Machine Learning Reveals Intrinsic Determinants of siRNA Efficacy

Christian Mandelli[1,✉] **and Giulia Crippa**[2]

[1]Department of Horticulture, Oregon State University, Corvallis, OR, United States.
[2]Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ, United States.

Small interfering RNAs (siRNAs) are widely used in therapeutics and agriculture for sequence-specific gene silencing. However, siRNA efficacy remains difficult to predict due to complex dependencies on sequence, structure, and thermodynamic properties. Existing computational tools largely rely on heuristic rules or pre-scored features, limiting generalizability and biological interpretability. Here, we present a machine learning model to predict siRNA efficacy directly from intrinsic antisense sequence features. Using a dataset of 2,428 experimentally validated siRNAs, we developed a comprehensive feature set that encompasses sequence composition, regulatory motifs, thermodynamic parameters, and structural complexity. We trained and evaluated multiple models for both regression and classification tasks. Support Vector Regression (SVR) achieved the best regression performance overall, with a predictive accuracy of R = 0.719 and R^2 = 0.516, while logistic regression achieved the best classification results with ROC = 0.886 and F1 = 0.809 using a combination of composition, motif, and structural features. Among all features, position-specific nucleotides were the strongest predictors of efficacy, with a uracil at the 5′ antisense end (P1_U) and an adenine at the 3′ end (P19_A) showing the highest importance, consistent with known mechanisms of strand selection and RISC loading. Our approach improves both predictive power and biological interpretability compared to existing methods, eliminating reliance on external scoring functions. The resulting framework supports rational siRNA design for therapeutic applications, functional genomics, and non-transgenic crop protection strategies.

Keywords: siRNA | RNAi | Machine Learning | Support Vector Regression | Crop Protection | Drug Design

Correspondence: christian.mandelli@oregonstate.edu

Introduction

RNA interference (RNAi) is a conserved, sequence-specific gene regulatory mechanism found across eukaryotes, serving essential roles in antiviral defense, epigenetic modulation, and post-transcriptional gene silencing. Central to RNAi are small interfering RNAs (siRNAs), which originate from the cleavage of long double-stranded RNA (dsRNA) by Dicer-like (DCL) enzymes into 21- and 24-nucleotide duplexes. These siRNAs are incorporated into Argonaute (AGO) proteins to form the RNA-induced silencing complex (RISC), which guides mRNA degradation or transcriptional silencing via RNA-directed DNA methylation (RdDM) in plants or translational repression in animals (1, 2). These processes are not only fundamental to eukaryotic gene regulation but also represent powerful entry points for targeted genetic intervention.

Initially developed in model systems, siRNAs have become a clinically validated therapeutic platform, with several approved drugs—such as patisiran for hereditary transthyretin amyloidosis (3)—and expanding pipelines targeting previously inaccessible genes (2). By acting at the mRNA level, siRNAs provide a mechanism different from traditional small molecules or antibodies, enabling precise, transient, and programmable gene knockdown. In agriculture, RNAi technologies are used to suppress pathogens and modulate plant traits through methods such as host-induced gene silencing (HIGS) and virus-induced gene silencing (VIGS). More recently, spray-induced gene silencing (SIGS) has emerged as a promising non-transgenic approach, allowing topical application of double-stranded RNA (dsRNA) to provide protection without genetically modifying the host plant. These applications demonstrate the broad potential of RNAi translation—yet they also reveal a key challenge: siRNA design still relies on an empirical, unpredictable process.

A major obstacle to effectively using siRNA-based technologies—whether for therapy or agriculture—is the inconsistent effectiveness of siRNA. Although sequence complementarity is necessary for target binding, it does not ensure successful silencing. DCL processing of dsRNA can produce diverse siRNA pools with different strengths, and small changes in siRNA sequences can influence strand selection by RISC, leading to lower effectiveness or unintended off-target effects (4, 5). These effects reflect deeply embedded biophysical and biochemical limitations that affect RNAi efficiency at various stages, from initial duplex processing to target access and AGO loading. Additionally, siRNA activity is affected by a complex interplay of sequence, thermodynamics, structure, and context, many of which are hard to predict with traditional heuristic or rule-based design methods. As a result, designing effective siRNAs remains a lengthy, error-prone process, which limits wider use and scaling of RNAi in both medicine and agriculture.

Several computational tools for siRNA design have been developed—including siDirect, i-Score, RNAxs, Biopredsi, dsCheck, and si-Fi—which rely largely on empirically derived design rules or linear models emphasizing sequence complementarity, GC content, or thermodynamic asymmetry

(6–10). These tools have provided valuable guidance, particularly in early development, but as highlighted in (2), they often fail to account for the complex biology behind RNAi. Notably missing from many platforms are considerations of mRNA target accessibility, local RNA secondary structure, siRNA duplex stability, and off-target interactions—all of which influence silencing efficiency. Without incorporating these biological factors, such tools find it difficult to perform well across different organisms, cell types, or experimental conditions (9, 11–13).

A significant breakthrough was achieved with the $psRNAit$ platform (5), which integrates multiple siRNA prediction tools within a cascade support vector machine (SVM) framework. More recently, deep learning–based methods such as *DeepSilencer* (14) have further advanced the field by leveraging transformer–CNN hybrid architectures and multi-task learning, achieving state-of-the-art performance for siRNA knockdown prediction. However, these approaches often require large datasets and are less interpretable, limiting their applicability for mechanistic understanding and rational siRNA design.

In this work, we present a machine learning framework that significantly enhances siRNA efficacy prediction by leveraging only intrinsic 21 nt antisense siRNA sequence features, eliminating reliance on external scoring systems. Our approach integrates rigorous feature engineering—encompassing sequence composition, regulatory motifs, thermodynamic properties (e.g., ΔG, melting temperature), and structural complexity from RNAfold—with advanced nonlinear models, including Support Vector Regression (SVR) and XGBoost. Our results demonstrate high predictive accuracy for both regression ($R = 0.719$, $R^2 = 0.516$) and classification tasks ($ROC = 0.886$, $F1 = 0.809$) using a combination of composition, motif, and structural features, outperforming linear models and single-feature approaches, while maintaining full biological interpretability and computational efficiency.

Materials and Methods

Data and codes are available on Github.

Dataset. For our model to be comparable with existing studies, the well-established siRNA efficacy datasets available from (15) was used. In detail, our dataset comprises 2,428 experimentally tested siRNAs targeting 34 mRNAs, derived from a reporter gene assay and employed in training the BIOPREDsi model.

For classification-based evaluations, two efficacy thresholds — 0.5 and 0.7 — were applied to convert continuous efficacy scores into binary labels for high versus low activity.

Features Extraction. To enable predictive modeling of siRNA efficacy, a custom Python-based feature extraction pipeline that processes each 21-nucleotide antisense siRNA strand and encodes it into a structured numerical format suitable for our methodology was developed. A description of all features can be found in Table SN.1.

Recent work has shown that distinct classes of sequence-derived features contribute differentially to siRNA efficacy, with some influencing strand selection, others affecting target accessibility, and still others modulating off-target risk (2). Therefore, features from four primary categories were extracted: (1) sequence composition, (2) regulatory and structural motifs, (3) thermodynamic parameters, and (4) structural complexity.

The composition-based features include both position-specific nucleotide identities and global nucleotide distributions. These capture local sequence preferences known to influence strand selection and siRNA functionality (15). Global composition metrics such as GC content and skew between the $5'$ and $3'$ ends were included due to their relevance to thermodynamic asymmetry and RISC incorporation (16, 17).

To account for regulatory and structural motifs, patterns at both termini of the siRNA duplex were identified, including recurring sequence motifs, palindromic structures, and simple nucleotide repeats. These elements have been previously associated with enhanced silencing efficiency, reduced off-target potential, or increased structural instability (4, 18–20).

Thermodynamic features were computed to capture both strand and duplex stability. This includes the free energy of hybridization with the target, the overall folding energy of the antisense strand, melting temperature estimates, and nearest-neighbor stacking energies. Local asymmetries across specific regions of the duplex were also evaluated using sliding-window calculations of free energy differences (18, 21–23).

Lastly, structural complexity features from RNA secondary structure predictions were generated using RNAfold (version 2.6.4) from the ViennaRNA Package (24). These include metrics related to folding energy, base-pairing density, loop topology, and nucleotide-level entropy. These properties inform on the accessibility and rigidity of the siRNA structure, both of which impact RISC loading and functional efficiency (2, 25).

All categorical features were one-hot encoded to ensure compatibility with machine learning models. The resulting feature matrix was exported as a CSV file and used as input for downstream model training and testing, for both regression and classification tasks.

Dataset Preprocessing and Analysis. Prior to model training, standard cleaning procedures were applied to ensure data integrity and consistency. All infinite values were replaced with missing values and subsequently dropped. A small number of sequences exhibiting extreme global folding energy (Global $\Delta G > 100$ or < -100 kcal/mol) were removed, as these likely represent non-biological outliers or computational artifacts from RNA secondary structure prediction. The dataset was then reindexed to ensure continuous indexing after row removal. This reduced the dataset from 2,428 to 2,408 siRNAs.

All numerical features were scaled, and categorical variables were one-hot encoded to ensure compatibility with machine learning models. The final preprocessed dataset contained 190 features, including one-hot encodings of positional nucleotides and selected di- and trinucleotide motifs. This matrix was used as input for all downstream modeling tasks. The structure of key thermodynamic and compositional features after preprocessing is shown in Figure SN.1, which illustrates distinct pairwise correlations and unimodal distributions, supporting the biological relevance and diversity of the cleaned dataset. Lastly, to address class imbalance in the binary classification task, a label balancing procedure was applied prior to training. The dataset was re-sampled using either undersampling or oversampling strategies, depending on the modeling objective. Specifically, an equal number of positive and negative examples was randomly selected to form a balanced training set, preserving the original label distribution proportions where possible.

Machine Learning Model Training and Evaluation. To assess the predictive power of the extracted siRNA features, a comprehensive machine learning framework encompassing both regression and classification tasks was implemented. All models were trained and evaluated using a unified pipeline based on scikit-learn, ensuring consistency in preprocessing, feature transformation, and validation procedures. Prior to model fitting, the full dataset was randomly split into training (80%) and testing (20%) subsets, stratified by efficacy label to preserve class proportions. Continuous features were standardized using StandardScaler, while binary (categorical) features were passed through without transformation using a ColumnTransformer.

For regression analysis, the target variable was treated as continuous, and three predictive models were evaluated: ordinary least squares linear regression (LR), support vector regression (SVR) with a radial basis function kernel, and gradient-boosted regression trees (XGBoost). Model performance was assessed using 5-fold cross-validation repeated across all training data partitions. The evaluation metrics included mean absolute error (MAE), root mean squared error (RMSE), coefficient of determination (R^2), and Pearson's correlation coefficient (R). In addition to the cross-validation metrics, each model was also evaluated on the held-out test set to assess its generalization capacity.

To investigate the contribution of different feature classes to model performance, models were systematically trained on all $2^k - 1$ non-empty combinations of $k = 4$ biologically motivated feature categories: sequence composition, regulatory motifs, thermodynamic stability, and structural complexity. For each feature subset, the corresponding columns were extracted from the training and test data, processed independently, and passed to the model pipeline. This exhaustive combinatorial design enabled direct quantification of the marginal and joint predictive utility of each biological feature group across both regression and classification objectives.

A parallel classification pipeline was implemented to predict binary siRNA efficacy labels (effective vs. ineffective). For this task, four classifiers were trained: logistic regression (LR), support vector machines (SVM), random forests (RF), and multi-layer perceptron neural networks (NN). As in the regression setting, models were trained using 5-fold stratified cross-validation, with performance evaluated using accuracy, precision, recall, F1 score, and the area under the receiver operating characteristic curve (ROC). Class imbalance was mitigated using undersampling, ensuring an equal representation of positive and negative examples in the training set. Test performance was computed separately on the untouched 20% holdout set.

All models were implemented using standard Python libraries, including scikit-learn for linear models, SVMs, and neural networks, and XGBoost for gradient-boosted regression. The evaluation pipeline and model training loops were designed to support modular experimentation across feature sets, model types, and target modalities. This framework enabled a rigorous comparison of both regression and classification strategies and allowed us to quantify how different combinations of siRNA features influence predictive accuracy.

Results

In this section, we discuss the results of several machine learning models for siRNA efficacy using both regression and classification tasks. Models were trained on combinations of four biologically feature categories: Composition (C), Motifs (M), Thermodynamic (T), and Structural (S) features.

Nonlinear Models and Composition Features Drive Superior siRNA Efficacy Prediction. Regression models were trained to predict continuous efficacy scores. Among the three tested algorithms – Linear Regression (LR), Support Vector Regression (SVR), and XGBoost – nonlinear methods substantially outperformed linear ones. Table 1 summarizes the test set performance of the top-performing models and feature combinations.

Among all algorithms, SVR achieved the highest performance ($R = 0.719$, $R^2 = 0.516$) using composition, motif, and structural feature (C+M+S). XGBoost and LR models performed slightly below SVR, particularly when structural and thermodynamic features were used alone.

Across all methods, models trained on composition features (C) consistently outperformed those trained on motif (M), thermodynamic (T), or structural (S) features alone. For instance, in the XGBoost model, using only composition features outperforms ($R = 0.689$, $R^2 = 0.743$) all other features composition.

Linear models achieved slightly lower predictive performance overall. The best linear model required all features ($R = 0.676$, $R^2 = 0.451$), underscoring the utility of nonlinear learners for capturing complex interactions between siRNA sequence and functions.

Table 1. Results summary (test set performance). Models trained on different feature group combinations, sorted by R Pearson coefficient. Best test R^2 and R values are highlighted. C: Composition features; M: Motif and regulatory elements; T: Thermodynamic stability features; S: Structural complexity features.

Task	Model	Feature Set	MAE	R^2	R
Regression	SVR	C + M + S	0.106	0.516	0.719
	SVR	C + M + T + S	0.106	0.515	0.719
	SVR	C + M + T	0.107	0.513	0.718
	XGBoost	C	0.110	0.473	0.689
	LR	C + M + T + S	0.113	0.451	0.676

Task	Model	Feature Set	Accuracy	F1	ROC
Classification *Label threshold: 0.5*	Logit	C + M + S	0.801	0.809	0.886
	Logit	C + M	0.801	0.809	0.885
	Logit	C + M + T	0.795	0.809	0.882
	RF	C + M + T + S	0.740	0.741	0.839
	SVM	C + M	0.708	0.729	0.806
Classification *Label threshold: 0.7*	SVM	C + S	0.760	0.763	0.844
	SVM	C	0.754	0.759	0.844
	SVM	C + T	0.760	0.759	0.843
	Logit	C + M + T + S	0.740	0.741	0.839
	RF	C + M	0.765	0.75	0.822

Composition-Based Features Enable Interpretable Models to Outperform Deep Learning in siRNA Efficacy Classification. We evaluated three classifiers – Logistic Regression (LR), Support Vector Machines (SVM), and Random Forests (RF) – to predict binary siRNA efficacy at two different thresholds levels (0.5 and 0.7). Performance was benchmarked using test F1-score and ROC.

As shown in Table 1, logistic regression models trained on composition, motif, and structure features (C+M+S) yielded the best performance at the 0.5 threshold (ROC = 0.886; F1 = 0.809). On the other hand, SVM trained on composition and structure (C+S) was the top performer at the 0.7 threshold (ROC = 0.844; F1 = 0.763). As per the regression results, features sets lacking composition showed substantially reduced accuracy (ROC < 0.73), confirming that sequence composition remains the dominant predictor of efficacy.

Compared to the recently proposed deep learning model DeepSilencer (AUC = 0.820, F1 = 0.775 in cross-dataset evaluation) (14), our logistic regression model achieved higher AUROC (0.886) and F1 (0.809) on the Huesken dataset. These results demonstrate that a feature-engineered and interpretable model can match or even outperform state-of-the-art deep learning approaches on this benchmark dataset while remaining computationally efficient and biologically transparent.

Position-Specific Nucleotides and Local Motifs Drive siRNA Efficacy. Feature importance analysis of the best-performing regression model (SVR trained on composition, motif, and structural features; C+M+S) revealed that position-specific nucleotides are the strongest predictors of siRNA silencing efficiency (Figure 1). The most informative feature was a uracil at the 5′ antisense end (P1_U), followed by an adenine at the 3′ antisense end (P19_A). Other important terminal positions included guanine at position 1 (P1_G) and cytosine or guanine at position 19 (P19_C, P19_G). Short sequence motifs also contributed to prediction accuracy: the UCG trinucleotide motif (UCG%) and the UG dinucleotide (UG%) frequencies ranked within the top 10 features, suggesting that local nucleotide context beyond single positions can modulate siRNA activity. Structural and compositional features such as global GC content (GC_Content) and GC skew between the 5′ and 3′ ends (GC_skew_5p_vs_3p) had moderate predictive value, while AU content at the 5′ end (five_prime_AU) was among the lower but still relevant contributors. Overall, these results indicate that effective siRNAs are primarily defined by specific terminal base identities (especially P1_U and P19_A) together with favorable short sequence motifs and moderate GC-related features, rather than by global thermodynamic properties alone.

Discussion

Our machine learning framework for predicting siRNA efficacy represents a significant advancement over existing computational tools, achieving high performance in both regression and classification tasks. For continuous efficacy prediction, Support Vector Regression (SVR) trained on a combined feature set of sequence composition, regulatory motifs, and structural complexity (C+M+S) achieved the best performance ($R = 0.719$, $R^2 = 0.516$). For binary classification, logistic regression trained on the same feature set outperformed competing models ($ROC = 0.886$, $F1 = 0.809$).

Across both regression and classification tasks, sequence composition features emerged as the most informative and biologically meaningful predictors. Notably, P1_U (uracil at the 5′ end of the antisense strand) and P19_A (adenine at the 3′ end) were the top contributors to model performance, as evidenced by their consistently high permutation importance across 10 random shuffles in our best-performing regression model (Figure 1).

Importantly, this pattern was not unique to SVR: analyses

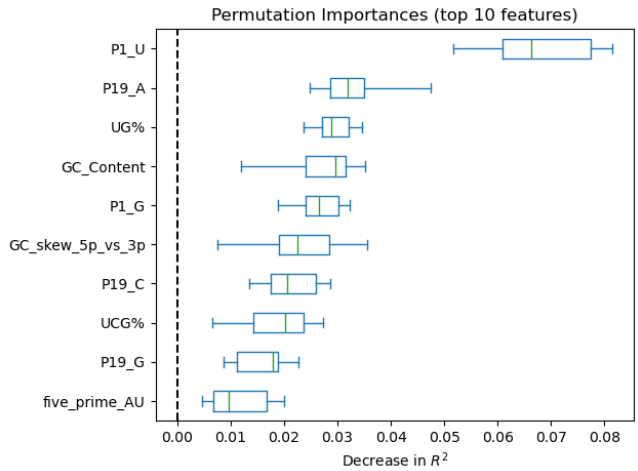

Fig. 1. Top 10 most important features ranked by permutation importance from the best-performing regression model: SVR trained on composition, motif, and structural features (C+M+S). Error bars represent standard deviation across 10 shuffled folds. Results from Ridge Regression, XGBoost, and classification tasks show consistent patterns, highlighting the robustness of position-specific features such as P1 and P19.

of other regression and classification models (Appendix Figures SN.2 and SN.3) revealed a strikingly similar ranking of features, with P1_U and P19_A repeatedly among the top predictors regardless of the algorithm[1]. Additional terminal nucleotides, such as P19_C and P1_G, also ranked highly in several models, further supporting the functional significance of both ends of the antisense strand.

The importance of P1_U aligns with well-established RNAi principles. Thermodynamic asymmetry at siRNA duplex ends facilitates strand selection by RISC. A uracil at the 5′ antisense end decreases local duplex stability, thereby promoting antisense strand loading. This mechanism is supported by prior studies showing that a less stable 5′ antisense end—enriched in U or A bases—favors guide strand selection (16, 26). Similarly, Sakurai et al. (27) demonstrated that Dicer discriminates effective siRNAs by sensing thermodynamic properties of the terminal base pairs, underscoring the functional relevance of position 1 nucleotides.

By contrast, P19_A highlights a less-characterized but potentially important determinant of siRNA efficacy. Our model suggests that adenine at the 3′ of the antisense end enhances RISC loading, likely through improved binding to the PAZ domain. This finding is consistent with Alagia et al. (28), who showed that optimal PAZ–overhang interactions promote siRNA potency and specificity. The strong predictive contribution of P19_A suggests that 3′-end composition plays a more active role in RISC engagement than previously appreciated. Consistent with this, other nucleotides at position 19, including cytosine (P19_C) and guanine (P19_G), also showed high importance (Figure 1), further supporting a key role of terminal base identity in siRNA efficacy.

Interestingly, global GC content was only moderately predictive (≈ 0.03 importance), indicating that while duplex stability influences silencing, position-specific sequence features are more informative for accurate efficacy prediction. Beyond composition, regulatory motifs such as the UCG trinucleotide motif (UCG%) and the UG dinucleotide frequency (UG%) were also ranked among the top predictors. These motifs may enhance RISC loading or siRNA–mRNA pairing, though experimental evidence remains limited. Prior studies have noted that such motifs can affect both immunogenicity and off-target effects (29, 30), suggesting a dual role in efficacy and safety.

In addition to sequence motifs, several composition-derived metrics, including GC_skew_5p_vs_3p and five_prime_AU, contributed moderately to prediction accuracy. These features likely capture differences in end stability that facilitate strand selection and guide strand incorporation.

Structural complexity features, derived from RNAfold predictions, ranked just below composition and motifs. These features capture local base-pairing density and accessibility, which likely influence guide strand unwinding and target hybridization. We propose that effective siRNAs balance structural rigidity and flexibility, enabling stable RISC incorporation while preserving target accessibility. In contrast, thermodynamic features such as global folding free energy (ΔG) were among the least predictive. Although earlier studies linked thermodynamic stability to silencing efficiency (26), our findings suggest that global thermodynamic metrics lack positional specificity and generalizability (28).

Overall, our model advances both predictive accuracy and biological interpretability. The consistent importance of P1_U and P19_A, together with informative short motifs and compositional metrics such as UCG% and GC_skew, highlights their potential as rational design principles for siRNA optimization. Their prevalence in effective siRNAs across a dataset of 2,428 siRNAs suggests broad generalizability.

Our findings have practical relevance for both therapeutic and agricultural RNAi applications. In medicine, they provide a biologically grounded framework for rational siRNA drug design, potentially reducing the need for extensive experimental screening. In agriculture, they can guide SIGS strategies, enabling non-transgenic crop protection (31). The interpretability of our feature set also facilitates mechanistic insights into siRNA function, which can improve understanding of RNAi dynamics.

Despite these advances, several limitations remain. Our training data were derived primarily from in vitro reporter sys-

[1] While permutation-based feature importance provides a model-agnostic and interpretable approach for quantifying the contribution of each input variable, its application to ordinary least squares (OLS) linear regression proved unstable in our setting. Specifically, the absence of regularization in OLS makes the model highly sensitive to collinearity and small perturbations in the feature space. As permutation importance involves randomly shuffling feature values, this instability led to inflated and erratic importance scores when using unregularized linear regression. To address this, we replaced OLS with Ridge regression (L2-regularized linear regression), which introduces a penalty on large coefficients and is known to be more robust to multicollinearity. We evaluated multiple values of the regularization parameter α, and found that Ridge regression with $\alpha = 5$ consistently produced stable results across cross-validation folds and yielded interpretable permutation importance profiles. We therefore report results using Ridge regression throughout.

tems, which may not fully capture the complexity of endogenous transcript environments. The strong influence of sequence composition might also reflect dataset-specific biases. Future work should focus on (i) validation across diverse biological systems, including primary cells and whole plants; (ii) integration of higher-order transcriptomic features, such as RNA-binding protein occupancy and epitranscriptomic modifications (m^6A-seq), which are known to affect RNA folding, stability, and accessibility (32, 33); and (iii) iterative refinement with experimental feedback, ensuring that computational predictions translate effectively to real-world applications.

Finally, we propose that siRNA efficacy reflects a multi-layered code integrating sequence, structure, and dynamic transcriptomic context. Achieving higher predictive power will likely require models that combine data-driven learning with biophysical principles. Physics-informed machine learning (PIML) (34, 35) offers a promising approach by embedding mechanistic constraints (e.g., RNA thermodynamics and RISC assembly rules) directly into statistical models. Such hybrid frameworks could improve generalization across biological systems while retaining mechanistic interpretability, bridging the current gap between computational performance and biological validity.

In summary, our study (i) identifies P1_U and P19_A as the most robust and biologically meaningful predictors of siRNA efficacy, (ii) demonstrates that position-specific sequence features outperform global metrics like GC content, (iii) provides a biologically interpretable model that can guide rational siRNA design for therapeutic and agricultural applications, and (iv) outlines a path toward hybrid biophysical–statistical models that enable context-aware and mechanistically grounded RNAi prediction.

Bibliography

1. Shijun Mi, Tao Cai, Yugang Hu, Yemiao Chen, Emily Hodges, Fangrui Ni, Liang Wu, Shan Li, Huanyu Zhou, Chengzu Long, et al. Sorting of small rnas into arabidopsis argonaute complexes is directed by the 5 terminal nucleotide. *Cell*, 133(1):116–127, 2008.
2. Dominic D. Martinelli. From sequences to therapeutics: Using machine learning to predict chemically modified sirna activity. *Genomics*, 116(2):110815, 2024.
3. U.S. Food and Drug Administration. Drugs@FDA: FDA-Approved Drugs – ONPATTRO (patisiran). https://www.accessdata.fda.gov/scripts/cder/daf/index.cfm?event=BasicSearch.process, 2018. Accessed July 2025.
4. Dianne S Schwarz, György Hutvágner, Tingting Du, Zuoshang Xu, Neil Aronin, and Phillip D Zamore. Asymmetry in the assembly of the rnai enzyme complex. *Cell*, 115(2):199–208, 2003.
5. Firoz Ahmed, Muthappa Senthil-Kumar, Xinbin Dai, Vemanna S Ramu, Seonghee Lee, Kirankumar S Mysore, and Patrick Xuechun Zhao. pssrnait: a web server for designing effective and specific plant sirnas with genome-wide off-target assessment. *Plant Physiology*, 184(1):65–81, 2020.
6. Stefanie Lück, Tino Kreszies, Marc Strickert, Patrick Schweizer, Markus Kuhlmann, and Dimitar Douchkov. sirna-finder (si-fi) software for rnai-target design and off-target prediction. *Frontiers in Plant Science*, 10:1023, 2019. doi: 10.3389/fpls.2019.01023.
7. Vijaya Sai Ayyagari. Design of sirna molecules for silencing of membrane glycoprotein, nucleocapsid phosphoprotein, and surface glycoprotein genes of sars-cov2. *Journal of Genetic Engineering and Biotechnology*, 20(1):1–12, 2022. doi: 10.1186/s43141-022-00346-z.
8. Aparna Chaudhary, Sonam Srivastava, and Sanjeev Garg. Development of a software tool and criteria evaluation for efficient design of small interfering rna. *Biochemical and Biophysical Research Communications*, 404(1):313–320, 2011. doi: 10.1016/j.bbrc.2010.11.114.
9. Sebastian Petri and Gunter Meister. sirna design principles and off-target effects. In *RNA therapeutics*, pages 49–60. Springer, 2013. doi: 10.1007/978-1-62703-311-4_4.
10. Luquan Wang and Forest Y Mu. A web-based design center for vector-based sirna and sirna cassette. *Bioinformatics*, 20(11):1818–1820, 2004. doi: 10.1093/bioinformatics/bth164.
11. Jiawen Chen and Wenbing Zhang. Kinetic analysis of the effects of target structure on sirna efficiency. *Journal of Chemical Physics*, 137(22):225101, 2012. doi: 10.1063/1.4769821.
12. Sebastian Petri, Anne Dueck, Gerhard Lehmann, Nicholas Putz, Sabine Rüdel, Elisabeth Kremmer, and Gunter Meister. Increased sirna duplex stability correlates with reduced off-target and elevated on-target effects. *RNA*, 17(4):737–749, 2011. doi: 10.1261/rna.2348111.
13. Steffen Schubert, Arnold Grünweller, Volker A Erdmann, and Jens Kurreck. Local rna target structure influences sirna efficacy: systematic analysis of intentionally designed binding regions. *Journal of Molecular Biology*, 348(4):883–893, 2005. doi: 10.1016/j.jmb.2005.03.011.
14. Wangdan Liao and Weidong Wang. Deepsilencer: A novel deep learning model for predicting sirna knockdown efficiency. *bioRxiv*, 2025. doi: 10.1101/2025.03.06.123456. doi pending publication.
15. Dieter Huesken, Joerg Lange, Craig Mickanin, Jan Weiler, Fred Asselbergs, Justin Warner, Brian Meloon, Sharon Engel, Avi Rosenberg, Dalia Cohen, et al. Design of a genome-wide sirna library using an artificial neural network. *Nature Biotechnology*, 23(8):995–1001, 2005.
16. Kumiko Ui-Tei, Yuki Naito, Fumitaka Takahashi, Takeshi Haraguchi, Hiroko Ohki-Hamazaki, Aya Juni, Ryu Ueda, and Kaoru Saigo. Guidelines for the selection of highly effective sirna sequences for mammalian and chick rna interference. *Nucleic Acids Research*, 32(3):936–948, 2004.
17. Huang Huang, Renping Qiao, Deyao Zhao, Tong Zhang, Youxian Li, Fan Yi, Fangfang Lai, Junmei Hong, Xianfeng Ding, Zhenjun Yang, et al. Profiling of mismatch discrimination in rnai enabled rational design of allele-specific sirnas. *Nucleic Acids Research*, 37(22):7560–7569, 2009.
18. Svetlana A Shabalina, Alexey N Spiridonov, and Aleksey Y Ogurtsov. Computational models with thermodynamic and composition features improve sirna design. *BMC Bioinformatics*, 7(1):65, 2006.
19. Sabine Klingelhoefer et al. Differential activity of sirna sequences in mammalian cells. *RNA Biology*, 6(4):321–327, 2009.
20. Xin Zhao, Wenli Xu, Jianming Fu, et al. A sirna database and design tool for rna interference. *Nucleic Acids Research*, 33:e127–e127, 2005.
21. Masaki Ichihara, Yoshiki Murakumo, et al. Design of effective sirna sequences with high functionality. *Nucleic Acids Research*, 35(3):e27, 2007.
22. Andrew S. Peek. Predicting sirna efficiency with thermodynamic features and support vector machines. *BMC Bioinformatics*, 8:104, 2007.
23. Istvan Ladunga. More effective sirna design through sifting of overlapping sirnas that match identical targets. *Nucleic Acids Research*, 35(4):e27, 2006.
24. Ronny Lorenz, Stephan H Bernhart, Christian Honer Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. Viennarna package 2.0. *Algorithms for Molecular Biology*, 6(1):26, 2011. doi: 10.1186/1748-7188-6-26.
25. Zhichao J. Lu and David H. Mathews. Rna secondary structure prediction with free energy models. *Methods in Molecular Biology*, 395:3–26, 2007.
26. Anastasia Khvorova, Alexa Reynolds, and Seth D Jayasena. Functional assessment of sirna duplexes identifies thermodynamic characteristics that distinguish effective and non-effective sirnas. *Cell*, 115(2):209–216, 2003.
27. Katsutomo Sakurai, Sachi Yamasaki, Yoshiko Nitanai, et al. A role for human dicer in pre-risc loading of sirnas. *Nucleic Acids Research*, 38:e116, 2010. PMC3045585.
28. Andreia Alagia and Andreia F Jorge. Exploring paz/3-overhang interaction to improve sirna specificity: insights from atomistic simulations and free energy calculations. *ACS Medicinal Chemistry Letters*, 5(7):670–675, 2014.
29. Yuriy Fedorov, Emily M Anderson, Amanda Birmingham, Angela Reynolds, Jon Karpilow, Kathryn Robinson, Devin Leake, William S Marshall, and Anastasia Khvorova. Off-target effects by sirna can induce toxic phenotype. *RNA*, 12(7):1188–1196, 2006. doi: 10.1261/rna.30706.
30. Aimee L Jackson and Peter S Linsley. Recognizing and avoiding sirna off-target effects for target identification and therapeutic application. *Nature Reviews Drug Discovery*, 9(1):57–67, 2010. doi: 10.1038/nrd3010.
31. Satyanarayana Gouthu, Christian Mandelli, Britt A Eubanks, and Laurent G Deluc. Transgene-free genome editing and rnai ectopic application in fruit trees: potential and limitations. *Frontiers in Plant Science*, 13:979742, 2022. doi: 10.3389/fpls.2022.979742. PMID: 36325537.
32. Jonathan Mauer, Xiaolan Luo, Antoine Blanjoie, and et al. Reversible methylation of m6am in the 5 cap controls mrna stability. *Nature*, 541:371–375, 2017.
33. Nian Liu and Tao Pan. Merip-seq reveals the broad scope of m6a regulation in arabidopsis. *Nature Plants*, 1(11):1–12, 2015.
34. George E Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3:422–440, 2021.
35. Michael Richter and Alem Admasu. sirna features—automated machine learning of 3d molecular fingerprints and structures for therapeutic off-target data. 26(14):6795, 2025. doi: 10.3390/ijms26146795.

Supplementary Note 1: Dataset

Table SN.1. Summary of features used for siRNA efficacy modeling, grouped by category. Full details and abbreviations are provided in the Methods section.

Feature Category	Feature Description
Sequence Composition	P1–P21 (position-specific nucleotides)
	Mononucleotide frequencies (e.g., A%, C%, G%, U%)
	Dinucleotide frequencies (e.g., GC%, AU%, etc.)
	Trinucleotide frequencies (e.g., AAG%, GCU%, etc.)
	GC content
	GC skew between 5′ and 3′ ends
	Global AU/GC ratio
	AU/GC ratio at 5′ end
	AU/GC ratio at 3′ end
	Shannon entropy
	G excluded at position 13
Regulatory and Structural Motifs	Palindromic sequence motif
	Known enhancer motifs (e.g., UCU, UCCG, AAU, AUU, UUA, AUA)
	Known inhibitory or toxic motif (e.g., GUGG)
Thermodynamic Features	Global folding free energy (ΔG)
	Melting temperature (Tm)
	Duplex binding free energy
	Stacking energy (nearest-neighbor model)
Structural Complexity	Minimum free energy from RNAfold
	Number of base pairs
	Number of structural loops
	Energy per nucleotide

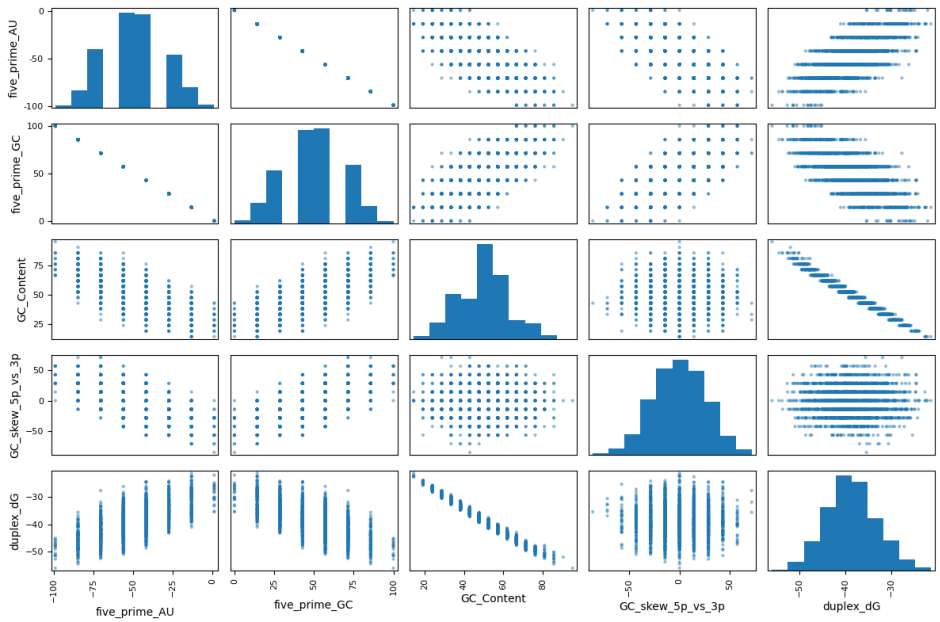

Fig. SN.1. Pairwise scatterplots and marginal distributions for selected thermodynamic and compositional features. Features shown include AU and GC content at the 5′ end, overall GC content, GC skew between the 5′ and 3′ ends, and duplex binding free energy (ΔG).

Supplementary Note 2: Regression Results

Table SN.2. Linear Regression Performance. Models trained on different feature group combinations, sorted by R Pearson coefficient. Best test R^2 and R values are highlighted. C: Composition features; M: Motif and regulatory elements; T: Thermodynamic stability features; S: Structural complexity features.

	Train				Test			
Feature Group	MAE	RMSE	R^2	R	MAE	RMSE	R^2	R
C + M + T + S	0.111	0.139	0.524	0.724	**0.113**	**0.021**	**0.451**	**0.676**
C + M + T	0.112	0.140	0.519	0.720	0.113	0.021	0.449	0.674
C + M + S	0.112	0.140	0.523	0.723	0.113	0.021	0.448	0.674
C + M	0.112	0.140	0.518	0.720	0.113	0.021	0.448	0.673
C + T + S	0.112	0.140	0.521	0.722	0.114	0.021	0.447	0.673
C + T	0.112	0.140	0.517	0.719	0.114	0.021	0.445	0.672
C + S	0.112	0.140	0.521	0.721	0.114	0.021	0.444	0.671
C	0.112	0.141	0.516	0.718	0.114	0.021	0.444	0.671
M + T + S	0.143	0.179	0.213	0.462	0.138	0.030	0.203	0.454
M + T	0.144	0.180	0.207	0.455	0.139	0.031	0.199	0.450
M + S	0.145	0.181	0.195	0.441	0.142	0.031	0.177	0.426
M	0.147	0.183	0.179	0.423	0.143	0.032	0.161	0.409
T + S	0.153	0.190	0.112	0.335	0.144	0.033	0.147	0.388
T	0.154	0.192	0.101	0.318	0.145	0.033	0.142	0.382
S	0.158	0.196	0.062	0.250	0.151	0.035	0.077	0.282

Table SN.3. Single Value Regression. Models trained on different feature group combinations, sorted by R Pearson coefficient. Best test R^2 and R values are highlighted. C: Composition features; M: Motif and regulatory elements; T: Thermodynamic stability features; S: Structural complexity features.

	Train				Test			
Feature Group	MAE	RMSE	R^2	R	MAE	RMSE	R^2	R
C + M + S	0.0106	0.0141	0.9951	0.998	**0.106**	**0.0185**	**0.516**	**0.719**
C + M + T + S	0.0107	0.0147	0.9947	0.997	0.106	0.0185	0.515	0.719
C + M + T	0.0107	0.0150	0.9945	0.997	0.107	0.0186	0.513	0.718
C + S	0.0105	0.0136	0.9955	0.998	0.107	0.0186	0.513	0.718
C + T + S	0.0107	0.0141	0.9951	0.998	0.107	0.0186	0.513	0.717
C + M	0.0106	0.0145	0.9949	0.997	0.107	0.0186	0.513	0.717
C + T	0.0107	0.0144	0.9949	0.997	0.108	0.0188	0.508	0.715
C	0.0106	0.0139	0.9953	0.998	0.108	0.0188	0.508	0.714
M + T	0.1182	0.1617	0.3598	0.600	0.145	0.0330	0.135	0.407
T	0.1474	0.1882	0.1331	0.365	0.146	0.0334	0.125	0.359
M + S	0.1153	0.1599	0.3738	0.611	0.149	0.0356	0.067	0.358
M	0.1348	0.1745	0.2546	0.505	0.148	0.0344	0.099	0.351
M + T + S	0.1037	0.1495	0.4527	0.673	0.151	0.0366	0.040	0.345
T + S	0.1392	0.1812	0.1967	0.444	0.149	0.0359	0.060	0.285
S	0.1531	0.1925	0.0934	0.306	0.152	0.0361	0.054	0.245

Table SN.4. Gradient Boosting. Models trained on different feature group combinations, sorted by R Pearson coefficient. Best test R^2 and R values are highlighted. C: Composition features; M: Motif and regulatory elements; T: Thermodynamic stability features; S: Structural complexity features.

	Train				Test			
Feature Group	MAE	RMSE	R^2	R	MAE	RMSE	R^2	R
C	0.0778	0.0981	0.7643	0.8743	**0.1097**	**0.0201**	**0.4731**	**0.6894**
C + T + S	0.0747	0.0940	0.7835	0.8852	0.1109	0.0203	0.4691	0.6863
C + M + T + S	0.0748	0.0943	0.7824	0.8845	0.1115	0.0205	0.4625	0.6809
C + S	0.0766	0.0962	0.7735	0.8795	0.1117	0.0207	0.4580	0.6776
C + M + T	0.0759	0.0956	0.7763	0.8811	0.1120	0.0210	0.4493	0.6709
C + M + S	0.0760	0.0954	0.7773	0.8817	0.1125	0.0210	0.4497	0.6709
C + T	0.0767	0.0965	0.7722	0.8787	0.1121	0.0210	0.4490	0.6704
C + M	0.0773	0.0974	0.7680	0.8764	0.1125	0.0211	0.4463	0.6690
M + T	0.1138	0.1434	0.4964	0.7045	0.1381	0.0302	0.2074	0.4582
M + T + S	0.1088	0.1377	0.5355	0.7318	0.1398	0.0311	0.1844	0.4350
M	0.1403	0.1750	0.2500	0.5000	0.1439	0.0326	0.1458	0.3923
M + S	0.1213	0.1533	0.4250	0.6519	0.1440	0.0338	0.1148	0.3688
T	0.1247	0.1563	0.4019	0.6340	0.1453	0.0330	0.1349	0.3675
T + S	0.1193	0.1500	0.4490	0.6701	0.1455	0.0334	0.1240	0.3535
S	0.1391	0.1740	0.2587	0.5087	0.1536	0.0374	0.0210	0.2093

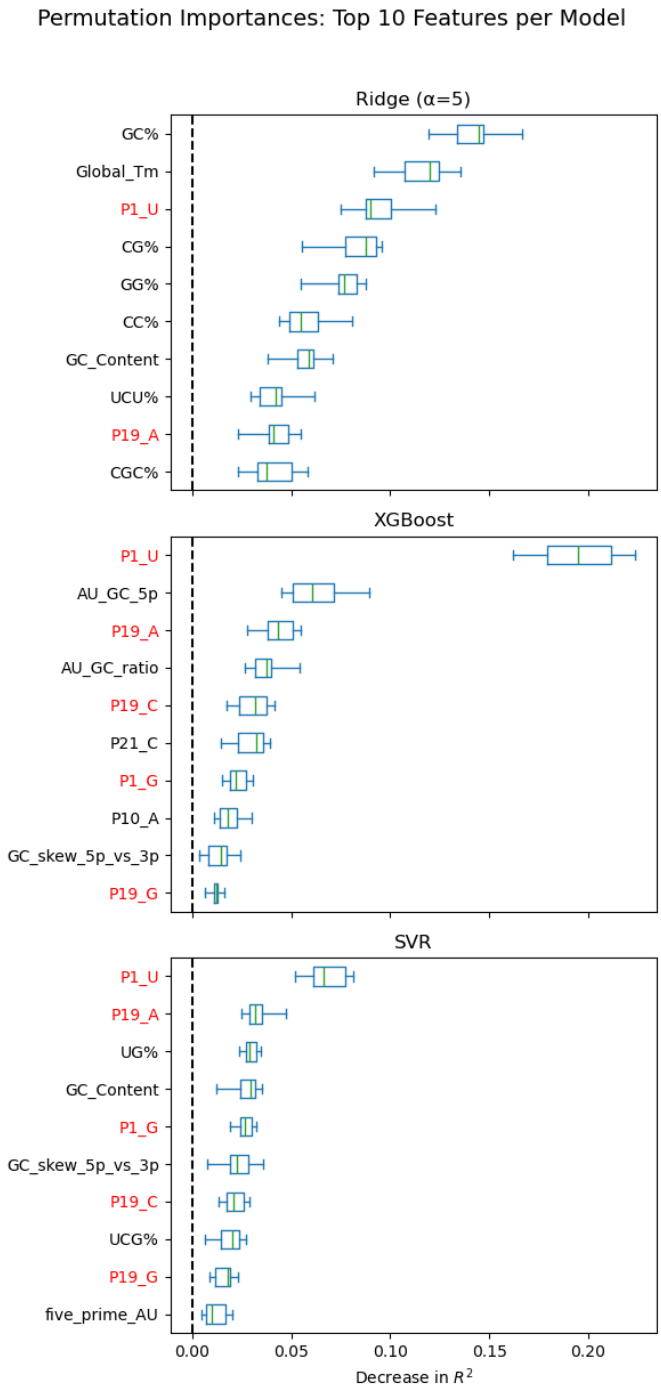

Fig. SN.2. Permutation importances for the top 10 features from three regression models (Ridge regression with $\alpha = 5$, XGBoost, and SVR), each trained on the combined composition, motif, and structural feature set (C+M+S). Feature importance was measured as the mean decrease in R^2 across 10 permutations, with error bars representing standard deviation. Position-specific nucleotides (e.g., P1_U, P19_A) consistently ranked among the most important features across models, supporting the robustness of the SVR results presented in the main text (Figure 1).

Supplementary Note 3: Classification Results

Table SN.5. Logistic regression. Comparison of models trained on different feature group combinations using two label thresholds (0.5 and 0.7), sorted by Test ROC. C: Composition; M: Motifs and regulatory elements; T: Thermodynamic stability; S: Structural complexity.

	Train			Test		
Feature Group	Accuracy	F1	ROC	Accuracy	F1	ROC
Label threshold: 0.5						
C + M + T	0.860	0.859	0.931	**0.801**	**0.809**	**0.886**
C + M	0.859	0.858	0.931	0.801	0.809	0.885
C + M + T	0.861	0.860	0.933	0.795	0.809	0.882
C + M + S	0.860	0.859	0.933	0.795	0.809	0.882
C	0.852	0.851	0.927	0.795	0.802	0.870
C + T	0.853	0.853	0.928	0.807	0.811	0.868
C + T + S	0.859	0.859	0.930	0.795	0.800	0.865
C + S	0.858	0.858	0.929	0.795	0.800	0.864
M + T + S	0.658	0.662	0.725	0.673	0.678	0.721
M + T	0.656	0.657	0.720	0.637	0.640	0.717
M + S	0.650	0.652	0.713	0.661	0.667	0.707
M	0.642	0.644	0.696	0.643	0.647	0.703
T	0.586	0.615	0.634	0.579	0.613	0.636
T + S	0.604	0.643	0.638	0.573	0.614	0.630
S	0.578	0.618	0.599	0.503	0.581	0.588
Label threshold: 0.7						
C + M + T + S	0.805	0.805	0.888	**0.740**	**0.741**	**0.839**
C + M + S	0.803	0.804	0.888	0.744	0.750	0.838
C + M + T	0.804	0.805	0.886	0.740	0.750	0.836
C + M	0.804	0.805	0.886	0.738	0.750	0.836
C + T + S	0.802	0.803	0.886	0.725	0.725	0.836
C + S	0.802	0.803	0.886	0.733	0.733	0.835
C + T	0.801	0.801	0.884	0.735	0.742	0.833
C	0.801	0.801	0.884	0.731	0.738	0.833
M + T + S	0.663	0.669	0.733	0.663	0.663	0.731
M + T	0.661	0.665	0.729	0.654	0.654	0.722
M + S	0.654	0.660	0.721	0.673	0.667	0.719
M	0.654	0.647	0.715	0.648	0.596	0.707
T + S	0.612	0.636	0.661	0.638	0.733	0.665
T	0.607	0.621	0.656	0.610	0.658	0.653
S	0.564	0.608	0.598	0.573	0.671	0.615

Table SN.6. Support Vector Machine (SVM) classification. Comparison of models trained on different feature group combinations under two label thresholds (0.5 and 0.7), sorted by Test ROC. C: Composition; M: Motifs and regulatory elements; T: Thermodynamic stability; S: Structural complexity.

	Train			Test		
Feature Group	Accuracy	F1	ROC	Accuracy	F1	ROC
Label threshold: 0.5						
C + M	0.937	0.937	0.984	**0.708**	**0.729**	**0.806**
C + M + T	0.936	0.936	0.983	0.719	0.741	0.805
C + M + S	0.933	0.933	0.983	0.702	0.718	0.801
C	0.938	0.937	0.987	0.713	0.718	0.800
C + M + T + S	0.932	0.932	0.982	0.702	0.706	0.800
C + T	0.937	0.937	0.986	0.702	0.706	0.799
C + T + S	0.938	0.938	0.985	0.702	0.706	0.794
M + T	0.714	0.713	0.794	0.643	0.647	0.681
M + T + S	0.734	0.731	0.814	0.614	0.593	0.676
M	0.679	0.640	0.741	0.602	0.547	0.674
M + S	0.720	0.711	0.793	0.585	0.548	0.660
T + S	0.648	0.672	0.702	0.544	0.571	0.575
T	0.617	0.660	0.662	0.550	0.613	0.613
S	0.603	0.677	0.645	0.550	0.645	0.588
C + S	0.928	0.928	0.983	**0.760**	**0.763**	**0.844**
C	0.926	0.926	0.982	0.754	0.759	0.844
C + T	0.923	0.923	0.981	0.754	0.759	0.843
C + T + S	0.925	0.925	0.981	0.758	0.761	0.843
C + M	0.921	0.921	0.980	0.754	0.756	0.840
C + M + S	0.925	0.925	0.980	0.758	0.761	0.840
C + M + T + S	0.921	0.921	0.979	0.758	0.761	0.839
C + M + T	0.918	0.918	0.979	0.760	0.763	0.839
M + T	0.695	0.696	0.767	0.658	0.651	0.708
M + T + S	0.714	0.720	0.793	0.652	0.650	0.707
M	0.679	0.680	0.736	0.642	0.633	0.693
M + S	0.709	0.717	0.776	0.642	0.648	0.690
T + S	0.637	0.669	0.685	0.625	0.667	0.659
T	0.613	0.637	0.660	0.621	0.649	0.653
S	0.593	0.651	0.623	0.621	0.670	0.617

Table SN.7. Random Forest. Models trained on different group combinations, sorted by test ROC. Rows are grouped by label threshold (0.5 or 0.7). C: Composition; M: Motifs and regulatory elements; T: Thermodynamic stability; S: Structural complexity.

	Train			Test		
Feature Group	Accuracy	F1	ROC	Accuracy	F1	ROC
Label threshold: 0.5						
C + M + T + S	0.805	0.805	0.888	**0.740**	**0.741**	**0.839**
C + M + S	0.803	0.804	0.888	0.744	0.750	0.838
C + M + T	0.804	0.805	0.886	0.740	0.750	0.836
C + M	0.804	0.805	0.886	0.738	0.750	0.836
C + T + S	0.802	0.803	0.886	0.725	0.725	0.836
C + S	0.802	0.803	0.886	0.733	0.733	0.835
C + T	0.801	0.801	0.884	0.735	0.742	0.833
C	0.801	0.801	0.884	0.731	0.738	0.833
M + T + S	0.663	0.669	0.733	0.663	0.663	0.731
M + T	0.661	0.665	0.729	0.654	0.654	0.722
M + S	0.654	0.660	0.721	0.673	0.667	0.719
M	0.654	0.647	0.715	0.648	0.596	0.707
T + S	0.612	0.636	0.661	0.638	0.733	0.665
T	0.607	0.621	0.656	0.610	0.658	0.653
S	0.564	0.608	0.598	0.573	0.671	0.615
Label threshold: 0.7						
C + M	1.000	1.000	1.000	**0.765**	**0.750**	**0.822**
C + S	1.000	1.000	1.000	0.731	0.700	0.816
C + T + S	1.000	1.000	1.000	0.738	0.713	0.811
C	1.000	1.000	1.000	0.715	0.696	0.810
C + M + T + S	1.000	1.000	1.000	0.742	0.729	0.806
C + M + T	1.000	1.000	1.000	0.719	0.692	0.802
C + M + S	1.000	1.000	1.000	0.735	0.704	0.801
C + T	1.000	1.000	1.000	0.727	0.721	0.800
M + T + S	0.999	0.999	1.000	0.673	0.667	0.719
M + T	0.999	0.999	1.000	0.681	0.671	0.706
M + S	0.983	0.983	0.999	0.635	0.629	0.698
M	0.708	0.701	0.793	0.631	0.604	0.675
T + S	0.999	0.999	1.000	0.600	0.621	0.665
T	0.999	0.999	1.000	0.615	0.621	0.654
S	0.862	0.867	0.949	0.563	0.575	0.589

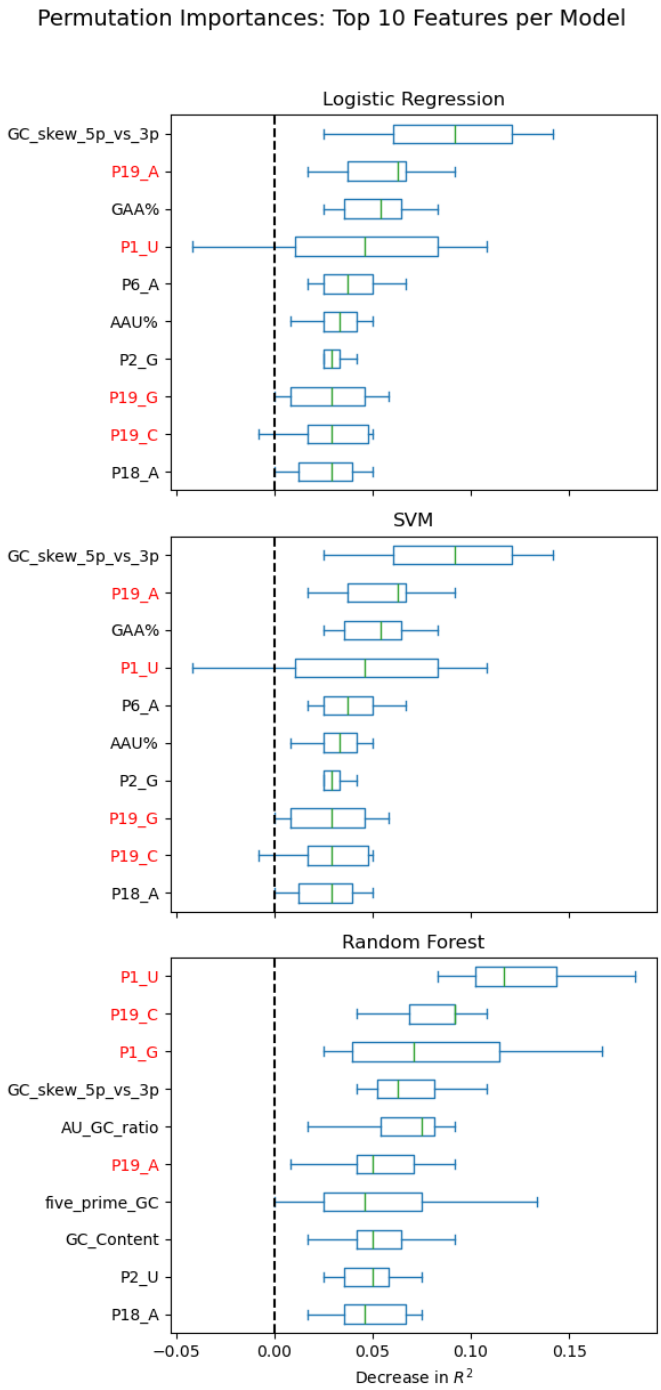

Fig. SN.3. Permutation importances for the top 10 features from three classification models (Logistic Regression, SVM, and Random Forest), each trained on the combined composition, motif, and structural feature set (C+M+S). Feature importance was measured as the mean decrease in R^2 across 10 permutations, with error bars representing standard deviation. Consistent with regression models, position-specific nucleotides (especially P1_U and P19_A) were among the top predictors, indicating that these features are robust determinants of siRNA efficacy regardless of model architecture.