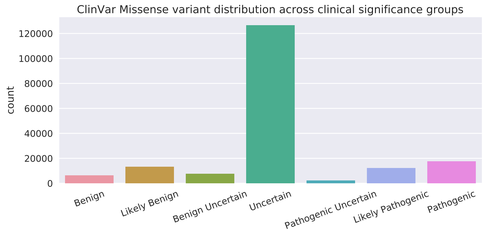
Missense Variant Pathogenicity Modelling with C-MIP

Philip Fradkin† and Daniele Merico†‡

†Deep Genomics ‡Hospital for Sick Kids

Novel Missense Variant Classification

More than a third of rare and pathogenic mendelian variants are missense, resulting in an amino acid substitution. However, not all missense mutations are deleterious and accurately deciphering variant impact on protein function remains a challenge. In this work we outline technologies leading to the development of new methods, assess performance of novel missense impact predictors, and propose our own predictor, ClinVar Missense Impact Predictor (C-MIP).

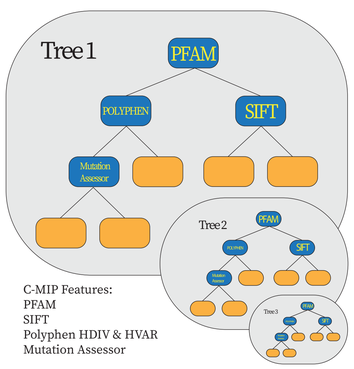


Break Throughs In Missense Pathogenicity Modelling

Three core ideas have driven the development of these tools: multiple sequence alignments, prediction ensembles, and learning complex motifs patterns using neural networks. We evaluated 9 predictors, 5 of which were released in the last two years utilizing these methods.

Predictor	Sequence Alignment	Prediction Ensemble	Neural Network
EVmutation	✓		
PrimateAI			✓
Envision		✓	
MCAP	✓	✓	
C-MIP		✓	

C-MIP



ClinVar Missense Impact Predictor (C-MIP) is a supervised gradient boosting machine. It was trained on ClinVar Pathogenic & Benign variants that have over two submitters. Its features are SIFT [4], Polyphen2 HDIV, Polyphen2 HVAR [1], Pfam [2] and Mutation Assessor [5]. Gradient boosting machines work by iteratively building trees to predict the residual of the previous tree ensemble.

Results

EVmutation [3] is an unsupervised method which captures co-dependencies between protein positions from multiple sequence alignments. This approach however requires a significant amount of sequences and is only available for a fraction of the proteome. To assess accuracy of this tool we divided our datasets into EVmutation defined and undefined regions.

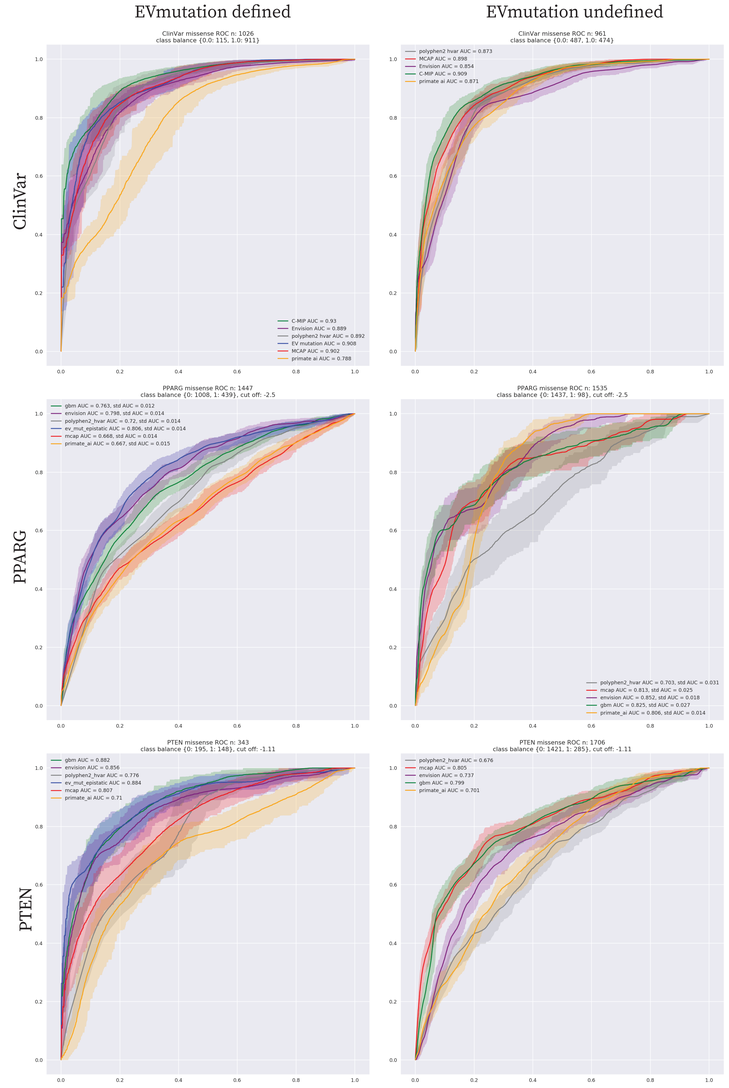


Fig. 3: ROC missense impact predictor performance

EVmutation and C-MIP on average have the highest true positive rates at 5% false positive threshold in EVmutation defined and EVmutation undefined regions respectively.

Confounding Factors

Dataset confounding factors, such as variant allele frequency can obscure a predictor's true performance. There can be significant differences in these variables between positive and negative classes leading a predictor to pick up irrelevant features. This can lead the predictor to misclassify variants which resemble the opposite class only in terms of the confounding variable.

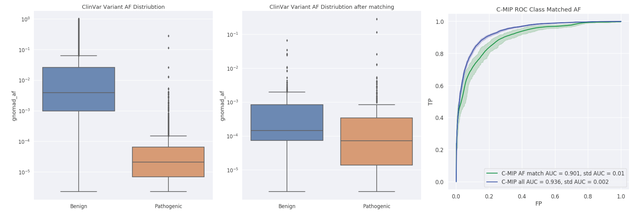


Fig. 4: Allele frequency as a confounding factor

Performance of CMIP is robust to balancing ClinVar, the pathogenicity dataset, by variant allele frequency as extracted from gnomAD.

Remarks & Conclusions

There is high variability in predictor performance across datasets. This work can benefit from adding more datasets and explicitly controlling for confounding variables such as allele frequency and conservation. However the three major take-aways from this work are:

1. For novel missense variant prediction use EVmutation when available. If not use C-MIP.
2. Confounding factors such as allele frequency or conservation can greatly influence predictor analysis.
3. There is a significant opportunity for use of neural networks in missense pathogenicity modelling.

References

[1] Ivan A Adzhubei et al. "A method and server for predicting damaging missense mutations". In: *Nature methods* (Apr. 2010).

[2] Robert D Finn et al. "Pfam: the protein families database". In: *Nucleic acids research* (Jan. 2014).

[3] Thomas A Hopf et al. "Mutation effects predicted from sequence co-variation". In: *Nature biotechnology* (Feb. 2017).

[4] Prateek Kumar, Steven Henikoff, and Pauline C Ng. "Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm". In: *Nature protocols* (2009).

[5] Boris Reva, Yevgeniy Antipin, and Chris Sander. "Determinants of protein function revealed by combinatorial entropy optimization". In: *Genome biology* (2007).