FusionPath: Gene fusion pathogenicity prediction using protein structural data and contextual protein embeddings

Nadine S. Kurz[1,*], Irem Berna Güven[1], Tim Beißbarth[1,2,3,4], and Jürgen Dönitz[1,2,3]

[1]Department of Medical Bioinformatics, University Medical Center Göttingen, 37075 Göttingen, Niedersachsen, Germany
[2]Comprehensive Cancer Center Göttingen (G-CCC), Göttingen, 37075, Germany
[3]Campus-Institut Data Science (CIDAS), Section Medical Data Science (MeDaS)
[4]Lower Saxony Center for AI and Causal Methods in Medicine (CAIMed)
[*]Corresponding author: nadine.kurz@bioinf.med.uni-goettingen.de

Abstract

Accurate prediction of gene fusion pathogenicity is critical for understanding oncogenic mechanisms and advancing precision oncology. While existing computational methods provide valuable insights, their performance remains limited by incomplete integration of multi-scale biological features and insufficient model interpretability for clinical translation.. We present FusionPath, a novel deep learning framework for gene fusion pathogenicity prediction. FusionPath uniquely integrates embeddings from multiple pretrained protein language models, including FusON-pLM and ProtBERT, alongside retained protein domains and Gene Ontology (GO) functional annotations. The model was trained and validated on a large-scale dataset of annotated pathogenic and benign fusions. The model was trained and validated on a rigorously curated dataset of 100,433 gene fusions (78,115 benign, 22,318 pathogenic) derived from FusionPDB, ChimerDB4.0, and 27 RNA-seq datasets of normal tissues. FusionPath significantly outperformed state-of-the-art methods, achieving higher AUC scores of 0.95 and 0.87 on independent test sets. By synergistically leveraging sequence, structural, and functional information with explicit modeling of wild-type sequence context, FusionPath establishes a new standard for gene fusion pathogenicity prediction by effectively leveraging complementary sequence, structural, and functional information.
Keywords: Gene fusions, Gene fusion pathogenicity, deep learning, protein language models, precision medicine

1 Introduction

Gene fusions represent structural genomic rearrangements yielding chimeric transcripts from two previously distinct protein-coding genes. Identified in ∼17% of all cancers [1], they are a well-established class of oncogenic drivers across diverse cancers [2, 3], including *BCR-ABL1* in chronic myeloid leukemia, and *EML4-ALK* in non-small cell lung carcinoma. Critically, the clinical success of kinase-inhibiting therapies targeting fusion oncoproteins, such as imatinib for *BCR-ABL1* [4], has established them as promising targets in precision medicine [11]. Comprehensive reviews have documented the expanding landscape of targetable oncogenic fusions across diverse cancer types, highlighting their established role as therapeutic targets in precision oncology [5]. However, only a subset of these fusions are pathogenic, while others represent passenger events or occur in normal physiology [6].

Recent studies reveal that fusions tend to preserve protein function and contain specific domain architectures [7], while over half of fusion oncoproteins form biomolecular condensates with distinct functional consequences for oncogenesis [3]. These findings underscore the critical need for pathogenicity prediction frameworks that integrate protein structural and functional context to enable precision oncology applications. Furthermore, fusion oncoproteins have been associated with additional characteristics, such as RNA polymerase II partitioning [8].

Various databases have aggregated data on NGS-detected gene fusions in pan-cancer and normal tissue, including FusionGDB [9] and FusionPDB [10], and the KuNG FU (KiNase Gene FUsion) database [11]. Fusion annotation tools, including FusionAnnotator, annoFuse [12] and FusionInspector [13] focus on structural validation and basic feature annotation, but lack explicit pathogenicity scoring. Computational methods for gene fusion pathogenicity prediction include 1) ChimerDriver [14], a Multi-Layer Perceptron (MLP) integrating transcriptional and post-transcriptional information including microRNAs for oncogenic potential classification, 2)

DEEPrior [15], a model combining Convolutional Neural Networks (CNN) and bidirectional Long-Short Term Memory (LSTM), 3) Pegasus [16], an approach using gradient tree boosting for binary classification of oncogenic potential, and OncoFuse [17], a Naive Bayes Network Classifier that predicts the oncogenic potential of fusion genes. Despite these advances, current methods remain limited by incomplete integration of multi-scale biological features and lack of interpretability, highlighting the need for more sophisticated approaches that can capture the complex molecular determinants of fusion pathogenicity [18].

We present FusionPath, an interpretable framework for gene fusion pathogenicity prediction. FusionPath integrates three complementary biological scales, including fusion-specific protein embeddings via language models, domain architecture alterations, and Gene Ontology (GO) [19] term enrichment. By unifying sequence-level, structural, and systems-level features, FusionPath provides biologically grounded pathogenicity scores with mechanistic insights.

2 Methods

2.1 Datasets

We compiled gene fusion data from two primary sources: FusionPDB, a structural database of experimentally validated oncogenic fusions, and ChimerDB4.0 [20], which integrates literature-mined, cancer-type-specific fusions. To retrieve additional gene fusions from normal tissues, we utilized 27 RNA-seq datasets of the European Nucleotide Archive (ENA), including fusions from 95 healthy donors of 27 tissues were downloaded from PRJEB4337 [21], and fusions from cancerous and normal breast cancer tissue were downloaded from PRJNA975550 [22], among others. The full list of datasets is available at Supplementary Table S1.

2.2 Preprocessing and annotation

All genomic coordinates were standardized to GRCh38 using the SeqCAT LiftOver API [23]. Fusion calling for normal tissues was performed with STARFusion [24, 25], yielding 5.403 fusion candidates. Wild-type protein sequences of the fusions' head and tail genes were queried using SeqCAT's protein sequence API endpoint, retrieving the genes' MANE Select transcripts [26]. Protein domain and GO terms were retrieved using InterProScan 5 [27]. Known gene roles of tumor suppressor and oncogenes were retrieved from Onkopus [28] gene role Application Programming Interface (API) endpoint, where we selected the classification by Vogelstein et al. (2013) [29]. Gene-disease associations were retrieved from Clingen [30].

2.3 The FusionPath model

For each fusion protein sequence, FusionPath extracts the wild-type protein sequences of the affected genes, as well as the retained protein domains and GO terms identified in the fusion sequence. The fusion amino acid sequence was encoded using FusON-pLM [31]. Wild-type head-/tail sequences were independently embedded via ProtBERT [32]. Concatenated sequence embeddings underwent batch normalization and dropout. Concurrently, retained protein domains and Gene Ontology terms were tokenized into a vocabulary, projected to 128-dim embeddings, and processed via additive attention [33]. To dynamically weight the contribution of each domain and GO term based on its relevance to pathogenicity, we employ an additive attention mechanism. Given an embedding matrix $\mathbf{E} \in \mathbb{R}^{x \times d}$ containing the d-dimensional embeddings of n retained protein domains and GO terms, the attention weights are computed as:

$$\mathbf{a} = \text{softmax}\left(\mathbf{v}^\top \tanh(\mathbf{W}_1\mathbf{E} + \mathbf{b}_1)\right), \quad \mathbf{e}_{\text{dom}} = \mathbf{a}^\top \mathbf{E} \quad (1)$$

where $\mathbf{W}_1 \in \mathbb{R}^{h \times d}$ is a learnable weight matrix, $b_1 \in \mathbb{R}^h$ is a bias term, $\mathbf{v} \in \mathbb{R}^h$ is a context vector, and h is the hidden dimension. The processed sequence embeddings and functional annotation embeddings are then concatenated into a unified feature vector $\mathbf{z} \in \mathbb{R}^D$, where D is the total dimensionality of the integrated representation. Integrated features fed a 3-layer Multi-layer perceptron (MLP), with sigmoid output for pathogenicity probability.

$$\begin{aligned} \mathbf{h}_1 &= \text{ReLU}(\mathbf{W}_2\mathbf{z} + \mathbf{b}_2) \\ \mathbf{h}_2 &= \text{ReLU}(\mathbf{W}_3\mathbf{h}_1 + \mathbf{b}_3) \\ p &= \sigma(\mathbf{w}_4^\top \mathbf{h}_2 + b_4) \end{aligned} \quad (2)$$

where $\mathbf{W}_2 \in \mathbb{R}^{d_1 \times D}$ and $\mathbf{W}_3 \in \mathbb{R}^{d_2 \times d_1}$ are weight matrices for the hidden layers with dimensions $d_1 = 512$ and $d_2 = 256$, $\mathbf{b}_2$ and $\mathbf{b}_3$ are corresponding bias vectors, $\mathbf{w}_4 \in \mathbb{R}^{d_2}$ is the output layer weight vector, b_4 is the output bias term, and $\sigma(\cdot)$ denotes the sigmoid function. The model was trained as a binary classifier, optimizing binary cross-entropy loss with logits (BCEWithLogitsLoss) using the AdamW optimizer. Given N training samples with true labels $y_i \in \{0, 1\}$ and raw model outputs (logits) z_i, the loss function is defined as:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\sigma(z_i)) + (1 - y_i)\log(1 - \sigma(z_i))\right] \quad (3)$$

where z_i are the raw logits. Training proceeded with standard forward-backward passes, where the model outputs were passed through a sigmoid function.

2.4 Evaluation metrics

We employed receiver operating characteristic (ROC) curve analysis and the area under the ROC curve (AUC) for evaluating the model's performance compared to prior methods. The impact of protein domains and GO terms was assessed by computing attention weights. SHAP

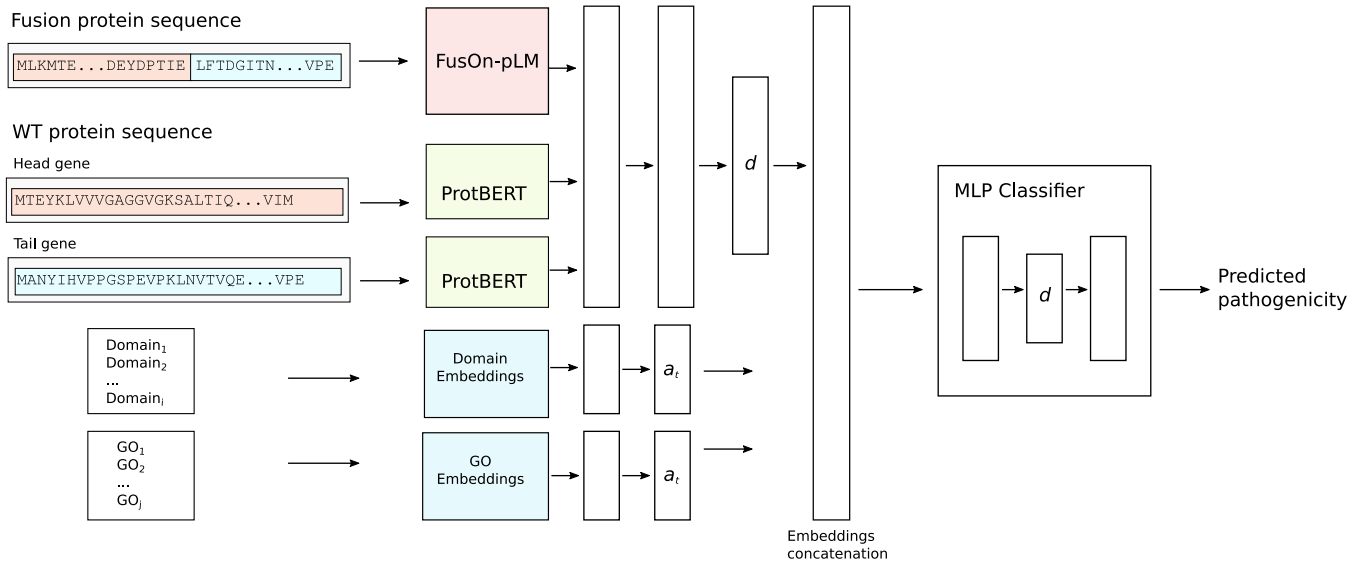

Figure 1: The FusionPath architecture. The fusion amino acid sequence is processed by FusON-pLM, generating contextual embeddings. The wild-type head and tail gene sequences are independently encoded using ProtBERT, yielding separate embedding tensors for each sequence. Embeddings from all sequence streams are concatenated along the feature dimension, followed by batch normalization and dropout regularization. Protein domains and and GO terms associated with the fusion are mapped to discrete tokens using a curated biological vocabulary. Tokenized inputs are then projected into a dense embeddings and processed through a dedicated attention layer, enabling context-aware weighting of functionally relevant domains and GO annotations. The processed sequence embeddings and functional annotation embeddings are then concatenated into a unified feature vector. This integrated representation is passed through a multilayer perceptron (MLP) classifier comprising ReLU activation and dropout. The final layer uses a sigmoid activation to output a pathogenicity probability score.

(SHapley Additive exPlanations) [34] values were computed using the KernelExplainer with 500 background samples drawn from the test dataset, with 2000 Monte Carlo iterations for precision. Attention weights were computed using an additive attention mechanism [33], where feature importance was quantified as the mean attention weight across high-probability pathogenic samples. Only features appearing in $\geq$5 samples were included in the final ranking, with results sorted by mean attention weight to identify the most influential protein domains and GO terms.

3 Results

3.1 The FusionPath architecture and dataset

FusionPath implements a multimodal model that combines protein structural information with the contextual embeddings of pretrained protein language models. The model processes amino acid sequences of fusion proteins as input and extracts the wild-type sequences of the head and tail genes, as well as the retained protein domains and GO terms (Figure 1). The model incorporates two specialized components: 1) FusON-pLM, a protein language model fine-tuned exclusively on oncoprotein sequences to process fusion protein contexts, and 2) ProtBERT, used to extract implicit functional knowledge from wild-type sequences. In addition, the model processes the retained protein domains as well as identified GO terms as input.

We curated a dataset of 100.433 gene fusions from cancerous and normal tissue databases, supplemented by 27 manually validated normal tissue RNA-seq datasets. For classifying gene fusions, we annotated the head and tail genes of the fusions with their known role in carcinogenesis (tumor suppressors and oncogenes), known gene-disease associations. Fusions were labeled as pathogenic if they were associated with cancer, involved a known oncogene or tumor suppressor, or a gene-disease association. Fusions were labeled as benign if they did not include a known tumor suppressor or oncogene, had no known gene-disease association, and whose fusion protein as well as the head or tail gene had no known association as cancer biomarker. Additionally, all fusions that were called from normal tissue were labeled as benign. The final dataset comprised 78.115 benign (77.8%) and 22.318 pathogenic (22.2%) fusions.

3.2 FusionPath improves pan-cancer gene fusion pathogenicity prediction

We benchmarked FusionPath against the three state-of-the-art methods ChimerDriver, DEEPrior, and Pegasus. Using our curated dataset of 19.934 validated pathogenic and benign gene fusions, we assessed the model performance. FusionPath achieved significantly higher predic-

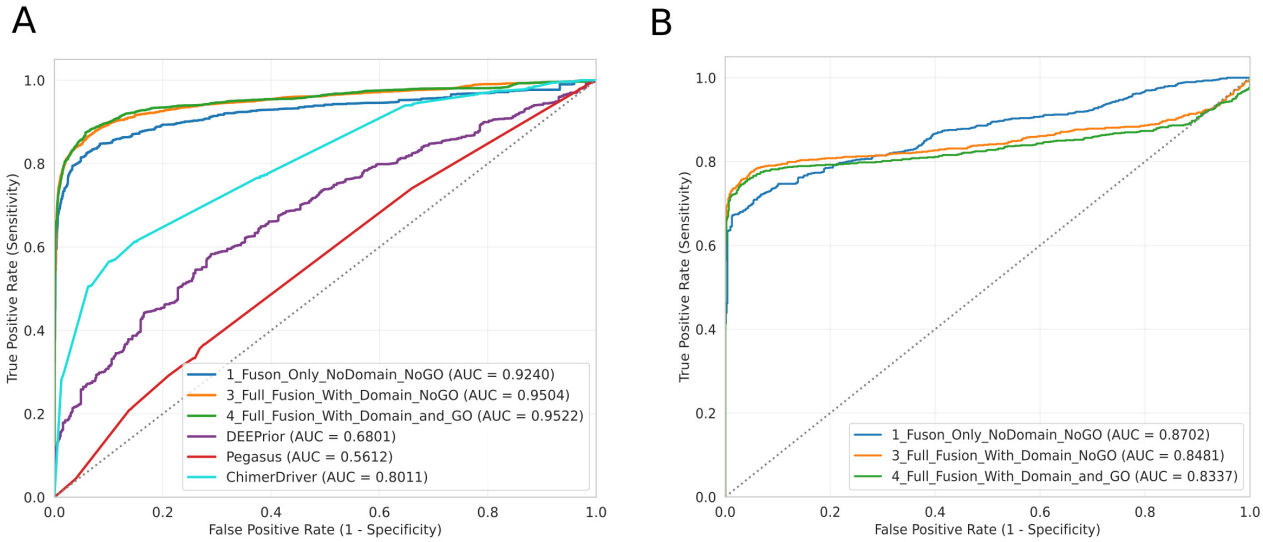

Figure 2: Performance comparison and classification results of FusionPaths. **A** ROC/AUC curves of different models of FusionPath on the test set containing 19.934 fusion protein, including models with and without the inclusion of the ProtBERT models wild-type sequence embeddings, as well as with and without the inclusion of protein structure. The full model (integrating wild-type sequence embeddings, protein structure, domains, and GO terms) achieved the highest AUC (0.95). All models significantly outperformed ChimerDriver (AUC=0.8011), DEEPrior (AUC=0.68), and Pegasus (AUC=0.56). **B** ROC/AUC curve of FusionPath tested on the DEEPrior training dataset, achieving an AUC of 0.87.

tion accuracy compared to the other methods, with an AUC of 0.95% on the test data set (Figure 2A). To assess generalizability, we further evaluated FusionPath on the DEEPrior training dataset as an external validation cohort, achieving an AUC of 0.87 (Figure 2B), confirming robustness across diverse data sources.

3.3 FusionPath provides insights in the role of protein domains and functions in fusion proteins

Beyond predictive accuracy, FusionPath provides biologically interpretable insights through SHAP analysis as well as its attention mechanism, which quantify feature contributionss, including the protein domains and GO terms, to pathogenicity predictions. This enables identification of molecular determinants driving oncogenicity, facilitating hypothesis generation for experimental validation. Analysis of the test set revealed cancer-type-specific pathogenicity patterns, with the highest average pathogenicity in Acute Myeloid Leukemia (LAML) and Mesothelioma (MESO), and the lowest in Adrenocortical Carcinoma (ACC), Kidney Renal Clear Cell Carcinoma (KIRC), and Uveal Melanoma (UVM) (Figure 3A). Critically, SHAP analysis identified protein kinase domains (e.g., serine-threonine/tyrosine kinases) as the strongest positive contributors to pathogenicity, consistent with known oncogenic mechanisms like constitutive kinase activation. Conversely, domains such as fibroblast growth factor receptors (FGFRs) and immunoglobulin-like domains exhibited the most negative SHAP values, suggesting protective or neutral roles in fusion contexts. This domain-level resolution reveals how specific functional modules dictate fusion-driven oncogenesis.

4 Discussion

Accurate identification of pathogenic gene fusions remains a critical need in precision oncology, where misclassification may lead to missed driver mutations, ineffective therapies or overlooked therapeutic opportunities. We present FusionPath, a multimodal framework integrating protein language modeling with functional genomic context to predict fusion pathogenicity. Our results demonstrate that FusionPath significantly outperforms state-of-the-art methods in both pan-cancer and tissue-specific benchmarks. FusionPath's integration of protein structural data with contextual embeddings and functional annotations enables more comprehensive assessment across diverse fusion types, addressing a key limitation identified in the benchmark where tools struggled with fusions lacking obvious cancer-related gene associations [18]. This capability is particularly valuable for discovering novel fusion drivers beyond the well-characterized kinase fusions. FusionPath's performance stems from its integration of complementary biological signals through three synergistic components: A FusON-pLM model embedding the chimeric protein sequence to capture novel structural and functional features induced by the fusion junction, dual

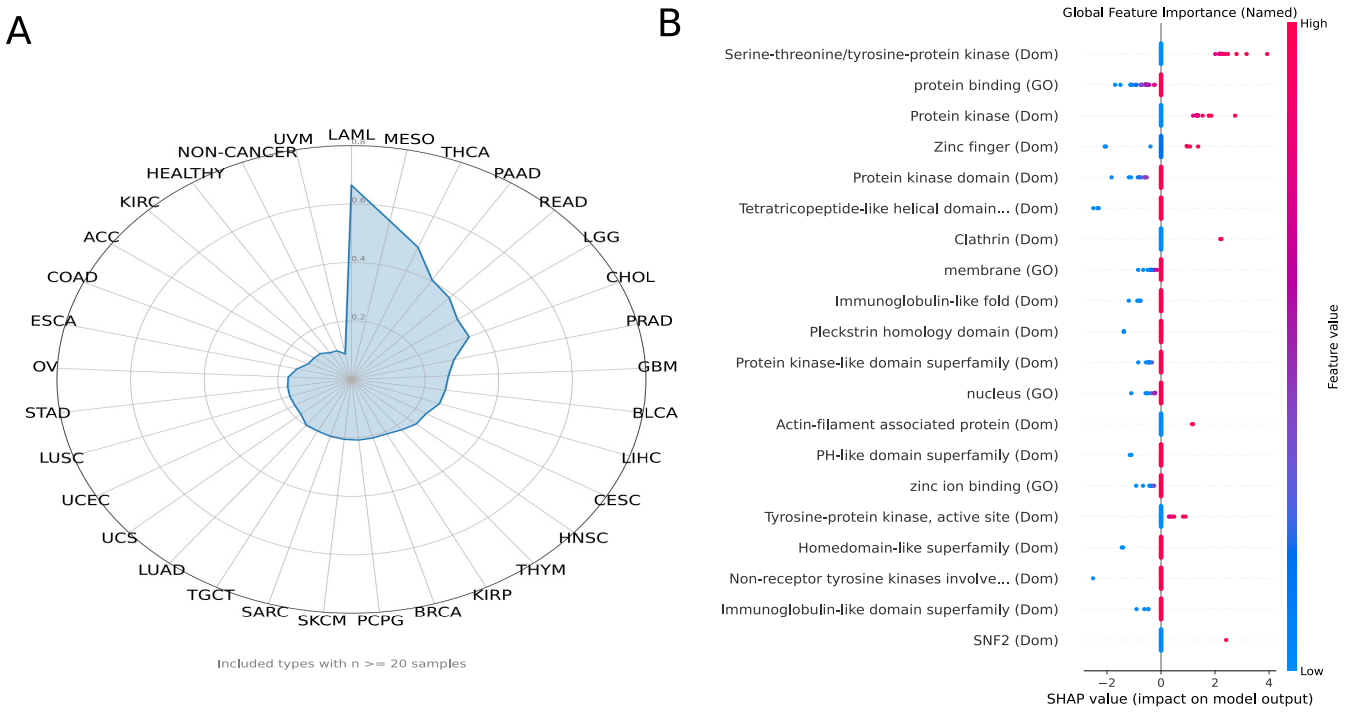

Figure 3: Analysis of the effect of protein domains and GO terms on gene fusion pathogenicity. **A** Average predicted gene fusion pathogenicity per cancer type. Acute Myeloid Leukemia (LAML), Mesothelioma (MESO) and Thyroid carcinoma (THCA) exhibited the highest average fusion protein pathogenicity, while the most benign fusions could be identified in Adrenocortical carcinoma (ACC), kidney renal clear cell carcinoma (KIRC), and uveal melanoma (UVM). **B** SHAP values for FusionPath detecting how individual protein domains and GO terms shift the model's prediction. A positive SHAP value indicates that the presence of a domain/GO term increases the predicted pathogenicity, while a negative SHAP value indicates that its presence decreases pathogenicity.

ProtBERT encoders processing the wild-type parental sequences to preserve baseline functional context disrupted by the fusion, and layers integrating protein domain architectures and GO terms to explicitly model functional consequences, e.g. loss of regulatory domains, or gain of dimerization motifs.

A major challenge in predicting the pathogenicity of gene fusions is determining the ground truth dataset that serves as the basis for training and validating the model. While two human-curated datasets, ClinVar [35] and HGMD [36], are available for missense variants and are often employed for model validation for missense variant pathogenicity prediction, there is no comparable equivalent for gene fusions or structural variants. This challenge was starkly demonstrated by Wei et al., who systematically re-evaluated 2,727 reported prostate cancer fusions using TCGA RNA-seq data and found that over 90% could not be validated in independent cohorts, with only 4% showing tumor-specific occurrence [37]. Existing fusion databases, e.g. COSMIC [38], catalog occurrences, but lack systematic pathogenicity labels. To overcome this, we constructed a rigorously curated dataset of cancer-derived and normal-tissue gene fusions, where fusions were labeled pathogenic if they involved a known tumor suppressor gene or oncogene, included at least one gene disease association. This approach corresponds to biological reality, as most gene fusions represent passenger rather than driver mutations. However, a critical consideration is the class imbalance in the dataset (77.8% benign fusions), which inherently biases models toward benign predictions. While our biologically informed curation criteria (cancer recurrence, tumor suppressor gene/oncogene involvement, disease association) provide a pragmatic and interpretable proxy for pathogenicity, it has to be pointed out that this approach does not imply with certainty that all fusions are correctly classified.

Notably, FusionPath revealed cancer-type-specific pathogenicity patterns, with LAML, MESO, and THCA exhibiting the highest scores, and ACC, KIRC, and UVM showing the lowest. While requiring experimental validation, these patterns align with known oncogenic mechanisms. For instance, high scores in THCA reflect the established role of RET and NTRK fusions in subsets of papillary thyroid carcinoma, which are actionable targets for kinase inhibitors [cite specific reviews, e.g., Cancer Discov 2018]. As these are tumor types with fusions established as therapeutic targets, e.g. *BCR::ABL1*, *RUNX1::RUNX1T1* or *ETV6::NTRK3* in LAML [39, 40], FusionPath has direct clinical rele-

vance and can contribute to improved targeted therapies for fusion oncoproteins. Conversely, the low average pathogenicity in UVM is consistent with its primary dependence on GNAQ/GNA11 mutations rather than fusion drivers [41].

SHAP interpretability analysis revealed kinase-related domains, including receptor tyrosine kinases, fibroblast growth factor receptors, and serine-threonine/tyrosine kinases, as the dominant predictors of pathogenicity, directly aligning with established oncogenic mechanisms where fusion-driven kinase activation drives malignancy. Critically, the model prioritizes therapeutically targetable kinase fusions (e.g. FGFR, ALK) with high confidence, enabling direct translation to biomarker-guided therapy selection. Despite known database overrepresentation of kinase fusions, rigorous validation on kinase-depleted cohorts confirmed robust generalizability to non-kinase pathogenic events. This kinase-centric interpretability establishes our tool as a biologically grounded clinical decision aid for precision oncology.

In summary, FusionPath establishes a new paradigm for fusion pathogenicity prediction by unifying deep protein language modeling with explicit functional genomics. By providing a transparent, biologically grounded dataset and a model that reveals fusion protein pathogenicity, we offer a tool that can accelerate both discovery of novel fusion drivers and clinical decision-making in precision oncology.

5 Code and data availability

The source code of FusionPath is available at https://gitlab.gwdg.de/MedBioinf/mtb/fusionpath.

6 Competing interests

No competing interest is declared.

7 Author contributions

NSK conceptualized the method. NSK and IBG implemented the model. NSK wrote the manuscript. All authors proofread the manuscript.

8 Acknowledgments

The authors thank the Internal Max Planck Research School for Genome Science (IMPRS-GS).

9 Funding

This work was supported by the Gemeinsamer Bundesauschuss (01NVF20006), the Deutsche Krebshilfe (70114018), the Deutsche Forschungsgemeinschaft (KFO5002), Niedersachsen Ministerium für Wissenschaft und Kultur (VWZN4257, 11-76251-12-1/19 ZN3421), the Internal Max Planck Research School for Genome Science (IMPRS-GS), and the Bundesministerium für Bildung und Forschung (BMBF; 01KD2437, 01KD2401B, 01KD2208A, 01KD2414A).

References

[1] Gao, Q., Liang, W., Foltz, S., Mutharasu, G., Jayasinghe, R., Cao, S., Liao, W., Reynolds, S., Wyczalkowski, M., Yao, L. & Others Driver fusions and their implications in the development and treatment of human cancers. *Cell Reports*. **23**, 227-238 (2018)

[2] Mitelman, F., Johansson, B. & Mertens, F. The impact of translocations and gene fusions on cancer causation. *Nature Reviews Cancer*. **7**, 233-245 (2007)

[3] Tripathi, S., Shirnekhi, H., Gorman, S., Chandra, B., Baggett, D., Park, C., Somjee, R., Lang, B., Hosseini, S., Pioso, B. & Others Defining the condensate landscape of fusion oncoproteins. *Nature Communications*. **14**, 6008 (2023)

[4] Stransky, N., Cerami, E., Schalm, S., Kim, J. & Lengauer, C. The landscape of kinase fusions in cancer. *Nature Communications*. **5**, 4846 (2014)

[5] Liu, S., Nagasaka, M., Atz, J., Solca, F. & Müllauer, L. Oncogenic gene fusions in cancer: from biology to therapy. *Signal Transduction And Targeted Therapy*. **10**, 111 (2025)

[6] Singh, S., Qin, F., Kumar, S., Elfman, J., Lin, E., Pham, L., Yang, A. & Li, H. The landscape of chimeric RNAs in non-diseased tissues and cells. *Nucleic Acids Research*. **48**, 1764-1778 (2020)

[7] Latysheva, N. & Babu, M. Discovering and understanding oncogenic gene fusions through data intensive computational approaches. *Nucleic Acids Research*. **44**, 4487-4503 (2016)

[8] Lyons, H., Pradhan, P., Prakasam, G., Vashishtha, S., Li, X., Eppert, M., Fornero, C., Tcheuyap, V., McGlynn, K., Yu, Z. & Others RNA polymerase II partitioning is a shared feature of diverse oncofusion condensates. *Cell*. (2025)

[9] Kim, P., Tan, H., Liu, J., Lee, H., Jung, H., Kumar, H. & Zhou, X. FusionGDB 2.0: fusion gene annotation updates aided by deep learning. *Nucleic Acids Research*. **50**, D1221-D1230 (2022)

[10] Kumar, H., Tang, L., Yang, C. & Kim, P. FusionPDB: a knowledgebase of human fusion proteins. *Nucleic Acids Research*. **52**, D1289-D1304 (2024)

[11] Somaschini, A., Di Bella, S., Cusi, C., Raddrizzani, L., Leone, A., Carapezza, G., Mazza, T., Isacchi, A. & Bosotti, R. Mining potentially actionable kinase gene fusions in cancer cell lines with the KuNG FU database. *Scientific Data*. **7**, 420 (2020)

[12] Gaonkar, K., Marini, F., Rathi, K., Jain, P., Zhu, Y., Chimicles, N., Brown, M., Naqvi, A., Zhang, B., Storm, P. & Others annoFuse: an R Package to annotate, prioritize, and interactively explore putative oncogenic RNA fusions. *BMC Bioinformatics*. **21**, 577 (2020)

[13] Haas, B., Dobin, A., Ghandi, M., Van Arsdale, A., Tickle, T., Robinson, J., Gillani, R., Kasif, S. & Regev, A. Targeted in silico characterization of fusion transcripts in tumor and normal tissues via FusionInspector. *Cell Reports Methods*. **3** (2023)

[14] Lovino, M., Montemurro, M., Barrese, V. & Ficarra, E. Identifying the oncogenic potential of gene fusions exploiting miRNAs. *Journal Of Biomedical Informatics*. **129** pp. 104057 (2022)

[15] Lovino, M., Ciaburri, M., Urgese, G., Di Cataldo, S. & Ficarra, E. DEEPrior: a deep learning tool for the prioritization of gene fusions. *Bioinformatics (Oxford, England)*. **36**, 3248-3250 (2020)

[16] Abate, F., Zairis, S., Ficarra, E., Acquaviva, A., Wiggins, C., Frattini, V., Lasorella, A., Iavarone, A., Inghirami, G. & Rabadan, R. Pegasus: a comprehensive annotation and prediction tool for detection of driver gene fusions in cancer. *BMC Systems Biology*. **8** pp. 97 (2014)

[17] Shugay, M., Mendíbil, I., Vizmanos, J. & Novo, F. Oncofuse: a computational framework for the prediction of the oncogenic potential of gene fusions. *Bioinformatics (Oxford, England)*. **29**, 2539-2546 (2013)

[18] Miccolis, F., Lovino, M. & Ficarra, E. Prioritizing gene fusions: a benchmark study of computational tools. *Bioinformatics*. **14513** pp. 202-211 (2025)

[19] Consortium, G. & Others The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research*. **49**, D325 (2020)

[20] Jang, Y., Jang, I., Kim, S., Cho, S., Kim, D., Kim, K., Kim, J., Hwang, J., Kim, S., Kim, J. & Others ChimerDB 4.0: an updated and expanded database of fusion genes. *Nucleic Acids Research*. **48**, D817-D824 (2020)

[21] Fagerberg, L., Hallström, B., Oksvold, P., Kampf, C., Djureinovic, D., Odeberg, J., Habuka, M., Tahmasebpoor, S., Danielsson, A., Edlund, K. & Others Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Molecular & Cellular Proteomics*. **13**, 397-406 (2014)

[22] Li, S., Hammarlund, J., Wu, G., Lian, J., Howell, S., Clarke, R., Adamson, A., Gonçalves, C., Hogenesch, J., Anafi, R. & Others Tumor circadian clock strength influences metastatic potential and predicts patient prognosis in luminal A breast cancer. *Proceedings Of The National Academy Of Sciences*. **121**, e2311854121 (2024)

[23] Kornrumpf, K., Kurz, N., Drofenik, K., Krauß, L., Schneider, C., Koch, R., Beißbarth, T. & Dönitz, J. Seqcat: Sequence conversion and analysis toolbox. *Nucleic Acids Research*. **52**, W116-W120 (2024)

[24] Haas, B., Dobin, A., Stransky, N., Li, B., Yang, X., Tickle, T., Bankapur, A., Ganote, C., Doak, T., Pochet, N. & Others STAR-Fusion: fast and accurate fusion transcript detection from RNA-Seq. *BioRxiv*. pp. 120295 (2017)

[25] Haas, B., Dobin, A., Li, B., Stransky, N., Pochet, N. & Regev, A. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biology*. **20**, 213 (2019)

[26] Morales, J., Pujar, S., Loveland, J., Astashyn, A., Bennett, R., Berry, A., Cox, E., Davidson, C., Ermolaeva, O., Farrell, C. & Others A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature*. **604**, 310-315 (2022)

[27] Jones, P., Binns, D., Chang, H., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G. & Others InterProScan 5: genome-scale protein function classification. *Bioinformatics*. **30**, 1236-1240 (2014)

[28] Kurz, N., Kornrumpf, K., Tucholski, T., Drofenik, K., König, A., Beißbarth, T. & Dönitz, J. Onkopus: precise interpretation and prioritization of sequence variants for biomedical research and precision medicine. *Nucleic Acids Research*. pp. gkaf376 (2025)

[29] Vogelstein, B., Papadopoulos, N., Velculescu, V., Zhou, S., Diaz Jr, L. & Kinzler, K. Cancer genome landscapes. *Science*. **339**, 1546-1558 (2013)

[30] Rehm, H., Berg, J., Brooks, L., Bustamante, C., Evans, J., Landrum, M., Ledbetter, D., Maglott, D., Martin, C., Nussbaum, R. & Others ClinGen—the clinical genome resource. *New England Journal Of Medicine*. **372**, 2235-2242 (2015)

[31] Vincoff, S., Goel, S., Kholina, K., Pulugurta, R., Vure, P. & Chatterjee, P. FusOn-pLM: a fusion oncoprotein-specific language model via adjusted rate masking. *Nature Communications*. **16**, 1436 (2025)

[32] Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M. & Others Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions On Pattern Analysis And Machine Intelligence*. **44**, 7112-7127 (2021)

[33] Bahdanau, D. Neural machine translation by jointly learning to align and translate. *ArXiv Preprint ArXiv:1409.0473*. (2014)

[34] Lundberg, S. & Lee, S. A unified approach to interpreting model predictions. *Advances In Neural Information Processing Systems*. **30** (2017)

[35] Landrum, M., Chitipiralla, S., Brown, G., Chen, C., Gu, B., Hart, J., Hoffman, D., Jang, W., Kaur, K., Liu, C. & Others ClinVar: improvements to accessing data. *Nucleic Acids Research*. **48**, D835-D844 (2020)

[36] Stenson, P., Mort, M., Ball, E., Chapman, M., Evans, K., Azevedo, L., Hayden, M., Heywood, S., Millar, D., Phillips, A. & Others The Human Gene Mutation Database (HGMD®): optimizing its use in a clinical diagnostic or research setting. *Human Genetics*. **139**, 1197-1207 (2020)

[37] Wei, T., Lu, J., Ma, T., Huang, H., Kocher, J. & Wang, L. Re-evaluate fusion genes in prostate cancer. *Cancer Informatics*. **20** pp. 11769351211027592 (2021)

[38] Sondka, Z., Dhir, N., Carvalho-Silva, D., Jupe, S., Madhumita, N., McLaren, K., Starkey, M., Ward, S., Wilding, J., Ahmed, M. & Others COSMIC: a curated database of somatic variants and clinical data for cancer. *Nucleic Acids Research*. **52**, D1210-D1217 (2024)

[39] Nimmanapalli, R. & Bhalla, K. Novel targeted therapies for Bcr–Abl positive acute leukemias: beyond STI571. *Oncogene*. **21**, 8584-8590 (2002)

[40] Gozgit, J., Wong, M., Wardwell, S., Tyner, J., Loriaux, M., Mohemmad, Q., Narasimhan, N., Shakespeare, W., Wang, F., Druker, B. & Others Potent activity of ponatinib (AP24534) in models of FLT3-driven acute myeloid leukemia and other hematologic malignancies. *Molecular Cancer Therapeutics*. **10**, 1028-1035 (2011)

[41] Carvajal, R., Sosman, J., Quevedo, J., Milhem, M., Joshua, A., Kudchadkar, R., Linette, G., Gajewski, T., Lutzky, J., Lawson, D. & Others Effect of selumetinib vs chemotherapy on progression-free survival in uveal melanoma: a randomized clinical trial. *Jama*. **311**, 2397-2405 (2014)