# Transcriptomics Data Generation with Deep Generative Models

## Séminaire Biopuces - 13 Fev. 2025

**Alice Lacan**

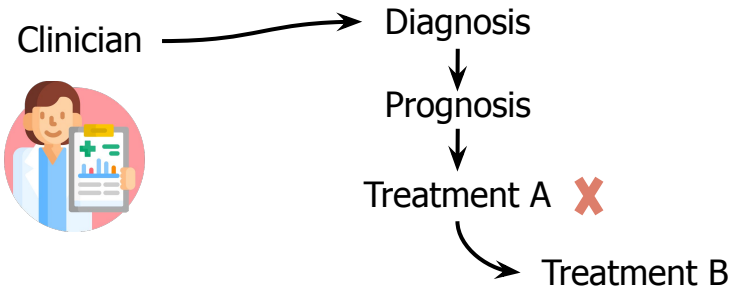
**Supervision:**

*Blaise Hanczar, IBISC, Univ. Evry*
*Michèle Sebag, LISN-CNRS-Inria, Univ. Paris-Saclay*

1. **Context**
   - 1.1. Omics and precision medicine
   - 1.2. Transcriptomics
   - 1.3. Machine Learning for cancer prediction
   - 1.4. Small *n*, large *p*
   - 1.5. Data augmentation
2. **State-of-the-art deep generative models**
3. **Contribution 1: Realistic generation with AttGAN**
4. **Contribution 2: Diversity with diffusion models**
5. **Contribution 3: Trade-off with GMDA**
6. **Conclusion & perspectives**

# Omics and precision medicine

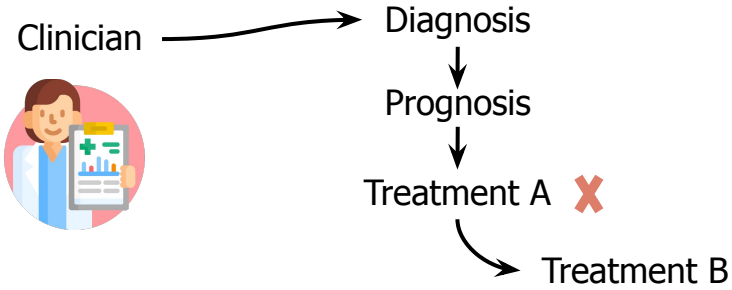
## Traditional medicine

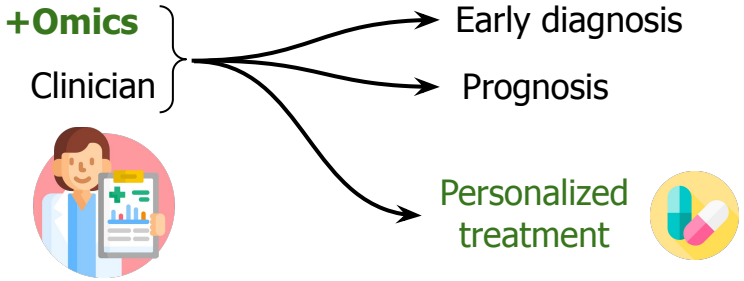

Source: BD "Les Trois Mousquetaires", inspired from Alexandre Dumas' work

# Omics and precision medicine

## Traditional medicine

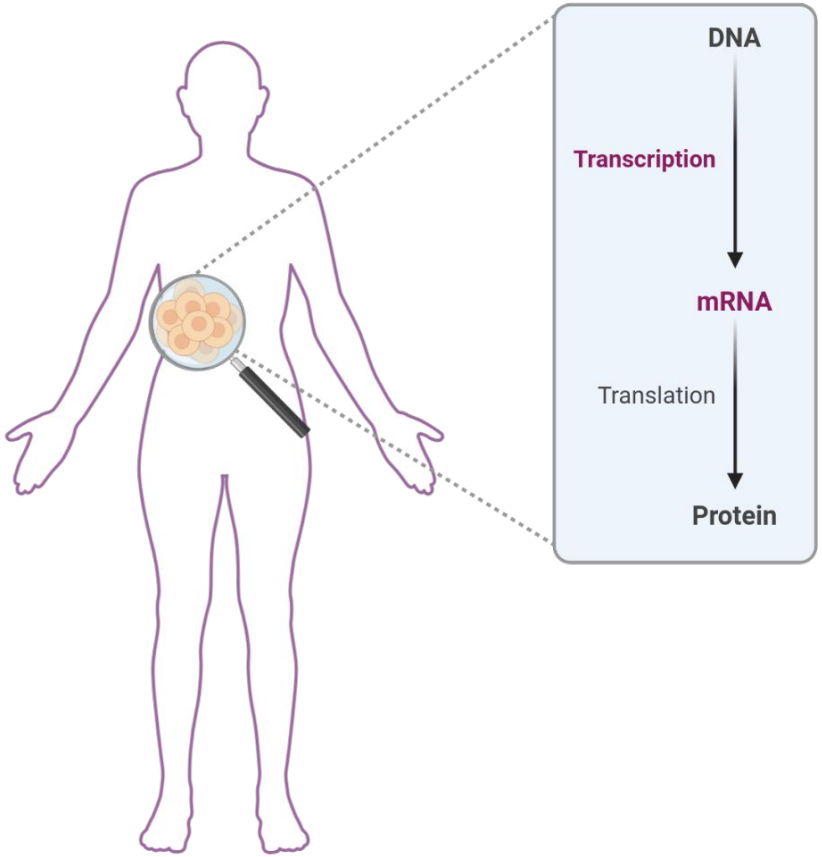




## Precision medicine


Source: BD "Les Trois Mousquetaires", inspired from Alexandre Dumas' work

# Transcriptomics




Dynamic information at a given time in a given cell

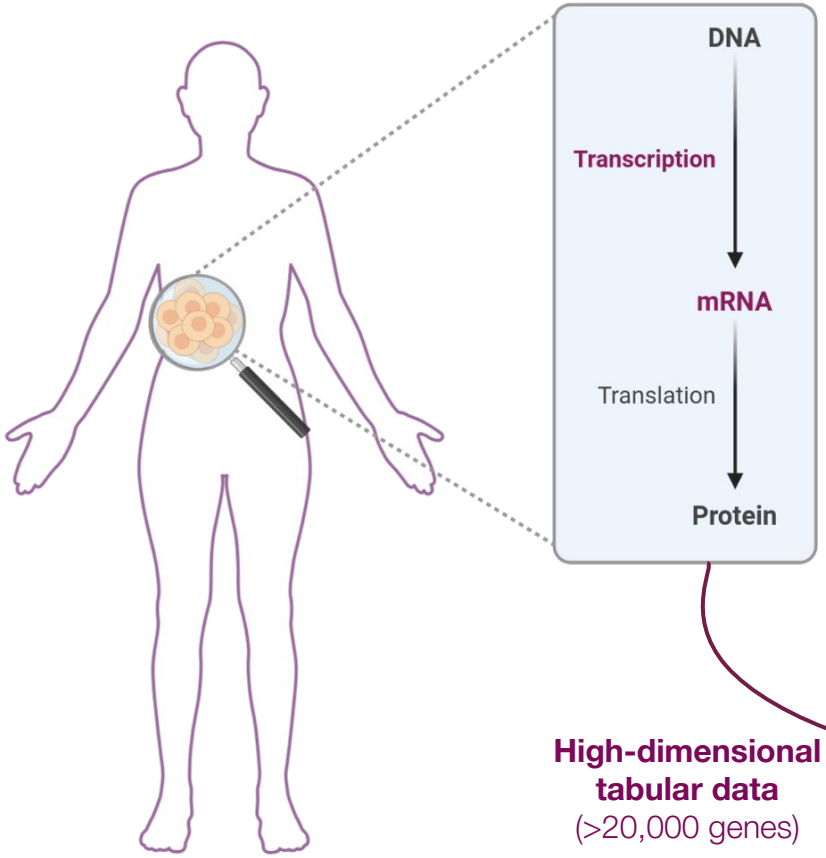
**Why?** To understand gene expression dynamics

**What for?**

- ❏ Disease biomarkers identification
- ❏ Capturing drugs effects on gene expression

Source: BioRender

# Transcriptomics




Dynamic information at a given time in a given cell

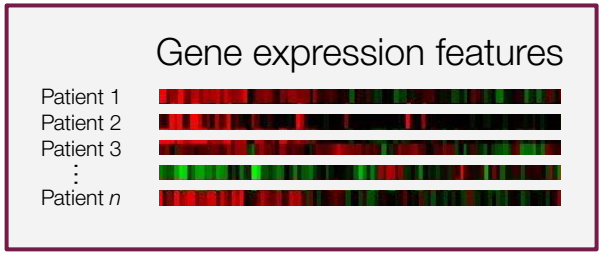
**Why?** To understand gene expression dynamics

**What for?**

- ❏ Disease biomarkers identification
- ❏ Capturing drugs effects on gene expression

**Context: transcriptomic data became more available** (sequencing advances, etc.)

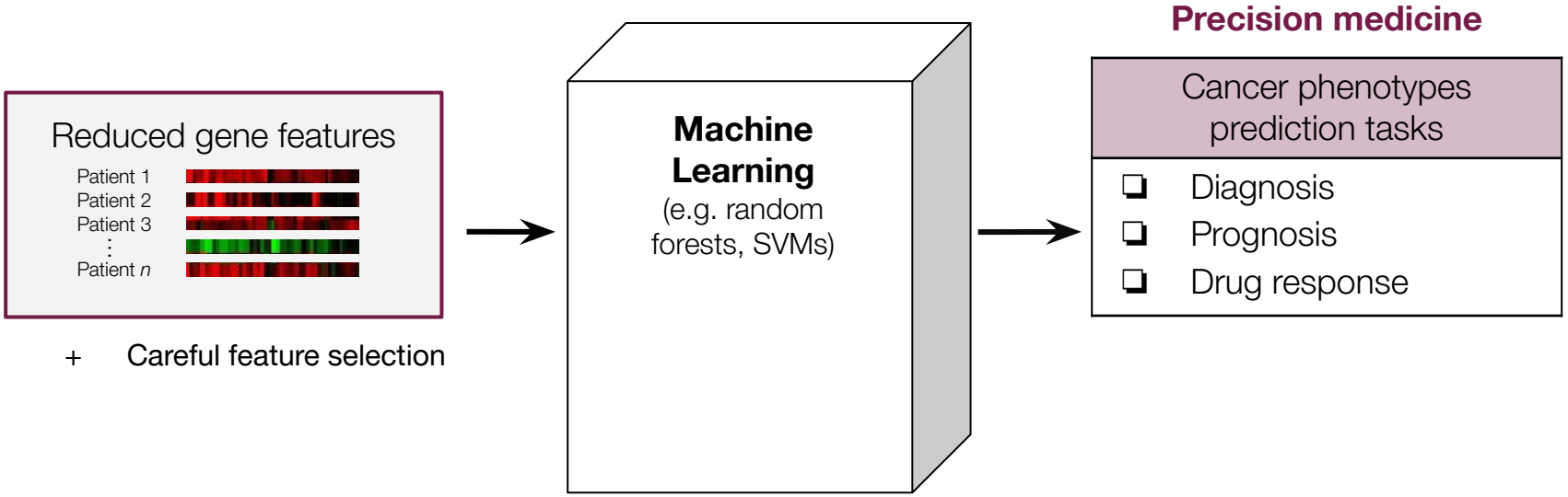
**High-dimensional tabular data**
(>20,000 genes)




Source: BioRender

# Machine Learning for cancer prediction

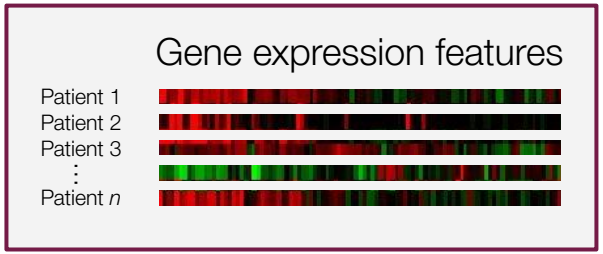
Gene expression data as input for data-driven models:

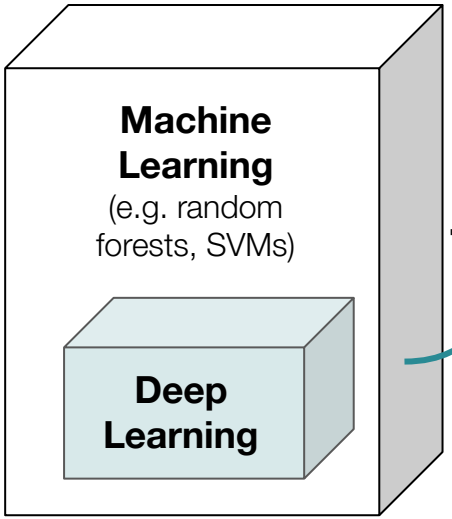



Kourou et al. *Machine learning applications in cancer prognosis and prediction* (Computational and structural biotechnology journal 2015)

# Machine Learning for cancer prediction

Gene expression data as input for data-driven models:




**Precision medicine**

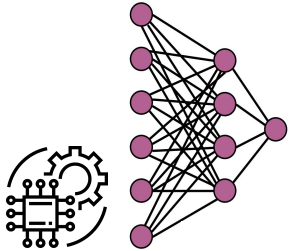
Cancer phenotypes prediction tasks

- ❑ Diagnosis
- ❑ Prognosis
- ❑ Drug response

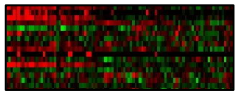
- Deep abstract representation learning
- Non linear modeling
- Complex feature relationships

Kourou et al. *Machine learning applications in cancer prognosis and prediction* (Computational and structural biotechnology journal 2015)

# Small *n*, large *p*



Small dataset



**Deep Learning**

overfitting

**Precision medicine**

| Cancer phenotypes prediction tasks |
| --- |
| ❑ Diagnosis |
| ❑ Prognosis |
| ❑ Drug response |

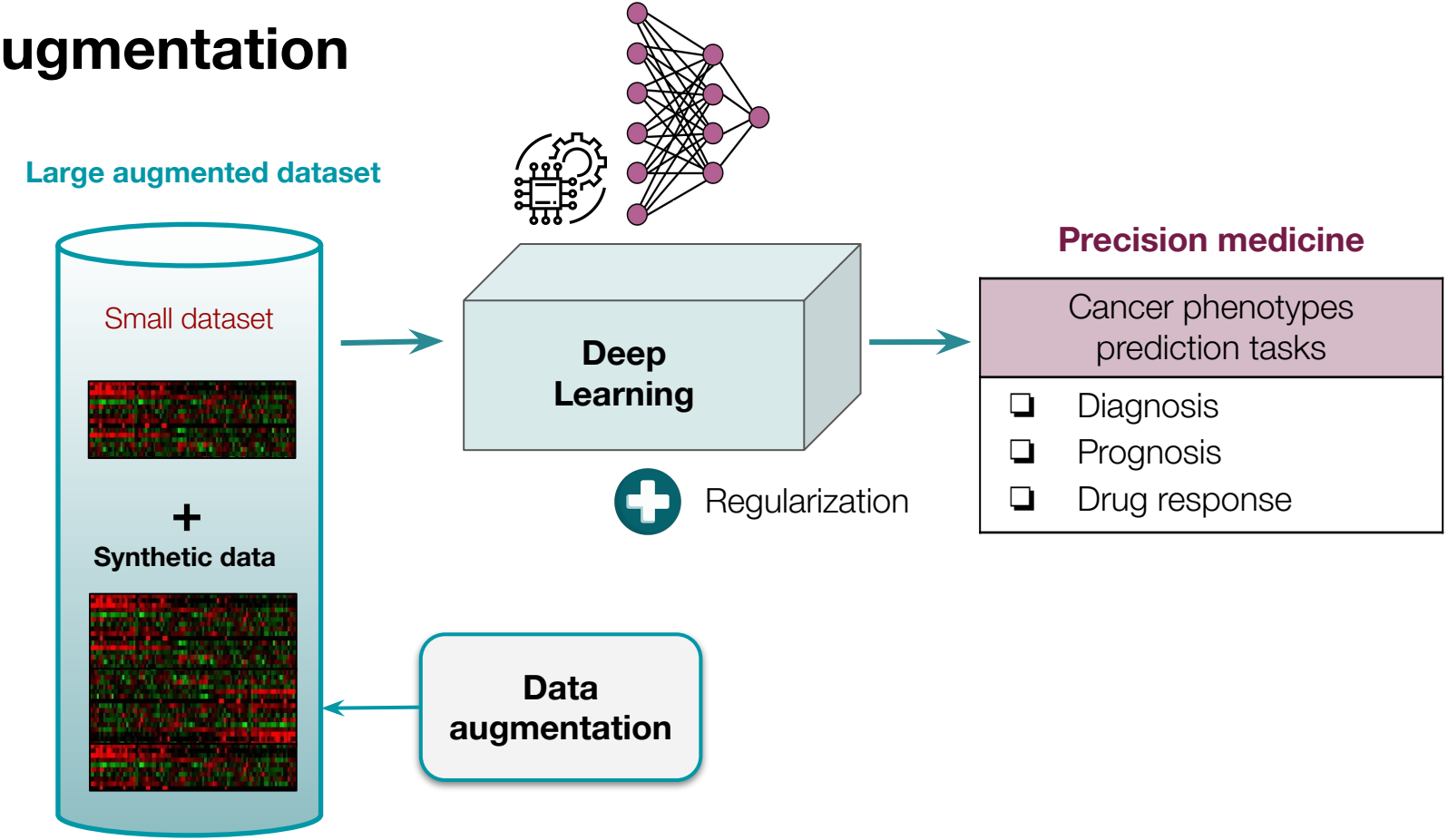
➔ **Scarcity:** cost, lack of integration, few patients (~100-1,000)

➔ **High dimensionality** (*+20,000* features)

➔ **Few studies on bulk RNA:** AEs for dimensionality reduction, multilayer perceptrons (MLPs)

Halevy et al. *The unreasonable effectiveness of data* (IEEE Intell Syst. 2009)

# Data augmentation

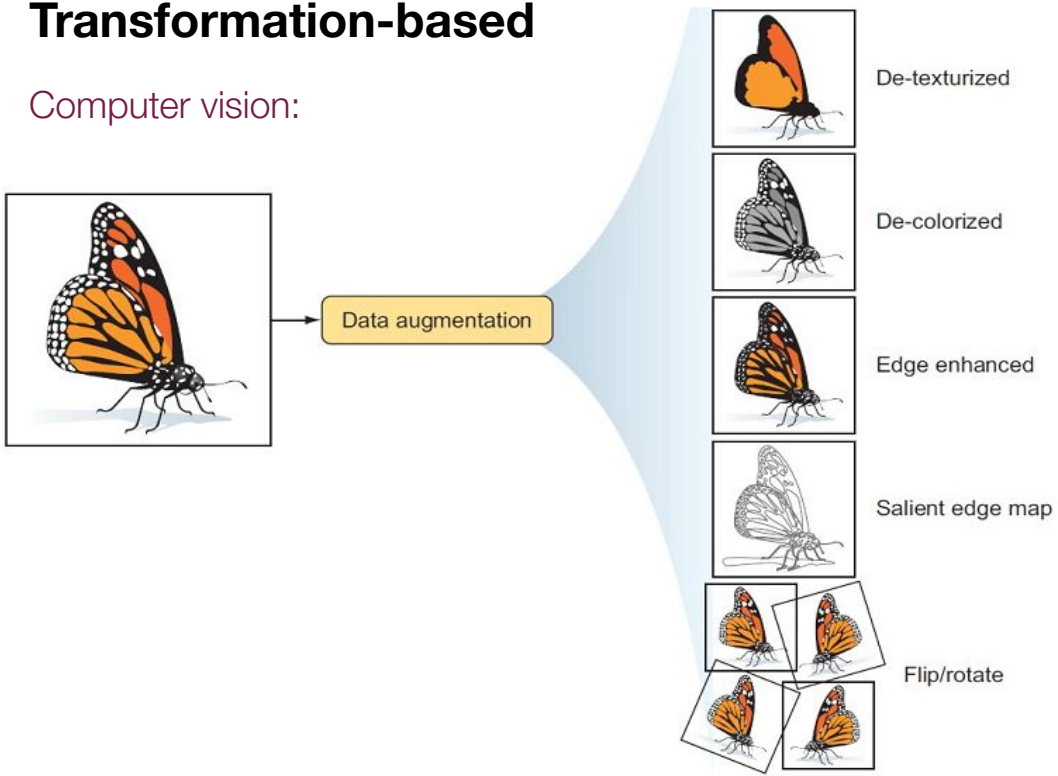


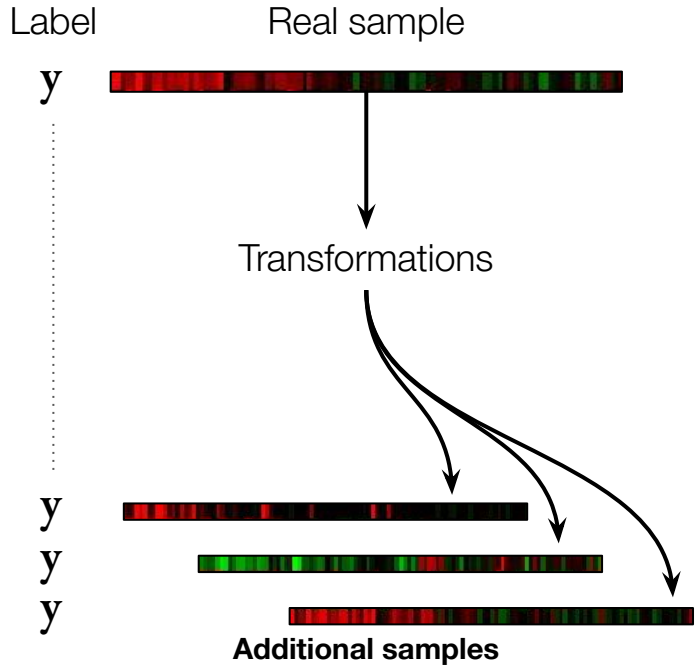

Halevy et al. *The unreasonable effectiveness of data* (IEEE Intell Syst. 2009)
Hanczar et al. *Phenotypes Prediction from Gene Expression Data with Deep Multilayer Perceptron and Unsupervised Pre-training* (International Journal of Bioscience 2018)
Bourgeais et al. *Deep GONet: self-explainable deep neural network based on Gene Ontology for phenotype prediction from gene expression data* (BMC Bioinformatics 2021)

# Data augmentation

## Transformation-based

Computer vision:




*What label-invariant transformations for gene expression data?*

Y. Lecun et al. *Gradient-based learning applied to document recognition* (IEEE 1998)
Elgendy, *Deep Learning for Vision Systems* (2020)

# Data augmentation

## Model-based

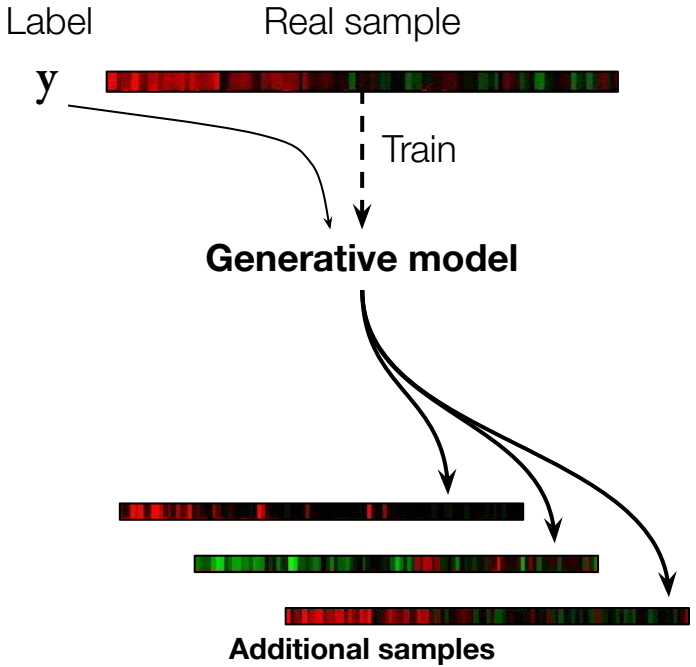
Deep generative models (DGMs):

- ❑ Variational autoencoders (VAEs)
- ❑ Generative Adversarial Networks (GANs)
- ❑ Diffusion models (DMs)
- ❑ Large Language Models (LLMs)

*Can we adapt DGMs in this small n, large p scenario?*

Source: *thispersondoesnotexist.com*

Source: DALL-E 2





Welling et al. *Auto-encoding variational bayes* (ICLR 2014)
Karras et al. *Analyzing and improving the image quality of stylegan* (*IEEE/CVF* 2020)
Rames et al. *Hierarchical text-conditional image generation with clip latents* (2022)

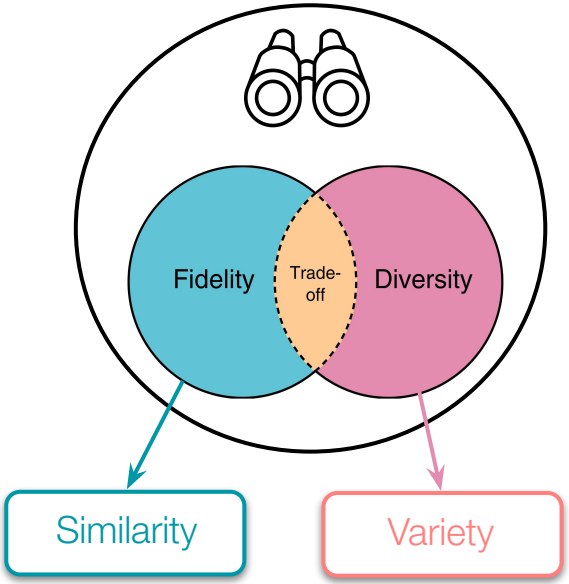
# Objectives of the thesis

**Scope:** precision medicine with ML and transcriptomics

*Can we leverage data augmentation with DGMs to enhance deep learning classification performance?*

**Objectives:**

- ❑ Adapt and extend DGMs for transcriptomics
- ❑ Proper data quality evaluation
- ❑ Data augmentation methodology




**Challenges for DGMs:**

- High dimensional data distribution (~20,000 features)
- Tabular features (less explored in DL)
- Data evaluation

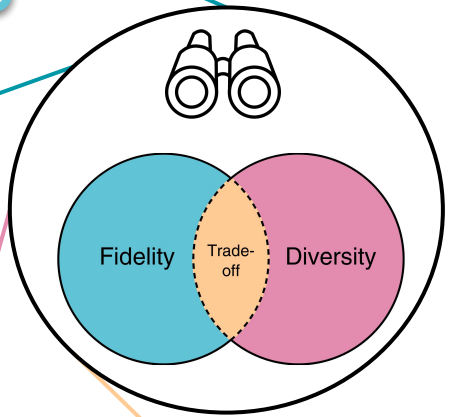
# Contributions of the thesis

**AttGAN:**
A. Lacan, M. Sebag and B. Hanczar. "GAN-based data augmentation for transcriptomics : survey and comparative assessment". In: ISMB, June 2023.

**GANs for microarray data:**
A. Alsamadi, A. Lacan, B. Hanczar and M. Sebag. "Identifying GANs Blind Spots in Transcriptomic Data Generation". In: JDSE, September 2024.

**Diffusion for transcriptomics (preprint):**
A. Lacan, R. André, M. Sebag and B. Hanczar. "In Silico Generation of Gene Expression profiles using Diffusion Models". In: bioRxiv, 2024.

**GMDA:**
A. Lacan, B. Hanczar and M. Sebag.
"Frugal Generative Modeling for Tabular Data". In: ECML-PKDD, September 2024.



# Contributions of the thesis

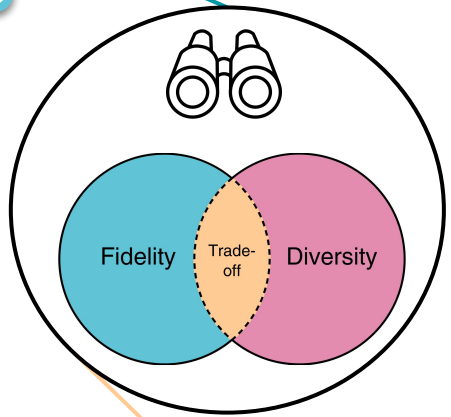
**AttGAN:**
A. Lacan, M. Sebag and B. Hanczar. "GAN-based data augmentation for transcriptomics : survey and comparative assessment". In: ISMB, June 2023.

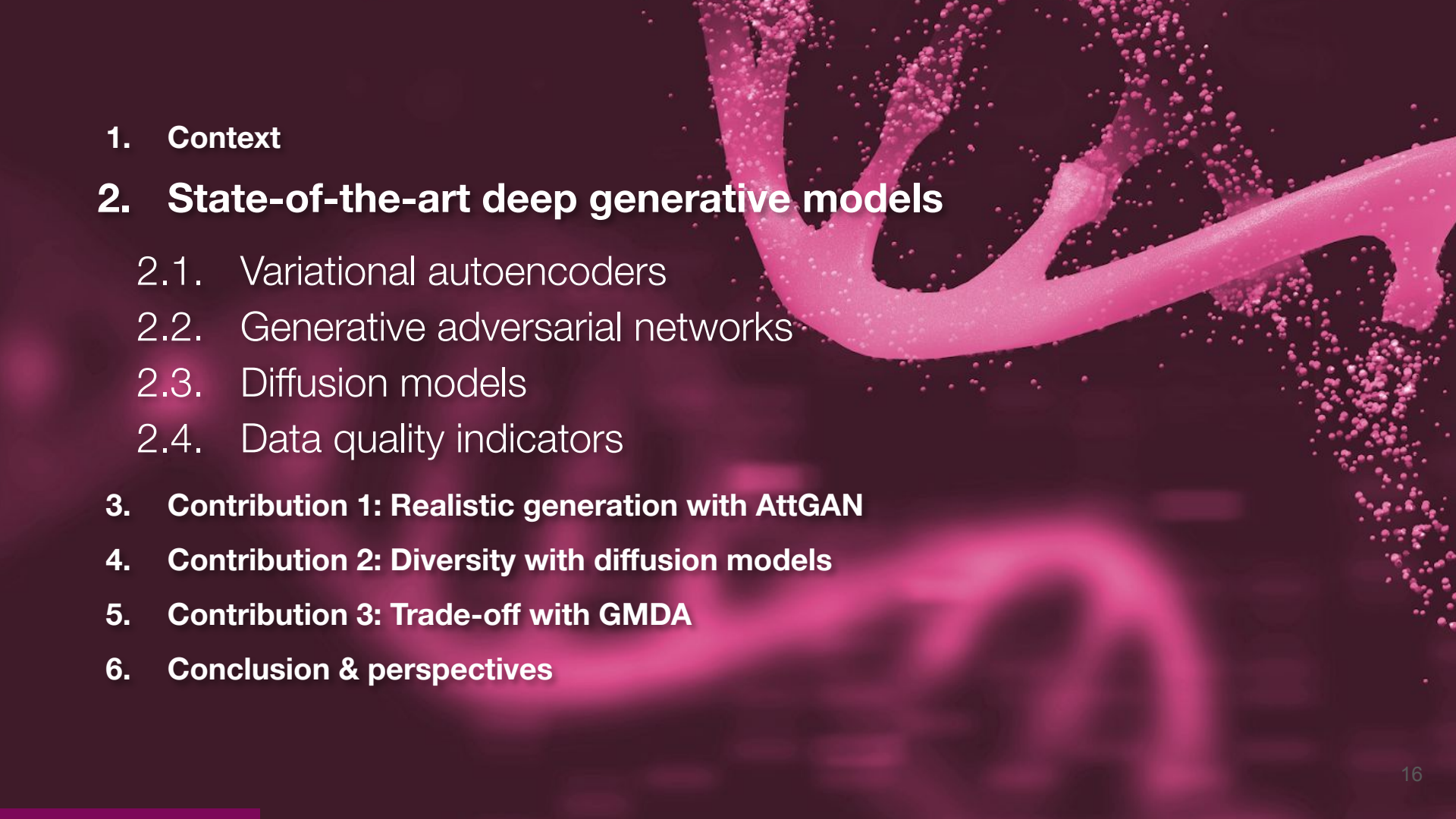


**GANs for microarray data:**
A. Alsamadi, A. Lacan, B. Hanczar and M. Sebag. "Identifying GANs Blind Spots in Transcriptomic Data Generation". In: JDSE, September 2024.

**Diffusion for transcriptomics (preprint):**
A. Lacan, R. André, M. Sebag and B. Hanczar. "In Silico Generation of Gene Expression profiles using Diffusion Models". In: bioRxiv, 2024.

**GMDA:**
A. Lacan, B. Hanczar and M. Sebag.
"Frugal Generative Modeling for Tabular Data". In: ECML-PKDD, September 2024.

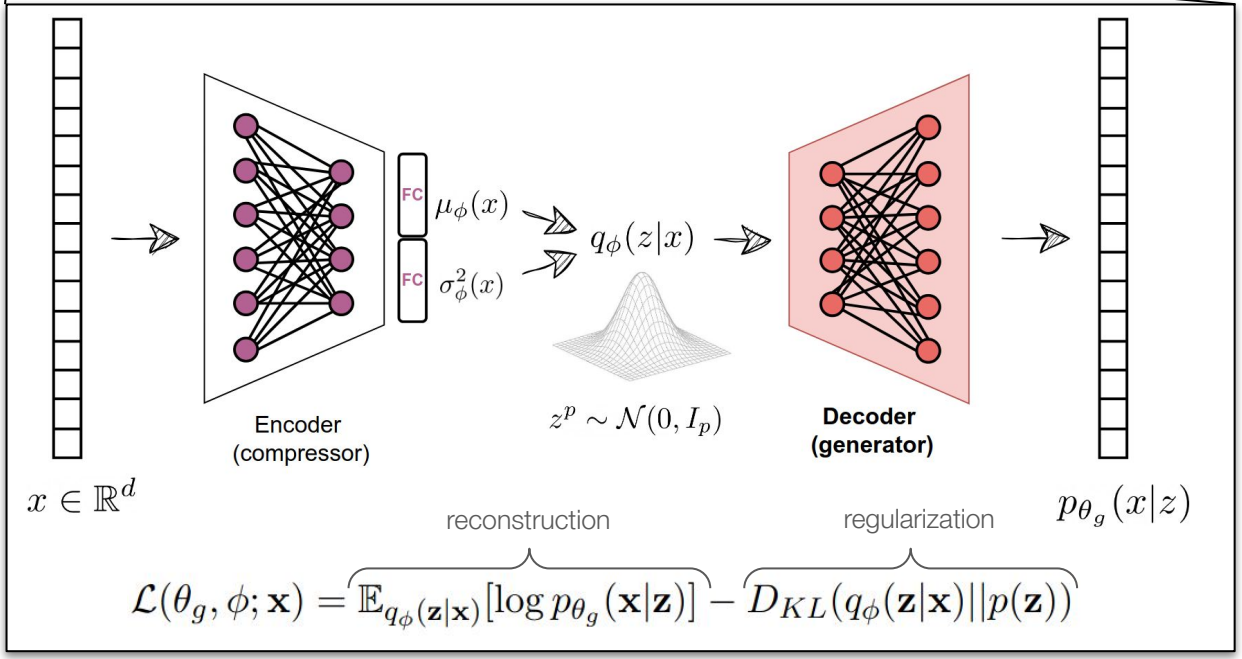




1. **Context**
2. **State-of-the-art deep generative models**
   - 2.1. Variational autoencoders
   - 2.2. Generative adversarial networks
   - 2.3. Diffusion models
   - 2.4. Data quality indicators
3. **Contribution 1: Realistic generation with AttGAN**
4. **Contribution 2: Diversity with diffusion models**
5. **Contribution 3: Trade-off with GMDA**
6. **Conclusion & perspectives**

# State-of-the-art deep generative models

Variational autoencorders (VAEs)




Diverse outputs

Meaningful regularized latent space

Lack of detailed outputs

Simple prior

**Tabular:** TVAE (Xu et al. NeurIPS 2019), GOGGLE (Liu et al. ICLR 2023)

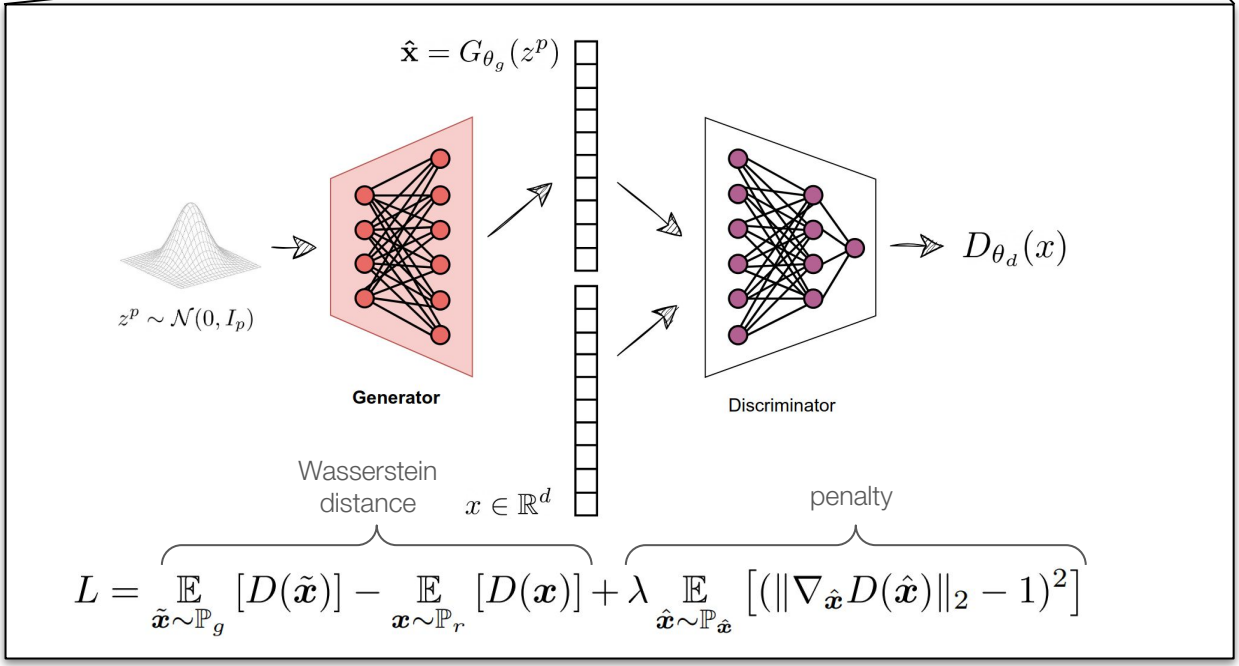
**Transcriptomics:** Single-cell applications (Lopez et al. Nat Methods 2018)

Welling et al. *Auto-encoding variational bayes* (ICLR 2014)

# State-of-the-art deep generative models

Variational autoencorders (VAEs)

Generative adversarial networks (GANs)




Goodfellow et al. *Generative adversarial nets.* (NeurIPS 2014)
Gulrajani et al. *Improved training of Wasserstein GANs* (NeurIPS 2017)

⇨ Wasserstein GAN with Gradient Penalty (WGAN-GP)

➕ High-quality and realistic outputs

➖ Mode collapse

➖ Convergence issues

**Tabular:** CTGAN (Xu et al. NeurIPS 2019), PATE-GAN (Yoon et al. ICLR 2019)

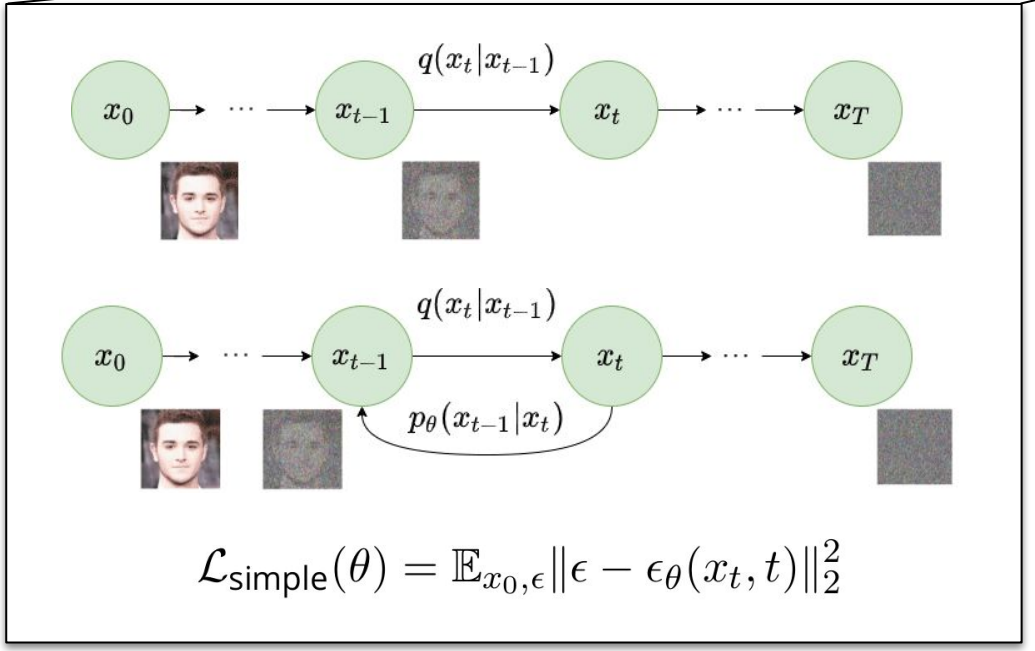
**Transcriptomics:**
Chaudhari et al. Soft Computing 2019;
Vinas et al. Bioinformatics 2021;
Li et al. BMC Bioinformatics 2023.

# State-of-the-art deep generative models

Variational autoencorders (VAEs)

Generative adversarial networks (GANs)

Diffusion models (DMs)




e.g., Denoising Diffusion Probabilistic Models (DDPMs), Denoising Diffusion Implicit Models (DDIMs)

High-quality, realistic and diverse outputs

Computational complexity

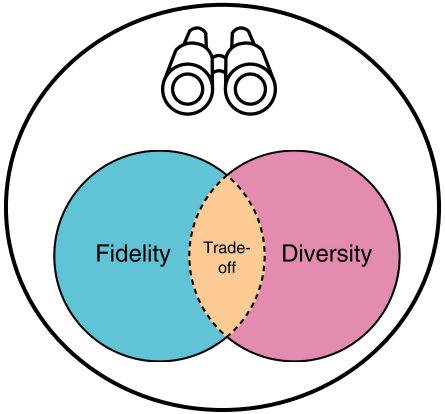
Inference instability

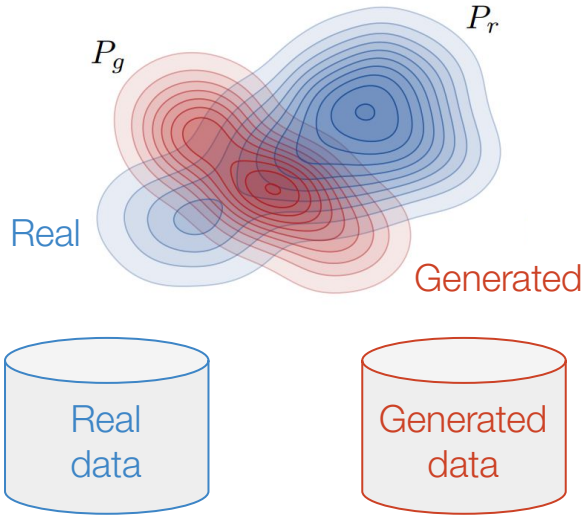
**Tabular:** TabDDPM (Kotelnikov et al. ICML 2023), TabSYN (Zhang et al. ICLR 2024)

**Transcriptomics:** scDiffusion (Luo et al. Bioinformatics 2024)

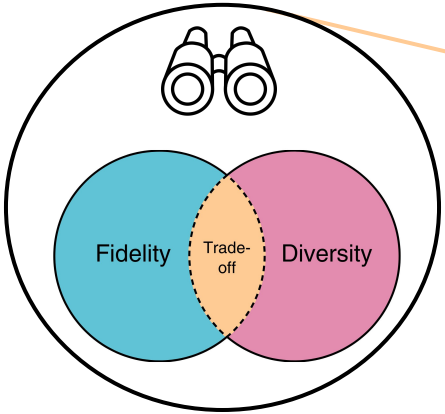
Ho et al. *Denoising Diffusion Probabilistic Models* (NeurIPS 2020)
Song et al. *Denoising diffusion implicit models* (ICLR, 2021)

# Data quality indicators

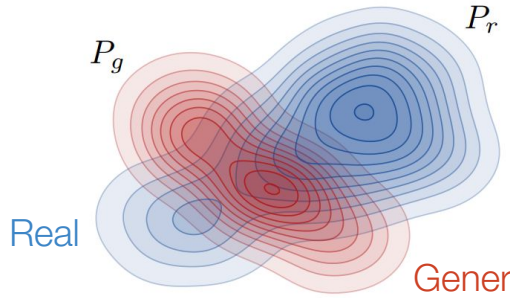


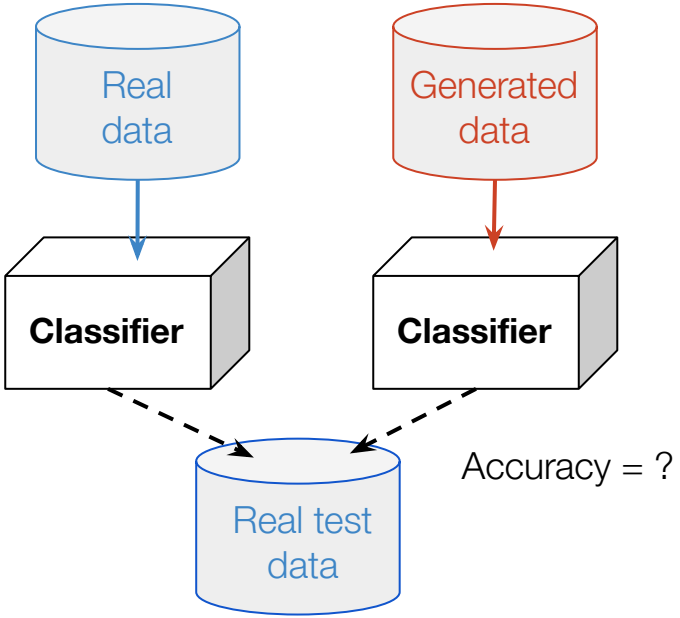



# Data quality indicators

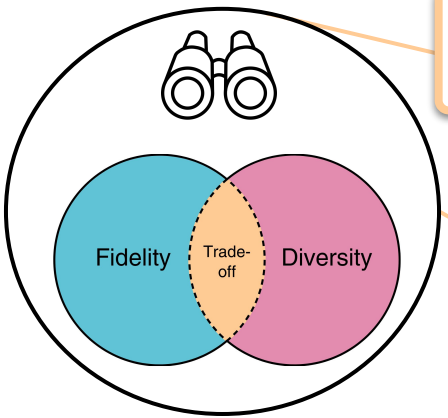



**Machine Learning efficiency (MLE)** or reverse validation
= knowledge preservation



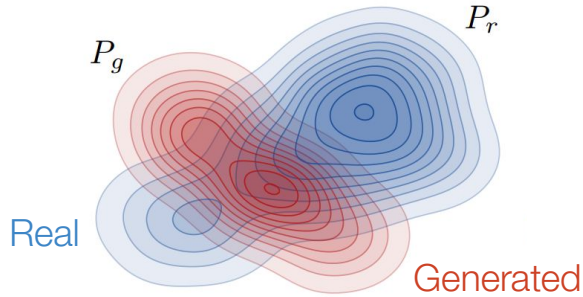



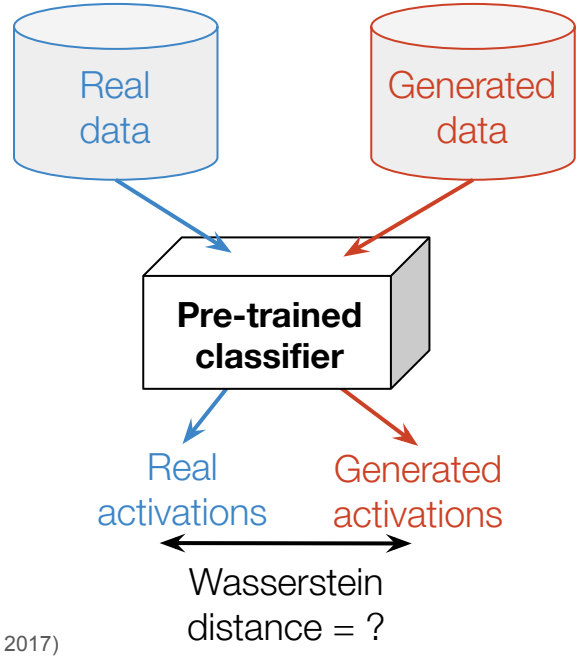
# Data quality indicators




Machine Learning efficiency (MLE) or reverse validation
= knowledge preservation

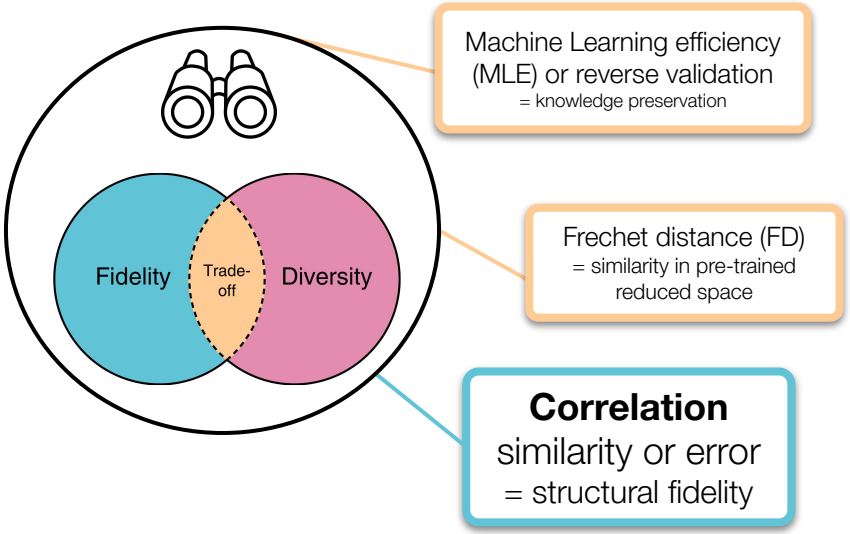
**Frechet distance (FD)**
= similarity in pre-trained reduced space




Heusel et al., *GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium* (NeurIPS 2017)

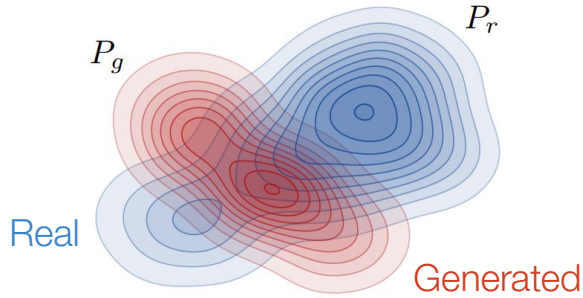
# Data quality indicators

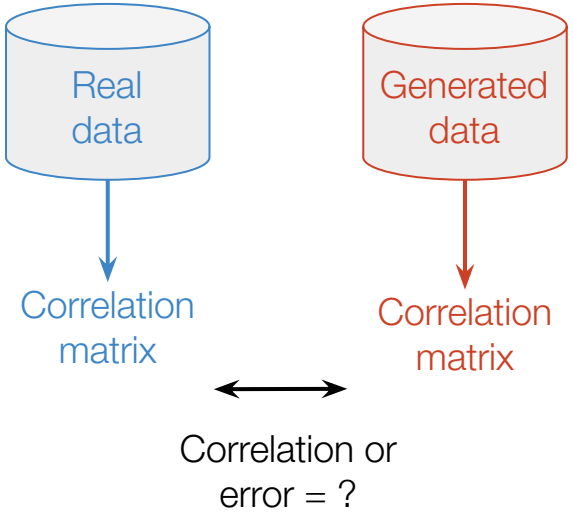



Machine Learning efficiency (MLE) or reverse validation
= knowledge preservation

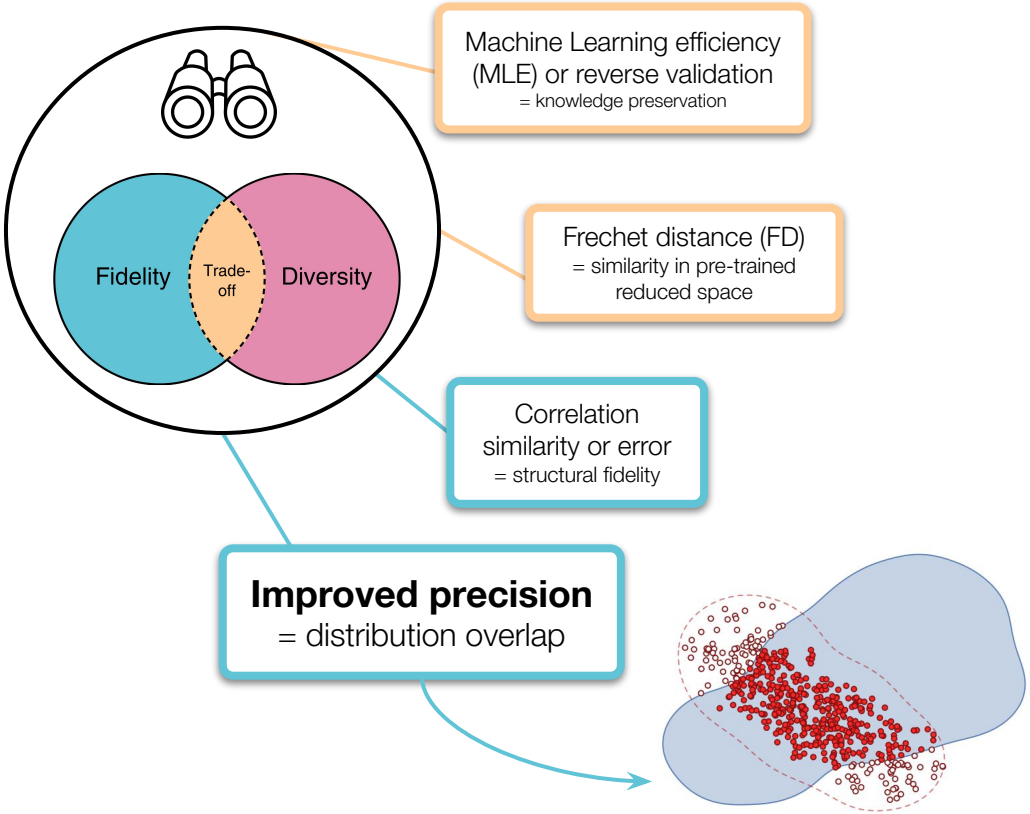
Frechet distance (FD)
= similarity in pre-trained reduced space

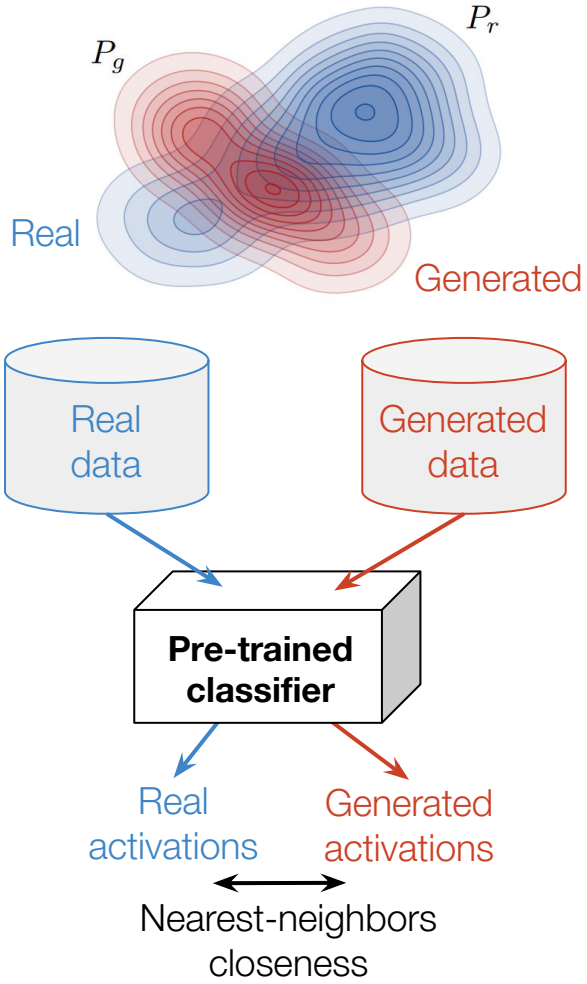
**Correlation**
similarity or error
= structural fidelity



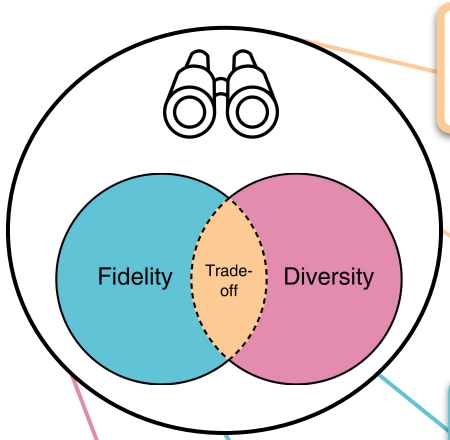
# Data quality indicators




Kynkäänniemi et al. *Improved Precision and Recall Metric for Assessing Generative Models* (NeurIPS 2019)

# Data quality indicators



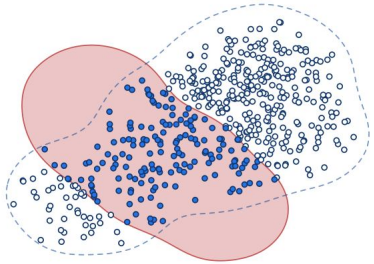

Machine Learning efficiency (MLE) or reverse validation
= knowledge preservation

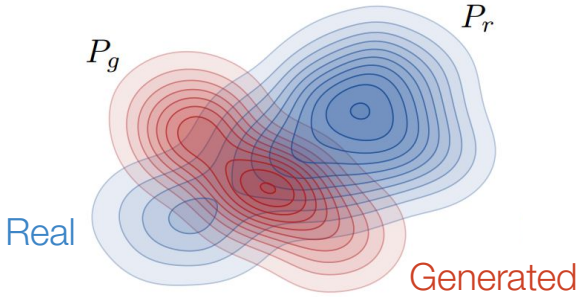
Frechet distance (FD)
= similarity in pre-trained reduced space

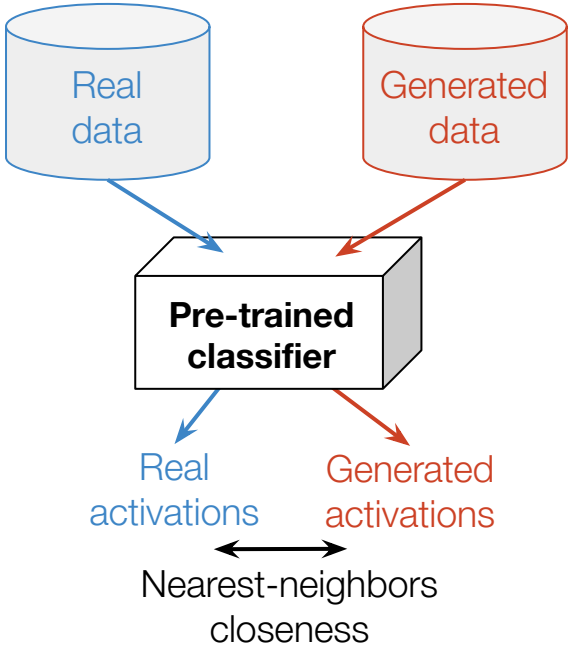
Correlation similarity or error
= structural fidelity

Improved precision
= distribution overlap

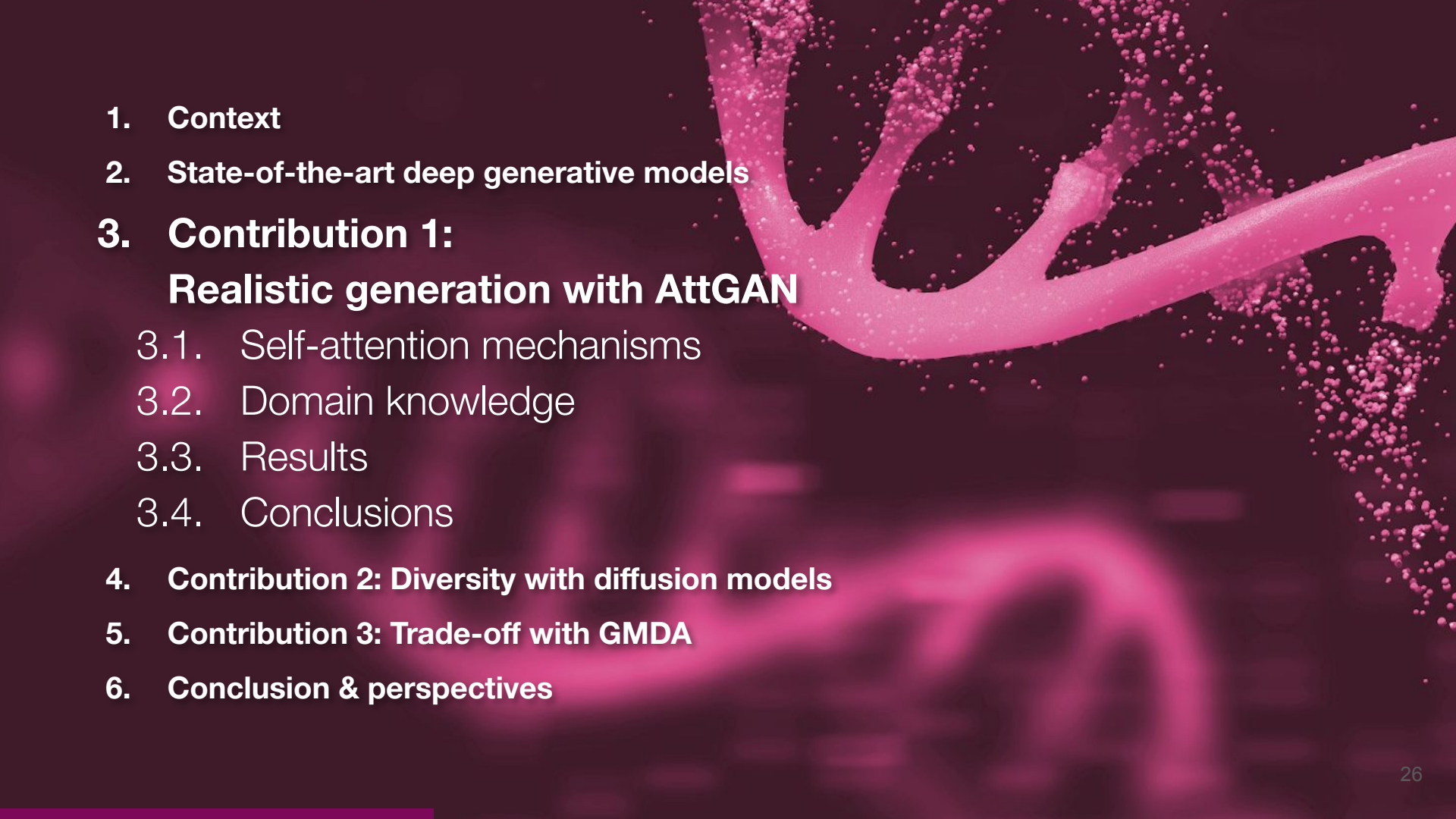
**Improved recall**
= distribution coverage





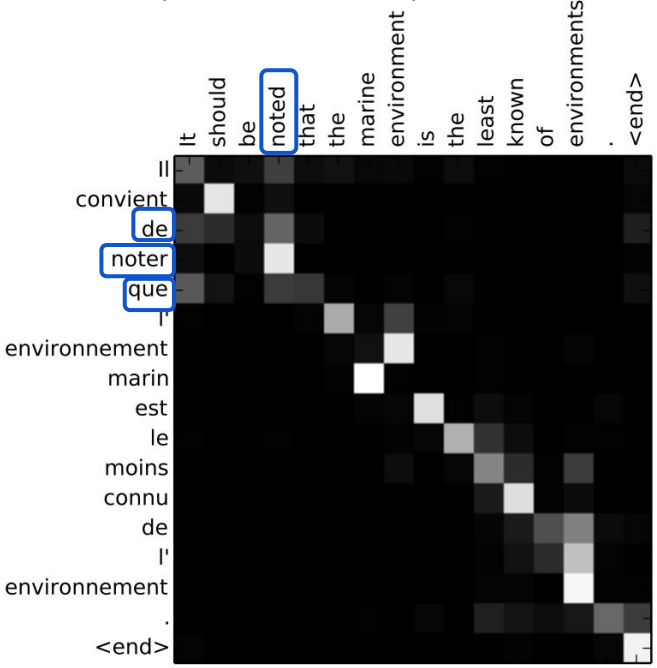

Kynkäänniemi et al. *Improved Precision and Recall Metric for Assessing Generative Models* (NeurIPS 2019)



1. Context
2. State-of-the-art deep generative models
3. **Contribution 1: Realistic generation with AttGAN**
   - 3.1. Self-attention mechanisms
   - 3.2. Domain knowledge
   - 3.3. Results
   - 3.4. Conclusions
4. Contribution 2: Diversity with diffusion models
5. Contribution 3: Trade-off with GMDA
6. Conclusion & perspectives

# Self-attention mechanisms

**Attention:** the relevance of each word in the input to the final output words




Attention map (Bahdanau et al., 2014)

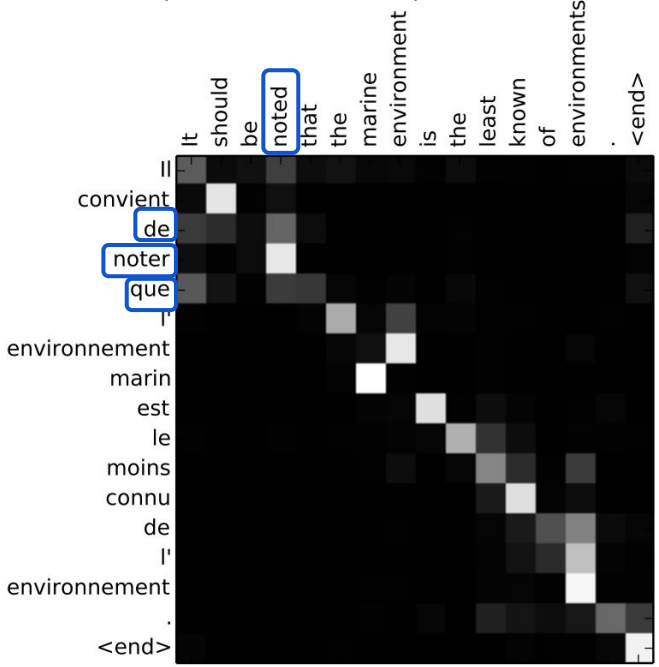
$$\text{Attention}(\mathbf{q}, \mathbf{K}, \mathbf{V}) = \sum_{i}^{n} \text{softmax}\left(\frac{\mathbf{q}\mathbf{K}^T}{\sqrt{d_k}}\right)_i \mathbf{V}_i$$

Focus on relevant elements = **context**

Vaswani et al. *Attention Is All You Need* (NeurIPS 2017)

# Self-attention mechanisms

**Attention:** the relevance of each word in the input to the final output words

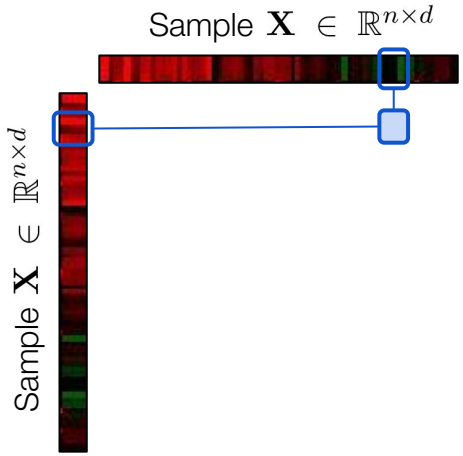



Attention map (Bahdanau et al., 2014)

$$\text{Attention}(\mathbf{q}, \mathbf{K}, \mathbf{V}) = \sum_i^n \text{softmax}\left(\frac{\mathbf{q}\mathbf{K}^T}{\sqrt{d_k}}\right)_i \mathbf{V}_i$$

Focus on relevant elements = **context**

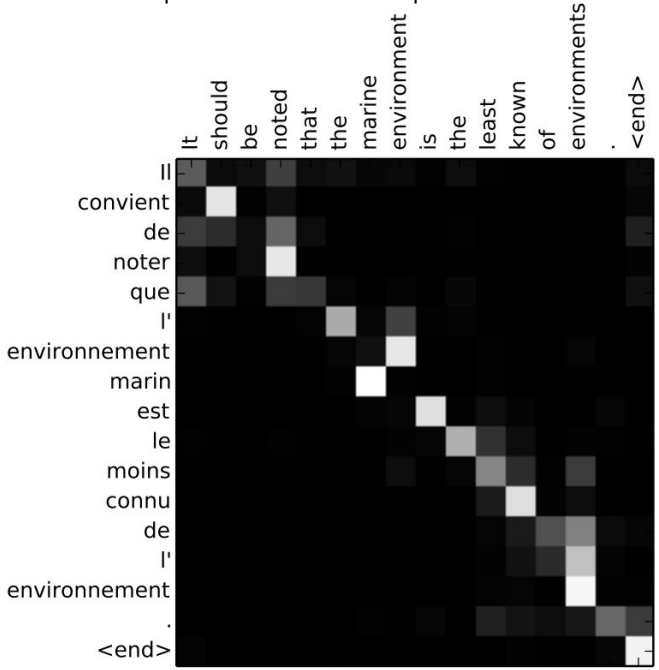
*Can we learn the context of each gene?*




Vaswani et al. *Attention Is All You Need* (NeurIPS 2017)

# Self-attention mechanisms

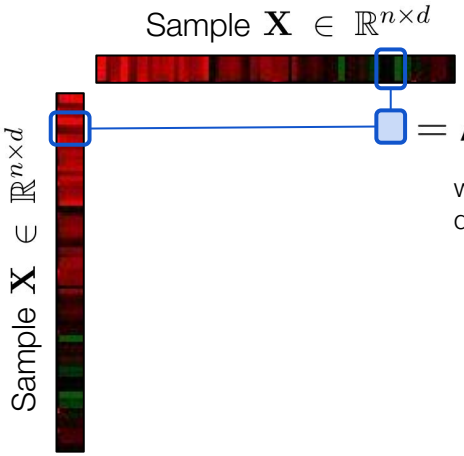
**Attention:** the relevance of each word in the input to the final output words




Attention map (Bahdanau et al., 2014)

$$\text{Attention}(\mathbf{q}, \mathbf{K}, \mathbf{V}) = \sum_i^n \text{softmax}\left(\frac{\mathbf{q}\mathbf{K}^T}{\sqrt{d_k}}\right)_i \mathbf{V}_i$$

**Self-attention:** captures the relevance of each element among other elements **within** a sequence




$$= \text{Attention}(\mathbf{X}\mathbf{W}_Q, \mathbf{X}\mathbf{W}_K, \mathbf{X}\mathbf{W}_V)$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ are queries, keys and values linear projection matrices

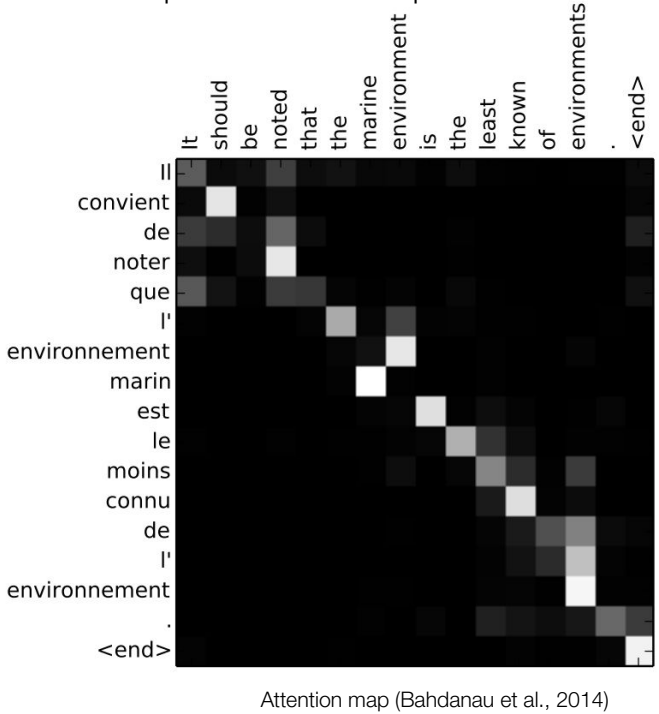
Our adaptation to tabular data:

$$Attention(\mathbf{q}, \mathbf{K}, \mathbf{V}) = \sum_{i=1}^{n} softmin\left(|\mathbf{q} - \mathbf{K}_i|\right)_i \mathbf{V}_i$$

Vaswani et al. *Attention Is All You Need* (NeurIPS 2017)

# Self-attention mechanisms

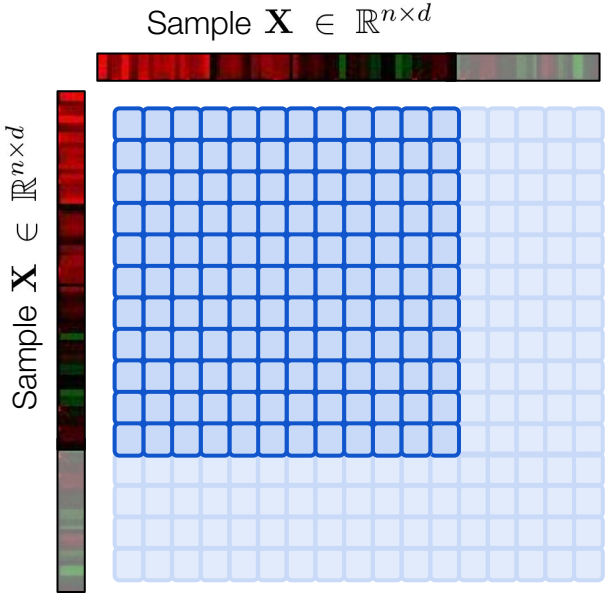
**Attention:** the relevance of each word in the input to the final output words




Attention map (Bahdanau et al., 2014)

$$\text{Attention}(\mathbf{q}, \mathbf{K}, \mathbf{V}) = \sum_i^n \text{softmax}\left(\frac{\mathbf{q}\mathbf{K}^T}{\sqrt{d_k}}\right)_i \mathbf{V}_i$$

**Self-attention:** captures the relevance of each element among other elements **within** a sequence

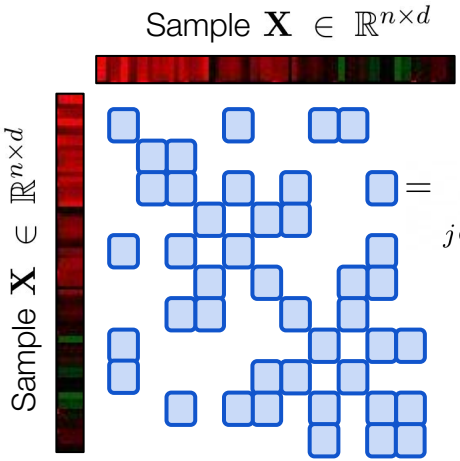



*How to handle quadratic complexity?*

Vaswani et al. *Attention Is All You Need* (NeurIPS 2017)

# Including sparse domain knowledge

**AttGAN:** WGAN-GP + self-attention module based on domain knowledge




$$= \sum_{j \in \mathcal{G}_{\mathbf{i};\mathbf{PPI}}} \alpha_{j,i} \mathbf{W}_V^{\mathbf{PPI}} x_j$$

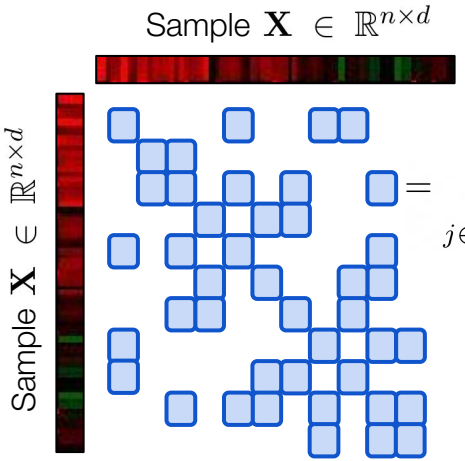
with $\mathcal{G}_{\mathbf{i};\mathbf{PPI}}$ the set of genes interacting the most with gene ***i*** and $\alpha$ the attention score

- **Co-expression (Co-exp):** statistical view
- **Protein-protein interactions (PPI):** functional view

Zhang et al. *Self-Attention Generative Adversarial Networks* (ICML 2019)

# Including sparse domain knowledge

**AttGAN:** WGAN-GP + self-attention module based on domain knowledge




$$= \sum_{j \in \mathcal{G}_{\mathbf{i};\mathbf{PPI}}} \alpha_{j,i} \mathbf{W}_V^{\mathbf{PPI}} x_j$$

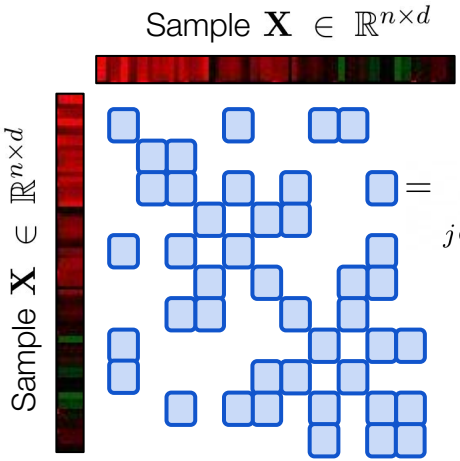
with $\mathcal{G}_{\mathbf{i};\mathbf{PPI}}$ the set of genes interacting the most with gene ***i*** and $\alpha$ the attention score

- **Co-expression (Co-exp):** statistical view
- **Protein-protein interactions (PPI):** functional view
- **Lesion study:** random interaction graph of same density

Zhang et al. *Self-Attention Generative Adversarial Networks* (ICML 2019)

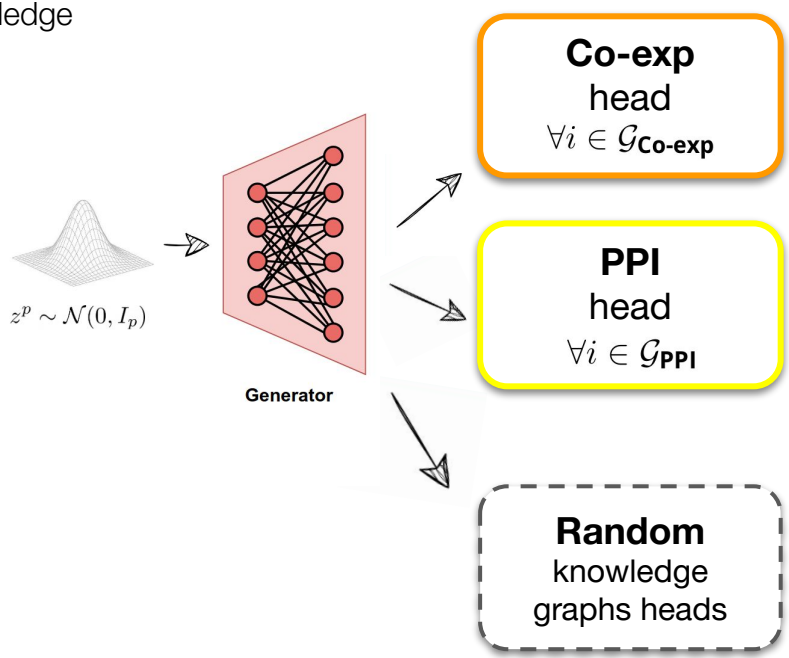
# Including sparse domain knowledge

**AttGAN:** WGAN-GP + self-attention module based on domain knowledge




$$= \sum_{j \in \mathcal{G}_{\mathbf{i};\mathbf{PPI}}} \alpha_{j,i} \mathbf{W}_V^{\mathbf{PPI}} x_j$$

with $\mathcal{G}_{\mathbf{i};\mathbf{PPI}}$ the set of genes interacting the most with gene ***i*** and $\alpha$ the attention score

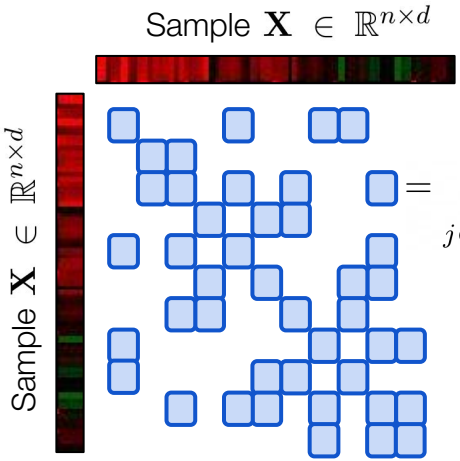



- **Co-expression (Co-exp):** statistical view
- **Protein-protein interactions (PPI):** functional view
- **Lesion study:** random interaction graph of same density

Zhang et al. *Self-Attention Generative Adversarial Networks* (ICML 2019)

# Including sparse domain knowledge

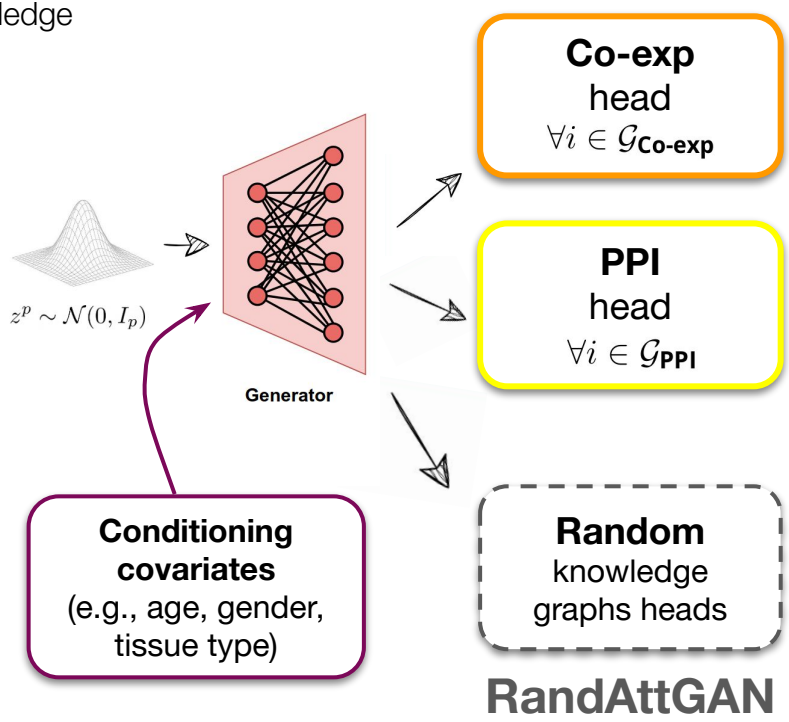
**AttGAN:** WGAN-GP + self-attention module based on domain knowledge




$$= \sum_{j \in \mathcal{G}_{\mathbf{i};\mathbf{PPI}}} \alpha_{j,i} \mathbf{W}_V^{\mathbf{PPI}} x_j$$

with $\mathcal{G}_{\mathbf{i};\mathbf{PPI}}$ the set of genes interacting the most with gene ***i*** and $\alpha$ the attention score

- **Co-expression (Co-exp):** statistical view
- **Protein-protein interactions (PPI):** functional view
- **Lesion study:** random interaction graph of same density



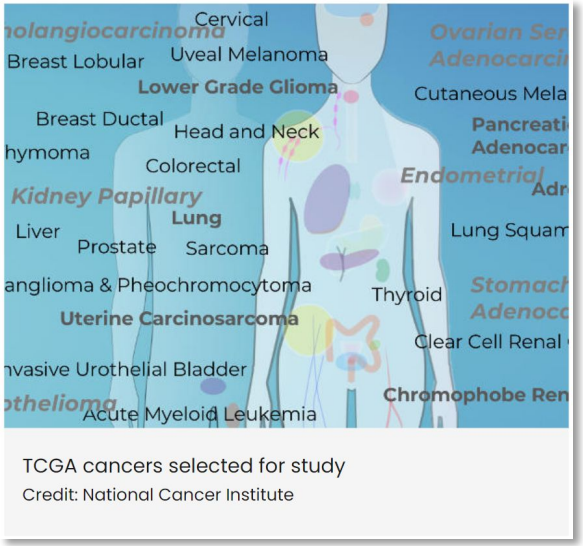

Zhang et al. *Self-Attention Generative Adversarial Networks* (ICML 2019)

# Evaluation of AttGAN and RandAttGAN

**Baselines:**
GAN, WGAN-GP

**Performance indicators:**
Correlations, Prediction performance (MLE)

**Adapted indicators:**
Fréchet distance (2 MLPs pre-trained for binary/multiclass tasks), precision/recall



TCGA cancers selected for study
Credit: National Cancer Institute

**Benchmark dataset:**
The Pan-Cancer Genome Atlas (TCGA) with 20,531 genes and ~10k samples
*Covariates: patient age, gender, tissue type, cancer target*

Weinstein et al. *The Cancer Genome Atlas Pan-Cancer analysis project.* (Nat Genet 2013)

# Visual evaluation

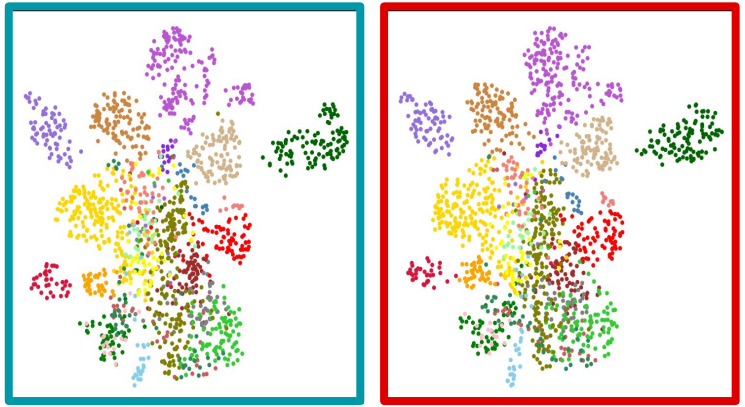
UMAP GAN

Generated

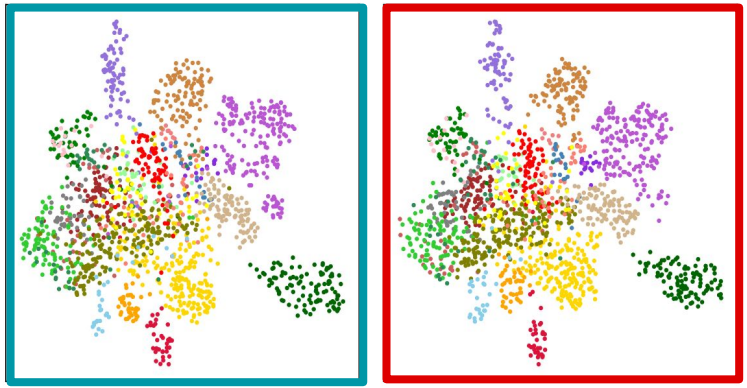
GAN fails

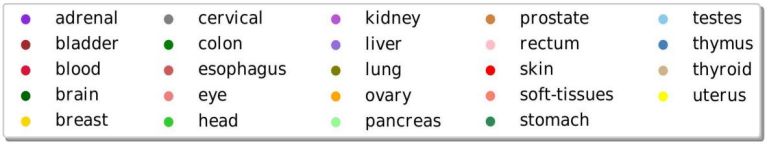
WGAN-GP and AttGAN preserve tissue clusters

UMAP WGAN-GP




UMAP AttGAN



# Data quality indicators

| Model | Corr. ↑ | Precision ↑ | Recall ↑ | FD binary ↓ | FD tissue ↓ |
|---|---|---|---|---|---|
| GAN | 14.40 | 80.3 ± 0.27 | 0.0 ± 0.0 | 1506121 ± 2617 | 96611 ± 99 |
| WGAN-GP | **90.98** | **99.21 ± 0.09** | 49.32 ± 0.24 | **16452 ± 1531** | 638 ± 36 |
| AttGAN PPI + CoExp + $\gamma$ fixed | 86.21 | 79.45 ± 0.29 | **72.03 ± 0.28** | 32507 ± 1119 | **556 ± 14** |

- WGAN-GP outperforms on most indicators
- Mode collapse issue

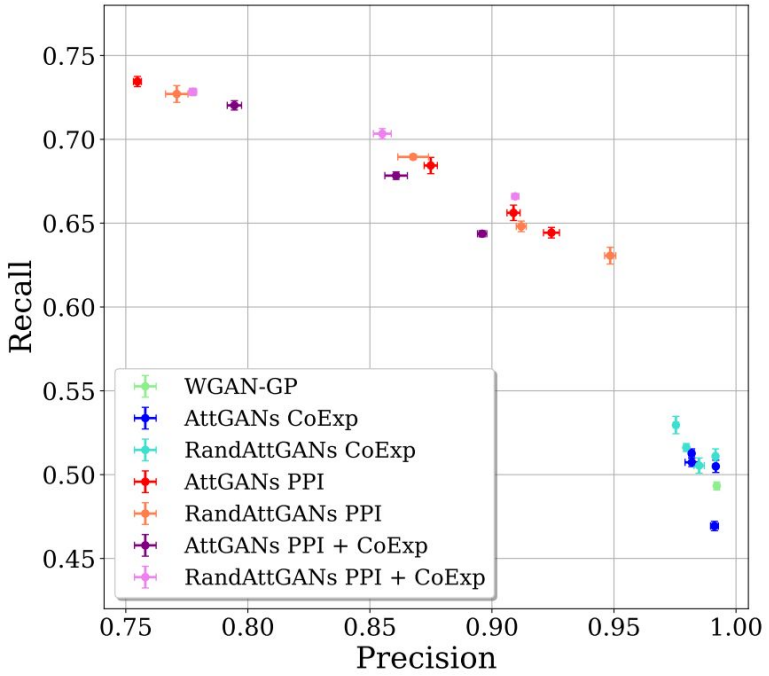
# Data quality indicators

| Model | Corr. ↑ | Precision ↑ | Recall ↑ | FD binary ↓ | FD tissue ↓ |
|---|---|---|---|---|---|
| GAN | 14.40 | 80.3 ± 0.27 | 0.0 ± 0.0 | 1506121 ± 2617 | 96611 ± 99 |
| WGAN-GP | **90.98** | **99.21 ± 0.09** | 49.32 ± 0.24 | **16452 ± 1531** | 638 ± 36 |
| AttGAN PPI + CoExp + $\gamma$ fixed | 86.21 | 79.45 ± 0.29 | **72.03 ± 0.28** | 32507 ± 1119 | **556 ± 14** |

WGAN-GP outperforms on most indicators

Mode collapse issue

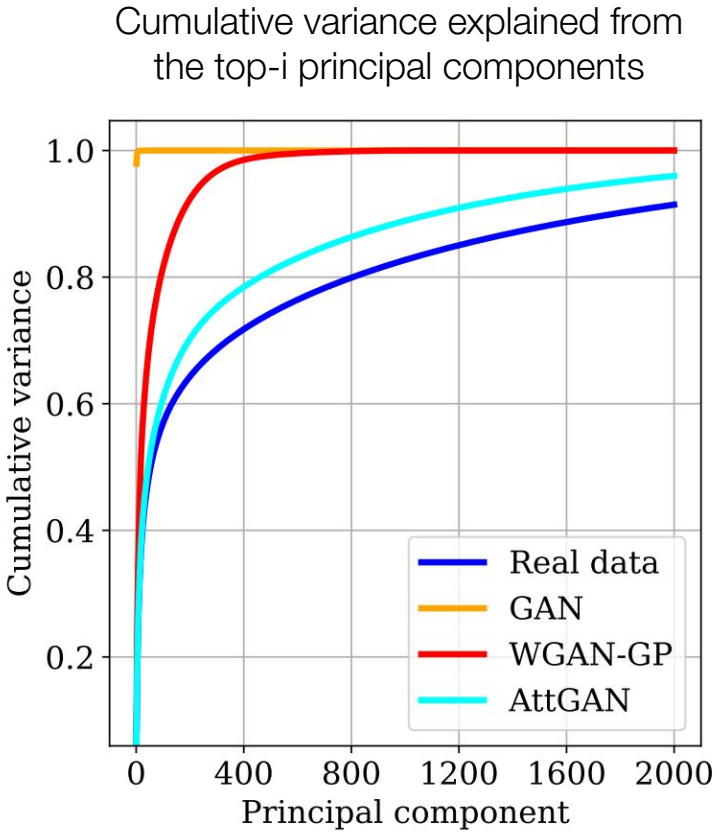
**Precision (fidelity) vs. recall (diversity) trade-off:**
AttGANs with different attention masks
(PPI, CoExp, CoExp-PPI, Random) and settings

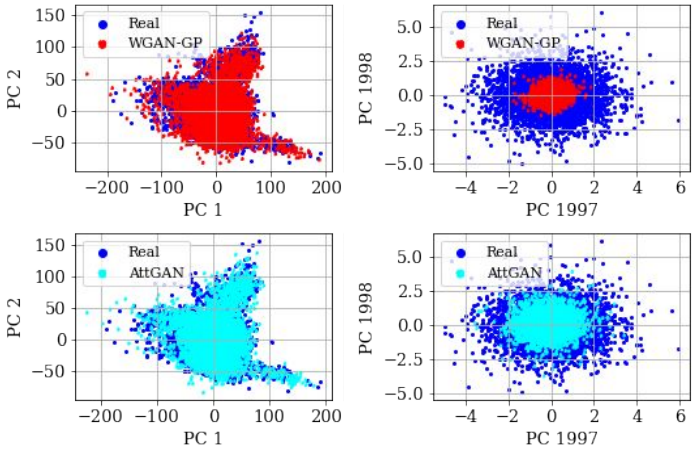



AttGAN allows to play on this trade-off

Lesion study: no impact of random attention

# PCA comparison



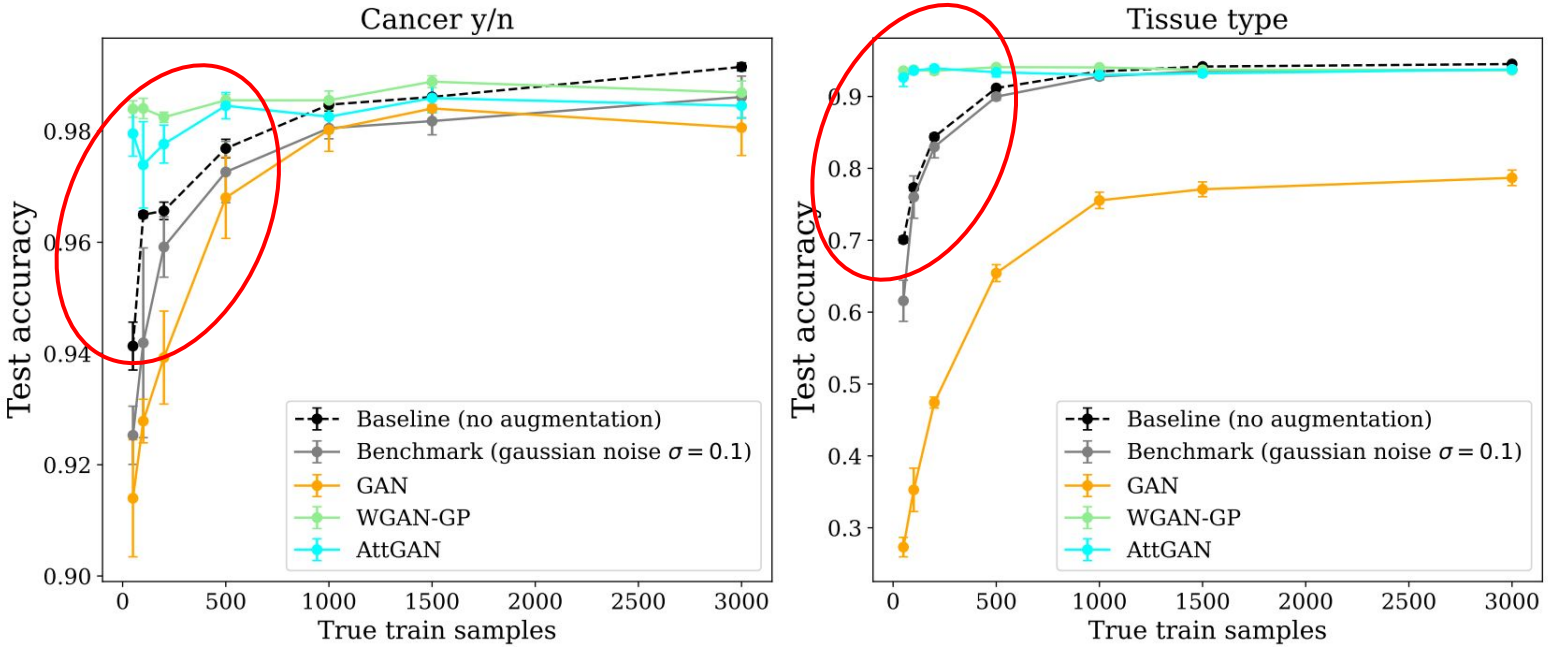




Variance explained by less than 500 PCs in WGAN-GP generated data

AttGAN preserves more information w.r.t. PCA

# Predictive accuracy with Data Augmentation

Test accuracy of a MLP trained with *N* true samples + 8000 augmented samples




- WGAN-GP and AttGAN reach significantly higher accuracy with fewer true training samples

# Partial conclusions

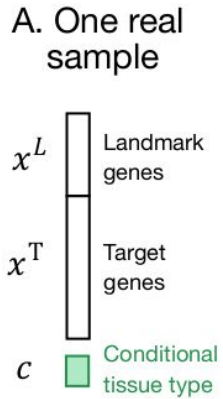
## Take-aways

- ❑ **Evaluation:** a multi-objective problem
- ❑ **Performance:** WGAN-GP outperforms AttGAN on fidelity but lacks diversity (e.g., recall and PCA), while AttGAN reaches better fidelity-diversity trade-off
- ❑ **Lesion study:** attention performance depends on the additional expressivity not the injected knowledge
- ❑ **Data augmentation yields very good performance** with very limited true data: gain of ~4%/20% accuracy for binary/multiclass tasks

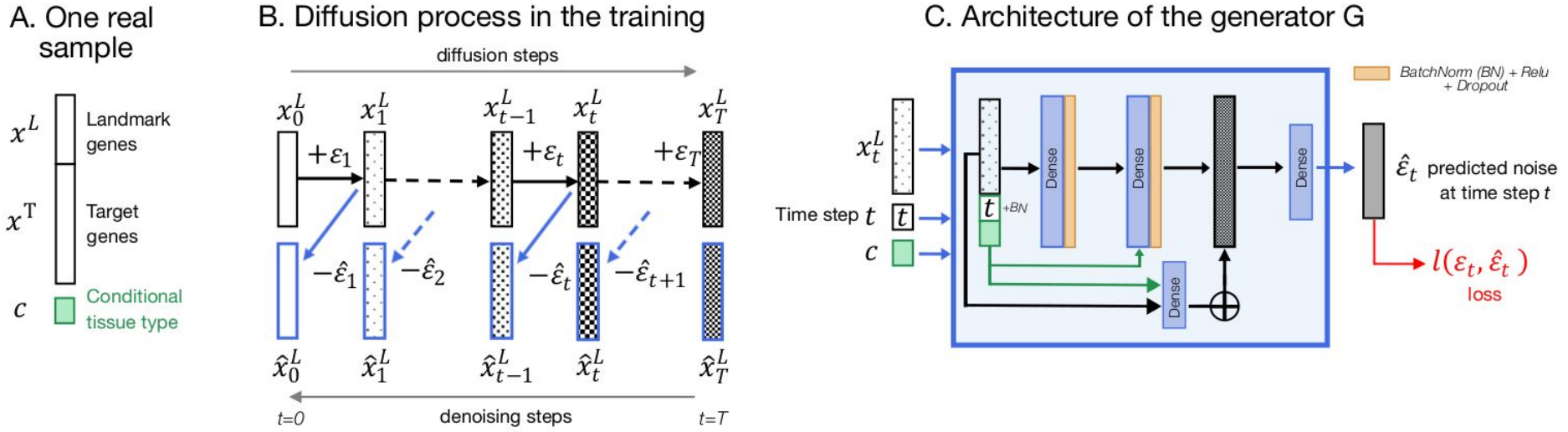
**Publication:**
A. Lacan, M. Sebag and B. Hanczar. "GAN-based data augmentation for transcriptomics : survey and comparative assessment". In: ISMB, June 2023.

1. **Context**
2. **State-of-the-art deep generative models**
3. **Contribution 1: Realistic generation with AttGAN**
4. **Contribution 2: Diversity with diffusion models**
   - 4.1. Diffusion framework
   - 4.2. Interpolation strategy
   - 4.3. Results
   - 4.4. Conclusions
5. **Contribution 3: Trade-off with GMDA**
6. **Conclusion & perspectives**

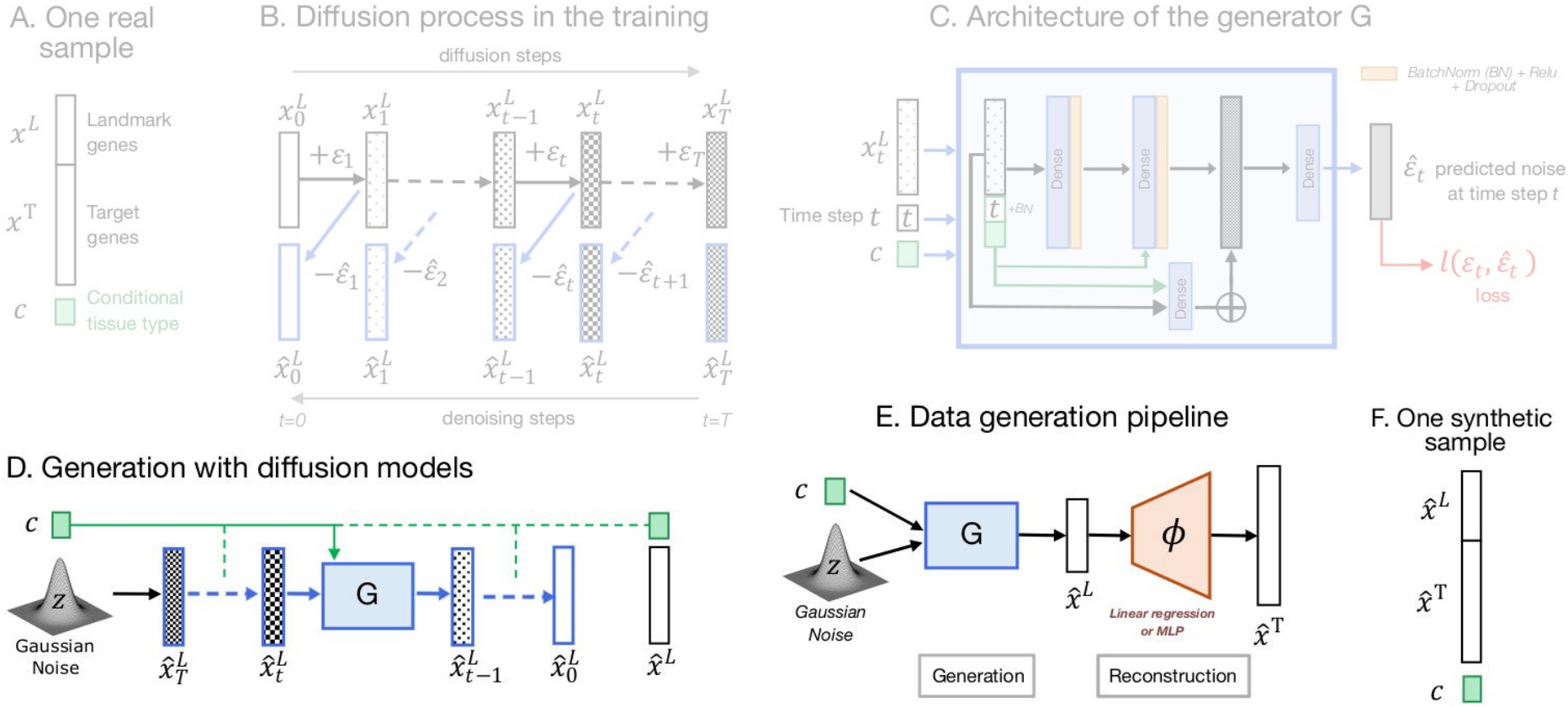
# Diffusion framework: DDPM and DDIM



# Diffusion framework: DDPM and DDIM

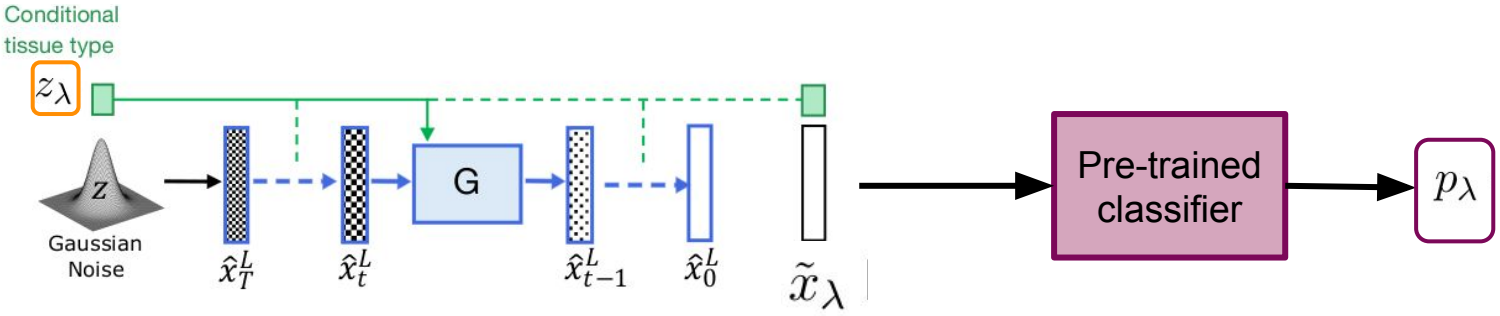


# Diffusion framework: DDPM and DDIM



# Interpolation strategy

*Can we improve diversity with interpolated data?*

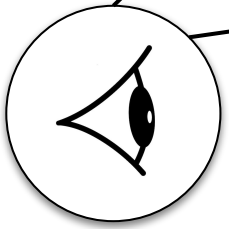



**Definition** Let $\lambda \in \{0., 0.25, 0.5, 0.75, 1.\}$ be the interpolation weight. Let $t_1$ and $t_2$ be tissue types conditional variables. Let $z_\theta(t)$ be the conditional embedding obtained from input tissue $t$ and parameterized by the embedding layer parameters $\theta$. The final LERP embedding is obtained as follows :

$$z_\lambda = \lambda z_\theta(t_2) + (1 - \lambda) z_\theta(t_1)$$

# Evaluation of DDPM and DDIM

**Baselines:**
VAE, WGAN-GP



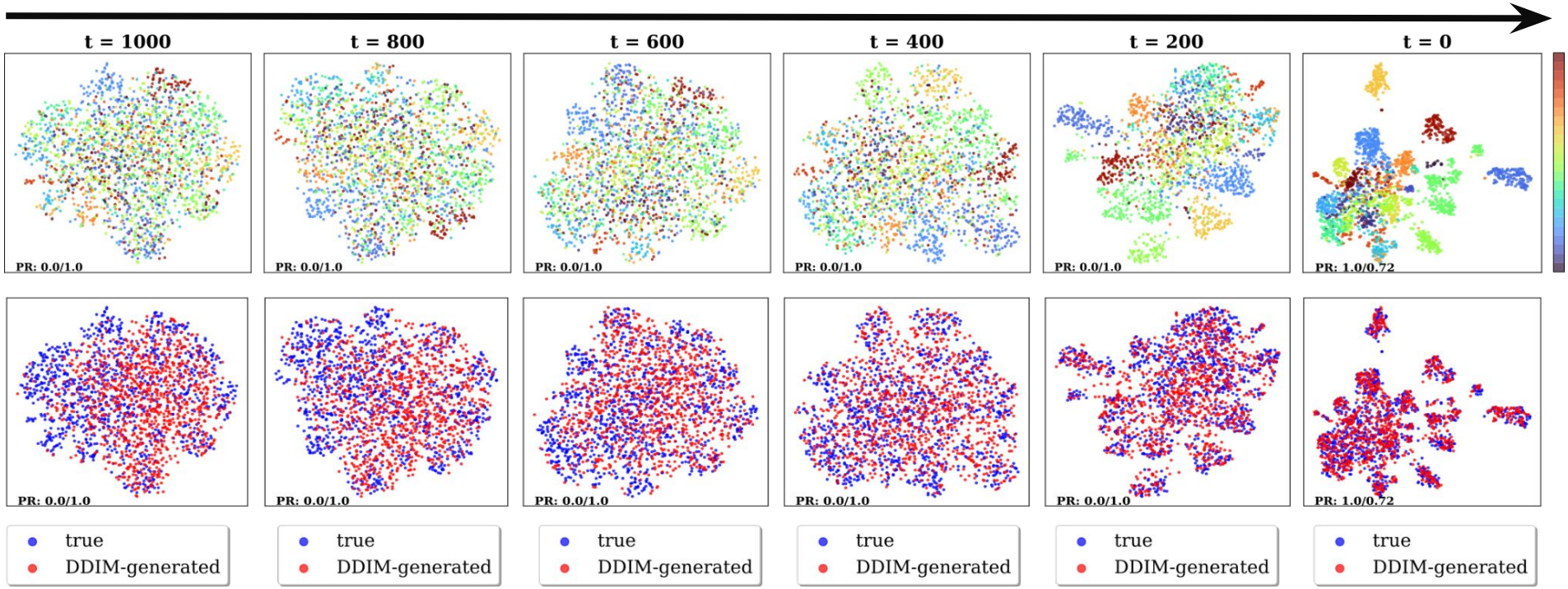
**Performance indicators:**
Correlations error, Prediction performance (MLE), Precision/recall (F1 score)

**Benchmark datasets:**

- The Pan-Cancer Genome Atlas (TCGA) with *d*=1,000 genes and *n*=~10k.
  *Covariate: tissue type*
- The Genotype-Tissue Expression (GTEx) with *d*=1,000 genes and *n*=~17k.
  *Covariate: tissue type*

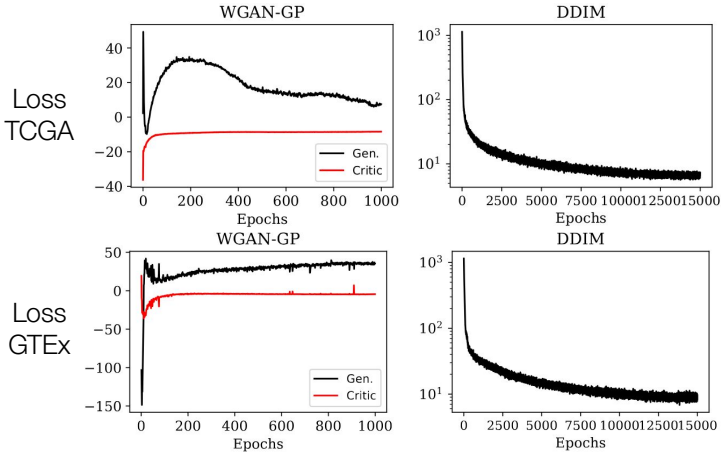
Lonsdale et al. *The genotype-tissue expression* (Nature Genetics 2013)

# Visualizing diffusion process



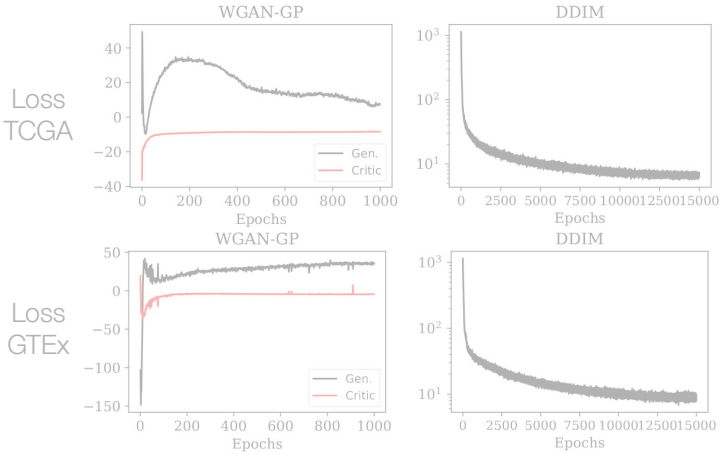

*Top: UMAP of TCGA generated data with tissue coloring.*
*Bottom: Same UMAPs with true vs. generated coloring.*

# Results



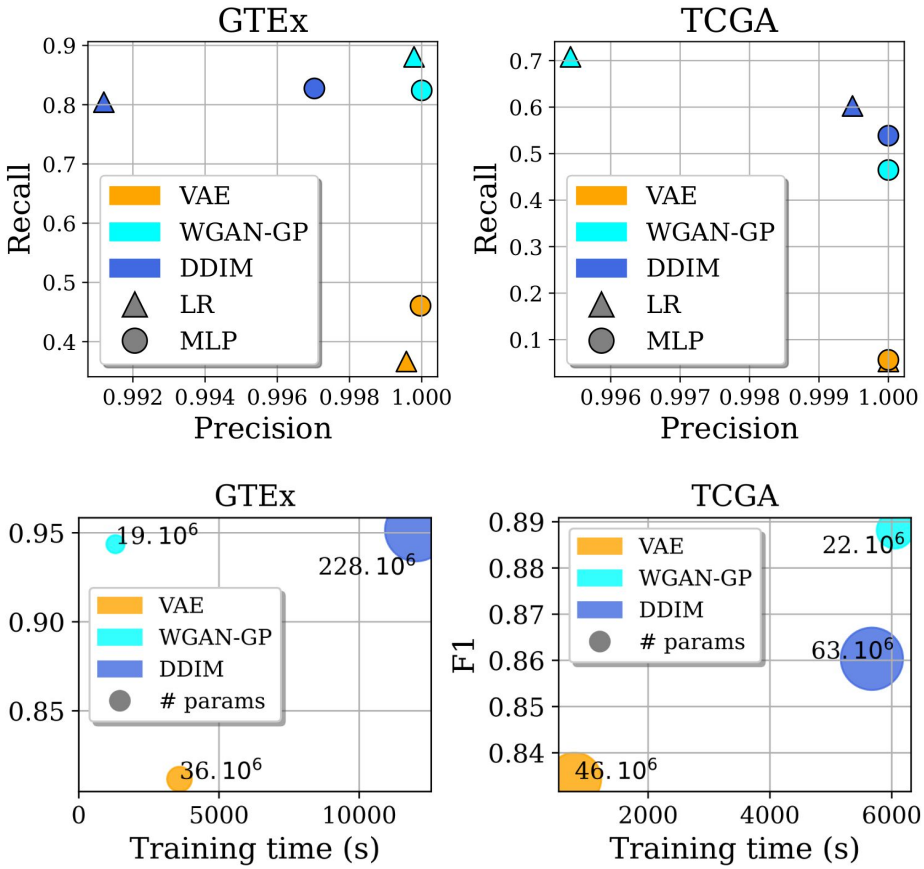

DDIM training objective is more stable

# Results




DDIM training objective is more stable

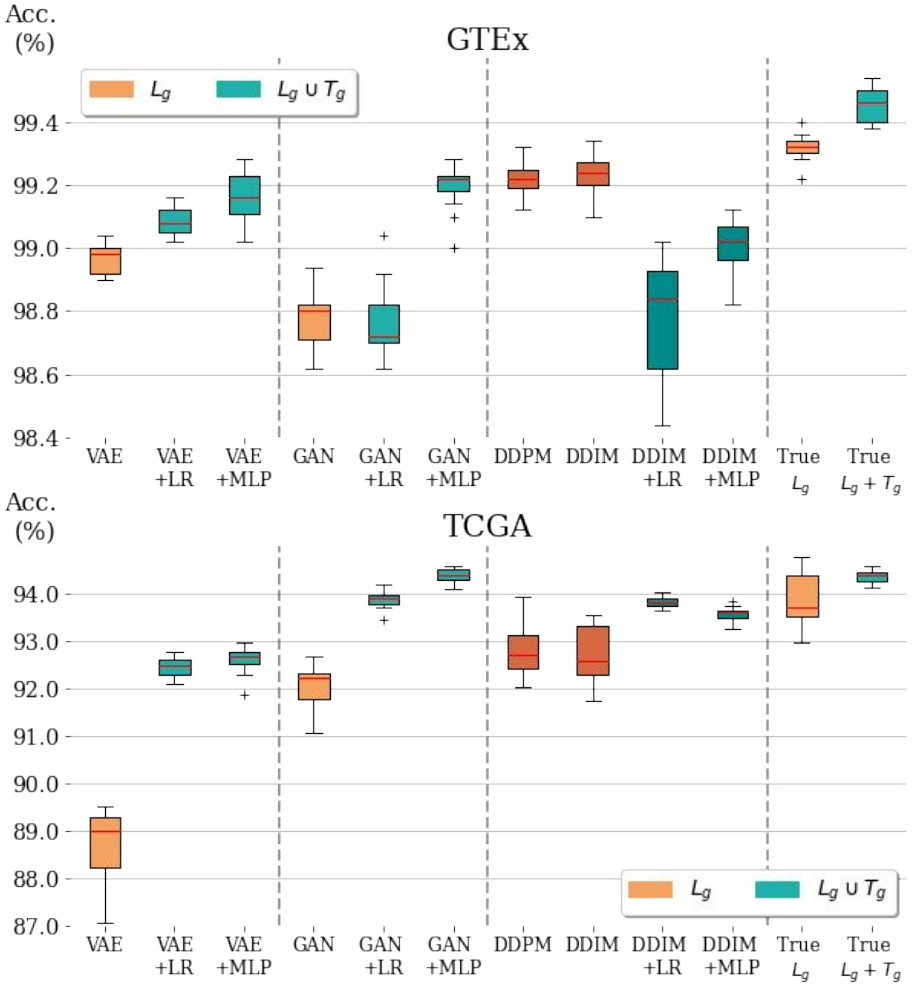
WGAN-GP reaches the best performance on the pareto front




DDIM needs 12x (resp. 3x) more parameters to reach the same performance on GTEx (resp. TCGA)

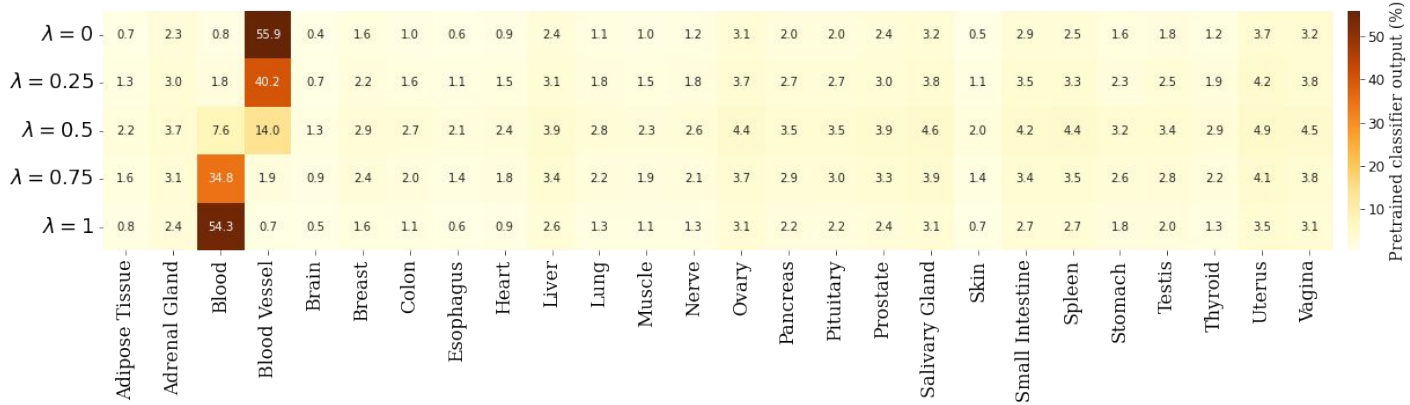
# Prediction performance

- DDPM and DDIM reach the best accuracy in reduced L1000 space
- WGAN-GP remains very competitive with state-of-the-art results in reconstructed space

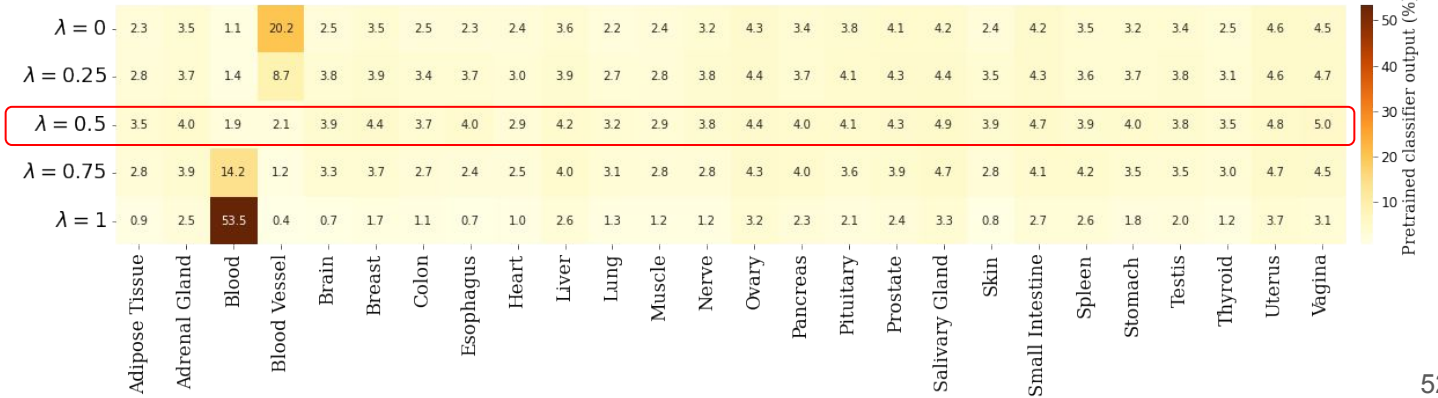


# Interpolation results

***WGAN-GP:*** *Interpolations between "blood vessel" and "blood" tissues in GTEx-generated data.*




***DDIM***




➔ DDIM latent space ill-suited for linear interpolations

# Partial conclusions

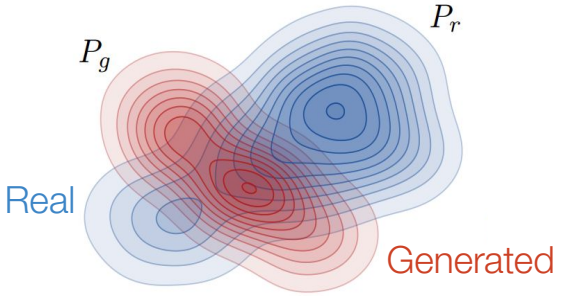
## Take-aways

- ❏ **Performance:** WGAN-GP outperforms DMs except on MLE in reduced space where DDIM ranks first
- ❏ **Complexity:** diffusion is 3-12x more complex while reaching similar fidelity-diversity trade-off
- ❏ **Interpolations:** DMs latent space is less suited for linear interpolations than the one of WGAN-GP

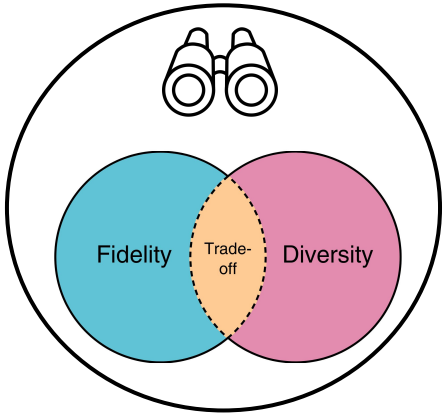
**Diffusion for transcriptomics (preprint):**
A. Lacan, R. André, M. Sebag and B. Hanczar. "In Silico Generation of Gene Expression profiles using Diffusion Models". In: bioRxiv, 2024.

# Can we improve the generated data quality?

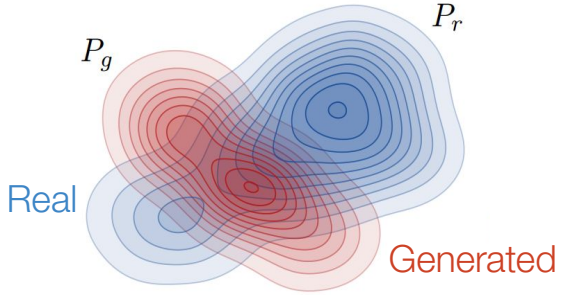
- ➔ High-dimensional distribution remains complex
- ➔ Sophisticated architectures required
- ➔ Distributions comparison is unanswered (loss is decorrelated from indicators)



# Can we improve the generated data quality?



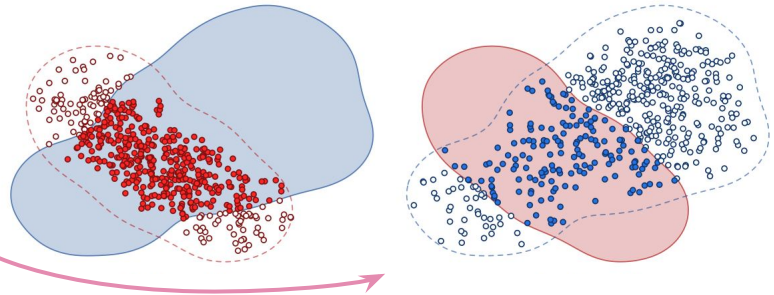

- ➔ High-dimensional distribution remains complex
- ➔ Sophisticated architectures required
- ➔ Distributions comparison is unanswered (loss is decorrelated from indicators)




**Motivation:**
Frugal approximation and comparison with Precision/Recall

**Improved precision**
= distribution overlap

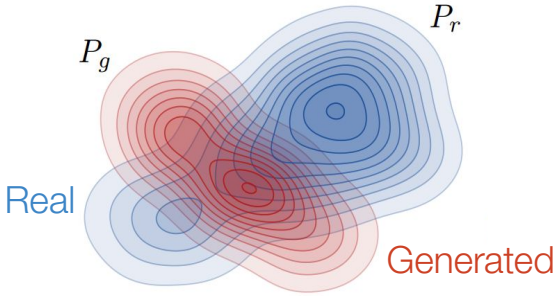
**Improved recall**
= distribution coverage

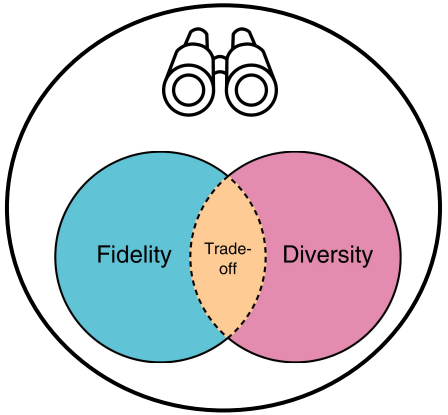


*Can we use such comparison as a training objective?*

# Can we improve the generated data quality?

- ➔ High-dimensional distribution remains complex
- ➔ Sophisticated architectures required
- ➔ Distributions comparison is unanswered (loss is decorrelated from indicators)

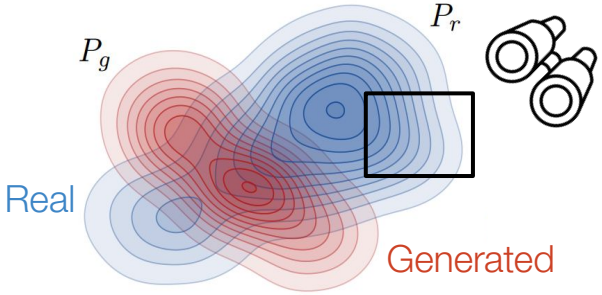


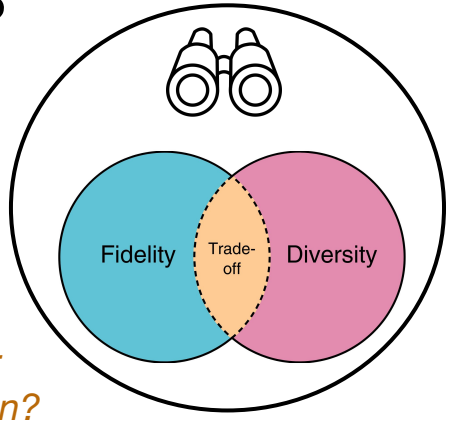

**Intuition:**
Two distributions are similar if same support within **any region** of the space

# Can we improve the generated data quality?

- ➔ High-dimensional distribution remains complex
- ➔ Sophisticated architectures required
- ➔ Distributions comparison is unanswered (loss is decorrelated from indicators)

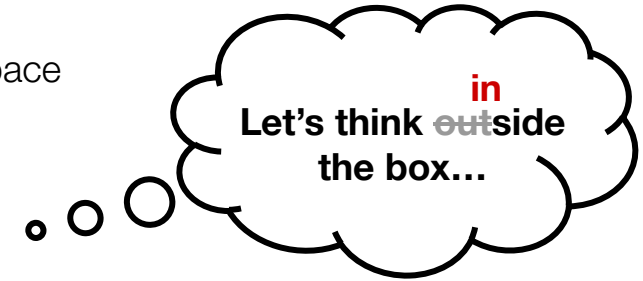



*Do we have similar densities in this region?*



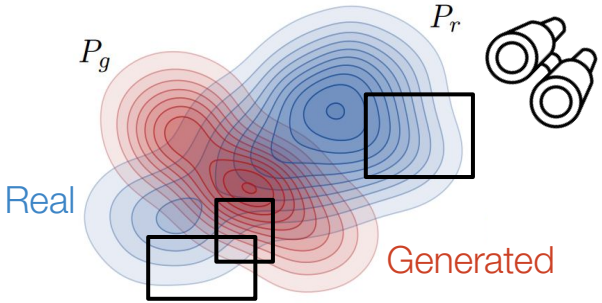

**Intuition:**
Two distributions are similar if same support within **any region** of the space

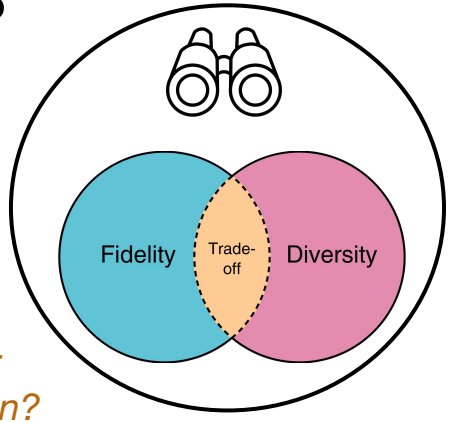
**Proposition:**
Frugal regions of 2D-3D rectangles uniformly sampled (density probes)



# Can we improve the generated data quality?

- ➔ High-dimensional distribution remains complex
- ➔ Sophisticated architectures required
- ➔ Distributions comparison is unanswered (loss is decorrelated from indicators)



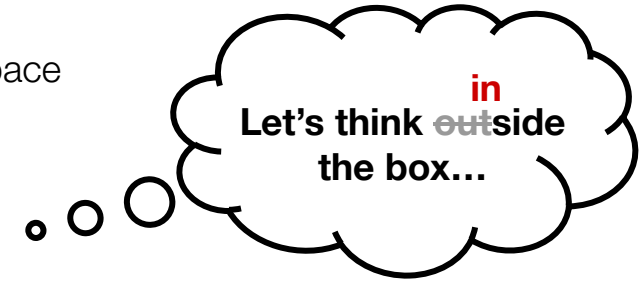



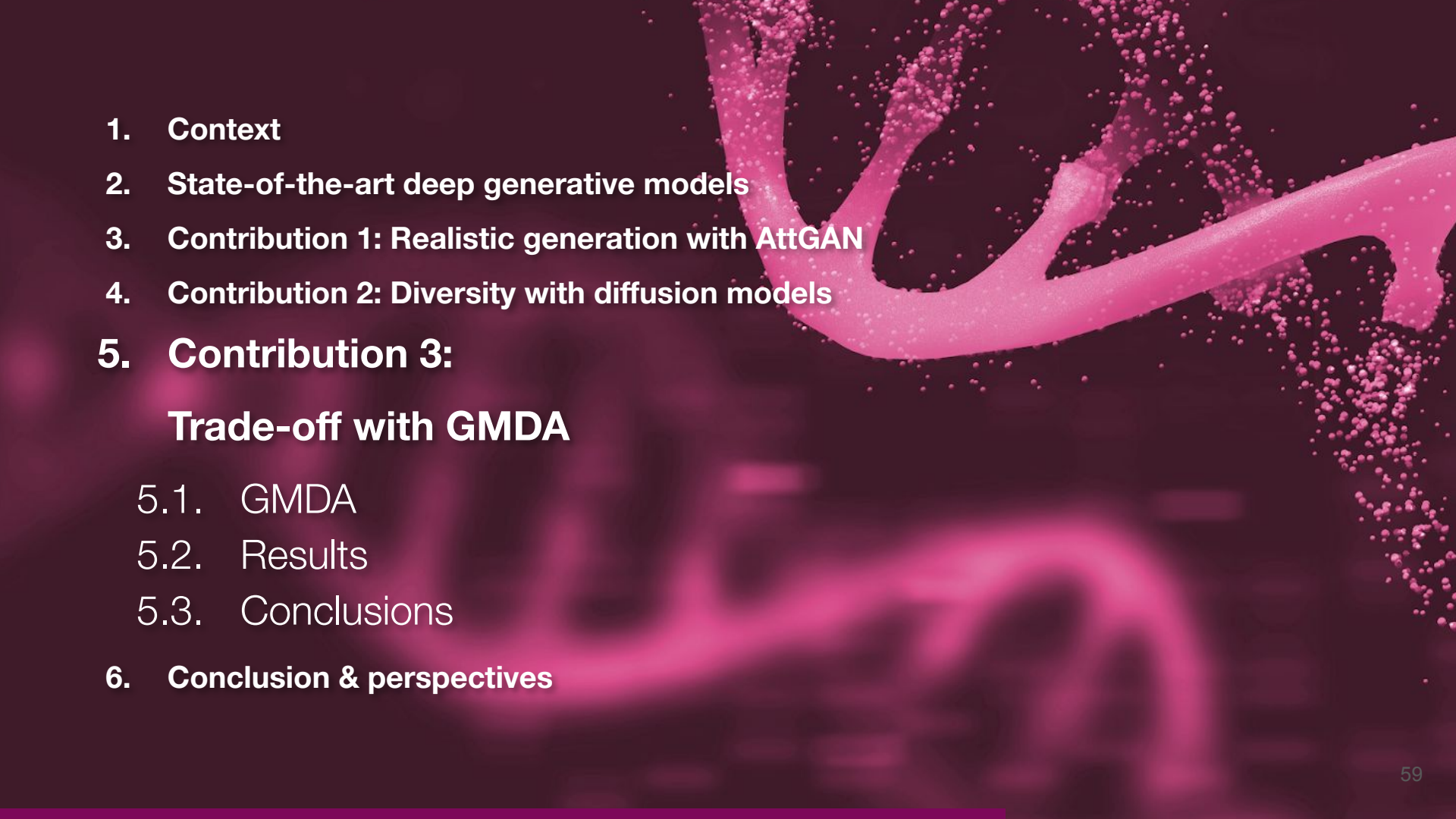

*Do we have similar densities in this region?*

*Do we have similar densities in all regions?*

**Intuition:**
Two distributions are similar if same support within **any region** of the space

**Proposition:**
Frugal regions of 2D-3D rectangles uniformly sampled (density probes)

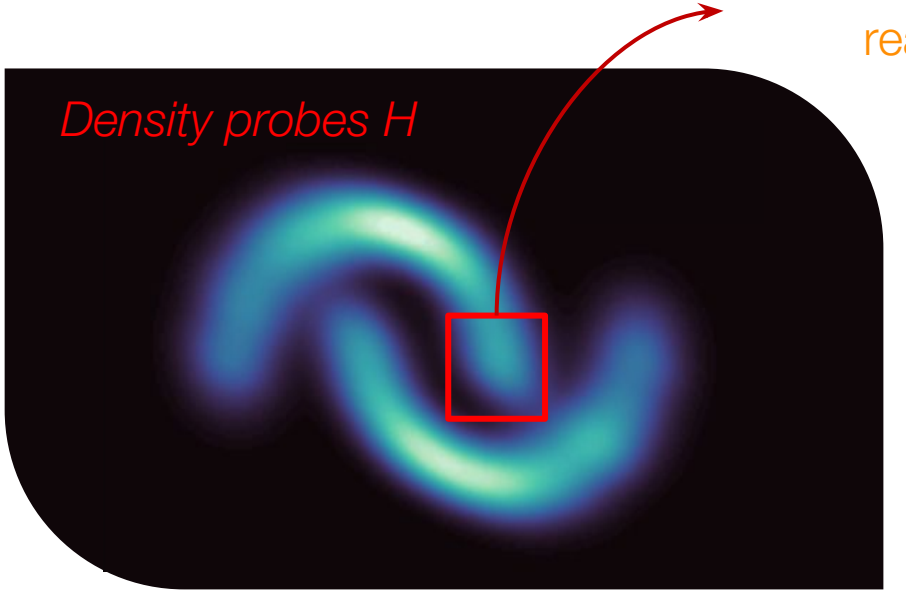




1. Context
2. State-of-the-art deep generative models
3. Contribution 1: Realistic generation with AttGAN
4. Contribution 2: Diversity with diffusion models
5. **Contribution 3: Trade-off with GMDA**
   - 5.1. GMDA
   - 5.2. Results
   - 5.3. Conclusions
6. Conclusion & perspectives

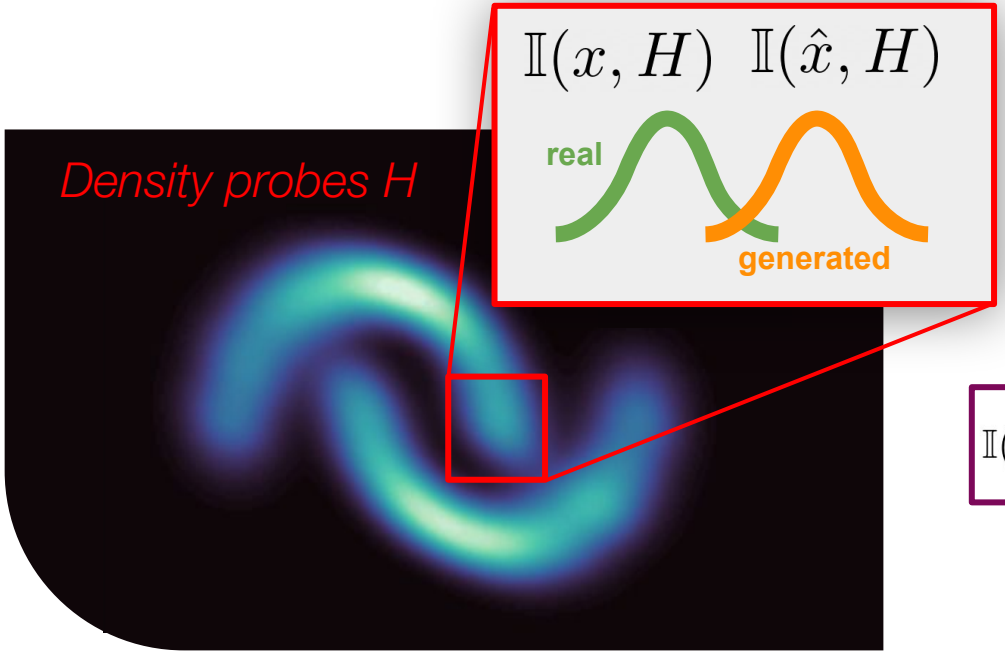
# Generative Modeling with Density Alignment (GMDA)

Do we have the same density of real and fake samples?




- **Milestone 1:**
  Differentiable density approximation
- **Milestone 2:**
  Enforce local and global density alignment
- **Milestone 3:**
  Stochastic density probes (no trainable adversary)

# Density approximation



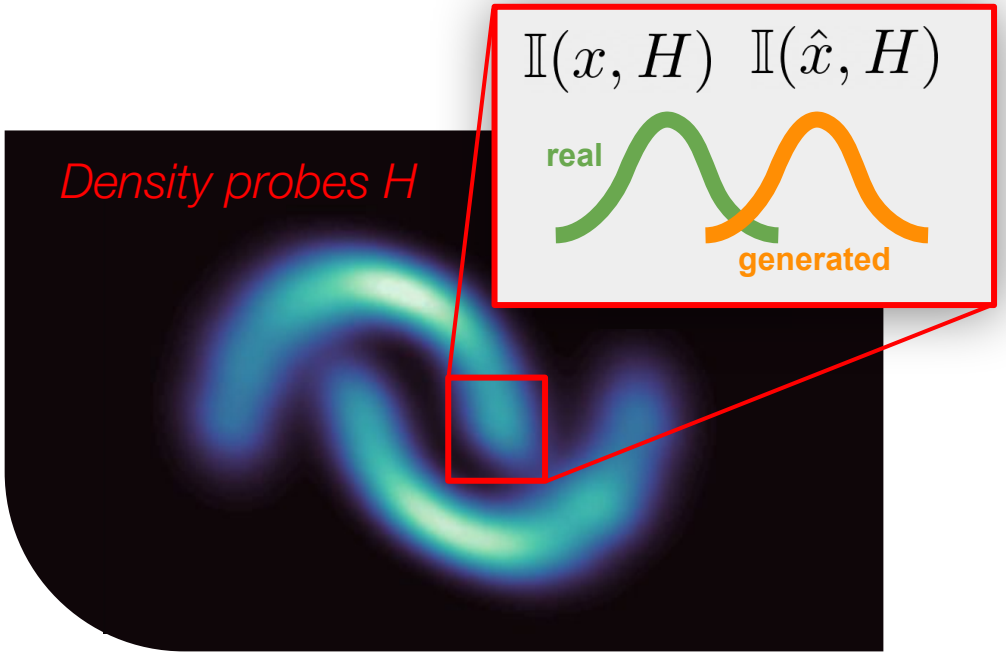

Given $\mathbf{d}$ the dimension of $\mathbf{x}$, $\mathbf{a}$ and $\mathbf{b}$ the probe's interval, we approximate the indicator with **sigmoids** parameterized by $\boldsymbol{\lambda}$:

$$\mathbb{I}(x, [a, b]) = \left(\frac{1}{1 + e^{-\lambda(x-a)}}\right) \times \left(\frac{1}{1 + e^{-\lambda(b-x)}}\right)$$

**Differentiable indicator function:**
cartesian product between 2 or 3 intervals (uniformly sampled)

$$\mathbb{I}(\mathbf{x}, H) = \prod_{i=1}^{d} I(x_i, [a_i, b_i])$$

# Density alignment




➔ Generator $G_\theta$ takes local density as feedback (**no adversary**)

Density alignment with **learning criterion**:

$$\mathcal{L} = \sum_{i=1}^{K} \mathcal{L}(H_i) + w_{DH}\mathcal{L}(DH)$$

Loss per probe

Dark probe

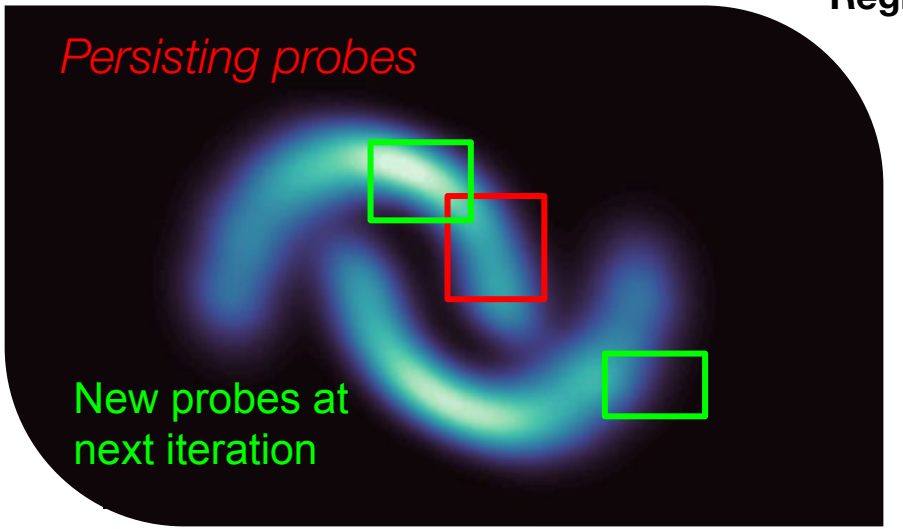
Regions not visited by current probes:

$$\prod_{i=1}^{K}(1 - \mathbb{I}(\mathbf{x}, H_i))$$

**Discrepancy** within probe:

$$\mathcal{L}(H) = \frac{\left|\sum_{\mathbf{x}\in\mathcal{D}} I(\mathbf{x}, H) - \sum_{\mathbf{x}'\in\mathcal{G}} \mathbb{I}(\mathbf{x}', H)\right|}{\log\left(1 + \max\left(\sum_{\mathbf{x}\in\mathcal{D}} \mathbb{I}(\mathbf{x}, H), \sum_{\mathbf{x}'\in\mathcal{G}} \mathbb{I}(\mathbf{x}', H)\right)\right)}$$

# Density probes sampling




**Initialization:**

**Regions:** probes centered on real points uniformly sampled

**Width:** based on desired density rate $\boldsymbol{\delta}$

If fixed probes, inefficient loss

If stochastic probes, highly varying loss

**Mixed strategy:**

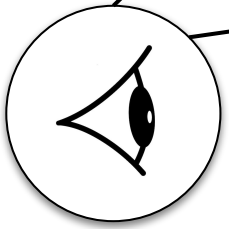
**Exploitation:** persistence of $\eta$ % of probes with high loss

**vs.**

**Exploration:** (1-$\eta$ )% of stochastic probes sampling

# Evaluation of GMDA

**Baselines:**
TVAE, CTGAN, TabDDPM
*Transcriptomics:* VAE, WGAN-GP



**Performance indicators:**
Correlations error, Prediction performance (MLE), Precision/recall (F1 score)

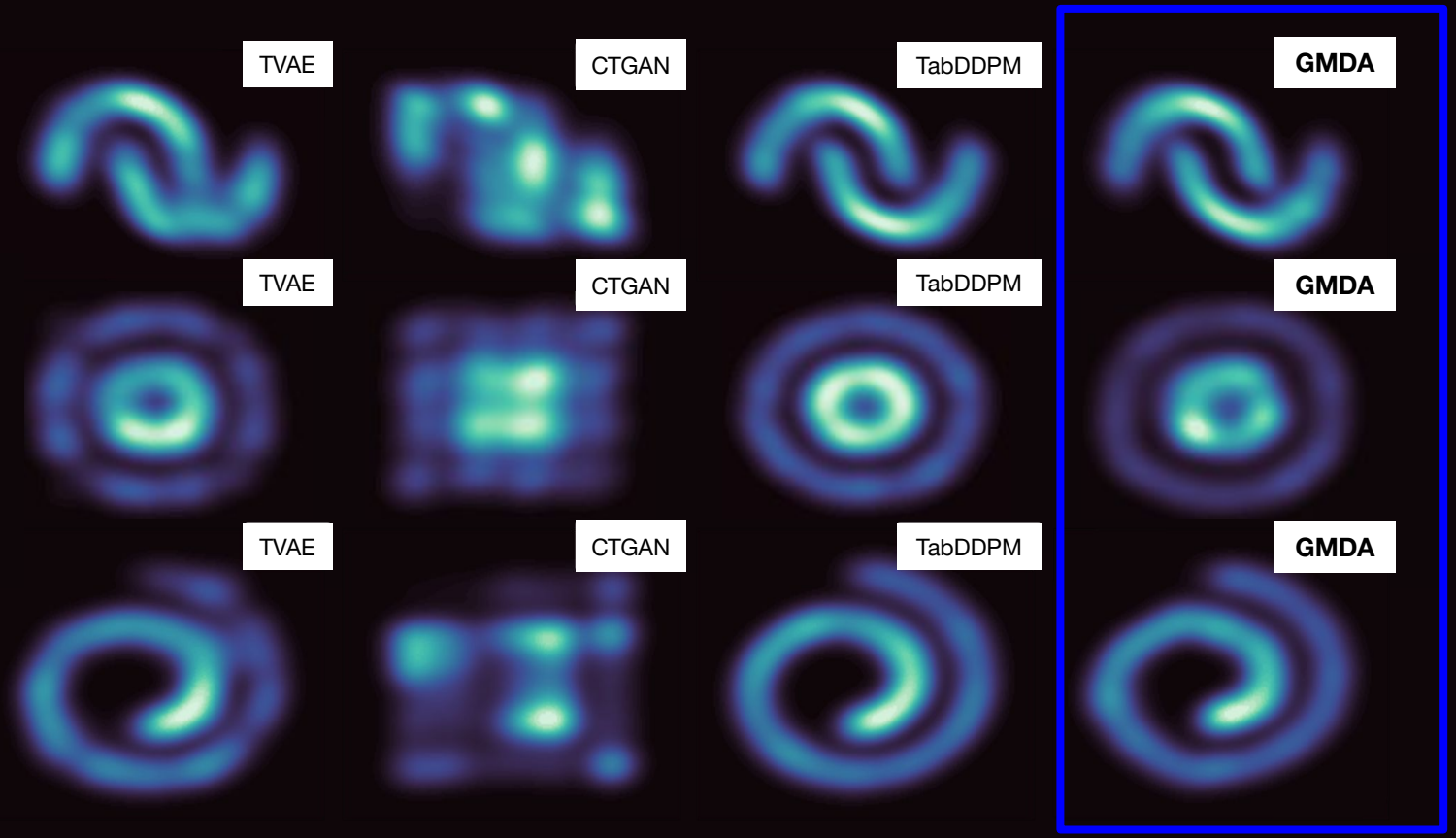
**Benchmark datasets:**

- ❏ 3 2D toy datasets
- ❏ 4 medium size tabular datasets (*d*= 5 to 32 features, *n*=500 to 15k)
- ❏ The Pan-Cancer Genome Atlas (TCGA) with *d*=1,000 genes and *n*=~10k.
  *Covariate: tissue type*
- ❏ The Genotype-Tissue Expression (GTEx) with *d*=1,000 genes and *n*=~17k.
  *Covariate: tissue type*

Lonsdale et al. *The genotype-tissue expression* (Nature Genetics 2013)
Xu, *et al. Modeling tabular data using conditional GAN* (NeurIPS 2019)
Kotelnikov *et al*. *TabDDPM: Modelling tabular data with diffusion models* (ICML 2023)

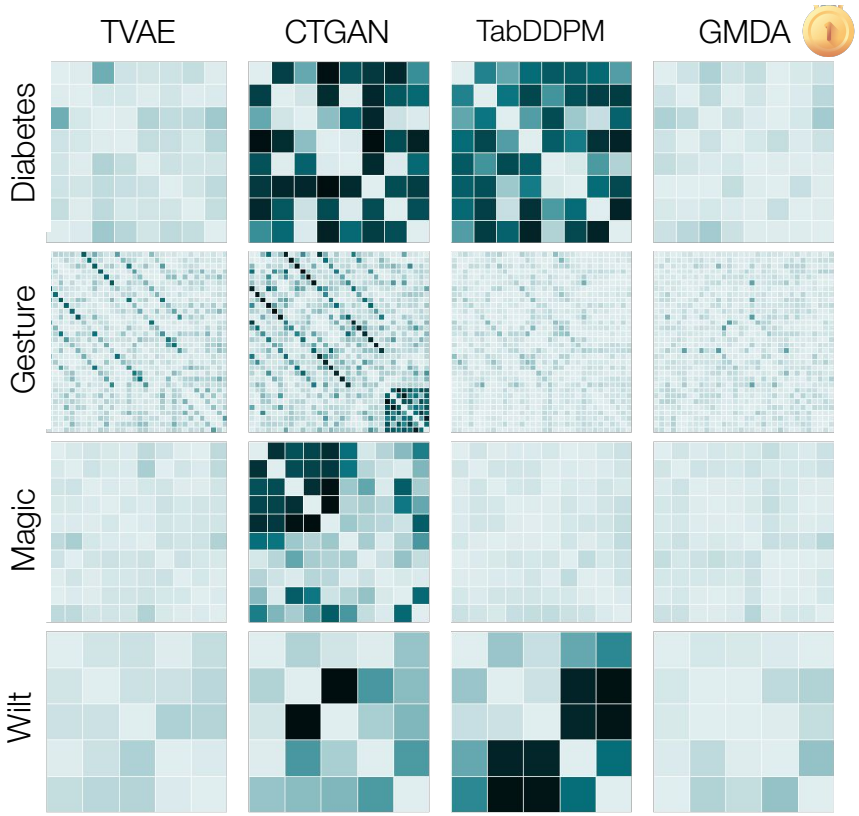
# Results on toy datasets

TabDDPM > GMDA > TVAE > CTGAN



# Results on real datasets

Correlation errors *(the lighter, the better):*




Fidelity-diversity trade-off

| Model | Avg. |
|---|---|
| TVAE | 80.43% |
| CTGAN | 13.44% |
| TabDDPM | **89.51%** |
| GMDA | 88.09% |

MLE

| Model | Avg. |
|---|---|
| Baseline | - |
| TVAE | 67.45% |
| CTGAN | 41.02% |
| TabDDPM | **75.32%** |
| GMDA | 72.42% |

- GMDA ranks 1st on correlation errors (medium-size)
- GMDA ranks second-best after TabDDPM on other indicators (medium-size)

# Scaling to high dimensions

Correlation errors *(the lighter, the better)*:


Fidelity-diversity trade-off

| Model | Avg. |
|---|---|
| TVAE | 80.43% |
| CTGAN | 13.44% |
| TabDDPM | **89.51%** |
| GMDA | 88.09% |

MLE

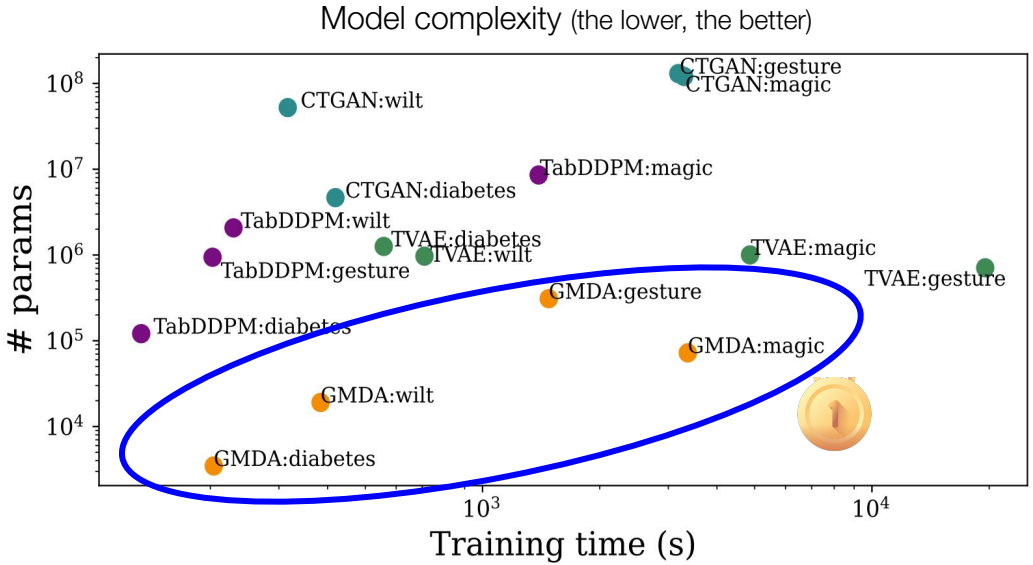
| Model | Avg. |
|---|---|
| Baseline | - |
| TVAE | 67.45% |
| CTGAN | 41.02% |
| TabDDPM | **75.32%** |
| GMDA | 72.42% |

MLE and fidelity-diversity trade-off (F1)

| Model | GTEx | | TCGA | |
|---|---|---|---|---|
| | MLE | F1 (PR) | MLE | F1 (PR) |
| MLP Class. | 99.32±0.04 | - | 93.59±0.6 | - |
| VAE | **98.98±0.05** | 74.28±0.1 | 88.36±0.97 | 82.36±0.06 |
| WGAN-GP | 98.76±0.09 | **94.66±0.1** | **92.04±0.46** | **93.17±0.17** |
| GMDA | 98.4±0.6 | 79.86±0.25 | 89.68±0.4 | 83.27±0.34 |

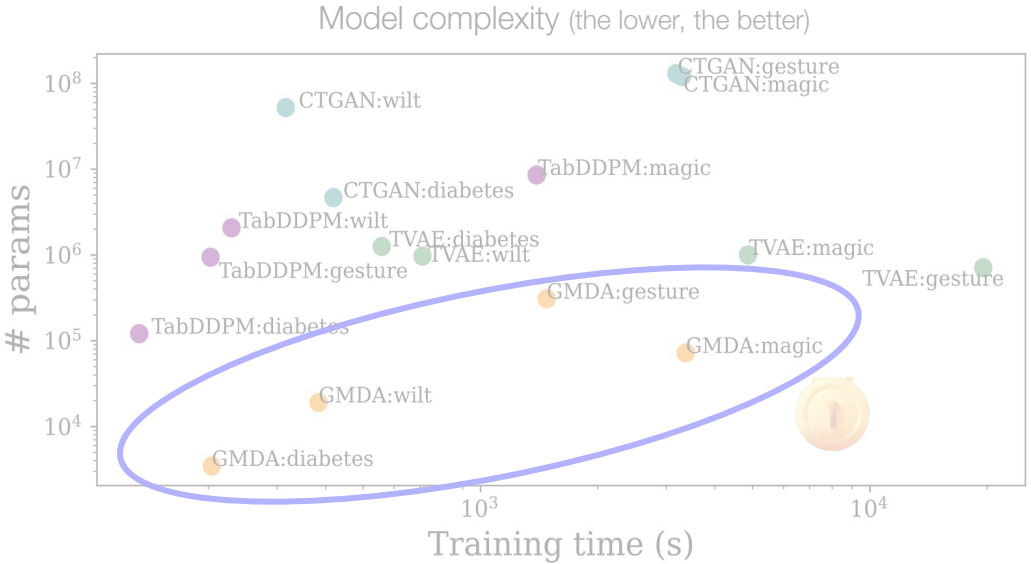
- GMDA ranks 1st on correlation errors (medium-size)
- GMDA ranks second-best after TabDDPM on other indicators (medium-size)
- GMDA ranks second-best on transcriptomic data

# Frugality and robustness

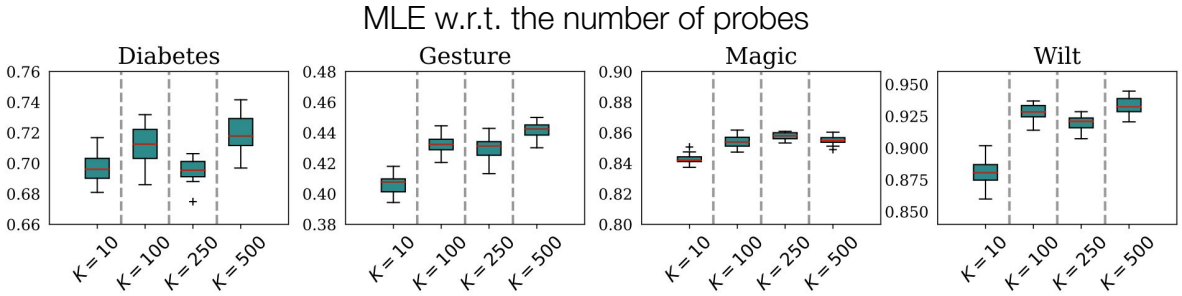



GMDA is at least one order of magnitude smaller

# Frugality and robustness




- GMDA is at least one order of magnitude smaller
- GMDA is not significantly sensitive to its hyper-parameters



# Partial conclusions

## Take-aways

- ❑ **Performance:** GMDA is competitive on different datasets size and complexity
- ❑ **Dimensionality:** GMDA scales up to 1,000 dimensions
- ❑ **Frugality:** significant complexity gain by at least one order of magnitude
- ❑ **Robustness:** to few hyper-parameters, to small training sets

**GMDA:**
A. Lacan, B. Hanczar and M. Sebag.
"Frugal Generative Modeling for Tabular Data". In: ECML-PKDD, September 2024.

# Conclusion & perspectives

# Contributions

**AttGAN:**
A. Lacan, M. Sebag and B. Hanczar. "GAN-based data augmentation for transcriptomics : survey and comparative assessment". In: ISMB, June 2023.

→ Survey on DA feasibility, inclusion of self-attention and domain knowledge

**GANs for microarray data:**
A. Alsamadi, A. Lacan, B. Hanczar and M. Sebag. "Identifying GANs Blind Spots in Transcriptomic Data Generation". In: JDSE, September 2024.

→ Evaluation methodology for GANs limitations

**Diffusion for transcriptomics (preprint):**
A. Lacan, R. André, M. Sebag and B. Hanczar. "In Silico Generation of Gene Expression profiles using Diffusion Models". In: bioRxiv, 2024.

→ Computational requirements of diffusion models

**GMDA:**
A. Lacan, B. Hanczar and M. Sebag.
"Frugal Generative Modeling for Tabular Data". In: ECML-PKDD, September 2024.

→ Alternative frugal generative model

# Further work

**Representation**

➔ Operating in **latent space**

➔ Bypass decoder with **Optimal Transport** mapping

# Further work

**Representation**

- ➔ Operating in **latent space**
- ➔ Bypass decoder with **Optimal Transport** mapping

**Efficiency**

- ➔ **Algorithmic**: better suited **attention and conditioning** strategy
- ➔ **Theoretical** analysis: formal analysis of GMDA

# Further work

**Representation**

- ➔ Operating in **latent space**
- ➔ Bypass decoder with **Optimal Transport** mapping

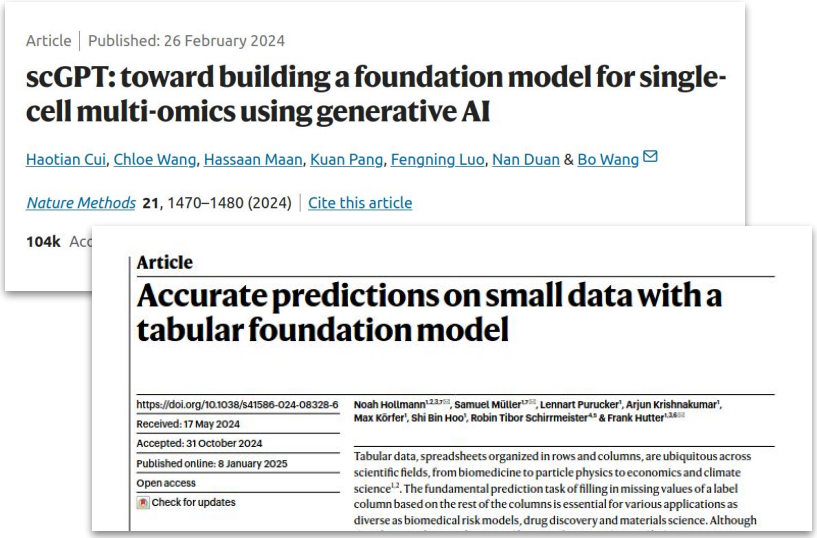
**Efficiency**

- ➔ **Algorithmic**: better suited **attention and conditioning** strategy
- ➔ **Theoretical** analysis: formal analysis of GMDA

**Interpretability for transcriptomics**

- ➔ Interpreting attention maps and GMDA's probes to **identify biomarkers**
- ➔ **Interpolation strategy:** valuable biological pathways

# Much ado about Large Language Models



Article | Published: 26 February 2024

**scGPT: toward building a foundation model for single-cell multi-omics using generative AI**

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan & Bo Wang

*Nature Methods* **21**, 1470–1480 (2024) | Cite this article

**104k** Acc

Article

**Accurate predictions on small data with a tabular foundation model**

https://doi.org/10.1038/s41586-024-08328-6

Received: 17 May 2024

Accepted: 31 October 2024

Published online: 8 January 2025

Open access

Check for updates

Noah Hollmann[1,2,3,4], Samuel Müller[3,4], Lennart Purucker[1], Arjun Krishnakumar[1], Max Körfer[1], Shi Bin Hoo[1], Robin Tibor Schirrmeister[4,5] & Frank Hutter[1,3,6]

Tabular data, spreadsheets organized in rows and columns, are ubiquitous across scientific fields, from biomedicine to particle physics to economics and climate science[1,2]. The fundamental prediction task of filling in missing values of a label column based on the rest of the columns is essential for various applications as diverse as biomedical risk models, drug discovery and materials science. Although

**Advantages:**

Foundation models (FMs) learn **high-level transferable representation** (e.g., downstream transfer learning, few-shot learning, domain adaptation)

**Intuition:**

Hybrid integration of tabular FMs and omic-specific fine-tuning could help **generation** and **predictions**

**Challenges:**

- Tabular data lacks consistent patterns
- Develop large-scale datasets
- Scalability with dimensions (e.g., converting to sentences)


Thank you for your attention!

# Appendix

# Benchmark Datasets

**The Cancer Genome Atlas (TCGA):**

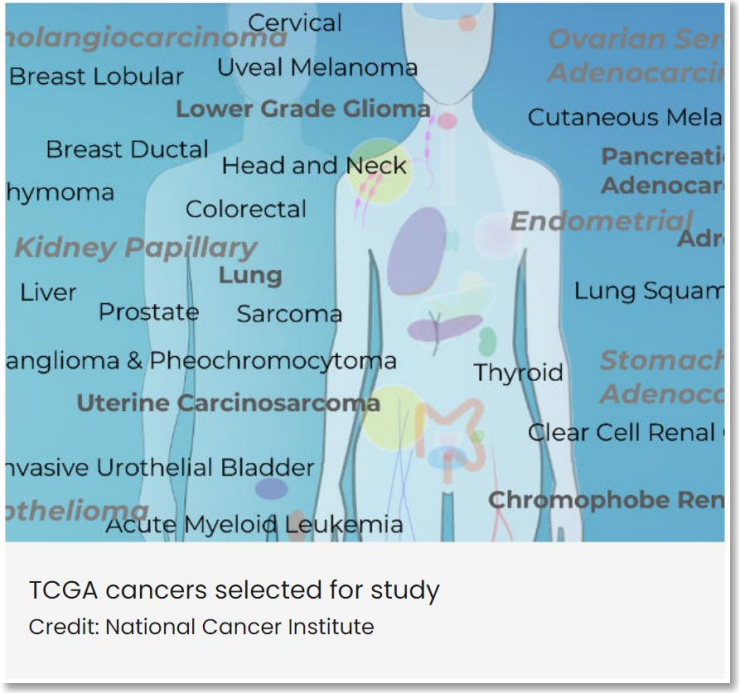
- 9,749 cancerous/non cancerous bulk RNA-seq samples
- 20,531 genes
- 24 tissue types

*Clinical covariates:* age, gender, cancer (y/n), tissue type

**The Genotype-Tissue Expression (GTEx):**

- 17,244 non cancerous bulk RNA-seq samples
- 18,691 genes
- 26 tissue types

*Clinical covariates:* age, gender, tissue type



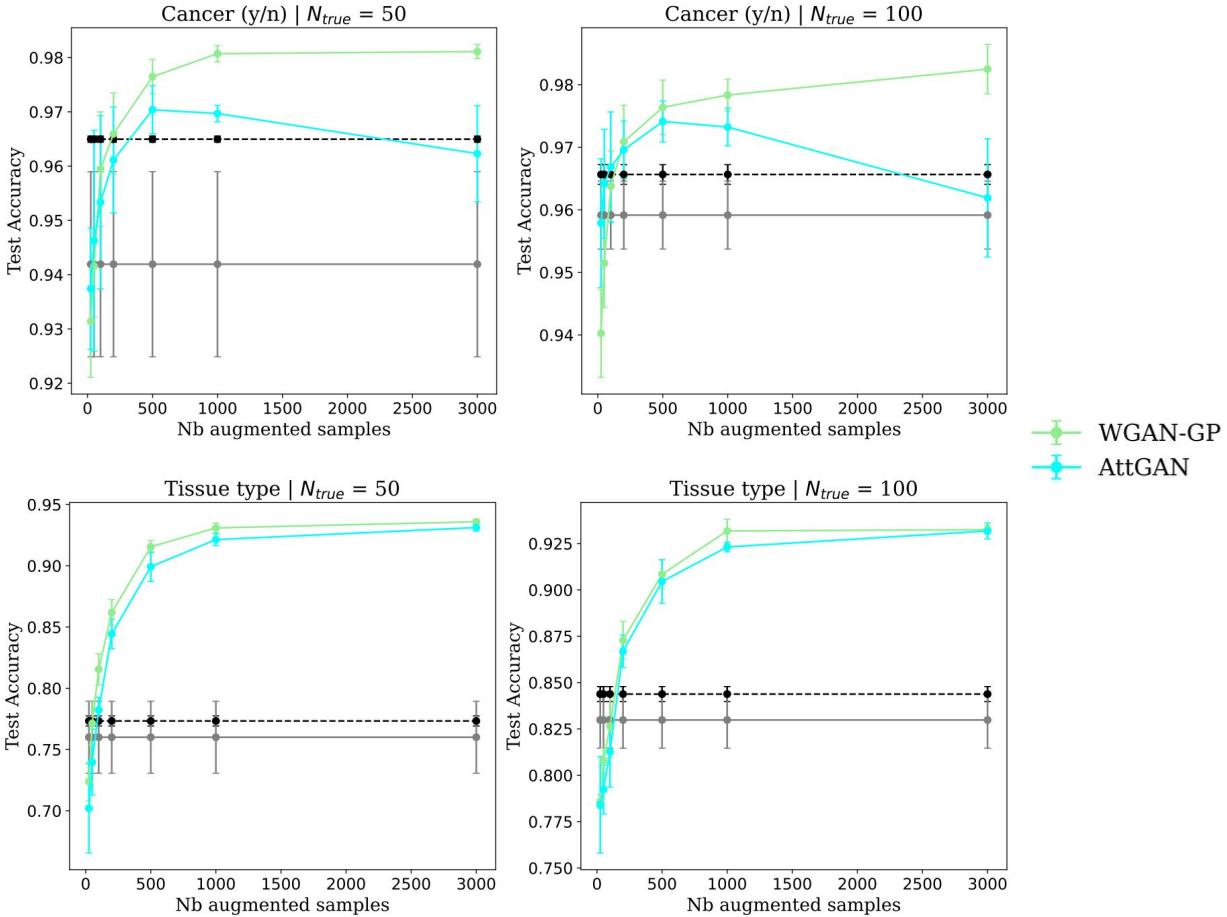

TCGA cancers selected for study
Credit: National Cancer Institute

Cancer Genome Atlas Research Network, Weinstein et al. *The Cancer Genome Atlas Pan-Cancer analysis project.* (Nature Genetics 2013)
Lonsdale et al. *The genotype-tissue expressio (GTEx) project.* (Nature Genetics 2013)

# Data Augmentation on limited real samples

Test accuracy of a MLP trained on true samples with a **varying** number of augmented samples

Performance reach a plateau after adding >1,000 augmented samples



# Data Augmentation Results per Tissue

| Tissue | $N_{true} = 50$ | $N_{true} = 50$ $N_{fake} = 1000$ | $N_{true} = 50$ $N_{fake} = 3000$ |
|---|---|---|---|
| adrenal | 0.81 ± 0.1245 | **0.9 ± 0.0** | 0.84 ± 0.0652 |
| bladder | 0.4838 ± 0.2805 | 0.9324 ± 0.0214 | **0.9568 ± 0.0113** |
| blood | 0.9172 ± 0.0866 | 0.9517 ± 0.0523 | **0.9724 ± 0.0289** |
| brain | 0.9133 ± 0.1097 | **0.98 ± 0.0** | 0.9787 ± 0.003 |
| breast | 0.8 ± 0.0726 | **0.9913 ± 0.0043** | 0.9905 ± 0.0036 |
| cervical | 0.4323 ± 0.2999 | **0.8677 ± 0.021** | 0.8516 ± 0.0177 |
| colon | 0.5846 ± 0.2851 | 0.92 ± 0.1204 | **0.9815 ± 0.0069** |
| esophagus | 0.2103 ± 0.2289 | **0.4103 ± 0.1612** | 0.1795 ± 0.1612 |
| eye | **0.8824 ± 0.2038** | 0.6 ± 0.5477 | 0.5882 ± 0.5375 |
| head | 0.7453 ± 0.2219 | 0.9302 ± 0.0338 | **0.9453 ± 0.0301** |
| kidney | 0.7586 ± 0.201 | 0.9841 ± 0.0024 | **0.9877 ± 0.0048** |
| liver | 0.8378 ± 0.0469 | 0.94 ± 0.0186 | **0.9556 ± 0.0** |
| lung | 0.4141 ± 0.2504 | **0.9507 ± 0.0079** | 0.9498 ± 0.0158 |
| ovary | 0.6281 ± 0.2778 | **1.0 ± 0.0** | 0.993 ± 0.0096 |
| pancreas | 0.551 ± 0.2126 | 0.9306 ± 0.0341 | **0.9714 ± 0.0112** |
| prostate | 0.8695 ± 0.1879 | **1.0 ± 0.0** | 0.9966 ± 0.0076 |
| rectum | **0.2273 ± 0.085** | 0.1364 ± 0.2802 | 0.0 ± 0.0 |
| skin | 0.628 ± 0.2728 | 0.9699 ± 0.0118 | **0.9785 ± 0.0108** |
| soft-tissues | 0.6226 ± 0.183 | **0.9581 ± 0.0216** | 0.9323 ± 0.0385 |
| stomach | 0.4088 ± 0.3586 | 0.6176 ± 0.1348 | **0.8353 ± 0.1197** |
| testes | 0.8529 ± 0.11 | **0.9235 ± 0.0161** | 0.9118 ± 0.036 |
| thymus | 0.725 ± 0.282 | 0.8167 ± 0.0475 | **0.8667 ± 0.0349** |
| thyroid | 0.923 ± 0.0598 | **0.9967 ± 0.0045** | 0.9885 ± 0.011 |
| uterus | 0.6739 ± 0.1969 | **0.9261 ± 0.0607** | 0.9043 ± 0.0917 |

**Table 5.** Test accuracy per tissue type after training a MLP (tissue type classification) on either 50 or 100 true samples ($N_{true}$) and 0, 1000 and 3000 augmented samples ($N_{fake}$) generated by our best AttGAN model. Best accuracy (given a number of true samples $N_{true}$) in bold.

# Additional results

## Reverse validation:

| Model | Test accuracy cancer (y/n) | Test accuracy tissue type |
|---|---|---|
| GAN | 0.8228 ±0.007 | 0.0856 ±0.0028 |
| WGAN-GP | 0.9839 ±0.0022 | 0.9361 ±0.0027 |
| RandAttGAN PPI + pretrain | 0.9858 ±0.0013 | 0.9333 ±0.0019 |
| AttGAN PPI + pretrain | 0.9826 ±0.0017 | 0.934 ±0.0038 |
| AttGAN PPI + CoExp + pretrain | 0.9811 ±0.0033 | 0.9276 ±0.0034 |
| RandAttGAN PPI + CoExp + pretrain | 0.9812 ±0.0039 | 0.9334 ±0.0025 |
| AttGAN PPI + CoExp + gamma fixed | 0.9843 ±0.0009 | 0.9361 ±0.0026 |
| Baseline on true data | **0.9916 ± 0.0015** | **0.9469 ± 0.0029** |

**Table 2.** Reverse validation results: test accuracy of binary and multiclass MLPs trained on 8000 generated samples only and tested on true TCGA test set. Although a MLP trained on generated data only does not outperform the state-of-the-art resuts (baseline results on last row), it reaches very close accuracy results with data generated by the WGAN-GP and the five AttGANs versions.

## Label knowledge preservation:

| Model | Test accuracy cancer (y/n) | Test accuracy tissue type |
|---|---|---|
| GAN | 0.9296 | 0.092 |
| WGAN-GP | 0.9823 | 0.9621 |
| RandAttGAN PPI + pretrain | 0.9867 | 0.969 |
| AttGAN PPI + pretrain | 0.9847 | 0.969 |
| AttGAN PPI + CoExp + pretrain | 0.9852 | 0.9695 |
| RandAttGAN PPI + CoExp + pretrain | 0.9838 | 0.9646 |
| AttGAN PPI + CoExp + gamma fixed | 0.9877 | 0.9621 |

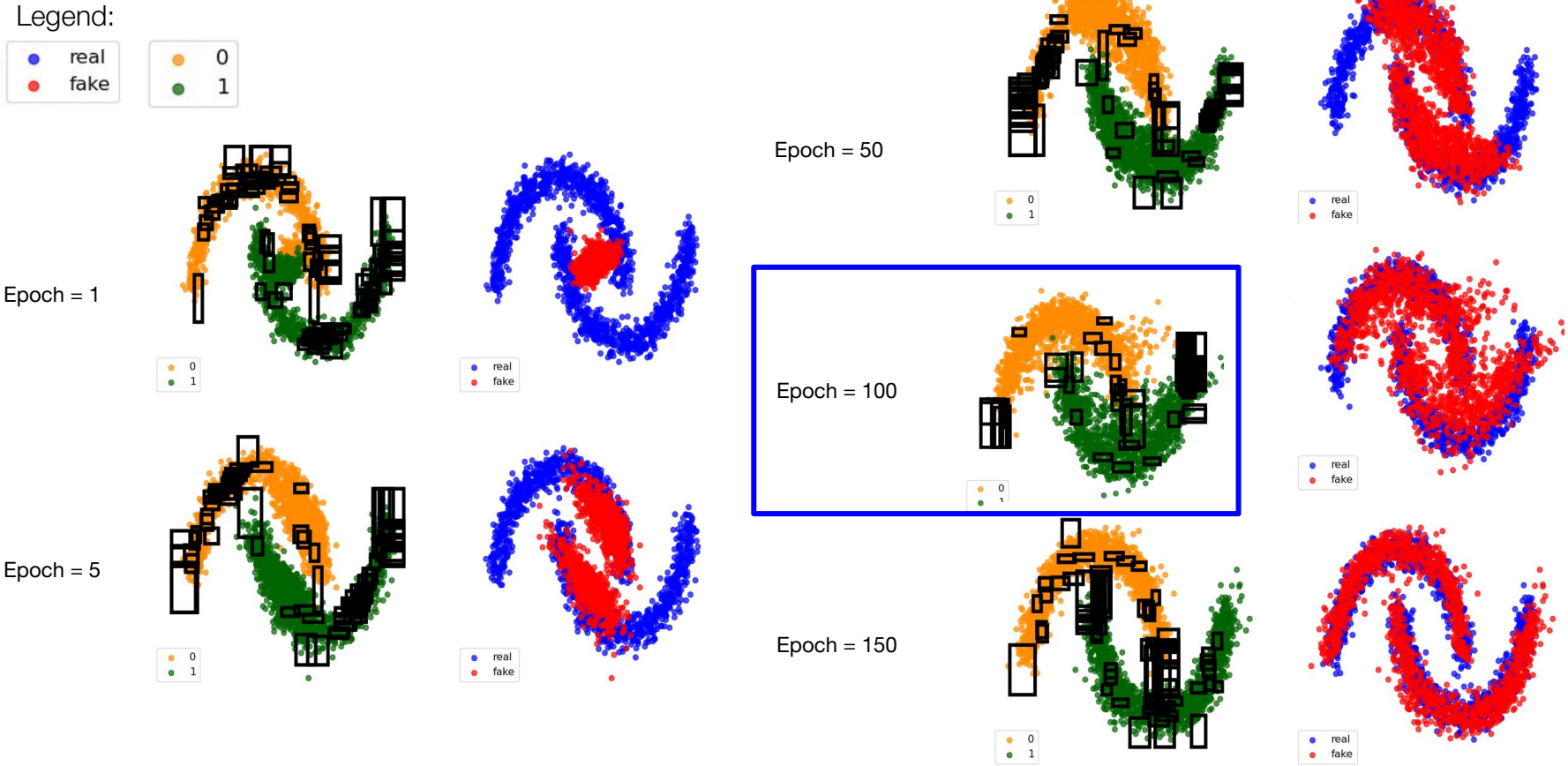
**Table 3.** Test accuracy of binary and multiclass MLPs pretrained on true TCGA data and tested on generated data. We observe that the pretrained MLPs are able to correctly label the data generated by the WGAN-GP and the five AttGANs versions (with a higher test accuracy of 96% than the 94% accuracy on true data for tissue classification). However, the accuracy drops for both binary and tissue classification with data generated by the GAN. It seems the generated data does not violate cancer/tissue information except for the GAN.

# Knowledge Graphs

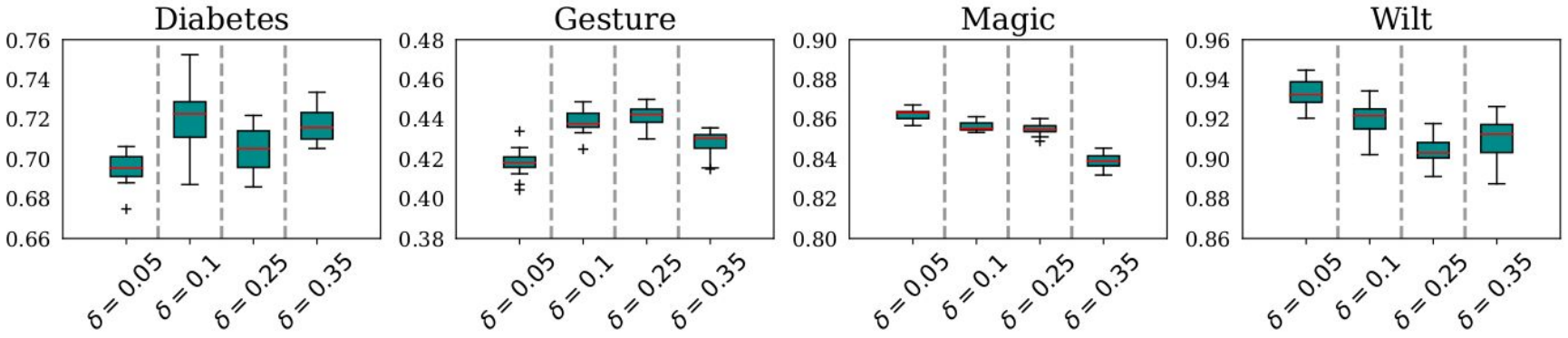
| Threshold $\tau$ | $I$ (PPI) | $N$ (PPI) | $I$ (Co-exp.) | $N$ (Co-exp.) |
|---|---|---|---|---|
| 0.7 | 401,370 | 14,044 | 132,588 | 5,872 |
| 0.8 | 287,970 | 12,057 | 33,006 | 2,648 |
| 0.9 | 199,621 | 10,187 | 8,014 | **758** |
| 0.95 | 92,549 | 8,369 | 3,302 | 335 |
| 0.98 | 54,432 | 6,591 | 1,170 | 148 |
| 0.99 | 39,472 | 5,502 | 526 | 99 |
| 0.995 | 29,928 | **4,670** | 218 | 55 |

Table 5.1 – Number $I$ of interactions and number $N$ of genes retained depending on the threshold level $\tau$. Given thresholds correspond to the percentage of lowest interactions removed from the knowledge graph (left : PPI; right : Co-exp). The number of genes retained in the experiments is highlighted in bold.

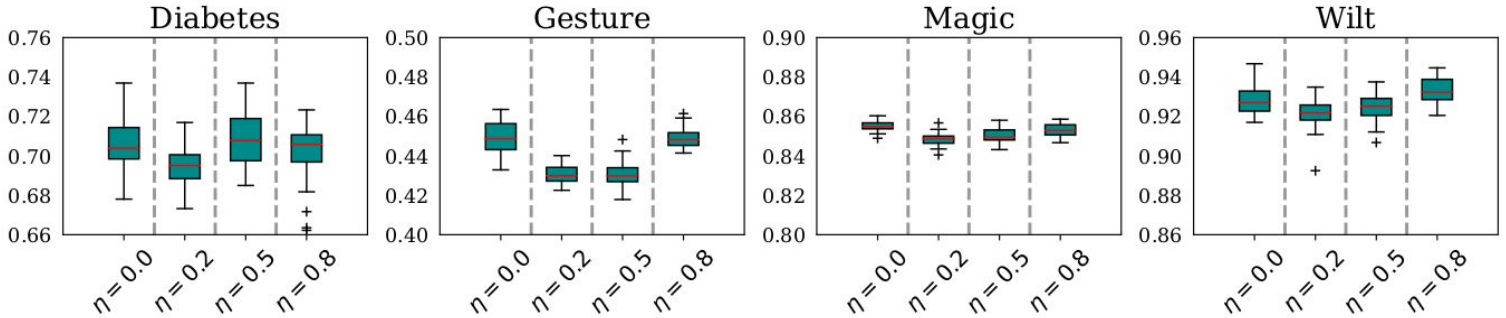
# GMDA: evolution on 2D moons



# GMDA: sensitivity study




MLE w.r.t. the density rate




MLE w.r.t. the persistence rate