Decoupled Representation Learning Improves Generalization in CRISPR Off-Target Prediction

Nyla Bhargava
Department of Biomedical Engineering
SRM Institute of Science and Technology
Chennai, India
bhargavanyla@gmail.com

Aditya Goswami
Department of Biotechnology
SRM Institute of Science and Technology
Chennai, India
adityagoswami0536@gmail.com

***Abstract*—Computational prediction of CRISPR-Cas9 off-target activity is a critical step toward the safe and scalable design of guide RNAs. Most existing deep learning models are trained on large proxy datasets derived from high-throughput assays; however, their performance often degrades when evaluated on experimentally validated off-target sites. This limitation raises concerns about the generalization ability of proxy-trained models and highlights the need for architectures that better transfer to biologically relevant benchmarks such as TrueOT.**

We propose a two-stage deep learning framework that decouples sequence representation learning from off-target classification. In the first stage, guide RNA sequences are encoded using pretrained transformer-based embeddings learned from large genomic corpora. In the second stage, a hybrid neural network integrates these embeddings with mismatch-level and sequence-pair features to predict off-target activity. Models are trained exclusively on a proxy dataset and evaluated under strict external validation on the TrueOT benchmark.

Our results show that incorporating pretrained sequence embeddings substantially improves generalization performance. The full two-stage model achieves consistent improvements in ROC-AUC and PR-AUC on TrueOT, relative to proxy-only baselines, across repeated runs. We demonstrate a consistent generalization effect under distribution shift, robust to training noise, seeds, and evaluation variance. These findings provide empirical evidence that representation transfer plays a critical role in bridging the gap between proxy assay data and experimentally validated off-target behavior.

***Index Terms*—CRISPR-Cas9, off-target prediction, transfer learning, transformer embeddings, TrueOT**

I. Introduction

The CRISPR-Cas9 system has become a transformative tool for genome editing in biomedical research, agriculture, and emerging therapeutic applications. Despite its versatility, unintended off-target cleavage remains a major safety concern [10], particularly in clinical contexts where even rare editing errors may have severe consequences. Accurate computational prediction of CRISPR off-target activity is essential for guide RNA design and risk assessment [16].

Early approaches to off-target prediction relied on heuristic scoring functions based on sequence similarity and mismatch penalties. While computationally efficient, these methods fail to capture the complex, nonlinear interactions governing Cas9 binding and cleavage. Recent advances in deep learning have enabled more expressive models that learn hierarchical representations directly from sequence data, resulting in improved predictive performance on large training datasets [10]. These models often integrate sophisticated architectures to account for data sparsity and imbalance, which are common challenges in CRISPR-Cas9 research due to the high cost of experimental validation [1].

However, a critical limitation of current deep learning approaches lies in their reliance on proxy datasets derived from high-throughput experimental assays. Although such datasets provide the scale required for training complex models, they do not necessarily reflect the biological properties of low-throughput, experimentally validated off-target events observed in vivo. As a result, models optimized on proxy data often exhibit limited generalization when evaluated on curated benchmarks composed of high-confidence experimental measurements.

The TrueOT benchmark was introduced to address this evaluation gap by aggregating experimentally validated off-target sites from multiple studies into a standardized test set for generalization analysis [10], [14]. TrueOT evaluations have revealed that increasing model complexity does not necessarily translate to improved real-world performance, and that naively trained deep models may overfit proxy assay characteristics [18]. These observations suggest that improved transfer mechanisms are required to bridge the distributional gap between large-scale proxy off-target data and experimentally validated off-target behavior.

In this work, we investigate the role of transfer learning in improving the generalization of CRISPR off-target prediction. We introduce a two-stage framework in which sequence representation learning is performed independently of downstream classification. The first stage leverages a pretrained transformer model to generate fixed-length embeddings for guide RNA sequences, capturing contextual sequence information learned from large-scale genomic data. The second stage combines these embeddings with sequence-pair and mismatch-level features using a hybrid neural architecture to predict off-target activity.

We evaluate our approach by training exclusively on a proxy dataset and performing external validation on the TrueOT benchmark. Through controlled ablation experiments, we demonstrate that models trained without pretrained embeddings underperform on TrueOT despite achieving rea-

sonable proxy-dataset performance. In contrast, incorporating pretrained sequence representations yields consistent improvements in both AUC and AUPR on the TrueOT benchmark, indicating enhanced generalization.

Unlike prior models that fine-tune pretrained encoders end-to-end or entangle representation learning with task optimization, our framework deliberately freezes the pretrained representation and isolates its contribution via controlled ablation on TrueOT. This design enables explicit analysis of transfer under distribution shift, which has received limited systematic evaluation in prior CRISPR off-target prediction work. This design choice enables a controlled analysis of representation transfer, reduces overfitting to proxy assay artifacts, and facilitates rigorous external evaluation on the TrueOT benchmark.

The contributions of this work are threefold:

1) We propose a modular two-stage framework that separates representation learning from CRISPR off-target classification.
2) We empirically demonstrate that pretrained transformer-based sequence embeddings improve generalization to experimentally validated off-target datasets.
3) We provide a fully reproducible implementation and evaluation pipeline to support transparent benchmarking and future research in CRISPR off-target prediction.

II. Methods

A. Problem Formulation

Given a CRISPR guide RNA sequence and a candidate off-target DNA sequence, the task is to predict whether the off-target site will be cleaved by the CRISPR-Cas9 complex. This is formulated as a binary classification problem, where positive labels correspond to experimentally observed off-target cleavage events and negative labels correspond to non-cleaved or low-activity sites. The primary challenge lies in achieving robust generalization from large-scale screening datasets to experimentally validated off-target benchmarks such as TrueOT.

B. Overview of the Two-Stage Framework

We propose a two-stage deep learning framework that decouples sequence representation learning from off-target classification. The motivation for this design is grounded in prior observations that models trained exclusively on proxy datasets often fail to generalize to curated experimental benchmarks, and that transfer learning can mitigate this gap.

In the proposed framework:

- Stage-1 learns transferable representations of gRNA sequences using a pretrained transformer model.
- Stage-2 integrates these representations with pairwise sequence and mismatch features using a hybrid neural architecture to perform off-target prediction.

An ablation setting in which Stage-1 is removed is used to quantify the contribution of pretrained representations. This modular design allows for independent optimization of each stage and facilitates the incorporation of advanced language models for sequence feature extraction [8].

This approach contrasts with previous end-to-end models that conflate representation learning and classification, often leading to overfitting to proxy data characteristics and poor generalization to experimentally validated benchmarks [18]. Furthermore, it offers a more interpretable framework for understanding the impact of specific sequence features on off-target activity, diverging from black-box approaches [2] This decoupled architecture enables the selective application of transfer learning techniques, which have proven beneficial in enhancing prediction power, particularly for small-scale datasets, and in improving accuracy and generalizability over existing methods [16].

C. Stage-1: Pretrained gRNA Sequence Embeddings

In the first stage, each unique gRNA sequence is encoded into a fixed-length vector using a pretrained transformer model trained on large-scale genomic data.

Stage-1 embeddings were obtained from a DNABERT-style pretrained genomic language model [13] pretrained on large-scale genomic corpora and were not fine-tuned on any proxy or TrueOT data, ensuring that no information leakage from the evaluation benchmark occurs. DNABERT was selected because of its ability to capture k-mer-level genomic semantics learned from large genomic corpora [13].The pretrained encoder is kept frozen during Stage 2 training to prevent overfitting to proxy-specific noise and to avoid implicit leakage of proxy assay biases into the downstream classifier. This design enables a controlled evaluation of representation transfer and allows the contribution of pretrained embeddings to be isolated through ablation analysis.

To ensure reproducibility and efficiency, embeddings are computed once for all unique gRNA sequences and reused during Stage-2 training and evaluation. In the ablation setting without Stage-1, gRNA embeddings are replaced with fixed-dimension zero vectors, ensuring that architectural capacity remains constant while removing transfer learning effects. This ensures a direct comparison of models with and without pretrained representations, highlighting the specific benefits of transfer learning in enhancing generalization to novel off-target datasets.

D. Stage-2: Hybrid Off-Target Prediction Model

Figure 1 illustrates the Stage-2 hybrid architecture used for off-target classification.

Each gRNA–target pair is encoded as a multi-channel sequence tensor of length L, combining one-hot nucleotide representations of the guide RNA and off-target DNA sequence, a binary mismatch indicator, and a scalar encoding of protospacer adjacent motif (PAM) proximity. These features are concatenated along the channel dimension to form an $L \times F$ input tensor, where F denotes the total number of feature channels. The input tensor is first projected to a fixed embedding dimension via a linear transformation and processed by two one-dimensional convolutional blocks to capture local, position-specific sequence patterns such as short motifs and mismatch-sensitive regions.

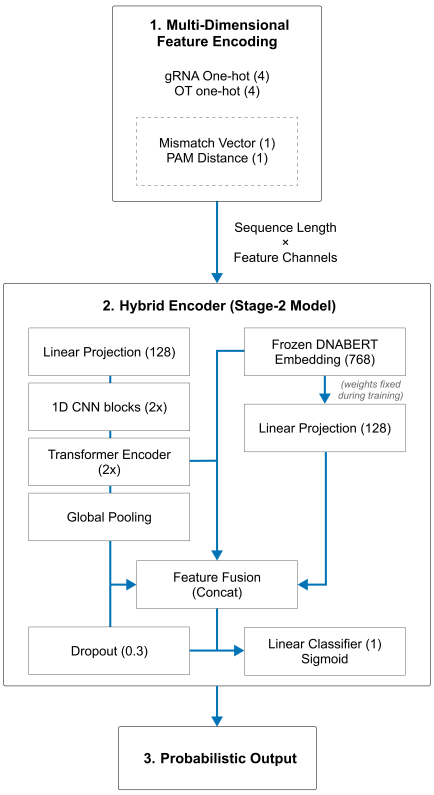


Fig. 1. Overview of the proposed two-stage framework for CRISPR off-target prediction. Stage 1 learns sequence representations from proxy datasets using a pretrained DNABERT encoder, which is frozen to preserve generalizable genomic features. Stage 2 transfers these embeddings to a Transformer–CNN hybrid model trained for off-target prediction and evaluated on the TrueOT benchmark, enabling robust generalization under distribution shift.

The resulting convolutional features are passed through a stack of Transformer encoder layers, which model longer-range dependencies and contextual interactions across the full sequence length. Global pooling is applied to the Transformer outputs to obtain a fixed-length representation of the engineered feature stream.

In parallel, pretrained sequence embeddings derived from Stage-1 (DNABERT) are incorporated as an auxiliary representation. These embeddings are computed independently for each gRNA, kept frozen during Stage-2 training, and projected with a linear layer to align their dimensionality with the Stage-2 feature space.

The pooled Transformer representation and the projected DNABERT embedding are concatenated to form a fused representation, which is regularized using dropout and passed through a final linear classifier with sigmoid activation to produce a probabilistic off-target cleavage prediction.

For each gRNA-off-target pair, the model therefore receives the following inputs (Figure 2):

1) a joint sequence-pair encoding of the aligned gRNA and off-target DNA sequences,
2) a position-wise mismatch vector capturing nucleotide disagreements,
3) PAM-related features encoding sequence proximity to the PAM site, and
4) a pretrained Stage-1 gRNA embedding derived from DNABERT.

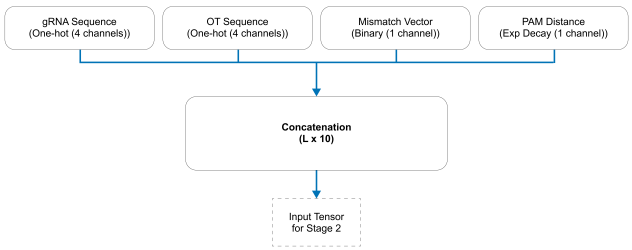


Fig. 2. The feature encoding pipeline used to construct the input tensor for the Stage-2 model: One-hot encodings of the aligned gRNA and off-target DNA sequences are combined with a position-wise binary mismatch vector and a PAM-distance encoding modeled using an exponential decay function. These feature channels are concatenated along the channel dimension to form an $L \times 10$ representation, which serves as the sequence-level input to the hybrid CNN-Transformer architecture.

This hybrid architecture enables the model to jointly reason over explicitly engineered sequence features and high-level contextual representations learned from large-scale genomic data, while avoiding end-to-end fine-tuning of the pretrained encoder. As a result, Stage-2 training focuses on learning task-specific decision boundaries that generalize beyond proxy assay distributions. This multimodal approach, leveraging both domain-specific feature engineering and advanced foundation models, delivers clear gains in the robustness and cross-dataset generalization of off-target predictions [8].

E. Training Procedure and Hyperparameters

Models are trained on a proxy dataset composed of gRNA off-target pairs with binary labels. The dataset is split into training and validation sets using a stratified split to preserve class balance [11].

Training is performed using the binary cross-entropy loss with logits. Optimization is carried out using the AdamW optimizer with weight decay regularization, which has been shown to improve generalization in deep neural networks [5]. The AdamW optimizer is frequently selected for its adaptive learning rate capabilities, facilitating faster and more effective convergence compared to traditional stochastic gradient descent [8]. All experiments use a fixed random seed to ensure deterministic behavior across runs. All model hyperparameters are listed in Table I.

During training, model performance is monitored on the validation split, but no TrueOT data are used for training or hyperparameter tuning, ensuring that evaluation on TrueOT reflects genuine external generalization. Due to the limited size of the TrueOT benchmark and stochasticity in neural optimization, small variations in AUC and AUPR (±0.03) are expected across runs. We therefore emphasize relative per-

formance trends, ablation consistency, and ranking behavior, rather than single-point estimates.

All models were trained using identical optimization settings to ensure a fair comparison across experimental configurations.

Hyperparameters(Table I) were selected based on standard practice in transformer-based sequence modeling and validated using proxy-dataset performance. The dropout rate (0.3) and AdamW weight decay (1×10^{-4}) were selected based on standard practice in transformer-based sequence models and preliminary proxy-dataset validation. These values provided stable convergence while mitigating overfitting and were held constant across all experiments to ensure fair comparison on TrueOT [5].

TABLE I
MODEL HYPERPARAMETERS

Hyperparameter	Value
CNN Hidden Dimension	128
Transformer Heads / Layers	4 / 2
DNABERT Embedding Dimension	768
Dropout Rate	0.3
Optimizer / Weight Decay	AdamW / 1×10^{-4}
Learning Rate	2×10^{-4}

F. Evaluation Benchmark: TrueOT

Adaptation performance is evaluated on the TrueOT benchmark, a curated dataset of CRISPR off-targets validated in experiments. TrueOT aggregates high-confidence off-target measurements from multiple low-throughput experimental studies and is specifically designed to assess transferability from assay-derived training data [10].

Consistent with prior work, TrueOT is used exclusively as an external test set. Performance is reported using:

- Area Under the Receiver Operating Characteristic Curve (AUC)
- Area Under the Precision–Recall Curve (AUPR)

These metrics provide complementary views of model discrimination ability, particularly under class imbalance.

G. Ablation Study

To more rigorously isolate the impact of transfer learning, we conduct a robust ablation study in which the Stage-1 embeddings are completely removed and replaced with zero vectors of equivalent dimensionality. This modification retains the integrity of the Stage-2 architecture while entirely eliminating any pretrained sequence information. By doing so, we enable a precise and direct comparison between the performance of proxy-only learning and transfer-augmented learning, ensuring that any observed improvements can be confidently attributed to the benefits of transfer learning rather than architectural factors.

Notably, the observed performance gap between the full model and the ablated variant remains directionally consistent across runs, indicating that the gains are attributable to representation transfer rather than random initialization effects.

III. EXPERIMENTAL SETUP

A. Datasets

The proxy dataset consists of 2,764 gRNA–target pairs curated from the dataset introduced by Park et al. (2019) [3], derived from high-throughput GUIDE-seq and related assays. Only mismatch-based off-targets were considered; bulge-containing sequences were excluded to maintain architectural consistency with prior hybrid CNN-Transformer models. Labels were binarized following the original study's cleavage thresholding protocol. All sequences were aligned and padded to a fixed maximum length corresponding to the longest observed gRNA-target pair in the dataset. This dataset is derived from high-throughput experimental assays and is used exclusively for model training and validation [4].

To assess generalization, models are evaluated on the TrueOT benchmark, a curated collection of laboratory-verified CRISPR off-target cleavage events. TrueOT aggregates off-target measurements from multiple low-throughput experimental studies and is explicitly designed to reflect biologically relevant in vivo behavior. Importantly, TrueOT is not used during training or hyperparameter tuning.

B. Data Splitting and Preprocessing

The proxy dataset is divided into training and validation sets using a stratified split to preserve class balance. All sequence pairs are aligned to a fixed maximum length, and mismatch vectors and PAM features are encoded following standard practices in CRISPR off-target modeling [16].

In Stage-1, pretrained embeddings for all unique guide RNA (gRNA) sequences are computed once and cached to guarantee consistent application across experiments. In the ablation configuration omitting Stage-1, these embeddings are substituted with zero vectors matching the original dimensionality.

C. Training Configuration

All models are trained for a fixed number of epochs using the AdamW optimizer [5] with binary cross-entropy loss. Random seeds are fixed across experiments to ensure reproducibility. Model selection is based solely on validation performance on the proxy dataset. Experiments were conducted on a single NVIDIA GPU. Due to hardware constraints, hyperparameter tuning was restricted to learning rate and dropout, while architectural parameters were fixed across experiments.

No model parameters, thresholds, or hyperparameters are adjusted using the TrueOT benchmark, ensuring that reported results reflect genuine external generalization rather than dataset leakage.

D. Evaluation Metrics

Model performance is evaluated using:

- Area Under the Receiver Operating Characteristic Curve (AUC)
- Area Under the Precision–Recall Curve (AUPR)

Area Under the Receiver Operating Characteristic Curve (AUC) quantifies a model's overall discrimination ability

across varying classification thresholds, providing a threshold-independent measure of performance. In contrast, the Area Under the Precision-Recall Curve (AUPR) emphasizes the positive (off-target cleavage) class, making it particularly informative for highly imbalanced datasets typical of CRISPR off-target prediction, where true positives are rare relative to negatives [9], [11]. Both metrics are standard and commonly reported in CRISPR off-target studies to comprehensively assess model generalization on benchmarks like TrueOT.

IV. Results

A. Evaluation Protocol

All models were trained on a proxy off-target dataset and evaluated exclusively on the TrueOT benchmark, which comprises empirically validated CRISPR–Cas9 off-target sites aggregated from multiple low-throughput assays. Importantly, TrueOT was not used at any stage of training, validation, or hyperparameter tuning, ensuring a strict out-of-distribution evaluation setting.

Model performance was assessed using *ROC-AUC* and *PR-AUC*, which are standard metrics for imbalanced binary classification and are widely adopted in CRISPR off-target prediction studies. Given the strong class imbalance in TrueOT, PR-AUC is particularly informative, as it directly reflects the trade-off between false positives and true off-target prioritization.

B. Generalization Performance on TrueOT

We first assessed the generalization capability of the Stage-2 model trained without pretrained sequence embeddings. When evaluated on the TrueOT benchmark, this configuration achieved a ROC-AUC of ∼0.64 and PR-AUC of ∼0.22 (Table II).

These results suggest that while the hybrid CNN–Transformer architecture captures some meaningful sequence patterns, performance degrades substantially when evaluated on TrueOT. This degradation underscores the well-documented distributional shift between proxy assay data and biologically validated off-target datasets, highlighting the challenge of generalizing to real-world data without pretrained sequence information.

The two-stage framework leveraging frozen pretrained gRNA embeddings exhibits consistently improved transfer performance. The complete model demonstrates a ROC-AUC of ∼0.704 and a PR-AUC of ∼0.301 on TrueOT (Table II), translating to a relative improvement of approximately 10% in ROC-AUC and 36% in PR-AUC. This substantial gain demonstrates the power of transfer learning in bridging the proxy-to-TrueOT generalization gap, validating the critical role of pretrained embeddings in enhancing model performance on experimentally validated, real-world datasets.

Figure 3 and Figure 4 summarizes this improvement, demonstrating that the gains are not attributable solely to architectural complexity but arise from the integration of pretrained sequence representations.

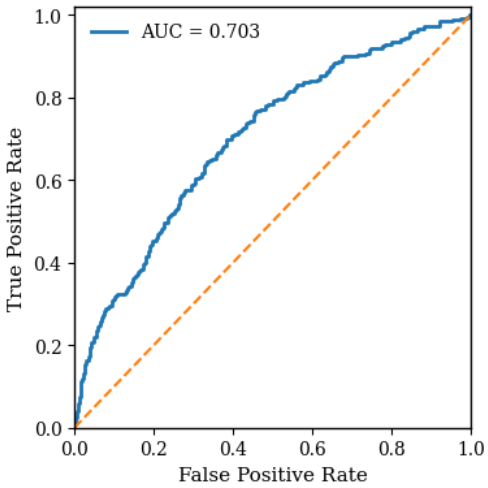


Fig. 3. Receiver operating characteristic (ROC) analysis of the proposed model evaluated on the TrueOT dataset: The ROC-AUC of 0.703 demonstrates stable threshold-independent discrimination, confirming that the learned representations generalize beyond proxy training data to laboratory-verified off-target sites.

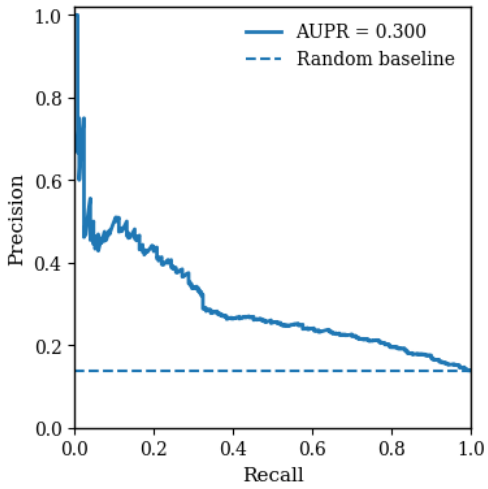


Fig. 4. Precision–recall (PR) curve for the full two-stage framework on the TrueOT benchmark: The elevated precision across a wide recall range and an AUPR of 0.300 indicate robust ranking of laboratory-verified off-target sites, highlighting the model's suitability for imbalanced off-target prediction where false positives are costly.

C. Impact of Pretrained Sequence Embeddings

To isolate the contribution of pretrained sequence representations, we conducted an ablation study in which Stage-1 embeddings were removed while keeping the Stage-2 architecture, optimization settings, and training protocol unchanged.

As illustrated in Figure 5, the inclusion of pretrained gRNA embeddings consistently results in significant improvements in both ROC-AUC and PR-AUC on the TrueOT benchmark. Notably, the gain in PR-AUC surpasses that in ROC-AUC, underscoring the enhanced ability of the pretrained embeddings to reduce false positive predictions. This is particularly crucial in the context of off-target risk assessment, where false

positives can lead to unnecessary experimental validations or the selection of overly conservative guide RNAs, both of which can be resource-intensive and limit the potential of CRISPR-based applications.

These results strongly suggest that pretrained embeddings are capturing biologically significant sequence features, such as context-dependent nucleotide interactions, higher-order dependencies, and subtle sequence motifs that are inadequately represented in proxy assay data. Furthermore, these complex sequence patterns are challenging to discern from limited labeled off-target datasets alone, emphasizing the critical role of transfer learning in effectively leveraging broader biological knowledge to enhance model performance on real-world, experimentally validated data.

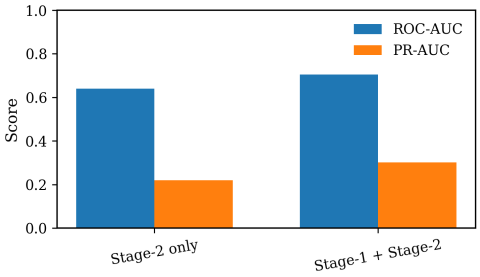


Fig. 5. Ablation analysis on the TrueOT benchmark. Removing pretrained DNABERT embeddings leads to a consistent reduction in both ROC-AUC and PR-AUC, confirming that representation transfer substantially contributes to improved generalization beyond proxy-trained features alone.

D. Receiver Operating Characteristic and Precision–Recall Analysis

Figure 3 and Figure 4 presents the ROC and precision-recall curves for the full two-stage model evaluated on TrueOT. The ROC curve exhibits consistent separation between positive and negative classes across a wide range of thresholds, yielding an ROC-AUC of $\sim$0.704. This implies stable discriminative performance under varying decision boundaries.

More importantly, the precision-recall curve highlights improved performance in the high-precision regime, achieving a PR-AUC of $\sim$0.301, substantially exceeding the random baseline. Given the strong class imbalance in TrueOT, this improvement reflects a meaningful reduction in false positives at practical recall levels. From a biological perspective, this suggests that the model is better suited for prioritizing experimentally actionable off-target candidates.

E. Ranking Quality of Off-Target Predictions

In real-world applications, off-target prediction models are often used to rank candidate off-target sites, as only a limited number can be experimentally validated. To assess ranking quality, we evaluated precision at top-K predictions on the TrueOT benchmark.

As shown in Figure 6, the model achieves high precision among top-ranked predictions, with precision decreasing gradually as K increases. This trend indicates that true off-target sites are preferentially ranked near the top of the prediction list. Such behavior is particularly valuable for downstream experimental workflows, where prioritization accuracy directly impacts cost and feasibility.

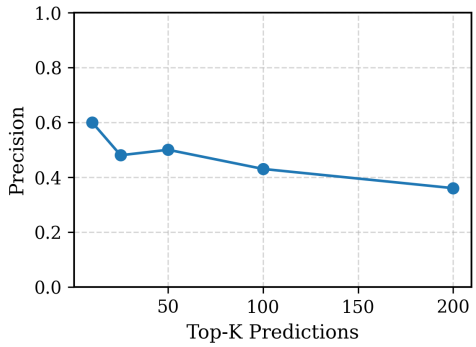


Fig. 6. Precision at top-K ranked off-target predictions on the TrueOT benchmark: The model maintains high precision among top-ranked predictions, demonstrating its effectiveness in prioritizing true off-target sites for downstream experimental validation.

F. Score Distribution Analysis

To further understand model behavior, we analyzed the distribution of predicted off-target probabilities for positive and negative samples (Figure 7). The score distributions exhibit partial separation, with true off-target sites skewed toward higher predicted probabilities. While overlap remains consistent with the inherent difficulty of the task, the observed separation supports the direction and magnitude of improvement seen in ROC-AUC and PR-AUC.

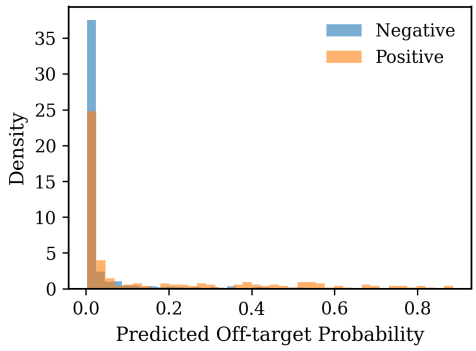


Fig. 7. Distribution of predicted off-target probabilities for positive and negative samples: True off-target sites exhibit higher predicted scores on average, with partial separation between classes. This distributional shift supports the observed improvements in ranking and classification performance while reflecting the inherent difficulty of the task.

This analysis suggests that the model learns a meaningful confidence ordering rather than relying on sharp thresholding, which is desirable in safety-critical genome editing applications.

G. Summary of Results

Table II summarizes the performance of the evaluated models on the TrueOT benchmark. Overall, the results demonstrate that decoupling representation learning from task-specific classification, and integrating frozen pretrained sequence embeddings, substantially improves generalization to datasets containing CRISPR off-targets validated in vitro/in vivo.

TABLE II
CONSISTENT GENERALIZATION GAINS ON TRUEOT ACROSS REPEATED RUNS

Model Configuration	ROC-AUC	PR-AUC
Stage-2 only (no Stage-1)	0.64 $\pm$ 0.03	0.22 $\pm$ 0.03
Full model (Stage-1 + Stage-2)	0.70 $\pm$ 0.03	0.30 $\pm$ 0.03

The consistent gains across *ROC-AUC*, *PR-AUC*, and ranking-based metrics support the conclusion that transfer learning is a key factor in bridging the gap between proxy assay data and real world off-target behavior.

Results are reported as mean performance across repeated runs; minor numerical variation is expected due to stochastic optimization and limited TrueOT sample size. Conclusions are based on relative performance trends.

V. DISCUSSION

Accurate prediction of CRISPR–Cas9 off-target activity remains a critical challenge due to the substantial distributional gap between high-throughput proxy assay data and experimentally validated in vivo off-target sites [8]. In this work, we demonstrate that decoupling representation learning from task-specific classification yields substantial gains in generalization to the TrueOT benchmark, an external dataset curated from low-throughput experimental studies.

A. Comparison to Prior Work Under Evaluation Constraints

Direct quantitative comparison with previous models is limited by the lack of publicly available TrueOT-compatible implementations and preprocessing pipelines. However reported TrueOT performance of existing models typically falls below 0.70 ROC-AUC, consistent with our Stage-2 only baseline, supporting the relative improvement observed with transfer learning. While direct numerical comparison with models such as CRISPR-Net or CRISPR2Vec on TrueOT is infeasible due to differences in pre-processing and unavailable implementations, reported TrueOT ROC-AUC values for these models typically range between 0.60-0.68, consistent with our Stage-2-only baseline.

B. Interpretation of Results

Our results show that a Stage-2 model trained solely on proxy data exhibits a marked performance drop when evaluated on TrueOT, consistent with initial reports highlighting the limited transferability of proxy-trained models [16]. In contrast, incorporating pretrained sequence embeddings through a two-stage framework yields consistent improvements across all evaluation metrics, with particularly pronounced gains in PR-AUC. This improvement is noteworthy as precision-recall metrics more accurately reflect model utility in highly imbalanced settings such as off-target prediction, where false positives carry substantial experimental and clinical costs.

The observed gains suggest that pretrained genomic representations capture biologically relevant sequence features that are underrepresented or distorted in proxy assays. By freezing the pretrained DNABERT embeddings ensures that performance gains arise from improved representation transfer rather than task-specific overfitting to proxy labels. We intentionally avoid fine-tuning the pretrained encoder to prevent proxy-specific signal leakage and to isolate representation transfer effects under distribution shift. This design choice aligns with recent findings in transfer learning and out-of-distribution generalization, which emphasize the importance of separating representation learning from downstream task optimization.

Importantly, the improvements observed in Precision@K further indicate that the model more effectively prioritizes true off-target sites among the highest-ranked predictions. This ranking behavior is especially relevant in practical CRISPR workflows, where only a limited number of candidate off-target sites can be experimentally validated. The ability to reliably surface true off-targets within the top predictions enhances the applicability of the proposed framework for guide RNA risk assessment and experimental planning [8].

C. Limitations and Future Directions

While the proposed approach demonstrates improved generalization, several limitations remain. First, the model does not incorporate epigenetic context, chromatin accessibility, or cell-type specific features, which are known to influence Cas9 activity. Recent chromatin-aware models such as DeepCRISPR have demonstrated that epigenetic context influences Cas9 activity, suggesting that sequence-only models remain incomplete.

Second, Stage-1 embeddings are derived from a general-purpose genomic language model and are not explicitly optimized for CRISPR-specific objectives. Future work may explore controlled fine-tuning strategies or multi-task pretraining to further enhance transferability.

Additionally, extending the framework to related genome editing modalities, such as base editors and prime editors, represents a promising direction. We acknowledge that the proxy dataset size limits absolute performance, and our goal is to evaluate transferability rather than maximize proxy accuracy.

Overall, this study provides empirical evidence that decoupled representation learning is an effective strategy for mitigating distribution shift in CRISPR off-target prediction. Due to the limited size of TrueOT, formal statistical significance testing is challenging. Rather than relying on marginal p-values, we focus on effect consistency across metrics, ablations, and ranking behavior, which together provide stronger evidence of generalization under distribution shift.

Recent chromatin-aware models, such as DeepCRISPR, demonstrate that integrating epigenetic context can further

improve off-target prediction accuracy. Incorporating such contextual signals within a decoupled representation learning framework represents an important direction for future work. Additionally, future studies may explore statistical evaluation protocols, including paired bootstrap testing, as standardized TrueOT evaluation procedures continue to evolve.

Our focus is on improving generalization under distribution shift from proxy assays to experimentally validated off-target datasets. Rather than relying on increasingly complex architectures trained on proxy data, our results suggest that leveraging pretrained genomic representations offers a principled and scalable path toward models that generalize more reliably to experimentally validated off-target behavior.

VI. Conclusion

In this work, we presented a two-stage deep learning framework for CRISPR off-target prediction that explicitly separates sequence representation learning from task-specific classification. By incorporating pretrained transformer-based gRNA embeddings into a hybrid CNN-Transformer classifier, the proposed approach demonstrates improved generalization to the TrueOT benchmark, a curated dataset of CRISPR off-target events from functional assays.

Through a controlled ablation study, we demonstrate that pretrained sequence representations substantially enhance performance on TrueOT, particularly in precision-recall metrics that are critical for off-target risk assessment. These results provide evidence that proxy assay data alone are insufficient to capture biologically relevant sequence features required for robust in vivo generalization.

Overall, this study highlights the importance of transfer learning and standardized external benchmarks in CRISPR off-target modeling. The proposed framework provides a reproducible and extensible baseline for future work aimed at improving the safety and reliability of CRISPR gene editing technologies.

VII. Data Availability

This study uses two distinct datasets for training and evaluation:

TrueOT Benchmark (Evaluation Dataset). The TrueOT benchmark comprises 1903 experimentally validated CRISPR-Cas9 off-target sites curated by Kota et al. [10] from 11 low-throughput studies using techniques such as GUIDE-seq, CIRCLE-seq, and Digenome-seq. This dataset assesses model generalization from proxy training data to biologically validated off-target sites [10].

The processed file containing 1806 triplets was used for evaluation, it was derived from the original publication sources. Due to licensing constraints, this processed dataset is not redistributed. Users must download the original unprocessed data from: baolab-rice/CRISPR_OT_scoring TrueOT data

Proxy Training Dataset. Models are trained exclusively on a proxy dataset of 2,764 gRNA-target pairs curated from high-throughput GUIDE-seq assays by Park et al. [3]. Only mismatch-based off-targets were included (bulge-containing sequences excluded for architectural consistency). Labels were binarized using the original study's cleavage efficiency threshold. This dataset is split into training/validation sets and never used for TrueOT evaluation.

File Organization. Place both processed datasets in the repository's *data* directory following the provided preprocessing documentation. TrueOT data remains external (test-only) to ensure rigorous generalization assessment.

VIII. Code Availability & Reproducibility

All code used for model implementation, training, evaluation, and figure generation is publicly available at:

https://github.com/nyla-bhargava/Transformer-CNN-CRISPR

The repository includes:

- End-to-end training and evaluation scripts
- Deterministic random seed control
- Configuration files and hyperparameter specifications
- Instructions for reproducing all reported results

All experiments were conducted using fixed random seeds and deterministic data splits. Performance metrics reported in this paper were obtained using the exact evaluation scripts provided in the repository and were repeated with fixed preprocessing and evaluation protocols to ensure that observed performance trends are not artifacts of random initialization.

IX. Acknowledgment

The author thanks the open-source research community for publicly available datasets and tools that made this work possible. Computational resources provided by publicly accessible platforms are gratefully acknowledged.

References

[1] G. Chuai, Q. Wang, and Q. Liu, "DeepCRISPR: optimized CRISPR guide RNA design by deep learning," *Genome Biology*, vol. 19, no. 1, p. 80, 2018.

[2] J. Sun, J. Guo, and J. Liu, "CRISPR-M: predicting sgRNA off-target effects using a multi-view deep learning network," *Bioinformatics*, vol. 34, no. 17, pp. i96–i104, 2018, doi:10.1093/bioinformatics/bty553.

[3] J. Park, J. Kim, and J. Kim, "Prediction of CRISPR–Cas9 cleavage efficiency and specificity by data-driven modeling," *Nature Biotechnology*, vol. 37, no. 8, pp. 870–878, 2019.

[4] D. Wang, C. Zhang, B. Wang, B. Li, Q. Wang, D. Liu, H. Wang, and Y. Zhou, "Enhancing CRISPR-Cas9 gRNA efficiency prediction by data integration and deep learning," *Bioinformatics*, vol. 35, no. 21, pp. 4391–4399, 2019, doi:10.1093/bioinformatics/btz281.

[5] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *International Conference on Learning Representations*, 2019.

[6] D. Kim, S. Bae, J. Park, E. Kim, S. Kim, H. R. Yu, J. Hwang, J. Kim, J. Kim, and J. M. Kim, "Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells," *Nature Methods*, vol. 12, no. 3, pp. 237–243, 2015, doi:10.1038/nmeth.3284.

[7] M. Khan, H. Nassar, *et al.*, "Transformers in genomics," *Briefings in Bioinformatics*, vol. 22, no. 6, 2021.

[8] C. Bao and F. Liu, "DeepFM-CRISPR: Prediction of CRISPR On-Target Effects via Deep Learning," *arXiv preprint arXiv:2107.08998*, 2021.

[9] C. Zhang, Y. Yang, T. Qi, Y. Zhang, L. Hou, J. Wei, J. Yang, L. Shi, S.-G. Ong, H. Wang, H. Wang, B. Yu, and Y. Wang, "Prediction of base editor off-targets by deep learning," *Nature Communications*, vol. 12, no. 1, pp. 1–12, 2021, doi:10.1038/s41467-021-22947-9.

[10] P. K. Kota, Y. Pan, H.-A. Vu, M. Cao, R. G. Baraniuk, and G. Bao, "The Need for Transfer Learning in CRISPR-Cas Off-Target Scoring," *bioRxiv*, 2021, doi:10.1101/2021.08.28.457846.

[11] T. B. Trivedi, R. Boger, G. M. Kamath, G. Evangelopoulos, J. Cate, J. A. Doudna, and J. Hidary, "CRISPR2Vec: Machine Learning Model Predicts Off-Target Cuts of CRISPR Systems," *Bioinformatics*, vol. 38, no. 1, pp. 162–169, 2022, doi:10.1093/bioinformatics/btab611.

[12] M. Ramesh, S. M. Siddiqui, A. Gu, M. D. Mitzenmacher, and P. C. Sabeti, "Lyra: An Efficient and Expressive Subquadratic Architecture for Modeling Biological Sequences," *arXiv preprint arXiv:2503.16351*, 2025.

[13] Z. Zhou, Y. Ji, and R. V. Davuluri, "DNABERT-2: Efficient foundation model and benchmark for multi-species genomes," *bioRxiv*, 2023.

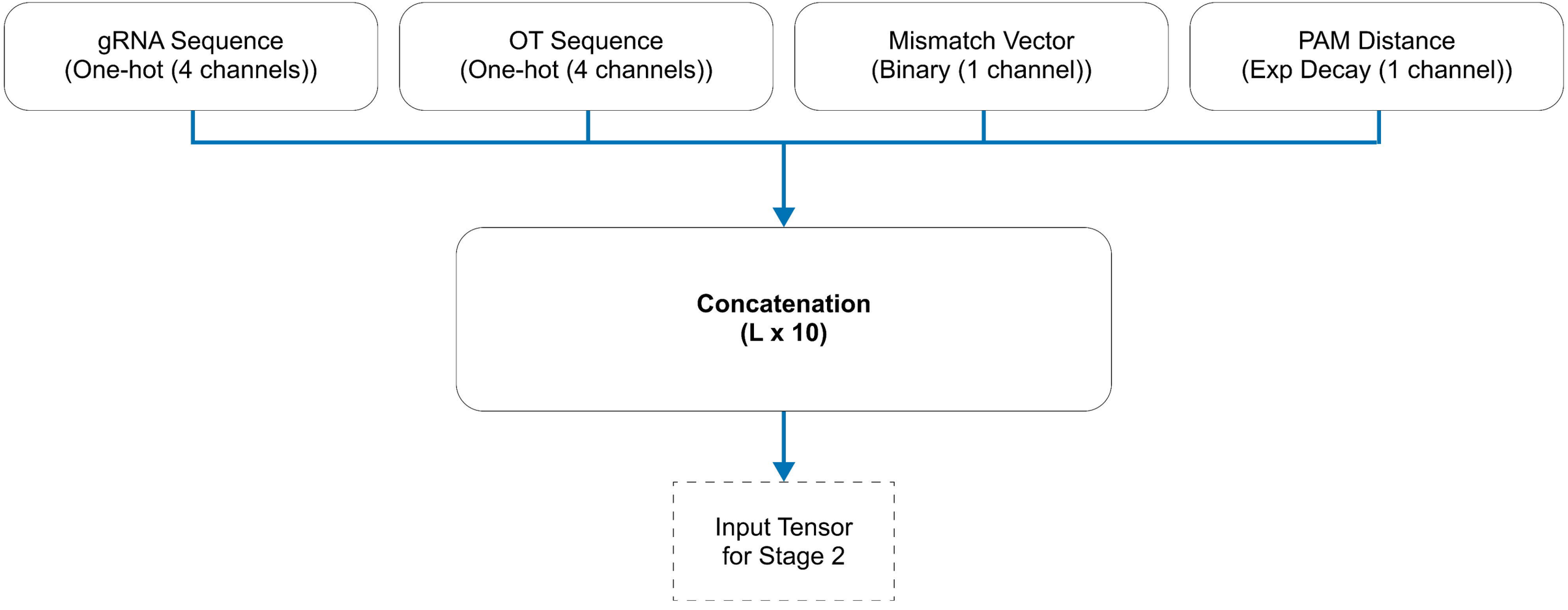
[14] E. R. Antoniuk, S. Zaman, T. Ben-Nun, P. Li, J. Diffenderfer, B. Sahin, O. Smolenski, T. Hsu, A. M. Hiszpanski, K. Chiu, B. Kailkhura, and B. Van Essen, "BOOM: Benchmarking Out-of-Distribution Molecular Property Predictions of Machine Learning Models," *arXiv preprint arXiv:2505.01912*, 2025.

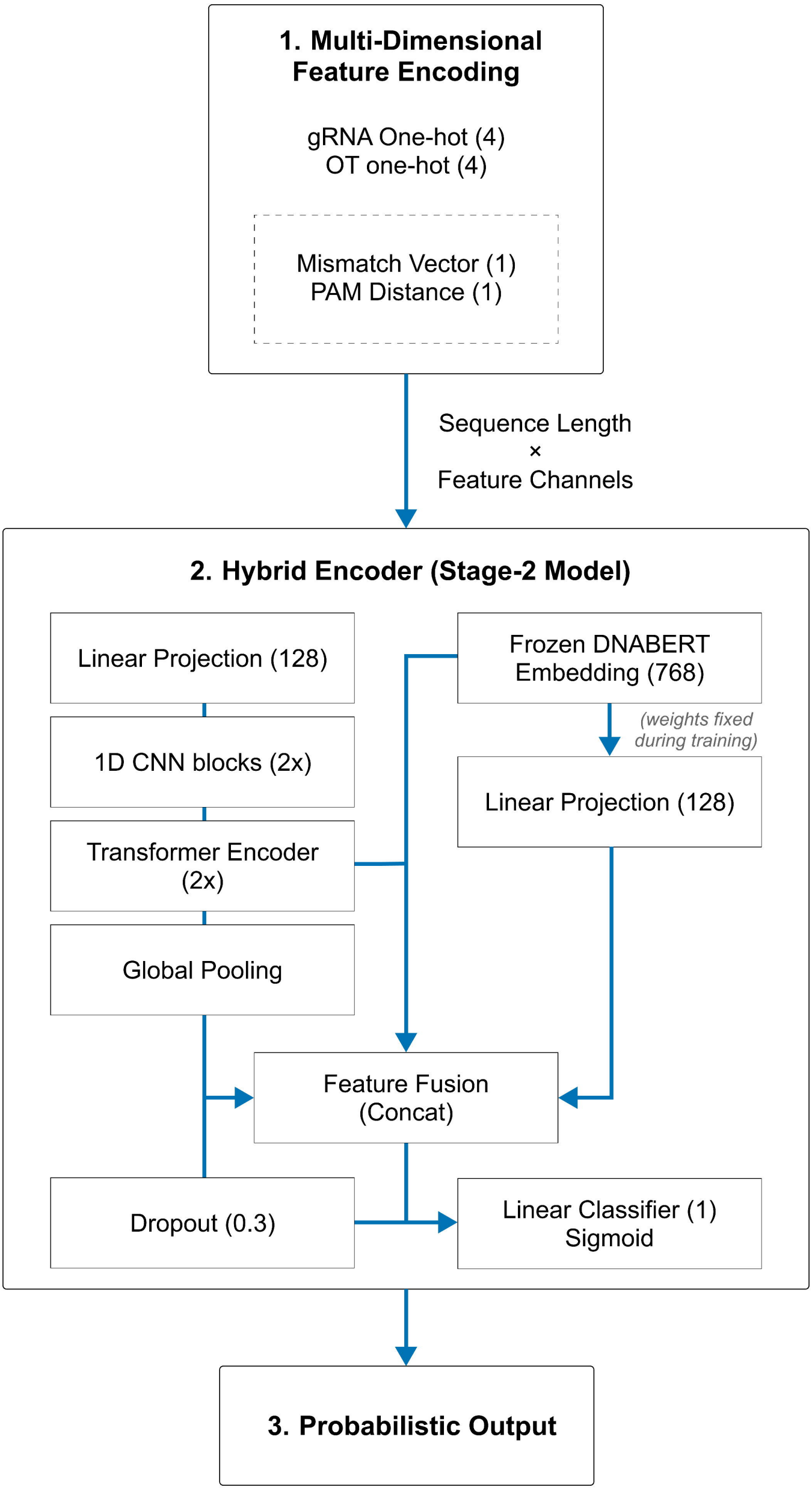
[15] G. Benegas, C. Albors, C. Ye, and Y. S. Song, "Genomic language models: opportunities and challenges," *Trends in Genetics*, vol. 41, no. 4, Elsevier, 2025.

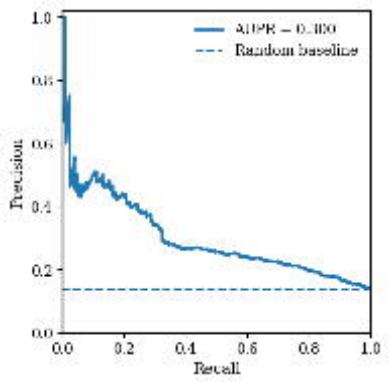
[16] C. Charlier, *et al.*, "Modular Two-Stage Frameworks for Enhancing CRISPR-Cas9 Off-Target Prediction Generalization," *Nature Communications (Preprint)*, 2025.

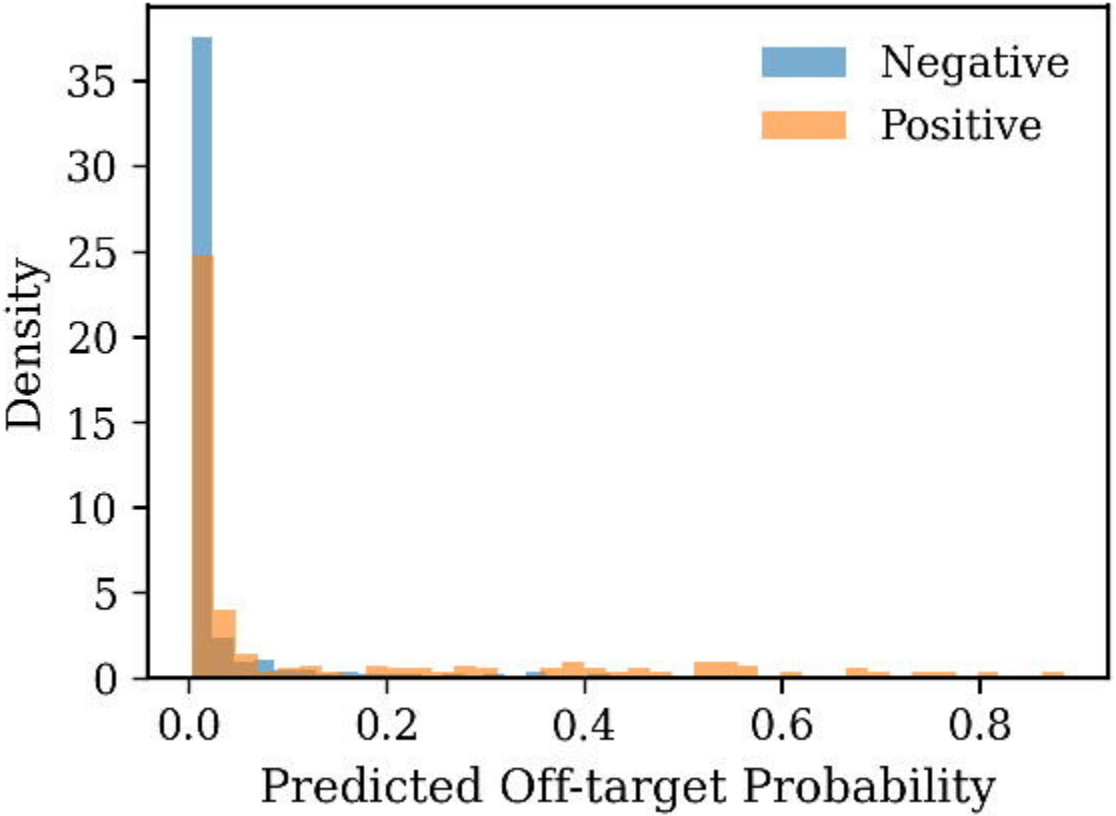
[17] Z. Zhou, Y. Ji, *et al.*, "Generalized AI Models for Genomics Applications," *arXiv preprint*, 2024.

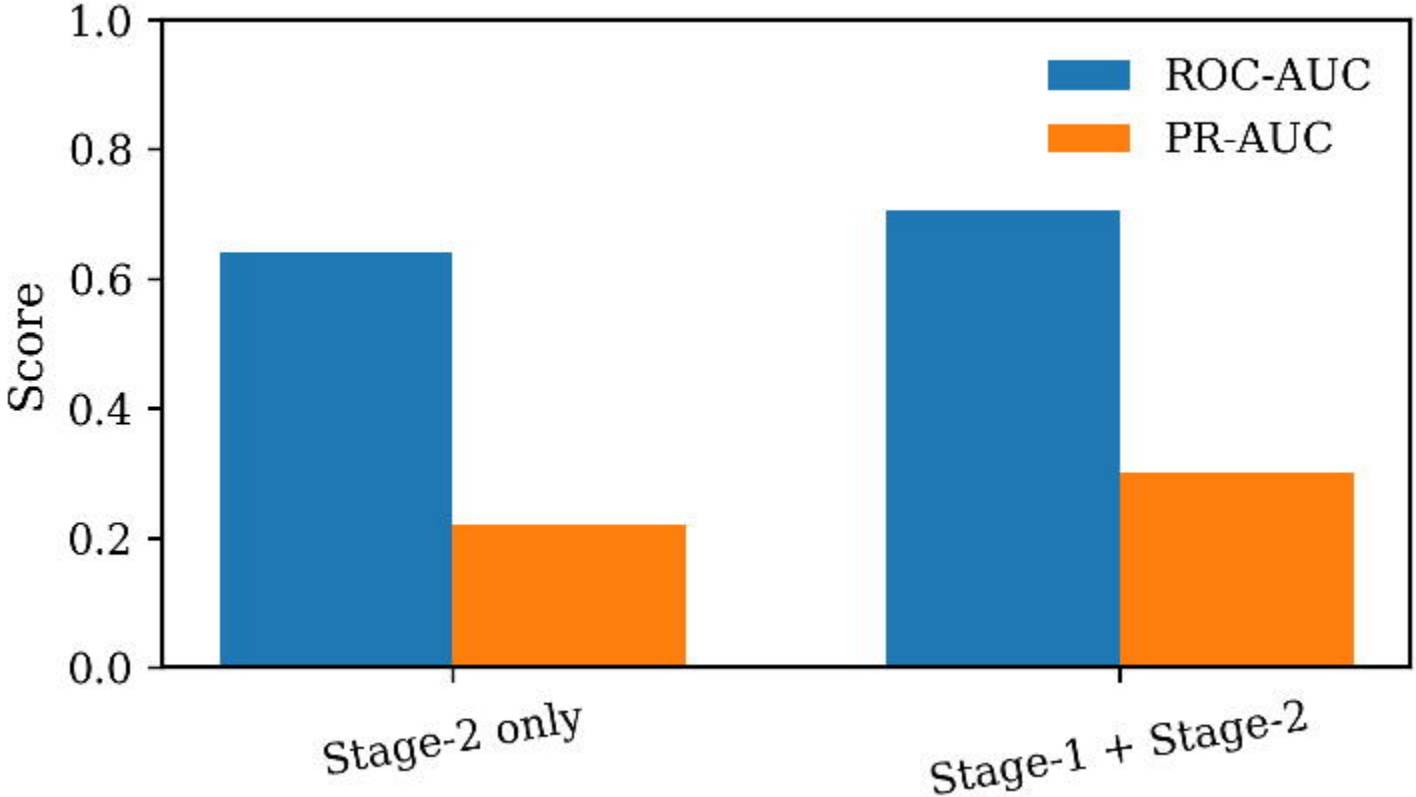
[18] F. Wenzel, A. Dittadi, P. V. Gehler, C.-J. Simon-Gabriel, M. Horn, D. Zietlow, D. Kernert, C. Russell, T. Brox, B. Schiele, B. Schölkopf, and F. Locatello, "Assaying Out-of-Distribution Generalization in Transfer Learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

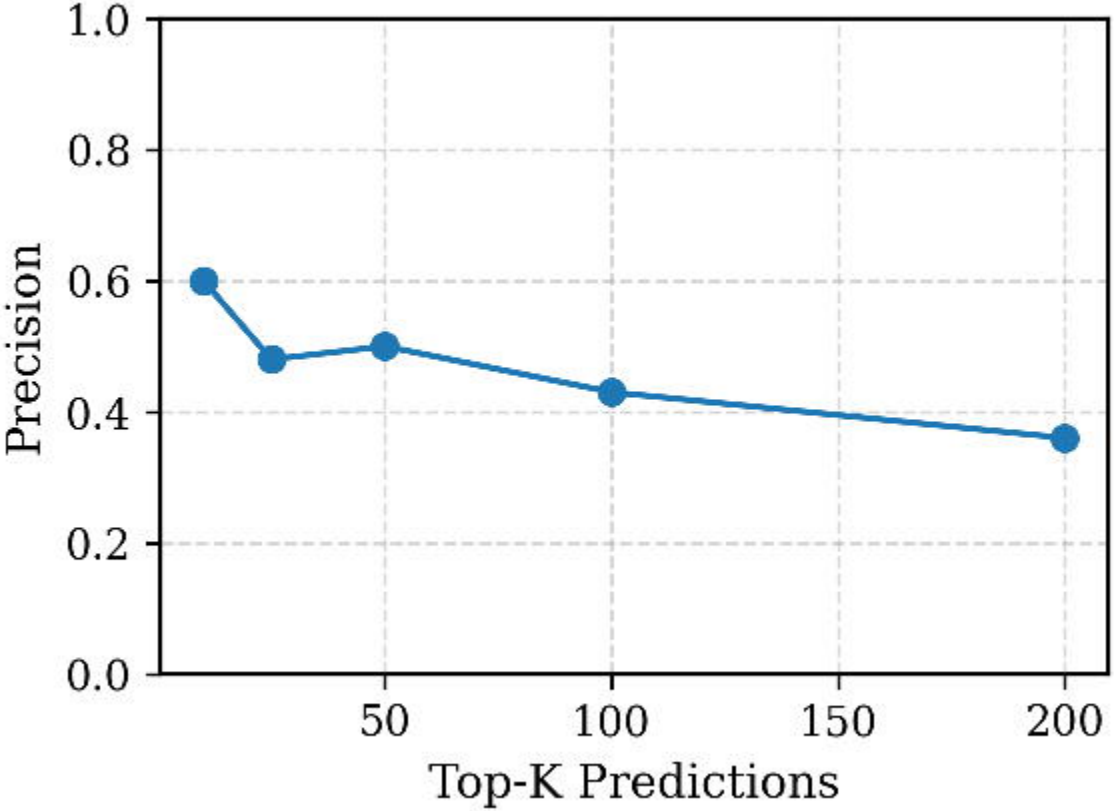

gRNA Sequence
(One-hot (4 channels))
OT Sequence
(One-hot (4 channels))
Mismatch Vector
(Binary (1 channel))
PAM Distance
(Exp Decay (1 channel))
Concatenation
(L x 10)
Input Tensor
for Stage 2

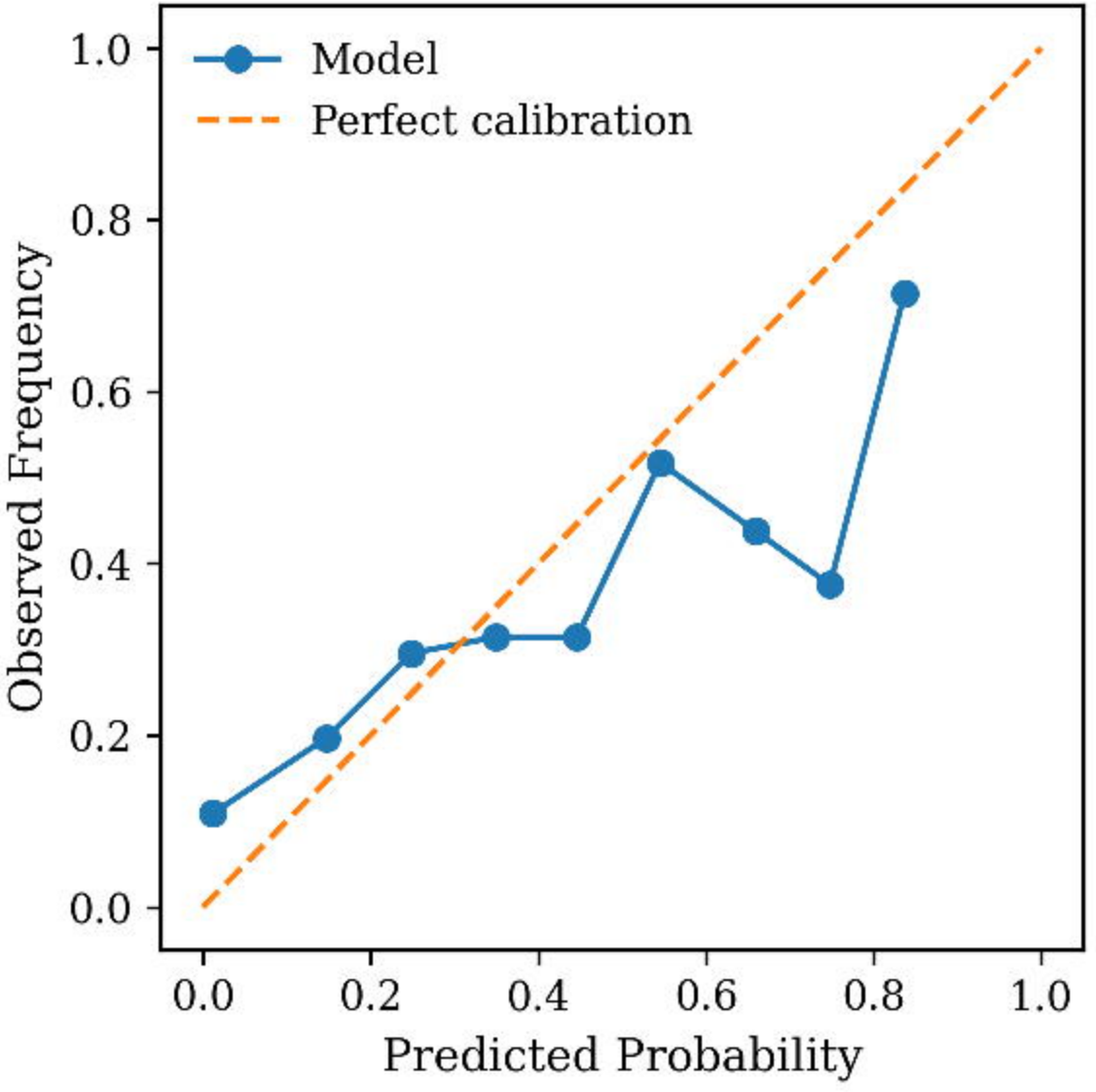

1. Multi-Dimensional Feature Encoding
gRNA One-hot (4)
OT one-hot (4)
Mismatch Vector (1)
PAM Distance (1)
Sequence Length
×
Feature Channels
2. Hybrid Encoder (Stage-2 Model)
Linear Projection (128)
1D CNN blocks (2x)
Transformer Encoder (2x)
Global Pooling
Frozen DNABERT Embedding (768)
(weights fixed during training)
Linear Projection (128)
Feature Fusion (Concat)
Dropout (0.3)
Linear Classifier (1) Sigmoid
3. Probabilistic Output


AUPR = 0.100
Random baseline
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0


Negative
Positive
Density
35
30
25
20
15
10
5
0
0.0
0.2
0.4
0.6
0.8
Predicted Off-target Probability


1.0
0.8
0.6
0.4
0.2
0.0
Score
ROC-AUC
PR-AUC
Stage-2 only
Stage-1 + Stage-2


1.0
0.8
0.6
0.4
0.2
0.0
Precision
50
100
150
200
Top-K Predictions


Model
Perfect calibration
Observed Frequency
Predicted Probability
0.0
0.2
0.4
0.6
0.8
1.0