

Introduction

The genes that contain the driver mutations that contribute to the disease process are traditionally classified either as ‘tumour suppressors’ (TS) or as oncogenes (OG), depending on their role in cancer development.

Gain-of-function (GOF) variants increase the activity of the mutated protein, while loss-of-function (LOF) variants result in partial or complete inactivation of the protein (Sevim Bayrak et al., 2021).

When mutations result in the loss of function (LOF) of the protein products of tumour suppressors, cancer progression occurs.

Methods and Materials

Database / Class	DataSet_1	DataSet_2	DataSet_3	DataSet_4	DataSet_5	DataSet_6	DataSet_7
Fathmm_cancer	5	5	5	5	5	5	5
Fathmm_Disease		1	1	1	1	1	1
NetSurfP3			8	8	8	8	8
AlphaMissense				1	1	1	1
Protnvar					1	1	1
MutationAssessor						1	1
PolyPhen2							6
Total_Features	5	6	14	15	16	17	23
GOF	1049	1049	952	901	723	605	452
LOF	747	747	737	652	608	553	436
NEU	861	861	857	652	548	521	292
GOF_Balanced	747	747	737	652	548	521	292
LOF_Balanced	747	747	737	652	548	521	292
NEU_Balanced	747	747	737	652	548	521	292

Table 2. Distribution of GOF, LOF, Neutral Datasets with Training Features

Results

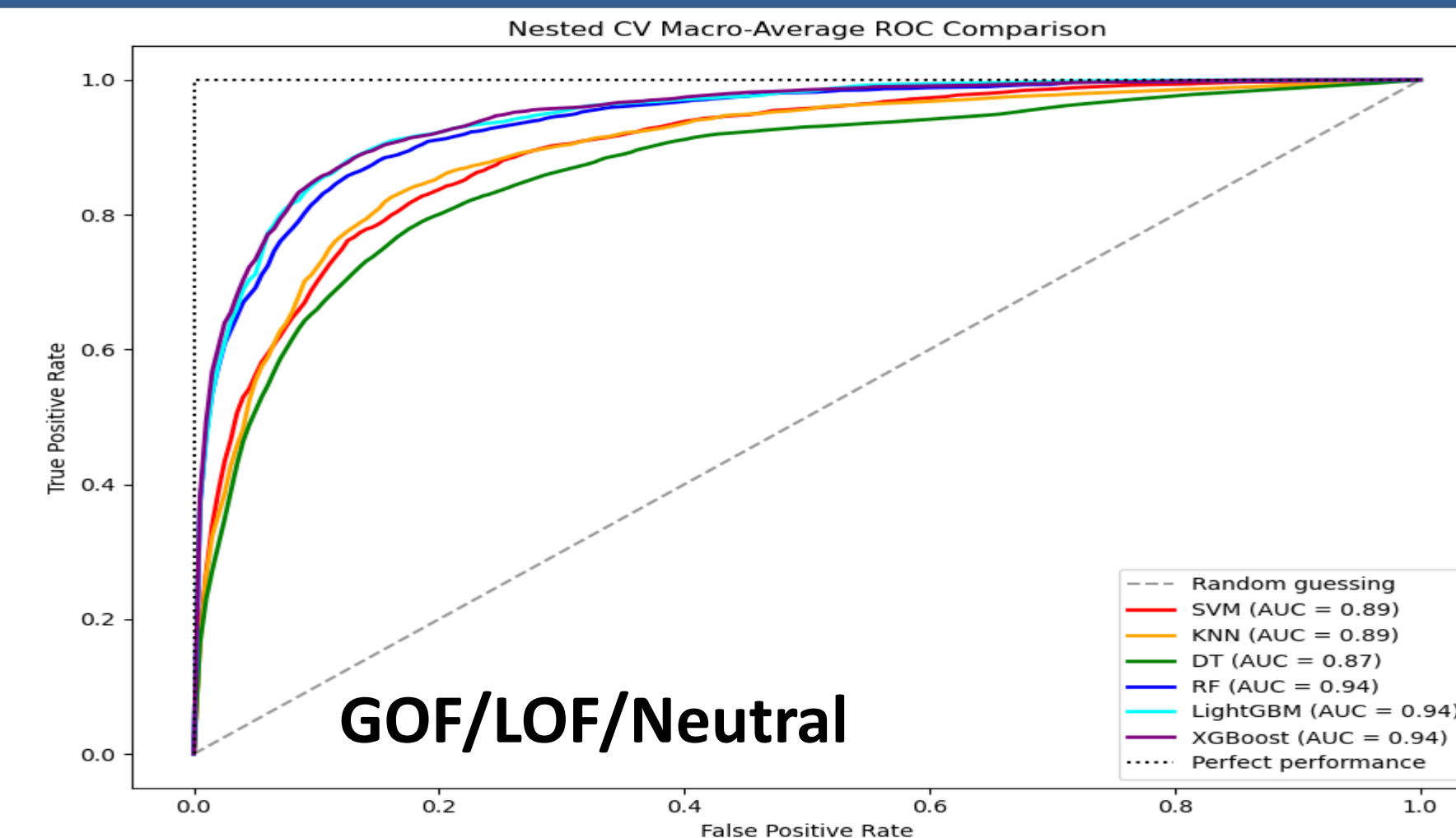


Figure 4. Ensemble Methods Outperform Single-Estimators in Predicting Functional Impact of Genetic Variants

Class	Dataset Split	Precision	Recall	F1-Score
LOF	Training	0.9	0.81	0.85
	Validation (Genes in training)	0.78	0.74	0.76
	Validation (Genes not in training)	0.88	0.55	0.67
LOF	Training	0.89	0.91	0.9
	Validation (Genes in training)	0.81	0.72	0.76
	Validation (Genes not in training)	0.28	0.59	0.38
NEU	Training	0.83	0.9	0.86
	Validation (Genes in training)	0.66	0.86	0.74
	Validation (Genes not in training)	0.67	0.86	0.7

Figure 5. LightGBM Performance on Training and Validation Gene Sets

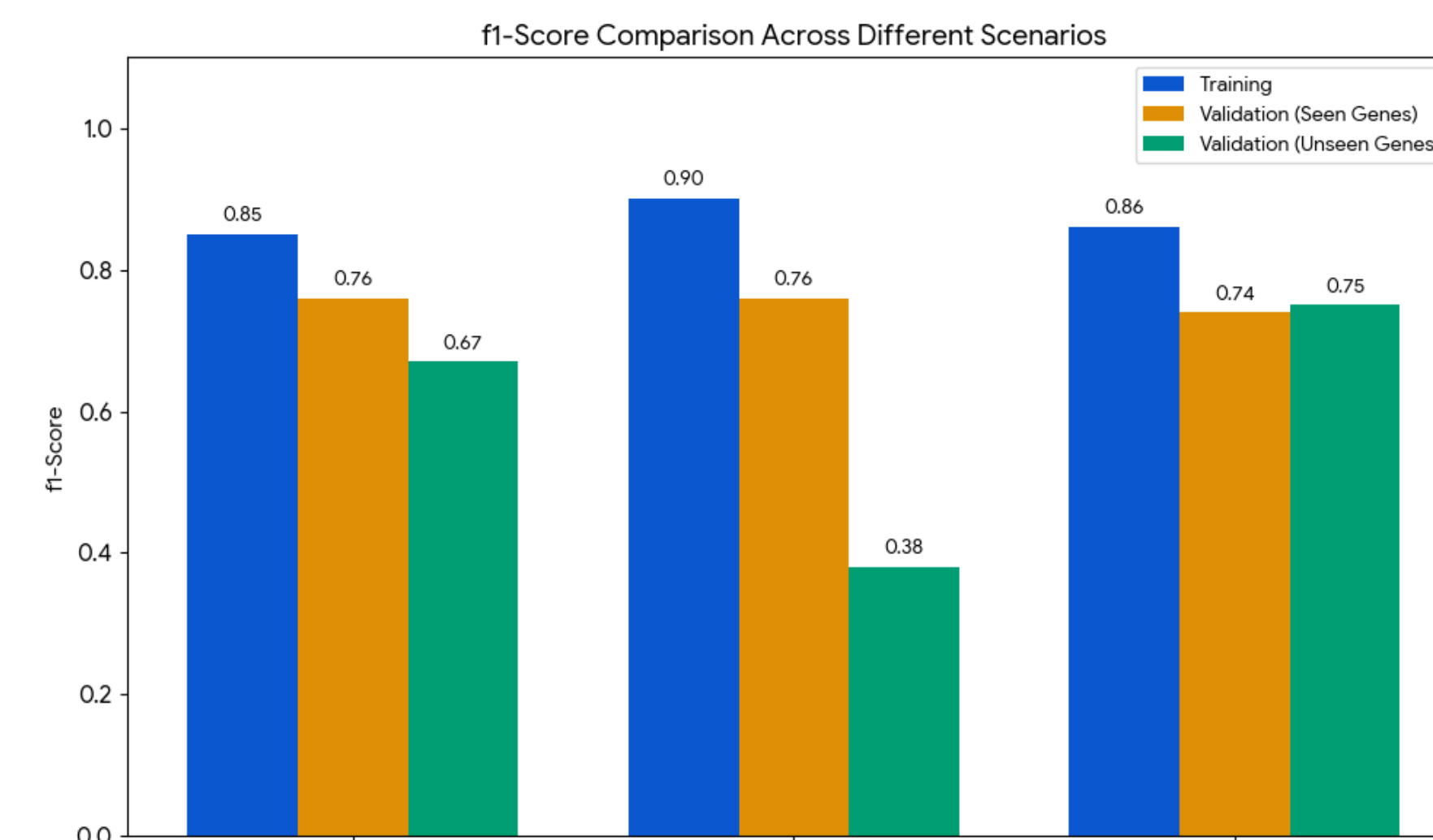


Figure 6. Comparison of F1 Score Among Training and Validation Sets

The LightGBM model demonstrates strong performance on the training set across all classes, with F1-scores of 0.85 (LOF), 0.90 (GOF), and 0.86 (NEU). Performance decreases on validation sets, particularly for unseen genes.

F1-score for validation GOF on unseen genes NEU generalises best (0.75), LOF shows a noticeable decline (0.67), and GOF drops sharply (0.38).

These results highlight that while the model reliably identifies NEU variants, caution is needed for GOF predictions on novel genes, and LOF predictions may require further refinement to help reduce overfitting.

Conclusion

- Our aim has been to analyze the characteristics of GOF and LOF variants, identify potential distinguishing features, and discuss the biological implications of these characteristics in terms of their contribution to protein gain and loss of function by using Machine Learning Methods.
- This study's findings are expected to enhance our understanding of how GOF and LOF variants affect protein function and contribute to disease
- Limitation of this study is using missense mutation data.

Methods and Materials

In this study, I utilised an inherited dataset consisting of 2908 missense variants and spanning across 243 genes. My LOF and GOF mutations were selected with a literature review. For the neutral dataset, I employed data from gnomAD. The goal was to identify a group of mutations that would likely have minimal impact on protein structure or function.

I used several tools and resources to annotate a broad set of features. I mainly used variant-level features and protein-level features.

In this study, we utilised the main existing prediction algorithms, namely FATHMM cancer, FATHMM disease, and Mutation Assessor, to derive features for our analysis. The details of the features can be seen in Table 1. These algorithms provided valuable insights into the potential impact and assessment of mutations in the context of cancer and disease.

I used 6 different types of classifiers on this work and compared their performance to discriminate between the different classes of 7 different datasets. The aim is to uncover molecular mechanisms linked to cancer diagnosis, prognosis, and treatment.

Source	Feature	Description
FATHMM-Cancer	HMM Weights D.	The pathogenicity weights of disease
	HMM Weights O.	The pathogenicity weights of passenger
	HMM Prob W.	The potential impact of wild type amino acid on protein function
	HMM Prob M.	The potential impact of mutated amino acid on protein function
FATHMM-Disease	HMM Weights D.	The pathogenicity weights of disease
Mutation Assessor (MA)	FI score	Functional impact combined score
PolyPhen-2	PSIC Score1	Position Specific Independent Counts score for wild type amino acid
	PSIC Score2	PSIC score for mutant amino acid residue
	dScore	The different between Score1 and Score2
	Transv	Assess whether a substitution is a transversion
	CodPos	Evaluate the position of the substitution within a codon
	CpG	Predicts whether a substitution changes the CpG context.
NetSurfP3	RSA	Relative solvent-accessible area
	ASA	absolute solvent-accessible area
	phi	the dihedral angles phi
	psi	the dihedral angles psi
	p[q3_H]	Prediction of three-state secondary structure for Helix conformation
	p[q3_E]	Prediction of three-state secondary structure for strand conformation
	p[q3_C]	Prediction of three-state secondary structure for Coil conformation
	Disorder	structural disorder
AlphaMissense	am_pathogenicity	AlphaMissense pathogenicity
Protnvar	Stability Change ΔΔG	The predicted free energy change in the protein when the reference amino acid is replaced with the variant.

Table 1. Prediction Databases for GOF/LOF/Neutral

Results

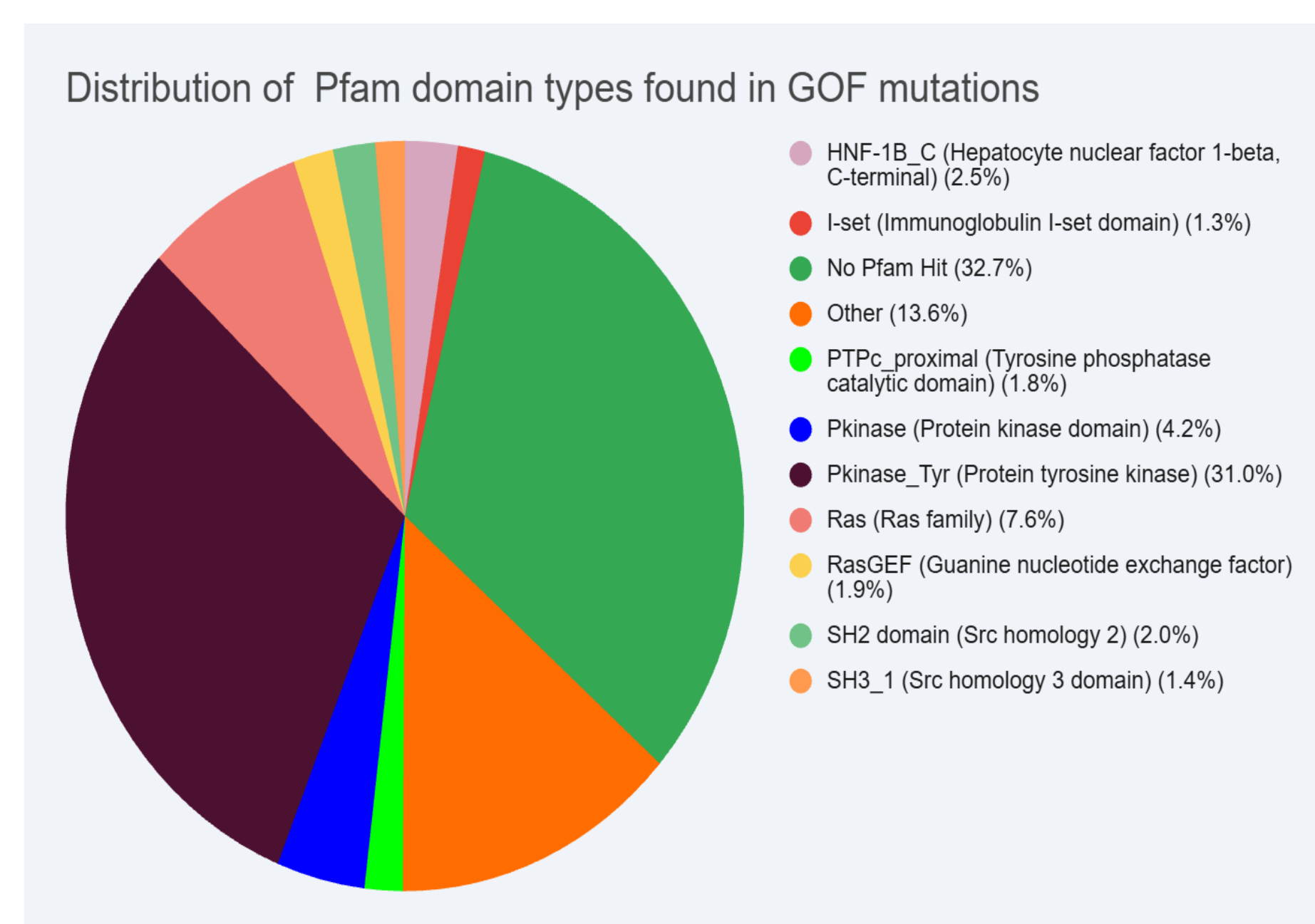


Figure 1. Distribution of Pfam domain types found in GOF mutations.

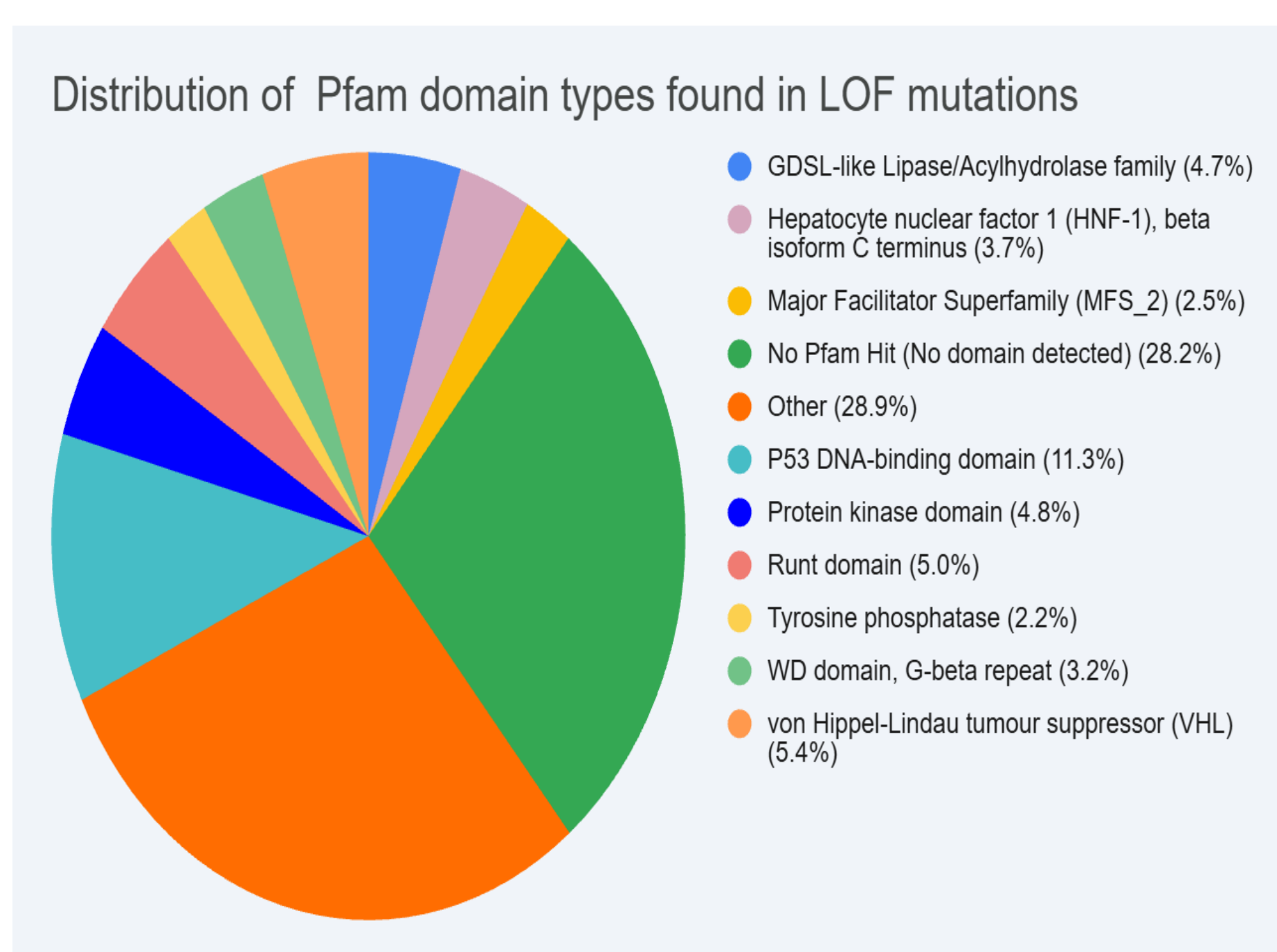


Figure 2. Distribution of Pfam domain types found in LOF mutations

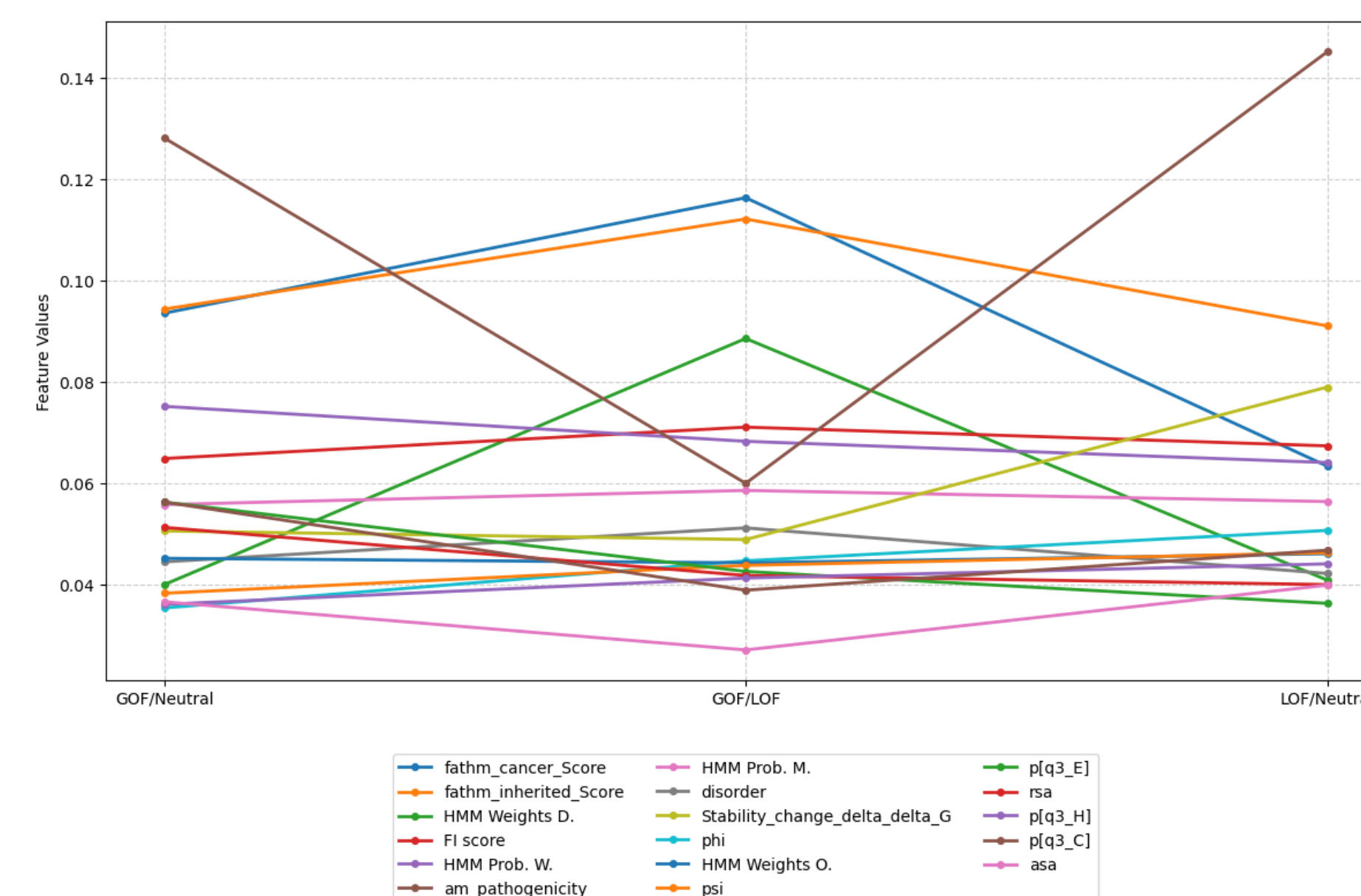


Figure 3. Relative importance of the features used in different classes

Contact

Bioinformatics Lab
School of Life Sciences
University of Sussex
Website: <http://www.sussex.ac.uk/lifesci/bioinformaticsclub/index>

References

- Sevim Bayrak, C., Stein, D., Jain, A., Chaudhary, K., Nadkarni, G.N., Van Vleck, T.T., Puel, A., Boisson-Dupuis, S., Okada, S., Stenson, P.D., Cooper, D.N., Schlessinger, A., Itan, Y., 2021. Identification of discriminative gene-level and protein-level features associated with pathogenic gain-of-function and loss-of-function variants. *The American Journal of Human Genetics* 108, 2301–2318. <https://doi.org/10.1016/j.ajhg.2021.10.007>
- Li, Y., Zhang, Y., Li, X., Yi, S., Xu, J., 2019. Gain-of-Function Mutations: An Emerging Advantage for Cancer Biology. *Trends Biochem Sci* 44, 659–674. <https://doi.org/10.1016/j.tibs.2019.03.009>
- Stein, D., Kars, M.E., Wu, Y. et al. Genome-wide prediction of pathogenic gain- and loss-of-function variants from ensemble learning of a diverse feature set. *Genome Med* 15, 103 (2023). <https://doi.org/10.1186/s13073-023-01261-9>