



Delft University of Technology

**Document Version**

Final published version

**Citation (APA)**

van den Berg, B., Thornton, JM., Reinders, M., de Ridder, D., & Beer, TAP. (2013). *Predicting functional effect of human missense mutations*. 1.

**Important note**

To cite this publication, please use the final published version (if applicable). Please check the document version above.

**Copyright**

In case the licence states "Dutch Copyright Act (Article 25fa)", this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership. Unless copyright is transferred by contract or statute, it remains with the copyright holder.

**Sharing and reuse**

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

**Takedown policy**

Please contact us and provide details if you believe this document breaches copyrights. We will remove access to the work immediately and investigate your claim.

*This work is downloaded from Delft University of Technology.*

# → Predicting functional effect of human missense mutations

B.A. van den Berg<sup>\*1,3,4</sup>, J.M. Thornton<sup>2</sup>, M.J.T. Reinders<sup>1,3,4</sup>, D. de Ridder<sup>1,3,4</sup>, and T.A.P. de Beer<sup>2</sup>

<sup>1</sup> Delft Bioinformatics Lab, Faculty of Electrical Engineering, Mathematics & Computer Science, Delft University of Technology, Delft, The Netherlands,

<sup>2</sup> European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK,

<sup>3</sup> Netherlands Bioinformatics Centre, Nijmegen, The Netherlands,

<sup>4</sup> Kluwer Centre for Genomics of Industrial Fermentation, Delft, The Netherlands

\* b.a.vandenberg@tudelft.nl

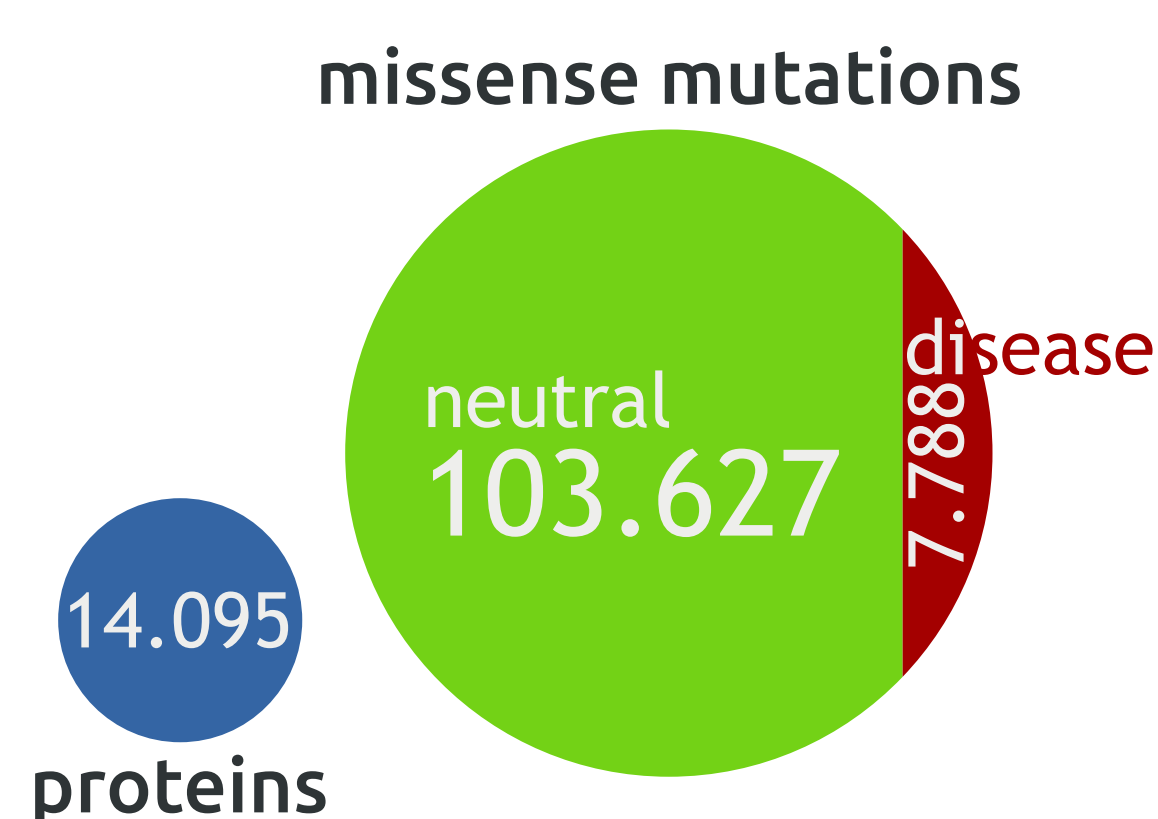
## Introduction

Our aim is to prioritize human missense mutations by their probability of being disease causing. Such a computational method could be used to obtain a reduced set of mutations with a relatively large fraction of disease related mutations, thereby aiding in the search for this type of mutation within a large mutation set.

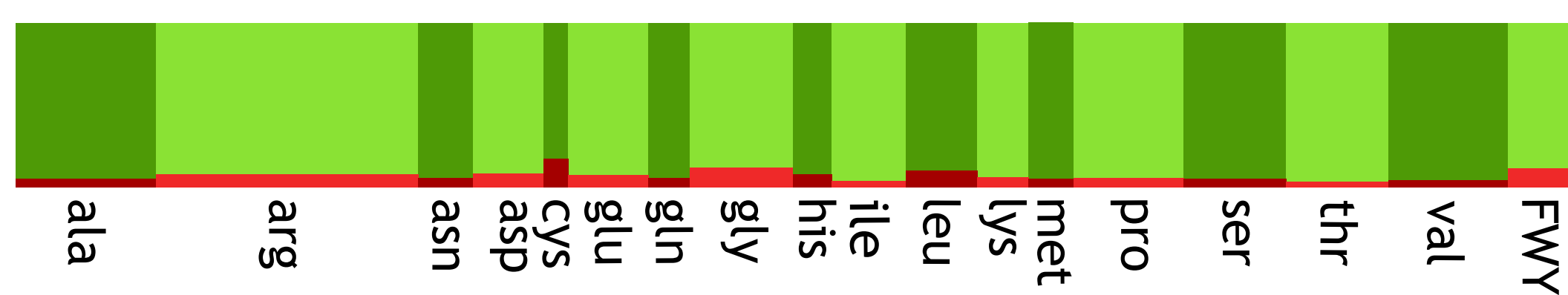
Whereas a range of methods is available for this purpose, only few employ the availability of the 1000G data to obtain a set of neutral mutations. The novelty of our approach is the use of separate classifiers that were trained on a subset of mutations from one amino acid to any other amino acid. The combined performance of these classifiers show an improved performance compared to the often used prediction method PolyPhen2.

## Data set

The data set is composed of in total 111.415 mutations in 14.095 proteins. The disease mutations were obtained from the OMIM database and the neutral mutations were obtained from the 1000 Genomes project.



Mutations were split into 20 (non-overlapping) subsets, with in each subset mutations from one amino acid to any other amino acid. The phenylalanine, tryptophan, and tyrosine subsets are combined into one set to increase the set size, resulting in a total of 18 subsets.



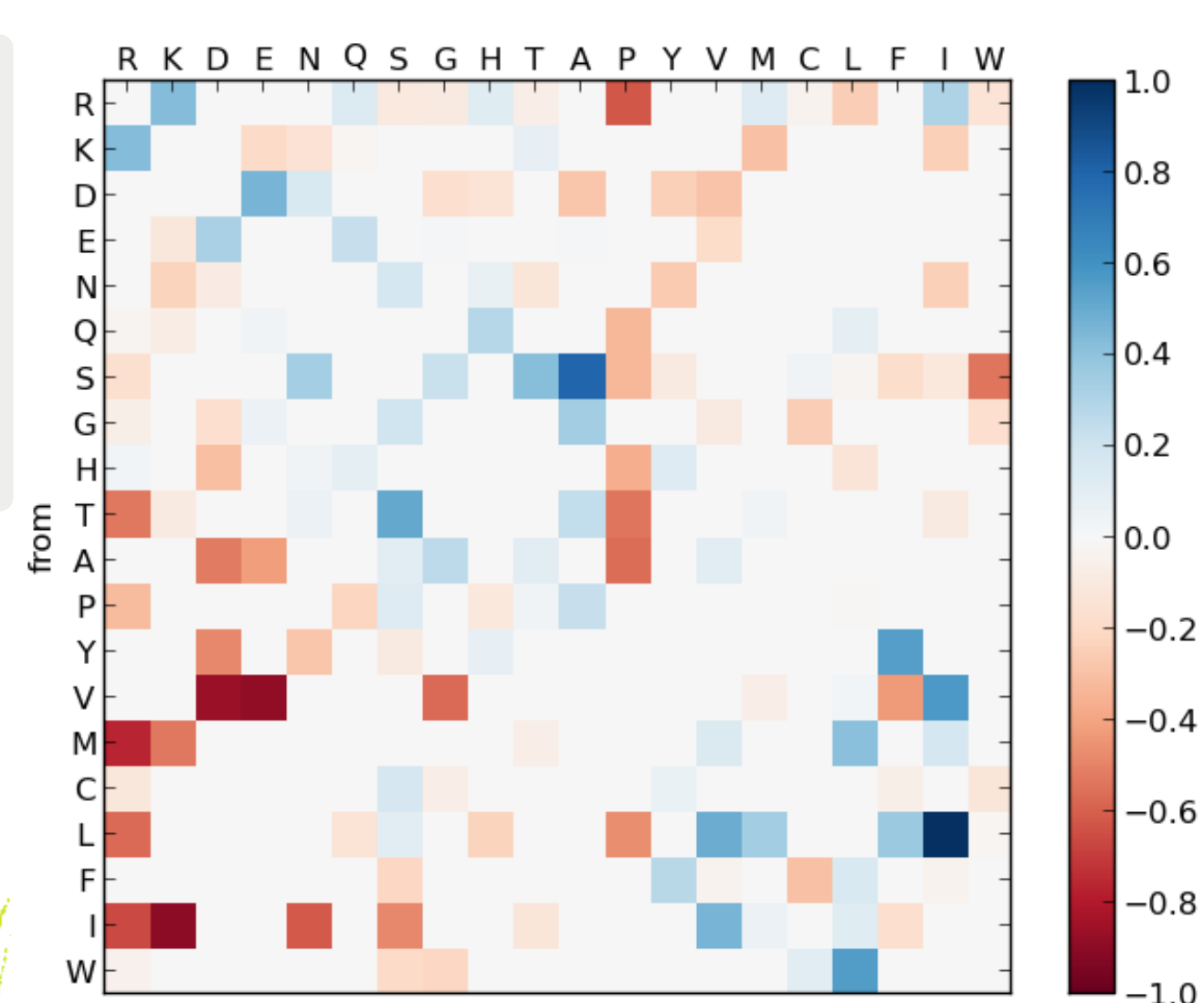
## Classification

Separate classifiers were trained on each of the eighteen mutation subsets using the settings below. For comparison, one classifier was trained on the entire dataset.

protocol: 10-fold cross-validation  
 classifier: linear discriminant analysis (LDA) classifier  
 measure: area under the receiver operator curve (AUROC)

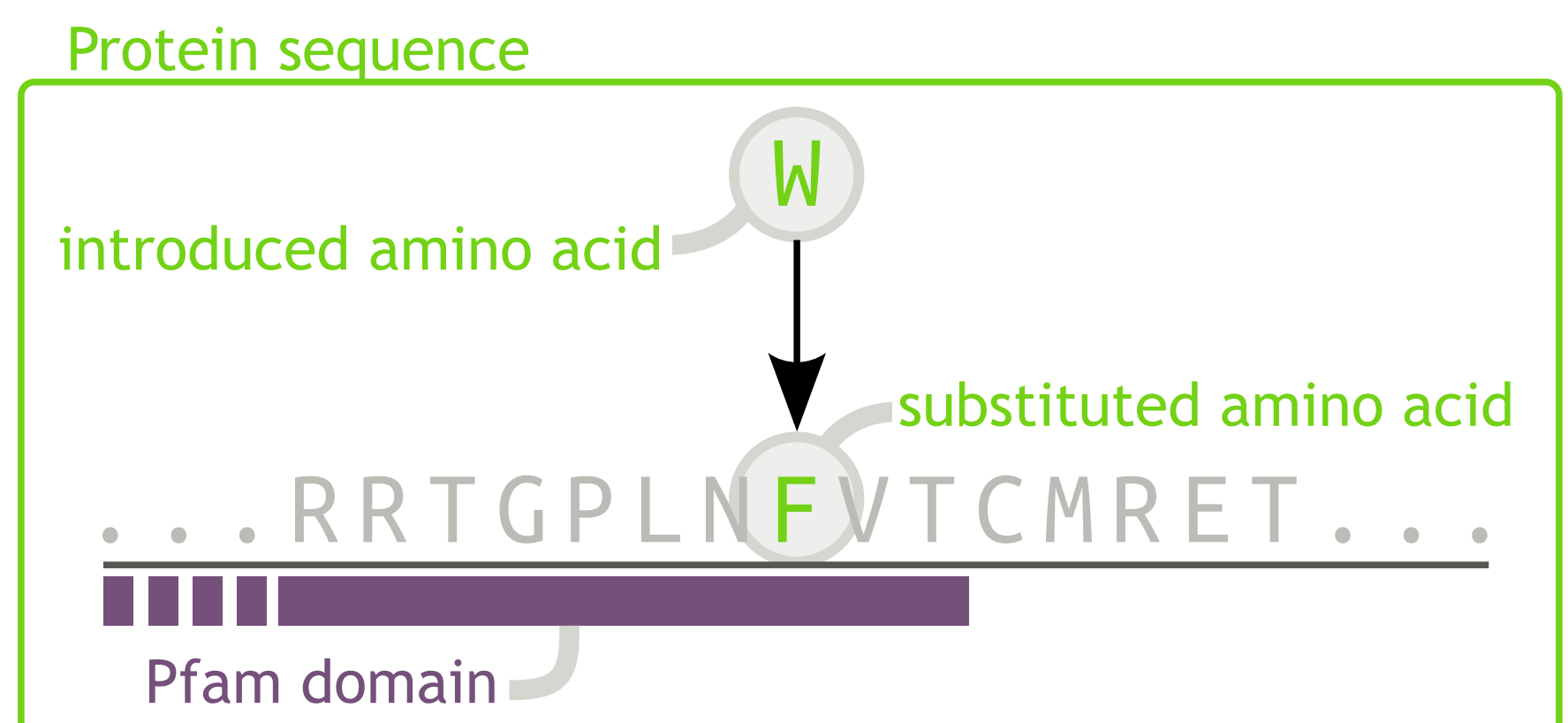
## Amino acid counts

Comparison of the occurrences in the neutral and disease set shows which mutations are relatively safe (blue) and dangerous (red).

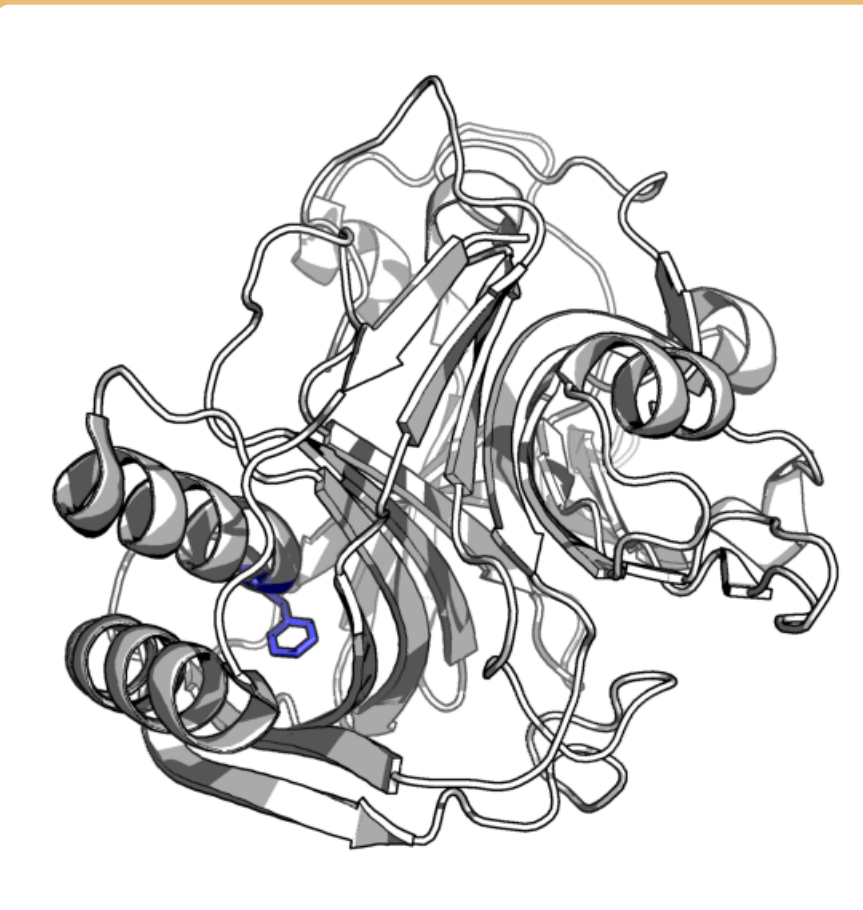


## Feature data sources

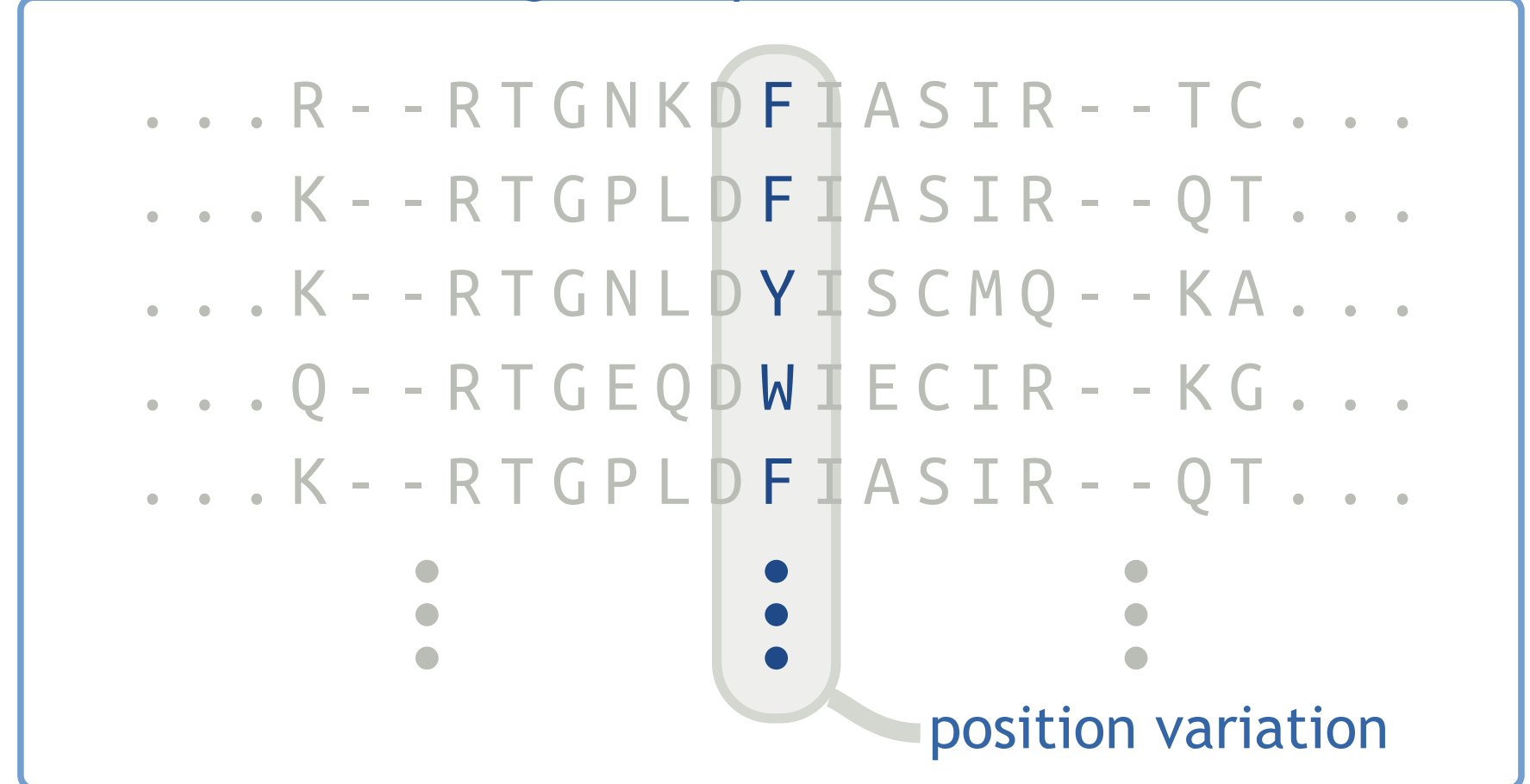
The data sources shown in the figures were employed for feature extraction. Because the availability of structure data is limited, structure-based features were only acquired for part of the mutations.



### Protein structure



### MSA with homologous sequences



## Features

### Missense mutation feature vector



Twenty features encode the missense mutation, each column representing one amino acid. The substituted amino acid is set to -1 and the introduced amino acid to 1. All other amino acids are 0.

The first feature is a conservation score based on the MSA with homologous sequences as obtained by the Evolutionary Trace Server. The second feature is a binary feature that indicates if the introduced amino acid is in the position variation or not.

Nineteen features that give the minimal 'characteristic' distance between the introduced amino acid and the amino acids in the position variation. The used characteristics are, for example, hydrophobicity, size, and isoelectric point.

Protein structure features: solvent exposed area and the three backbone angles.

Binary feature that indicates if the mutation falls within a Pfam domain or not.

## Results: classification performance

Most of the sub-classifiers as well as their combined result (green) show an improved performance compared to PolyPhen2 (blue). In particular, a striking improvement is observed for charged (arg, lys, asp, glu) and aliphatic (leu, val) sub-classifiers. The reduced performance of the classifier trained on the entire data set (purple) supports the use of sub-classifiers.

