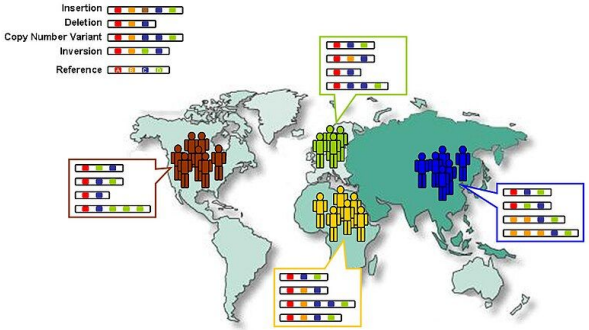
Curriculum representation learning with gene fusion evidence in cancer and healthy samples



Jacob Krol, 12/12/24
CSCI-7000-006, neurosymbolic-NLP

Motivation

- Rare cancers tragically are hard to diagnose and treat

How is chondromyxoid fibroma treated?

Treatment for each patient will be unique. Treatment options to discuss with your doctor include:

Surgery:

Once CMF is diagnosed, you may have surgery to remove the part of the bone with the tumor.

Cryotherapy:

Along with surgery, liquid nitrogen may be used to kill the tumor cells.

Phenol:

After surgery, phenol is a chemical which can be used to kill the remaining tumor cells.

Does chondromyxoid fibroma run in families?

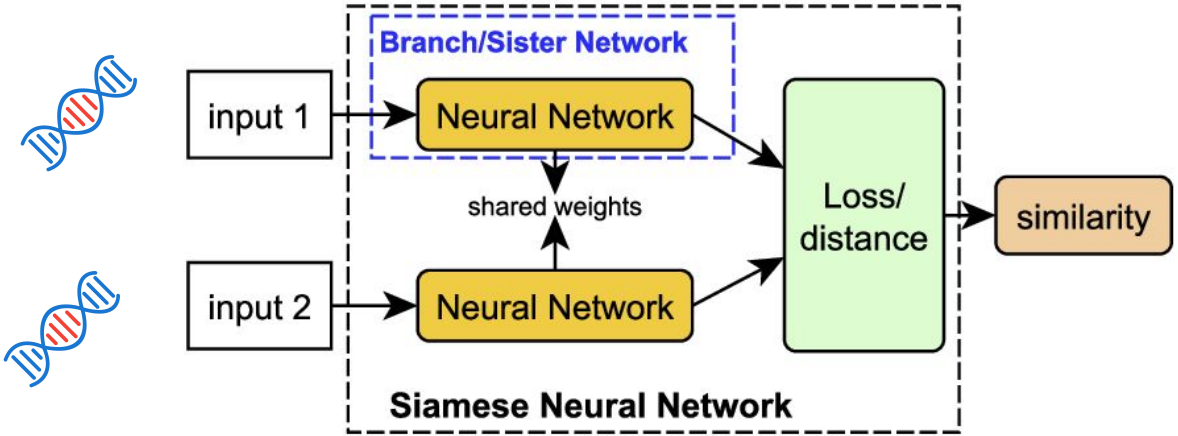
We do not know if it runs in families because there are so few people with CMF.

NIH NATIONAL CANCER INSTITUTE
MyPART

Motivation

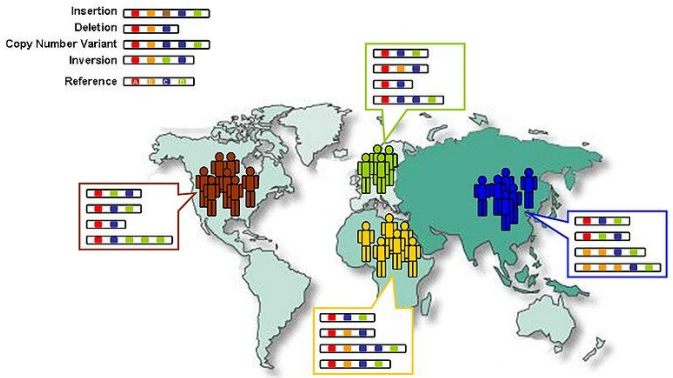
Siamese neural networks learn representations from paired examples

Q: Can low dimensional representations of human genomes support genetics-driven similarity searches to find similar patients?



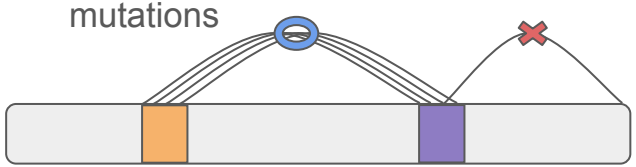
Data

1000 Genomes
(Healthy samples)



Pan cancer analysis whole genomes
(Cancer samples)

Gene2gene
mutations



Gene left

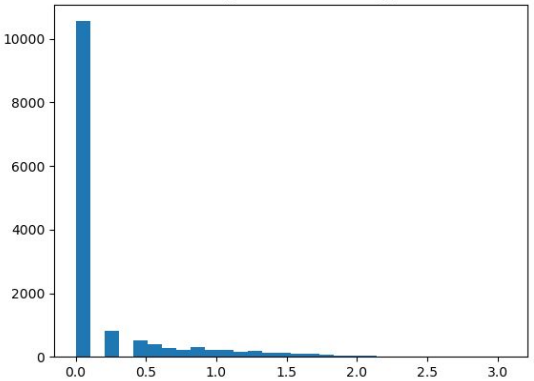
Gene right

A **364486500** length vector for each person

Dimensionality reduction

- Encoding is sparse and heavy-tailed

Gene2gene evidence counts (feature values)

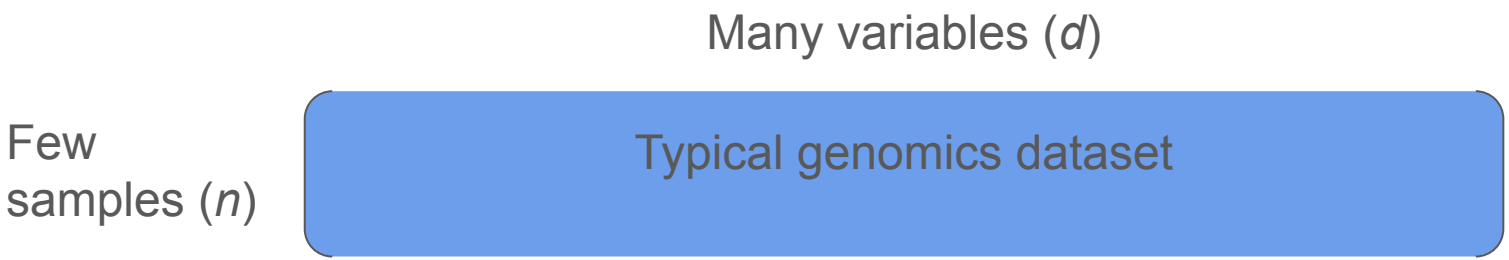


Log10(genefusion evidence+1)

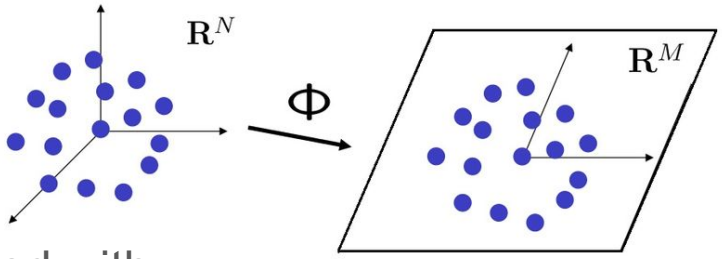
Johnson-lindenstrauss lemma

$$m = O(\frac{\log n}{\epsilon^2})$$

- JL is ideal for small, high dimensional datasets
  - Common in genomics
- Unfortunately, not an online model



Johnson-Lindenstrauss lemma (baseline)



- Guarantees pairwise distances are preserved with fractional error

n := number of points
d := original dimensionality
m := reduced dimensionality
ϵ := fractional error

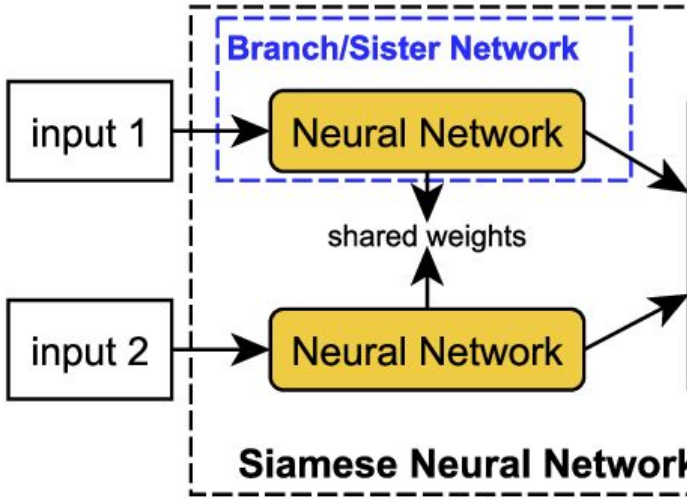
m Does not depend on *d*

$$m = O(\frac{\log n}{\epsilon^2})$$

$$(1-\epsilon)||x_i - x_j||^2 \leq ||f(x_i) - f(x_j)||^2 \leq (1+\epsilon)||x_i - x_j||^2$$

Siamese network with symbolic components

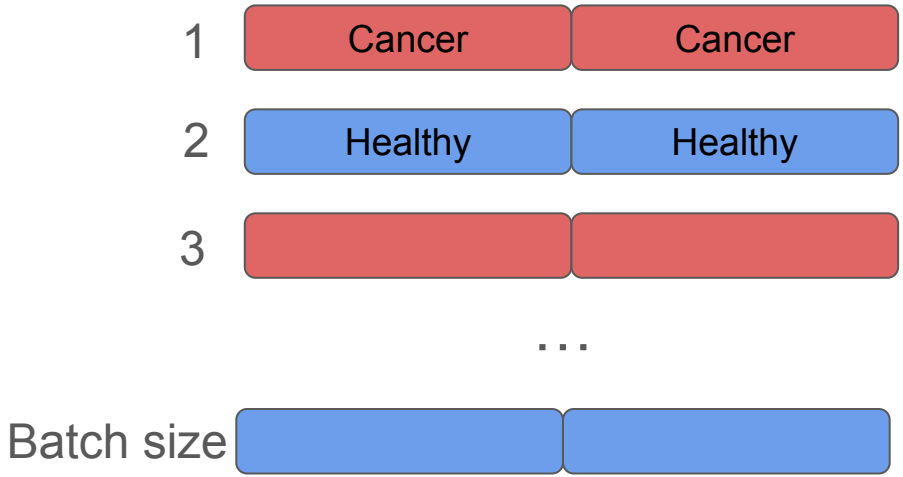
- Rule-based batch curation



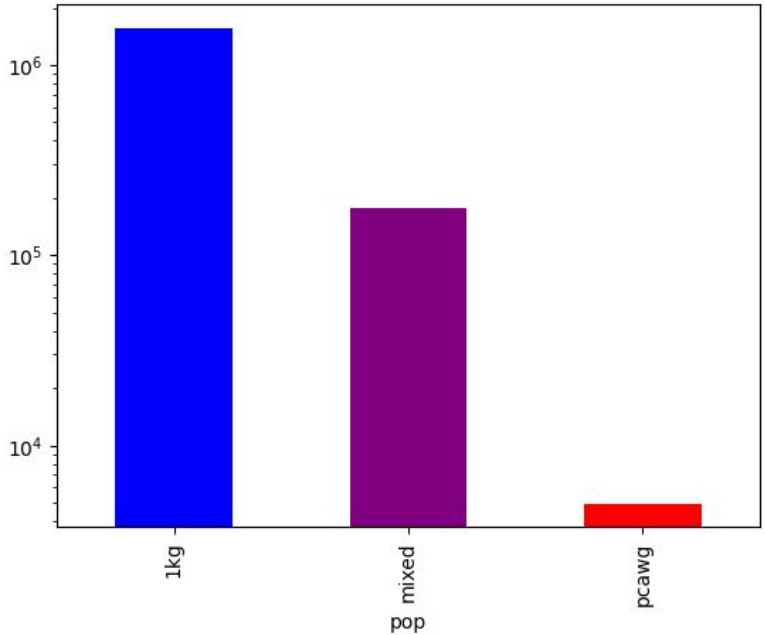
- Custom loss
- Curriculum learning

Symbolic component:
Curriculum 1 & population-aware batch construction

Curriculum 1: construct batch by alternating within population pairs



Count of population pairs is very imbalanced



$$L(x_i, x_j, x_i', x_j') = (||x_i' - x_j'||_2^2 - ||x_i - x_j||_2^2)$$
$$L_{batch} = MSE_L$$

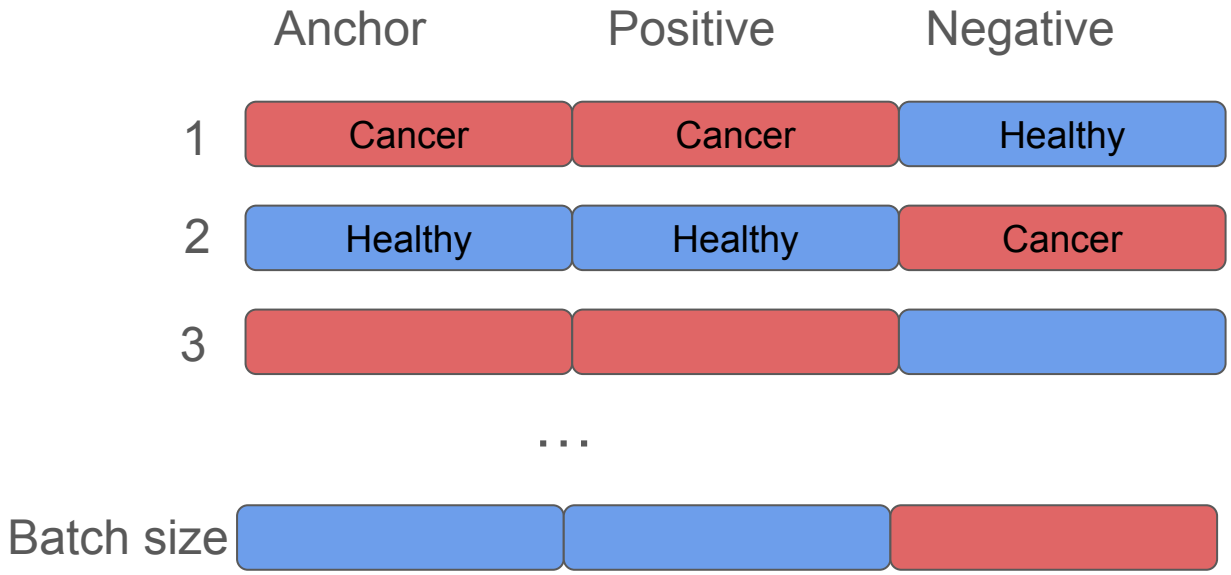
Curriculum 1: learn within population distances

$$L(x_i, x_j, x'_i, x'_j) = (||x'_i - x'_j||_2^2 - ||x_i - x_j||_2^2)$$
$$L_{batch} = MSE_L$$

Embedding distance

True distance

Symbolic component:
Curriculum 2 & contrast populations with triplet loss



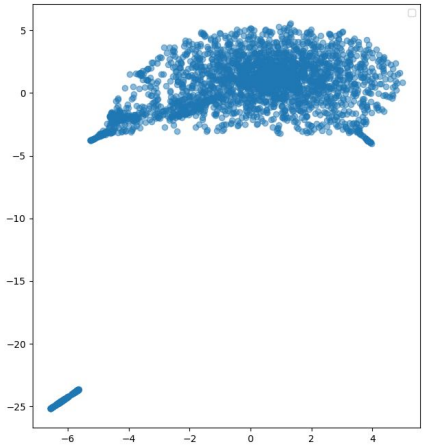
$$\mathcal{L}\left(A, P, N\right) = \max(\| \mathrm{f}(A) - \mathrm{f}(P) \|_2 - \| \mathrm{f}(A) - \mathrm{f}(N) \|_2 + \alpha, 0)$$

Model	**RMSE**	**Embedding size**
Siamese-C1	8.39	512
Siamese-C1+C2	7.3	512
Siamese-C2+C1	7.3	512
JL ($\epsilon = 0.1$)	4.95	2104
Siamese-C2	**4.88**	512

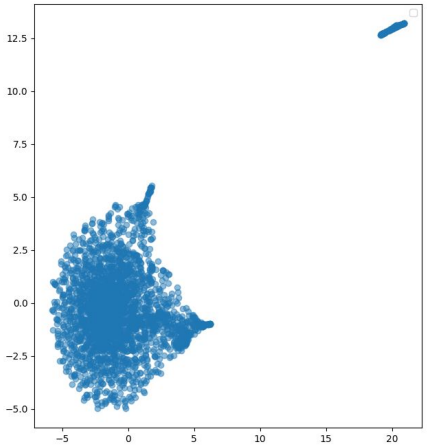
Table 1: Model performance and embedding size. C1 and C2 are curricula 1 and 2.

Qualitative results do not match RMSE

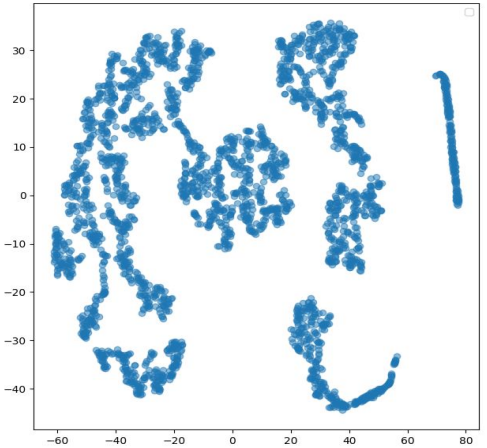
Untransformed



JL



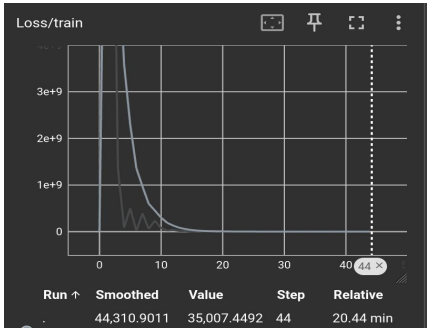
Siamese-C2

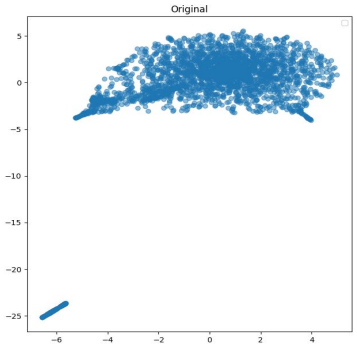


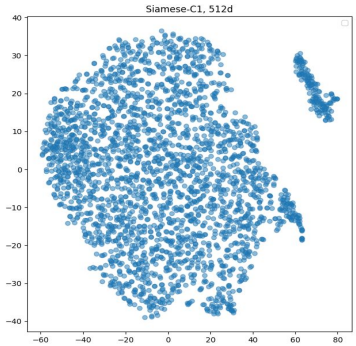
JL looks great

Training and qualitative analysis

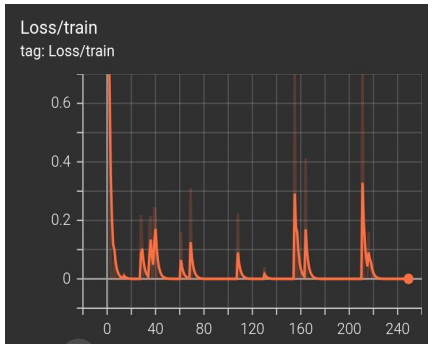
Curriculum 1

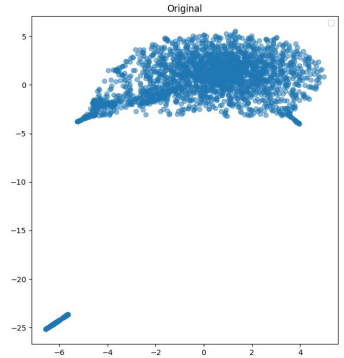


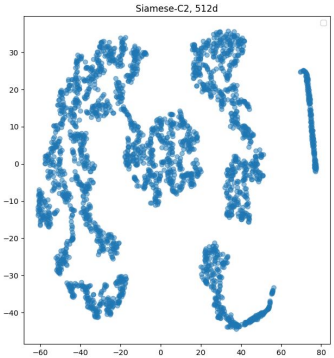




Curriculum 2

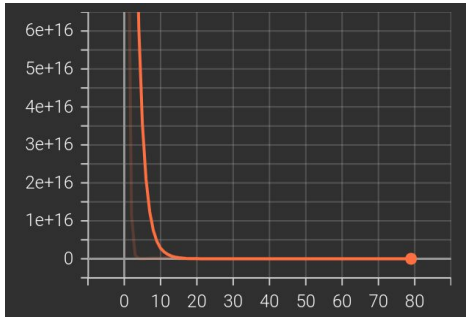


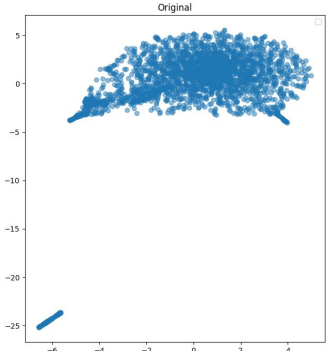


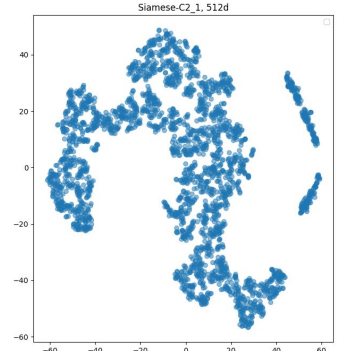


Training and qualitative analysis

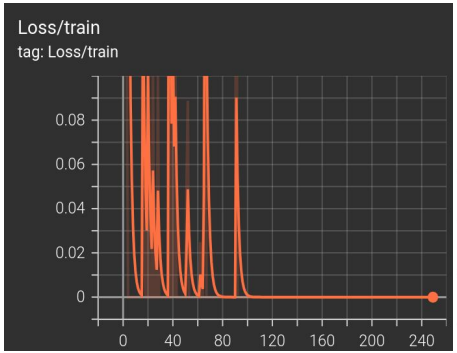
Curriculum 2-1

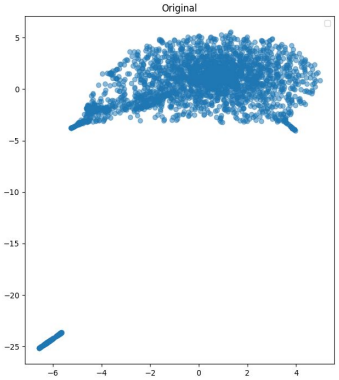


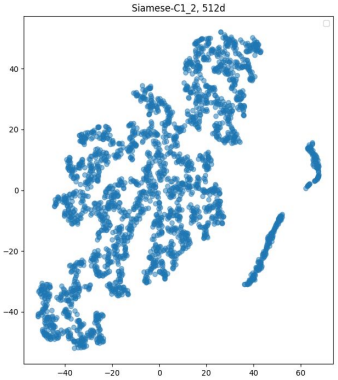




Curriculum 1-2







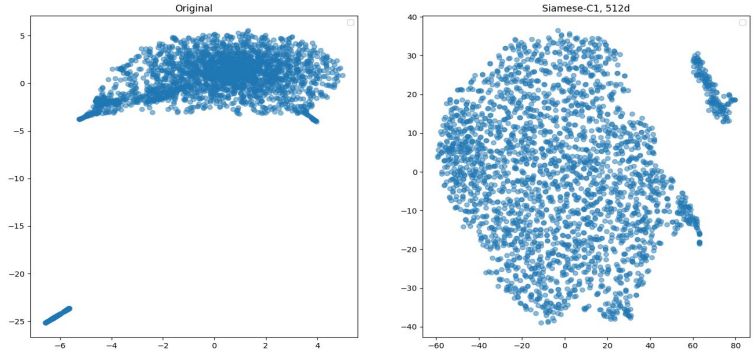
Conclusions

Model	RMSE	Embedding size
Siamese-C1	8.39	512
Siamese-C1+C2	7.3	512
Siamese-C2+C1	7.3	512
JL ($\epsilon = 0.1$)	4.95	2104
Siamese-C2	**4.88**	512

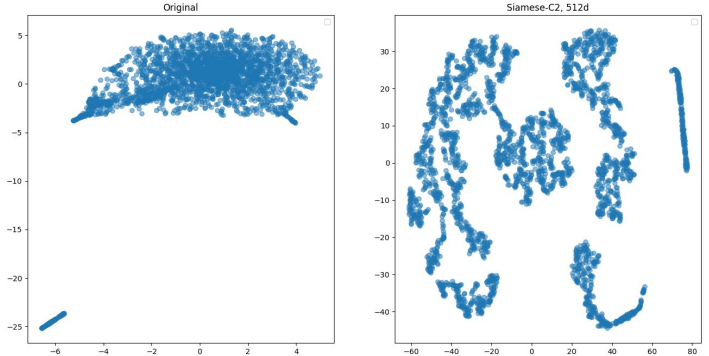
Table 1: Model performance and embedding size. C1 and C2 are curricula 1 and 2.

- Yes, low dimensional representations of sample-wise genefusions can be created with JL or learned by a siamese neural network
- JL is effective, yet impractical
- Siamese neural network learns population distances (triplet loss, curriculum 2)

Curriculum 1



Curriculum 2



Future directions

- Add other cancers to see if sub-cancer clusters appear
- Use normalized distance metric for ground truth (cosine)
- Siamese architecture tuning
  - Embedding size
  - Hidden layers
  - Activation functions
- Train and test with full data
  - Total number of sample pairs is 3643650
  - 70/30 train/test split
  - Further sampling the train and test splits for practicality
    - The subsamples for train and test were the same across methods
- Early stopping for Triplet objective
- Weight freezing for curricula
- Evaluation: compute loss separately for across and within pairs