Supplementary File 01: Obs/Exp ration of k-mers

LncRNA Counts		N	46672885					mRNA Counts		N	55601112				
K=1	Counts	K=2	Counts	Obs/Exp	K=3	Counts	Obs/Exp	K=1	Counts	K=2	Obs/Exp bibj	K=3	Counts	Obs/Exp
A	13069796	AA	4081993	1.115319	AAA	1434085	1.399258	A	14066297	AA	3959049	AAA	1269782	1.410447
C	10709264	AC	2463621	0.821505	AAC	684141	0.814662	C	14090007	AC	2882962	AAC	718558	0.796816
G	10699386	AG	3552272	1.185614	AAG	1038524	1.237797	G	14410331	AG	4265676	AAG	1181698	1.281268
T	12194439	AT	2949444	0.863722	AAT	912329	0.954072	T	13034477	AT	2940352	AAT	781218	0.936453
		CA	3539454	1.180246	ACA	912432	1.086507			CA	4116725	ACA	935363	1.037233
		CC	2937602	1.195469	ACC	632739	0.91953			CC	4225852	ACC	855462	0.947034
		CG	785615	0.320004	ACG	180479	0.262524			CG	1708894	ACG	330313	0.357543
		CT	3439099	1.229102	ACT	736143	0.93951			CT	4030780	ACT	760351	0.909906
		GA	3191789	1.065298	AGA	1145022	1.36473			GA	3946557	AGA	1277304	1.38493
		GC	2458974	1.001613	AGC	802910	1.167909			GC	3739766	AGC	1101564	1.192372
		GG	2816995	1.148505	AGG	920082	1.339582			GG	4073225	AGG	1130728	1.196733
		GT	2225604	0.796144	AGT	682297	0.871592			GT	2641097	AGT	752691	0.880717
		TA	2241369	0.656368	ATA	646108	0.675671			TA	2028569	ATA	525766	0.63024
		TC	2841874	1.015659	ATC	608063	0.776046			TC	3233196	ATC	686420	0.821433
		TG	3529222	1.262475	ATG	832838	1.063899			TG	4347114	ATG	958594	1.121642
		TT	3574632	1.121947	ATT	860515	0.964485			TT	3417915	ATT	767986	0.993467
					CAA	865328	1.030417					CAA	939476	1.041794
					CAC	761650	1.10687					CAC	866856	0.959647
					CAG	1123550	1.63431					CAG	1471107	1.592378
					CAT	784905	1.001743					CAT	835278	0.99957
					CCA	976163	1.418612					CCA	1242646	1.375663
					CCC	760502	1.34881					CCC	1183349	1.307815
					CCG	239308	0.424823					CCG	566706	0.61239
					CCT	959682	1.494774					CCT	1230813	1.470425
					CGA	164619	0.239454					CGA	348436	0.377159
					CGC	210335	0.37339					CGC	490726	0.530285
					CGG	236879	0.4209					CGG	569993	0.602251

CGT	173403	0.270337
CTA	510000	0.650892
CTC	875561	1.36375
CTG	1159822	1.808174
CTT	891766	1.219823
GAA	1055750	1.258328
GAC	570363	0.829647
GAG	935841	1.362526
GAT	627540	0.801643
GCA	762867	1.109662
GCC	723266	1.283953
GCG	206252	0.36648
GCT	765070	1.192752
GGA	923704	1.344855
GGC	671573	1.193288
GGG	700035	1.245009
GGT	520238	0.811805
GTA	410651	0.524581
GTC	487754	0.760414
GTG	736129	1.148692
GTT	589908	0.807665
TAA	723989	0.757115
TAC	444394	0.567162
TAG	448004	0.572297
TAT	621746	0.696867
TCA	885803	1.130515
TCC	819259	1.276055
TCG	158821	0.247604
TCT	975791	1.334759
TGA	954718	1.219593
TGC	770853	1.201768
TGG	954148	1.4889

CGT	299049	0.349326
CTA	517190	0.618917
CTC	1021329	1.220159
CTG	1552959	1.814045
CTT	937193	1.210313
GAA	1192688	1.293184
GAC	779006	0.843224
GAG	1228652	1.300374
GAT	742445	0.868728
GCA	990706	1.072375
GCC	1179736	1.274839
GCG	499562	0.527834
GCT	1068074	1.247641
GGA	1256675	1.330032
GGC	1107721	1.170411
GGG	1041307	1.075781
GGT	665078	0.759623
GTA	414648	0.485176
GTC	612105	0.715013
GTG	995083	1.13654
GTT	617877	0.780203
TAA	554259	0.664395
TAC	515711	0.617147
TAG	378367	0.442724
TAT	577541	0.747107
TCA	946395	1.132543
TCC	1005151	1.200832
TCG	310557	0.362768
TCT	968836	1.251177
TGA	1060758	1.241183
TGC	1035550	1.209649
TGG	1326582	1.515164

		TGT	847264	1.16002		TGT	921061	1.163039
		TTA	673820	0.755233		TTA	570467	0.737956
		TTC	869330	1.189134		TTC	912236	1.178083
		TTG	798676	1.093497		TTG	839169	1.059632
		TTT	1230496	1.478169		TTT	1093439	1.526443

Supplementary File 02: Details for parameter optimization for both human and mouse models

Details for parameter optimization (Human)

For Catboot we have used parameters of random_seed (0, 1, 42,100, 200, 300, 400, 500, 600).

For Decision Tree Classifier we have used criterion of (gini, entropy), splitter of (best, random), a max_depth of (1, 2, 3, 4, 5, 6, 7, 8, 9, 10) and min_samples_leaf of (1, 5, 10, 20, 50, 100).

For XGBootClassifier we have used max_depth range (2, 10, 1), n_estimators range (60, 220, 40), and learning_rate of (0.1, 0.01, 0.05).

For Random Forest Classifier we have used max_depth of (10, 20, None) max_features of (auto, sqrt], min_samples_leaf of (1,2, 3, 4, 5), min_samples_split of (2, 5, 10, 12) and n_estimators of (50, 80, 100).

For Support Vector Machines (SVM), we have used parameters C of (1), gamma of (0.1,0.01,0.001), kernel of (linear, rbf, poly, sigmoid)

For MLPClassifier we have used hidden_layer_sizes of [(50,50,50), (50,100,50), (100,)], activation of ('tanh', 'relu'), solver (sgd, adam), alpha of (0.0001, 0.05), learning_rate of (constant, adaptive)

List of parameters for ML models (Human)

CatBoost

- nan_mode: Min
- eval_metric: Logloss
- iterations: 100
- sampling_frequency: PerTree
- leaf_estimation_method': Newton
- grow_policy: SymmetricTree
- penalties_coefficient: 1
- boosting_type: Plain
- model_shrink_mode: Constant'
- feature_border_type: 'GreedyLogSum
- bayesian_matrix_reg: 0.10000000149011612
- l2_leaf_reg: 3
- random_strength: 1

- rsm: 1
- boost_from_average: False
- model_size_reg: 0.5
- pool_metainfo_options: tags: {}}
- subsample: 0.800000011920929
- class_names: [0, 1]
- random_seed: 42
- depth: 4
- posterior_sampling: False
- border_count: 254
- classes_count: 0
- auto_class_weights: None
- sparse_features_conflict_fraction: 0
- leaf_estimation_backtracking: AnyImprovement
- best_model_min_trees: 1
- model_shrink_rate: 0
- min_data_in_leaf: 1
- loss_function: Logloss
- learning_rate: 0.03999999910593033
- score_function: Cosine
- task_type:CPU
- leaf_estimation_iterations: 10
- bootstrap_type: MVS
- max_leaves: 16

DecisionTreeClassifier

- criterion=entropy
- max_depth=8
- min_samples_leaf=20

XGBClassifier

- base_score=0.5
- booster=gbtree
- colsample_bylevel=1
- colsample_bynode=1
- colsample_bytree=1
- eval_metric=mlogloss
- gamma=0
- gpu_id=-1
- importance_type=gain
- learning_rate=0.1

- max_delta_step=0
- max_depth=8
- min_child_weight=1
- missing=nan
- monotone_constraints=()
- n_estimators=180
- n_jobs=16
- num_parallel_tree=1
- random_state=0
- reg_alpha=0
- reg_lambda=1
- scale_pos_weight=1
- subsample=1
- tree_method=exact
- use_label_encoder=False
- validate_parameters=1

Support Vector Machines (SVM)

- C=1
- gamma=0.1
- kernel=poly
- probability=True

RandomForestClassifier

- Random_state=42

MLPClassifier

- hidden_layer_sizes=(50, 100, 50)
- learning_rate=adaptive
- max_iter=1000
- solver=sgd

Details for parameter optimization (Mouse)

For Catboot we have used parameters of random_seed (0, 1, 42,100, 200, 300, 400, 500, 600).

For Decision Tree Classifier we have used criterion of (gini, entropy), splitter of (best, random), a max_depth of (1, 2, 3, 4, 5, 6, 7, 8, 9, 10) and min_samples_leaf of (1, 5, 10, 20, 50, 100).

For XGBootClassifier we have used max_depth range (2, 10, 1), n_estimators range (60, 220, 40), and learning_rate of (0.1, 0.01, 0.05).

For Random Forest Classifier we have used max_depth of (10, 20, 40, None) max_features of (auto, sqrt], min_samples_leaf of (1,2, 3, 4, 5), min_samples_split of (2, 5, 10, 12) and n_estimators of (200, 400, 600).

For Support Vector Machines (SVM), we have used parameters C of (0.1,1, 10, 100), gamma of (1,0.1,0.01,0.001), kernel of (rbf, poly, sigmoid)

For MLPClassifier we have used hidden_layer_sizes of [(50,50,50), (50,100,50), (100,)], activation of ('tanh', 'relu'), solver (sgd, adam), alpha of (0.0001, 0.05), learning_rate of (constant, adaptive)

List of parameters for ML models (Mouse)

CatBoost

- nan_mode: Min
- eval_metric: Logloss
- iterations: 100
- sampling_frequency: PerTree
- leaf_estimation_method': Newton
- grow_policy: SymmetricTree
- penalties_coefficient: 1
- boosting_type: Plain
- model_shrink_mode: Constant'
- feature_border_type: 'GreedyLogSum
- bayesian_matrix_reg: 0.10000000149011612
- l2_leaf_reg: 3
- random_strength: 1
- rsm: 1
- boost_from_average: False
- model_size_reg: 0.5

- pool_metainfo_options: tags: {}}
- subsample: 0.800000011920929
- class_names: [0, 1]
- random_seed: 600
- depth: 4
- posterior_sampling: False
- border_count: 254
- classes_count: 0
- auto_class_weights: None
- sparse_features_conflict_fraction: 0
- leaf_estimation_backtracking: AnyImprovement
- best_model_min_trees: 1
- model_shrink_rate: 0
- min_data_in_leaf: 1
- loss_function: Logloss
- learning_rate: 0.03999999910593033
- score_function: Cosine
- task_type:CPU
- leaf_estimation_iterations: 10
- bootstrap_type: MVS
- max_leaves: 16

DecisionTreeClassifier

- criterion=entropy
- max_depth=7
- splitter=random

XGBClassifier

- base_score=0.5
- booster=gbtree
- colsample_bylevel=1
- colsample_bynode=1
- colsample_bytree=1
- eval_metric=mlogloss
- gamma=0
- gpu_id=-1
- importance_type=gain
- learning_rate=0.1
- max_delta_step=0
- max_depth=8
- min_child_weight=1

- missing=nan
- monotone_constraints=()
- n_estimators=180
- n_jobs=8
- num_parallel_tree=1
- random_state=0
- reg_alpha=0
- reg_lambda=1
- scale_pos_weight=1
- subsample=1
- tree_method=exact
- use_label_encoder=False
- validate_parameters=1

Support Vector Machines (SVM)

- C=10
- gamma=0.1
- kernel=poly
- probability=True

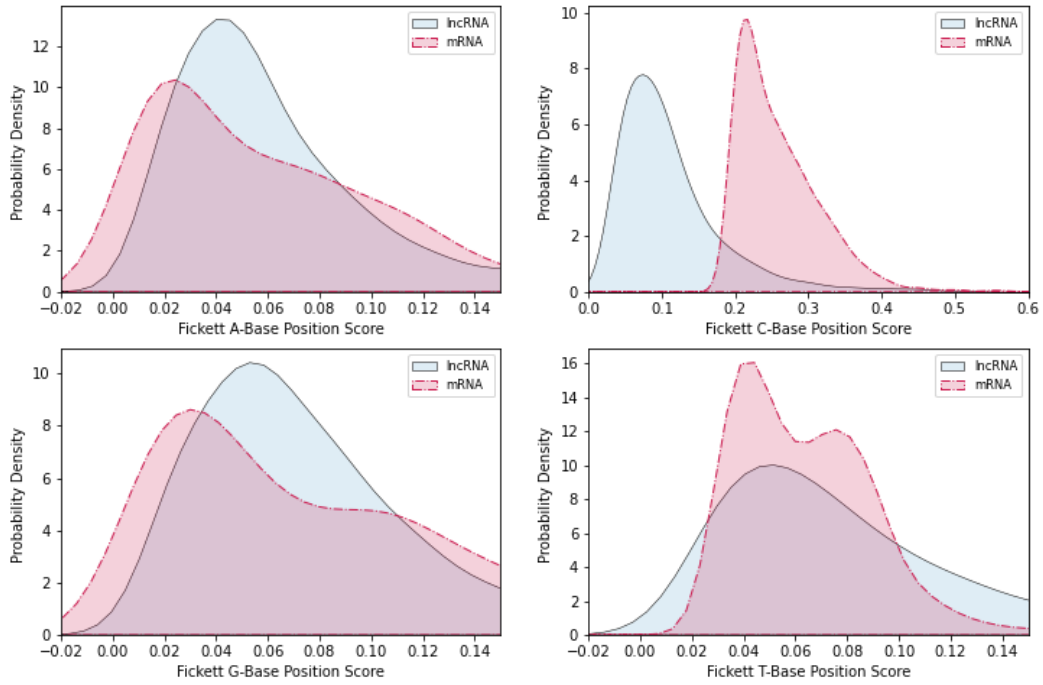
RandomForestClassifier

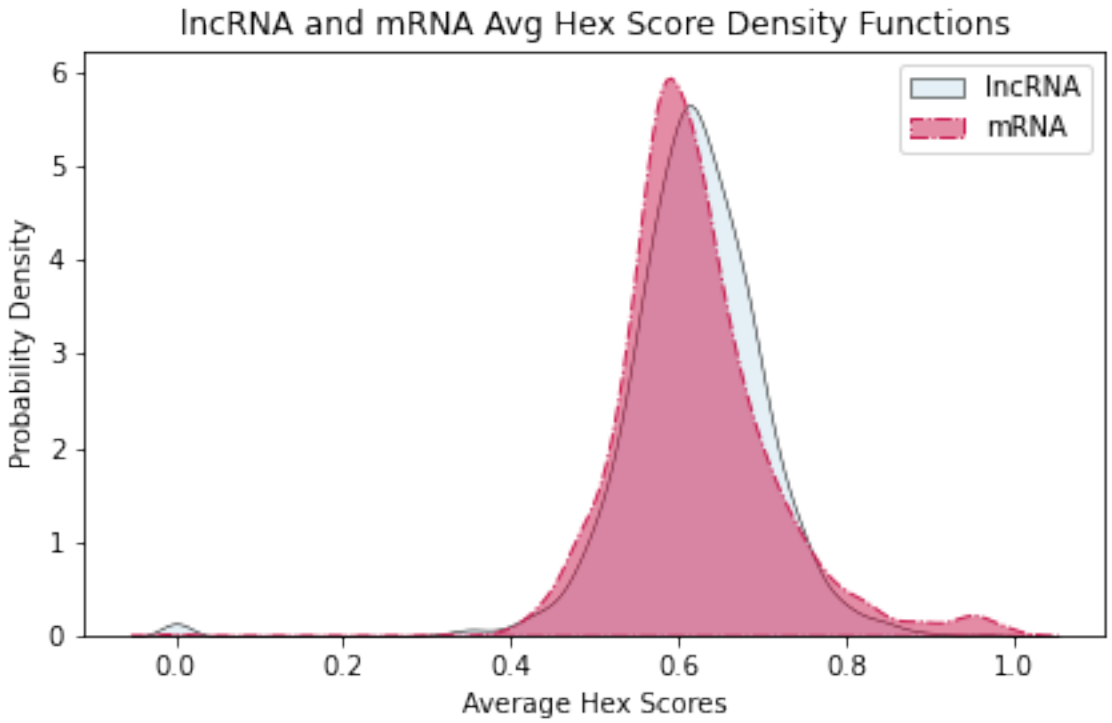
- max_depth=20
- n_estimators=200
- Random_state=42

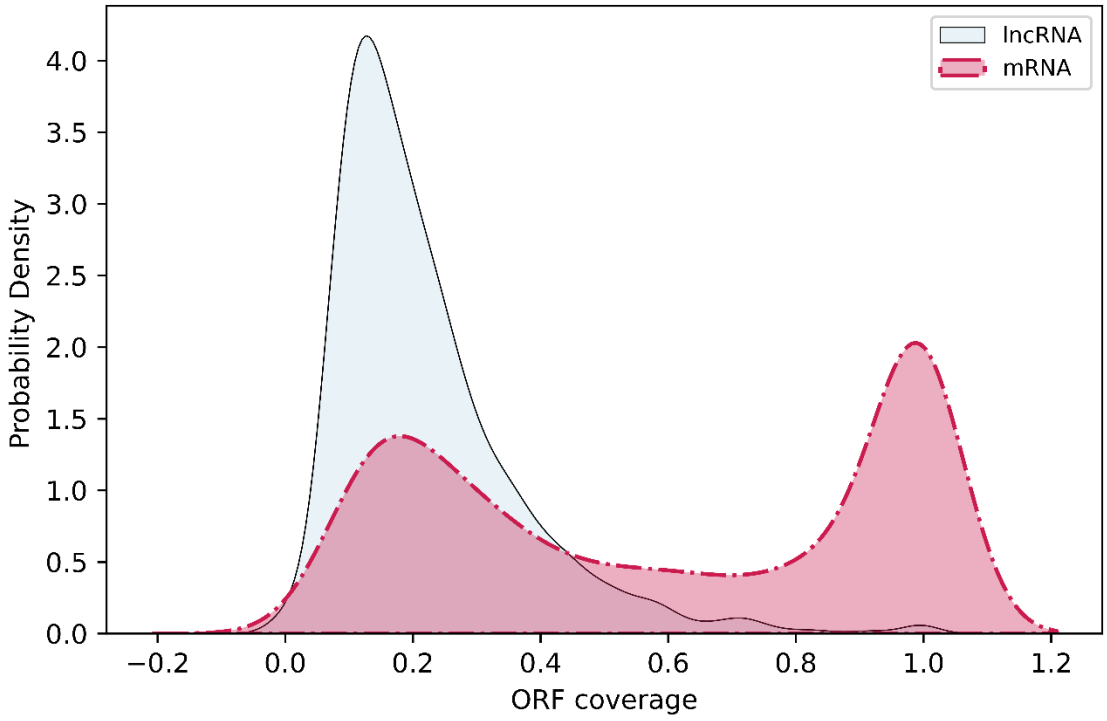
MLPClassifier

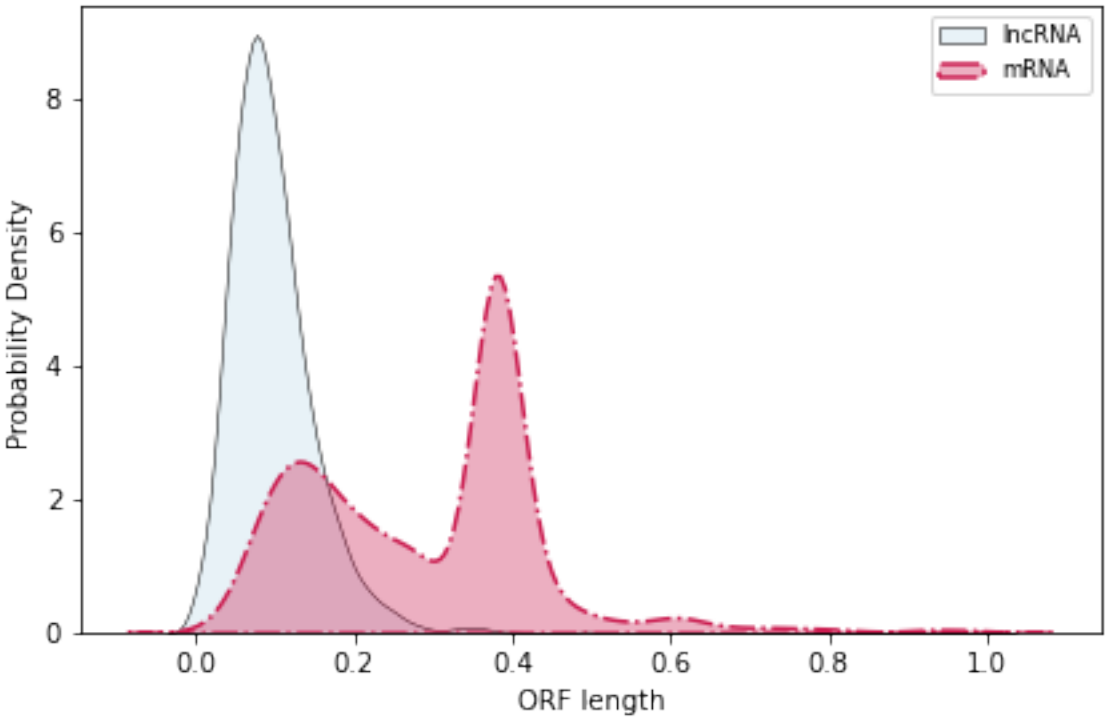
- alpha=0.05
- hidden_layer_sizes=(50, 100, 50)
- max_iter=1000

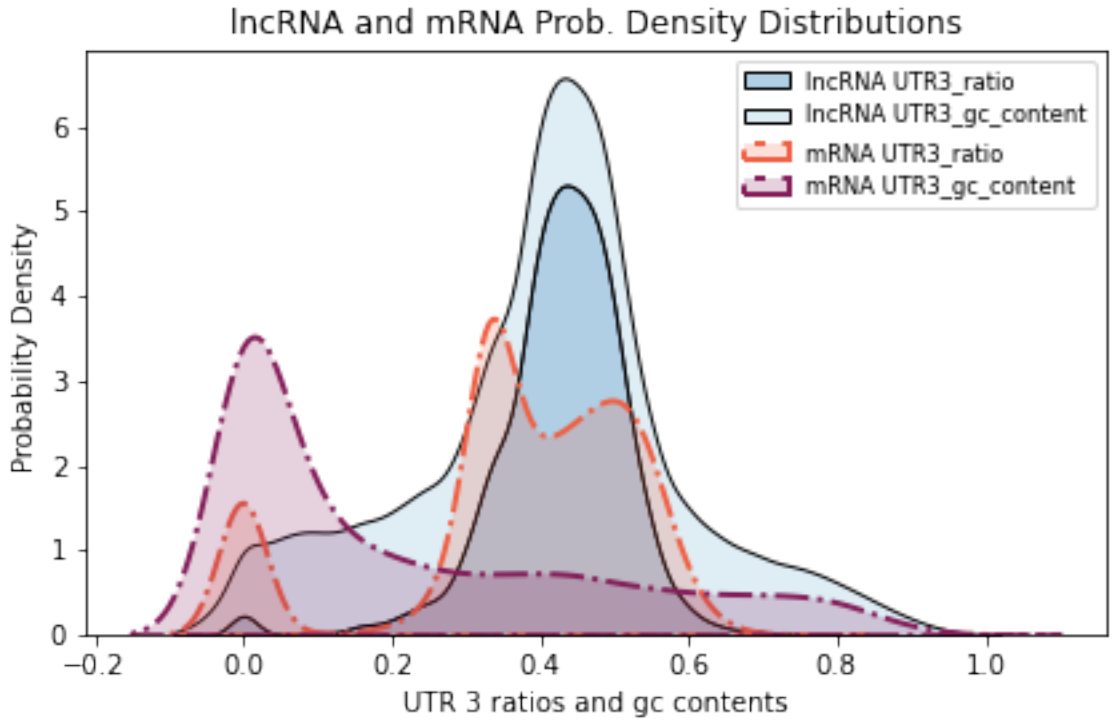
Supplementary File 03: Graphs represent probability density function for Fickett, Average HEX, ORF, UTR, and HMM Features and SHAP values in Mouse.

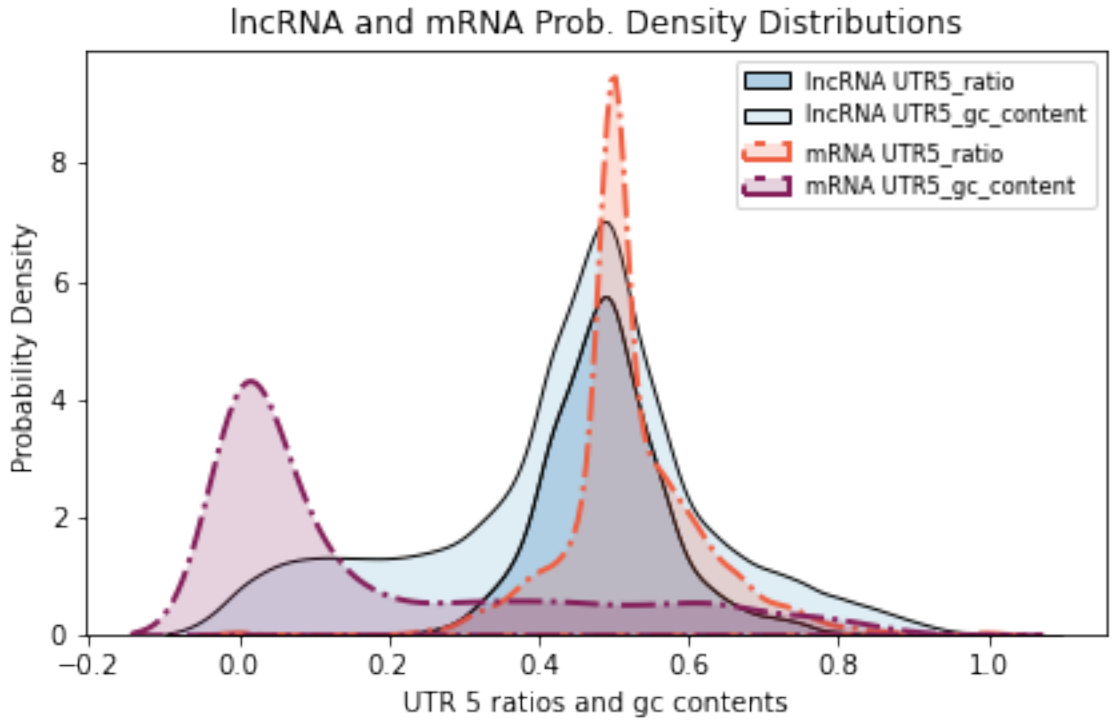


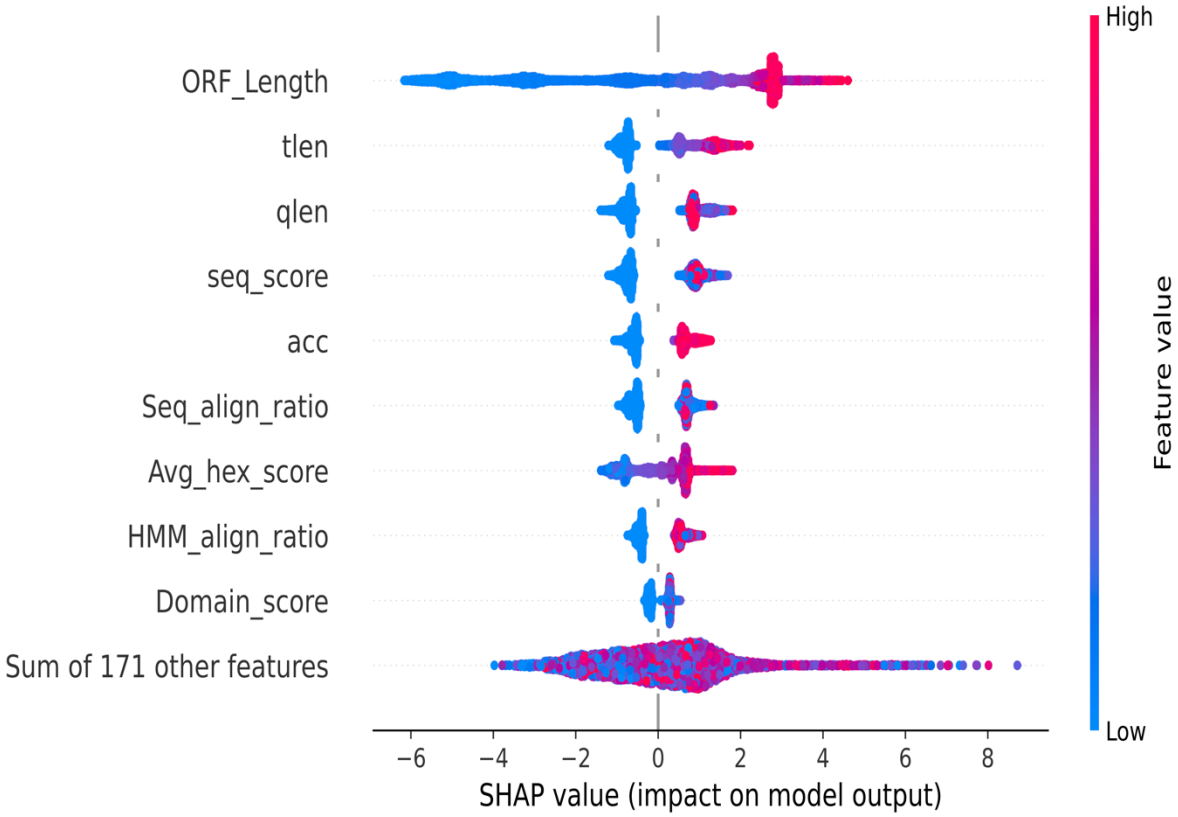



lncRNA
mRNA
Probability Density
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
ORF coverage


lncRNA
mRNA
Probability Density
8
6
4
2
0
0.0
0.2
0.4
0.6
0.8
1.0
ORF length


lncRNA and mRNA Prob. Density Distributions
lncRNA UTR3_ratio
lncRNA UTR3_gc_content
mRNA UTR3_ratio
mRNA UTR3_gc_content
Probability Density
UTR 3 ratios and gc contents


lncRNA and mRNA Prob. Density Distributions
lncRNA UTR5_ratio
lncRNA UTR5_gc_content
mRNA UTR5_ratio
mRNA UTR5_gc_content
Probability Density
UTR 5 ratios and gc contents


ORF_Length
tlen
qlen
seq_score
acc
Seq_align_ratio
Avg_hex_score
HMM_align_ratio
Domain_score
Sum of 171 other features
−6
−4
−2
0
2
4
6
8
SHAP value (impact on model output)
High
Feature value
Low