A Deep Learning Approach to Identifying Complex Structural Variation in the Genome

July 22nd, 2026

Bhargav Srinivasan | bds062@umd.edu

Computer Science + Biological Sciences Undergraduate Student, UMD

Summer Student, Kolmogorov Lab, National Cancer Institute

DEPARTMENT OF COMPUTER SCIENCE

What is structural variation?

- Large-scale mutations ranging from 50 bp to 1 Mb +

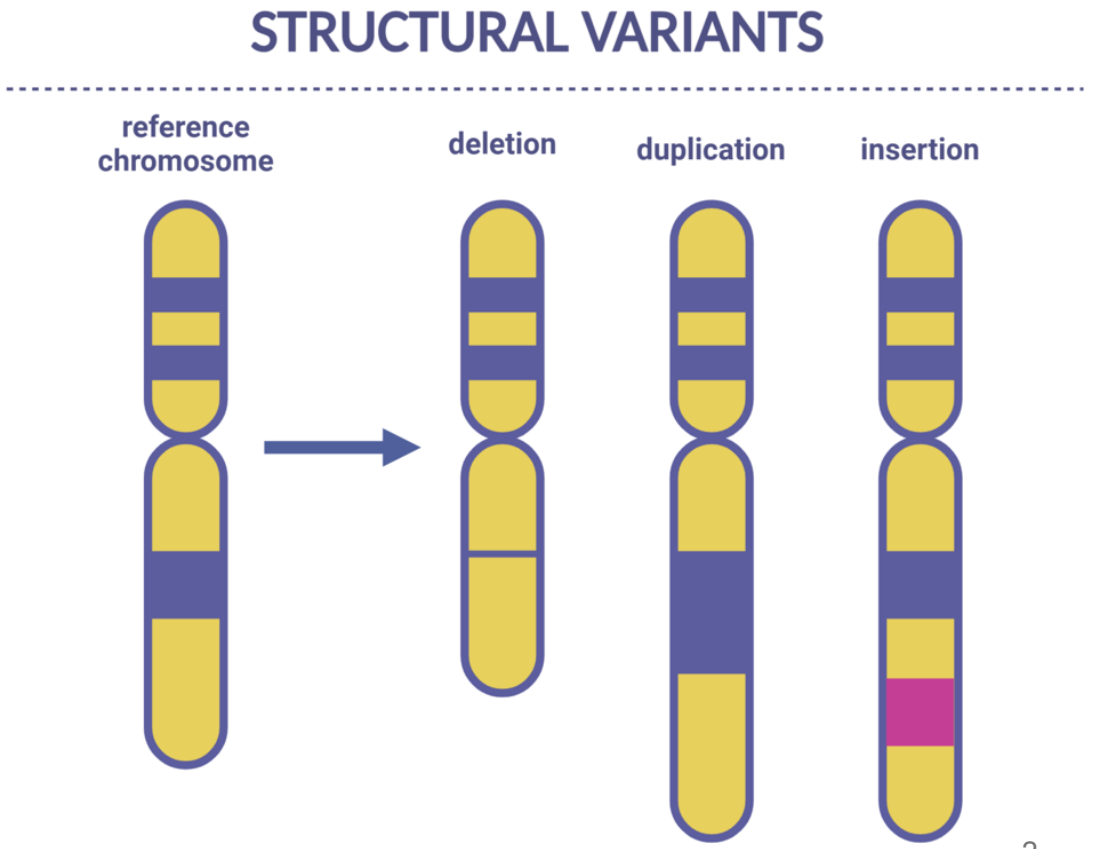
Many tools call SVs directly:

Long Reads:

- Severus
- Sniffles2
- Cue2

Short Reads:

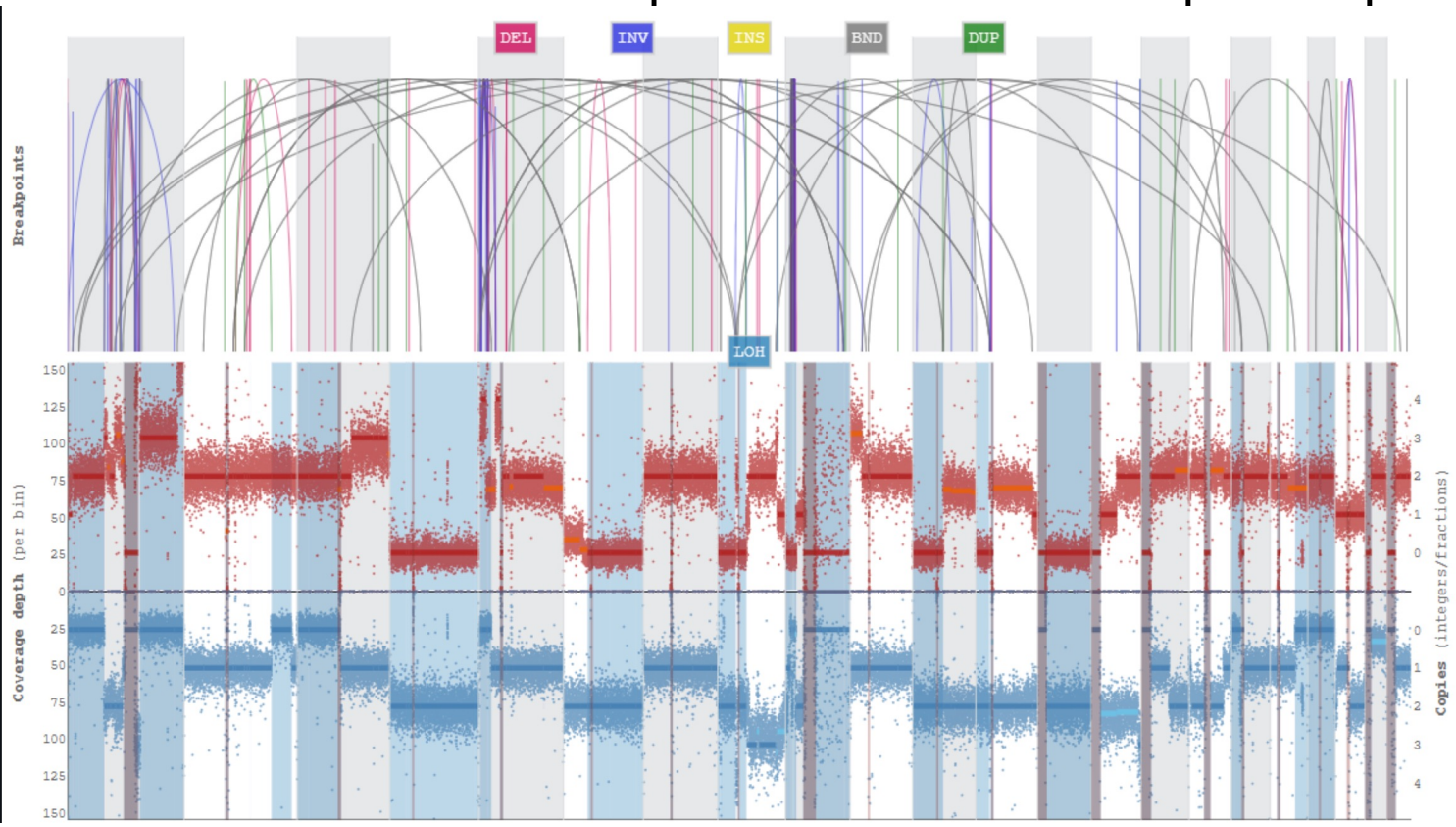
- Delly
- LUMPY
- Manta

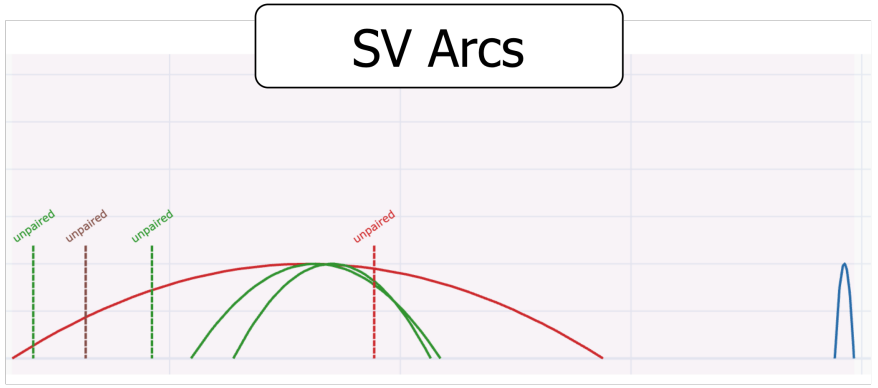


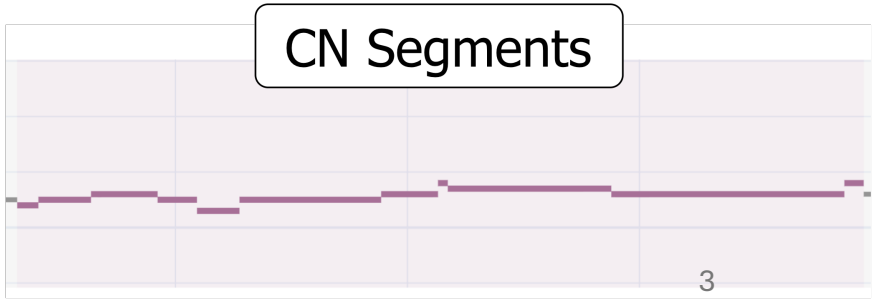
Structural variation causes copy number alterations

Deletions, Duplications, and Inversions **create complex copy number alterations** (CNAs)

- CNAs correspond to read depth, split by SV breakpoints

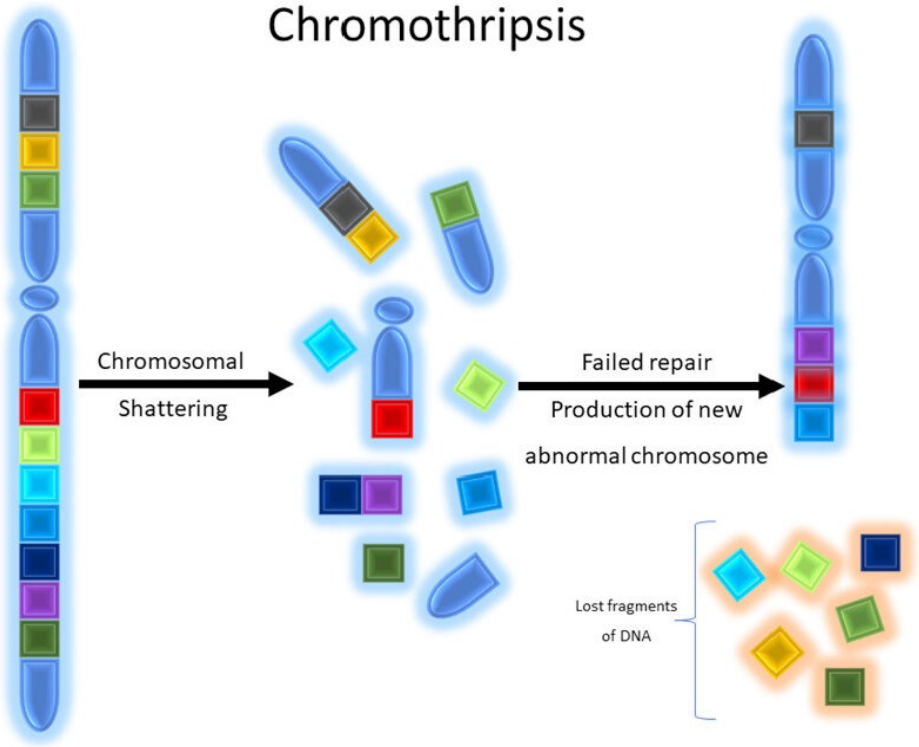


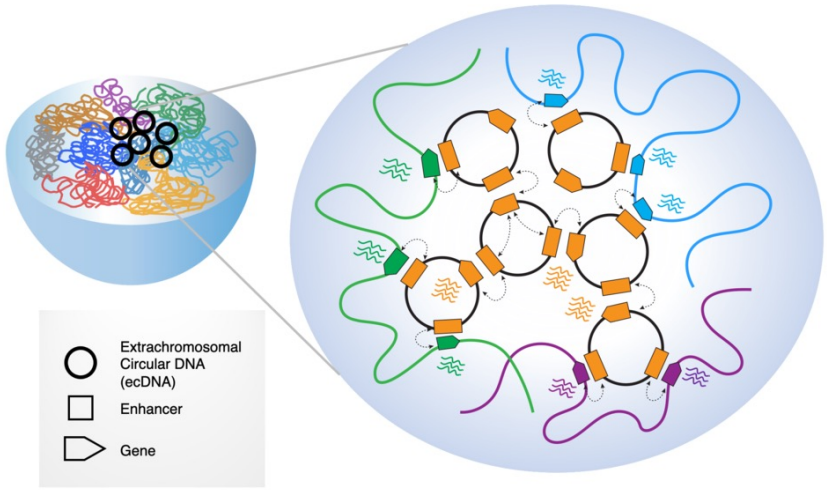




What is complex structural variation?

- Characterized by a cluster of specific SVs and CNAs
- Separately derived from individual biological processes
- Potential for oncogenesis

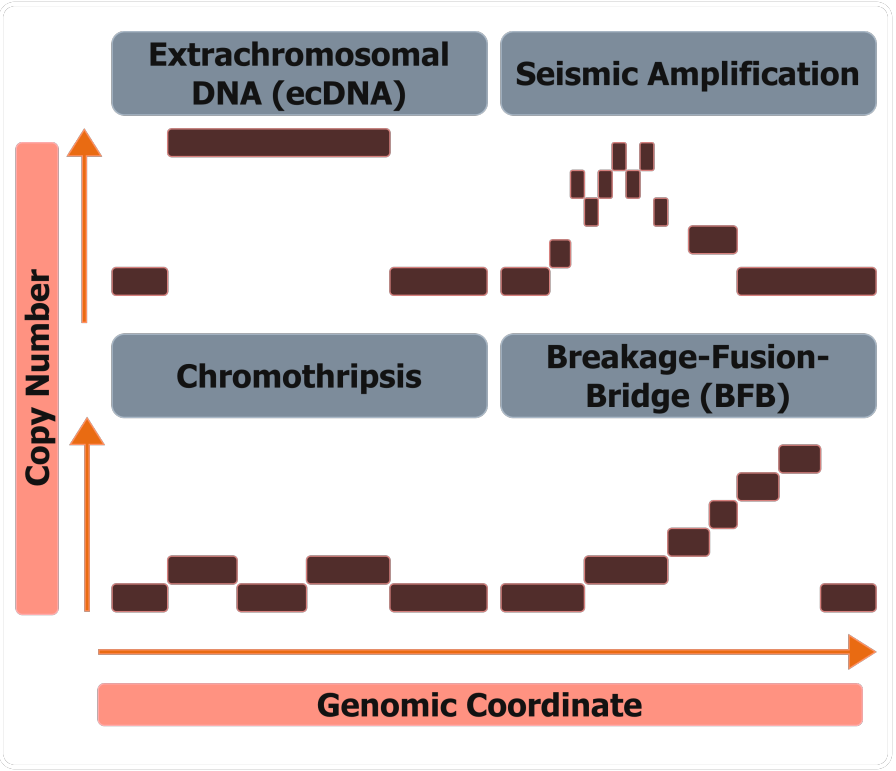




"EcDNAs are Extra Active and Terrifying" - NCI

Why study Complex Structural Variation?

- Drivers for oncogenesis
- Long Read Sequencing allows discovery
- Limited discriminatory models



- Broad definitions
- Intercorrelated events
- No tool can classify all four events

Can we create a deep learning approach to classifying and localizing complex SVs?

Project Goals:

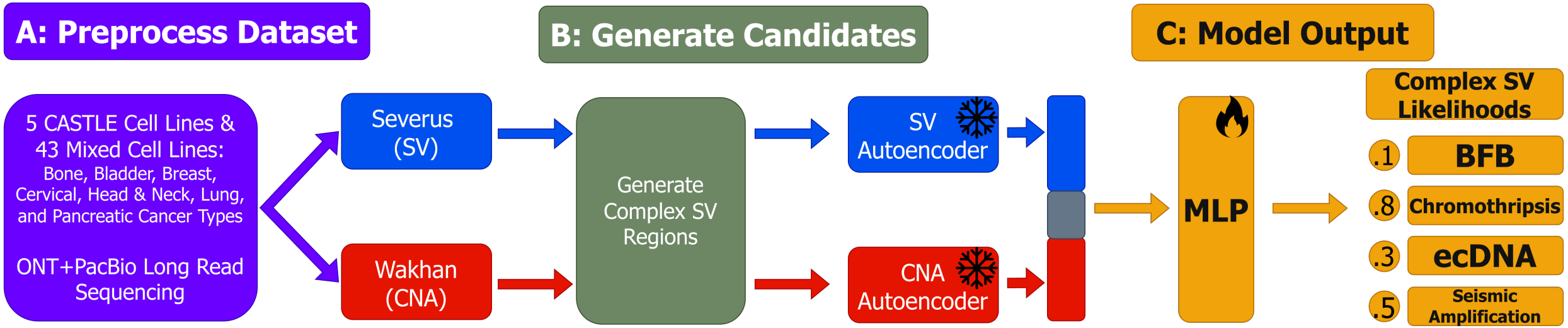
1)Classify complex SVs given a genome

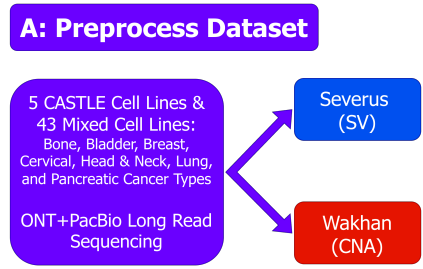
2)Identify subtypes of specific SVs: canonical vs noncanonical

3)Leverage unsupervised learning techniques to identify new forms of complex SVs

How can we predict complex SVs with deep learning?
Our Method:

A) Call SVs and CNAs
B) Featurize candidate regions with **SV+CNA aware embeddings**
C) Classify candidate regions





Label Generation

Event-specific callers produce training labels

Labels generated from 48 total Cell Lines:

- 9 Breast, 4 Pancreatic, 1 Bone, 1 Bladder, 3 Head & Neck, 28 Cervical, and 2 Lung Cancer Cell Lines

A. Candidate Region Generator

1) High Density SV Regions
2) Unusual Copy Number
3) Foldback Inversions
4) Distant SV Breakends

→ Candidate Region

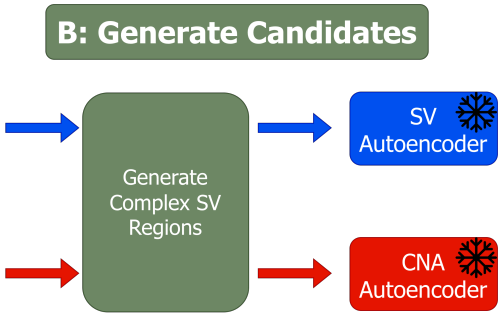
B. Method-Based Labeling Pipeline

CoRaL - **ecDNA**

BFBArchitect - **BFB**

Shatterseek - **Chromothripsis** + **Seismic Amplification**

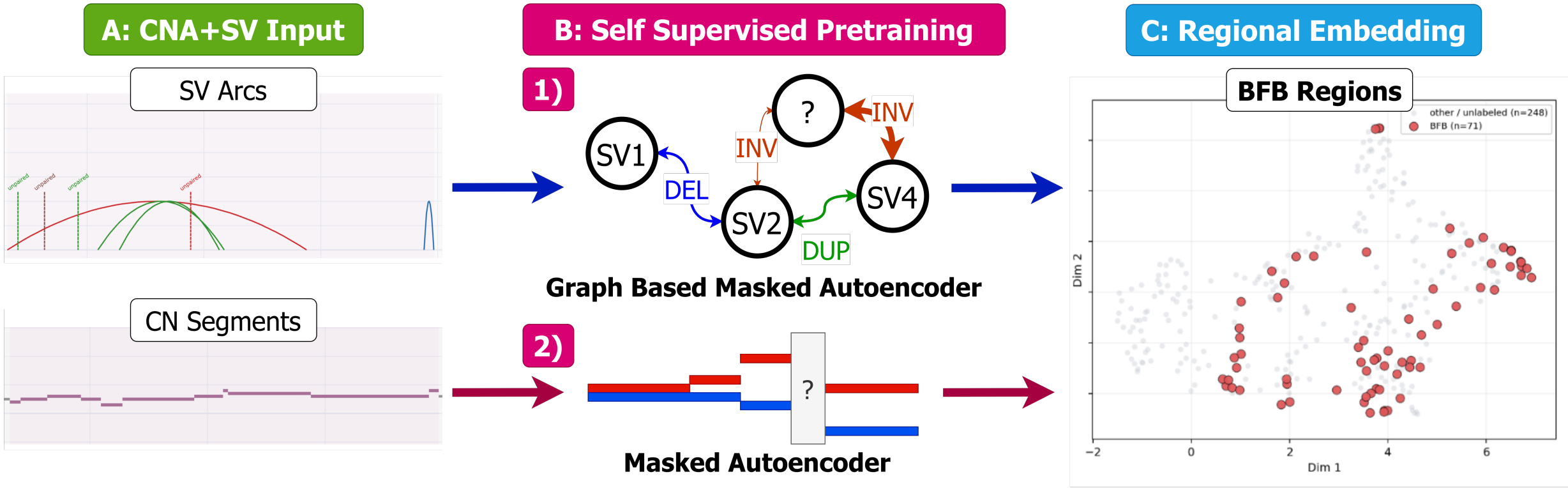
Featurizing Candidate Regions



A) Identify clustered CNAs and SVs

B) Train independent **masked autoencoders** for SVs and CNAs

C) **Concatenate embeddings** + regional statistics



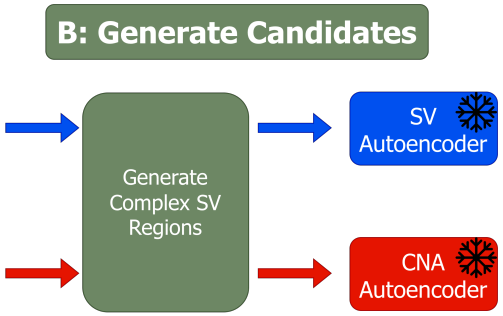
Featurizing Candidate Regions

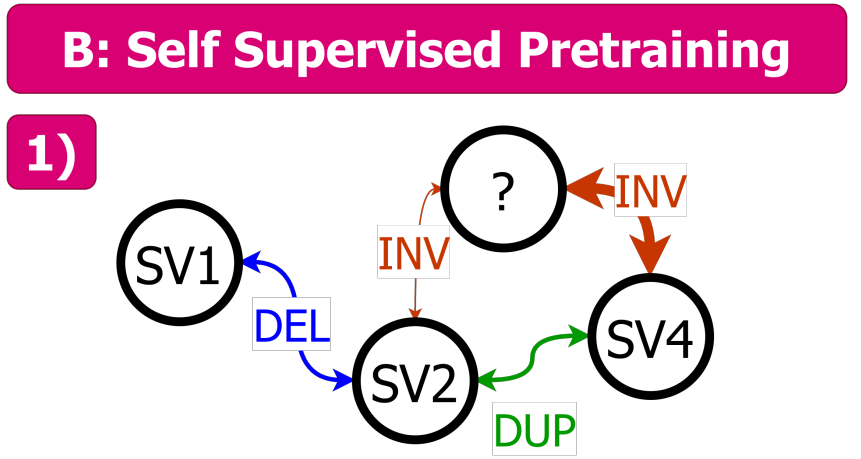
B1 – SV Autoencoder) GNN trained on breakpoint reconstruction

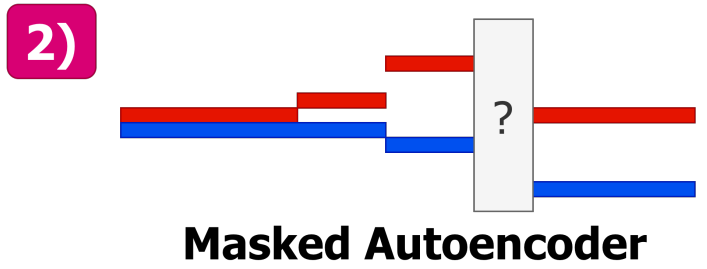
- Edges consist of phasing, locality, and breakpoints

B2 - CNA Autoencoder) **CN, read depth, and mapping quality** reconstruction

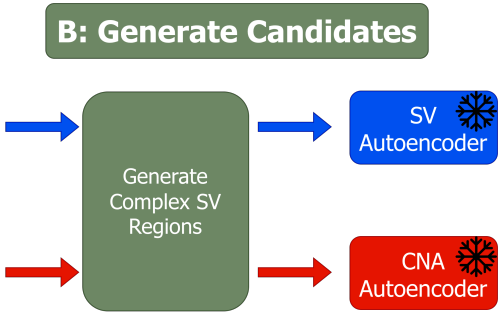
- Down sampled to fixed resolution

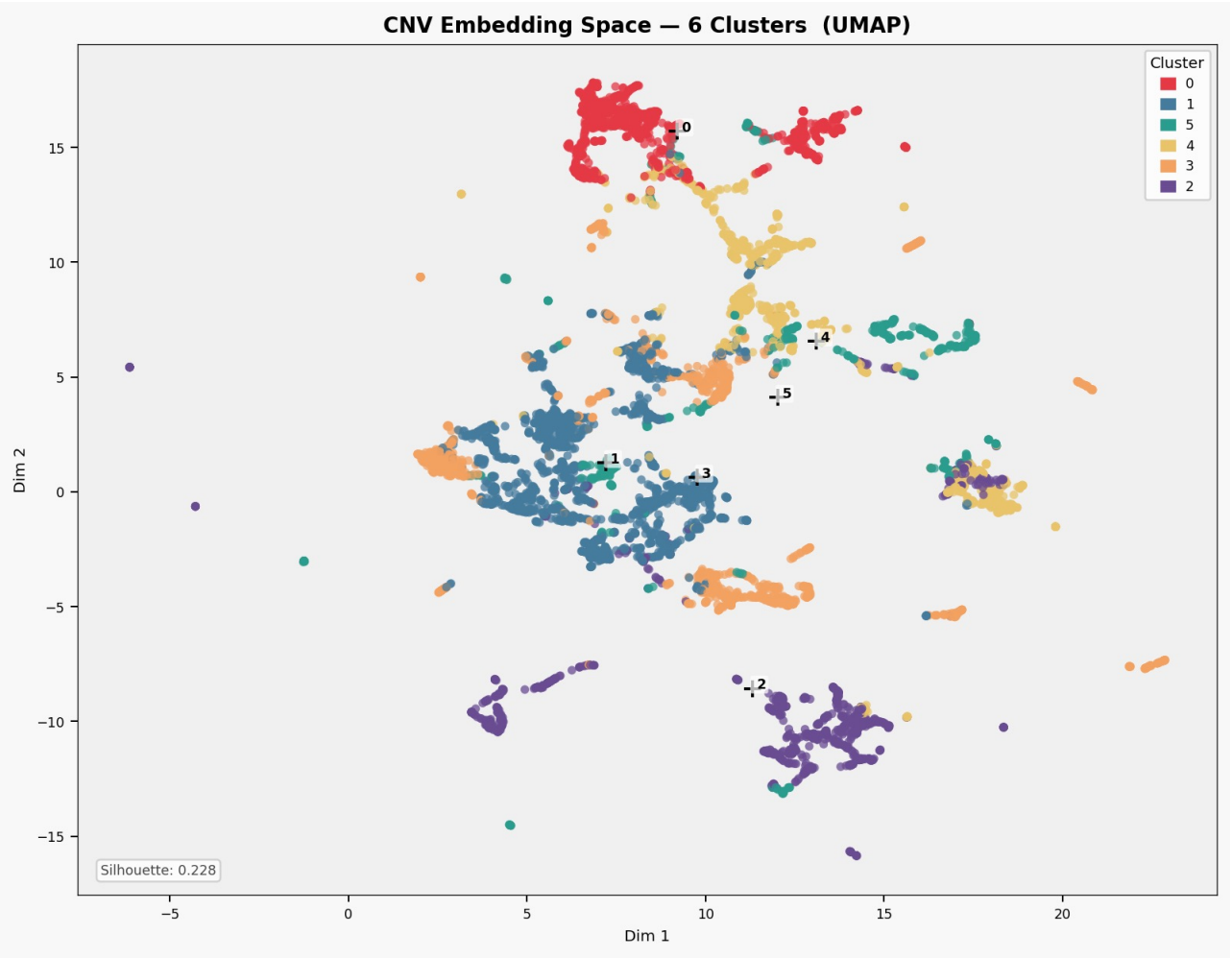


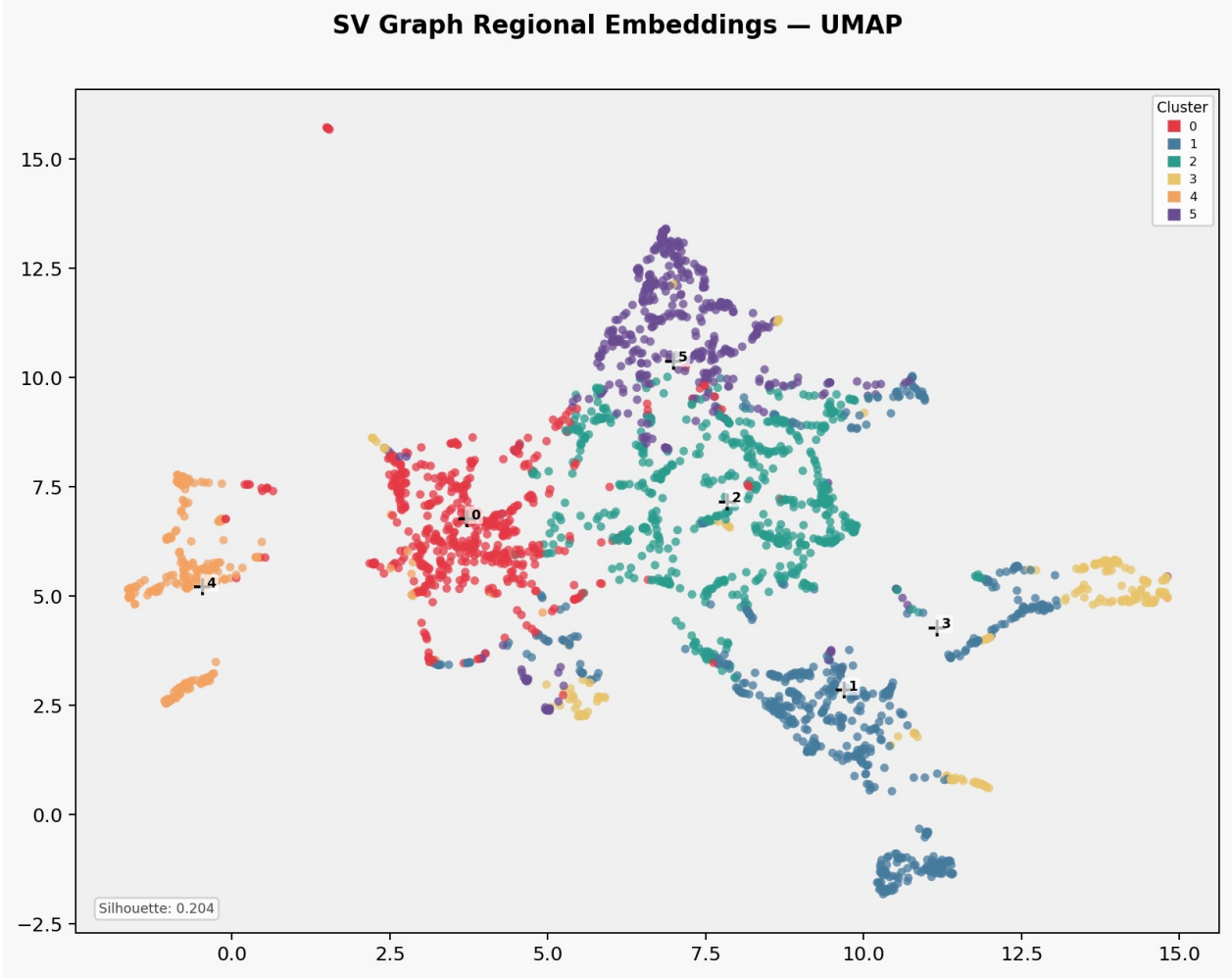


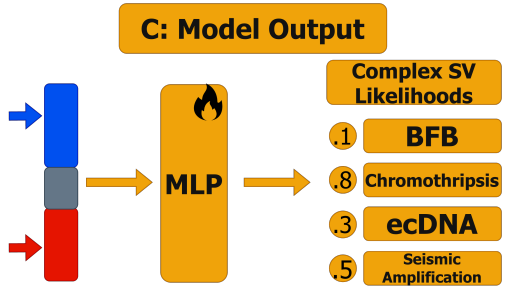


Featurizing Candidate Regions





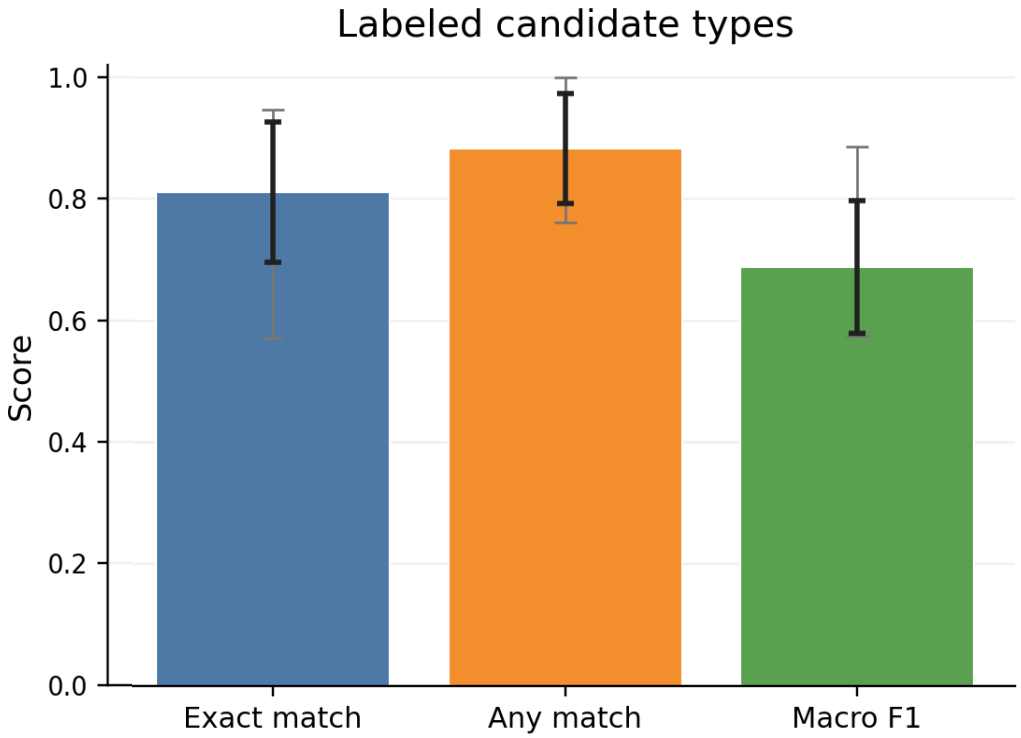


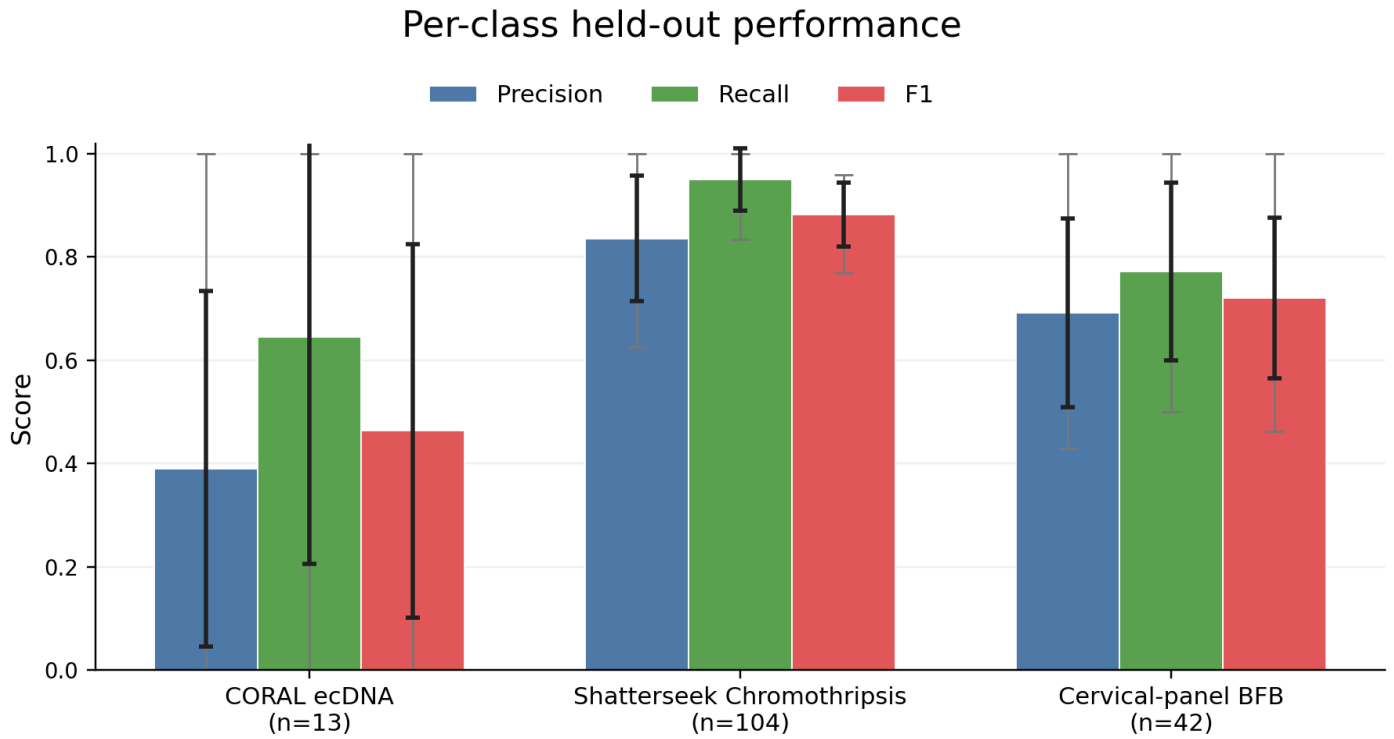


Classification Results

Model trained on output labels shows exact match accuracy of 0.81 and macro F1 of 0.69.

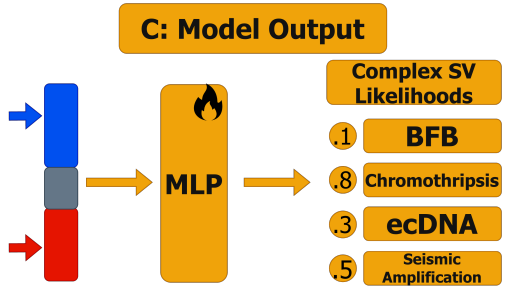
F1 and Exact-Match Accuracy across testing folds

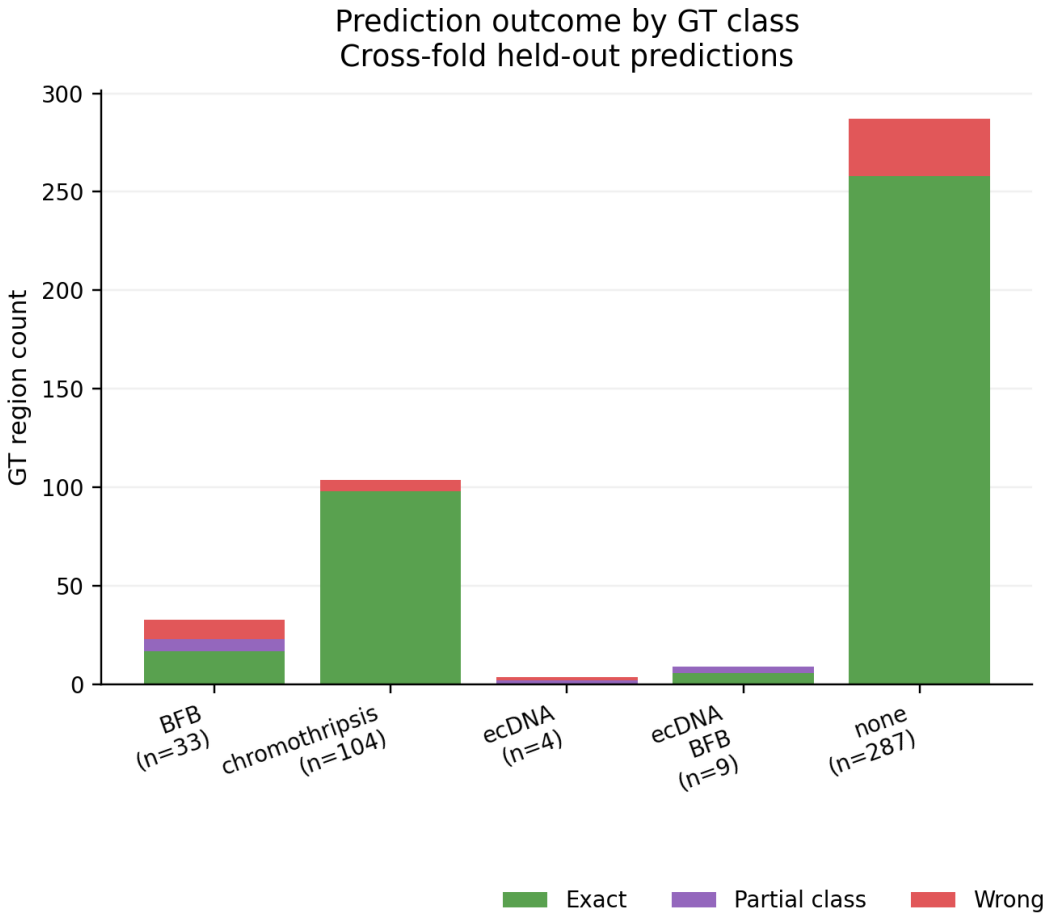


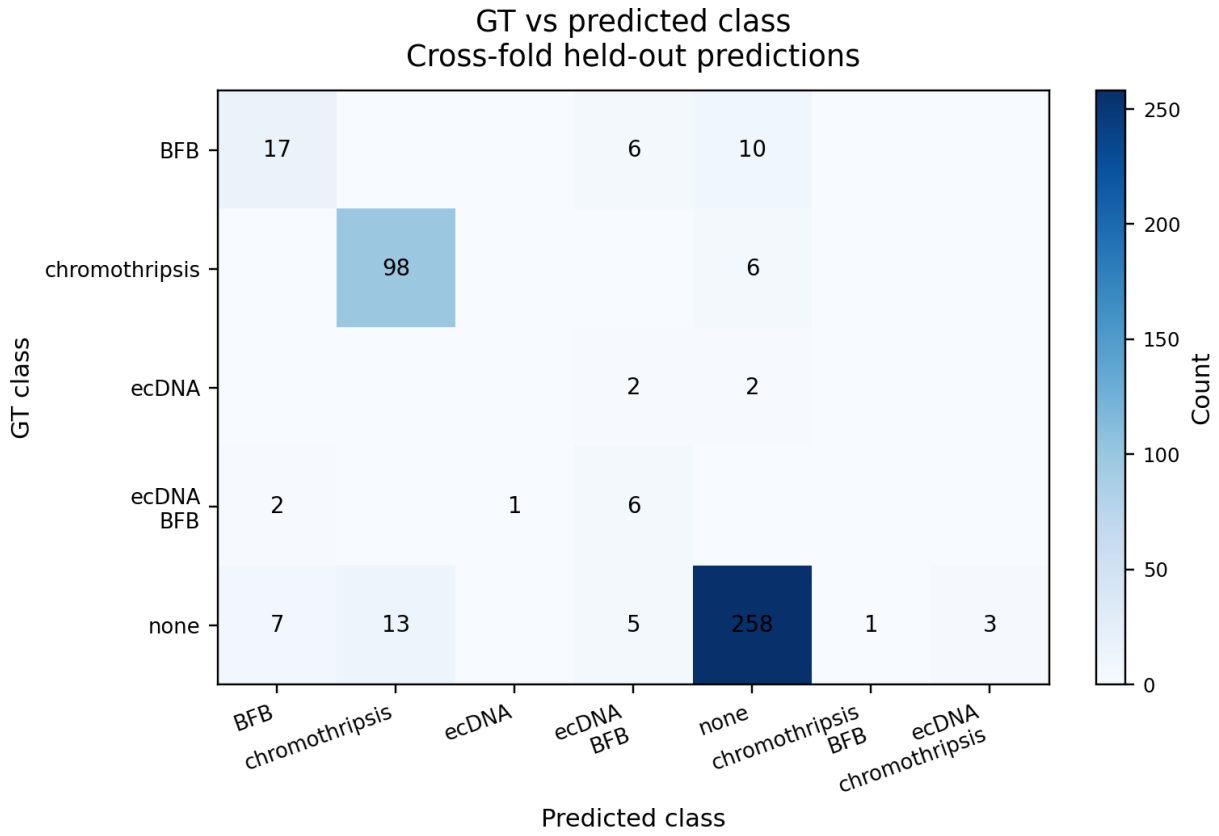


Classification Results

Given a candidate region, the model classifies accurately.

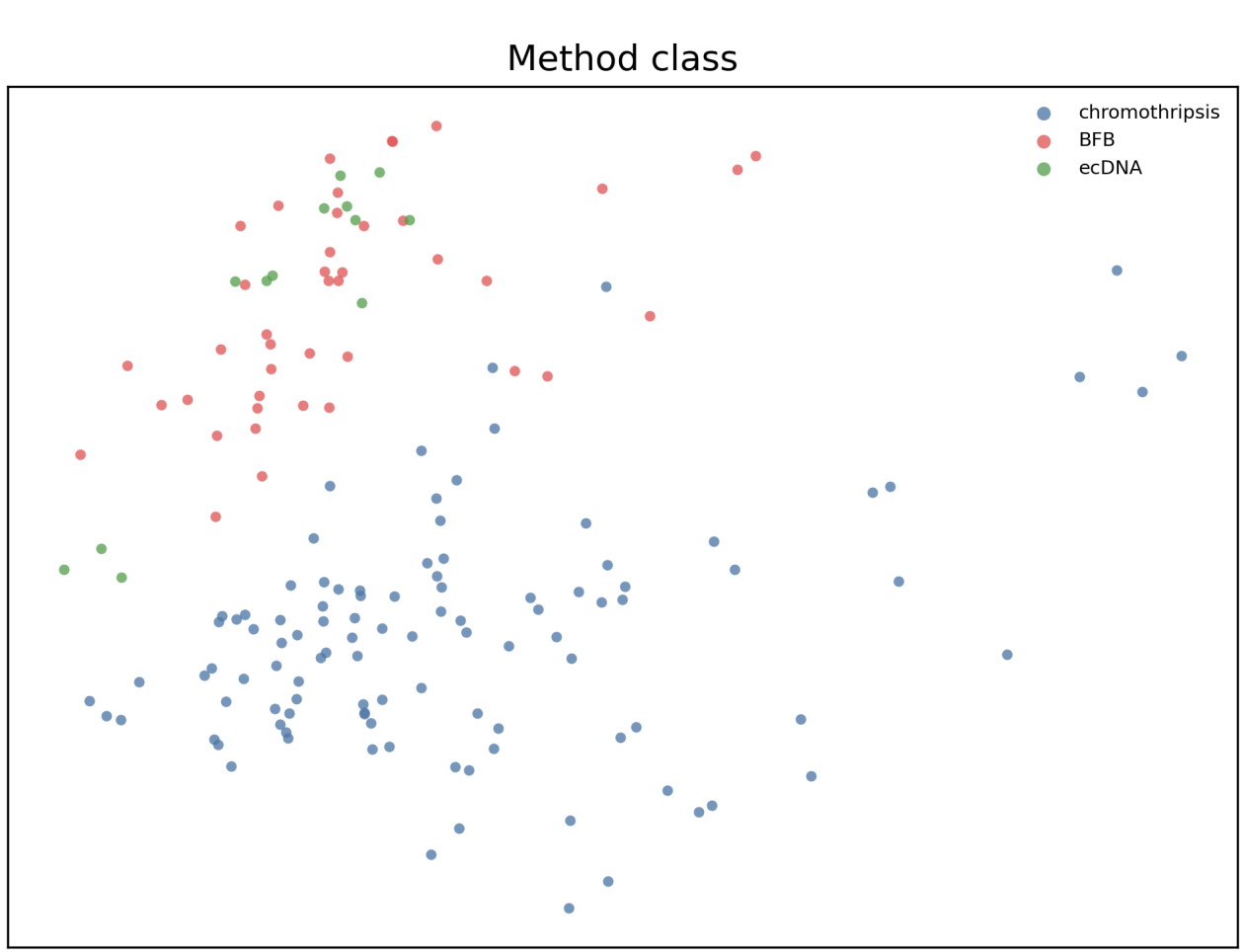


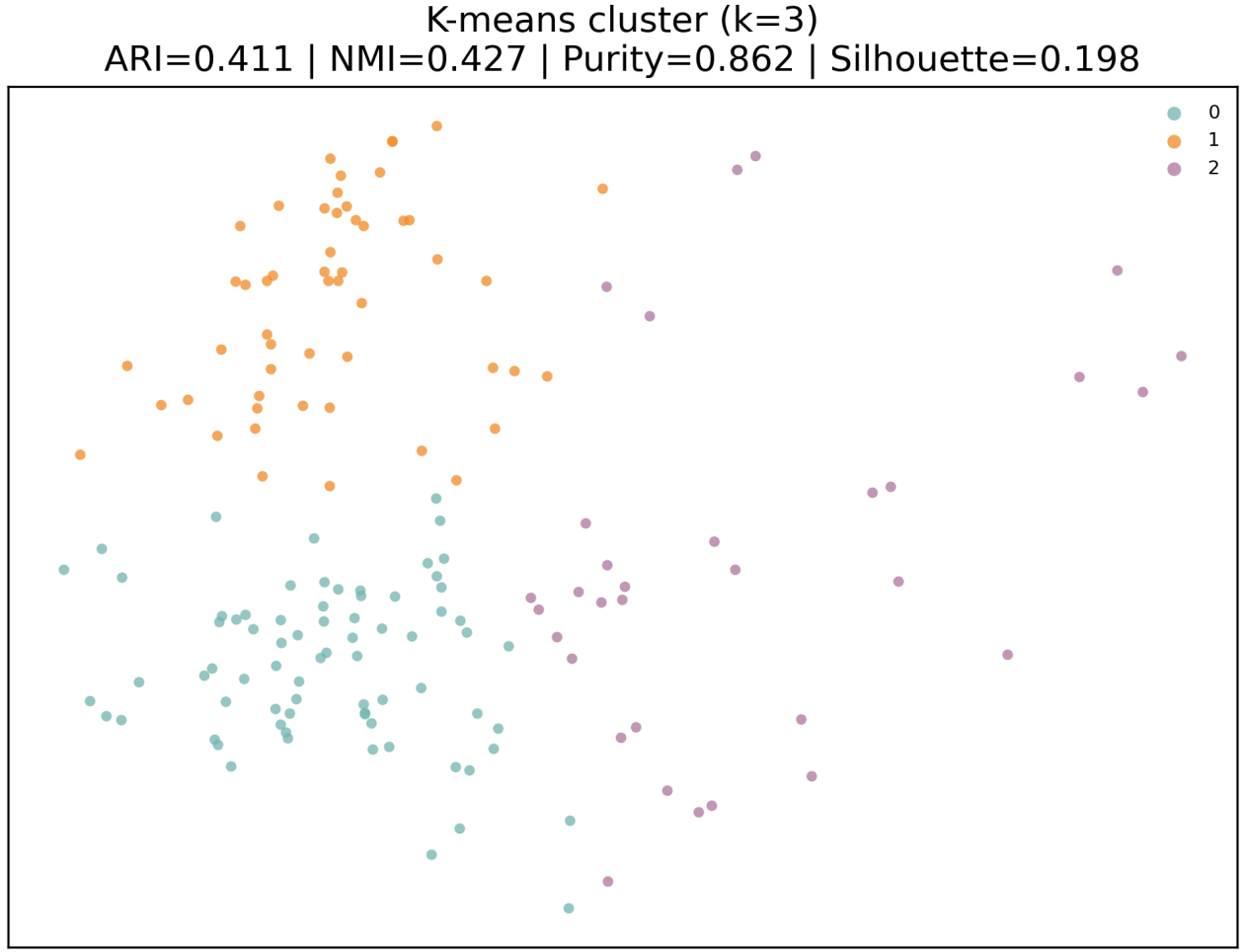




Discovery Results

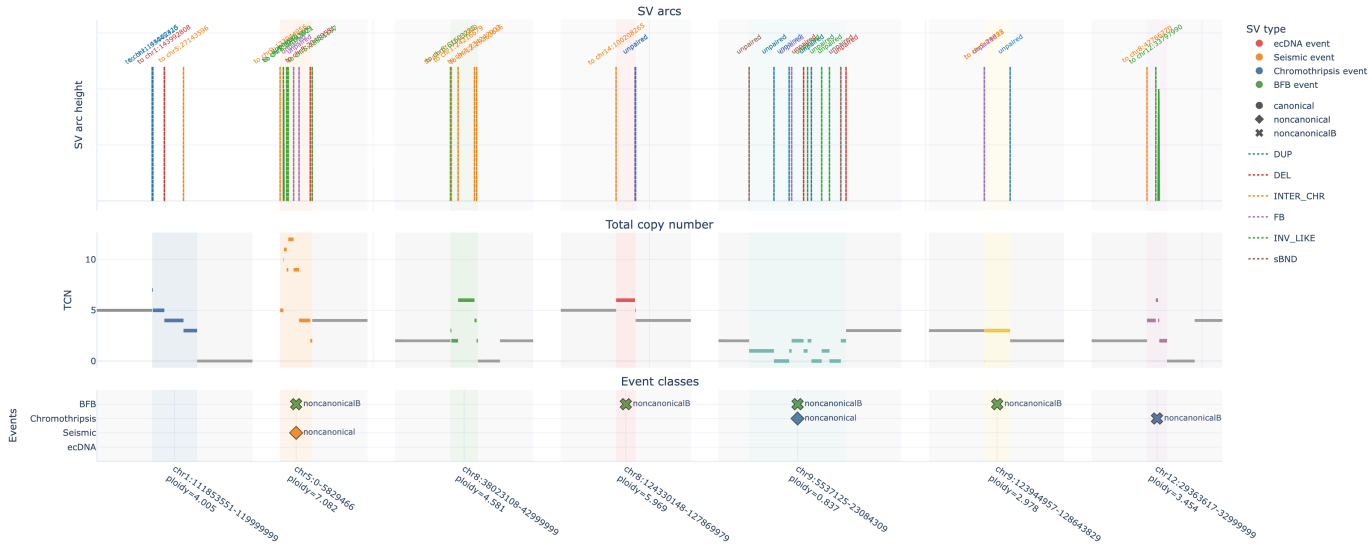
Can we use an unsupervised approach to identify new complex-SVs?

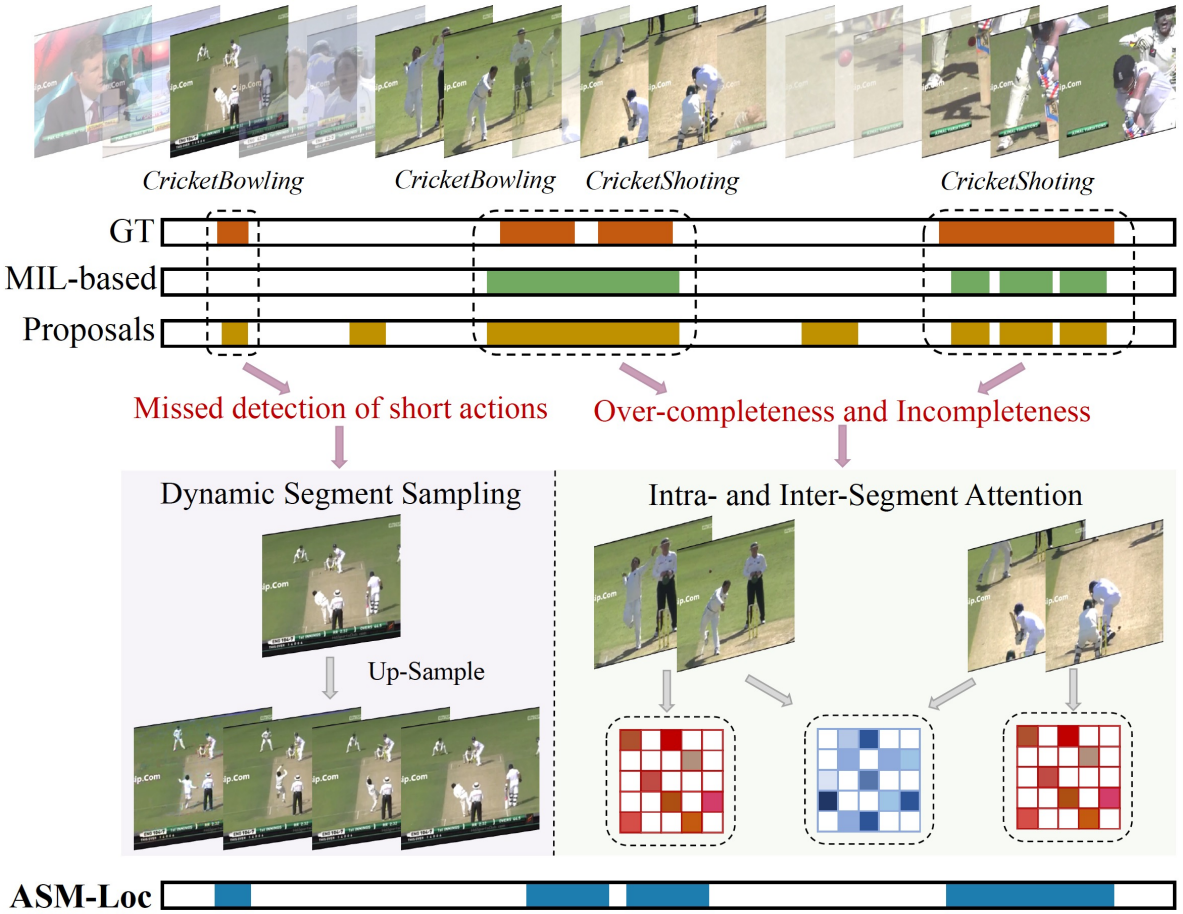




Future Works

1) Refine candidate generation
2) Weakly supervised localization:
3) Include patient data
4) Harmonize pipeline:





Acknowledgements

Kolmogorov Lab – CDSL:
Ataberk Domnez
Anton Goretsky
Ayse Keskus
Mikhail Kolmogorov
Rachel Parsons

Contact:
bds062@umd.edu

UNIVERSITY OF MARYLAND 18 56 DEPARTMENT OF COMPUTER SCIENCE

NIH NATIONAL CANCER INSTITUTE

A Deep Learning Approach to Identifying Complex Structural Variation in the Genome: Extra Slides

July 22nd, 2026

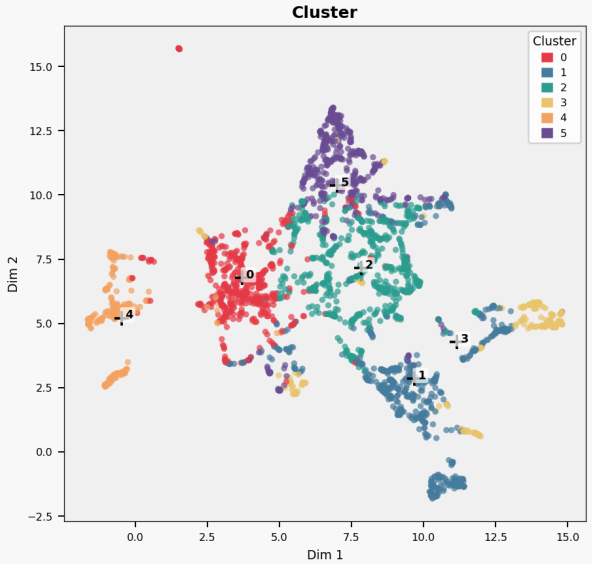
Bhargav Srinivasan | bds062@umd.edu

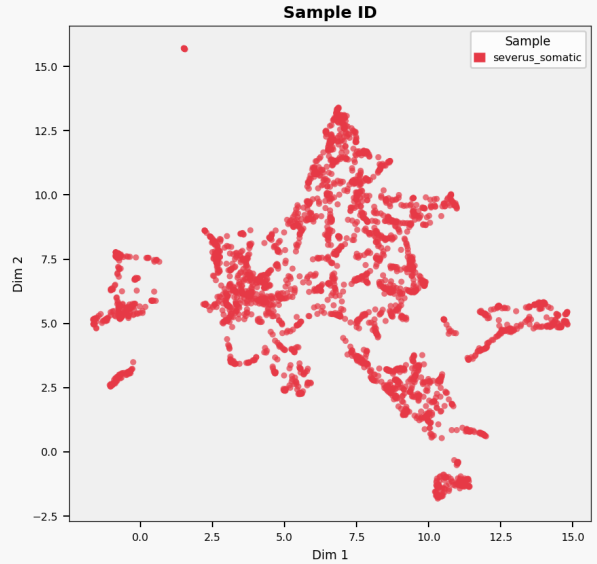
Computer Science + Biological Sciences Undergraduate Student, UMD

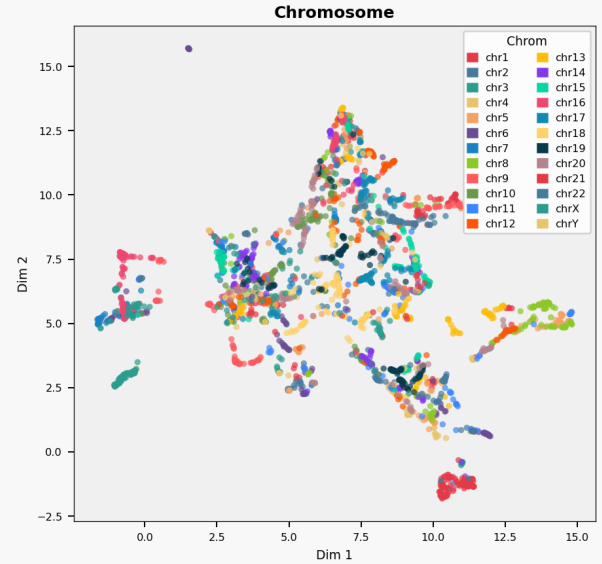
Summer Student, Kolmogorov Lab, National Cancer Institute

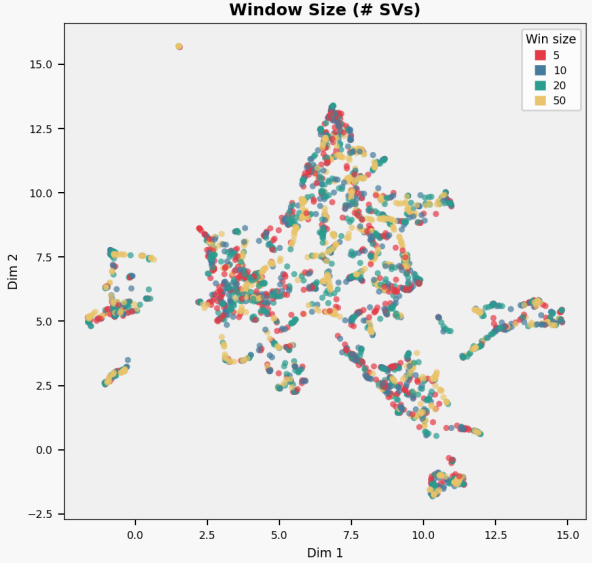
DEPARTMENT OF COMPUTER SCIENCE

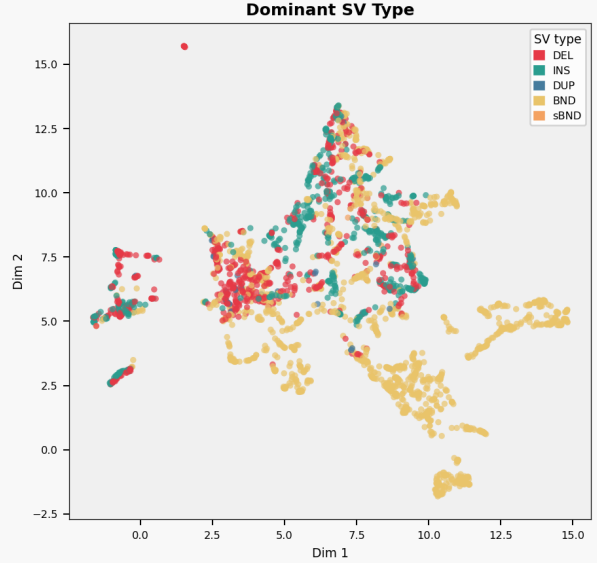
SV Embeddings — UMAP

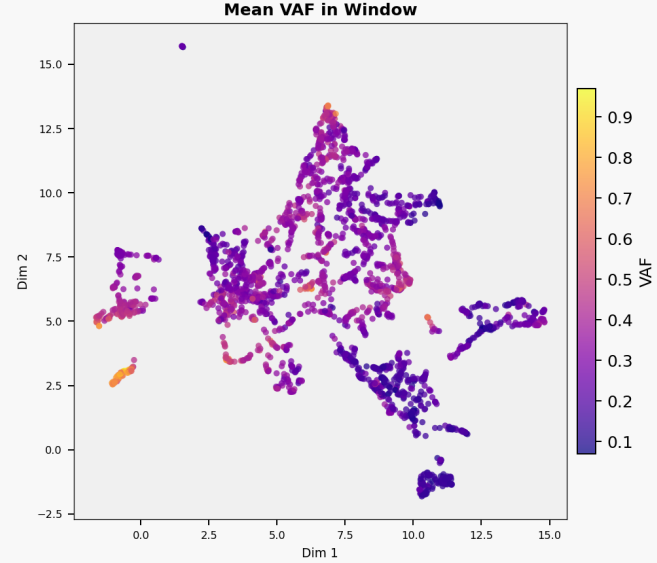




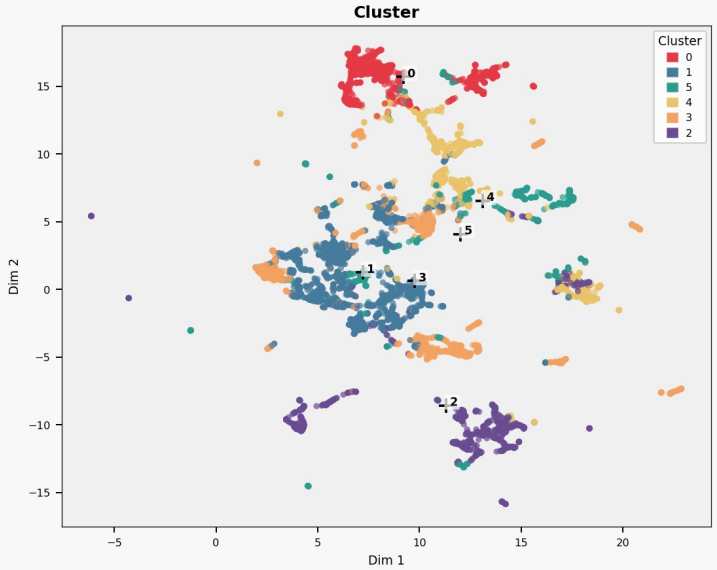


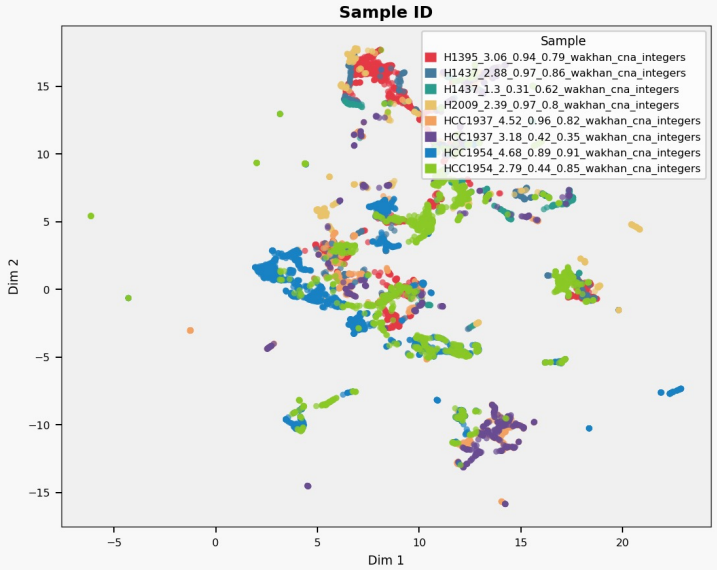


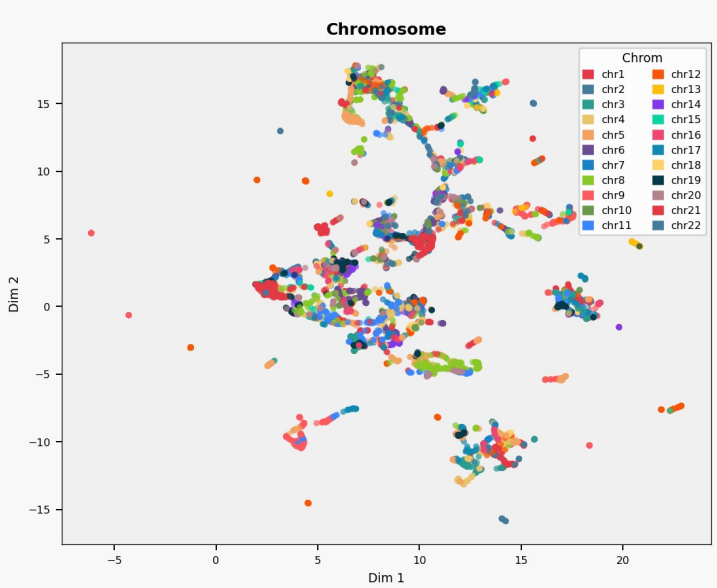


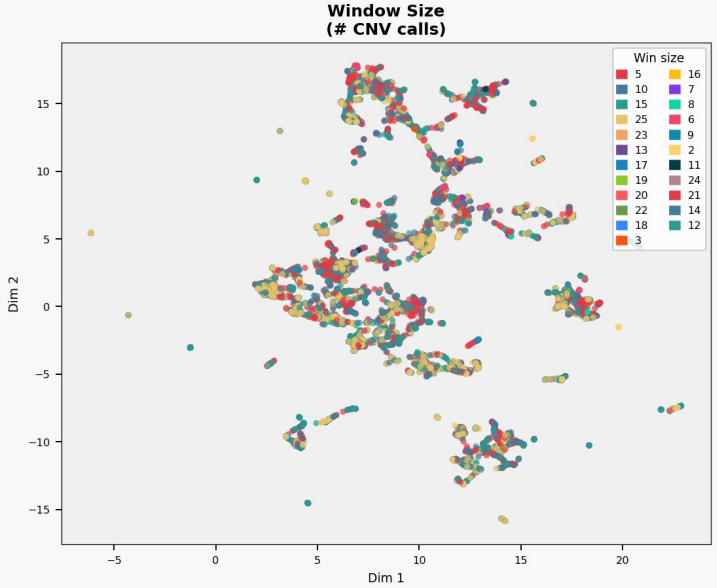


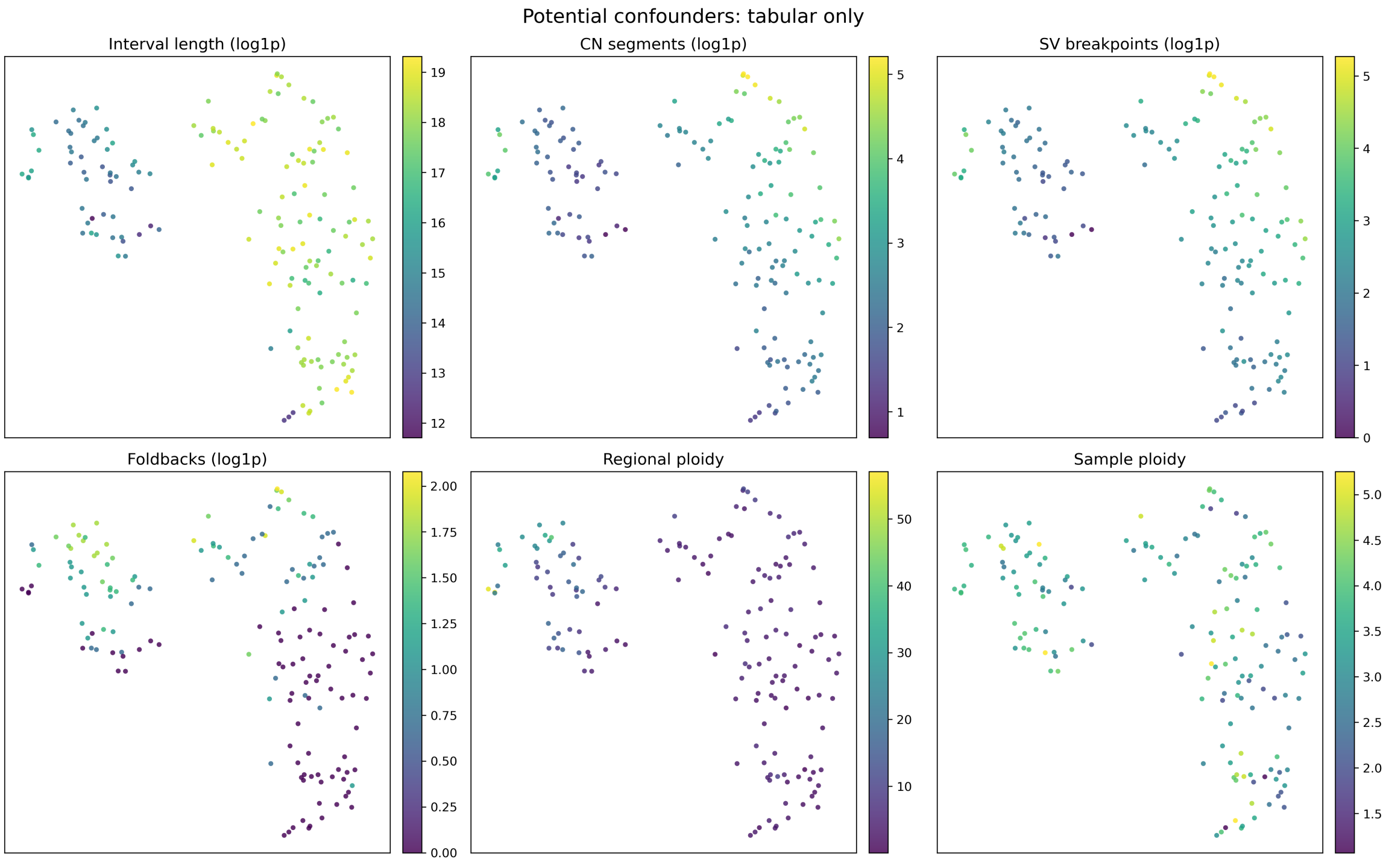
CNV Embeddings — UMAP

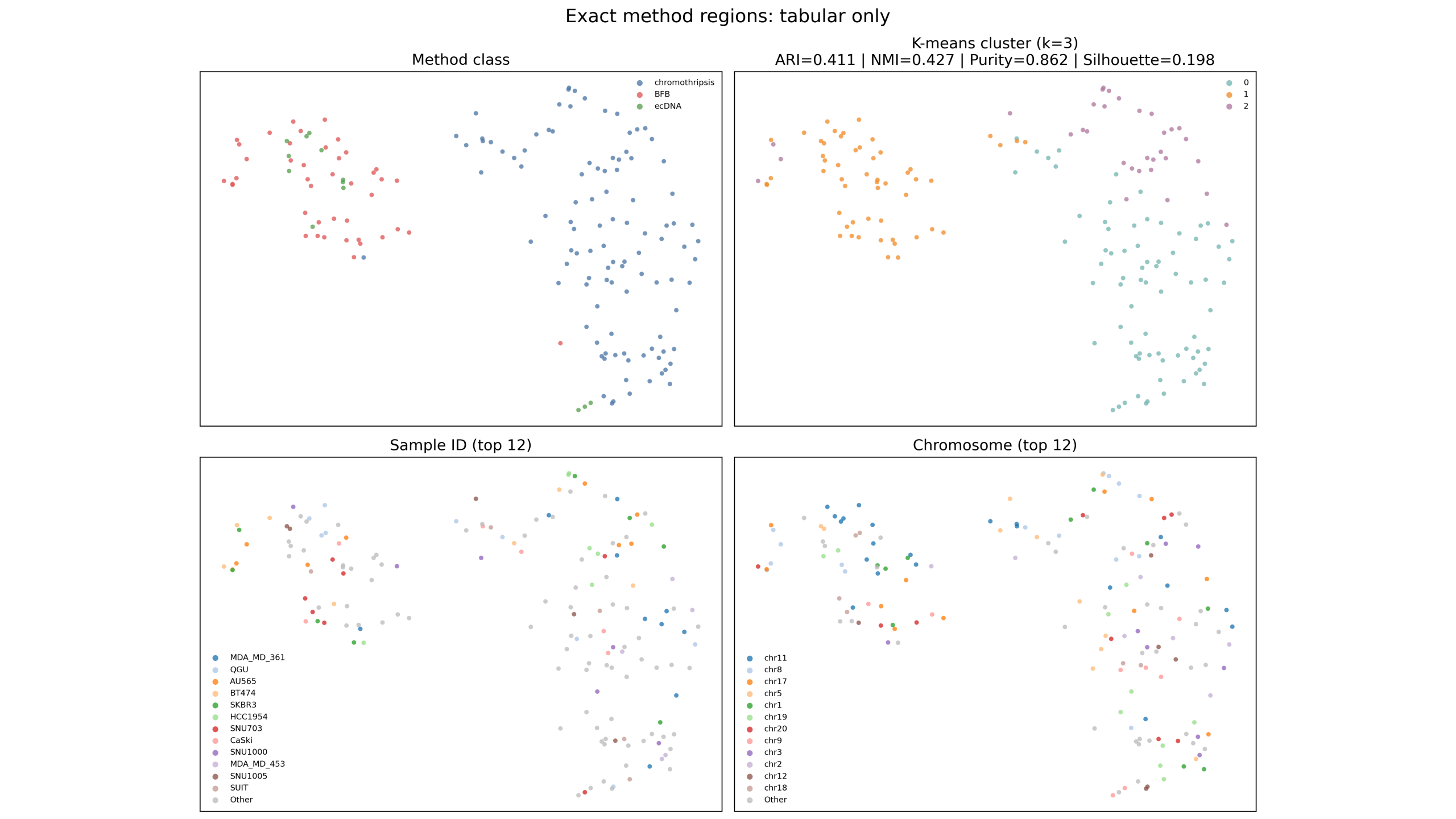


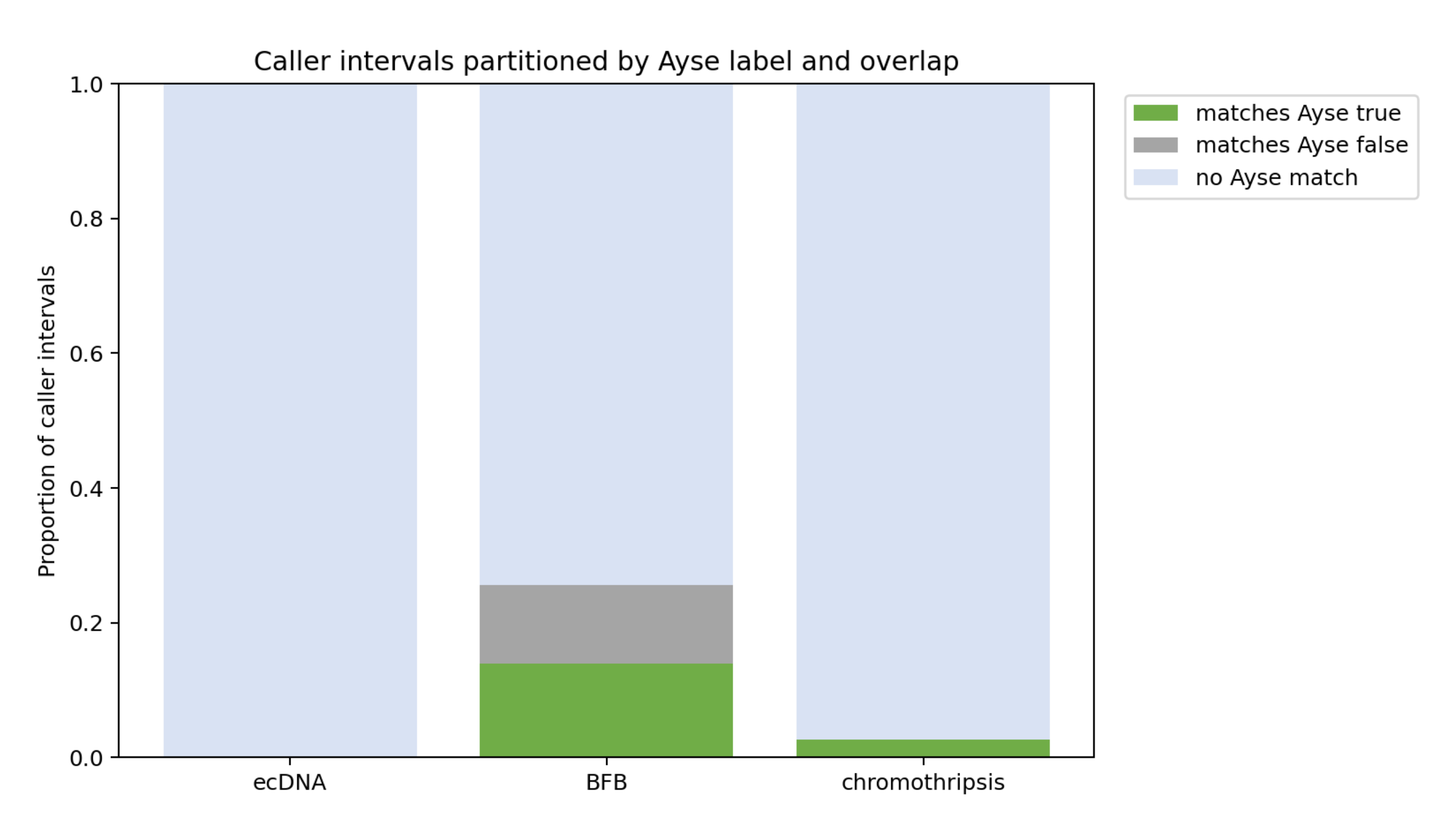







Potential confounders: tabular only
Interval length (log1p)
19
18
17
16
15
14
13
12
CN segments (log1p)
5
4
3
2
1
SV breakpoints (log1p)
5
4
3
2
1
0
Foldbacks (log1p)
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Regional ploidy
50
40
30
20
10
Sample ploidy
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5


Exact method regions: tabular only
Method class
chromothripsis
BFB
ecDNA
K-means cluster (k=3)
ARI=0.411 | NMI=0.427 | Purity=0.862 | Silhouette=0.198
0
1
2
Sample ID (top 12)
MDA_MD_361
QGU
AU565
BT474
SKBR3
HCC1954
SNU703
CaSki
SNU1000
MDA_MD_453
SNU1005
SUIT
Other
Chromosome (top 12)
chr11
chr8
chr17
chr5
chr1
chr19
chr20
chr9
chr3
chr2
chr12
chr18
Other


Caller intervals partitioned by Ayse label and overlap
Proportion of caller intervals
1.0
0.8
0.6
0.4
0.2
0.0
ecDNA
BFB
chromothripsis
matches Ayse true
matches Ayse false
no Ayse match