Can we use standard tools to predict functional effects of point variations outside conserved domains? TET2 example

B. Gemovic, V. Perovic, S. Glisic, N. Veljkovic

Centre for Multidisciplinary Research
VINCA Institute of Nuclear Sciences

TABIS, Belgrade, 2013

MOTIVATION

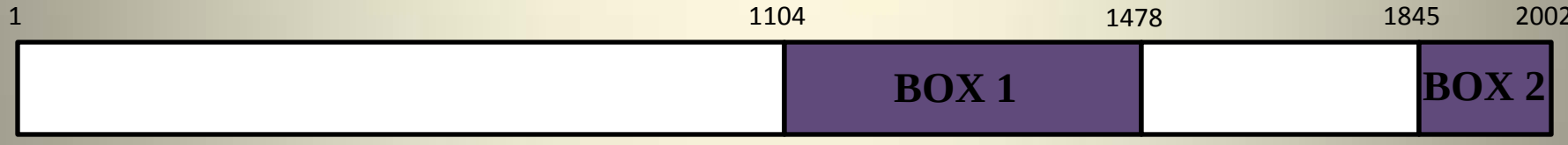
- Personalized medicine
  - Mutations are important clinical markers for diagnosis, prognosis and choice of therapy
- 3.7 million variations per human genome
  - >24 000 in coding regions, > 500 change protein sequence
  - Single nucleotide polymorphisms (SNPs) are recognized as the main cause of human genetic variability
- The main challenge ahead:
  - Differentiate between "neutral“ SNPs versus "functional" or "pathogenic" mutations that assign (positive or negative) susceptibility to Mendelian disorders, common complex diseases, cancers

MOTIVATION

- Most commonly used tools
  - Multiple Sequence Alignments (MSA), structural and functional information, physicochemical characteristics of amino acids
  - Predict mutations in conserved domains (CDs): affect important protein functions
- Mutations positioned outside CDs
  - Cancer, complex diseases

DATASET

- TET2
  - Epigenetic regulation
  - Mutated in all myeloid malignancies
  - Defined CDs



	CDs	nCDs
Mutations	94	27
SNPs	3	42

TOOLS

- **SIFT**
  - Basis: MSA
- **PolyPhen-2**
  - Basis: 8 sequence-based and 3 structure-based features, Naïve Bayes classifier
- **PhD-SNP**
  - Basis: MSA, sequence environment, SVM classifier
- **MutPred**
  - Basis: MSA and 14 structural and functional properties

Advantages:

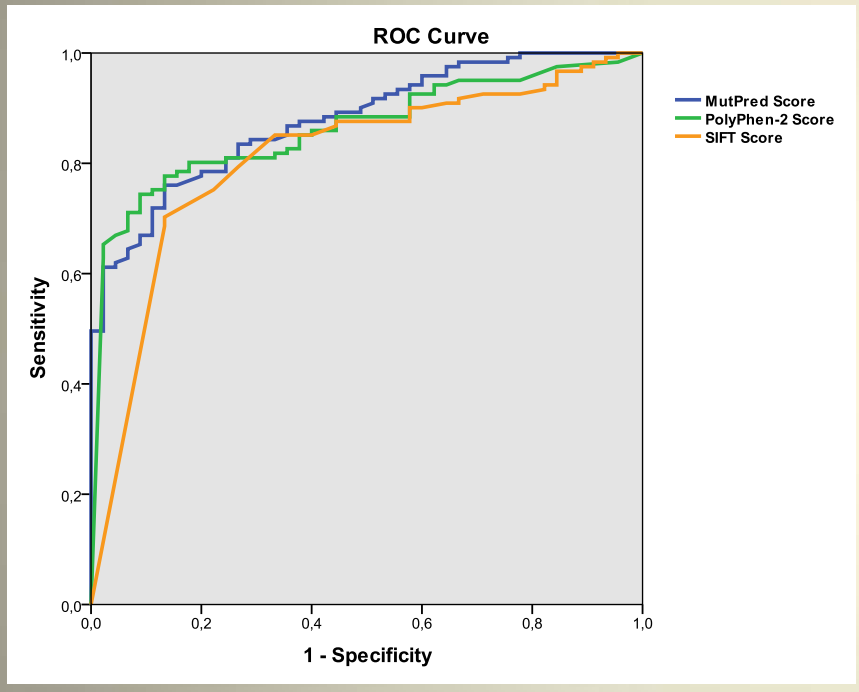
- Identification of mutations that affect conserved functional domains
- Use of structural and functional information
- Machine learning

Disadvantages:

- Insufficient sequences for MSA
- Unknown 3D structure

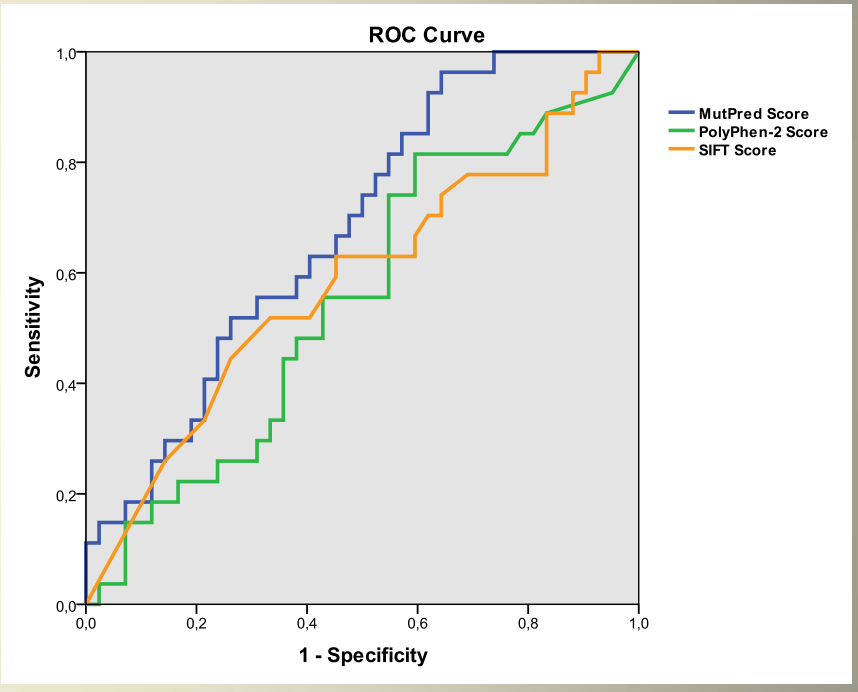
RESULTS - Scores

Whole dataset (CDs + nCDs)



Method	AUC
MutPred	0.879
PolyPhen-2	0.863
SIFT	0.810

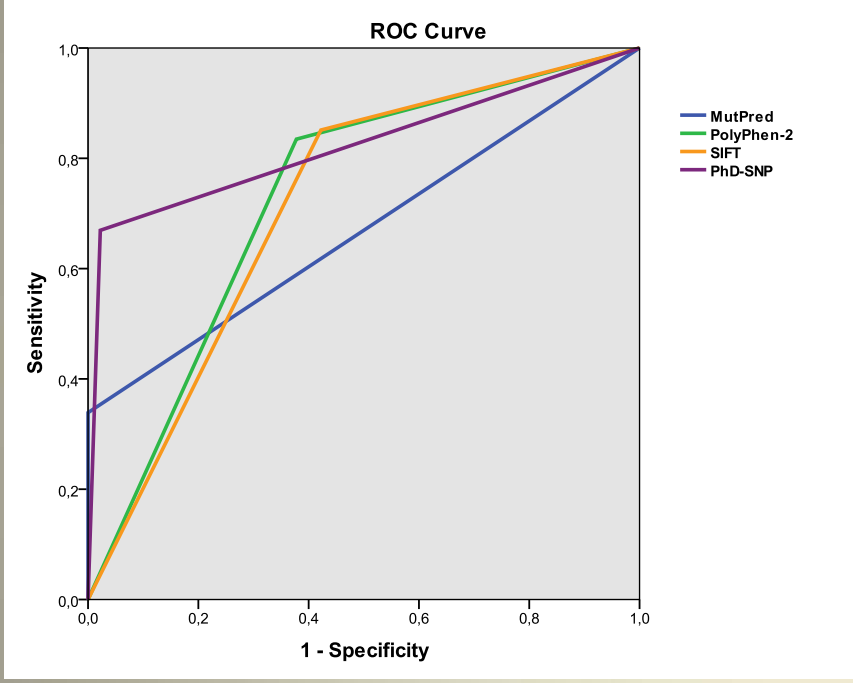
Subset (nCDs)



Method	AUC
MutPred	0.681
PolyPhen-2	0.552
SIFT	0.585

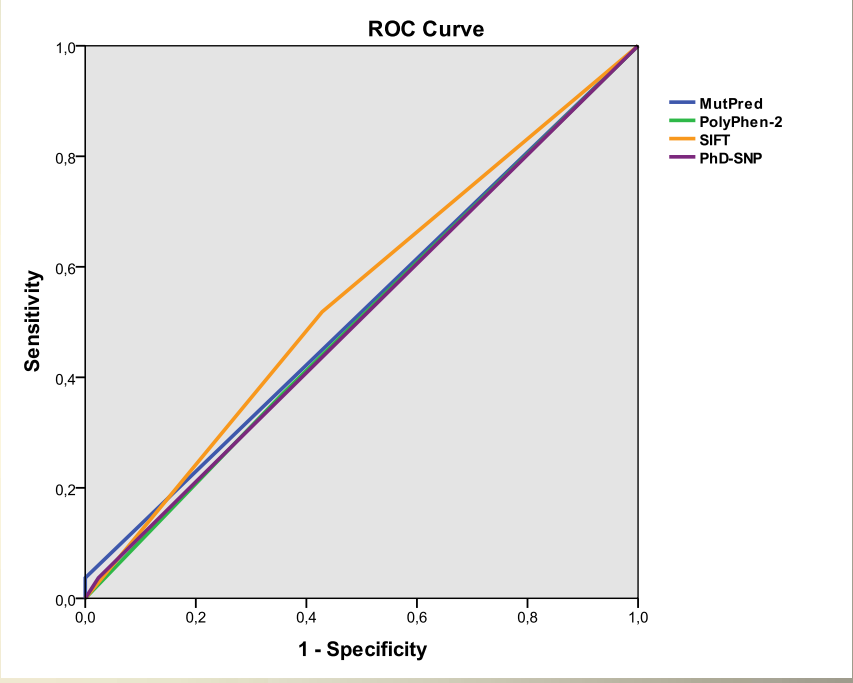
RESULTS – Binary classification

Whole dataset (CDs + nCDs)



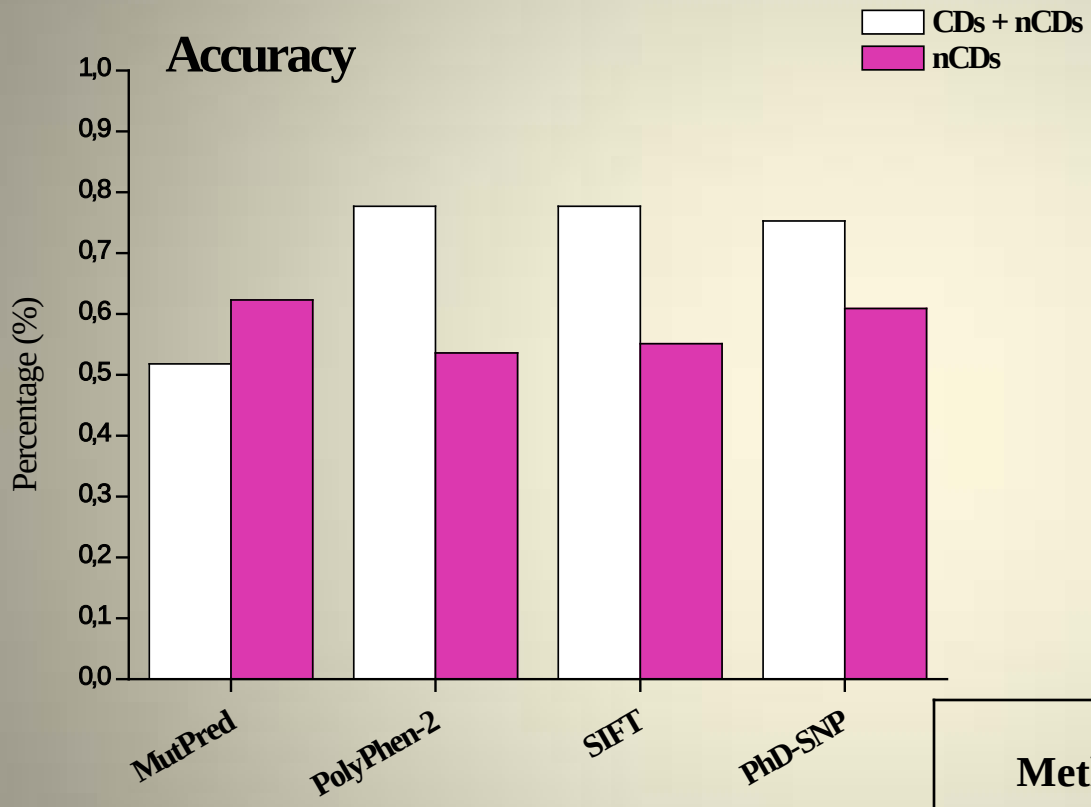
Method	AUC
PhD-SNP	0.824
PolyPhen-2	0.728
SIFT	0.715
MutPred	0.669

Subset (nCDs)



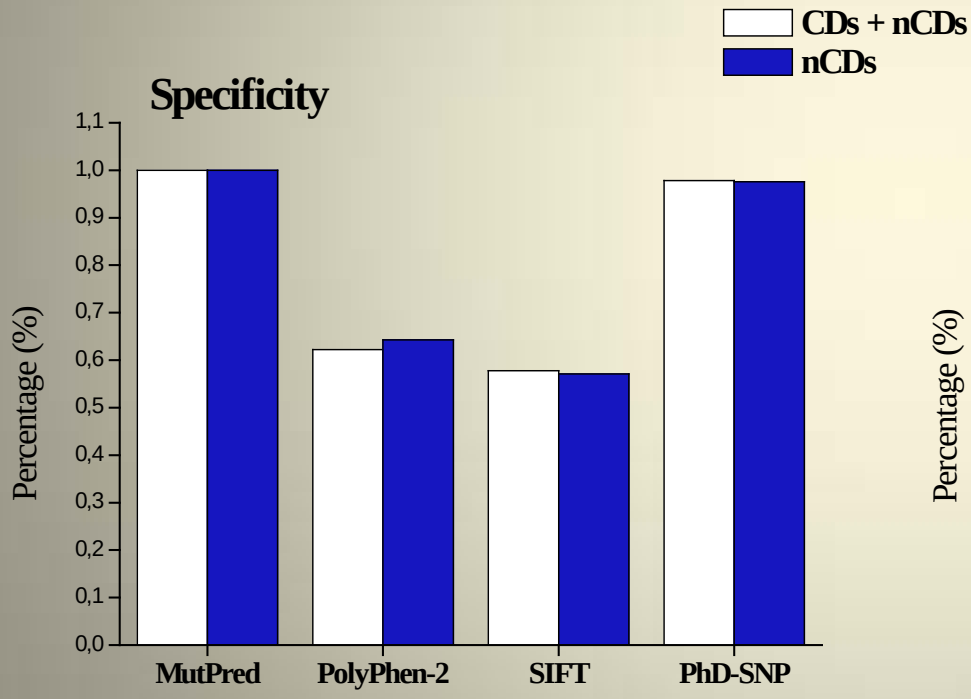
Method	AUC
PhD-SNP	0.507
PolyPhen-2	0.507
SIFT	0.545
MutPred	0.519

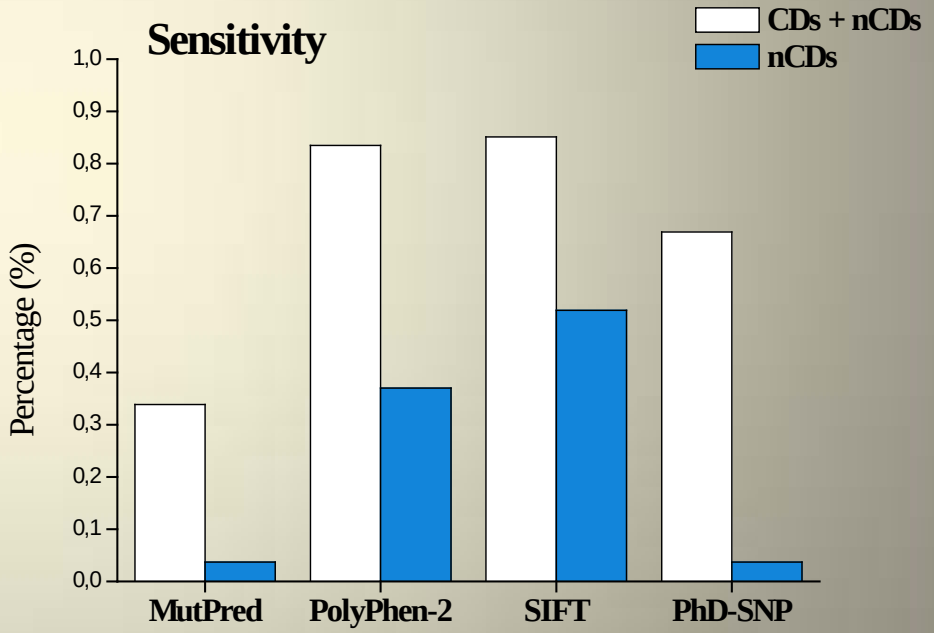
RESULTS – Binary classification



Method	Fisher's Exact Test	
	CDs + nCDs	nCDs
MutPred	0.000	0.391
PolyPhen-2	0.000	1.000
SIFT	0.000	0.621
PhD-SNP	0.000	1.000

RESULTS – Binary classification





SUMMARY

SNPs: source of genetic variability, clinical markers

TET2: epigenetic regulator, mutated in myeloid malignancies; 166 variations (69 outside CDs)

SIFT, PolyPhen-2, PhD-SNP, MutPred: MSA-based tools; plus structural and functional information

Scores: 20-30% lower AUC for nCDs variations compared with CDs variations

Binary: Decreased accuracy for nCDs variations compared with CDs variations, owing to decreased sensitivity

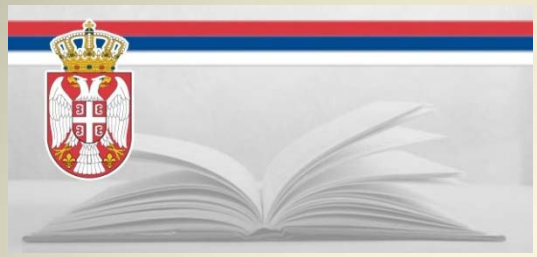
CONCLUSIONS

Can we use standard tools to predict functional effects of point variations outside conserved domains?

It seems… **NO**

We need new algorithms

ACKNOWLEDGMENTS



Ministry of Education, Science and Technological Development

EuGESMA: BM0801

Thank you for your attention!