Uncertainty-Guided Model Selection for Tabular Foundation Models in Biomolecule Efficacy Prediction

Jie Li
AIML
GSK
South San Francisco, CA, USA 94080
jerry.8.li@gsk.com

Andrew McCarthy
Molecular Modalities Discovery
GSK
Cambridge, MA, USA 02140
andrew.p.mccarthy@gsk.com

Zhizhuo Zhang
AIML
GSK
South San Francisco, CA, USA 94080
zhizhuo.x.zhang@gsk.com

Stephen Young
AIML
GSK
London, UK, N1C 4AG
stephen.r.young@gsk.com

Abstract

In-context learners like TabPFN are promising for biomolecule efficacy prediction, where established molecular feature sets and relevant experimental results can serve as powerful contextual examples. However, their performance is highly sensitive to the provided context, making strategies like post-hoc ensembling of models trained on different data subsets a viable approach. An open question is how to select the best models for the ensemble without access to ground truth labels. In this study, we investigate an uncertainty-guided strategy for model selection. We demonstrate on an siRNA knockdown efficacy task that a TabPFN model using simple sequence-based features can surpass specialized state-of-the-art predictors. We also show that the model's predicted inter-quantile range (IQR), a measure of its uncertainty, has a negative correlation with true prediction error. By selecting and averaging an ensemble of models with the lowest mean IQR, we achieve superior performance compared to naive ensembling or using a single model trained on all available data. This finding highlights model uncertainty as a powerful, label-free heuristic for optimizing biomolecule efficacy predictions.

1 Introduction

Predicting biomolecule efficacy is a central challenge in drug discovery, with the potential to significantly accelerate pharmaceutical development by prioritizing promising candidates before wet-lab validation. However, biomolecule efficacy datasets are often small, heterogeneous, and collected using varied experimental techniques[1]. Such inconsistency complicates the training of reliable predictive models, hindering the real-world influence of these models.

While domain-specific in-context learners (ICLs) have shown promise for few-shot molecular property prediction[2], recent general-purpose tabular models like TabPFN[3], TabPFNv2[4] and TabDPT[5] offer a new approach that can enhance biomolecule efficacy prediction by leveraging relevant data as context. However, choosing the right data is not always straightforward. Simply using more data does not guarantee better performance, and large datasets can exceed an ICL's practical limits. For example, TabPFN's computational cost scales quadratically with the number of training examples,

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: 2nd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences.

making it infeasible to process very large training sets[6]. Furthermore, using a context size that is out-of-distribution from the model's pre-training data can also lead to inferior results.

Strategies to address this include data retrieval and fine-tuning, as exemplified by LoCalPFN [7], or post-hoc ensembling [8], which combines predictions from models trained on different data subsets. However, effective ensembling still requires a reliable method for selecting best performing models. In this paper, we explored an uncertainty based strategy for this task. We propose that a model's predicted uncertainty can serve as a proxy for its true prediction error, allowing us to select and combine models with the lowest predicted uncertainty to form a final, more accurate ensemble. While prior work has used uncertainty of sensitive attributes to guide sample selection to improve fairness in ICLs [9], its use for improving model accuracy by model selection has not been explored to the best of our knowledge. We demonstrated on an siRNA knockdown efficacy task that TabPFN model with simple features surpasses specialized, state-of-the-art models. Furthermore, we show the model's inter-quantile range (IQR) reflects its prediction uncertainty and negatively correlates with true prediction error, making it a competitive metric for selecting models that yield optimal performance.

2 Dataset and Features

Small interfering RNAs (siRNAs) are a promising therapeutic modality that silences target genes by cleaving mRNA transcripts[10]. A key challenge is designing siRNAs with high knockdown efficacy. We used the dataset composed by Huesken et al.[11], which has been used to develop multiple siRNA efficacy predictors, including the state of the art model OligoFormer[12]. We divided the Huesken dataset (2361 data points) into 29 subsets, each corresponding to a different mRNA target. Additionally, we collected data on a novel target not present in the Huesken dataset (Target1) from public accessible patents from three institutions denoted as institution A, B and C, with 295, 366 and 9 data points, respectively. We also collected 252 data points from a patent published by institution D on a second novel target (Target2).

Following the protocol from OligoFormer, we represented siRNA molecules as 19-mers, and prepared the corresponding mRNA centered around the region that the siRNA reverse complements to, with a flanking region of 19 nucleotides on both sides. Our feature set includes: (1) one-hot representation for each nucleotide in the siRNA and mRNA sequences, (2) counts of all possible ribonucleotide trimers from both sequences, and (3) thermodynamic parameters that describe the siRNA stability and binding affinity of the siRNA-mRNA interactions. A detailed description of the feature formulation can be found in Appendix A. This resulted in 574 features per data point. Although the number of features exceeds TabPFN's pre-training limit, we found that using the full feature set with the `"ignore_pretraining_limits=True"` flag yielded higher performance than feature subsetting.

3 Experiment and Results

3.1 Comparing TabPFN Model with a State-of-the-Art Model

We validated the use of TabPFN with custom features by comparing its performance to OligoFormer[12], a state-of-the-art model for siRNA knockdown efficacy prediction. To assess in-distribution performance, we evaluated the mean absolute error (MAE) and Pearson correlation coefficient on the Huesken dataset using a 5-fold cross-validation approach for TabPFN. As shown in Table 1, the cross-validated TabPFN model surpasses the performance of OligoFormer evaluated on its own training set, supporting the use of TabPFN for this task.

Given that these models are usually used in out-of-distribution settings, we then explored the capability for the TabPFN model to generalize with few-shot data on another target (Target1) that is distinct from those in the Huesken dataset. We used the smallest context (subset C) with only 9 data points to evaluate prediction MAEs and correlation coefficients for the two larger subsets (A and B). The results are also provided in Table 1. Despite the lower performance compared with the Huesken dataset, it still does significantly better than OligoFormer. This highlights the real-world applicability of TabPFN and the potential limitations of specialized models like OligoFormer when faced with novel targets.

Table 1: Mean absolute errors (MAEs) and correlation coefficients for TabPFN [4] and OligoFormer [12]. In-distribution performance (Huesken) for TabPFN is reported as mean ± 95% confidence interval based on 5-fold cross validation. Out-of-distribution (Target1) performance for TabPFN is based on few-shot prediction using the same target as context. In both cases, TabPFN outperforms the specialized SOTA model OligoFormer.

	MAE (↓)		**Corr. Coef (↑)**	
Dataset	TabPFN	OligoFormer	TabPFN	OligoFormer
Huesken	**0.087 ±0.004**	0.096	**0.677± 0.042**	0.630
Target1 (A)	**0.245**	0.251	**0.244**	0.158
Target1 (B)	**0.159**	0.180	**0.200**	0.082

3.2 Using Model Inter-Quantile Range for Prediction Quality Estimation

A TabPFN regressor can output not just a point estimate, but it can generate a distribution of possible values and provide lower bound and upper bound estimations for a certain quantile level[3]. We confirmed that our model is well-calibrated (see Appendix B), meaning its quantile estimates accurately reflect the probability of the true value falling within a given range. In this study, we use the 15% and 85% quantiles, expecting a 70% chance of correctness when the data is in-distribution. The inter-quantile range (IQR) is then defined as the difference between 85% quantile and 15% quantile from model predictions. The IQR serves as a measure of the model's intrinsic uncertainty; we investigated whether it could be used to approximate the true prediction error.

First, we combined all available data, used a randomly selected 70% of data for training, and predicted efficacy values for the rest 30% of data using the TabPFN model. The MAEs are plotted against IQRs in Figure 1a for the test data. We could see a clear trend between the IQR and MAE, especially in the middle IQR range. A higher IQR leads to the distribution of MAE towards higher values, which means the predictions are less accurate. Because there is little data falling into the smallest and largest IQR data ranges, the estimates are noisy and inconclusive of the trend in these ranges. We would like to point out a single data point with lower IQR does not necessarily mean it will have high prediction accuracy, but rather it has a higher chance to be accurate compared to data with higher IQR.

We then explored whether the IQRs can be used to differentiate models with higher and lower accuracies. Using Target1 (A) as an example, we generated eight different random permutations of training subset orders, and trained a collection of models by progressively adding more training subsets as context following the predefined random order, until every training subset was included. Figure 1b plots the correlation coefficient of each model against the mean IQR of its predictions. An observable negative correlation (Pearson's $r = -0.42$) emerges, confirming that models exhibiting lower overall uncertainty (lower mean IQR) tend to produce more accurate predictions. Crucially, this IQR metric is calculated without knowledge of the ground truth labels, making it a viable heuristic for model selection.

3.3 Model Selection with Inter-Quantile Range

Post-hoc ensembling (PHE) is a common technique to improve model performance by averaging predictions from multiple models trained with different model hyperparameters and different datasets. Given that IQR is indicative of model performance, we investigated its use as a metric for model selection within a PHE framework. For each test set, we trained an ensemble of 400 models, where each model was trained on k randomly selected training subsets (k was chosen randomly between 1 and 20 for each model). The process was repeated 3 times using different random seeds. We then compared three strategies: (1) a single baseline model trained on all available training data (**All data single model**), (2) a naive ensemble that averages predictions from all 400 models (**Full ensemble mean**), and (3) our uncertainty-guided ensemble that averages predictions from only the top 10% of models with the lowest mean IQR (**Top 10% ensemble mean**). All results are given as means and 95% confidence intervals of 3 parallel runs.

The results are summarized in Table 2 (MAE) and Table 3 (Correlation). While our selection strategy has a marginal effect on MAE, it provides a notable improvement in correlation for two of the three test datasets. For Target1 (A), the top-10% ensemble yields a correlation of 0.278 ± 0.015, a

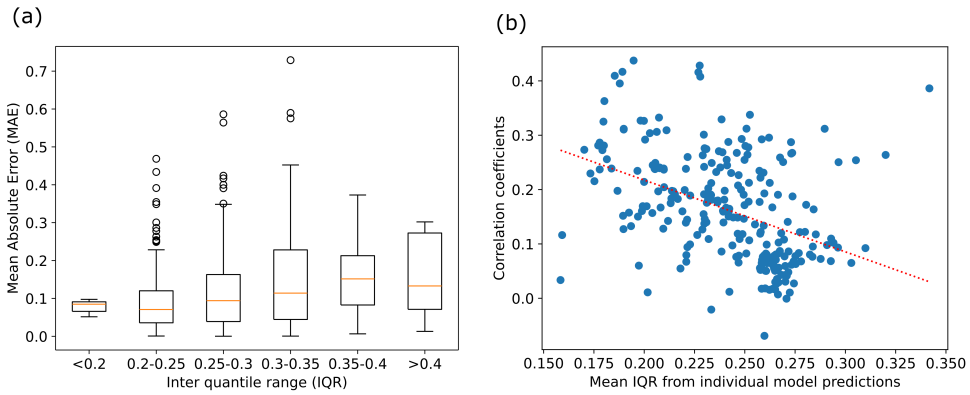

Figure 1: *Using inter-quantile ranges (IQRs) to evaluate prediction quality.* (a) Distribution of mean absolute error (MAE) categorized by (IQR) on randomly chosen held-out test set. Higher IQR is associated with higher error. (b) Scatter plot showing the negative correlation between a model's performance (correlation coefficient) and its mean prediction IQR on the Target1 (A) dataset. The dotted line shows the best linear fit ($r = -0.42$).

Table 2: Mean absolute errors (MAEs) for different datasets with different model selection strategies. For the ensemble strategies, results are reported as means and 95% confidence intervals of 3 individual runs using different random seeds.

Dataset	**Top 10% ensemble mean**	**Full ensemble mean**	**All data single model**	**Ensemble best** (oracle result)
Target1 (A)	0.270 ± 0.005	0.268 ± 0.002	0.278	0.197
Target1 (B)	0.174 ± 0.001	0.169 ± 0.001	0.172	0.149
Target2	0.185 ± 0.001	0.189 ± 0.001	0.186	0.161

Table 3: Correlation coefficients for different datasets with different model selection strategies. For the ensemble strategies, results are reported as means and 95% confidence intervals of 3 individual runs using different random seeds.

Dataset	**Top 10% ensemble mean**	**Full ensemble mean**	**All data single model**	**Ensemble best** (oracle result)
Target1 (A)	0.278 ± 0.015	0.257 ± 0.012	0.051	0.544
Target1 (B)	0.072 ± 0.005	0.086 ± 0.020	0.112	0.430
Target2	0.246 ± 0.015	0.230 ± 0.002	0.230	0.384

substantial improvement over the performance of a single model trained with all data (correlation of 0.051). Similarly, for Target2, the top-10% ensemble achieves the highest correlation, highlighting the effectiveness of our uncertainty-guided approach. While Target1 (B) remains a challenging case for all methods, these results demonstrate that IQR-based model selection is a promising strategy. We would like to emphasize that the extra amount of compute is acceptable, given that each model in the ensemble only operates on a limited size of data, and the inference with different models in the ensemble is embarrassingly parallelizable. Furthermore, this approach provides a direct way to handle large amount of context data that do not fit into a single forward pass of the TabPFN model. The "**Ensemble Best**" column shows the performance of the single best model in the ensemble from all 3 parallel runs. It is an oracle result which means there is no way to identify a single best model *a priori* for new test data. Comparing our approach with the "ensemble best" reveals the upper bound of performance and the remaining gap for improvement in model selection strategies.

4 Related Works

Tabular In-Context Learning The development of tabular foundation models began with TabPFN [3], a Transformer architecture pre-trained on synthetic data that established the viability of in-context learning for small tabular problems using deep learning models. This was followed by improvements in pre-training, such as using real-world data in TabDPT [5] for better generalization, and architectural refinements in TabPFN v2 [4]. A key line of subsequent research has focused on overcoming the scalability limitations imposed by the Transformer's quadratic complexity. Proposed solutions include retrieval and fine-tuning on local data subsets, as in LoCalPFN [7], and new architectures designed for larger contexts like TabICL [6]. Our work aligns with a more application focused research direction which adapts these general models for a specialized scientific domain.

Model Selection for Post-Hoc Ensembling Our work on uncertainty-guided model selection builds upon a rich history of post-hoc ensembling techniques. Foundational approaches include constructing ensembles from models saved during hyperparameter optimization, as demonstrated in Auto-sklearn [8], and dynamic weighted regressors that adjust model contributions based on performance metrics like RRMSE [13, 14].

More recently, sophisticated methods for model re-weighting and pruning have emerged. PSEO (Post-hoc Stacking Ensemble Optimization) frames model selection as a Bayesian hyperparameter optimization problem [15], while other work has employed end-to-end neural networks to dynamically re-weight base models [16]. Our method is particularly related to approaches that leverage model uncertainty. For instance, [17] used network uncertainty estimations to select a dynamic ensemble for improved adversarial robustness. Conformal prediction, which uses a small calibration set to provide statistically rigorous uncertainty bounds, has also been explored for model selection and aggregation [18–20].

siRNA Knockdown Efficacy Prediction Computational prediction of siRNA efficacy has been a long-standing goal to accelerate therapeutic development. Early efforts utilized classical machine learning models [21–24], and more recently a variety of neural architectures have been applied [11, 12, 25–27]. Currently, a state-of-the-art predictor is OligoFormer [12], which integrates thermodynamic features, pre-trained RNA-FM embeddings [28], and a custom Transformer-based encoder. Our work demonstrates that a general-purpose tabular foundation model can achieve competitive, and in some cases superior, performance compared to this specialized model.

5 Conclusion

We have shown that in-context learners like TabPFN, equipped with straightforward engineered features, could surpass specialized state-of-the-art models on siRNA knockdown efficacy prediction. Our primary contribution is the finding that the model's self-reported uncertainty, quantified by the inter-quantile range (IQR) of its predictions, is a potent, label-free heuristic for post-hoc model selection. We showed that an ensemble composed of models selected for their low prediction uncertainty can achieve superior performance, particularly in terms of correlation, compared to a single model using all available data or a naive ensemble average. This approach offers a practical strategy to enhance the reliability of in-context learners in real-world scenarios where ground truth labels are unavailable for model tuning. Future work should aim to narrow the gap to the oracle "ensemble best" by further iterating on the algorithm, and by validating this uncertainty-guided methodology across a broader range of biomolecule prediction tasks.

References

[1] Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, et al. Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery*, 18(6):463–477, 2019.

[2] Christopher Fifty, Jure Leskovec, and Sebastian Thrun. In-context learning for few-shot molecular property prediction, 2023. URL `https://arxiv.org/abs/2310.08863`.

[3] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second, 2023. URL `https://arxiv.org/abs/2207.01848`.

[4] Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

[5] Junwei Ma, Valentin Thomas, Rasa Hosseinzadeh, Hamidreza Kamkari, Alex Labach, Jesse C. Cresswell, Keyvan Golestan, Guangwei Yu, Anthony L. Caterini, and Maksims Volkovs. TabDPT: Scaling tabular foundation models on real data, 2025. URL https://arxiv.org/abs/2410.18164.

[6] Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICL: A tabular foundation model for in-context learning on large data, 2025. URL https://arxiv.org/abs/2502.05564.

[7] Valentin Thomas, Junwei Ma, Rasa Hosseinzadeh, Keyvan Golestan, Guangwei Yu, Maks Volkovs, and Anthony L. Caterini. Retrieval & fine-tuning for in-context tabular models. In *Advances in Neural Information Processing Systems*, volume 37, pages 108439–108467, 2024.

[8] Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum, and Frank Hutter. Efficient and robust automated machine learning. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

[9] Patrik Kenfack, Samira Ebrahimi Kahou, and Ulrich Aïvodji. Towards fair in-context learning with tabular foundation models, 2025. URL https://arxiv.org/abs/2505.09503.

[10] Bo Hu, Liping Zhong, Yuhua Weng, Ling Peng, Yuanyu Huang, Yongxiang Zhao, and Xing-Jie Liang. Therapeutic siRNA: state of the art. *Signal transduction and targeted therapy*, 5(1):101, 2020.

[11] Dieter Huesken, Joerg Lange, Craig Mickanin, Jan Weiler, Fred Asselbergs, Justin Warner, Brian Meloon, et al. Design of a genome-wide siRNA library using an artificial neural network. *Nature Biotechnology*, 23(8):995–1001, 2005.

[12] Yilan Bai, Haochen Zhong, Taiwei Wang, and Zhi John Lu. Oligoformer: an accurate and robust prediction method for siRNA design. *Bioinformatics*, 40(10):btae577, 2024.

[13] Niall Rooney, David Patterson, Sarab Anand, and Alexey Tsymbal. Dynamic integration of regression models. pages 164–173, 06 2004. ISBN 978-3-540-22144-9. doi: 10.1007/978-3-540-25966-4_16.

[14] Shikun Chen and Nguyen Manh Luc. RRMSE voting regressor: A weighting function based improvement to ensemble regression, 2022. URL https://arxiv.org/abs/2207.04837.

[15] Beicheng Xu, Wei Liu, Keyao Ding, Yupeng Lu, and Bin Cui. Pseo: Optimizing post-hoc stacking ensemble through hyperparameter tuning, 2025. URL https://arxiv.org/abs/2508.05144.

[16] Sebastian Pineda Arango, Maciej Janowski, Lennart Purucker, Arber Zela, Frank Hutter, and Josif Grabocka. Regularized neural ensemblers, 2025. URL https://arxiv.org/abs/2410.04520.

[17] Ruoxi Qin, Linyuan Wang, Xuehui Du, Xingyuan Chen, and Bin Yan. Dynamic ensemble selection based on deep neural network uncertainty estimation for adversarial robustness, 2023. URL https://arxiv.org/abs/2308.00346.

[18] Yachong Yang and A. Kuchibhotla. Selection and aggregation of conformal prediction sets. *Journal of the American Statistical Association*, 120:435 – 447, 2021. doi: 10.1080/01621459.2024.2344700.

[19] Ruiting Liang, Wanrong Zhu, and Rina Foygel Barber. Conformal prediction after data-dependent model selection, 2025. URL https://arxiv.org/abs/2408.07066.

[20] Matteo Gasparin and Aaditya Ramdas. Merging uncertainty sets via majority vote, 2024. URL https://arxiv.org/abs/2401.09379.

[21] Masatoshi Ichihara, Yoshiki Murakumo, Akio Masuda, Toru Matsuura, Naoya Asai, Mayumi Jijiwa, Maki Ishida, Jun Shinmi, Hiroshi Yatsuya, Shanlou Qiao, Masahide Takahashi, and Kinji Ohno. Thermodynamic instability of siRNA duplex is a prerequisite for dependable prediction of siRNA activities. *Nucleic Acids Research*, 35(18):e123–e123, 09 2007. ISSN 0305-1048. doi: 10.1093/nar/gkm699. URL https://doi.org/10.1093/nar/gkm699.

[22] Jean-Philippe Vert, Nicolas Foveau, Christian Lajaunie, and Yves Vandenbrouck. An accurate and interpretable model for siRNA efficacy prediction. *BMC Bioinformatics*, 7:520, Nov 2006. doi: 10.1186/1471-2105-7-520.

[23] Liangjiang Wang, Caiyan Huang, and Jack Y. Yang. Predicting siRNA potency with random forests and support vector machines. *BMC Genomics*, 11(3):S2, 2010. ISSN 1471-2164. doi: 10.1186/1471-2164-11-S3-S2. URL https://doi.org/10.1186/1471-2164-11-S3-S2.

[24] Kathryn R. Monopoli, Dmitry Korkin, and Anastasia Khvorova. Asymmetric trichotomous partitioning overcomes dataset limitations in building machine learning models for predicting siRNA efficacy. *Molecular Therapy Nucleic Acids*, 33:93–109, sep 2023. doi: 10.1016/j.omtn.2023.06.010. URL https://doi.org/10.1016/j.omtn.2023.06.010.

[25] Honggen Zhang, Xiangrui Gao, and Lipeng Lai. siDPT: siRNA efficacy prediction via debiased preference-pair transformer, 2025. URL https://arxiv.org/abs/2509.15664.

[26] Bin Liu, Huiya Huang, Weixi Liao, Xiaoyong Pan, Cheng Jin, and Ye Yuan. DeepSipred: A deep-learning-based approach on siRNA inhibition prediction. In *Proceedings of the 2024 4th International Conference on Bioinformatics and Intelligent Computing*, BIC '24, page 430–436, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400716645. doi: 10.1145/3665689.3665761. URL https://doi.org/10.1145/3665689.3665761.

[27] Wangdan Liao and Weidong Wang. DeepSilencer: A novel deep learning model for predicting siRNA knockdown efficiency, 2025. URL https://arxiv.org/abs/2503.04200.

[28] Jiayang Chen, Zhihang Hu, Siqi Sun, Qingxiong Tan, Yixuan Wang, Qinze Yu, Licheng Zong, Liang Hong, Jin Xiao, Tao Shen, Irwin King, and Yu Li. Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions, 2022. URL https://arxiv.org/abs/2204.00300.

[29] T. Xia, J. SantaLucia, M. E. Burkard, R. Kierzek, S. J. Schroeder, X. Jiao, C. Cox, and D. H. Turner. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson−Crick base pairs. *Biochemistry*, 37:14719–14735, 1998. doi: 10.1021/bi9809425.

A Details about data preparation and features used for TabPFN model

Small interfering RNA molecules (siRNAs) naturally exist as double stranded oligonucleotides, and we take the antisense strand (the strand that reverse complements to the mRNA transcript) as our input to the model. We ensure that siRNA molecules are 19 nucleotides (nts) in length. For any molecule longer than 19 nts from the patents, we take the first 19 nts after the leading uracil (U) as U is typically not part of the siRNA that matches to mRNA. We remove any siRNA shorter than 19 nts. For the mRNA, we take a slice from the transcript centered around the region that siRNA binds to, plus flanking regions of 19 nts on both sides. This adds up to 57 nts for the mRNA sequence input. When the flanking region goes beyond the transcript, "X"s are added in the places where nts are missing. Figure 2 gives a schematic explanation of the siRNA and mRNA sequences used to generate features.

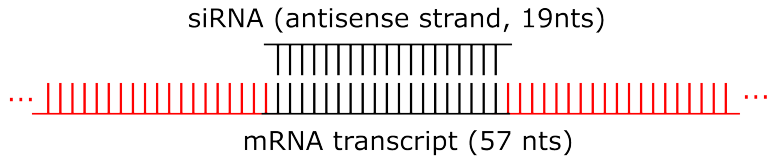

Figure 2: Schema of the siRNA and mRNA taken as input to the TabPFN model

We generated straightforward features for both the siRNA and mRNA sequences, which include one-hot features:

$$F_{\text{one-hot}} = \left(\bigoplus_{i=1}^{19} \text{OHE}_{\text{AUCG}}(S_{\text{siRNA}}[i]) \right) \oplus \left(\bigoplus_{j=1}^{57} \text{OHE}_{\text{AUCGX}}(S_{\text{mRNA}}[j]) \right)$$

where $S_{\text{siRNA}}[i]$ represents the ith position in the siRNA sequence, $S_{\text{mRNA}}[j]$ represents the jth position in the mRNA sequence, and $\text{OHE}(\mathbf{N})$ is the one-hot encoding for a nucleotide $\mathbf{N} \in \{\text{A, U, C, G}\}$ for siRNA and $\mathbf{N} \in \{\text{A, U, C, G, X}\}$ for mRNA. $\oplus$ represents feature concatenation. This corresponds to $4 \times 19 + 5 \times 57 = 361$ features.

We also include trimer count features for all possible trimers from the siRNA and mRNA vocabularies. Let T_{siRNA} be the set of all $4^3 = 64$ trimer combinations for siRNA and T_{mRNA} be the set of $5^3 = 125$ trimer combinations for mRNA, then the trimer count features are:

$$F_{\text{trimer}} = \left(\bigoplus_{t \in T_{\text{siRNA}}} \sum_{i=1}^{17} \mathbb{1}(S_{\text{siRNA}}[i:i+2] = t) \right) \oplus \left(\bigoplus_{t \in T_{\text{mRNA}}} \sum_{j=1}^{55} \mathbb{1}(S_{\text{mRNA}}[j:j+2] = t) \right)$$

where $S_{\text{siRNA}}[i:i+2]$ represents the three consecutive nucleotides in the siRNA sequence starting at position i and $S_{\text{mRNA}}[j:j+2]$ represents the three consecutive nucleotides in the mRNA sequence starting at position j. $\mathbb{1}(a = b)$ is 1 when $a = b$ evaluates to true, and 0 otherwise. This corresponds to $64 + 125 = 189$ features.

Finally, the thermodynamic parameters follows the same implementation as in OligoFormer[12], and are only calculated based on the siRNA sequence. These features include single and dinucleotide features:

$$\mathbf{N}(k) = \begin{cases} 0, & S_{\text{siRNA}}[k] \neq \mathbf{N} \\ 1, & S_{\text{siRNA}}[k] = \mathbf{N} \end{cases} \quad \mathbf{N} \in \{\mathbf{A}, \mathbf{U}, \mathbf{C}, \mathbf{G}\}, k \in [1, 19]$$

$$\mathbf{NM}(k) = \begin{cases} 0, & S_{\text{siRNA}}[k:k+1] \neq \mathbf{NM} \\ 1, & S_{\text{siRNA}}[k:k+1] = \mathbf{NM} \end{cases} \quad \textbf{N,M} \in \{\textbf{A, U, C, G}\}, k \in [1, 18]$$

$$\mathbf{N}(all) = \frac{\sum_{k=1}^{19}[\mathbf{N}(k)]}{19}, \quad \mathbf{N} \in \{A, U, C, G\}$$

$$\mathbf{NM}(all) = \frac{\sum_{k=1}^{18}[\mathbf{NM}(k)]}{18}, \quad \textbf{N,M} \in \{A, U, C, G\}$$

A second part of the thermodynamic parameters are calculated from the Gibbs free energy changes (ΔG) and enthalpy changes (ΔH). These are defined for each dinucleotide as [29]

Table 4: ΔG values for dinucleotides used for thermodynamic parameters calculation ($kcal/mol$)

First nt \ Second nt	A	U	C	G
A	-0.93	-1.10	-2.24	-2.08
U	-1.33	-0.93	-2.35	-2.11
C	-2.11	-2.08	-3.26	-2.36
G	-2.35	-2.24	-3.42	-3.26

Table 5: ΔH values for dinucleotides used for thermodynamic parameters calculation ($kcal/mol$)

First nt \ Second nt	A	U	C	G
A	-6.82	-9.38	-11.40	-10.48
U	-7.69	-6.82	-12.44	-10.44
C	-10.44	-10.48	-13.39	-10.64
G	-12.44	-11.40	-14.882	-13.39

There are also three special thermodynamic parameters, calculated as

$$\Delta G_{all} = \Delta G_{init} + \Delta G_{end} \times n_{\text{A/U end}} + \Delta G_{sym} + \sum_{i=1}^{18} \Delta G(S_{\text{siRNA}}[k:k+1])$$

$$\Delta H_{all} = \Delta H_{init} + \Delta H_{end} \times n_{\text{A/U end}} + \sum_{i=1}^{18} \Delta H(S_{\text{siRNA}}[k:k+1])$$

$$\Delta\Delta G_{all} = \Delta G(S_{\text{siRNA}}[1:2]) - \Delta G(S_{\text{siRNA}}[18:19]) + \Delta G_{end} \times n_{\text{A/U end}}$$

where $n_{\text{A/U end}} = \mathbf{A}(1) + \mathbf{U}(1) + \mathbf{A}(19) + \mathbf{U}(19)$ is the count of A and U nucleotides at the ends, $\Delta G_{init} = 4.09$, $\Delta G_{end} = 0.45$, $\Delta H_{init} = 3.61$, $\Delta H_{end} = 3.72$, and $\Delta G_{sym} = 0.43$ (all in $kcal/mol$) if the siRNA sequence is its own reverse complement, and zero otherwise.

The full set of thermodynamic features consist of

$$\begin{aligned} F_{thermo} = [&\Delta\Delta G_{all}, \Delta G(S_{\text{siRNA}}[1:2]), \Delta H(S_{\text{siRNA}}[1:2]), \mathbf{U}(1), \mathbf{G}(1), \\ &\Delta H_{all}, \mathbf{U}(all), \mathbf{UU}(1), \mathbf{G}(all), \mathbf{GG}(1), \mathbf{GC}(1), \mathbf{GG}(all), \\ &\Delta G(S_{\text{siRNA}}[2:3]), \mathbf{UA}(all), \mathbf{U}(2), \mathbf{C}(1), \mathbf{CC}(all), \\ &\Delta G(S_{\text{siRNA}}[18:19]), \mathbf{CC}(1), \mathbf{GC}(all), \mathbf{CG}(1), \\ &\Delta G(S_{\text{siRNA}}[13:14]), \mathbf{UU}(all), \mathbf{A}(19)] \end{aligned}$$

which has a length of 24.

The final input feature is the combination of one-hot features, trimer count features and thermodynamic features:

$$F_{input} = F_{one-hot} \oplus F_{trimer} \oplus F_{thermo}$$

and the length is $361 + 189 + 24 = 574$.

B TabPFN Model Calibration with siRNA Data

To show that the model is well calibrated with the siRNA data, we combined all data and performed a 70%-30% train-test split. We ran inference on the test data using the train data as context, obtaining lower and upper bound predictions at different quantile levels using `reg.predict(test_X, output_type="quantiles", quantiles=[l, u])`, where `reg` is the trained TabPFN model and l, u are the lower and upper quantile bounds. We then calculated the empirical coverage as the fraction of true labels that fall within the predicted quantile range:

$$\text{Coverage} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(y^{(i)}_{\text{pred,lower}} < y^{(i)}_{\text{true}} < y^{(i)}_{\text{pred,upper}}).$$

The empirical coverage was plotted against the expected coverage (the quantile range, $u - l$) in Figure 3. The curve closely follows the diagonal line, indicating the model is well-calibrated, meaning its uncertainty estimates are reliable.

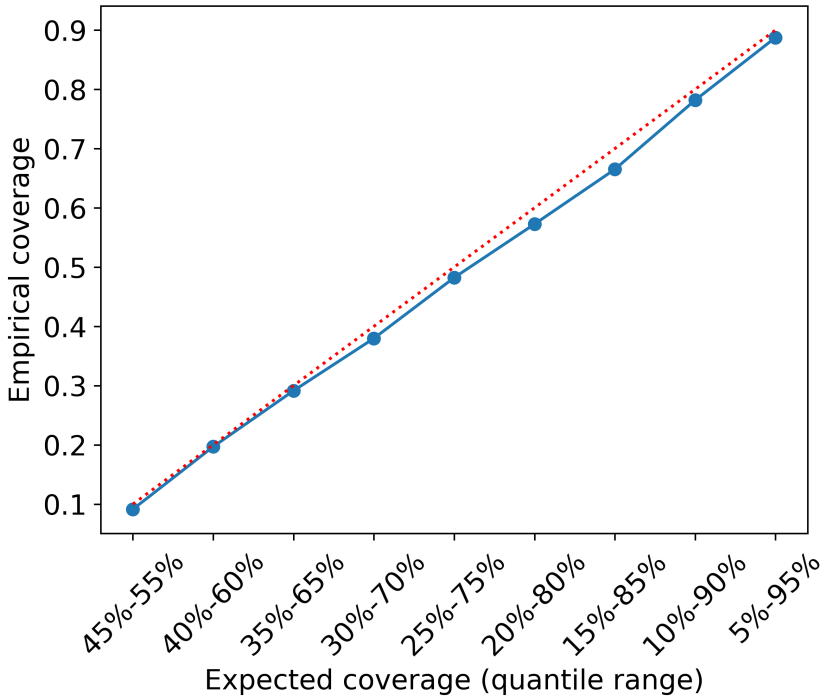

Figure 3: Relationship between empirical coverage and the expected quantile range provided to the model. The close adherence to the identity line ($y = x$) demonstrates that the model is well-calibrated.