

DeepRNAi: Toward Reducing Late-Stage Setbacks with a Frontier Model for siRNA Design

Nicolas Y. Masse, Tammy T.Y. Lau, Jeff Wintersinger, Mykhaylo Slobodyanyuk, Roujia Li, Vivian S. Chu, Helen Zhu, Bhargav Kanuparthi, Jonathan La, Alice J. Gao, Brodie Fischbacher, Krista Arseneault, Sydney Johnson, Wang-Ting Lu, Noah Lorincz-Comi, Emma Bohn, Christopher B. Cole, Gregory R. Hoffman, Nathan H. Lazar, Scott D. Findlay, Albi Celaj, Brendan J. Frey

¹Deep Genomics, Toronto, ON, Canada; Cambridge MA, USA
[†]Corresponding authors: albi@deepgenomics.com, scott.findlay@deepgenomics.com



1. Background and Objectives

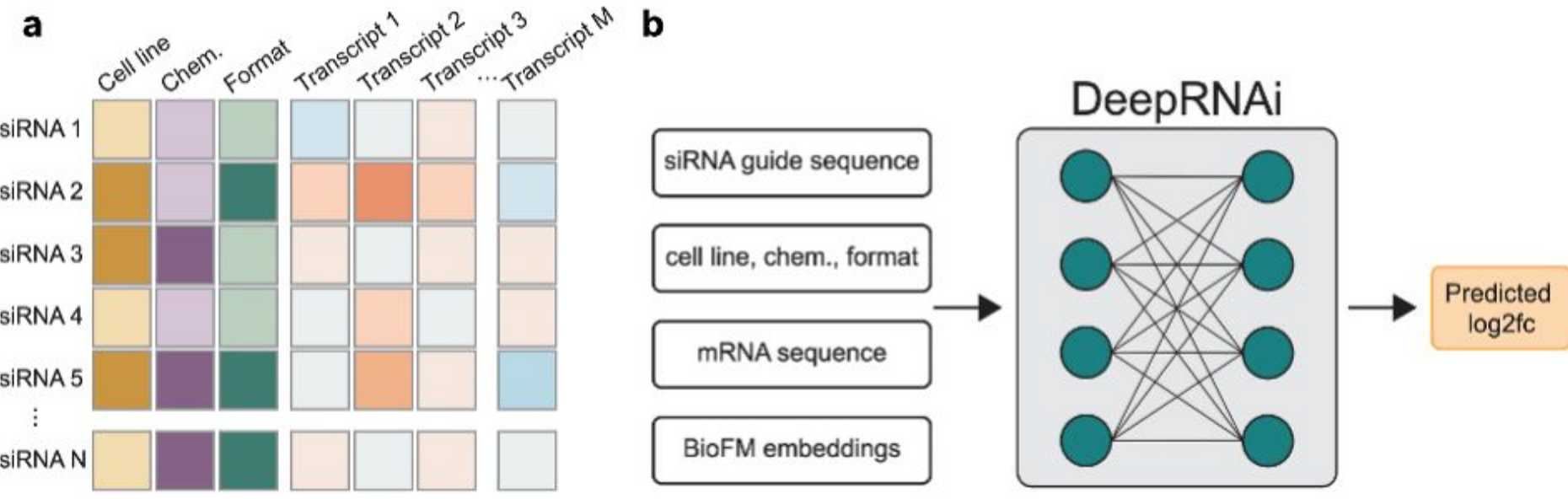
- **siRNA** has proven to be a successful oligo therapeutic modality, with eight siRNAs now approved for clinical use by the FDA.
- **Off-target liabilities** pose risks to the *efficient* development of putatively safe molecules to advance into the clinic:
 1. Unexpected off-target liabilities in animal models used for toxicology screening increase the risk of program failure.
 2. Molecules conferring knockdown of Near-Perfect Match (NPM) off-targets in *in vitro* human cell line screens are undesirable and pose regulatory hurdles.

Objectives:

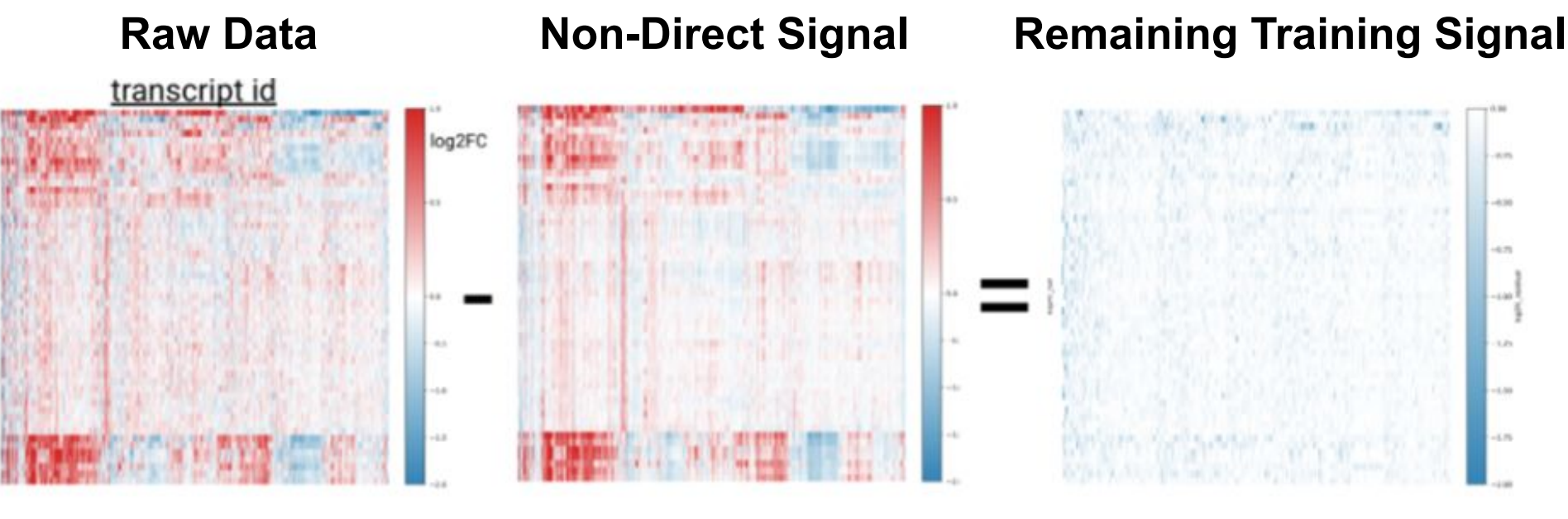
1. Generate experimental data and computational infrastructure to build, train, and evaluate a state-of-the-art model for predicting siRNA efficacy and specificity that leverages the power of Deep Genomics' Biology Foundation Models (BioFMs)—**DeepRNAi**.
2. Begin to use DeepRNAi to accelerate the preclinical development of siRNA therapeutics and increase the probability of clinical success.

2. Methods

- **Fit for purpose data generation**
 - A total of **2,043** fully chemically modified **siRNAs** were designed against 70 human target genes. These siRNAs were enriched for those with many Near-Perfect Matches (NPMs) to transcriptome-wide off-targets.
 - HepG2 or HuH7 cells were transfected with 10 nM of individual siRNAs, with RNA collected after 48 hours.
 - RNA-seq followed by differential expression analysis was used to assign log2 fold-change (log2FC) values to each siRNA-gene pair, using the most highly expressed transcript for each expressed protein-coding gene.
 - This resulted in a total of over **15 million siRNA-gene pairs** with experimentally observed log2FCs for **training** and **evaluation** of **DeepRNAi**.
- **Model training and inference**
 - Proprietary deep learning architecture
 - Inputs:
 - Cell line (1 of 2 categorical values)
 - siRNA guide strand sequence
 - siRNA chemistry/format (1 of 2 categorical values)
 - mRNA sequence
 - **Deep Genomics' Biology Foundation Model (BioFM) embeddings** [1,2]—rich representations of causal molecular biological features and processes with per nucleotide resolution, trained on trillions of data points.
 - Output:
 - Predicted log2FC (resulting from direct, hybridization dependent knockdown) for an arbitrary siRNA-transcript pair



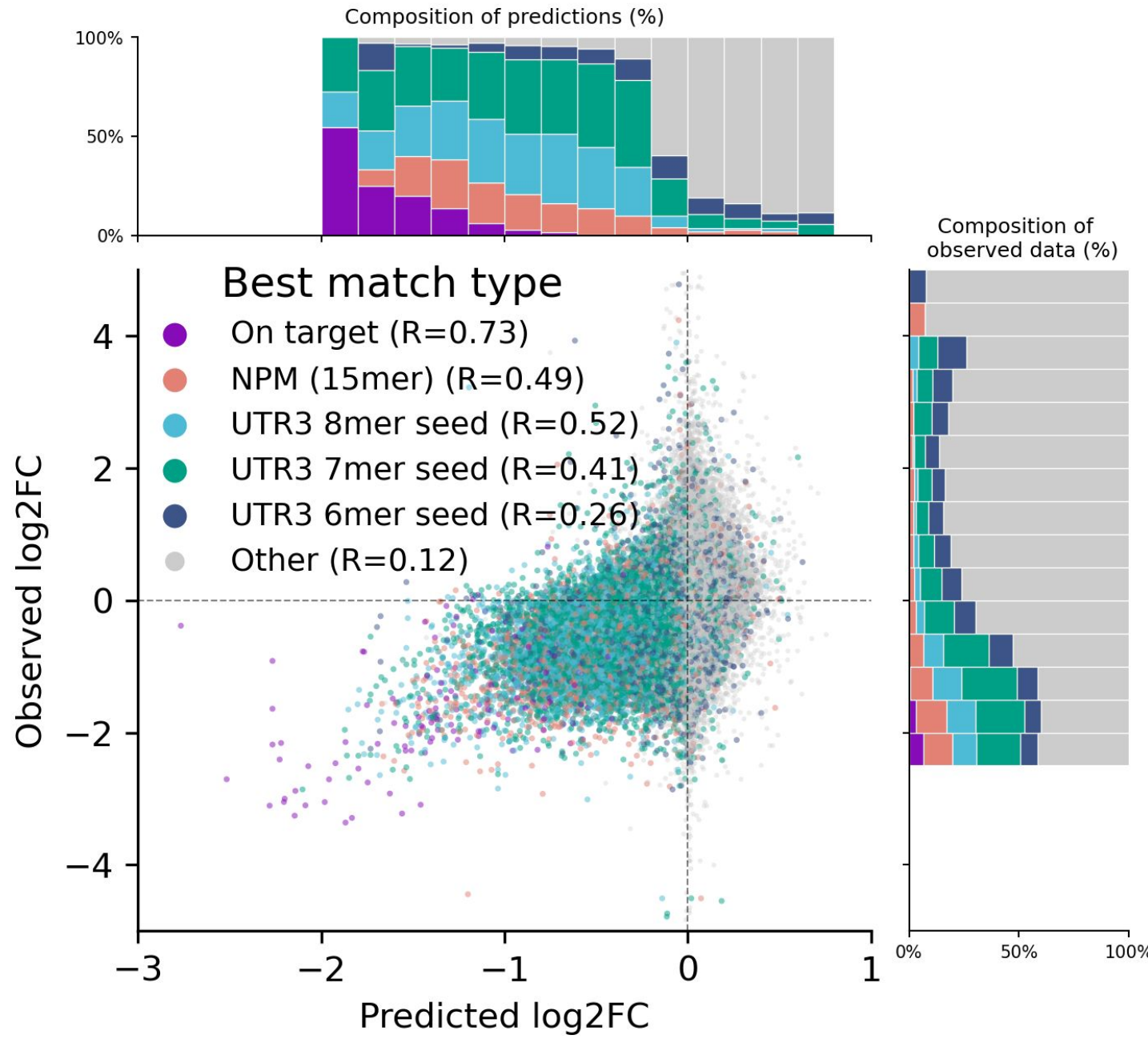
- **DIRECTOR**: DeepRNAi's solution to preprocessing the observed data to enable learning of direct, hybridization dependent effects to truly understand RISC biology:
 - Hybridization independent effects (e.g. secondary effects downstream of hybridization dependent knockdown, non-sequence specific oligo effects) dominate differential expression data (at least 2/3rd of effects).
 - **DIRECTOR** is a sequence blind, unsupervised approach to isolate direct effects for DeepRNAi training (*not* testing/evaluation).



- **Definitions**: siRNA-gene complementarity for evaluations:
 - **NPM**: Near-Perfect Match/complementarity between positions 2 and 16 (1-indexed, from 5' end, inclusive) of the siRNA guide strand and a mature mRNA, allowing an edit distance of up to 2.
 - **Seed match**: an miRNA-like 7mer-A1, 7mer-m8, or 8mer seed match between the guide strand seed region and the 3' UTR of a mature mRNA.

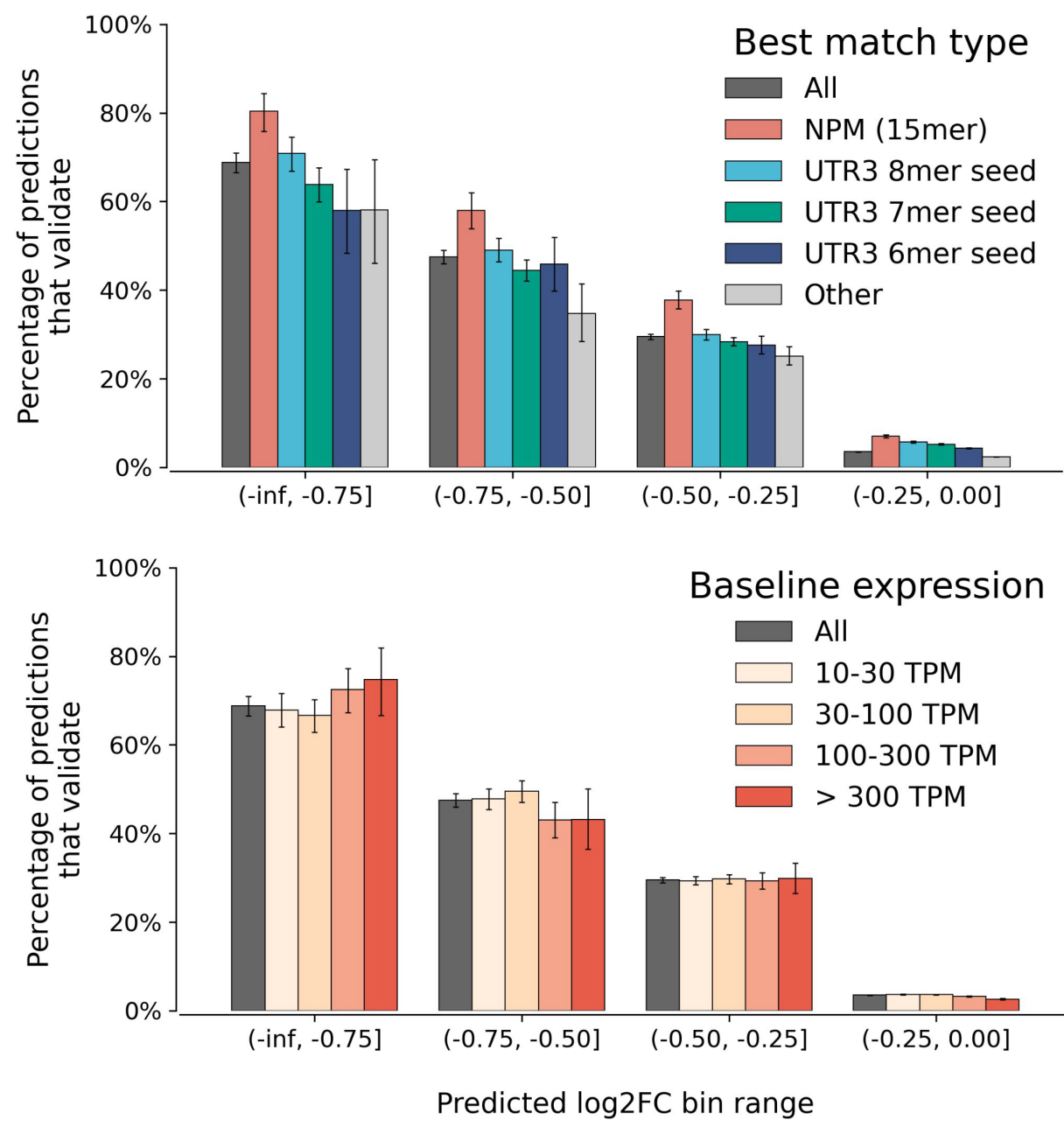
3. Results

A. DeepRNAi accurately predicts the efficacy and (per off-target gene) specificity of siRNAs against unseen targets



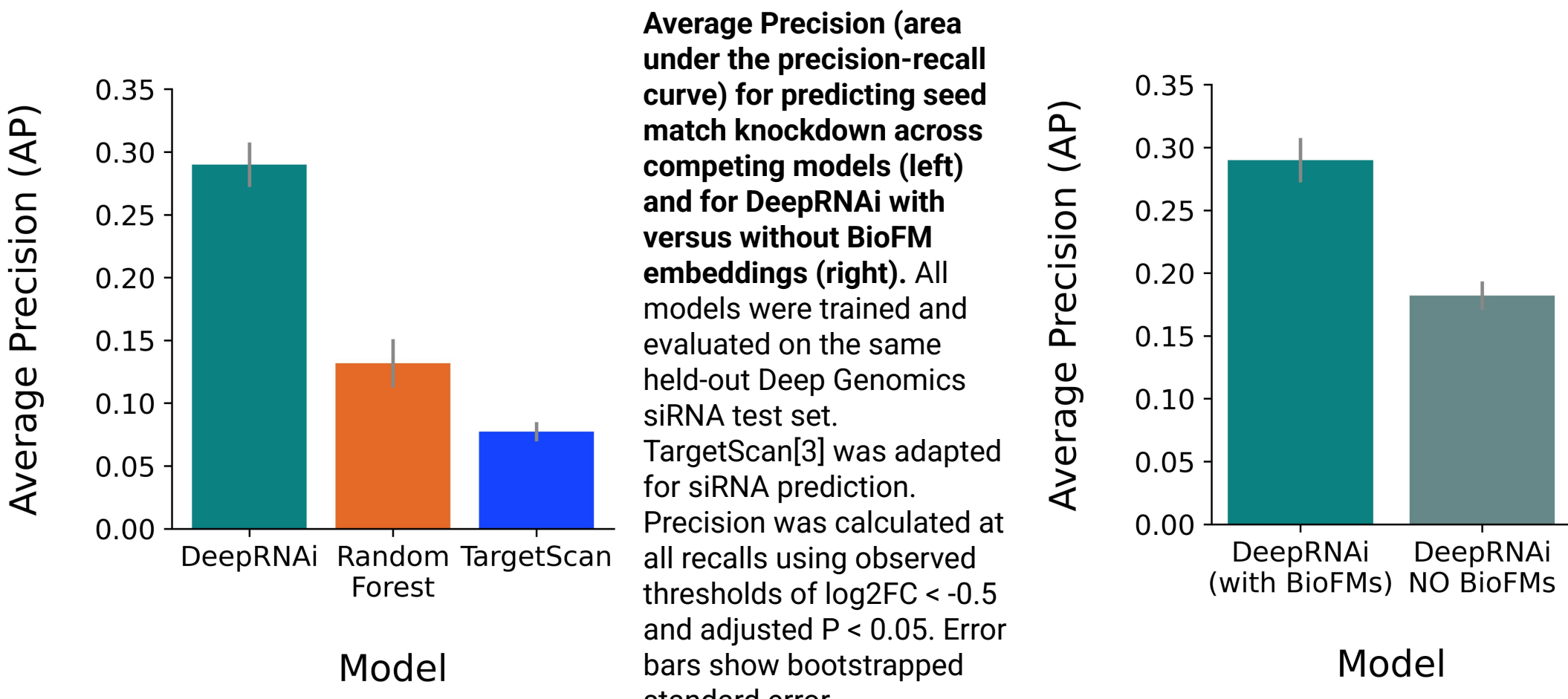
Predicted versus observed differential expression for 1.7 million siRNA-gene pairs spanning 240 unique siRNAs. The "best match type" for each siRNA-gene pair was based on the highest complementarity match between the siRNA guide strand and the gene's most highly expressed annotated transcript. Pairs labeled "7mer seed" includes both 7mer-m8 and 7mer-A1 site types. Pairs labeled "other" have no NPM or 3' UTR seed match. Pearson correlation coefficients ("R") are shown for each match type. log2FC = log2 Fold-Change; NPM = Near-Perfect Match.

B. DeepRNAi is accurate independent of the degree of siRNA-transcript complementarity, and transcript expression level

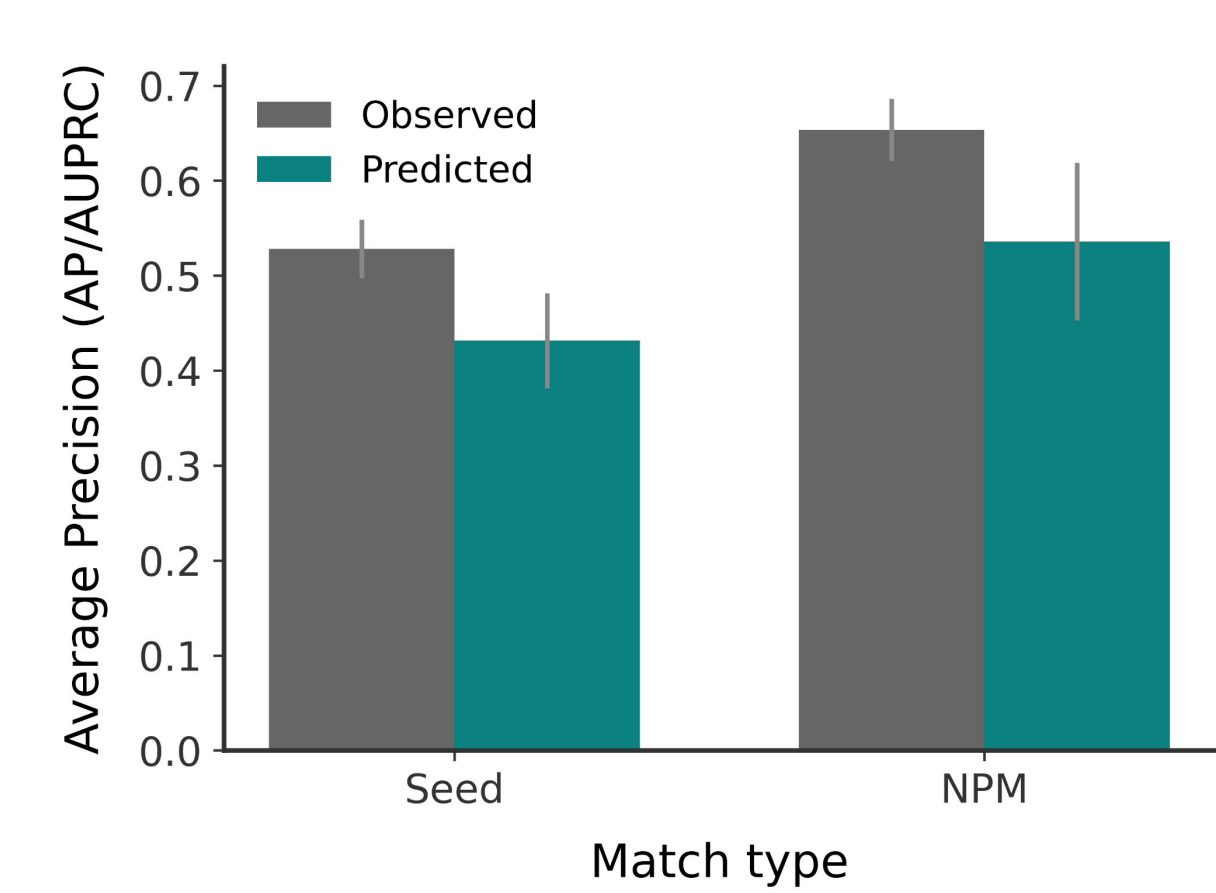


Precision of DeepRNAi predictions, binned by predicted magnitude of knockdown, stratified by degree of complementarity (top) or off-target transcript expression level (bottom). A prediction for an siRNA-gene pair log2FC is considered validated if the pair has an observed log2FC < -0.5 with adjusted P < 0.05. The "best match type" for each siRNA-gene pair was based on the highest complementarity match between the siRNA guide strand and the gene's most highly expressed annotated transcript. Pairs labeled "7mer seed" includes both 7mer-m8 and 7mer-A1 site types. Pairs labeled "other" have no NPM or 3' UTR seed match. Error bars show 95% Clopper-Pearson binomial confidence intervals. log2FC = log2 Fold-Change; NPM = Near-Perfect Match.

C. DeepRNAi consistently outperforms other benchmarks on predicting off-target liabilities; BioFMs are a differentiator

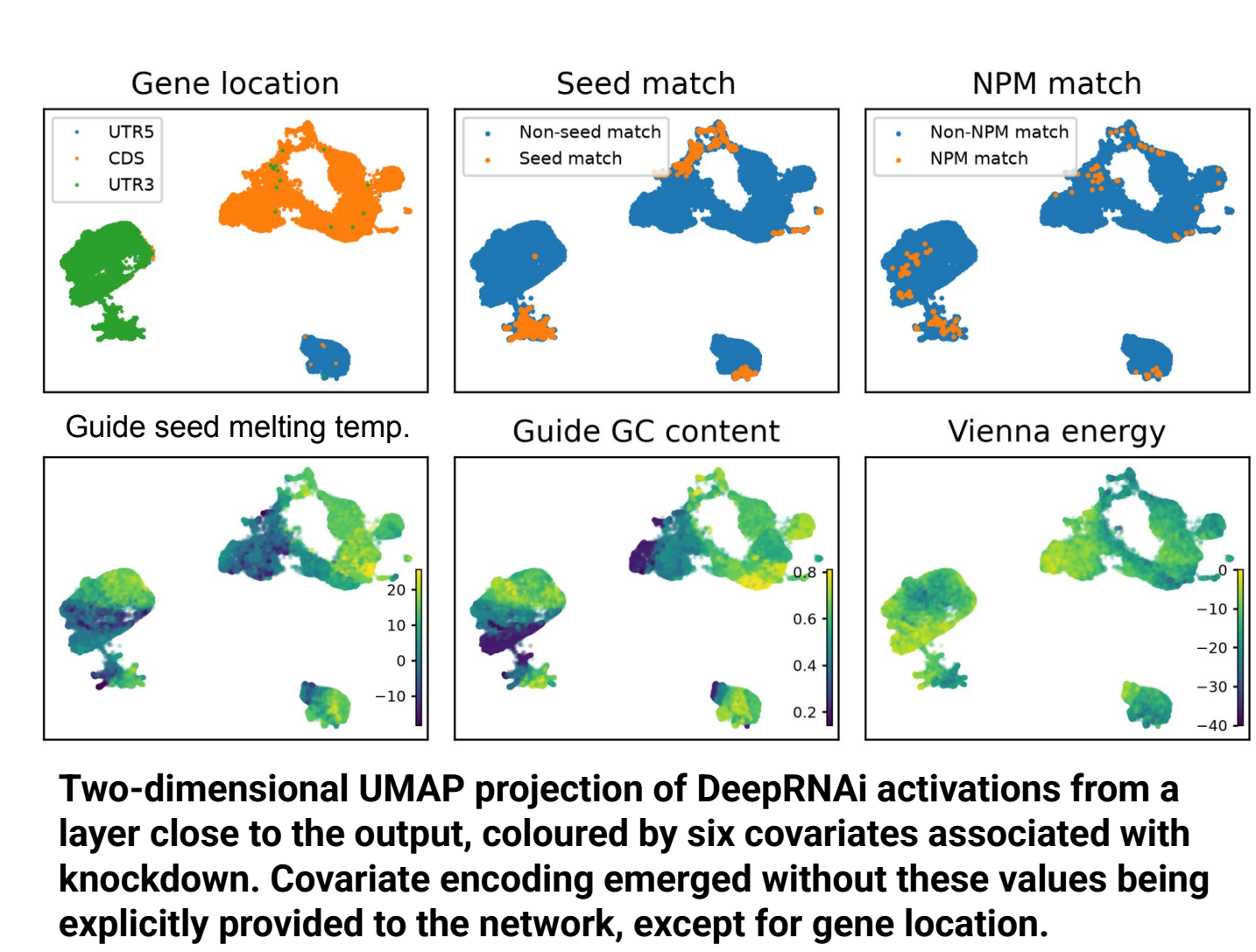


D. DeepRNAi performance approaches the limits of experimental reproducibility



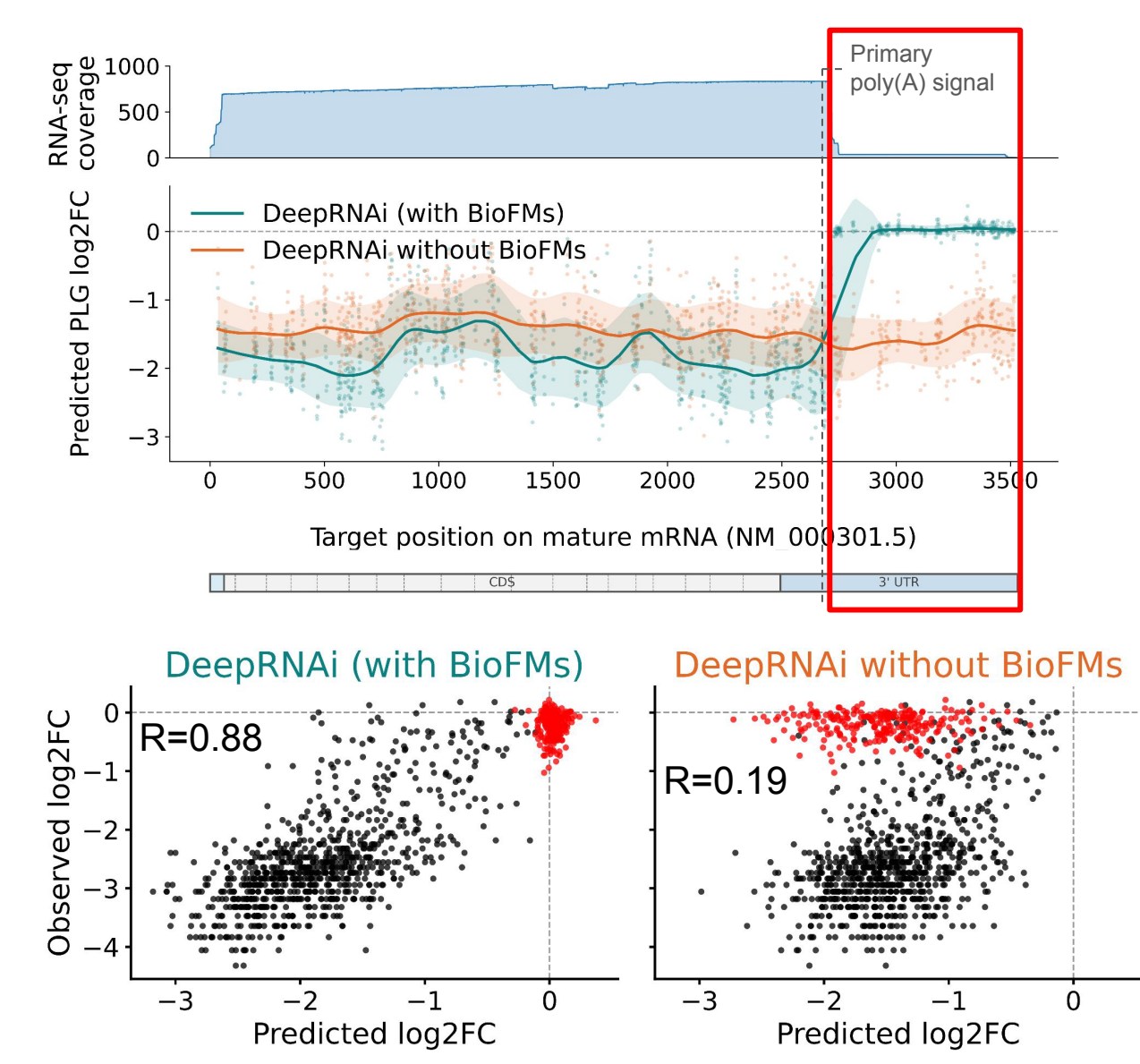
Average precision (AP/AUPRC) of DeepRNAi predictions, benchmarked against using observed data from one cell line to "predict" observed data in another cell line (HepG2 to HuH7 and HuH7 to HepG2). All observations and predictions are from 10 siRNAs in the Deep Genomics hold-out test set that were screened in both HepG2 and HuH7 cells. Precision was calculated based on observed thresholds of log2FC < -0.5 and adjusted P < 0.05. Error bars show bootstrapped standard error.

E. DeepRNAi has learned previously studied biophysical features of siRNAs and off-target transcripts, and accounts for them when making predictions



DeepRNAi has learned to account for many previously studied attributes of the siRNA guide strand (e.g. seed melting temperature, overall GC content) and mRNA target (e.g. secondary structure). For example, guide strand GC content accounted for 12% of variance in observed on-target log2FC (P=5e-5), but this was reduced to 0.49% (P=0.46) after subtracting DeepRNAi predictions using the partial correlation method. For off-targets, eSHAPE accessibility score accounted for 1.2% of variance in observed log2FC, which was reduced to 0.16% after subtracting DeepRNAi predictions using partial correlation. Together, these results indicate that DeepRNAi has learned what these features have to offer.

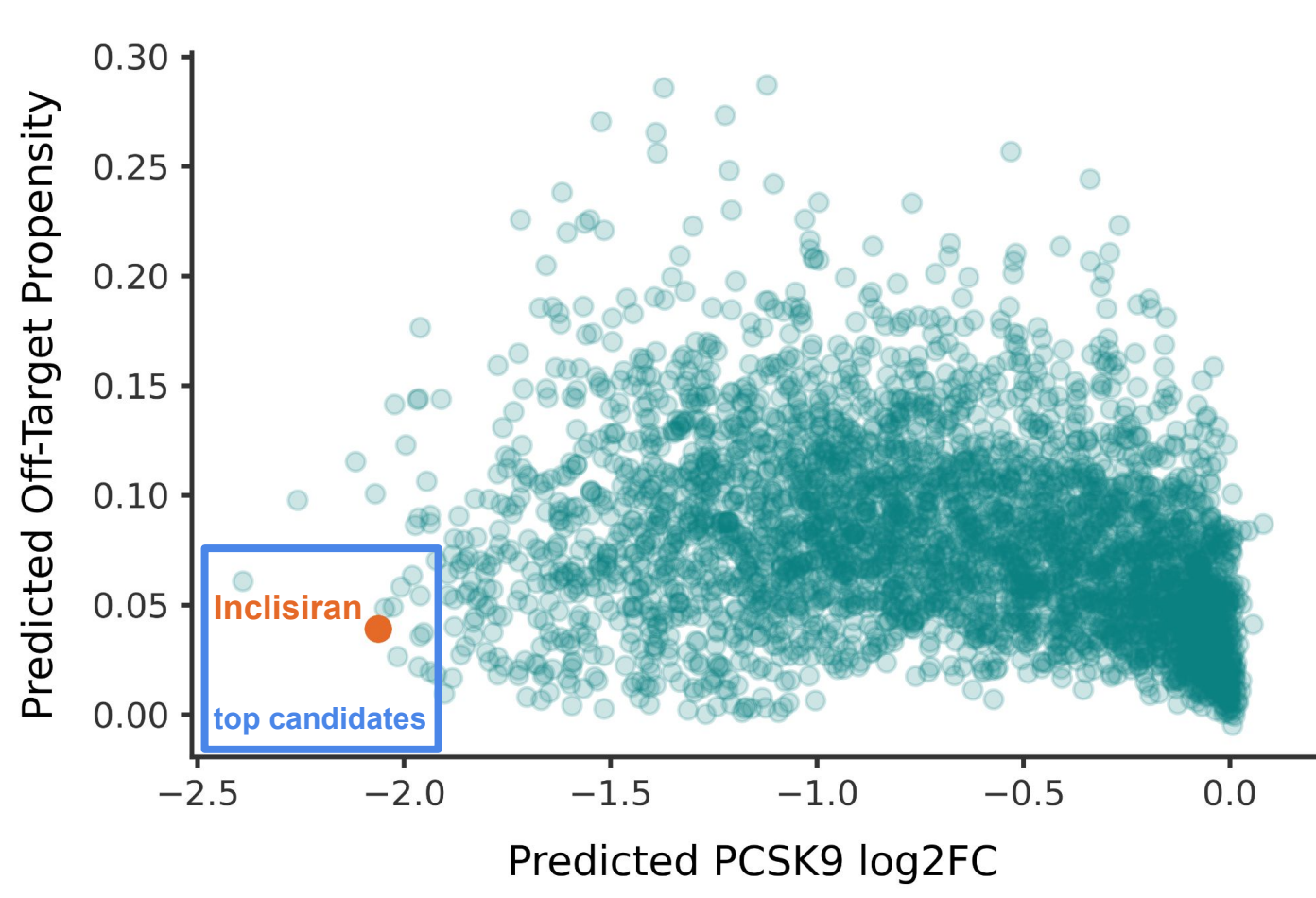
F. Deep Genomics' Biology Foundation Models (BioFMs) automatically account for (off)-target context



BioFMs enabled DeepRNAi to accurately predict that targeting the distal 3' UTR of human PLG is ineffective. siRNAs targeting this region were confirmed to be inactive in publicly available patent data [4] due to usage of an upstream poly(A) signal, confirmed by RNA-seq.

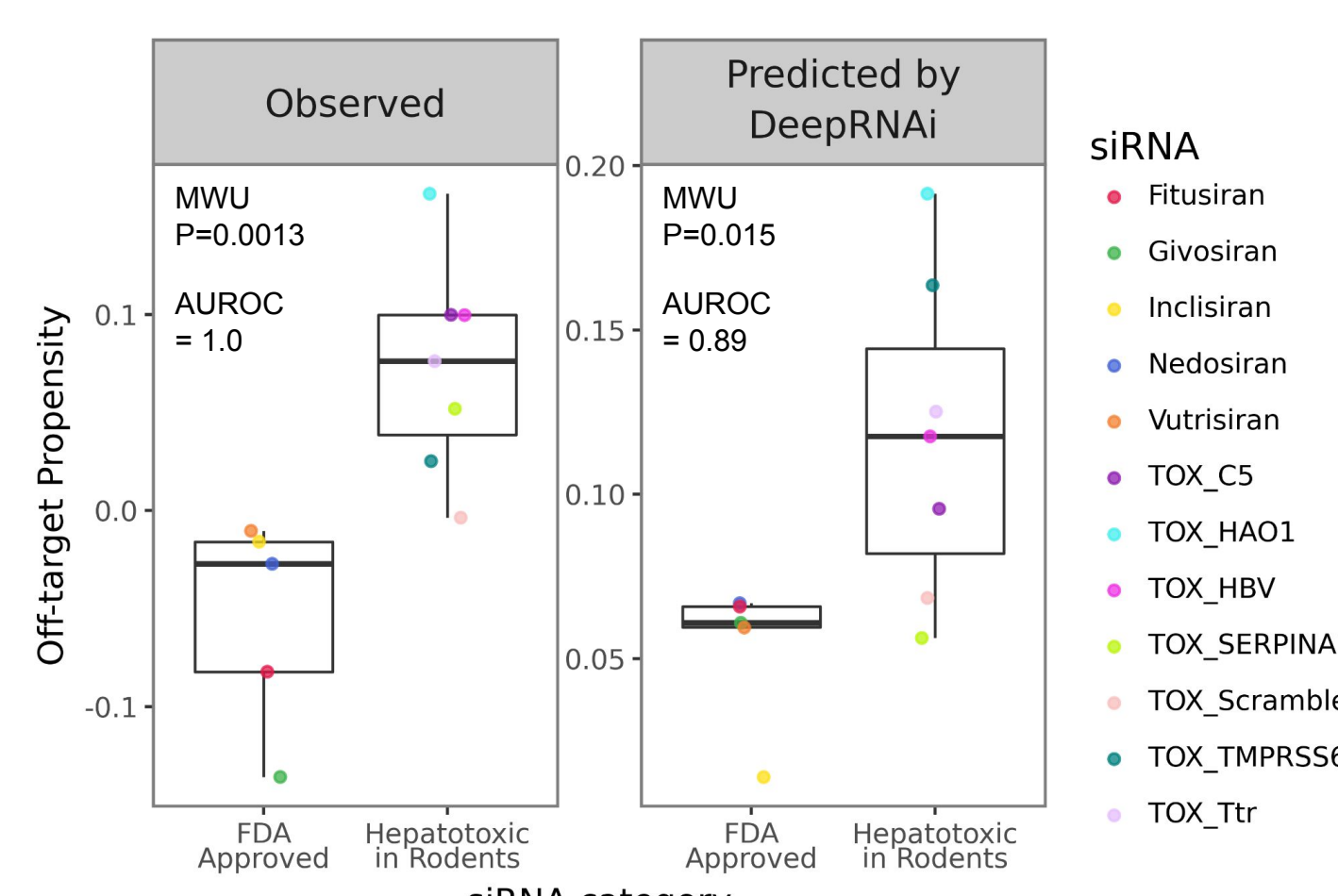
Top: Predictions are shown for siRNAs from Alnylam patent WO2024077148 [4] only, using DeepRNAi, or a version of DeepRNAi without BioFM embeddings. Bottom: Points highlighted in red target the region downstream of the primary poly(A) site. R = Pearson correlation coefficient.

G. In prospective design against unseen targets, DeepRNAi nominates clinically approved siRNAs as specific and efficacious



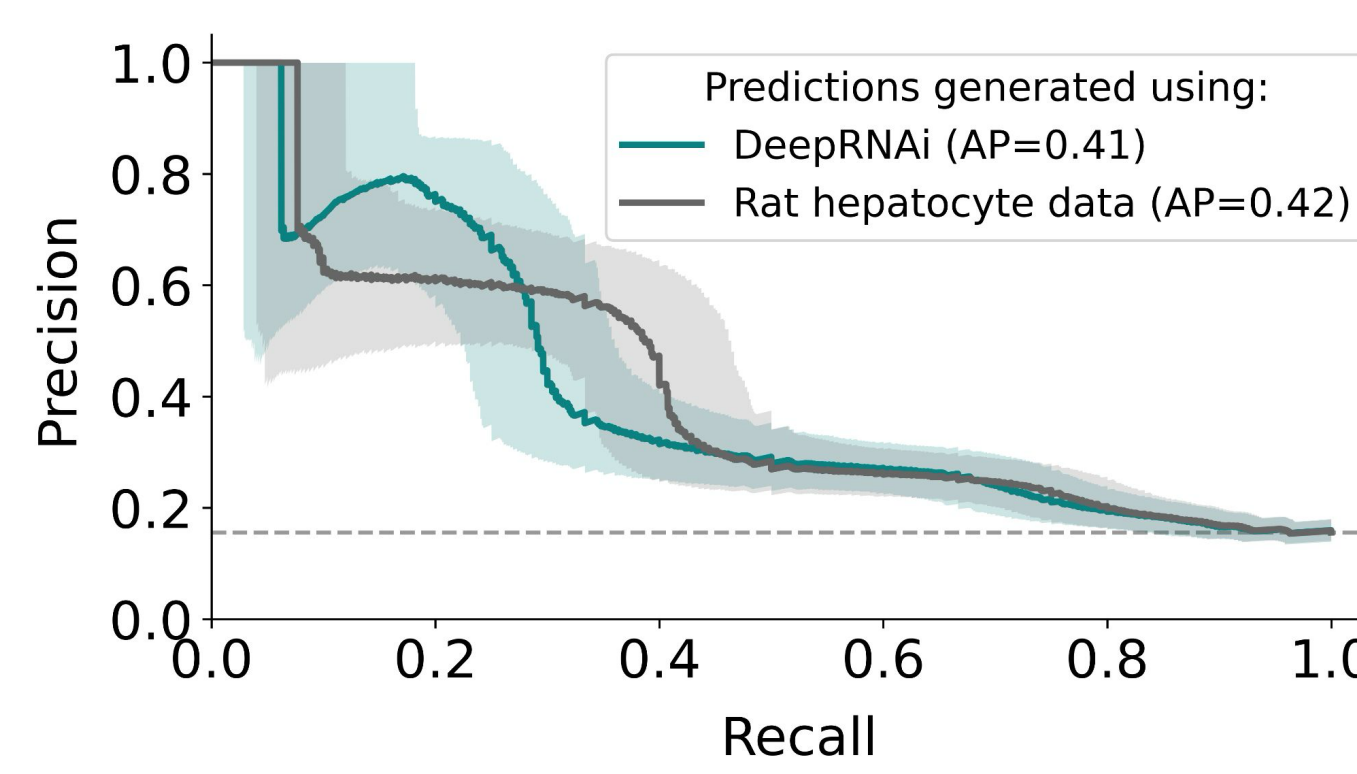
DeepRNAi nominates Incisiran, without having been trained on any siRNAs targeting PCSK9. In silico tiling of all 23-mer guide siRNAs targeting human PCSK9. Predicted efficacy and off-target propensity is shown for each siRNA (green dots; Incisiran sequence highlighted in orange). Off-target propensity = mean log2FC_{background} - mean log2FC_{seed/NPM}. Considering only pairs with log2FC < 0.

H. Despite never being trained on in vivo or clinical data, DeepRNAi can distinguish between FDA approved siRNAs and those toxic in rodents



Off-target propensity = mean log2FC_{background} - mean log2FC_{seed/NPM}. Considering only pairs with log2FC < 0. Note that not all eight FDA approved siRNAs were screened based on oligo availability at time of study. "Hepatotoxic in Rodents" siRNA sequences were taken from [5]. Observed values were from screening in HuH7 cells at 10 nM for 48 hours. MWU = Mann-Whitney U. AUROC = Area Under the Receiver Operating Characteristic curve.

I. Despite being trained on exclusively human data, predictions from DeepRNAi generalize across species and can help de-risk in vivo toxicology studies



DeepRNAi predicts in vivo off-target NPM knockdown, achieving similar performance to using observed data from in vitro hepatocyte screening. NPM knockdown was predicted using DeepRNAi or hepatocyte screening data (50 nM) and evaluated on in vivo rat liver data (30 mg/kg). All data was from [6], consisting of two siRNAs with a total of 173 NPMs. Positive hits were defined as log2FC < -0.5 and adjusted P < 0.05. Shading indicates 25th-75th percentile range from 10,000 bootstrap samples. AP = average precision.

4. Conclusions

- **DeepRNAi predictions generalize well across cell context, unseen targets, unseen species, and unseen chemistries, suggesting it has learned causal RISC biology and can serve as a reliable model to accelerate therapeutic siRNA development.**
- Crucial components that differentiate DeepRNAi include:
 - **i) Fit-for-purpose training data** enriched for high complementarity putative off-targets.
 - **ii) DIRECTOR** to enable learning of hybridization dependent effects and unlock predictions that generalize.
 - **iii) Incorporation of Deep Genomics' Biology Foundation Model (BioFM) embeddings** during training and inference to leverage a rich understanding of the molecular contexts of (off)-targets that impact siRNA activity.

5. References

- [1] Kanuparthi et al., 2025. *Sequence based prediction of cell type specific microRNA binding and mRNA degradation for therapeutic discovery.* bioRxiv.
- [2] Jung et al., 2025. *Enigma: An Efficient Model for Deciphering Regulatory Genomics.* bioRxiv.
- [3] Agarwal et al., 2015. *Predicting effective microRNA target sites in mammalian mRNAs.* eLife.
- [4] Patent WO2024077148, 2024. *Plasminogen (PLG) iRNA Compositions and Methods of Use Thereof.* Alnylam Pharmaceuticals, Inc.
- [5] Janas et al., 2018. *Selection of GalNAc-conjugated siRNAs with limited off-target-driven rat hepatotoxicity.* Nature Communications.
- [6] Schlegel et al., 2022. *From Bench to Bedside and Back: Improving the Safety of GalNAc-siRNA Conjugates Using GNA-Destabilized Seed-Pairing.* Nucleic Acids Research. (GSE184929).

Deep Genomics

Thank you for stopping by!

Please also check out our **DeepADAR poster (P-103)**

Interested to learn more?

Scan below!

