Imperial College London

Using SuSPect to Predict the Phenotypic Effects of Missense Variants

Chris Yates
UCL Cancer Institute
c.yates@ucl.ac.uk

Imperial College London

Outline

- SAVs and Disease
- Development of SuSPect
  - Features included
  - Feature selection
  - Performance
- Web-Server & Availability
  - Usage
  - Example results

Imperial College London

Outline

- SAVs and Disease
- Development of SuSPect
  - Features included
  - Feature selection
  - Performance
- Web-Server & Availability
  - Usage
  - Example results

Imperial College London

Background

- 10-15,000 single amino acid variants (SAVs) per exome.

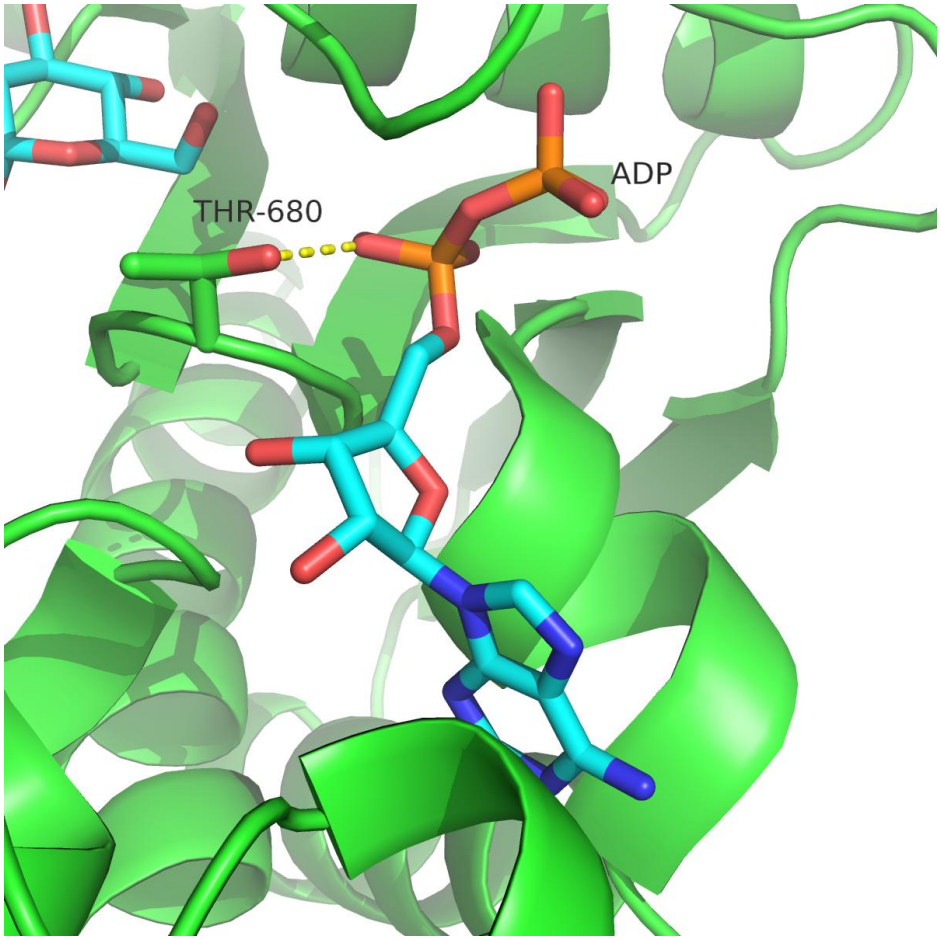
- Many variants are tolerated, but some SAVs cause disease.
  - Glu6Val in HBB causes sickle cell anæmia.

- Many mechanisms by which SAVs can impair function.
  - Decrease stability,
  - Change active site,
  - Protein-protein interaction.

- Need methods for predicting SAV effects
  - Sequence- and structure-based.

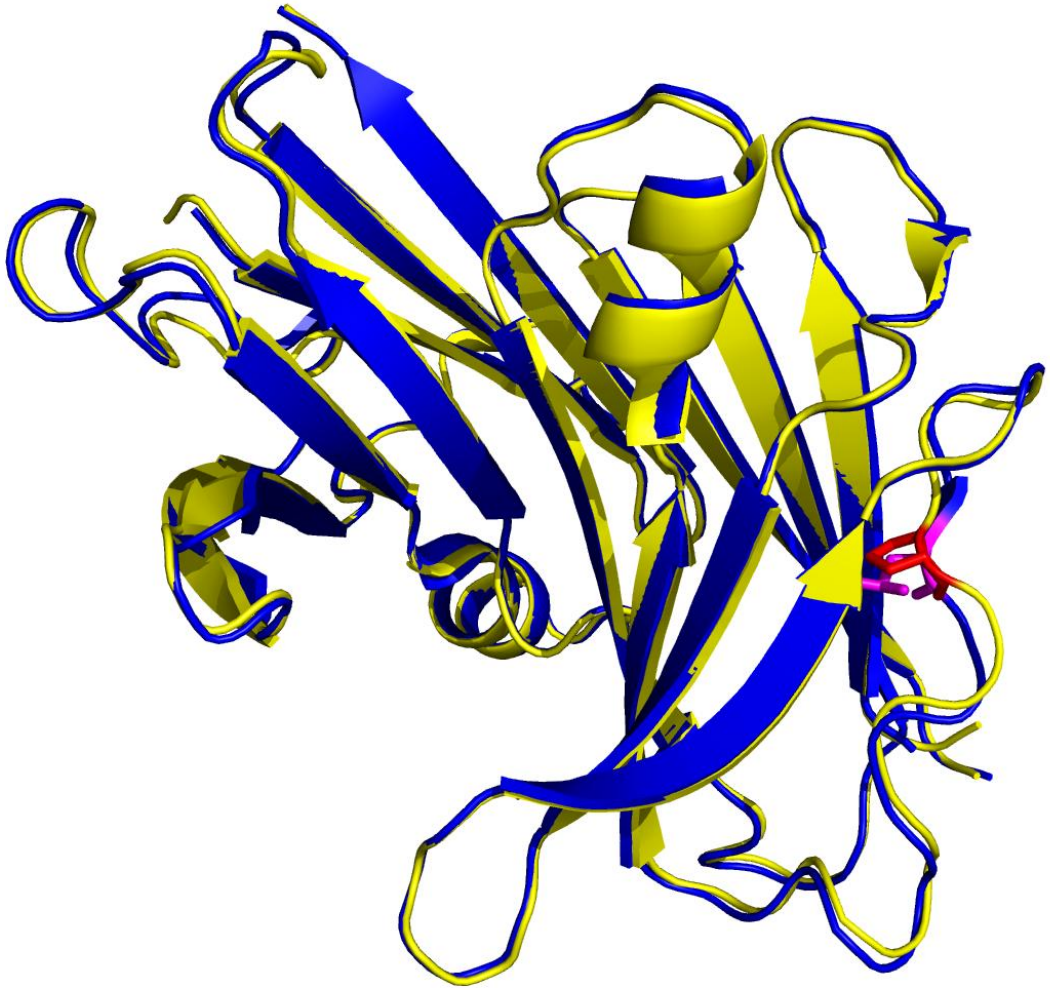
Imperial College
London

Hexokinase



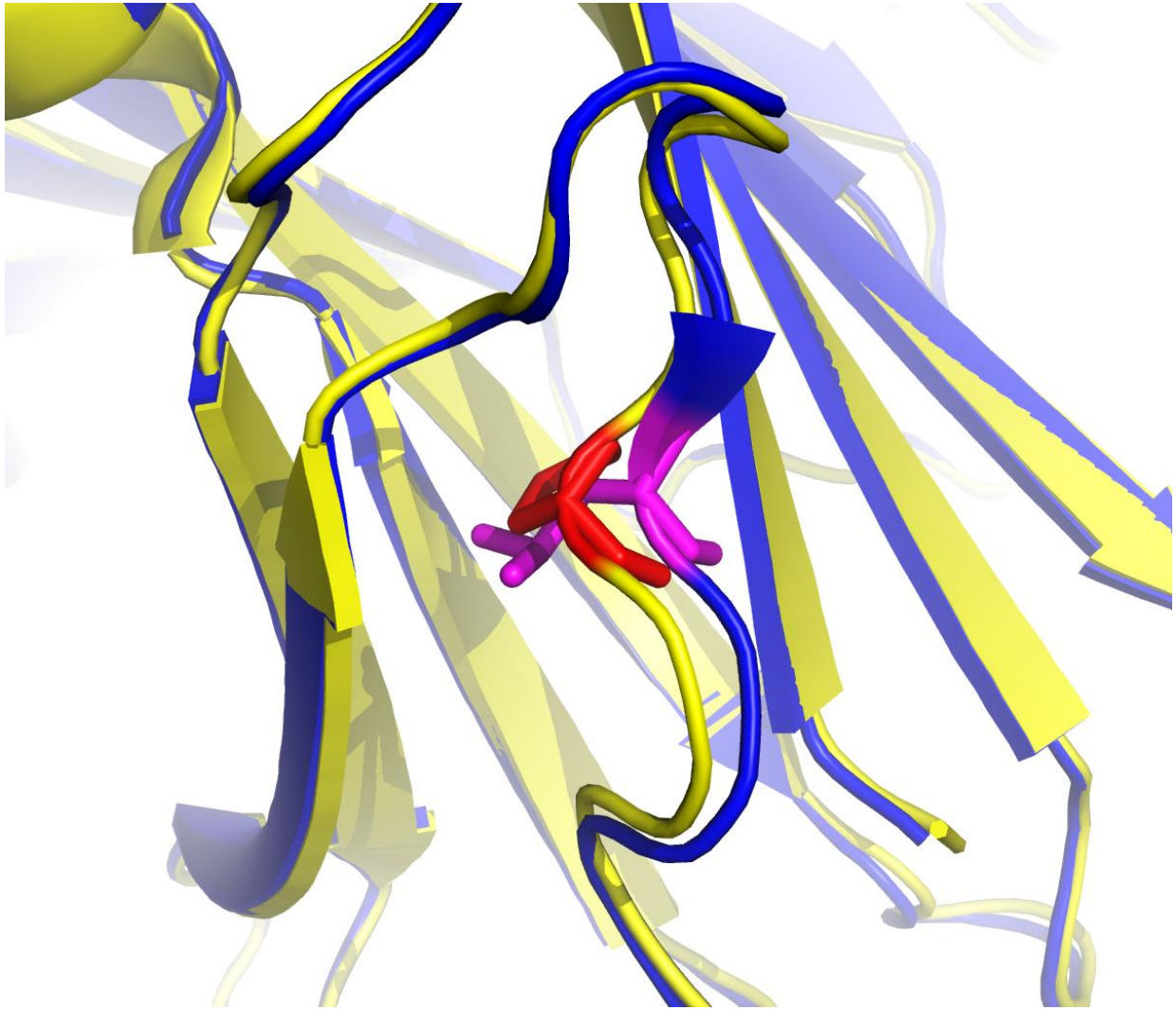
Imperial College
London

Transthyretin



Imperial College
London

Transthyretin



Imperial College London

Outline

- SAVs and Disease
- Development of SuSPect
  - Features included
  - Feature selection
  - Performance
- Web-Server & Availability
  - Usage
  - Example results

Imperial College London

Features

Sequence conservation

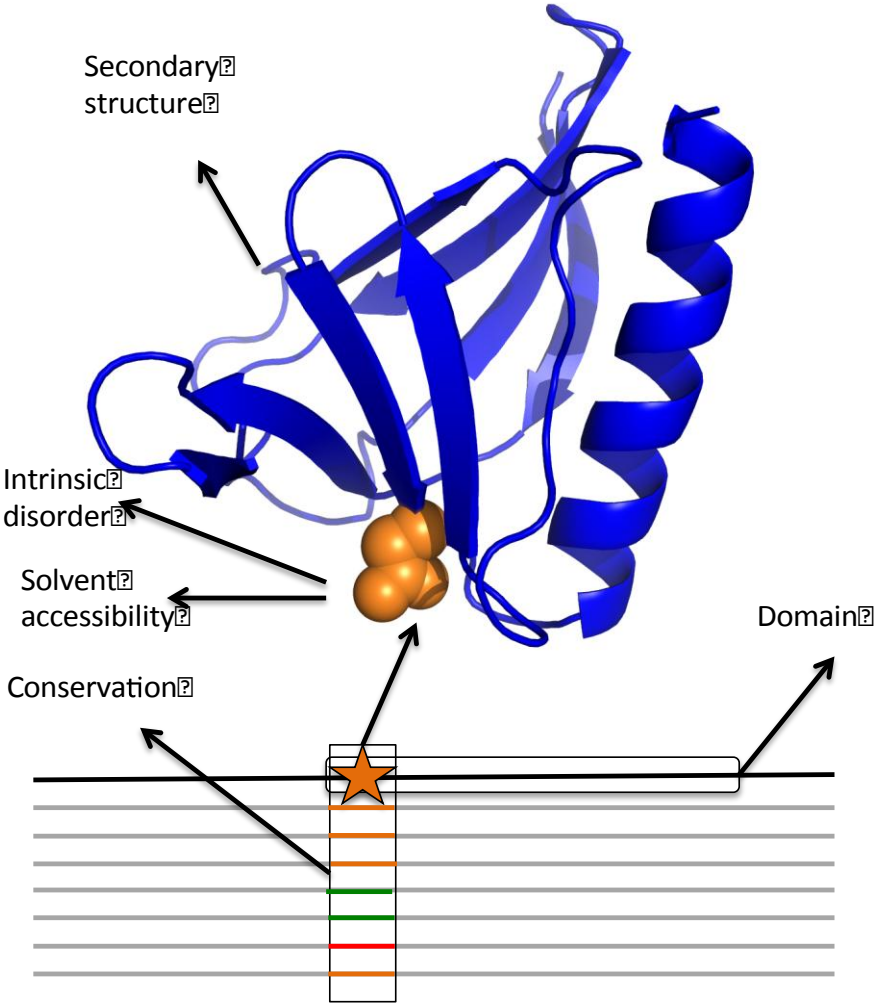
- Position-specific scoring matrix (PSI-BLAST)
- Pfam domain
- Jensen-Shannon divergence

Structural features

- From PDB or Phyre2 homology models where available.
- Secondary structure
- Solvent accessibility

Network features

- Protein-protein interaction (PPI)
- Domain-domain interaction (DDI)
- Domain bigram



Imperial College London

Features

Sequence conservation

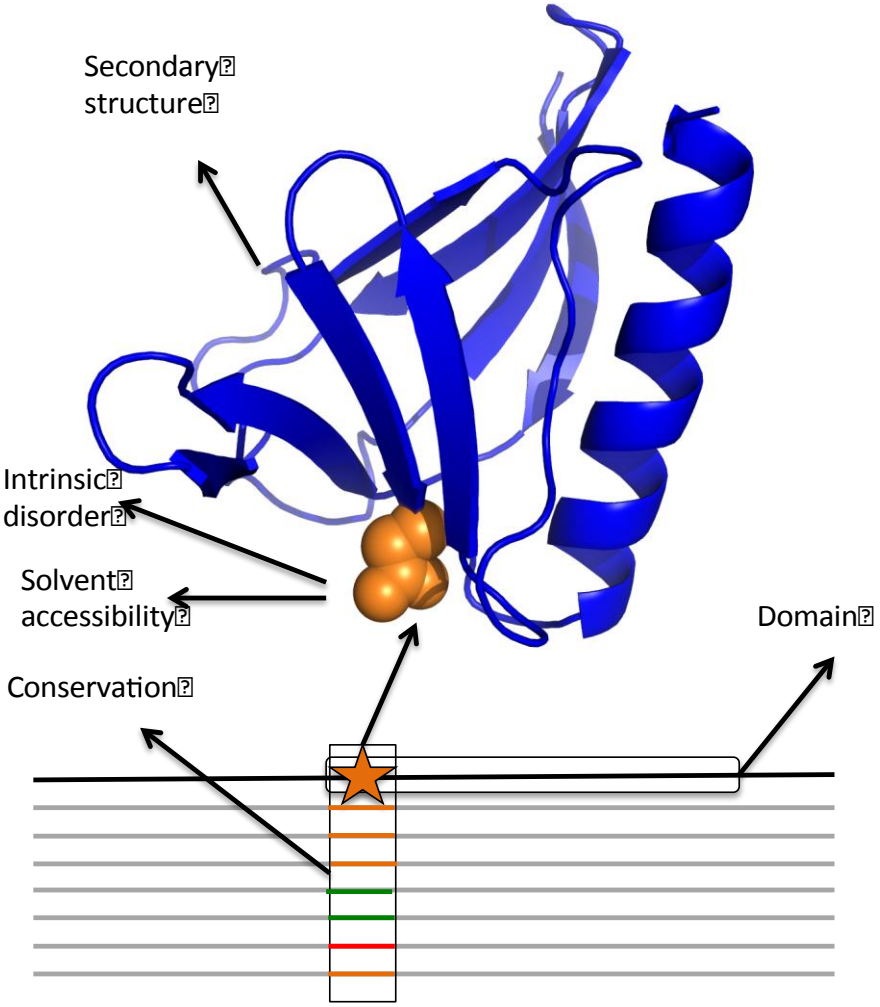
- Position-specific scoring matrix (PSI-BLAST)
- Pfam domain
- Jensen-Shannon divergence

Structural features

- From PDB or Phyre2 homology models where available.
- Secondary structure
- Solvent accessibility

Network features

- Protein-protein interaction (PPI)
- Domain-domain interaction (DDI)
- Domain bigram



Imperial College London

Features

Sequence conservation

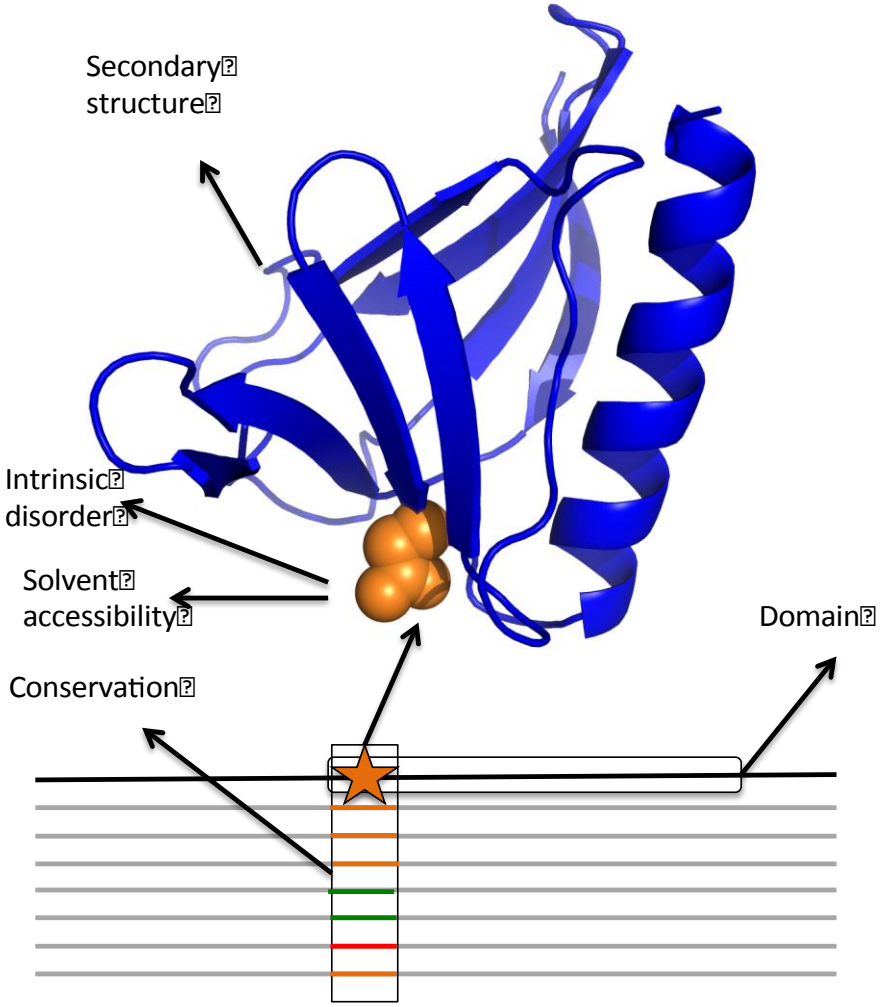
- Position-specific scoring matrix (PSI-BLAST)
- Pfam domain
- Jensen-Shannon divergence

Structural features

- From PDB or Phyre2 homology models where available.
- Secondary structure
- Solvent accessibility

Network features

- Protein-protein interaction (PPI)
- Domain-domain interaction (DDI)



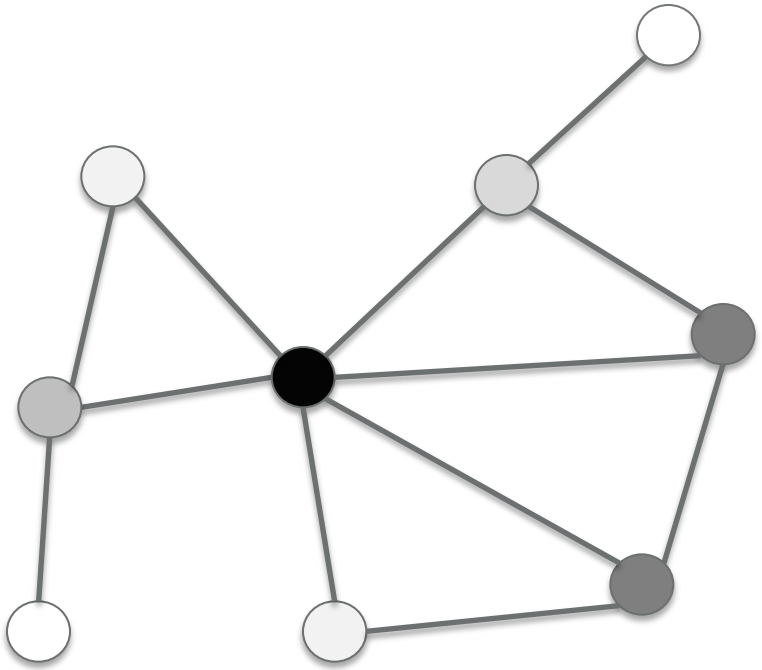
Imperial College London

Network Features

Change in protein function is not the same as causing disease.

More ‘important’ proteins are more likely to be involved in disease.

Centrality of a protein within a protein-protein interaction network can be used to measure ‘importance’.



Imperial College London

VariBench

Neutral and Pathogenic datasets obtained from VariBench (Thusberg *et al.* 2011).

Neutral SAVs from dbSNP version 131, filtered by allele frequency (>0.01) and chromosome count (>49).

- SAVs present in OMIM removed.

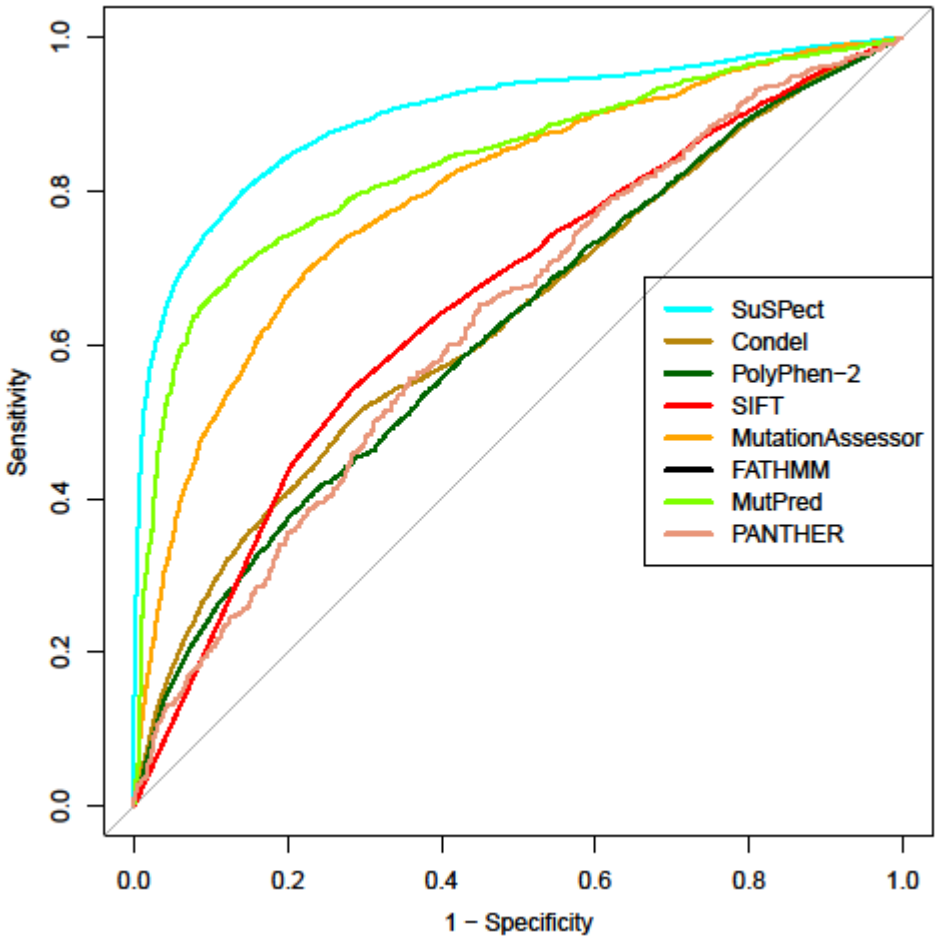
Pathogenic SAVs from PhenCode (2009).

VariBench datasets were filtered to remove any SAVs present in training data.

13,236	Neutral
5,397	Pathogenic

Imperial College London

VariBench



Method	AUC	Balanced Accuracy
SuSPect	**0.90**	**0.82**
MutPred	0.84	0.75
MutationAssessor	0.79	0.70
SIFT	0.65	0.63
FATHMM	0.63	0.63
Condel	0.63	0.61
PANTHER	0.63	0.59
PolyPhen-2	0.62	0.58

Imperial College London

Results – Take home messages

Feature selection improves performance

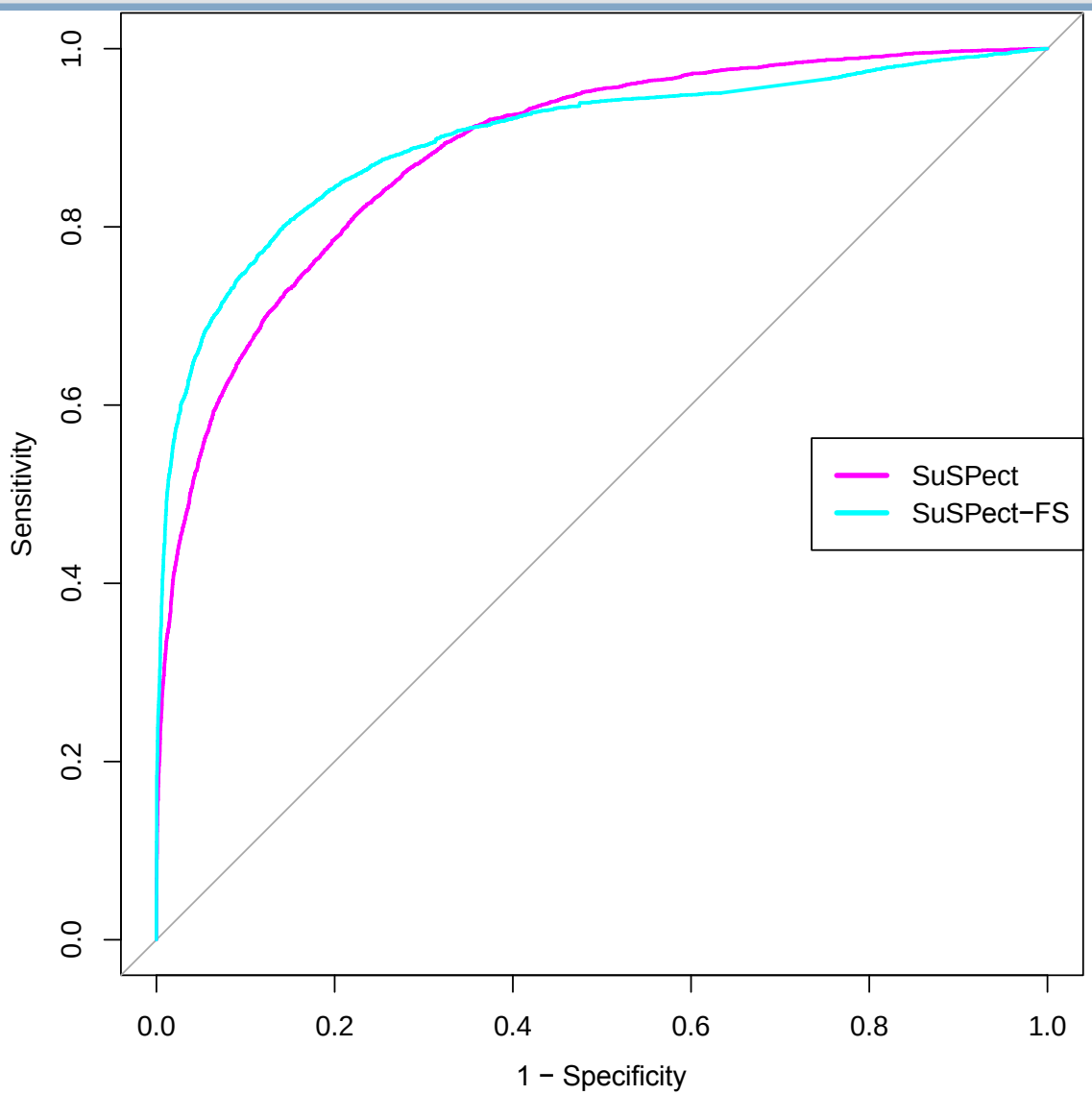
- Top 9 features selected.
  - Predicted relative solvent accessibility;
  - WT and Variant scores in PSSM, and their difference;
  - Number of UniProt annotations;
  - Difference in Pfam scores;
  - PPI network degree centrality;
  - Jensen-Shannon divergence;
  - Sequence identity with best-matching sequence to lack WT amino acid.

Network features are important

- Removal of network features drops AUC from 0.88 to 0.78.
- Removal of PPI centrality from SuSPect-FS gives drop from 0.90 to 0.74.
- Network centrality helps show the difference between variants affecting protein function and leading to disease.

Imperial College London

Results – Feature Selection



Imperial College London

Results – Take home messages

Feature selection improves performance

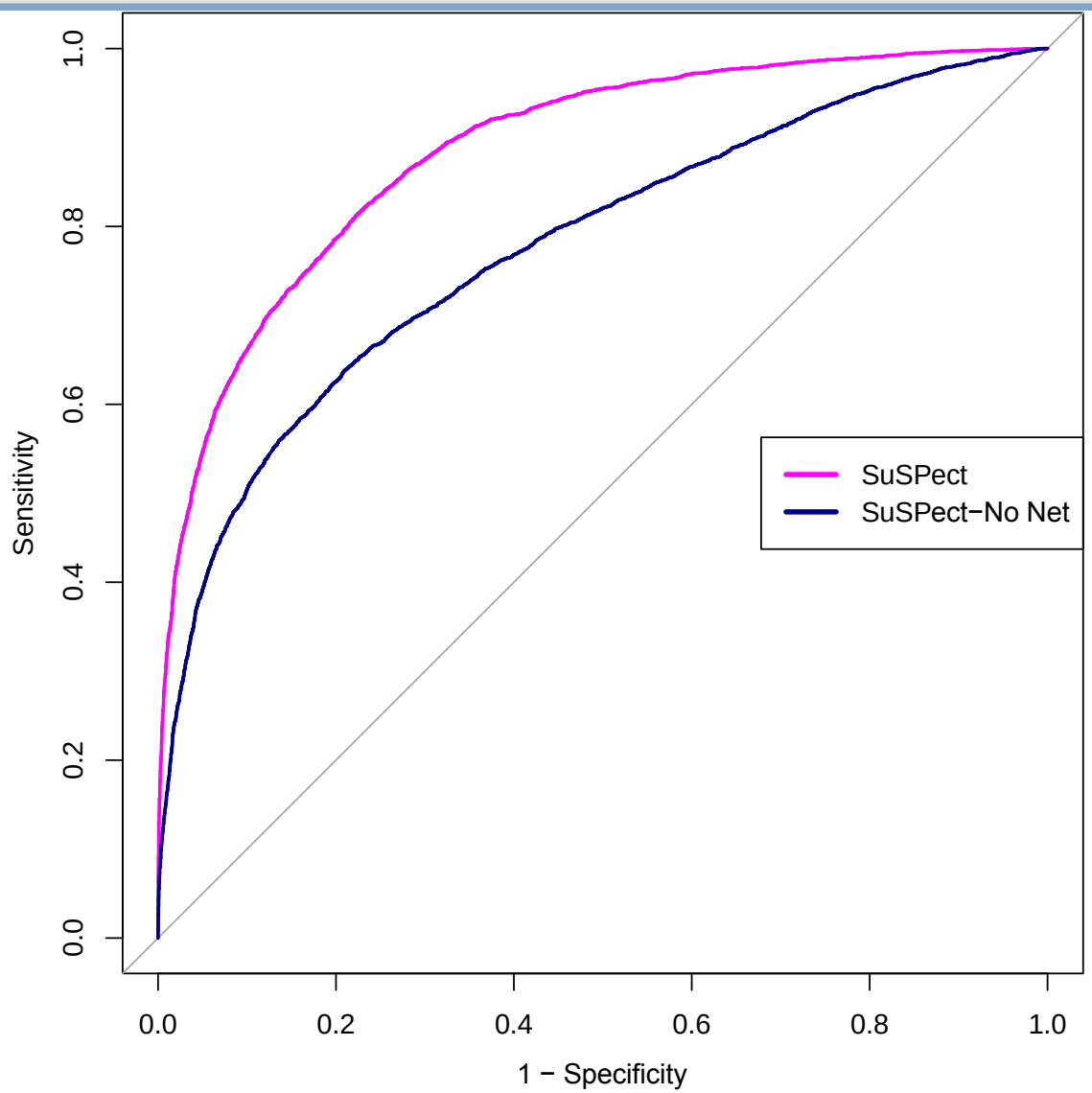
- Top 9 features selected.
  - Predicted relative solvent accessibility;
  - WT and Variant scores in PSSM, and their difference;
  - Number of UniProt annotations;
  - Difference in Pfam scores;
  - PPI network degree centrality;
  - Jensen-Shannon divergence;
  - Sequence identity with best-matching sequence to lack WT amino acid

Network features are important

- Removal of network features drops AUC from 0.88 to 0.78.
- Removal of PPI centrality from SuSPect-FS gives drop from 0.90 to 0.74.
- Network centrality helps show the difference between variants affecting protein function and leading to disease.

Imperial College London

Results – No Network Features



Imperial College London

Results – Take home messages

Feature selection improves performance

- Top 9 features selected.
  - Predicted relative solvent accessibility;
  - WT and Variant scores in PSSM, and their difference;
  - Number of UniProt annotations;
  - Difference in Pfam scores;
  - PPI network degree centrality;
  - Jensen-Shannon divergence;
  - Sequence identity with best-matching sequence to lack WT amino acid

Network features are important

- Removal of network features drops AUC from 0.88 to 0.78.
- Removal of PPI centrality from SuSPect-FS gives drop from 0.90 to 0.74.
- Network centrality helps show the difference between variants affecting protein function and leading to disease.

Imperial College London

Results - Prokaryotic Mutations

HIV-1 protease – Loeb *et al.* (1989)

- 225 deleterious
- 111 neutral

LacI repressor – Suckow *et al.* (1996)

- 1,774 deleterious
- 2,267 neutral

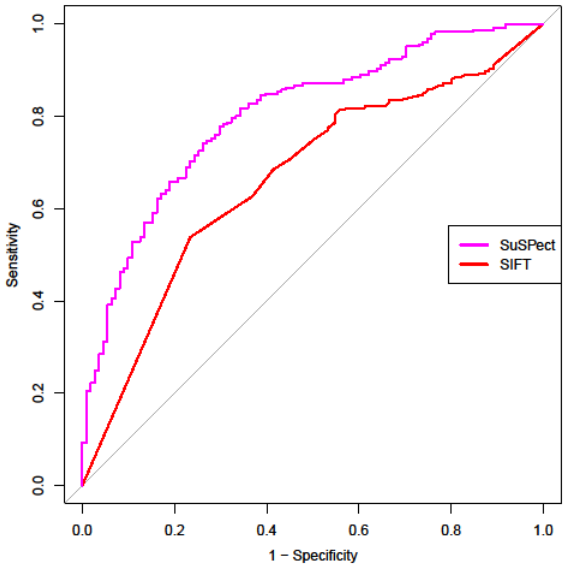
T4 lysozyme – Rennel *et al.* (1991)

- 638 deleterious
- 1,377 neutral

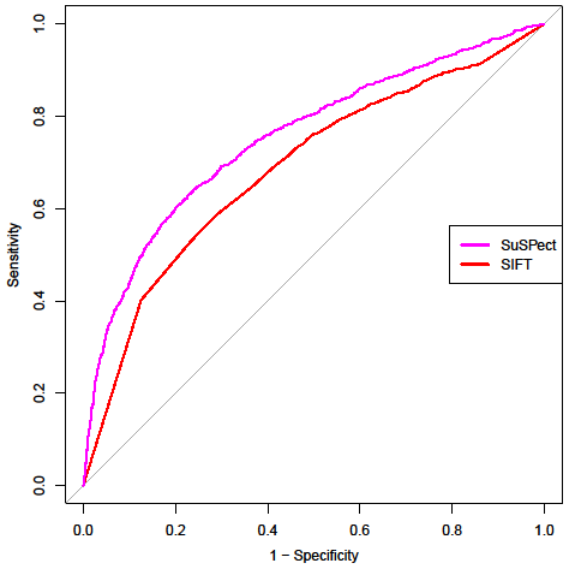
Imperial College London

Results - Prokaryotic Mutations

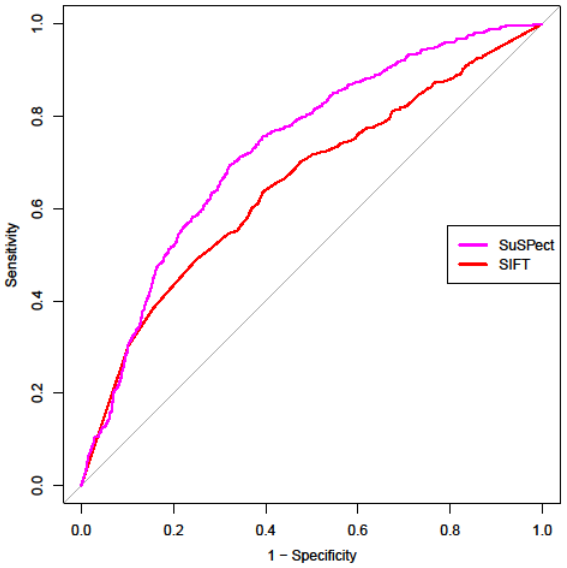
HIV-1 Protease



E. coli LacI repressor



T4 Lysozyme



Imperial College London

Outline

- SAVs and Disease
- Development of SuSPect
  - Features included
  - Feature selection
  - Performance
- Web-Server & Availability
  - Usage
  - Example results

Imperial College London

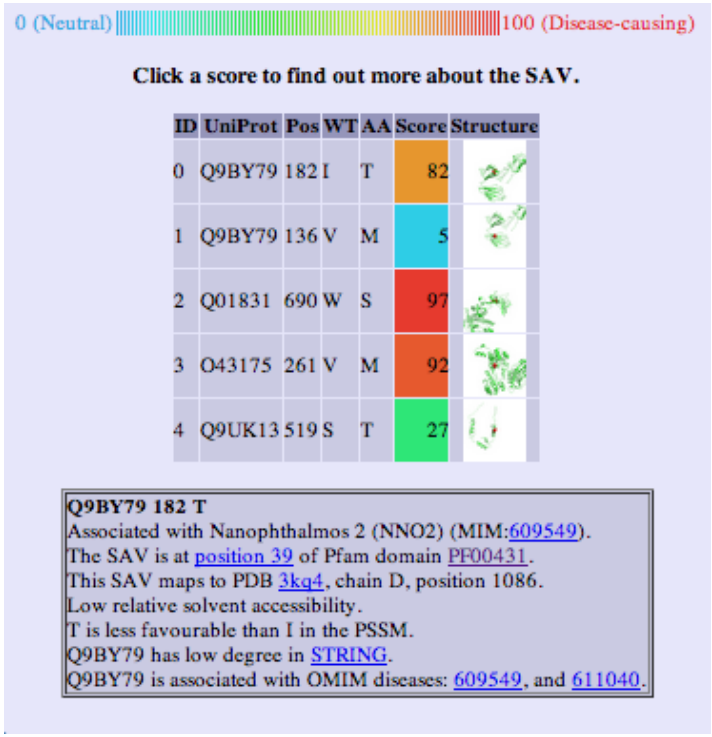
Web-Server & Download

Available at www.sbg.bio.ic.ac.uk/suspect

Upload list of SAVs or VCF file to obtain scores for human missense variants

- In addition to score, gives easily interpretable descriptions.
- Sequence conservation, structure, active site, and much more.
- Useful for interpretation of how variants can have their effects.

SuSPect Package – downloadable database of pre-calculated scores for all possible human missense variants.

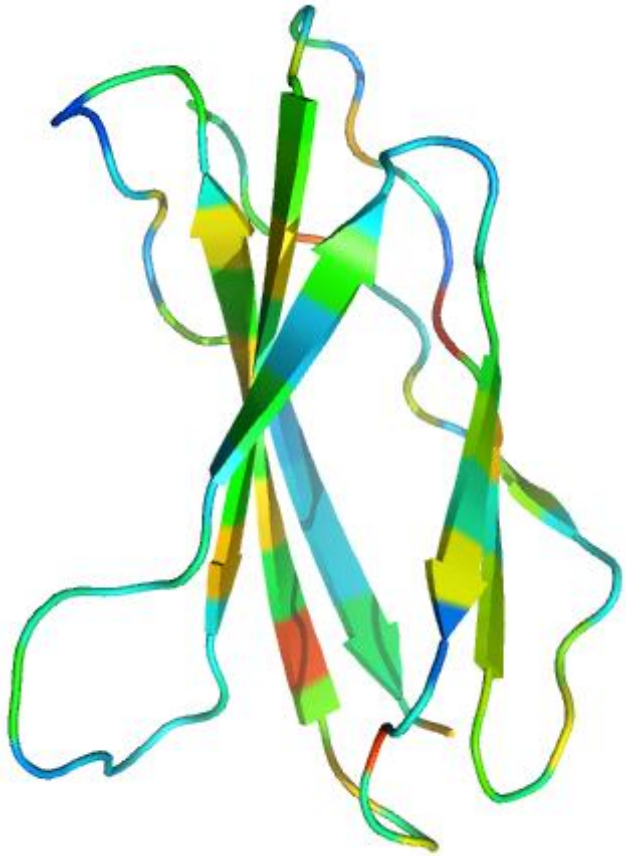


Imperial College London

Web-Server & Download

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
11	D	17	34	4	14	24	20	26	32	22	32	35	13	18	21	18	16	20	22	44	28	D
12	V	40	63	53	45	63	61	62	32	52	39	48	52	47	49	40	42	41	26	65	63	V
13	T	62	82	51	57	79	43	76	73	61	67	66	41	66	57	54	37	4	68	80	72	T
14	D	24	56	5	17	47	31	29	46	30	41	45	27	18	33	37	18	25	42	39	38	D
15	T	52	64	27	45	60	53	43	53	42	62	66	31	54	48	41	26	4	57	75	52	T
16	T	45	75	52	45	65	49	57	65	52	65	66	51	54	49	54	18	29	48	72	58	T
17	A	31	73	79	53	39	54	73	20	69	19	26	64	56	67	66	56	54	22	79	53	A
18	L	29	42	30	20	29	39	24	30	27	22	29	23	36	24	19	18	14	25	40	29	L
19	I	61	80	88	83	56	82	87	4	84	30	65	86	79	84	70	77	63	25	87	76	I
20	T	33	65	49	29	66	36	34	48	38	47	43	33	53	34	30	21	28	43	59	52	T
21	W	97	98	97	97	87	96	97	95	98	92	97	98	95	98	97	97	96	92	27	90	W
22	F	12	22	6	6	4	13	12	14	9	12	16	8	12	8	12	5	5	13	31	14	F
23	K	18	52	34	25	43	31	25	28	4	28	41	23	9	19	20	24	27	23	46	41	K
24	P	56	86	74	60	81	72	74	78	69	53	76	78	25	67	76	53	67	61	88	84	P

Launch JSmol | Generate Image | Clear All



Imperial College London

Web-Server & Download

Human Proteins

• Scores have been pre-calculated for the Mar-2013 release of UniProt.

• If human variants or proteins are uploaded (either as sequence, structure or ID), these pre-calculated scores are used.

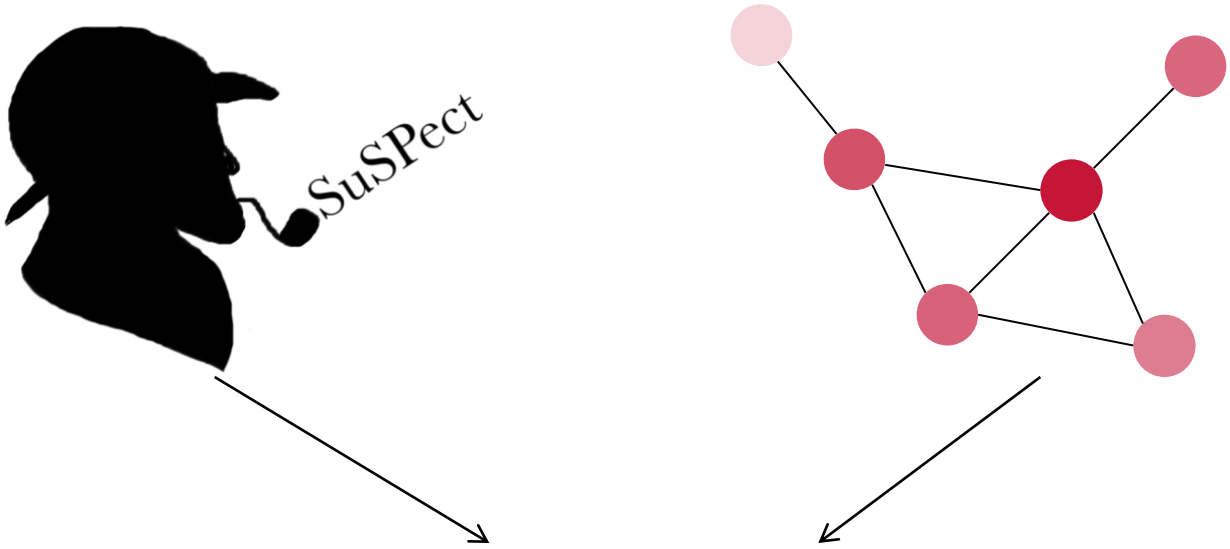
•These scores are calculated using SuSPect-FS, which is quicker and shows better performance than the full version.

Other Organisms

• For non-human proteins, scores are calculated on-the-fly, using a version of SuSPect including all features except the PPI network information and UniProt annotations.

Imperial College London

SuSPectP



Disease-specific scores associating SAVs with disease

Imperial College London

SuSPectP

Home | About | Contact

SuSPect-P is a new method giving disease-specific scores to rank variants for the specific disease you are interested in. Try SuSPect-P

Download scores

0 (Neutral) 100 (Disease-associated)

Click a score to find out more about the SAV.

ID	UniProt	Pos	WT	AA	Score
0	P04217	52	H	R	13
1	E5RG08	39	Q	P	22
1	Q9NQG7	44	Q	P	27

Imperial College London

SuSPectP

Home | About SuSPect | Contact | Download

SuSPect-P

Using SuSPect for **P**rioritisation

-SuSPect-P is an extension to SuSPect that ranks single amino acid variants according to how likely they are to be involved in a specific disease of interest.
- In simulations on whole exomes, SuSPect-P identifies the true causative variant around 50 times more often than using SuSPect alone.

Disease:

Filter by MAF (<0.02)

Run SuSPect-P

Imperial College London

SuSPectP

Protein	Variant	Score
P48382	R197Q	0.80
P13569	R75Q	0.41
-----**Confident** above here-----		
P51168	M51I	0.31
Q30201	C282Y	0.24
P15848	S384N	0.24
Q9ULD9	P1017S	0.20
O14836	V220A	0.18
P35670	V149M	0.18
P19438	R121Q	0.18
P30613	R510Q	0.15
P17813	E41K	0.14
P52333	P151R	0.13
-----**Possible** above here-----		
P06865	I436V	0.11
P11509	L160H	0.11
Q07021	T130M	0.10
Q9NQR1	L373P	0.10

Imperial College London

Acknowledgements & References

- Prof. Michael Sternberg
- Dr Ioannis Filippis
- Dr Lawrence Kelley
- Dr Suhail Islam

- Yates CM & Sternberg MJE (2013) Proteins and domains vary in their tolerance of non-synonymous single nucleotide polymorphisms. *J. Mol. Biol.*, **425**:1274-86
- Yates CM *et al.* (2014) SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features. *J. Mol. Biol.*, **426**:2692-701

MRC Medical Research Council

Cross-Validation

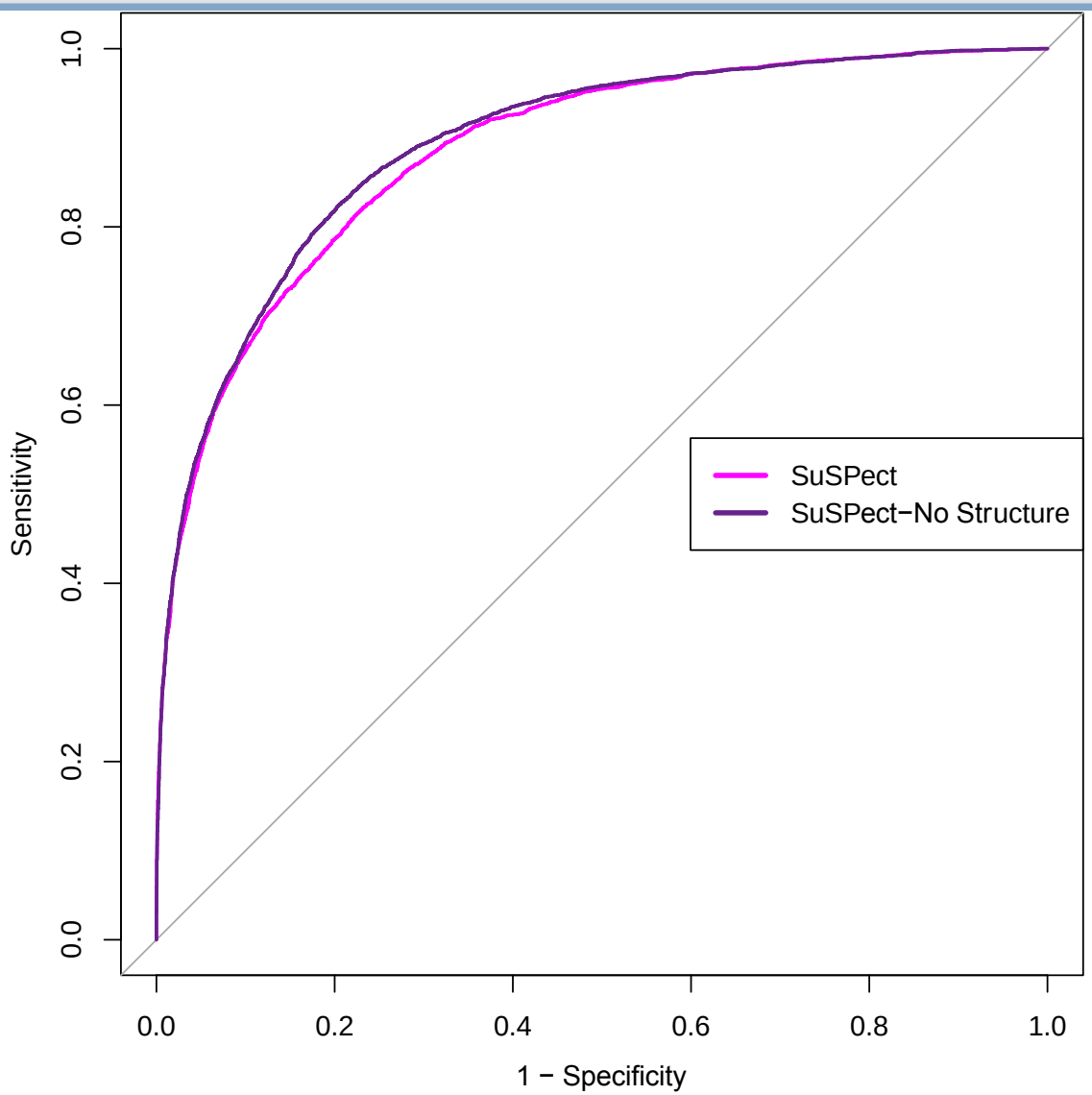
	Precision	Recall	MCC	Balanced Accuracy
SAV	0.81	0.75	0.66	0.83
Protein	0.80	0.72	0.64	0.81
Feature Selection	1.00	0.63	0.72	0.82

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN} \qquad BA = \frac{0.5 \times TP}{TP + FN} + \frac{0.5 \times TN}{TN + FP}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Imperial College
London

Results – No Structural Features



Imperial College
London

Results – No Network Features