Linear classifiers and t-SNE for understanding relationships across cancer types

Ali Yang

October 17th, 2021

Mentors: Alkis Gkotovos and Stefanie Jegelka

Background

In the US, cancer is the **second** leading cause of death.

How do we treat it?

Background

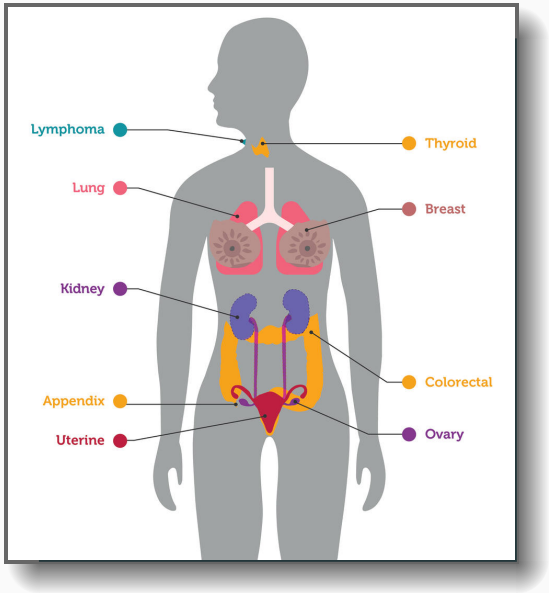
To treat cancers, we need to be able to classify them.

PD-1, a target for immunotherapies[1]

[1] Image source: Wikipedia

Background

Cancers were originally classified primarily based on the organ or cell they originated from, or growth patterns.



Some of the organs from which cancer can originate[2]

[2]Image source: National Cancer Institute

Background

Genetic sequencing has revealed many more cancer subtypes than previously thought.

CMS1 MSI immune	CMS2 Canonical	CMS3 Metabolic	CMS4 Mesenchymal
14%	37%	13%	23%
MSI, CIMP high, hypermutation	SCNA high	Mixed MSI status, SCNA low, CIMP low	SCNA high
BRAF mutations		*KRAS* mutations	
Immune infiltration and activation	WNT and MYC activation	Metabolic deregulation	Stromal infiltration, TGF-β activation, angiogenesis
Worse survival after relapse			Worse relapse-free and overall survival

The four consensus molecular subtypes of colorectal cancer[3]

[3]Image source: Guinney *et al.*, 2015

Cancer Type Prediction

How good are existing cancer types?

Cancer Type Prediction

The Cancer Genome Atlas (TCGA) Dataset

- 9051 patients with known types
- Data on mutation, amplification, and deletion for 763 genes

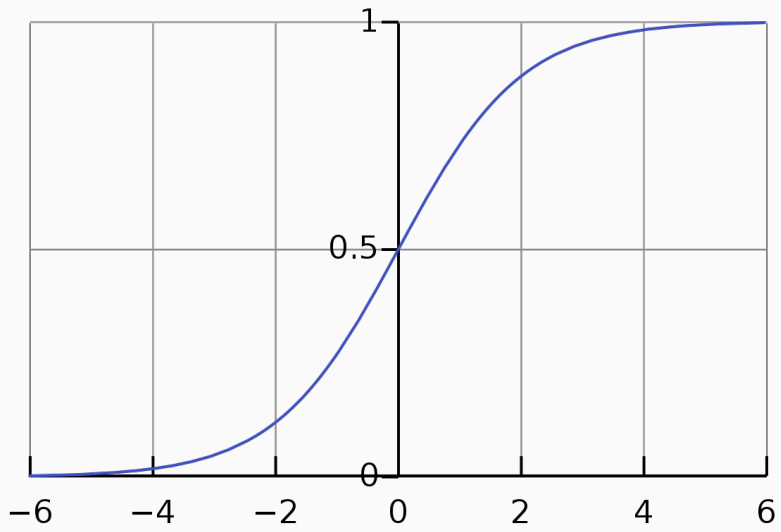
Cancer Type Prediction

- Train a model to predict type from genetics.
- If it has good accuracy, then the types are genetically meaningful.

Cancer Type Prediction

Logistic classifier:

$$f(\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{w}\cdot\boldsymbol{x}}}.$$

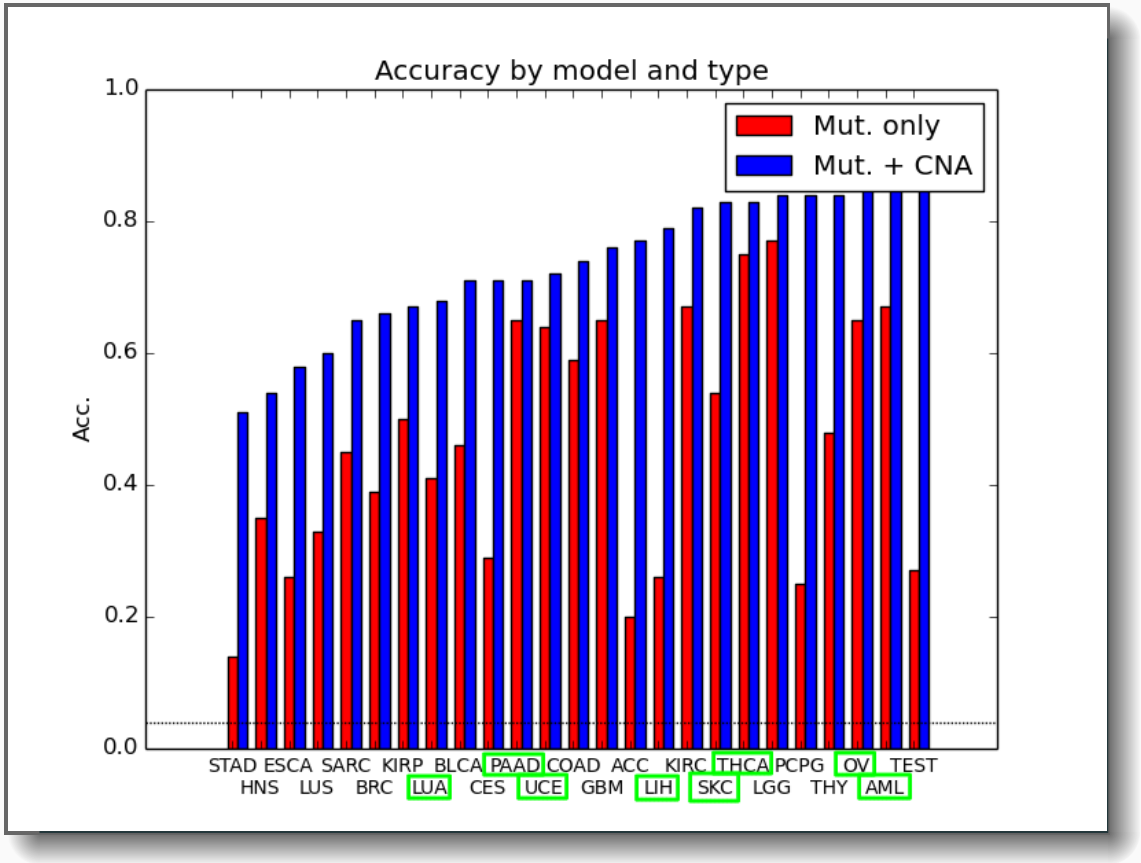


Shape of the logistic classifier's classification boundary[4]

[4]Image source: Wikipedia

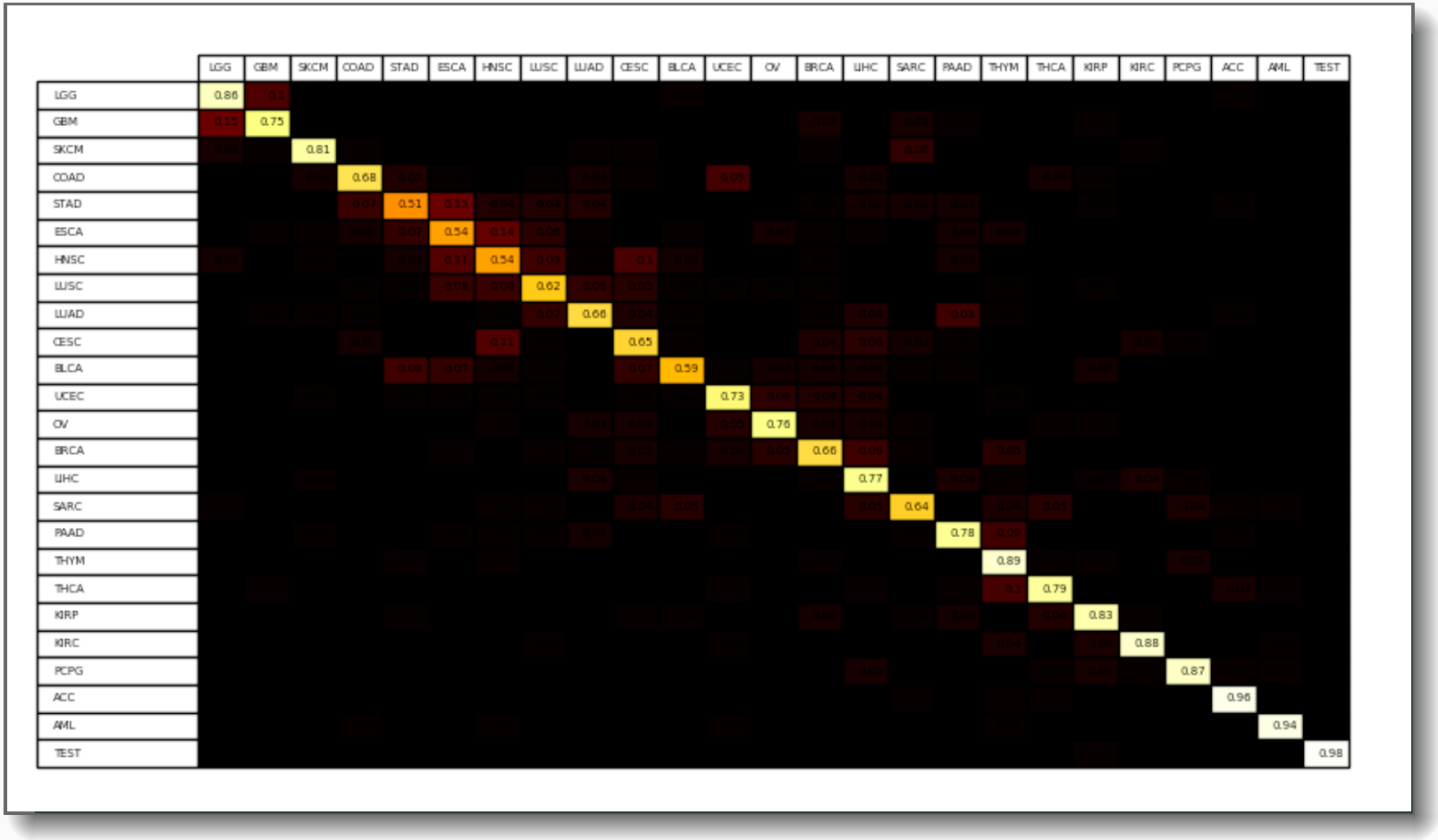
Cancer Type Prediction

Overall accuracy: **74.4%**



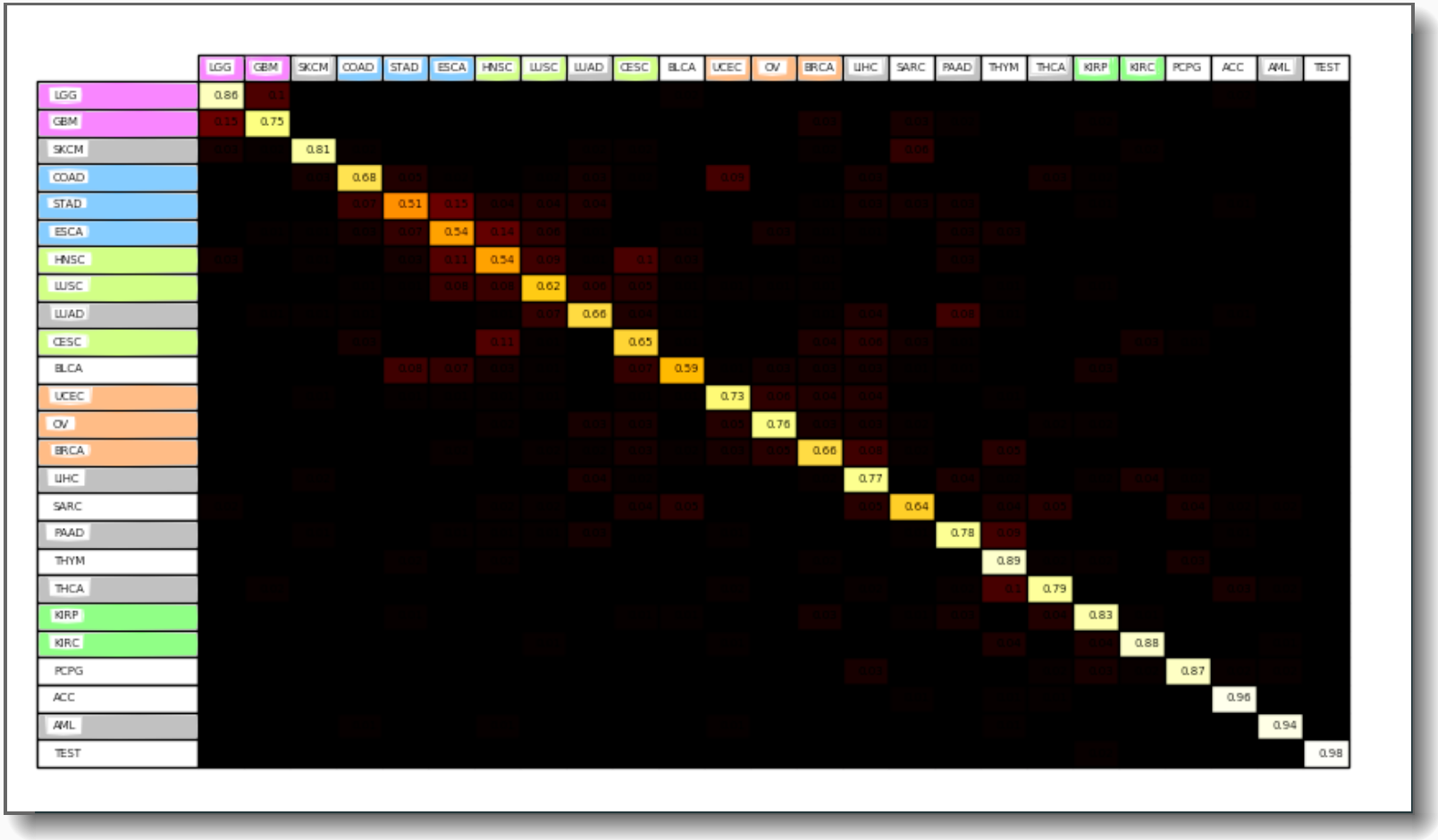
Logistic classifier accuracy on different cancer types

Cancer Type Prediction



Logistic classifier confusion matrix

Cancer Type Prediction

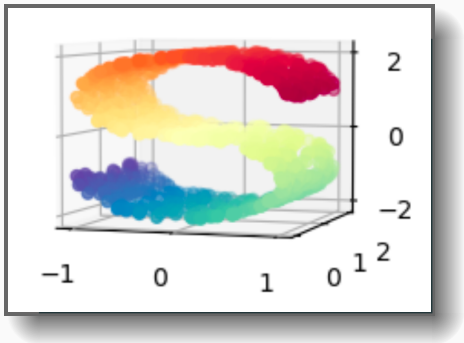


Logistic classifier confusion matrix—related organ groups colored in

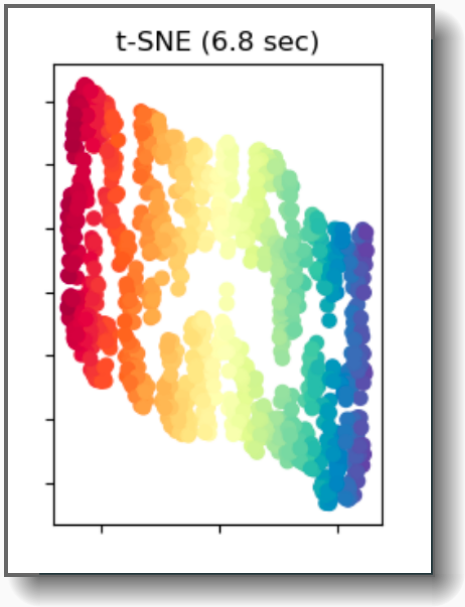
Classifying Cancers

What are other ways to classify cancers?

Dimensionality reduction with t-SNE

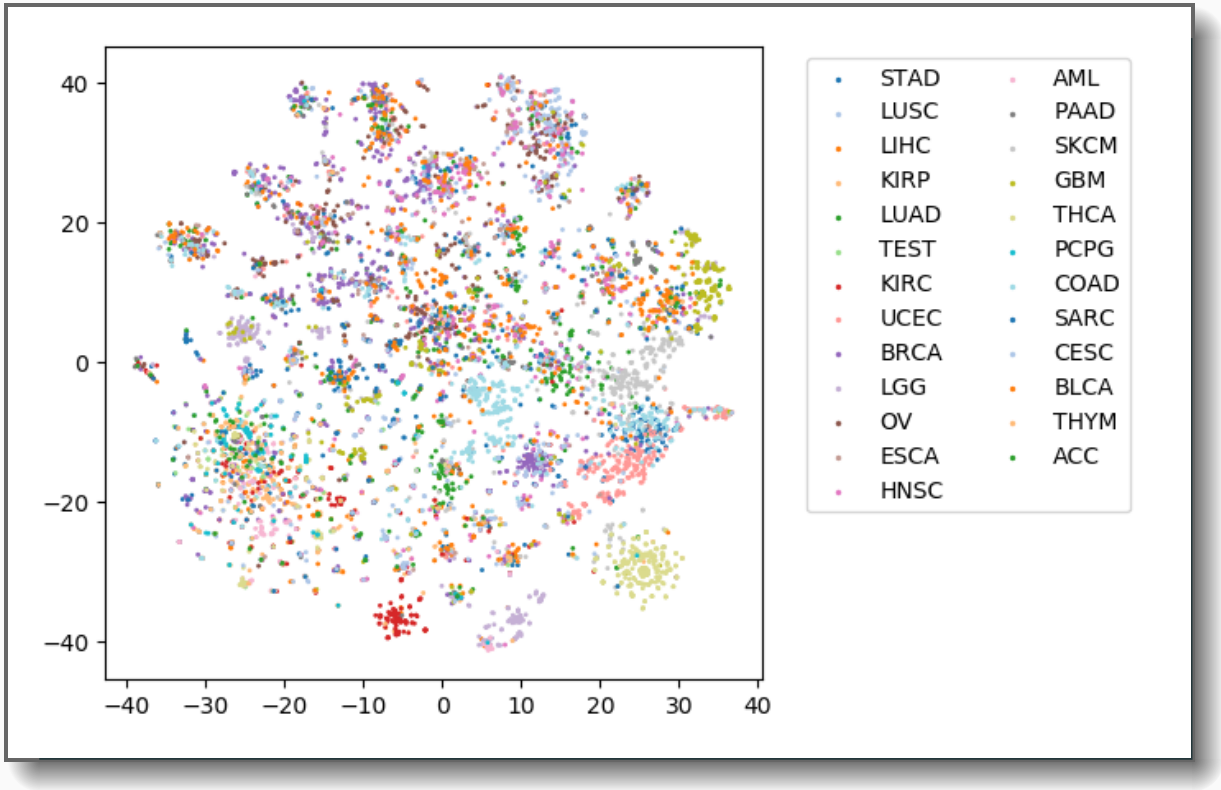


3D data



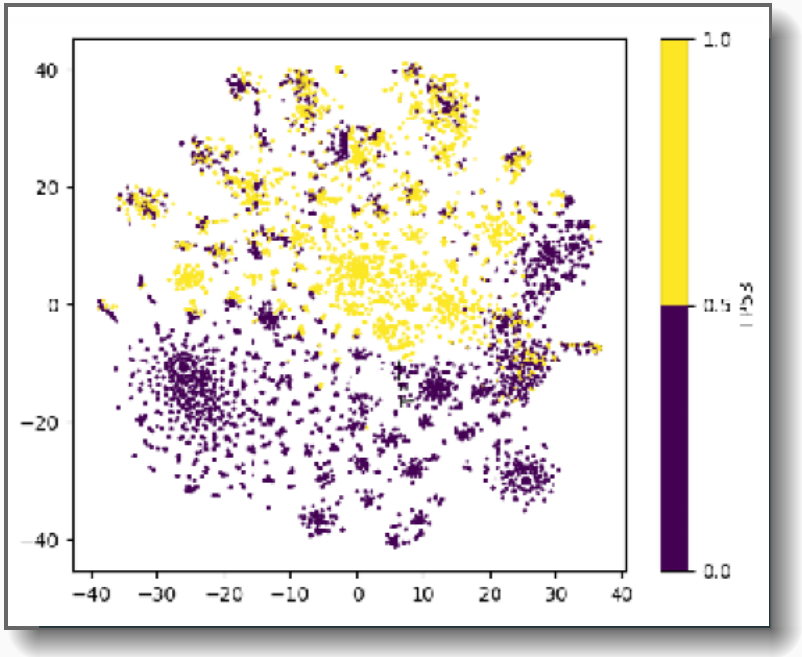
2D version with t-SNE

Classifying Cancers

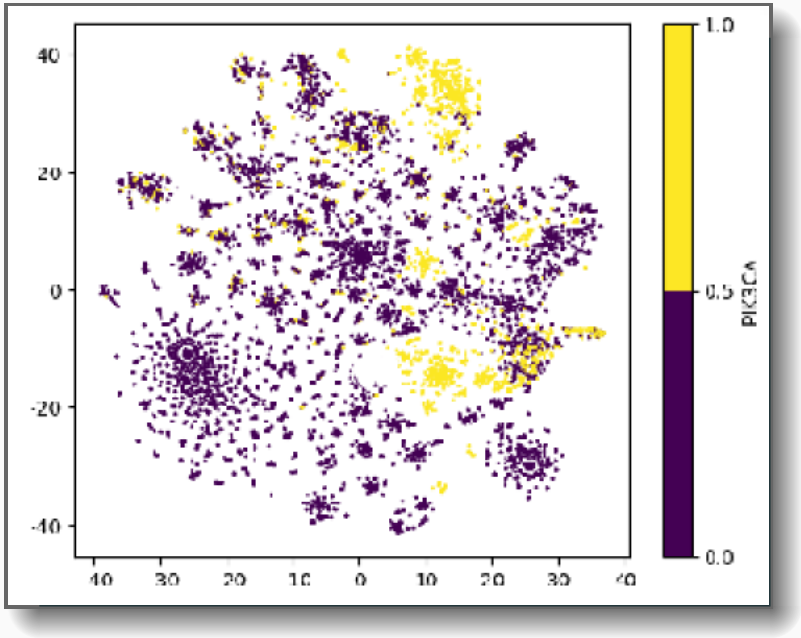


t-SNE of TCGA data, colored by type

Classifying Cancers

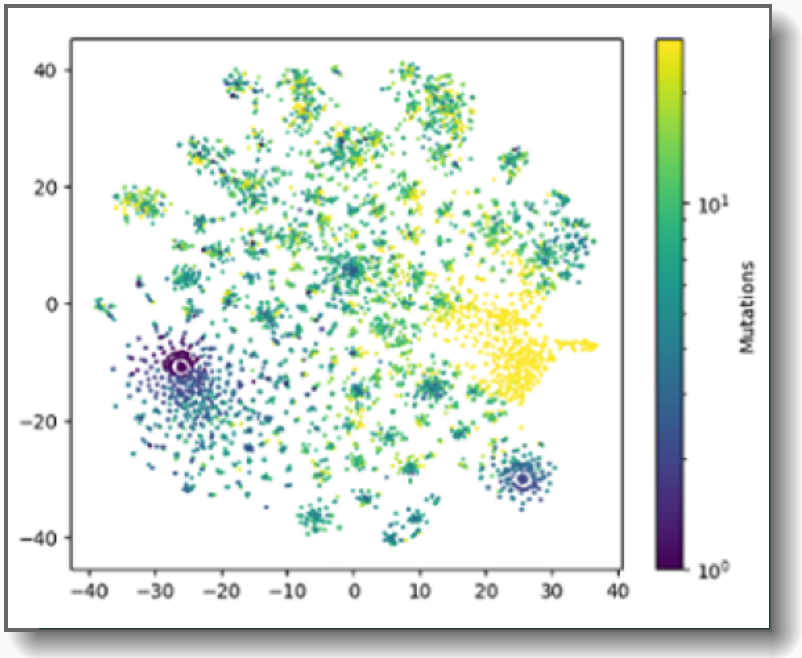


TP53

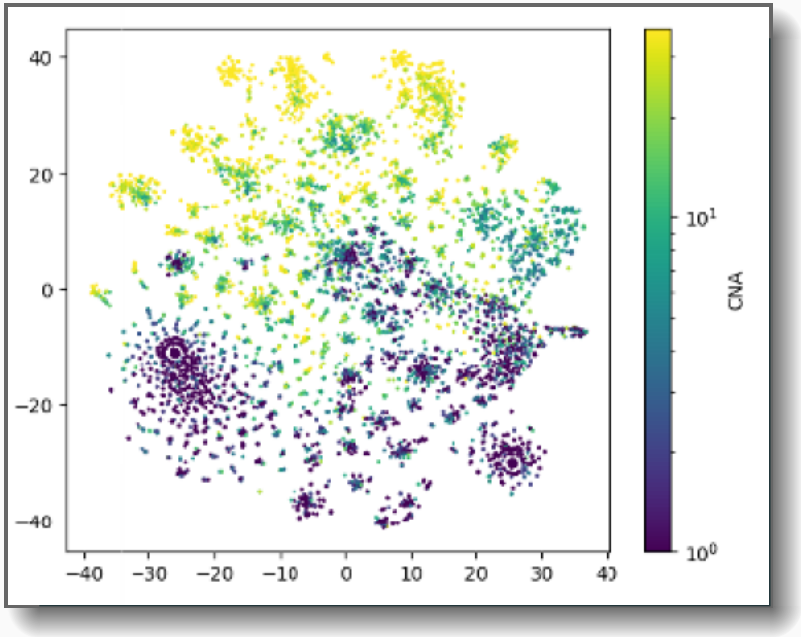


PIK3CA

Classifying Cancers

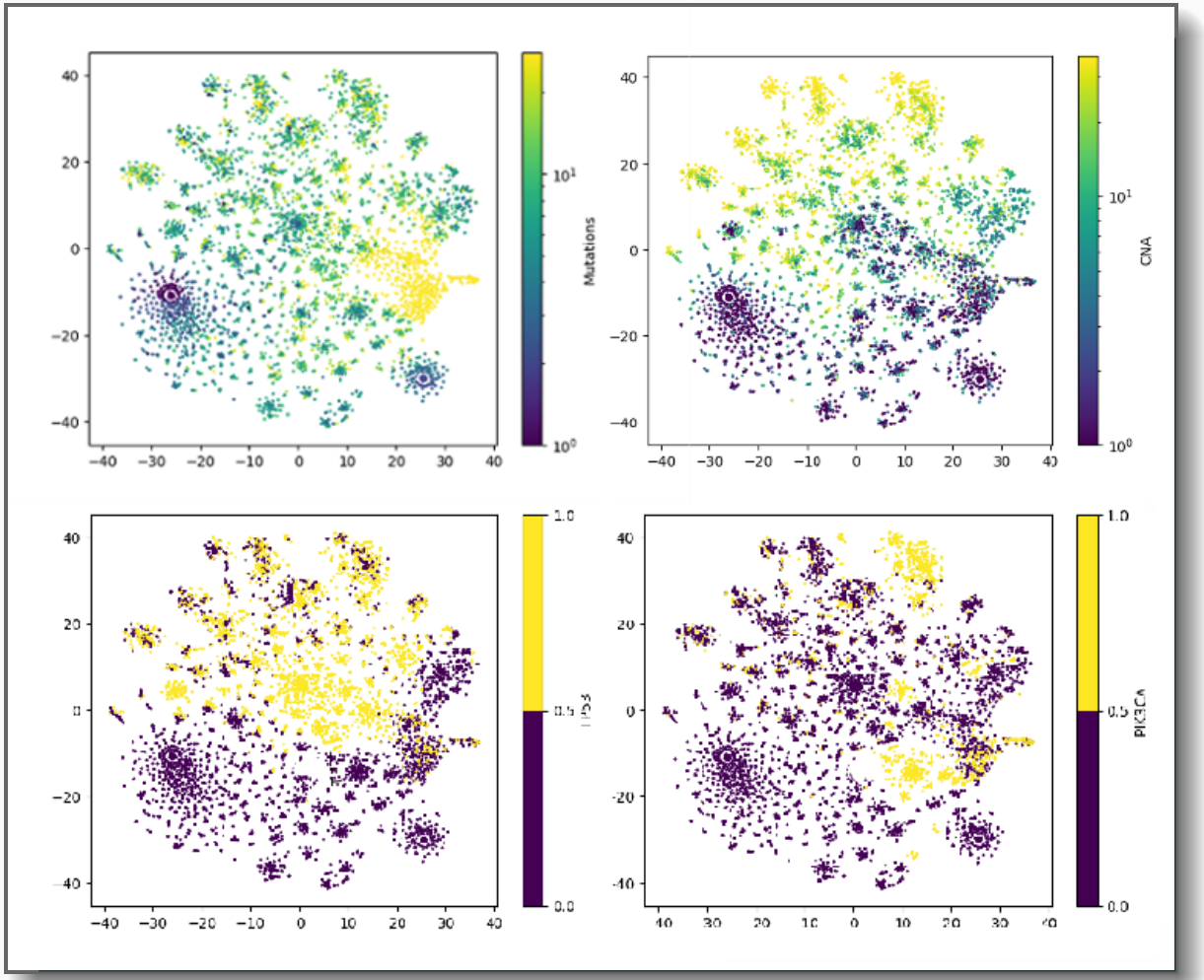


Mutations



Copy-number alterations

Classifying Cancers



t-SNE plot of TCGA data colored four different ways.

From top left clockwise: mutation number, copy-number alteration number, TP53 mutated, PIK3CA mutated

Conclusion

- Existing cancer types can be easily distinguished genetically.
- Cancers from related organs are similar.
- We can classify cancers by TP53 or PIK3CA mutations, number of mutations, and number of copy-number variations.

Acknowledgements

I would like to thank my mentors, Alkis Gkotovos and Professor Stefanie Jegelka, for their guidance throughout this project.

I would also like to thank Qinghong Yang for explanations of many of the biological concepts in this project.