

Predicting Cancer Phenotypes from Somatic Genomic Alterations via Genomic Impact Transformer

Yifeng Tao¹, Chunhui Cai², William W. Cohen^{1,*}, Xinghua Lu^{2,3,*}

¹School of Computer Science, Carnegie Mellon University

²Department of Biomedical Informatics, University of Pittsburgh

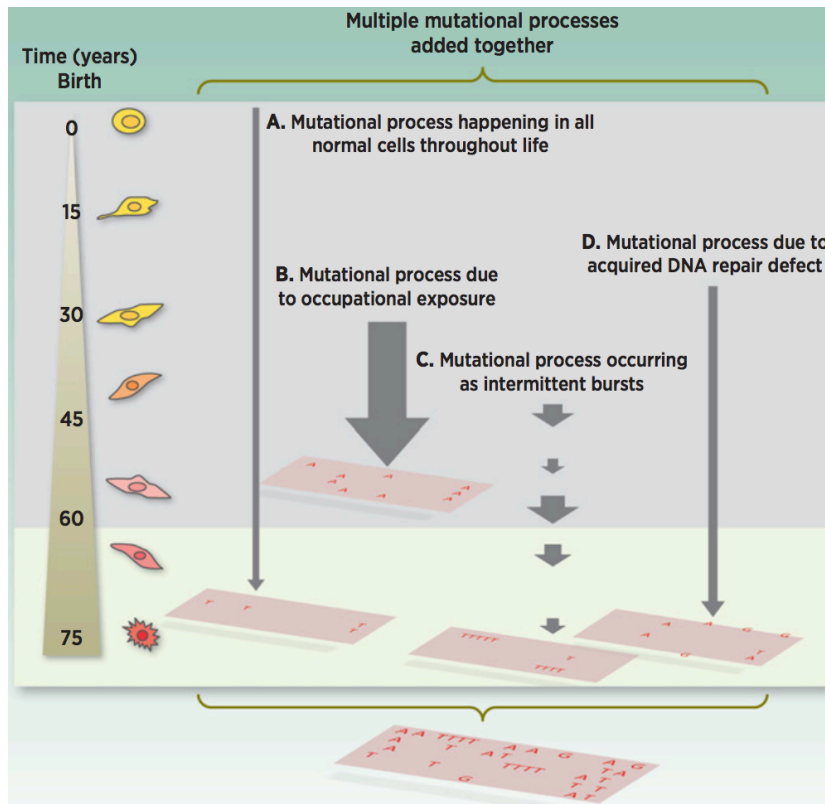
³Department of Pharmaceutical Sciences, School of Pharmacy, University of Pittsburgh

**Carnegie
Mellon
University**



Tumor origin and progression

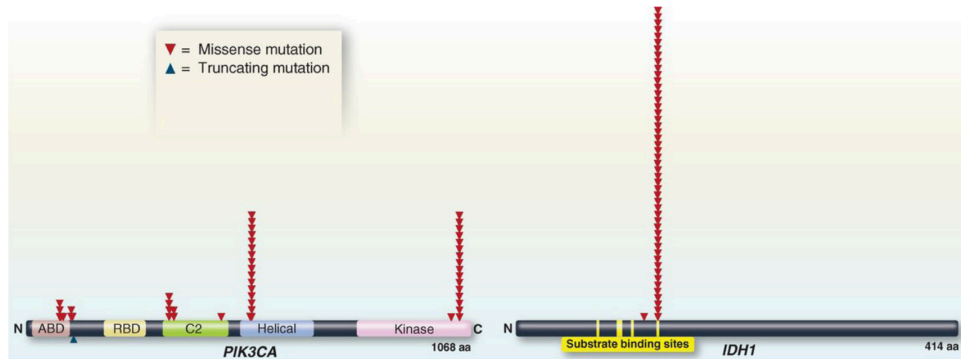
- Cancers are mainly caused by somatic genomic alterations (SGAs)
 - Driver SGAs (~10s/tumor): Promote tumor progression
 - Passenger SGAs (~100s/tumor): Neutral mutations
 - How to distinguish drivers from passengers?



S Nik-Zainal et al. 2017

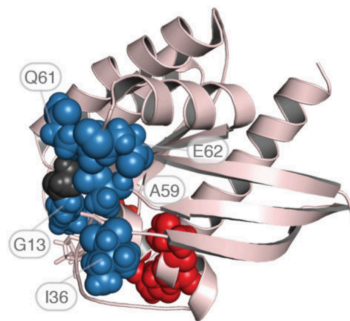
Cancer drivers

- How to distinguish drivers from passengers?
 - Frequency: recurrent mutations more likely to be drivers



B Vogelstein et al. 2013
ND Dees et al. 2012
MS Lawrence et al. 2013

- Conserved domain: protein function significantly disturbed

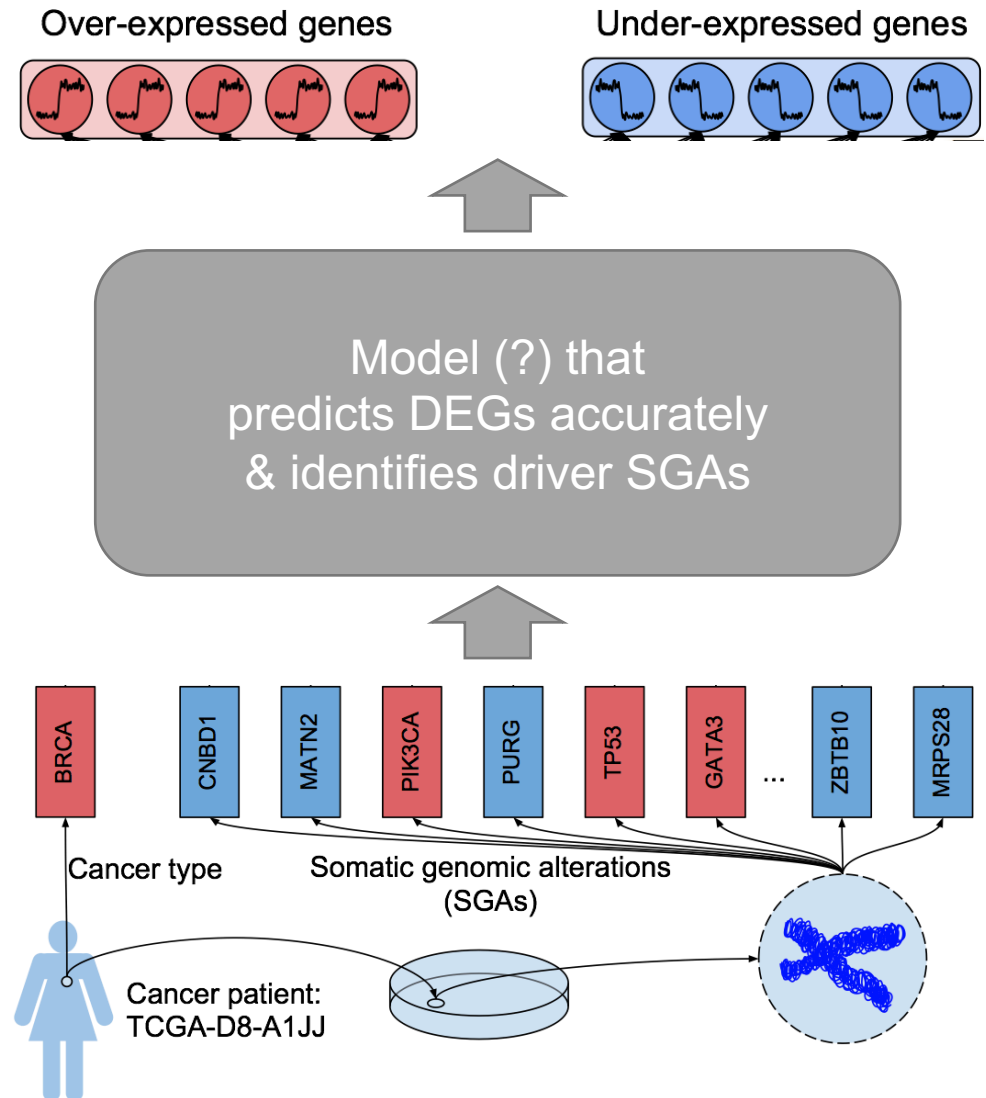


B Reva et al. 2011
B Niu et al. 2016

- All unsupervised. But drivers are defined as mutations that promote to tumor development...

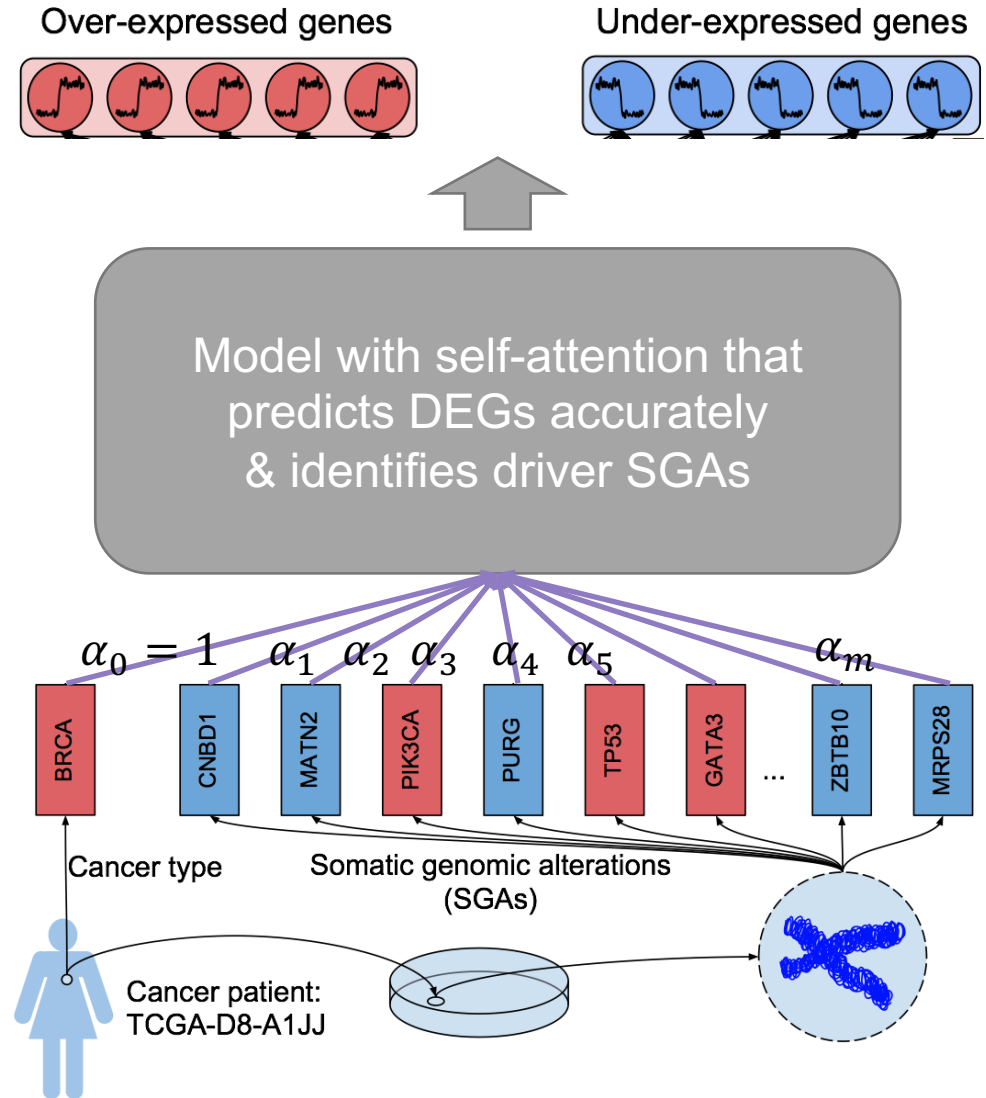
Cancer drivers

- Identify driver SGAs with supervision of downstream phenotypes
 - Change of RNA expression
 - Differentially expressed genes (DEGs)
- Candidate models
 - Bayesian model (C Cai et al. 2019)
 - Lasso/Elastic net (R Tibshirani 1994)
 - Multi-layer perceptrons (MLPs) (F Rosenblatt 1958)
 - Models do prediction & driver detection?



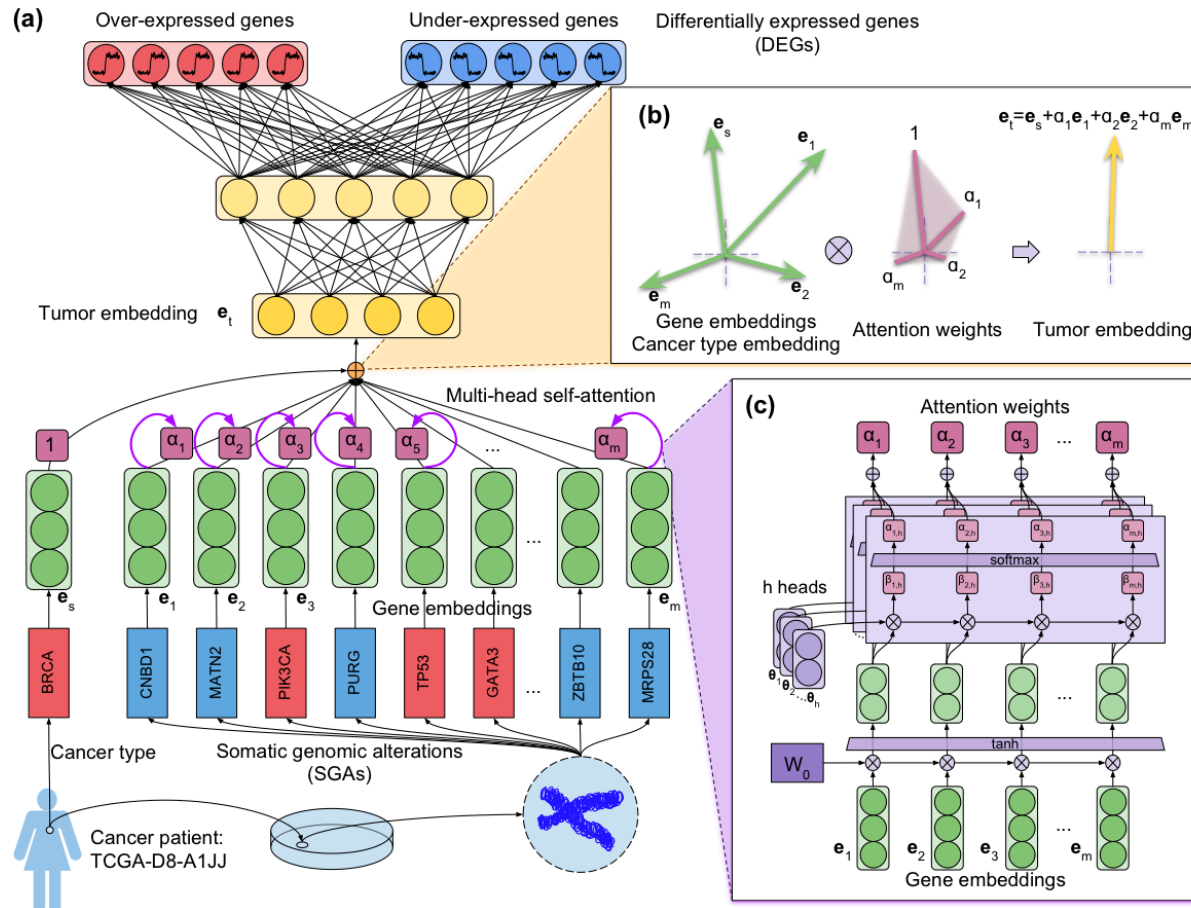
Self-attention mechanism

- Models do prediction & driver detection?
- Attention mechanism
 - Initially in CV (K Xu et al. 2015)/NLP (A Vaswani et al. 2017)
 - Better interpretability
 - Improves performance
- Self-attention mechanism (Z Yang et al. 2016)
 - Contextual deep learning framework: weights determined by all the input mutations



Genomic impact transformer (GIT)

- Transformer: encoder-decoder architecture
- Encoder: self-attention mechanism; Decoder: MLP



Encoder: Multi-head self-attention

- Tumor embedding is the weighted sum of gene embeddings:

$$\mathbf{e}_t = 1 \cdot \mathbf{e}_s + \sum_g \alpha_g \cdot \mathbf{e}_g$$

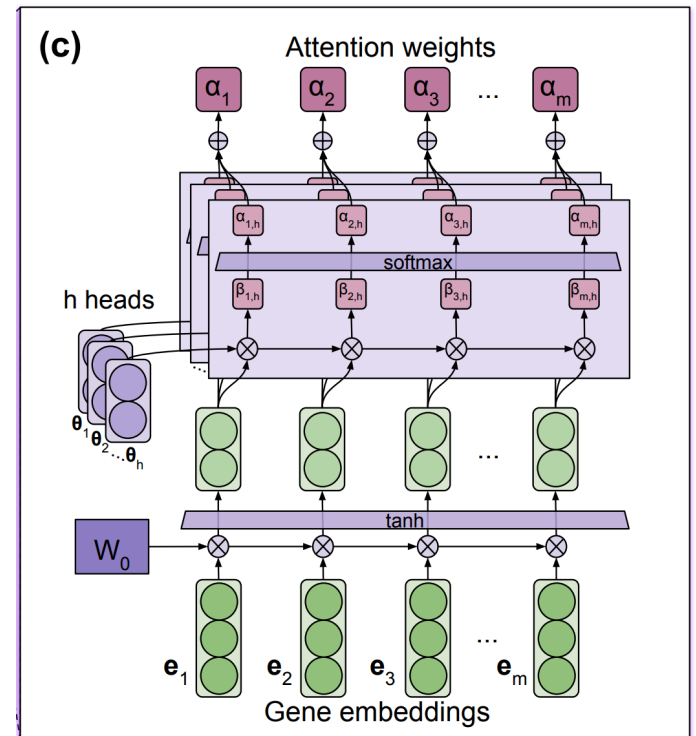
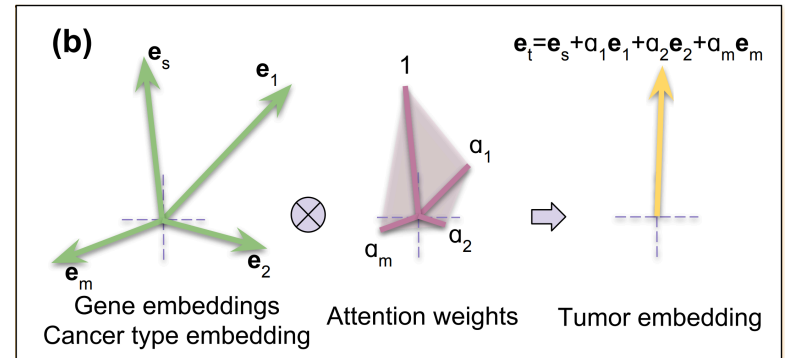
- Weights determined by input gene embeddings:

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

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

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

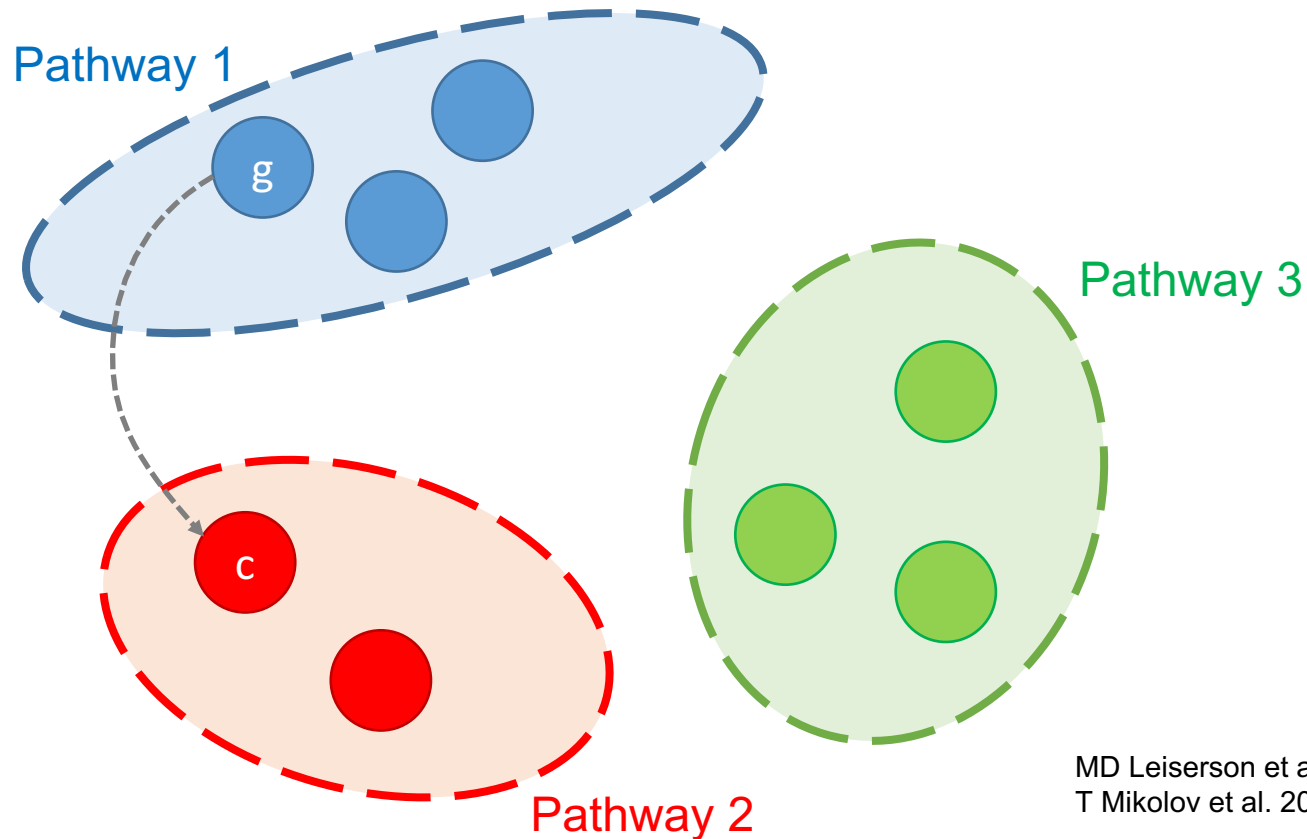
$$\beta_{g,j} = \boldsymbol{\theta}_j^T \cdot \tanh(W_0 \cdot \mathbf{e}_g), \quad g = 1, 2, \dots, m$$



Pre-training gene embedding: Gene2Vec

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

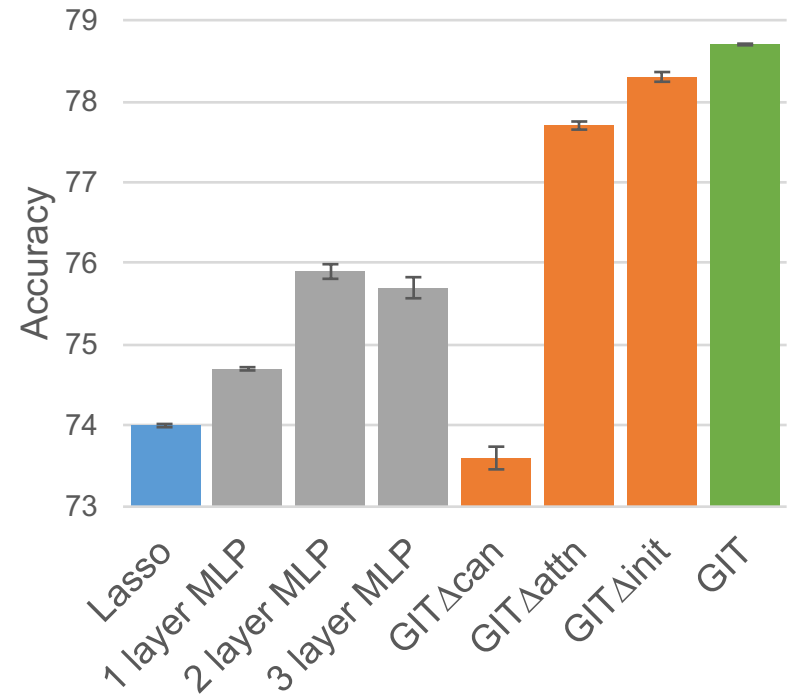
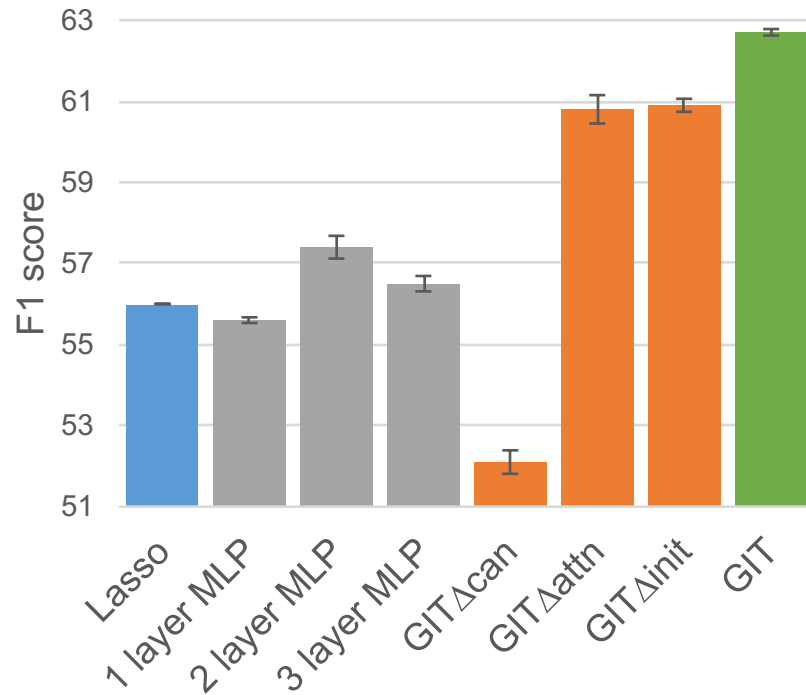
$$\Pr(c \in \text{Context}(g) \mid g) = \frac{\exp(\mathbf{e}_g^T \mathbf{v}_c)}{\sum_{c' \in \mathcal{G}} \exp(\mathbf{e}_g^T \mathbf{v}_{c'})}$$



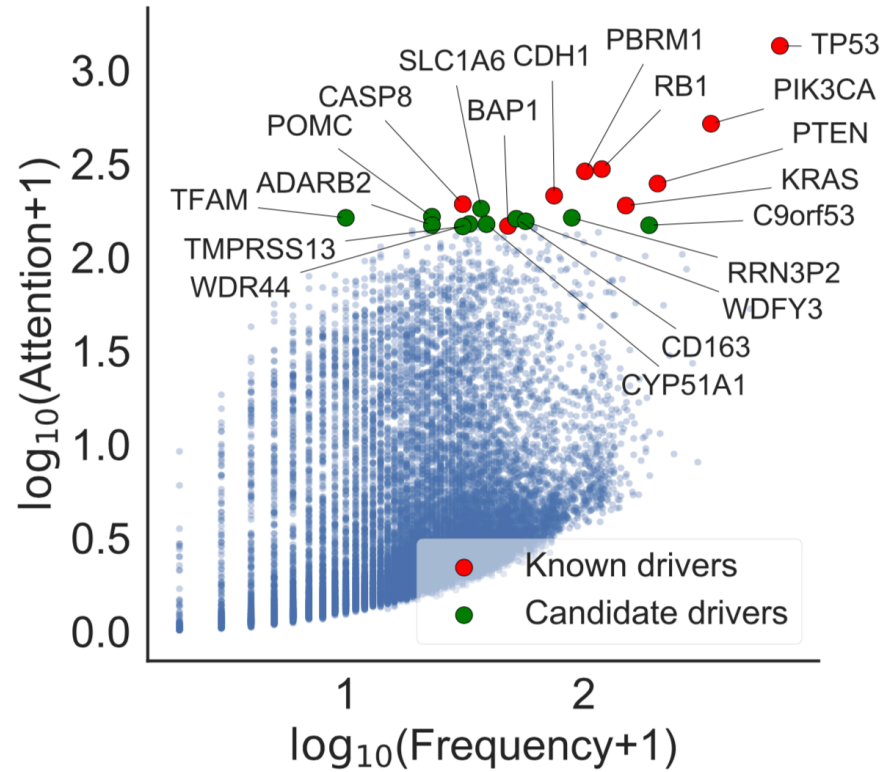
MD Leiserson et al. 2015
T Mikolov et al. 2013

Improved performance in predicting DEGs

- Predicting DEGs from SGAs
 - Conventional models
 - Ablation studies

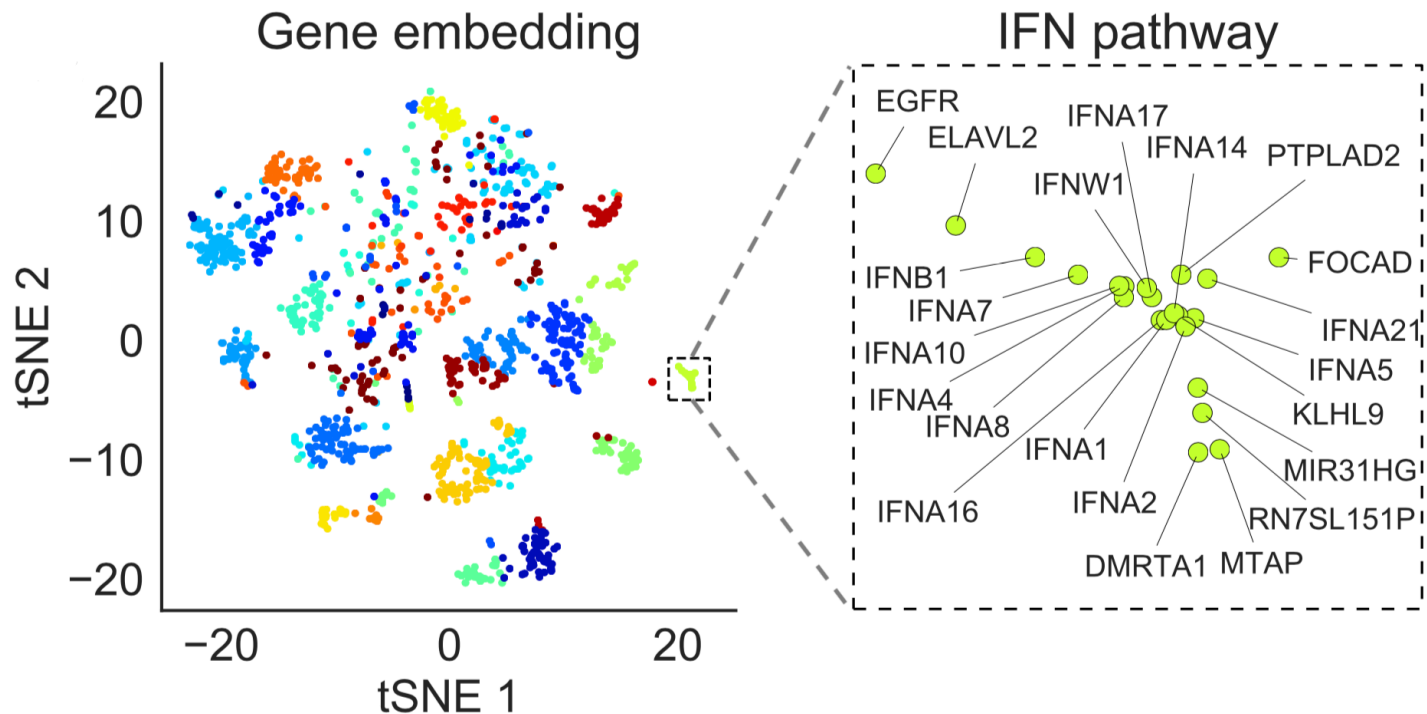


Candidate drivers via attention mechanism



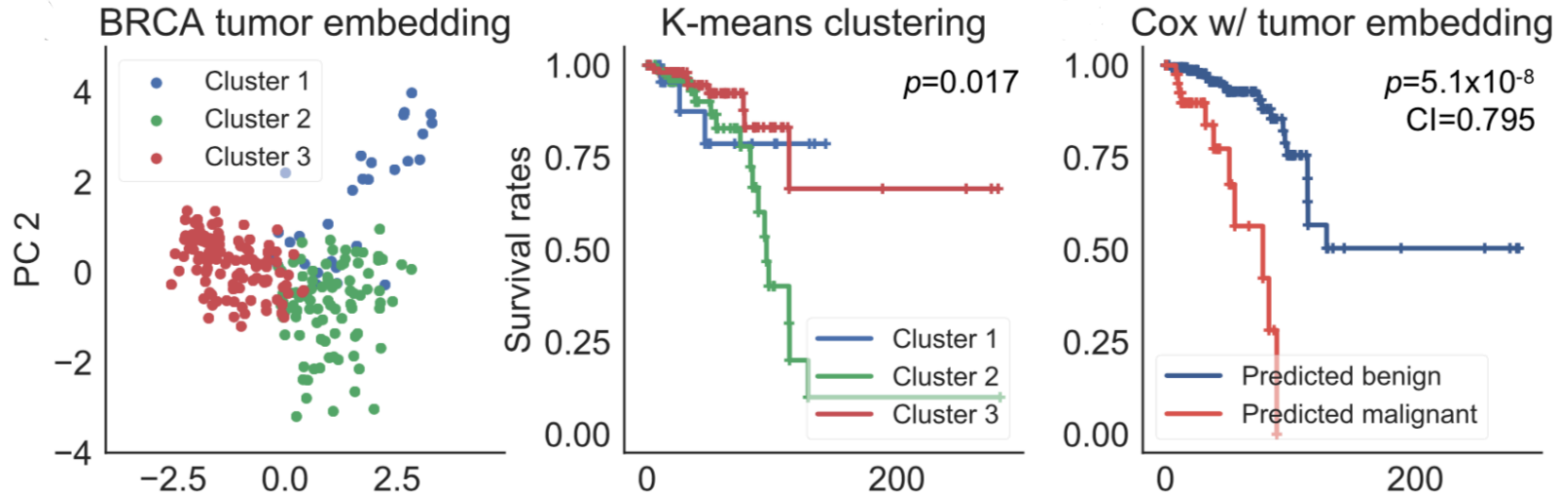
Gene embedding space

- Functionally similar genes are close in gene embedding space
 - Qualitatively and quantitatively (i.e., GO enrichment, NN accuracy)



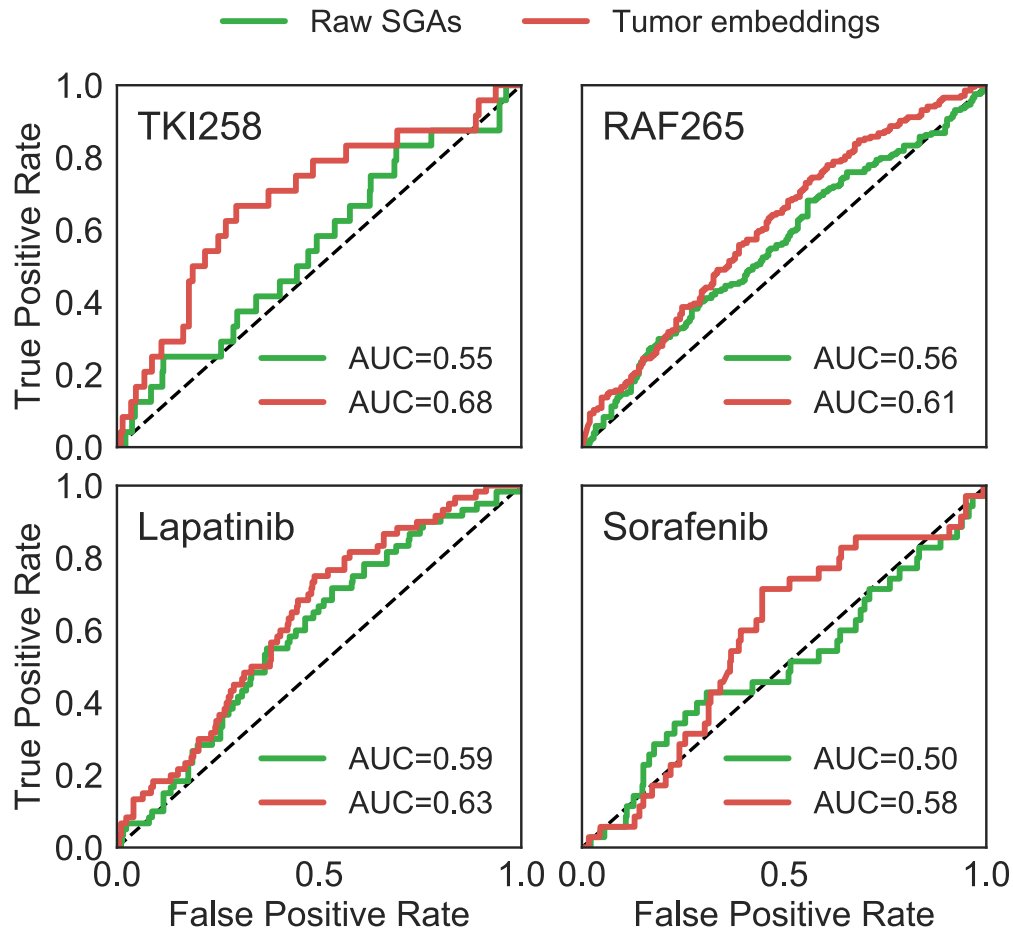
Tumor embedding: Survival analysis

- Tumor embeddings reveal distinct survival profiles



Tumor embedding: Drug response

- Tumor embeddings are predictive of drug response



Conclusions and future work

- Biologically inspired neural network framework
 - Identifying cancer drivers with supervision of DEGs
 - Accurate prediction of DEGs from mutations
- Side products
 - Gene embedding: informative of gene functions
 - Tumor embedding: transferable to other phenotype prediction tasks
- Code and pretrained gene embedding:
<https://github.com/yifengtao/genome-transformer>
- Future work
 - Fine-grained embedding representation in codon level
 - Tumor evolutionary features, e.g., hypermutability, intra-tumor heterogeneity

Acknowledgments

- Dr. Xinghua Lu
- Dr. William W. Cohen
- Dr. Chunhui Cai
- Michael Q. Ding
- Yifan Xue



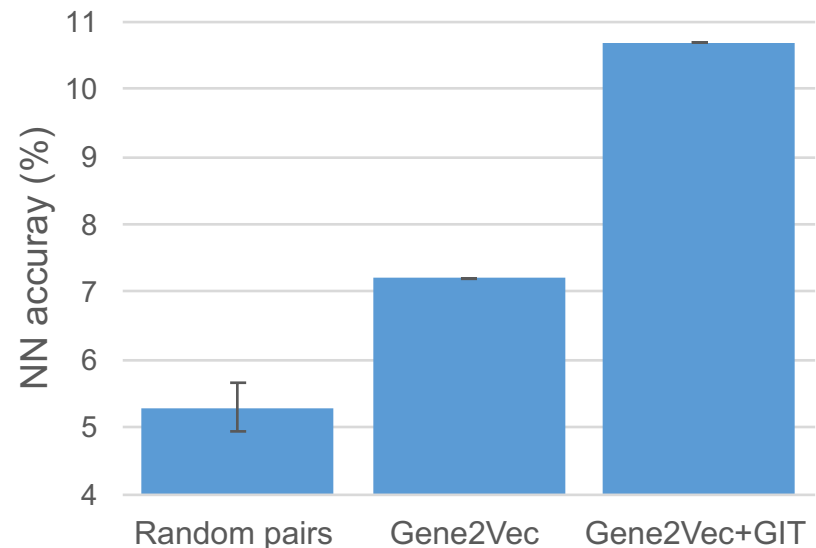
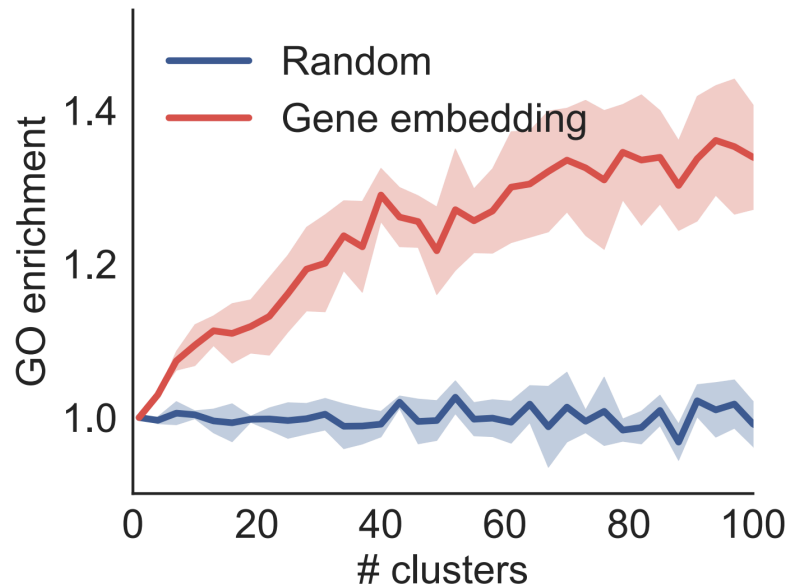
Quantitative measurement of gene embeddings

- Functional similar genes \rightarrow closer in embedding space
 - Go enrichment:

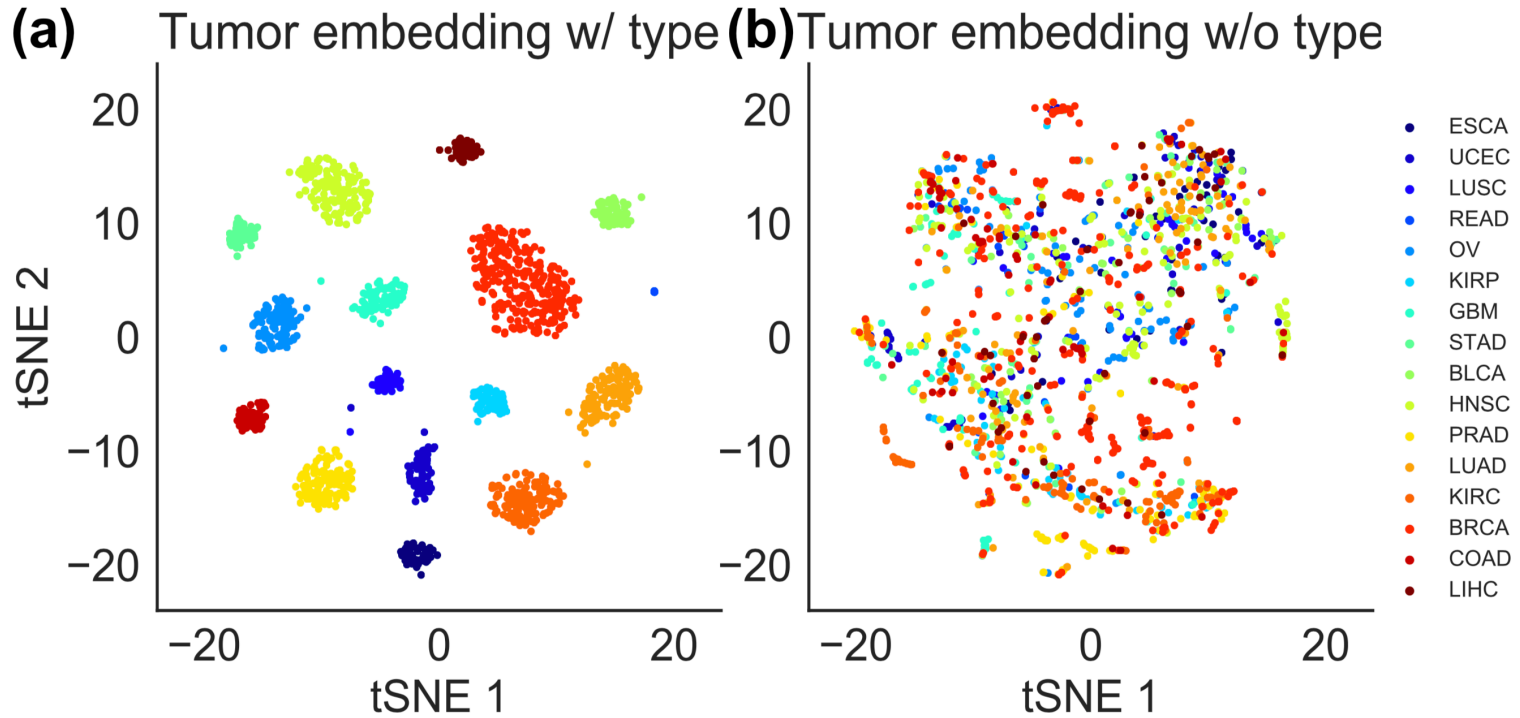
$$\text{enrichment} = \frac{\mathbb{E}_{\text{Clust}(\mathbf{e}_g) = \text{Clust}(\mathbf{e}_c)} [\mathbb{1}(\text{GO}(g) \cap \text{GO}(c) \neq \emptyset)]}{\mathbb{E}_{g, c \in \mathcal{G}} [\mathbb{1}(\text{GO}(g) \cap \text{GO}(c) \neq \emptyset)]}$$

- NN accuracy:

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



Tumor embedding space



Gene2Vec algorithm

Data: Genomic alterations in each tumor: $\mathcal{T} = \{T_i = \{g_{i1}, g_{i2}, \dots, g_{im(i)}\}\}_{i=1,2,\dots,N}$.

Result: Pretrained gene embedding of each gene:
 $\mathcal{E} = \{\mathbf{e}_g \in \mathbb{R}^n\}_{g \in \mathcal{G}}$.

Context gene embeddings:
 $\mathcal{V} = \{\mathbf{v}_g \in \mathbb{R}^n\}_{g \in \mathcal{G}}$.

$f(g) \leftarrow \frac{1}{Z} \sum_{i=1}^N \mathbb{1}(g \in T_i), g \in \mathcal{G};$ // Gene frequency

$f_n(g) \leftarrow \frac{1}{Z^n} f(g)^{3/4}, g \in \mathcal{G};$ // Normalized frequency

$\mathbf{e}_g \sim U\left(-\frac{0.5}{n}, \frac{0.5}{n}\right)^n, \mathbf{v}_g \leftarrow \mathbf{0}^n, g \in \mathcal{G};$ // Initialize gene embeddings and context embeddings

while *not converges* **do**

$l \leftarrow 0;$ // Total loss of a mini-batch samples

for $b = 1, 2, \dots, \text{batch_size}$ **do**

$g \sim f;$ // Sample a gene

$g_c \sim \text{Context}(g; \mathcal{T});$ // Sample a context gene

$g_{nr} \sim f_n, r = 1, 2, \dots, R;$ // Sample negative context genes

$l \leftarrow l + \text{NSLoss}(g, g_c, \{g_{nr}\}_{r=1}^R; \mathcal{E}, \mathcal{V});$ // Update

end

$(\mathcal{E}, \mathcal{V}) \leftarrow (\mathcal{E}, \mathcal{V}) - \eta \cdot \frac{\partial l}{\partial (\mathcal{E}, \mathcal{V})};$ // Gradient descent

end

Function $\text{Context}(g; \mathcal{T})$

$P_c \leftarrow U(\{g_c \mid g_c \in T_i, g \in T_i\}_{i=1,2,\dots,N});$ // Uniform distribution on sequence of adjacent mutations

return P_c

Function $\text{NSLoss}(g, g_c, \{g_{nr}\}_{r=1}^R; \mathcal{E}, \mathcal{V})$

$l \leftarrow \log \sigma(\mathbf{e}_g^\top \mathbf{v}_{g_c}) + \sum_{r=1}^R \log \sigma(-\mathbf{e}_g^\top \mathbf{v}_{g_{nr}});$ // Negative sampling loss of one sample

return l

Gene2Vec: Co-occurrence patterns

- Co-occurrence does not necessarily mean similar embeddings
 - Ex 1: two cats sit there .
 - Ex 2: two cats stand there .
 - Ex 3: two dogs sit there .

