nature research

https://doi.org/10.1038/s42256-020-0222-1

Supplementary information

Protein function prediction is improved by creating synthetic feature samples with generative adversarial networks

In the format provided by the authors and unedited

Supporting File 1

Protein function prediction is improved by creating synthetic feature samples with generative adversarial networks

Cen Wan [1,2] and David T. Jones [1,2]

[1]Biomedical Data Science Laboratory, The Francis Crick Institute, London, UK
[2]Department of Computer Science, University College London, London, UK

The supporting information includes

The 2D visualisation of real and synthetic protein feature samples. a-t, The *t*-SNE transformed 2D visualisation of real and synthetic protein feature samples obtained during different training epochs of FFPred-GAN by using positive protein feature samples for molecular function term GO:0000981 (a-e) and cellular component term GO:0000785 (f-j); and by using the negative protein feature samples for molecular function term GO:0046872 (k-o) and cellular component term GO:0016020 (p-t). (Fig. S1).

The scatter-plots about the MCC and AUROC values obtained by SMOTE augmented and original samples. a-f, The MCC and AUROC values obtained by the SMOTE augmented training samples and the benchmark combination of real protein samples for predicting three domains of GO terms by using SVM (a-e) and RF (f) classification algorithms. (Fig. S2).

The results of CTST on synthetic and real testing samples. a-c, The Leave One Out Cross-Validation (LOOCV) accuracy of CTST obtained by the real and synthetic testing protein samples for GO terms from Biological Process (BP) (a), Molecular Function (MF) (b) and Cellular Component (CC) (c) domains. (Fig. S3).

Summary of average ranks of 27 modelling strategies obtained by predicting BP, MF and CC domains of GO terms according to MCC and AUROC values (Tables S1 – S2).

Summary of F_{max} scores obtained by the FFPred-GAN augmented training samples, the SMOTE augmented training samples and the original training samples on the CAFA 3 targets (Table S3).

Summary of computational time cost, training sample size and their correlation coefficient values (Table S4).

Summary of protein sequence-derived biophysical feature groups information (Table S5).

Summary of hyper-parameters of different classification algorithms for conducting the 5-fold cross validation-based grid-search by using the Scikit-learn (Table S6).

Summary of hyper-parameters of logistic regression for conducting grid-search by using the Scikit-learn (Table S7).

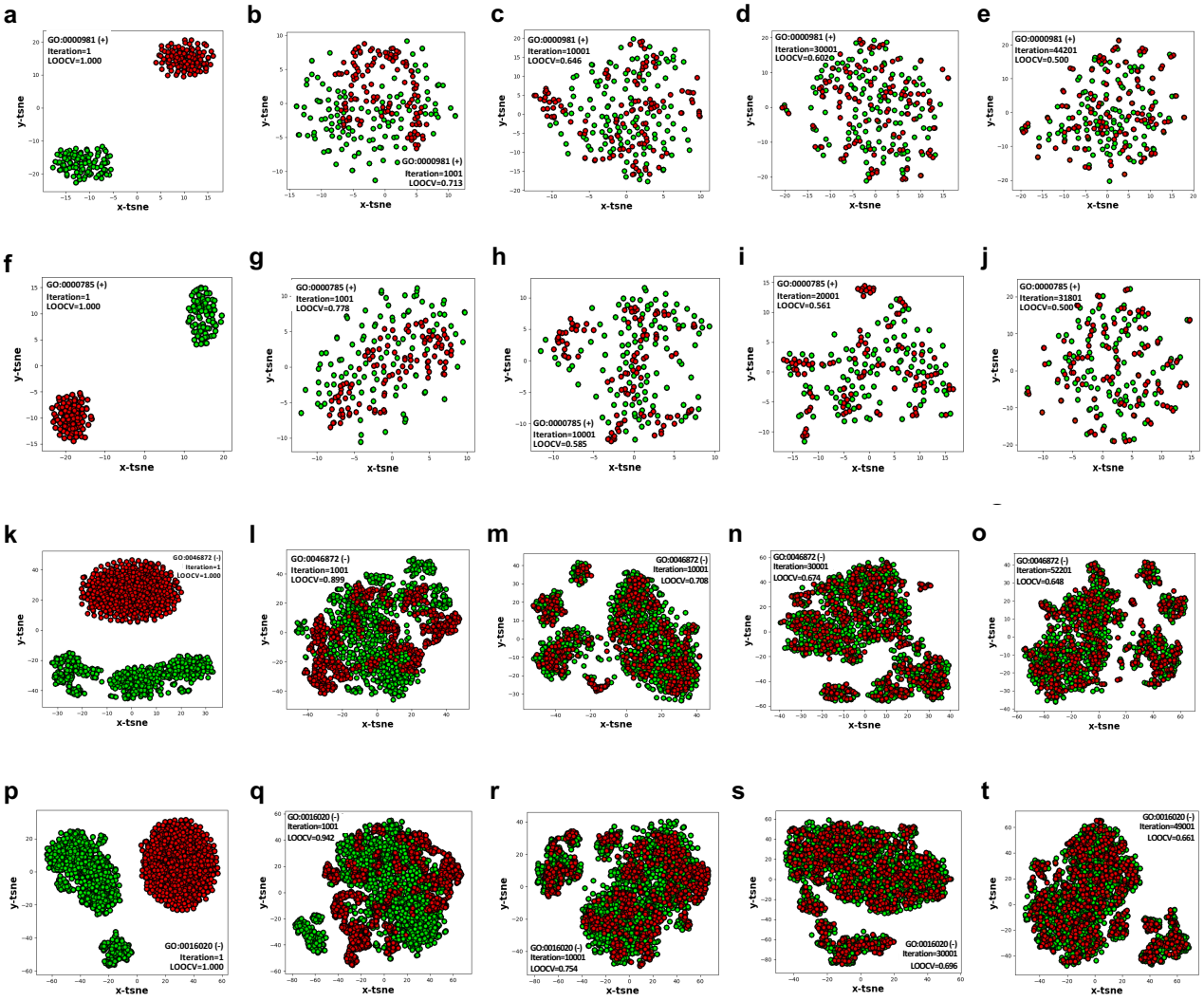

Fig. S1. The 2D visualisation of real and synthetic protein feature samples. a-t, The *t*-SNE transformed 2D visualisation of real and synthetic protein feature samples obtained during different training epochs of FFPred-GAN by using positive protein feature samples for molecular function term GO:0000981 (**a-e**) and cellular component term GO:0000785 (**f-j**); and by using the negative protein feature samples for molecular function term GO:0046872 (**k-o**) and cellular component term GO:0016020 (**p-t**).

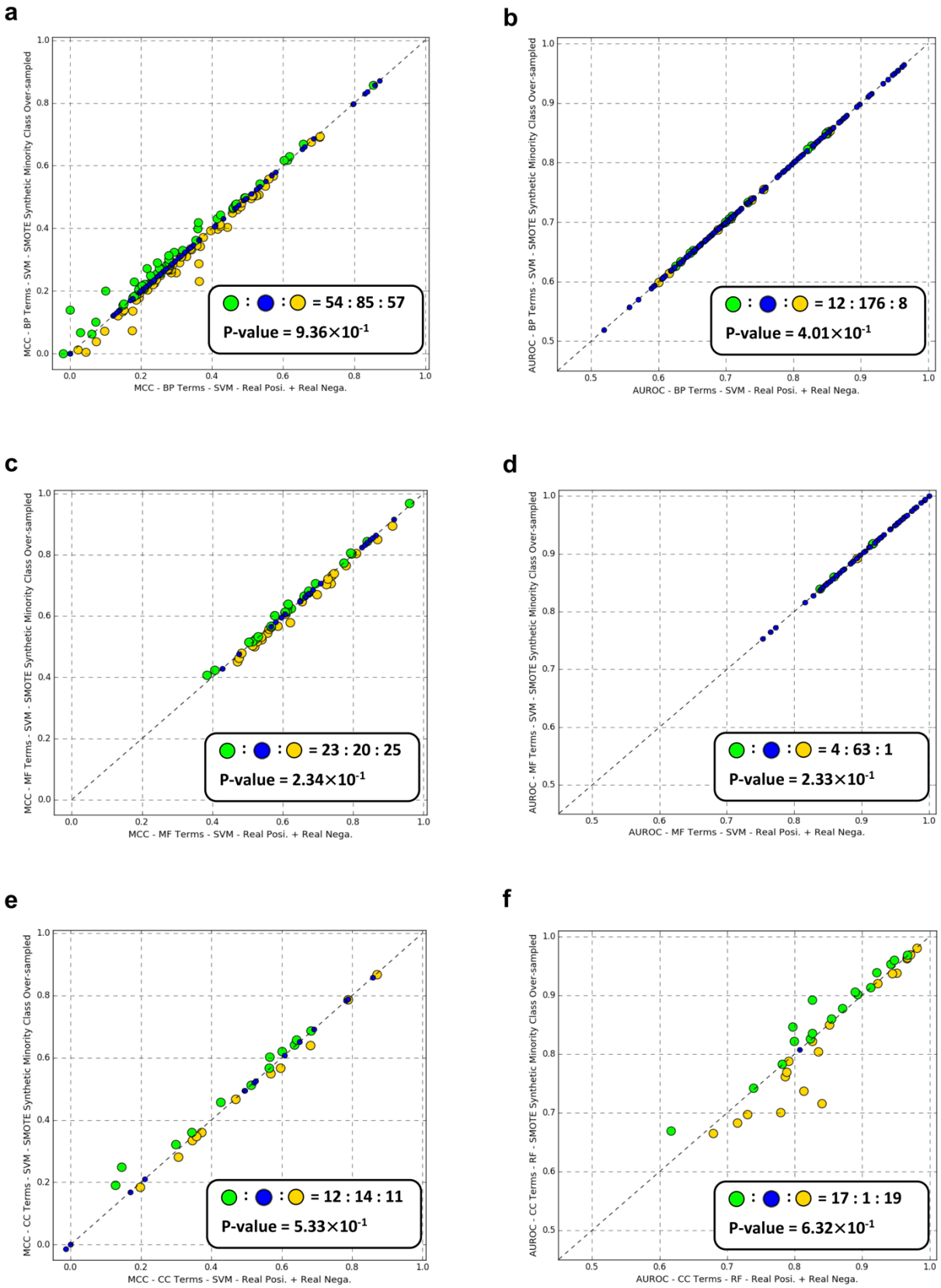

Fig. S2. The scatter-plots about the MCC and AUROC values obtained by SMOTE augmented and original samples. a-f, The MCC and AUROC values obtained by the SMOTE augmented training samples and the benchmark combination of real protein samples for predicting three domains of GO terms by using SVM (**a-e**) and RF (**f**) classification algorithms.

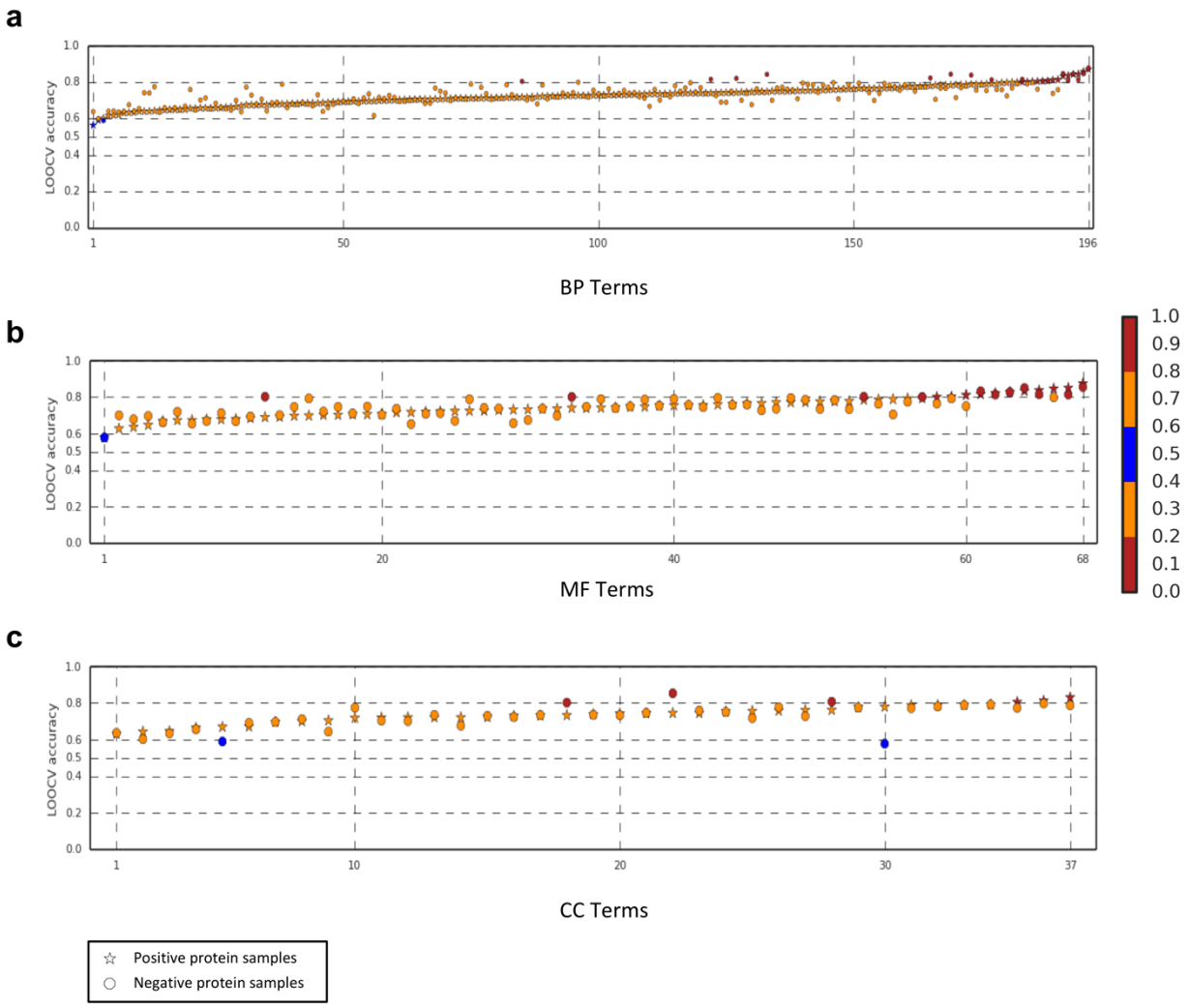

Fig. S3. The results of CTST on synthetic and real testing samples. a-c, The Leave One Out Cross-Validation (LOOCV) accuracy of CTST obtained by the real and synthetic testing protein samples for GO terms from Biological Process (BP) (**a**), Molecular Function (MF) (**b**) and Cellular Component (CC) (**c**) domains.

Table S1. Summary of average ranks of 27 modelling strategies obtained by predicting BP, MF and CC domains of GO terms according to MCC values.

Modelling Strategy	Ave. Rank	Modelling Strategy	Ave. Rank	Modelling Strategy	Ave. Rank
Biological Process (BP)		Molecular Function (MF)		Cellular Component (CC)	
SVM – Syn. Posi. & Real Posi. & Real Nega.	5.88	SVM – Syn. Posi. & Real Posi. & Real Nega.	3.82	SVM – Syn. Posi. & Syn. Nega. & Real Posi. & Real Nega.	4.95
SVM – Syn. Posi. & Syn. Nega. & Real Posi. & Real Nega.	5.95	SVM – Syn. Posi. & Syn. Nega. & Real Posi. & Real Nega.	4.13	SVM – Syn. Posi. & Real Posi. & Real Nega.	5.30
RF – Real Posi. & Real Nega. *	7.62	SVM – Real Posi. & Real Nega. *	4.60	SVM – Real Posi. & Real Nega. *	7.53
SVM – Real Posi. & Real Nega *.	7.92	SVM – Syn. Nega. & Real Posi. & Real Nega.	4.97	RF – Real Posi. & Real Nega. *	8.42
RF – Syn. Posi. & Syn. Nega. & Real Posi. & Real Nega.	8.29	RF – Real Posi. & Real Nega. *	7.48	RF – Syn. Nega. & Real Posi. & Real Nega.	8.69
RF – Syn. Nega. & Real Posi. & Real Nega.	8.85	RF – Syn. Nega. & Real Posi. & Real Nega.	7.69	RF – Syn. Posi. & Syn. Nega. & Real Posi. & Real Nega.	8.73
KNN – Syn. Posi. & Syn. Nega. & Real Posi. & Real Nega.	8.91	RF – Syn. Posi. & Real Posi. & Real Nega.	7.86	RF – Syn. Posi. & Real Posi. & Real Nega.	8.92
KNN – Syn. Posi. & Real Posi. & Real Nega.	9.04	RF – Syn. Posi. & Syn. Nega. & Real Posi. & Real Nega.	8.46	SVM – Syn. Nega. & Real Posi. & Real Nega.	9.09
KNN – Syn. Nega. & Real Posi. & Real Nega.	9.28	KNN – Syn. Posi. & Syn. Nega. & Real Posi. & Real Nega.	9.32	KNN – Syn. Posi. & Syn. Nega. & Real Posi. & Real Nega.	9.24
SVM – Syn. Nega. & Real Posi. & Real Nega.	9.86	KNN – Syn. Nega. & Real Posi. & Real Nega.	9.45	KNN – Syn. Nega. & Real Posi. & Real Nega.	9.43
KNN – Real Posi. & Real Nega. *	10.27	KNN – Syn. Posi. & Real Posi. & Real Nega.	10.77	KNN – Real Posi. & Real Nega. *	9.69
RF – Syn. Posi. & Real Posi. & Real Nega.	10.66	KNN – Real Posi. & Real Nega. *	11.15	KNN – Syn. Posi. & Real Posi. & Real Nega.	9.96
KNN - Syn. Posi. & Real Nega.	12.41	KNN - Syn. Posi. & Real Nega.	13.09	KNN – Syn. Nega. & Real Posi.	12.05
KNN – Syn. Nega. & Real Posi.	12.93	KNN – Syn. Nega. & Real Posi.	14.07	KNN – Syn. Posi. & Syn. Nega. & Real Posi.	12.30
KNN – Syn. Posi. & Syn. Nega. & Real Posi.	13.14	KNN – Syn. Posi. & Syn. Nega. & Real Posi.	14.71	SVM – Syn. Posi. & Syn. Nega.	15.23
KNN – Syn. Posi. & Syn. Nega. & Real Nega.	13.41	KNN – Syn. Posi. & Syn. Nega. & Real Nega.	14.78	KNN - Syn. Posi. & Real Nega.	15.58
KNN – Syn. Posi. & Syn. Nega.	14.85	SVM – Syn. Posi. & Syn. Nega. & Real Nega.	15.90	KNN – Syn. Posi & Syn. Nega.	16.05
SVM – Syn. Posi. & Syn. Nega.	15.95	KNN – Syn. Posi. & Syn. Nega.	16.65	KNN – Syn. Posi. & Syn. Nega. & Real Nega.	16.09
SVM – Syn. Posi. & Syn. Nega. & Real Posi.	17.51	SVM - Syn. Posi. & Real Nega.	17.26	SVM – Syn. Posi. & Syn. Nega. & Real Posi.	16.86
SVM – Syn. Nega. & Real Posi.	18.33	SVM – Syn. Posi. & Syn. Nega.	17.64	SVM – Syn. Nega. & Real Posi.	17.92
SVM – Syn. Posi. & Syn. Nega. & Real Nega.	18.65	SVM – Syn. Posi. & Syn. Nega. & Real Posi.	20.49	SVM – Syn. Posi. & Syn. Nega. & Real Nega.	18.81
SVM - Syn. Posi. & Real Nega.	19.05	SVM – Syn. Nega. & Real Posi.	21.15	SVM - Syn. Posi. & Real Nega.	19.58
RF – Syn. Posi. & Syn. Nega.	21.13	RF – Syn. Posi. & Syn. Nega.	21.43	RF – Syn. Posi. & Syn. Nega.	20.65
RF – Syn. Posi. & Syn. Nega. & Real Posi.	23.88	RF – Syn. Posi. & Syn. Nega. & Real Posi.	24.23	RF – Syn. Posi. & Syn. Nega. & Real Posi.	22.91
RF – Syn. Nega. & Real Posi.	24.13	RF – Syn. Nega. & Real Posi.	24.76	RF – Syn. Nega. & Real Posi.	24.15
RF – Syn. Posi. & Syn. Nega. & Real Nega.	25.04	RF – Syn. Posi. & Syn. Nega. & Real Nega.	25.99	RF – Syn. Posi. & Syn. Nega. & Real Nega.	24.76
RF - Syn. Posi. & Real Nega.	25.09	RF - Syn. Posi. & Real Nega.	26.16	RF - Syn. Posi. & Real Nega.	25.11

Table S2. Summary of average ranks of 27 modelling strategies obtained by predicting BP, MF and CC domains of GO terms according to AUROC values.

Modelling Strategy	Ave. Rank	Modelling Strategy	Ave. Rank	Modelling Strategy	Ave. Rank
Biological Process (BP)		Molecular Function (MF)		Cellular Component (CC)	
SVM – Syn. Posi. & Real Posi. & Real Nega.	4.84	SVM – Syn. Posi. & Real Posi. & Real Nega.	4.40	RF – Syn. Posi. & Syn. Nega. & Real Posi. & Real Nega.	4.70
RF – Syn. Posi. & Syn. Nega. & Real Posi. & Real Nega.	5.36	SVM – Syn. Posi. & Syn. Nega. & Real Posi. & Real Nega.	4.81	SVM – Syn. Posi. & Syn. Nega. & Real Posi. & Real Nega.	4.84
SVM – Real Posi. & Real Nega. *	5.66	SVM – Real Posi. & Real Nega. *	4.83	SVM – Syn. Posi. & Real Posi. & Real Nega.	4.92
SVM – Syn. Posi. & Syn. Nega. & Real Posi. & Real Nega.	5.67	RF – Real Posi. & Real Nega. *	5.01	SVM – Real Posi. & Real Nega. *	5.03
RF – Real Posi. & Real Nega. *	6.34	RF – Syn. Posi. & Syn. Nega. & Real Posi. & Real Nega.	5.51	SVM – Syn. Nega. & Real Posi. & Real Nega.	5.51
RF – Syn. Nega. & Real Posi. & Real Nega.	6.63	RF – Syn. Posi. & Real Posi. & Real Nega.	5.63	RF – Real Posi. & Real Nega. *	6.14
SVM – Syn. Nega. & Real Posi. & Real Nega.	7.26	SVM – Syn. Nega. & Real Posi. & Real Nega.	5.64	SVM – Syn. Posi. & Syn. Nega. & Real Posi.	7.30
RF – Syn. Posi. & Real Posi. & Real Nega.	8.49	RF – Syn. Nega. & Real Posi. & Real Nega.	5.77	RF – Syn. Posi. & Real Posi. & Real Nega.	7.57
SVM – Syn. Posi. & Syn. Nega. & Real Posi.	8.59	SVM – Syn. Posi. & Syn. Nega. & Real Posi.	8.93	RF – Syn. Nega. & Real Posi. & Real Nega.	7.59
KNN – Real Posi. & Real Nega. *	10.62	SVM – Syn. Posi. & Syn. Nega. & Real Nega.	10.60	SVM – Syn. Nega. & Real Posi.	10.88
SVM – Syn. Nega. & Real Posi.	11.19	SVM – Syn. Nega. & Real Posi.	10.74	SVM – Syn. Posi. & Syn. Nega.	12.34
SVM – Syn. Posi. & Syn. Nega.	12.71	SVM - Syn. Posi. & Real Nega.	11.01	SVM – Syn. Posi. & Syn. Nega. & Real Nega.	12.43
SVM – Syn. Posi. & Syn. Nega. & Real Nega.	12.92	SVM – Syn. Posi. & Syn. Nega.	12.56	SVM - Syn. Posi. & Real Nega.	12.68
SVM - Syn. Posi. & Real Nega.	13.34	KNN – Real Posi. & Real Nega. *	13.00	KNN – Real Posi. & Real Nega. *	13.18
KNN – Syn. Posi. & Real Posi. & Real Nega.	13.78	KNN – Syn. Posi. & Real Posi. & Real Nega.	16.82	KNN – Syn. Posi. & Real Posi. & Real Nega.	14.61
KNN – Syn. Posi. & Syn. Nega. & Real Posi. & Real Nega.	16.32	KNN – Syn. Posi. & Syn. Nega. & Real Posi. & Real Nega.	17.86	KNN – Syn. Posi. & Syn. Nega. & Real Posi.	18.19
KNN – Syn. Nega. & Real Posi. & Real Nega.	17.28	RF – Syn. Posi. & Syn. Nega. & Real Posi.	18.03	RF – Syn. Posi. & Syn. Nega. & Real Posi.	18.27
KNN – Syn. Posi. & Syn. Nega. & Real Posi.	17.40	KNN – Syn. Nega. & Real Posi. & Real Nega.	18.58	KNN – Syn. Posi. & Syn. Nega. & Real Posi. & Real Nega.	18.28
KNN – Syn. Nega. & Real Posi.	18.48	RF – Syn. Posi. & Syn. Nega.	18.74	KNN – Syn. Nega. & Real Posi. & Real Nega.	18.36
RF – Syn. Posi. & Syn. Nega.	18.66	KNN – Syn. Posi. & Syn. Nega. & Real Posi.	19.15	RF – Syn. Posi. & Syn. Nega.	18.73
RF – Syn. Posi. & Syn. Nega. & Real Posi.	20.39	KNN – Syn. Nega. & Real Posi.	19.58	KNN – Syn. Nega. & Real Posi.	18.92
KNN - Syn. Posi. & Real Nega.	20.68	RF – Syn. Nega. & Real Posi.	21.51	RF – Syn. Nega. & Real Posi.	19.84
RF – Syn. Nega. & Real Posi.	20.93	KNN - Syn. Posi. & Real Nega.	22.02	KNN - Syn. Posi. & Real Nega.	21.77
KNN – Syn. Posi. & Syn. Nega.	22.08	KNN – Syn. Posi. & Syn. Nega.	23.43	KNN – Syn. Posi. & Syn. Nega.	22.89
KNN – Syn. Posi. & Syn. Nega. & Real Nega.	22.36	KNN – Syn. Posi. & Syn. Nega. & Real Nega.	23.49	KNN – Syn. Posi. & Syn. Nega. & Real Nega.	23.45
RF – Syn. Posi. & Syn. Nega. & Real Nega.	24.64	RF – Syn. Posi. & Syn. Nega. & Real Nega.	24.62	RF – Syn. Posi. & Syn. Nega. & Real Nega.	24.65
RF - Syn. Posi. & Real Nega.	25.37	RF - Syn. Posi. & Real Nega.	25.72	RF - Syn. Posi. & Real Nega.	24.95

Table S3. Summary of F_{max} scores obtained by the FFPred-GAN augmented training samples, the SMOTE augmented training samples and the original training samples on the CAFA 3 targets.

F_{max}	BP	MF	CC
FFPred-GAN augmented training samples	**0.325**	**0.385**	**0.629**
SMOTE augmented training samples	0.307	0.383	0.606
Original training samples	0.308	0.381	0.605

Table S4. Summary of computational time cost, training sample size and their correlation coefficient values.

Training Samples	Computational time cost										Training sample size					Correlation Coefficient
	Minimum		1st Quartile		Median		3rd Quartile		Maximum		Minimum	1st Quartile	Median	3rd Quartile	Maximum	
	Second	Hour	Second	Hour	Second	Hour	Second	Hour	Second	Hour						
Positive-BP	2073.0	0.6	8079.4	2.2	20038.6	5.6	34556.1	9.6	139210.2	38.7	105.0	143.8	226.0	392.5	1939.0	0.521
Positive-MF	3105.7	0.9	10887.2	3.0	24187.2	6.7	37066.0	10.3	153056.2	42.5	109.0	139.5	234.0	444.0	2901.0	0.379
Positive-CC	3520.0	1.0	12390.5	3.4	20973.8	5.8	43664.0	12.1	186317.1	51.8	106.0	140.0	238.0	584.0	3200.0	0.900
Negative-BP	2991.8	0.8	17456.5	4.8	28624.6	8.0	51873.3	14.4	166180.2	46.2	436.0	795.0	873.0	915.0	960.0	0.321
Negative-MF	3615.1	1.0	18286.3	5.1	29401.6	8.2	50182.8	13.9	121113.6	33.6	525.0	805.0	875.0	916.3	974.0	0.349
Negative-CC	36847.1	10.2	75593.5	21.0	113777.8	31.6	132606.9	36.8	171204.0	47.6	579.0	1548.0	1680.0	1891.0	2020.0	0.140

Table S5. Summary of protein sequence-derived biophysical feature groups information.

Feature names	Dimensions
Secondary structure	49
Transmembrane segments	15
Amino acid composition	20
Intrinsically disordered regions	18
Signal peptides	8
Subcellular localization	12
Sequence features	17
PEST regions	12
Low complexity regions	12
Coiled coils	12
N-linked glycosylation sites	11
O-GalNAc-glycosylation sites	22
Phosphorylation sites	50

Table S6. Summary of hyper-parameters of different classification algorithms for conducting the 5-fold cross validation-based grid-search by using the Scikit-learn.

Classification algorithms	**Hyper-parameters**	
Support Vector Machine	*Kernel* = linear	C: [1, 10, 0.1, 100, 0.01, 1000, 0.001, 1e4, 1e-4]
	Kernel = rbf	C: [1, 10, 0.1, 100, 0.01, 1000, 0.001, 1e4, 1e-4]
		γ: [1, 0.5, 3, 0.2, 10, 0.1, 0.03, 0.01, 0.001, 1e-4]
k-Nearest Neighbours	*n_neighbors*: [1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29]	
	weights: ['uniform', 'distance']	
Random Forests	*n_estimators*: [10, 20, 30, 40, 50]	
	max_depth: [1, 2, 3, None]	
	max_features: [1, "auto", "log2", None],	
	criterion: ["gini", "entropy"]	

Table S7. Summary of hyper-parameters of logistic regression for conducting grid-search by using the Scikit-learn.

Solver = lbfgs	C: [1, 10, 0.1, 100, 0.01, 1000, 0.001, 1e4, 1e-4]	Penalty: ['l2', 'none']	class_weight:['balanced',None]
Solver = newton-cg	C: [1, 10, 0.1, 100, 0.01, 1000, 0.001, 1e4, 1e-4]	Penalty: ['l2', 'none']	class_weight:['balanced',None]
Solver = liblinear	C: [1, 10, 0.1, 100, 0.01, 1000, 0.001, 1e4, 1e-4]	Penalty: ['l1', l2]	class_weight:['balanced',None]
Solver = sag	C: [1, 10, 0.1, 100, 0.01, 1000, 0.001, 1e4, 1e-4]	Penalty: ['l2', 'none']	class_weight:['balanced',None]
Solver = saga	C: [1, 10, 0.1, 100, 0.01, 1000, 0.001, 1e4, 1e-4]	Penalty: [‘l1’, 'l2', 'none']	class_weight:['balanced',None]