Accurately Predicting Functional Fully Modified siRNAs with Advanced Machine Learning Techniques

Kathryn R. Monopoli,[1,2] Bailey Sostek,[3] RNA Therapeutics Institute siRNA screening consortium,[1] Dmitry Korkin,[2,3] and Anastasia Khvorova[1]
[1]RNA Therapeutics Institute, University of Massachusetts Chan Medical School, [2]Department of Bioinformatics & Computational Biology, Worcester Polytechnic Institute,
[3]Department of Computer Science, Worcester Polytechnic Institute

Introduction

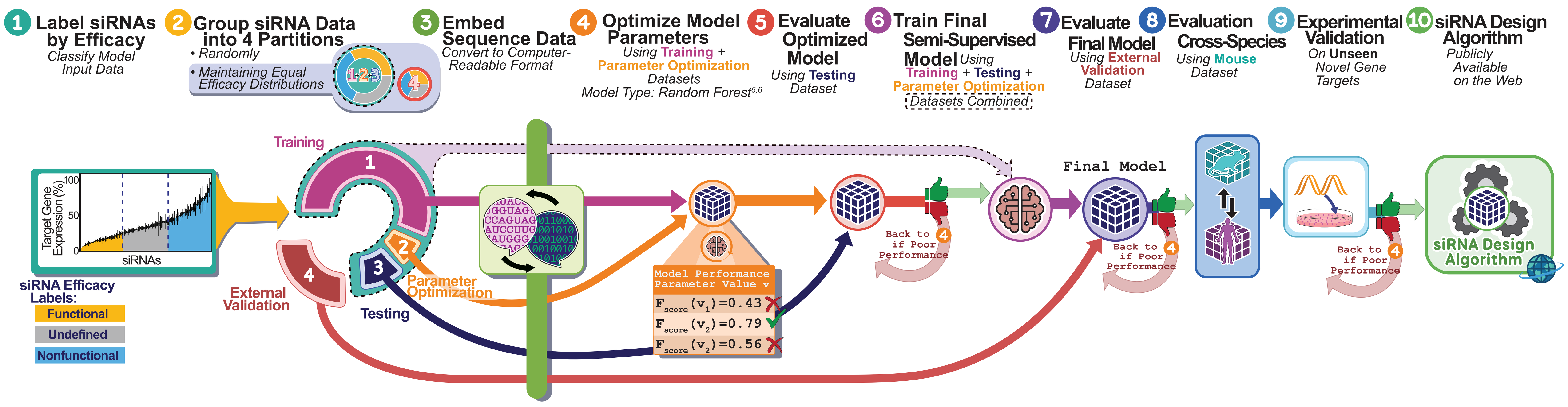
Problem: identifying potent therapeutic fully-modified siRNAs

- To be therapeutically active siRNAs must be fully modified[1]
- Existing methods for designing native siRNAs are not predictive for therapeutic, fully modified siRNAs[2]
- Both RISC (RNA induced silencing complex) interactions and target specific context define siRNA efficacy[2]
- Target specific features are not well defined for therapeutic siRNAs

Goal: develop algorithm to identify potent therapeutic siRNAs

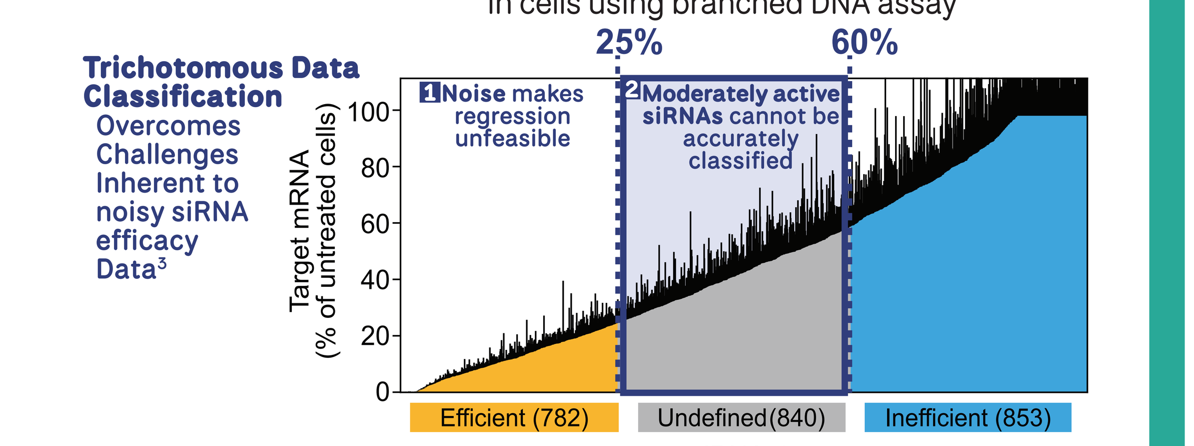
- Apply 1 large (**~5000 siRNAs**) therapeutic fully modified data-set to explore 2 target gene-aware real world design task 5
- Employ AI methods: 3 **deep learning**-based sequence embedding and 6 **semi-supervised learning** to capture both RISC and target-driven elements
- 7 markedly enhance prediction in 8 **species-specific** manner
- 4 Optimize & 9 **experimentally-validate** model on therapeutic design task across disease-relevant target genes
- Deploy 10 **siRNA Design Algorithm** as an easy-to-use **Web Portal** free for academic use[3,4]

Procedure for Building and Optimizing siRNA Efficacy Prediction Model

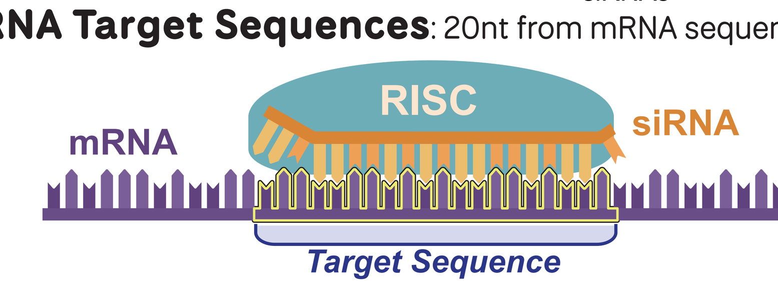


1 Data Used to Build Model

1 siRNA Efficacy Data: endogenous transcript knockdown evaluated in cells using branched DNA assay

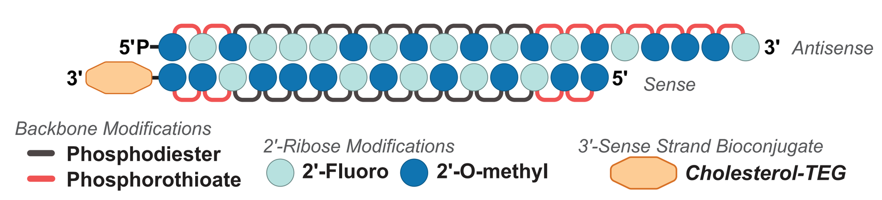


2 siRNA Target Sequences: 20nt from mRNA sequence



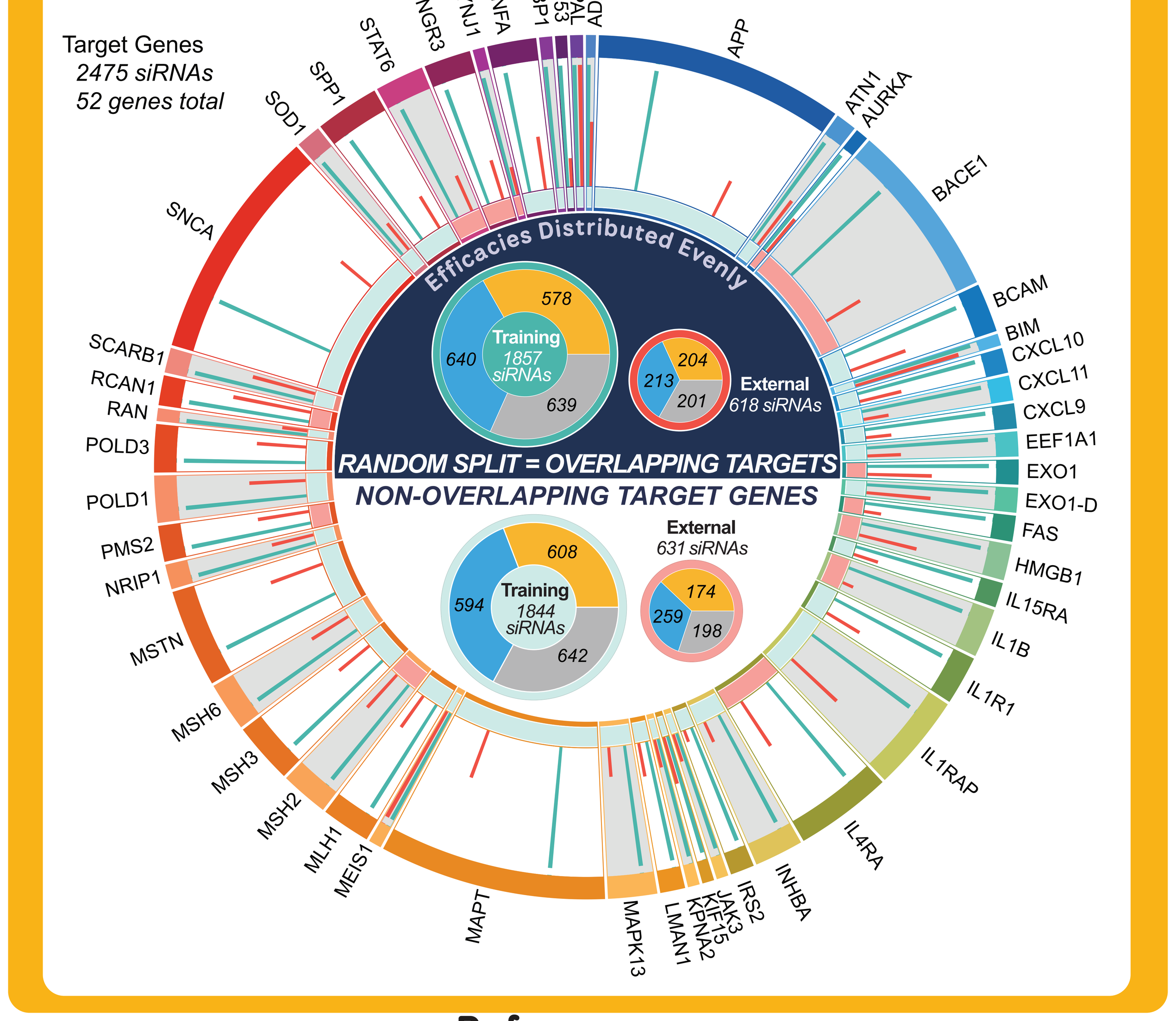
*2475 human siRNAs (shown), 2865 mouse, 5340 total

Chemical Scaffold of Fully Modified Therapeutic siRNA



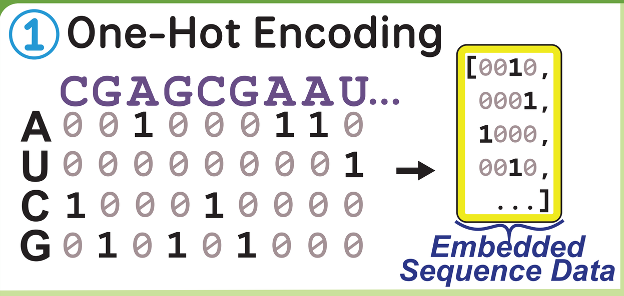
2 Data Partitioning: Random vs. by Target Gene

Target Gene split evaluates real-world siRNA design task

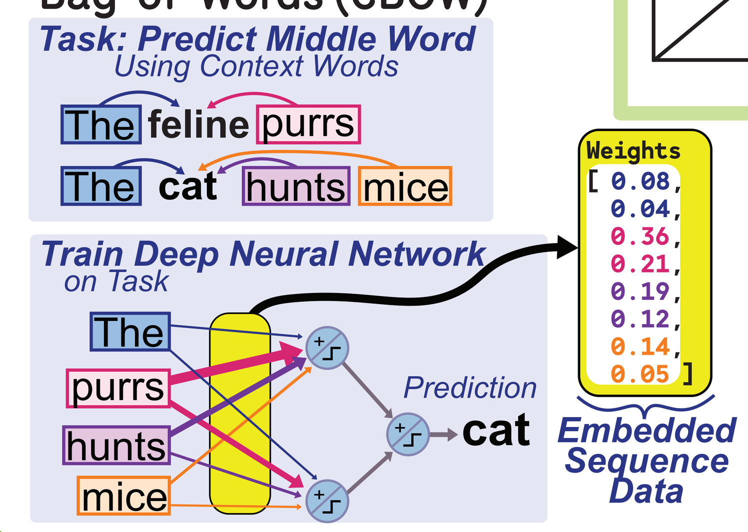


3 Sequence Feature Embedding Methods Significantly Impact Model Predictive Power

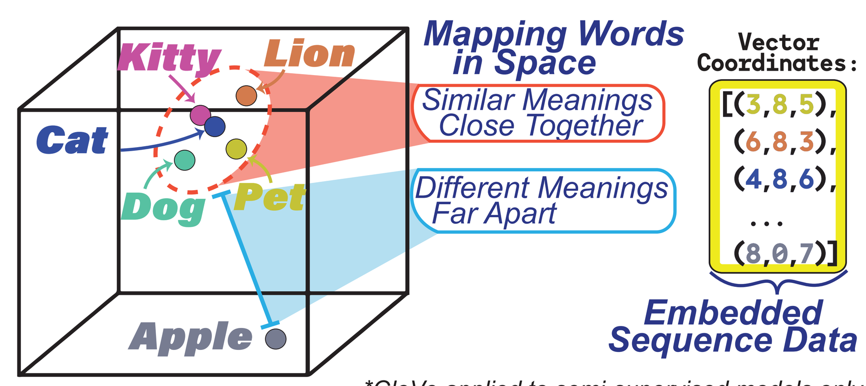
1 One-Hot Encoding



2 Deep Continuous Bag-of-Words (CBOW)[7]

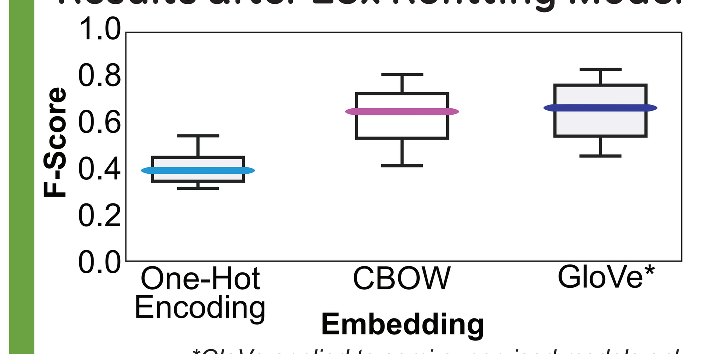


3 Global Vector Embedding[8]* (GloVe)



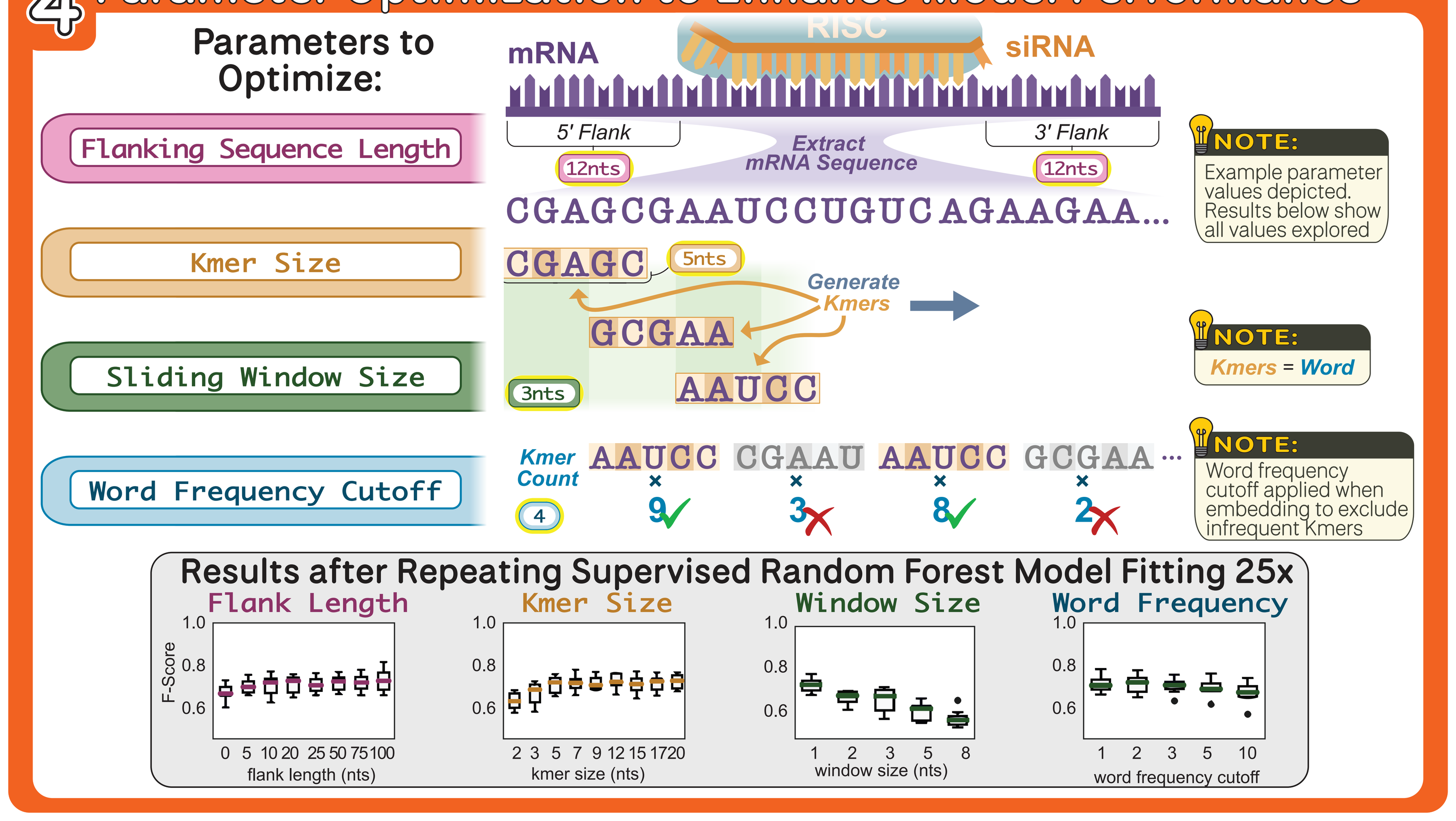
*GloVe applied to semi-supervised models only

Results after 25x Refitting Model

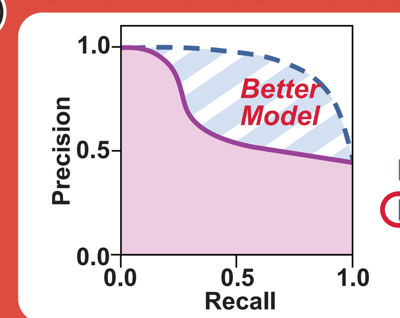


*GloVe applied to semi-supervised models only

4 Parameter Optimization to Enhance Model Performance



5 Performance Evaluation Metrics: Precision-Recall Curve & F-Score



$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

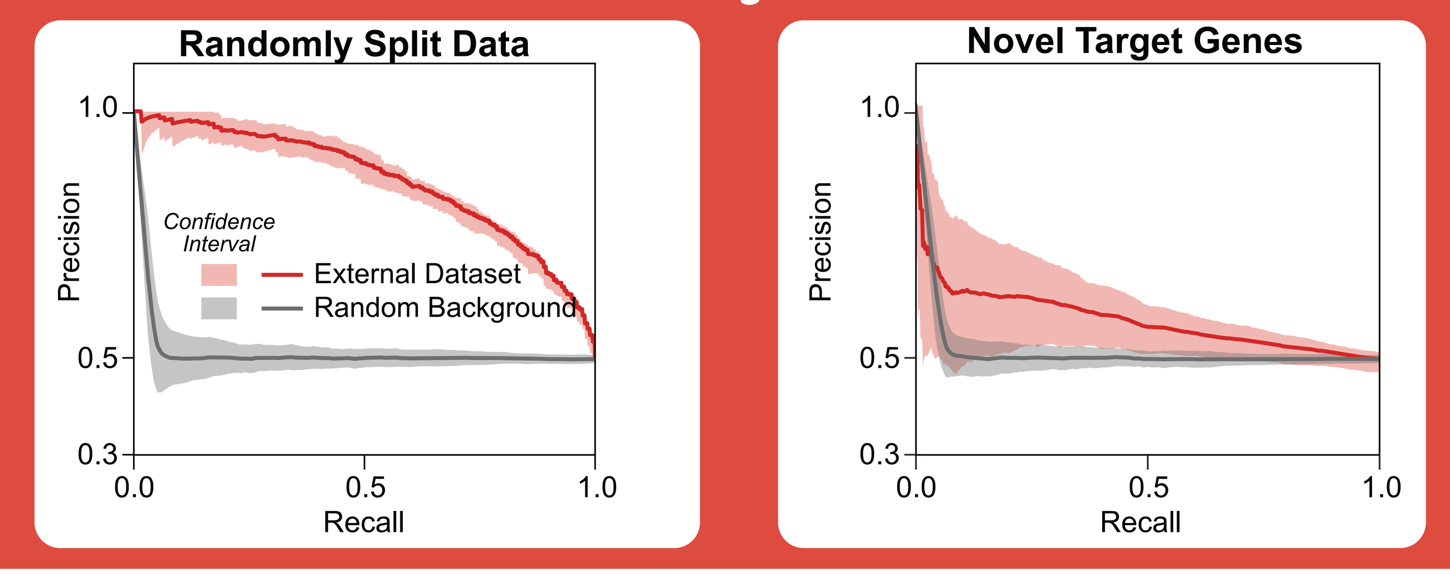
$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{F-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

F-Score combines Precision and Recall into a single metric

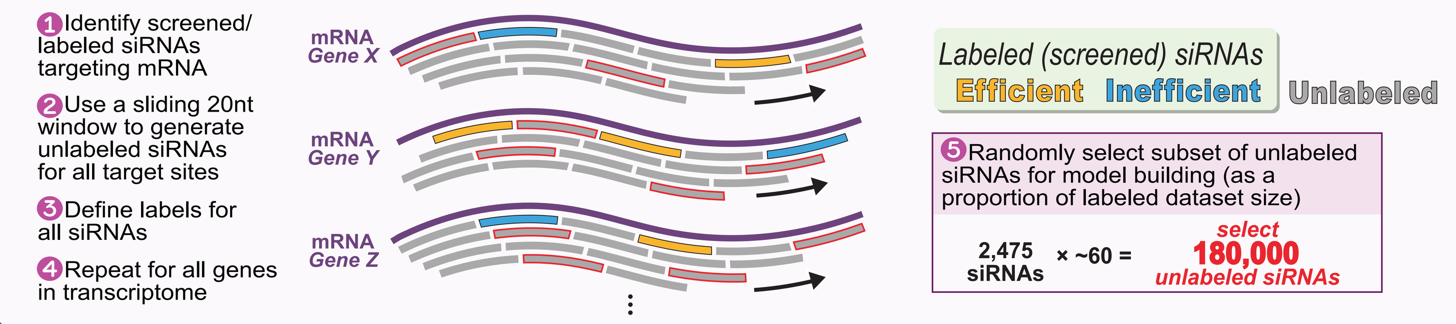
Useful for quick assessments (model optimization)

Supervised Model is Highly Predictive on Randomly-Split Data but not Novel Target Genes

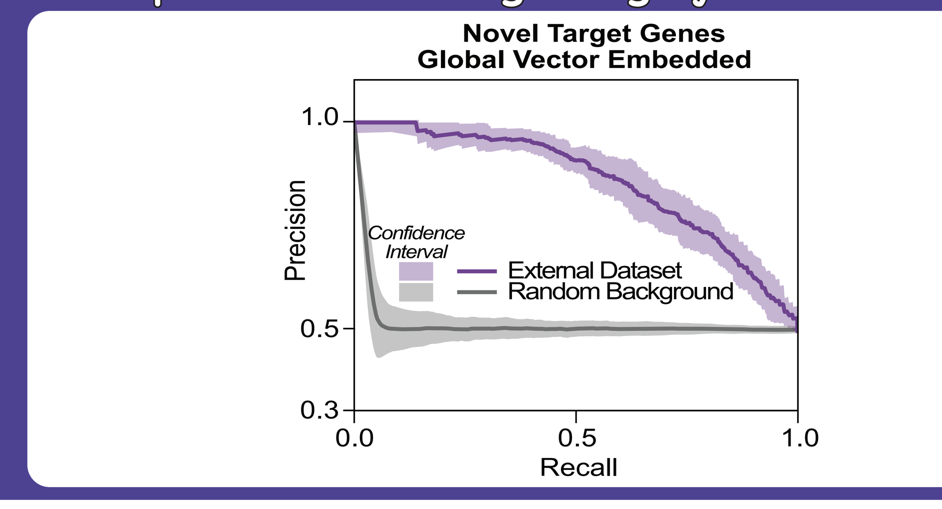


6 Semi-Supervised Learning Applies Unlabeled Data

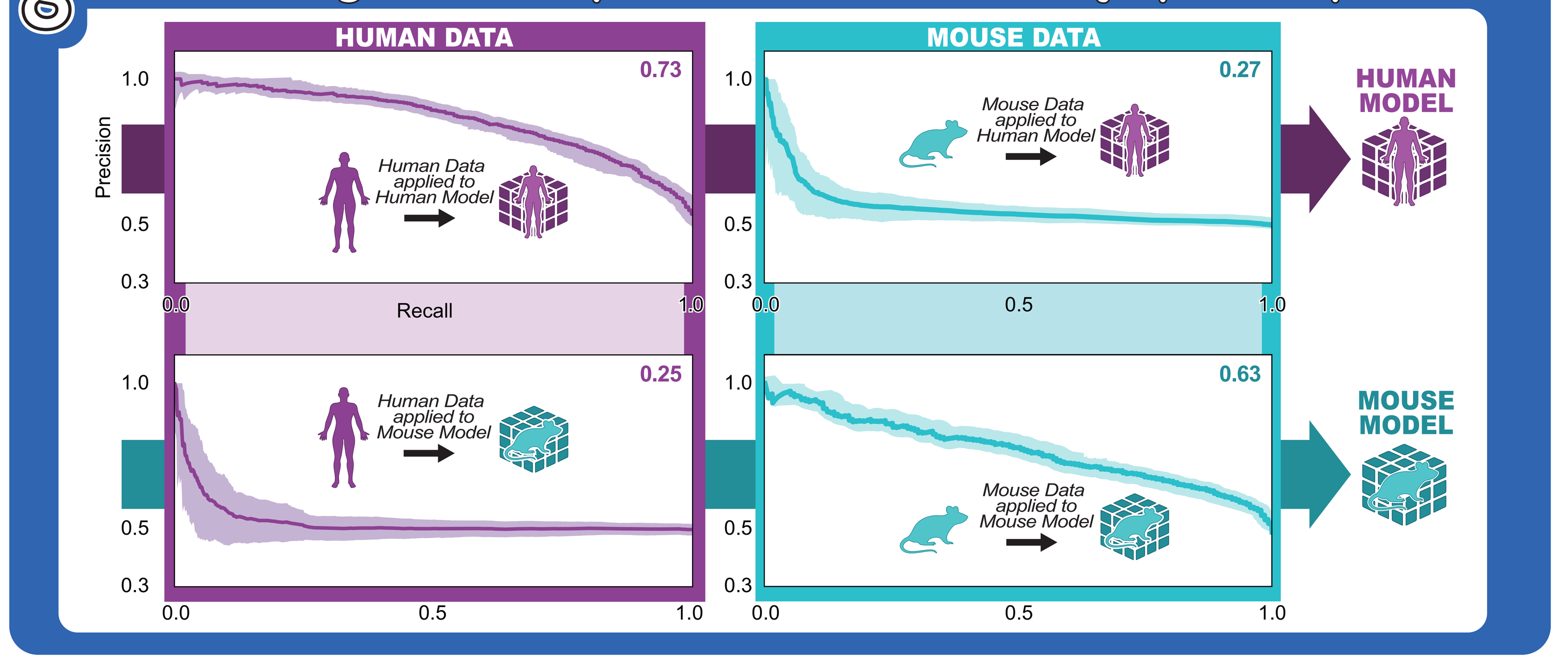
Generating Unlabeled Data: siRNAs with Unknown Efficacies



7 Semi-Supervised Learning is Highly Predictive on Novel Targets

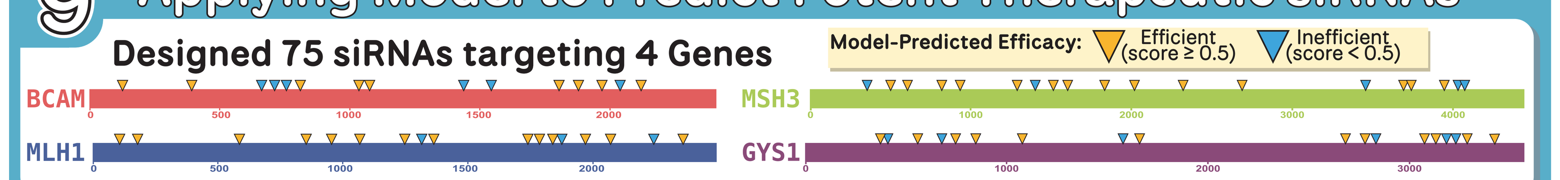


8 siRNA Design: Semi-Supervised Model is Strictly Species Specific



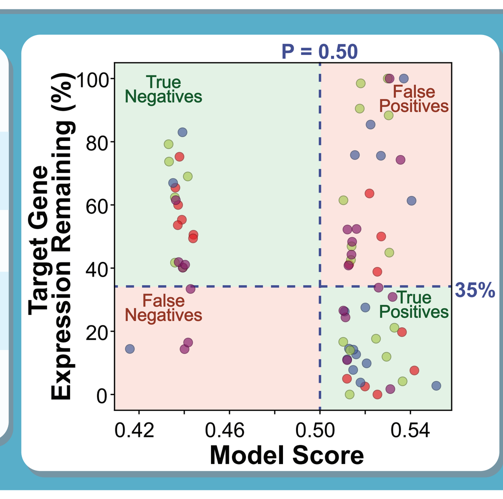
9 Applying Model to Predict Potent Therapeutic siRNAs

Designed 75 siRNAs targeting 4 Genes



Experimentally Validating siRNA Efficacy Prediction Model

1. Apply model to score siRNAs targeting 4 new genes not in dataset
2. Select top 15-18 scoring and 3-6 bottom-scoring compounds per gene
3. Synthesize 75 siRNAs in fully modified therapeutically permissive scaffold*
4. Evaluate endogenous target transcript knockdown in human HeLa cells



Semi-Supervised Model

	Inefficient (Predicted)	Efficient (Predicted)
Inefficient (Actual)	True Negatives 19	False Positives 22
Efficient (Actual)	False Negatives 5	True Positives 29

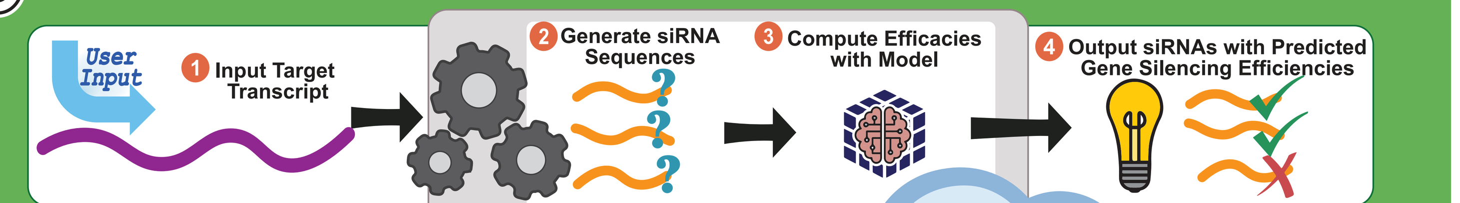
F-Score = 0.73
Accuracy = 64%

Linear Model[2]

	23	28
	11	13

F-Score = 0.40
Accuracy = 48%

10 Algorithm to Design Therapeutic siRNAs is Live!



Scan QR Code To Access Web Application

Questions?
Kathryn Monopoli
krmonopoli@gmail.com
https://kmonopoli.github.io

References

1. Hassler MR, Turanov AA, Alterman JF, et al. (2018). Comparison of partially and fully chemically-modified siRNA in conjugate-mediated delivery in vivo. *NAR*, 46, 2185–2196.
2. Shmushkovich T, Monopoli KR, Homsy D, Leyfer D, Betancur-Boissel M, Khvorova A, Wolfson AD. (2018) Functional features defining the efficacy of cholesterol-conjugated, self-deliverable, chemically modified siRNAs. *NAR*, 46, 10905–16.
3. Monopoli KR, Korkin D, Khvorova A. (2023) Asymmetric trichotomous partitioning overcomes dataset limitations in building machine learning models for predicting siRNA efficacy. *MTNA*.
4. Monopoli KR, Gross K, Sostek B, Korkin D, Khvorova A. Algorithm for selecting efficient siRNA through Deep Learning and Sequence feature representation. *In preparation for submission to Nature Methods.*
5. Breiman L. (2001) Random Forests. *Machine Learning* 45, 5–32.
6. Pedregosa et al. (2011) Scikit-learn: Machine Learning in Python, *JMLR* 12, 2825-2830.
7. Mikolov P, Sutskever I, Chen K, Corrado GS, Deanrt J. (2013) Distributed Representations of Words and Phrases and their Compositionality. Advances in Neural Information Processing Systems (NIPS), 26.
8. Pennington J, Socher J, Manning C. 2014. GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics,1532–1543.