Predicting Cancer Phenotypes based on Somatic Genomic Alterations via Genomic Impact Transformer

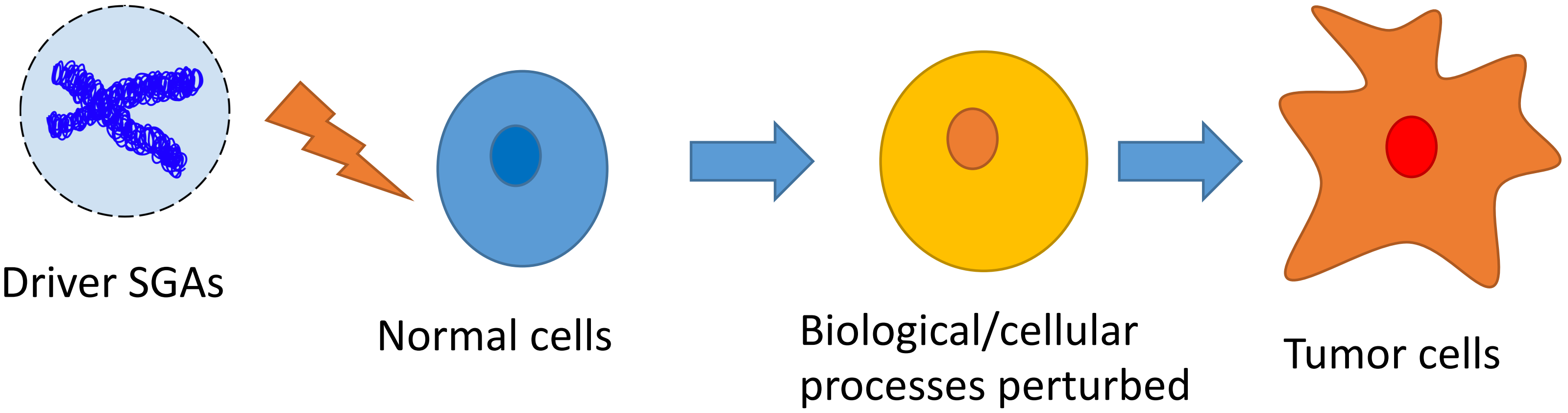
Yifeng Tao[1], Chunhui Cai[2], William W. Cohen[1*], Xinghua Lu[2*]

[1]Carnegie Mellon University

[2]University of Pittsburgh

Background

- Cancers are mainly caused by somatic genomic alterations (SGAs)
  - Driver SGAs → causal to tumor development
  - Passenger SGAs → neutral mutations

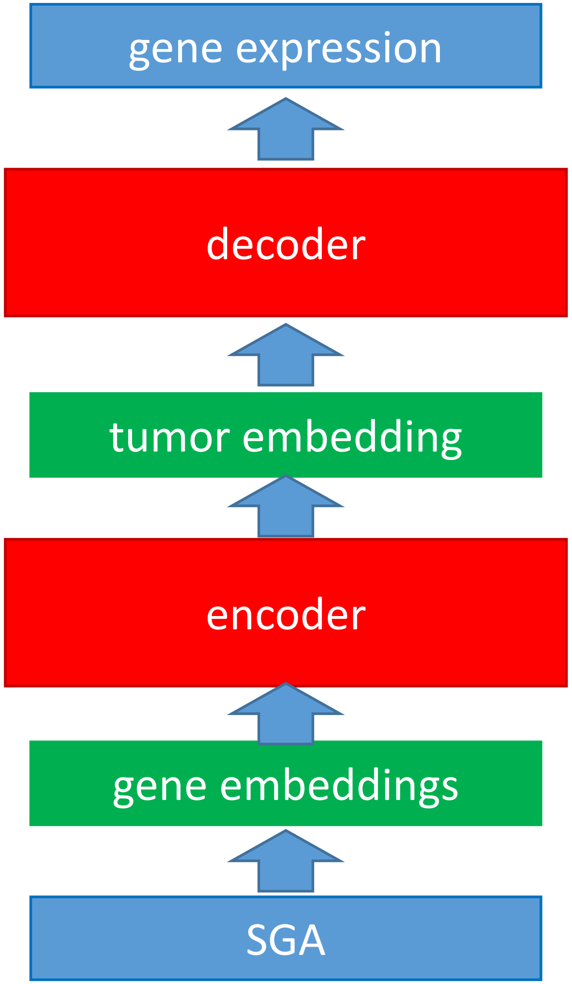


Challenges

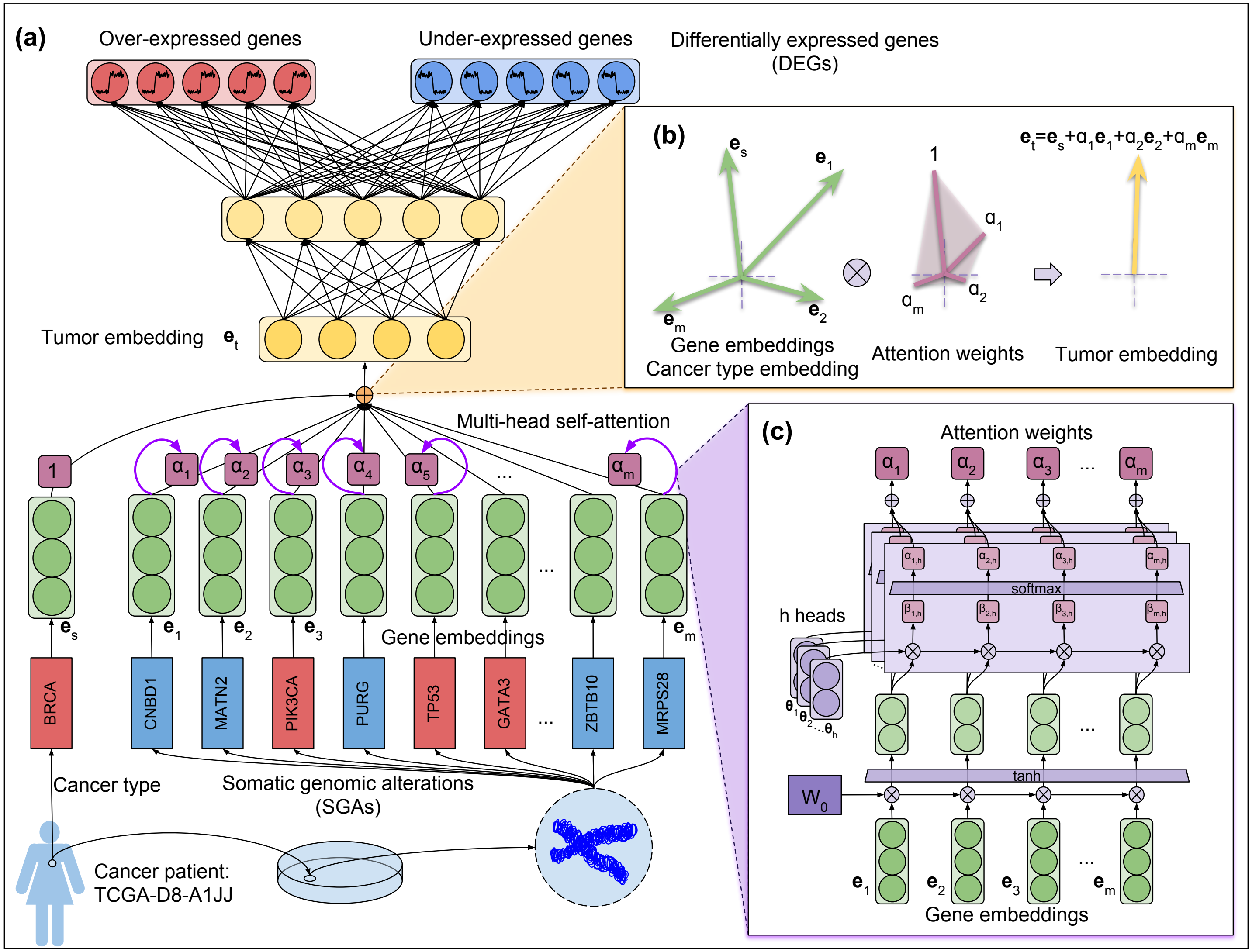
- Driver SGA detection
  - Solution 1: frequency
  - Solution 2: conserved domain of protein
  - Problem: downstream effect of SGAs
- SGA/tumor representation
  - Solution: a higher dimensional one-hot/sparse vector
  - Problem: little information/knowledge

Genomic Impact Transformer (GIT)

- GIT: encoder-decoder architecture
- Mimic cellular signaling process
- Driver SGA detection
  - Problem: downstream effect of SGAs
  - Solution: supervised by gene expressions
- SGA/tumor representation
  - Problem: little information/knowledge
  - Solution: gene/tumor embedding

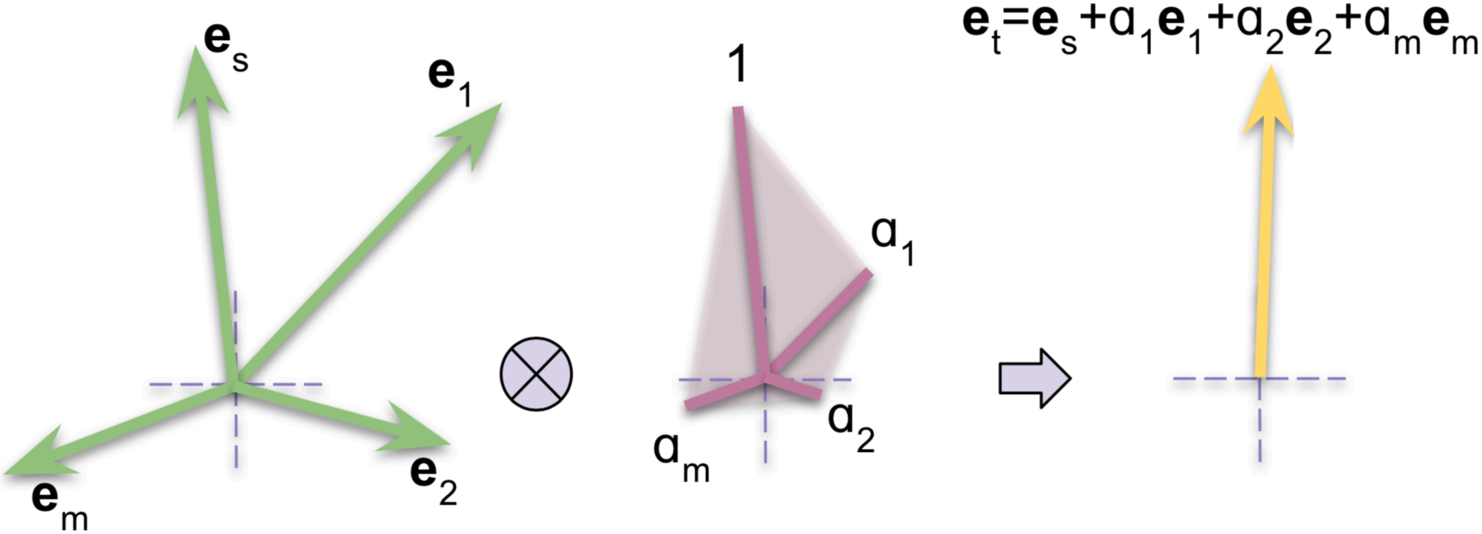


Genomic Impact Transformer (GIT)



Encoder: Attention

$$\mathbf{e}_t=\mathbf{e}_s+a_1\mathbf{e}_1+a_2\mathbf{e}_2+a_m\mathbf{e}_m$$



Gene embeddings
Cancer type embedding

Attention weights

Tumor embedding

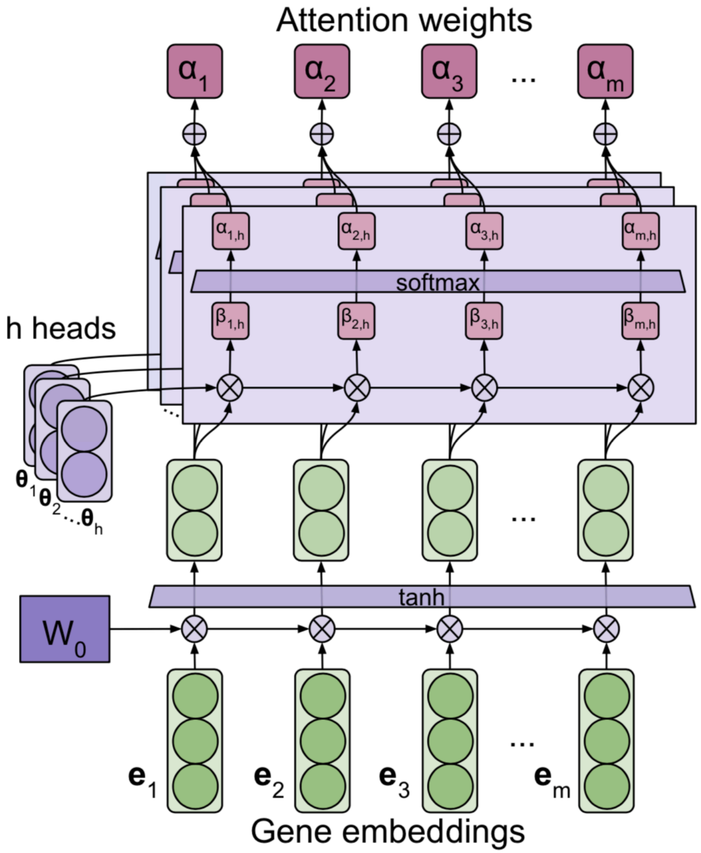
Encoder: Attention

$$\boxed{\alpha_1, \alpha_2, \ldots, \alpha_m} = \boxed{\text{Function}_{\text{Attention}}}(\boxed{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m})$$

$$\alpha_g = \sum_{j=1}^{h} \alpha_{g,j} = \alpha_{g,1} + \alpha_{g,2} + \ldots + \alpha_{g,h},\ g = 1, 2, \ldots, m$$

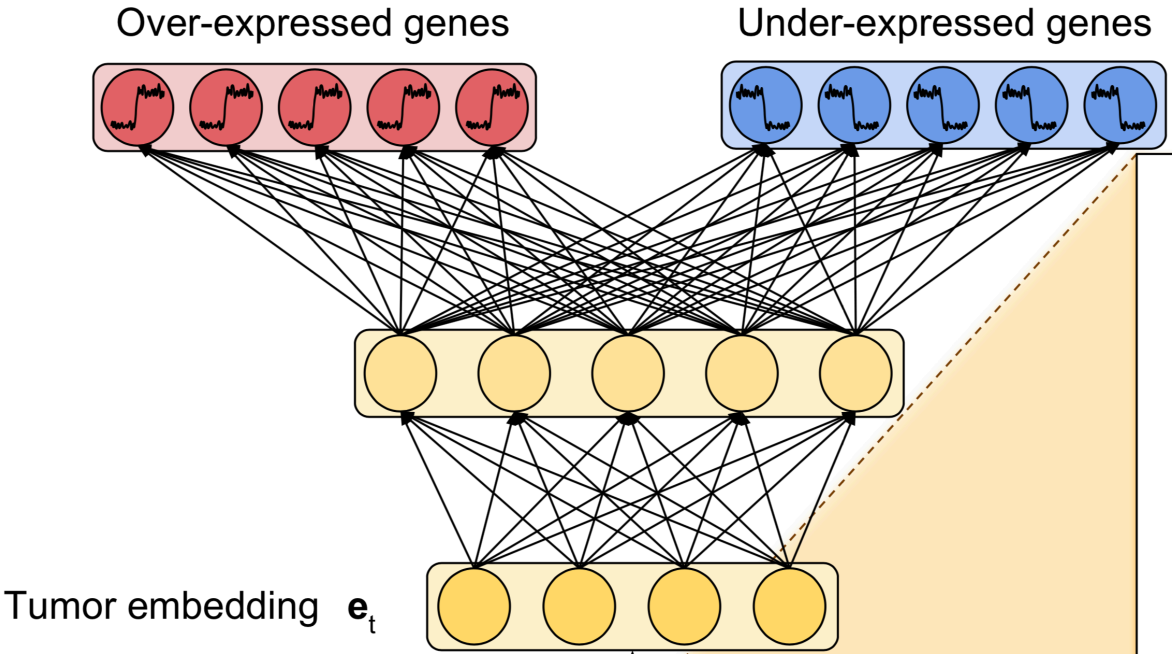
$$\alpha_{1,j}, \alpha_{2,j}, \ldots, \alpha_{m,j} = \text{softmax}(\beta_{1,j}, \beta_{2,j}, \ldots, \beta_{m,j})$$

$$\beta_{g,j} = \boldsymbol{\theta}_j^{\top} \cdot \tanh(W_0 \cdot \mathbf{e}_g),\ g = 1, 2, \ldots, m$$



Decoder: MLP

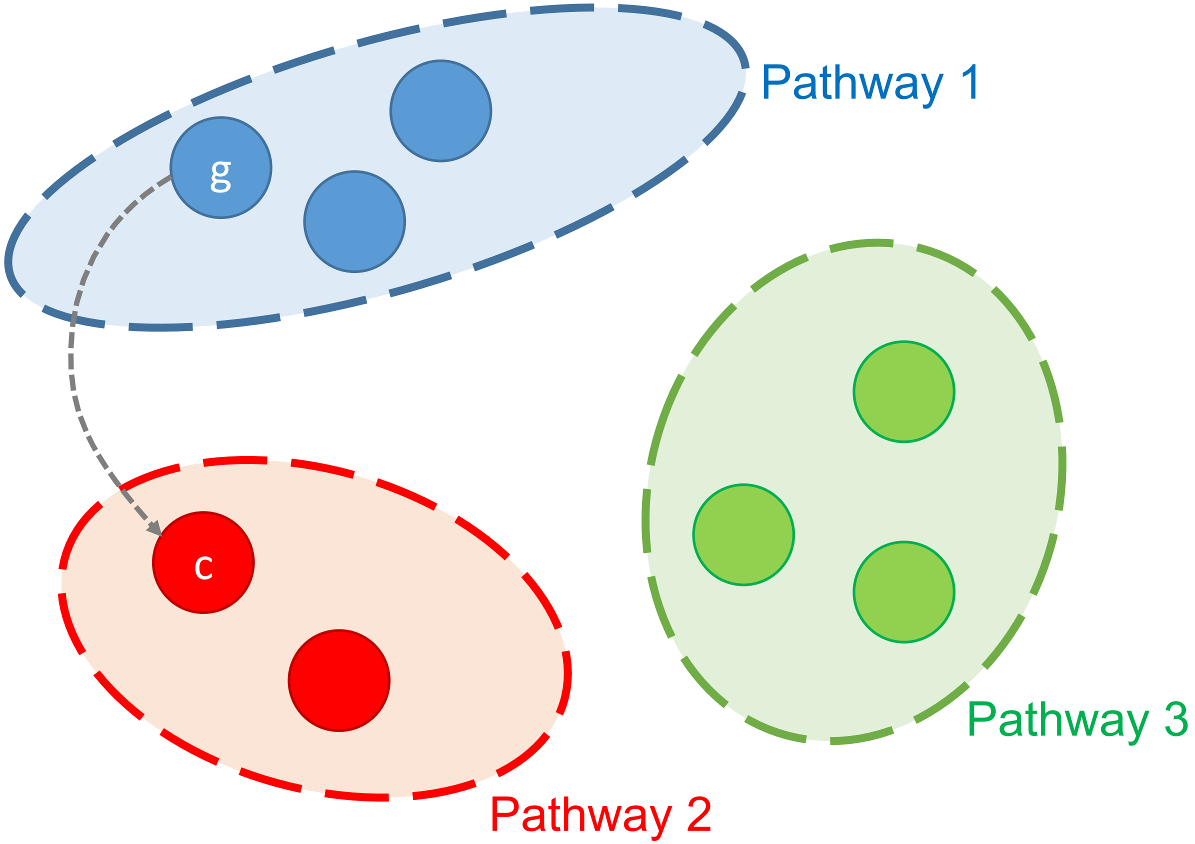
$$\hat{y} = \sigma(W_2 \cdot \mathrm{ReLU}(W_1 \cdot \mathrm{ReLU}(\mathbf{e}_t) + b_1) + b_2)$$



Pre-training Gene Embedding: Gene2Vec

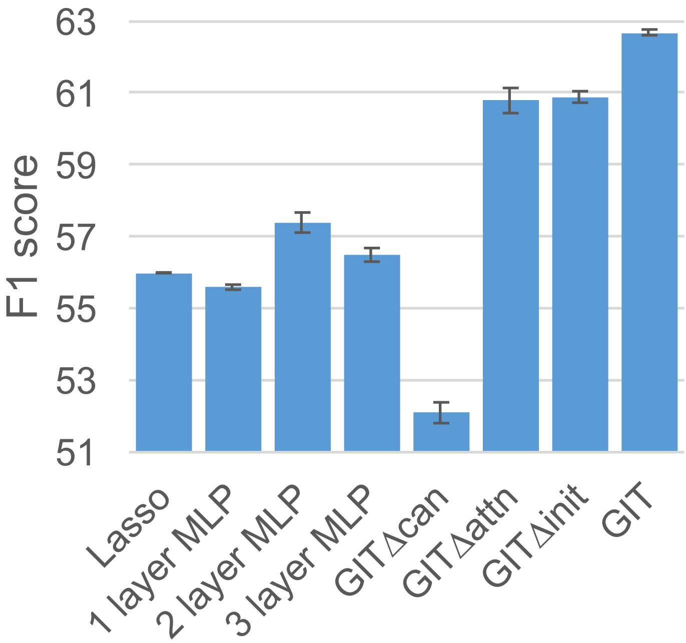
- Co-occurrence pattern

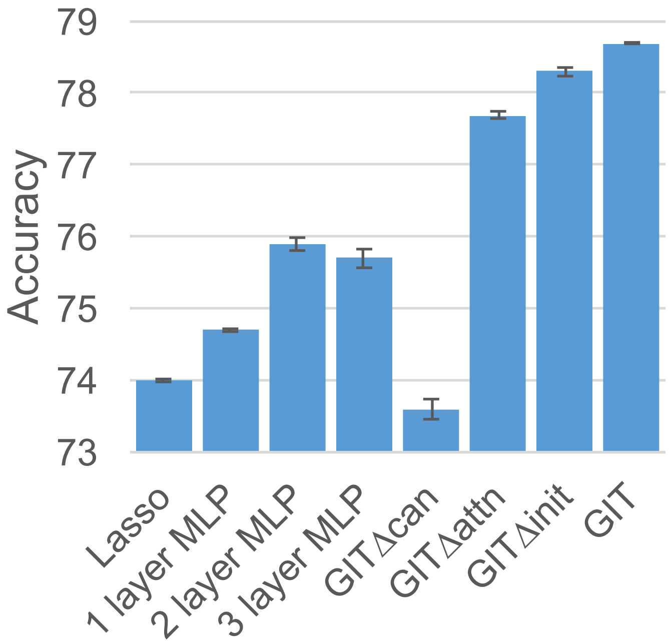
$$\Pr\left(c \in \mathrm{Context}(g) \mid g\right) = \frac{\exp\left(\mathbf{e}_g^{\intercal}\mathbf{v}_c\right)}{\sum_{c' \in \mathcal{G}} \exp\left(\mathbf{e}_g^{\intercal}\mathbf{v}_{c'}\right)}$$



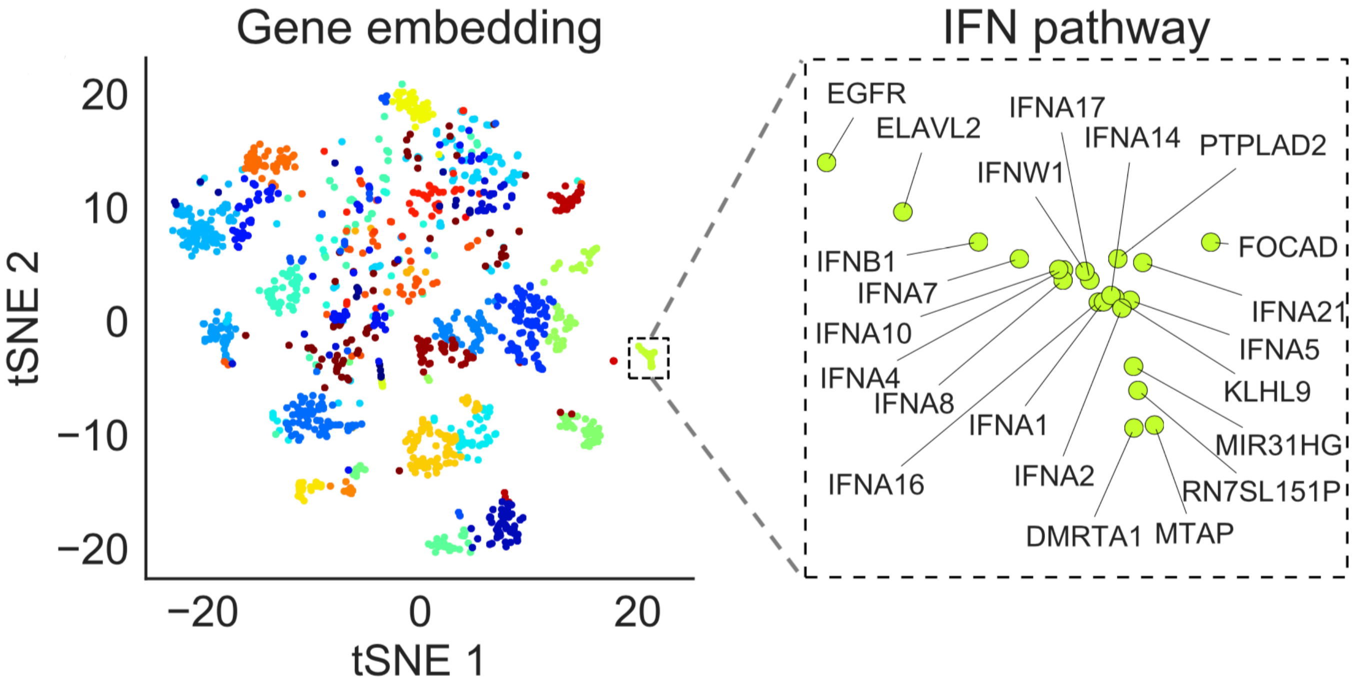
Performance

- Predicting gene expression using SGAs





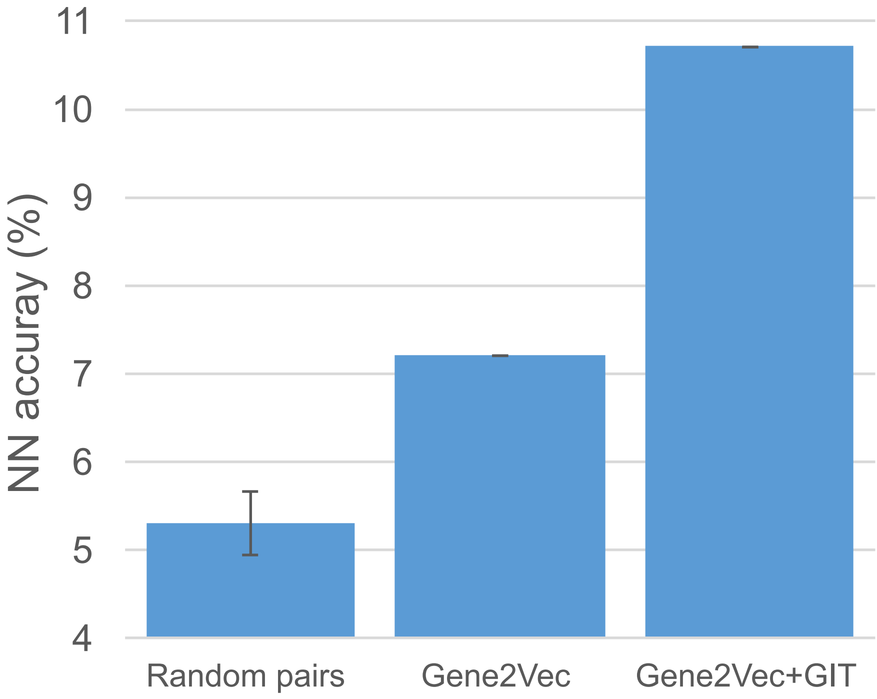
Gene Embedding



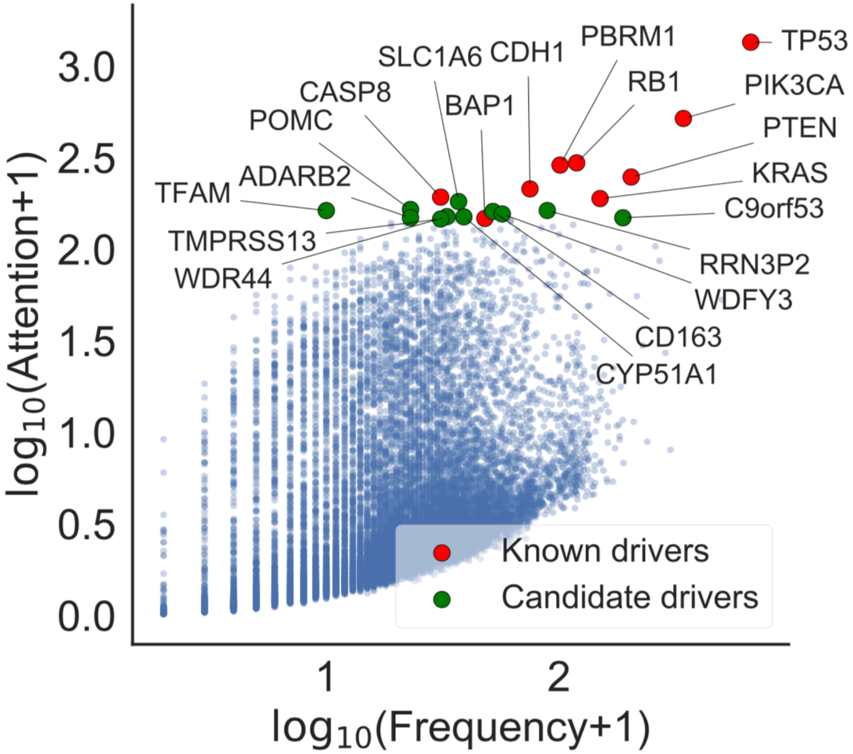
Gene Embedding

- NN accuracy: functional similar genes → closer in embedding space

$$\text{NN accuracy} = \mathbb{E}_{\mathbf{e}_c \in \text{NN}(\mathbf{e}_g)} \left[\mathbb{1}\left(\text{GO}(g) \cap \text{GO}(c) \neq \emptyset \right) \right]$$

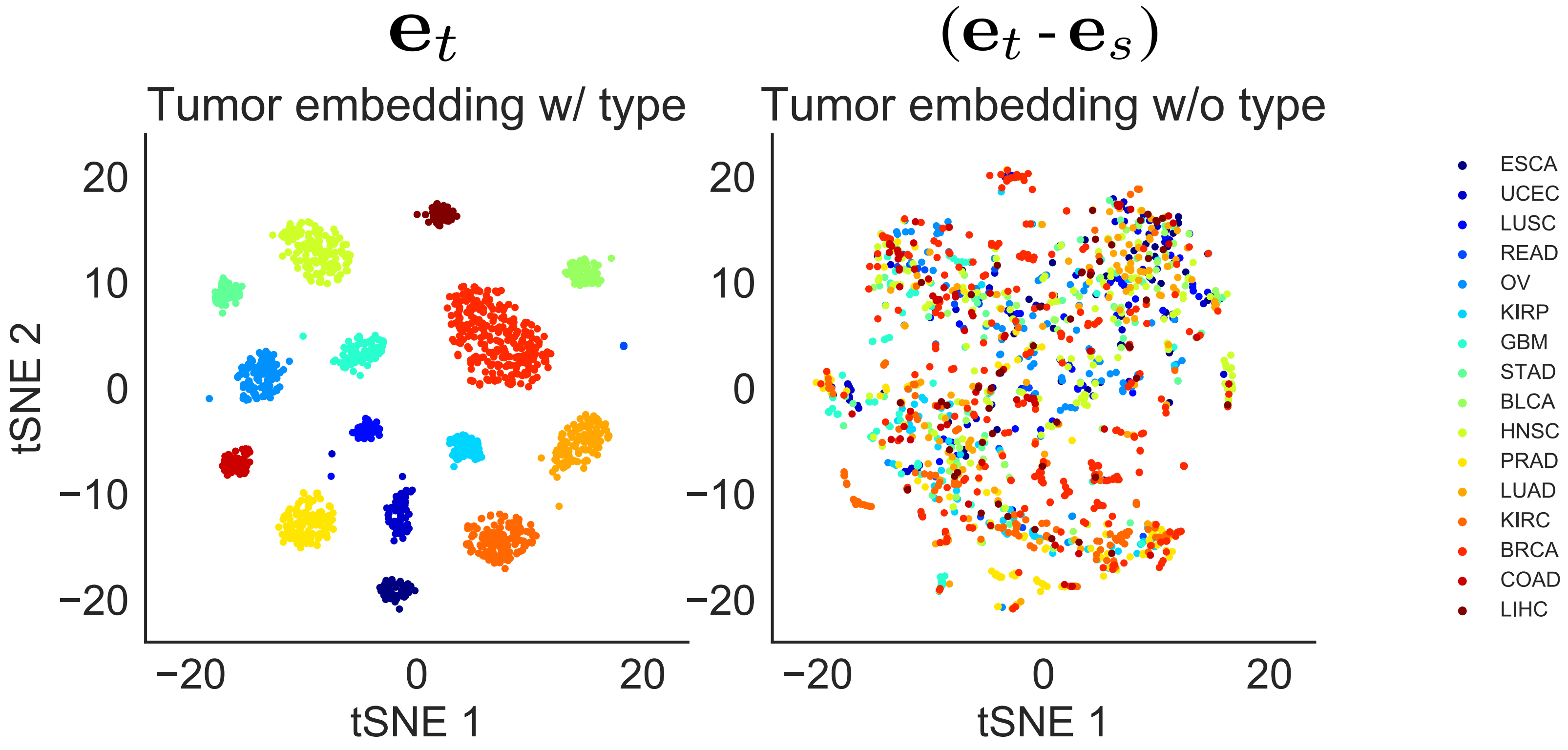


Candidate Drivers via Attention

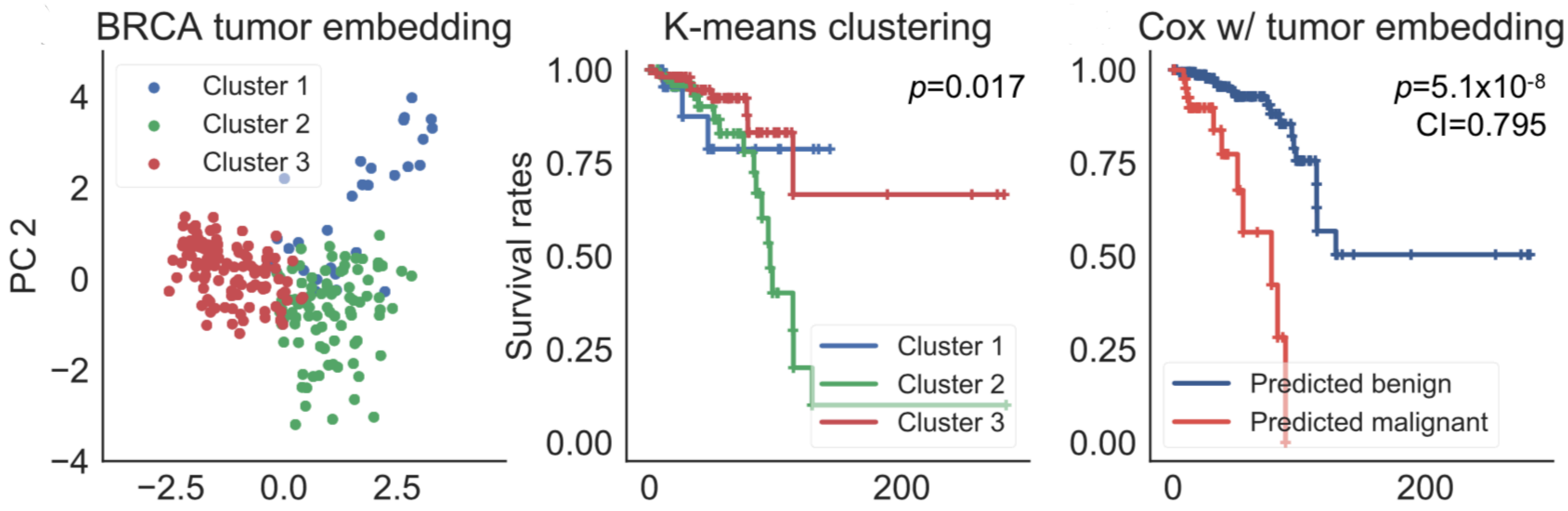


Tumor Embedding

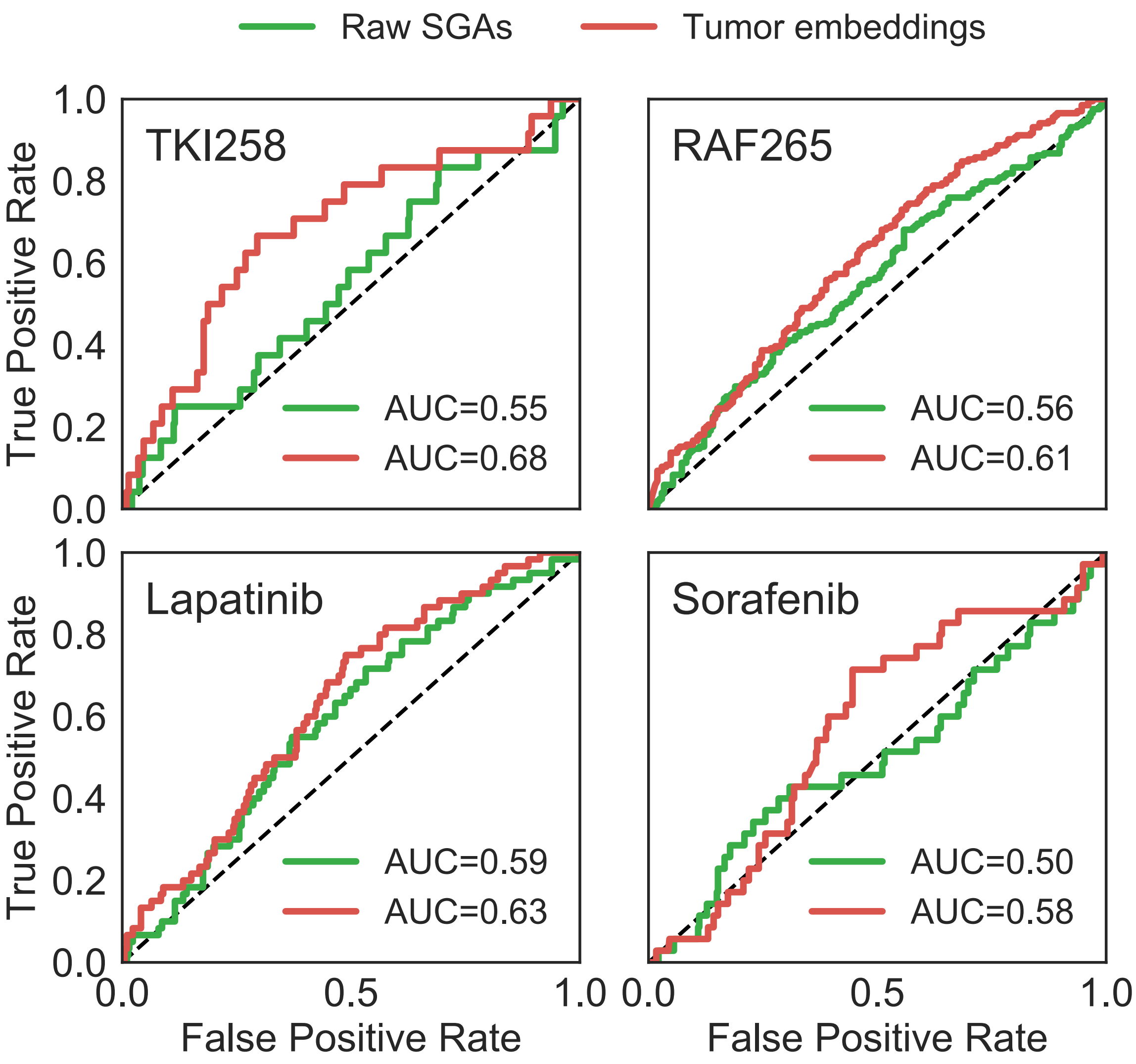
- Common cellular signaling process across cancer types



Application - Survival Analysis



Application – Drug Response



Summary

- Biological-inspired neural network to mimic cellular signaling
- Distinguish drivers from passengers with supervision of expression
- Gene embedding: informative of gene functions
- Tumor embedding: transferable to other phenotype prediction tasks
- Gene2Vec: speed up training and alleviate overfitting

Acknowledgements

- Lu Lab
  - Xinghua Lu
  - Chunhui Cai
  - Yifan Xue
  - Michael Q. Ding
- Cohen Lab
  - William W. Cohen
- Funding
  - NIH
  - Pennsylvania Department of Health