Pacific Symposium on Biocomputing 25:79-90(2020)

# From genome to phenome: Predicting multiple cancer phenotypes based on somatic genomic alterations via the genomic impact transformer

Yifeng Tao[1], Chunhui Cai[2], William W. Cohen[1,†] and Xinghua Lu[2,3,†]

[1]*School of Computer Science, Carnegie Mellon University,*
[2]*Department of Biomedical Informatics, University of Pittsburgh,*
[3]*Department of Pharmaceutical Sciences, School of Pharmacy, University of Pittsburgh, Pittsburgh, PA, USA*
[†]*To whom correspondence should be addressed. E-mail: wcohen@cs.cmu.edu, xinghua@pitt.edu*

Cancers are mainly caused by somatic genomic alterations (SGAs) that perturb cellular signaling systems and eventually activate oncogenic processes. Therefore, understanding the functional impact of SGAs is a fundamental task in cancer biology and precision oncology. Here, we present a deep neural network model with encoder-decoder architecture, referred to as genomic impact transformer (GIT), to infer the functional impact of SGAs on cellular signaling systems through modeling the statistical relationships between SGA events and differentially expressed genes (DEGs) in tumors. The model utilizes a multi-head self-attention mechanism to identify SGAs that likely cause DEGs, or in other words, differentiating potential driver SGAs from passenger ones in a tumor. GIT model learns a vector (gene embedding) as an abstract representation of functional impact for each SGA-affected gene. Given SGAs of a tumor, the model can instantiate the states of the hidden layer, providing an abstract representation (tumor embedding) reflecting characteristics of perturbed molecular/cellular processes in the tumor, which in turn can be used to predict multiple phenotypes. We apply the GIT model to 4,468 tumors profiled by The Cancer Genome Atlas (TCGA) project. The attention mechanism enables the model to better capture the statistical relationship between SGAs and DEGs than conventional methods, and distinguishes cancer drivers from passengers. The learned gene embeddings capture the functional similarity of SGAs perturbing common pathways. The tumor embeddings are shown to be useful for tumor status representation, and phenotype prediction including patient survival time and drug response of cancer cell lines.*

*Keywords*: Neural networks; Knowledge representation; Gene regulatory networks; Cancer.

## 1. Introduction

Cancer is mainly caused by the activation of oncogenes or deactivation of tumor suppressor genes (collectively called "driver genes") as results of somatic genomic alterations (SGAs),[1] including somatic mutations (SMs),[2,3] somatic copy number alterations (SCNAs),[4,5] DNA structure variations (SVs),[6] and epigenetic changes.[7] Precision oncology relies on the capability of identifying and targeting tumor-specific aberrations resulting from driver SGAs and

*Supplemental information (SI), GIT model, pre-processed TCGA data, and gene embeddings are available at https://github.com/yifengtao/genome-transformer.

© 2019 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

their effects on molecular and cellular phenotypes. However, our knowledge of driver SGAs and cancer pathways remains incomplete. Particularly, it remains a challenge to determine which SGAs (among often hundreds) in a specific tumor are drivers, which cellular signals or biological processes a driver SGA perturbs, and which molecular/cellular phenotypes a driver SGA affects.

Current methods for identifying driver genes mainly concentrate on identifying genes that are mutated at a frequency above expectation, based on the assumption that mutations in these genes may provide oncogenic advantages and thus are positively selected.[8,9] Some works further focus on the mutations perturbing conserved (potentially functional) domains of proteins as indications they may be driver events.[10,11] However, these methods do not provide any information regarding the functional impact of enriched mutations on molecular/cellular phenotypes of cells. Without the knowledge of functional impact, it is difficult to further determine whether an SGA will lead to specific molecular, cellular and clinical phenotypes, such as response to therapies. What's more, while both SMs and SCNAs may activate/deactivate a driver gene, there is no well-established frequency-based method that combines different types of SGAs to determine their functional impact.

Conventionally, an SGA event perturbing a gene in a tumor is represented as a "one-hot" vector spanning gene space, in which the element corresponding to the perturbed gene is set to "1". This representation simply indicates which gene is perturbed, but it does not reflect the functional impact of the SGA, nor can it represent the similarity of distinct SGAs that perturb a common signaling pathway. We conjecture that it is possible to represent an SGA as a low-dimensional vector, in the same manner as the "word embedding"[12–14] in the natural language processing (NLP) field, such that the representation reflects the functional impact of a gene on biological systems, and genes sharing similar functions should be closely located in such embedding space. Here the "similar function" is broadly defined, e.g., genes from the same pathway or of the same biological process.[15] Motivated by this, we propose a scheme for learning "gene embeddings" for SGA-affected genes, i.e., a mapping from individual genes to low-dimensional vectors of real numbers that are useful in multiple prediction tasks.

Based on the assumption that SGAs perturbing cellular signaling systems often eventually lead to changes in gene expression,[16] we introduce an encoder-decoder architecture neural network model called "genomic impact transformer" (GIT) to predict DEGs and detect potential cancer drivers with the supervision of DEGs. While deep learning models are being increasingly used to model different bioinformatics problems,[17,18] to our knowledge there are few studies using the neural network to model the relationships between SGAs and molecular/cellular phenotypes in cancers. The proposed GIT model has the following innovative characteristics: (1) The encoder part of the transformer[19] first uses SGAs observed in a tumor as inputs, maps each SGA into a gene embedding representation, and combines gene embeddings of SGAs to derive a personalized "tumor embedding". Then the decoder part decodes and translates the tumor embedding to DEGs. (2) A multi-head self-attention mechanism[20,21] is utilized in the encoder, which is a technique widely used in NLP to choose the input features that significantly influence the output. It differentiates SGAs by assigning different weights to them so that it can potentially distinguish SGAs that have an impact on DEG from those do

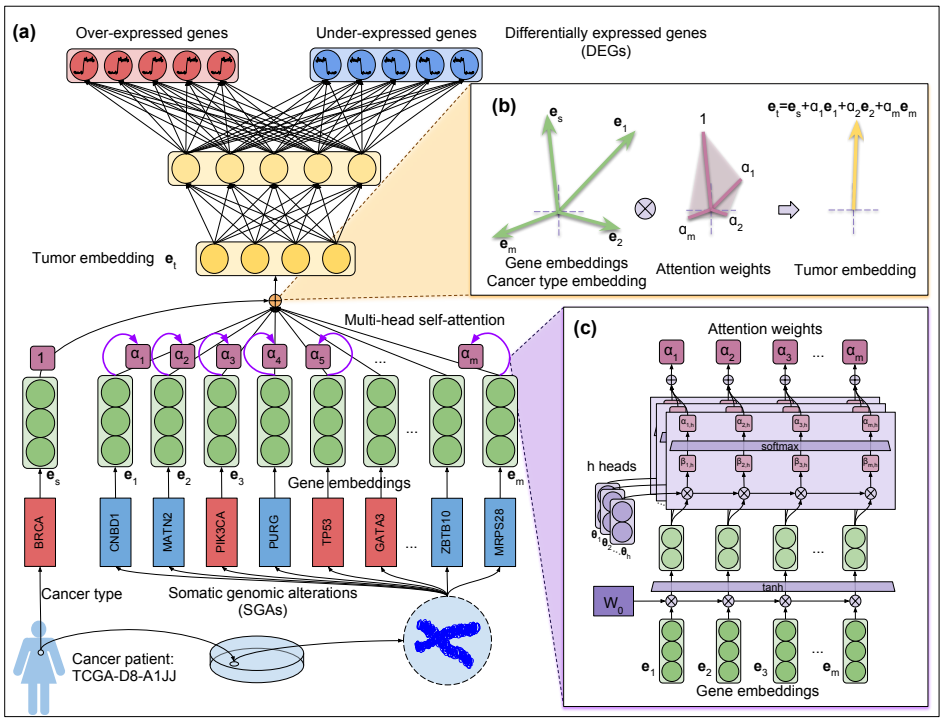


Fig. 1. **(a)** Overall architecture of GIT. An example case and its detected drivers are shown. **(b)** A two-dimensional demo that shows how attention mechanism combines multiple gene embeddings of SGAs $\{\mathbf{e}_g\}_{g=1}^{m}$ and cancer type embedding $\mathbf{e}_s$ into a tumor embedding vector $\mathbf{e}_t$ using attention weights $\{\alpha_g\}_{g=1}^{m}$. **(c)** Calculation of attention weights $\{\alpha_g\}_{g=1}^{m}$ using gene embeddings $\{\mathbf{e}_g\}_{g=1}^{m}$.

not, i.e., detecting drivers from passengers. (3) Pooling inferred weighted impact of SGAs in a tumor produces a personalized tumor embedding, which can be used as an effective feature to predict DEGs and other phenotypes. (4) Gene embeddings are pre-trained by a "Gene2Vec" algorithm and further refined by the GIT, which captures the functional impact of SGAs on the cellular signaling system. Our results and analysis indicate that above innovative approaches enable us to derive powerful gene embedding and tumor embedding representations that are highly informative of molecular, cellular and clinical phenotypes.

## 2. Materials and methods

### 2.1. *SGAs and DEGs pre-processing*

We obtained SGA data, including SMs and SCNAs, and DEGs of 4,468 tumors consisting of 16 cancer types directly from TCGA portal.[22] Details available in SI (Sec. S1).

### 2.2. *The GIT neural network*

#### 2.2.1. *GIT network structure: encoder-decoder architecture*

Figure 1a shows the general structure of the GIT model with an overall encoder-decoder architecture. GIT mimics hierarchically organized cellular signaling system,[23,24] in which a

neuron may potentially encode the signal of one or more signaling proteins. When a cellular signaling system is perturbed by SGAs, it often can lead to changes in measured molecular phenotypes, such as gene expression changes. Thus, for a tumor $t$, the set of its SGAs $\{g\}_{g=1}^{m}$ is connected to the GIT neural network as observed input (Fig. 1a bottom part squares). The impact of SGAs is represented as embedding vectors $\{\mathbf{e}_g\}_{g=1}^{m}$, which are further linearly combined to produce a tumor embedding vector $\mathbf{e}_t$ through an attention mechanism in the encoder (Fig. 1a middle part). We explicitly represent cancer type $s$ and its influence on encoding system $\mathbf{e}_s$ of the tumor because tissue type influences which genes are expressed in cells of specific tissue as well. Finally, the decoder module, which consists of a feed-forward multi-layer perceptron (MLP),[25] transforms the functional impact of SGAs and cancer type into DEGs of the tumor (Fig. 1a top part).

#### 2.2.2. *Pre-training gene embeddings using Gene2Vec algorithm*

In this study, we projected the discrete binary representation of SGAs perturbing a gene into a continuous embedding space, which we call "gene embeddings" of corresponding SGAs, using a "Gene2Vec" algorithm, based on the assumption of co-occurrence pattern of SGAs in each tumor, including mutually exclusive patterns of mutations affecting a common pathway.[26] These gene embeddings were further updated and fine-tuned by the GIT model with the supervision of affected DEGs. Algorithm details available in SI (Sec. S2).

#### 2.2.3. *Encoder: multi-head self-attention mechanism*

To detect the difference of functional impact of SGAs in a tumor, we designed a multi-head self-attention mechanism (Fig. 1a middle part). For all SGA-affected genes $\{g\}_{g=1}^{m}$ and the cancer type $s$ of a tumor $t$, we first mapped them to corresponding gene embeddings $\{\mathbf{e}_g\}_{g=1}^{m}$ and a cancer type embedding $\mathbf{e}_s$ from a look-up table $\mathcal{E} = \{\mathbf{e}_g\}_{g\in\mathcal{G}} \cap \{\mathbf{e}_s\}_{s\in\mathcal{S}}$, where $\mathbf{e}_g$ and $\mathbf{e}_s$ are real-valued vectors. From the implementation perspective, we treated cancer types in the same way as SGAs, except the attention weight of it is fixed to be "1". The overall idea of producing the tumor embedding $\mathbf{e}_t$ is to use the weighted sum of cancer type embedding $\mathbf{e}_s$ and gene embeddings $\{\mathbf{e}_g\}_{g=1}^{m}$ (Fig. 1b) :

$$\mathbf{e}_t = 1 \cdot \mathbf{e}_s + \sum\nolimits_g \alpha_g \cdot \mathbf{e}_g = 1 \cdot \mathbf{e}_s + \alpha_1 \cdot \mathbf{e}_1 + ... + \alpha_m \cdot \mathbf{e}_m. \tag{1}$$

The attention weights $\{\alpha_g\}_{g=1}^{m}$ were calculated by employing multi-head self-attention mechanism, using gene embeddings of SGAs $\{\mathbf{e}_g\}_{g=1}^{m}$ in the tumor: $\{\alpha_g\}_{g=1}^{m} = \text{Function}_{\text{Attention}}\left(\{\mathbf{e}_g\}_{g=1}^{m}; W_0, \Theta\right)$ (Fig. 1c). See SI (Sec. S3) for mathematical details. Overall we have three parameters $\{W_0, \Theta, \mathcal{E}\}$ to train in the multi-head attention module using back-propagation.[27] The look-up table $\{\mathbf{e}_g\}_{g\in\mathcal{G}}$ was initialized with Gene2Vec pre-trained gene embeddings and refined by GIT here.

#### 2.2.4. *Decoder: multi-layer perceptron (MLP)*

For a specific tumor $t$, we fed tumor embedding $\mathbf{e}_t$ into an MLP with one hidden layer as the decoder, using non-linear activation functions and fully connected layers, to produce the final

predictions $\hat{y}$ for DEGs $y$; (Fig. 1a top part):

$$\hat{y} = \sigma(W_2 \cdot \mathrm{ReLU}(W_1 \cdot \mathrm{ReLU}(\mathbf{e}_t) + b_1) + b_2). \tag{2}$$

where $\mathrm{ReLU}(x) = \max(0, x)$ is rectified linear unit, and $\sigma(x) = (1+\exp(-x))^{-1}$ is sigmoid activation function. The output of the decoder and actual values of DEGs were used to calculate the $\ell_2$-regularized cross entropy, which was minimized during training: $\min_{\mathcal{W},\mathcal{E},\Theta,b} \mathrm{CrossEnt}(y, \hat{y}) + \ell_2(\mathcal{W}, \mathcal{E}, \Theta; \lambda_2)$, where $\mathcal{W} = \{W_l\}_{l=0}^{2}$, cross entropy loss defined as $\mathrm{CrossEnt}(y, \hat{y}) = -\sum_i [(1 - y_i)\log(1 - \hat{y}_i) + y_i \log \hat{y}_i]$, and $\ell_p$ regularizer defined as $\ell_p(\mathcal{W}; \lambda) = \lambda \cdot \sum_l \|W_l\|_p, p \in \{1, 2\}$.

### 2.3. *Training and evaluation*

We utilized PyTorch (https://pytorch.org/) to train, validate and test the Gene2Vec, GIT (variants) and other conventional models (Lasso and MLPs; Section 3.1). The training, validation and test sets were split in the ratio of 0.33:0.33:0.33 and fixed across different models. The hyperparameters were tuned over the training and validation sets to get best F1 scores, trained on training and validation sets, and finally applied to the test set for evaluation if not further mentioned below. The models were trained by updating parameters using backpropagation,[27] specifically, using mini-batch Adam[28] with default momentum parameters. Gene2Vec used mini-batch stochastic gradient descent (SGD) instead of Adam. Dropout[29] and weight decay ($\ell_p$-regularization) were used to prevent overfitting. We trained all the models 30 to 42 epochs until they fully converged. The output DEGs were represented as a sparse binary vector. We utilized various performance metrics including accuracy, precision, recall, and F1 score, where F1 is the harmonic mean of precision and recall. The training and test were repeated for five runs get the mean and variance of evaluation metrics. We designed two metrics in the present work for evaluating the functional similarity among genes sharing similar gene embedding: "nearest neighborhood (NN) accuracy" and "GO enrichment". See SI (Sec. S4) for the definition and meaning of them.

## 3. Results

### 3.1. *GIT statistically detects real biological signals*

Table 1. Performances of GIT (variants) and baseline methods.

| Methods | Precision | Recall | F1 score | Accuracy |
|---|---|---|---|---|
| Lasso | 59.6±0.05 | 52.8±0.03 | 56.0±0.01 | 74.0±0.02 |
| 1 layer MLP | 61.9±0.09 | 50.4±0.17 | 55.6±0.07 | 74.7±0.02 |
| 2 layer MLP | 64.2±0.39 | 52.0±0.66 | 57.4±0.28 | 75.9±0.09 |
| 3 layer MLP | 64.2±0.37 | 50.5±0.30 | 56.5±0.19 | 75.7±0.13 |
| GIT - can | 60.5±0.34 | 45.8±0.38 | 52.1±0.29 | 73.6±0.14 |
| GIT - attn | 67.6±0.32 | 55.3±0.77 | 60.8±0.35 | 77.7±0.05 |
| GIT - init | **69.8**±0.28 | 54.1±0.37 | 60.9±0.16 | 78.3±0.06 |
| GIT | 69.5±0.09 | **57.1**±0.18 | **62.7**±0.08 | **78.7**±0.01 |

The task of GIT is to predict DEGs (dependent variables) using SGAs as input (independent variables). Our results of GIT performance on both real and shuffled data demonstrates that GIT is able to capture real statistical relationships between SGAs and DEGs from the noisy biological data (SI: Sec. S5).

As a comparison, we also trained and tested the Lasso

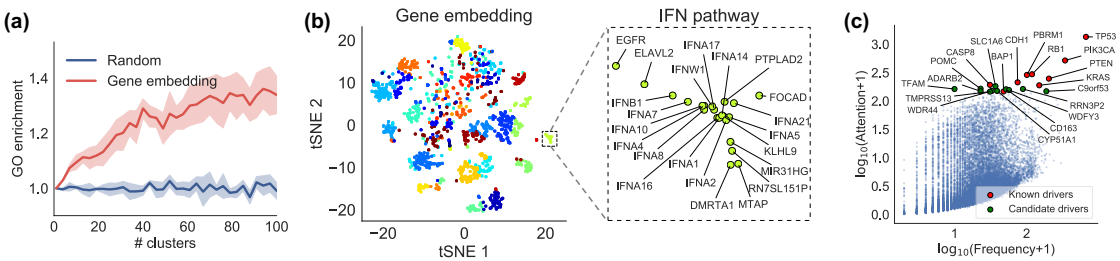


Fig. 2. **(a)** GO enrichment of vs. number of groups in $k$-means clustering. **(b)** t-SNE visualization of gene embeddings. The different colors represent $k$-means (40 clusters) clustering labels. An enlarged inset of a cluster is shown, which contains a set of closely related genes which we refer to "IFN pathway". **(c)** Landscape of attention of SGAs based on attention weights and frequencies.

(multivariate regression with $\ell_1$-regularization)[30] and MLPs[25] as baseline prediction models to predict DEGs based on SGAs. The Lasso model is appealing in our setting because, when predicting a DEG, it can filter out most of the irrelevant input variables (SGAs) and keep only the most informative ones, and it is a natural choice in our case where there are 19.8k possible SGAs. However, in comparison to MLP, it lacks the capability of portraying complex relationships between SGAs and DEGs. On the other hand, while conventional MLPs have sufficient power to capture complex relationships–particularly, the neurons in hidden layers may mimic signaling proteins[24]–they can not utilize any biological knowledge extracted from cancer genomics, nor do they explain the signaling process and distinguish driver SGAs. We employed the precision, recall, F1 score, as well as accuracy to compare GIT and traditional methods (Table 1: 1st to 4th, and last rows). One can conclude that GIT outperforms all these other conventional baseline methods for predicting DEGs in all metrics, indicating the specifically designed structure of GIT is able to soar the performance in the task of predicting DEGs from SGAs.

In order to evaluate the utility of each module (procedure) in GIT, we conducted ablation study by removing one module at a time: the cancer type input ("can"), the multi-head self-attention module ("attn"), and the initialization with pre-trained gene embeddings ("init"). The impact of each module can be detected by comparing to the full GIT model. All the modules in GIT help to improve the prediction of DEGs from SGAs in terms of overall performance: F1 score and accuracy (Table 1: 5th to last rows).

### 3.2. *Gene embeddings compactly represent the functional impact of SGAs*

We examined whether the gene embeddings capture the functional similarity of SGAs, using mainly two metrics: NN accuracy and GO enrichment (Defined in SI Sec. S4). **NN accuracy:** By capturing the co-occurrence pattern of somatic alterations, the Gene2Vec pre-trained gene embeddings improve 36% in NN accuracy over the random chance of any pair of the genes sharing Gene Ontology (GO) annotation[15] (Table 2). The fine-tuned embeddings by GIT further show a one-fold increase in NN accuracy. These results indicate that the learned gene embeddings are consistent with the gene functions, and they map the discrete binary SGA representation into a meaningful and compact space. **GO enrichment:** We performed

clustering analysis of SGAs in embedding space using $k$-means clustering, and calculated GO enrichment, and we varied the number of clusters ($k$) to derive clusters with different degrees of granularity (Fig. 2a). As one can see, when the genes are randomly distributed in the embedding space, they get GO enrichment of 1. However, in the gene embedding space, the GO enrichment increases fast until the number of clusters reaches 40, indicating a strong correlation between the clusters in embedding space and the functions of the genes.

Table 2. NN accuracy with respect to GO in different gene embedding spaces.

| Gene embeddings | NN accuracy | Improvement |
|---|---|---|
| Random pairs | 5.3±0.36 | – |
| Gene2Vec | 7.2 | 36% |
| Gene2Vec + GIT | **10.7** | 100% |

To visualize the manifold of gene embeddings, we grouped the genes into 40 clusters, and conducted the t-SNE[31] of genes (Fig. 2b left panel). Using PANTHER GO enrichment analysis,[32] 12 out of 40 clusters are shown to be enriched in at least one biological process (SI Sec. S6). Most of the gene clusters are well-defined and tight located in the projected t-SNE space. As a case study, we took a close look at one cluster (Fig. 2b right panel), which contains a set of functionally similar genes, such as that code a protein family of type I interferons (IFNs), which are responsible for immune and viral response.[33]

### 3.3. *Self-attention reveals impactful SGAs on cancer cell transcriptome*

Table 3. Top five SGA-affected genes ranked according to attention weight.

| Rank | PANCAN | BRCA | HNSC | LUAD | GBM | BLCA |
|---|---|---|---|---|---|---|
| 1 | ***TP53*** | ***TP53*** | ***TP53*** | ***STK11*** | ***TP53*** | ***TP53*** |
| 2 | ***PIK3CA*** | ***PIK3CA*** | ***CASP8*** | ***TP53*** | ***PTEN*** | ***FGFR3*** |
| 3 | ***RB1*** | ***CDH1*** | ***PIK3CA*** | ***KRAS*** | *C9orf53* | ***RB1*** |
| 4 | ***PBRM1*** | ***GATA3*** | ***CYLD*** | *CYLC2* | ***RB1*** | ***HSP90AA1*** |
| 5 | ***PTEN*** | *MED24* | *RB1* | ***KEAP1*** | *CHIC2* | ***STAG2*** |

While it is widely accepted that cancer is mainly caused by SGAs, but not all SGAs observed in a cancer cell are causative.[1] Previous methods mainly concentrate on searching for SGAs with higher than expected frequency to differentiate candidate drivers SGAs from passenger SGAs. GIT provides a novel perspective to address the problem: identifying the SGAs that have a functional impact on cellular signaling systems and eventually lead DEGs as the *tumor-specific* candidate drivers. Here we compare the relationship of overall attention weights (inferred by GIT model) and the frequencies of somatic alterations (used as the benchmark/control group) in all the cancer types (Pan-Cancer) from our test data (Fig. 2c). In general, the attention weights are correlated with the alteration frequencies of genes, e.g., common cancer drivers such as *TP53* and *PIK3CA* are the top two SGAs selected by both methods.[2] However, our self-attention mechanism assigns high weights to many of genes previously not designated as drivers, indicating these genes are potential cancer drivers although their roles in cancer development remain to be further studied. Table 3 lists top SGAs ranked according to GIT attention weights in pan-cancer and five selected cancer types, where known cancer drivers from TumorPortal[3] and IntOGen[34] are marked as bold font. Apart from *TP53* and *PIK3CA*

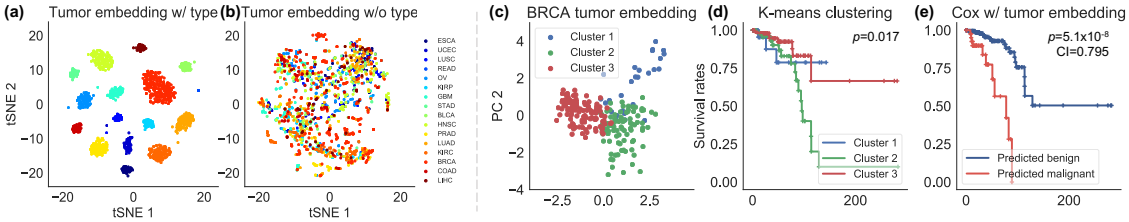


Fig. 3. **(a)** t-SNE of full tumor embedding $\mathbf{e}_t$. **(b)** t-SNE of stratified tumor embedding ($\mathbf{e}_t$-$\mathbf{e}_s$). **(c)** PCA of tumor embedding shows internal subtype structure of BRCA tumors. Color lablels the group index of $k$-means clustering. **(d)** KM estimators of the three breast cancer groups. **(e)** Cox regression using tumor embeddings.

as drivers in the pan-cancer analysis,[2] we also find the top cancer drivers in specific cancer types consistent with our knowledge of cancer oncology. For example, *CDH1* and *GATA3* are drivers of breast invasive carcinoma (BRCA),[35] *CASP8* is known driver of head and neck squamous cell carcinoma (HNSC),[36] *STK11*, *KRAS*, *KEAP1* are known drivers of lung adenocarcinoma (LUAD),[37] *PTEN* and *RB1* are drivers of glioblastoma (GBM),[38] and *FGFR3*, *RB1*, *HSP90AA1*, *STAG2* are known drivers in urothelial bladder carcinoma (BLCA).[39] In contrast, the most frequently mutated genes (control group) are quite different from that using attention mechanism (experiment group), and only a few of them are known drivers (SI Sec. S7).

### 3.4. *Personalized tumor embeddings reveal distinct survival profiles*

Besides learning the specific biological function impact of SGAs on DEGs, we further examined the utility of tumor embeddings $\mathbf{e}_t$ in two perspectives: (1) Discovering patterns of tumors potentially sharing common disease mechanisms across different cancer types; (2) Using tumor embedding to predict patient survival.

We first used the t-SNE plot of tumor embeddings to illustrate the common disease mechanisms across different cancer types (Fig. 3a). When cancer type embedding $\mathbf{e}_s$ is included in full tumor embedding $\mathbf{e}_t$, which has a much higher weight than any individual gene embedding (Fig. 1b, Eq. 1) and dominates the full tumor embedding, tumor samples are clustered according to cancer types. This is not surprising as it is well appreciated that expressions of many genes are tissue-specific.[40] To examine the pure effect of SGAs on tumor embedding, we removed the effect of tissue by subtracting cancer type embeddings $\mathbf{e}_s$, followed by clustering tumors in the stratified tumor embedding space (Fig. 3b). It is interesting to see that each dense area (potential tumor clusters) includes tumors from different tissues of origins, indicating SGAs in these tumors may reflect shared disease mechanisms (pathway perturbations) among tumors, warranting further investigations.

The second set of experiments was to test whether differences in tumor embeddings (thereby difference in disease mechanisms) are predictive of patient clinical outcomes. We conducted unsupervised $k$-means clustering using only breast cancer tumors from our test set, which reveals 3 three groups (Fig. 3c) with significant difference in survival profiles evaluated by log-rank test[41] (Fig. 3d; $p$-value=0.017). In addition, using tumor embeddings as input fea-

tures, we trained $\ell_{1,2}$-regularized (elastic net)[42] Cox proportional hazard models[43] in a 10-fold cross-validation (CV) experiment. This led to an informative ranked list of tumors according to predicted survivals/hazards evaluated by the concordance index (CI) value (CI=0.795), indicating that the trained model is very accurate. We further split test samples into two groups divided by the median of predicted survivals/hazards, which also yields significant separation of patients in survival profiles (Fig. 3e; $p$-value=$5.1 \times 10^{-8}$), indicating that our algorithm has correctly ranked the patients according to characteristics of the tumor.

As shown above, distinct SGAs may share similar embeddings if they share similar functional impact. Thus, two tumors may have similar tumor embeddings even though they do not share any SGAs, as long as the functional impact of distinct SGAs from these tumors are similar. Therefore, tumor embedding makes it easier to discover common disease mechanisms and their impact on patient survival. To further test this, we also performed clustering analysis on breast cancer tumors represented in original SGA space, followed similar survival analysis as described in the previous paragraph (SI Sec. S8).

### 3.5. *Tumor embeddings are predictive of drug responses of cancer cell lines*

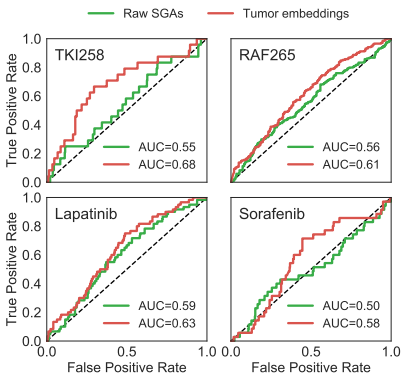


Fig. 4. ROC curves and the areas under the curve (AUCs) of Lasso models trained with original SGAs and tumor embeddings representations on predicting responses to four drugs.

Precision oncology concentrates on using patient-specific omics data to determine optimal therapies for a patient. We set out to see if SGA data of cancer cells can be used to predict their sensitivity to anti-cancer drugs. We used the CCLE dataset,[44] which performed drug sensitivity screening over hundreds of cancer cell lines and 24 anti-cancer drugs. The study collects genomic and transcriptomic data of these cell lines, but in general, the genomic data (except the molecularly targeted genes) from a cell line are not sufficient to predict sensitivity its sensitivity to different drugs.

We discretized the response of each drug following the procedure in previous research.[44,45] Since CCLE only contains a small subset of mutations in TCGA dataset (around 1,600 gene mutations), we retrained the GIT with this limited set of SGAs in TCGA, using default hyperparameters we set before. Cancer type input was removed as well, which is not explicitly provided in CCLE dataset. The output of tumor embeddings $\mathbf{e}_t$ was then extracted as feature. We formulated drug response prediction as a binary classification problem with $\ell_1$-regularized cross entropy loss (Lasso), where the input can be raw sparse SGAs or tanh-curved tumor embeddings $\tanh(\mathbf{e}_t)$. Following previous work,[44] we performed 10-fold CV experiment training Lasso using either inputs to test the drug response prediction task of four drugs with distinct targets. Lasso regression using tumor embeddings consistently outperforms the models trained with original SGAs as inputs (Fig. 4). Specifically, in the case of Sorafenib, the raw mutations just give random prediction results, while the tumor embedding is able to give predictable results. It should be noted that it is possible that certain

cancer cells may host SGAs along the pathways related to FGFR, RAF, EGFR, and RTK, rendering them sensitive to the above drugs. Such information can be implicitly captured and represented by the tumor embeddings, so that the information from raw SGAs are captured and pooled to enhance classification accuracy.

## 4. Conclusion and Future Work

Despite the significant advances in cancer biology, it remains a challenge to reveal disease mechanisms of each individual tumor, particularly which and how SGAs in a cancer cell lead to the development of cancer. Here we propose the GIT model to learn the general impact of SGAs, in the form of gene embeddings, and to precisely portray their effects on the downstream DEGs with higher accuracy. With the supervision of DEGs, we can further assess the importance of an SGA using multi-head self-attention mechanisms in each individual tumor. More importantly, while the tumor embeddings are trained with predicting DEGs as the task, it contains information for predicting other phenotypes of cancer cells, such as patient survival and cancer cell drug sensitivity. The key advantage of transforming SGA into a gene embedding space is that it enables the detection and representation of the functional impact of SGAs on cellular processes, which in turn enables detection of common disease mechanisms of tumors even if they host different SGAs. We anticipate that GIT, or other future models like it, can be applied broadly to gain mechanistic insights of how genomic alterations (or other perturbations) lead to specific phenotypes, thus providing a general tool to connect genome to phenome in different biological fields and genetic diseases. One should also be careful that despite the correlation of genomic alterations and phenotypes such as survival profiles and drug response, the model may not fully reveal the causalities and there may exist other confounding factors not considered.

There are a few future directions for further improving the GIT model. First of all, decades of biomedical research has accumulated a rich body of knowledge, e.g., Gene Ontology and gene regulatory networks, which may be incorporated as the prior of the model to boost the performance.[46] Secondly, we expect that by getting a larger corpus of tumor data with mutations and gene expressions, we will be able to train better models to minimize potential overfitting or variance. Lastly, more clinically oriented investigations are warranted to examine, when trained with a large volume of tumor omics data, the learned embeddings of SGAs and tumors may be applied to predict sensitivity or resistance to anti-cancer drugs based SGA data that are becoming readily available in contemporary oncology practice.

## Acknowledgments

We would like to thank Yifan Xue and Michael Q. Ding for providing the processed TCGA data and discretized CCLE data. We also thank to the helpful suggestions from anonymous reviewers.

## Funding

This work has been partially supported by the following NIH grants: R01LM012011, R01LM010144, and U54HG008540, and it has also been partially supported by the Grant

#4100070287 awarded by the Pennsylvania Department of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the above funding agencies.

## References

1. B. Vogelstein, N. Papadopoulos, V. E. Velculescu *et al.*, Cancer genome landscapes, *Science* **339**, 1546 (2013).
2. C. Kandoth, M. D. McLellan, F. Vandin *et al.*, Mutational landscape and significance across 12 major cancer types, *Nature* **502**, p. 333 (oct 2013).
3. M. S. Lawrence, P. Stojanov, C. H. Mermel *et al.*, Discovery and saturation analysis of cancer genes across 21 tumour types, *Nature* **505**, p. 495 (jan 2014).
4. G. Ciriello, M. L. Miller, B. A. Aksoy *et al.*, Emerging landscape of oncogenic signatures across human cancers, *Nat. Genet.* **45**, p. 1127 (sep 2013).
5. T. I. Zack, S. E. Schumacher, S. L. Carter *et al.*, Pan-cancer patterns of somatic copy number alteration, *Nat. Genet.* **45**, p. 1134 (sep 2013).
6. N. Stransky, E. Cerami, S. Schalm, J. L. Kim and C. Lengauer, The landscape of kinase fusions in cancer, *Nat. Commun.* **5**, p. 4846 (sep 2014).
7. P. A. Jones and S. B. Baylin, The fundamental role of epigenetic events in cancer, *Nat. Rev. Genet.* **3**, p. 415 (jun 2002).
8. N. D. Dees, Q. Zhang, C. Kandoth *et al.*, MuSiC: identifying mutational significance in cancer genomes., *Genome Res.* **22**, 1589 (aug 2012).
9. M. S. Lawrence, P. Stojanov, P. Polak *et al.*, Mutational heterogeneity in cancer and the search for new cancer-associated genes, *Nature* **499**, p. 214 (jun 2013).
10. B. Reva, Y. Antipin and C. Sander, Predicting the functional impact of protein mutations: application to cancer genomics, *Nucleic Acids Res.* **39**, e118 (sep 2011).
11. B. Niu, A. D. Scott, S. Sengupta *et al.*, Protein-structure-guided discovery of functional mutations across 19 cancer types, *Nat. Genet.* **48**, p. 827 (jun 2016).
12. T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean, Distributed representations of words and phrases and their compositionality, in *Proc. of NIPS*, 2013.
13. J. Pennington, R. Socher and C. D. Manning, GloVe: global vectors for word representation., in *Proc. of EMNLP*, 2014.
14. Y. Tao, B. Godefroy, G. Genthial and C. Potts, Effective feature representation for clinical text concept extraction, in *Proc. of Clinical NLP Workshop*, June 2019.
15. M. Ashburner, C. A. Ball, J. A. Blake *et al.*, Gene Ontology: tool for the unification of biology, *Nature Genet.* **25**, p. 25 (may 2000).
16. C. Cai, G. F. Cooper, K. N. Lu *et al.*, Systematic discovery of the functional impact of somatic genome alterations in individual tumors through tumor-specific causal inference, *PLOS Comput. Biol.* **15**, p. e1007088 (jul 2019).
17. B. Lee, S. Min and S. Yoon, Deep learning in bioinformatics, *Brief. Bioinform.* **18**, 851 (2016).
18. K. Lan, D. Wang, S. Fong *et al.*, A survey of data mining and deep learning in bioinformatics, *J. Med. Syst.* **42**, p. 139 (2018).
19. A. Vaswani, N. Shazeer, N. Parmar *et al.*, Attention is all you need, in *Proc. of NIPS*, 2017
20. D. Bahdanau, K. Cho and Y. Bengio, Neural machine translation by jointly learning to align and translate, in *Proc. of ICLR*, 2015.
21. K. Xu, J. Ba, R. Kiros *et al.*, Show, attend and tell: neural image caption generation with visual attention, in *Proc. of ICML*, 07–09 Jul 2015.
22. T. C. G. A. R. Network *et al.*, The cancer genome atlas pan-cancer analysis project, *Nat. Genet.* **45**, p. 1113 (sep 2013).

23. L. Chen, C. Cai, V. Chen and X. Lu, Trans-species learning of cellular signaling systems with bimodal deep belief networks., *Bioinformatics* **31**, 3008 (sep 2015).
24. L. Chen, C. Cai, V. Chen and X. Lu, Learning a hierarchical representation of the yeast transcriptomic machinery using an autoencoder model., *BMC Bioinformatics* **17 Suppl 1**, p. 9 (jan 2016).
25. F. Rosenblatt, The perceptron: a probabilistic model for information storage and organization in the brain, *Psychol. Rev.* , 65 (1958).
26. F. Vandin, E. Upfal and B. J. Raphael, De novo discovery of mutated driver pathways in cancer., *Genome Res.* **22**, 375 (feb 2012).
27. D. E. Rumelhart, G. E. Hinton and R. J. Williams, Learning representations by back-propagating errors, *Nature* **323**, p. 533 (oct 1986).
28. D. P. Kingma and J. L. Ba, Adam: a method for stochastic optimization, in *Proc. of ICLR*, 2015.
29. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, *J. of Mach. Learn. Res.* **15**, 1929 (2014).
30. R. Tibshirani, Regression shrinkage and selection via the lasso, *J. R. Stat. Soc. B* **58**, 267 (1994).
31. L. Maaten and G. Hinton, Visualizing high-dimensional data using t-SNE, *J. Mach. Learn. Res.* **9**, 2579 (01 2008).
32. H. Mi, A. Muruganujan, J. T. Casagrande and P. D. Thomas, Large-scale gene function analysis with the PANTHER classification system, *Nat. Protoc.* **8**, p. 1551 (jul 2013).
33. N. A. de Weerd and T. Nguyen, The interferons and their receptors–distribution and regulation, *Immunol. Cell Biol.* **90**, 483 (2012).
34. A. Gonzalez-Perez, C. Perez-Llamas, J. Deu-Pons *et al.*, IntOGen-mutations identifies cancer drivers across tumor types, *Nat. Methods* **10**, p. 1081 (sep 2013).
35. T. C. G. A. Network *et al.*, Comprehensive molecular portraits of human breast tumours, *Nature* **490**, p. 61 (sep 2012).
36. N. Stransky, A. M. Egloff, A. D. Tward *et al.*, The mutational landscape of head and neck squamous cell carcinoma, *Science* **333**, 1157 (aug 2011).
37. T. C. G. A. R. Network *et al.*, Comprehensive molecular profiling of lung adenocarcinoma, *Nature* **511**, p. 543 (jul 2014).
38. C. W. Brennan, R. G. W. Verhaak, A. McKenna *et al.*, The somatic genomic landscape of glioblastoma, *Cell* **155**, 462 (oct 2013).
39. T. C. G. A. R. Network *et al.*, Comprehensive molecular characterization of urothelial bladder carcinoma, *Nature* **507**, p. 315 (jan 2014).
40. K. A. Hoadley, C. Yau, D. M. Wolf *et al.*, Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin., *Cell* **158**, 929 (aug 2014).
41. N. Mantel, Evaluation of survival data and two new rank order statistics arising in its consideration, *Cancer Chemoth. Rep.* **50**, 163 (March 1966).
42. H. Zou and T. Hastie, Regularization and variable selection via the elastic net, *J. R. Stat. Soc. B* **67**, 301 (2005).
43. D. R. Cox, *Regression Models and Life-Tables* (Springer New York, New York, NY, 1992), New York, NY, pp. 527–541.
44. J. Barretina, G. Caponigro, N. Stransky *et al.*, The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity, *Nature* **483**, p. 603 (mar 2012).
45. M. Q. Ding, L. Chen, G. F. Cooper, J. D. Young and X. Lu, Precision oncology beyond targeted therapy: combining omics data with machine learning matches the majority of cancer cells to effective therapeutics, *Mol. Cancer Res.* (2017).
46. J. Ma, M. K. Yu, S. Fong *et al.*, Using deep learning to model the hierarchical structure and function of a cell, *Nat. Methods* **15**, p. 290 (mar 2018).