FusOn-pLM: A Fusion Oncoprotein-Specific Language Model via Adjusted Rate Masking

Sophia Vincoff
Department of Biomedical Engineering
Duke University
Durham, NC 27707
sophia.vincoff@duke.edu

Shrey Goel
Department of Computer Science
Duke University
Durham, NC 27707
shrey.goel@duke.edu

Kseniia Kholina
Department of Biomedical Engineering
Duke University
Durham, NC 27707
kseniia.kholina@duke.edu

Rishab Pulugurta
Department of Biomedical Engineering
Duke University
Durham, NC 27707
rishab.pulugurta@duke.edu

Pranay Vure
Department of Biomedical Engineering
Duke University
Durham, NC 27707
pranay.vure@duke.edu

Pranam Chatterjee
Department of Biomedical Engineering
Department of Computer Science
Department of Biostatistics and Bioinformatics
Duke University
Durham, NC 27707
pranam.chatterjee@duke.edu

Abstract

Fusion oncoproteins, a class of chimeric proteins arising from chromosomal translocations, are major drivers of various cancers, particularly in children. These proteins are intrinsically disordered, large, and lack well-defined druggable pockets, making them highly challenging therapeutic targets for both small molecule-based and structure-based approaches. Protein language models (pLMs) have recently emerged as powerful tools for capturing protein sequence features, enabling downstream applications such as disorder prediction, binding site identification, and therapeutic design. However, existing pLMs, including ESM-2 and ProtT5, have not been trained on fusion oncoprotein sequences, limiting their effectiveness for this class of proteins. In this work, we introduce FusOn-pLM, a fine-tuned pLM specifically trained on a newly-curated, comprehensive set of fusion oncoprotein sequences, FusOn-DB. FusOn-pLM employs a novel cosine-scheduled masked language modeling (MLM) strategy, dynamically varying the masking rate from 15% to 40% during training, to balance feature extraction and representation quality. Our model demonstrates improved performance against baseline embeddings on fusion-specific tasks, including fusion oncoprotein localization and puncta formation propensity, as well as strong prediction of intrinsically disordered residues and properties. Furthermore, as a case study of its biological relevance, we show that FusOn-pLM is uniquely capable of predicting drug-resistant mutations in fusion oncoproteins, offering a framework for therapeutic design that anticipates resistance mechanisms. By leveraging these capabilities, FusOn-pLM provides biologically relevant representations for advancing therapeutic discovery in fusion-driven cancers.

Machine Learning for Structural Biology Workshop, NeurIPS 2024.

1 Introduction

Fusion oncoproteins arise from chromosomal rearrangements that fuse segments of two distinct genes (Figure 1A).(1) The resulting mutants contain unrelated functional domains connected by long regions of disorder.(2) This flexible configuration promotes constitutive activation or aberrant regulation of the fusion proteins, driving oncogenic transformation and tumor development.(3) Thousands of unique fusion oncoproteins have been discovered by sequencing patient tumors, and several common culprits such as EWSR1::FLI1 in Ewing's sarcoma,(4) PAX3::FOXO1 in alveolar rhabdomyosarcoma (ARMS)(5), SS18::SSX1 in synovial sarcoma (6), and EML4::ALK proteins in non-small-cell lung cancer (7) are well characterized in the literature. However, even the best understood fusion oncoproteins have proven to be elusive drug targets due to their structural instability and absence of defined binding pockets.(2) For small molecules that are able to bind fusion oncoproteins, such as EWSR1::FLI1,(8; 9) these compounds do not achieve strict fusion specificity, binding to one of their head or tail protein counterparts that are often critical regulators of cellular homeostasis. As such, biologics, such as antibodies, miniproteins, and peptides, represent attractive therapeutic alternatives, but necessitate advanced design approaches for specific targeting to these undruggable proteins.(10; 11; 12; 13)

Recently, structure-based prediction and design models, such as AlphaFold and RFDiffusion,(14; 15; 16) have accelerated the design of biologics targeting pathogenic proteins. These tools, by default, fail to accurately capture the structure of numerous conformationally unstable proteins, limiting their usefulness for fusion oncoprotein targeting.(17) Meanwhile, protein language models (pLMs), such as ESM-2 and ProtT5, have been trained on millions of protein sequences, from the exceedingly stable to the intrinsically disordered.(18; 19) They capture physicochemical, structural, and functional properties of proteins from their sequence alone, and have even been extended to design novel proteins (20; 21) and binders.(22; 23) However, these models were not trained on fusion oncoprotein sequences, which are functionally and structurally distinct from their wild-type counterparts due to their altered binding sites and unique breakpoint junctions.(24)

To fill this critical gap, we fine-tune the state-of-the-art ESM-2 pLM on over 44,414 fusion oncoprotein sequences collected from the FusionPDB and FOdb databases, collectively termed the new FusOn-DB database.(2; 25) Training on FusOn-DB data, we unfreeze all of the weights of the final eight layers of the ESM-2-650M model and fine-tune these parameters using a masked language modeling (MLM) head. To enhance the model's ability to learn the unique properties of fusion oncoproteins, we introduce a novel cosine-scheduled masking strategy, dynamically varying the masking rate from 15% to 40% during training. This approach enables our top-performing model, FusOn-pLM, to capture the distinct structural and functional features of fusion oncoproteins. As evidence, our results demonstrate that FusOn-pLM outperforms baseline embeddings on diverse fusion-specific tasks, including puncta formation propensity and the prediction of intrinsic disorder. Moreover, we showcase its utility in identifying drug-resistant mutations in fusion oncoproteins, highlighting its biological relevance and potential for advancing therapeutic design.

2 Results

2.1 Fusion oncoproteins comprise a distinct and diverse sequence dataset

ESM-2 was pretrained on 65 million sequences from UniRef50, a database which includes over 9,000 wild-type proteins known to act as the head or tail components of fusion oncoproteins.(26) However, ESM-2 was not trained on the fusions themselves (Figure 1B).(18) By collecting fusion oncoprotein sequences from the FusionPDB and FOdb databases,(2; 25) two complementary resources that provide experimentally validated and computationally predicted fusion proteins with clinical or biological relevance, we assembled FusOn-DB, a comprehensive and non-redundant dataset of 44,414 fusion oncoprotein sequences. Running BLAST between FusOn-DB and SwissProt(27) revealed a wide distribution of sequence homology. On average, fusion oncoproteins shared 71.0% identity with the top-aligning SwissProt sequence, which corresponded to either the head or tail protein in 87% of cases. Over 12,000 fusion oncoproteins had <60% maximum identity, and over 5,000 had <50% maximum identity.

Fusion oncoproteins are also characterized by a high level of structural disorder. AlphaFold2 structures of four highly-studied fusion oncoproteins (PAX3::FOXO1, EWSR1::FLI1, EML4::ALK, and

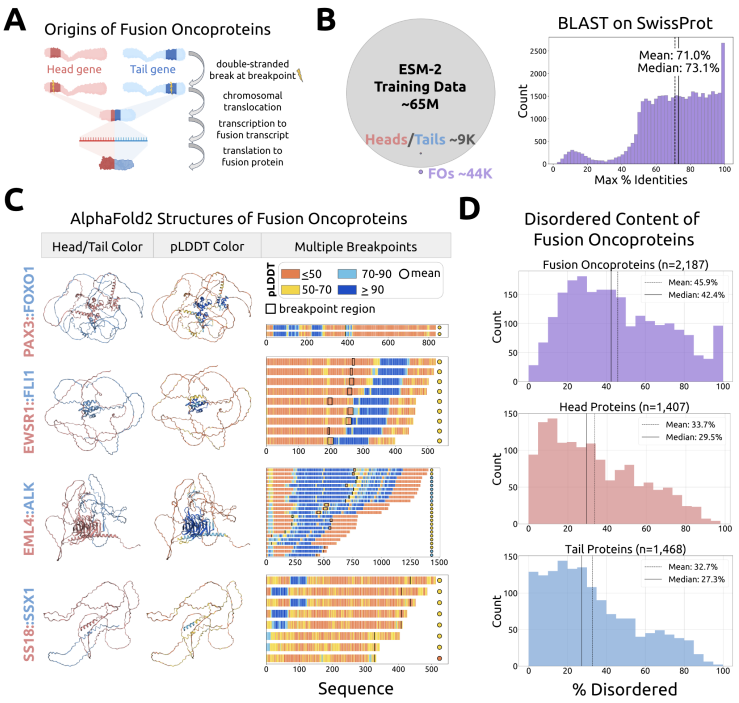

Figure 1: Overview of fusion oncoproteins (FOs). **A** FOs are formed by chromosomal rearrangements between two independent genes, the 5' head gene and 3' tail gene. **B** ESM-2 training data included the wild-type head and tail proteins involved in FOs, but not FOs themselves. FOs were compared to SwissProt, a representative subset of ESM-2's training data, via BLAST. The best alignments for each FO are shown (% identity = total identities / length of FO sequence). **C** AlphaFold2 structures of four well-studied fusion oncoproteins: PAX3::FOXO1, EWSR1::FLI1, EML4::ALK, and SS18::SSX1. Structures are colored by composition (red = head, blue = tail) and pLDDT, AlphaFold2's primary confidence metric. Each FO has multiple known breakpoints, producing different amino acid sequences. Breakpoint regions (rectangle), per-residue pLDDTs (bar coloring), and average pLDDTs (colored circle) are shown for each sequence. **D** The percentage of disordered residues per sequence for FOs and their respective heads and tails. Average disorder content is 45.9% for FOs, 33.7% for head proteins, and 32.7% for tail proteins. Only FOs with AlphaFold2 structures available on FusionPDB are included.

SS18::SSX1) largely exhibit low (50-70) and very low (< 50) confidence pLDDT scores, indicating extensive intrinsic disorder (Figure 1C). These structural trends are consistent across various sequences for the same fusion genes, arising from different breakpoints (Figure 1C). To quantify the difference in disorder between fusion oncoproteins and wild-type proteins, we used a well-validated threshold to assign disorder labels (pLDDT < 68.8 = disordered)(17) to each residue in a set of fusion oncoproteins. Fusion oncoproteins were 45.9% disordered on average, while head proteins were 33.7% disordered and tail proteins were 32.7% disordered (Figure 1D). Similarly to fusion oncoproteins, the gold-standard disorder dataset Disorder-NOX(28) had a greater proportion of near-fully disordered proteins than fusion heads and tails (Supplementary Figure S1). In contrast, fusion oncoproteins had a more right-skewed distribution (Supplementary Figure S1). In total, these findings highlight the distinct sequence and structural characteristics of fusion oncoproteins, underscoring the need for better representations tailored to their properties.

2.2 Cosine-scheduled masking enables accurate fusion oncoprotein sequence recovery

Having curated a diverse dataset of fusion oncoproteins, we sought to fine-tune the standard ESM-2-650M model via an MLM objective (Figure 2A).(18) This classic training approach forces the model to reconstruct masked tokens from sequence context, refining representations to emphasize unique physicochemical properties of fusion oncoproteins. Fixed-rate masking at 15% is the established standard in most BERT-based MLM architectures,(18; 29) but fusion oncoproteins' intrinsic complexity prompted us to explore higher and variable masking rates. Recent findings from Wettig et al.,

have demonstrated that increasing masking rates (up to 40%) improves performance by forcing the model to rely more heavily on sequence context for token reconstruction.(30) Additionally, varying the masking rate during training balances representation learning (improved by lower masking rates) with reconstruction quality (improved by higher masking rates).(31) Motivated by these findings, we fine-tuned the final eight layers of ESM-2-650M using a cosine scheduler to dynamically adjust the masking rate from 15% to 40% across each training epoch (Figure 2A), hypothesizing that this approach would maximize model performance by gradually increasing the difficulty of the reconstruction task.

Our results strongly validated this hypothesis. When evaluated on a 15% masked, held-out test set from FusOn-DB, our fine-tuned model consistently outperformed both fixed-rate masking strategies and the non-fine-tuned ESM-2-650M baseline (Figure 2B). ESM-2-650M, which uses a static 15% masking rate during pre-training,(18) performed poorly, with a loss of 1.83 and a pseudo-perplexity of 6.24. As a note, pseudo-perplexity (pPL) is a metric adapted from language modeling to evaluate how well a model predicts masked tokens, with lower values indicating better reconstruction performance and overall sequence comprehension. While far better, fine-tuning FusOn-pLM with fixed masking rates of 15%, 20%, and 25% produced progressively higher loss and pPL values, reflecting the difficulty of optimizing both sequence reconstruction and representation learning with static masking. In contrast, cosine-scheduled masking achieved better performance across all tested ranges, with the best results observed for a masking range of 15%-40% (loss: 1.29; pPL: 3.62). Further exploration of different adjusted-rate masking schedulers, including log-linear and stepwise strategies, demonstrated that the cosine scheduler still remained optimal, achieving the lowest loss and pPL values (1.28 and 3.61, respectively) (Figure 2B).

2.3 FusOn-pLM generates fusion oncoprotein-relevant representations

To determine if FusOn-pLM produces relevant embeddings, we sought to evaluate its performance on downstream fusion oncoprotein-specific tasks. We first assessed the embeddings' ability to accurately predict the formation and localization of puncta, which are critical in driving cancer pathology.(2) Many fusion oncoproteins have been shown to form puncta via phase separation, and these condensates may localize to the nucleus and/or cytoplasm (Figure 3A).(2) Experimental data describing the puncta formation and localization of 178 fusion oncoproteins were used to train three FusOn-pLM-Puncta models, consisting of FusOn-pLM embeddings fed into a gradient boosting (XGBoost) classifier (Figure 3B). For puncta formation, FusOn-pLM embeddings outperform ESM-2-650M, ProtT5, and FOdb physicochemical embeddings on four relevant classification metrics across the entire held-out test dataset (Figure 3C). We observed similar results when predicting localization to the nucleus, the primary location of fusion oncoproteins (Figure 3D).(3) While manually-curated FOdb embeddings perform strongly on cytoplasm localization prediction, FusOn-pLM embeddings prove most effective on critical metrics, such as AUROC (Figure 3E). In total, these results indicate that FusOn-pLM learns representations capturing key properties encoded in fusion oncoprotein sequences.

2.4 FusOn-pLM can accurately predict disordered content in wild-type and fusion oncoproteins

Given that fusions are structurally disordered, we hypothesized that FusOn-pLM's embeddings may encode information pertinent to the properties of intrinsically disordered regions (IDRs). Specifically, we sought to predict: 1. Asphericity, which quantifies a protein's ensemble shape and molecular conformation, 2. End-to-end radius (R_e), the average distance between the N-terminal and C-terminal residue, 3. Radius of gyration (R_g), the average distance between a protein's residues and its center of mass, and 4. Polymer scaling exponent, which describes an IDR's behavior when solvated in water.(32) Individual FusOn-pLM-IDR regressors were trained on non-fusion IDR sequences for each property, using multi-layer perceptron (MLP) heads to predict the property values directly from FusOn-pLM embeddings (Figure 4A). We demonstrate that FusOn-pLM-IDR models achieve a high coefficient of determination (R2) on all four properties, indicating a strong fit (Figure 4B). We also find that FusOn-pLM and ESM-2-650M embeddings achieve nearly equivalent performance, signaling that FusOn-pLM did not overfit on fusion oncoproteins and lose ESM-2's intrinsic ability to represent a wide range of proteins (Supplementary Figure S2).

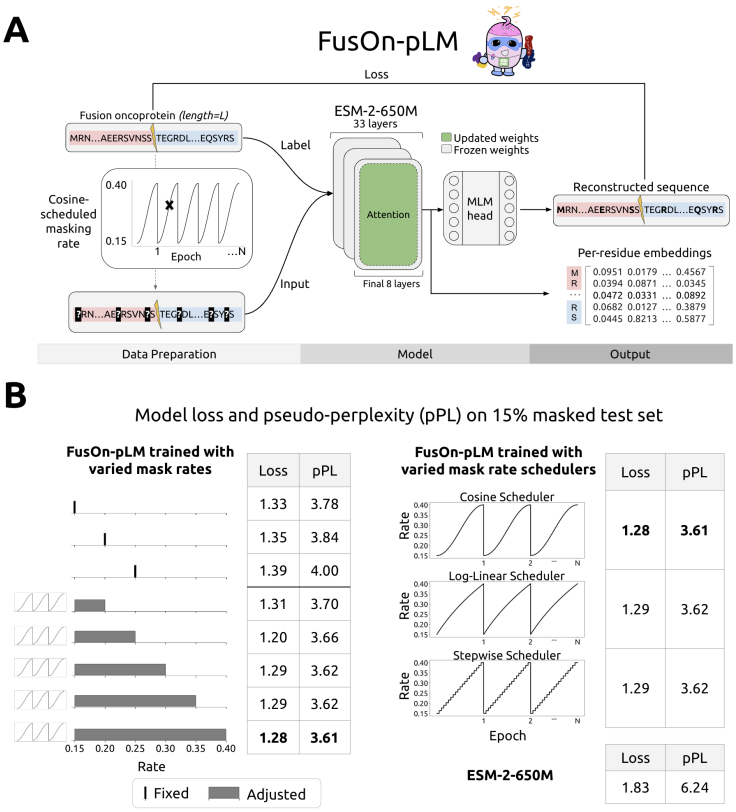

Figure 2: FusOn-pLM **A** Model pipeline. *Data preparation:* Fusion oncoprotein sequences (length L) undergo random masking, where each amino acid has equal likelihood of selection. The masking rate increases from 15% to 40% throughout each epoch according to a cosine scheduler. The masked sequence is fed as input and the original sequence as label into the *model*: 33-layer ESM-2-650M with an MLM head. The final eight layers are unfrozen for fine-tuning. *Output*: the MLM head outputs an attempted reconstruction of the original sequence, which is compared with the label to calculate loss. FusOn-pLM embeddings, of shape $[L, 1280]$, are extracted from the final layer of the ESM-2-650M encoder stack. **B** Test set loss and perplexity (pPL) for various masking strategies. Fixed-rate masking is tested at three rates, and adjusted-rate masking is tested in five ranges. At the top-performing range (15%-40%), three schedulers are tested (cosine, log-linear, stepwise).

Next, we sought to assess FusOn-pLM's ability to identify IDR regions within protein sequences. The FusOn-pLM-Diso model was trained to predict per-residue probabilities of disorder directly from FusOn-pLM embeddings (Figure 4C). When evaluated on the Disorder-NOX dataset used in the CAID2 competition,(33) FusOn-pLM achieved an AUROC of 0.825. Compared with a parallel architecture trained on ESM-2 embeddings (ESM-2-650M-Diso) and fourteen CAID2 competitors, FusOn-pLM-Diso ranked in the top 5 of all models (Figure 4D).(28) We then questioned whether FusOn-pLM embeddings could accurately distinguish between structured and disordered residues in fusion oncoproteins, specifically. On a set of proteins from FusOn-pLM's test set, FusOn-pLM-Diso achieved average accuracy, precision, recall, F1, and AUROC metrics all above 0.9. We also observed a strong correlation (R^2 = 0.83) between the disorder percentages predicted by FusOn-pLM-Diso and that of AlphaFold-pLDDT (Figure 4E), further supporting the notion that FusOn-pLM embeddings capture the disorder properties of fusion oncoproteins. When visualizing the per-residue disorder probabilities for six well-studied fusion oncoproteins, we observe differential coloring between disordered and structured residues. We establish that FusOn-pLM correctly identifies structure in the α-helix and β-sheet-rich regions, coloring these areas dark blue (Figure 4F). Overall, our results suggest that FusOn-pLM accurately encodes disorder-related information in its embeddings. Given that fusion oncoproteins are characterized by their disordered regions, we reason that FusOn-pLM embeddings more effectively represent fusion oncoproteins.

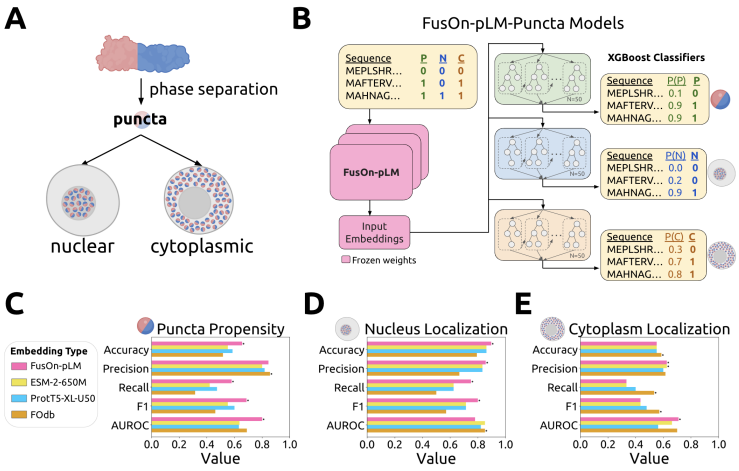

Figure 3: FusOn-pLM embedding benchmarks on puncta prediction tasks. **A** Certain FOs form puncta (condensates) via phase separation. Puncta may localize to the nucleus, cytoplasm, or both. **B** Three XGBoost classifiers are trained on FusOn-pLM-embedded FOs. One predicts formation of puncta (puncta propensity); one predicts formation of nuclear puncta (nucleus localization); one predicts formation of cytoplasmic puncta (cytoplasm localization). **C-E** Performance on a held-out test set when predictors are trained on FusOn-pLM, ESM-2-650M, ProtT5-XL-U50, and FOdb embeddings.

2.5 FusOn-pLM embeddings enable zero-shot discovery of relevant mutations

Fusion oncoproteins themselves are mutants, but they also have the potential to acquire additional mutations which can alter their structure, function, and druggability.(34) Beyond property and disorder prediction, we sought to establish the biological utility and relevance of FusOn-pLM by performing zero-shot discovery via its MLM head, which can sequentially unmask each position in an input sequence, outputting residue probabilities per unmasked position (Figure 5A). As with any pLM, within evolutionarily conserved domains, the logits corresponding to the original residue are much higher than for any alternate residue. For example, in the TF::Kinase fusion TRIM24::RET, FusOn-pLM correctly identifies TRIM24's zinc finger domains and RET's kinase domain as highly conserved (Figure 5B). FusOn-pLM also identifies that the EWSR1 activation domain and FLI1 DNA-binding domain in EWSR1::FLI1 are unlikely to mutate (Figure 5B). In PAX3::FOXO1, the DNA-binding domains of PAX3 are highly conserved, but the truncated DNA-binding domain of FOXO1 (25/75 amino acids) is less strongly conserved (Figure 5B), corroborated by studies showing FOXO1's DNA binding activity is not critical for fusion function.(35; 36) This result indicates that FusOn-pLM has implicitly captured the function of fusion oncoproteins, which is further strengthened by the observation of clear differences between TF::TF and Kinase::Kinase fusions in its latent space (Supplementary S3A).

Although FusOn-pLM may not predict that change is likely within a conserved domain, its logits still provide rank-ordered, possible mutations within these regions. This feature holds promise for discovering potential drug resistance mutations, as small molecule drugs are designed to interact with well-structured, conserved binding pockets like kinase active sites.(37) Fusion oncoprotein mutations causing drug resistance have been identified in a small number of studies on kinase-containing fusions.(38; 39; 40) We sought to determine whether FusOn-pLM prioritizes the resistance-causing mutations discovered in patients with fusion-driven cancers. In EML4::ALK, a set of 14 mutation sites were linked to resistance to at least one of five drugs: Crizotinib, Ceritinib, Alectinib, Brigatinib, and Lorlatinib.(38) FusOn-pLM successfully predicted at least one true resistance mutation among the top three mutation logits for 12/14 sites (Figure 5C). In BCR::ABL, whose sequence is nearly twice as long as EML4::ALK, a set of 28 mutation sites were linked to imatinib resistance.(39) FusOn-pLM recovered drug resistance mutations in 13 of these locations (Figure 5C). Finally, we selected ETV6::NTRK3 as a case study for recovering known drug resistance mutations and investigating potential mutations away from the active site. FusOn-pLM successfully prioritized two resistance mutations in the NTRK3 kinase domain,(40) assigned high conservation probability throughout the kinase domain, and predicted the most volatile positions to be in the disordered region from head protein ETV6 (Figure 5D). In total, these results highlight FusOn-pLM's potential as a biologically-

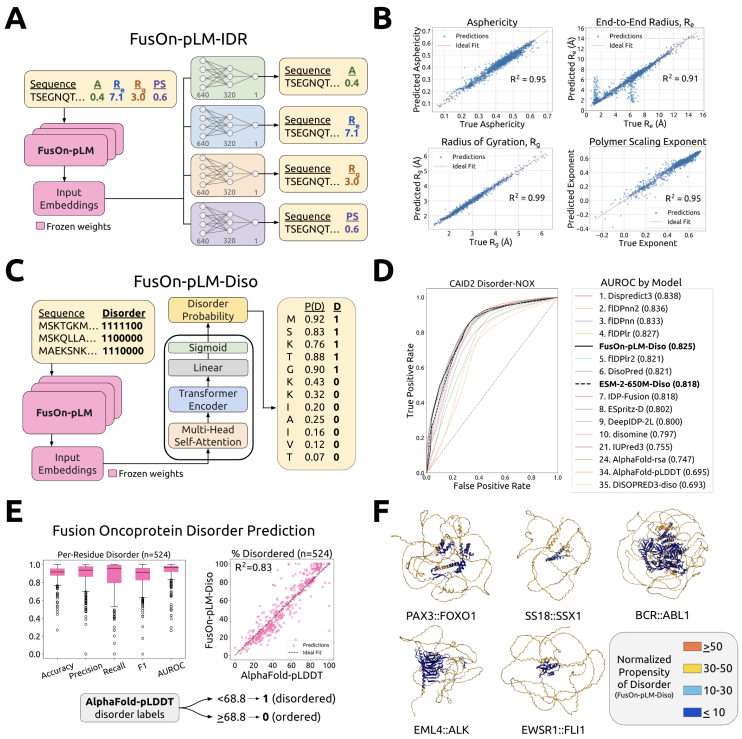

Figure 4: FusOn-pLM prediction of IDR properties and regions. **A** FusOn-pLM-IDR models predict asphericity (A), end-to-end radius (R_e), radius of gyration (R_g), and polymer scaling exponent (PS) by feeding FusOn-pLM embeddings through an MLP classification head. **B** FusOn-pLM-IDR predictions vs. true values. The coefficient of determination (R^2) between predictions and labels was calculated for each model to assess goodness of fit. **C** FusOn-pLM-Diso utilizes a Transformer architecture to predict per-residue disorder labels from FusOn-pLM embeddings. **D** Disorder predictor performance in CAID2 competition when trained on FusOn-pLM vs. ESM-2-650M embeddings.(33) **E** FusOn-pLM-Diso performance on test set fusion oncoproteins, based on AlphaFold-pLDDT-derived disorder labels. The coefficient of determination (R^2) between predictions and labels was calculated for each model to assess goodness of fit. **F** Visualization of FusOn-pLM embedding predictions of disorder propensity on AlphaFold2-predicted structure. Disorder probabilities are shaded according to the legend for interpolation.

relevant tool for predicting resistance mutations both within conserved domains and in disordered regions critical to therapeutic outcomes.

3 Discussion

In this work, we introduce FusOn-pLM, an ESM-2-based pLM fine-tuned to generate fusion oncoprotein-specific embeddings. We further provide a newly-curated, comprehensive dataset, FusOn-DB, consisting of over 44,000 annotated fusion oncoprotein sequences. To our knowledge, no pLM has explicitly sought to learn the unique characteristics of fusion oncoproteins, which differ from most proteins due to their highly disordered nature and altered structural and functional properties. Our benchmarking results establish that via a novel cosine-scheduled MLM training strategy, FusOn-pLM embeddings outperform those of the original ESM-2-650M model,(18) the ProtT5 model,(19) as well baseline FOdb descriptor embeddings,(2) on fusion oncoprotein-related tasks, while retaining distinct representations of fusion proteins from their head and tail counterparts (Supplementary Figure S3B). We further demonstrate that by training on fusion oncoprotein sequences, which represent a large class of IDR-containing proteins, FusOn-pLM embeddings rank highly on the CAID-2 benchmark for IDR detection(33) and strongly predict IDR properties themselves. Finally, as a demonstration of the model's biological relevance, we show that FusOn-pLM uniquely enables

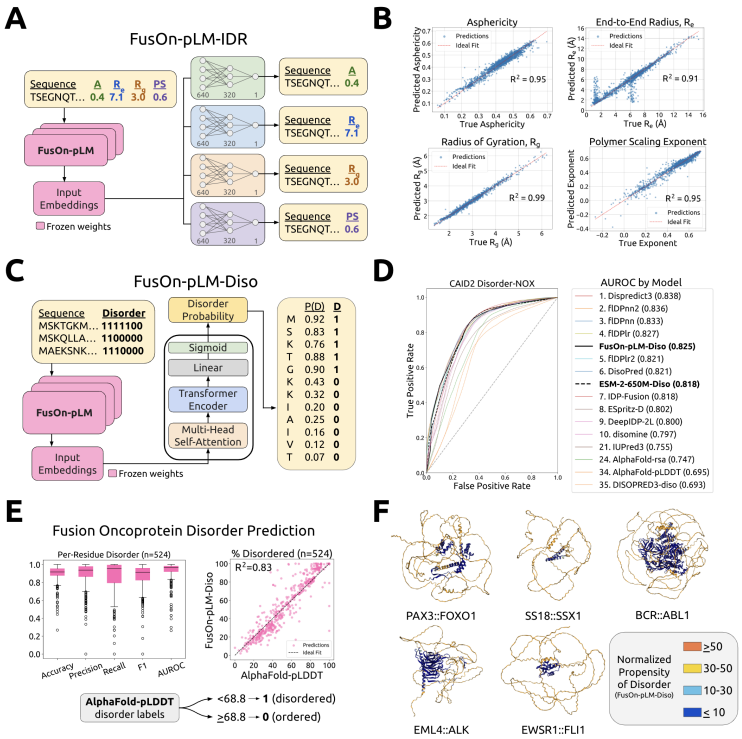

Figure 5: Zero-shot mutation prediction. **A** FusOn-pLM performs zero-shot mutation discovery via its MLM head through sequential unmasking of individual residues. Potential mutations are ranked by their logit values. **B** FusOn-pLM logits for the longest EWSR1::FLI1, PAX3::FOXO1, and TRIM24::RET sequences in FusOn-DB. Yellow regions are considered highly conserved domains. **C** Recovery of mutations found to cause drug resistance in patients with EML4::AK and BCR::ABL1-driven cancers. **D** Case study on kinase fusion ETV6::NTRK3 (647 amino acids), which drives various cancers. FusOn-pLM predictions of NTRK3 kinase domain mutations identified in ETV6::NTRK3+ cancer patients with drug resistance are shown in the table. Based on logit values, disordered residues from the head protein ETV6 are indicated.

the prediction of current and future drug-resistant mutations in fusion oncoproteins, highlighting its potential for informing therapeutic strategies and anticipating resistance mechanisms.

While FusOn-pLM represents an important advancement, there are several limitations to address. First, despite leveraging over 44,000 fusion oncoprotein sequences, the diversity of the FusOn-DB dataset may not fully capture all fusion variants, particularly rare or less well-characterized fusions. Additional data, particularly from emerging databases and clinical studies, would further enhance the model's generalizability. Second, due to GPU memory constraints, proteins longer than 2,000 amino acids were excluded during training. While such cases are rare among known fusion oncoproteins, this limitation may exclude certain outliers with repetitive domains or extensive intrinsically disordered regions. Future optimizations in tokenization or memory-efficient architectures could enable the inclusion of these sequences, ensuring comprehensive coverage of fusion oncoprotein diversity. Third, while FusOn-pLM provides strong predictions for intrinsic disorder and drug-resistant mutations, its ability to predict driver mutations or to connect sequence embeddings with regulatory elements such as enhancers or transcription factors remains unexplored.(34) Future efforts could involve developing models that integrate FusOn-pLM embeddings with regulatory sequence data to elucidate mechanisms underlying oncogenesis.(41) Most importantly, experimental validation of FusOn-pLM's predictions, including drug resistance mechanisms and therapeutic design tasks, will be essential to confirm its utility in practical settings.

Recently, our lab has trained ESM-2-based models to generate peptides provided only the sequence of the target protein, facilitating the design of peptide-E3 ubiquitin ligase fusions for the proteasomal degradation of diverse protein substrates.(22; 23; 42) As our main objective is to enable the degrada-

tion of fusion oncoproteins, our next steps will be to replace ESM-2 embeddings in these models with FusOn-pLM embeddings, enabling fusion-specific degrader design. Since post-translational modifications (PTMs) are also well known to affect the oncogenic activity of fusion oncoproteins (43; 44; 45), we plan to retrain FusOn-pLM with our recent PTM-Mamba pLM,(46) which effectively tokenizes PTMs, enabling both fusion- and PTM-specific therapeutic design. Finally, by leveraging recent advancements in gene delivery, such as lipid nanoparticles (LNPs) and adeno-associated viral (AAV) vectors(47; 48), we envision that fusion-specific biologics may eventually serve as safe and efficacious therapeutics for fusion-positive cancer patients. Overall, the results of our study motivate the use of FusOn-pLM embeddings for downstream fusion oncoprotein design tasks, serving as a major step toward this goal.

References

[1] T. H. Rabbitts, "Chromosomal translocations in human cancer," *Nature*, vol. 372, no. 6502, pp. 143–149, 1994.

[2] S. Tripathi, H. K. Shirnekhi, S. D. Gorman, B. Chandra, D. W. Baggett, C.-G. Park, R. Somjee, B. Lang, S. M. H. Hosseini, B. J. Pioso *et al.*, "Defining the condensate landscape of fusion oncoproteins," *Nature communications*, vol. 14, no. 1, p. 6008, 2023.

[3] S. D. A. Angione, A. Y. Akalu, J. Gartrell, E. P. Fletcher, G. J. Burckart, G. H. Reaman, R. Leong, and C. F. Stewart, "Fusion oncoproteins in childhood cancers: Potential role in targeted therapy," *The Journal of Pediatric Pharmacology and Therapeutics*, vol. 26, no. 6, p. 541–555, Aug. 2021. [Online]. Available: http://dx.doi.org/10.5863/1551-6776-26.6.541

[4] O. Delattre, J. Zucman, B. Plougastel, C. Desmaze, T. Melot, M. Peter, H. Kovar, I. Joubert, P. de Jong, G. Rouleau, A. Aurias, and G. Thomas, "Gene fusion with an ets dna-binding domain caused by chromosome translocation in human tumours," *Nature*, vol. 359, no. 6391, p. 162–165, Sep. 1992. [Online]. Available: http://dx.doi.org/10.1038/359162a0

[5] C. M. Linardic, "Pax3–foxo1 fusion gene in rhabdomyosarcoma," *Cancer Letters*, vol. 270, no. 1, p. 10–18, Oct. 2008. [Online]. Available: http://dx.doi.org/10.1016/j.canlet.2008.03.035

[6] M. J. McBride, J. L. Pulice, H. C. Beird, D. R. Ingram, A. R. D'Avino, J. F. Shern, G. W. Charville, J. L. Hornick, R. T. Nakayama, E. M. Garcia-Rivera, D. M. Araujo, W.-L. Wang, J.-W. Tsai, M. Yeagley, A. J. Wagner, P. A. Futreal, J. Khan, A. J. Lazar, and C. Kadoch, "The ss18-ssx fusion oncoprotein hijacks baf complex targeting and function to drive synovial sarcoma," *Cancer Cell*, vol. 33, no. 6, pp. 1128–1141.e7, Jun. 2018. [Online]. Available: http://dx.doi.org/10.1016/j.ccell.2018.05.002

[7] M. Soda, Y. L. Choi, M. Enomoto, S. Takada, Y. Yamashita, S. Ishikawa, S.-i. Fujiwara, H. Watanabe, K. Kurashina, H. Hatanaka, M. Bando, S. Ohno, Y. Ishikawa, H. Aburatani, T. Niki, Y. Sohara, Y. Sugiyama, and H. Mano, "Identification of the transforming eml4–alk fusion gene in non-small-cell lung cancer," *Nature*, vol. 448, no. 7153, p. 561–566, Jul. 2007. [Online]. Available: http://dx.doi.org/10.1038/nature05945

[8] H. V. Erkizan, Y. Kong, M. Merchant, S. Schlottmann, J. S. Barber-Rotenberg, L. Yuan, O. D. Abaan, T.-h. Chou, S. Dakshanamurthy, M. L. Brown, A. Üren, and J. A. Toretsky, "A small molecule blocking oncogenic protein ews-fli1 interaction with rna helicase a inhibits growth of ewing's sarcoma," *Nature Medicine*, vol. 15, no. 7, p. 750–756, Jul. 2009. [Online]. Available: http://dx.doi.org/10.1038/nm.1983

[9] T. Vital, A. Wali, K. V. Butler, Y. Xiong, J. P. Foster, S. S. Marcel, A. W. McFadden, V. U. Nguyen, B. M. Bailey, K. N. Lamb, L. I. James, S. V. Frye, A. L. Mosely, J. Jin, S. G. Pattenden, and I. J. Davis, "Ms0621, a novel small-molecule modulator of ewing sarcoma chromatin accessibility, interacts with an rna-associated macromolecular complex and influences rna splicing," *Frontiers in Oncology*, vol. 13, Jan. 2023. [Online]. Available: http://dx.doi.org/10.3389/fonc.2023.1099550

[10] P. J. Carter and A. Rajpal, "Designing antibodies as therapeutics," *Cell*, vol. 185, no. 15, pp. 2789–2805, 2022.

[11] B. L. Hie, V. R. Shanker, D. Xu, T. U. Bruun, P. A. Weidenbacher, S. Tang, W. Wu, J. E. Pak, and P. S. Kim, "Efficient evolution of human antibodies from general protein language models," *Nature Biotechnology*, vol. 42, no. 2, pp. 275–283, 2024.

[12] J. M. Ham, M. Kim, T. Kim, S. E. Ryu, and H. Park, "Structure-based de novo design for the discovery of miniprotein inhibitors targeting oncogenic mutant braf," *International Journal of Molecular Sciences*, vol. 25, no. 10, p. 5535, 2024.

[13] S. M. P. Vadevoo, S. Gurung, H.-S. Lee, G. R. Gunassekaran, S.-M. Lee, J.-W. Yoon, Y.-K. Lee, and B. Lee, "Peptides as multifunctional players in cancer therapy," *Experimental & Molecular Medicine*, vol. 55, no. 6, pp. 1099–1109, 2023.

[14] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, p. 583–589, Jul. 2021. [Online]. Available: http://dx.doi.org/10.1038/s41586-021-03819-2

[15] J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C.-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis, and J. M. Jumper, "Accurate structure prediction of biomolecular interactions with alphafold3," *Nature*, May 2024. [Online]. Available: http://dx.doi.org/10.1038/s41586-024-07487-w

[16] J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, N. Hanikel, S. J. Pellock, A. Courbet, W. Sheffler, J. Wang, P. Venkatesh, I. Sappington, S. V. Torres, A. Lauko, V. De Bortoli, E. Mathieu, S. Ovchinnikov, R. Barzilay, T. S. Jaakkola, F. DiMaio, M. Baek, and D. Baker, "De novo design of protein structure and function with rfdiffusion," *Nature*, vol. 620, no. 7976, p. 1089–1100, Jul. 2023. [Online]. Available: http://dx.doi.org/10.1038/s41586-023-06415-8

[17] D. Piovesan, A. M. Monzon, and S. C. Tosatto, "Intrinsic protein disorder and conditional folding in alphafolddb," *Protein Science*, vol. 31, no. 11, p. e4466, 2022.

[18] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli *et al.*, "Evolutionary-scale prediction of atomic-level protein structure with a language model," *Science*, vol. 379, no. 6637, pp. 1123–1130, 2023.

[19] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, and B. Rost, "Prottrans: Toward understanding the language of life through self-supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, p. 7112–7127, Oct. 2022. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2021.3095381

[20] N. Ferruz, S. Schmidt, and B. Höcker, "Protgpt2 is a deep unsupervised language model for protein design," *Nature Communications*, vol. 13, no. 1, Jul. 2022. [Online]. Available: http://dx.doi.org/10.1038/s41467-022-32007-7

[21] A. Madani, B. Krause, E. R. Greene, S. Subramanian, B. P. Mohr, J. M. Holton, J. L. Olmos, C. Xiong, Z. Z. Sun, R. Socher, J. S. Fraser, and N. Naik, "Large language models generate functional protein sequences across diverse families," *Nature Biotechnology*, vol. 41, no. 8, p. 1099–1106, Jan. 2023. [Online]. Available: http://dx.doi.org/10.1038/s41587-022-01618-2

[22] G. Brixi, T. Ye, L. Hong, T. Wang, C. Monticello, N. Lopez-Barbosa, S. Vincoff, V. Yudistyra, L. Zhao, E. Haarer *et al.*, "Salt&peppr is an interface-predicting language model for designing peptide-guided protein degraders," *Communications Biology*, vol. 6, no. 1, p. 1081, 2023.

[23] S. Bhat, K. Palepu, L. Hong, J. Mao, T. Ye, R. Iyer, L. Zhao, T. Chen, S. Vincoff, R. Watson *et al.*, "De novo design of peptide binders to conformationally diverse targets with contrastive language modeling," *bioRxiv*, pp. 2023–06, 2024.

[24] S. K. Verma, K. L. Witkin, A. Sharman, and M. A. Smith, "Targeting fusion oncoproteins in childhood cancers: challenges and future opportunities for developing therapeutics," *JNCI: Journal of the National Cancer Institute*, p. djae075, 2024.

[25] H. Kumar, L.-Y. Tang, C. Yang, and P. Kim, "Fusionpdb: a knowledgebase of human fusion proteins," *Nucleic acids research*, vol. 52, no. D1, pp. D1289–D1304, 2024.

[26] B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu, and U. Consortium, "Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches," *Bioinformatics*, vol. 31, no. 6, pp. 926–932, 2015.

[27] "Uniprot: the universal protein knowledgebase in 2023," *Nucleic acids research*, vol. 51, no. D1, pp. D523–D531, 2023.

[28] M. Necci, D. Piovesan, and S. C. Tosatto, "Critical assessment of protein intrinsic disorder prediction," *Nature methods*, vol. 18, no. 5, pp. 472–481, 2021.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018. [Online]. Available: https://arxiv.org/abs/1810.04805

[30] A. Wettig, T. Gao, Z. Zhong, and D. Chen, "Should you mask 15% in masked language modeling?" *arXiv preprint arXiv:2202.08005*, 2022.

[31] S. S. Sahoo, M. Arriola, Y. Schiff, A. Gokaslan, E. Marroquin, J. T. Chiu, A. Rush, and V. Kuleshov, "Simple and effective masked diffusion language models," *arXiv preprint arXiv:2406.07524*, 2024.

[32] J. M. Lotthammer, G. M. Ginell, D. Griffith, R. Emenecker, and A. S. Holehouse, "Direct prediction of intrinsically disordered protein conformational properties from sequence," *Biophysical Journal*, vol. 123, no. 3, p. 43a, 2024.

[33] A. D. Conte, M. Mehdiabadi, A. Bouhraoua, A. Miguel Monzon, S. C. Tosatto, and D. Piovesan, "Critical assessment of protein intrinsic disorder prediction (caid)-results of round 2," *Proteins: Structure, Function, and Bioinformatics*, vol. 91, no. 12, pp. 1925–1934, 2023.

[34] R. Zhang, L. Dong, and J. Yu, "Concomitant pathogenic mutations and fusions of driver oncogenes in tumors," *Frontiers in Oncology*, vol. 10, p. 544579, 2021.

[35] Y. Asante, K. Benischke, I. Osman, Q. A. Ngo, J. Wurth, D. Laubscher, H. Kim, B. Udhayakumar, M. I. H. Khan, D. H. Chin *et al.*, "Pax3-foxo1 uses its activation domain to recruit cbp/p300 and shape rna pol2 cluster distribution," *Nature Communications*, vol. 14, no. 1, p. 8361, 2023.

[36] L. E. Crose, K. A. Galindo, J. G. Kephart, C. Chen, J. Fitamant, N. Bardeesy, R. C. Bentley, R. L. Galindo, J.-T. A. Chi, C. M. Linardic *et al.*, "Alveolar rhabdomyosarcoma–associated pax3-foxo1 promotes tumorigenesis via hippo pathway suppression," *The Journal of clinical investigation*, vol. 124, no. 1, pp. 285–296, 2014.

[37] P. Cohen, D. Cross, and P. A. Jänne, "Kinase drug discovery 20 years after imatinib: progress and future directions," *Nature reviews drug discovery*, vol. 20, no. 7, pp. 551–569, 2021.

[38] M. Elshatlawy, J. Sampson, K. Clarke, and R. Bayliss, "Eml4-alk biology and drug resistance in non-small cell lung cancer: a new phase of discoveries," *Molecular Oncology*, vol. 17, no. 6, pp. 950–963, 2023.

[39] T. O'Hare, C. A. Eide, and M. W. Deininger, "Bcr-abl kinase domain mutations, drug resistance, and the road to a cure for chronic myeloid leukemia," *Blood, The Journal of the American Society of Hematology*, vol. 110, no. 7, pp. 2242–2249, 2007.

[40] A. Drilon, T. W. Laetsch, S. Kummar, S. G. DuBois, U. N. Lassen, G. D. Demetri, M. Nathenson, R. C. Doebele, A. F. Farago, A. S. Pappo *et al.*, "Efficacy of larotrectinib in trk fusion–positive cancers in adults and children," *New England Journal of Medicine*, vol. 378, no. 8, pp. 731–739, 2018.

[41] C. Vicente-Garcia, B. Villarejo-Balcells, I. Irastorza-Azcarate, S. Naranjo, R. D. Acemel, J. J. Tena, P. W. Rigby, D. P. Devos, J. L. Gomez-Skarmeta, and J. J. Carvajal, "Regulatory landscape fusion in rhabdomyosarcoma through interactions between the pax3 promoter and foxo1 regulatory elements," *Genome Biology*, vol. 18, pp. 1–18, 2017.

[42] T. Chen, S. Pertsemlidis, R. Watson, V. S. Kavirayuni, A. Hsu, P. Vure, R. Pulugurta, S. Vincoff, L. Hong, T. Wang *et al.*, "Pepmlm: Target sequence-conditioned generation of peptide binders via masked language modeling," *ArXiv*, 2023.

[43] L. Yu, I. J. Davis, and P. Liu, "Regulation of ewsr1-fli1 function by post-transcriptional and post-translational modifications," *Cancers*, vol. 15, no. 2, p. 382, Jan. 2023. [Online]. Available: http://dx.doi.org/10.3390/cancers15020382

[44] V. Thalhammer, L. A. Lopez-Garcia, D. Herrero-Martin, R. Hecker, D. Laubscher, M. E. Gierisch, M. Wachtel, P. Bode, P. Nanni, B. Blank *et al.*, "Plk1 phosphorylates pax3-foxo1, the inhibition of which triggers regression of alveolar rhabdomyosarcoma," *Cancer research*, vol. 75, no. 1, pp. 98–110, 2015.

[45] S. Pan and R. Chen, "Pathological implication of protein post-translational modifications in cancer," *Molecular Aspects of Medicine*, vol. 86, p. 101097, Aug. 2022. [Online]. Available: http://dx.doi.org/10.1016/j.mam.2022.101097

[46] Z. Peng, B. Schussheim, and P. Chatterjee, "Ptm-mamba: A ptm-aware protein language model with bidirectional gated mamba blocks," Feb. 2024. [Online]. Available: http://dx.doi.org/10.1101/2024.02.28.581983

[47] X. Hou, T. Zaks, R. Langer, and Y. Dong, "Lipid nanoparticles for mrna delivery," *Nature Reviews Materials*, vol. 6, no. 12, pp. 1078–1094, 2021.

[48] J.-H. Wang, D. J. Gessler, W. Zhan, T. L. Gallagher, and G. Gao, "Adeno-associated virus as a delivery vector for gene therapy of human diseases," *Signal Transduction and Targeted Therapy*, vol. 9, no. 1, p. 78, 2024.

[49] M. Steinegger and J. Söding, "Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets," *Nature Biotechnology*, vol. 35, no. 11, p. 1026–1028, Oct. 2017. [Online]. Available: http://dx.doi.org/10.1038/nbt.3988

[50] Y. Liu, X. Wang, and B. Liu, "Idp–crf: intrinsically disordered protein/region identification based on conditional random fields," *International journal of molecular sciences*, vol. 19, no. 9, p. 2483, 2018.

[51] G. Hu, A. Katuwawala, K. Wang, Z. Wu, S. Ghadermarzi, J. Gao, and L. Kurgan, "fldpnn: Accurate intrinsic disorder prediction with putative propensities of disorder functions," *Nature communications*, vol. 12, no. 1, p. 4438, 2021.

[52] K. Salokas, R. G. Weldatsadik, and M. Varjosalo, "Human transcription factor and protein kinase gene fusions in human cancer," *Scientific Reports*, vol. 10, no. 1, p. 14169, 2020.

[53] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler *et al.*, "Api design for machine learning software: experiences from the scikit-learn project," *arXiv preprint arXiv:1309.0238*, 2013.

[54] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

A Supplemental material

A.1 Methods

A.1.1 Model Training Set Curation

Model training data was curated from FusionPDB and FOdb to create FusOn-DB, a dataset of 44,414 fusion oncoprotein sequences representing 16,364 unique head::tail fusions. FusionPDB contributed 41,456 unique sequences,(25) including AlphaFold2 predictions for 3.5K proteins, while FOdb added 4,537 unique sequences derived largely from patient data.(2) After removing duplicates, sequences longer than 2,000 amino acids were excluded, leaving 42,141 sequences for training. To create train-validation-test splits with low sequence homology, sequences were clustered using MMSeqs2 with a 30% sequence identity and 80% coverage threshold.(49) The test set included 250 sequences: 195 with experimental puncta data from FOdb and sequences for four well-studied fusions (EWSR1::FLI1, PAX3::FOXO1, BCR::ABL1, and EML4::ALK). Clusters overlapping these sequences were manually assigned to the test set, with the remaining clusters split into training (33,719 sequences, 80.01%), validation (4,214 sequences, 10.00%), and testing (4,208 sequences, 9.99%) sets.

A.1.2 BLAST and Breakpoint Mapping

To estimate sequence homology between FusOn-DB and SwissProt, local blastp (v2.16.0) was used. Head and tail gene names from FOdb and FusionPDB were mapped to UniProt IDs using the UniProt ID Mapping tool. Of the 44,414 fusion sequences, 44,257 had both head and tail components mapped, and 157 had one unmapped component (43 head, 114 tail). Both SwissProt and TrEMBL IDs were stored. For each fusion oncoprotein, three alignments were extracted: the top overall alignment, the top alignment corresponding to the head gene, and the top alignment corresponding to the tail gene. Alignments included all isoforms. Maximum percent identity was calculated as the number of identical amino acids in the alignment divided by the length of the fusion sequence. BLAST alignments were also used to determine breakpoints by identifying the indices corresponding to the top head and tail alignments. Overlapping regions were labeled as breakpoint regions, and specific loci were manually annotated where applicable for visualization purposes.

A.1.3 Benchmarking Dataset Curation

To evaluate FusOn-pLM, datasets were curated for three benchmarking tasks: puncta formation and localization, IDR ensemble dimensions, and intrinsic disorder prediction. Data for puncta formation and localization were collected from FOdb,(2) which includes 178 fusion oncoproteins with experimentally validated results. Train-test splits from FOdb were used, with 149 sequences for training and 29 for testing across three tasks: puncta formation propensity, nuclear localization, and cytoplasmic localization. Class distributions were maintained as reported in FOdb.(2) For IDR ensemble dimensions, 47,114 IDR sequences from synthetic and natural proteins were sourced from a published dataset.(32) Labels included asphericity, end-to-end radius (R_e), radius of gyration (R_g), and polymer scaling exponent. Sequences were clustered using MMSeqs2 with a minimum sequence identity of 30% and split into training (80%), validation (10%), and testing (10%) sets.(49) Data distributions were normalized as needed, and sequences with multiple labels for the same property were averaged. Final dataset sizes were 47,114 for asphericity, 42,868 for Re, 22,912 for Rg, and 40,637 for the scaling exponent. For disorder prediction, training data included 5,273 sequences from IDP-CRF(50) and 545 sequences from flDPnn after cleaning and deduplication.(51) The testing dataset comprised 210 gold-standard sequences from the CAID2 Disorder-NOX dataset with per-residue annotations indicating disorder (1) or structure (0).(28; 33) FusOn-pLM-Diso was trained on the combined dataset and benchmarked on Disorder-NOX. To analyze disorder in fusion oncoproteins, pseudo-labels were generated using AlphaFold-pLDDT scores, where residues with pLDDT < 68.8 were labeled as disordered.(17) Structures for 524 fusion oncoproteins in the FusOn-pLM test set were obtained from FusionPDB.(25) The BeautifulSoup package in python was used to scrape FusionPDB for structure download links.

A.1.4 Embedding Exploration Dataset Curation

FusOn-pLM embeddings of transcription factor (TF) and kinase fusions were visualized in 2D plots. To efficiently determine which fusion oncoproteins possessed TF heads and tails or kinase heads and tails, a categorized list of fusion head and tail genes was consulted.(52) 524 fusion oncoproteins from FusOn-DB (364 TF::TF and 231 Kinase::Kinase) were identified.

A.2 Model Architecture and Training

FusOn-pLM is based on ESM-2-650M, a 33-layer transformer model pre-trained on UniRef50, and was fine-tuned to generate fusion oncoprotein-specific representations. To adapt ESM-2-650M for this task without overfitting, the final eight layers of the model were selectively fine-tuned. Specifically, the key, query, and value weight matrices of the self-attention mechanism in these layers were unfrozen, while earlier layers remained fixed. The multi-head self-attention mechanism is parameterized such that the attention output is computed as a weighted sum of values V, where the weights are derived from the scaled dot-product of queries: $Q = W_q h$ and keys: $K = W_k h$. For fine-tuning, the learnable parameters W_q, W_k, and W_v in the last eight layers were updated, enabling task-specific adaptation to fusion oncoproteins while preserving the general-purpose representations learned during pre-training.

Specifically, a cosine-scheduled masking strategy was employed during training to dynamically vary the masking rate.

Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ be the input amino acid sequence of length n. Define M as the set of masked positions such that $|M| = \lceil r \cdot n \rceil$, where the masking rate r varies within each training epoch according to a cosine schedule. The masking rate at step t within an epoch of T steps is given by:

$$r_t = r_{min} + \frac{1}{2}(r_{max} - r_{min})(1 - cos(\frac{t\pi}{T})) \tag{1}$$

where $r_{min} = 0.15$ and $r_{max} = 0.40$. At the start of each epoch, r_t is reset to r_{min}, increasing to r_{max} and cycling back to r_{min} at the beginning of the next epoch.

Masked positions are selected uniformly at random from the set {1, 2, …, n} without replacement. Mathematically, the selection of M is described as:

$$M \sim \text{Uniform}(\{1, 2, \ldots, n\}, \lceil r \cdot n \rceil) \tag{2}$$

All selected positions are replaced with a special mask token. The MLM objective is computed as:

$$\mathcal{L}_{\text{MLM}} = -\sum_{i \in \mathcal{M}} \log P(x_i | x_{\setminus \mathcal{M}}) \tag{3}$$

where x_i is the true amino acid at position i, and $x_{\setminus \mathcal{M}}$ represents the sequence with masked tokens excluded. A visualization of the masking strategy is shown in Figure 1.

FusOn-pLM was trained on one NVIDIA H100 GPU with 80 GB of VRAM each for 30 epochs with batch size of 8 and learning rate of 3e-4. The Adam optimizer was utilized with no weight decay. Only fusion oncoproteins of length 2000 or shorter were used for training; short sequences were padded to this maximal length.

A.2.1 Fusion Oncoprotein Property Benchmarks

Embedding performance on predicting the propensity of puncta formation, as well as predicting if puncta form in the nucleus or cytoplasm, were evaluated. Here, sequences from FOdb with conclusive experimental data on puncta formation were utilized for pLM embedding evaluation.(2) FOdb tested 195 total FOs for puncta formation, but only used the 178 with conclusive results to train the FO-Puncta ML model. Puncta formation and localization predictions were treated as a binary class, where label 0 or 1 represented a lack or presence of puncta formation in a given area. FusOn-pLM embeddings were compared against three others: 1) Base wild-type ESM-2-650M embeddings, 2) ProtT5-XL-UniRef50 embeddings,(19) and 3) FOdb embeddings,(2) which are 25 physicochemical

features manually curated by FOdb for only the 195 proteins. The standard binary cross-entropy loss function was minimized for each task using the XGBoost model with 50 trees via scikit-learn(53). The binary cross-entropy loss is defined as:

$$\text{BCE}(y, \hat{y}) = -\left(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})\right) \tag{4}$$

A.2.2 Disorder Property Benchmark

Disorder properties were evaluated by training regression models that used FusOn-pLM embeddings of IDRs, to predict four ensemble features: asphericity, R_e, R_g, and polymer scaling exponent.(32) For each property, a separate FusOn-pLM-IDR regression model was trained. These models fed FusOn-pLM embeddings through a multi-layer perceptron (MLP) network with three fully connected layers (Figure 4E). The input layer performed dimensionality reduction to hidden dimension 640 and passed the output through a ReLU activation function, followed by layer normalization and dropout regularization with a probability of 0.2. This structure was repeated for two more iterations, shrinking the hidden dimension to 320 and finally culminating in a single neuron: the predicted value of the property. Each model was trained to minimize the mean square error (MSE), and early stopping was implemented to prevent overfitting. The MSE loss function is defined by:

$$\text{MSE}(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{5}$$

Models were evaluated on a held-out test set by predicting each property given the sequence embedding alone. The coefficient of determination (R^2) between predictions and labels was calculated for each model to assess goodness of fit. In order to maximize R^2, a hyperparameter screen across two batch sizes (32, 64) and five learning rates (1e-5, 3e-4, 1e-4, 3e-3, 1e-3) was performed. The true values and predicted values were plotted in Matplotlib, with an ideal fit line included for reference. The entire process was repeated using ESM-2-650M embeddings rather than FusOn-pLM embeddings (Supplementary Figure S2).

A.2.3 CAID Benchmark

FusOn-pLM's ability to predict intrinsic disorder was evaluated using a per-residue disorder prediction benchmark based on the CAID2 Disorder-NOX dataset.(28; 33) Binary labels indicating whether each residue is disordered (1) or structured (0) were used to train FusOn-pLM-Diso, a per-residue disorder predictor. The predictor employs a multi-head self-attention Transformer architecture, minimizing binary cross-entropy loss. Hyperparameter optimization was performed for the number of attention heads (5, 8, 10), Transformer layers (2, 4, 6), and dropout rates (0.2, 0.5). Models were trained for 2 epochs with a learning rate of 5e-5, and optimal hyperparameters were selected by maximizing AUROC. An equivalent model, ESM-2-650M-Diso, was trained using ESM-2-650M embeddings for comparison. Both models were trained and evaluated on the CAID2 Disorder-NOX dataset,(28; 33) with per-residue predictions used for benchmarking. Predicted per-residue disorder probabilities were computed for each input sequence, and binary predictions were made using thresholds selected to optimize classification performance metrics. To extend the analysis to fusion oncoproteins, per-residue disorder predictions were made for sequences with available AlphaFold2 structures.(14) Percentage disorder was calculated by dividing the number of predicted disordered residues by sequence length. Additionally, predicted per-residue disorder probabilities were mapped onto 3D protein structures for visualization. AlphaFold2's pLDDT metric was used as a reference for structural disorder to aid in the assessment of predicted regions.(14)

A.3 Embedding Exploration

To explore how FusOn-pLM embeddings capture the physicochemical and functional properties of fusion oncoproteins, we first conducted a dimensionality reduction analysis on both fusion oncoprotein embeddings and/or their head and tail proteins using Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP)(54) via the umap module. The FusOn-pLM embeddings of six highly-studied fusion oncoproteins (EWSR1::FLI1, PAX3::FOXO1, BCR::ABL1, CIC::DUX4, SS18::SSX1, and EML4::ALK) and their respective head and tail proteins (derived

from the BLAST against SwissProt) were transformed by UMAP and plotted (Supplementary Figure S4A). Additionally, 364 transcription factor (TF), where both head and tail were TFs, and 231 kinase fusions, where both head and tail were kinases, were embedded and plotted in UMAP coordinates (Supplementary Figure S4B).

A.3.1 Zero-Shot Mutation Prediction

Zero-shot mutation prediction was performed on a set of fusion oncoproteins. For each protein, the sequence was input to FusOn-pLM with its MLM head L times, where L is the protein length. During each iteration, a single <mask> token was introduced at a different position in the sequence, and only this position was unmasked. The raw logits for each of the twenty amino acids at the masked position were recorded. These logits were ranked in descending order, creating a list of the most to least likely amino acids predicted at that position. The top three predicted amino acids, based on their logits, were considered the "top 3 mutations."

Heatmaps of the logits for the original amino acid at each position were constructed for representative fusion oncoproteins: EWSR1::FLI1, PAX3::FOXO1, and TRIM24::RET. Functional domains were identified using UniProt annotations for the reviewed SwissProt accession corresponding to the head and tail genes. Residue positions for these domains were converted from their coordinates on the original head or tail protein to their corresponding positions on the fusion protein using string indexing in Python. A binary conservation label was applied to logits, with values < 0.7 designated as non-conserved (0) and values > 0.7 as conserved (1).

Sequences for EML4::ALK and BCR::ABL1 were generously provided by the authors of Elshatlawy, et al.,(38) and O'Hare, et al.,(39) and were screened through the zero-shot mutation pipeline. Positions corresponding to known drug resistance mutations, as reported in the literature, were evaluated to determine whether one of the top three predicted amino acids matched a reported mutation ("hit") or did not ("miss"). For positions where the original amino acid was among the top three predicted tokens, an additional token was included in the analysis. Structural models for these sequences were folded in AlphaFold2 and visualized using PyMOL.

Potential mutations in ETV6::NTRK3 were also predicted using the zero-shot prediction pipeline.(40) Literature-reported mutations in NTRK3 coordinates were converted to the corresponding positions in ETV6::NTRK3 coordinates. For example, NTRK3 G623R and G696A became ETV6::NTRK3 G504A and G431R. These positions were evaluated as "hit" or "miss" based on whether the top three predicted mutations included the correct token. Structural predictions were obtained from FusionPDB and visualized in PyMOL. Additionally, the top five mutations were identified as those with the smallest logits for the original amino acid.

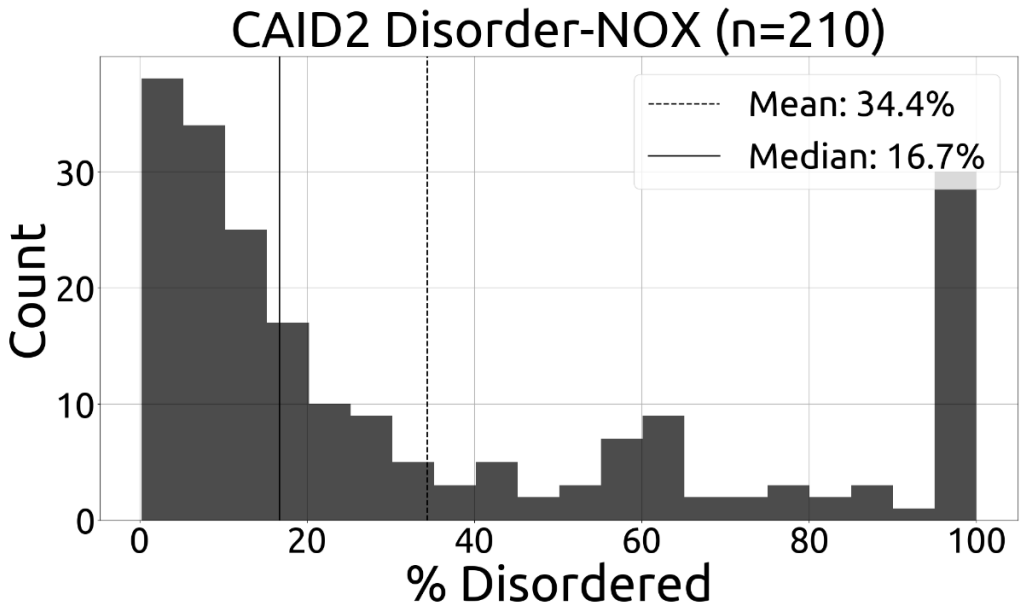

Figure S1: Disorder distribution of CAID2 Disorder-NOX set (210 amino acid sequences). Percentage disordered is defined as the number of residues labeled 1 (disordered), divided by total sequence length.

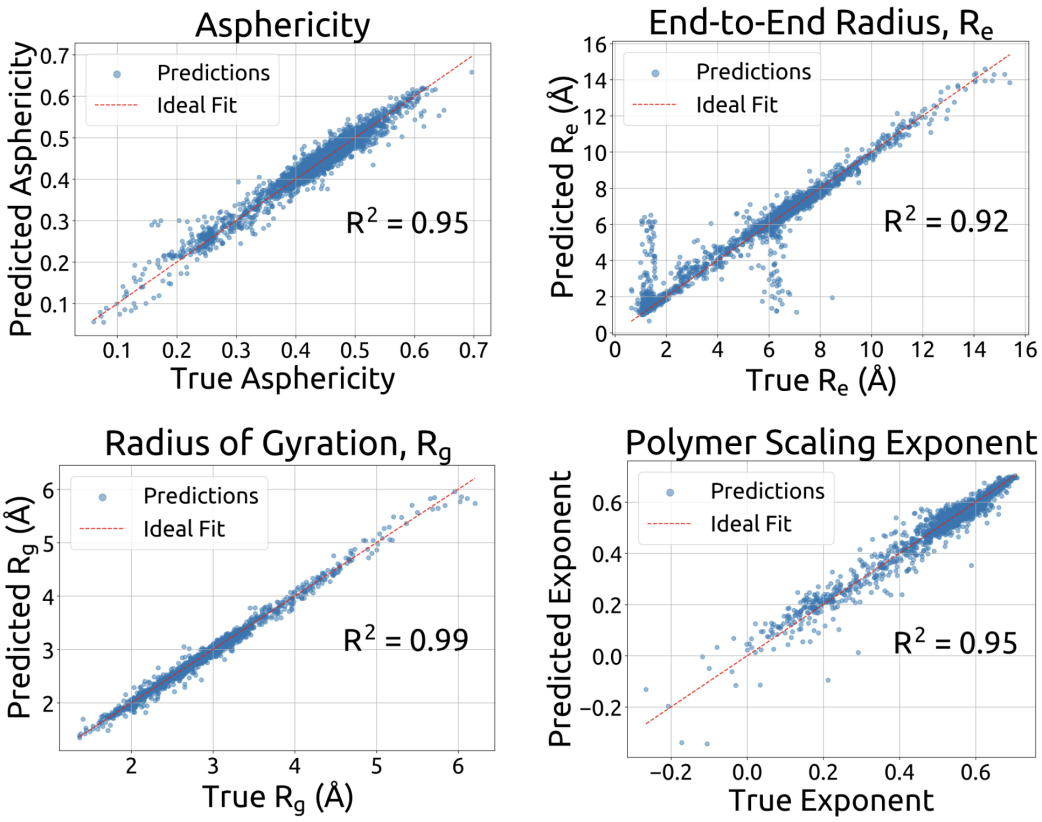

Figure S2: ESM-2-650M-IDR performance on IDR property prediction benchmarking task. Details on data curation and model training for this benchmark can be found in the Methods section of the main manuscript file.

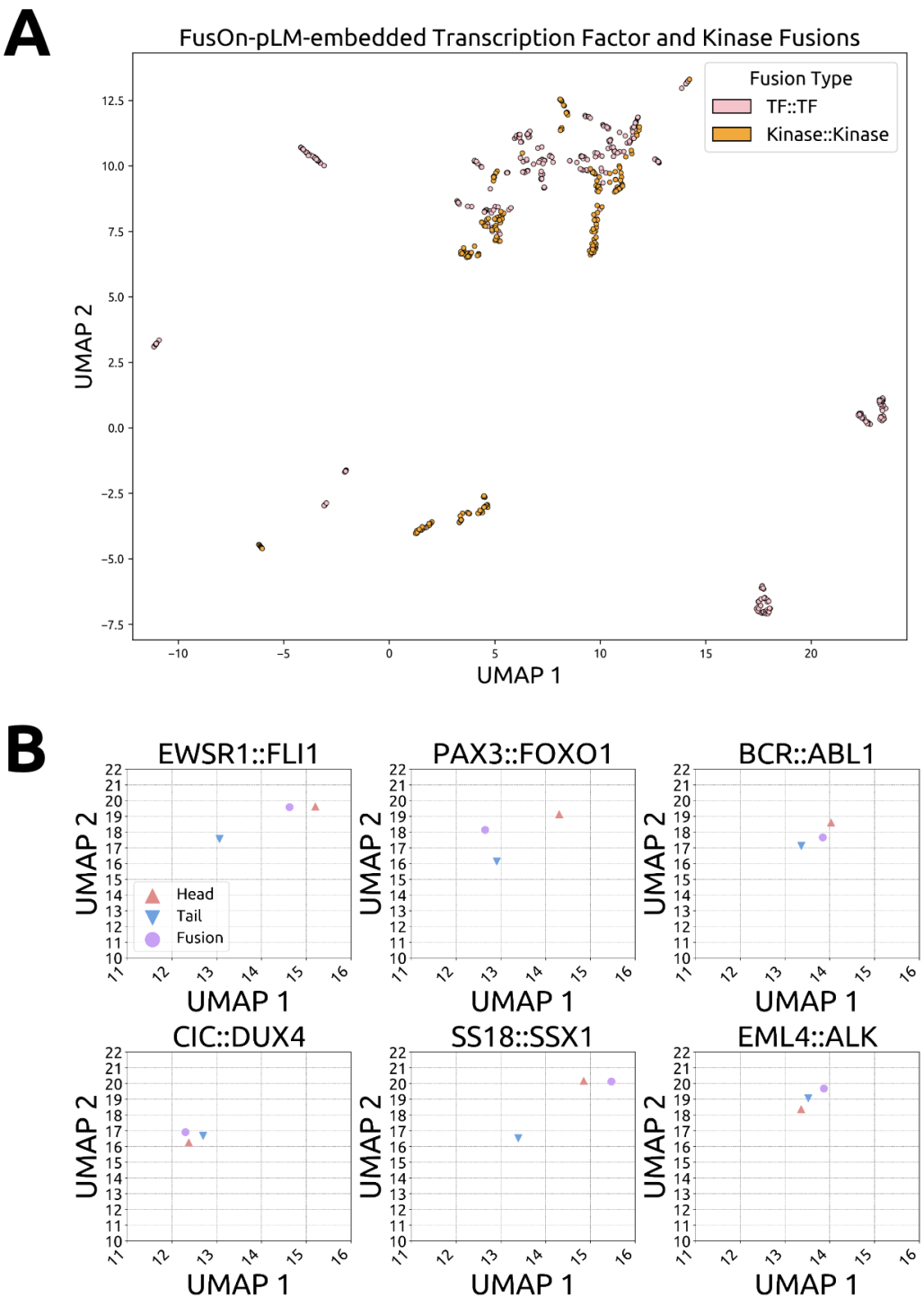

Figure S3: Exploration of FusOn-pLM embeddings for fusion oncoproteins. **A** FusOn-pLM embeddings of the head, tail, and fusion oncoprotein for EWSR1::FLI1, PAX3::FOXO1, BCR::ABL1, CIC::DUX4, SS18::SSX1, and EML4::ALK. **B** FusOn-pLM embeddings of transcription factor (TF) and kinase fusions. TF::TF fusions have TFs as both head and tail; Kinase::Kinase fusions have kinases as both head and tail.

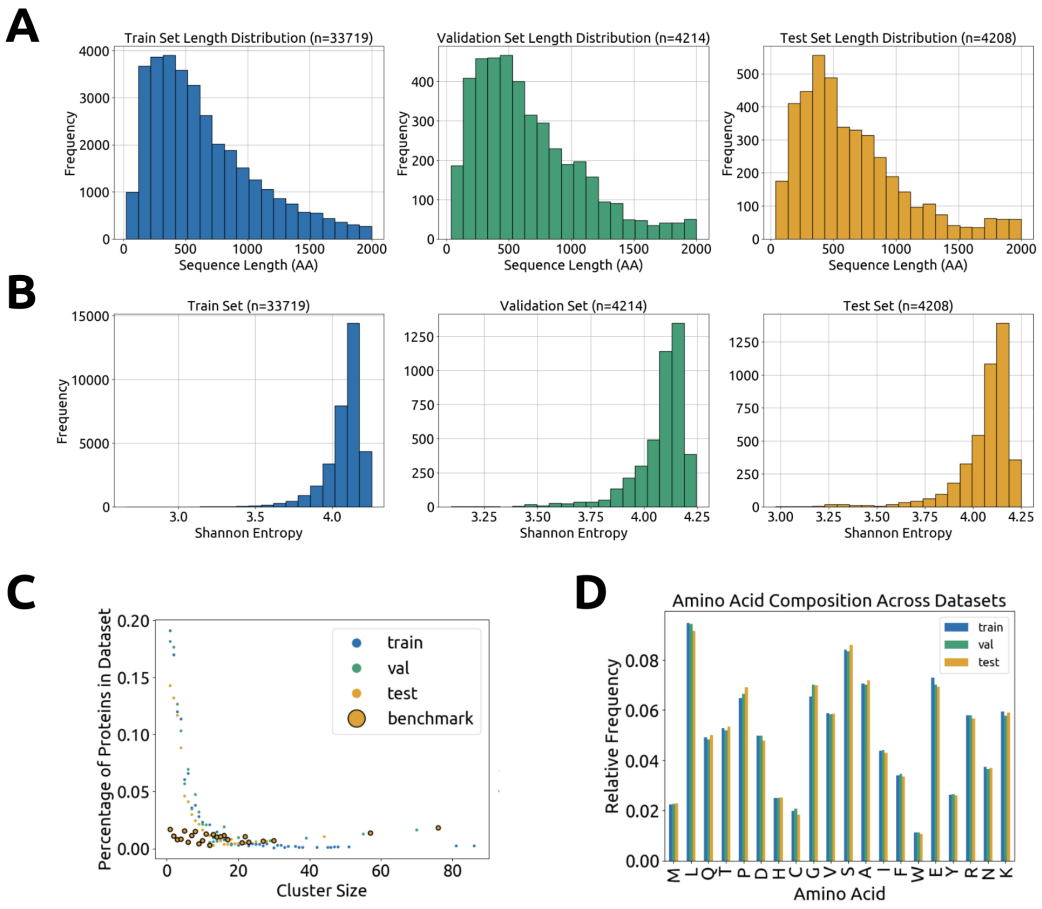

Figure S4: Composition of the FusOn-pLM train, validation and test sets. **A** Distribution of sequence lengths. **B** Distribution of Shannon Entropy scores, which indicate sequence diversity. **C** Clusters for each dataset broken down by cluster size. Benchmark clusters were manually chosen for the test set, while other test clusters were randomly assigned to the test set. **D** Relative frequency of each amino acid across datasets.