Genome-Driven Personalized Medicine of Cancer via Machine Learning and Phylogenetic Models

Yifeng Tao

August 11, 2021

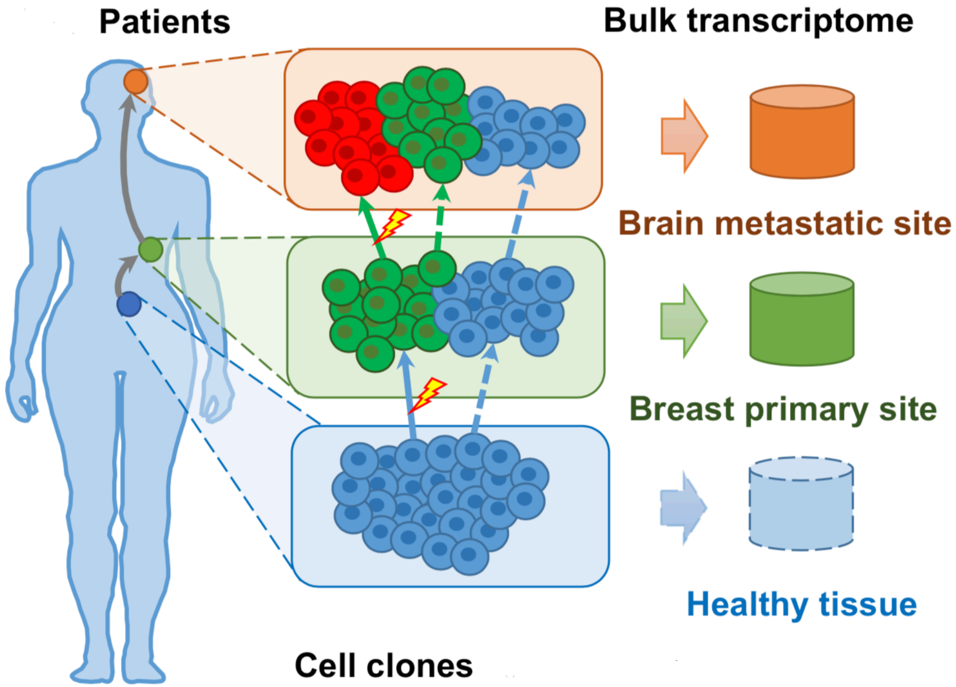
Computational Biology Department | Carnegie Mellon University School of Computer Science

Tumor heterogeneity and personalized medicine

- Cancer is a disease caused by aberrant mutations in genome.
- It develops via an evolutionary process into mixture of heterogeneous populations.

GOAL: To understand mechanism of tumor evolution and utilize genomic data for personalized medicine.

- Phenotype inference of cancer.
- Mechanism of tumor progression.
- Machine learning on evolutionary features.



OUTLINE

1

Reliable phenotype inference of cancer through well-designed interpretable machine learning models

2

Revealing intra-/inter-tumor heterogeneity and mechanism of tumor progression via robust deconvolution and phylogenetic algorithms

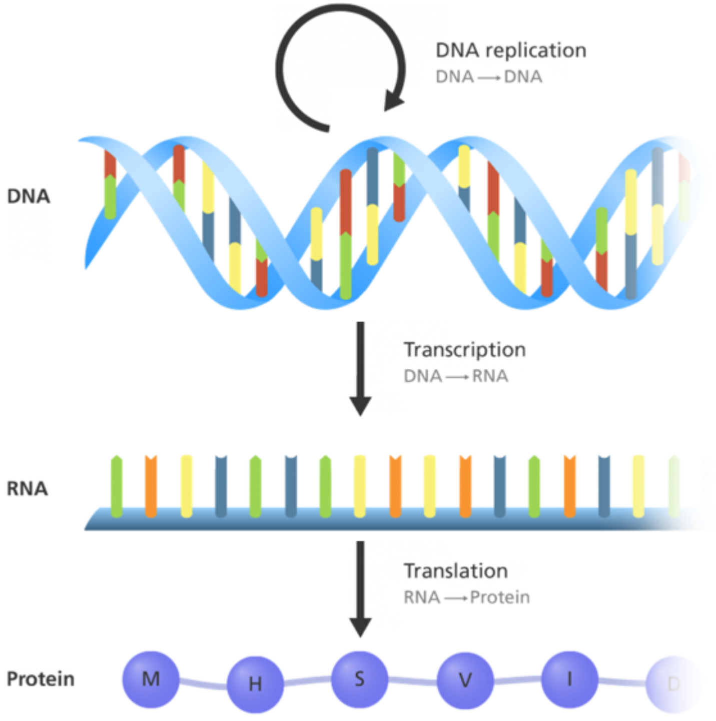
3

Improving prognostic prediction of cancer by incorporating machine learning and evolutionary methods

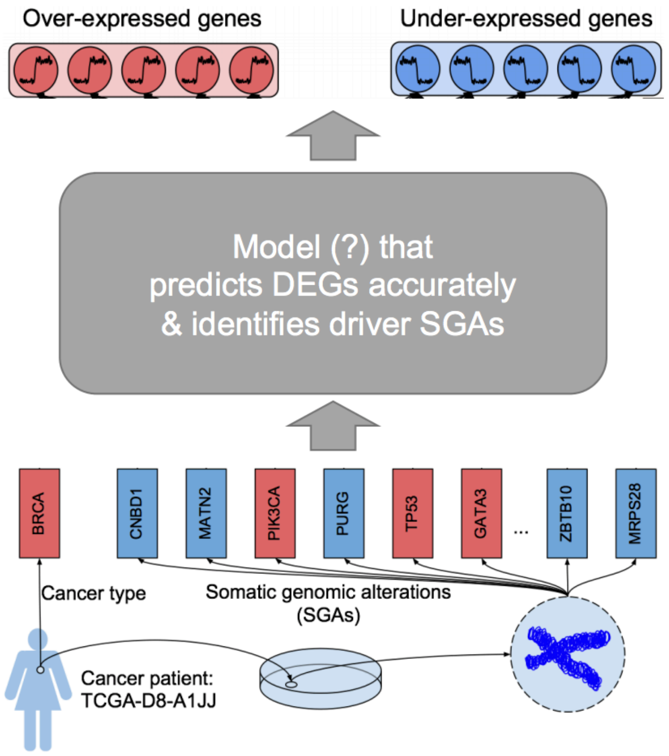
Inference of RNA expression from mutated genes

Central dogma: DNA → mRNA → protein.

- RNA is the bridge between mutations and downstream phenotypes.
- Identify driver mutations in cancer with the supervision of mRNA.



https://www.yourgenome.org/facts/what-is-the-central-dogma



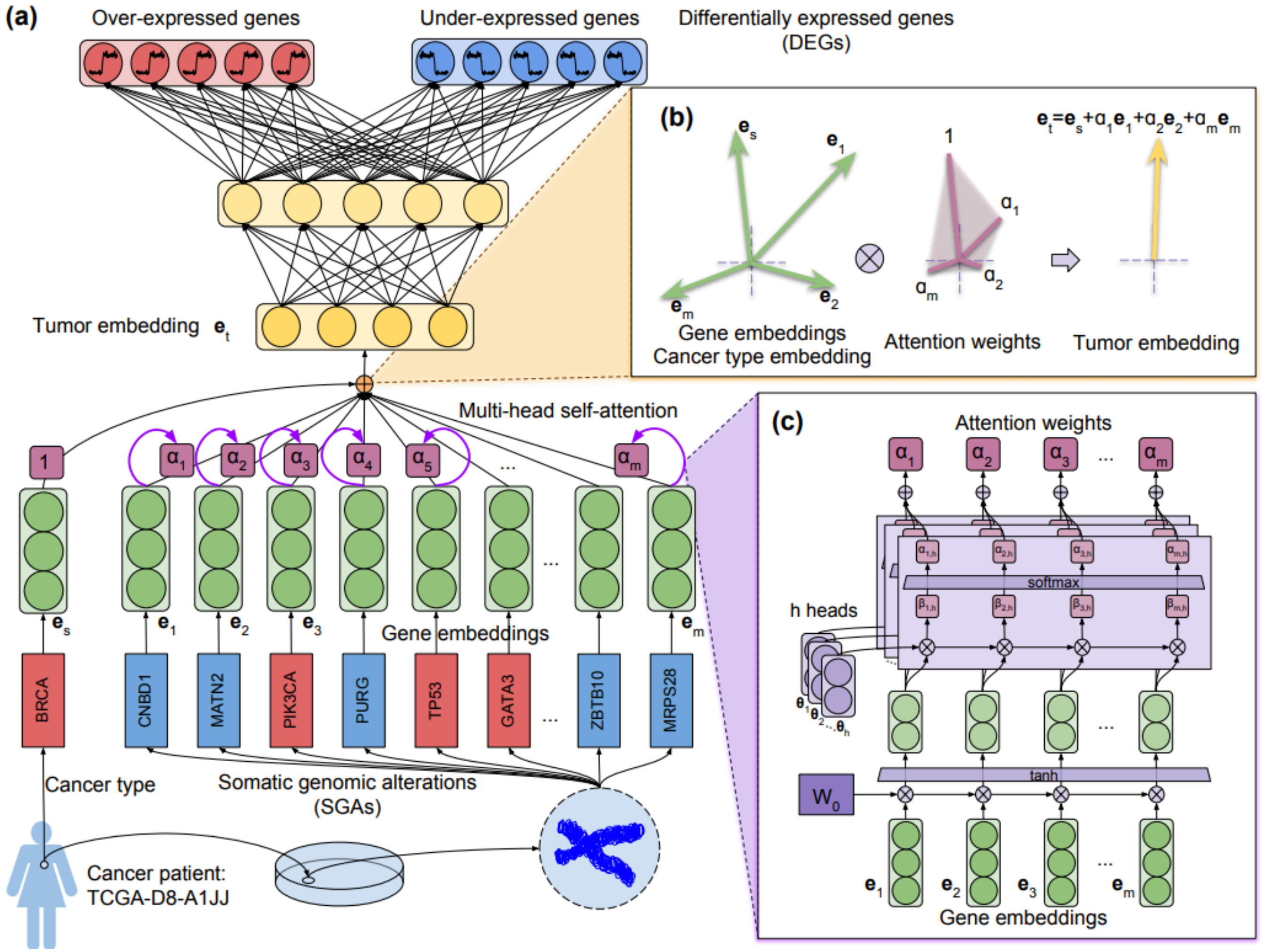
Inference of RNA expression from mutated genes

GIT: Genomic Impact Transformer

- Autoencoder architecture.
- Input: bag of mutated genes.
- Output: differentially expressed genes.

Self-attention: capture contextual impact of input mutated genes.

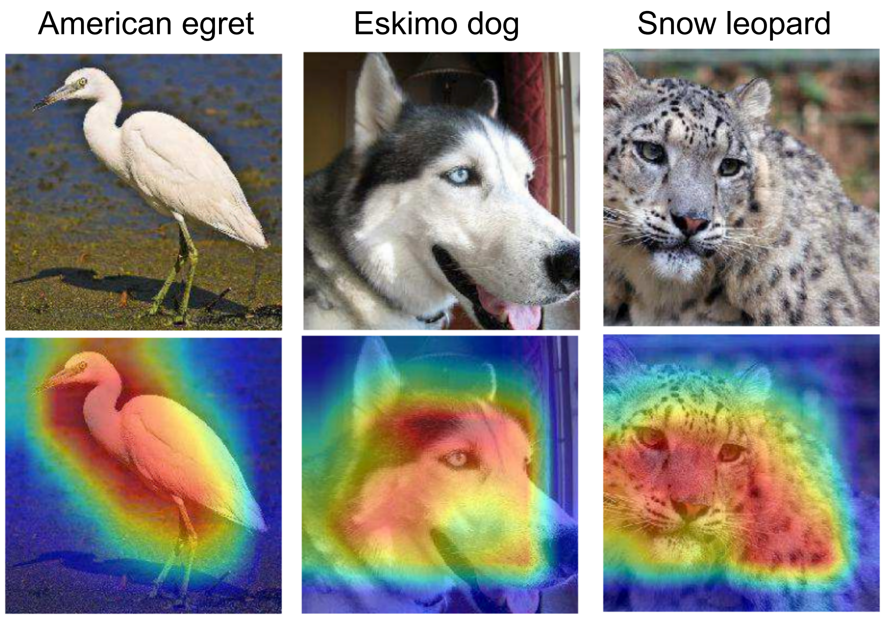
- Widely used in CV/NLP.
- Performance.
- Interpretability.



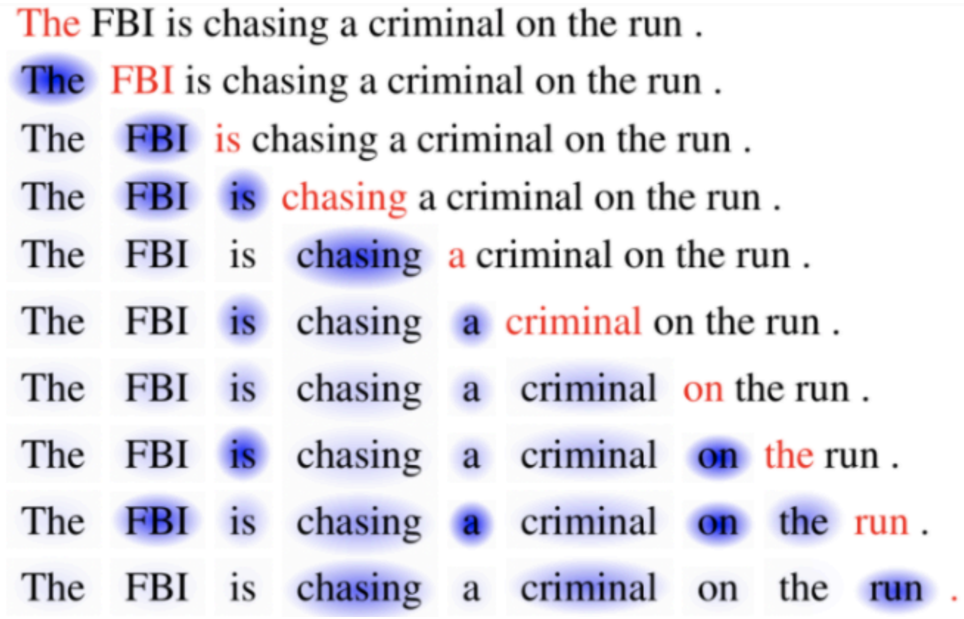
Tao Y et al. PSB. 2020.

Examples of self-attention applications

- Computer Vision and Natural Language processing



S Woo et al. *ECCV*. 2018.



The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

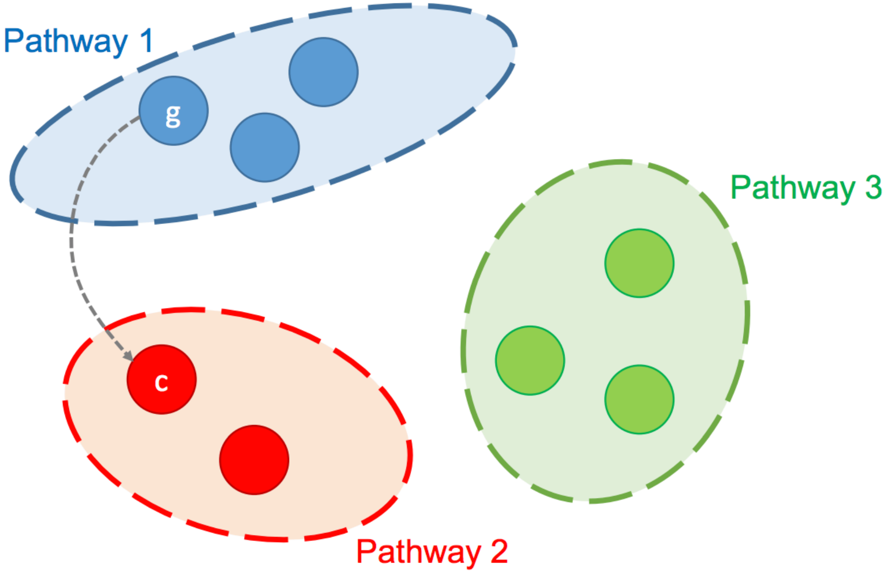
The FBI is chasing a criminal on the run .

J Cheng et al. *EMNLP*. 2016.

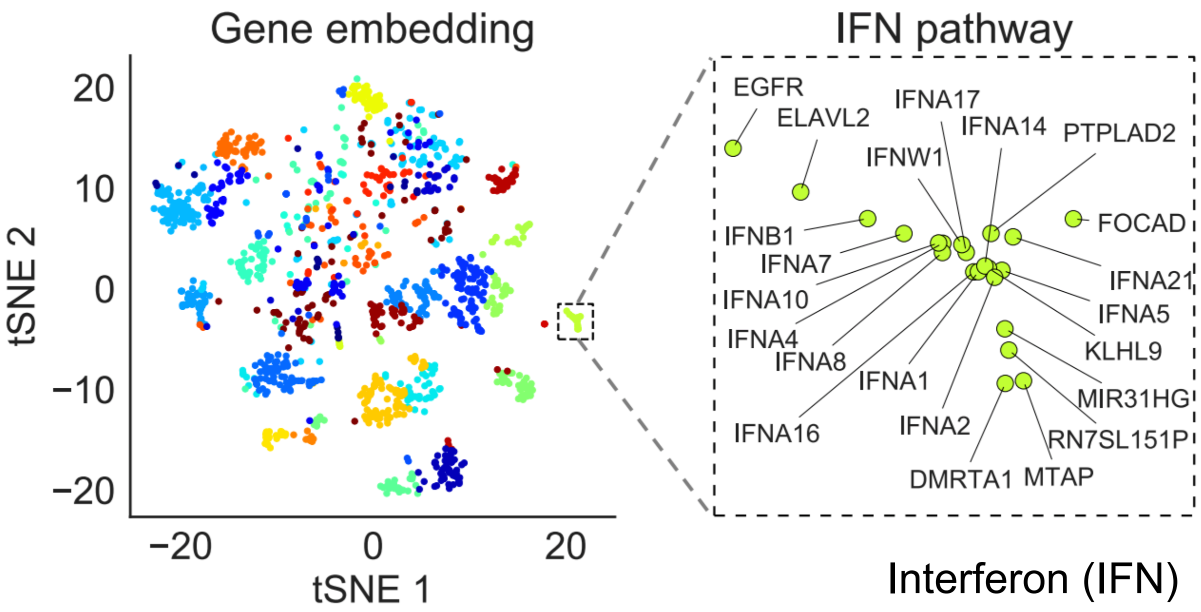
Pretraining gene embeddings: Gene2Vec

Co-occurrence pattern (e.g., mutually exclusive alterations)

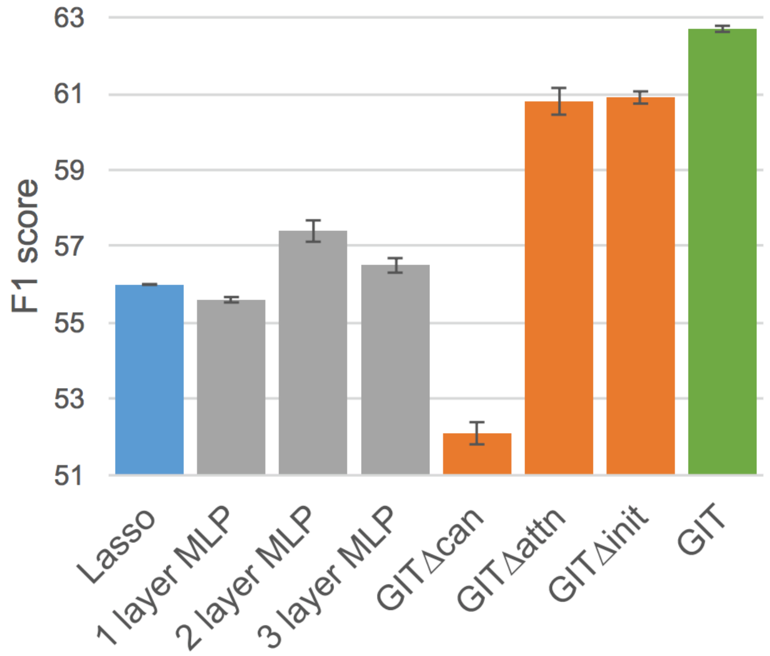
$$\Pr\left(c \in \text{Context}(g) \mid g\right) = \frac{\exp\left(\mathbf{e}_g^{\intercal}\mathbf{v}_c\right)}{\sum_{c' \in \mathcal{G}} \exp\left(\mathbf{e}_g^{\intercal}\mathbf{v}_{c'}\right)}$$



Leiserson MD et al. Nature. 2015.
Mikolov T et al. NeurIPS. 2013.



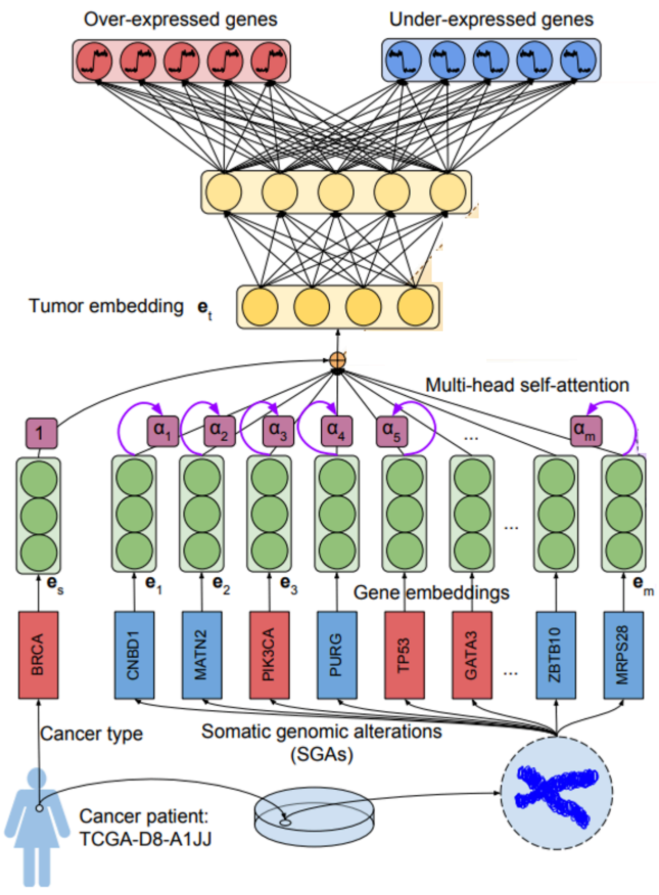
Performance of GIT and competitors



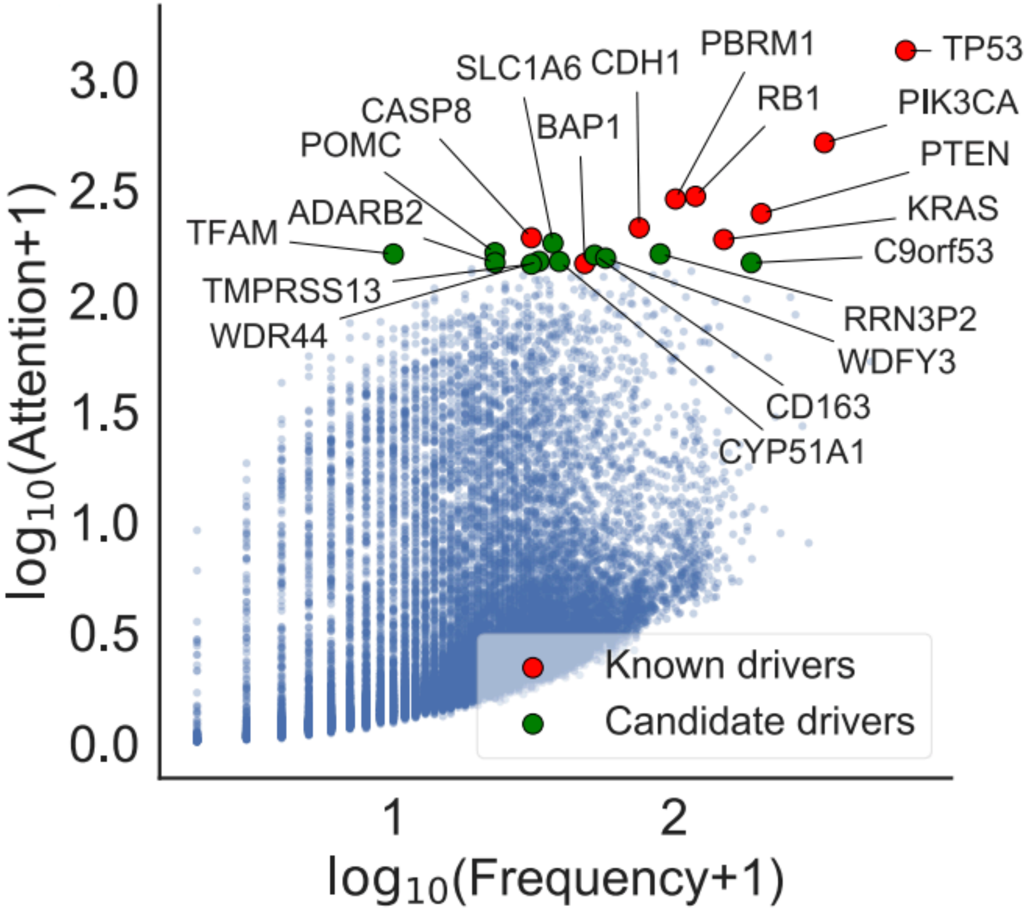
Deeper MLP is not always better.

Essential modules:

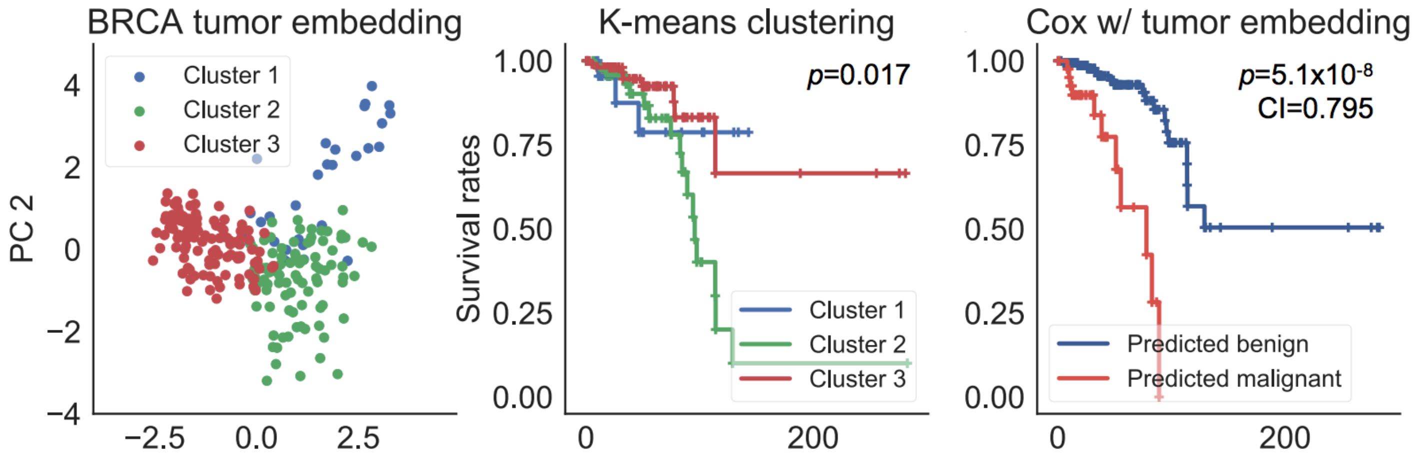
- 🟪 attn: attention mechanism
- 🟢 init: gene embeddings
- 🟥 can: cancer type input

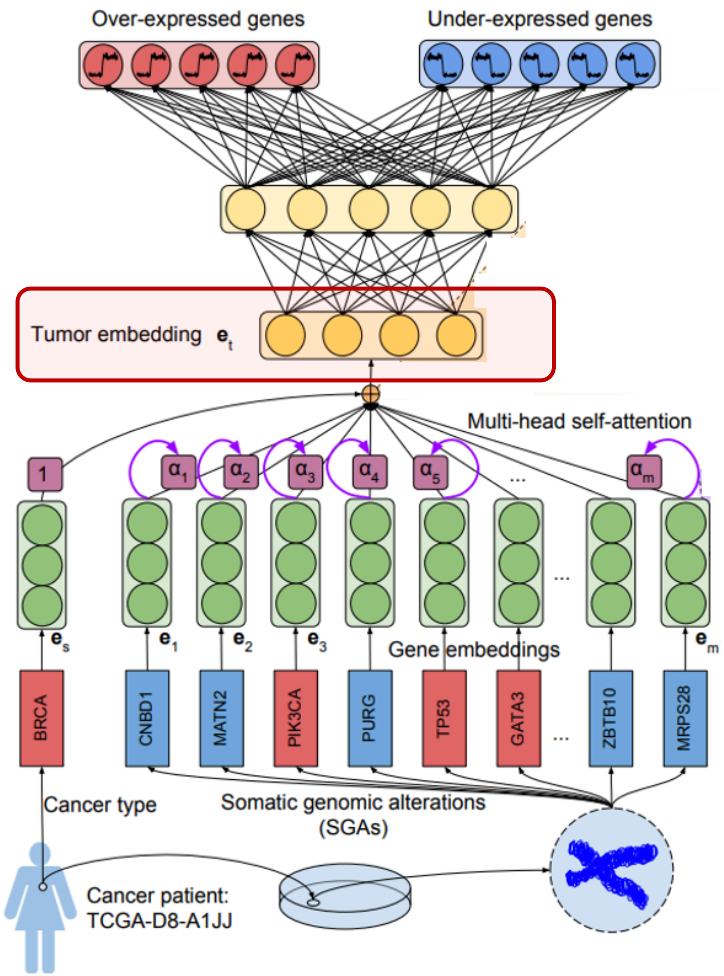


Personalized attention weights



Survival profiles encoded by tumor embeddings



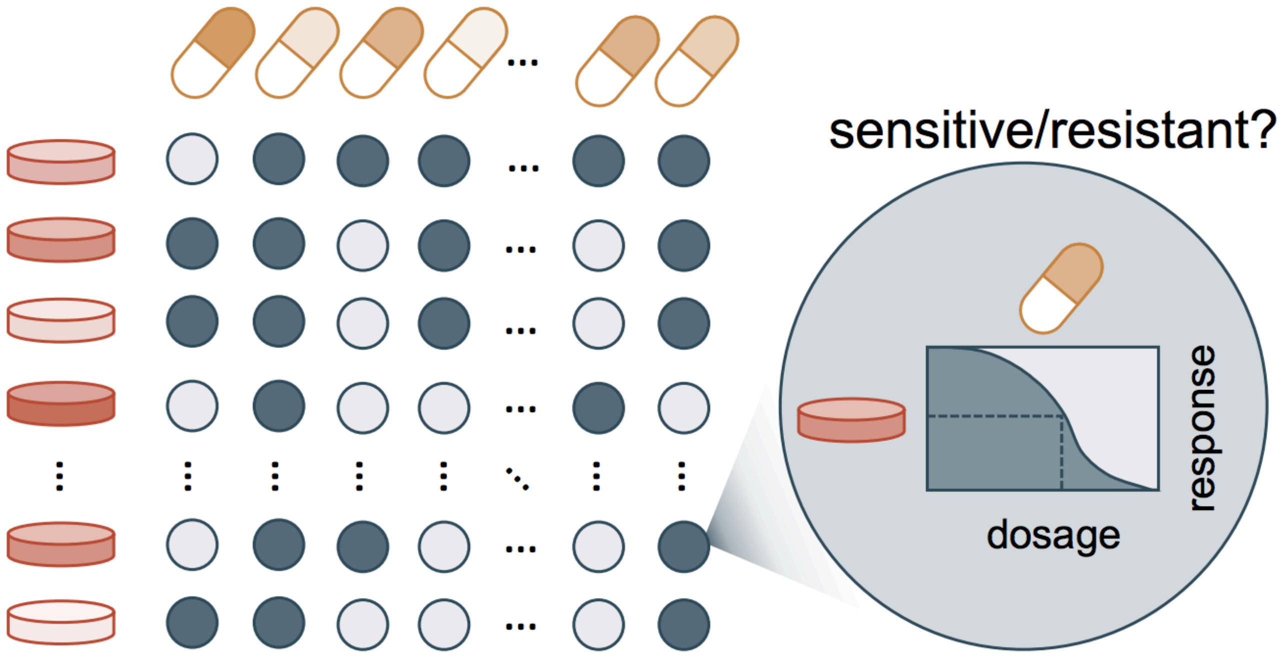


Normally impossible by only using mutation data due to sparsity.

Inference of drug response

Challenges in predicting drug response of cancer cell lines

- **Robustness:** noise.
- **Contextual effects:** gene interactions.
- **Interpretability:** biomarkers.

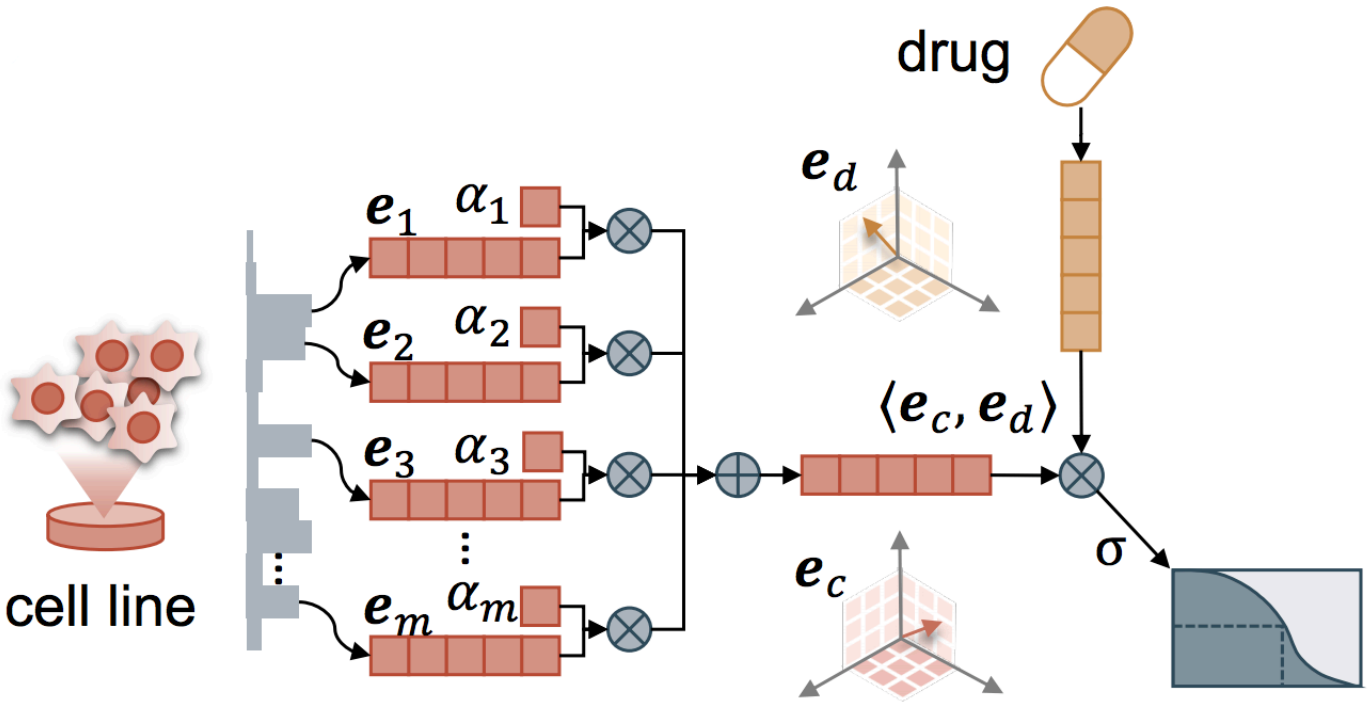


CADRE: Contextual Attention-based Drug REsponse

Collaborative filtering: copes with noisy data.

Contextual attention mechanism: improves interpretability and performance.

Pretrained gene embeddings: boosts performance further.



Tao Y et al. PMLR. 2020.

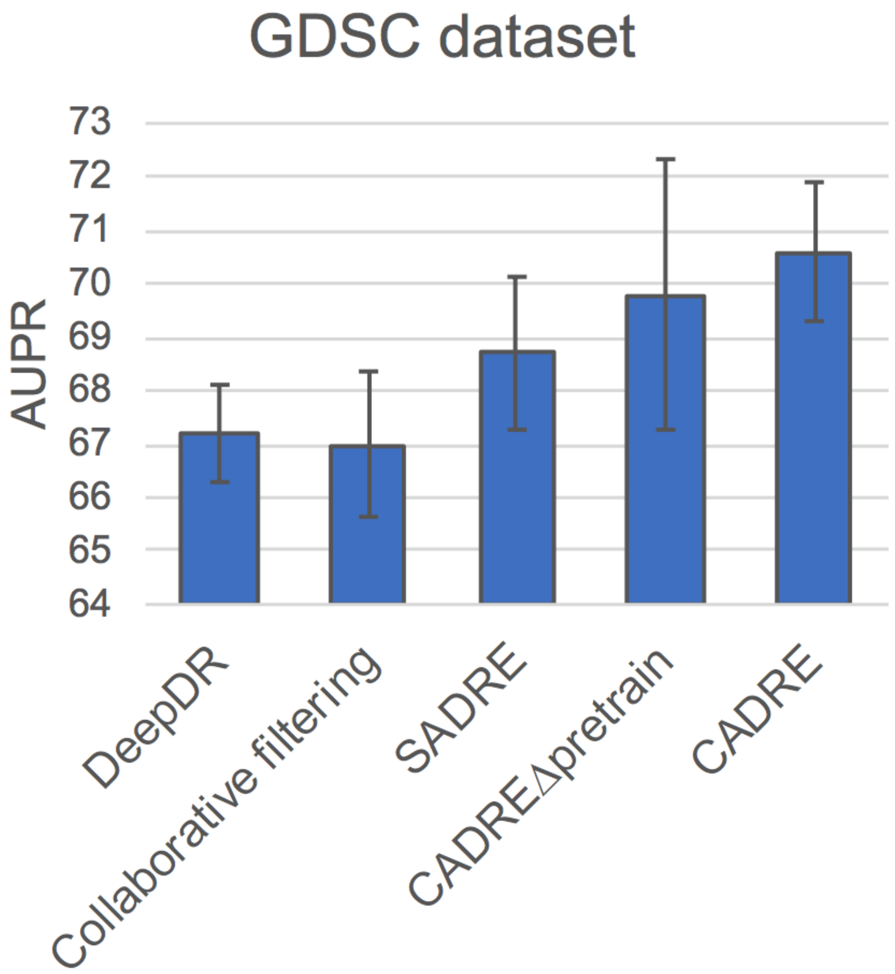
Performance of CADRE and competitors

Traditional algorithm: collaborative filtering

Deep learning: DeepDR

SADRE: self-attention

CADRE: contextual-attention



OUTLINE

1

Reliable phenotype inference of cancer through well-designed interpretable machine learning models

2

Revealing intra-/inter-tumor heterogeneity and mechanism of tumor progression via robust deconvolution and phylogenetic algorithms

3

Improving prognostic prediction of cancer by incorporating machine learning and evolutionary methods

Deconvolution of bulk RNA

Heterogeneous tumor populations/clones even from same tissue.

scRNA not available, e.g., FFPE tissue of breast cancer / immune cells.

Deconvolution of bulk tumor samples is essential.

Bulk tissue

Image credit to Bo Xia.

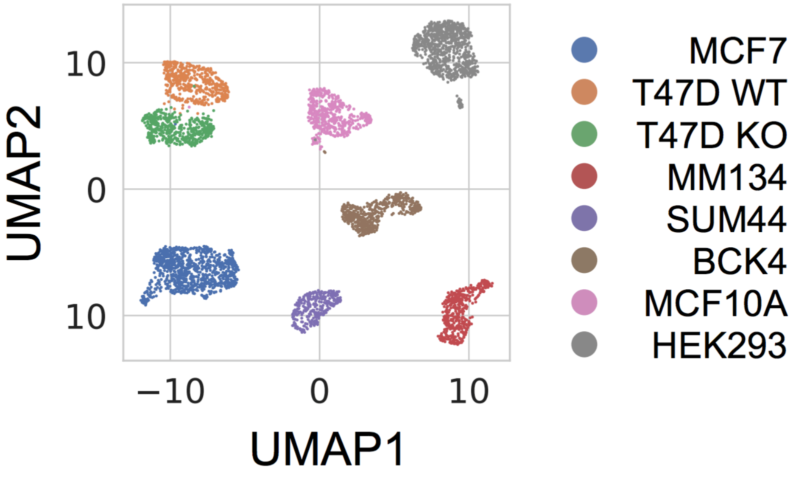
Bulk seq

Deconvolution

Cell clones

scSeq

Cell clones



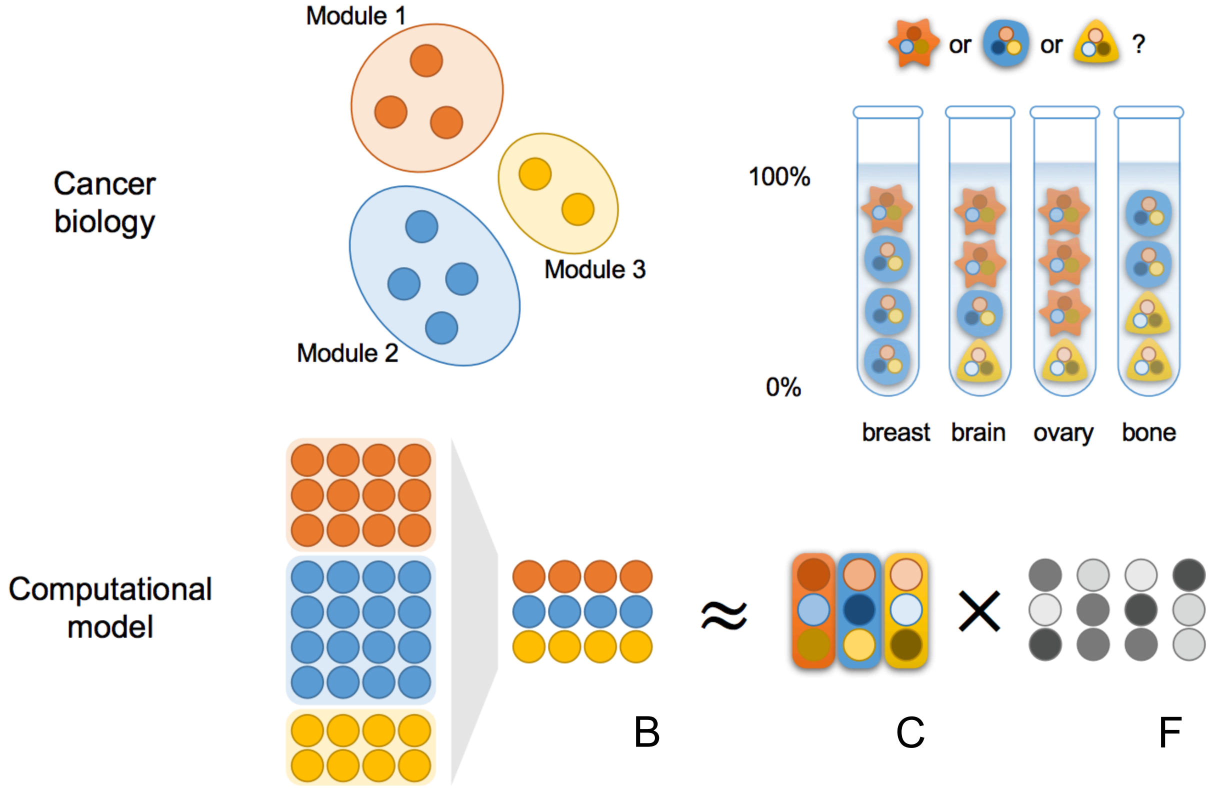
Chen F et al. Cancer Research. 2020.

Mathematical formulation of deconvolution problem

Matrix factorization.

Additional constraints make it hard to solve.

$$
\begin{aligned}
\min_{\mathbf{C},\mathbf{F}} \quad & \|\mathbf{B}-\mathbf{CF}\|_{\mathrm{Fr}}^{2}\,, \\
\text{s.t.} \quad & \mathbf{F}_{lj} \geq 0, \\
& \sum\nolimits_{l=1}^{k} \mathbf{F}_{lj} = 1
\end{aligned}
$$



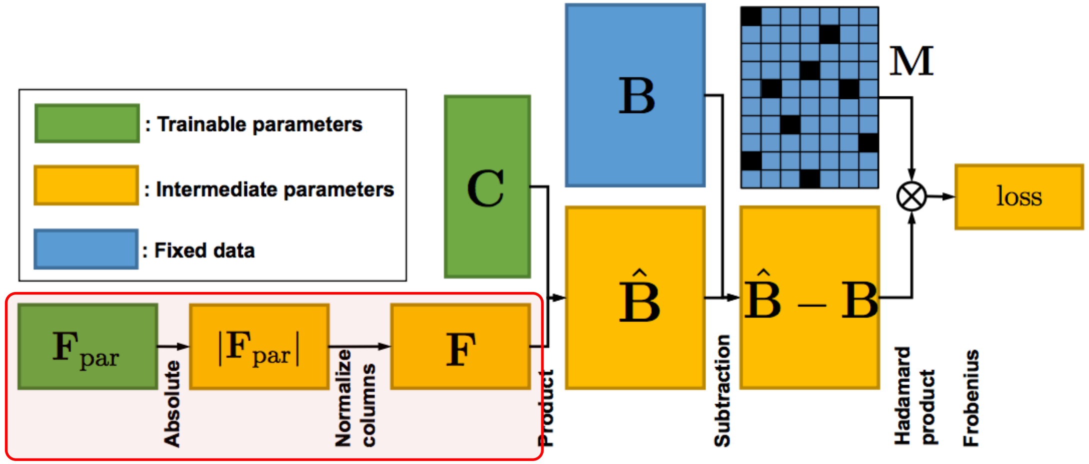
Solution 1: Gradient descent / backpropagation

NND: Neural Network Deconvolution

- Equivalently transfer the problem into a neural network (w/o input).
- Solved with backpropagation.
- Easily adapted when constraints change.

Limitations

- Need to choose learning rate.
- Computationally slow.
- Accuracy is moderate.



$$\begin{aligned} \min_{\mathbf{C},\mathbf{F}} \quad & \|\mathbf{B} - \mathbf{CF}\|_{\mathrm{Fr}}^2 , \\ \text{s.t.} \quad & \boxed{\begin{aligned} & \mathbf{F}_{lj} \geq 0, \\ & \textstyle\sum_{l=1}^{k} \mathbf{F}_{lj} = 1 \end{aligned}} \end{aligned}$$

$$\begin{aligned} \min_{\mathbf{C},\mathbf{F}_{\mathrm{par}}} \quad & \|\mathbf{B} - \mathbf{CF}\|_{\mathrm{Fr}}^2 , \\ \text{s.t.} \quad & \boxed{\mathbf{F} = \mathrm{cwn}\left(|\mathbf{F}_{\mathrm{par}}|\right)} \end{aligned}$$

cwn: column-wise normalization

Tao Y et al. Frontiers in Physiology. 2020.

Solution 2: Hybrid optimizer

RAD: Robust and Accurate Deconvolution

Fast and accurate by utilizing a hybrid optimizer w/ three phases.

Almost no parameters need to choose manually.

Phase 1: Multiplicative update of C and F until convergence

Fast to converge to a reasonable solution.

Phase 2: Coordinate descent of C and F until convergence

Further reduces loss by ~5-30%.

Phase 3: Minimum similarity selection of C

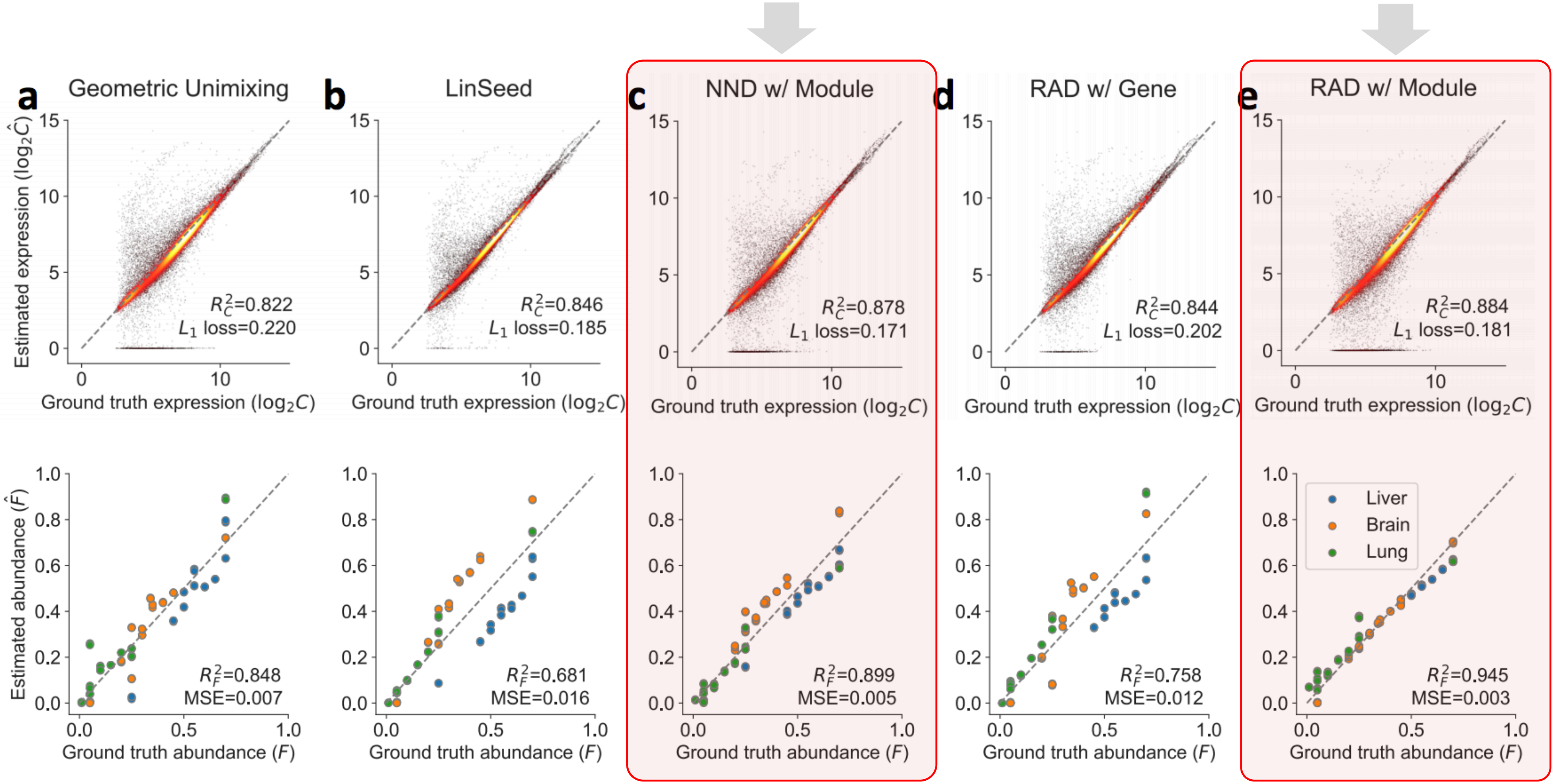
Select biologically meaningful solutions.

$$
\begin{aligned}
\min_{\mathbf{C},\mathbf{F}} \quad & \|\mathbf{B}-\mathbf{CF}\|_{\mathrm{Fr}}^2\,, \\
\text{s.t.} \quad & \mathbf{C}_{il} \geq 0, \quad i=1,\dots,m,\ l=1,\dots,k \\
& \mathbf{F}_{lj} \geq 0, \quad l=1,\dots,k,\ j=1,\dots,n \\
& \textstyle\sum_{l=1}^{k} \mathbf{F}_{lj} = 1, \qquad j=1,\dots,n
\end{aligned}
$$

Tao Y et al. Bioinformatics. 2020.

Performance of NND and RAD

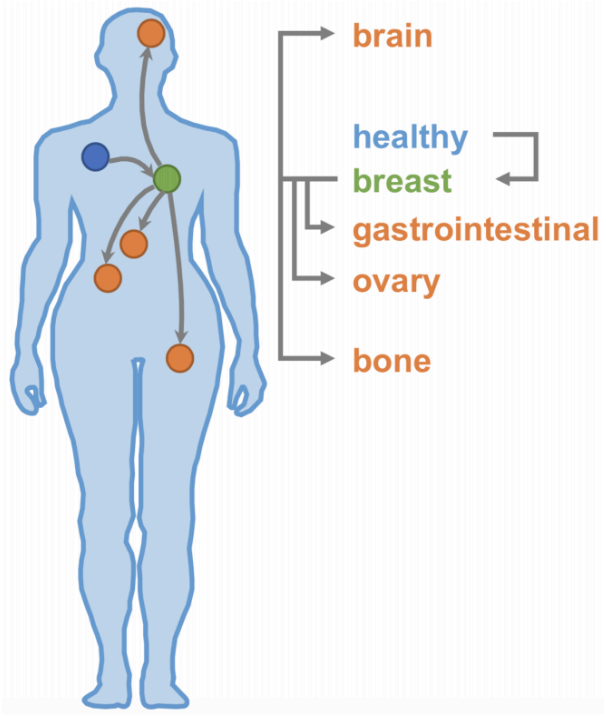
GSE19830 dataset: mixture of liver, brain and lung cells (Shen-Orr et al. Nature Methods. 2010)



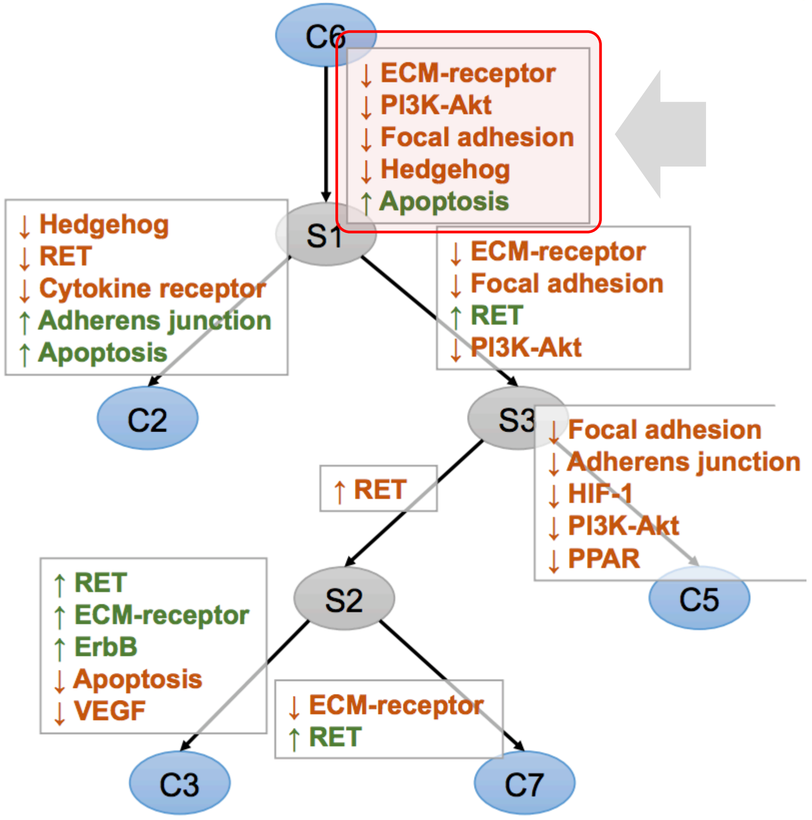
Common evolutionary mechanism

Dataset:

- Matched bulk RNA-Seq
- Breast cancer metastasis patients



Zhu L et al. Journal for ImmunoTherapy of Cancer. 2019.



Infer phylogenies from **RAD-unmixed populations**

Common early pathway-level events:

- ↓ PI3K-Akt
- ↓ Extracellular matrix (ECM)-receptor interaction
- ↓ Focal adhesion

OUTLINE

1

Reliable phenotype inference of cancer through well-designed interpretable machine learning models

2

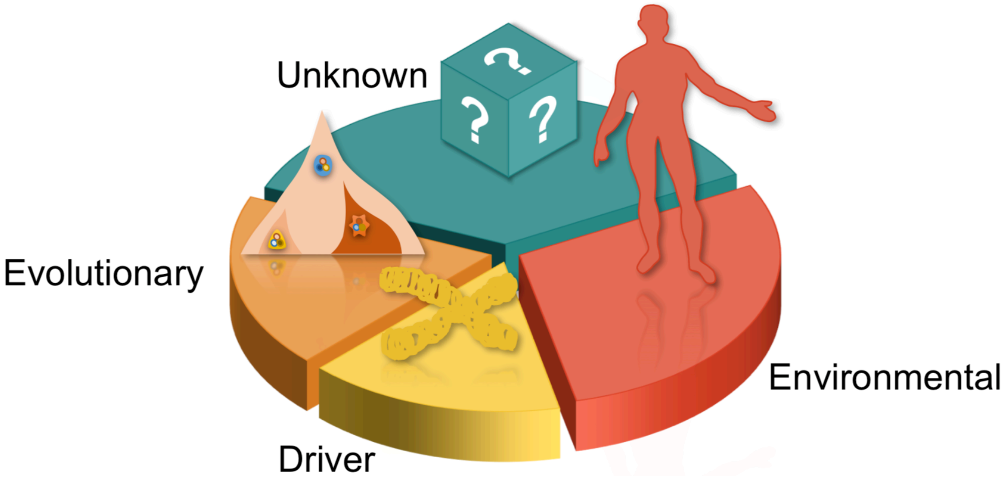
Revealing intra-/inter-tumor heterogeneity and mechanism of tumor progression via robust deconvolution and phylogenetic algorithms

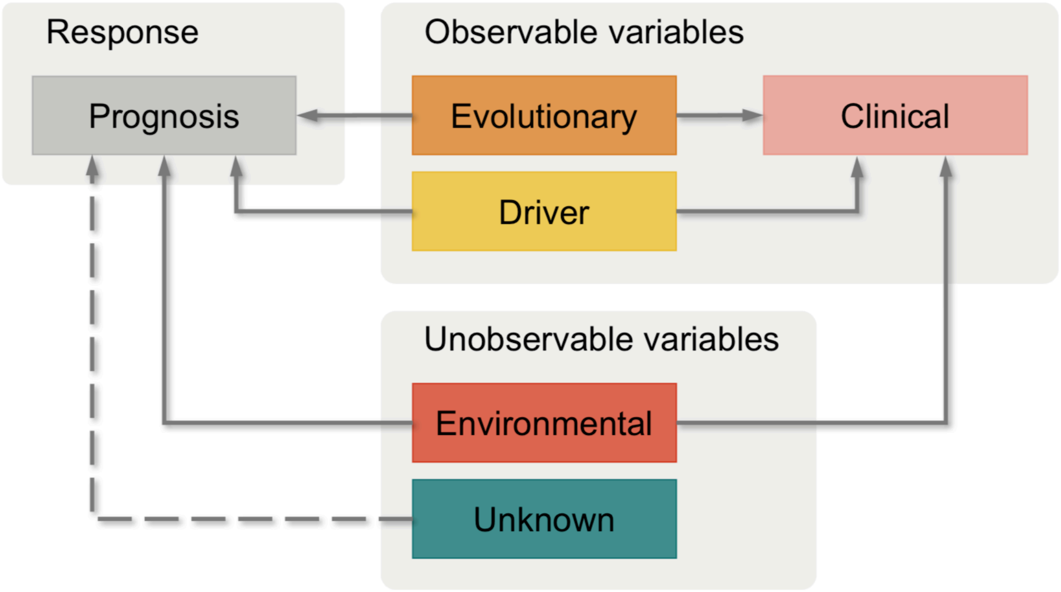
3

Improving prognostic prediction of cancer by incorporating machine learning and evolutionary methods

Factors affecting tumor prognosis

- Clinical and driver-level genomic factors are well-studied.
  - TNM pathological stage, driver mutations in BRCA1/2, PIK3CA, etc.
- Impact of evolutionary features are little known.
- Different types of factors are correlated.





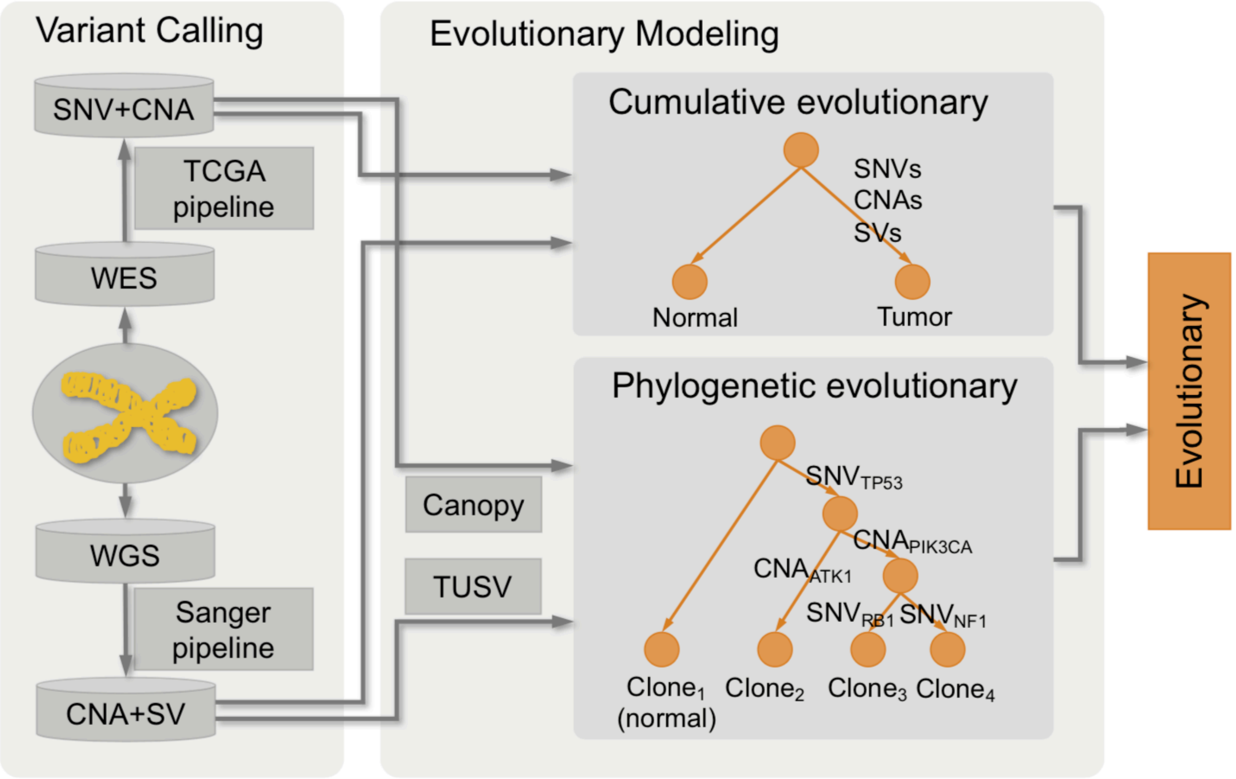
Pipeline of extracting evolutionary features

Mutational signatures:

- e.g., *T→A*, *GTC→GAC*, mutation rate of CNAs.

Topological structures of phylogenies:

- e.g., height, average branch length.



Phylo-risk: Evaluating contribution of evolutionary features to tumor progression risk

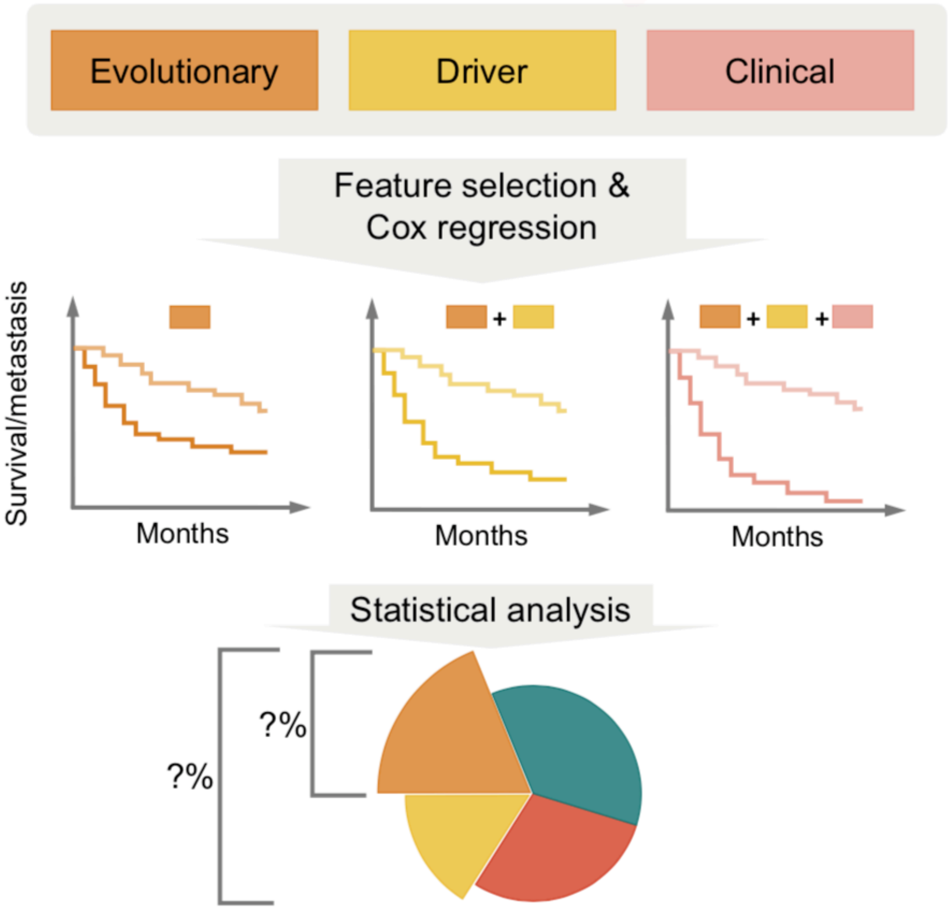
- **Employ L0-regularizaed Cox regression**

$$\min_{\beta}\ l\left(\beta \,\middle|\, \{(X_i, y_i, \delta_i)\}_{i=1}^{N}\right), \quad \text{s.t. } \|\beta\|_0 \le k$$

  - Solved heuristically through step-wise feature selection

- **Risk evaluated in the log-scale HR (hazard ratio)**

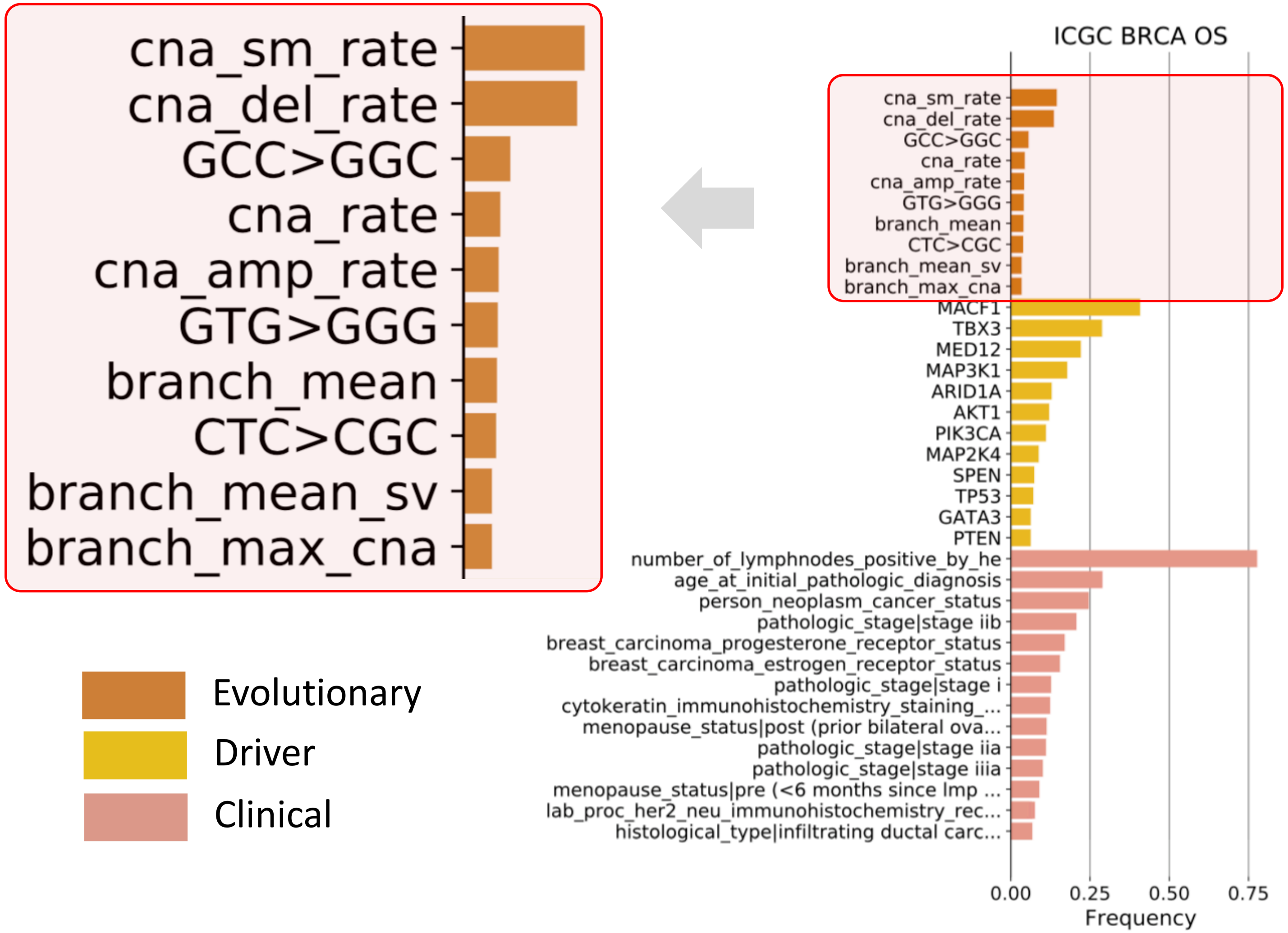
$$\text{fraction(evolutionary)} = \frac{\log \text{HR(evolutionary)}}{\log \text{HR(evolutionary+driver+clinical)}}$$



Tao Y et al. PLOS Computational Biology. 2021.

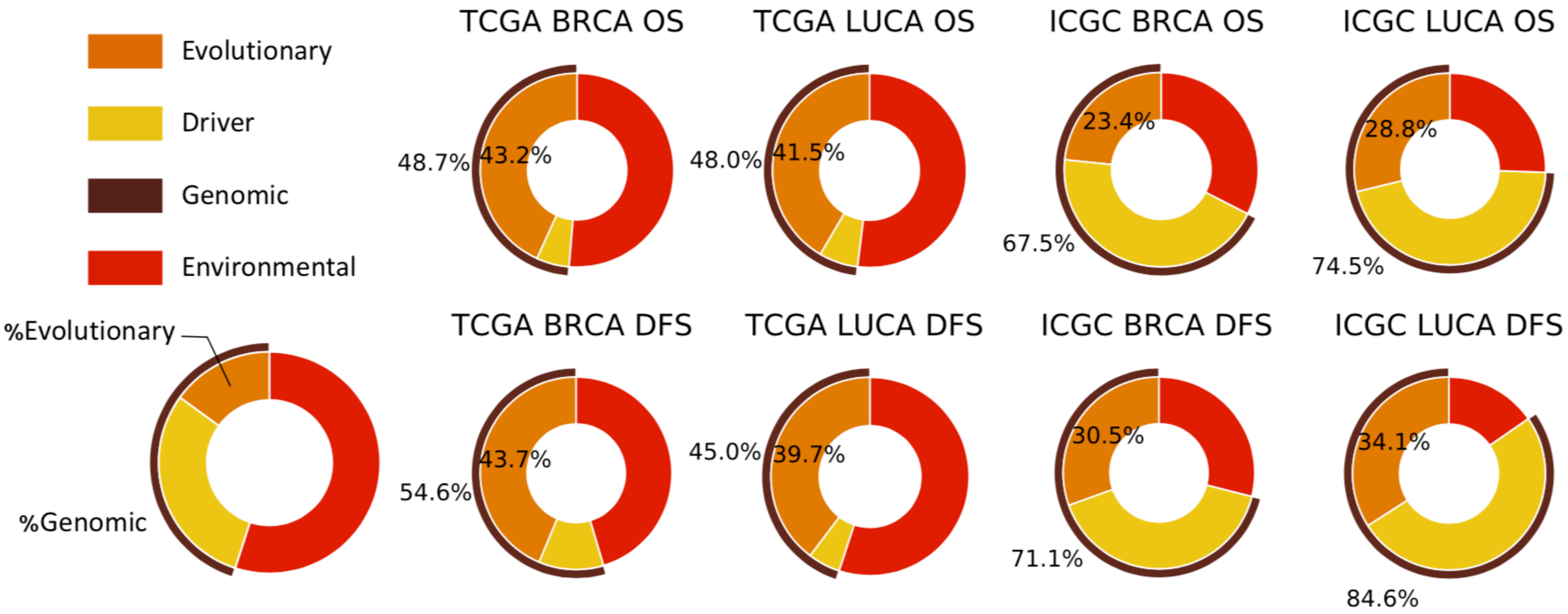
Important evolutionary features

- Trinucleotide SNV rate
- Features related to CNAs and SVs
  - CNA duplication/deletion rates
  - Rates of CNA above/below 500k nt
- Average branch length in unit of SV rates



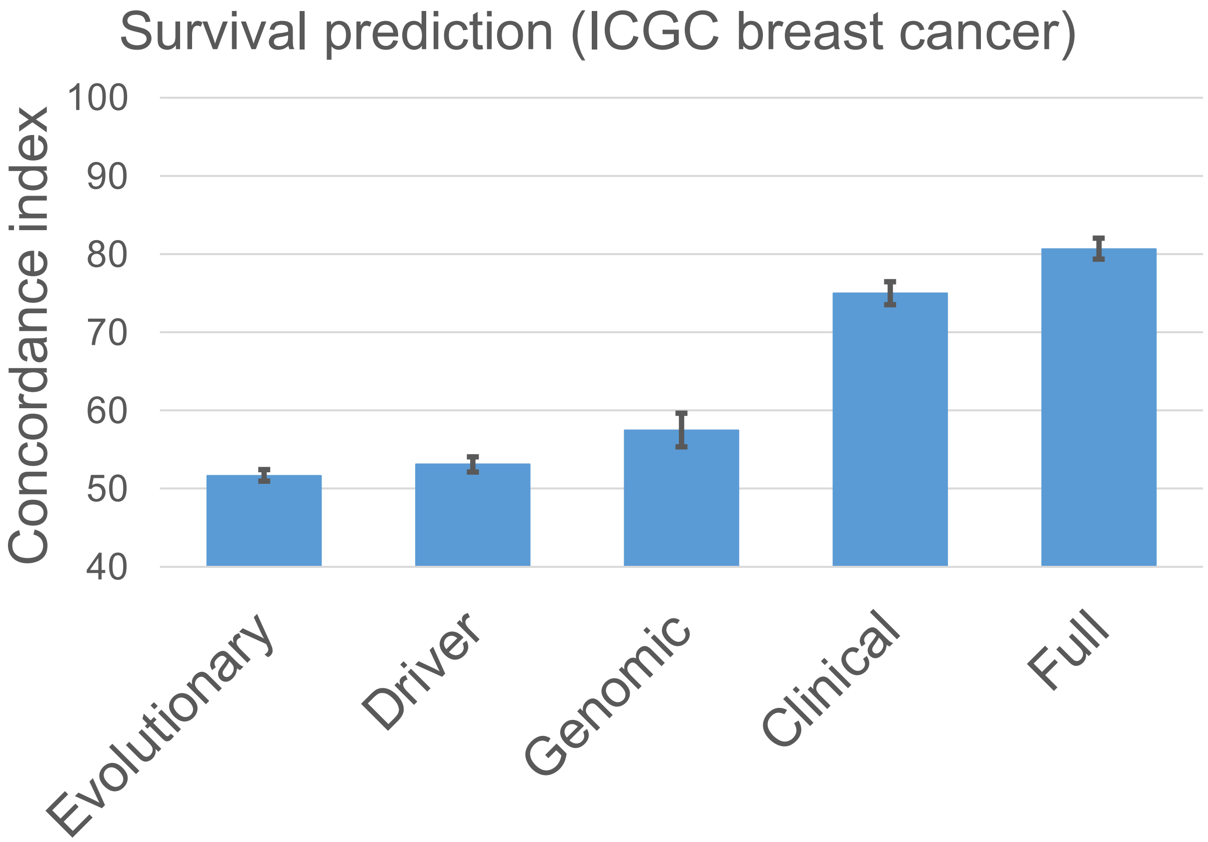
Contribution of evolutionary factors

- Two datasets (TCGA/ICGC); Two cancer types (BRCA/LUCA); Two tasks (OS/DFS).
- Evolutionary features account for around 1/3 of the overall risk.

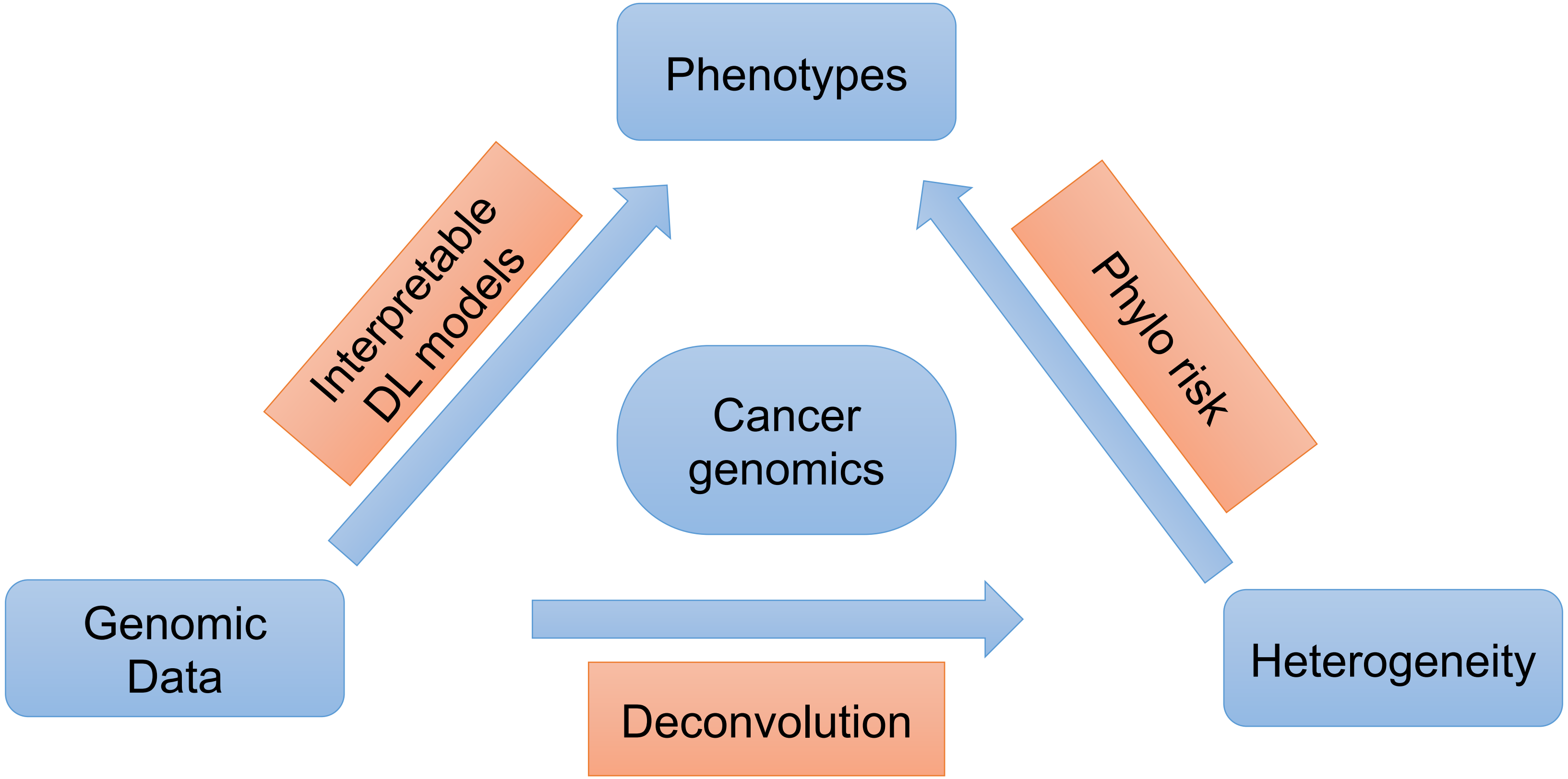


Improving prognostic prediction using evolutionary features

- Performance evaluated through nested cross-validation.

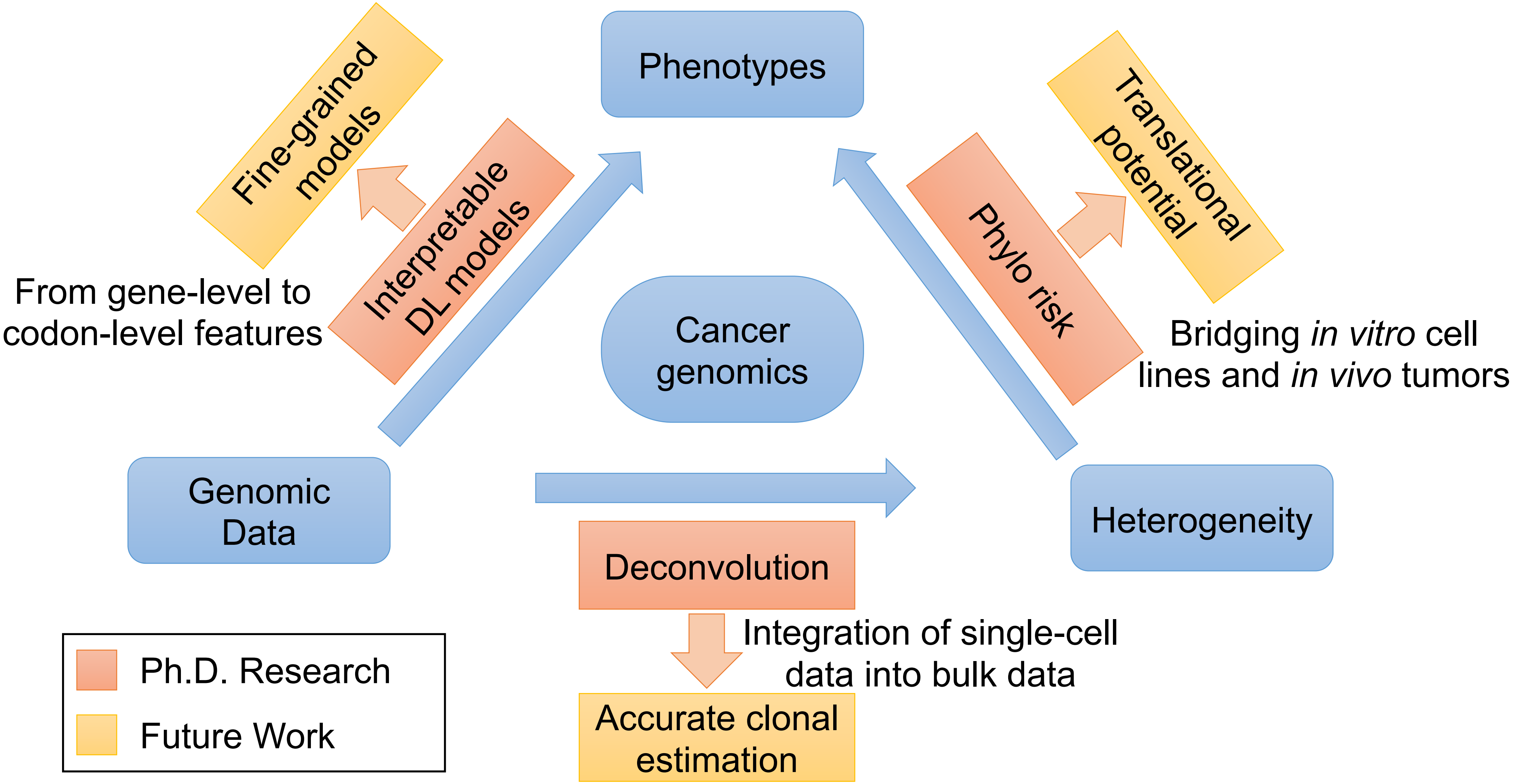


Conclusions



- Interpretable deep learning models for accurate phenotype inference of tumor.
- Deconvolution of bulk breast cancer samples discovers early pathway-level event of metastasis.
- Trinucleotide mutation rates, CNAs, and SVs contribute to around 1/3 of the tumor progression risk.

Future work



Phenotypes

Cancer genomics

Genomic Data

Heterogeneity

Interpretable DL models

Fine-grained models

From gene-level to codon-level features

Phylo risk

Translational potential

Bridging *in vitro* cell lines and *in vivo* tumors

Deconvolution

Accurate clonal estimation

Integration of single-cell data into bulk data

🟧 Ph.D. Research

🟨 Future Work

Acknowledgments

Prof. Russell Schwartz

Haoyun Lei

Xuecong Fu

Marcus Thomas

Arjun Srivatsa

Xiaoyue Cui

Ziyi Cui

Jesse Eaton

Hannah Kim

Haoran Chen

Yuanqi Zhao

Alex Guo

Rishi Verma

Alyssa Lee

Prof. Hatice Ulku Osmanbeyoglu

Xiaojun Ma

Drake Palmer

Dr. William W. Cohen

Prof. Jian Ma

Ashok Rajaraman

Yang Zhang

Prof. Xinghua Lu

Shuangxia Ren

Chunhui Cai

Michael Q. Ding

Yifan Xue

Xueer Chen

Qiao Jin

Prof. Adrian V. Lee

Kai Ding

Fangyuan (Chelsea) Chen

Prof. Eric P. Xing

Haohan Wang

Benjamin Lengerich

Pengtao Xie

NIH National Institutes of Health
Turning Discovery Into Health

pennsylvania
DEPARTMENT OF HEALTH

Pittsburgh Health Data Alliance
Carnegie Mellon University University of Pittsburgh UPMC

Center for Machine Learning and Health
Carnegie Mellon University

66 MARIO LEMIEUX FOUNDATION

SUSAN G. KOMEN FOR THE CURE

aws

BREAST CANCER ALLIANCE