RNA-to-Image synthesis: generating synthetic digital pathology tiles based on NGS data using deep generative models

Francisco Carrillo-Perez, Ph.D.
Stanford Center for Biomedical Informatics Research (BMIR), Stanford University, School of Medicine.

Gevaert's Lab

Outline

- Introduction
  - Digital pathology
  - Next-Generation Sequencing (NGS) data
  - Variational Autoencoders (VAEs)
  - Generative Adversarial Networks (GANs)
  - Diffusion Models
- Text-to-image models
  - Recent advances in text-to-image
  - Can this be applied to biomedical data?
- Synthetic whole-slide image tile generation with gene expression profiles infused deep generative models
- Generation of synthetic whole-slide image tiles of tumours from RNA-sequencing data via cascaded diffusion models
- Future Directions

My background

Universidad de Granada

(Sept 2013- March 2018)
Bachelor in Computer Science

mDurance

Universidad de Granada

(Sept 2018- July 2019)
Master in Data Science and Computer Engineering / Data Scientist

Universidad de Granada

(Nov 2019- Jan 2023)
Ph.D. in Machine Learning applied to Bioinformatics

FULBRIGHT

(Sept 2021- Sep 2022)
Awarded one of the 18 predoctoral Fulbright scholarships in Spain

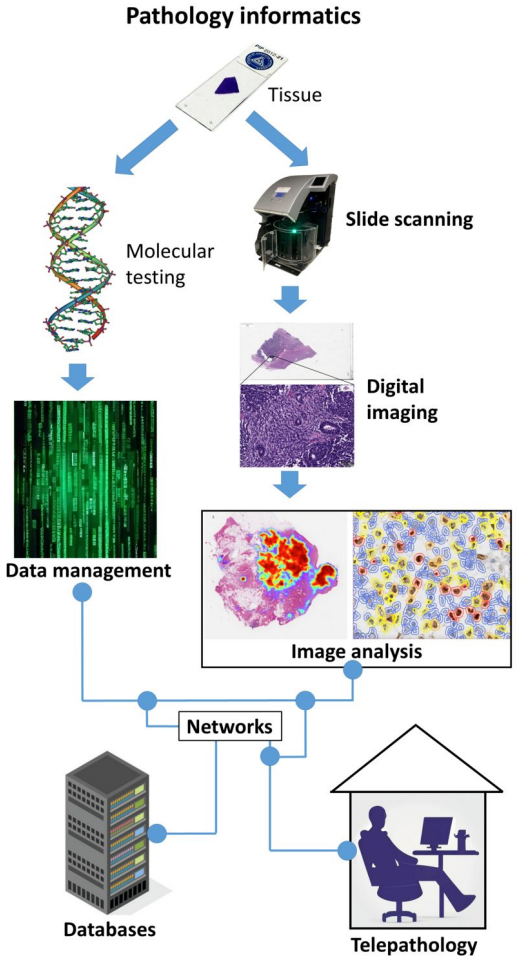
Stanford MEDICINE

(March 2023-)
Postdoctoral Researcher (remotely)

Introduction

Digital pathology

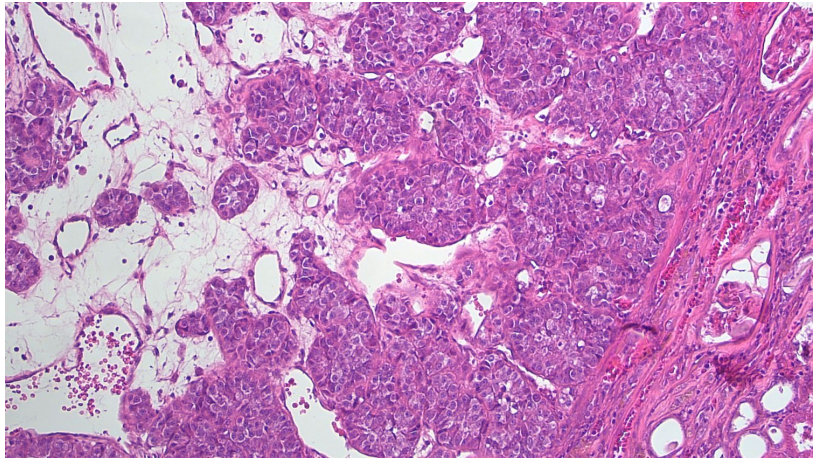
Digital pathology is a sub-field of pathology that focuses on data management based on information generated from digitized specimen slides.



https://en.wikipedia.org/wiki/Digital_pathology#/media/File:Major_topics_of_pathology_informatics.png

Next-generation sequencing data

Cancer tissue stained with hematoxylin & eosin (H&E) stain

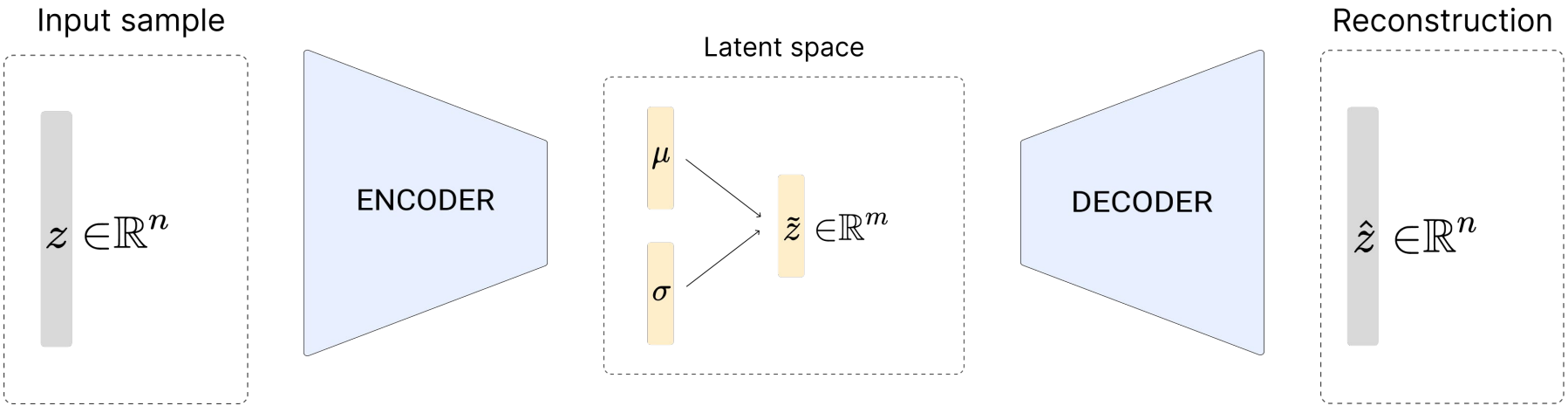


HTTPS://FOCUSONTOXPATH.COM/WP-CONTENT/UPLOADS/TOXICOLOGIC-PATHOLOGY-TISSUE-SLIDE.JPG

RNA-Seq sequencing

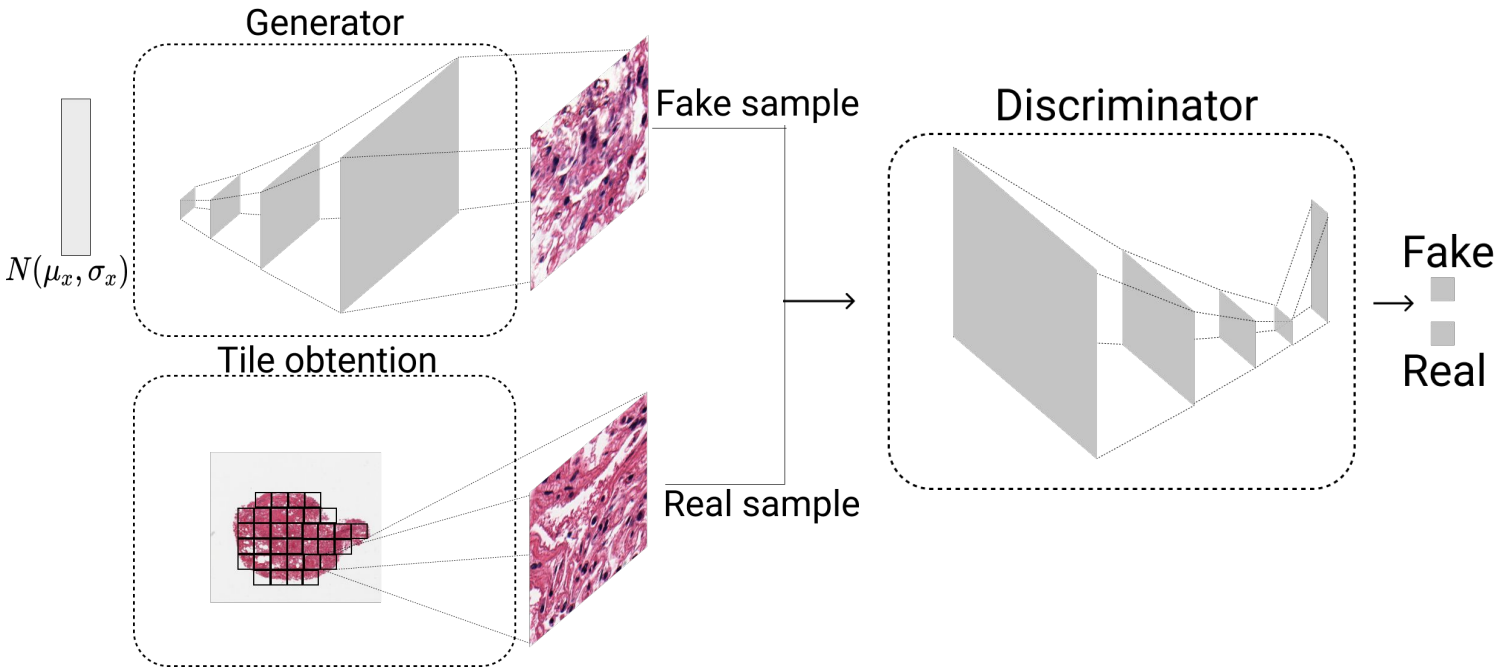
0.5	0.6	0.9	0.1	.....	1.1
TP53	*EGFR*	*KRAS*	*ALB*		*IDH1*

Variational Autoencoders (VAEs)



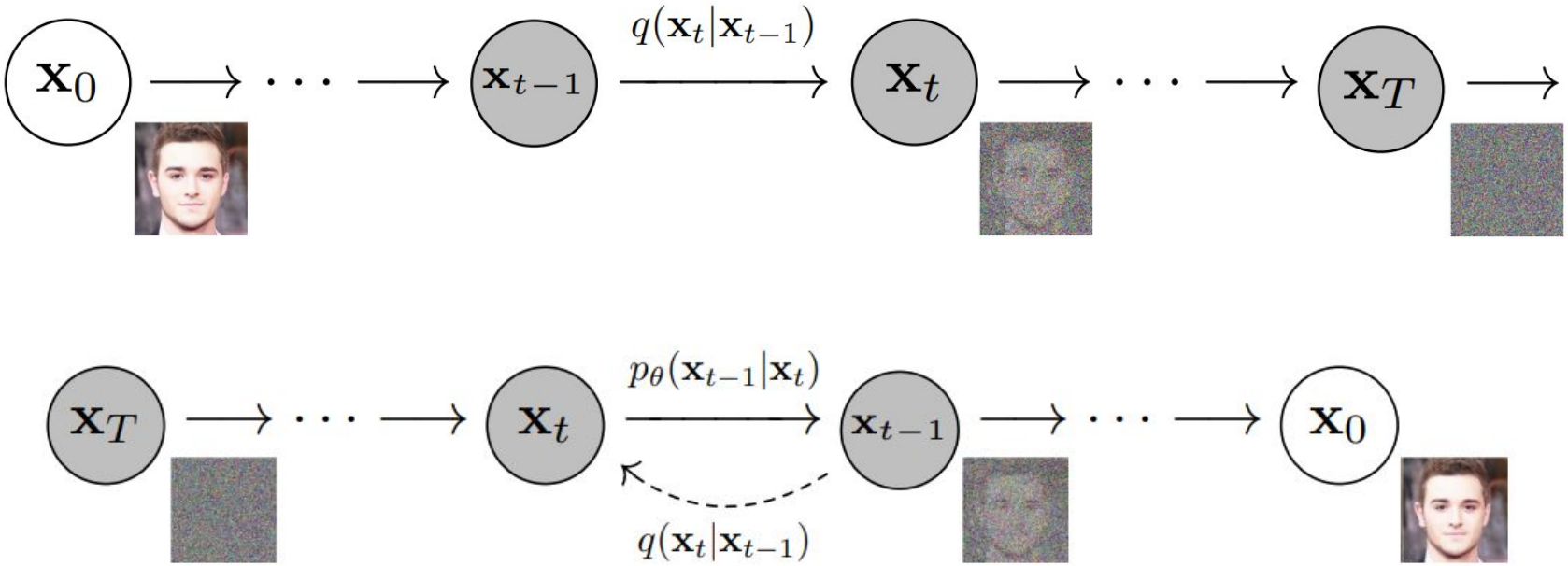
Bibliography: Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Generative Adversarial Networks (GANs)



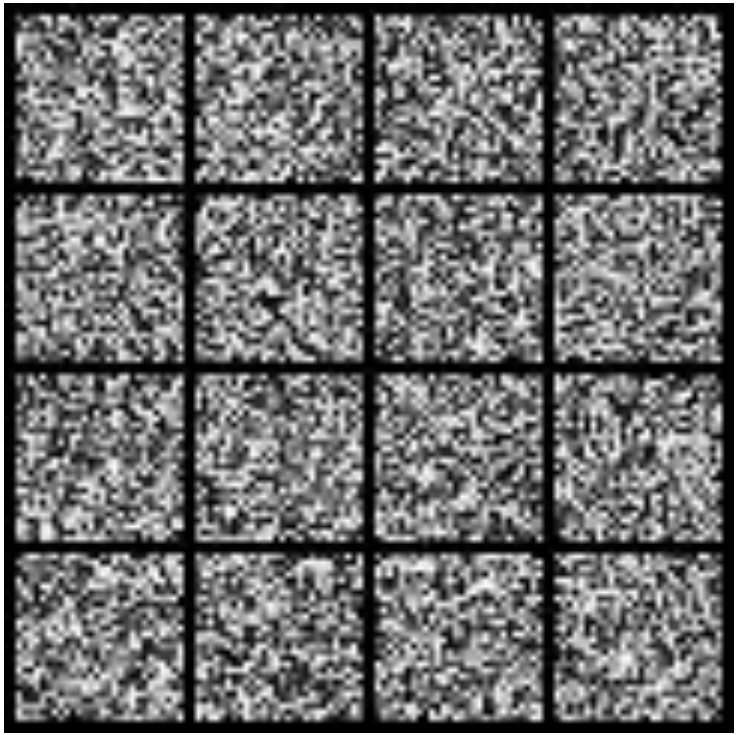
Bibliography: Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139-144.

Diffusion models



Bibliography: Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, *33*, 6840-6851.

Diffusion models



Bibliography: https://www.assemblyai.com/blog/diffusion-models-for-machine-learning-introduction/

Text-to-image models

Recent advances in text-to-image

DALLE-3 (Open AI)

Imagen (Google)

Midjourney (Midjourney)

Stable Diffusion 3 (Stability AI)

Many more...

How do they work?

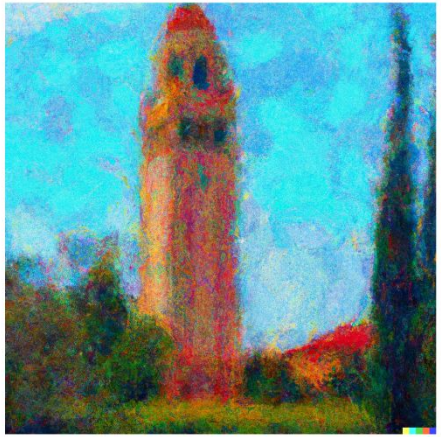
Text prompt

"A painting of the Stanford Hoover Tower in the style of Monet"

→

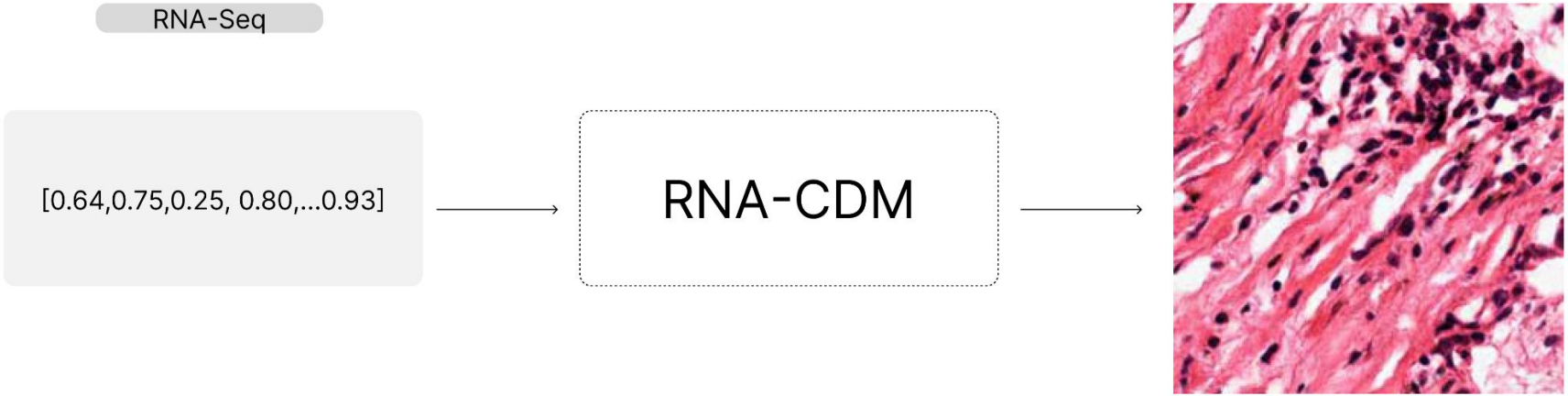
DALL-E 2

→



Can this be applied to biomedical data?

It is known that gene expression has an effect on tissue morphology (Fu et al. 2020;Schmauch et al. (2020); Zheng et al. (2023)). Tissue tiles (Formalin-Fixed Paraffin-Embedded (FFPE) tissue specimens) are routinely obtained, and a single bulk RNA-Seq expression is obtained for the whole FFPE



Cell Reports Methods

Volume 3, Issue 8, 28 August 2023, 100534

Article

Synthetic whole-slide image tile generation with gene expression profile-infused deep generative models

Francisco Carrillo-Perez [1,2], Marija Pizurica [1,3], Michael G. Ozawa [4], Hannes Vogel [4], Robert B. West [4], Christina S. Kong [4], Luis Javier Herrera [2], Jeanne Shen [4], Olivier Gevaert [1,5,6]

Motivation and objectives

Motivation:

- Several works presented single-modality generative models (Quiros et al. 2019; Marouf et. al. 2020)
- Not all datasets have both modalities, or have a scarce number of samples

Objectives: Create an RNA-to-image synthesis model to fight data scarcity on healthy tissues using GANs

- Obtain a informative latent representation of the RNA-Seq using a VAE
- Generate high-quality tissue tiles an RNA-informed GAN and a traditional GAN
- Compare the quality of tiles between using an RNA-informed GAN and a traditional GAN

Data acquisition

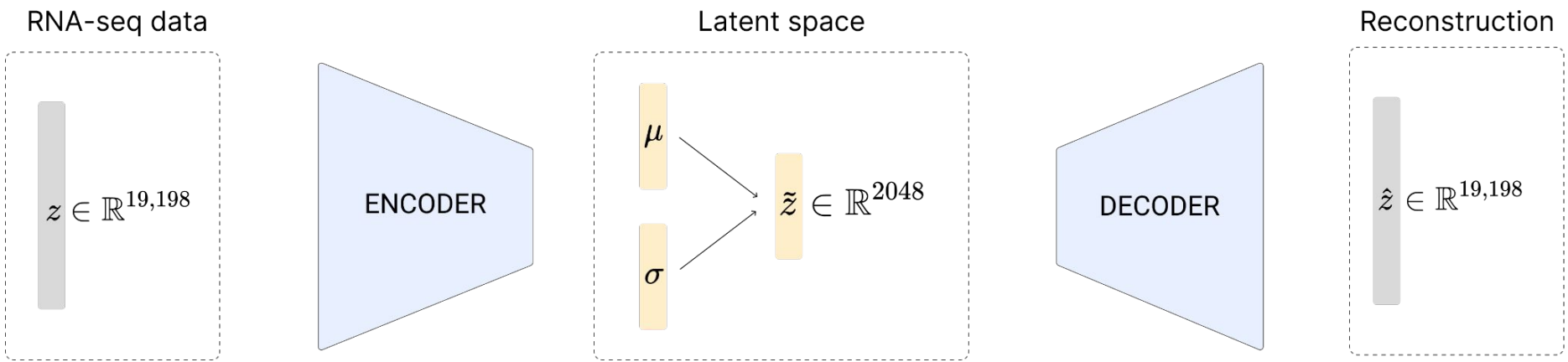
- RNA-Seq (more than 60,000 genes) and WSI obtained from The Genotype-Tissue Expression (GTEx) project

- 246 samples of brain cortex, 562 samples of lung tissue, 328 samples of pancreas tissue, 356 samples of stomach tissue, and 226 samples of liver tissue

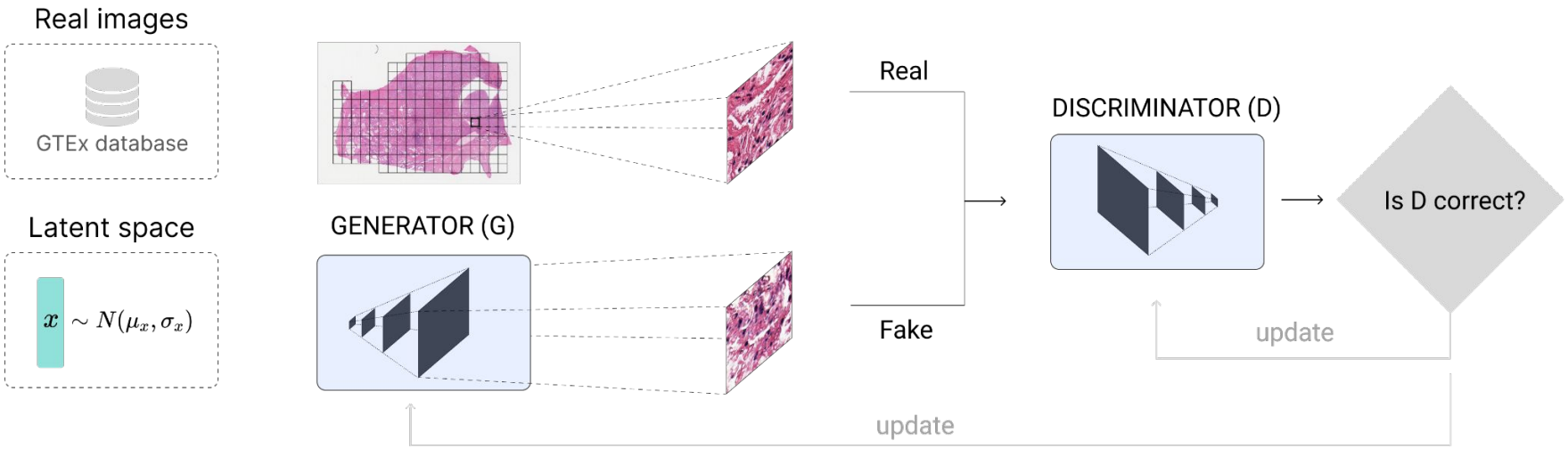
- Lung and brain cortex tissue used from the GEO serie 120795 for generalization capabilities

- We focused on generating two tissues: lung and brain cortex.

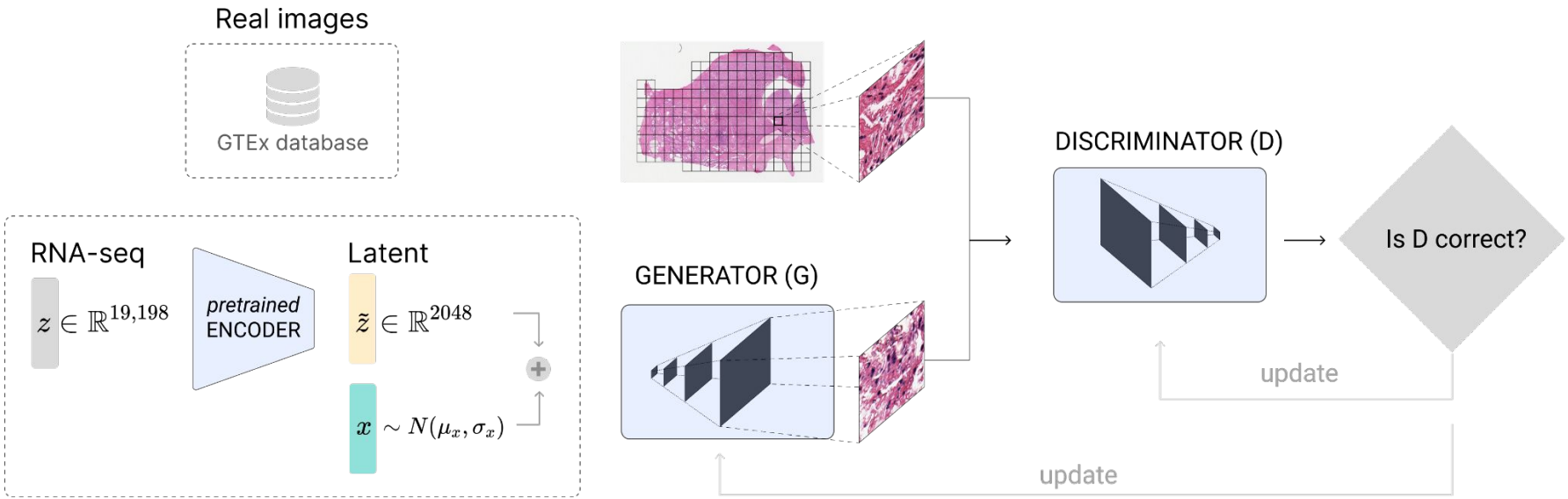
Methodology: VAE



Methodology: GAN



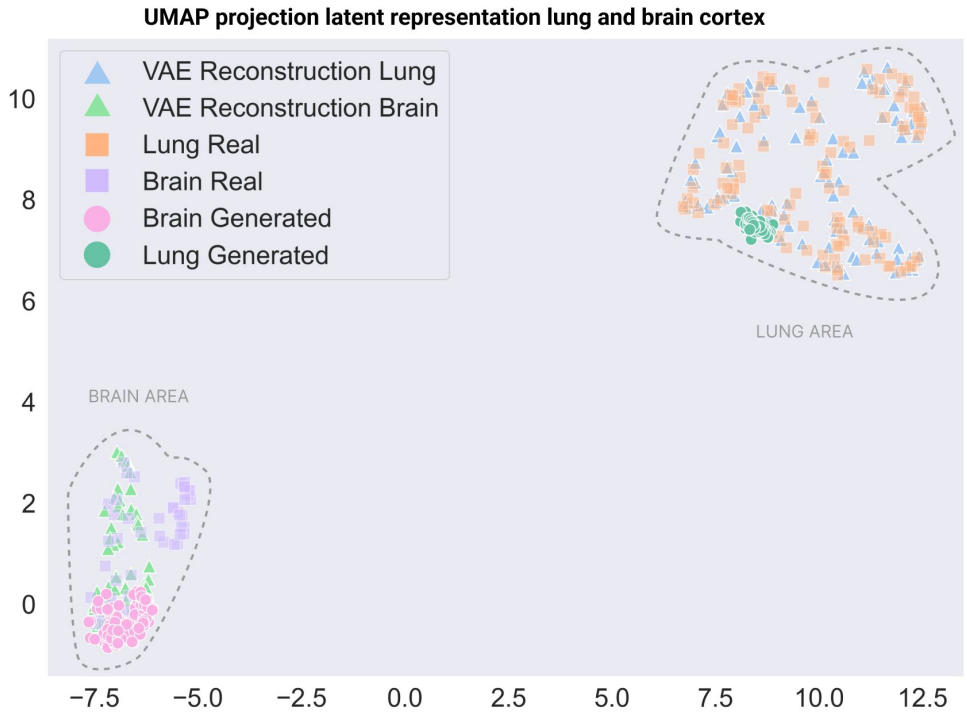
Methodology: RNA-GAN



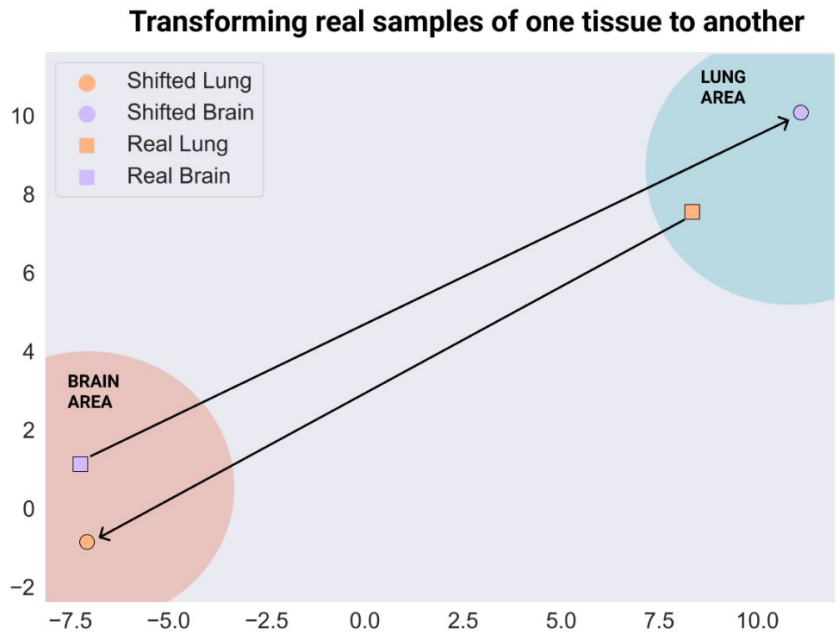
Results: VAE

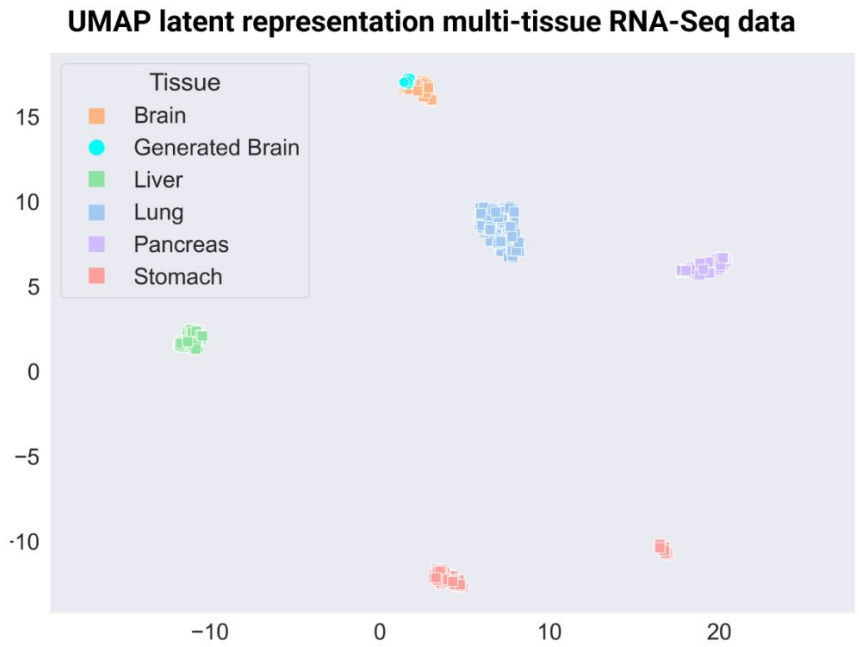
Reconstruction

$\hat{z} \in \mathbb{R}^{19,198}$



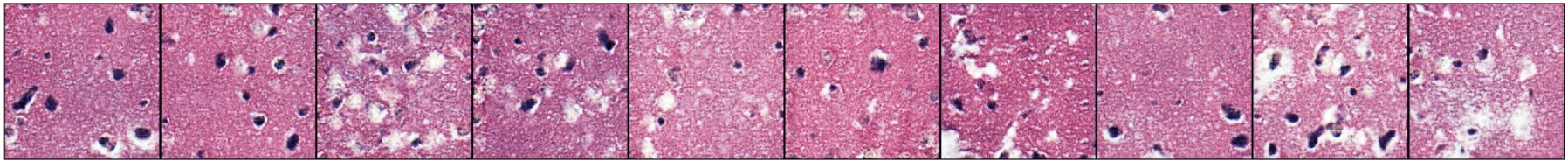
Results: VAE


Transforming real samples of one tissue to another
Shifted Lung
Shifted Brain
Real Lung
Real Brain
LUNG AREA
BRAIN AREA
10
8
6
4
2
0
−2
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

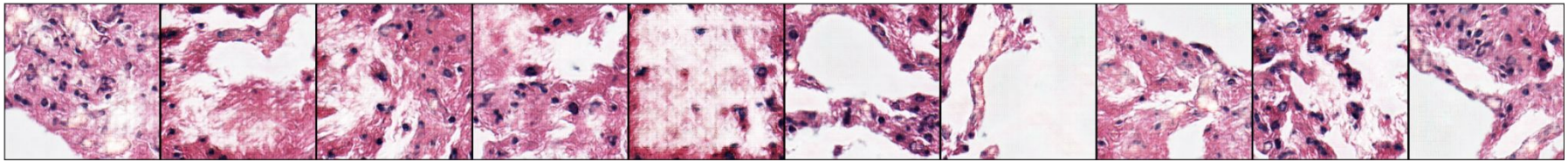

UMAP latent representation multi-tissue RNA-Seq data
Tissue
Brain
Generated Brain
Liver
Lung
Pancreas
Stomach
15
10
5
0
−5
-10
−10
0
10
20

Results: GAN

GAN Brain

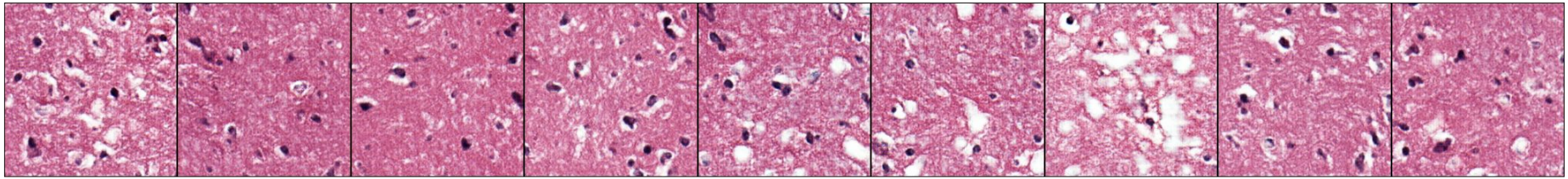


GAN Lung

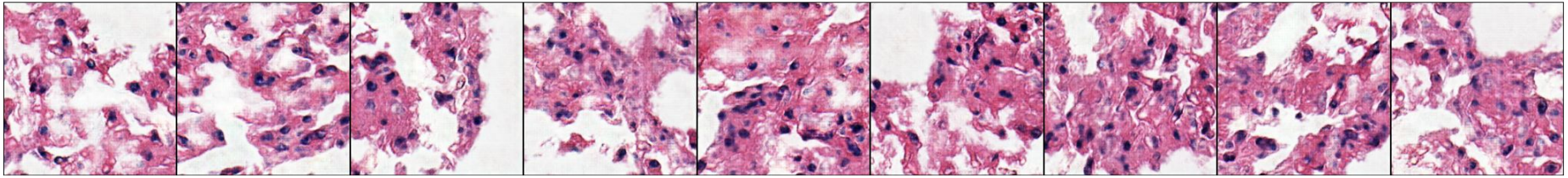


Results: RNA-GAN

RNA-GAN Brain



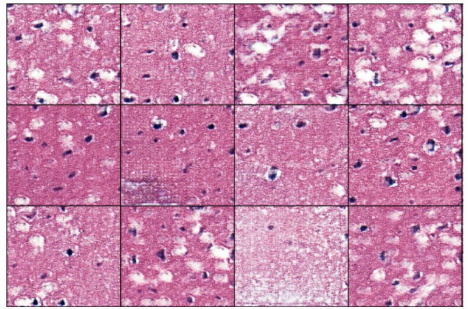
RNA-GAN Lung



Results: Training time comparison

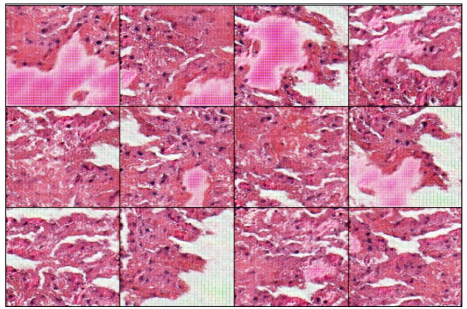
Epoch 24/39

GAN Brain



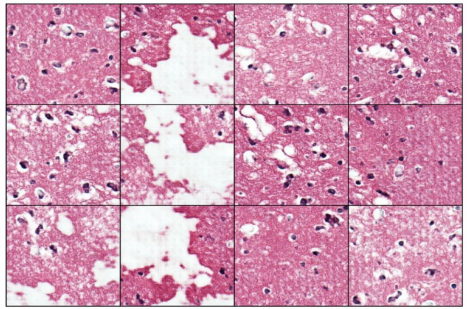
Epoch 11/91

GAN Lung



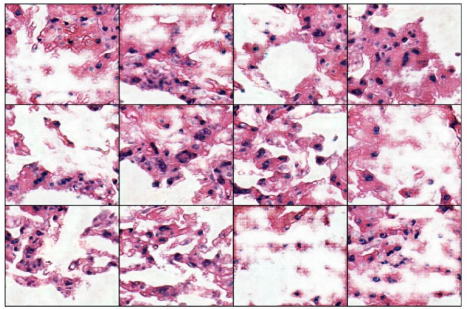
Epoch 24/24

RNA-GAN Brain



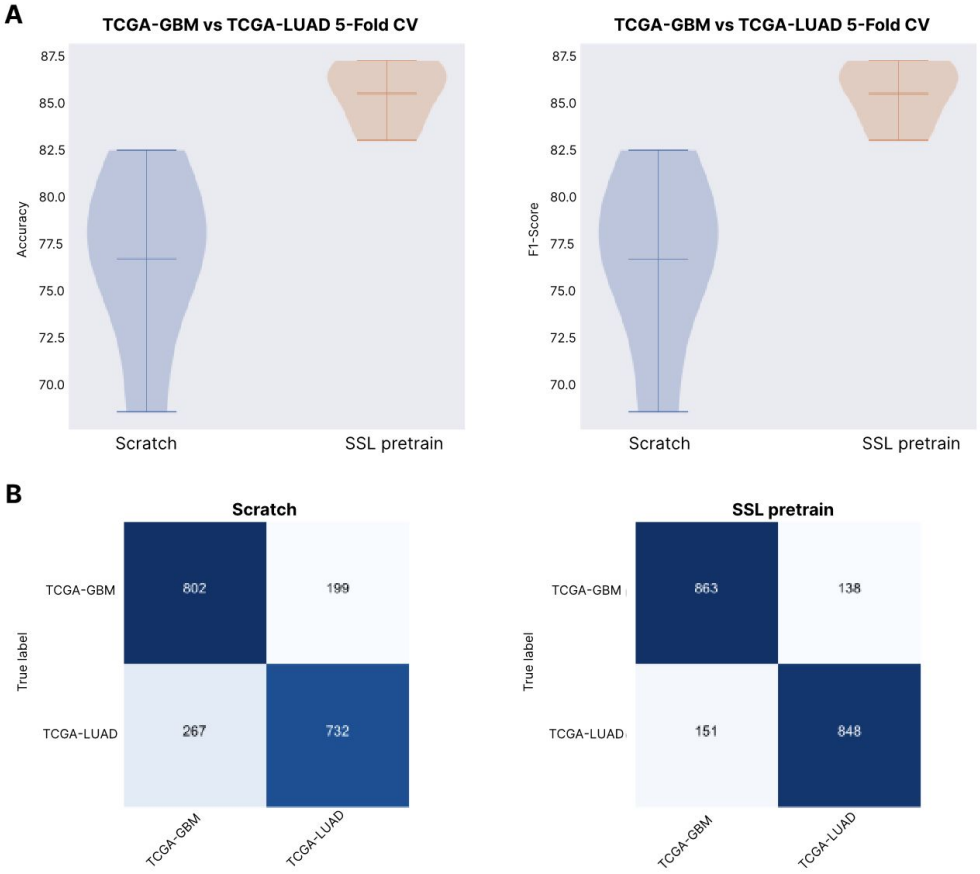
Epoch 11/11

RNA-GAN Lung

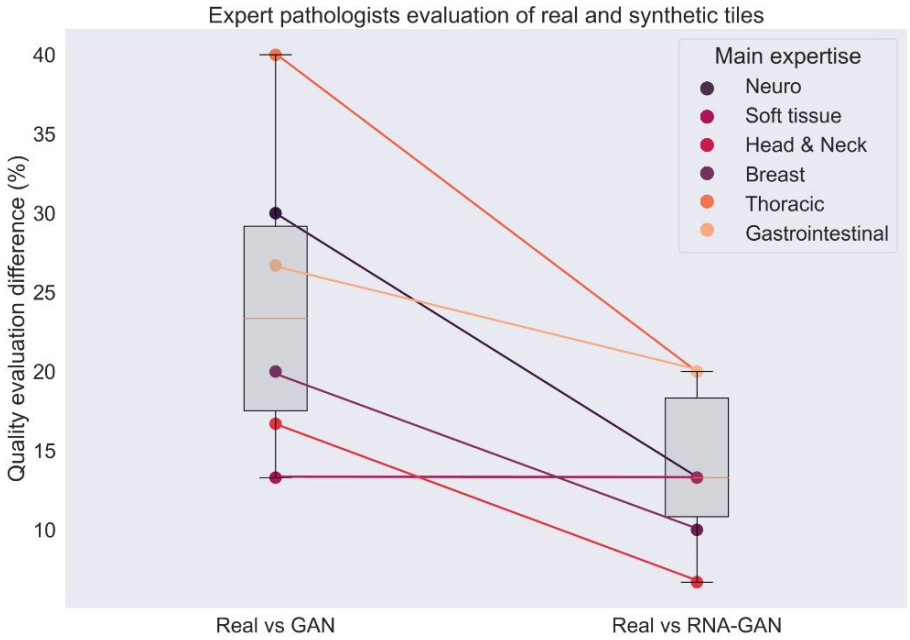


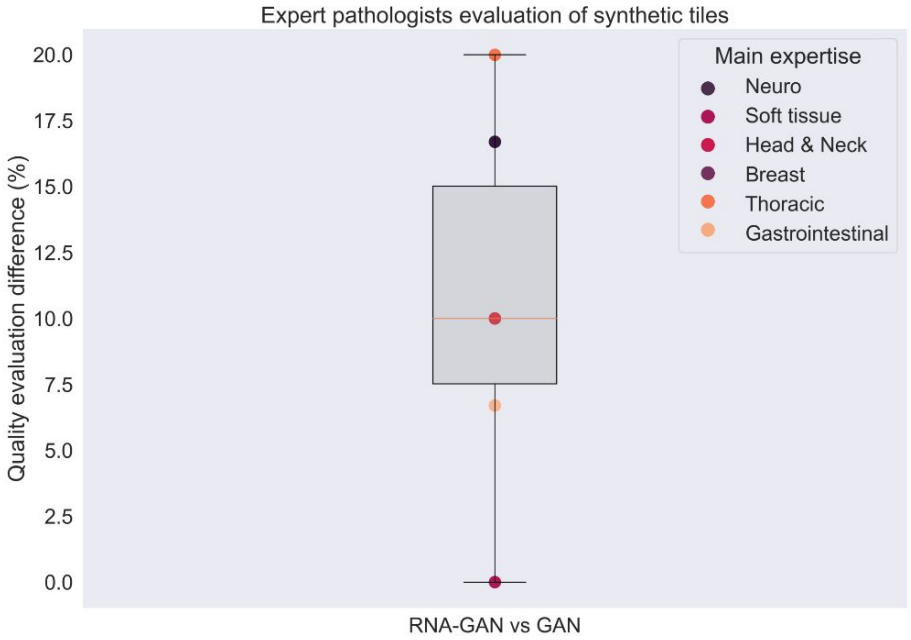
Results: Self-supervised learning

- We pre-trained a ResNet-18 with simCLR using only synthetic tiles, and compare the performance with a model trained from scratch.

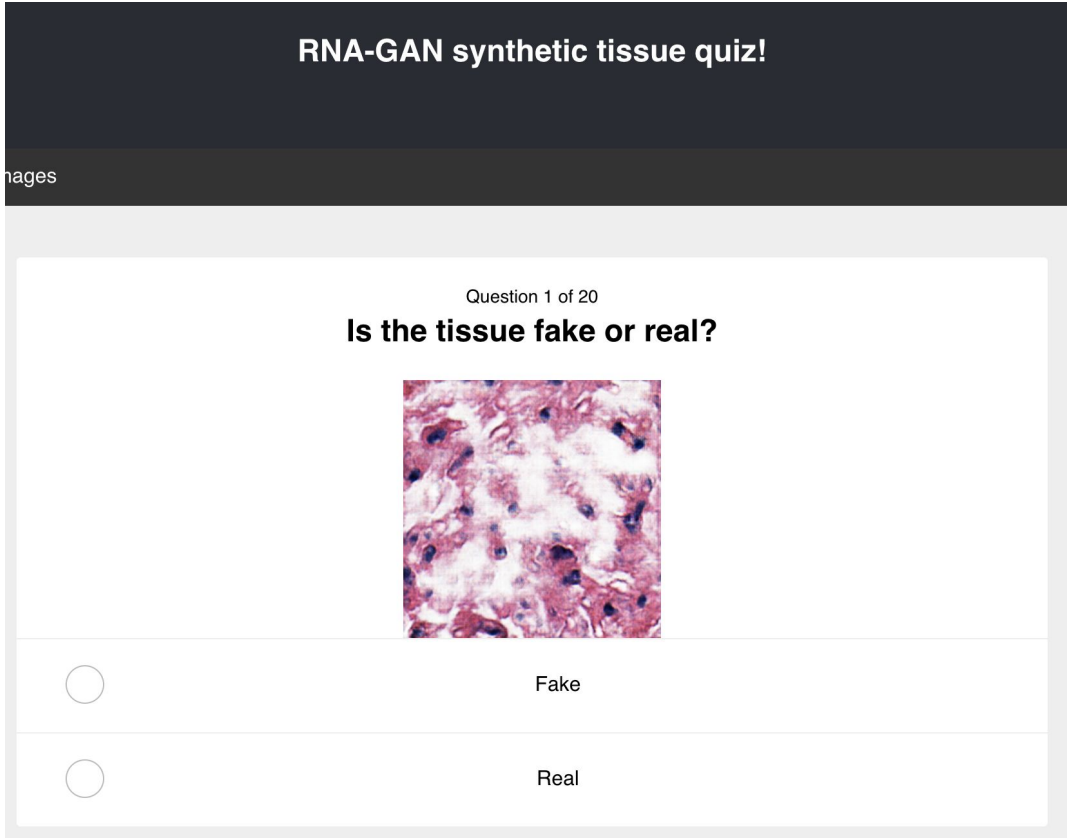


Results: Pathologists evaluation





You can play!



Quiz:
https://rna-gan.stanford.edu/

Conclusions

- RNA-GAN produces more realistic samples and trains faster than a traditional GAN approach
- It can be used for imputing missing FFPE tiles, in those datasets with only RNA-Seq available
- However, tissue quality can be improved. Another drawback is that a different model needs to be trained per tissue.
- The code and models are available at: https://github.com/gevaertlab/RNA-GAN

Article | Published: 21 March 2024

Generation of synthetic whole-slide image tiles of tumours from RNA-sequencing data via cascaded diffusion models

Francisco Carrillo-Perez, Marija Pizurica, Yuanning Zheng, Tarak Nath Nandi, Ravi Madduri, Jeanne Shen & Olivier Gevaert

Nature Biomedical Engineering (2024) | Cite this article

Motivation and objectives

Motivation:

- In recent years text-to-image models have been presented based on diffusion models (Saharia et. al. 2022; Ramesh et. al. 2022)
- GANs can be used for RNA-to-image generation, but they have multiple drawbacks

Objectives: Create a multi-cancer RNA-to-Image model that preserve cancer-specific characteristics

- Use a single architecture to generate tiles from 5 different cancer types
- Test that cancer-specific characteristics are preserved, by using cell counts (which cell types are found in the tiles) and cell proliferation based on deconvolved RNA-Seq
- The synthetic tiles can substitute real data to pre-train machine learning models

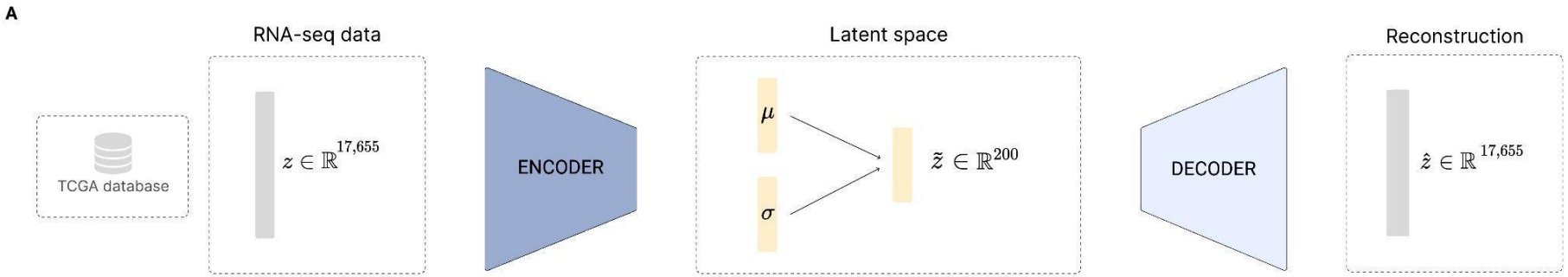
Data acquisition

Project Code	Cancer Type	Number of 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

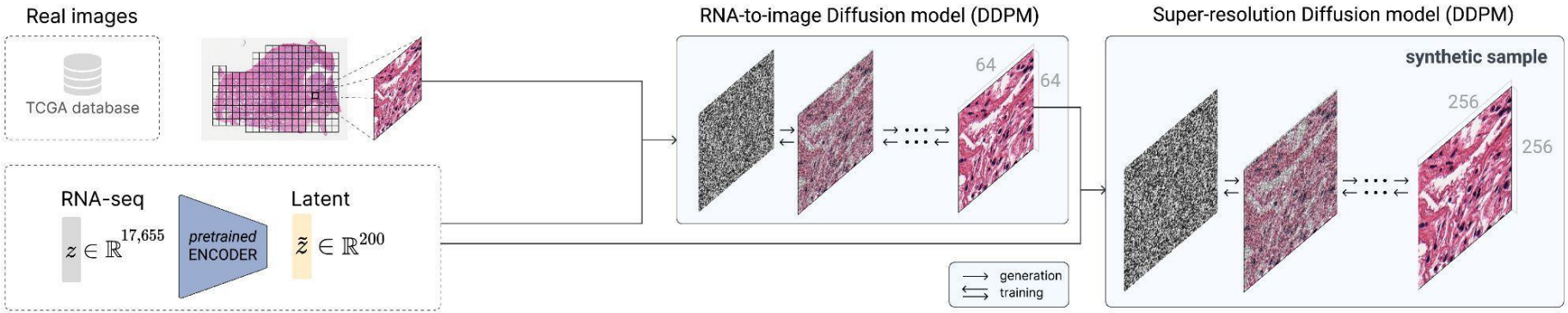
Data acquisition

- For further experiments, bulk RNA-Seq was deconvolved using CIBERSORTx (Newmann et al. 2015; Newman et al. 2019) into four cell-types: epithelial, endothelial, fibroblast, and haematopoietic.
- Two series were downloaded from GEO for generalization experiments: GSM1228184 (Kim et al. 2014) , and GSE226069 (Quintana et al. 2021)

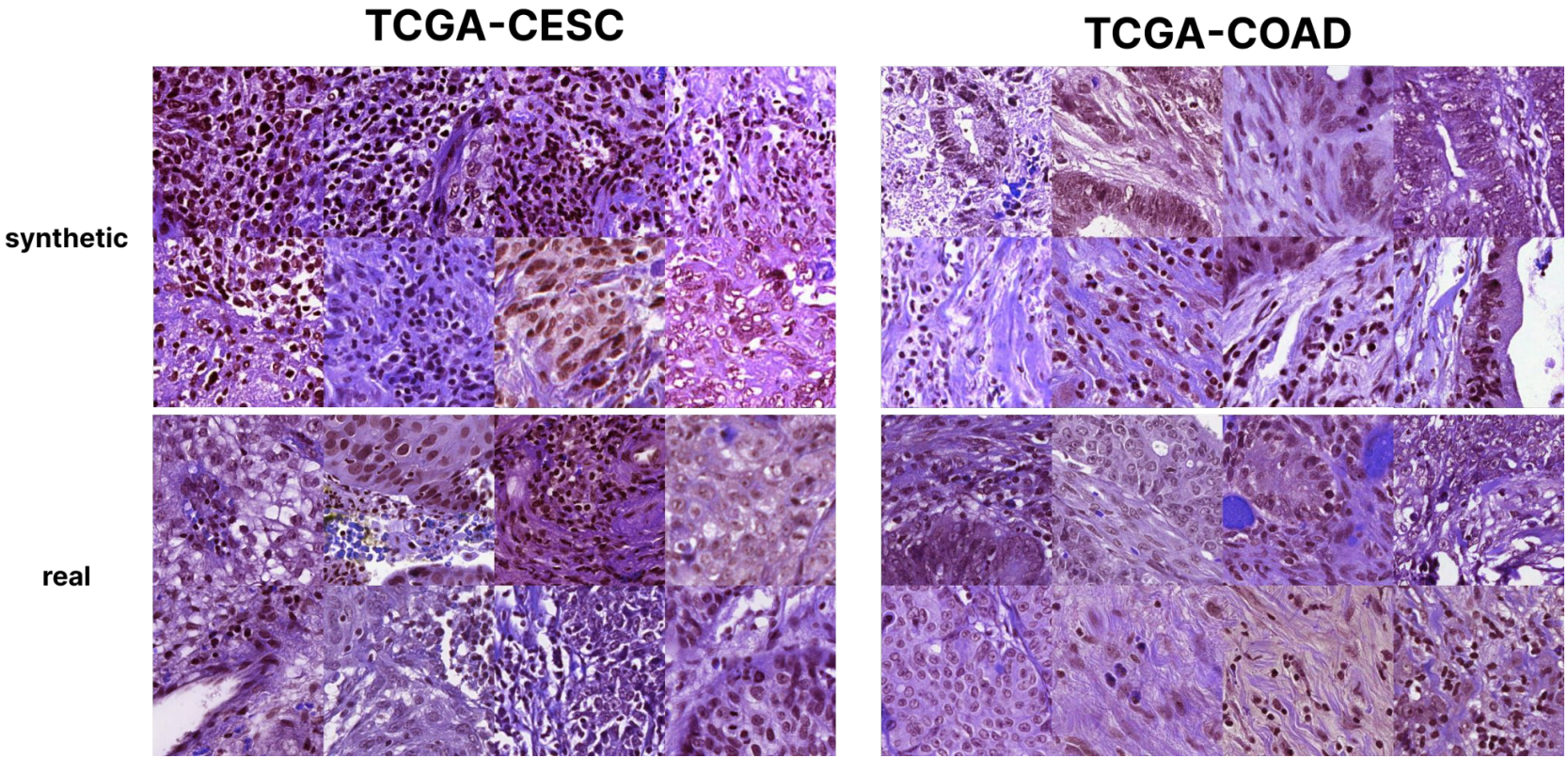
Methodology: VAE



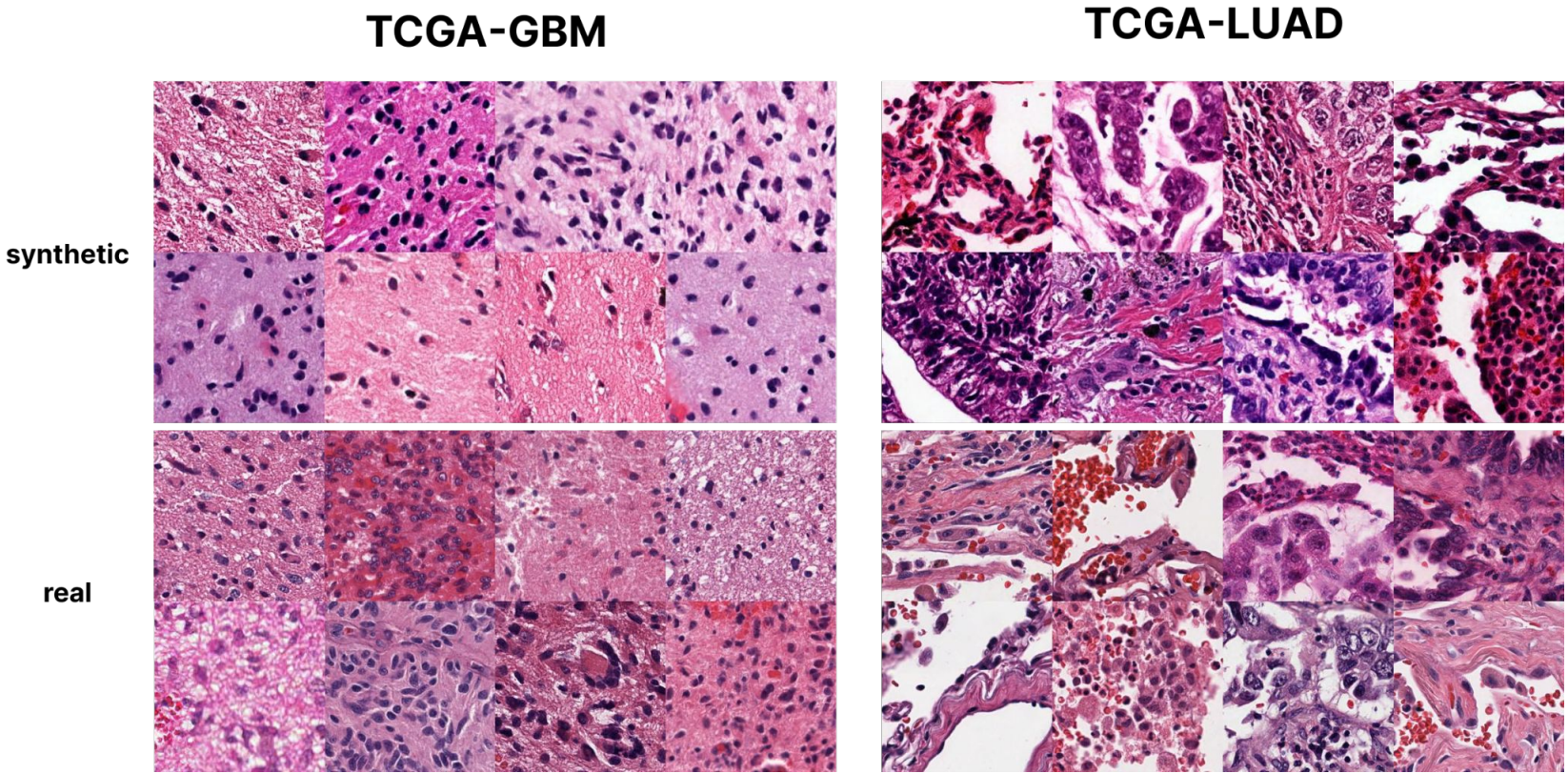
Methodology: RNA-CDM



Results: Tile generation



Results: Tile generation

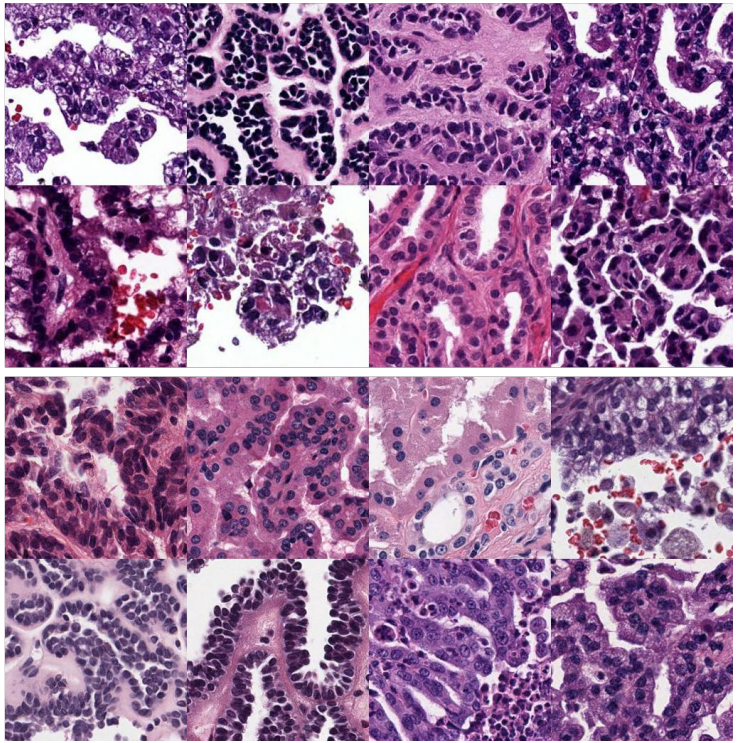


Results: Tile generation

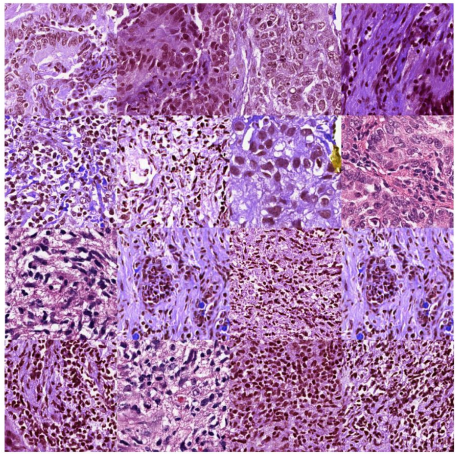
TCGA-KIRP

synthetic

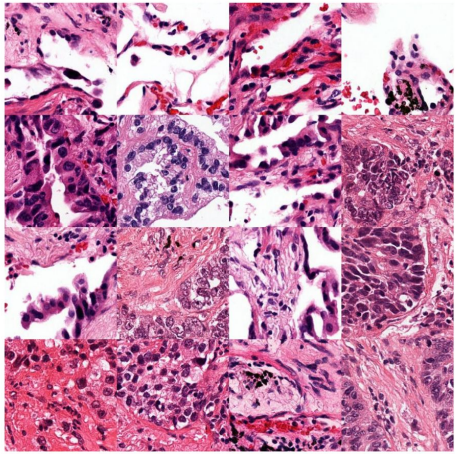
real



Colorectal cancer patient H&E tiles generated from RNA-Seq

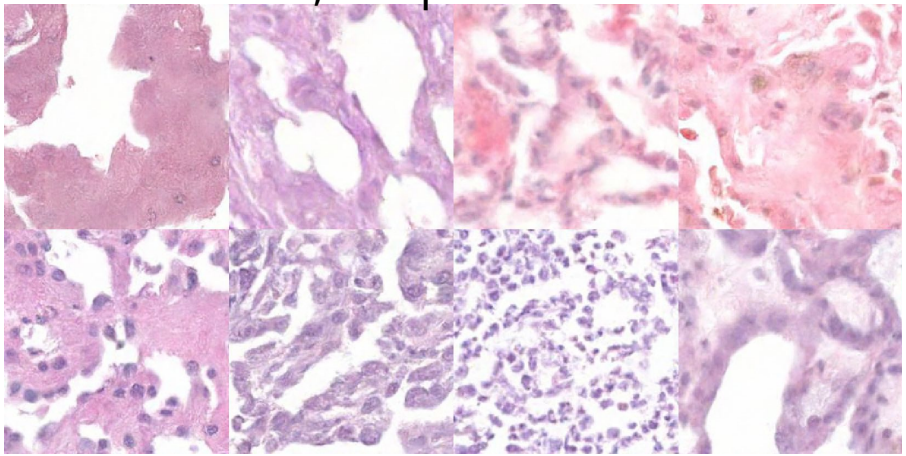


Lung cancer patient H&E tiles generated from RNA-Seq

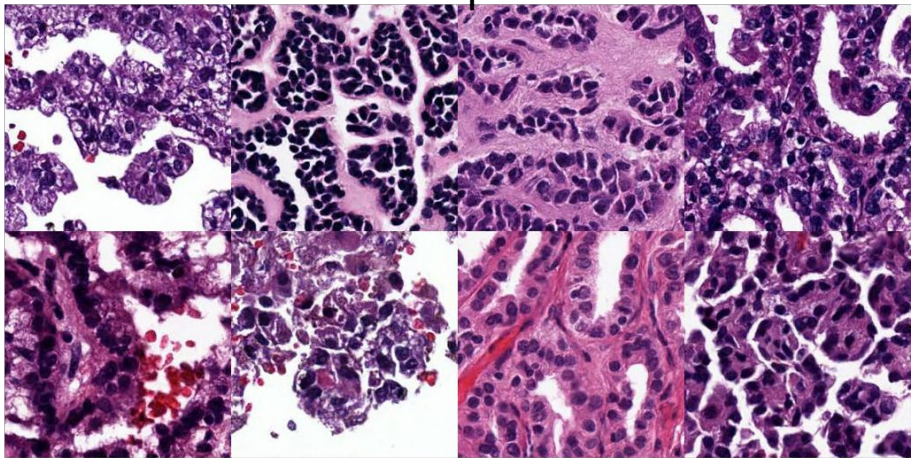


Results: Model size matters

500,000 parameters

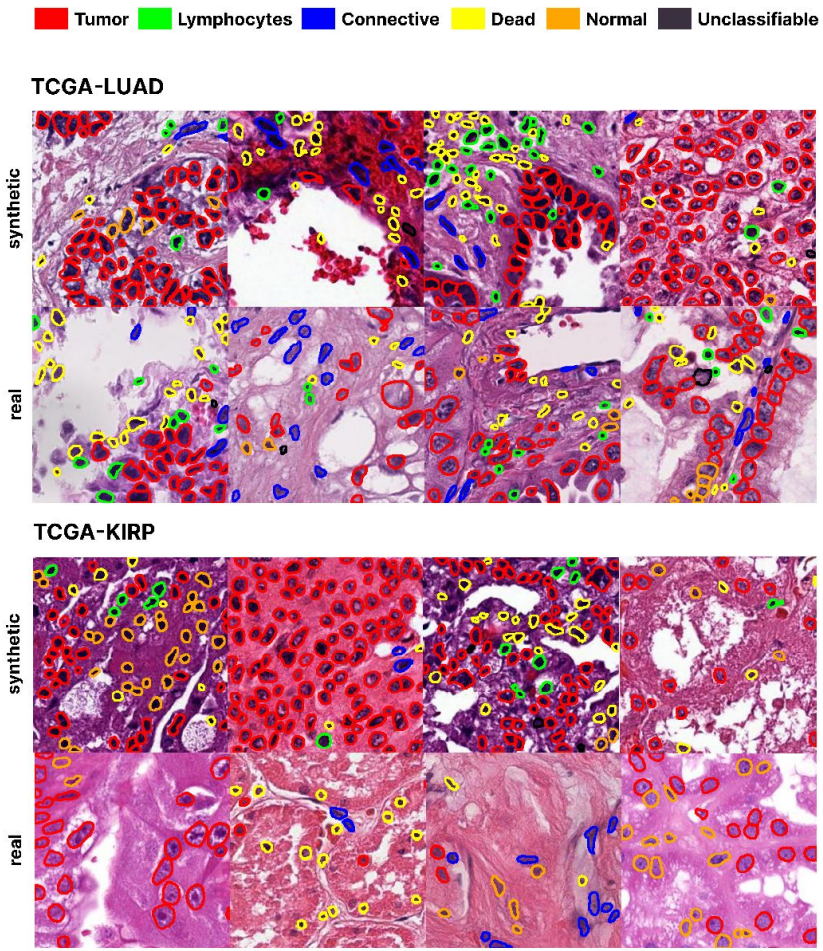


1.8 billion parameters



Results: Cell distribution using Hovernet

We generated 50.000 synthetic tiles and obtained the same amount of real tiles (10.000 per cancer type)
We ran HoverNet (cell identification and segmentation model) over the real and synthetic tiles.
We computed the percentage per cell within each tile, and compare the cell distribution.

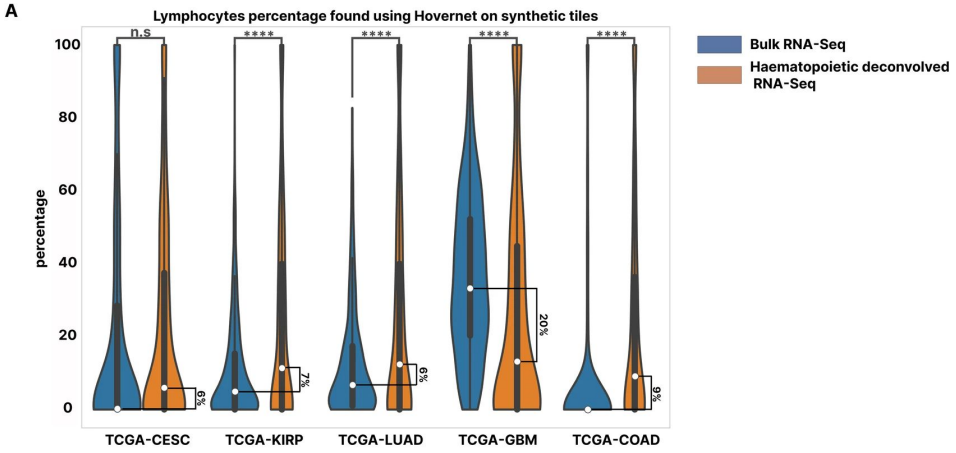


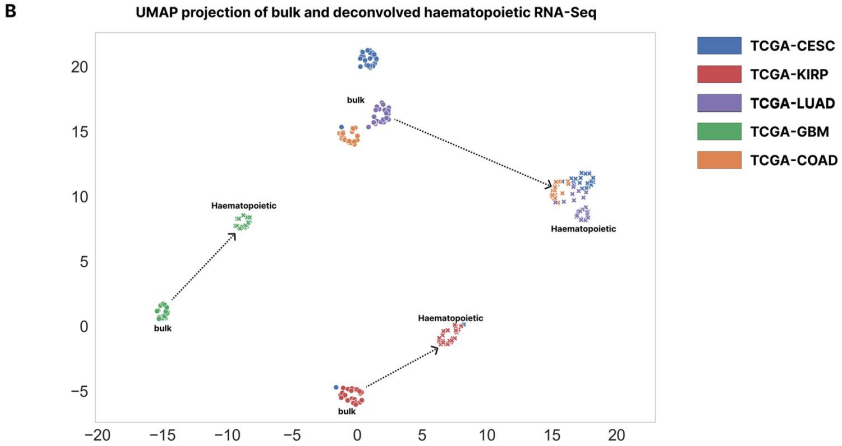
Results: Cell distribution using Hovernet

	Tile type	Tumour	Lymphocytes	Connective	Dead	Normal
TCGA-COAD	Real	47.44 ± 43.12	7.69 ± 20.43	10.48 ± 24.58	4.21 ± 14.66	3.94 ± 14.51
	Synthetic	61.78 ± 42.43	6.11 ± 17.88	2.69 ± 13.27	1.33 ± 7.36	6.87 ± 17.89
TCGA-GBM	Real	22.57 ± 25.83	17.66 ± 20.25	18.54 ± 21.74	26.20 ± 26.20	12.5 ± 21.34
	Synthetic	9.18 ± 16.15	35.09 ± 24.89	17.11 ± 20.33	22.89 ± 24.57	11.99 ± 19.76
TCGA-LUAD	Real	37.40 ± 32.49	8.12 ± 11.72	15.36 ± 19.22	35.15 ± 28.07	3.95 ± 9.78
	Synthetic	27.74 ± 28.62	12.93 ± 15.15	11.00 ± 15.19	41.30 ± 28.98	3.31 ± 9.86
TCGA-KIRP	Real	48.14 ± 33.02	7.39 ± 12.11	12.57 ± 21.83	19.20 ± 21.75	10.37 ± 18.35
	Synthetic	40.85 ± 32.25	12.92 ± 17.84	8.17 ± 17.38	20.59 ± 23.95	12.34 ± 21.84
TCGA-CESC	Real	45.52 ± 43.65	13.61 ± 27.54	5.34 ± 17.52	4.59 ± 15.31	2.85 ± 10.87
	Synthetic	45.82 ± 42.78	15.48 ± 28.17	1.52 ± 10.29	1.70 ± 7.74	7.89 ± 19.97

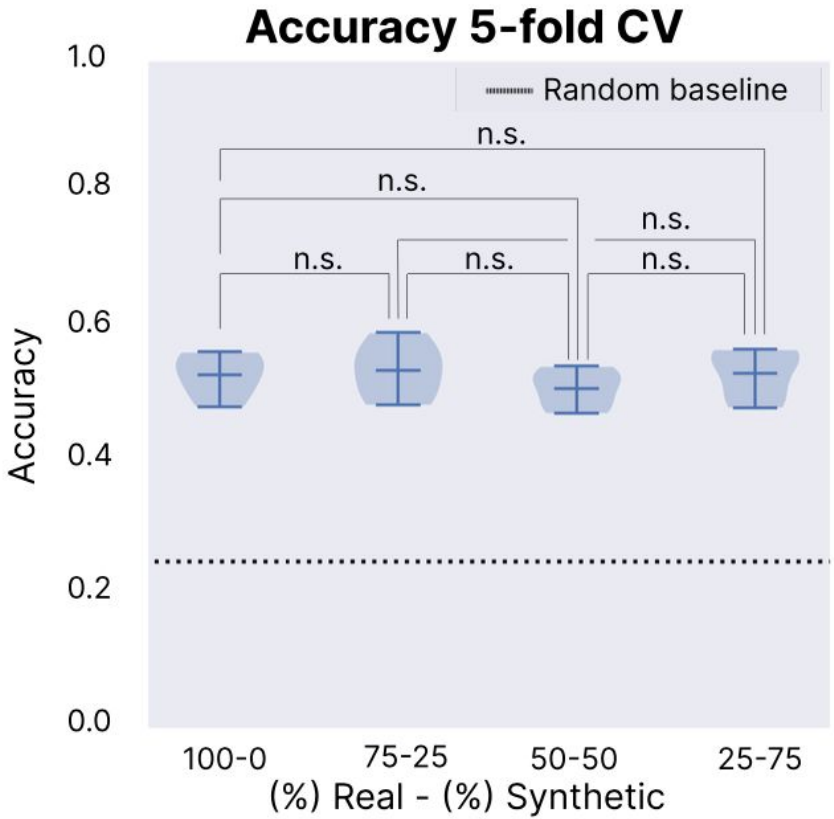
Results: Differences in tiles between bulk and deconvolved

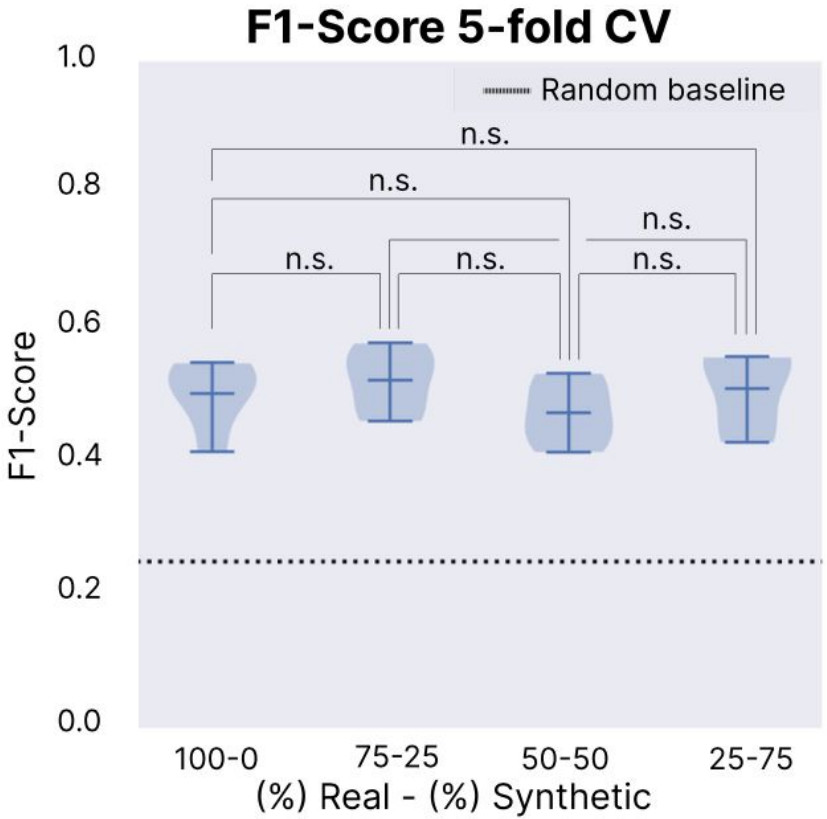
If we generate synthetic tiles using the haematopoietic deconvolved RNA-Seq, we find more lymphocytes in the tiles



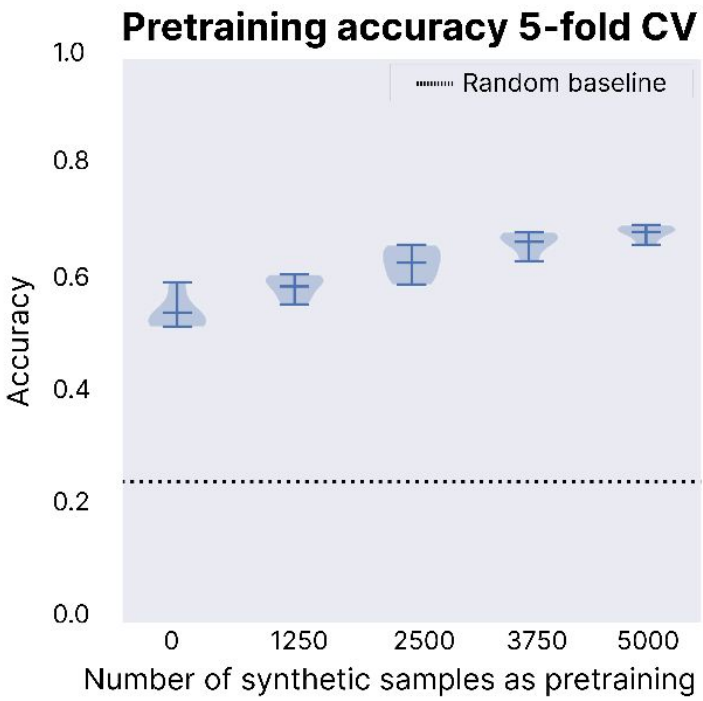


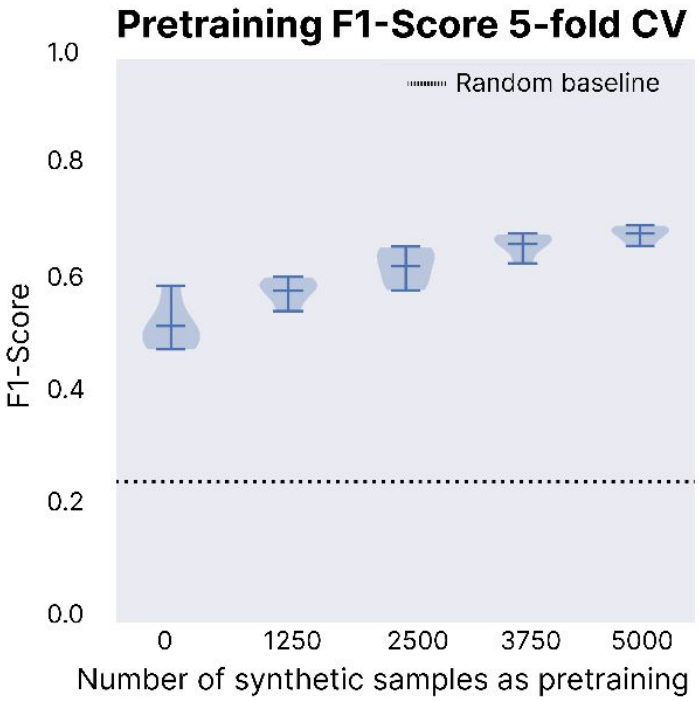
Results: Synthetic data can substitute real data



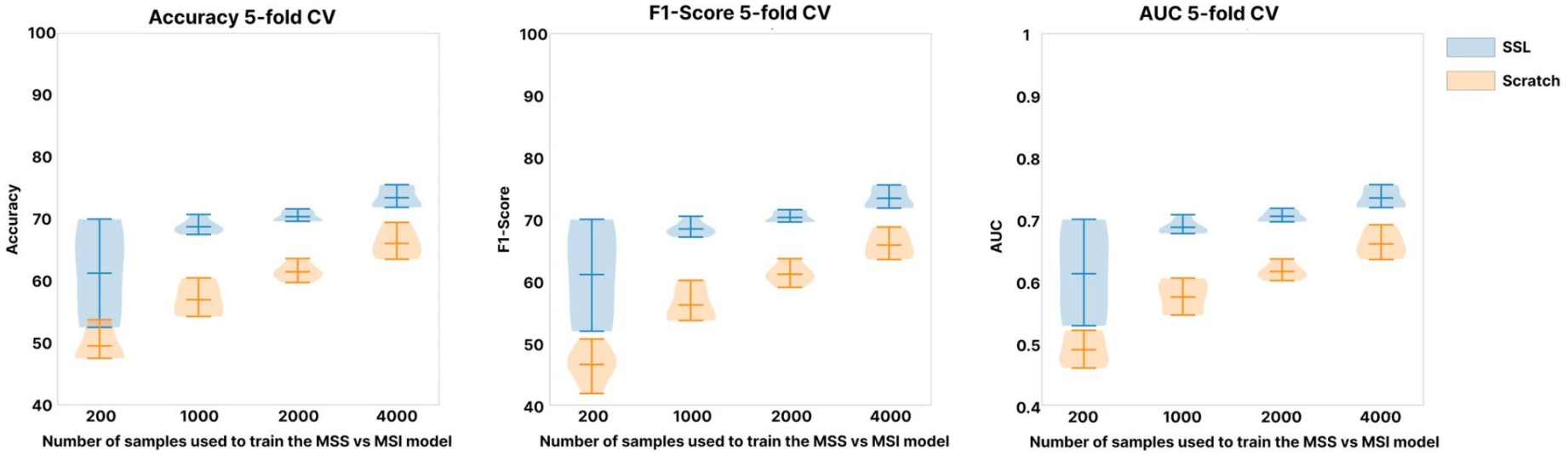


Results: Synthetic data can be used to pretrain ML models





Results: Microsatellite instability status prediction



Results: Prognosis prediction in pediatric gliomas

- We compared the performance of our pre-trained model on synthetic data for prognosis prediction in pediatric gliomas.
- The model outperformed those results obtained in literature, while also reducing the overfitting.

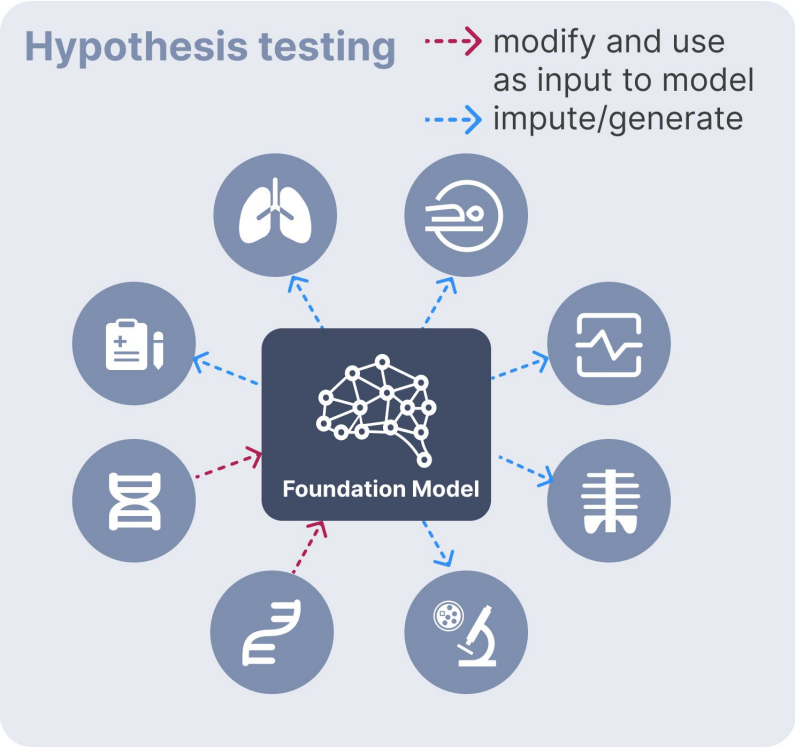
	Train CS (mean ± std)	**Val CS (mean ± std)**	**Test CS**
Steyaert et al.	0.900 ± 0.010	0.792 ± 0.070	0.854
Ours	0.806 ± 0.029	0.805 ± 0.058	**0.871**

Conclusions

- RNA-CDM, with a single architecture, produces realistic FFPE tiles from five different cancer types
- The cell fraction proportions are preserved in the synthetic tiles. In addition, higher fraction of specific cell types affect the synthetic tissue generated
- The synthetic tiles do not damage the performance of machine learning models, and can be used as pretraining to improve the classification metrics
- We released 1 million synthetic tiles (QR code)
- The code is available under academic-use license only at https://rna-cdm.stanford.edu

Future directions

Future directions: Synthetic multi-modal modelling



Thanks to my collaborators!

Thanks for your attention!

Any questions?

Email: fcperez@stanford.edu
ORCID: 0000-0003-0974-4092
Webpage: https://pacocp.es/
Twitter: @pacocp9
Github: @pacocp