Supplementary information

QuoteTarget: A sequence-based transformer protein language model to identify potentially druggable protein targets

Jiaxiao Chen[1], Zhonghui Gu[2], Youjun Xu[3], Minghua Deng[1,4,5], Luhua Lai[1,2,6,7], Jianfeng Pei[1,7*]

[1]Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, 100871, China

[2]Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, 100871, China

[3]Infinite Intelligence Pharma, Beijing 100083, China

[4]School of Mathematical Sciences, Peking University, Beijing 100871, China

[5]Center for Statistical Science, Peking University, Beijing 100871, China

[6]BNLMS, College of Chemistry and Molecular Engineering, Peking University, Beijing, 100871, China

[7]Research Unit of Drug Design Method, Chinese Academy of Medical Sciences (2021RU014), Beijing 100871, China

***Correspondence:**

Jianfeng Pei, Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, 100871, China.

Email: jfpei@pku.edu.cn

Supplementary Table 1: Classification results on all drug target dataset with the same number of positive and negative samples

	Dataset	Acc	Precision	F1	Mcc	Sensitivity	Specificity
5-fold cross-validation	All-Pfam	0.95±0.01	0.96±0.01	0.94±0.01	0.89±0.01	0.93±0.01	0.97±0.01
	All-Evalue0.001	0.95±0.00	0.97±0.01	0.94±0.00	0.89±0.00	0.92±0.01	0.97±0.01
	All-Evalue1	0.95±0.00	0.97±0.01	0.95±0.00	0.90±0.01	0.93±0.01	0.97±0.01
	All-Evalue10	0.94±0.00	0.96±0.01	0.95±0.00	0.89±0.00	0.94±0.01	0.95±0.01
External Test	All-Pfam	0.94±0.00	0.96±0.01	0.94±0.00	0.88±0.01	0.92±0.01	0.97±0.01
	All-Evalue0.001	0.94±0.00	0.97±0.00	0.94±0.00	0.89±0.00	0.92±0.00	0.97±0.00
	All-Evalue1	0.94±0.00	0.95±0.01	0.94±0.00	0.88±0.01	0.92±0.00	0.95±0.01
	All-Evalue10	0.94±0.00	0.95±0.01	0.95±0.00	0.89±0.00	0.94±0.01	0.94±0.01

The values in the table are the mean values and standard deviation are from the 5-fold cross-validation.

Abbreviations: Acc, accuracy; F1, F1-score; Mcc, Matthews correlation coefficient.

Supplementary Table 2: Classification results on FDA approved drug target dataset with the same number of positive and negative samples

	Dataset	Acc	Precision	F1	Mcc	Sensitivity	Specificity
5-fold cross-validation	App-Pfam	0.90±0.01	0.92±0.01	0.91±0.00	0.81±0.01	0.89±0.01	0.92±0.02
	App-Evalue0.001	0.91±0.01	0.93±0.01	0.90±0.01	0.81±0.01	0.88±0.02	0.93±0.01
	App-Evalue1	0.91±0.01	0.91±0.01	0.91±0.01	0.82±0.02	0.91±0.01	0.91±0.01
	App-Evalue10	0.90±0.01	0.90±0.02	0.91±0.01	0.81±0.02	0.92±0.00	0.89±0.02
External Test	App-Pfam	0.91±0.01	0.93±0.01	0.91±0.01	0.82±0.01	0.89±0.02	0.93±0.01
	App-Evalue0.001	0.89±0.00	0.91±0.01	0.89±0.00	0.79±0.01	0.88±0.01	0.91±0.02
	App-Evalue1	0.91±0.01	0.93±0.01	0.91±0.01	0.82±0.01	0.89±0.01	0.93±0.01
	App-Evalue10	0.90±0.01	0.91±0.01	0.90±0.01	0.80±0.01	0.90±0.01	0.90±0.01

The values in the table are the mean values and standard deviation are from the 5-fold cross-validation.

Supplementary Table 3: Classification results with different classifiers on the All-Pfam dataset with ablation of ESM1b pretraining

	Classifier	Acc	Precision	F1	Mcc	Sensitivity	Specificity
Cross-validation	DT	0.55±0.03	0.42±0.02	0.42±0.04	0.05±0.02	0.44±0.09	0.62±0.10
	GNB	0.55±0.03	0.42±0.03	0.46±0.04	0.07±0.06	0.50±0.08	0.57±0.07
	K-NEST	0.59±0.02	0.44±0.04	0.36±0.05	0.07±0.05	0.30±0.06	0.76±0.05
	LR	0.62±0.01	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	1.00±0.00
	RF	0.56±0.04	0.43±0.05	0.38±0.11	0.06±0.07	0.40±0.21	0.66±0.17
	SVM	0.58±0.04	0.24±0.20	0.17±0.19	0.01±0.05	0.18±0.25	0.83±0.22
	GCN	0.72±0.01	0.70±0.02	0.57±0.02	0.39±0.02	0.48±0.03	0.87±0.01
External Test	DT	0.55±0.03	0.40±0.01	0.41±0.04	0.05±0.01	0.43±0.10	0.62±0.10
	GNB	0.53±0.02	0.39±0.02	0.43±0.04	0.04±0.03	0.48±0.09	0.56±0.07
	K-NEST	0.58±0.02	0.42±0.02	0.35±0.04	0.06±0.03	0.30±0.06	0.75±0.05
	LR	0.63±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	1.00±0.00
	RF	0.55±0.04	0.40±0.02	0.36±0.11	0.04±0.06	0.39±0.20	0.65±0.17
	SVM	0.59±0.05	0.22±0.19	0.17±0.18	0.01±0.03	0.18±0.25	0.83±0.23
	GCN	0.72±0.00	0.66±0.02	0.56±0.01	0.37±0.00	0.49±0.03	0.85±0.02

The values in the table are the mean values and standard deviations are from the 5-fold cross-validation.

Abbreviations: DT, Decision Tree; GNB, Gaussion Naive Bayes; K-NEST, K-nearest Neighbor; LR, Logistic Regression; RF, Random Forest; SVM, Support Vector Machine; GCN, Graph Convolutional Network.

Supplementary Table 4: Classification results with different classifiers on the App-Pfam dataset with ablation of ESM1b pretraining

	Classifier	Acc	Precision	F1	Mcc	Sensitivity	Specificity
Cross-validation	DT	0.65±0.03	0.23±0.02	0.25±0.04	0.03±0.03	0.28±0.08	0.75±0.06
	GNB	0.59±0.07	0.25±0.03	0.31±0.04	0.07±0.04	0.46±0.14	0.62±0.12
	K-NEST	0.77±0.01	0.32±0.05	0.15±0.06	0.07±0.03	0.10±0.05	0.94±0.03
	LR	0.79±0.01	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	1.00±0.00
	RF	0.76±0.03	0.22±0.10	0.11±0.07	0.02±0.06	0.08±0.06	0.94±0.04
	SVM	0.79±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	1.00±0.00
	GCN	0.81±0.01	0.62±0.02	0.29±0.04	0.26±0.02	0.19±0.04	0.97±0.01
External Test	DT	0.66±0.03	0.25±0.02	0.27±0.05	0.05±0.04	0.30±0.10	0.75±0.06
	GNB	0.60±0.07	0.27±0.04	0.34±0.04	0.10±0.05	0.49±0.13	0.63±0.12
	K-NEST	0.77±0.01	0.37±0.06	0.17±0.07	0.10±0.04	0.12±0.06	0.94±0.03
	LR	0.79±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	1.00±0.00
	RF	0.76±0.02	0.27±0.07	0.11±0.08	0.03±0.05	0.08±0.07	0.94±0.04
	SVM	0.79±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	1.00±0.00
	GCN	0.80±0.00	0.62±0.02	0.29±0.06	0.26±0.03	0.19±0.05	0.97±0.01

The values in the table are the mean values and standard deviation are from the 5-fold cross-validation.

Supplementary Table 5: Classification results based on BLAST on datasets of All-Pfam and App-Pfam

Dataset	E-value	Acc	Precision	F1	Mcc	Sensitivity	Specificity
All-Pfam	e-3	0.30	0.81	0.43	-0.39	0.00	0.81
	e-6	0.28	0.84	0.42	-0.35	0.00	0.84
	e-10	0.27	0.86	0.41	-0.33	0.00	0.86
App-Pfam	e-3	0.17	0.78	0.27	-0.43	0.00	0.78
	e-6	0.16	0.82	0.27	-0.39	0.00	0.82
	e-10	0.16	0.84	0.26	-0.37	0.00	0.84

Supplementary Table 6: Classification results of 5-fold cross-validation on the datasets of *H. sapiens* in Figure 5

	Dataset	Acc	Precision	F1	Mcc	Sensitivity	Specificity
Cross-validation	A+B	0.95±0.00	0.97±0.00	0.94±0.01	0.90±0.01	0.92±0.01	0.97±0.00
	B+C	0.95±0.01	0.97±0.01	0.95±0.01	0.90±0.01	0.92±0.01	0.98±0.01
	A+C	0.95±0.01	0.97±0.01	0.95±0.01	0.90±0.01	0.93±0.02	0.97±0.01

The values in the table are the mean values and standard deviation are from the 5-fold cross-validation.

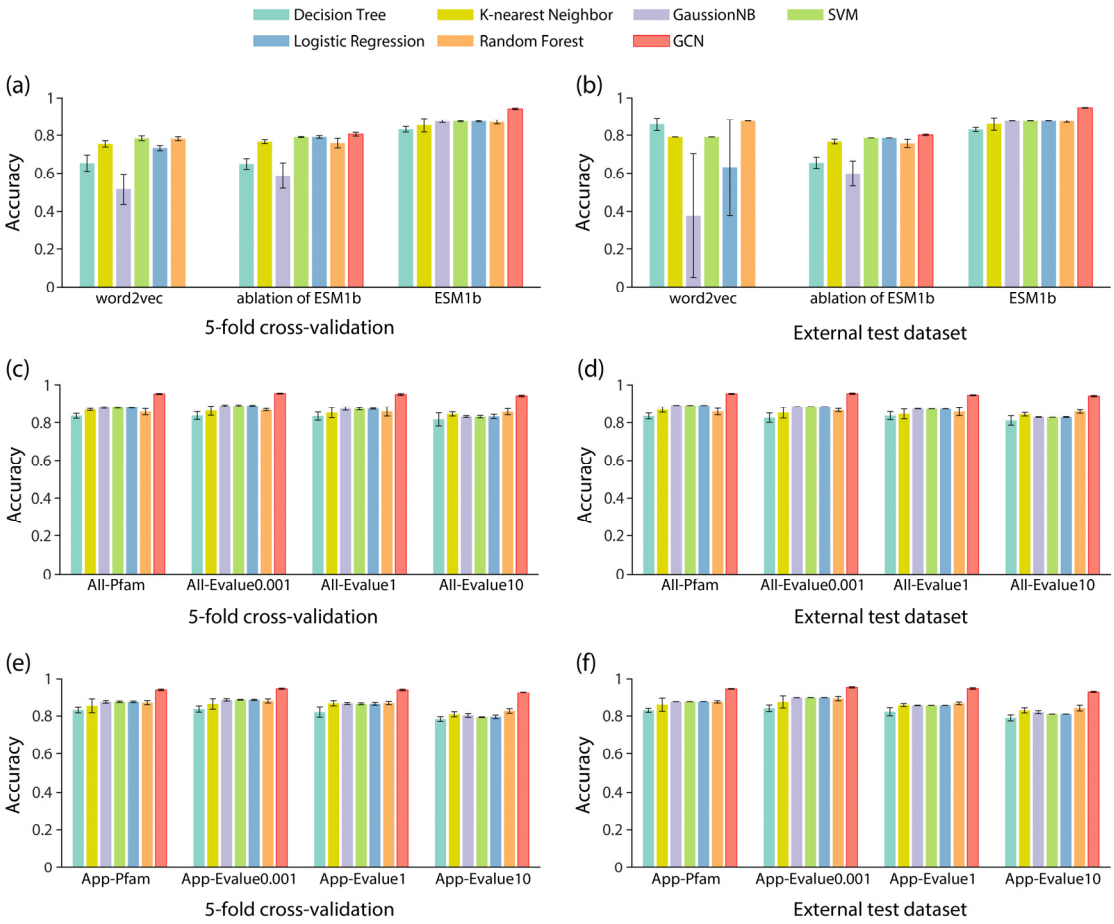
.

Supplementary Table 7: The list of undeveloped drug target proteins in *H. sapiens* from our model

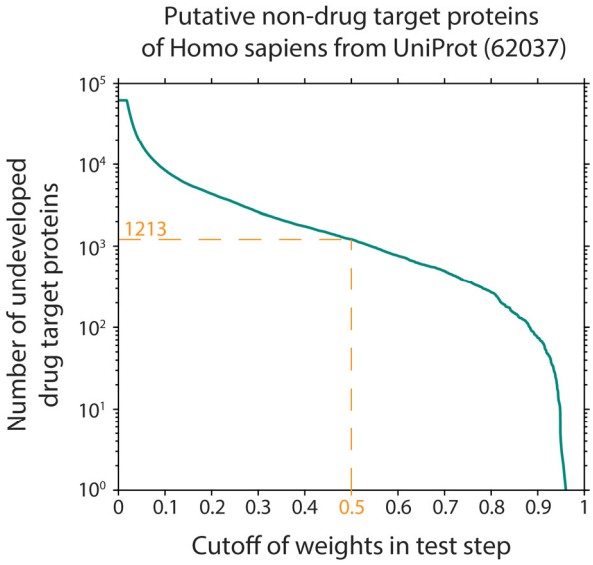
UniProt ID	Weight	UniProt ID	Weight	UniProt ID	Weight	UniProt ID	Weight
P22749	0.974902	P0DPI2	0.880529	Q96MA6	0.709324	P50395	0.566302
P30085	0.974829	P38117	0.879324	Q14390	0.702568	Q96JN8	0.56408
Q14353	0.97479	Q9Y6K8	0.877921	P30040	0.698446	Q9H1P3	0.561815
P00568	0.974782	O95433	0.87548	P36268	0.688868	Q9Y3C7	0.561717
P22061	0.974778	A0A0B4J2D5	0.874937	Q16134	0.678327	Q6ZV89	0.561366
Q8IVH4	0.974714	Q8TB22	0.866978	Q96GG9	0.675535	Q9BZF2	0.557529
O95865	0.974699	P46777	0.866216	O75365	0.660316	Q2Y0W8	0.5571
Q86Y39	0.974695	Q02878	0.861963	Q7KZN9	0.65383	Q15435	0.55285
Q05932	0.974678	Q8NBJ7	0.861302	P60059	0.652227	Q9BX51	0.552098
Q92506	0.974668	P0CG22	0.859744	Q5TA45	0.651076	Q9HBB8	0.55052
Q9Y234	0.974643	P22087	0.856953	O60507	0.649644	Q6P988	0.548566
Q9BTZ2	0.974618	Q9BPX1	0.849202	Q9NY12	0.640061	Q8IXN7	0.548503
P02818	0.974602	A0A075B6L2	0.84816	A6NHQ2	0.63703	Q8WWA0	0.543998
P51161	0.974543	Q13268	0.840222	O00767	0.635873	Q99571	0.543037
Q6NUM9	0.974536	P0DTE2	0.837805	Q9Y5Z7	0.635163	B5MD39	0.539385
Q8WVQ1	0.974515	Q2TBF2	0.836365	Q00005	0.63295	Q86Y79	0.537537
Q6N063	0.973995	Q9NQ11	0.835728	Q9Y2T4	0.631297	O95825	0.537398
P19404	0.973749	A6NGU5	0.83498	A3KN83	0.630856	P63173	0.53619
O94760	0.973668	Q5VTL7	0.832507	Q5TC63	0.628693	P54840	0.534449
Q7Z4W1	0.9734	O00625	0.822298	O60704	0.628103	Q9Y693	0.532006
P50440	0.973379	Q9NPF4	0.803374	Q6NXG1	0.625893	P23511	0.53034
P19440	0.973303	Q9Y237	0.799958	Q9BS92	0.61443	Q9H299	0.530029
P78559	0.973235	A6NEY8	0.796118	Q06203	0.61165	Q9NR22	0.529451
P15502	0.972725	Q5W064	0.792034	P62805	0.609764	Q8NI37	0.52902
P08590	0.949492	Q96AT9	0.784862	P51452	0.609156	Q9BUE6	0.527773
P12829	0.949006	P49326	0.782469	Q9BY49	0.608191	Q9HAT2	0.52768
Q8TDX5	0.947522	Q9Y5T4	0.774123	A5YM72	0.607706	Q6DKI1	0.526922
P60660	0.941082	Q15198	0.771987	Q9Y282	0.606879	Q96PE7	0.526015
P14649	0.940274	Q6UY09	0.767729	Q9Y6R1	0.606444	Q9H6T0	0.517768
Q9BUT1	0.937981	Q2QD12	0.76768	Q92597	0.602533	Q96FJ2	0.517127
Q9H1K1	0.93696	Q9HD20	0.764682	Q9UHN6	0.600543	Q9Y696	0.515069
P07148	0.929724	Q6JVE6	0.763513	Q9UFN0	0.600139	P46783	0.514568
Q6P1N9	0.922464	Q8ND71	0.751415	Q8WUX2	0.595898	Q9UGK3	0.513458
Q13867	0.915804	Q9HAV7	0.748875	Q5HYK3	0.589915	Q6PH85	0.512657
P04632	0.91014	Q9BY44	0.746898	Q8IXL6	0.587731	Q8N131	0.508171
Q8N4T8	0.90646	P36269	0.744304	Q96BS2	0.584984	Q13228	0.507639
P13804	0.906407	Q8NBK3	0.739423	P18124	0.584575	Q9Y215	0.505059
Q9Y639	0.900539	Q8WUR7	0.732628	Q99460	0.58156	Q9BUX1	0.502935
P05976	0.887143	P18077	0.71383	O75175	0.580931	Q9H3J6	0.50277
P35613	0.886707	P0DPI4	0.712282	Q66LE6	0.580389	Q86VZ6	0.501325
Q16698	0.882575	P32969	0.712167	O95445	0.568671	O75063	0.500905

Supplementary Table 8: Hyperparameters of traditional machine learning classification algorithms

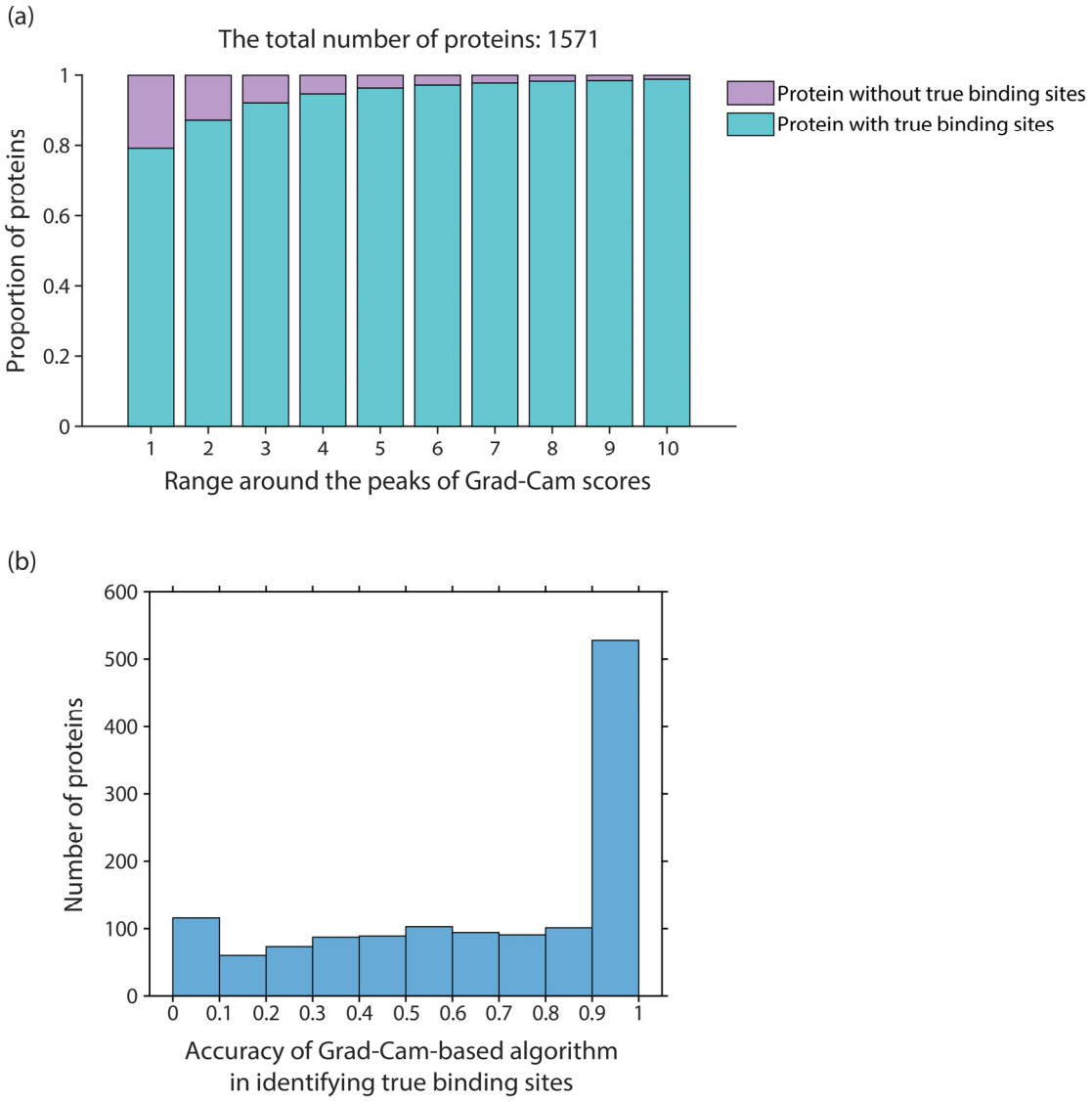
Algorithm	Super parameters
Decision Tree	Criterion = gini; splitter = best; max_depth = None; min_samples_split = 2; min_samples_leaf = 1
K-nearest Neighbor	n_neighbors = 5; weights = uniform
GaussionNB	var_smoothingfloat = 1e-9
SVM	Regularization parameter = 1; kernel = rbf; degree = 3
Logistic Regression	penalty = l2
Random Forest	The number of trees = 100; Criterion = gini; max_depth = None; min_samples_split = 2; min_samples_leaf = 1



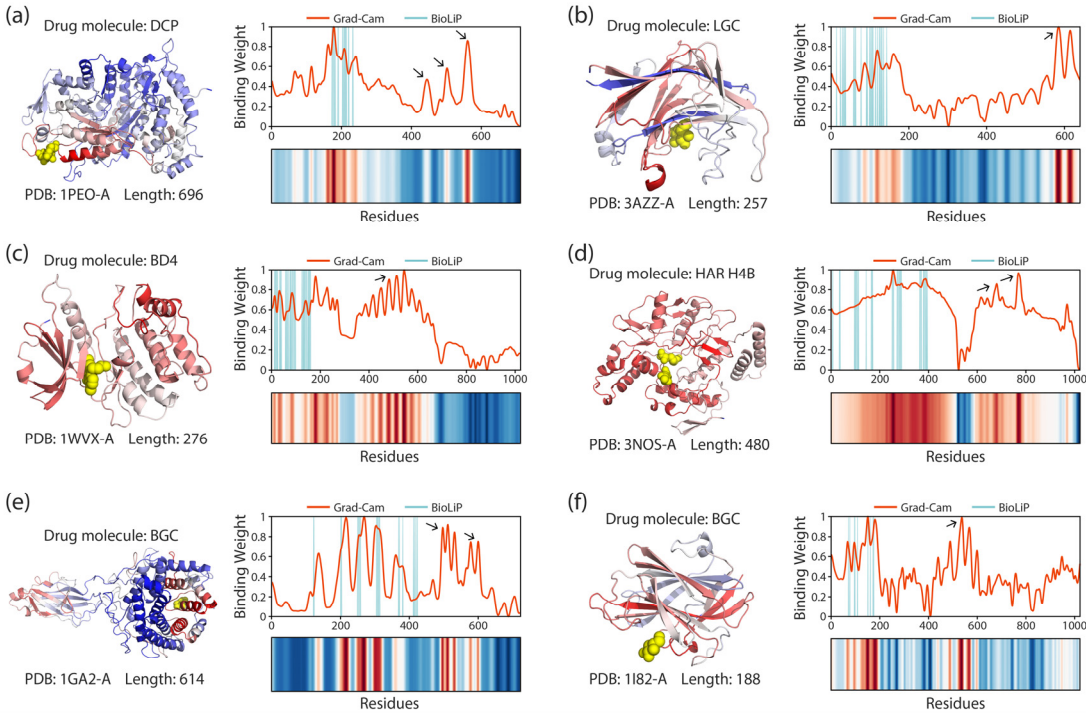
Supplementary Figure 1. Comparison of different protein encoding methods combined with different classification algorithms. (a) Classification results on 5-fold cross-validation with word2vec, randomly initialized ESM1b, and complete ESM1b protein-encoding, respectively. App-Pfam dataset was used for this analysis. (b) Data analyzed as in (a) but showing results from the external test dataset. (c) Classification results on 5-fold cross-validation with the different classification algorithms combined with ESM1b. Datasets of All-Pfam, All-Evalue0.001, All-Evalue1, and All-Evalue10 were used for this analysis. (d) Data analyzed as in (c) but showing results from the external test dataset. (e) Classification results on 5-fold cross-validation with the different classification algorithms combined with the protein-encoding method based on ESM1b. Datasets of App-Pfam, App-Evalue0.001, App-Evalue1, and App-Evalue10 were used for this analysis. (f) Data analyzed as in (e) but showing results from the external test dataset.



Supplementary Figure 2. Identification of undeveloped drug target proteins of Homo sapiens from big dataset. The number of proteins obtained by different cutoffs of weights in test step was shown.



Supplementary Figure 3. Statistical of the results of Grad-Cam scoring on the whole dataset. (a) The horizontal axis represents the residue range around the peak of Grad-Cam scoring. The vertical axis represents the proportion of proteins that have the true binding site within this range around the scoring peaks. (b) To calculate the accuracy for a single protein, the denominator is the number of all true binding sites of the protein, and the numerator is the number of true binding sites successfully identified by the Grad-Cam-based algorithm. The criterion for successful identification was the presence of sites with a Grad-Cam score of more than 0.5 within a range of five amino acids in the vicinity of the binding site.



Supplementary Figure 4. Examples of potential drug molecule binding sites from Grad-Cam. (a) 3D conformation of protein 1PEO-A binding to drug molecule (left panel). Yellow spheres represent drug molecules. Color of the cartoon conformation represents the residue binding weight calculated by Grad-Cam. And comparison of residue binding weights calculated by Grad-Cam with experimentally confirmed binding sites from BioLip (right panel). Gaussian smoothing was performed for the residue binding weight curve. The black arrows indicate the locations with high Grad-Cam scores but no experimentally confirmed binding sites. (b-f) Same as (a) but showing the results of protein 3AZZ-A, 1WVX-A, 3NOS-A, 1GA2-A, 1I82-A, respectively.