Evaluation of 50 Computational Tools for Predicting Pathogenicity of Genetic Variants: Unveiling representativeness issues on public datasets

Ragousandirane Radjasandirane, Jean-Christophe Gelly, Julien Diharce, Alexandre de Brevern

Evaluation of 50 Computational Tools for Predicting Pathogenicity of Genetic Variants: Unveiling representativeness issues on public datasets

Ragousandirane Radjasandirane, Jean-Christophe Gelly, Julien Diharce and Alexandre G. de Brevern

Université Paris Cité and Université de la Réunion and Université des Antilles, INSERM, BIGR, DSIMB Bioinformatics Team, F-75014 Paris, France

INTRODUCTION

Amino acid substitutions on protein sequences are generally harmless, but a significant proportion of them can cause diseases. Accurately predicting the effect of these genetic variants can be crucial for clinicians, it can potentially speed up the diagnosis of patients having missense variants that are likely to lead to disease. Today, a variety of computational tools have been developed to predict the pathogenicity of genetic variants using numerous methodologies. The most well-known tools are SIFT [1] and PolyPhen [2], each of them accumulated more than 10,000 citations. More recently, many tools have been developed using Artificial Intelligence and other innovative approaches. It is important to evaluate and rank the performance of these different computational tools in order to guide future users and clinicians.

In this study, we rigorously evaluated 50 tools using quality data and measures for each computational method. In addition, we carried out a detailed analysis of the available data on genetic variants to highlight a problem inherent in public databases: the prediction quality is significantly impacted by the different variant datasets.

Our results show that variants from ClinVar appear to be easy to predict, whereas variants from other data sources are more difficult to predict. We show that the predictability of variants can be divided into two distinct categories: (i) Easy and (ii) Difficult to predict. We have therefore developed a neural network model capable of classifying variants into these categories and tested the model on cancer datasets to demonstrate its potential use.

DATA

Variant datasets used :

- ClinVar [3] and ClinGen [4] : Connecting human variation and observed health status
- Clinical dataset [5] : Data collected from patient
- 1000Genomes [6] : Catalogue of common human genetic variation across the world from sequencing

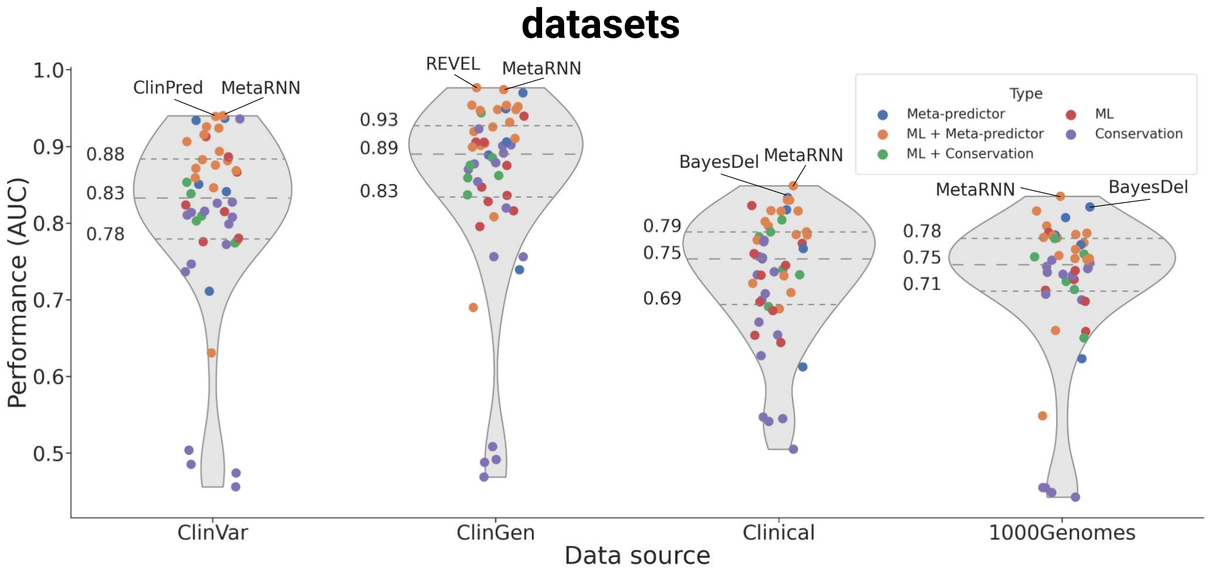
METHODS

A total of 50 computational methods have been assessed. They are based on various algorithms among : Machine Learning (ML), Meta-prediction, Conservation and combination of algorithms (e.g., ML + Meta-prediction)

The performance of each method has been evaluated using the Area Under the Curve (AUC) metric.

RESULTS

Performances of 50 computational methods on 4 different datasets

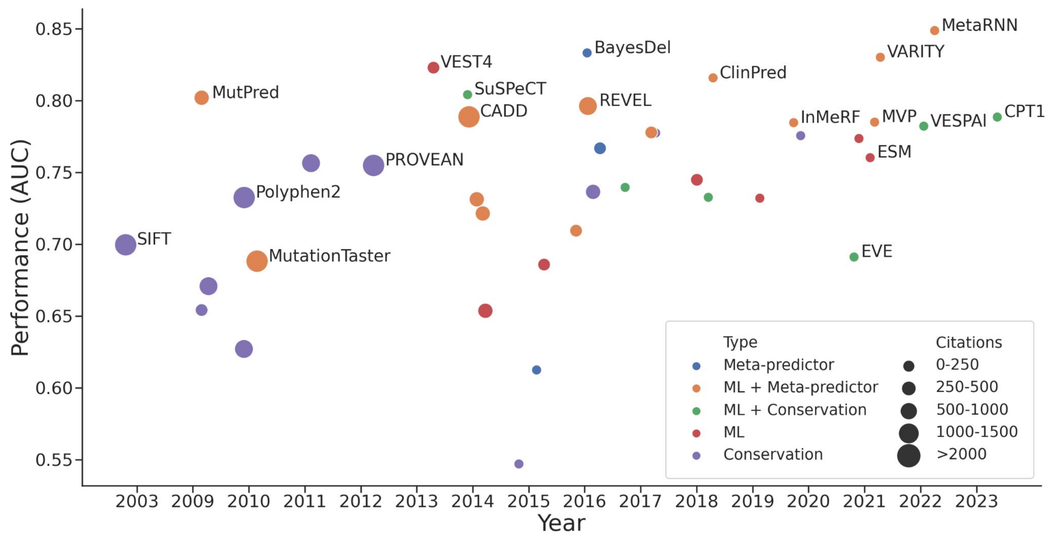


Computational methods show robust performances on **ClinVar** and **ClinGen** datasets with an median AUC of 0.89 and 0.84 respectively

However, performances on **Clinical** and **1000Genomes** datasets are lower with an median AUC of 0.75

Method performances are highly dependent to the **nature** of the dataset

Performances on the clinical dataset

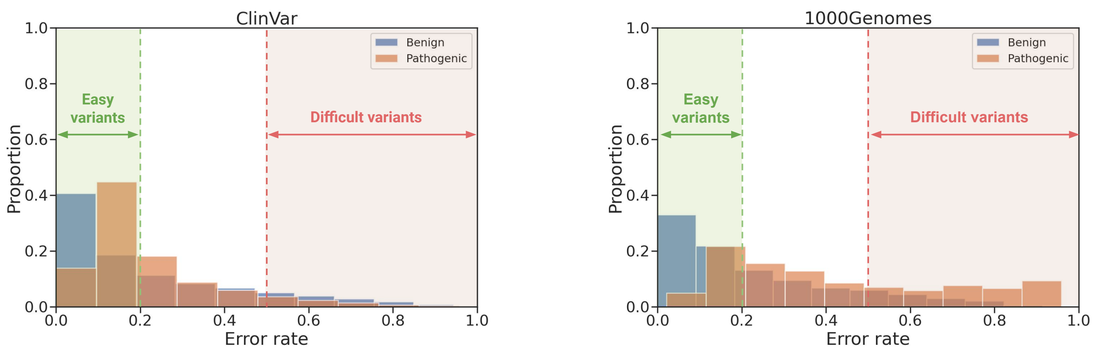


Best performing methods with an AUC > 0.8 are methods based on **ML** and/or **Meta-prediction** (e.g., MetaRNN [7] with an AUC of 0.84)

Conservation based methods are likely to be **outperformed** by machine learning algorithms

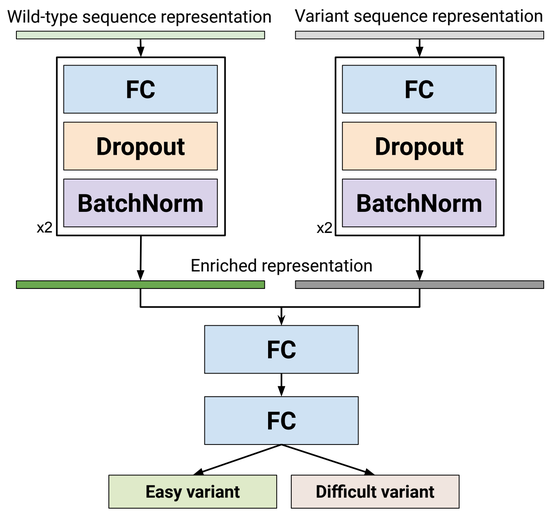
Difficulty of variant analysis

For each variant, we can compute the proportion of tools that incorrectly predicted the variant annotation (Benign or Pathogenic), giving an error rate per variant

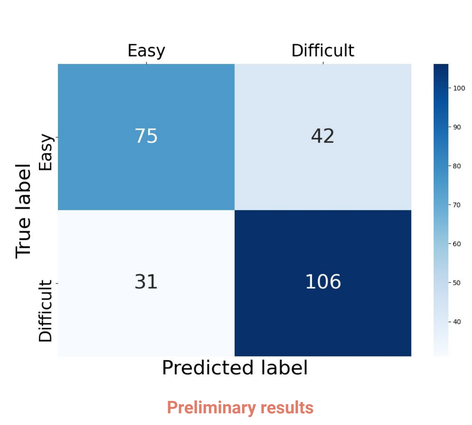


According to error rates, variants can be categorized into two classes: **Easy** to predict and **Difficult** to predict

Neural network model to predict the difficulty of variants



We have developed a siamese network to compare wild-type and variant sequences



The model is capable of correctly classify the difficulty of variants from a cancer dataset [8] with an sensitivity of **77%** and specificity of **64%**

REFERENCES

[1] Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic acids research, 31(13), 3812-3814.
[2] Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., ... & Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. Nature methods, 7(4), 248-249.
[3] Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., & Maglott, D. R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic acids research, 42(D1), D980-D985.
[4] Rehm, H. L., Berg, J. S., Brooks, L. D., Bustamante, C. D., Evans, J. P., Landrum, M. J., ... & Watson, M. S. (2015). ClinGen—the clinical genome resource. New England Journal of Medicine, 372(23), 2235-2242.
[5] Gunning, A. C., Fryer, V., Fasham, J., Crosby, A. H., Ellard, S., Baple, E. L., & Wright, C. F. (2021). Assessing performance of pathogenicity predictors using clinically relevant variant datasets. Journal of medical genetics, 58(8), 547-555.
[6] 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. Nature, 526(7571), 68.
[7] Li, C., Zhi, D., Wang, K., & Liu, X. (2022). MetaRNN: differentiating rare pathogenic and rare benign missense SNVs and InDels using deep learning. Genome Medicine, 14(1), 115.
[8] Goncearenco, A., Rager, S. L., Li, M., Sang, Q. X., Rogozin, I. B., & Panchenko, A. R. (2017). Exploring background mutational processes to decipher cancer genetic heterogeneity. Nucleic acids research, 45(W1), W514-W522.

CONCLUSION

- One of the **biggest benchmark** currently available
- Significant variation in performance depending on the **nature of the dataset**
- Possibility to **improve** prediction of variant effects
- Difficulty of variants can be a criterion for generating **representative datasets** that can improve model performance