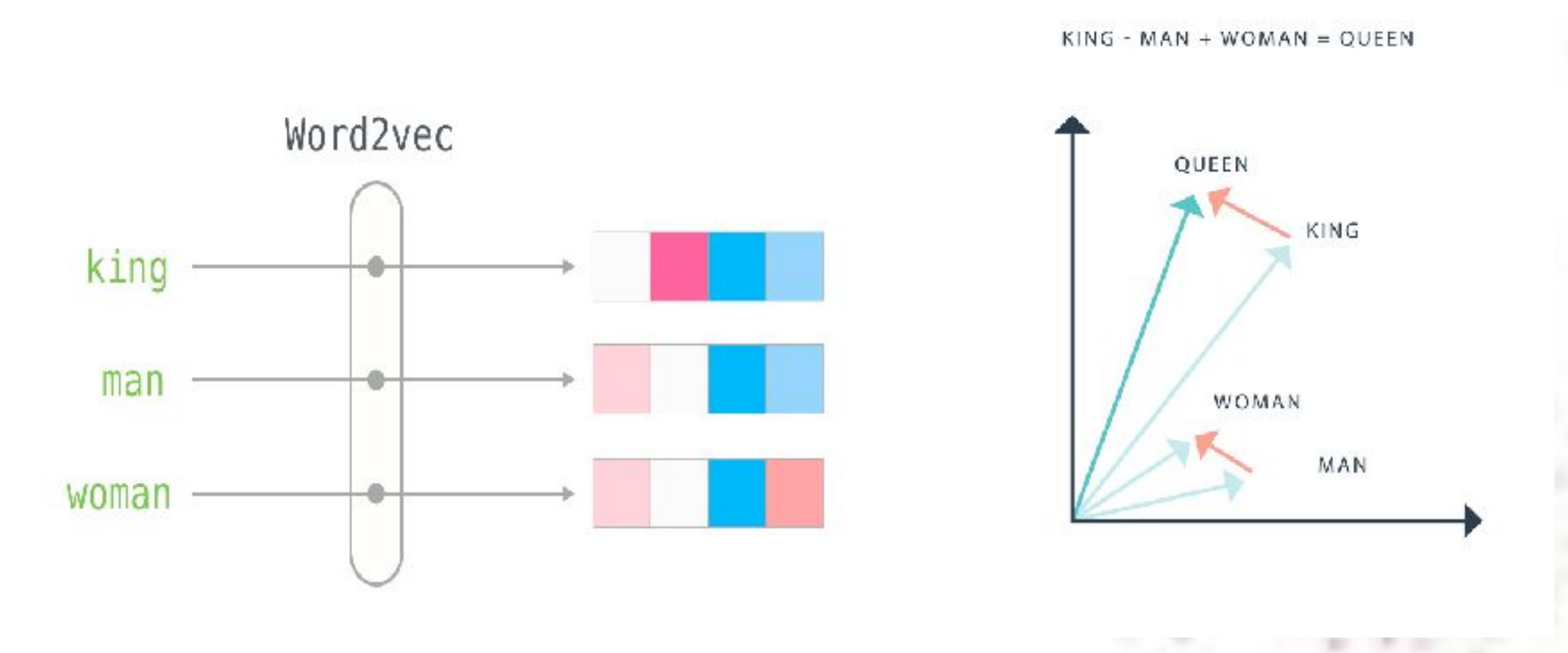
PREDICTING THE IMPACT OF MISSENSE MUTATIONS VIA GRAPH EMBEDDINGS

BÜŞRA KULOĞLU/Sabancı University
EZGİ ŞEN/Bilkent University

SUPERVISOR:
ÖZNUR TAŞTAN

ABSTRACT



A missense mutation is a mutation that occurs in a single nucleotide, which results in the alteration of an amino acid in the resulting protein. Determination of the impact of a missense mutation is closely related to disease diagnosis. Having a variety of possibilities in the case of amino acid substitution, predicting the impact of a missense mutation remains to be a challenge.

An embedding translates a relatively low dimensional space into high dimensional vectors.

Our aim, in this study, is **to test whether including structural information derived from protein structures as node embeddings improves prediction of the functional impact of missense mutations**

OBJECTIVES

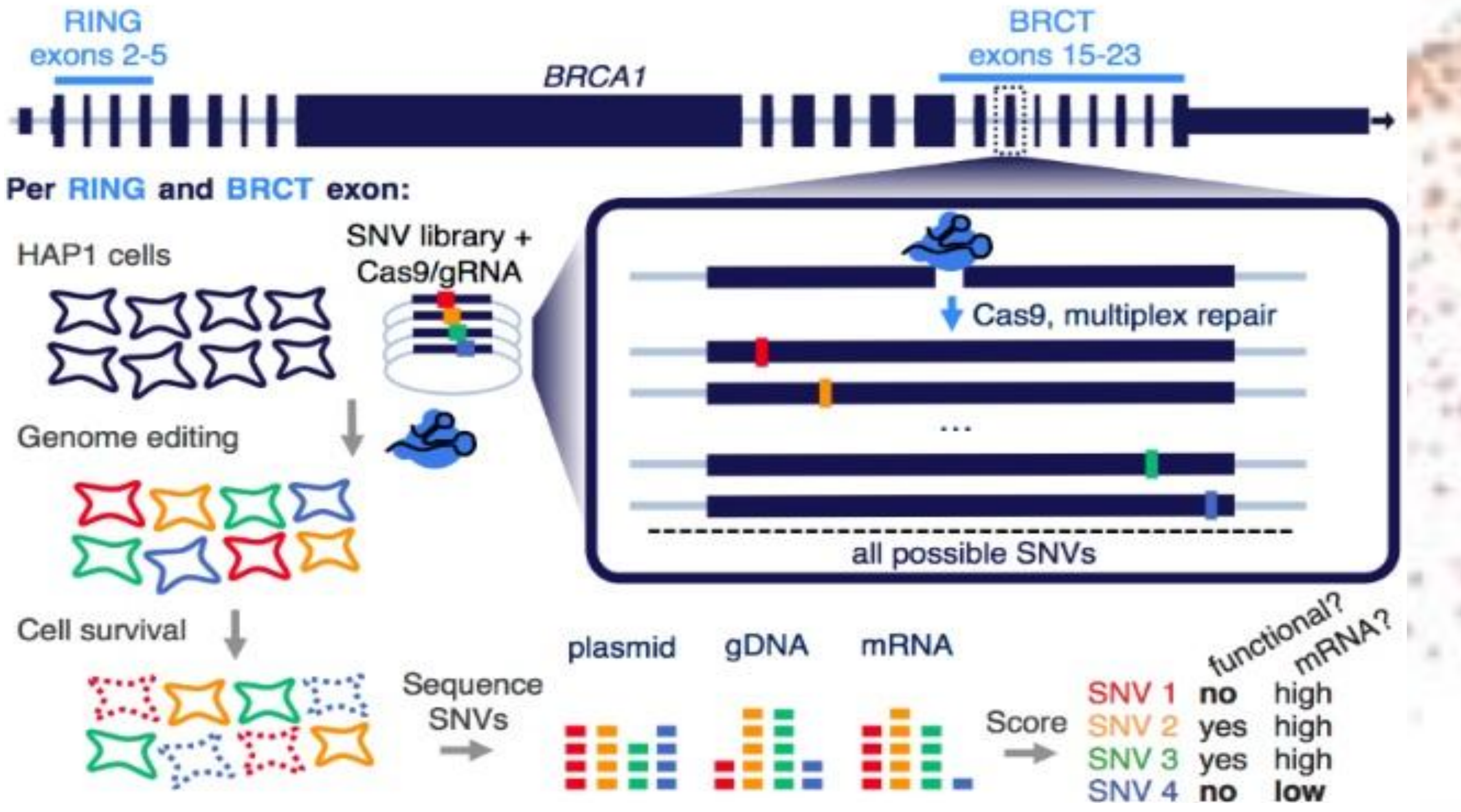
★ Mutation caused changes in proteins
→ may be recognized by the cells or not
→ the mutations that not noticed by cells
- may cause diseases like cancer.

★ Predicting the functional consequence of a mutation
→ very critical for health care
→ diagnosis.

★ *Our research focus:* finding the ways of predicting missense mutations' impacts on proteins.

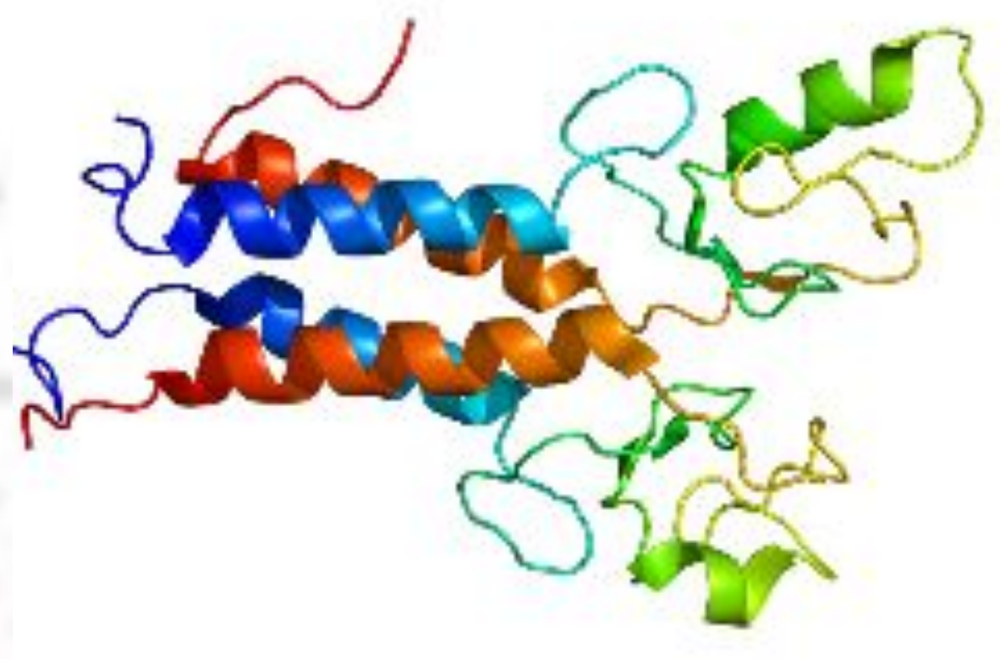
★ The potential benefits of variant analysis
→ improving patient care
→ surveillance
→ treatment outcomes.

DATASET

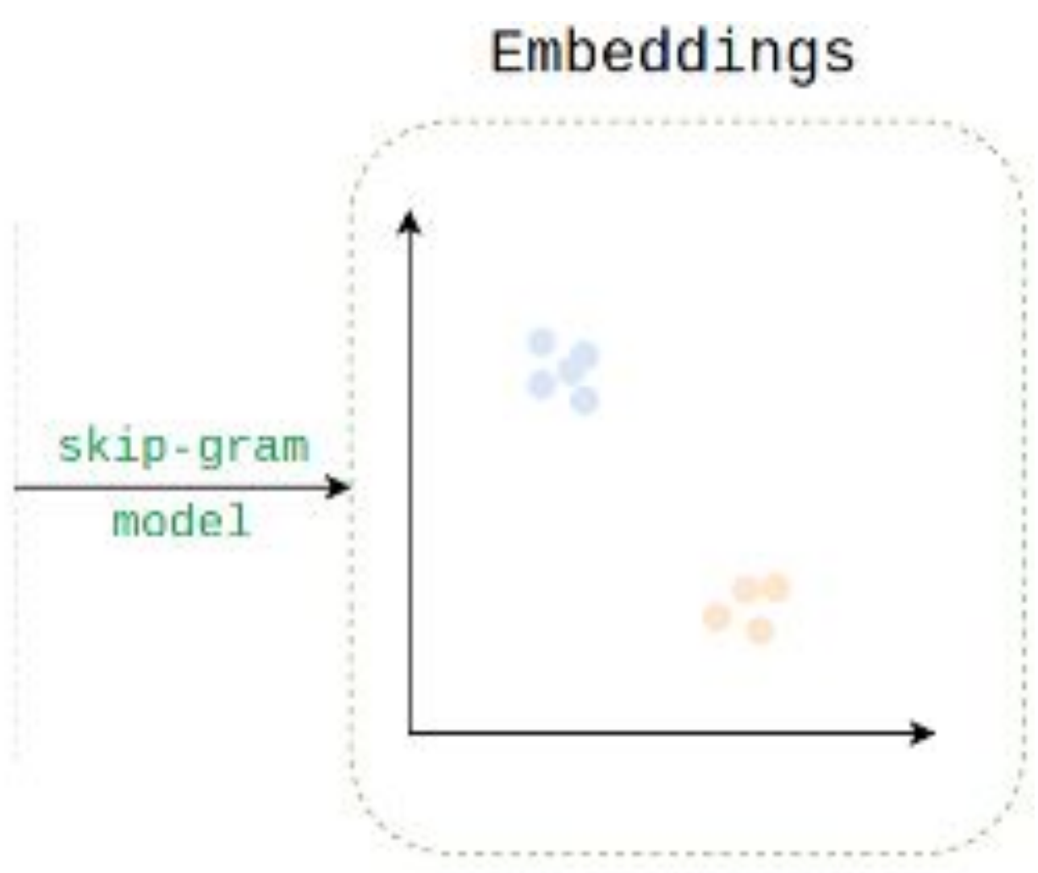


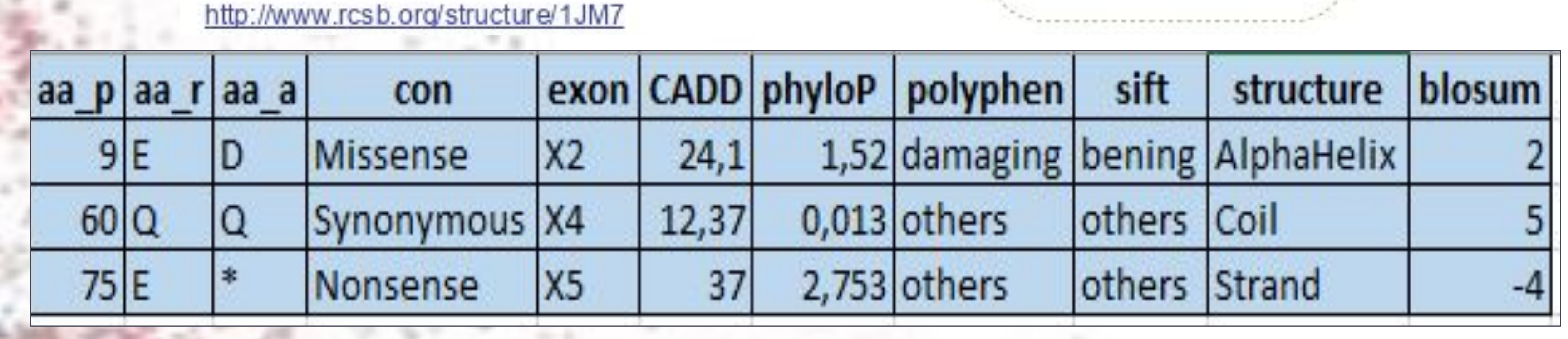
The raw data included 3892 mutations. Deleting the non-coding regions of the raw data, 2769 mutation data is left in the used dataset.

METHODS



http://www.rcsb.org/structure/1JM7





aa_p	aa_r	aa_a	con	exon	CADD	phyloP	polyphen	sift	structure	blosum
9	E	D	Missense	X2	24,1	1,52	damaging	bening	AlphaHelix	2
60	Q	Q	Synonymous	X4	12,37	0,013	others	others	Coil	5
75	E	*	Nonsense	X5	37	2,753	others	others	Strand	-4

aa_p → Amino acid position
aa_r → Amino acid reference
aa_a → Amino acid altered
con → Consequence

CADD, phyloP, polyphen, sift and blosum represent the scores.

★ As features,
  ★ amino acid position, reference amino acid, altered amino acid, consequence of mutation, exon number, CADD score, phyloP score, polyPhen2 score, SIFT score, secondary structure information for mentioned domains, BLOSUM62 score

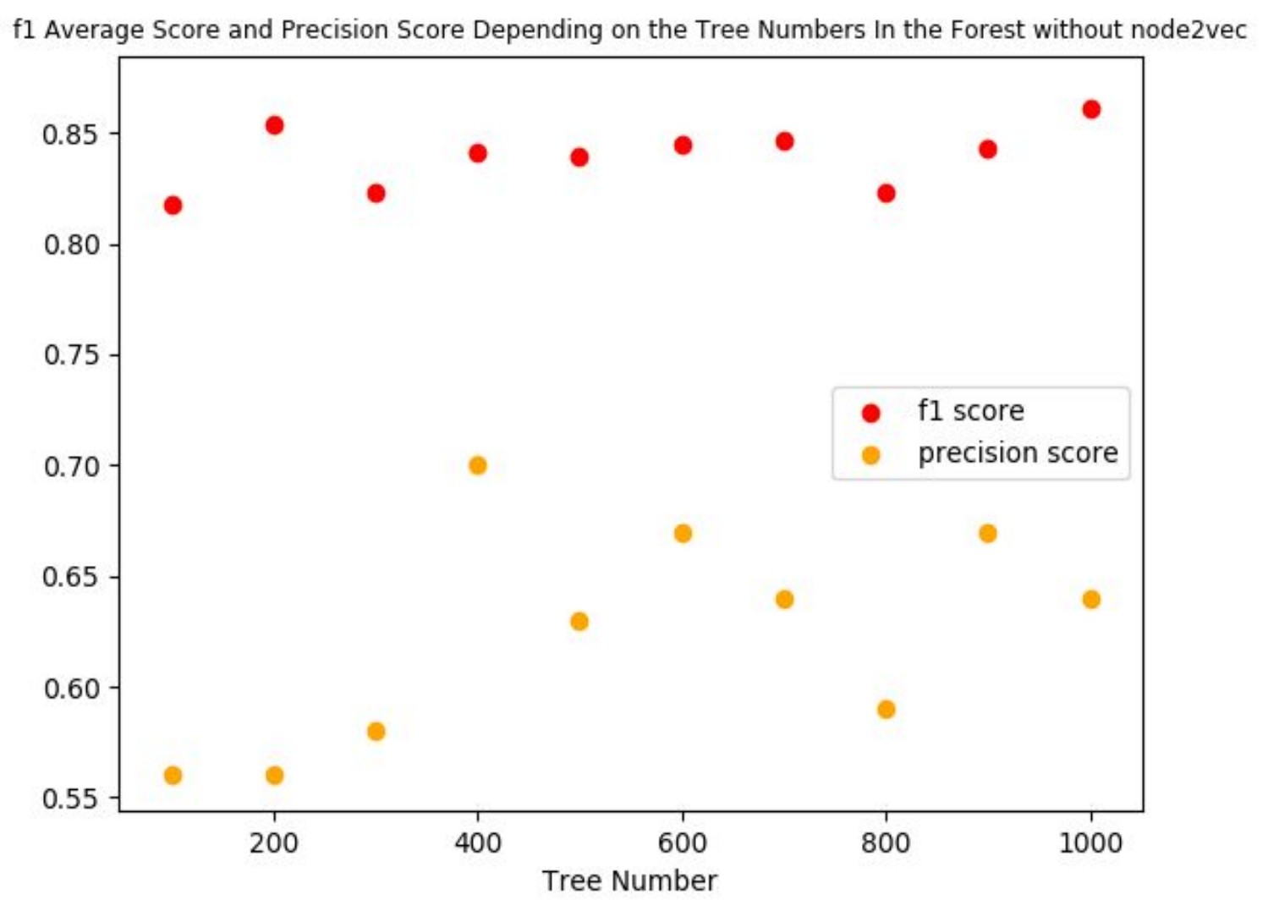
★ LoF information is used as label

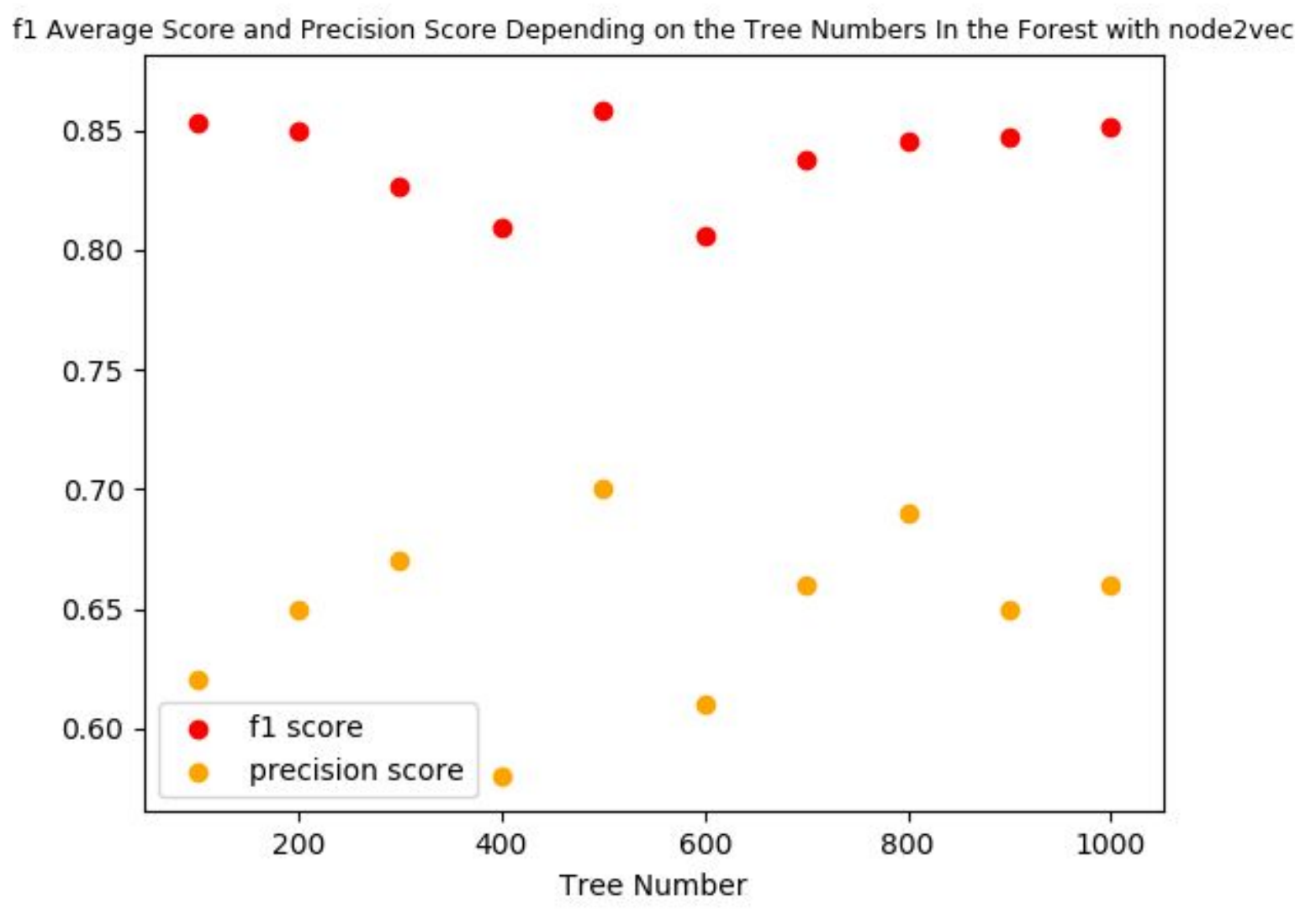
★ Random Forest Classifier is chosen

★ Oversampling on train data was performed for balancing the data

RESULTS & CONCLUSION

★ The random forest classifier performed by using the feature columns gave the best performances with 1000 trees
- ■ without the usage of node embedding features
- ■ f1 score → 86.1%
  - observed with 1000 tree number
  - most effective feature → CADD score
    CADD score is one of the features mostly related to the LoF consequence, thus carrying the most informative feature overall.
- ■ with node embedding features
- ■ dataset altered via node2vec
  - ○ 65 new feature columns added
  - performance increased in overall





REFERENCES

1. National center for biotechnology information, 12 2018. https://www.ncbi.nlm.nih.gov.
2. RCSB protein data bank, 12 2018. https://www.rcsb.org
3. Findlay, et al. Accurate classification of brca1 variants with saturation genome editing. Nature, 562(7726):217,2018.