poses and M-loop responses in β-tubulin.

70. Jiménez VA, Alderete JB, Navarrete KR. Molecular modeling study on the tubulin-binding modes of epothilone derivatives: insight into the structural basis for epothilones activity. Chem Biol Drug Des. 2017;90(6):1247-1259. MD over 20 epothilones links pose preferences and M-loop structuring to activity.

71. Churchill CDM, Klobukowski M, Tuszynski JA. The unique binding mode of laulimalide to two tubulin protofilaments. Chem Biol Drug Des. 2015;86(2):190-199. Extended microtubule model + 135 ns MD reveal a distinct, water-mediated laulimalide pose spanning adjacent β-tubulins.

72. Chen X. et al. Exploring tubulin–paclitaxel binding modes through molecular dynamics simulations. J. Mol. Biol. 437, 118973 (2025).

73. Narasimhan S. et al. Frontotemporal dementia P301L mutation potentiates tau fibrillization. Nat. Commun. 14, 5132 (2023).

74. Haupt S. et al. How does p53 induce apoptosis? Genes Cancer 8, 72–81 (2017).

75. Gilmore T.D. NF-κB: activation and regulation in immune responses. PLoS Biol. 12, e1001839 (2014).

76. Müller P.A.J., Vousden K.H. p53 mutations in cancer. Nature Cell Biology 15, 2–8 (2013).

77. Attardi LA, Donehower LA. p53 and tumor suppression: it takes a network. Trends in Cell Biology 31(4):298–310 (2021).

78. Freed-Pastor W.A., Prives C. Mutant p53: one name, many proteins. Cell 137, 413–431 (2012).

79. Sung H., Ferlay J., Siegel R.L., et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians 71, 209–249 (2021).

80. Lewis J., Dickson D.W., Lin W.L., et al. Enhanced neurofibrillary degeneration in transgenic mice expressing mutant (P301L) tau protein. Neuron 28, 617–631 (2000).

81. Jha, A., Aicher, J. K., Gazzara, M. R., Singh, D., & Barash, Y. (2020). Enhanced Integrated Gradients: improving interpretability of deep learning models using splicing codes as a case study. Genome Biology, 21, 149.

82. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science 369, 1318-1330 (2020).

83. Tabula Sapiens Consortium. The Tabula Sapiens: a single-cell transcriptomic atlas of multiple human organs. Science 376, eabl4896 (2022).

84. Bushweller J.H. Targeting transcription factors in cancer—from undruggable to reality. Nature Reviews Cancer 19, 611–624 (2019).

85. Lambert M., Jambon S., Depauw S., David-Cordonnier M.H. Targeting transcription factors for cancer treatment. Nature Reviews Drug Discovery 17, 997–1012 (2018).

86. Evans R.M., Mangelsdorf D.J. Nuclear receptors, RXR, and the big bang. Cell 157, 255–266 (2014).

87. Darnell J.E. Transcription factors as targets for drug therapy. Nature Reviews Cancer 2, 740–749 (2002).

88. Sundararajan M., Taly A., Yan Q. Axiomatic Attribution for Deep Networks. ICML 2017 (Proc. Mach. Learn. Res. 70, 3319-3328).

89. DeepLIFT (Shrikumar et al., ICML 2017) — Backprop-based attributions relative to a reference input; predates IG and motivates completeness-style axioms.

90. SmoothGrad (Smilkov et al., 2017) — Noise-averaged saliency maps that sharpen gradient attributions; often paired with IG/DeepLIFT.

91. Spillantini, M. G., & Goedert, M. (2013). Tau pathology and neurodegeneration. Lancet Neurology, 12(6), 609–622.

92. Hutton, M. et al. (1998). Association of missense and 5′-splice-site mutations in tau with the inherited dementia FTDP-17. Nature, 393(6686), 702–705.

93. Lewis, J. et al. (2000). Neurofibrillary tangles, amyotrophy and progressive motor disturbance in mice expressing mutant (P301L) tau protein. Nature Genetics, 25(4), 402–405.

94. SantaCruz et al., 2005 (Science) — rTg4510 (tet-off P301L) mice: progressive NFTs, neuron loss, and impairments; suppressing tau rescues memory even while NFTs keep accumulating.

95. Götz et al., 2001 (Science) — In P301L mice, intracerebral Aβ42 fibrils drive ~5× more NFTs near projection fields—linking amyloid to tau pathology in vivo.

96. Ramsden et al., 2005 (J. Neurosci.) — Age-dependent NFT formation and neuron loss in forebrain-directed P301L expression; strong behavioral decline.

97. Sahara, N. et al. (2002). “Assembly of tau in transgenic animals expressing P301L tau.” Biochemical dissection of JNPL3 (P301L) mice showing how soluble tau transitions into sarkosyl-insoluble aggregates—useful for staging pathology and mechanistic readouts beyond histology.

98. Li Y., Xie P., Lu L., Wang J., et al. “An integrated bioinformatics platform for investigating the human E3 ubiquitin ligase–substrate interaction network.” Nature Communications **8** (2017): 347. https://doi.org/10.1038/s41467-017-00299-9

99. Park S, Khan S, Wahab A. E3-targetPred: Prediction of E3-Target Proteins Using Deep Latent Space Encoding. arXiv [Preprint]. 2020; arXiv:2007.12073.

100. Shu Y, Hai Y, Cao L, Wu J. Deep-learning based approach to identify substrates of human E3 ubiquitin ligases and deubiquitinases. Comput Struct Biotechnol J. 2023;21:1014-1021. doi:10.1016/j.csbj.2023.01.021

101. Li Y, Xie P, Lu L, Wang J, Diao L, Liu Z, et al. An integrated bioinformatics platform for investigating the human E3 ubiquitin ligase-substrate interaction network. Nature Communications. 2017;8(1):347. doi:10.1038/s41467-017-00299-9.

102. Lin A., Ye J., Qi C., et al. Bridging artificial intelligence and biological sciences: a comprehensive review of large language models in bioinformatics. Briefings in Bioinformatics. 2025;26(4):bbaf357. doi:10.1093/bib/bbaf357

103. Lu et al., 2024 — "Large language models in biomedicine and health." Broad survey of clinical, biomedical NLP, and governance issues, with practical considerations.

104. Created in BioRender. Patjoshi, S. (2025) https://BioRender.com/p81l653

105. Created in BioRender. Patjoshi, S. (2025) https://BioRender.com/k25j412

106. Wang et al., 2025 — "Large Language Models in Bioinformatics: A Survey." Bio sequence–centric review (DNA/RNA/protein, single-cell) with datasets and evaluation discussion.

107. Created in BioRender. Patjoshi, S. (2025) https:// BioRender.com/1umdzlv

108. Created in BioRender. Patjoshi, S. (2025) https://BioRender.com/o1f61yk

109. Liu L-M, Tang Q, Hu X, Zhao J-J, Zhang Y, Ying G-G, Zhang F. Arginine Methyltransferase PRMT1 Regulates p53 Activity in Breast Cancer. Life. 2021;11(8):789.

110. Abdeen S K, Aqeilan R I. Decoding the link between WWOX and p53 in aggressive breast cancer. Cell Cycle. 2019;18(11):1177-1186.

111. Kim H J, Moon S J, Kim J H. Mechanistic insights into the dual role of CCAR2/DBC1 in cancer. Experimental & Molecular Medicine. 2023;55(8):1691-1701.

112. Kondo S, Lu Y, Debbas M, et al. Characterization of cells and gene-targeted mice deficient for the p53-binding kinase HIPK1. Proceedings of the National Academy of Sciences USA. 2003;100(9):5431-5436.

113. Das S, Raj L, Zhao B, et al. Hzf determines cell survival upon genotoxic stress by modulating p53 transactivation. Cell. 2007;130(4):624-637.

114. Rui Y, Xu Z, Lin S, et al. Axin stimulates p53 functions by activation of HIPK2 kinase through multimeric complex formation. EMBO Journal. 2004;23(23):4583-4594.

115. Wright P E, Dyson H J. Intrinsically disordered proteins in cellular signalling and regulation. Nature Reviews Molecular Cell Biology. 2015;16(1):18-29.

116. Maruyama Y. et al. Ubiquitin ligase Kf-1 is involved in the endoplasmic-reticulum-associated degradation pathway. Biochem Biophys Res Commun 374, 737–741 (2008).

117. Van Roey K. et al. Short linear motifs: ubiquitous and functionally diverse protein-interaction modules directing cell regulation. Chem Rev 114, 6733–6778 (2014).

118. Brayer K. J. & Segal D. J. Keep your fingers off my DNA: protein-protein interactions mediated by C2H2 zinc-finger domains. Cell Biochem Biophys 50, 111–131 (2008).

119. Duda D. M., Scott D. C., Calabrese M. F. *et al.* Structural insights into the autoinhibition and activation of Cbl E3 ubiquitin ligases. Nat. Struct. Mol. Biol. 19, 184–192 (2012).

120. Spratt D. E., Wu K., Kovacev J. *et al.* RING-finger E3 ubiquitin ligases: autoinhibition, substrate recognition and regulation. Q. Rev. Biophys. 46, 483–517 (2013).

121. Clackson, T., & Wells, J. A. (1995). A hot spot of binding energy in a hormone-receptor interface. *Science, 267*(5196), 383–386. https://doi.org/10.1126/science.7529940

122. Cukuroglu, E., Engin, H. B., Gursoy, A., & Keskin, O. (2014). Hot spots in protein–protein interfaces: Towards drug discovery. Progress in Biophysics and Molecular Biology, 116(2–3), 165–173. https://doi.org/10.1016/j.pbiomolbio.2014.06.003

123. Burgoyne, N. J., & Jackson, R. M. (2006). Predicting protein interaction sites: Binding hot-spots in protein–protein and protein–ligand interfaces. *Bioinformatics, 22*(11), 1335–1342. https://doi.org/10.1093/bioinformatics/btl079

124. McCoy, A. J., Epa, V. C., & Colman, P. M. (1997). Electrostatic complementarity at protein/protein interfaces. *Journal of Molecular Biology, 268*(2), 570–584. https://doi.org/10.1006/jmbi.1997.0987

125. Zhou, H. X., & Pang, X. (2018). Electrostatic interactions in protein structure, folding, binding, and condensation. Chemical Reviews, 118(2), 1691–1741. https://doi.org/10.1021/acs.chemrev.7b00305

126. Li, Y., Huang, Y., Swaminathan, C. P., Smith-Gill, S. J., & Mariuzza, R. A. (2005). Magnitude of the hydrophobic effect at central versus peripheral sites in protein-protein interfaces. Structure, 13(2), 297-307. https://doi.org/10.1016/j.str.2004.12.012