Figure 1

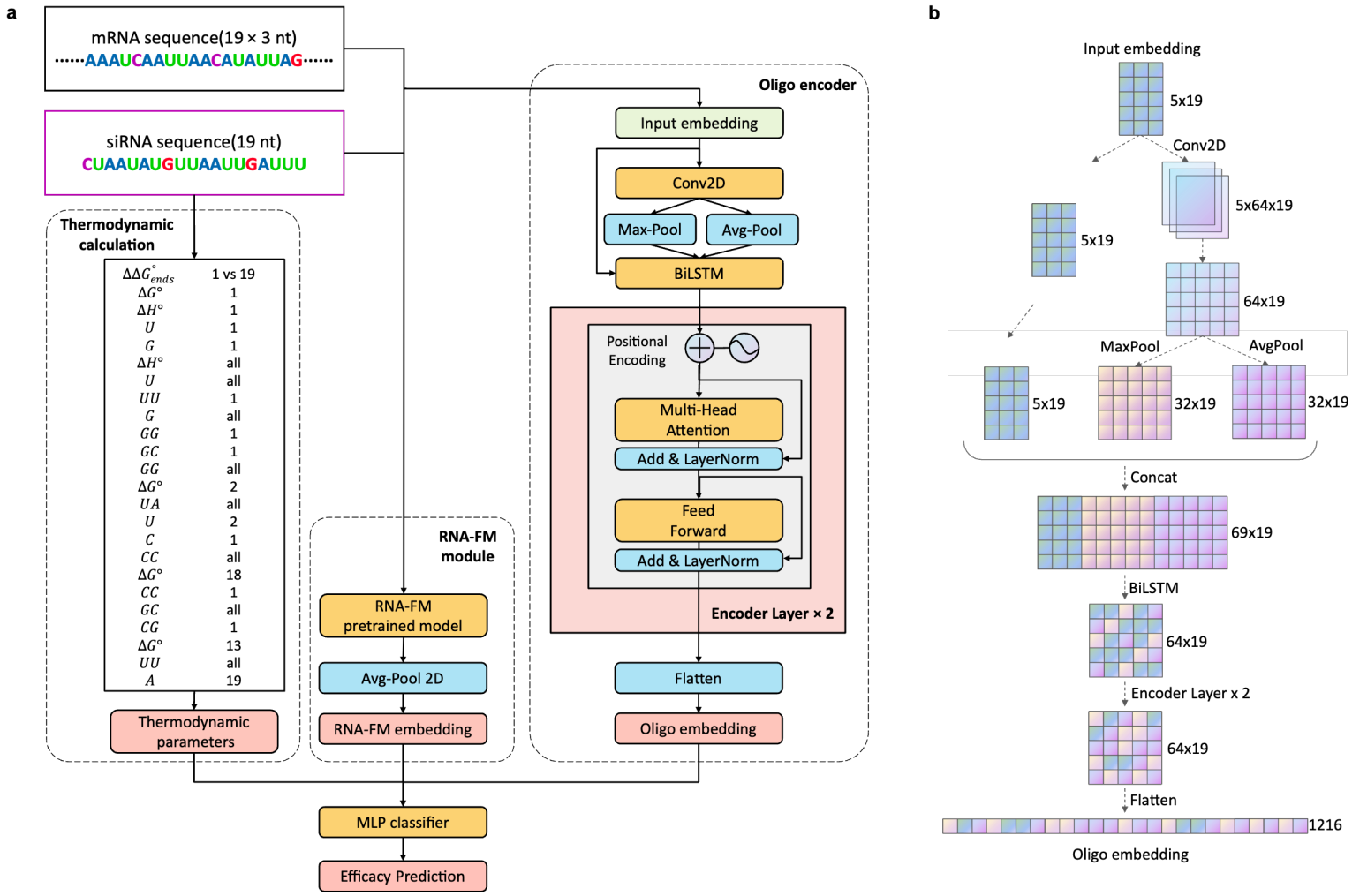
a
mRNA sequence(19 × 3 nt)
·····AAAUCAAUUAACAUAUUAG·····
siRNA sequence(19 nt)
CUAAUAUGUUAAUUGAUUU
Thermodynamic calculation
ΔΔG°ends 1 vs 19
ΔG° 1
ΔH° 1
U 1
G 1
ΔH° all
U all
UU 1
G all
GG 1
GC 1
GG all
ΔG° 2
UA all
U 2
C 1
CC all
ΔG° 18
CC 1
GC all
CG 1
ΔG° 13
UU all
A 19
Thermodynamic parameters
RNA-FM module
RNA-FM pretrained model
Avg-Pool 2D
RNA-FM embedding
Oligo encoder
Input embedding
Conv2D
Max-Pool
Avg-Pool
BiLSTM
Positional Encoding
Multi-Head Attention
Add & LayerNorm
Feed Forward
Add & LayerNorm
Encoder Layer × 2
Flatten
Oligo embedding
MLP classifier
Efficacy Prediction
b
Input embedding
5x19
Conv2D
5x64x19
5x19
64x19
MaxPool
AvgPool
5x19
32x19
32x19
Concat
69x19
BiLSTM
64x19
Encoder Layer x 2
64x19
Flatten
1216
Oligo embedding

Figure 2

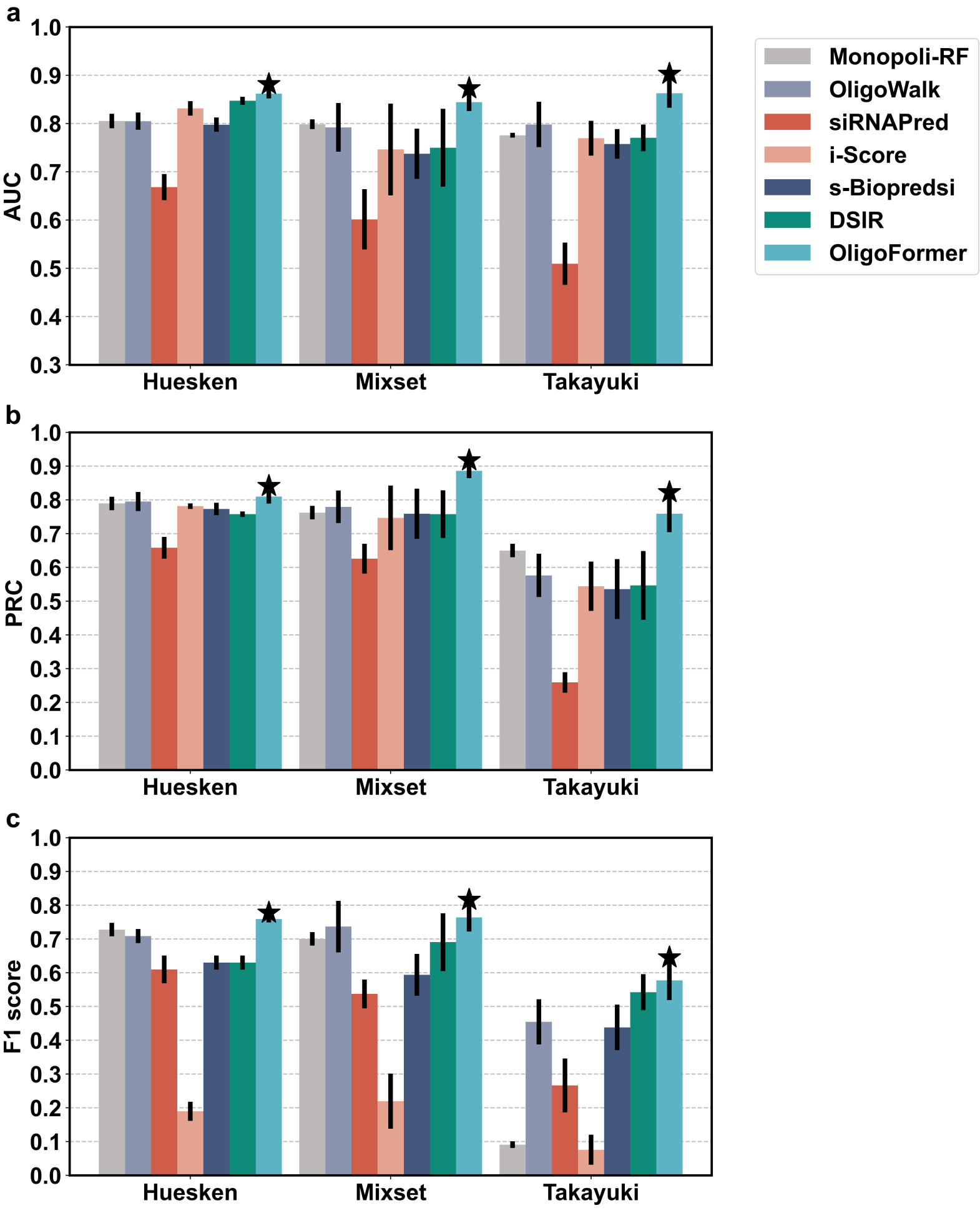
a
AUC
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
Huesken
Mixset
Takayuki
b
PRC
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Huesken
Mixset
Takayuki
c
F1 score
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Huesken
Mixset
Takayuki
Monopoli-RF
OligoWalk
siRNAPred
i-Score
s-Biopredsi
DSIR
OligoFormer

Figure 3

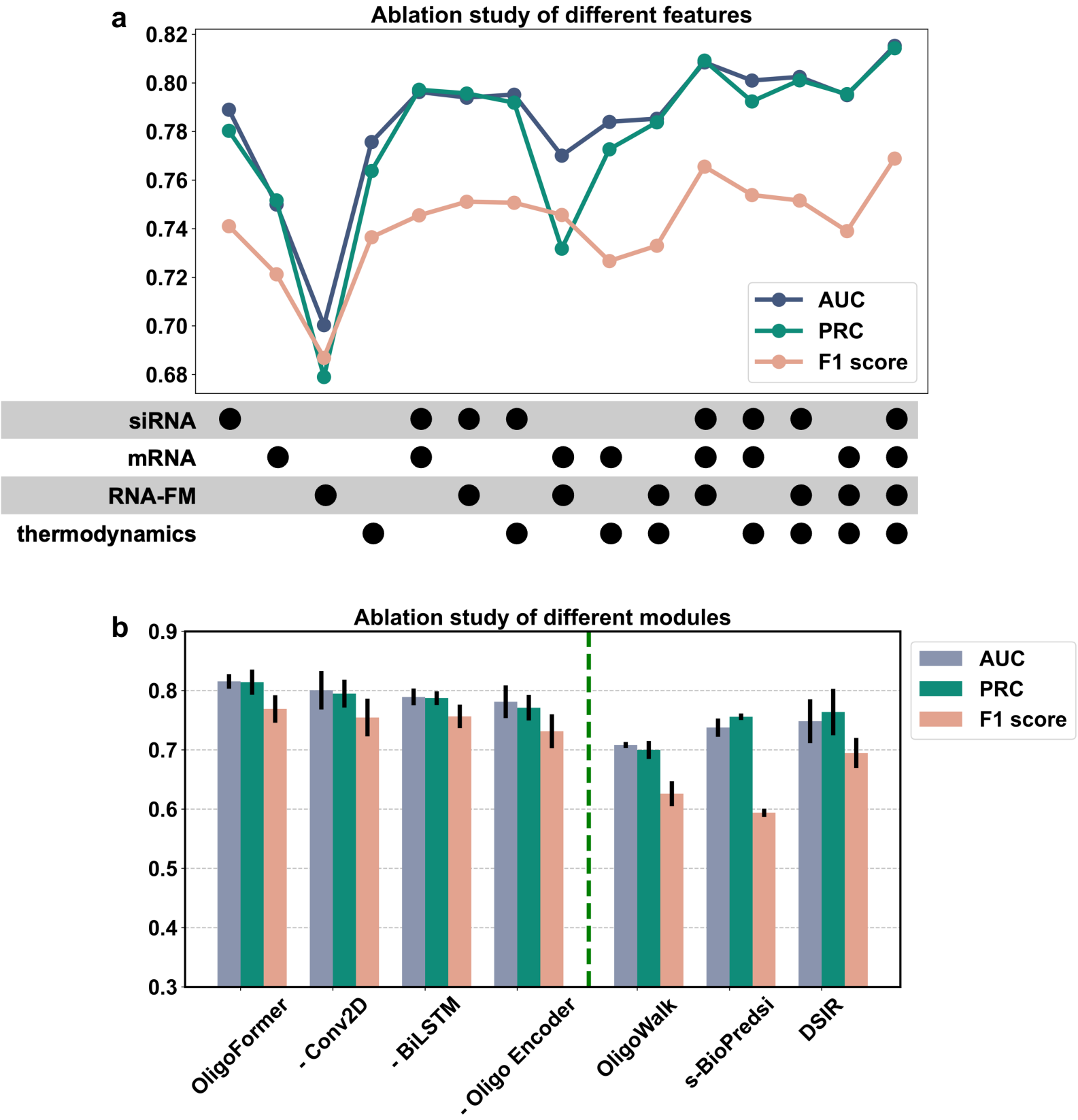
a
Ablation study of different features
0.82
0.80
0.78
0.76
0.74
0.72
0.70
0.68
AUC
PRC
F1 score
siRNA
mRNA
RNA-FM
thermodynamics
b
Ablation study of different modules
0.9
0.8
0.7
0.6
0.5
0.4
0.3
AUC
PRC
F1 score
OligoFormer
- Conv2D
- BiLSTM
- Oligo Encoder
OligoWalk
s-BioPredsi
DSIR

Figure 4

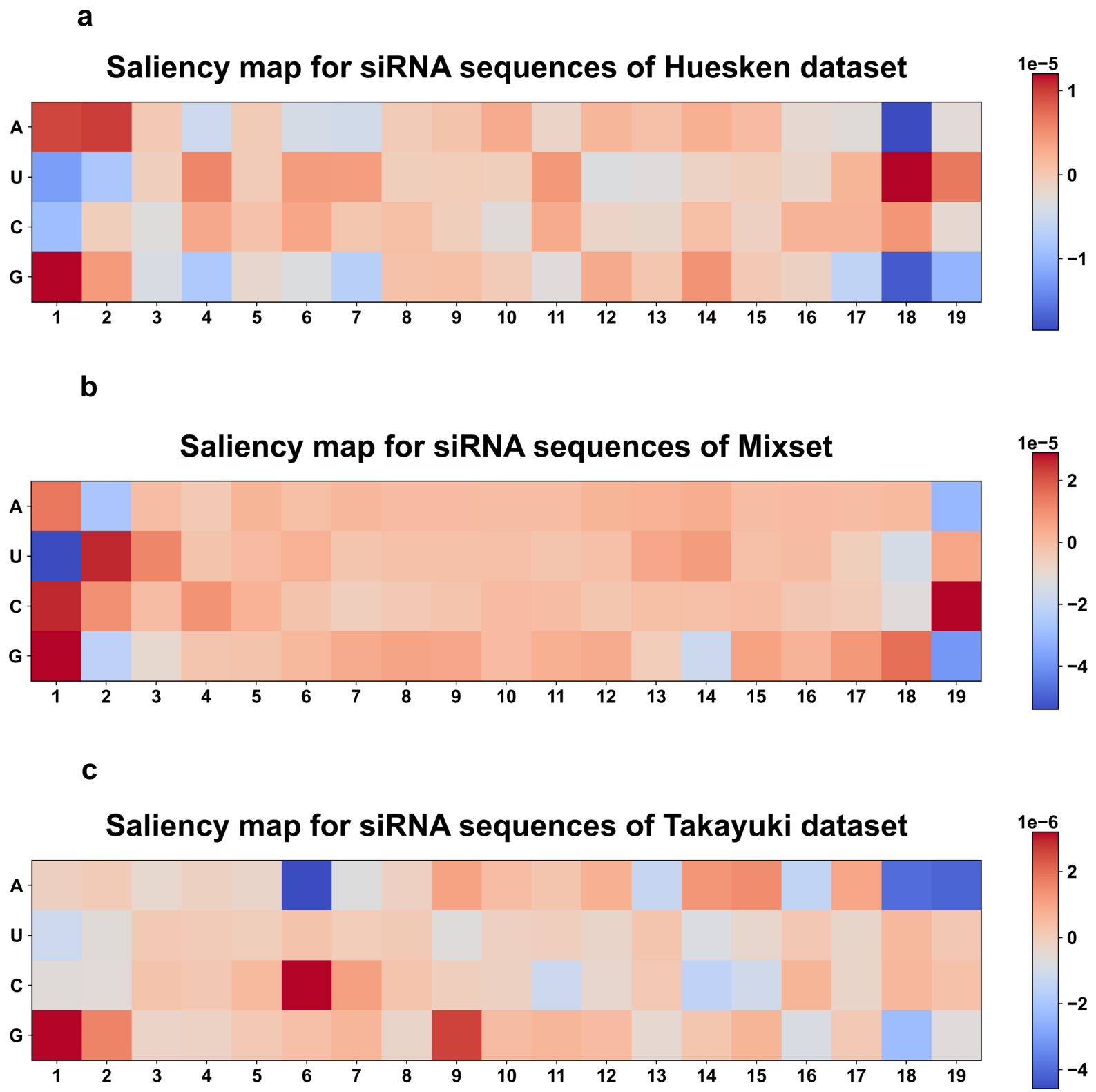
a
Saliency map for siRNA sequences of Huesken dataset
1e−5
1
0
−1
A
U
C
G
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
b
Saliency map for siRNA sequences of Mixset
1e−5
2
0
−2
−4
A
U
C
G
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
c
Saliency map for siRNA sequences of Takayuki dataset
1e−6
2
0
−2
−4
A
U
C
G
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

Figure 5

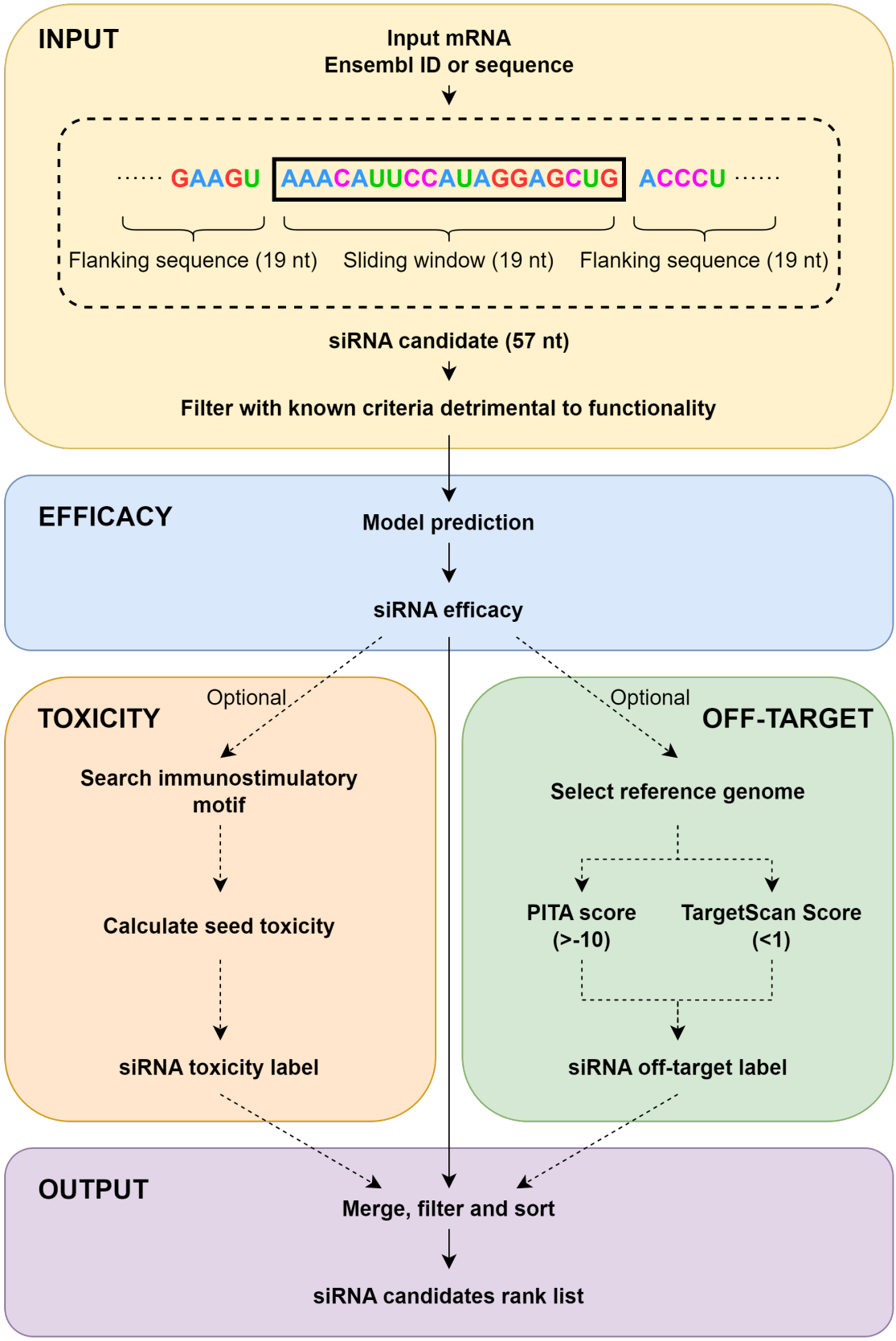
INPUT
Input mRNA
Ensembl ID or sequence
..... GAAGU AAACAUUCCAUAGGAGCUG ACCCU .....
Flanking sequence (19 nt)
Sliding window (19 nt)
Flanking sequence (19 nt)
siRNA candidate (57 nt)
Filter with known criteria detrimental to functionality
EFFICACY
Model prediction
siRNA efficacy
TOXICITY
Optional
Search immunostimulatory motif
Calculate seed toxicity
siRNA toxicity label
OFF-TARGET
Optional
Select reference genome
PITA score (>-10)
TargetScan Score (<1)
siRNA off-target label
OUTPUT
Merge, filter and sort
siRNA candidates rank list

Table 1. The datasets of siRNAs

Sources	siRNA[1]	mRNA[1]	Cell line[2]	Datasets
Huesken	2431	34	H1299	Huesken
Takayuki	702	1	HeLa	Takayuki
Amarzguioui	46	4	HaCaT	
Haborth	44	1	HeLa	
Hsieh	108	22	HEK293T	
Khvorova	14	1	HEK293	Mixset
Reynolds	240	7	HEK293	
Vickers	76	2	T24	
Ui-Tei	53	3	HeLa	

[1]This represents the number of siRNAs or mRNAs;[2]only the most representative cell line was selected

Table 2. Performance comparison among different siRNA design data methods for intra-dataset and inter-dataset validation

Methods	Huesken dataset[1]			Mixset			Takayuki dataset			Inter-dataset		
	AUC	PRC	F1 score	AUC	PRC	F1 score	AUC	PRC	F1 score	AUC	PRC	F1 score
Monopoli-RF[2]	0.805	0.7893	**0.7276**	0.7984	0.7621	0.7001	0.7756	**0.6498**	0.0909	0.7439	0.7388	**0.7002**
OligoWalk	0.8034	**0.795**	0.7083	**0.7919**	**0.7793**	**0.7364**	**0.7979**	0.5764	0.4548	0.708	0.6998	0.6258
siRNAPred	0.6681	0.658	0.6097	0.6014	0.6257	0.5371	0.5094	0.2595	0.2662	0.5782	0.618	0.5245
iScore	0.8314	0.7813	0.1896	0.7462	0.7465	0.2197	0.7695	0.5443	0.07572	0.7462	0.7565	0.2197
s-Biopredsi	0.7975	0.7731	0.6299	0.7374	0.7589	0.5936	0.7576	0.5358	0.4379	0.7374	0.7557	0.5936
DSIR	**0.8472**	0.7575	0.6299	0.7498	0.7575	0.6905	0.7702	0.5469	0.5422	**0.7483**	**0.7637**	0.6945
OligoFormer	**0.8619**	**0.8099**	**0.7584**	**0.8454**	**0.8858**	**0.7641**	**0.8628**	**0.7586**	**0.5769**	**0.8163**	**0.8143**	**0.769**

[1]The first nine columns represent intra-dataset training and each value is the mean of 5-fold cross-validation;[2]using two partition method for Monopoli-RF training