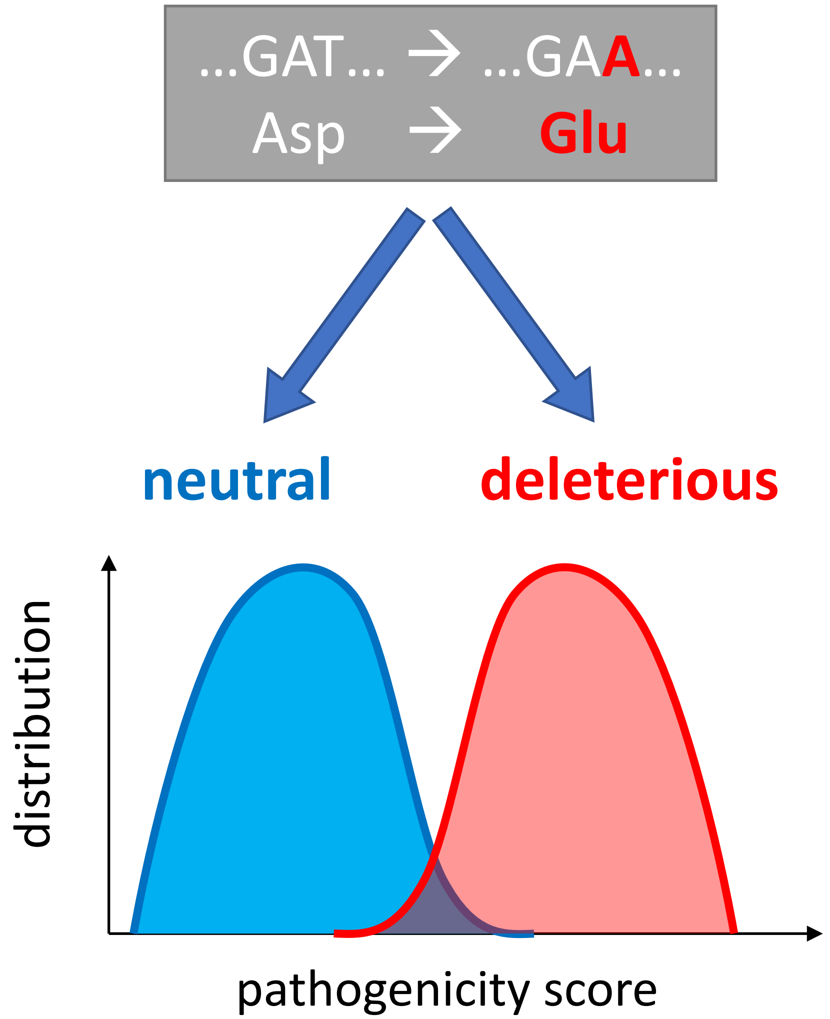
Structural dynamics is a determinant of the functional significance of missense variants

Luca Ponzoni, Ivet Bahar

Department of Computational and Systems Biology,
School of Medicine

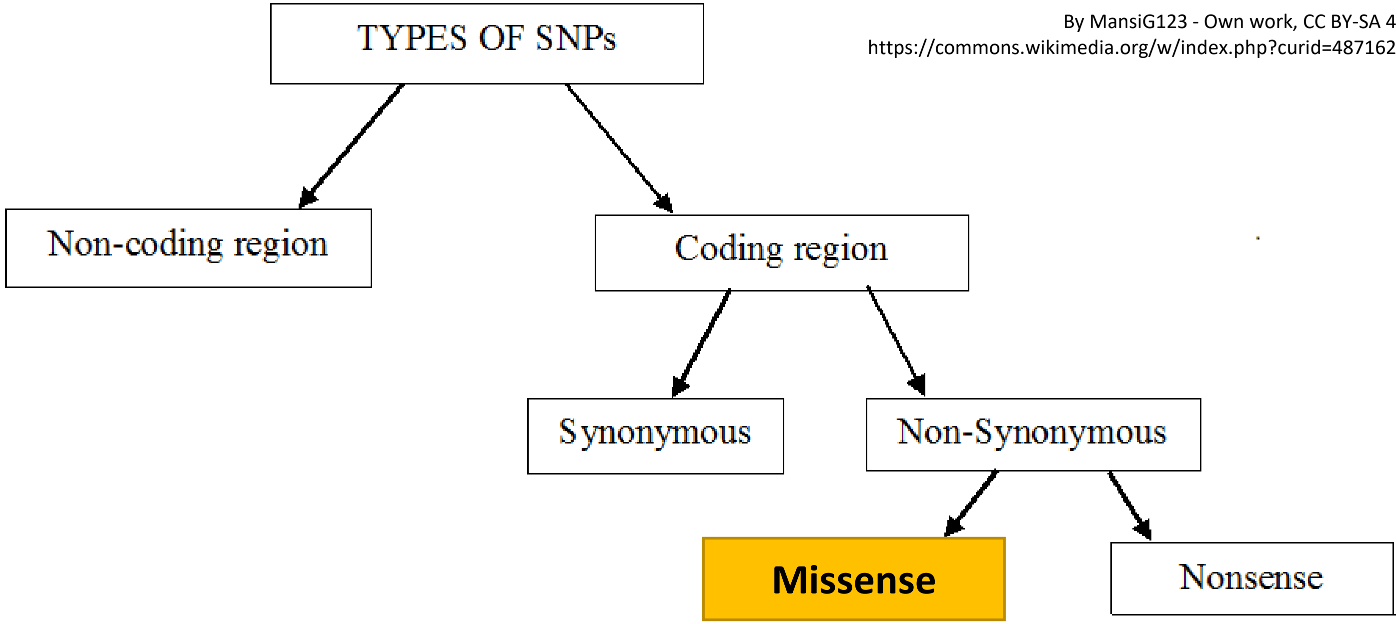
Classification of Single Amino acid Variants (SAVs)



...GAT... → ...GA**A**...

Asp → **Glu**

neutral **deleterious**



TYPES OF SNPs

Non-coding region

Coding region

Synonymous

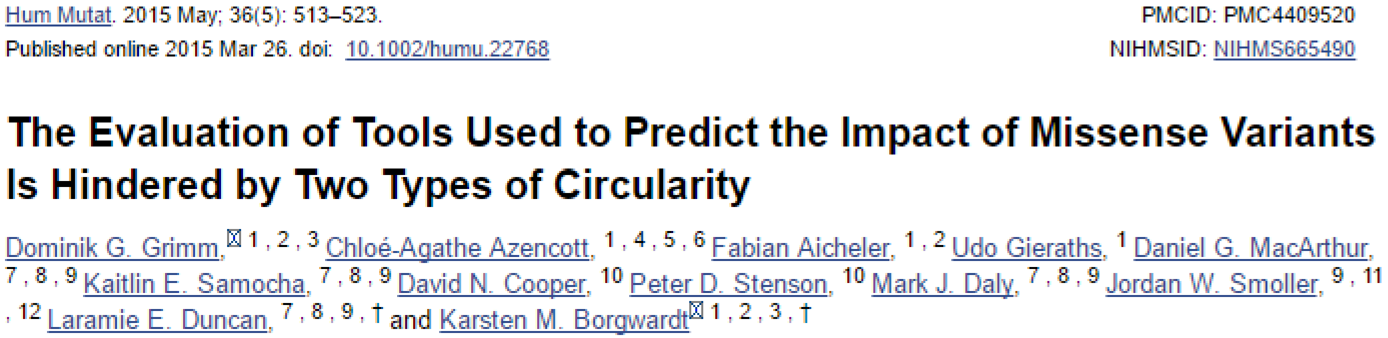
Non-Synonymous

Missense

Nonsense

By MansiG123 - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=48716253

- hereditary (or germline) mutations → genetic disease
- acquired (or somatic) mutations → cancer



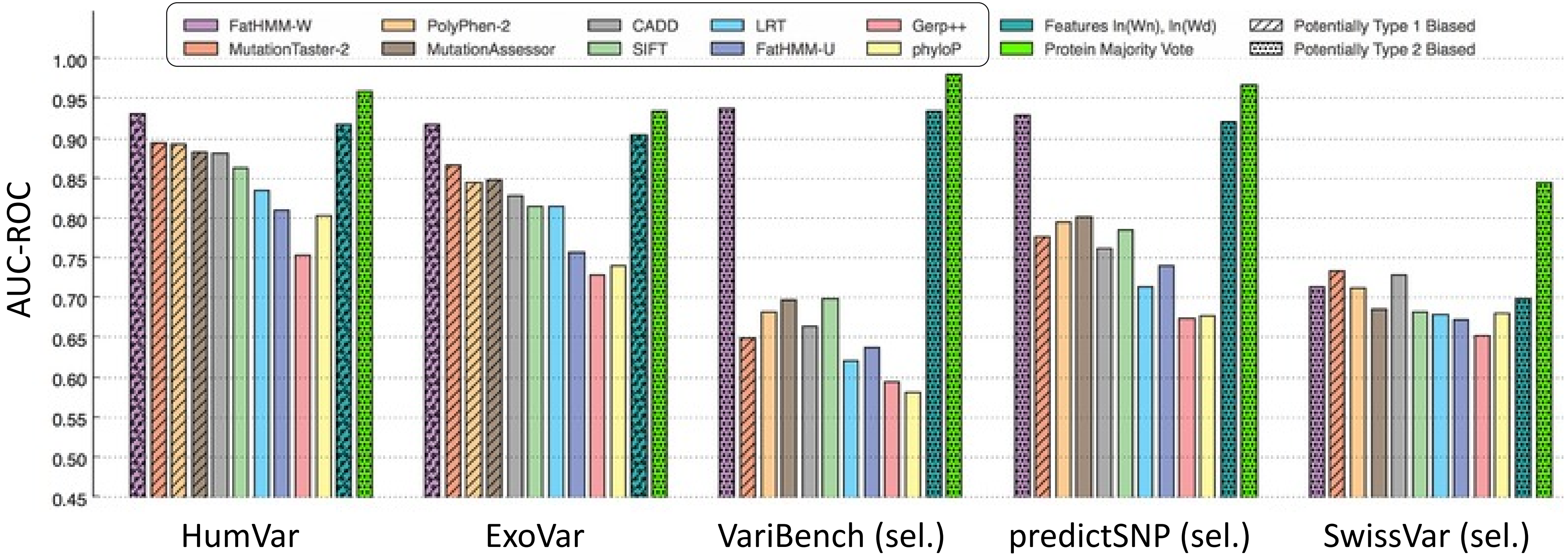
Hum Mutat. 2015 May; 36(5): 513–523.
Published online 2015 Mar 26. doi: 10.1002/humu.22768

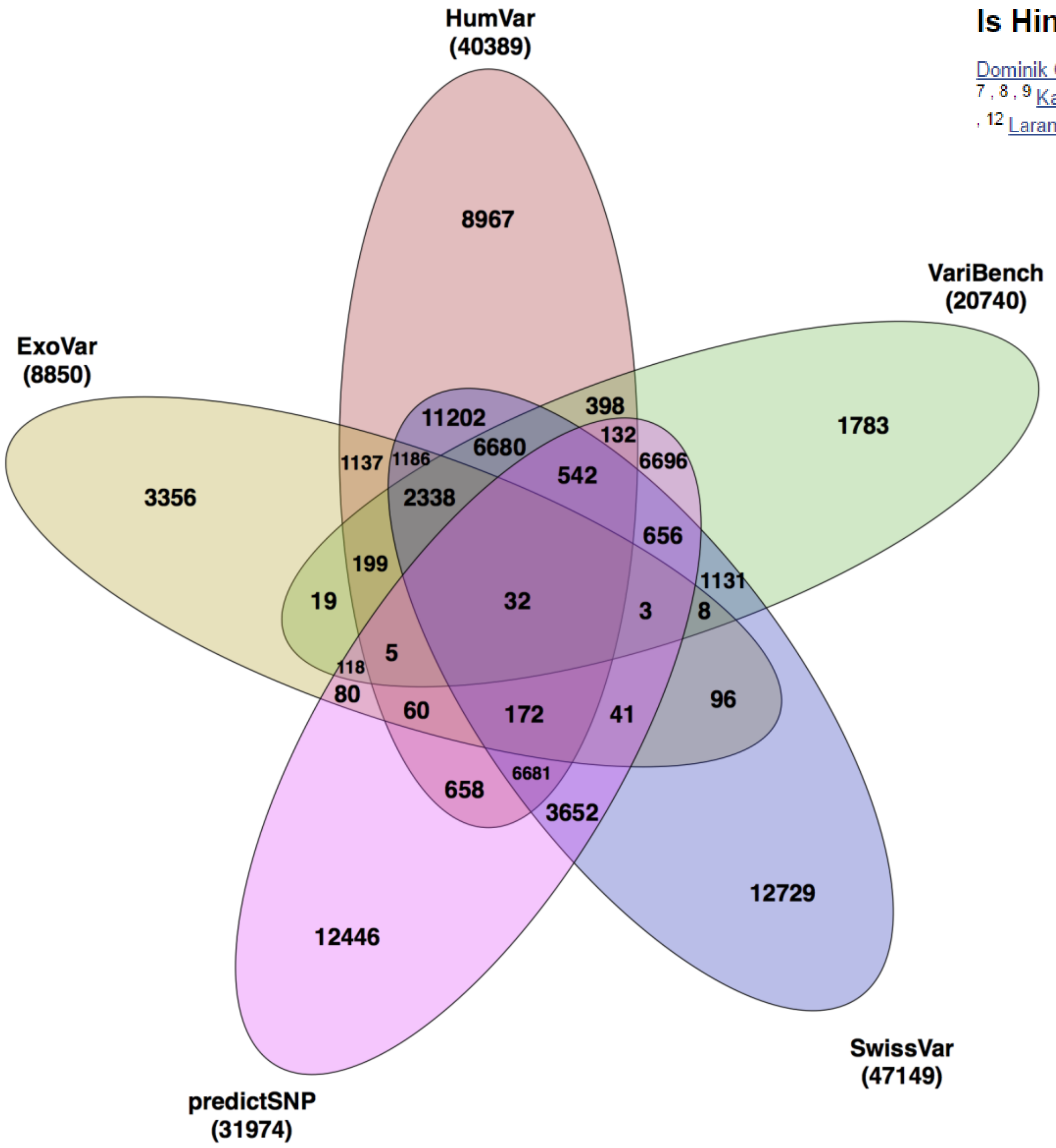
PMCID: PMC4409520
NIHMSID: NIHMS665490

The Evaluation of Tools Used to Predict the Impact of Missense Variants Is Hindered by Two Types of Circularity

Dominik G. Grimm,⊠[1, 2, 3] Chloé-Agathe Azencott, [1, 4, 5, 6] Fabian Aicheler, [1, 2] Udo Gieraths, [1] Daniel G. MacArthur, [7, 8, 9] Kaitlin E. Samocha, [7, 8, 9] David N. Cooper, [10] Peter D. Stenson, [10] Mark J. Daly, [7, 8, 9] Jordan W. Smoller, [9, 11, 12] Laramie E. Duncan, [7, 8, 9, †] and Karsten M. Borgwardt⊠[1, 2, 3, †]

prediction tools

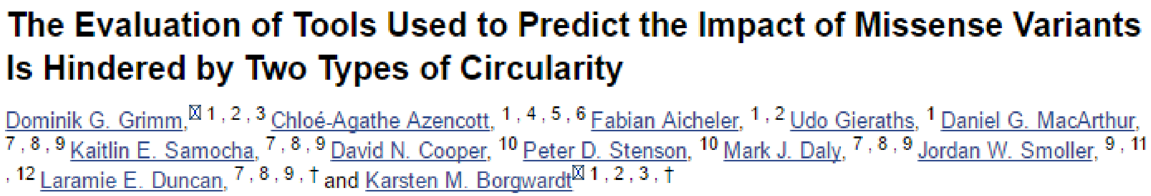




The Evaluation of Tools Used to Predict the Impact of Missense Variants Is Hindered by Two Types of Circularity

Dominik G. Grimm,⊠ [1,2,3] Chloé-Agathe Azencott, [1,4,5,6] Fabian Aicheler, [1,2] Udo Gieraths, [1] Daniel G. MacArthur, [7,8,9] Kaitlin E. Samocha, [7,8,9] David N. Cooper, [10] Peter D. Stenson, [10] Mark J. Daly, [7,8,9] Jordan W. Smoller, [9,11,12] Laramie E. Duncan, [7,8,9,†] and Karsten M. Borgwardt⊠ [1,2,3,†]

Venn diagram showing the overlap between five datasets used in this study.
VariBenchSelected (10266 variants) is the part of *VariBench* not overlapping with *HumVar* nor *ExoVar*. *predictSNPSelected* (16098 variants) is the part of *predictSNP* not overlapping with *HumVar*, *ExoVar* nor *VariBench.* *SwissVarSelected* (12729 variants) is the part of *SwissVar* that does not overlap with *HumVar*, *ExoVar*, *VariBench, nor predictSNP*.



The Evaluation of Tools Used to Predict the Impact of Missense Variants Is Hindered by Two Types of Circularity

Dominik G. Grimm,☒[1,2,3] Chloé-Agathe Azencott,[1,4,5,6] Fabian Aicheler,[1,2] Udo Gieraths,[1] Daniel G. MacArthur,[7,8,9] Kaitlin E. Samocha,[7,8,9] David N. Cooper,[10] Peter D. Stenson,[10] Mark J. Daly,[7,8,9] Jordan W. Smoller,[9,11,12] Laramie E. Duncan,[7,8,9,†] and Karsten M. Borgwardt☒[1,2,3,†]

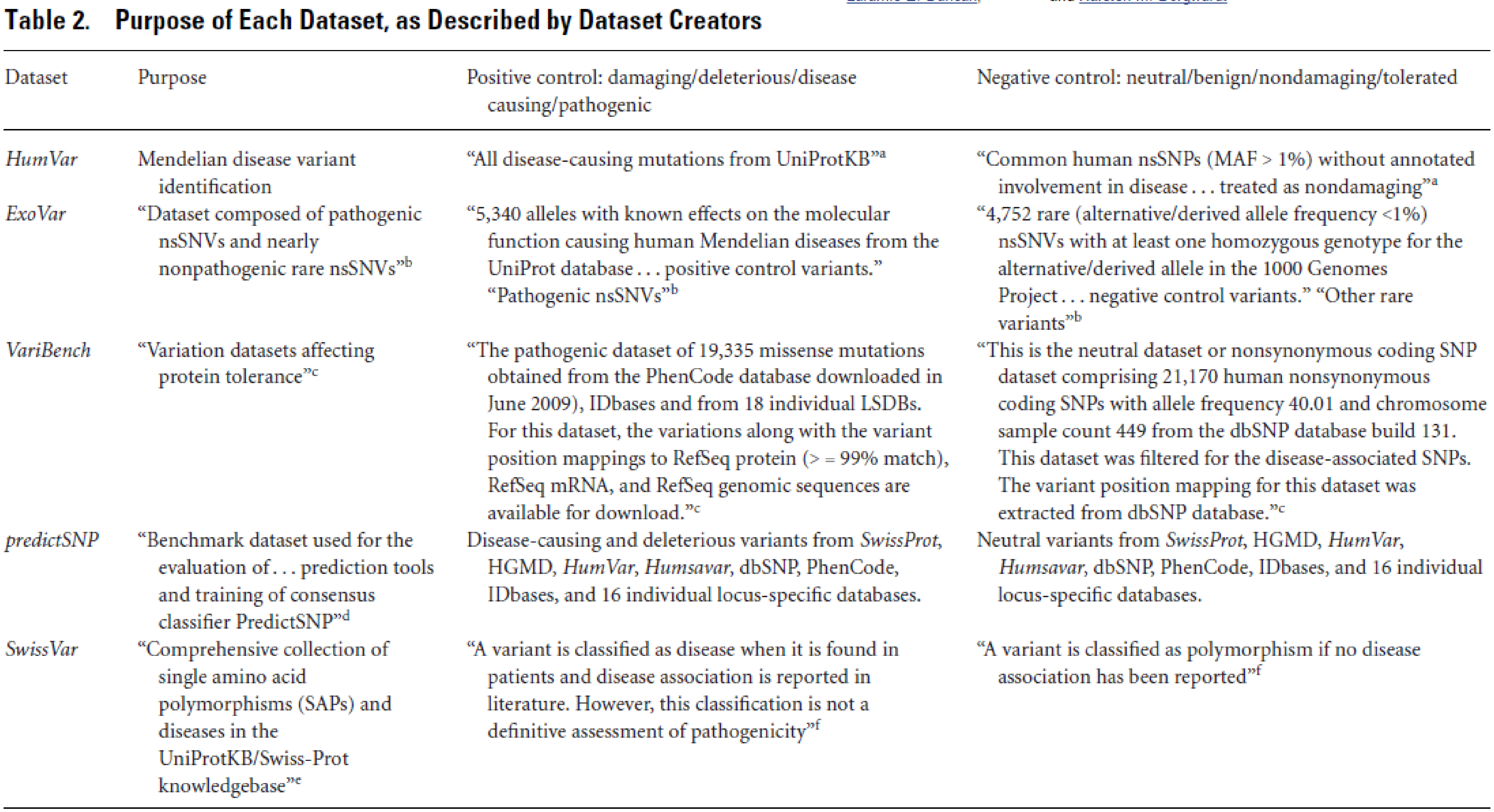


Table 2. Purpose of Each Dataset, as Described by Dataset Creators

Dataset	Purpose	Positive control: damaging/deleterious/disease causing/pathogenic	Negative control: neutral/benign/nondamaging/tolerated
HumVar	Mendelian disease variant identification	“All disease-causing mutations from UniProtKB”[a]	“Common human nsSNPs (MAF > 1%) without annotated involvement in disease . . . treated as nondamaging”[a]
ExoVar	“Dataset composed of pathogenic nsSNVs and nearly nonpathogenic rare nsSNVs”[b]	“5,340 alleles with known effects on the molecular function causing human Mendelian diseases from the UniProt database . . . positive control variants.” “Pathogenic nsSNVs”[b]	“4,752 rare (alternative/derived allele frequency <1%) nsSNVs with at least one homozygous genotype for the alternative/derived allele in the 1000 Genomes Project . . . negative control variants.” “Other rare variants”[b]
VariBench	“Variation datasets affecting protein tolerance”[c]	“The pathogenic dataset of 19,335 missense mutations obtained from the PhenCode database downloaded in June 2009), IDbases and from 18 individual LSDBs. For this dataset, the variations along with the variant position mappings to RefSeq protein (> = 99% match), RefSeq mRNA, and RefSeq genomic sequences are available for download.”[c]	“This is the neutral dataset or nonsynonymous coding SNP dataset comprising 21,170 human nonsynonymous coding SNPs with allele frequency 40.01 and chromosome sample count 449 from the dbSNP database build 131. This dataset was filtered for the disease-associated SNPs. The variant position mapping for this dataset was extracted from dbSNP database.”[c]
predictSNP	“Benchmark dataset used for the evaluation of . . . prediction tools and training of consensus classifier PredictSNP”[d]	Disease-causing and deleterious variants from *SwissProt*, HGMD, *HumVar*, *Humsavar*, dbSNP, PhenCode, IDbases, and 16 individual locus-specific databases.	Neutral variants from *SwissProt*, HGMD, *HumVar*, *Humsavar*, dbSNP, PhenCode, IDbases, and 16 individual locus-specific databases.
SwissVar	“Comprehensive collection of single amino acid polymorphisms (SAPs) and diseases in the UniProtKB/Swiss-Prot knowledgebase”[e]	“A variant is classified as disease when it is found in patients and disease association is reported in literature. However, this classification is not a definitive assessment of pathogenicity”[f]	“A variant is classified as polymorphism if no disease association has been reported”[f]

Novel Approach

Features used for classification

SEQuence-based features:

- conservation
- Δ conservation (wt *vs* mutated allele)

STRuctural feature:

- Solvent Accessible Surface Area

DYNamical features:

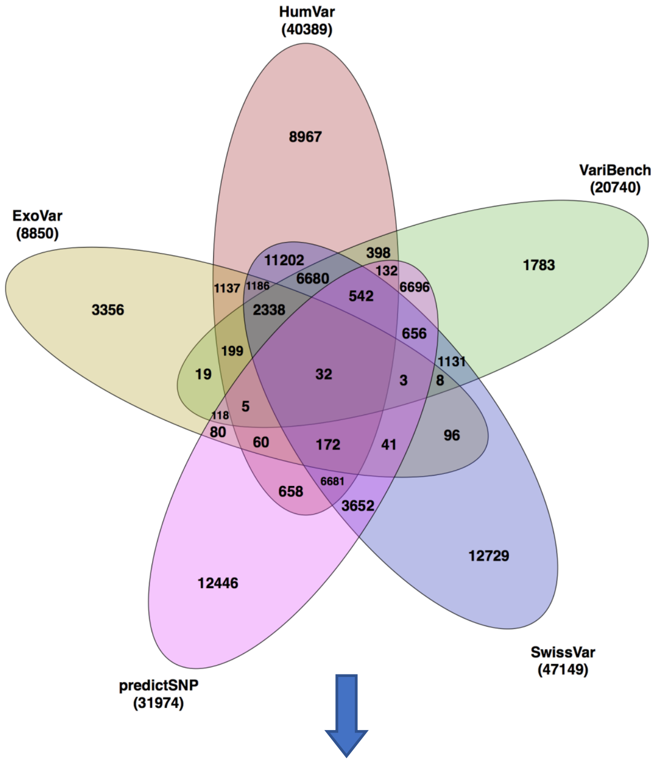
- GNM Mean Squared Fluctuations
- PRS analysis (effectors/sensors)
- Mechanical Bridging Score
- MechStiff

Random Forest classification

- trained on 20,000 annotated human variants
- 10-fold cross-validation procedure

Aims:

1. **estimate accuracy attainable by combining SEQ-STR-DYN features**
2. **quantify contribution of dynamical features**



Integrated Dataset

~ 20,000

unique SAVs with known PDB structure

Integrated Dataset

Dataset	original size [a]	SAVs with PDB structure [b]	% deleterious SAVs	% same-site SAVs [c]
HumVar (Adzhubei et al. 2010)	40,389	10,973	83.9 %	23.0 %
ExoVar (Li et al. 2013)	8,850	3,053	90.4 %	8.9 %
VariBenchSelected (Nair and Vihinen 2013)	10,266	3,286	82.3 %	40.3 %
predictSNPSelected (Bendl et al. 2014)	16,098	3,893	85.4 %	10.3 %
SwissVarSelected (Mottaz et al. 2010)	12,729	2,033	38.2 %	2.4 %
Union of all datasets [d]	-	**20,413**	**78.4 %**	18.6 %

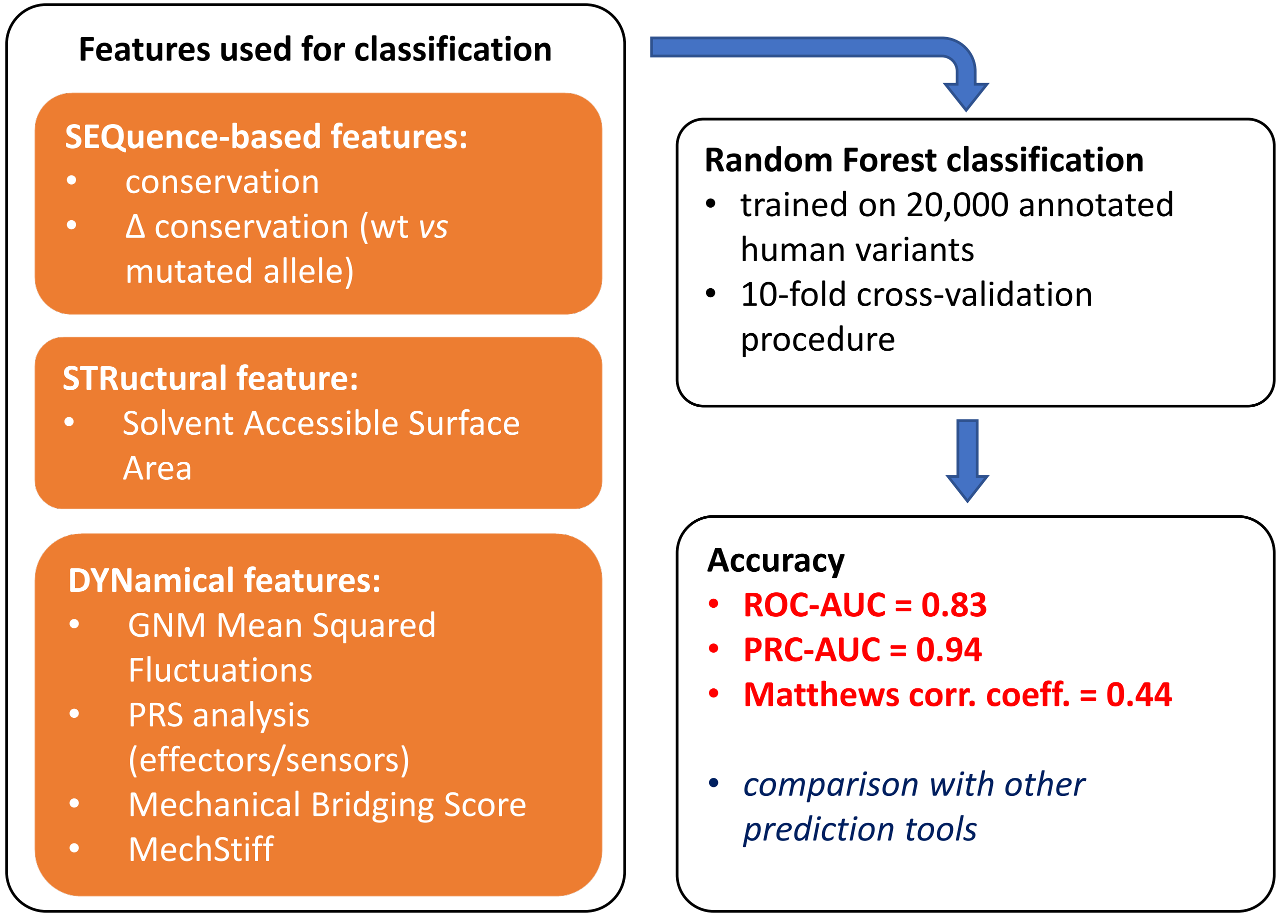
[a] The original 5 datasets have been extracted from (13). The three "Selected" datasets have been cleared from SAVs already present in HumVar and ExoVar.

[b] Only the SAVs in proteins for which a PDB structure has been reported (according to Uniprot website) have been considered. In parenthesis, we show the number of SAVs used in our analysis, after excluding duplicates and the cases where structural data were insufficient to compute all DYN features.

[c] Percentage of SAVs for which at least one other variant at the same sequence position, but with different substitutions, is reported in the dataset. Such same-site variants (e.g. S100A and S100R in given protein) are distinguished by SEQ features only. For this reason, for training/testing of the DYN classifier, we retained only a single representative for each group of same-site variants.

[d] When combining the five datasets, duplicates have been eliminated.

Novel Approach



Features used for classification

SEQuence-based features:

- conservation
- Δ conservation (wt *vs* mutated allele)

STRuctural feature:

- Solvent Accessible Surface Area

DYNamical features:

- GNM Mean Squared Fluctuations
- PRS analysis (effectors/sensors)
- Mechanical Bridging Score
- MechStiff

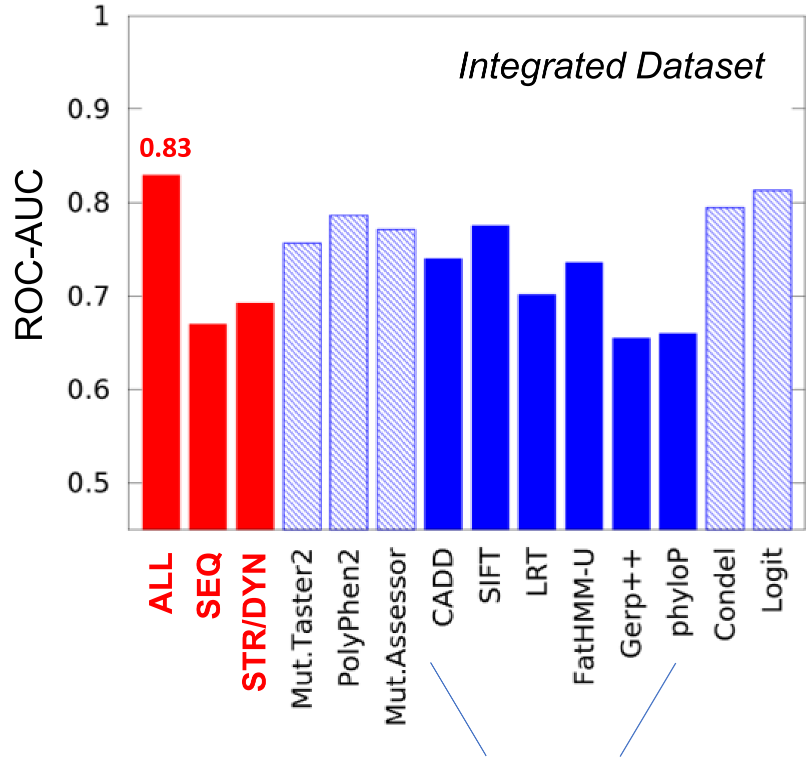
Random Forest classification

- trained on 20,000 annotated human variants
- 10-fold cross-validation procedure

Accuracy

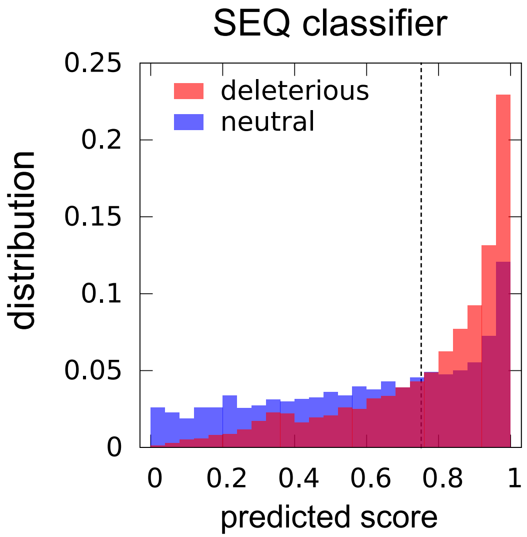
- **ROC-AUC = 0.83**
- **PRC-AUC = 0.94**
- **Matthews corr. coeff. = 0.44**

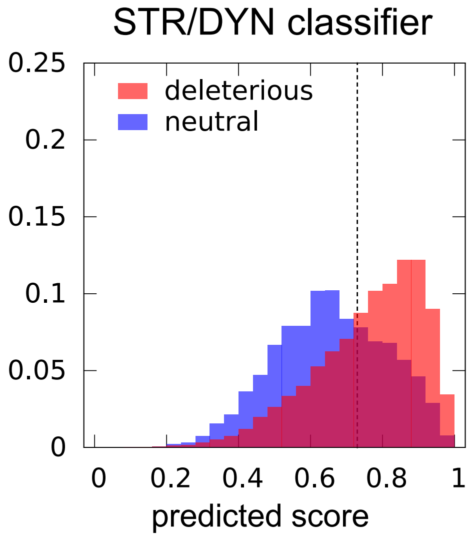
- *comparison with other prediction tools*

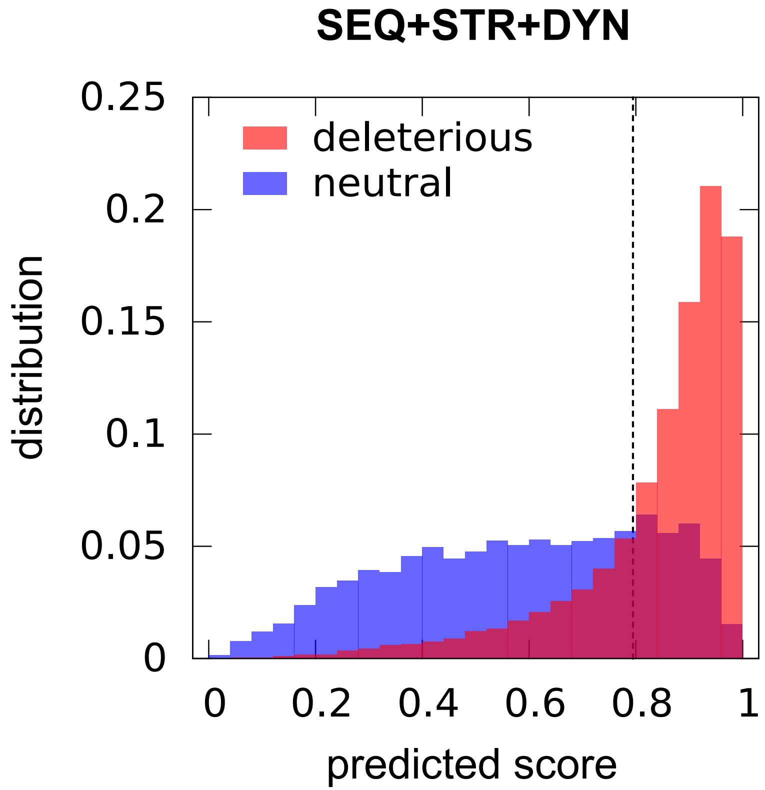


training bias

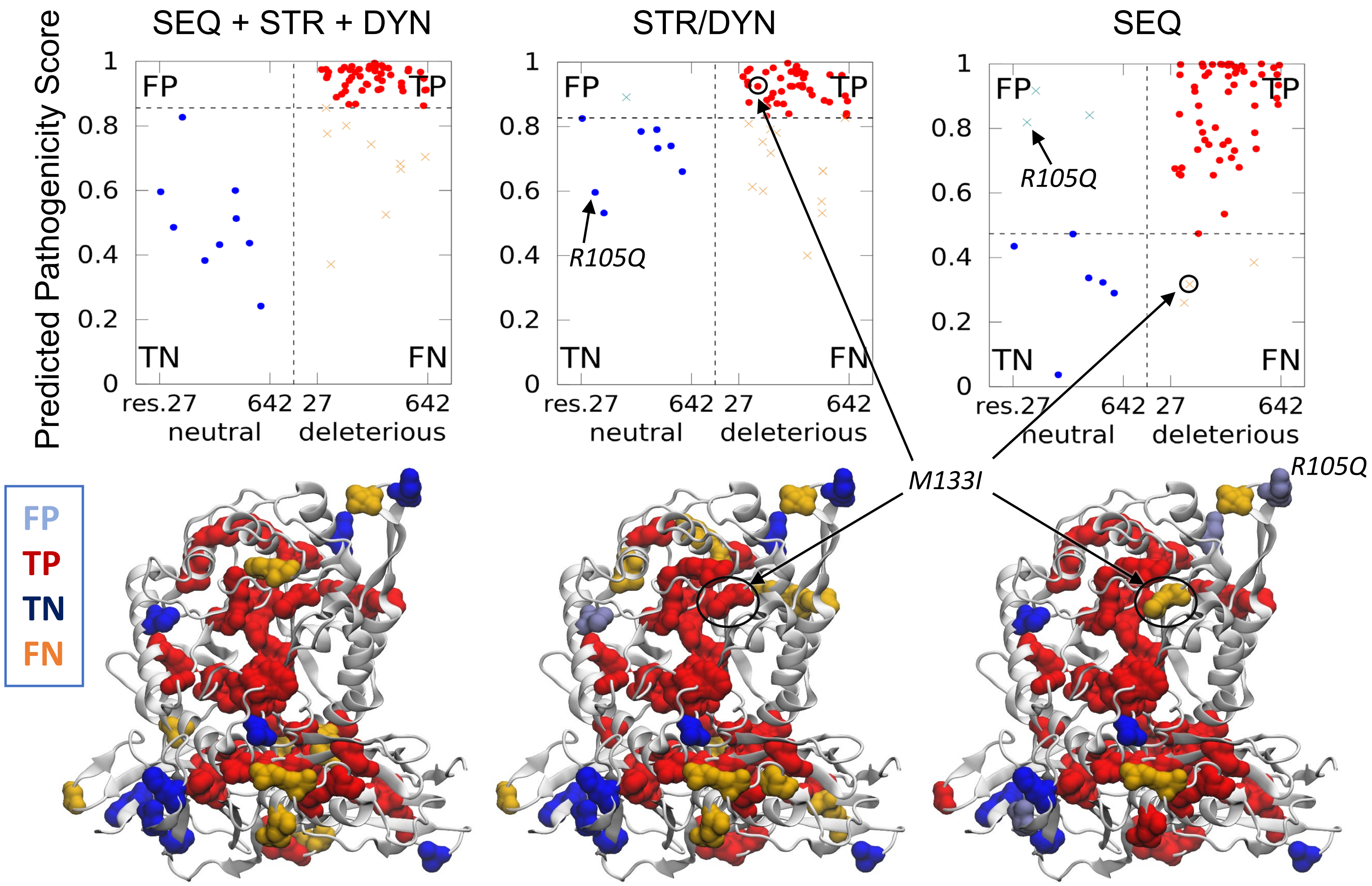
Increased accuracy by combining SEQ + STR + DYN features

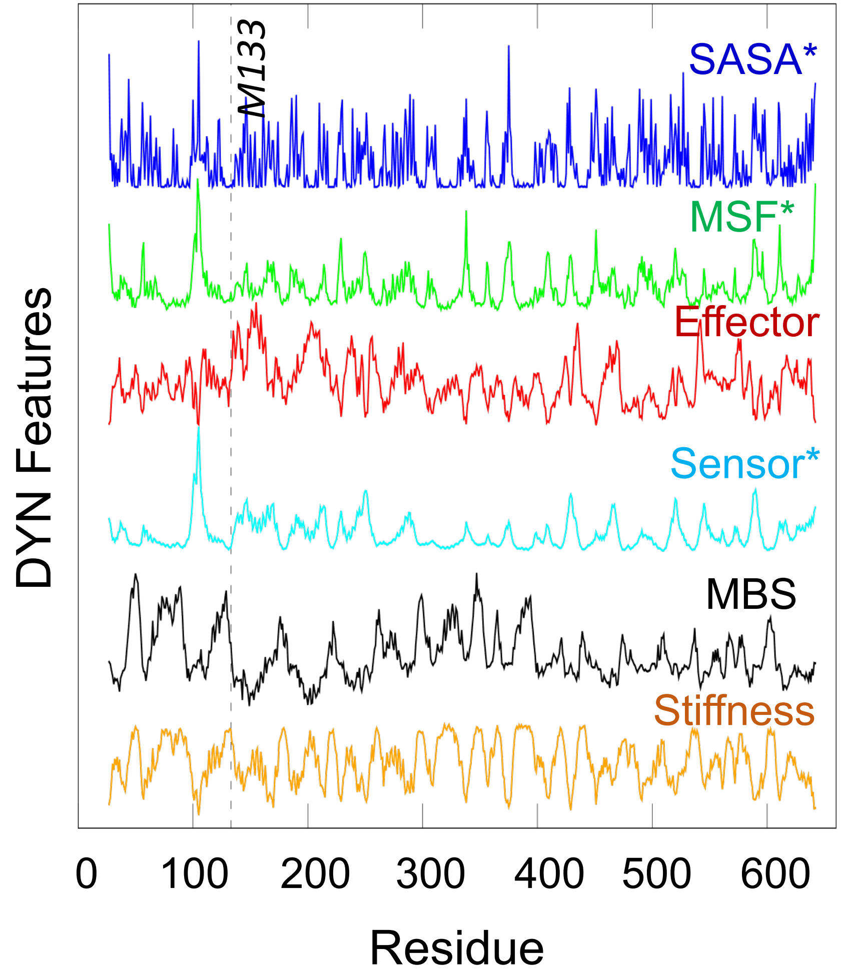

SEQ classifier


STR/DYN classifier


SEQ+STR+DYN

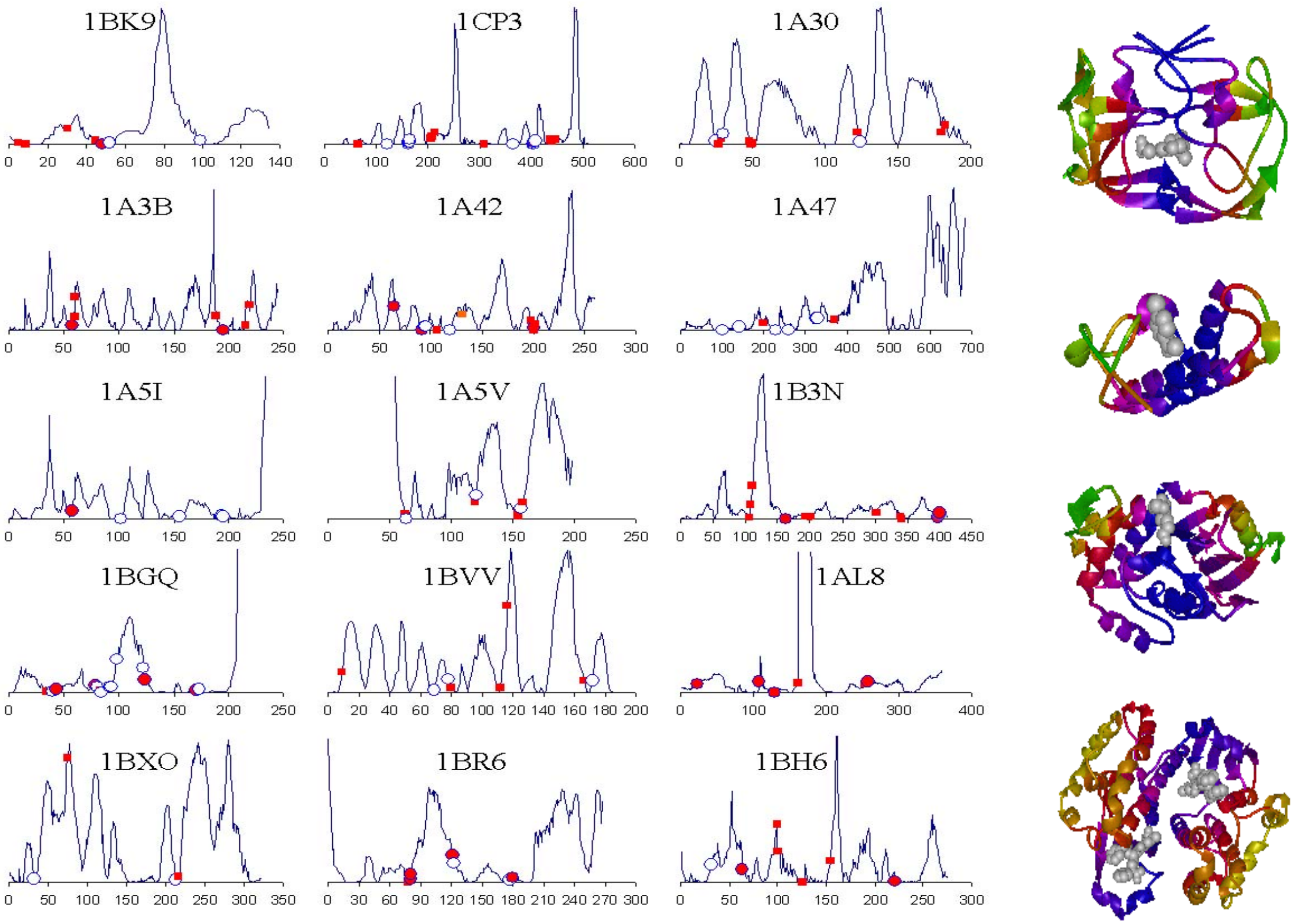
Example: human α-L-iduronidase





Coupling between global mechanics & catalysis

Catalytic sites coincide/communicate with global hinge centers



Global mode shapes for 15 PDB structures. Residues forming the catalytic active sites are marked as (**O**), inhibitors binding sites as (■), and both as (●).

Lee-Wei Yang & Bahar (2005) Structure 13, 893-904.

Novel Approach

Features used for classification

SEQuence-based features:
- conservation
- Δ conservation (wt *vs* mutated allele)

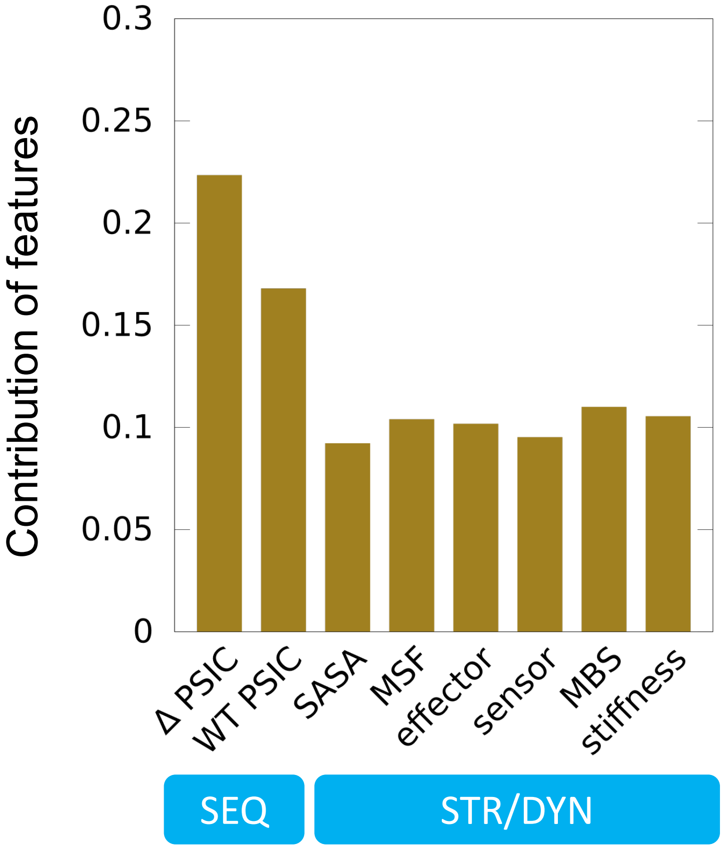
STRuctural feature:
- Solvent Accessible Surface Area

DYNamical features:
- GNM Mean Squared Fluctuations
- PRS analysis (effectors/sensors)
- Mechanical Bridging Score
- MechStiff

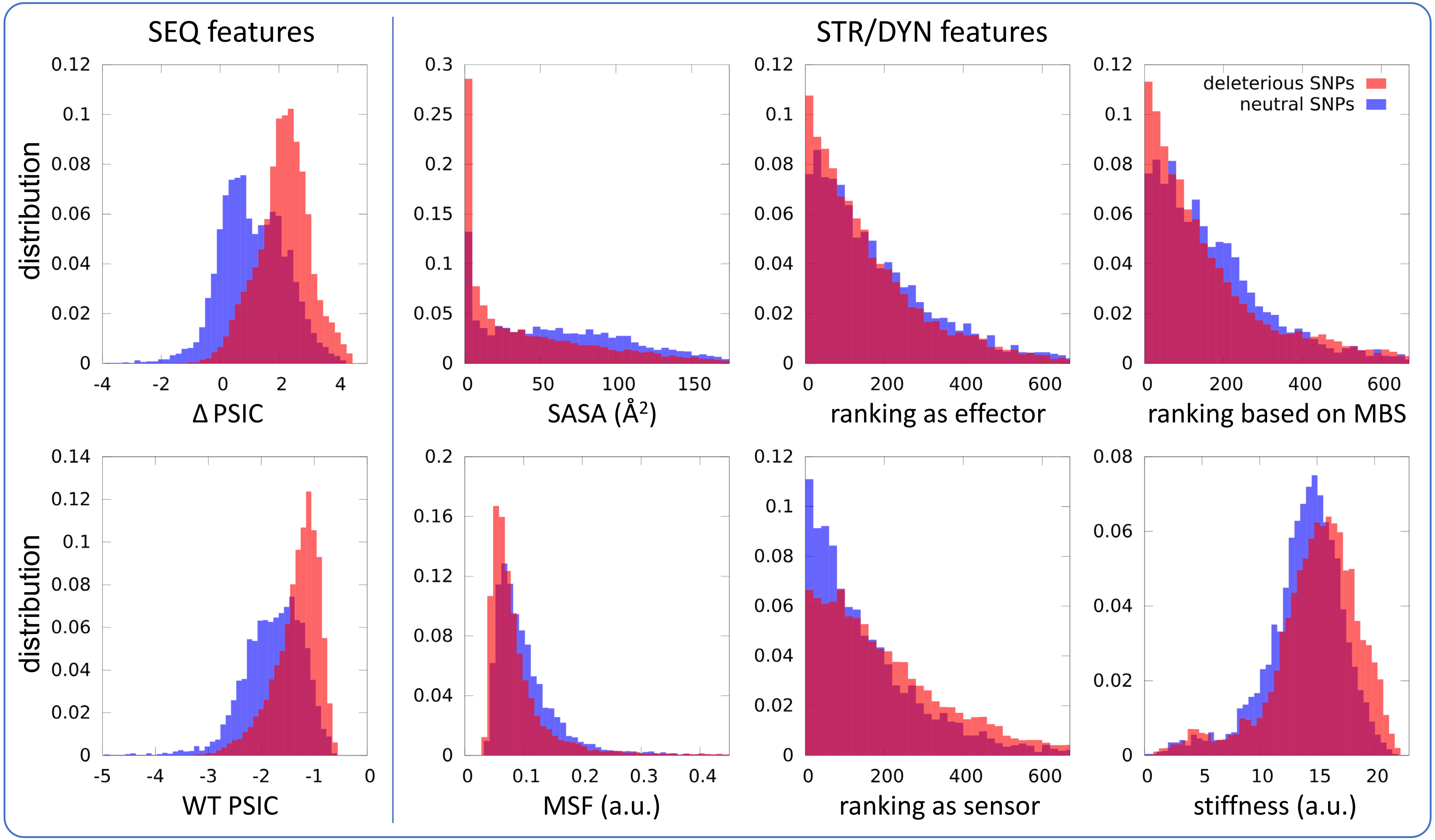
Random Forest classification
- trained on 20,000 annotated human variants
- 10-fold cross-validation procedure

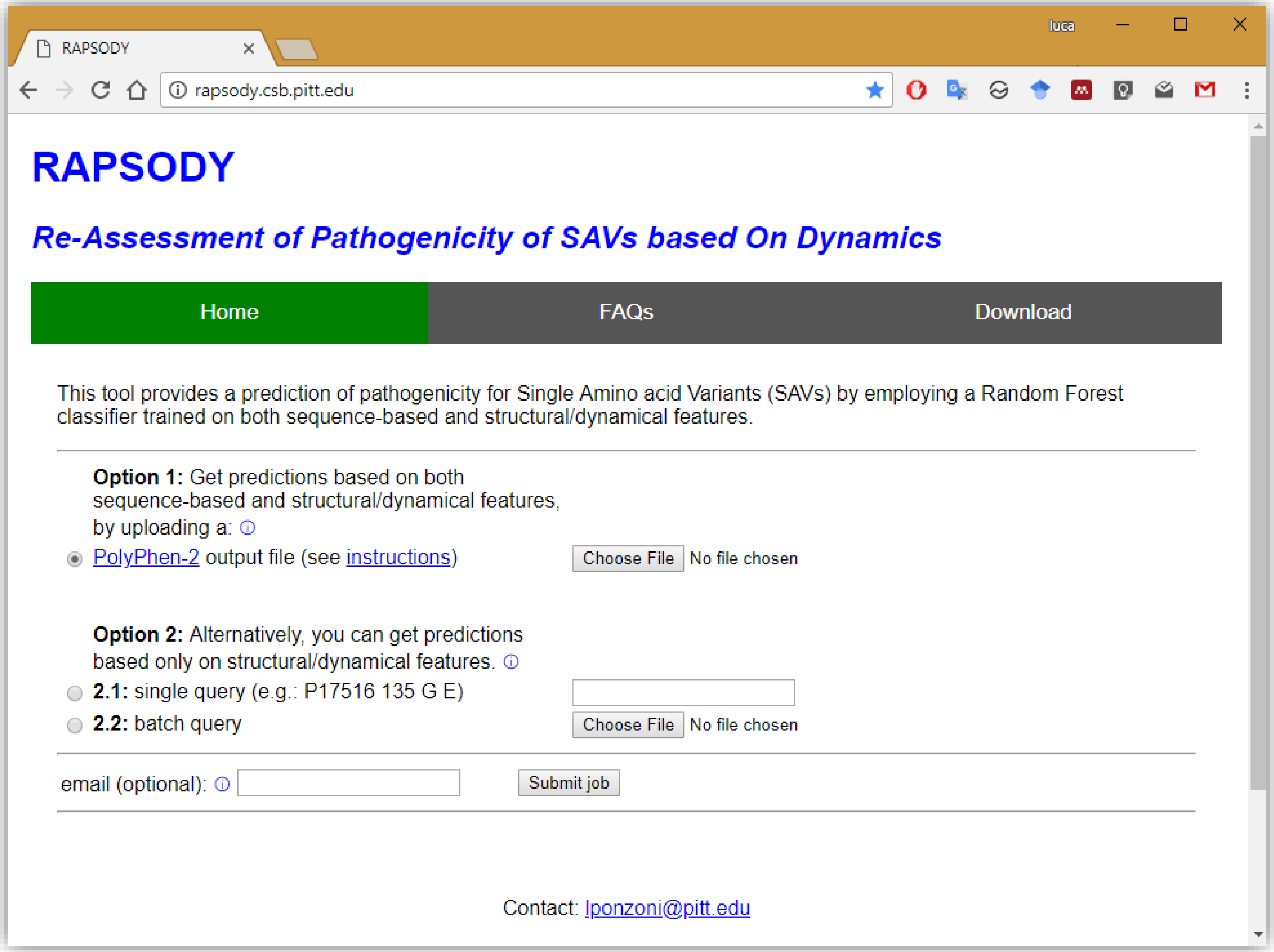
Aims:

1. estimate accuracy attainable by combining SEQ-STR-DYN features
2. **quantify contribution of dynamical features**



Discriminatory power of individual features





RAPSODY

Re-Assessment of Pathogenicity of SAVs based On Dynamics

Home | FAQs | Download

This tool provides a prediction of pathogenicity for Single Amino acid Variants (SAVs) by employing a Random Forest classifier trained on both sequence-based and structural/dynamical features.

Option 1: Get predictions based on both sequence-based and structural/dynamical features, by uploading a: ⓘ

● PolyPhen-2 output file (see instructions) [Choose File] No file chosen

Option 2: Alternatively, you can get predictions based only on structural/dynamical features. ⓘ

○ **2.1:** single query (e.g.: P17516 135 G E) [ ]

○ **2.2:** batch query [Choose File] No file chosen

email (optional): ⓘ [ ] [Submit job]

Contact: lponzoni@pitt.edu

ProDy

Protein Dynamics Analysis in Python