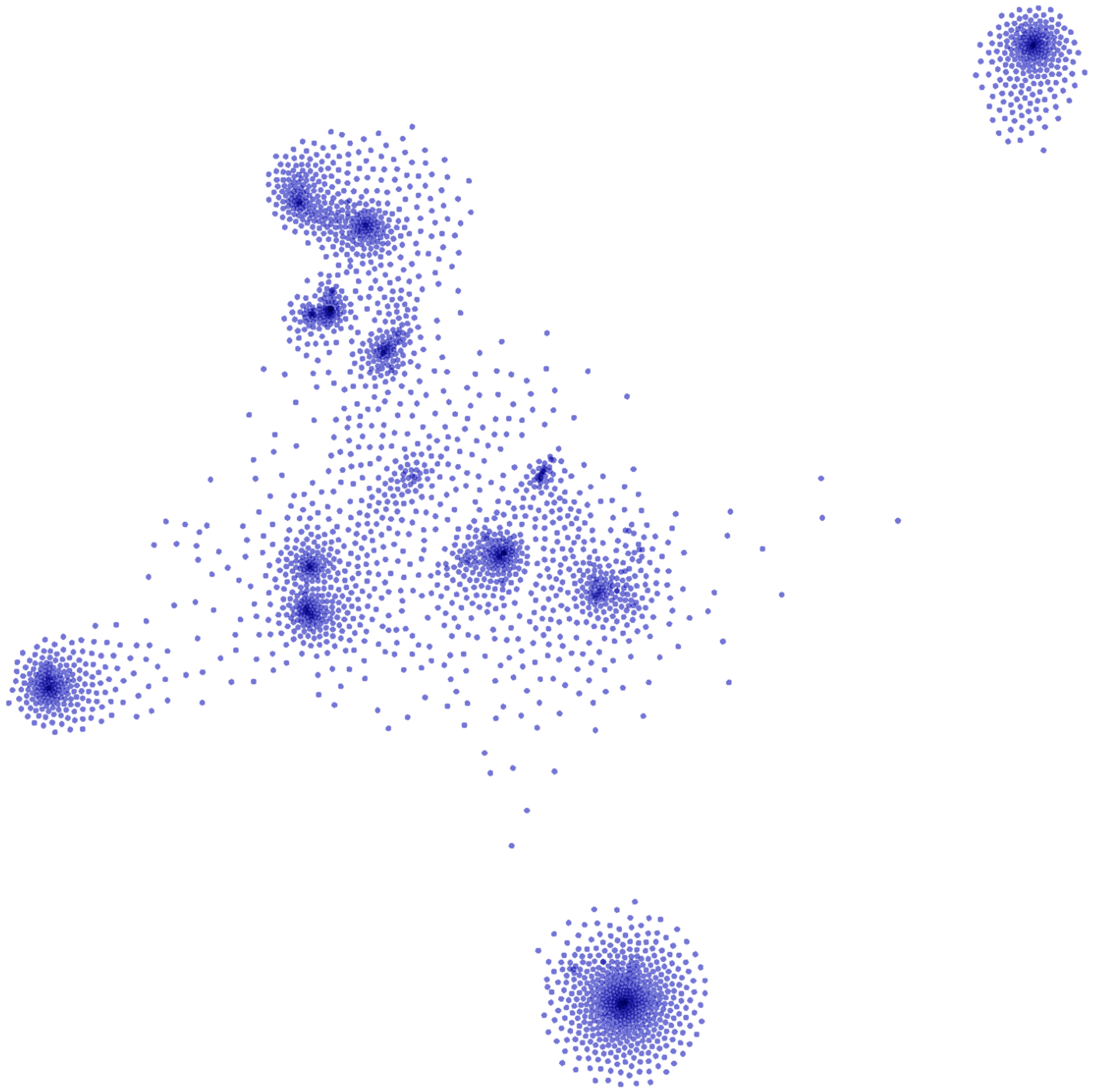
2D Visualization of the CNV Training Dataset



Predicting the Pathogenicity of Copy Number Variations

BCB330Y1 - 2018 Summer Project

Yoonsik Park, Researcher
Bank Engchuan, Supervisor
Brett Trost, Supervisor

Copy Number Variations

- Copy Number Variations (CNV) are defined as structural genomic variants, either duplications or deletions of sequences larger than 50 base pairs (bp)

- Associated with neurodevelopmental diseases such as autism and schizophrenia as well as other diseases.

- Some commonly occurring and/or large CNVs that have been described in research

Research Purpose

- The purpose of this project is to use state-of-the-art machine learning models to help clinicians and researchers quickly and confidently screen out non-pathogenic CNVs

- Furthermore, this project aims to understand the diversity of CNVs and CNV features, and understand how they play a role in pathogenicity using both visualization techniques and feature-importances from the trained models

Two Data Sources for CNVs in Population

- *DECIPHER* for pathogenic and non-pathogenic CNVs
  - Database of genomic variants that are clinically relevant or associated with rare diseases
- *DGV* for non-pathogenic CNVs
  - Database from The Hospital for Sick Children, consisting of mainly controls

Initial Dataset

- Sequence "loss" or "gain" converted to a binary variable:
  - `−1` for `loss`
  - `+1` for `gain`

- Pathogenicity description converted to a binary variable:
  - `−1` for `{benign, likely benign}`
  - `+1` for `{pathogenic, likely pathogenic}`

chr	start	end	size	gain_loss	pathogenicity
chr1	49911	222421	172510	-1	-1
chr1	542945	673049	130104	-1	1
chr1	837847	1477469	639622	1	-1
chr1	862453	1069517	207064	1	1

Extracted Feature 1: Gene Annotations

- Annotations from NCBI RefSeq file: `hg19_ncbi_refseq.txt`

- Converted NCBI accession numbers to Entrez IDs

- Determined if the CNV start and end intervals intersected the gene's `txStart/txEnd` intervals

- In the case of multiple annotations for the same gene —> used the widest possible interval

Extracted Feature 1: Gene Annotations cont.

chr	start
chr1	837847
chr1	536263
chr1	862453
chr1	668630

…

genes_overlapping	number_of_genes
83858;126789;81669;29101;339453;9636;126792;219293;84808;64856;26155;6339;55052;254173;80772;375790;388581;54587;54973;84069;148398;401934;54998;116983;339451;643965;51150;7293;118424;8784;55210;54991;1855;83756;441869;57801	36
81399	1
375790;84808;84069;148398;54991;401934;26155;339451;9636;57801	10
	0

Baseline Correlations

Feature	Pearson Correlation w/ Pathogenicity
size	0.649
number_of_genes	0.588
gain_loss	0.058

Extracted Feature 2: Mouse Phenotype Ontology (MPO)

- After converting human gene numbers to the mouse homologue, the MPO database describes the variety of mice phenotypes are associated with each gene

- Each phenotype is a column, with the count of every gene associated with it

- Finally, created a column for the total number of phenotypes associated

chr	genes_in_proximity		adipose tissue phenotype	behavior/ neurological phenotype	cardiovascular system phenotype	cellular phenotype		taste/ olfaction phenotype	vision/eye phenotype
16	11273;27040;79874;728	…	0	1	1	2	…	0	0
2	23040		0	0	0	0		0	0
10	196792;253738;1755;84		5	11	12	14		0	6

Extracted Feature 3: Online Mendelian Inheritance in Man (OMIM)

- The OMIM database simply describes if a gene is associated with a disease or not
- Feature is based on the number of CNV associated genes that appear in the OMIM Database

chr	start	…	genes_in_proximity	omim_num_diseases
10	123150811		196792;253738;1755;8433;3998	11
10	102969339		27343;8945;6468;10660;25911	0
22	23717624		266747;4320;4282;3543;7621;5	4
19	30379880		57616;8725;9745;22847;100507	0
13	93422696		10082;2262	1

Extracted Feature 4: pLI / Intolerance from ExAC

- Exome Aggregation Consortium (ExAC) has computed the probability, ranging from 0.0 to 1.0, that a gene is intolerant to a loss of function gene mutation

- Each CNV associated gene is added to bins according to its pLI value:
  {0.0-0.1}, {0.1-0.2}, {0.2-0.3}, {0.3-0.4},{0.4-0.5}, {0.5-0.6},
  {0.6-0.7}, {0.7-0,8}, {0.8-0.9}, {0.9-1.0}

chr	genes_in_proximity
16	11273;27040;79874;7284;9
2	23040
21	149998;54033;64092;6782
10	196792;253738;1755;8433;

…

pli_0.0_to_0.1	pli_0.1_to_0.2	pli_0.2_to_0.3	pli_0.3_to_0.4	pli_0.4_to_0.5	pli_0.5_to_0.6	pli_0.6_to_0.7	pli_0.7_to_0.8	pli_0.8_to_0.9	pli_0.9_to_1.0
4	0	0	0	0	0	1	0	1	3
0	0	0	0	0	0	0	0	0	1
4	0	0	0	0	0	0	0	0	0
47	4	0	5	2	4	3	3	1	4

Extracted Feature 5: Repetitive Elements

- This feature was inspired by Hehir-Kwa, J. Y. *et al.*, as they found the number and density of repetitive elements helped predict neurodevelopment pathogenicity

- Repetitive elements describe DNA patterns that repeat many times in the genome

- Two examples of repetitive elements, LINEs (Long Interspersed Nuclear Elements) and SINEs (Short Interspersed Nuclear Elements) account for at least 30% of human genomic DNA

- The number of repetitive elements intersecting the CNV start and stop locations were counted and categorized by type

Extracted Feature 5: Repetitive Elements cont.

Gap	Homo polymer	LINE	LTR	Low complexity	RNA	SINE	Satellite	Segmental duplication	Simple repeat	Trans posable element
0	1	84	51	17	2	353	0	0	29	29
0	2	113	37	58	0	129	0	2	49	34
0	2	172	87	80	0	155	0	0	57	49
4	84	5067	2362	989	24	5628	1	23	1608	1745

Extracted Feature 6: Densities

- The gene density was calculated as

$$\text{gene\_density} = \frac{\text{\# of genes}}{\text{size of CNV (kb)}}$$

- The density for each repetitive element:

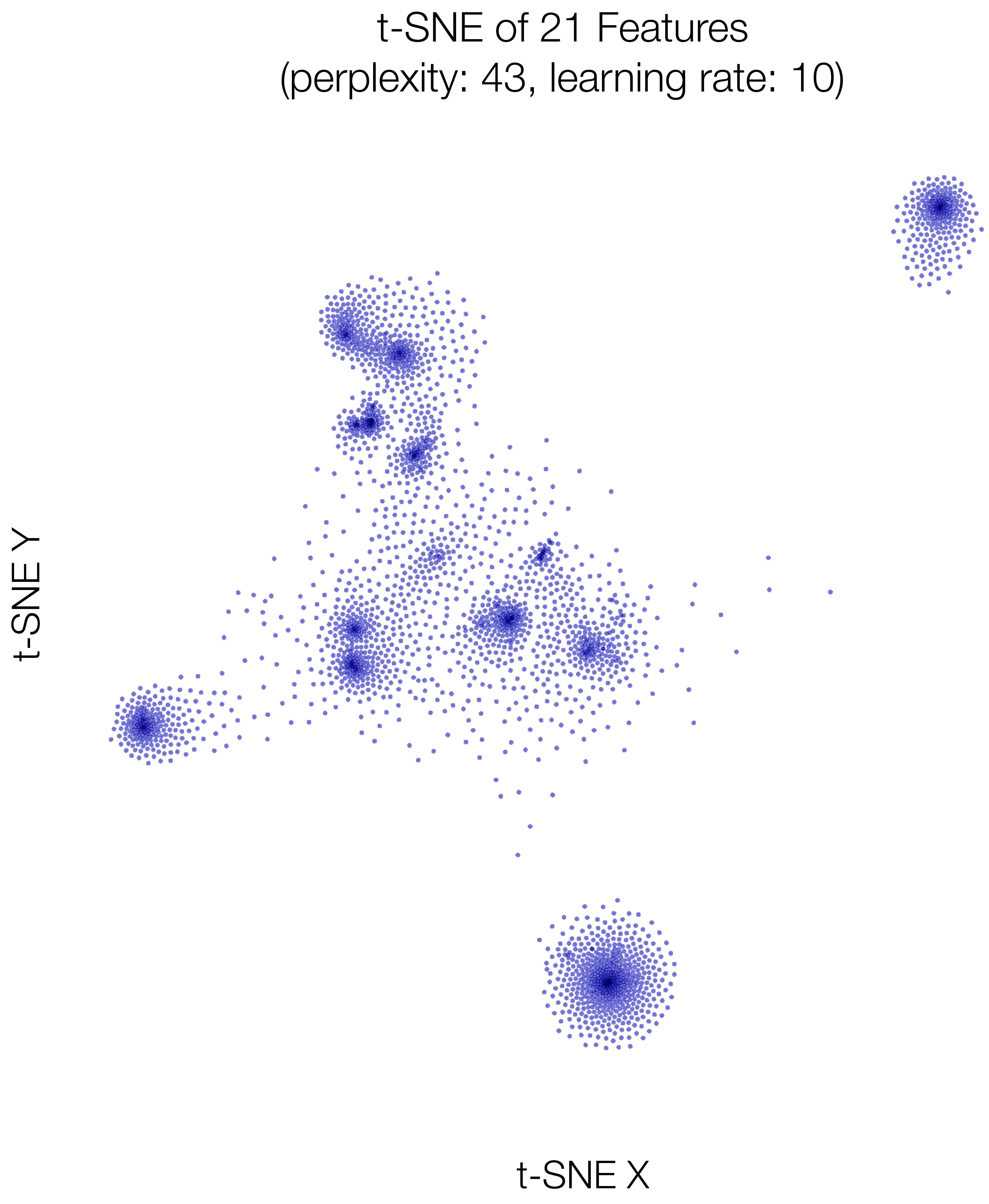
$$\text{repeat\_density} = \frac{\text{\# of repetitive elements}}{\text{size of CNV (kb)}}$$

Data Exploration

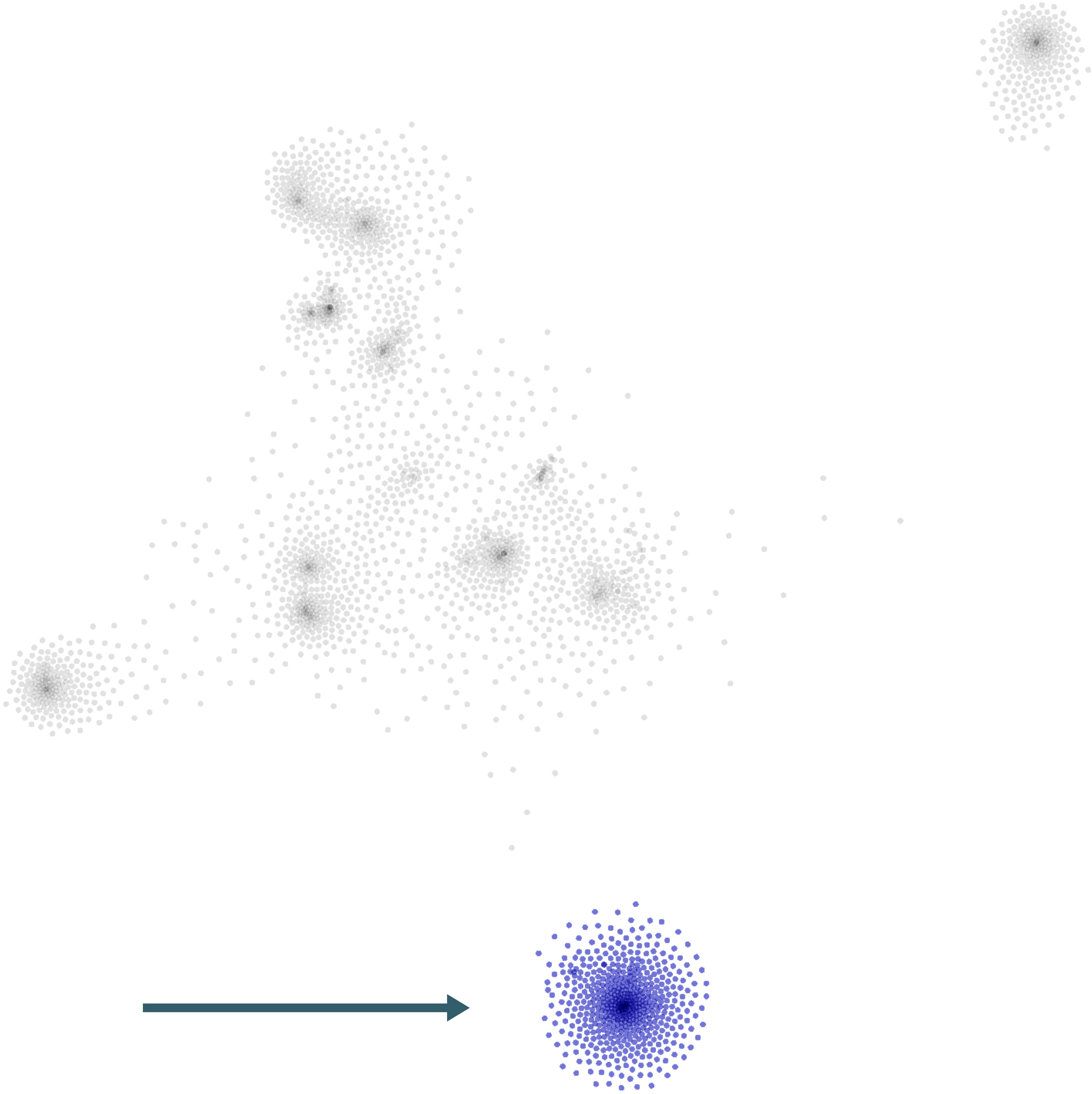
Dimensionality Reduction

- t-SNE constructs a probability distribution for each data point and its neighbours in both high dimensional and 2D space, then minimizes the divergence of the two distributions

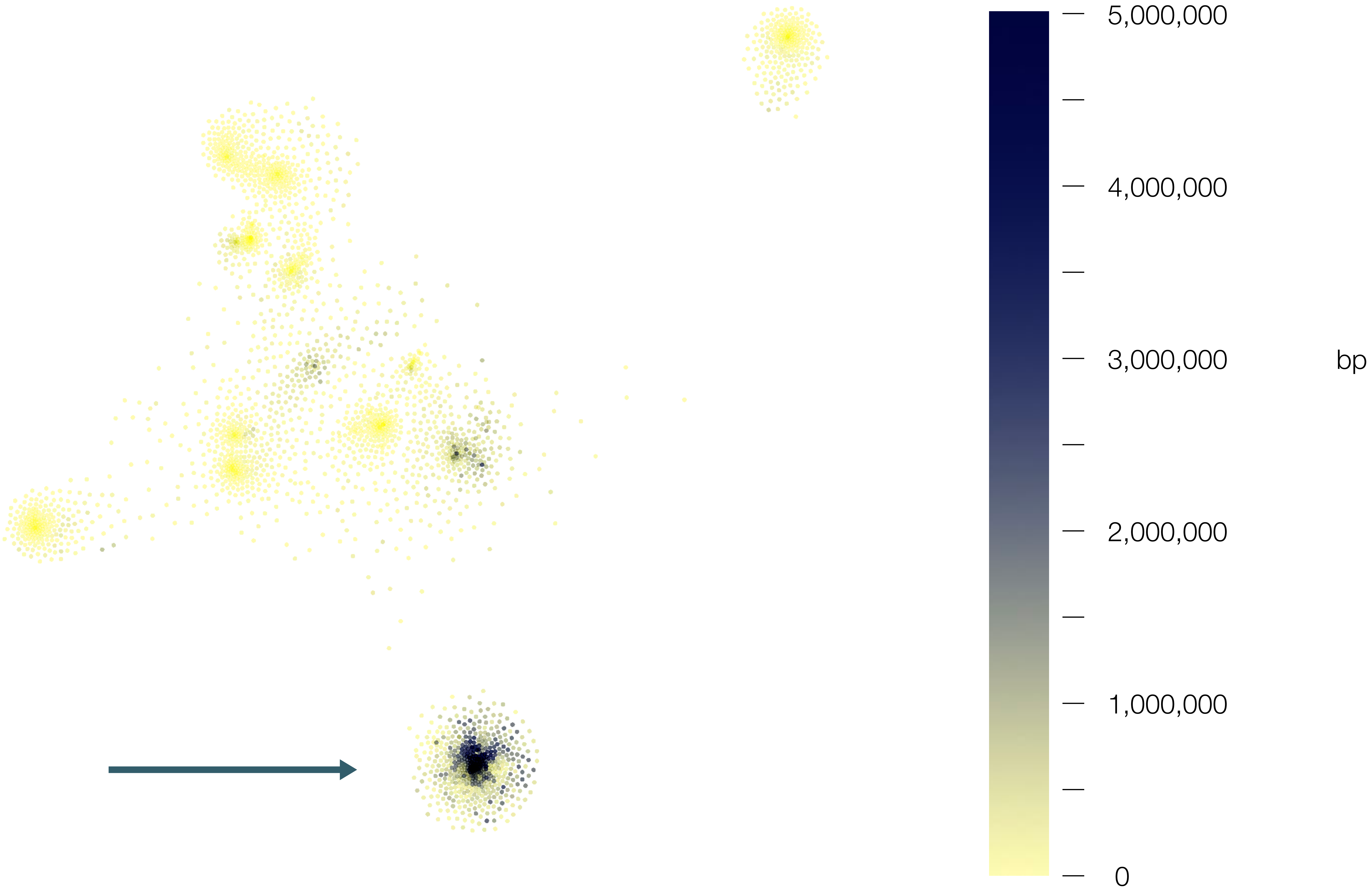
- Important structures and geometries emerge in the resulting t-SNE visualization



Cluster of Interest

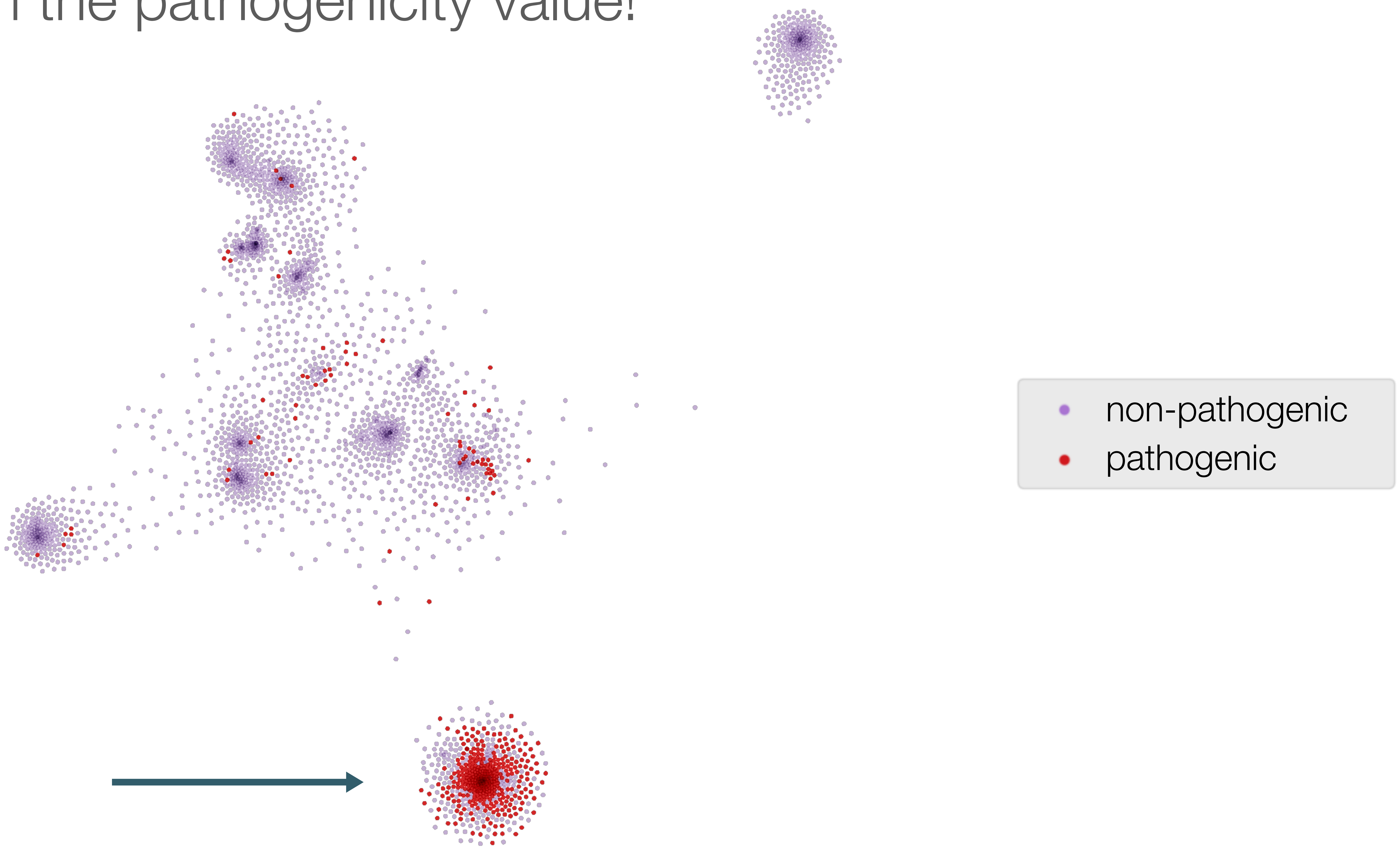


Coloured by: `size`



Coloured by: `pathogenicity`

- t-SNE was never given the pathogenicity value!

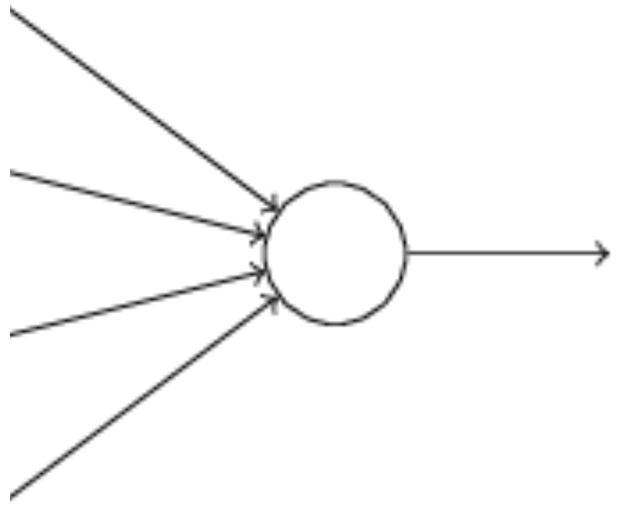


Models and Design

Machine Learning Methods Used

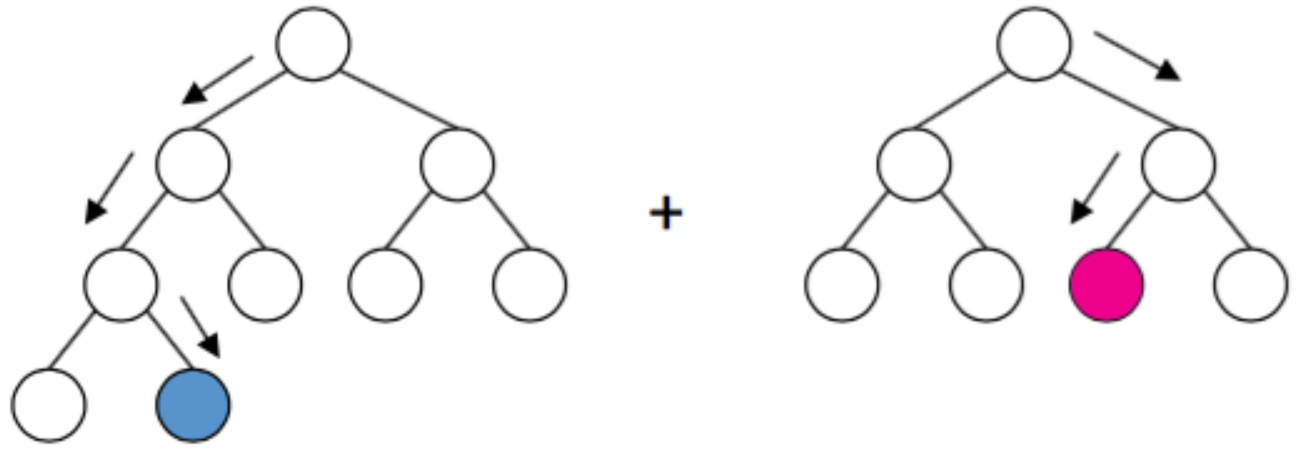
Logistic Regression

- Fast and great baseline
- Coefficients provide insight into important features



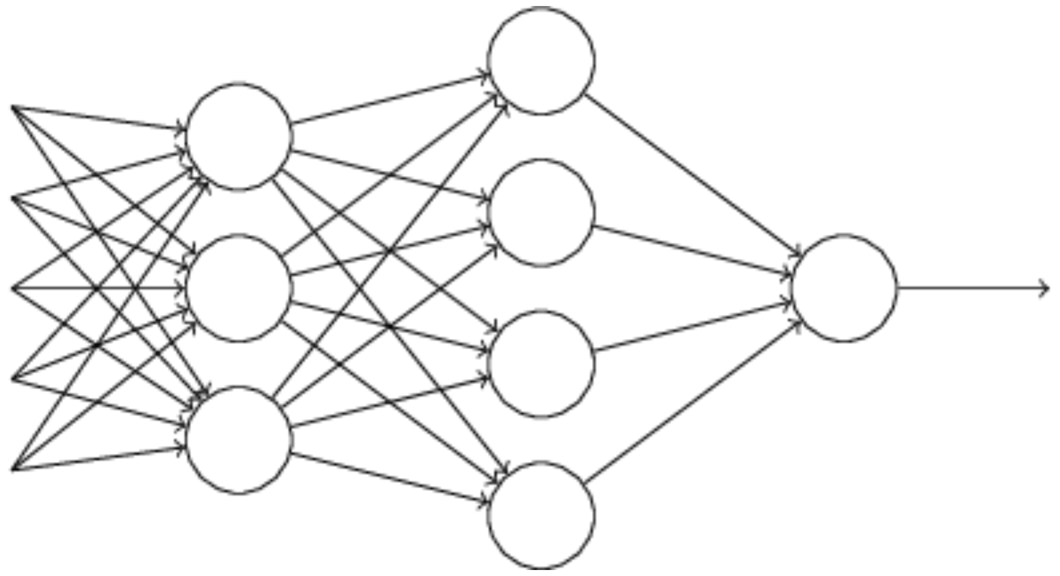
Gradient-Boosted Trees (XGBoost)

- Also fast, but many parameters to tune
- Provides a ranking of feature importances



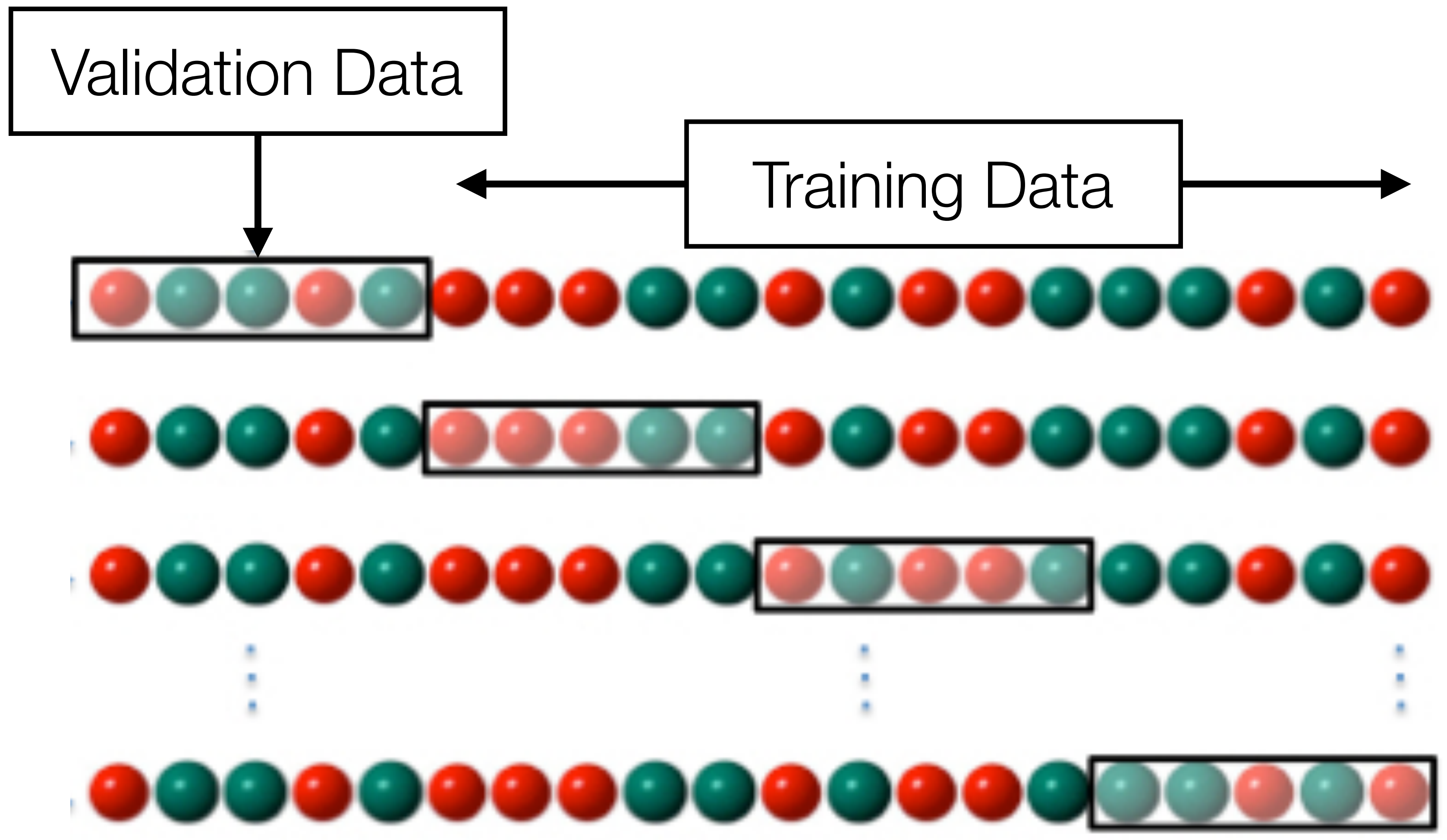
Fully Connected Neural Networks

- Slow, and many parameters to tune
- Black box, opaque model



Model Training and Testing Methodology

- 5-fold cross validation was used to assess the performance metrics of each model during the training phase. For each run:
  - Set aside 20% of the data as validation data
  - Use remaining 80% of the data for model training
  - Use validation data to assess the model's performance
  - Repeat with a new set of validation data
- After the training phase is complete, the models are tested on ClinVar, an independent testing set



Testing Data

Feature Selection Explained

- Currently there are 66 features based on the CNV

- If we can reduce the number of features, maybe this will make the models more generalizable and understandable

- Using the feature importance values generated by the XGBoost models, choose the top 10 overall features, and the top 4 from each "feature category" if possible

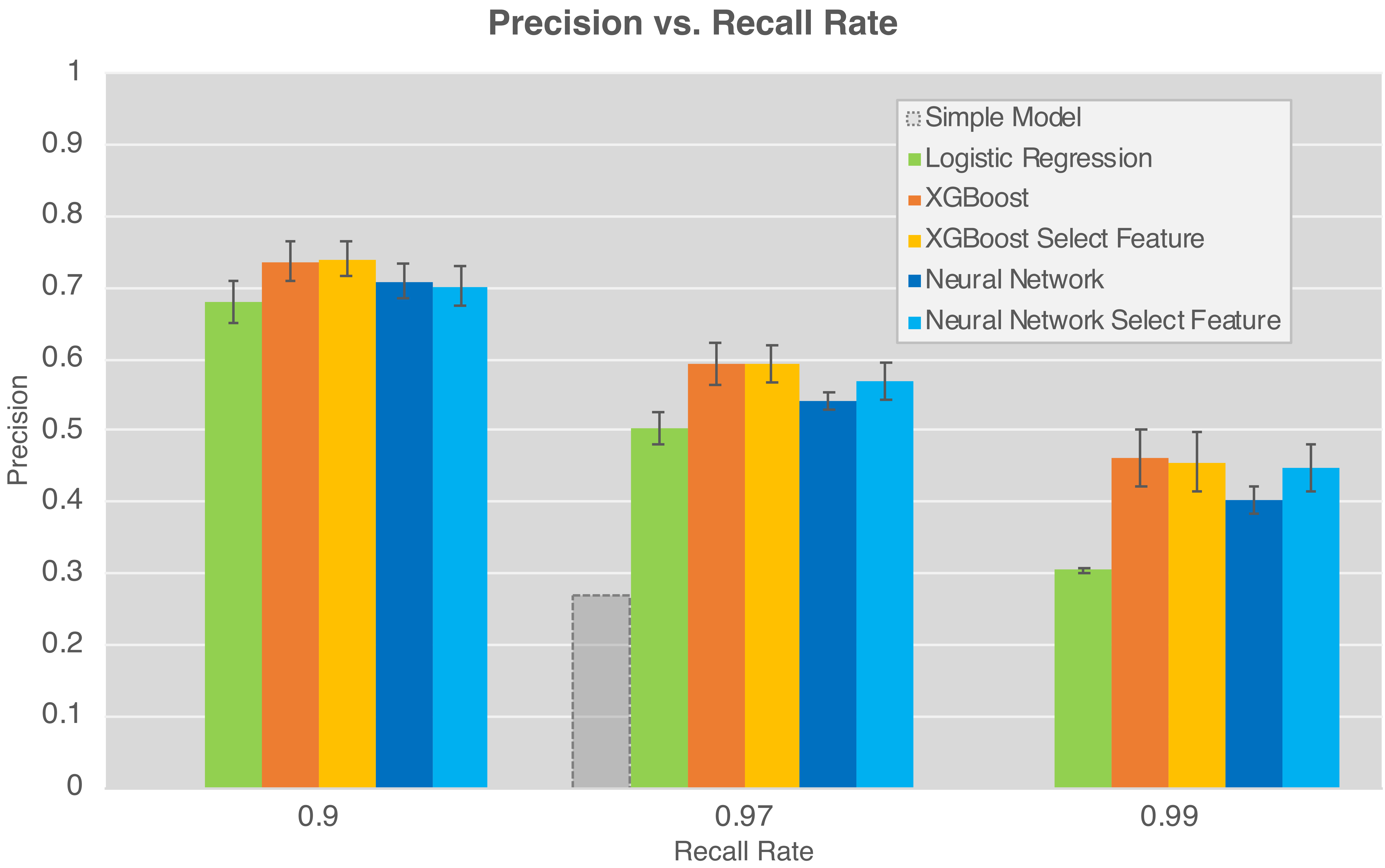
List of Features after Selection (ordered by importance)

- size
- repeat_LTR_density
- repeat_Simple_repeat_density
- repeat_SINE_density
- repeat_Transposable_element_density
- repeat_Low_complexity_density
- repeat_LINE_density
- repeat_Segmental_duplication_density
- repeat_LINE
- gene_density
- mpo_num_phenotypes
- gain_loss
- pli_0.9_to_1.0
- omim_num_diseases
- number_of_genes_in_proximity
- mpo_num_phenotypes_using_thresh
- pli_0.0_to_0.1
- mpo_behavior/neurological_phenotype
- mpo_growth/size/body_region phenotype
- pli_0.8_to_0.9
- pli_0.3_to_0.4

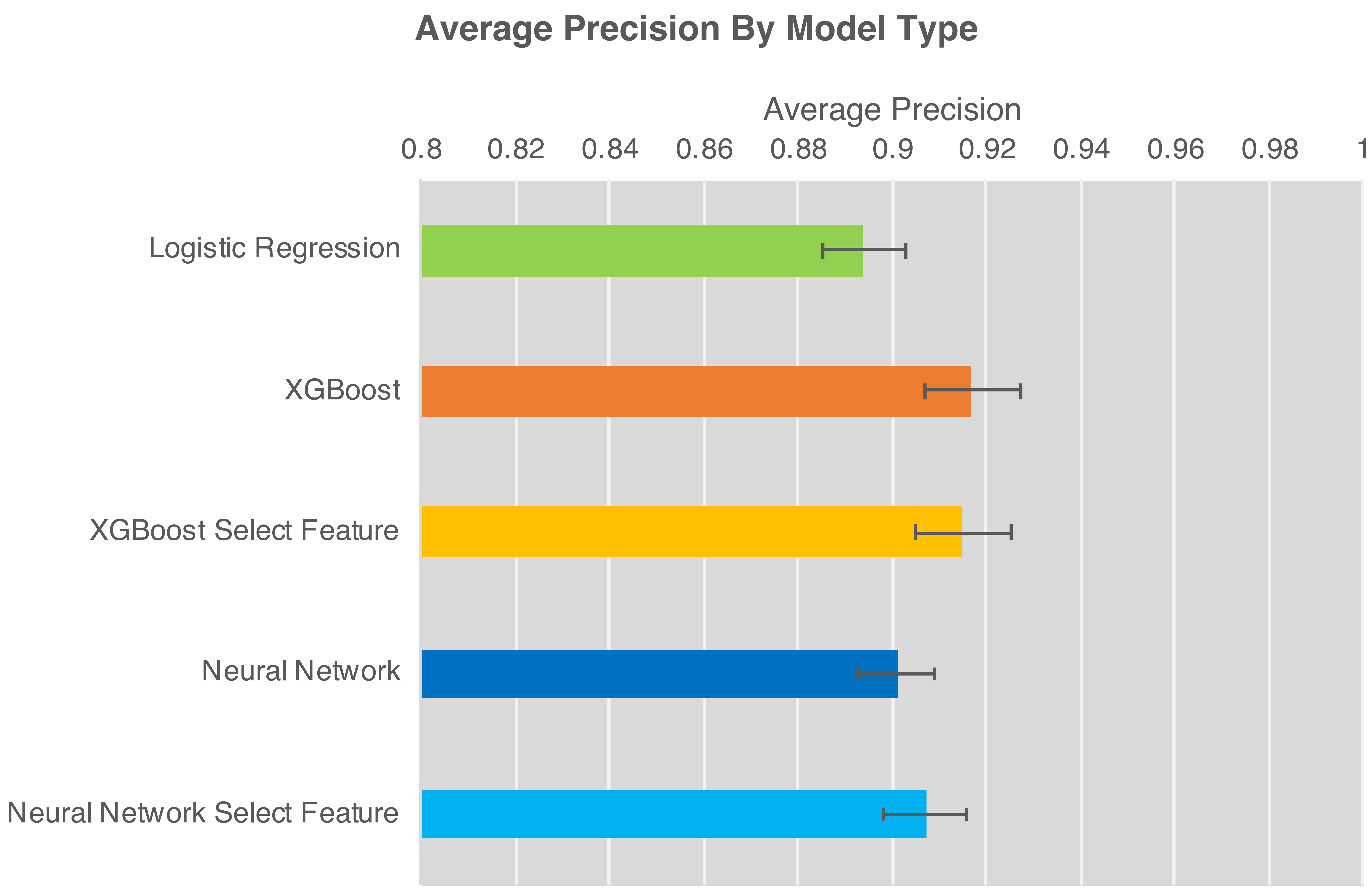
21 features!

Results

Precision at 90%, 97%, and 99% Recall



Average Precision



ClinVar Test Results

- All models were tested on an independent CNV database from ClinVar
- The database contains 15,000 CNVs that are known to be definitively benign or pathogenic

ClinVar Precision Test using "90% Recall" models

Neural Network All Features:	72.8% Precision, 95.8% Recall
XGBoost All Features:	69.7% Precision, 95.5% Recall
XGBoost Select Features:	65.2% Precision, 95.6% Recall
Neural Network Select Features:	63.2% Precision, 96.2% Recall

Pathogenicity Prediction Summary

- The XGBoost models performed best during the training phase, achieving up to ~59% Precision at a 97% Recall rate
- However, the Neural Network "all features" model performed best on ClinVar, achieving 73% Precision at a 96% Recall rate
- On ClinVar, "All Features" tested much better than "Select Features" (up to 73% vs 65% precision), indicating that information useful to generalization was lost during feature selection

Important Features

- Size of the CNV was the most important feature, although not all large CNVs are pathogenic
- Repetitive element densities were the next most important features
  - A variety of benign CNVs have high repetitive element densities, except ...
  - Many pathogenic CNVs have a high SINE density!
- Gene density was an important feature
- The number of MPO phenotypes associated per CNV was also important