

SUPPORTING INFORMATION

Meta-learning on property matrices and LLM embeddings enables state-of-the-art prediction of gene knockdown by modified siRNAs

Ivan Golovkin,^{1†} Denis Shatkovskii,^{1†} Nikita Serov^{1*}

¹ International Institute “Solution Chemistry of Advanced Materials and Technologies”,
ITMO University, 9 Lomonosov street, Saint-Petersburg

[†] Authors contributed equally to this work.

* e-mail: serov@scamt-itmo.ru

SUPPLEMENTARY TABLES

Table S1. Correlations of numerical variables with the target variable.

	siRNA concentration	Duration after transfection	Efficacy, %
siRNA concentration	1.0	0.029	-0.014
Duration after transfection	0.029	1.0	-0.109
Efficacy, %	-0.014	-0.109	1.0

Table S2. Correlations of categorical variables with the target variable.

	Efficacy, %
Experiment of activity measurement	0.062
Cell/Organism type	0.126
Transfection method	0.116

Table S3. Binary classification performance on the train and test set.

	Precision	Recall	F1-score	Accuracy	Matthews correlation coefficient (MCC)
Train	0.97	0.99	0.98	0.97	0.92
Test	0.92	0.93	0.92	0.90	0.77

Table S4. Multiclass classification performance on the test set. 1, 2, 3, and 4 stand for the ranges 0-23%, 27-47%, 52%-73%, and 77-100% knockdown efficacy, respectively.

	Class	Precision	Recall	F1-score
Train	1	<u>0.97</u>	0.95	0.96
	2	<u>0.95</u>	0.89	0.92
	3	0.95	0.92	0.94
	4	0.95	<u>0.92</u>	<u>0.97</u>
Test	1	0.75	0.73	0.74
	2	0.53	0.36	0.43
	3	0.62	0.65	0.63
	4	<u>0.81</u>	<u>0.82</u>	<u>0.85</u>

Table S5. Performance evolution of regressive ML models on unnormalized data during feature selection and hyperparameter optimization.

Stage	Train		Test	
	RMSE	R ²	RMSE	R ²
All features	11.29	0.85	17.11	0.65
Top-100 features + Correlation threshold	11.04	0.85	14.54	0.75
Top-100 features + Correlation threshold + Optuna optimization	<u>8.34</u>	<u>0.92</u>	<u>13.81</u>	<u>0.78</u>

Table S6. Performance comparison of LLM embeddings for binary classification of target genes.

	Precision	Recall	F1-score	Accuracy	Matthews correlation coefficient (MCC)
Mistral	<u>0.91</u>	0.94	0.92	0.90	0.77
MistralDNA	<u>0.91</u>	0.94	<u>0.93</u>	0.90	0.77
HyenaDNA	<u>0.91</u>	<u>0.95</u>	<u>0.93</u>	<u>0.91</u>	<u>0.79</u>

Table S7. Performance comparison of LLM embeddings for multiclass classification of target genes.

	Class	Precision	Recall	F1-score
MistralDNA	1	<u>0.77</u>	<u>0.74</u>	<u>0.76</u>
	2	0.51	0.39	0.44
	3	0.63	<u>0.65</u>	0.64
	4	<u>0.82</u>	0.88	<u>0.85</u>
<u>Mistral</u>	1	0.75	0.73	0.74
	2	0.53	<u>0.36</u>	<u>0.43</u>
	3	0.62	0.65	<u>0.63</u>
	4	0.81	<u>0.89</u>	<u>0.85</u>
HyenaDNA	1	0.71	0.73	0.72
	2	<u>0.56</u>	0.36	0.44
	3	<u>0.71</u>	0.60	0.65
	4	0.78	0.93	<u>0.85</u>

Table S8. Performance comparison of LLM embeddings for regression of target genes.

Type	Mistral			MistralDNA			HyenaDNA		
Metrics	RMSE train	RMSE test	R ² test	RMSE train	RMSE test	R ² test	RMSE train	RMSE test	R ² test
Value	4.42	12.27	0.84	6.67	12.17	0.84	5.21	12.78	0.81

Table S9. Performance comparison of meta-learning approach and regular regression with gene embeddings.

Regression model				Meta-learning approach			
RMSE test	R ² test	RMSE train	R ² train	RMSE test	R ² test	RMSE train	R ² train
12.71	0.81	7.83	0.93	<u>12.27</u>	<u>0.84</u>	<u>7.27</u>	<u>0.94</u>

SUPPLEMENTARY FIGURES

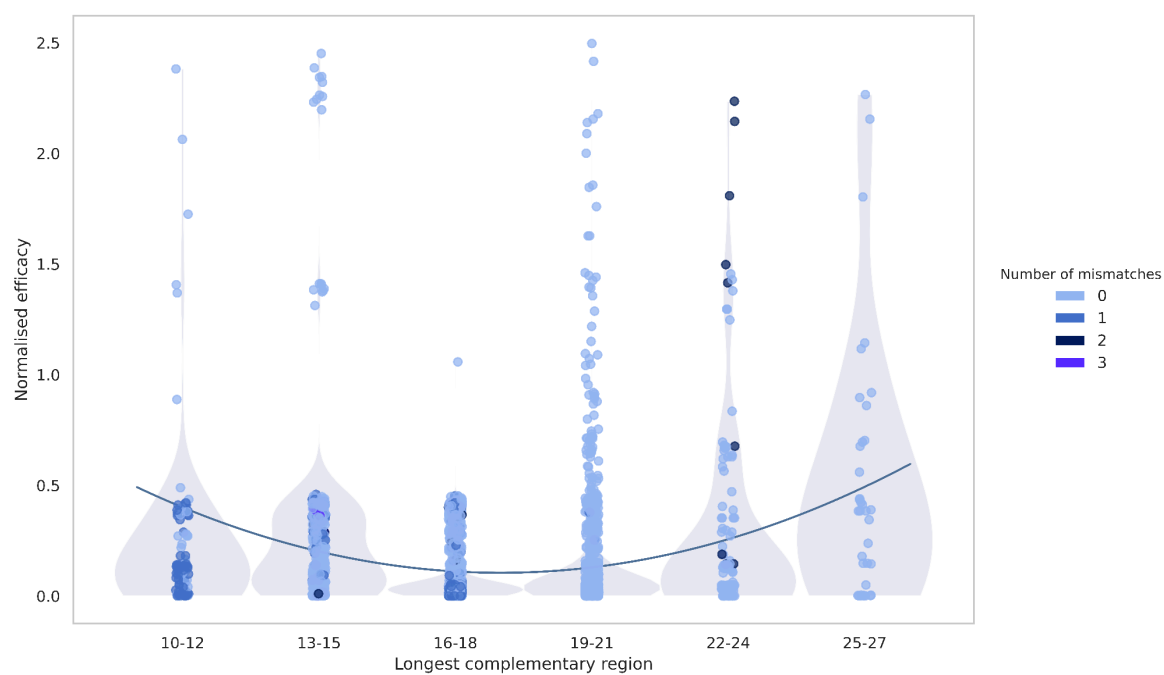


Figure S1. Distribution of normalised efficacy of sequences depending on the presence of mismatches with trendline inset (probability density plots shaded).