Article

Not peer-reviewed version

# Metastatic Cancer Expression Generator (MetGen): A Generative Contrastive Learning Framework for Metastatic Cancer Generation

Zhentao Liu , Yu-Chiao Chiu , Yidong Chen * , Yufei Huang *

Posted Date: 5 April 2024

doi: 10.20944/preprints202404.0397.v1

Keywords: metastatic cancer; deep learning; contrastive learning; tumor microenvironment

*Article*

# Metastatic Cancer Expression Generator (MetGen): A Generative Contrastive Learning Framework for Metastatic Cancer Generation

**Zhentao Liu [1,2], Yu-Chiao Chiu [3,4], Yidong Chen [5,6,*] and Yufei Huang [1,2,4,*]**

[1] Department of Electrical and Computer, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
[2] Cancer Virology Program, UPMC Hillman Cancer Center, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA
[3] Cancer Therapeutics Program, UPMC Hillman Cancer Center, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA
[4] Department of Medicine, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA
[5] Greehey Children Cancer Research Institute, The University of Texas Health Science Center at San Antonio, San Antonio, Texas, USA
[6] Department of Population Health Science, The University of Texas Health Science Center at San Antonio, San Antonio, Texas, USA
* Correspondence: cheny8@uthscsa.edu (Y.C.); yuh119@pitt.edu (Y.H.)

**Abstract:** Despite significant advances in tumor biology and clinical therapeutics, metastasis remains the primary cause of cancer-related deaths. While RNA-seq technology has been used extensively to study metastatic cancer characteristics, challenges persist in acquiring adequate transcriptomic data. To overcome this challenge, we propose MetGen, a generative contrastive learning based on deep learning model. MetGen generates synthetic metastatic cancer expression profiles using primary cancer and normal tissue expression data. Our results demonstrate that MetGen generates comparable samples to actual metastatic cancer samples, and we discuss the learning mechanism of MetGen. Additionally, we demonstrate MetGen's interpretability using metastatic prostate cancer and metastatic breast cancer. MetGen has learned highly relevant signatures in cancer, tissue, and tumor microenvironment, such as immune response and the metastasis process, which potentially fosters a more comprehensive understanding of metastatic cancer biology. The development of MetGen represents a significant step toward the study of metastatic cancer biology by providing a generative model that identifies candidate therapeutic targets for the treatment of metastatic cancer.

**Keywords:** metastatic cancer; deep learning; contrastive learning; tumor microenvironment

## 1. Introduction

Tumor metastasis is the movement of tumor cells spreading from a primary site and progressively colonizing distant organs. When cancer cells disperse from a primary tumor, they travel to other parts of the body through the blood or lymph system. In over 90% of cases, the metastatic spread of tumor cells is the greatest contributor to deaths from cancer [1]. Treating metastatic cancer is an enormous challenge. For all types of cancer, metastatic tumor patients are often unresponsive to existing corresponding primary cancer treatment. It could be due to the basic biology of metastatic tumors and the lack of treatments targeting their biology. The genetic characteristics of metastatic tumors make them highly resistant to standard treatments for each tumor may have a unique tumor microenvironment (TME) and respond differently to the treatment [2]. It is necessary to characterize the complicated molecular mechanism in order to choose appropriate treatment strategies.

Recently, RNA-seq transcriptome data have been widely used to study the nature of cancer metastasis [3–5]. High-throughput technology frees the potential exploration of transcriptome changes at the whole-genome level which would be used for detecting biomarkers for diagnosis and treatment.

One critical challenge of metastatic cancer study is the difficulty of collecting data due to the complicated characteristics of metastatic cancer. Metastatic cancer is quite late-stage cancer, therefore, patients' survival rate is significantly lower than primary cancer patients which makes the participants in any clinical trial much less [6]. The diagnosis of metastatic cancer is also relatively difficult since the primary site and biopsy site are separated which could be confusing at some points. Therefore, building an Atlas for metastatic cancer could be costly. Even some big projects such as The Cancer Genome Atlas (TCGA) or Integrative Clinical Genomics of Metastatic Cancer (MET500) only store a few hundred metastatic samples, including the tumor types for which the primary tumor is rarely diagnosed such as in melanoma [7,8]. It leaves great potential for metastatic cancer data augmentation.

Generative models, such as Generative Adversarial Networks (GANs), have achieved remarkable successes in various fields, including computer vision and natural language processing [9]. However, despite their potential, GANs suffer from several major challenges. 1) unstable training. their training can be unstable due to the simultaneous training of the generator and discriminator models in an adversarial game, where improvements to one model come at the expense of the other, resulting in difficulty in achieving convergence. 2) mode collapse. GANs are prone to mode collapse, where the generator model can learn to produce samples within a particular mode, resulting in a lack of diversity in the generated data. 3) Massive data requirement. GANs require large amounts of data to produce good results. 4) poor interpretability. The generator has no encouragement on giving interpretable and meaningful representation. In the context of metastatic data, GANs face additional challenges due to the limited availability of samples, especially for rare metastatic cancers where only a few samples are available, resulting in severe unstable training and mode collapse. Furthermore, biological data is complex and carries rich entangled information, requiring an interpretable model for downstream analysis.

To this end, we proposed a metastatic cancer expression generator (MetGen), a contrastive learning based generative model that can generate metastatic expression profiles from primary cancer and normal tissue expression samples. Generating metastatic samples from primary cancer and tissue samples is a challenging task, due mainly to limited metastatic cancer samples and corresponding primary cancer & tissue samples. For example, the MET500 dataset has in total of 868 samples from 22 different metastatic cancer types, where most of the cancer types have less than 30 samples[8]. If we further divide one metastatic cancer into subtypes based on tissue sites, the sample size of each subtype would be even smaller which causes mode collapse and poor generating performance. To resolve this issue, contrastive learning is adopted in our model. The goal of contrastive learning is to learn such an embedding space in which similar sample pairs stay close to each other while dissimilar ones are far apart [10]. Contrastive learning can be applied to both supervised and unsupervised settings. In the context of our study, our objective is to measure the similarity between a generated sample and an actual metastatic sample. Specifically, we aim to minimize the distance between samples that originate from the same type of metastatic cancer, while maximizing the distance between samples from different types of metastatic cancers.

**MetGen avoids mode collapse.** Mode collapse is caused by the model finding a trivial solution that only fits majority single mode samples but ignores the samples in other modes [11]. In our model, modes are regulated by not only MET500 target samples but also TCGA source samples. All the modes, and metastatic cancer types in our scenario, are predefined and learned evenly. Therefore, there is no way that model ignores any of them.

**MetGen ensures stable training.** In contrastive learning, samples are fed to the model in pairs[10]. By pairing cancer, tissue, and metastatic sample recursively, we could obtain a massive training dataset easily. For example, one breast cancer sample can pair with multiple different liver tissue samples which results in multiple sample pairs for training. In this way, even rare metastatic cancer types could be trained stable and sufficient.

**MetGen has interpretable latent space.** MetGen has many advantages in model interpretation. Firstly, MetGen model's latent code is obtained by Variational autoencoder, and to a large degree, the learned code components are disentangled [12,13]. Each component or component cluster represents different functions. Secondly, the autoencoder gives a convenient connection between latent space and gene expression space. The changes in latent space could be reflected in gene expression space by reconstructing latent code. Therefore, the biological function of each code components could be learned in expression space easily by masking them out in code space.

In summary, our proposed model integrates primary cancer information into the context of metastasized tissue site during the generative process. This allows for differential analysis and interpretable visualization in a latent space. We anticipate that our model can facilitate the identification of functional pathways associated with cancer metastasis and contribute to the investigation of their roles in this process.

## 2. Results and Discussion

**Overview of the MetGen framework for generating metastatic expression profiles.** The MetGen framework consists of two model components, MetVAE and MetGen. Establishing the MetGen framework includes three stages, i.e., 1) Training MetVAE to obtain latent metastatic cancer codes for MET500 samples, 2) Training MetGen to generate metastatic cancer codes through contrastive mixing of cancer and tissue expressions, and 3) Reconstructing metastatic cancer expression from the generated metastatic cancer code (Figure 1).

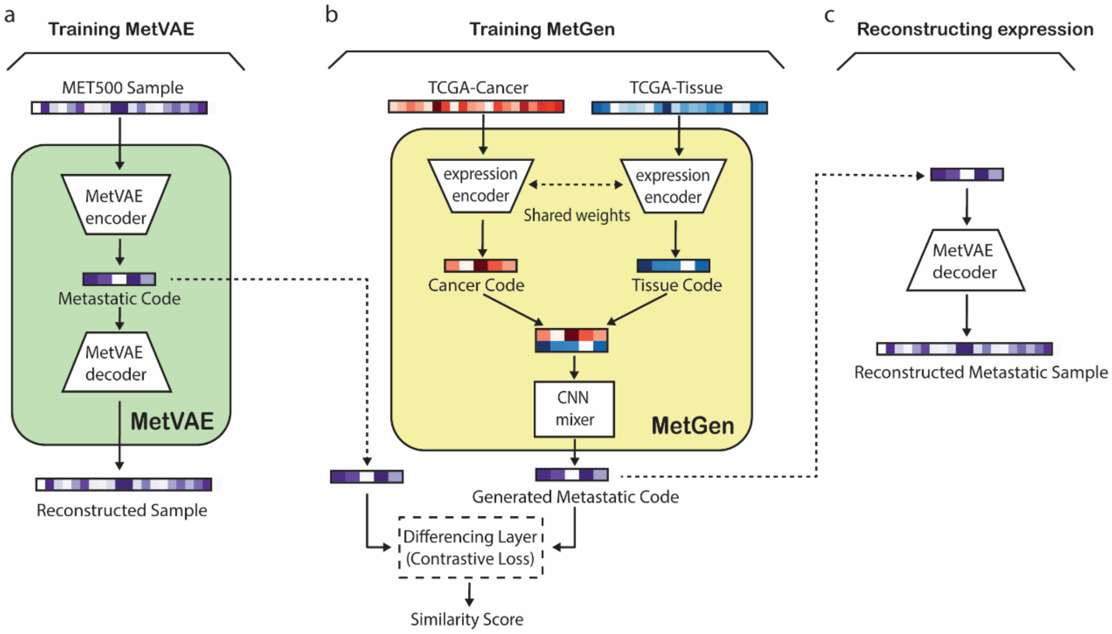

**Figure 1.** MetGen overview. Overview of MetGen workflow. solid arrows indicate model training path while dash arrows indicate workflow path. a) MetVAE model training illustration b) MetGen model learns metastatic code using contrastive learning c) metastatic sample reconstruction using MetVAE decoder.

In the first stage, MetVAE was trained using gene expression profiles from the Integrative Clinical Genomics of Metastatic Cancer (MET500). MetVAE is a variational autoencoder (VAE) that includes an encoder and a decoder (Figure 1a). MetVAE encoder converts a metastatic gene expression sample to a latent metastatic code distribution, which represents a low dimensional representation of the input expression. MetVAE decoder takes a sample from the latent code distribution and reconstructs the input gene expression. MetVAE was trained to minimize the VAE loss that enforces the reconstructed sample to be close to input while minimizing the KL divergence between the learned latent code distribution and standard Gaussian distribution. Therefore, MetVAE learns to produce good latent representations of input MET500 metastatic samples.

In the second stage, MetGen was trained using TCGA primary cancer and tissue expression samples to generate the latent metastatic code of a metastatic cancer type from the corresponding paired primary cancer and tissue expression data (Figure 1b). MetVAE includes two expression encoders with shared weights followed by a CNN mixer (Figure 1b). The two expression encoders convert TCGA primary cancer and tissue gene expression data to primary cancer and tissue latent codes, respectively. The two encoders, which are shared-weights convolutional neural networks (CNNs), extract cancer and tissue features and filter useless information for generating metastatic cancer samples. The CNN mixer integrates cancer and tissue codes to generate the corresponding metastatic cancer code, i.e., code for the corresponding input primary cancer spreading to the input tissue. At the input of the CNN mixer are stacked cancer and tissue codes (Figure 1b), which are compressed into a 1D code by multiple 2 × 1 kennels, each mixing them into different metastatic cancer related features. Two fully connected layers are followed to produce nonlinear combinations of these features and generate the metastatic cancer code. The expression encoder and CNN mixer are trained with a contrastive learning loss to make the generated codes follow the distribution of MetVAE codes of the corresponding metastatic cancer in MET500. The contrastive learning loss compares the generated code and a MET500 metastatic code and minimizes the distances between the codes of matching metastatic cancer types (e.g., when the MetGen output code is generated from TCGA breast cancer and live tissue and MetVAE code is from MET500 breast cancer in liver) but keeps the distance of codes from mismatched types at a preset margin. When the training converges, MetGen would generate metastatic cancer codes as close to MET500 codes of the matching metastatic cancer type as possible but not to be confused with other metastatic cancer types.

In the third stage, the MetGen generated metastatic code is fed into the MetVAE decoder to reconstruct the corresponding metastatic cancer gene expression profile (Figure 1c).

**Training and testing data preparation.** The dataset MET500 and TCGA used to train MetGen are downloaded from UCSC XenaHub. For the sake of discussion, we use the term 'cancer type' for TCGA primary cancer and metastatic cancer types; 'tissue type' for TCGA normal tissue and metastatic biopsy sites; 'metastatic subtype' for specific primary cancer migrated in specific tissue site.

TCGA and MET500 have 20 common cancer types. Our goal is to learn metastatic cancer samples, therefore we want to keep target information in training as much as possible. We picked the top 6 largest sample cancer types in MET500. Common tissue types shared by TCGA and MET500 are 16 in total, 6 of them are associated with selected cancer types. Thus, we have 6 cancer types, 6 tissue types and 21 metastatic subtypes left. We further removed two metastatic subtypes, breast cancer migrated from one to another and lung cancer migrated from one to another, from MET500 data, since it's difficult to define such a metastasis in our model when there are no left and right labels for breasts and lungs in datasets. We filtered datasets based on selected cancer, tissue types and metastatic subtypes. 3052 TCGA cancer samples in 6 cancer types, 338 TCGA tissue samples in 6 tissue types and 219 MET500 samples in 19 metastatic subtypes are kept (Table 1).

**Table 1.** Data table.

| TCGA cancer type | Number of samples | TCGA normal tissue sites | Number of samples |
|---|---|---|---|
| BRCA | 159 | bladder | 8 |
| CHOL | 45 | breast | 5 |
| HNSC | 45 | liver | 243 |
| LUNG | 52 | lung | 75 |
| PRAD | 155 | pancreas | 1 |
| SARC | 100 | skin | 40 |

We randomly take one sample from each TCGA primary cancer, TCGA normal tissue and MET500 metastatic cancer. Three samples together are called a pair. In each pair, if TCGA samples and MET500 samples matches in both cancer type and tissue type, we call it positive pair. Otherwise, it's called negative pair. For instance, a pair including TCGA breast cancer, TCGA normal liver tissue and MET500 breast cancer in liver is a positive pair while a pair including TCGA breast cancer, TCGA lung tissue and MET500 breast cancer in bladder is negative pair, etc. We generated 1000 positive pairs for each metastatic subtype and three times negative pairs, in total 76,000 pairs for training and validating. We also generate another 3800 pairs sample, 200 pairs for each metastatic subtypes, as testing data which is held out during training.

Expressions in all samples were normalized and transformed into log2(FPKM +1) and then scaled between 0 and 1 by the min-max scaler. 7,312 genes were selected for this study. The data preparation is explained in greater detail in Method.

**MetVAE codes capture the cancer and tissue features of metastatic samples.** We first examined if MetVAE codes could represent metastatic samples in lower dimensions without compromising feature patterns. The expressions and MetVAE latent codes are visualized using t-SNE [14] (Figure 2). In the expression space, unlike primary tumors [15] (Supplementary Figure S3), metastatic samples do not form distinguish clusters, though most of the cancer types are loosely grouped such as BRCA and PRAD (Figure 2a). On the other hand, tissue types are not segregated based on biopsy sites, with the exception of liver [8] (Figure 2b). In code space, cancer types show similar pattern to data space sample (Figure 2c), however, more cancer groups are more clustered than in gene expressions. For example, BRCA, HNSC and PRAD are further separated from other cancer groups and formed clear clusters. For tissue types t-SNE (Figure 2d), no significant change discovered. We still observe liver tissues clustered together while other tissue types loosely distributed. In general, MET500 samples group in cancer types, and within cancer clusters we can observe multiple tissue sites. This is expected because metastatic cancer is the primary tumor cell metastasized to other tissue sites. They are still considered same cancer type as primary tumor, so the overall characteristics is similar. In summary, our MetVAE model learns faithful pattern of the cancer types, tissue types and metastatic subtypes in code space compared to gene expression space if not better.

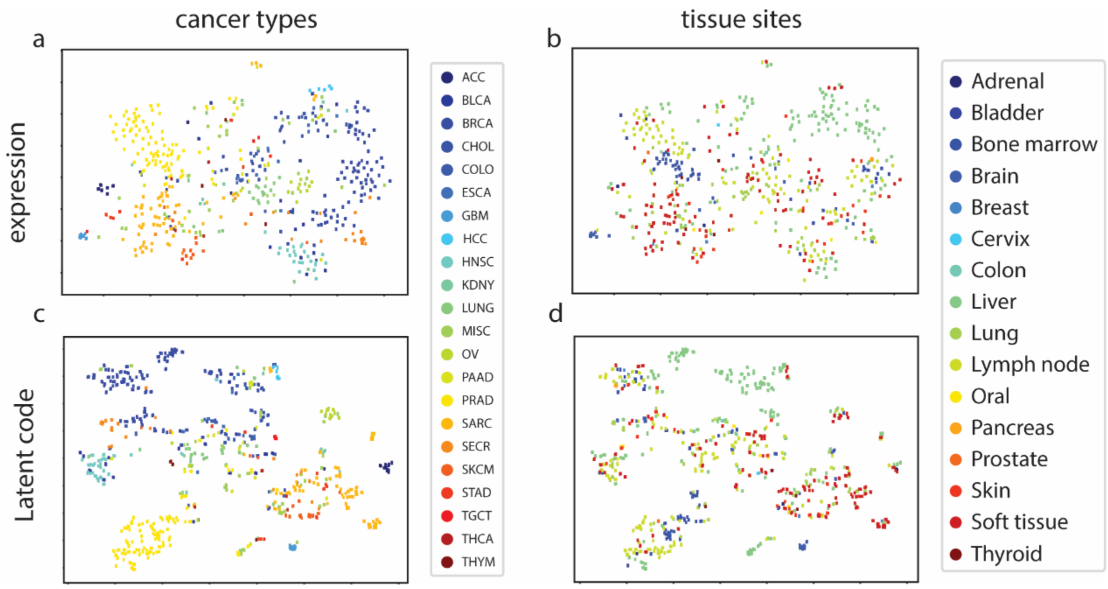

**Figure 2.** MetVAE capture the cancer and tissue pattern of metastatic cancer. MetVAE learns cancer and tissue pattern. a, b) MET500 samples in expression space colored by cancer types and tissue sites c, d) MET500 samples in latent code space colored by cancer types and tissue sites.

To further quantify the extent to which the latent metastatic cancer codes captured both cancer and tissue features, we trained two classifiers on MET500 codes to predict cancer types and tissue types, respectively. The cancer classifier and tissue classifier achieved 98.3 ± 0.7% and 92.3 ± 0.6% ROC AUCs, indicating that MetVAE could capture both cancer and tissue information of MET500 samples correctly.

**MetGen generates high-quality metastatic cancer samples that reveal the distinct patterns of metastatic cancer types.** We generated 200 samples for every 19 metastatic subtypes, resulting in 3,800 samples, using testing data. To visualize the results, t-SNE was applied to reduce the feature dimensions. We observed 19 metastatic subtypes clusters while original MET500 data does not form such clear cancer and tissue clusters (Figure 3a). To test if the generated codes reserve the cancer and tissue information, we tested 3,800 generated codes on trained cancer classifier and tissue classifier. The performance yields to 99.9% and 96.6% for cancer type and tissue type respectively. Nearly perfect classification performance suggests the generated code has very similar cancer and tissue feature to true MET500 code, or classifiers would not be able to classify them correctly.

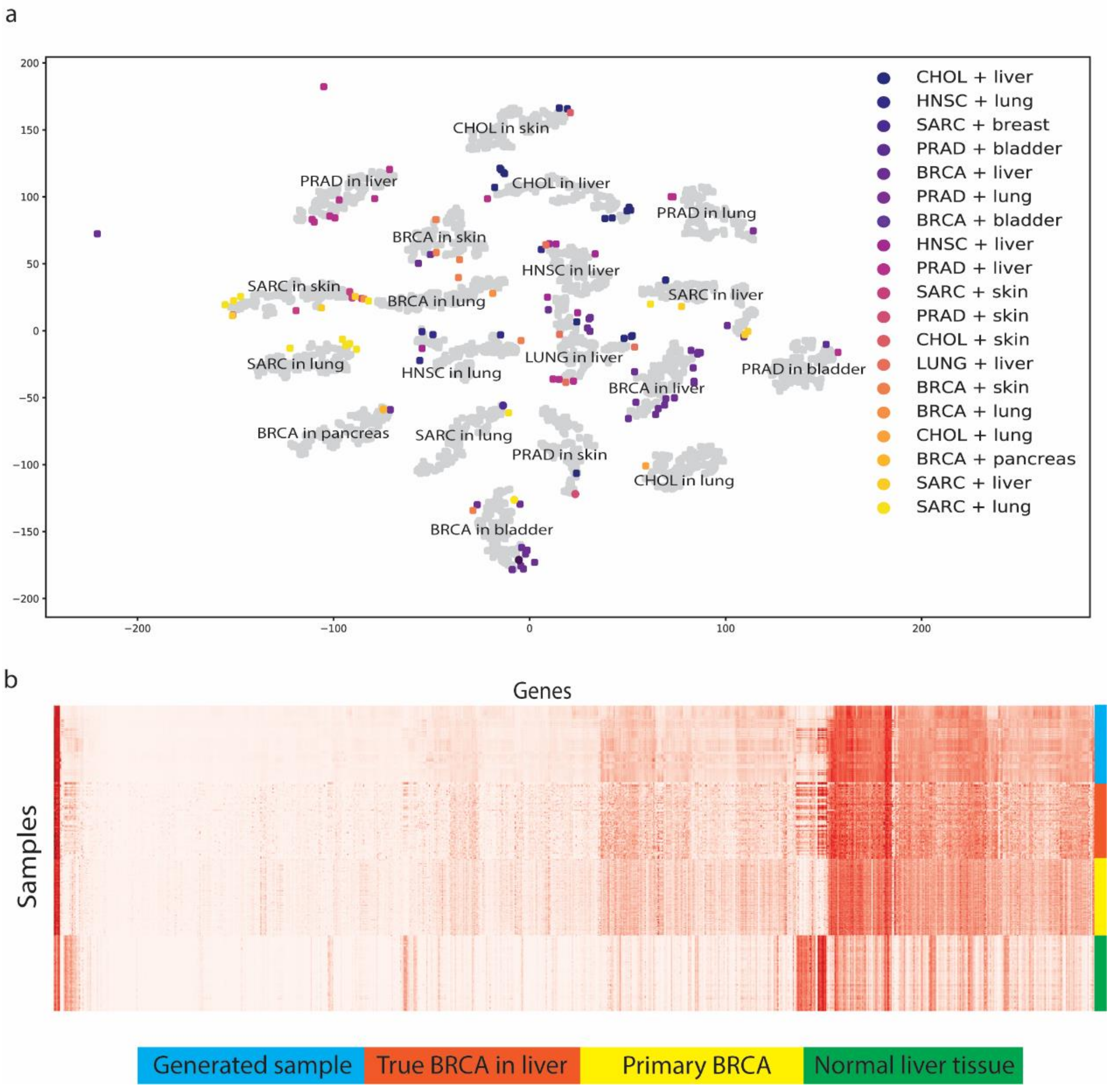

**Figure 3.** MetVAE capture the cancer and tissue pattern of metastatic cancer. MetGen generates metastatic cancer samples. a) t-SNE plot of generated metastatic samples (grey) are MET500 samples (colored). Metastatic types of generated sample are marked in text. b) expression of generated BRCA in liver samples compare to true BRCA in liver, primary BRCA and normal liver tissue.

To better show how MetGen reveals metastatic cancer distribution, we projected true MET500 sample to generated metastatic samples' t-SNE plot (Figure 3a). We found original data points and generated data points have significant overlaps. To be more specific, for each generated cluster, there are original metastatic samples enclosed. That is to say, MetGen generated new samples around the original metastatic target samples and impute space in between. The more data generated the better distribution pattern would show in t-SNE plot. To justify the results quantitatively, we trained a classifier on our testing dataset to predict metastatic subtypes. The idea is to see whether learned knowledge from generated metastatic data can apply on true metastatic samples. Unsurprisingly, the model shows good predicting power for trueMET500 samples with ROC AUC score 97.6%, which furtherly demonstrates MetGen's power in generating high quality metastatic samples from primary cancer and normal tissues.

To investigate one metastatic subtype specifically, we used breast cancer in liver for example. We first retrieved all breast cancer in liver samples from MET500, breast cancer samples and normal liver tissue samples from TCGA, then compared them with 1000 generated breast cancer in liver sample in gene expression heatmaps (Figure 3b). Generated samples do not only have similar discriminative power but also look like true samples in gene expressions. Some genes are not active in primary breast cancer but found high expression level in metastatic breast cancer. Those genes are also upregulated in liver tissues. It implies MetGen learns metastatic cancer by merging primary cancer information into tissue environment.

**MetGen model learns metastatic prostate cancer characteristics.** The development of metastatic cancer requires cancer cells to leave their primary sites and eventually acclimate to new cellular surroundings in tissue sites. It is rather complicated than simply 'put tumor to target location'. We used metastatic prostate cancer for example to illustrate how MetGen takes advantage of its architecture to capture metastatic cancer information.

We generated 4000 metastatic prostate samples in four different tissue sites (lung, liver, bladder, skin) using TCGA-PRAD and TCGA normal samples. Another 4000 samples were generated similarly except the MetGen cancer branch was muted, thus generated samples do not contain prostate cancer information, namely masked cancer samples (Figure 4A). To study the characteristics of the samples robustly, we use gene set variation analysis (GSVA) [16] to perform single sample gene set analysis on 186 KEGG and 50 HALLMARK gene sets downloaded from MSigDB collection [17,18]. GSVA transforms gene expression files into 236 gene set signatures. Eventually, we obtain a GSVA score matrix of 8000 by 236. Each number is a sample's enrichment score for one specific gene set (Figure 4b). Metastatic prostate cancer and masked cancer metastatic sample shows clear different patterns. We anticipate the difference is mainly contributed by the cancer information in metastatic prostate cancer.

We performed differential analysis to identify the difference of two groups (Figure 4c). KEGG androgen response (AR), a key pathway in prostate cancer, ranked high in upregulation pathways. Multiple studies identified it in the development of the normal prostate and in the progression from primary to metastasis, low AR transcriptional activity resulted in upregulated AR expression [19–21]. Some other well-known metastatic prostate pathways such as purine metabolism [22–24], complement and coagulation cascades [25,26] were also successfully detected by our model. Cancer intrinsic pathways like P53 signaling along with WNT β catenin signaling were found more activate in masked cancer samples which also makes sense since prostate cancer are usually suppressive [27,28]. Furthermore, tumor microenvironment (TME) pathways ECM receptor interaction [29,30] and epithelial mesenchymal transition [31,32] found significantly differential expressed in our comparison. The metastatic cascade is dependent on the loss of adhesion between cells which results in the dissociation of the cell from the primary tumor, and subsequently the ability of the cell to attain a motile phenotype via changes in cell to matrix interaction (Figure 4c, Supplementary Table S1).

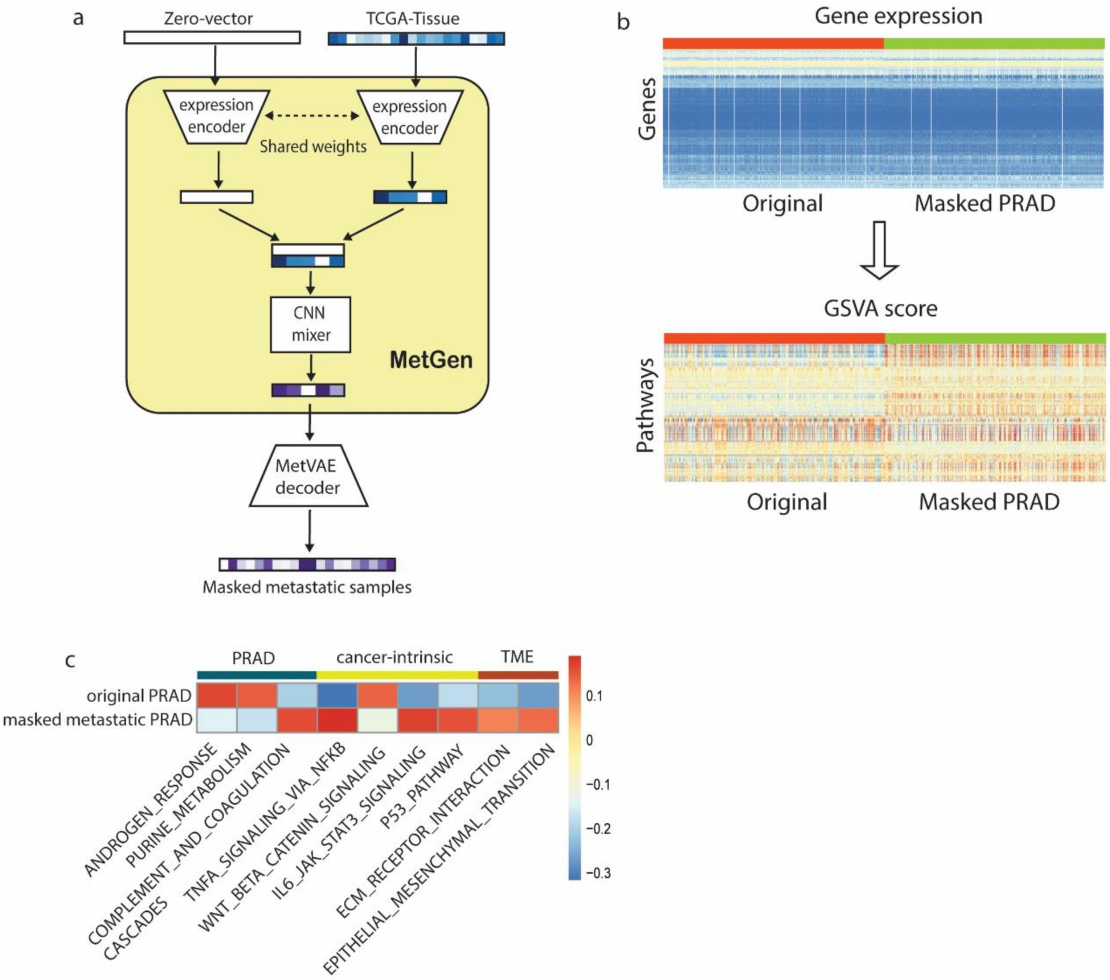

**Figure 4.** MetGen model learns metastatic prostate cancer characteristics. MetGen model learns metastatic prostate cancer characteristics. a) workflow of study metastatic prostate cancer b) converting gene expression profiles to pathway expression, red color is original generated samples, green is masked samples c) metastatic cancer related differential pathways.

MetGen not only ranked prostate cancer related pathways highly but also detected multiple cancer intrinsic and metastatic TME pathways correctly (Figure 4c). Such results prove our model could leverage cancer information from primary tumor and then embed it to local tumor microenvironment.

**MetGen latent components learns functional clusters in metastatic breast cancer in bladder.** The latent code leaned from MetVAE reserves key metastatic cancer information, however due to a limited number of metastatic cancer samples, interpret the code statistically is often difficult. We use metastatic breast cancer in bladder to illustrate how our model helps in latent functional interpretation.

We generated 1000 metastatic breast cancer in bladder latent code using TCGA-BRCA and TCGA normal bladder tissues. Then, the hierarchical cluster was used to characterize 100 code components into clusters and high concordance was shown among 3 functional clusters (Figure 5a).

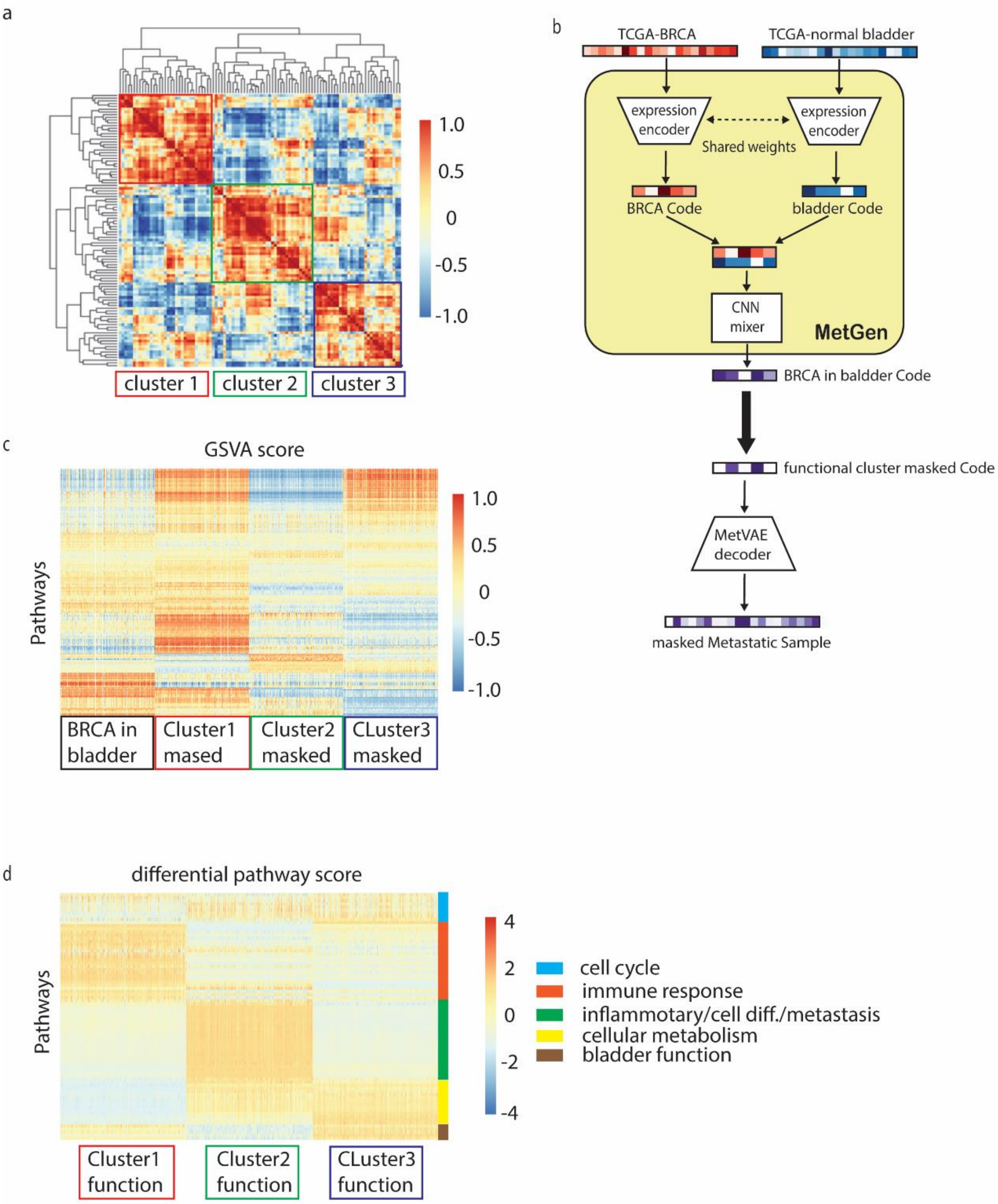

**Figure 5.** MetGen learns functional components in metastatic breast cancer in bladder. **MetGen learns functional components in metastatic breast cancer in bladder.** a) hierarchical clustermap of generated metastatic breast cancer in bladder code. b) analysis workflow of functional components. c) heatmap of GSEA score for 4 groups samples d) selected differential pathways. Padj < 0.05 was used for significant cutoff. Top 10 up or down regulated pathways in fold change are selected for analysis.

To discover the pathway enriched in these functional clusters, we use similar masking approach from previous section. For instance, to get cluster 1 pathways, we set the cluster 1code components to zero, keep the rest components value as they are, then fed the masked code to MetVAE decoder to obtain masked metastatic cancer samples. Later on, masked metastatic cancer samples will be compared to non-masked metastatic cancer samples, therefor the difference is mainly contributed by cluster 1 components (Figure 5b).

Similar to metastatic prostate cancer analysis, we first transform our samples into KEGG and HALLMARK gene set signatures using GSVA. Since we have three cluster masked samples and non-masked samples, 4000 by 236 GSVA score matrix was obtained. Figure 5c shows the heatmap of activate level for gene sets in different samples.

To identify the cluster information, three differential comparisons were performed using LIMMA, for instance, metastatic breast cancer samples vs. each masked cluster samples. Top 20 of positive and negative differential pathways from each comparison were selected for analysis (Supplementary Table S2). From the results (Figure 5d), we noticed functional clusters often represents multiple functions. We roughly group them into 5 signatures.

**Cell cycle**. All three clusters detected G2M checkpoint, E2F targets and MYC targets V1 upregulated. Those pathways are closely related to **cell cycle process** which is commonly active in biological samples [33–36].

**Immune response.** Immune response related pathways are mostly located in cluster 2 and 3 components. In this group we found interferon alpha response, phosphatidylinositol signaling system and PIK3, AKT, mTOR signaling which are actively regulated in immune activation and cell migration in the innate immune system [37,38]. Additionally, estrogen response has been proved with immunoenhancing effect on the immune system [39]. Above pathways together have been shown to regulate immune response by impairing negative selection of high affinity auto-reactive B cells, modulating B cell function, and leading to Th2 response [40].

**Inflammatory/Cell differentiation/metastasis.** This group has rather homogeneous pathways score pattern but very complicated signature. Cluster 1 and 3 detected both Hedgehog and Notch signaling pathways. Hedgehog is a signaling pathway that transmits information to embryonic cells required for proper **cell differentiation.** Notch signaling promotes proliferative signaling during neurogenesis, and its activity is inhibited by Numb to promote neural differentiation. They play a major role in the regulation of embryonic development [41–43]. Interestingly, cluster 1 and 3 also detected olfactory signaling and melanogenesis pathways, along with Hedgehog and Notch signaling pathways, are often found in melanoma tumor metastasis [44]. We also found several pathways related to specifical T cell response and immune cell trafficking in cluster 1 and 3. Antigen processing and presentation is a well-known pathway which can digest protein they encounter and display peptide fragment from them on their surface for another immune cell to recognize [45,46]. Cytosolic DNA sensing pathway are responsible for detecting foreign DNA from invading microbes or host cells and generating innate immune responses [47]. Besides those pattern recognition receptors, immune cells such as T cell receptor and NK cell mediated cytotoxicity signals are also activated which strongly implies this group is for inflammatory response.

**Cellular metabolism**. Cellular metabolism involves complex sequences of controlled biochemical reactions, better known as metabolic pathways. These processes allow organisms to grow and reproduce, maintain their structures, and respond to environmental changes. We found two important cell metabolism target files, lipid metabolism and cytochrome P450 (CYP450). CYP450 are usually membrane-bound and localized to the inner mitochondrial or endoplasmic reticular membrane [48]. Lipid metabolism are important sources of cellular energy, stored as triglycerides[49]. Both play critical roles in oxygenase activity [50].

**Bladder function**. Cluster 2 detected a few non cancer related pathways such as vascular smooth muscle contraction, tight junction and calcium signaling pathways [51,52]. Those pathways are highly correlated to muscle activities of bladder. We believe metastatic tissue site information are mainly stored in this cluster.

Above results show MetGen latent code components give great interpretability. Each functional code cluster captured multiple functional processes which are highly correlated to metastatic breast cancer. The generated samples reveal the characteristics of metastasis, target tissue sites and regular cancer related immune responses.

## 3. Method

### 3.1. Data Preprocessing

Both TCGA and MET500 has huge number of genes, most of those genes have very low expression and discriminative power, thus they induce noisy background which essentially hurt the model training. Also, two datasets have different labeling. Since the goal is to generate MET500 samples from TCGA samples, we need to match two datasets in genes and labels.

To unify TCGA and MET500 genes, we first filtered out genes in MET500 with mean and standard deviation of less than -5 and 1.5 respectively. 6032 genes were kept after filtering. Similarly, we use mean and standard deviation to filter TCGA data and 4096 genes were kept. We took union of two gene sets, then removed the genes which do not exist in both TCGA and MET500. Eventually, 7312 genes were kept for this study.

To unify TCGA and MET500 labels, we first combined similar cancer types in TCGA then renamed all the TCGA cancer label to the corresponding cancer label in MET500. Specifically, four primary cancer groups in TCGA known as [(COAD, READ), (KIRP, KIRC, KICH), (LUAD, LUSC), LIHC] were renamed as their corresponding metastatic labels in MET500 [COLO, KDNY, LUNG, HCC], respectively. After relabeling, TCGA and MET500 have in total 20 common cancer types. We pick the 6 top number common cancer types in MET500 as the total cancer types in this study since we want to have target information as much possible. Corresponding tissue types in both TCGA and MET500 dropped to 6 after cancer types filtering, we kept all of them as total tissue types in this study.

**Training and testing data preparation**. Training data has in total 6 cancer types and 6 tissue types in both TCGA samples and MET500 samples. After removing breast cancer in breast and lung cancer in lung from MET500 data, we obtained 219 samples in 19 metastatic subtypes such as BRCA in bladder, SARC in lung, etc. For each combination, we sampled pairs from TCGA cancer, normal tissue and MET500 samples. If TCGA cancer plus TCGA normal tissue match with MET500 sample, we call it positive pair, otherwise it's called negative pair. For example, TCGA breast cancer, TCGA liver tissue and MET500 breast cancer in liver is a positive pair while TCGA breast cancer, TCGA lung tissue, and MET500 breast cancer in bladder are negative pairs, etc. Negative pairs are critical for contrastive learning, increasing number of negative pairs improves the downstream task performance[53]. We generated 1000 positive pairs and 3000 negative pairs for each combination in a total of 76,000 sample pairs and split them into training data and validation data. We also created another 3800 sample pairs, 200 pairs for each combination, as testing data which is held out during training.

### 3.2. MetVAE Model

In this section, we explain the variational autoencoder used in MET500 code extraction in detail. Unlike classic autoencoder, variational autoencoder does not just learn representative code but the prior distribution of data [13]. We modeled MET500 gene expression $x_{MET}$ from a Gaussian distribution $p(x_{MET}|z)$. Vanilla VAE encourages posterior distribution over the $p(x_{MET}|z)$ to be isotropic Gaussian $p(x_{MET}|z)$ thus the latent code $z$ could have disentangled features which is convenient for further interpretation work.

Our VAE network has a standard architecture that consists of an encoder of two layers, a loss function, and a decoder of two layers. The number of neurons per hidden layer in the encoder used was 1000 and 100 with a symmetric decoder. All neural networks use batch normalization layer. The activation function between two hidden layers is ReLU. We use a sigmoid activation in the decoder.

The goal of inference is to estimate the maximum likelihood of $p(x_{MET}|z)$ which is naturally defined by decoder. We cannot apply Bayes theorem directly because in general it is an intractable computation. However, we could use the variational inference technique to low-bound the likelihood which can be expressed as

$$logp(x_{MET}|z) \geq E_{q(z|x_{MET})}logp(x_{MET}|z) - D_{KL}(q(z|x_{MET})||p(z)) \quad (1)$$

Where $q(z|x)$ is predefined to be Gaussian with diagonal covariance matrix. Mean and covariance are learned from the encoder network.

We use Adam optimizer to train the model with an initial learning rate of 0.0001 and a decay rate of 1e-6. The model is trained in batch size 32 and optimizes the objective function. When the loss is stable for more than 20 epochs, we stop the training since the model is converged. In our case, it usually takes around 200 epochs. After the model is well trained, we save code $z$ of MET500 for further studies.

### *3.3. Contrastive Learning for MetGen*

In this section we explain the MetGen network in greater details. The whole framework consists of three modules, TCGA encoder, CNN mixer and contrastive learner.

*TCGA encoder.* The performance of machine learning algorithms can be impacted dramatically by too many features, generally referred as 'curse of dimensionality'. The reason for the curse of dimensionality is essentially caused by complication of large data space which contains noise and unnecessary information. When dealing with high dimensional data, it is often efficient to reduce the number of features by projecting the data to a lower dimensional space:

$$code(x_{cancer}, x_{tissue}; \theta) = E\left(x_{cancer,} x_{tissue}; \theta\right) \quad (2)$$

Where $x_{cancer}$, $x_{tissue}$ are input TCGA samples, $E$ is the encoder, $\theta$ represents the parameters of the encoder.

To capture features of TCGA data, we adopted the architecture from published work [54] which is built by two CNN layers and one fully connected layer. Previous work used all 33 cancer types and normal tissue for training, since our data are relatively small, we reduced the number of parameters of the previous model to have smaller model space.

*CNN mixer.* CNN mixer is the model merging cancer code and tissue code to generate metastatic cancer code. It learns the mechanism of tissue environment activation for cancer invasion.

CNN mixer consists of two CNN layers and one fully connected layer. First, the module stacks the cancer code and tissue code to 2 × 512 matrix, then feed the matrix to first CNN layer. We used 64 kernels in 2 × 1 to learn the merging process. Since cancer code and tissue code are in same space, 2 × 1 kernel gives us better biological meaning which is helpful in interpretation. The second CNN layer and fully connected layer are used to project the output code to a dimension of 100 so that we can compare it with corresponding MET500 code.

$$code\left(x_{MET_{mix}}; \omega\right) = f\left(E(x_{cancer}), E(x_{tissue}); \omega\right) \quad (3)$$

Where $f$ is the CNN mixer module, $\omega$ represents the parameters of CNN mixer.

*Contrastive learner.* Contrastive loss takes the output of a positive example and calculates its distance to an example in the same class and contrasts it with the distance of negative examples[10]. In our case, the network will calculate the Euclidian distance of the generated code and MET500 code. When the labels of two codes match, the Euclidian distance should be minimized to zero, otherwise, the distance would be pushed to a margin. Thus, our final loss function is:

$$L_{contrastive}(x_{cancer}, x_{tissue}, z; \theta, \omega) = y_{true} \times ED^2\left(code\left(x_{MET_{mix}}; \omega\right), z\right) +$$

$$(1 - y_{true}) \times \max(margin - ED(code(x_{MET}; \omega), z), 0) \quad (4)$$

Where $ED(,)$ is the Euclidian distance, $y_{true}$ is the label of sample pair, 1 and 0 for positive and negative respectively.

### *3.4. Standard DNN Classifiers*

The classifier used in all the classification tasks adopted the same architecture. We construct a two layers MLP neural network which has 60 and 40 neurons in each layer. The activation function is chosen as ReLU. Then model feeds extracted low dimensional features to SoftMax classifier for prediction.

All the deep learning models are trained on Keras API with TensorFlow backend [55].

### *3.5. GSVA*

Gene set variation analysis (GSVA) is a nonparametric unsupervised analysis method mainly used to evaluate the gene set enrichment of single samples [16]. The expression matrix of genes is transformed into the expression matrix of pathways in different samples to evaluate different pathways enriched in different samples. It stabilizes the gene expression signals in robust high-level functions. In this study, we perform GSVA using GSVA package in R software. KEGG and HALLMARK gene sets are downloaded from MSigDB collection [56].

### *3.6. Differential Analysis*

The Bioconductor Linear Model for Microarray Analysis (LIMMA) package was used to calculate the differential pathway expression of each cancer sample in the present study. LIMMA remains highly recommended for such analyses since the pathway expressions from GSVA are continuous number which is more similar to microarray analysis [57].

## 4. Conclusions

We proposed an interpretable deep-learning framework inspired by contrastive learning for metastatic cancer sample generation using primary tumor and normal tissue samples. The pipeline consists of two model components, MetVAE and MetGen. We comprehensively investigated our models' performance and interpretability using TCGA and MET500 data. We show that

- MetVAE can encode metastatic cancer and tissue site information faithfully into latent code. We investigated MetVAE for cancer and tissue type classification using MET500 data and our model gives good performance on both tasks.
- MetGen can generate metastatic cancer expressions from primary tumor and normal tissues. We generated 19 metastatic cancer types using TCGA data. Our generated samples reserved essential metastatic cancer information and achieved good performance in multiple classification tests.
- We demonstrated the interpretability of our models using metastatic prostate cancer and metastatic breast cancer in bladder. Highly relevant functions are learned from primary cancer and tissue sites which further affirmed the power of our model.

Inevitably, there are still many challenges to be addressed. Metastatic tumors are highly microenvironment dependent, right now our model only provide a general environment tissue site. For same reason, the model cannot deal with the scenario such as cancer cell travels from left breast to right breast either. Research into solutions to these challenges will be our future focus.

**Supplementary Materials:** Supplementary Table S1 – metastatic prostate cancer pathways; Supplementary Table S2 - metastatic breast cancer in bladder functional cluster pathways.

**Funding:** This research was supported by Cancer Prevention and Research Institute of Texas (RP160732 and RP190346), and National Institutes of Health (K99CA248944 and R00CA248944).

**Author Contributions:** Conceptualization, Y.H. and Z.L.; methodology, Y.H., Y.-C.C. and Z.L.; formal analysis, Z.L.; writing—original draft preparation, Z.L.; writing—review and editing, Y.H., Y.C.; visualization, Z.L.; supervision, Y.H.; project administration, Y.H.; funding acquisition, Y.C. and Y.H. All authors have read and agreed to the published version of the manuscript.

**Data Availability Statement:** All data used in this study are publicly available on XenaHub; The code is available at: https://github.com/Fclef/MetGen.

## References

1. Seyfried, T.N. and L.C. Huysentruyt, *On the origin of cancer metastasis.* Crit Rev Oncog, 2013. **18**(1-2): p. 43-73.
2. Quail, D.F. and J.A. Joyce, *Microenvironmental regulation of tumor progression and metastasis.* Nat Med, 2013. **19**(11): p. 1423-37.
3. Crea, F., et al., *The non-coding transcriptome as a dynamic regulator of cancer metastasis.* Cancer and metastasis reviews, 2014. **33**: p. 1-16.
4. Crnic, I. and G. Christofori, *Novel technologies and recent advances in metastasis research.* International Journal of Developmental Biology, 2004. **48**(5-6): p. 573-581.
5. Welch, D.R., *Microarrays bring new insights into understanding of breast cancer metastasis to bone.* Breast Cancer Research, 2004. **6**(2): p. 1-4.
6. Caswell-Jin, J.L., et al., *Change in Survival in Metastatic Breast Cancer with Treatment Advances: Meta-Analysis and Systematic Review.* JNCI Cancer Spectrum, 2018. **2**(4).
7. 53, D.C.C.B.R.J.M.A.K.A.P.T.P.D.W.Y. and T.S.S.L.D.A. 68, *The cancer genome atlas pan-cancer analysis project.* Nature genetics, 2013. **45**(10): p. 1113-1120.
8. Robinson, D.R., et al., *Integrative clinical genomics of metastatic cancer.* Nature, 2017. **548**(7667): p. 297-303.
9. Durgadevi, M. *Generative Adversarial Network (GAN): a general review on different variants of GAN and applications*. in *2021 6th International Conference on Communication and Electronics Systems (ICCES)*. 2021. IEEE.
10. Chicco, D., *Siamese neural networks: An overview.* Artificial neural networks, 2021: p. 73-94.
11. Kushwaha, V. and G. Nandi. *Study of prevention of mode collapse in generative adversarial network (gan)*. in *2020 IEEE 4th Conference on Information & Communication Technology (CICT)*. 2020. IEEE.
12. Higgins, I., et al. *beta-vae: Learning basic visual concepts with a constrained variational framework*. in *International conference on learning representations*. 2017.
13. Kingma, D.P. and M. Welling, *Auto-encoding variational bayes.* arXiv preprint arXiv:1312.6114, 2013.
14. Van der Maaten, L. and G. Hinton, *Visualizing data using t-SNE.* Journal of machine learning research, 2008. **9**(11).
15. Zhang, Z., et al., *Uniform genomic data analysis in the NCI Genomic Data Commons.* Nature communications, 2021. **12**(1): p. 1226.
16. Hänzelmann, S., R. Castelo, and J. Guinney, *GSVA: gene set variation analysis for microarray and RNA-seq data.* BMC bioinformatics, 2013. **14**: p. 1-15.
17. Liberzon, A., et al., *The molecular signatures database hallmark gene set collection.* Cell systems, 2015. **1**(6): p. 417-425.
18. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic acids research, 2000. **28**(1): p. 27-30.
19. Alumkal, J.J., et al., *Transcriptional profiling identifies an androgen receptor activity-low, stemness program associated with enzalutamide resistance.* Proceedings of the National Academy of Sciences, 2020. **117**(22): p. 12315-12323.
20. Formaggio, N., M.A. Rubin, and J.-P. Theurillat, *Loss and revival of androgen receptor signaling in advanced prostate cancer.* Oncogene, 2021. **40**(7): p. 1205-1216.
21. Davies, A., et al., *An androgen receptor switch underlies lineage infidelity in treatment-resistant prostate cancer.* Nature cell biology, 2021. **23**(9): p. 1023-1034.
22. Yin, J., et al., *Potential mechanisms connecting purine metabolism and cancer therapy.* Frontiers in immunology, 2018. **9**: p. 1697.
23. De Vitto, H., et al., *The intersection of purine and mitochondrial metabolism in cancer.* Cells, 2021. **10**(10): p. 2603.
24. Ahmad, F., M.K. Cherukuri, and P.L. Choyke, *Metabolic reprogramming in prostate cancer.* British journal of cancer, 2021. **125**(9): p. 1185-1196.
25. Revel, M., et al., *Complement system: promoter or suppressor of cancer progression?* Antibodies, 2020. **9**(4): p. 57.
26. Stallone, G., et al., *Modulation of complement activation by pentraxin-3 in prostate cancer.* Scientific Reports, 2020. **10**(1): p. 1-11.
27. Zhang, Y. and X. Wang, *Targeting the Wnt/β-catenin signaling pathway in cancer.* Journal of hematology & oncology, 2020. **13**: p. 1-16.
28. Wen, X., et al., *New advances in canonical Wnt/β-catenin signaling in cancer.* Cancer Management and Research, 2020. **12**: p. 6987.
29. Brassart-Pasco, S., et al., *Tumor microenvironment: extracellular matrix alterations influence tumor progression.* Frontiers in oncology, 2020. **10**: p. 397.
30. Bao, Y., et al., *Transcriptome profiling revealed multiple genes and ECM-receptor interaction pathways that may be associated with breast cancer.* Cellular & molecular biology letters, 2019. **24**(1): p. 1-20.
31. Jing, Y., et al., *Epithelial-Mesenchymal Transition in tumor microenvironment.* Cell & bioscience, 2011. **1**: p. 1-7.
32. Gao, D. and V. Mittal, *Tumor microenvironment regulates epithelial–mesenchymal transitions in metastasis.* Expert review of anticancer therapy, 2012. **12**(7): p. 857-859.

33. Stark, G.R. and W.R. Taylor, *Analyzing the G2/M checkpoint.* Checkpoint Controls and Cancer: Volume 1: Reviews and Model Systems, 2004: p. 51-82.
34. Eastman, A., *Cell cycle checkpoints and their impact on anticancer therapeutic strategies.* Journal of cellular biochemistry, 2004. **91**(2): p. 223-231.
35. Ren, B., et al., *E2F integrates cell cycle progression with DNA repair, replication, and G2/M checkpoints.* Genes & development, 2002. **16**(2): p. 245-256.
36. Dang, C.V., *c-Myc target genes involved in cell growth, apoptosis, and metabolism.* Molecular and cellular biology, 1999. **19**(1): p. 1-11.
37. Xie, S., et al., *Identification of a role for the PI3K/AKT/mTOR signaling pathway in innate immune cells.* PloS one, 2014. **9**(4): p. e94496.
38. Belardelli, F., et al., *Interferon-alpha in tumor immunity and immunotherapy.* Cytokine & growth factor reviews, 2002. **13**(2): p. 119-134.
39. Cutolo, M., et al., *Estrogens, the immune response and autoimmunity.* Clinical and experimental rheumatology, 1995. **13**(2): p. 217-226.
40. Kovats, S., *Estrogen receptors regulate innate immune cells and signaling pathways.* Cellular immunology, 2015. **294**(2): p. 63-69.
41. Lin, G.L. and K.D. Hankenson, *Integration of BMP, Wnt, and notch signaling pathways in osteoblast differentiation.* Journal of cellular biochemistry, 2011. **112**(12): p. 3491-3501.
42. Jia, Y., Y. Wang, and J. Xie, *The Hedgehog pathway: role in cell differentiation, polarity and proliferation.* Archives of toxicology, 2015. **89**: p. 179-191.
43. Lawson, N.D., A.M. Vogel, and B.M. Weinstein, *sonic hedgehog and vascular endothelial growth factor act upstream of the Notch pathway during arterial endothelial differentiation.* Developmental cell, 2002. **3**(1): p. 127-136.
44. Zhang, H., et al., *The role of calcium signaling in melanoma.* International journal of molecular sciences, 2022. **23**(3): p. 1010.
45. Jurewicz, M.M. and L.J. Stern, *Class II MHC antigen processing in immune tolerance and inflammation.* Immunogenetics, 2019. **71**: p. 171-187.
46. Alberts, B., et al., *T cells and MHC proteins*, in *Molecular Biology of the Cell. 4th edition*. 2002, Garland Science.
47. Barber, G.N., *Innate immune DNA sensing pathways: STING, AIMII and the regulation of interferon production and inflammatory responses.* Current opinion in immunology, 2011. **23**(1): p. 10-20.
48. McDonnell, A.M. and C.H. Dang, *Basic review of the cytochrome p450 system.* Journal of the advanced practitioner in oncology, 2013. **4**(4): p. 263.
49. Liu, K. and M. Czaja, *Regulation of lipid stores and metabolism by lipophagy.* Cell Death & Differentiation, 2013. **20**(1): p. 3-11.
50. Nelson, D.R., *The cytochrome p450 homepage.* Human genomics, 2009. **4**(1): p. 1-7.
51. Carattino, M.D., et al., *Bladder filling and voiding affect umbrella cell tight junction organization and function.* American Journal of Physiology-Renal Physiology, 2013. **305**(8): p. F1158-F1168.
52. Kuo, I.Y. and B.E. Ehrlich, *Signaling in muscle contraction.* Cold Spring Harbor perspectives in biology, 2015. **7**(2): p. a006023.
53. Chen, T., et al., *A Simple Framework for Contrastive Learning of Visual Representations*, in *Proceedings of the 37th International Conference on Machine Learning*, D. Hal, III and S. Aarti, Editors. 2020, PMLR: Proceedings of Machine Learning Research. p. 1597--1607.
54. Mostavi, M., et al., *Convolutional neural network models for cancer type prediction based on gene expression.* BMC medical genomics, 2020. **13**: p. 1-13.
55. Chollet, F., *Keras: Deep Learning Library. 2015.* URL https://github. com/fchollet/keras.
56. Liberzon, A., et al., *Molecular signatures database (MSigDB) 3.0.* Bioinformatics, 2011. **27**(12): p. 1739-40.
57. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Res, 2015. **43**(7): p. e47.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.