# RNA-to-image multi-cancer synthesis using cascaded diffusion models

Francisco Carrillo-Perez[1,2], Marija Pizurica[1,3], Yuanning Zheng[1], Jeanne Shen[4] and Olivier Gevaert[1,5*]

[1]Stanford Center for Biomedical Informatics Research (BMIR), Stanford University, School of Medicine, 1265 Welch Rd, Stanford, 94305-547, CA, USA.
[2]Department of Architecture and Computer Technology (ATC), University of Granada, C. Periodista Daniel Saucedo Aranda, s/n, Granada, 18014, Granada, Spain.
[3]Internet technology and Data science Lab (IDLab), Ghent University, Technologiepark-Zwijnaarde 126, Gent, 9052, Gent, Belgium.
[4]Department of Pathology, Stanford University School of Medicine, 300 Pasteur Dr, Palo Alto, 94304, CA, USA.
[5]Department of Biomedical Data Science, Stanford University, School of Medicine, Medical School Office Building (MSOB), 1265 Welch Rd, Stanford, 94305-547, CA, USA.

*Corresponding author(s). E-mail(s): ogevaert@stanford.edu;

**Abstract**

Synthetic data generation offers a solution to the data scarcity problem in biomedicine where data are often expensive or difficult to obtain. By increasing the dataset size, more powerful and generalizable machine learning models can be trained, improving their performance in clinical decision support systems. The generation of synthetic data for cancer diagnosis has been explored in the literature, but typically in the single modality setting (e.g. whole-slide image tiles or RNA-Seq data). Given the success of text-to-image synthesis models for natural images, where one modality is used to generate a related one, we propose RNA-to-image synthesis (RNA-CDM) in a multi-cancer context. First, we trained a variational auto-encoder in order to reduce the dimensions of the patient's gene expression profile, showing that this can accurately differentiate between different cancer types. Then, we trained a cascaded diffusion model to synthesize realistic whole-slide image tiles using the latent representation of the patient's RNA-Seq data. We show that generated tiles preserved the cell-type distribution found in real-world data, with important cell types detectable by a state-of-the-art cell identification model in the synthetic samples. Next, we successfully used this synthetic data to pretrain a multi-cancer classification model, observing an improvement in performance after training from scratch across 5-fold cross validation. Our results demonstrate the potential utility of synthetic data for developing multi-modal machine learning models in data scarce

settings, as well as the possibility of imputing missing data modalities by leveraging the information present in available data modalities.

**Keywords:** Diffusion models, multi-modal synthetic data, data imputation

# Introduction

Cancer is one of the leading causes of death worldwide, just behind cardiovascular diseases [1]. A physician usually carries out multiple screenings to diagnose the disease in a clinical setting, such as visually examining a digitized tissue slide or finding specific up or down regulations in the patient's gene expression. Even though these screenings are routinely used at hospitals, rarely all of them are performed on the same patient due to monetary or logistical constraints. Cancer is a multi-scale and multi-factorial disease, and its effects can be recognized at multiple levels [2–4]. For instance, genetic alterations in tumor cells and cells from the tumor microenvironment lead to functional changes which in turn can influence their cellular physiology. [5–8]. Therefore, by not having all screening modalities available, we are losing a part of the picture that can lead to early detection.

Machine learning (ML), and specifically deep learning (DL), has shown tremendous potential for cancer detection and classification in recent years. By using different modalities, such as RNA-Seq, whole-slide-imaging (WSI), miRNA-Seq, or DNA methylation, clinical decision support systems have been created [8–12]. However, two problems are present when working with cancer data. First, DL models are known for being data hungry, needing huge amounts of data to be properly trained. Second, even though the combination of biological data types in a multimodal setting has shown to be superior for cancer detection and prognosis [13–19], unfortunately, the majority of available datasets are incomplete, missing some modalities. Although certain projects make efforts to get a complete picture of the disease by collecting all modalities, such as The Cancer Genome Atlas (TCGA) [20] project or the UK Genomics Pathology Imaging Collection [21], the number of single-modality datasets available greatly surpass them. For instance, there are thousands of gene expression series in the Gene Expression Omnibus (GEO) platform [22] for which the corresponding tissue slide is not available, limiting the potential for creating multimodal ML models.

To address these issues, generative models have been presented as a solution in the literature for cancer data. Specifically, generative adversarial networks (GANs) and Variational Autoencoders (VAE) have been used for generating synthetic WSI and RNA-Seq data. GANs have shown their abilities for modeling cancer characteristics across multiple cancer types, successfully generating synthetic tiles [23, 24] and synthetic gene expression profiles that closely resemble real profiles and capture biological information [25]. VAEs have been successfully applied to gene expression data, showing synthetic generation

capabilities in a temporal way and they have been used for data imputation [26, 27]. However, these models present some significant drawbacks when dealing with image data. While sampling is fast for GANs and although they can generate high-quality data, their training is unstable and prone to model collapse, leading to loss of diversity in the generated samples [28–30]. Given these problems, usually a different model is trained for each tissue per cancer type, which is not practical in a real-world scenario. In the case of VAEs, even though their training is much more stable, the generated samples are often blurry compared to those coming from GANs because of the injected noise and the imperfect reconstruction [31]. Furthermore, despite models have been presented for the synthetic generation of both WSI and gene expression data, the multi-scale nature of cancer is usually not taken into account for the generation. Nevertheless, to the best of our knowledge, RNA-to-image synthesis has not been yet explored for cancer tissue.

Text-to-image models have recently caught the attention of the world, given their incredible generative performance. Dall-E 2 and Imagen presented incredible results in image synthesis based on textual input, surpassing all previous methods in quality metrics and synthesis capabilities [32, 33]. They rely on a novel paradigm that is different from previous GAN and VAE-based approaches. Instead, they make use of diffusion models, a DL technique firstly introduced by Sohl-Dickstein et al. [34]. Diffusion models are based on Langevin dynamics, which is an approach in physics for the mathematical modelling of the dynamics of molecular systems. Random noise is added to the data, thereby 'destroying' it until an isotropic random Gaussian distribution is left. Then, a model is trained that learns to reverse this process. Once the model is trained, new images can be synthesized from noise, yet resemble the images from the training data. Different approaches are used for conditioning the model to generate different images based on the given text. Authors of Imagen used a pre-trained large language model to obtain an embedding of the text to condition or 'guide' the model generation. Similarly, Dall-E 2 authors obtained a CLIP image-text embedding [35] to condition the diffusion model during the generation. New images are created using these two approaches, capturing the context of the text and portraying it in the generated image.

In this work, we explore the relation between cancer tissue and its gene expression profile in an RNA-to-image synthesis problem, with the goal of using synthetic cancer images to pretrain DL models and impute missing data modalities. We present RNA-CDM, a single-cascaded diffusion-based model that is able to synthesize routine hematoxylin and eosin (H&E)-stained tissue image tiles for different cancer tissues without specifying the tissue label, by conditioning on a latent representation of the gene expression profile obtained using a $\beta$-VAE. Using a state-of-the-art cell segmentation model, we show that the generated tiles maintain the cell distribution of the real data, preserving cell morphology. In addition, we prove that the synthetic data can be used for

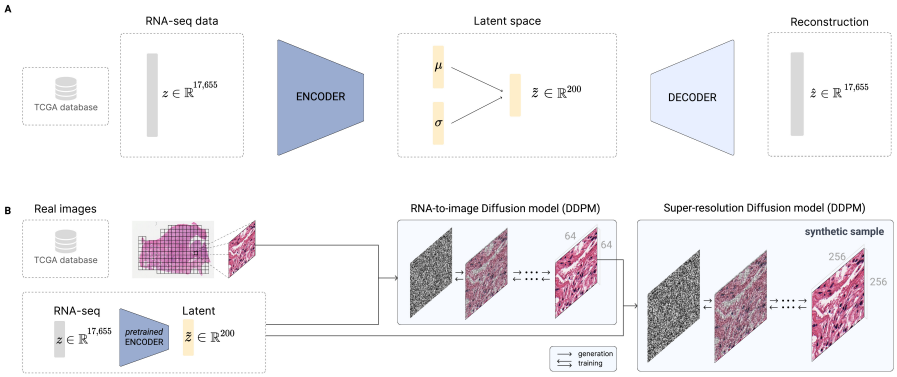


**Fig. 1 RNA-CDM model architecture used for the generation of RNA-Seq embeddings and synthetic WSI tiles using diffusion models**. **Panel A:** $\beta$VAE architecture for the generation of gene expression embeddings. The model uses as input the expression of $17,655$ genes. Both the encoder and the decoder are formed by two linear layers of $6,000$ and $4000$ respectively. The latent $\mu$ and $\sigma$ vectors have a feature size of 200. **Panel B:** RNA-CDM architecture for the generation of synthetic multi-cancer tiles. It is formed by two Deep Probabilistic Diffusion Model (DDPM) models, one acting as a RNA-to-image model and the second one as a super-resolution model. The pre-trained encoder is used to obtain the latent representation of the RNA-Seq data. A corresponding tile from the patient is obtained. During the training phase, noise is applied to the tile according to the given noise scheduler at a timestep $t$, and it is resized from 256x256 to 64x64. Then, the first DDPM learns to reduce the noise by using as input the noisy image, the timestep $t$, and the gene expression embedding. The noise predicted is removed from the noisy image, having a denoised tile of 64x64. Then, a second DDPM takes the denoised image, the noisy 256x256 image at timestep $t$, the timestep $t$, and the gene expression embedding and predicts the added noise again. Then, the noise is removed from the 256x256 tile and the image is compared to the original tile. For generating a new image, the process is the same, but we start from total random noise until we have a synthetic tile whose generation has been guided by the gene expression embedding.

pretraining models which boost the performance on a multi-cancer classification task, thereby showing the potential of generated synthetic tiles for both boosting the pretraining DL models and imputing missing modalities.

# Results

## RNA-CDM can perform realistic RNA-to-image multi-cancer synthesis

Given the high dimensionality of the RNA-Seq data, directly using it for conditioning the diffusion process is not possible. Therefore, inspired by text-to-image synthesis work, we trained a $\beta$VAE to project the RNA-Seq data from twelve different cancer tissues into a lower-dimensional latent space (Table 1). We decided to use a $\beta$VAE, given its superior performance in the literature, particularly in comparison to standard autoencoders [27]. The $\beta$VAE obtained a root mean squared error of 0.2475 and a mean absolute error of 0.1471 in

the test set (see Methods for details). We visualized the reconstructed RNA-Seq data using the Uniform Manifold Approximation and Projection (UMAP) algorithm [36] showing different clusters for each cancer tissue (Figure 2).

Next, we trained the RNA-CDM model using multimodal data from five cancer tissues: lung adenocarcinoma (LUAD), kidney renal papillary cell carcinoma (KIRP), cervical squamous cell carcinoma (CESC), colon adenocarcinoma (COAD) and glioblastoma (GBM). We trained the model for $> 50k$ steps and decided to stop the training when no visual quality gains were obtained based on the observable quality of the tissue and distinct morphological characteristics. Our RNA-CDM model was able to accurately generate tiles from the five cancer types without any tissue label information and only conditioning on the latent representation of the RNA-Seq data (Figure 2). Various tissue morphology can be observed in the generated tiles, from more homogeneous and cell-abundant (e.g. all the different types of glial cells) GBM tissue to the more sparse LUAD tissue. For all generated tiles, cell nuclei and muscle fibers can be distinguished in the tiles.

## RNA-CDM tiles maintain cell distribution of real tiles

Next, we used HoverNet [37], a state-of-the-art cell segmentation and classification model, to detect different cell types in cancer tissues. We used the weights trained on the PanNuke dataset to detect the following cell types: tumor, lymphocytes, connective, dead, and normal cells. We generated 1000 synthetic samples using multiple RNA-Seq profiles and obtained 1000 tiles from real whole-slide images for each of the five tumor tissues. We found that the distribution of cells was similar across the five cancer tissues (Figure 3 A). For example, the normal cell distribution is comparable across the cancer types, showing that a low percentage of normal cells is found in both real and synthetic tiles. Nevertheless, the percentage of normal cells was lower in synthetic tiles for all cancer types except TCGA-COAD and TCGA-CESC (TCGA-CESC 2.85% $\pm$ 10.87 real, 4.38% $\pm$ 17.57 synthetic; TCGA-KIRP 10.37% $\pm$ 18.35, 2.12% $\pm$ 9.18; TCGA-LUAD 3.65% $\pm$ 9.18, 0.85% $\pm$ 5.41; TCGA-GBM 12.50% $\pm$ 31.34, 0.30% $\pm$ 2.14; TCGA-COAD 3.94% $\pm$ 14.41, 5.68% $\pm$ 21.24). Next, we observed that slightly more tumor cells were found in real tiles, and more connective tissue was found in the synthetic tiles. Even though HoverNet was only trained on real data, cell types were correctly detected in the synthetic tiles, showing that RNA-CDM can produce samples with realistic cell morphology (Figure 3 B).

## RNA-CDM synthetic samples can be used as pretraining to improve classification performance

Next, we experimented with the utility of synthetic data for training DL models in settings of varying availability of real training data. To do so, we used $5,000$ tiles, consisting either of 100% real tiles or of which a certain percentage (25%, 50%, 75%) is replaced with synthetic tiles (simulating situations with fewer

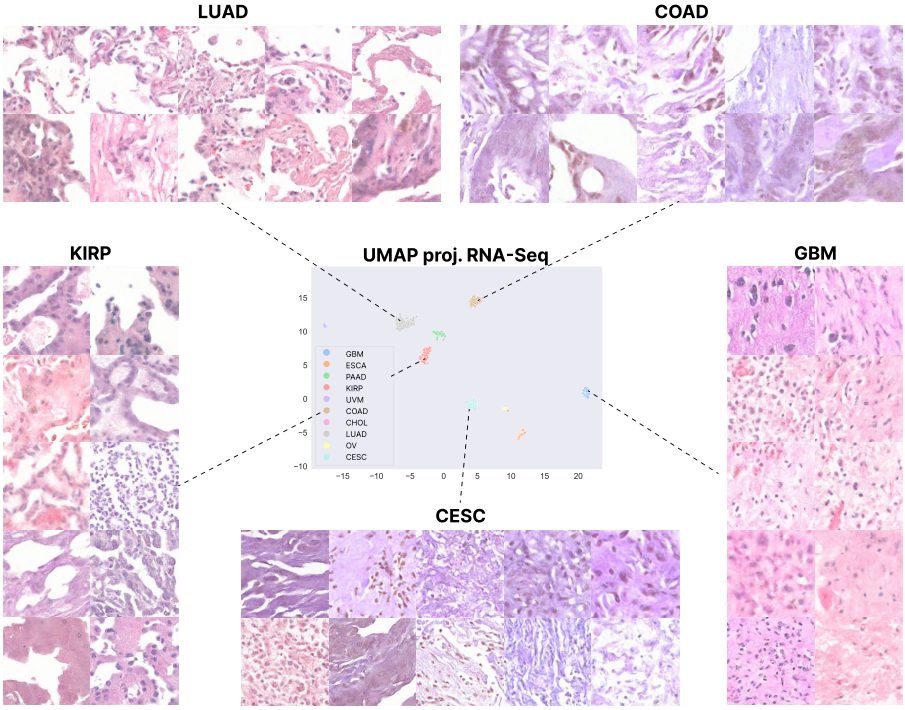


**Fig. 2 RNA-to-image multi-cancer synthetic samples generated by conditioning on the gene expression latent representation**. The $\beta$VAE is able to obtain an accurate representation of the multi-cancer gene expression profiles. Once this model has been trained, the RNA-CDM architecture (see Figure 1 and Subsection 5) is trained and conditioned on the latent representation of the RNA-Seq data over five different cancer types: lung adenocarcinoma (LUAD), kidney renal papillary cell carcinoma (KIRP), cervical squamous cell carcinoma (CESC), colon adenocarcinoma (COAD) and glioblastoma (GBM)). The model accurately synthesizes the samples, capturing the distinct morphologic characteristics of each cancer type.

available real data). We evaluated the performance of classifying each tile into one of five different tissue classes using a ResNet18 model and 5-fold cross validation (5-Fold CV).

We show that replacing 25% of real data with synthetic data does not reduce the accuracy of a five-way multi-class classification problem with mean accuracy of $0.535 \pm 0.037$ vs. $0.501 \pm 0.024$ (Figure 4 A). For the scenarios with 50% and 75% synthetic data we see an 9% and $8,7\%$ mean decrease in the classification performance respectively. This finding illustrates the potential of adding synthetic data to supplement real data for cancer type classification.

Next, we also experimented with using exclusively synthetic data for pre-training a classification model. We used four different sizes for the pretraining dataset 25% (1,250 samples), 50% (2,500 samples), 75% (3,750 samples) and 100% (5,000 samples) and compared how this affected the classification performance on a real sample dataset.

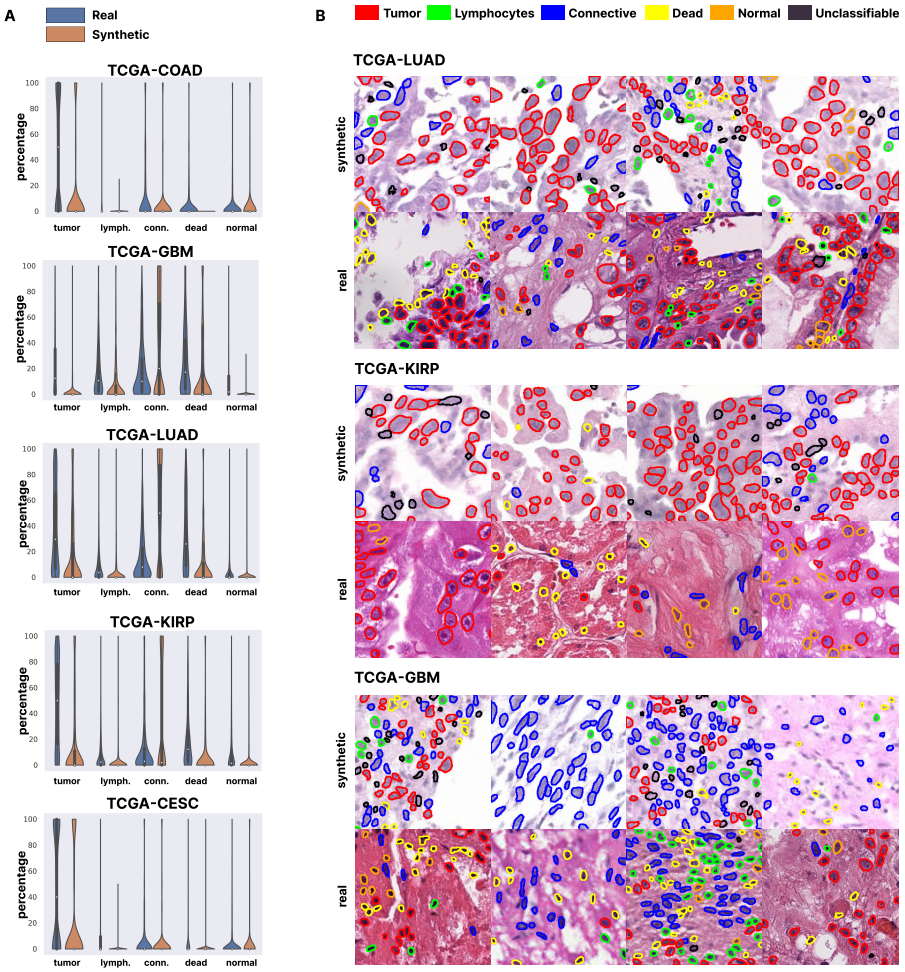


**Fig. 3 Synthetic samples maintain the cell distributions observed in real world data Panel A:** Distribution of cell types is maintained across the different cancer types in both real and synthetic tiles generated using RNA-CDM. We used HoverNet pretrained weights on the PanNuke dataset to detect five different cell types (tumor, lymphocytes, connective, necrotic, and normal). Normal in the PanNuke dataset includes cells from normal to degenerative, metaplastic, atypia etc. $10,000$ tiles were used in total ($5,000$ real and $5,000$ synthetic), with $1,000$ tiles per cancer type class. **Panel B:** Examples of cells detected in real and synthetic tiles by HoverNet. Cells are detected in both cases, showing that cell morphology is maintained in the RNA-CDM generated tiles.

We found that using the synthetic data for pretraining improved the model's performance over the baseline, regardless of the size of the pretraining dataset, for both the accuracy and F1-score (see Figure 4 B). The performance of the model also increased with the size of the synthetic dataset, showing that pretraining with more synthetic samples boosted the model's performance on the real dataset. To further quantify the observed improvement, we calculated

confusion matrices over all of the real samples for the model with and without pretraining (Figure 4 C).

## Discussion

Association of molecular and morphologic data of cancer tissues is emerging as an important research area in computational pathology and oncology [38]. Quantitative models have been developed to predict molecular biomarker status directly from H&E-stained WSI (including prediction of micro-statellite instability for colorectal cancer [39] and EGFR mutations for lung cancer [8]) opening an exciting avenue for the development of "digital biomarker" tests which might obviate the need for more expensive and time-consuming assays. However, a significant barrier to progress in the development of such tests is the relative paucity of tumor H&E specimens with corresponding ground truth labels availability for model training. While most patients have H&E tumor specimens collected as part of routine cancer diagnostic workflows, far fewer patients receive biomarker testing on those specimens, and when such biomarker tests are performed, an accompanying H&E-stained slide corresponding to the test sample might not be generated (for example, if the entire sample is exhausted during performance of the molecular assay). The ability of the RNA-CDM model in this study to generate synthetic but realistic H&E training images only from RNA-Seq expression data, and the demonstration that the resultant synthetic training data can be used to improve the performance of a downstream tumor classification model, offers a promising solution to the data scarcity problem.

While most models have focused on the detection of either discrete molecular changes, such as mutations in a specific gene or expression of a single protein, or classification of tumors into a limited number of discrete classes such as consensus molecular subtypes [40], few attempts have been made to explore how changes in the expression of a 1000s of genes might be reflected in routine H&E tumor histomorphology. Although spatial profiling methods (such as Akoya, 10x Genomics, etc) allow for some correlation between gene expression profiles and histology, these methods are expensive, currently impractical on a large scale and rely mainly on identification of pre-determined/handcrafted features or cell types that are recognizable to the human eye. Therefore, biologically important sub-visual morphologic features present on slides might be missed. Aside from their practical utility in data augmentation, models such as the RNA-CDM model developed in this study, which utilize latent representations of an entire RNA-Seq profile, might allow for the identification of novel morphologic features associated with clinically relevant molecular biological states that are currently unrecognized by the human eye.

Previous approaches for WSI cancer tile generation relied on GAN architectures and were able to generate realistic-appearing tissue images [23, 24]. However, GANs are prone to mode collapse and generate a low diversity of samples, limiting their potential for synthetic data generation that reflects

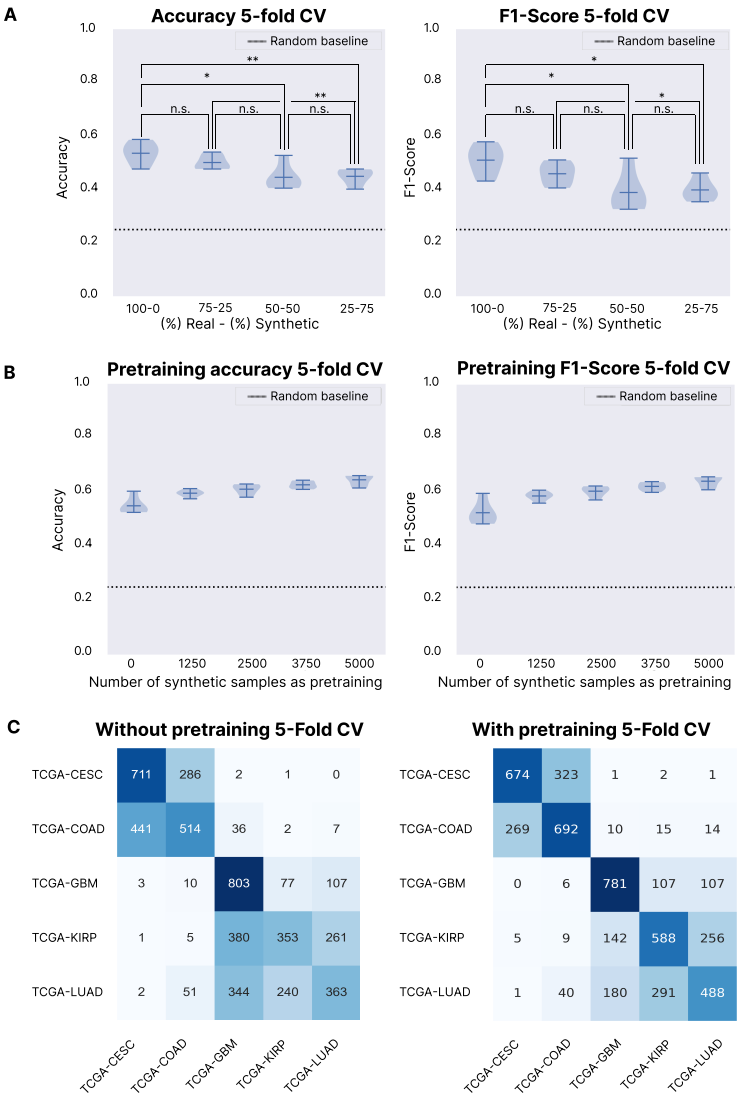


**Fig. 4 Pretraining on synthetic samples improves classification performance in a multi-cancer classification problem Panel A:** We substituted different percentages of the training data with synthetic samples generated by RNA-CDM in a 5-Fold CV over the real data (classification between the five different cancer types). The performance decreases over using 100% real data, but not significantly. Statistical significance is denoted between the groups within the plot (n.s. states for not significant, * states for p-value $< 0.05$, and ** states for p-value $< 0.01$). **Panel B:** We used different number of synthetic data samples as pretraining dataset. In both accuracy and F1-Score, increasing the number of samples in the pretraining dataset increases the performance in a 5-Fold CV over the real data. **Panel C:** (On the left) Confusion matrix of all the test sets in the 5-Fold CV without using any pretraining. (On the right) Confusion matrix of all the test sets in the 5-Fold CV using all the synthetic samples as pretraining. The model with the pretrained weights improves the classification performance and does not introduce any bias or new errors between the classes.

real-world data distributions [28–30]. Unlike natural images, cancer tissues are generally more homogeneous, requiring training of a different model for each cancer type in order to avoid mode collapse to a single class, which makes GANs impractical for generating cancer images in real-world scenarios. Furthermore, previous approaches for generating synthetic WSI have not made use of RNA expression data. Here, we present RNA-CDM, a single architecture based on cascaded diffusion models capable of performing RNA-to-image multi-cancer synthesis without any explicit label information (e.g. the cancer type), just using the expression profile of the patient. Because RNA-CDM uses only the latent representation obtained using the $\beta$VAE model (Figure 1 A) and requires only a single architecture to do this, it is more efficient and practical for synthetic data generation. This means that multiple models do not need to be implemented, saving on computational and storage costs.

Next, we show that the cell distribution trends in the synthetic and real tiles are maintained, showing more tumor cells and almost no normal cells in synthetic tiles. However, in the TCGA dataset, not all of the tissue contained tumor cells, and therefore, normal cells could be present in a tile. Our model was also able to synthesize normal cells, as they exist in real data (Figure 3 B). Nevertheless, RNA-CDM synthesizes normal cells in less proportion than they are found in real tiles, which is desirable. We hypothesize this can be due to the latent presentation that is obtained with the $\beta$VAE. To obtain a representative latent space, the model likely is using differentially expressed genes between the different cancer types. These genes can represent tumor changes in each tissue, and therefore, the RNA-to-image synthesis shows fewer normal cells than in real tiles.

Next, we focused on one of the main utilities of synthetic data, increasing the size of small datasets to boost the performance of machine learning models [41]. We first tested whether substituting real data with synthetic data impacted the performance of a multi-cancer classification model. There was no significant difference between using 100% real data and 75%real + 25% synthetic data in the training set (Figure 4 A). We also evaluated the benefit of using synthetic data for model pretraining. Both accuracy and F1-Score increased, irrespective of the size of the pretraining dataset (Figure 4 B). These results are promising and show that increasing the number of synthetic samples in the pretraining set directly improves the classification performance. The classification confusion matrices also showed that the fine-tuned model improves the model trained from scratch while maintaining the class-specific misclassifications, so we are not introducing new biases or errors with the synthetic data (Figure 4). Note that, theoretically, we can generate infinite number of samples as such the full potential of using synthetic data is likely still unknown. This work complements other approaches that are being explored for learning in scarce data settings, such as meta-learning or self-supervised learning techniques [12, 42].

In summary, we have proposed a solution for data scarcity in machine learning problems by using RNA-to-image synthetic data. How changes in gene

expression affect the tissue can also be studied in-silico by using RNA-CDM, exploring the interactions between the two modalities without the need of generating or finding new samples. New technologies are emerging, like spatial transcriptomics [43], that allow having a spatial map of the gene expression across the tissue. Unfortunately, these data are scarce, but we expect that it will be more accessible in the future and will expand the capabilities of RNA-CDM.

# Materials and Methods

## Data acquisition and data pre-processing

Data were obtained from the TCGA project database, which contains paired samples of RNA-Seq and WSIs from patients. For the WSI data, only diagnostic slides were considered, given the superior histologic quality of the sample. For the gene expression data, we downloaded the raw files for later pre-processing steps. Twelve cancer types were considered for the training of the $\beta$-VAE, while five were used during the training and validation of RNA-CDM (TCGA-LUAD, TCGA-GBM, TCGA-KIRP, TCGA-CESC, TCGA-COAD). The cancer types used and the number of available samples are presented in Table 1.

**Table 1** TCGA projects used for the training of the $\beta$-VAE and the RNA-CDM model.

| Project Code | Cancer type | Num. samples |
|---|---|---|
| TCGA-LUAD | Lung Adenocarcinoma | 520 |
| TCGA-KIRP | Kidney renal papillary cell carcinoma | 298 |
| TCGA-COAD | Colon adenocarcinoma | 289 |
| TCGA-CESC | Cervical squamous cell carcinoma and endocervical adenocarcinoma | 277 |
| TCGA-GBM | Glioblastoma multiforme | 212 |
| TCGA-PAAD | Pancreatic adenocarcinoma | 202 |
| TCGA-ESCA | Esophageal carcinoma | 156 |
| TCGA-OV | Ovarian serous cystadenocarcinoma | 83 |
| TCGA-UVM | Uveal Melanoma | 80 |
| TCGA-CHOL | Cholangiocarcinoma | 36 |

Once the data were downloaded, we proceeded with the pre-processing steps. For the gene expression data, we followed the pre-processing steps described by Zheng et al. [44]. The raw sequencing reads were aligned to the human transcriptome and quantified using the Kallisto-Sleuth algorithm proposed by Bray et al. [45]. NaN values were removed and we selected those genes the different cancer types have in common. These pre-processing steps left us with a total of $17,655$ genes that were used as input to the model. The gene expression was normalized with the z-score transformation from the training set values and log-transformed.

WSIs were acquired in SVS format and downsampled to $20\times$ magnification ($0.5\mu m$ px$^{-1}$). The size of WSIs is usually over $10k \times 10k$ pixels, and therefore,

cannot be directly used for training machine learning models to generate the data. Instead, tiles of a certain dimension are taken from the tissue, and they are used to train the models, which is consistent with related work in state-of-the-art WSI processing [8, 46, 47]. In our work, we took non-overlapping tiles of $256 \times 256$ pixels. Firstly, a mask of the tissue in the higher resolution of the SVS file was obtained using the Otsu threshold method [48]. Tiles containing more than 60% of the background and with low contrast were discarded. A maximum of $4,000$ tiles were taken from each slide. For the preprocessing of the images we relied on the python package openslide [49], which allows us to efficiently work with WSI images. The tiles were saved in an LMDB database using as an index the number of the tile. This approach enables us to reduce the number of generated files, and structure the tiles in an organized way for a faster reading while training. Tiles containing pen marks or other artifacts were filtered during the reading phase.

## $\beta$-VAE for multi-cancer latent embedding generation

We chose the $\beta$-VAE model for the creation of a latent embedding [50]. The $\beta$-VAE model is an extension of the VAE where a $\beta$ parameter is introduced in the loss function. VAEs are an modification of the original autoencoder, but they are both formed by two networks, the encoder and the decoder. The idea behind the autoencoder is to learn a smaller representation of the input data by learning the function $h_\theta(x) \approx x$ being $\theta$ the parameters of the neural network. We obtain a lower dimensionality representation of the data, and then we reconstruct the input using the decoder. Thus, we want to minimize the reconstruction error between the input and the output. On the other hand, the VAE extends this approach to learn a probability distribution of the latent space. The assumption of the VAE is that the distribution of the data $x$, $P(x)$ is related to the distribution of the latent variable $z$, $P(z)$. The loss function of the VAE, which is the negative log-likelihood with a regularizer is as follows:

$$L_i(\theta, \phi) = -\mathbb{E}_{z \sim q_\theta(z|x_i)}[\log p_\phi(x_i|z)] + \mathbb{KL}(q_\theta(z|x_i) \| p(z)) \tag{1}$$

where the first term is the reconstruction loss and the second term is the Kullback-Leibler (KL) divergence between the encoder's distribution $q_\theta(z|x)$ and $p(z)$ which is defined as the standard normal distribution $p(z) = N(0, 1)$.

For the $\beta$VAE we introduce the parameter $\beta$, which controls the effect of the KL divergence part of the equation:

$$L_i(\theta, \phi) = -\mathbb{E}_{z \sim q_\theta(z|x_i)}[\log p_\phi(x_i|z)] + \beta \times \mathbb{KL}(q_\theta(z|x_i) \| p(z)) \tag{2}$$

If $\beta = 1$, we have the standard loss of the VAE. If $\beta = 0$, we would only focus on the reconstruction loss, approximating the model to a normal autoencoder. For the rest of the values, we are regularizing the effect of the KL divergence on the training of the model, making the latent space smoother and more disentangled [50].

For the final architecture, we empirically determined to use two hidden layers of $6,000$ and 4000 neurons each for both the encoder and the decoder,

and size of 200 for the latent dimension, similar to the architecture proposed by Qiu et al. [27]. We used batch norm between the layers and the LeakyReLU as the activation function. A $\beta = 0.005$ was used in the loss function. We used the Adam optimizer for the training with a learning rate equal to $3 \times 10^{-3}$, along with a warm-up and a cosine learning rate scheduler and the mean square error as the loss function. We trained the model for 250 epochs with early stopping based on the validation set loss, and a batch size of 128. A schema of the architecture is presented in Figure 1 A. We divided the dataset into 60-20-20 % training, validation and test stratified splits, and we trained the model with twelve different cancer types in order to obtain a better and more general latent space (see Table 1).

## RNA-CDM: A Cascaded diffusion model for multi-cancer RNA-to-image synthesis

We present RNA-CDM, a cascaded diffusion model for multi-cancer RNA-to-image synthesis. Diffusion models are a kind of score-based generative models that model the gradient of the log probability density function using score matching [51, 52]. The idea for diffusion models is to learn a series of state transitions to map noise $\epsilon$ from a known prior distribution to $x_0$ from the data distribution. Firstly, we define an additive noise forward process from $x_0$ to $x_t$ defined as:

$$x_t = \sqrt{\gamma(t)}x_0 + \sqrt{1-\gamma(t)}\epsilon \tag{3}$$

where $\epsilon \sim \mathcal{N}(0, I)$, $t \sim \mathcal{U}(0, T)$, and $\gamma(t)$ is a monotonically decreasing function from 1 to 0. Then, we learn a neural network, $\{(x_t, t)$ to reverse this process by predicting $x_0$ (or $\epsilon$) from $x_t$. The training of the neural network is based on denoising with a $l_2$ regression loss:

$$\mathcal{L}_{x_0} = \mathbb{E}_{t\sim\mathcal{U}(0,T),\epsilon\sim\mathcal{N}(0,I)} \|f(\sqrt{\gamma(t)}x_0 + \sqrt{1-\gamma(t)}\epsilon, t) - \epsilon\|^2 \tag{4}$$

Once we have learned a model, new samples can be generated by reversely going from $x_t \rightarrow x_{t-n} \rightarrow ... \rightarrow x_0$. This can be achieved by applying the denoising function $f$ to the samples to obtain $x_0$, and then make the transition to $x_{t-n}$ by using the predicted $\hat{x}_0$ [53].

Cascaded diffusion models were proposed by Ho et al. [54] as a way to improve sample quality. Having high-resolution data $x_0$ and a low-resolution version $z_0$, we have a diffusion model at the low resolution $p_\theta(z_0)$, and a super-resolution diffusion model $p_\theta(x_0|z_0)$. The cascading pipeline forms a latent variable model for high resolution data, that can also be extended to conditioning to the class (or the gene expression latent representation in our case):

$$p_\theta(x_0) = \int p_\theta(x_0|z_0, c)p_\theta(z_0|c)dz_0 \tag{5}$$

where $c$ is the gene expression latent representation. Also, as in normal diffusion models, we condition on the timesteps both for the training and the sampling.

In this work, we used the extension presented by Karras et al. [55], where the training and sampling methods are modified. They proposed a stochastic sampling method, that hugely speeds the sampling process, one of the bottlenecks of diffusion models. More information about their approach can be read in Sections 4 and 5 of their paper [55]. We empirically selected a value of $\rho = 7$, $S_{churn} = 80$, $S_{tmin} = 0.05$, $S_{tmax} = 50$, $S_{noise} = 1.003$ and 32 steps for the stochastic sampling.

For the hyperparameters and architectures used, we followed those presented in the Imagen paper [33]. The Unet model [56] was chosen for the diffusion models, using a dimension of 128 for both the low-resolution and the super-resolution diffusion models. Attention and skip connections are used across the Unet layers. Adam was used as the optimizer with a learning rate equal to $1e^{-4}$, and an exponential moving average (EMA) was used during the training. The low-resolution diffusion models generate an image of 64x64, and it is conditioned on the RNA-Seq latent representation and the timestep. Then, a gaussian blur is applied to that image and it serves as input to the super-resolution diffusion model (along with the RNA-Seq latent embedding and the timestep), which generates an image of 256x256. For both Unets we used $timesteps = 1000$ with a linear gaussian diffusion process. The model was trained for 53k steps, deciding to stop the training based on the visual quality of the generated tissue. We used a maximum of randomly chosen 200 tiles per slide to train the model, in order to reduce the training time. A schematic representation of the training and sampling pipeline is presented in Figure 1 B.

For the ML experiments presented in the results section, we used a Resnet-50 pretrained on Imagenet (Figure 4 A) and a Resnet-18 architecture trained from scratch (Figure 4 B and C). For the first set of experiments, the model was fine-tuned for 20 epochs using AdamW optimizer with a learning rate equal to $3e^{-3}$. For the second set of experiment, the model was trained firstly on the synthetic data for 50 epochs using an early stopping strategy using the same optimizer and learning rate value. Then, it was fine-tuned in the real data for 20 epochs using a learning rate value of $3e^{-5}$.

**Acknowledgments.** The results published here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. FCP was supported by the Spanish Ministry of Sciences, Innovation and Universities under Projects RTI-2018-101674-B-I00 and PID2021-128317OB-I00, the project from Junta de Andalucia P20-00163 and a Predoctoral scholarship from the Fulbright Spanish Commission. Research reported here was further supported by the National Cancer Institute (NCI) under award: R01 CA260271. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. MP was supported by a fellowship from the Belgian American Educational Foundation.

# References

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2021;71(3):209–249.

[2] Jones PA, Baylin SB. The epigenomics of cancer. Cell. 2007;128(4):683–692.

[3] Lujambio A, Lowe SW. The microcosmos of cancer. Nature. 2012;482(7385):347–355.

[4] Frangioni JV. New technologies for human cancer imaging. Journal of clinical oncology. 2008;26(24):4012.

[5] Williams BJ, Bottoms D, Treanor D. Future-proofing pathology: the case for clinical adoption of digital pathology. Journal of clinical pathology. 2017;70(12):1010–1018.

[6] Heindl A, Nawaz S, Yuan Y. Mapping spatial heterogeneity in the tumor microenvironment: a new era for digital pathology. Laboratory investigation. 2015;95(4):377–384.

[7] Cheng J, Mo X, Wang X, Parwani A, Feng Q, Huang K. Identification of topological features in renal tumor microenvironment associated with patient survival. Bioinformatics. 2018;34(6):1024–1030.

[8] Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nature medicine. 2018;24(10):1559–1567.

[9] Castillo D, Gálvez JM, Herrera LJ, Román BS, Rojas F, Rojas I. Integration of RNA-Seq data with heterogeneous microarray data for breast cancer profiling. BMC bioinformatics. 2017;18(1):1–15.

[10] Yu D, Liu Z, Su C, Han Y, Duan X, Zhang R, et al. Copy number variation in plasma as a tool for lung cancer prediction using Extreme Gradient Boosting (XGBoost) classifier. Thoracic cancer. 2020;11(1):95–102.

[11] Maros ME, Capper D, Jones DT, Hovestadt V, von Deimling A, Pfister SM, et al. Machine learning workflows to estimate class probabilities for precision cancer diagnostics on DNA methylation microarray data. Nature protocols. 2020;15(2):479–512.

[12] Chen RJ, Chen C, Li Y, Chen TY, Trister AD, Krishnan RG, et al. Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised

Learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022. p. 16144–16155.

[13] Carrillo-Perez F, Morales JC, Castillo-Secilla D, Gevaert O, Rojas I, Herrera LJ. Machine-Learning-Based Late Fusion on Multi-Omics and Multi-Scale Data for Non-Small-Cell Lung Cancer Diagnosis. Journal of Personalized Medicine. 2022;12(4):601.

[14] Lee C, van der Schaar M. A variational information bottleneck approach to multi-omics data integration. In: International Conference on Artificial Intelligence and Statistics. PMLR; 2021. p. 1513–1521.

[15] Chen RJ, Lu MY, Wang J, Williamson DF, Rodig SJ, Lindeman NI, et al. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. IEEE Transactions on Medical Imaging. 2020;.

[16] Cheerla A, Gevaert O. Deep learning with multimodal representation for pancancer prognosis prediction. Bioinformatics. 2019;35(14):i446–i454.

[17] Chen RJ, Lu MY, Williamson DF, Chen TY, Lipkova J, Noor Z, et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. Cancer Cell. 2022;40(8):865–878.

[18] Vanguri RS, Luo J, Aukerman AT, Egger JV, Fong CJ, Horvat N, et al. Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L) 1 blockade in patients with non-small cell lung cancer. Nature cancer. 2022;p. 1–14.

[19] Lipkova J, Chen RJ, Chen B, Lu MY, Barbieri M, Shao D, et al. Artificial intelligence for multimodal data integration in oncology. Cancer Cell. 2022;40(10):1095–1110.

[20] Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, et al. The cancer genome atlas pan-cancer analysis project. Nature genetics. 2013;45(10):1113–1120.

[21] Jennings CN, Humphries MP, Wood S, Jadhav M, Chabra R, Brown C, et al. Bridging the gap with the UK Genomics Pathology Imaging Collection. Nature Medicine. 2022;p. 1–2.

[22] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic acids research. 2012;41(D1):D991–D995.

[23] Quiros AC, Murray-Smith R, Yuan K. PathologyGAN: Learning deep representations of cancer tissue. arXiv preprint arXiv:190702644. 2019;.

[24] Quiros AC, Murray-Smith R, Yuan K. Learning a low dimensional manifold of real cancer tissue with PathologyGAN. arXiv preprint arXiv:200406517. 2020;.

[25] Viñas R, Andrés-Terré H, Liò P, Bryson K. Adversarial generation of gene expression data. Bioinformatics. 2022;38(3):730–737.

[26] Mitra R, MacLean AL. RVAgene: generative modeling of gene expression time series data. Bioinformatics. 2021;37(19):3252–3262.

[27] Qiu YL, Zheng H, Gevaert O. Genomic data imputation with variational auto-encoders. GigaScience. 2020;9(8):giaa082.

[28] Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC. Improved training of wasserstein gans. Advances in neural information processing systems. 2017;30.

[29] Metz L, Poole B, Pfau D, Sohl-Dickstein J. Unrolled generative adversarial networks. arXiv preprint arXiv:161102163. 2016;.

[30] Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X. Improved techniques for training gans. Advances in neural information processing systems. 2016;29.

[31] Zhao S, Song J, Ermon S. Infovae: Balancing learning and inference in variational autoencoders. In: Proceedings of the aaai conference on artificial intelligence. vol. 33; 2019. p. 5885–5892.

[32] Ramesh A, Dhariwal P, Nichol A, Chu C, Chen M. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:220406125. 2022;.

[33] Saharia C, Chan W, Saxena S, Li L, Whang J, Denton E, et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. arXiv preprint arXiv:220511487. 2022;.

[34] Sohl-Dickstein J, Weiss E, Maheswaranathan N, Ganguli S. Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning. PMLR; 2015. p. 2256–2265.

[35] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. PMLR; 2021. p. 8748–8763.

[36] McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:180203426.

2018;.

[37] Graham S, Vu QD, Raza SEA, Azam A, Tsang YW, Kwak JT, et al. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. Medical Image Analysis. 2019;58:101563.

[38] Lehrer M, Powell RT, Barua S, Kim D, Narang S, Rao A. Radiogenomics and histomics in glioblastoma: the promise of linking image-derived phenotype with genomic information. In: Advances in Biology and Treatment of Glioblastoma. Springer; 2017. p. 143–159.

[39] Yamashita R, Long J, Longacre T, Peng L, Berry G, Martin B, et al. Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study. The Lancet Oncology. 2021;22(1):132–141.

[40] Marisa L, de Reyniès A, Duval A, Selves J, Gaub MP, Vescovo L, et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. PLoS medicine. 2013;10(5):e1001453.

[41] Chen RJ, Lu MY, Chen TY, Williamson DF, Mahmood F. Synthetic data in machine learning for medicine and healthcare. Nature Biomedical Engineering. 2021;5(6):493–497.

[42] Qiu YL, Zheng H, Devos A, Selby H, Gevaert O. A meta-learning approach for genomic survival analysis. Nature communications. 2020;11(1):1–11.

[43] Dries R, Chen J, Del Rossi N, Khan MM, Sistig A, Yuan GC. Advances in spatial transcriptomic data analysis. Genome research. 2021;31(10):1706–1718.

[44] Zheng H, Brennan K, Hernaez M, Gevaert O. Benchmark of long non-coding RNA quantification for RNA sequencing of cancer samples. GigaScience. 2019;8(12):giz145.

[45] Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nature biotechnology. 2016;34(5):525–527.

[46] Lu MY, Chen TY, Williamson DF, Zhao M, Shady M, Lipkova J, et al. AI-based pathology predicts origins for cancers of unknown primary. Nature. 2021;594(7861):106–110.

[47] Lu MY, Williamson DF, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering. 2021;5(6):555–570.

[48] Otsu N. A threshold selection method from gray-level histograms. IEEE transactions on systems, man, and cybernetics. 1979;9(1):62–66.

[49] Goode A, Gilbert B, Harkes J, Jukic D, Satyanarayanan M. OpenSlide: A vendor-neutral software foundation for digital pathology. Journal of pathology informatics. 2013;4.

[50] Higgins I, Matthey L, Pal A, Burgess CP, Glorot X, Botvinick MM, et al. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In: ICLR; 2017. .

[51] Hyvärinen A, Dayan P. Estimation of non-normalized statistical models by score matching. Journal of Machine Learning Research. 2005;6(4).

[52] Vincent P. A connection between score matching and denoising autoencoders. Neural computation. 2011;23(7):1661–1674.

[53] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems. 2020;33:6840–6851.

[54] Ho J, Saharia C, Chan W, Fleet DJ, Norouzi M, Salimans T. Cascaded Diffusion Models for High Fidelity Image Generation. J Mach Learn Res. 2022;23:47–1.

[55] Karras T, Aittala M, Aila T, Laine S. Elucidating the Design Space of Diffusion-Based Generative Models. arXiv preprint arXiv:220600364. 2022;.

[56] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer; 2015. p. 234–241.