Computational Biomedicine
https://doi.org/10.70401/cbm.2026.0016

Supplementary information

PIONEER: A structure-informed graph neural network for PE/PPE protein identification

Heyun Sun, Xudong Guo, Yi Hao, Xiaomin Li, Wenmin Li, Ming Liu, Fuyi Li

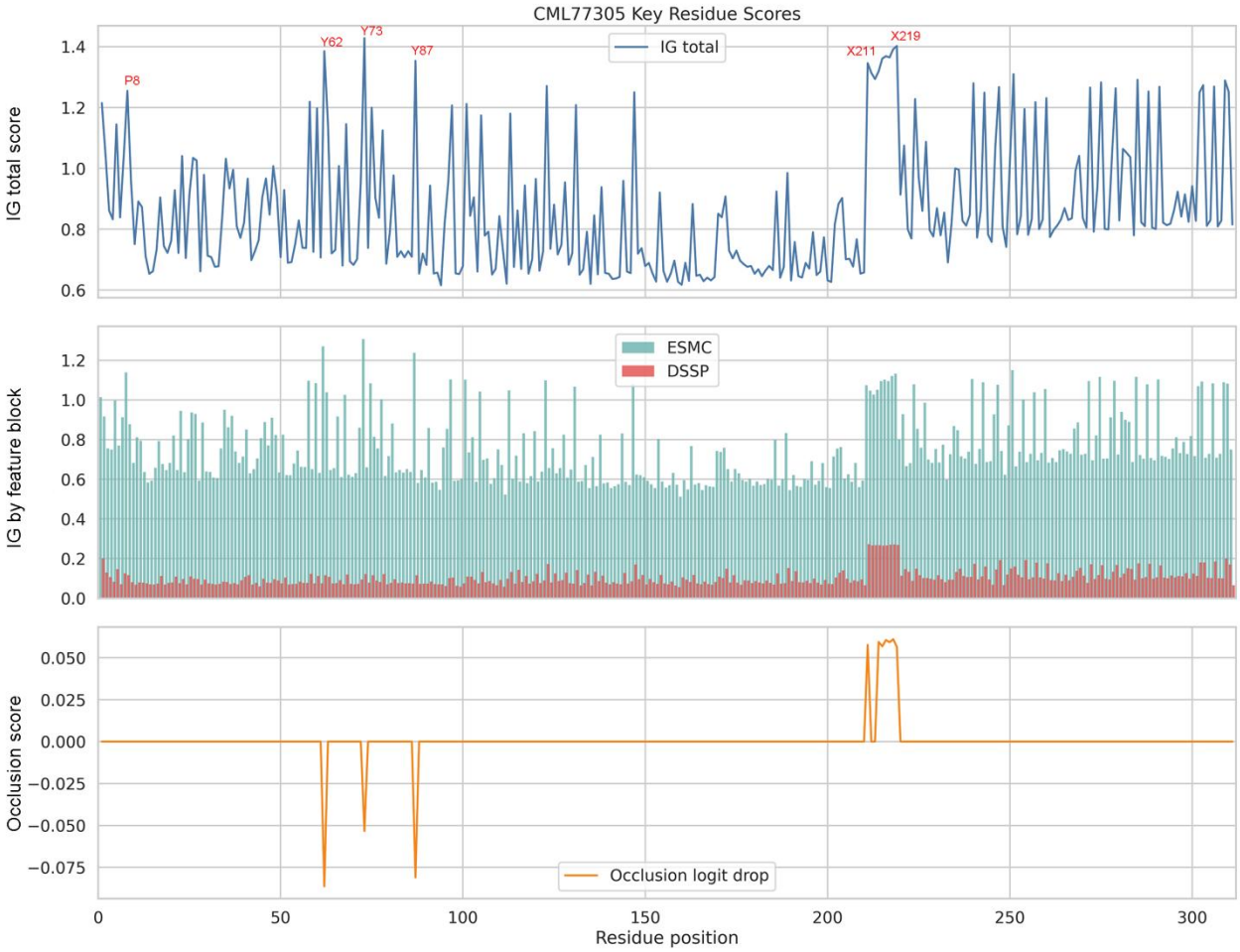

Figure S1. Analysis of Key Residues in CML77305. This study assesses the model's sensitivity to each input feature by computing the gradient of the output with respect to the input, thereby identifying the residues most critical to the current prediction. Furthermore, the influence of each residue on the model's input is evaluated through a residue-masking approach.

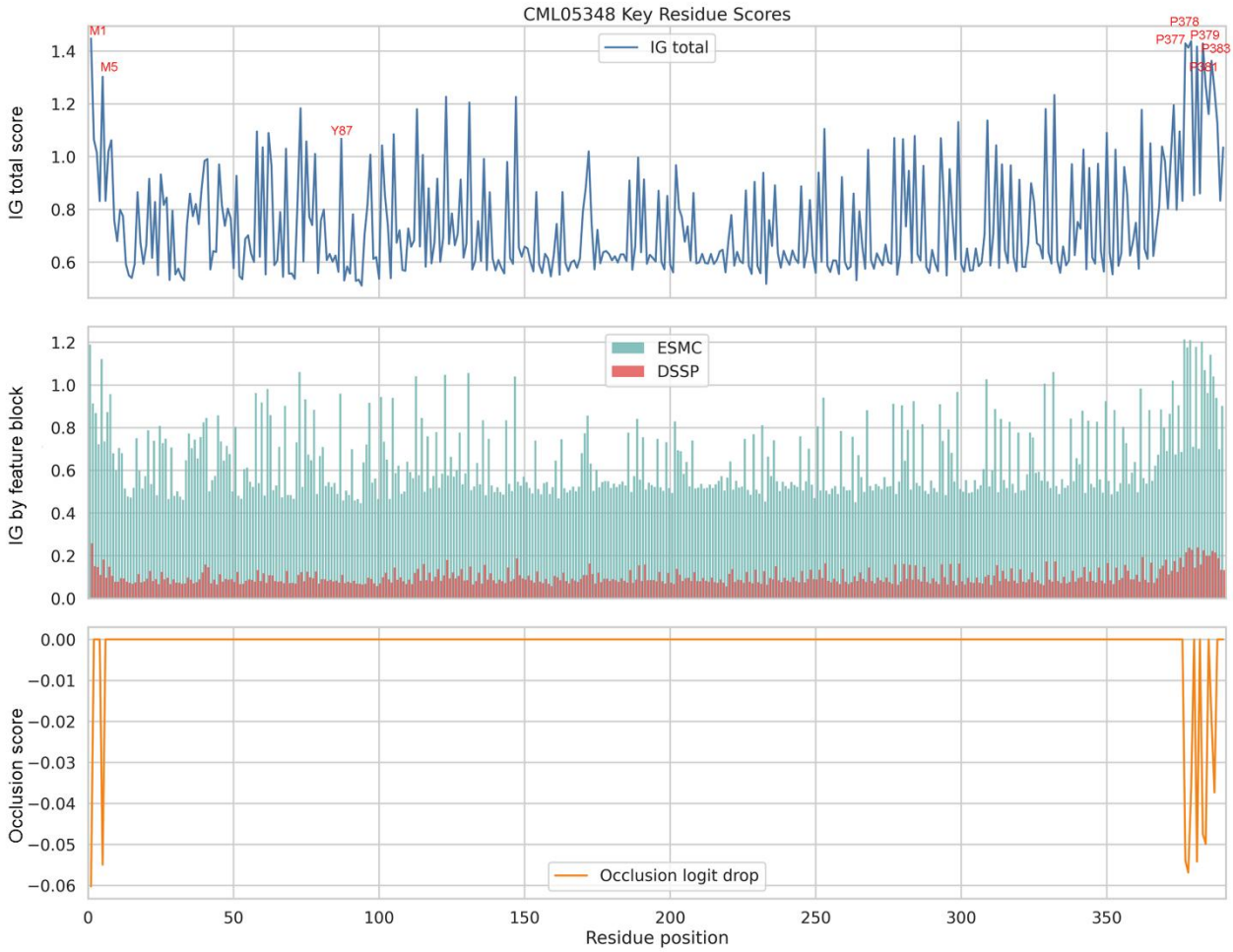

Figure S2. Analysis of Key Residues in CML05348. This study assesses the model's sensitivity to each input feature by computing the gradient of the output with respect to the input, thereby identifying the residues most critical to the current prediction. Furthermore, the influence of each residue on the model's input is evaluated through a residue-masking approach.

Table S1. The three-class classification performance of 16 mainstream machine learning algorithms.

	Model	Accuracy	AUC	Recall	Precision	F1 Score	MCC
LightGBM	Light Gradient Boosting Machine	0.9110 ± 0.0029	0.9810 ± 0.0009	0.9142 ± 0.0029	0.9170 ± 0.0024	0.9118 ± 0.0028	0.8680 ± 0.0042
XGBoost	Extreme Gradient Boosting	0.9094 ± 0.0032	0.9807 ± 0.0006	0.9124 ± 0.0035	0.9152 ± 0.0033	0.9102 ± 0.0033	0.8655 ± 0.0049
Catboost	CatBoost Classifier	0.9083 ± 0.0040	0.9793 ± 0.0008	0.9111 ± 0.0039	0.9138 ± 0.0038	0.9091 ± 0.0039	0.8636 ± 0.0058
GBC	Gradient Boosting Classifier	0.9034 ± 0.0026	0.9778 ± 0.0007	0.9057 ± 0.0026	0.9085 ± 0.0025	0.9042 ± 0.0026	0.8561 ± 0.0037
RF	Random Forest Classifier	0.8920 ± 0.0046	0.9697 ± 0.0012	0.8956 ± 0.0044	0.8993 ± 0.0041	0.8930 ± 0.0046	0.8400 ± 0.0067
ET	Extra Trees Classifier	0.8845 ± 0.0033	0.9678 ± 0.0011	0.8897 ± 0.0030	0.8946 ± 0.0030	0.8857 ± 0.0034	0.8301 ± 0.0048
Ridge	Ridge Classifier	0.8308 ± 0.0051	0	0.8259 ± 0.0058	0.8328 ± 0.0050	0.8307 ± 0.0052	0.7446 ± 0.0077
LR	Logistic Regression	0.8164 ± 0.0037	0.9224 ± 0.0016	0.8180 ± 0.0038	0.8196 ± 0.0040	0.8167 ± 0.0037	0.7238 ± 0.0056
DT	Decision Tree Classifier	0.8077 ± 0.0053	0.8548 ± 0.0037	0.8025 ± 0.0061	0.8087 ± 0.0055	0.8074 ± 0.0054	0.7087 ± 0.0081
KNN	K Neighbors Classifier	0.7486 ± 0.0036	0.8805 ± 0.0033	0.7653 ± 0.0035	0.7811 ± 0.0039	0.7484 ± 0.0040	0.6407 ± 0.0052
Ada	Ada Boost Classifier	0.7356 ± 0.0181	0.8290 ± 0.0066	0.7452 ± 0.0145	0.7414 ± 0.0179	0.7296 ± 0.0189	0.6073 ± 0.0259
SVM	SVM - Linear Kernel	0.6417 ± 0.0349	0	0.6298 ± 0.0516	0.7287 ± 0.0185	0.6139 ± 0.0538	0.4951 ± 0.0556

NB	Naive Bayes	0.6316 ± 0.0060	0.7963 ± 0.0024	0.6297 ± 0.0053	0.6179 ± 0.0092	0.6047 ± 0.0097	0.4525 ± 0.0073
Lda	Linear Discriminant Analysis	0.5559 ± 0.0089	0.6661 ± 0.0075	0.5529 ± 0.0095	0.5592 ± 0.0099	0.5564 ± 0.0090	0.3296 ± 0.0143
Dummy	Dummy Classifier	0.3921 ± 0.0028	0.5000 ± 0.0000	0.3333 ± 0.0000	0.1537 ± 0.0022	0.2208 ± 0.0027	0
Qda	Quadratic Discriminant Analysis	0.3762 ± 0.0406	0.5248 ± 0.0227	0.3649 ± 0.0261	0.2800 ± 0.1749	0.2293 ± 0.0563	0.0702 ± 0.0552

Table S2. Parameters of five machine learning algorithms.

Algorithms	Parameters	Value
LGBMClassifier	learning_rate	0.1
	max_depth	-1
	min_child_samples	20
	min_child_weight	0.001
	n_estimators	100
	num_leaves	31
XGBClassifier	colsample_bytree	1.0
	learning_rate	0.300000012
	max_bin	256
	max_depth	6
	n_estimators	100
CatBoostClassifier	border_count	254
	learning_rate	0.1
GradientBoostingClassifier	max_depth	3
	n_estimators	100
RandomForestClassifier	n_estimators	100
	min_samples_split	2
	min_samples_leaf	1

Table S3. Comparison of per-class performance.

Metrics	Label	Methods	Value (mean ± std)
Precision	Non-PE/PPE	PIONEER	0.974 ± 0.006
		Digerati	0.867 ± 0.005
		LightGBM	0.822 ± 0.005
	PPE	PIONEER	0.951 ± 0.004
		Digerati	0.898 ± 0.008
		LightGBM	0.902 ± 0.009
	PE-PGRS	PIONEER	0.961 ± 0.006
		Digerati	0.987 ± 0.001
		LightGBM	0.972 ± 0.002
Recall	Non-PE/PPE	PIONEER	0.973 ± 0.008
		Digerati	0.966 ± 0.007
		LightGBM	0.936 ± 0.008
	PPE	PIONEER	0.958 ± 0.004
		Digerati	0.944 ± 0.003
		LightGBM	0.904 ± 0.007
	PE-PGRS	PIONEER	0.956 ± 0.006
		Digerati	0.864 ± 0.011
		LightGBM	0.872 ± 0.002
F1-score	Non-PE/PPE	PIONEER	0.973 ± 0.003
		Digerati	0.913 ± 0.003
		LightGBM	0.875 ± 0.004
	PPE	PIONEER	0.955 ± 0.003
		Digerati	0.920 ± 0.005
		LightGBM	0.903 ± 0.005
	PE-PGRS	PIONEER	0.959 ± 0.002
		Digerati	0.921 ± 0.006
		LightGBM	0.919 ± 0.001