An improved machine learning-based model for the classification of off-targets in the CRISPR/Cpf1 system

Pragya Kesarwani[1,2], Dhvani Sandip Vora[2] and Durai Sundar[2,3,*]

[1]Regional Centre for Biotechnology, NCR Biotech Science Cluster, 3rd Milestone, Faridabad-Gurgaon Expressway, Faridabad 121001, Haryana, India
[2]Department of Biochemical Engineering and Biotechnology, Indian Institute of Technology (IIT) Delhi, New Delhi 110016, India
[3]Yardi School of Artificial Intelligence, Indian Institute of Technology (IIT) Delhi, New Delhi 110016, India

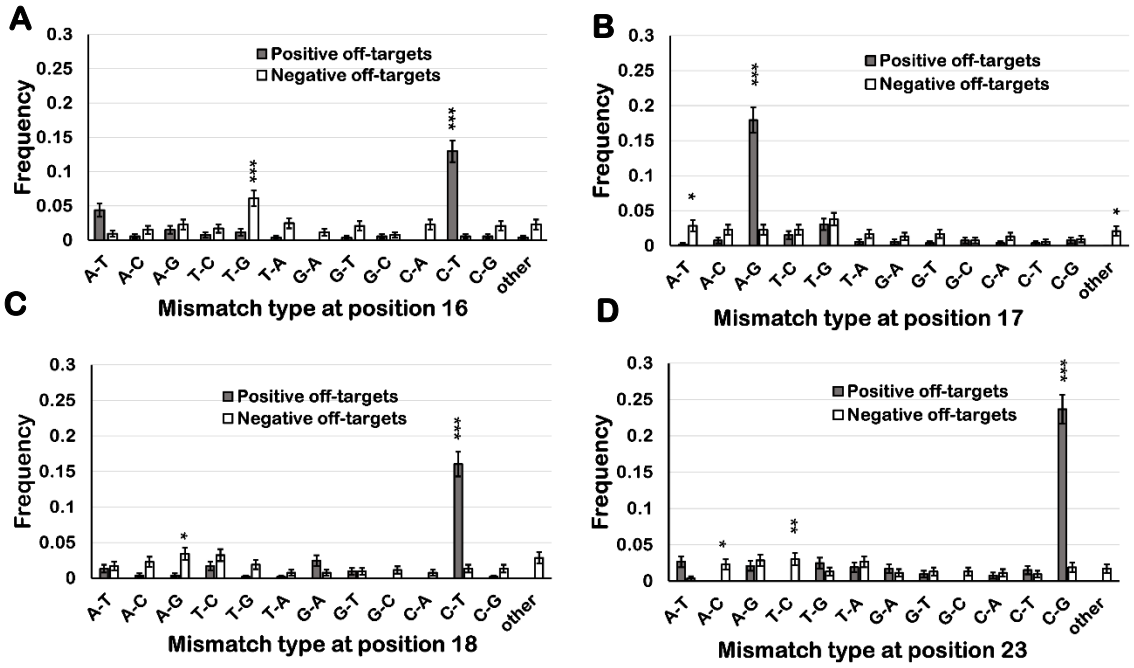
[*]Correspondence: DS: sundar@dbeb.iitd.ac.in

PK: pragya.kesarwani@rcb.res.in

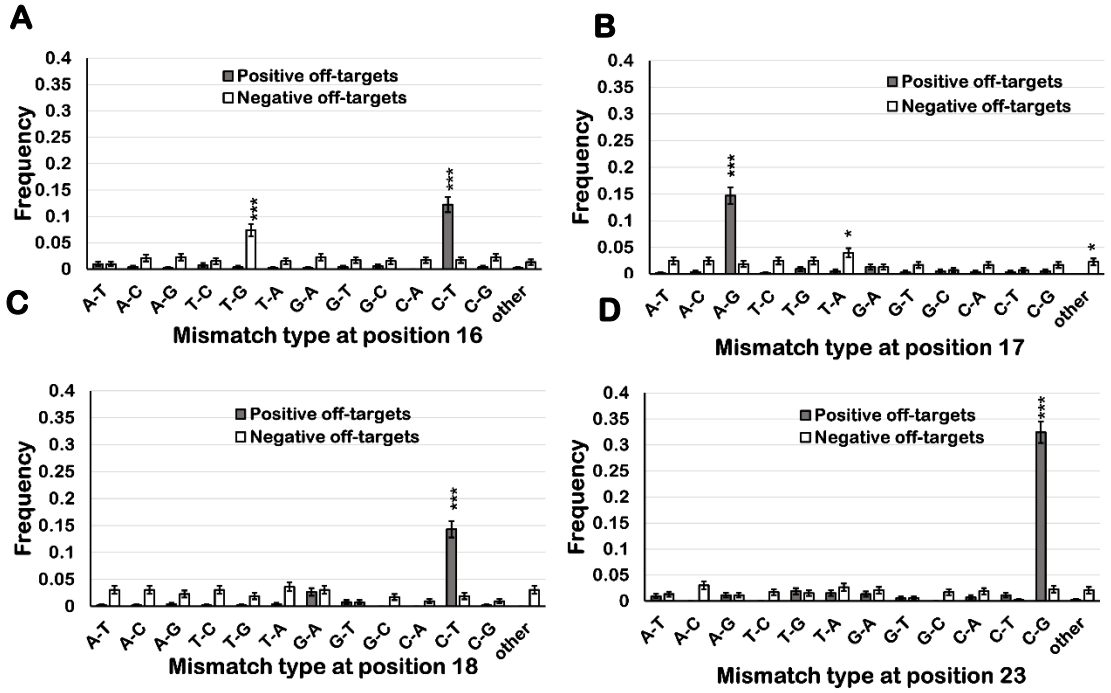
DSV: dhvani.vora@dbeb.iitd.ac.in

DS: sundar@dbeb.iitd.ac.in

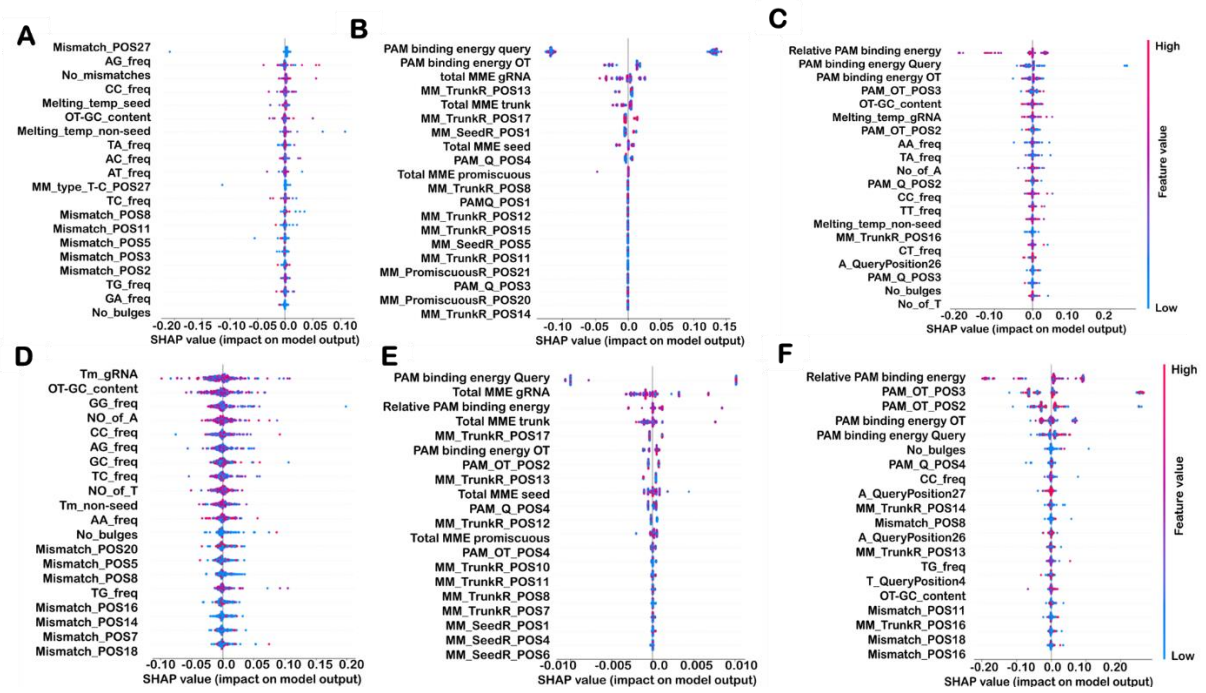
List of Supplementary Data



Supplementary Figure 1: Presence of one-base mismatch type in positive and negative off-targets of AsCpf1 at position (A) 16, (B)17, (C) 18 and (D) 23 from PAM.; * denotes the P-value >= 0.0005 whereas *** denotes the P-value < 0.0005; error bars, s.e.m.



Supplementary Figure 2: Presence of one-base mismatch type in positive and negative off-targets of LbCpf1 at position (A) 16, (B)17, (C) 18 and (D) 23 from PAM.; * denotes the P-value >= 0.0005 whereas *** denotes the P-value < 0.0005; error bars, s.e.m.



Supplementary Figure 3: Summary plot depicting top 20 features of best models on hybrid feature sets. Top 20 features based on mean SHAP values of models trained on AsCpf1 dataset using (A) sequence-based feature set (B) base-dependent binding energy associated feature set and (C) sequence and base-dependent binding energy associated feature set. Similarly, the Summary plot is based on mean SHAP values for the LbCpf1 dataset using (D) sequence-based feature set (E) base-dependent binding energy associated feature set, and (F) sequence and base-dependent binding energy associated feature set. The data points lying in the left quadrant represents the negative off-targets and data points lying in the right quadrant represents positive off-targets. The color of the data points represents the feature value. Blue color represents the minimum value of the feature and magenta indicates the maximum feature value.

Supplementary Table 1: Hyperparameters tuned recursively to optimize machine learning models for both AsCpf1 and LbCpf1

ML model	Hyperparameters
MLPClassifier	Seed, weight optimization, alpha, activation function, number of hidden layers, size of hidden layers, maximum number of iterations, validation set, initial learning rate, learning rate schedule, random state, tolerance for optimization, maximum number of epochs without any improvement in tolerance, validation set, epsilon, beta1 and beta2
AdaBoostClassifier	Base estimator, number of estimators, learning rate, random state, seed value
LinearSVC	Regularization parameter (C), class weights, loss function, penalty, dual, tolerance, seed value
LogisticRegression	Class weights, inverse of regularization strength (C), maximum number of iterations, algorithms, seed value
DecisionTreeClassifier	Class weights, number of features, maximum depth of the tree, minimal cost_complexity pruning (ccp_alpha), number of leaf nodes, random state, Seed value

*** The optimized values for the above mentioned parameters have been provided in the associated Jupyter Notebook.**

Supplementary Table 2: Comparison of prediction errors of best performing models of three different feature sets with and without undersampling.

Species	Model	Without under sampling			With under sampling		
		Bias	**Variance**	**MSE**	**Bias**	**Variance**	**MSE**
AsCpf1	SeqClassifier	0.003	0.001	0.004	0.008	0.015	0.023
	MMEClassifier	0.003	0.000	0.003	0.032	0.023	0.054
	CombClassifier	0.003	0.001	0.003	0.015	0.017	0.032
LbCpf1	SeqClassifier	0.002	0.001	0.002	0.039	0.036	0.075
	MMEClassifier	0.002	0.001	0.002	0.063	0.029	0.92
	CombClassifier	0.004	0.007	0.010	0.032	0.043	0.074

Supplementary Table 3: Performances of optimized machine learning models on 25% test split of AsCpf1 dataset with hybrid feature set

Feature sets	Models	Precision	Recall	F1 score	MCC
Sequence associated feature set	**AdaBoostClassifier**	**0.92**	**0.90**	**0.91**	**0.815**
	LinearSVC	0.85	0.91	0.87	0.749
	LogisticRegression	0.83	0.93	0.87	0.746
	DecisionTreeClassifier	0.81	0.92	0.86	0.722
	MLPClassifier	0.87	0.89	0.88	0.766
Mismatch energy associated feature set	**AdaBoostClassifier**	**0.95**	**0.89**	**0.92**	**0.844**
	LinearSVC	0.88	0.83	0.86	0.715
	LogisticRegression	0.91	0.88	0.89	0.789
	DecisionTreeClassifier	0.97	0.87	0.92	0.837
	MLPClassifier	0.97	0.86	0.91	0.827
Sequence and mismatch energy associated feature set	AdaBoostClassifier	0.94	0.86	0.89	0.790
	LinearSVC	0.95	0.86	0.90	0.806
	LogisticRegression	0.91	0.86	0.88	0.770
	DecisionTreeClassifier	0.67	0.84	0.72	0.477
	MLPClassifier	**0.92**	**0.91**	**0.92**	**0.832**

Supplementary Table 4: Performances of optimized machine learning models on 25% test split of LbCpf1 dataset with hybrid feature set

Feature sets	Models	Precision	Recall	F1 score	MCC
Sequence associated feature set	AdaBoostClassifier	0.95	0.85	0.89	0.791
	LinearSVC	0.87	0.93	0.89	0.789
	LogisticRegression	0.88	0.93	0.90	0.800
	DecisionTreeClassifier	0.86	0.90	0.88	0.756
	MLPClassifier	**0.92**	**0.90**	**0.91**	**0.824**
Mismatch energy associated feature set	**AdaBoostClassifier**	**0.94**	**0.90**	**0.92**	**0.834**
	SVC	0.97	0.86	0.91	0.818
	LogisticRegression	0.93	0.87	0.89	0.791
	DecisionTreeClassifier	0.81	0.90	0.85	0.714
	MLPClassifier	0.96	0.86	0.91	0.815
Sequence and mismatch energy associated feature set	AdaBoostClassifier	0.96	0.84	0.89	0.784
	LinearSVC	0.88	0.91	0.89	0.789
	LogisticRegression	0.90	0.92	0.91	0.818
	DecisionTreeClassifier	0.86	0.85	0.86	0.713
	MLPClassifier	**0.91**	**0.92**	**0.91**	**0.821**

Supplementary Table 5: Comparison of performance metrics of best performing model with the existing models

Dataset	Models	Precision	Recall	F1 score	auc_roc	MCC
AsCpf1	CombClassifier	0.92	0.91	0.92	0.915	0.832
	SeqClassifier	0.92	0.90	0.91	0.899	0.815
	MMEClassifier	0.95	0.89	0.92	0.892	0.820
	CRISPR-DT	0.50	0.45	0.37	0.483	0.016
	CINDEL	0.50	0.48	0.49	0.476	0.017
	DeepCpf1	0.68	0.57	0.62	0.546	0.116
LbCpf1	CombClassifier	0.91	0.92	0.91	0.915	0.821
	SeqClassifier	0.92	0.90	0.91	0.903	0.824
	MMEClassifier	0.94	0.90	0.92	0.899	0.833
	CRISPR-DT	0.50	0.48	0.49	0.483	0.017
	CINDEL	0.49	0.35	0.40	0.348	0.076
	DeepCpf1	0.63	0.54	0.58	0.537	0.120

Supplementary Table 6: Highly significant position-specific mononucleotides with enrichment score and P-values calculated using a Welch *t-test*.

Position	Nucleotide	P-value (LbCPf1)	enrichment ratio (LbCPf1)	P-value (AsCpf1)	Enrichment ratio (AsCpf1)
1	C	2.67E-15	0.163355	5.86E-24	0.048169
	G	6.69E-17	3.193583	6.07E-07	2.178062
	T	0.003058	1.155534	5.71E-14	1.38697
	A	5.01E-16	0.117872	9.63E-16	0.068624
2	C	5.75E-24	0.07221	1.80E-28	0.041746
	G	3.93E-11	0	1.91E-12	0.020038
	T	8.52E-63	1.752491	1.72E-66	1.837477
	A	8.08E-20	0	4.34E-16	0
3	C	1.54E-07	0.216098	1.99E-12	0.05465
	G	1.61E-17	0.02745	6.26E-16	0.015655
	T	4.09E-65	1.824905	8.07E-51	1.577262
	A	4.88E-31	0.00828	7.31E-18	0
4	C	0.00053	1.559933	4.63E-07	1.900679
	G	1.22E-07	1.535392	9.84E-05	1.39396
	T	1.52E-18	0.341432	1.27E-21	0.304929
	A	0.603652	0.938725	0.63144	1.041985
5	C	0.007723	1.355523	1.49E-07	1.75334
	G	9.00E-30	2.298496	7.16E-15	1.846801
	T	5.08E-16	0.338224	3.14E-13	0.378204
	A	4.74E-19	0.178306	6.32E-20	0.14313
6	C	0.000893	0.590013	0.522355	0.882278
	G	1.67E-08	1.520338	0.000748	1.340789
	T	0.222999	1.118259	0.000136	1.335878
	A	1.76E-09	0.359909	6.81E-18	0.202066
7	C	0.541413	0.925573	0.939707	0.972149
	G	3.53E-53	3.59111	1.38E-33	2.435533
	T	3.86E-25	0.333969	2.84E-19	0.354843
	A	5.41E-10	0.337276	4.71E-10	0.222646
8	C	4.56E-15	0.363192	8.99E-11	0.400763
	G	1.88E-10	0.150286	0.008993	0.550067
	T	9.62E-14	1.830825	4.71E-09	1.651583
	A	0.004876	1.271171	0.115496	1.121183
9	C	0.00046	0.660679	1.52E-05	0.577743
	G	1.58E-20	2.416975	1.92E-15	2.162013
	T	0.611726	0.955308	0.297628	1.077524
	A	5.47E-12	0.313096	7.75E-12	0.255806
10	C	0.003389	0.692376	0.684719	0.931066
	G	8.73E-08	1.748428	2.07E-09	1.830016
	T	1.16E-25	0.157835	1.74E-25	0.158923
	A	1.92E-08	1.579167	0.00334	1.274537
11	C	0.000532	0.592966	0.012353	0.624378
	G	6.43E-05	1.558524	1.11E-05	1.603053

	T	0.002283	1.259083	0.170254	0.881487
	A	3.03E-06	0.559036	0.368631	0.882022
12	C	4.80E-21	2.755248	3.43E-11	2.168514
	G	0.003885	0.667939	2.96E-05	0.520599
	T	0.910877	0.989837	0.017804	1.196218
	A	3.90E-11	0.471829	2.91E-08	0.497546
13	C	0.112851	1.166508	0.00331	1.276404
	G	0.000221	0.619065	2.80E-06	0.500954
	T	1.24E-10	0.459996	4.25E-10	0.446797
	A	8.95E-15	2.145501	1.65E-12	2.18137
14	C	0.929228	1.009442	0.01234	1.278297
	G	1.86E-09	1.7291	0.004636	1.31083
	T	0.000226	0.651591	0.050202	0.779262
	A	0.003157	0.687259	0.000842	0.640564
15	C	1.93E-10	0.405352	1.21E-07	0.472901
	G	7.48E-14	0.410138	3.22E-16	0.318487
	T	1.24E-05	0.466004	0.065348	0.751431
	A	1.72E-49	2.699171	1.68E-37	2.567916
16	C	5.64E-36	3.143919	9.54E-25	2.641395
	G	2.74E-07	0.430821	0.013905	0.635012
	T	0.004482	0.746764	0.014993	0.775286
	A	4.26E-10	0.500954	7.16E-11	0.450859
17	C	2.44E-19	2.428539	1.85E-14	2.295512
	G	0.475457	1.113232	0.714755	0.932008
	T	2.25E-16	0.288319	4.49E-06	0.509743
	A	0.023411	0.835693	0.009386	0.791752
18	C	7.03E-43	3.245312	2.15E-31	2.790261
	G	0.006899	0.7121	0.004309	0.661638
	T	1.17E-11	0.442561	4.63E-05	0.627001
	A	1.50E-08	0.48319	2.45E-09	0.42717
19	C	0.015501	1.354853	0.002588	1.409278
	G	9.31E-08	0.51487	2.69E-10	0.411231
	T	9.22E-07	0.512977	0.000341	0.589909
	A	2.36E-11	1.651293	9.71E-09	1.549003
20	C	1.13E-07	0.304929	2.87E-05	0.369124
	G	6.86E-06	0.612277	5.36E-06	0.581107
	T	0.418795	0.922287	0.441785	1.05464
	A	2.33E-16	1.923955	2.12E-09	1.667238
21	C	1.33E-23	2.460998	6.91E-13	1.965282
	G	0.114503	1.18258	0.010224	1.313181
	T	2.04E-05	0.477654	0.150574	0.780708
	A	2.00E-17	0.414251	3.44E-17	0.403754
22	C	1.40E-14	0.36198	1.09E-08	0.468518
	G	1.78E-05	0.471486	0.010011	0.59129
	T	4.49E-06	1.616286	3.08E-06	1.54075
	A	9.92E-08	1.488384	0.021302	1.189012
23	C	1.14E-05	1.51069	0.015363	1.265122
	G	0.006918	0.664309	0.011354	0.667939

	T	0.027167	1.282817	0.15793	1.134974
	A	2.77E-05	0.678215	0.152308	0.848282
24	C	1.06E-11	1.81448	0.000106	1.416761
	G	0.112259	1.295571	0.101062	1.288168
	T	0.000417	0.601145	0.001285	0.635138
	A	4.40E-07	0.663657	0.07593	0.841174
25	C	0.628107	0.948777	0.161395	1.189766
	G	0.000619	0.646393	0.839927	0.945197
	T	2.20E-06	0.593723	0.42042	0.873757
	A	0.865976	1.007944	0.722505	0.96139
26	C	0.867066	1.020292	1.94E-05	2.027117
	G	1.66E-10	1.518043	3.78E-11	1.767594
	T	0.01275	0.751431	0.954895	0.989539
	A	5.08E-33	0.398153	4.73E-23	0.3983
27	C	0.063181	1.217194	8.95E-07	1.9243
	G	3.46E-16	1.76218	9.86E-18	2.241383
	T	4.77E-06	0.577489	0.887903	0.960162
	A	1.04E-42	0.33715	2.41E-37	0.28626

Supplementary Table 7: Complete set of features used for training model for the prediction of on-target efficiencies.

Feature set 3	Feature set 1	Position-Specific nucleotide composition	Position-specific mononucleotides in target (27*4=108)
			Position-specific mononucleotides in off-target (27*4=108)
		Position non-specific nucleotide composition	Mononucleotides counts (4)
			Dinucleotides frequencies (16)
			GC content (1)
		Repetitive sequences	Presence of AAAA (1), TTTT (1), GGGG (1), CCCC(1)
		Mismatches	Number of mismatches (1)
			Position-specific mismatches (27)
			Position-specific mismatch types (16*27=432)
		Bulges	Number of bulges (1)
			Position-specific bulges (27)
		Thermodynamics Related features	Minimum free energy (1)
			Melting temperature (Seed region (1) , non-seed region (1) and gRNA (1))
	Feature set 2		Base-dependent binding energy of PAM region (8) [1]
			PAM binding energy of target (1) and off-target (1)
			Relative PAM binding energy (1)
			Position-specific mismatch-dependent binding energy weights to protospacer region (23) [1]
			Total base-dependent binding energy of seed region, non-seed region and gRNA (3) [1]

Reference

1. Specht, D. A.; Xu, Y.; Lambert, G., Massively parallel CRISPRi assays reveal concealed thermodynamic determinants of dCas12a binding. *Proceedings of the National Academy of Sciences* **2020,** *117* (21), 11274-11282.