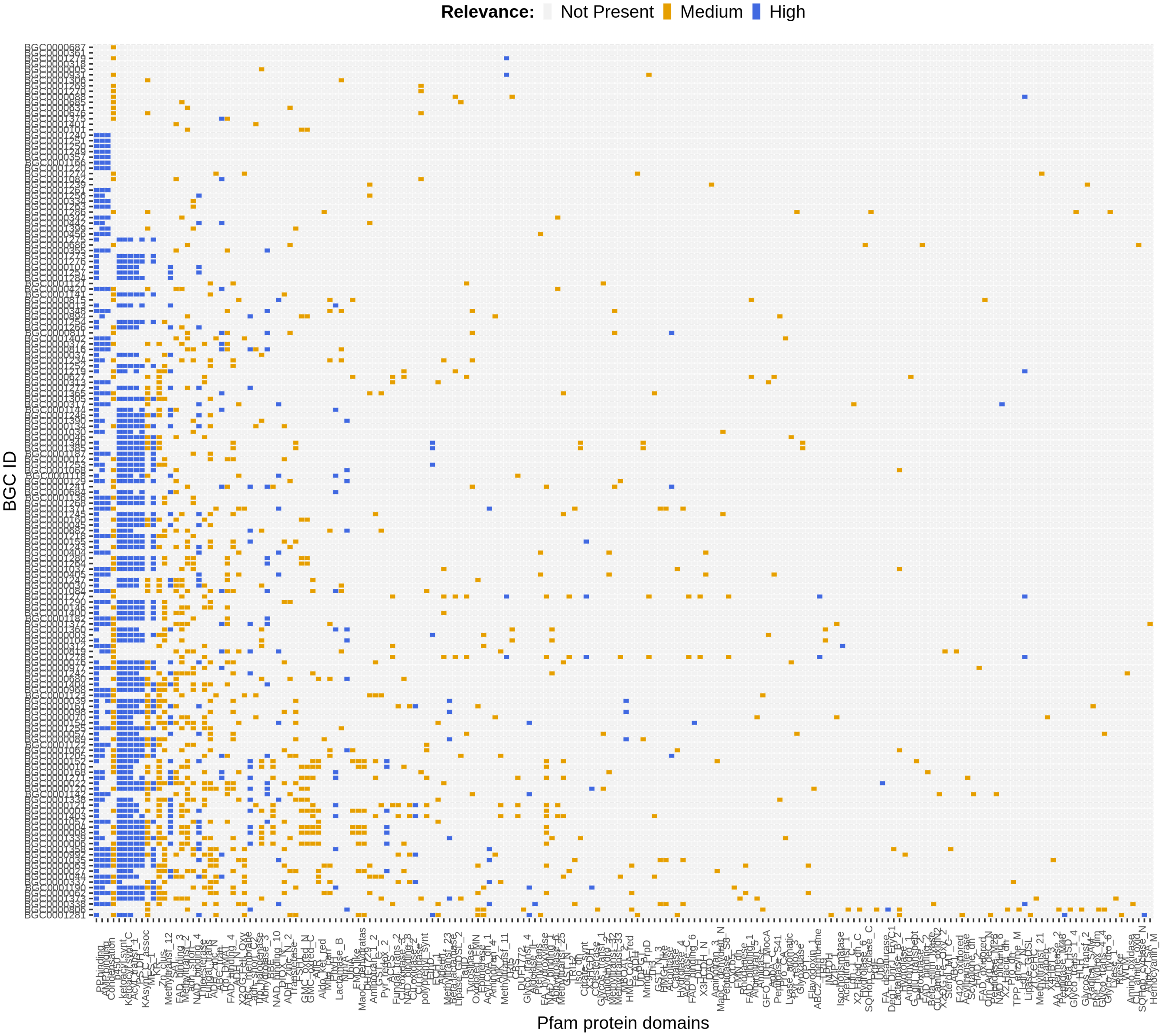
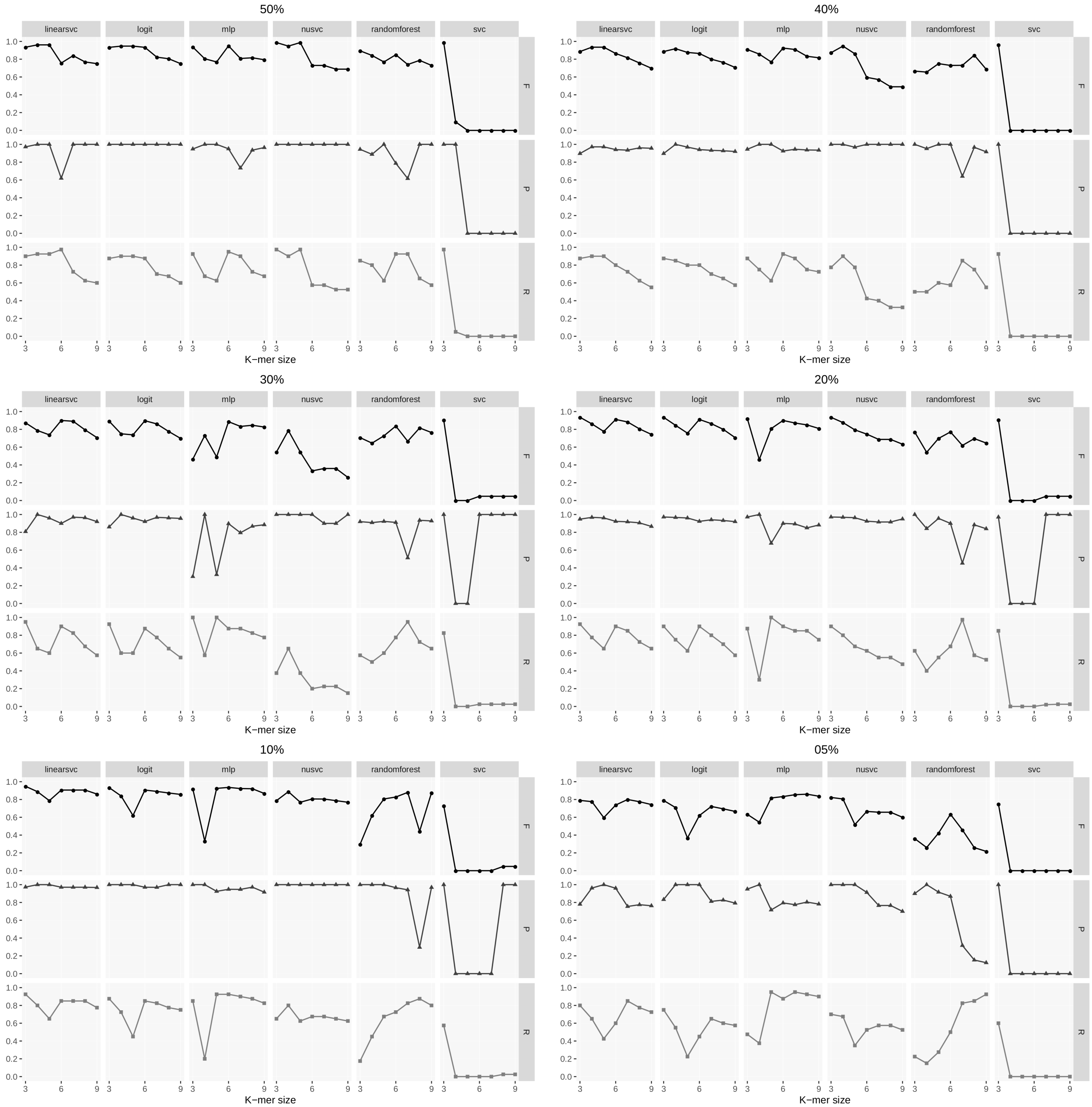
Supplementary Figure 1: Presence of Pfam protein domains annotated as *high* (usually present in BGCs) and *medium* (usually present, but not limited to BGCs) in our datatset positive instances. Each positive instance in our datasets is represented in a row. The columns represent the absence or presence of a *high* or *medium* Pfam protein domains, sorted by occurrence. The distribution of *high* and *medium* protein domains among positive instances shows that a structural pattern is shared by different BGC IDs.



Supplementary Figure 2: P, R, and F-m for classifiers on each validation set for $3 \leq K \leq 9$



Supplementary Table 1: Distribution of positive and negative instances across fungal BGC datasets, from completely balanced (50% positive, 50% negative) to most imbalanced (05% positive, 95% negative). Each dataset was split between train and validation subsets during the training phase.

Dataset	**Train**		**Validation**		**Total**	
distribution	**Pos**	**Neg**	**Pos**	**Neg**	**Pos**	**Neg**
50% - 50%	160	160	40	40	200	200
40% - 60%	160	240	40	60	200	300
30% - 70%	160	373	40	93	200	466
20% - 80%	160	640	40	160	200	800
10% - 90%	160	1,440	40	360	200	1,800
05% - 95%	160	3,040	40	760	200	3,800

Supplementary Table 2: Pfam domains annotated as *high* (usually present in BGCs) in our dataset positive instances.

Pfam ID	**Domain**	**Pfam ID**	**Domain**
PF00389	2-Hacid_dh	PF00378	ECH_1
PF01073	3Beta_HSD	PF00487	FA_desaturase
PF00725	3HCDH	PF00551	Formyl_trans_N
PF00583	Acetyltransf_1	PF00368	HMG-CoA_red
PF01648	ACPS	PF16197	KAsynt_C_assoc
PF00698	Acyl_transf_1	PF00109	ketoacyl-synt
PF13561	adh_short_C2	PF02801	Ketoacyl-synt_C
PF00578	AhpC-TSA	PF08659	KR
PF00596	Aldolase_II	PF00753	Lactamase_B
PF01063	Aminotran_4	PF00657	Lipase_GDSL
PF00501	AMP-binding	PF12013	OrsD
PF08031	BBE	PF00550	PP-binding
PF00144	Beta-lactamase	PF00432	Prenyltrans
PF00199	Catalase	PF14765	PS-DH
PF00135	COesterase	PF16073	SAT
PF00668	Condensation	PF00975	Thioesterase
PF00394	Cu-oxidase	PF06330	TRI5
PF01041	DegT_DnrJ_EryC1	PF08195	TRI9
PF14226	DIOX_N	PF11991	Trp_DMAT
PF01738	DLH	PF01040	UbiA

Supplementary Table 3: Pfam domains annotated as *medium* (usually present, but not limited to BGCs) in our datatset positive instances.

Pfam ID	Domain	Pfam ID	Domain	Pfam ID	Domain
PF02826	2-Hacid_dh_C	PF00970	FAD_binding_6	PF00891	Methyltransf_2
PF10014	2OG-Fe_Oxy_2	PF12831	FAD_oxidored	PF05050	Methyltransf_21
PF03171	2OG-FeII_Oxy	PF18325	Fas_alpha_ACP	PF13489	Methyltransf_23
PF02737	3HCDH_N	PF18314	FAS_I_H	PF13649	Methyltransf_25
PF13622	4HBT_3	PF17951	FAS_meander	PF13679	Methyltransf_32
PF13520	AA_permease_2	PF17828	FAS_N	PF10017	Methyltransf_33
PF00664	ABC_membrane	PF00465	Fe-ADH	PF07690	MFS_1
PF00005	ABC_tran	PF01613	Flavin_Reduct	PF00153	Mito_carr
PF03109	ABC1	PF00258	Flavodoxin_1	PF03972	MmgE_PrpD
PF01061	ABC2_membrane	PF01070	FMN_dh	PF00175	NAD_binding_1
PF07859	Abhydrolase_3	PF00743	FMO-like	PF13460	NAD_binding_10
PF08386	Abhydrolase_4	PF03959	FSH1	PF07993	NAD_binding_4
PF12697	Abhydrolase_6	PF04082	Fungal_trans	PF08030	NAD_binding_6
PF00330	Aconitase	PF11951	Fungal_trans_2	PF13450	NAD_binding_8
PF00694	Aconitase_C	PF01019	G_glu_transpept	PF05368	NmrA
PF00441	Acyl-CoA_dh_1	PF00117	GATase	PF03169	OPT
PF01553	Acyltransferase	PF01408	GFO_IDH_MocA	PF02784	Orn_Arg_deC_N
PF08240	ADH_N	PF01341	Glyco_hydro_6	PF00724	Oxidored_FMN
PF00106	adh_short	PF13692	Glyco_trans_1_4	PF00067	p450
PF00107	ADH_zinc_N	PF13632	Glyco_trans_2_3	PF04389	Peptidase_M28
PF13602	ADH_zinc_N_2	PF13579	Glyco_trans_4_4	PF01432	Peptidase_M3
PF08493	AflR	PF13439	Glyco_transf_4	PF01435	Peptidase_M48
PF00171	Aldedh	PF00534	Glycos_transf_1	PF02129	Peptidase_S15
PF00248	Aldo_ket_red	PF00535	Glycos_transf_2	PF03572	Peptidase_S41
PF01425	Amidase	PF00903	Glyoxalase	PF00082	Peptidase_S8
PF01979	Amidohydro_1	PF05199	GMC_oxred_C	PF08530	PepX_C
PF04909	Amidohydro_2	PF00732	GMC_oxred_N	PF01328	Peroxidase_2
PF01593	Amino_oxidase	PF00043	GST_C	PF07976	Phe_hydrox_dim
PF00155	Aminotran_1_2	PF02798	GST_N	PF05721	PhyH
PF00202	Aminotran_3	PF13417	GST_N_3	PF00348	polyprenyl_synt
PF00266	Aminotran_5	PF08759	GT-D	PF00484	Pro_CA
PF12796	Ank_2	PF13419	HAD_2	PF01619	Pro_dh
PF08546	ApbA_C	PF00372	Hemocyanin_M	PF04303	PrpF
PF00026	Asp	PF00132	Hexapep	PF07992	Pyr_redox_2
PF01212	Beta_elim_lyase	PF00010	HLH	PF13738	Pyr_redox_3
PF00170	bZIP_1	PF00682	HMGL-like	PF14027	Questin_oxidase
PF00571	CBS	PF18558	HTH_51	PF04055	Radical_SAM
PF00285	Citrate_synt	PF00702	Hydrolase	PF00581	Rhodanese
PF01179	Cu_amine_oxid	PF12146	Hydrolase_4	PF00355	Rieske
PF02727	Cu_amine_oxidN2	PF13344	Hydrolase_6	PF02982	Scytalone_dh
PF07731	Cu-oxidase_2	PF01231	IDO	PF13243	SQHop_cyclase_C
PF07732	Cu-oxidase_3	PF00478	IMPDH	PF13249	SQHop_cyclase_N
PF00173	Cyt-b5	PF00180	Iso_dh	PF08498	Sterol_MT_C
PF01266	DAO	PF00857	Isochorismatase	PF02668	TauD
PF01323	DSBA	PF12706	Lactamase_B_2	PF00205	TPP_enzyme_M
PF08354	DUF1729	PF02866	Ldh_1_C	PF02458	Transferase
PF08592	DUF1772	PF00056	Ldh_1_N	PF06609	TRI12
PF06441	EHN	PF02900	LigB	PF07428	Tri3
PF01370	Epimerase	PF13472	Lipase_GDSL_2	PF04820	Trp_halogenase
PF07110	EthD	PF00206	Lyase_1	PF00264	Tyrosinase
PF03807	F420_oxidored	PF00221	Lyase_aromatic	PF01977	UbiD
PF04116	FA_hydroxylase	PF13452	MaoC_dehydrat_N	PF00201	UDPGT
PF00667	FAD_binding_1	PF01575	MaoC_dehydratas	PF08325	WLM
PF00890	FAD_binding_2	PF13813	MBOAT_2	PF00096	zf-C2H2
PF01494	FAD_binding_3	PF08241	Methyltransf_11	PF00098	zf-CCHC
PF01565	FAD_binding_4	PF08242	Methyltransf_12	PF001728	Zn_clus

Supplementary Table 4: Unique features per training dataset distribution from completely balanced (50% positive, 50% negative) to most imbalanced (05% positive, 95% negative). Number of unique features (#) and feature percentage (%) is shown per each feature type for the total number of features in each dataset. K-mer features are shown for $K = 6$, the best performing K value in our study.

Dataset	**K-mers (K=6)**		**Pfam domains**		**GO terms**		**Total**
distribution	**#**	**%**	**#**	**%**	**#**	**%**	**#**
50% - 50%	45,874	(95.41)	1,866	(3.88)	341	(0.71)	48,081
40% - 60%	59,040	(96.59)	2,370	(3.87)	286	(0.46)	61,124
30% - 70%	80,604	(96.17)	2,885	(3.44)	323	(0.38)	83,812
20% - 80%	160,750	(97.38)	3,975	(2.41)	340	(0.20)	165,065
10% - 90%	559,708	(98.61)	7,524	(1.33)	354	(0.06)	567,586
05% - 95%	1,826,067	(98.97)	18,307	(0.99)	562	(0.03)	1,844,936

Supplementary Table 5: Validation performance on fixed train and validation sets per classifier. Models were built using all feature types combined.

Dataset	**Classifier**	**P**	**R**	**F-m**	**Average**
					F-m
50-50%	`lsvc`	1	0.925	0.961	**0.755**
50-50%	`logit`	1	0.925	0.961	**0.755**
40-60%	`mlp`	0.951	0.975	0.962	0.715
30-70%	`logit`	0.947	0.9	0.923	0.693
20-80%	`lsvc`	0.925	0.925	0.925	0.732
20-80%	`mlp`	0.925	0.925	0.925	0.732
10-90%	`mlp`	0.948	0.925	0.936	0.738
05-95%	`lsvc`	0.941	0.8	0.864	0.655

Supplementary Table 6: Validation performance on 5-fold CV per classifier on the completely balanced (50% positive, 50% negative) dataset. Models were built using all feature types combined.

Dataset	**Classifier**	**P**	**R**	**F-m**
50-50%	`lsvc`	0.934	0.925	0.929
50-50%	`logit`	0.922	0.935	0.928
50-50%	`mlp`	0.948	0.910	0.928
50-50%	`nusvc`	0.708	0.750	0.723
50-50%	`randomf`	0.944	0.900	0.919
50-50%	`svc`	0.911	0.900	0.904

Supplementary Table 7: DeepBGC original and fungal optimized hyperparameters applied during evaluation

Parameter	**Original**	**Fungal**
batch_size	64	16
hidden_size	128	128
timesteps	256	256
num_epochs	328	50
dropout	0.2	0.2
optimizer	adam	adam
learning_rate	1e-4	1e-4
loss	weighted binary cross-entropy	weighted binary cross-entropy

Supplementary Table 8: TOUCAN best performing hyperparameters to maximize F-m for each classifier.

`lsvc`	C = 0.01, loss = squared_hinge, penalty = l2
`logit`	penalty = l1, C=10, solver = saga
`mlp`	activation = relu, batch_size = 256, hidden_layer_sizes= 256, learning_rate = 'adaptive', solver = 'adam'
`nusvc`	coef0 = 0.01, gamma = 0.01, kernel = sigmoid
`randomf`	bootstrap = False, criterion = entropy, max_features= log2, n_estimators = 1000
`svc`	C = 100, gamma = 0.001, kernel = rbf