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

Bhargav Srinivasan[1], Ayse Keskus[1], Mikhail Kolmogorov[1]
[1]Cancer Data Science Laboratory, Center for Cancer Research, National Cancer Institute, NIH, Bethesda, MD, USA

Contact:
bds062@umd.edu

1. Goal

Leverage deep learning to annotate localized structural variation in the cancer genome.

2. Motivation

- Complex Structural Variation (SV) events may **drive oncogenesis** but are missed by most event-based SV callers
- Current methods are **heuristic based**, **limited** to specific complex SV types, or **computationally expensive**
- Many complex SV events have **arbitrary definitions** from different tools

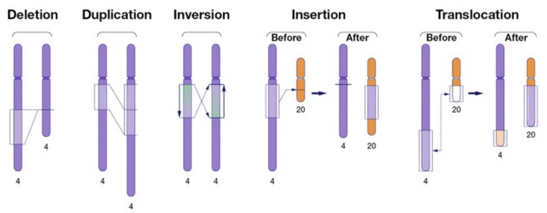

Figure 1. Structural variations lead to copy number alterations in the genome.

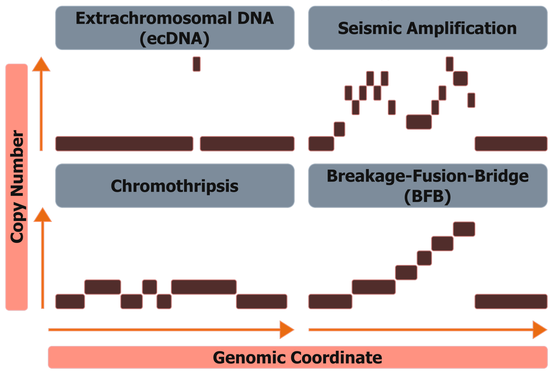

Figure 2. Complex SV events can be loosely identified from a copy number signature.

4. Our Solution

- **Deep Learning Pipeline** – localization and classification of complex SV regions
- **Autoencoder Pretraining** – CNA and SV self-supervised autoencoder pretraining
- **Multilabel Architecture** – lightweight model appended to autoencoders with object-ness and classification heads

5. Preprocessing and Label Generation

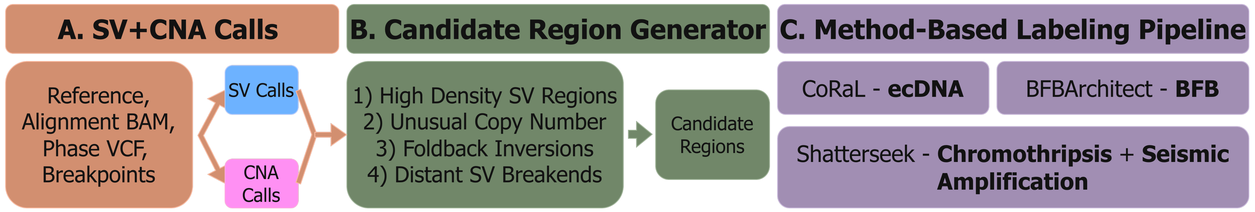

Figure 3. We employ a set of heuristics to identify candidate complex regions given an input genome

(A) An input genome must have phased CNA and SV calls, our pipeline uses Wakhan and Severus
(B) We use biological heuristics to generate candidate complex SV regions to be run through our model
(C) Method-based calls with >50% overlap to candidate regions are given labels for supervised training

6. Featurization of Copy Number and Structural Variation Calls

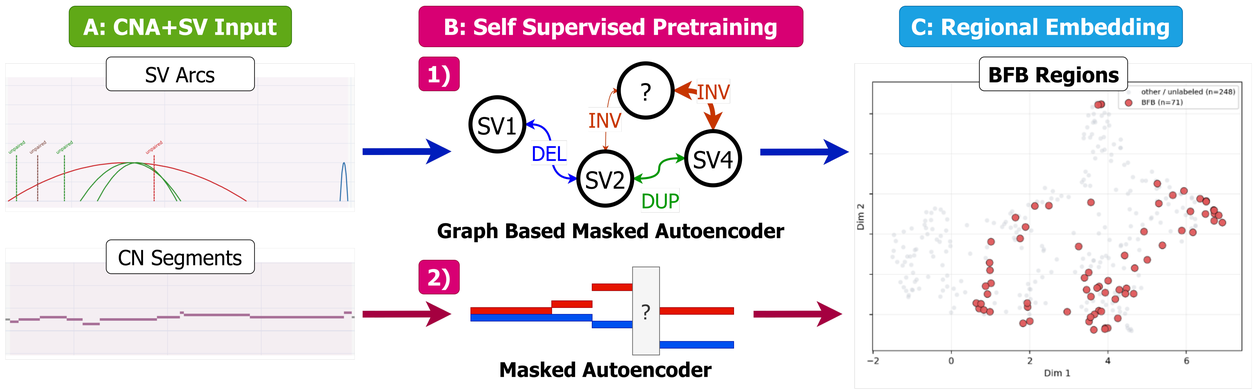

Figure 4. Our method includes two pretrained autoencoders to generate **CNA and SV aware features**.

(A) CNA and SV calls are parsed from candidate regions along with supporting summary statistics
(B1) We pretrain a **graph autoencoder** to learn SV based phasing, breakpoint, and regional relations.
(B2) We pretrain a **masked autoencoder** to learn CNA based read depth and copy number relations.
(C) We concatenate embeddings from (B1) and (B2) to create **CNA and SV aware local embeddings**.

7. Model Training

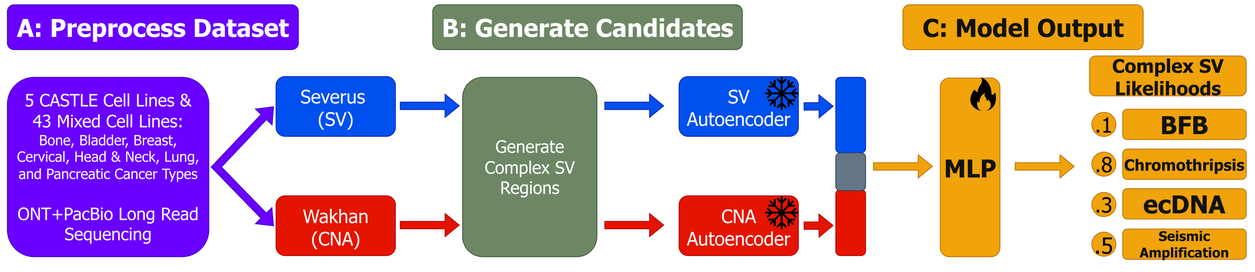

Figure 5. Our method uses a **machine learning approach** to classify complex SVs.

(A) We generate SV and CNA calls for 48 cell lines spread to identify and cluster candidate regions
(B) Candidate regions are input into **pretrained autoencoders** (Fig. 1) for context-aware featurization
(C) Train a **single hidden layer MLP** (54k parameters) to classify complex SVs from candidate regions

8. Results

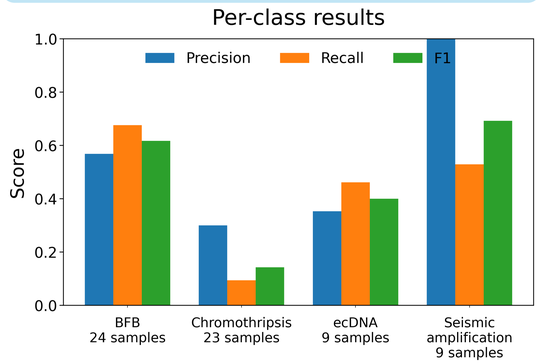

Figure 6. We observe variable results across classes with a **Macro-F1 of 0.46**.

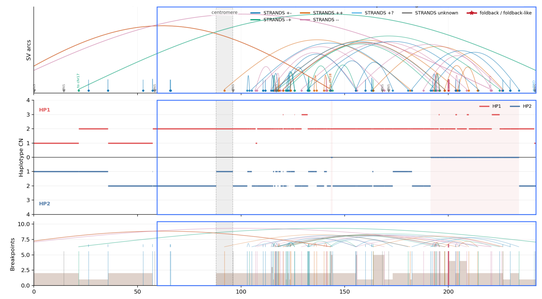

Figure 7. QGU2 is predicted to be chromothripsis but labeled none. Many False Positives still show complex events.

9. Unsupervised Discovery

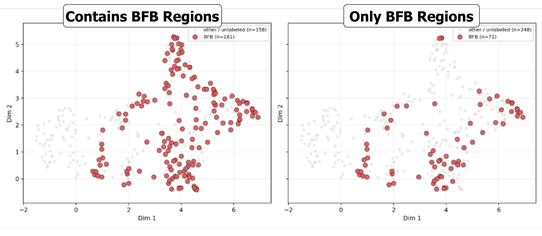

Figure 8. Unsupervised learning can cluster potentially new complex SV events.

10. Future Work

- **Include patient data** and more cell line data in training dataset
- Extend featurization approach to **Short Read** SV and CNA callers
- Postprocess results for **phasing labels and further localization**
- Refine autoencoder objective functions