# Predikčná metóda na určenie klinického dopadu štrukturálnej genomickej variability

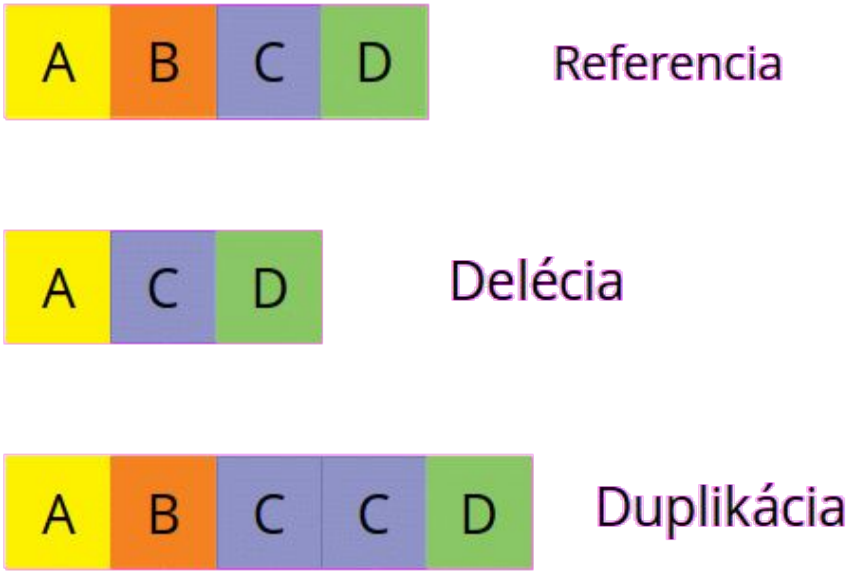
Michaela Gažiová

Školiteľ: Mgr. Jaroslav Budiš, PhD.

# Štrukturálne varianty

CNV (Copy number variations)

➔ Dĺžka ≥ 1000 bp



# CNV môžu prispieť k rozvoju ochorení

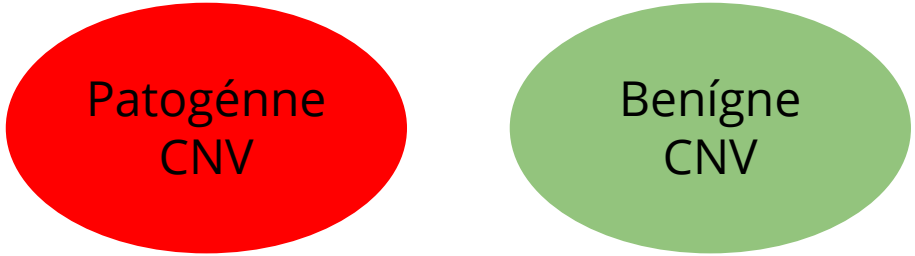
➔ narušením samotného génu

➔ zasiahnutím regulačnej oblasti

- zmena množstva produktu génu

Ochorenia spojené s vývinovými, kardiovaskulárnymi, neurodegeneratívnymi, autoimunitnými poruchami

# CNV

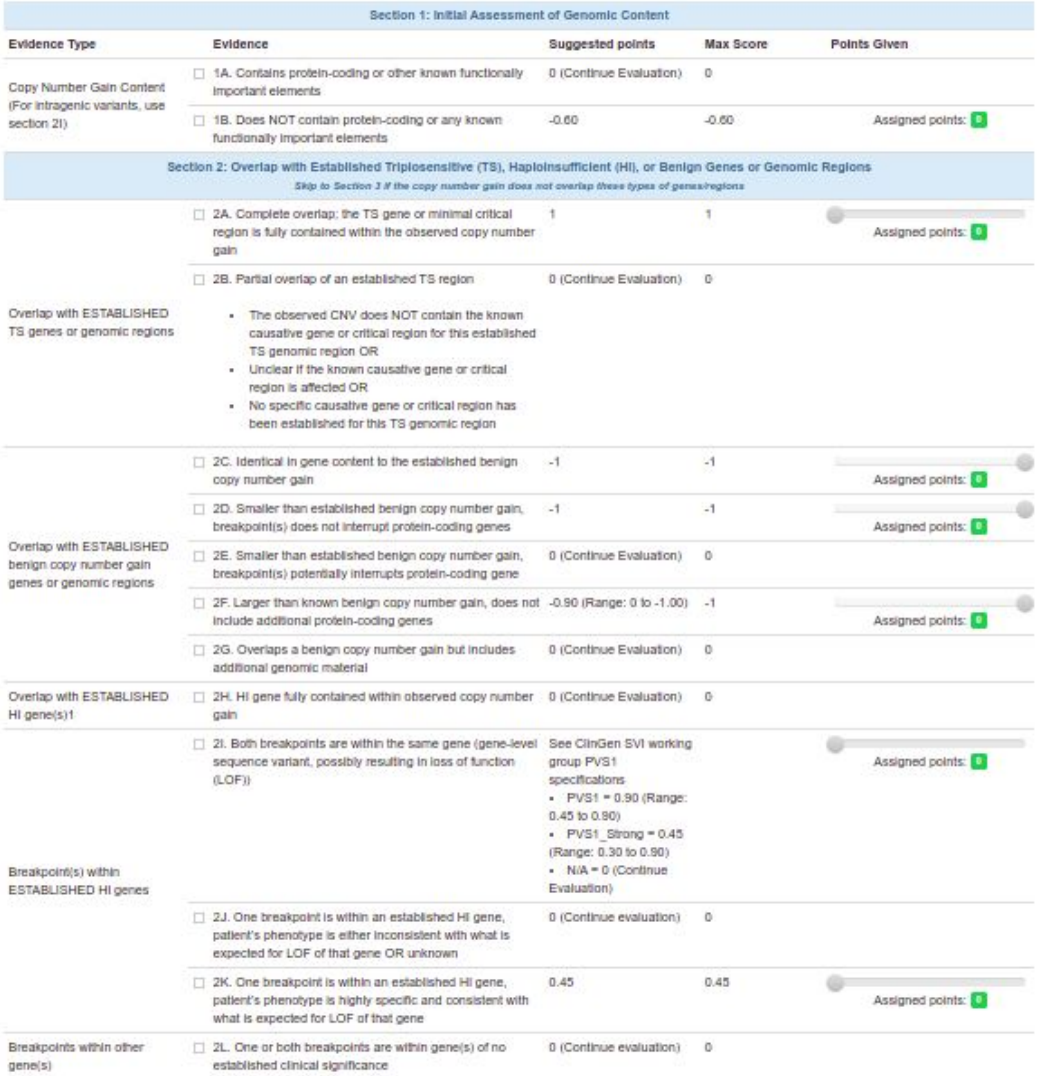



## Motivácia:

- ➔ vopred identifikovať CNV zapríčiňujúci chorobný prejav
- ➔ určenie klinického dopadu prítomnosti CNV
- ➔ včas zahájiť a zacieliť liečbu

# Súčasný stav riešenej problematiky

ACMG kritériá



**Section 1: Initial Assessment of Genomic Content**

| Evidence Type | Evidence | Suggested points | Max Score | Points Given |
|---|---|---|---|---|
| Copy Number Gain Content (For intragenic variants, use section 2I) | ☐ 1A. Contains protein-coding or other known functionally important elements | 0 (Continue Evaluation) | 0 | |
| | ☐ 1B. Does NOT contain protein-coding or any known functionally important elements | -0.60 | -0.60 | Assigned points: 0 |

**Section 2: Overlap with Established Triplosensitive (TS), Haploinsufficient (HI), or Benign Genes or Genomic Regions**

*Skip to Section 3 if the copy number gain does not overlap these types of genes/regions*

| | | | | |
|---|---|---|---|---|
| Overlap with ESTABLISHED TS genes or genomic regions | ☐ 2A. Complete overlap: the TS gene or minimal critical region is fully contained within the observed copy number gain | 1 | 1 | Assigned points: 0 |
| | ☐ 2B. Partial overlap of an established TS region<br>• The observed CNV does NOT contain the known causative gene or critical region for this established TS genomic region OR<br>• Unclear if the known causative gene or critical region is affected OR<br>• No specific causative gene or critical region has been established for this TS genomic region | 0 (Continue Evaluation) | 0 | |
| Overlap with ESTABLISHED benign copy number gain genes or genomic regions | ☐ 2C. Identical in gene content to the established benign copy number gain | -1 | -1 | Assigned points: 0 |
| | ☐ 2D. Smaller than established benign copy number gain, breakpoint(s) does not interrupt protein-coding genes | -1 | -1 | Assigned points: 0 |
| | ☐ 2E. Smaller than established benign copy number gain, breakpoint(s) potentially interrupts protein-coding gene | 0 (Continue Evaluation) | 0 | |
| | ☐ 2F. Larger than known benign copy number gain, does not include additional protein-coding genes | -0.90 (Range: 0 to -1.00) | -1 | Assigned points: 0 |
| | ☐ 2G. Overlaps a benign copy number gain but includes additional genomic material | 0 (Continue Evaluation) | 0 | |
| Overlap with ESTABLISHED HI gene(s)1 | ☐ 2H. HI gene fully contained within observed copy number gain | 0 (Continue Evaluation) | 0 | |
| Breakpoint(s) within ESTABLISHED HI genes | ☐ 2I. Both breakpoints are within the same gene (gene-level sequence variant, possibly resulting in loss of function (LOF)) | See ClinGen SVI working group PVS1 specifications<br>• PVS1 = 0.90 (Range: 0.45 to 0.90)<br>• PVS1_Strong = 0.45 (Range: 0.30 to 0.90)<br>• N/A = 0 (Continue Evaluation) | | Assigned points: 0 |
| | ☐ 2J. One breakpoint is within an established HI gene, patient's phenotype is either inconsistent with what is expected for LOF of that gene OR unknown | 0 (Continue evaluation) | 0 | |
| | ☐ 2K. One breakpoint is within an established HI gene, patient's phenotype is highly specific and consistent with what is expected for LOF of that gene | 0.45 | 0.45 | Assigned points: 0 |
| Breakpoints within other gene(s) | ☐ 2L. One or both breakpoints are within gene(s) of no established clinical significance | 0 (Continue evaluation) | 0 | |

# Súčasný stav riešenej problematiky

## Existúce nástroje:

➔ *AnnotSV*

- anotačný a klasifikačný nástroj

➔ *ClassifyCNV*

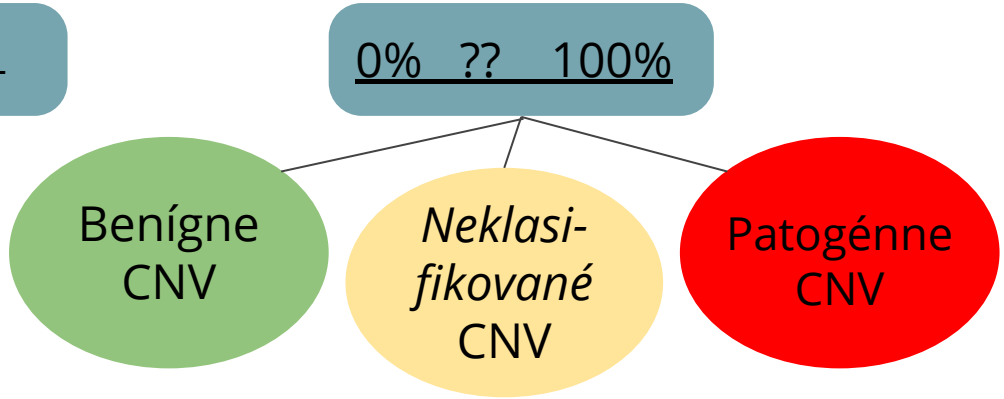
- presný
- veľkú časť klasifikuje ako *neklasifikované* (68% z delécií a 72% duplikácií na testovacích dátach)

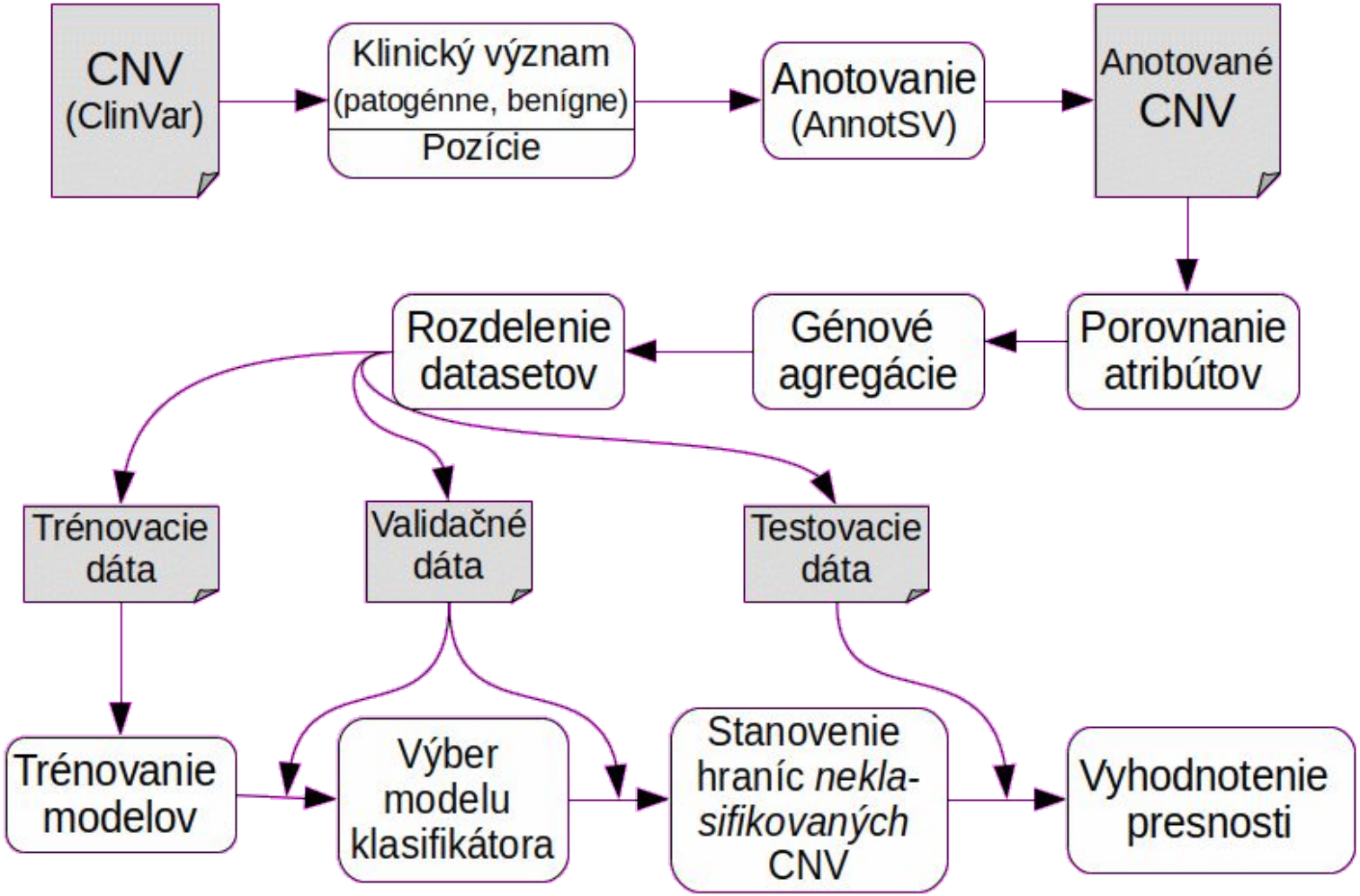
# Cieľ bakalárskej práce

Vytvorenie predikčnej metódy

➔ Predikcia pravdepodobnosti patogenicity CNV z koordinátov CNV (číslo chromozómu, počiatočná a koncová pozícia)

chrX:103,776,505-103,792,618 DEL




CNV
(ClinVar)
Klinický význam
(patogénne, benígne)
Pozície
Anotovanie
(AnnotSV)
Anotované
CNV
Porovnanie
atribútov
Génové
agregácie
Rozdelenie
datasetov
Trénovacie
dáta
Validačné
dáta
Testovacie
dáta
Trénovanie
modelov
Výber
modelu
klasifikátora
Stanovenie
hraníc *nekla-
sifikovaných*
CNV
Vyhodnotenie
presnosti

# Dáta

ClinVar databáza - klinický význam ľudských variantov

| | Delécie | | | Duplikácie | | |
|---|---|---|---|---|---|---|
| Dataset | Benígne | Patogénne | Spolu | Benígne | Patogénne | Spolu |
| Trénovací | 7608 | 4180 | 11788 | 2019 | 6417 | 8436 |
| Validačný | 1622 | 904 | 2526 | 1385 | 423 | 1808 |
| Testovací | 1635 | 891 | 2526 | 1380 | 428 | 1808 |
| | | | 16840 | | | 12052 |

# Výsledky

# Použité atribúty

Dĺžka CNV, Počet génov

Počet ’morbídnych’ génov

Konzervovanosť génov

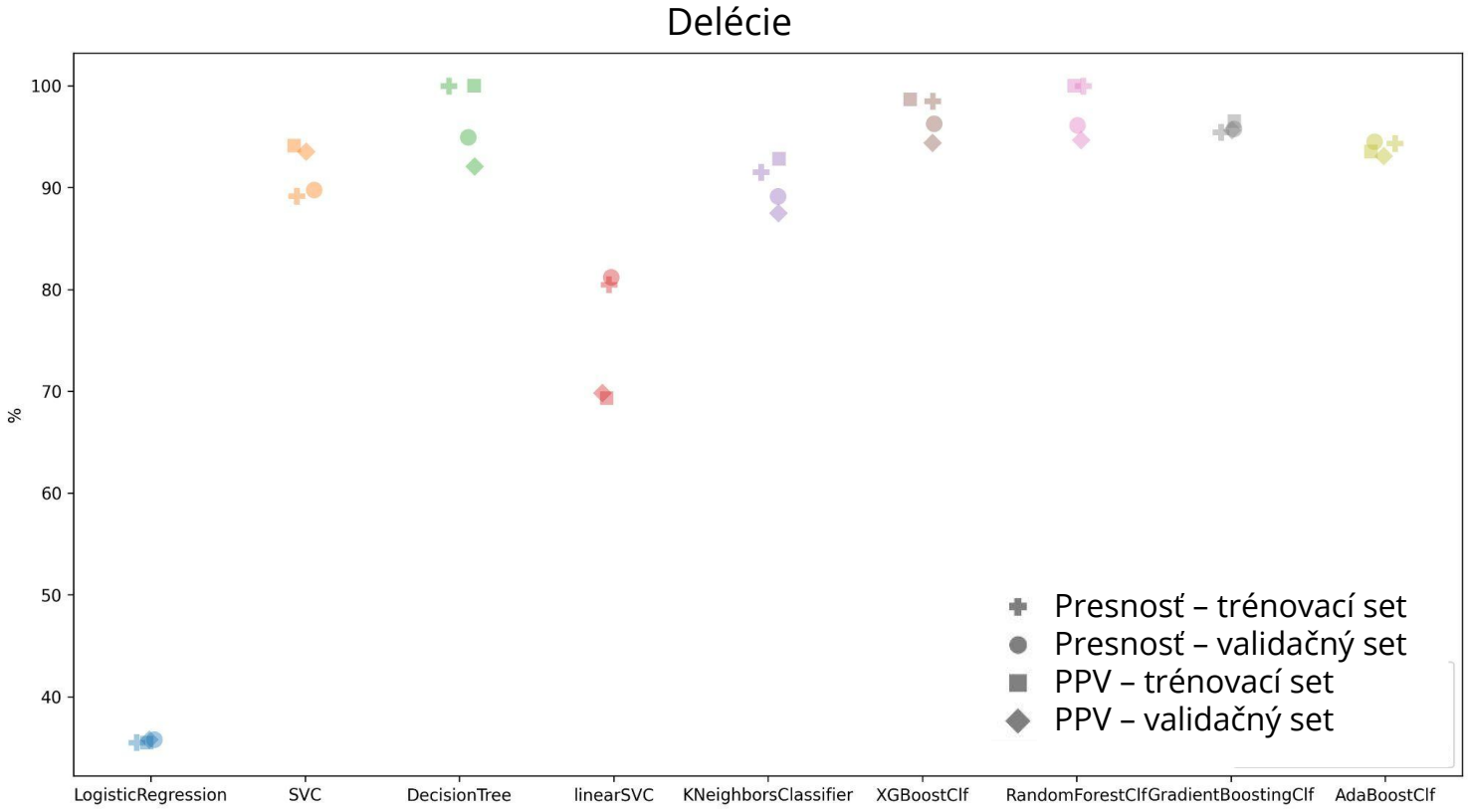
Skóre netolerancie straty funkcie

Max ExAC skóre nefunkčnosti génu

Skóre haploinsuficiencie z DDD

Skóre haploinsufficiencie z ClinGen-u

# Výber modelu



# Stanovenie hraníc *neklasifikovaných* CNV


➔ dôveryhodné výsledky vzhľadom

na citlivosť diagnostických analýz

➔ odstrániť chybné predikcie

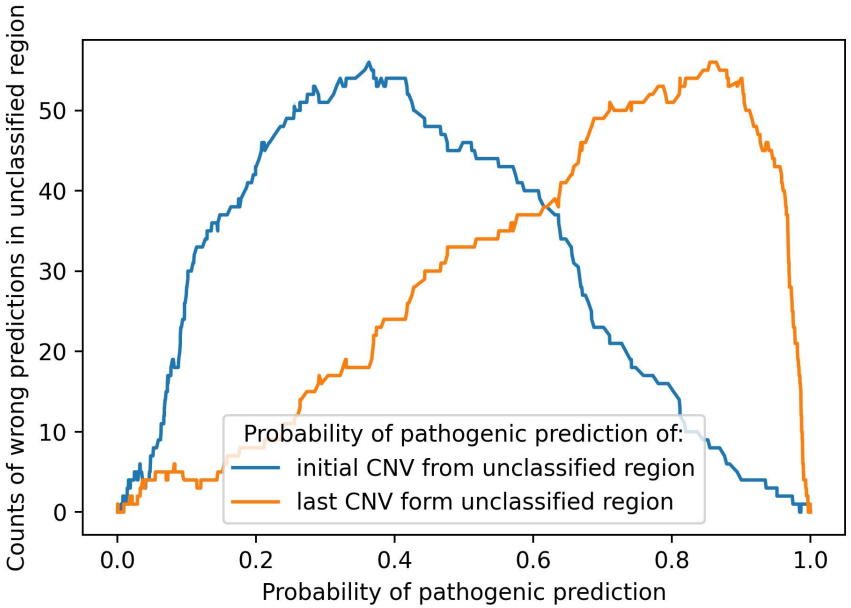
# Stanovenie hraníc *neklasifikovaných* CNV

Hranice pravdepodobnosti patogenicity neklasifikovaných CNV:

pre 5%-ný interval:

➔ 35% – 85% pre delécie

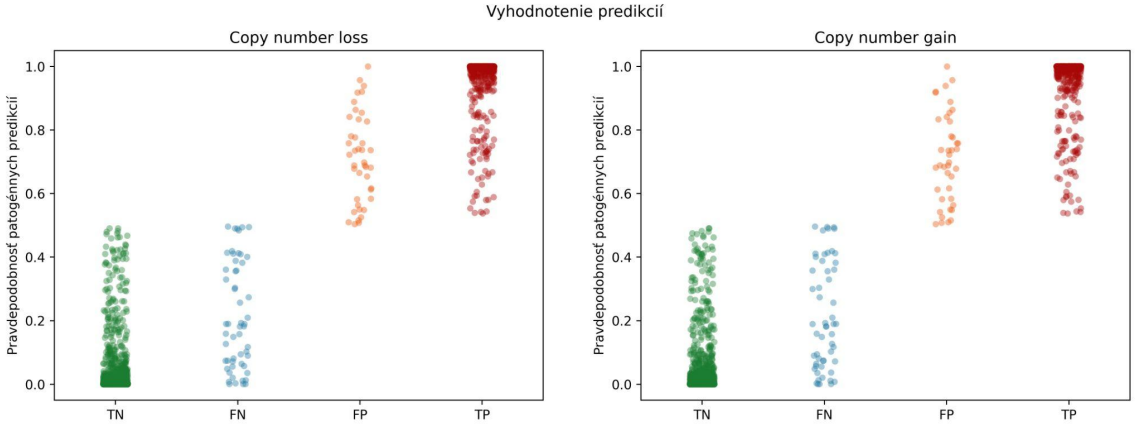
➔ 10% – 85% pre duplikácie




interval si moze uzivatel stanovit

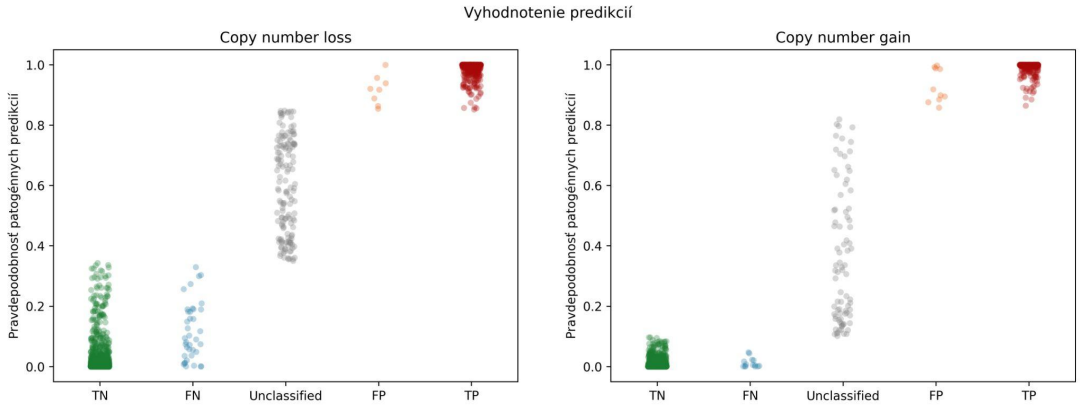
# Presnosti výsledného modelu

| CNV | Presnosť |
|---|---|
| Delécie | 96.04% |
| Duplikácie | 97.34% |



# Presnosti výsledného modelu

| CNV | Presnosť<br>5%-ný interval neklasifikovaných CNV |
|---|---|
| Delécie | 98.06% |
| Duplikácie | 98.61% |



# Porovnanie s dostupnými metódami

| | Presnosť (testovací set) | | Neklasifikované CNV | |
|---|---|---|---|---|
| Nástroj | Delécie | Duplikácie | Delécie | Duplikácie |
| AnnotSV | 73% | 65% | 0% | 0% |
| ClassifyCNV | 98.73% | 99.38% | 68% | 73% |
| Výsledný model | 96.04% | 97.34% | 0% | 0% |
| | 98.06% | 98.61% | 5% | 5% |

# Zhrnutie

- ➔ uzitočná pre klinickú diagnostiku
  - vyhodnocuje všetky CNV
  - možnosť určiť, koľko % z dát sme ochotní neklasifikovať
- ➔ veľké množstvo CNV z reálneho sveta
- ➔ úplne automatizovaná

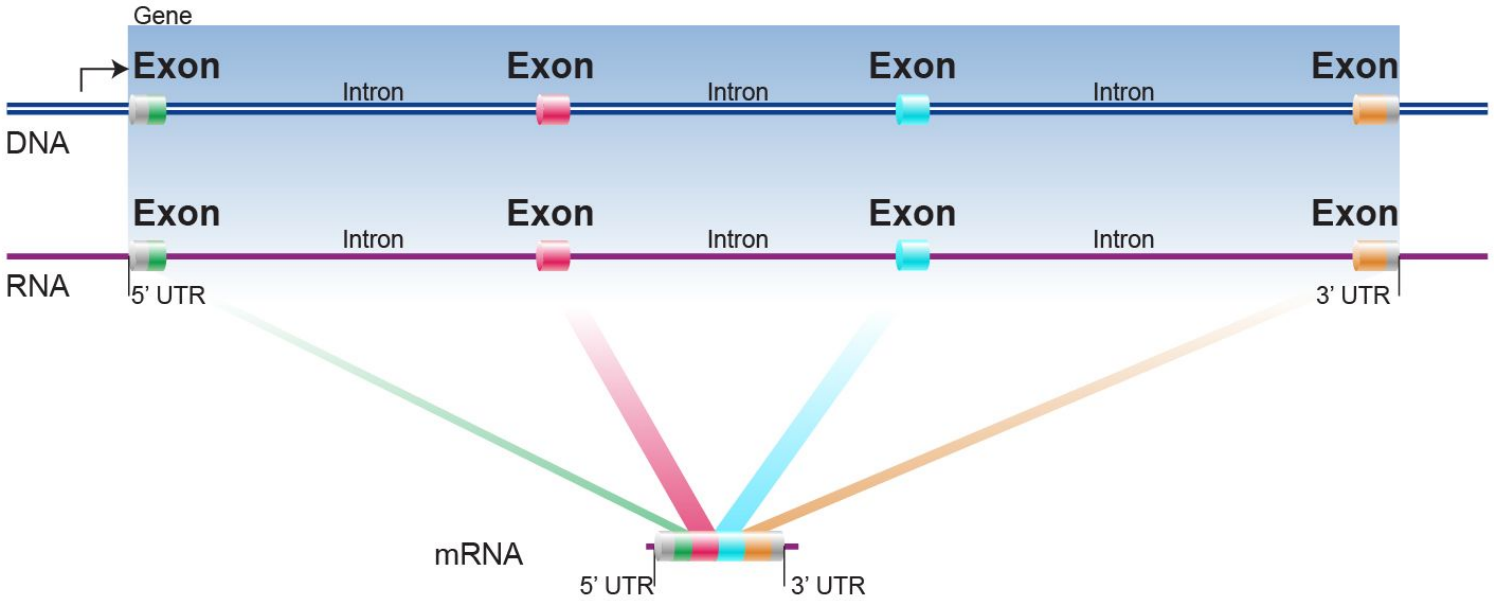
# Možnosti ďalšej práce

- ➔ Pridanie nových atribútov z ďalších genomických databáz
- ➔ Nastavenie vhodných parametrov modelu
- ➔ CNV ohodnotené ako likely pathogenic a likely benign v ClinVar
  - ako ich natrénovný model vyhodnotí?
  - natrénovať nový model
- ➔ CNV spôsobujúce závažné syndrómy a ochorenia z OMIM, DECIPHER
- ➔ benígne CNV s vysokou populačnou frekvenciou z gnomAD-u.

# Ďakujem za pozornosť

# Otázky od oponenta

➔ Vysvetlite rozdiel medzi pojmami exón a exóm (anglicky exon a exome). V práci používate slovo exóm aj namiesto slova exón



➔ Vysvetlite rozdiel medzi pojmami exón a exóm (anglicky exon a exome). V práci používate slovo exóm aj namiesto slova exón

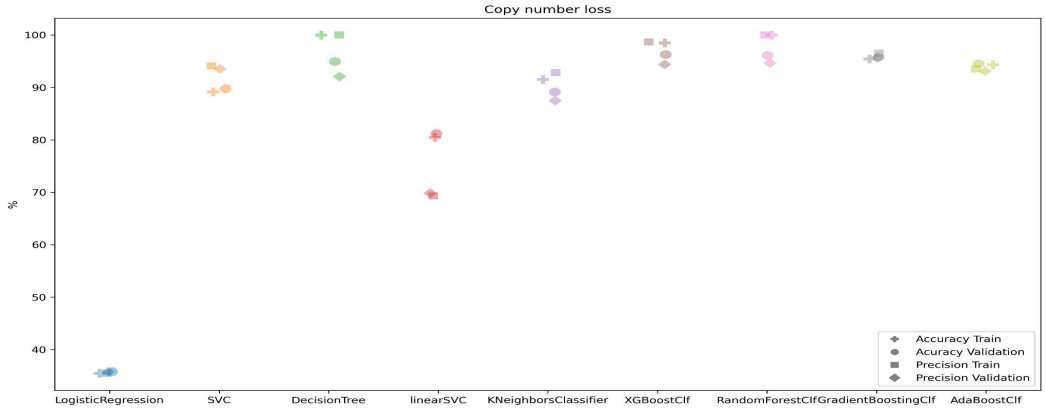
Exón -- kódujúcou sekvenciou génu

Exóm -- časť genómu tvorená exónmi

-- DNA, prepísaná do mRNA

# Otázky od oponenta

➔ Ako vysvetľujete, že napríklad pre duplikácie má logistická regresia presnosť menej ako 30%, kým XGBoost s iba jedným atribútom má presnosť 94%? (Obr. 3.2 a 3.4).



➔ Akým spôsobom boli klasifikované varianty v databáze ClinVar na patogénne, benígne a podobne? Považujete tieto údaje za dôveryhodné?

- poskytnutie metodiky určovania klinického významu (podporujúce dôkazy, zdôvodnenie klasifikácie)
- ACMG, AMP kritériá
- Sequence Variant Interpretation WG

➔ Posudzované *ClinGen Steering Committee*

| Section 1: Initial Assessment of Genomic Content | | | | |
|---|---|---|---|---|
| **Evidence Type** | **Evidence** | **Suggested points** | **Max Score** | **Points Given** |
| Copy Number Gain Content (For intragenic variants, use section 2I) | ☐ 1A. Contains protein-coding or other known functionally important elements | 0 (Continue Evaluation) | 0 | |
| | ☐ 1B. Does NOT contain protein-coding or any known functionally important elements | -0.60 | -0.60 | Assigned points: 0 |
| **Section 2: Overlap with Established Triplosensitive (TS), Haploinsufficient (HI), or Benign Genes or Genomic Regions** *Skip to Section 3 if the copy number gain does not overlap these types of genes/regions* | | | | |
| Overlap with ESTABLISHED TS genes or genomic regions | ☐ 2A. Complete overlap: the TS gene or minimal critical region is fully contained within the observed copy number gain | 1 | 1 | Assigned points: 0 |
| | ☐ 2B. Partial overlap of an established TS region<br>• The observed CNV does NOT contain the known causative gene or critical region for this established TS genomic region OR<br>• Unclear if the known causative gene or critical region is affected OR<br>• No specific causative gene or critical region has been established for this TS genomic region | 0 (Continue Evaluation) | 0 | |
| Overlap with ESTABLISHED benign copy number gain genes or genomic regions | ☐ 2C. Identical in gene content to the established benign copy number gain | -1 | -1 | Assigned points: 0 |
| | ☐ 2D. Smaller than established benign copy number gain, breakpoint(s) does not interrupt protein-coding genes | -1 | -1 | Assigned points: 0 |
| | ☐ 2E. Smaller than established benign copy number gain, breakpoint(s) potentially interrupts protein-coding gene | 0 (Continue Evaluation) | 0 | |
| | ☐ 2F. Larger than known benign copy number gain, does not include additional protein-coding genes | -0.90 (Range: 0 to -1.00) | -1 | Assigned points: 0 |
| | ☐ 2G. Overlaps a benign copy number gain but includes additional genomic material | 0 (Continue Evaluation) | 0 | |
| Overlap with ESTABLISHED HI gene(s)1 | ☐ 2H. HI gene fully contained within observed copy number gain | 0 (Continue Evaluation) | 0 | |
| Breakpoint(s) within ESTABLISHED HI genes | ☐ 2I. Both breakpoints are within the same gene (gene-level sequence variant, possibly resulting in loss of function (LOF)) | See ClinGen SVI working group PVS1 specifications<br>• PVS1 = 0.90 (Range: 0.45 to 0.90)<br>• PVS1_Strong = 0.45 (Range: 0.30 to 0.90)<br>• N/A = 0 (Continue Evaluation) | | Assigned points: 0 |
| | ☐ 2J. One breakpoint is within an established HI gene, patient's phenotype is either inconsistent with what is expected for LOF of that gene OR unknown | 0 (Continue evaluation) | 0 | |
| | ☐ 2K. One breakpoint is within an established HI gene, patient's phenotype is highly specific and consistent with what is expected for LOF of that gene | 0.45 | 0.45 | Assigned points: 0 |
| Breakpoints within other gene(s) | ☐ 2L. One or both breakpoints are within gene(s) of no established clinical significance | 0 (Continue evaluation) | 0 | |