FusOn-pLM: A Fusion Oncoprotein-Specific Language Model via Focused Probabilistic Masking

Anonymous Authors[1]

Abstract

Fusion oncoproteins, a class of chimeric proteins arising from chromosomal translocations, drive and sustain various cancers, particularly those impacting children. Unfortunately, due to their intrinsically disordered nature, large size and lack of well-defined, druggable pockets, they have historically been challenging to target therapeutically: neither small molecule-based methods nor structure-based approaches for binder design are strong options for this class of molecules. Recently, protein language models (pLMs) have demonstrated success at representing protein sequences with information-rich embeddings, enabling downstream design applications from sequence alone. However, no current pLM has been trained with fusion oncoprotein sequences and thus may not produce optimal representations for these proteins. In this work, we introduce FusOn-pLM, a novel pLM that fine-tunes ESM-2 embeddings on fusion oncoprotein sequences via masked language modeling (MLM). We specifically introduce a novel MLM strategy, employing a binding-site probability predictor to focus masking on key amino acid residues, thereby generating more optimal fusion oncoprotein-aware ESM-2 embeddings. Our model improves performance on fusion oncoprotein-specific benchmarks in comparison to baseline representations, including biophysical embeddings as well as base ESM-2 embeddings, motivating downstream usage of FusOn-pLM embeddings for therapeutic design tasks targeting these fusions.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the Workshop on Accessible and Efficient Foundation Models for Biological Discovery at ICML 2024. Do not distribute.

1. Introduction

Fusion oncoproteins arise from chromosomal rearrangements that fuse segments of two distinct genes. The resulting mutants contain unrelated functional domains connected by long regions of disorder. This flexible configuration promotes constitutive activation or aberrant regulation of the fusion proteins, driving oncogenic transformation and tumor development. Thousands of unique fusion oncoproteins have been discovered by sequencing patient tumors, and several common culprits such as EWS::FLI1 in Ewing's sarcoma, PAX3::FOXO1 in alveolar rhabdomyosarcoma (ARMS), and MLL-fusion proteins in leukemia 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 (Tripathi et al., 2023). For small molecules that are able to bind fusion oncoproteins, for example EWS::FLI1, these compounds do not achieve strict fusion specificity, binding to one of their head or tail protein counterparts that are often critical transcription factors for cellular homeostasis (Erkizan et al., 2009; Vital et al., 2023). As such, biologics, such as antibodies, miniproteins, and peptides, represent attractive therapeutic alternatives, but necessitate advanced design approaches for targeting to these undruggable proteins.

Recently, structure-based prediction and design models, such as AlphaFold and RFDiffusion (Jumper et al., 2021; Abramson et al., 2024; Watson et al., 2023), 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 (Piovesan et al., 2022). Meanwhile, protein language models (pLMs), such as ESM-2 and ProtT5, have been trained on the amino acid sequences of over 250 million proteins, from the exceedingly stable to the intrinsically disordered (Lin et al., 2023; Elnaggar et al., 2022). They capture physicochemical, structural, and functional properties of proteins from their sequence alone, and have even been extended to designing novel proteins and binders (Brixi et al., 2023; Bhat et al., 2023; Chen et al., 2023). 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.

To fill this critical gap, we fine-tune the state-of-the-art ESM-2 model on over 35,000 fusion oncoprotein sequences collected from the FusionPDB and FOdb databases (Kumar et al., 2024; Tripathi et al., 2023). To do this, we unfreeze the query weights and biases of the final eleven layers of the ESM-2 model and fine-tune these parameters via a masked language modeling (MLM) head (Figure 1). To encourage our model to learn the distinct features of fusion oncoproteins responsible for their function and interaction, we introduce a novel masking strategy, where we apply our recent SaLT&PepPr model to predict and bias masking toward residues most likely to participate in protein-protein interactions (PPIs) (Brixi et al., 2023) (Figure 1). Our results demonstrate that the output embeddings from our SaLT&PepPr-based masking strategy strongly outperform baseline embeddings on diverse fusion oncoprotein-specific tasks, while distinctly representing the fusion oncoproteins from their original head and tail protein counterparts. In total, these results motivate the application of our fusion-specific embeddings for therapeutic design tasks.

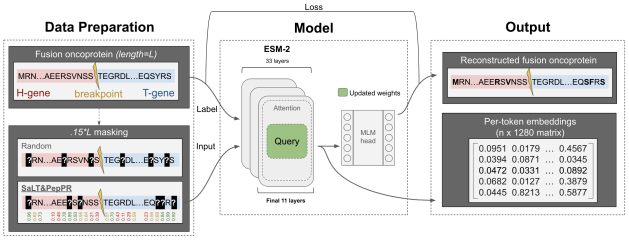

Figure 1. Overview of FusOn-pLM. **Data preparation**: Fusion oncoprotein sequences (length *L*) undergo 15% masking by either: (1) random masking, where each amino acid has equal likelihood of selection, or (2) SaLT&PepPr-based masking, where potential binding sites on the fusion oncoprotein are more likely to be masked. SaLT&PepPr-based masking produced the optimal FusOn-pLM. The masked sequence is fed as input and the original sequence as label into the **model**: 33-layer ESM-2-650M with a MLM head. In the top third of the model (final eleven layers), the query weights are unfrozen for finetuning. **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 encoder stack.

2. Methods

2.1. Amino Acid Masking Strategies

Dataset curation of fusion oncoprotein sequences are described in the Appendix. To force comprehension of physicochemical features of fusion oncoproteins, we employ a focused probabilistic masking strategy on input amino acid sequences. Specifically, we mask 15% of the full sequence, as this percentage has performed well in prior studies (Devlin et al., 2018). Since fusion oncoproteins represent the interaction of two distinct proteins, we masked amino acids that are likely to participate in PPIs as determined by the output probabilities of SaLT&PepPr (Brixi et al., 2023), which predicts a per-amino acid probability of binding. Our masking strategy is as follows:

Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ be the input amino acid sequence of length n, and p_i be the probability that the amino acid x_i participates in a PPI as predicted by SaLT&PepPr. Define M as the set of masked positions such that $|M| = \lceil 0.15n \rceil$.

We select M using the following probabilistic strategy:

1. Compute the probability distribution from SaLT&PepPr: $\boldsymbol{P} = (p_1, p_2, \ldots, p_n)$.

2. Normalize the probabilities to ensure that the sum is 1:

$$\hat{p}_i = \frac{p_i}{\sum_{j=1}^{n} p_j}$$

3. Sample M by selecting $\lceil 0.15n \rceil$ positions according to the normalized probabilities $\hat{\boldsymbol{P}} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n)$.

Mathematically, the selection of M can be described as:

$$M \sim \text{Multinomial}\left(\lceil 0.15n \rceil, \hat{\boldsymbol{P}}\right)$$

Alternatively, for the random 15% masking, we uniformly sample M from the set $\{1, 2, \ldots, n\}$ without replacement:

$$M_{\text{random}} \sim \text{Uniform}\left(\{1, 2, \ldots, n\}, \lceil 0.15n \rceil\right)$$

A visualization of the masking strategy is shown in Figure 1.

2.2. FusOn-pLM

2.2.1. Model Architecture and Training

FusOn-pLM is a fine-tuned encoder on curated fusion oncoprotein sequences trained via a MLM task to create fusion oncoprotein-aware embeddings (Figure 1). To preserve comprehension of wild-type proteins, we train FusOn-pLM with a MLM head on ESM-2-650M (Lin et al., 2023), where amino acid tokens (masked using the respective masking strategy) are passed into ESM-2-650M to retrieve its output embeddings. The MLM loss function $\mathcal{L}_{\text{MLM}}$ is defined as:

$$\mathcal{L}_{\text{MLM}} = -\sum_{i \in \mathcal{M}} \log P(x_i | x_{\setminus \mathcal{M}}) \quad (1)$$

where $\mathcal{M}$ represents the set of masked positions in the input sequence, x_i is the true amino acid token at position i, and $x_{\setminus \mathcal{M}}$ denotes the sequence with the masked tokens excluded.

FusOn-pLM was trained on a NVIDIA H100 GPU with 80 GB of VRAM for 14 epochs with batch size of 8 and learning rate of 5e-5. 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.

To optimize performance while avoiding overfitting on our new sequences, we unfroze only the query weights in a fraction of ESM-2 layers and benchmarked the ensuing models at each epoch (Figure 1). Using random masking, we trained models with a minimum of three and maximum of seventeen unfrozen terminal layers, to avoid sacrificing on batch size (Figure 1).

2.2.2. Benchmarking on Experimental Data

In recent works, certain fusion oncoproteins have been shown to form puncta, which form via phase separation and are a hallmark pathology preceding cancer phenotypes and tumor proliferation (Jiang et al., 2020). To determine if our FusOn-pLM embeddings produce accurate numerical representations of fusion oncoproteins, we evaluated the embeddings' performance on predicting the propensity of puncta formation, and predicting if puncta form in the nucleus or cytoplasm. Here, we utilized 177 sequences from FOdb with experimental data on puncta formation for pLM embedding evaluation (Tripathi et al., 2023). Cancer associations from FusionPDB were further used to evaluate FusOn-pLM's ability to distinguish fusion proteins that drive different malignancies.

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. For the cancer association task, two binary classes were defined for 1,072 test-set proteins: BRCA (class 0) and STAD (class 1). We compare FusOn-pLM embeddings against three others: 1) Base wild-type ESM-2-650M embeddings, 2) FOdb embeddings, which are 25 physicochemical features manually curated by FOdb for these 177 proteins, and finally, 3) Basic one-hot embeddings. We leverage the standard binary cross-entropy loss function and minimize this loss function for each task using the XGBoost model with 50 trees via `scikit-learn` (Buitinck et al., 2013).

3. Results

3.1. Probabilistic masking enables focused training

First, we sought to identify which masking strategy obtains optimal fusion oncoprotein embeddings. Our training results demonstrate that while both SaLT&PepPr-based and random masking produced similar training results with low perplexity values (Table 1), optimal results on preliminary benchmarking were reached before the model converged or displayed evidence of overfitting, indicating that training loss alone cannot be relied upon to choose the final model. As such, our final, optimal model was trained with 11 unfrozen layers using SaLT&PepPr-based masking. By freezing the weights in the remaining 22 layers of ESM-2 and the random MLM head, we enable efficient adaptation to fusion oncoproteins with a small set of trainable parameters. In total, our final FusOn-pLM model consists of 651,163,541 parameters in the ESM-2 encoder stack (18,036,480 of which are trainable parameters) and 1,684,513 parameters in its MLM head.

Table 1. FusOn-pLM perplexities at different training stages. The optimal model, SaLT&PepPr-masked and trained for 14 epochs, does not display minimal perplexities.

Masking	Epoch	Train pPL	Val pPL	Test pPL
SaLT&PepPr 15%	14	4.731	4.827	4.851
SaLT&PepPr 15%	20	4.455	4.598	4.607
Random 15%	14	4.620	4.700	4.840
Random 15%	20	4.342	4.475	4.506

3.2. FusOn-pLM provides fusion oncoprotein-relevant representations

To determine if FusOn-pLM produces relevant embeddings, we next sought to evaluate its performance on downstream fusion oncoprotein-specific tasks. We first assessed the embeddings' ability to accurately predict the propensity and localization of puncta, critical formations driving cancer pathology (Tripathi et al., 2023). From our classification metrics on puncta formation propensity, we demonstrate that FusOn-pLM embeddings strongly outperform ESM-2-650M, FOdb, and one-hot embeddings on all relevant classification metrics across the entire held-out test dataset (Figure 2A), which is also the case for predicting localization to the nucleus, the primary location of fusion oncoproteins (Angione et al., 2021) (Figure 2B). While FOdb embeddings perform strongly on cytoplasm localization prediction, FusOn-pLM proves most effective on the critical AUROC metric (Figure 2C), and comparatively outperforms all other embeddings for the prediction of carcinoma class (Figure 2D). In total, these results indicate that FusOn-pLM learns representations capturing key semantics and properties encoded in fusion oncoprotein sequences.

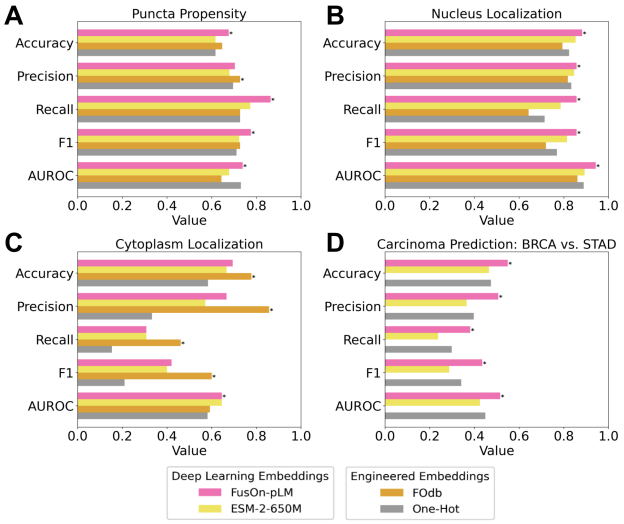

Figure 2. FusOn-pLM embeddings robustly outperform ESM-2-650M in predicting experimentally validated properties of fusion oncoproteins. **A-C)** XGBoost binary classifiers utilize FusOn-pLM, ESM-2-650M, FOdb, and one-hot embeddings to predict **A** propensity of puncta formation, **B** puncta localization to the nucleus, and **C** puncta localization to the cytoplasm. **D)** XGBoost binary classifiers utilize FusOn-pLM, ESM-2-650M, and one-hot embeddings to classify fusion oncoproteins as causing BRCA (breast invasive carcinoma) or STAD (stomach adenocarcinoma). FOdb embeddings not available.

3.3. FusON-pLM embeddings discriminate fusion oncoprotein from head and tail proteins

The primary objective of FusOn-pLM is to provide feature-rich but distinct representations of fusion oncoproteins, which will enable fusion-specific binder design applications. Given this aim, we visualize FusOn-pLM embeddings in a two-dimensional context to concretely assess the model's capability in achieving embedding differentiation (Figure 3). Via t-SNE visualization of the generated embeddings, we clearly observe distinct separation between FusOn-pLM fusion embeddings and embeddings of the head and tail proteins for well-studied fusion oncoproteins EWS::FLI1, PAX3::FOXO1, BCR::ABL1, CIC::DUX4, SS18::SSX1, and EML4::ALK. The distance between the final embeddings suggest that FusOn-pLM learns fusion oncoprotein-specific information in its embeddings that yield distinct, yet accurate, numerical representations of these sequences (Figure 3).

4. Discussion

In this work, we introduce FusOn-pLM, the first protein language model (pLM) fine-tuned to specifically represent fusion oncoproteins. To our knowledge, no pLM has explicitly sought to learn unique characteristics of fusion oncoproteins, which differ from most proteins due to their highly disordered nature and altered structural and functional properties driving oncogenic transformation. Our benchmarking results demonstrate that FusOn-pLM embeddings outperform those of the original ESM-2-650M model (Lin et al., 2023), as well baseline FOdb descriptor embeddings (Tripathi et al., 2023), on fusion oncoprotein-related tasks, and retain distinct representations of fusion proteins from their head and tail counterparts. While FOdb embeddings do perform strongly on certain tasks, such as cytoplasm localization, their inherent static nature precludes their application to design tasks via methods such as contrastive learning, autoregressive generation, and diffusion.

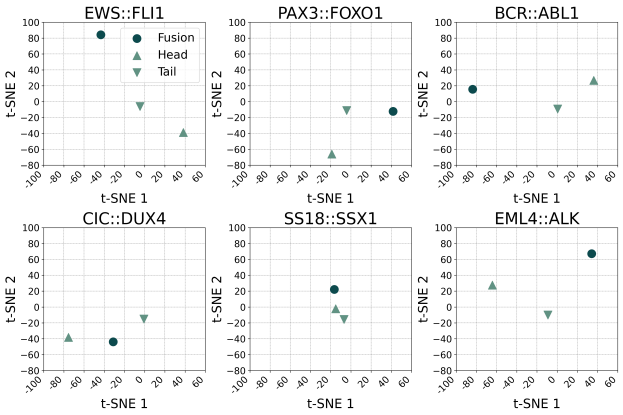

Figure 3. FusOn-pLM embeddings distinguish fusion oncoproteins from their constituent parts (Head and Tail). Six of the most common fusion oncoproteins are included: EWS::FLI, PAX3::FOXO1, BCR::ABL1, CIC::DUX4, SS18::SSX1, EML4::ALK.

Recently, our lab has trained ESM-2-based models to generate peptides given only the sequence of the target protein, facilitating the design of peptide-guided E3 ubiquitin ligases for target-specific proteasomal degradation (Bhat et al., 2023; Chen et al., 2023). As such, our next steps will be to replace ESM-2 embeddings in these models with FusOn-pLM embeddings, enabling fusion-specific degrader design. By leveraging recent advances in gene delivery, such as lipid nanoparticles (LNPs) and adeno-associated viral (AAV) vectors, 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

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C.-C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., Beattie, C., Bertolli, O., Bridgland, A., Cherepanov, A., Congreve, M., Cowen-Rivers, A. I., Cowie, A., Figurnov, M., Fuchs, F. B., Gladman, H., Jain, R., Khan, Y. A., Low, C. M. R., Perlin, K., Potapenko, A., Savy, P., Singh, S., Stecula, A., Thillaisundaram, A., Tong, C., Yakneen, S., Zhong, E. D., Zielinski, M., Žídek, A., Bapst, V., Kohli, P., Jaderberg, M., Hassabis, D., and Jumper, J. M. Accurate structure prediction of biomolecular interactions with alphafold3. *Nature*, May 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07487-w. URL http://dx.doi.org/10.1038/s41586-024-07487-w.

Angione, S. D. A., Akalu, A. Y., Gartrell, J., Fletcher, E. P., Burckart, G. J., Reaman, G. H., Leong, R., and Stewart, C. F. Fusion oncoproteins in childhood cancers: Potential role in targeted therapy. *The Journal of Pediatric Pharmacology and Therapeutics*, 26 (6):541–555, August 2021. ISSN 1551-6776. doi: 10.5863/1551-6776-26.6.541. URL http://dx.doi.org/10.5863/1551-6776-26.6.541.

Bhat, S., Palepu, K., Yudistyra, V., Hong, L., Kavirayuni, V. S., Chen, T., Zhao, L., Wang, T., Vincoff, S., and Chatterjee, P. De novogeneration and prioritization of target-binding peptide motifs from sequence alone. June 2023. doi: 10.1101/2023.06.26.546591. URL http://dx.doi.org/10.1101/2023.06.26.546591.

Brixi, G., Ye, T., Hong, L., Wang, T., Monticello, C., Lopez-Barbosa, N., Vincoff, S., Yudistyra, V., Zhao, L., Haarer, E., et al. Salt&peppr is an interface-predicting language model for designing peptide-guided protein degraders. *Communications Biology*, 6(1):1081, 2023.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., and Varoquaux, G. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108–122, 2013.

Chen, T., Pertsemlidis, S., Watson, R., Kavirayuni, V. S., Hsu, A., Vure, P., Pulugurta, R., Vincoff, S., Hong, L., Wang, T., Yudistyra, V., Haarer, E., Zhao, L., and Chatterjee, P. Pepmlm: Target sequence-conditioned generation of peptide binders via masked language modeling, 2023. URL https://arxiv.org/abs/2310.03842.

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

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., and Rost, B. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, October 2022. ISSN 1939-3539. doi: 10.1109/tpami.2021.3095381. URL http://dx.doi.org/10.1109/TPAMI.2021.3095381.

Erkizan, H. V., Kong, Y., Merchant, M., Schlottmann, S., Barber-Rotenberg, J. S., Yuan, L., Abaan, O. D., Chou, T.-h., Dakshanamurthy, S., Brown, M. L., Üren, A., and Toretsky, J. A. A small molecule blocking oncogenic protein ews-fli1 interaction with rna helicase a inhibits growth of ewing's sarcoma. *Nature Medicine*, 15(7): 750–756, July 2009. ISSN 1546-170X. doi: 10.1038/nm.1983. URL http://dx.doi.org/10.1038/nm.1983.

Jiang, S., Fagman, J. B., Chen, C., Alberti, S., and Liu, B. Protein phase separation and its role in tumorigenesis. *Elife*, 9:e60264, 2020.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. Highly accurate protein structure prediction with alphafold. *Nature*, 596 (7873):583–589, July 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL http://dx.doi.org/10.1038/s41586-021-03819-2.

Kumar, H., Tang, L.-Y., Yang, C., and Kim, P. Fusionpdb: a knowledgebase of human fusion proteins. *Nucleic acids research*, 52(D1):D1289–D1304, 2024.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130, 2023.

Piovesan, D., Monzon, A. M., and Tosatto, S. C. Intrinsic protein disorder and conditional folding in alphafolddb. *Protein Science*, 31(11):e4466, 2022.

Steinegger, M. and Söding, J. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, October 2017. ISSN 1546-1696. doi: 10.1038/nbt.3988. URL http://dx.doi.org/10.1038/nbt.3988.

Tripathi, S., Shirnekhi, H. K., Gorman, S. D., Chandra, B., Baggett, D. W., Park, C.-G., Somjee, R., Lang, B., Hosseini, S. M. H., Pioso, B. J., et al. Defining the condensate landscape of fusion oncoproteins. *Nature communications*, 14(1):6008, 2023.

Vital, T., Wali, A., Butler, K. V., Xiong, Y., Foster, J. P., Marcel, S. S., McFadden, A. W., Nguyen, V. U., Bailey, B. M., Lamb, K. N., James, L. I., Frye, S. V., Mosely, A. L., Jin, J., Pattenden, S. G., and Davis, I. J. 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*, 13, January 2023. ISSN 2234-943X. doi: 10.3389/fonc.2023.1099550. URL http://dx.doi.org/10.3389/fonc.2023.1099550.

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Hanikel, N., Pellock, S. J., Courbet, A., Sheffler, W., Wang, J., Venkatesh, P., Sappington, I., Torres, S. V., Lauko, A., De Bortoli, V., Mathieu, E., Ovchinnikov, S., Barzilay, R., Jaakkola, T. S., DiMaio, F., Baek, M., and Baker, D. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, July 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06415-8. URL http://dx.doi.org/10.1038/s41586-023-06415-8.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120, September 2013. ISSN 1546-1718. doi: 10.1038/ng.2764. URL http://dx.doi.org/10.1038/ng.2764.

Appendix

Dataset curation

Model training data was curated from the FusionPDB and the Fusion Oncoprotein Database (FOdb) (Kumar et al., 2024; Tripathi et al., 2023). Specifically, 41,420 FusionPDB and 4,536 FOdb unique amino acid sequences containing only the 20 natural amino acids were collected for downstream model training. Proteins longer than 2000 amino acids were removed due to GPU memory limits. 1,308 duplicates from database overlap were removed, and 177 FOdb sequences were held out for benchmarking tasks. All remaining sequences were clustered using MMSeqs2 easy clustering module with a minimum sequence identity threshold of 30% and a coverage threshold of 80% (Steinegger & Söding, 2017). The resulting clusters were split at 80/10/10 train/test/val ratio into a training set (31,788 proteins, 79.8%), validation set (4,030 proteins, 10.1%), and testing set (4,013 proteins, 10.1%).

Datasets for the three puncta-related benchmarking tasks were collected from FOdb (Tripathi et al., 2023). 177 FOdb sequences were held out for three classification tasks concerning the tendency of fusion oncoproteins to form condensates (puncta) and the cellular localizations of these puncta. These sequences were clustered using MMSeqs2 easy clustering module with a minimum sequence identity threshold of 30% and a coverage threshold of 30% (larger coverage thresholds led to formation of very few clusters). For each task, the clusters were split at 80/20 ratio into train and test sets with similar ratios of class 0 to class 1. For puncta propensity of formation, there were 143 train sequences (80.8% of total; 35.7%-64.3% class 0-1) and 34 test sequences (19.2% of total; 35.3%-64.7% class 0-1). For puncta localization to the nucleus, there were 143 train sequences (80.8%; 59.4%-40.6% class 0-1) and 34 test sequences (19.2%; 58.8%-41.2% class 0-1). For puncta localization to the cytoplasm, there were 141 train sequences (79.7%; 64.5%-35.5% class 0-1)) and 36 test sequences (20.3%; 63.9%-36.1% class 0-1)).

The fourth benchmarking task involved predicting fusion oncoprotein disease outcomes. Cancer associations for the test set (4,013 proteins) were extracted from FusionPDB (Kumar et al., 2024). This data was originally collected from The Cancer Genome Atlas (TCGA), which provided full definitions of each cancer acronym (Weinstein et al., 2013). The top two cancer types were breast invasive carcinoma (BRCA, 583 sequences) and stomach adenocarcinoma (STAD, 489 sequences). Fusion oncoproteins causing these diseases were extracted and clustered using MMSeqs2 easy clustering module with a minimum sequence identity threshold of 30% and a coverage threshold of 80%. These clusters were split into train and test sets: 859 train (80.13%; 54.4%-45.6% BRCA-STAD), 213 test (19.87%; 54.5%-45.5% BRCA-STAD).