Universitat Oberta de Catalunya

uoc.edu

Generación de datos multiómicos sintéticos con Generative Adversarial Networks

Tutor: Esteban Vegas
TFM de Federico Lara Salas
14/1/2024

Índice

Introducción

Importancia datos multi-ómicos

- Comprensión holística de los sistemas biológicos
- Comprensión de enfermedades
- Desarrollo de terapias más efectivas

Desafíos

- Volumen suficiente de datos
- Calidad de los datos
- Integración de los datos

Objetivos Iniciales y Desviaciones

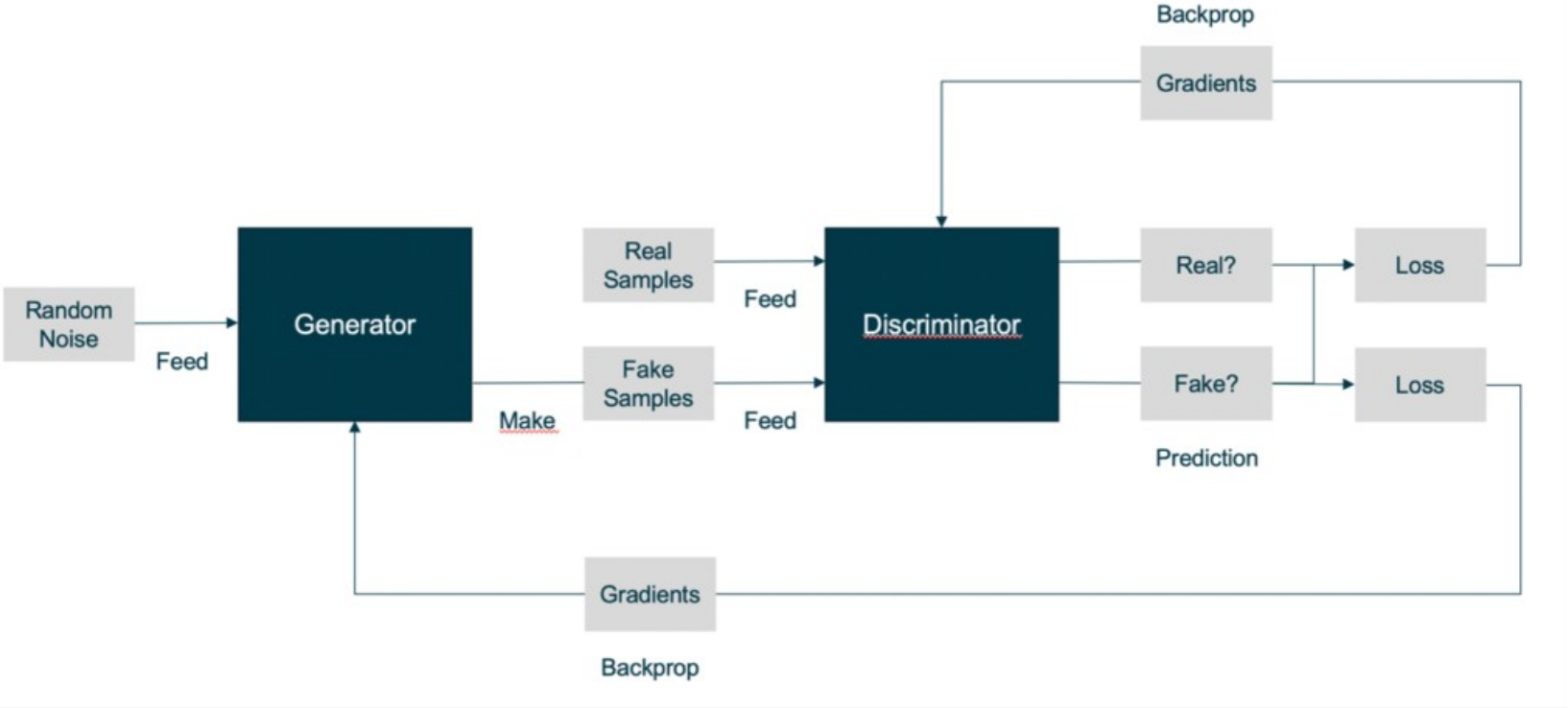
Objetivos Iniciales

- Comprender la integración de datos multi-ómicos y la generación de datos con GANs
- Implementar y validar modelo GAN que genere datos multi-ómicos y comparar con VAE
- Mejorar modelo anterior usando Transformer

Desviaciones

- Cambiamos a exploración modelos de conversión entre datos ómicos
- No podemos explorar modelos con datos reales en profundidad

Que es una GAN?



WGAN y WGAN-GP

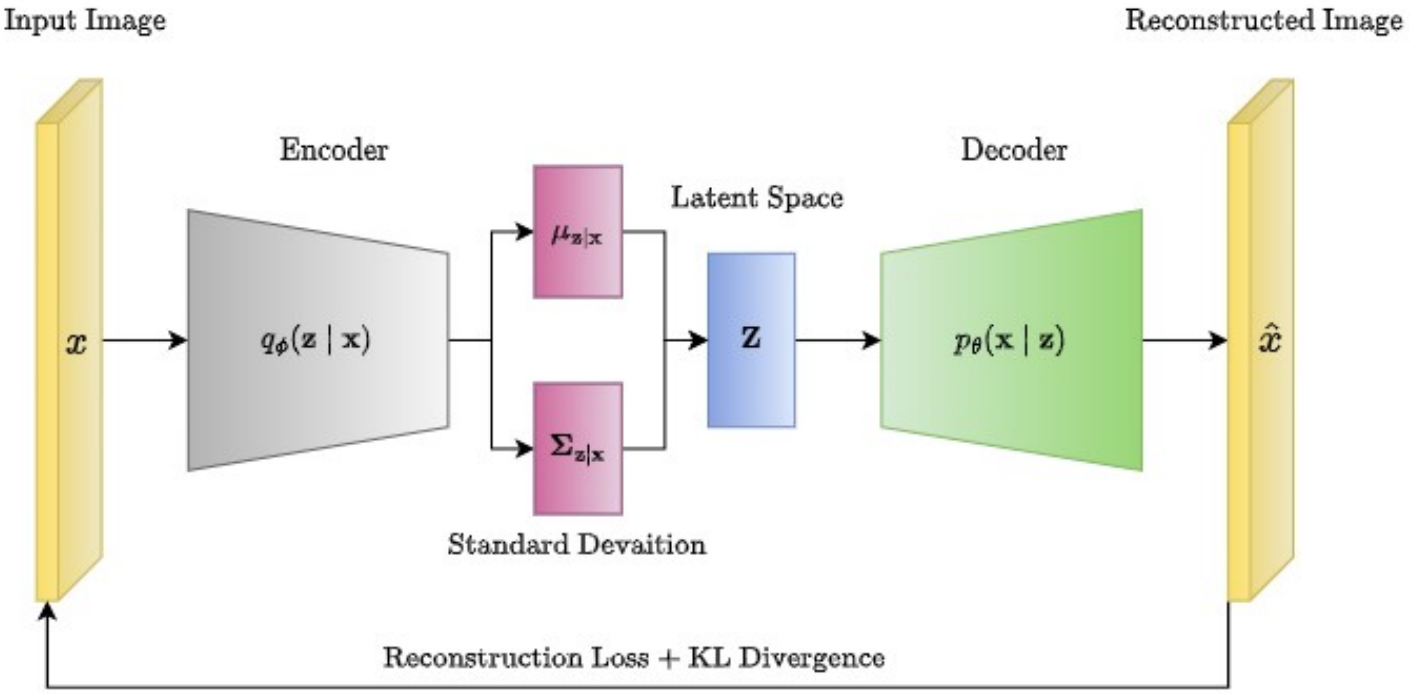
WGAN

- Función de pérdida de Wasserstein
- Crítico 1-Lipschitz
- Recorte de pesos

WGAN-GP

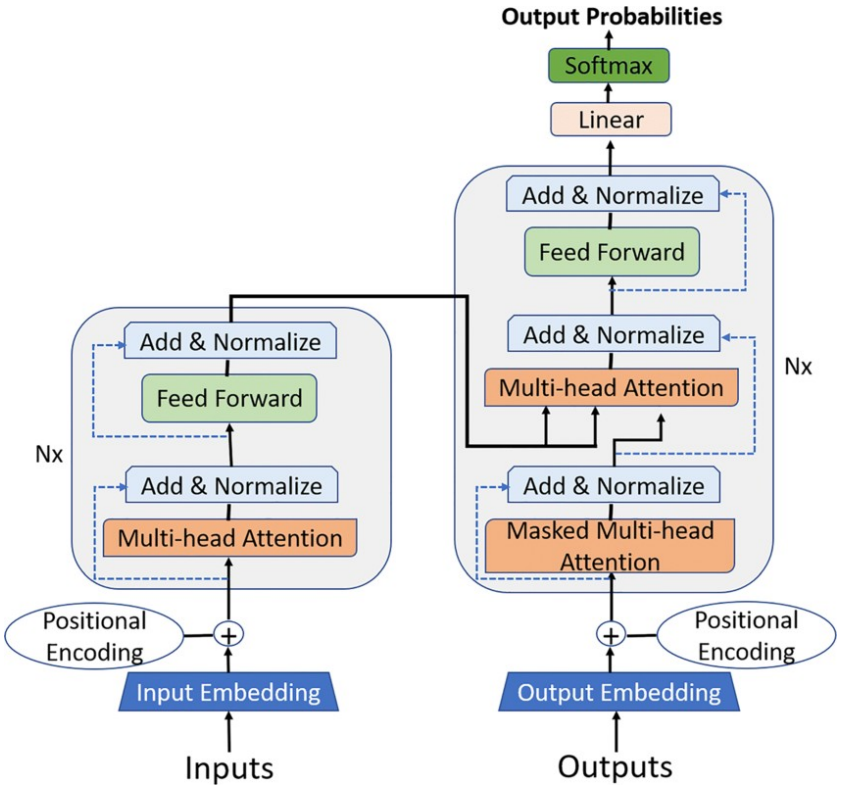
- Penalización de gradiente
- Actualización más a menudo del Crítico

Que es un VAE?

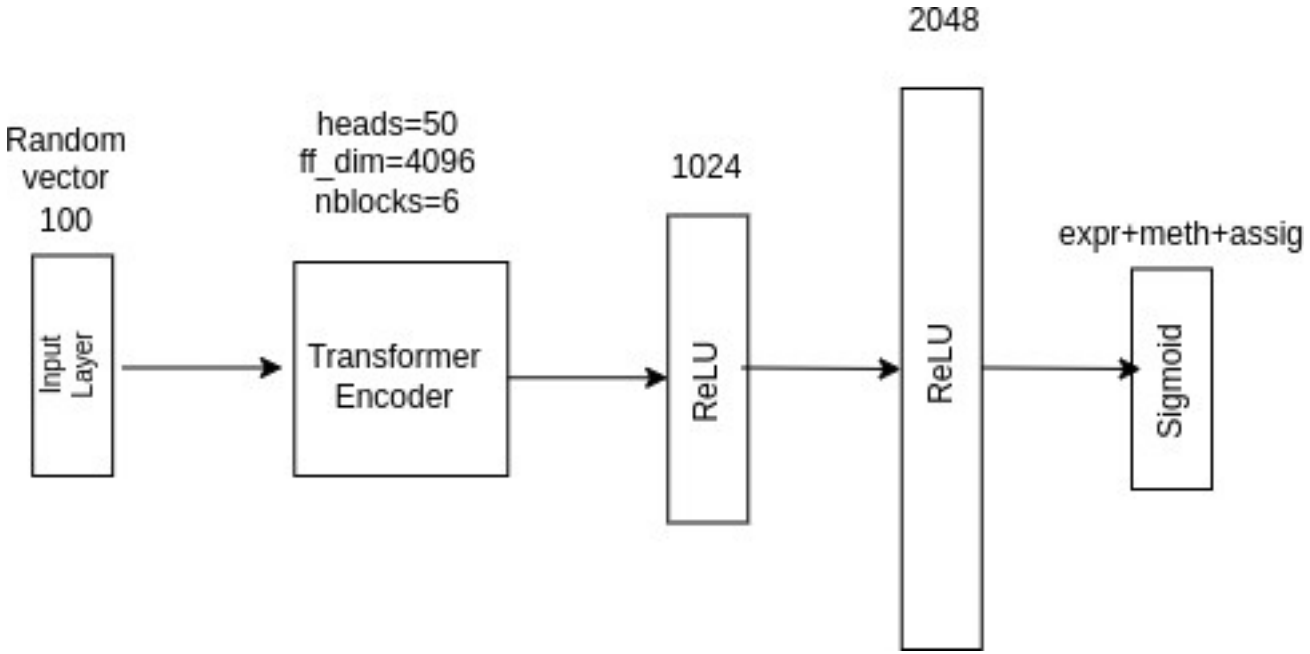

Input Image
Reconstructed Image
Encoder
Decoder
Latent Space
x
$q_\phi(\mathbf{z} \mid \mathbf{x})$
$\mu_{\mathbf{z}|\mathbf{x}}$
$\Sigma_{\mathbf{z}|\mathbf{x}}$
$\mathbf{z}$
$p_\theta(\mathbf{x} \mid \mathbf{z})$
$\hat{x}$
Standard Devaition
Reconstruction Loss + KL Divergence

Que es un Transformer?

- Query, Key, Value
- Cada cabeza enfoca en diferentes aspectos de la entrada
- No input embedding
- No positional encoding
- Encoder Only



GANFORMER


Random vector 100
Input Layer
heads=50
ff_dim=4096
nblocks=6
Transformer Encoder
1024
ReLU
2048
ReLU
expr+meth+assig
Sigmoid

Datos simulados

- Paquete de R InterSIM
- Expresión y metilación interelacionado basado en estudio cáncer de ovario de TCGA
- 2 grupos, 80% sanos y 20% cáncer
- Datasets de 500, 1k, 3k ,10k y más
- 131 variables expresión
- 367 variables metilación

Datos reales

- Obtenido de https://adex.genyo.es/
- Serie GSE117931 datos de expresión y metilación pacientes con esclerosis sistémica
- 2 grupos, 19 sanos y 18 con esclerosis
- 14760 variables de expresión tras preprocesamiento
- 416660 variables de metilación tras preprocesamiento
- También datos reales con datos en 90% SD

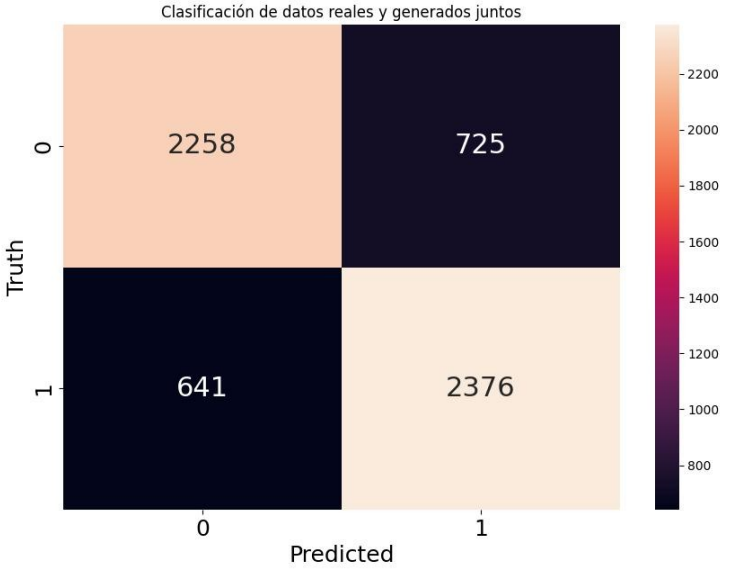
Exploración de modelos

- Número de capas ocultas
- Número de neuronas por capa
- Tasa de aprendizaje
- Optimizadores
- Función de perdida
- Tamaño de lote
- Número de épocas de entrenamiento
- Tipo de normalización
- Coeficiente de penalización de gradiente
- Número de veces de actualización de Discriminador
- Tamaño de espacio latente
- Coeficiente de perdida de reconstrucción
- Número de cabezas de atención
- Tamaño neuronas capa densa Transf.
- Número de bloques de Encoder

Métricas usadas para datos sintéticos generados

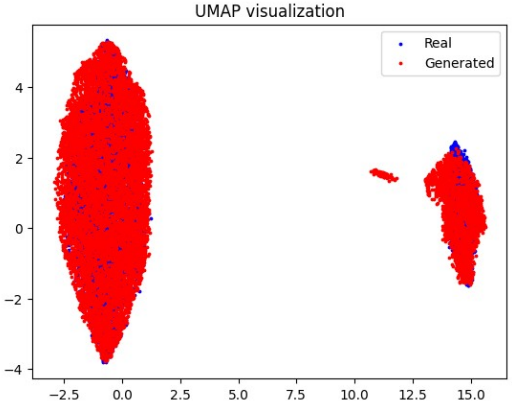
- Distancia de Wasserstein
- Test Kolmogorov-Smirnov
- Distancia euclidiana promedio
- Matrices de correlación de datos reales y generados
- Clasificación con SVC de datos reales y generados
- Clasificación con SVC de datos generados
- Dendogramas jerárquicos de datos reales y generados
- Clasificación con k-NN
- t-SNE
- UMAP
- PCA
- KDE de distintas variables al azar

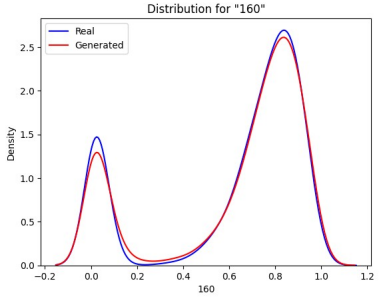
Resultados GAN datos simulados

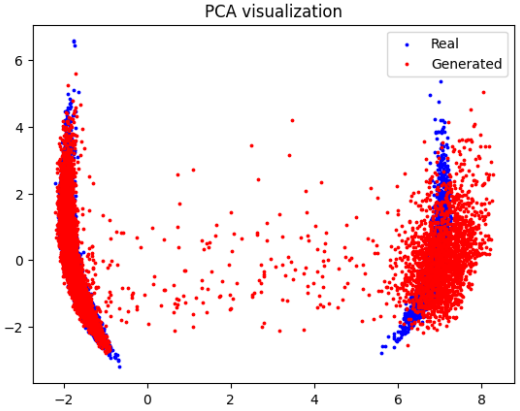


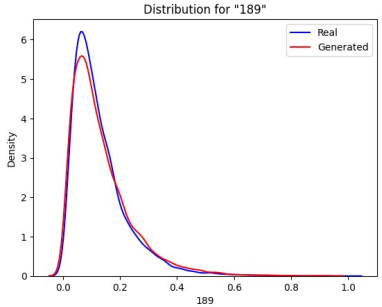
Generado=0, Real=1

78% precisión

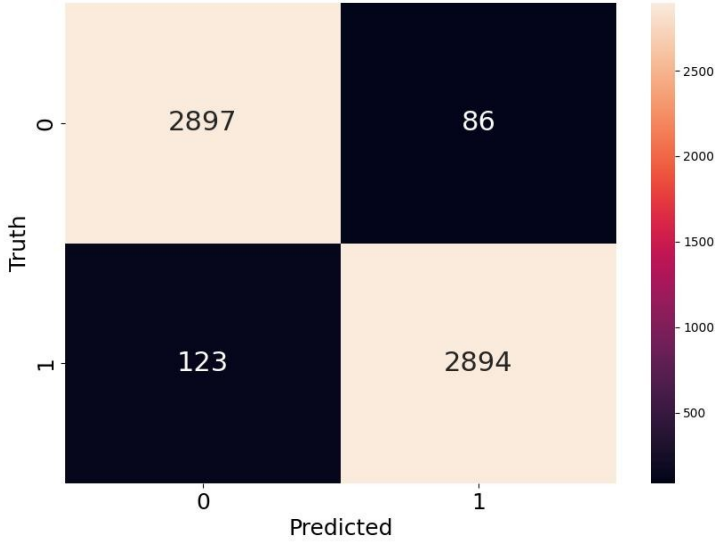






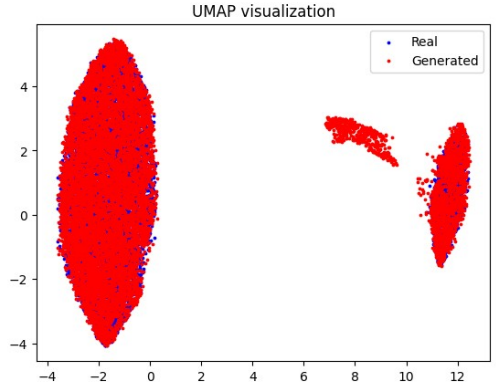


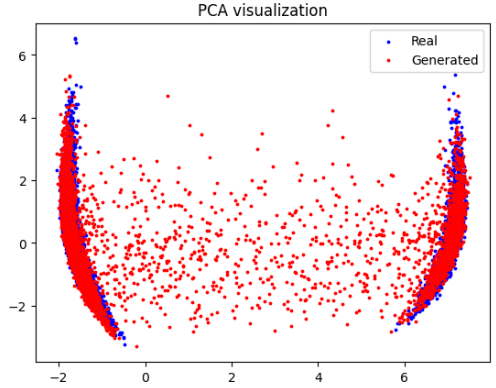
Resultados VAE datos simulados

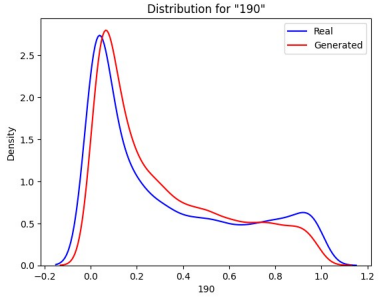


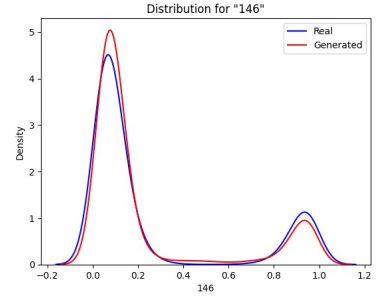
Generado=0, Real=1

96% precisión



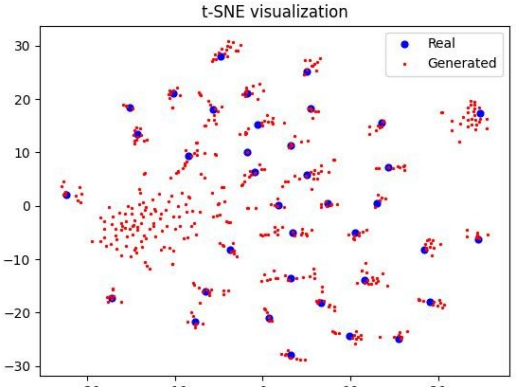


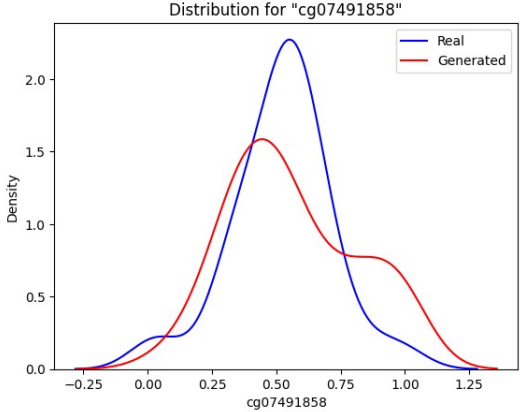




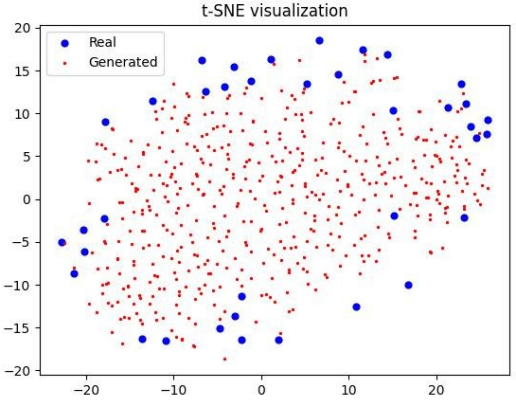
Datos reales

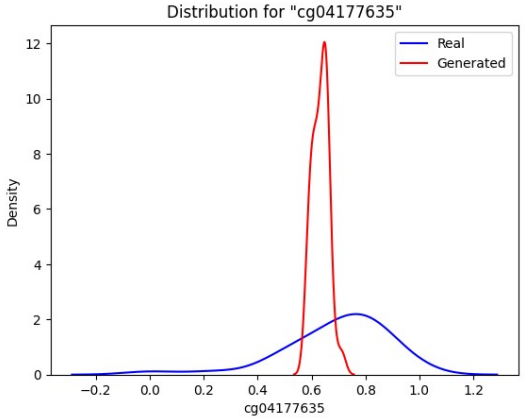
GAN



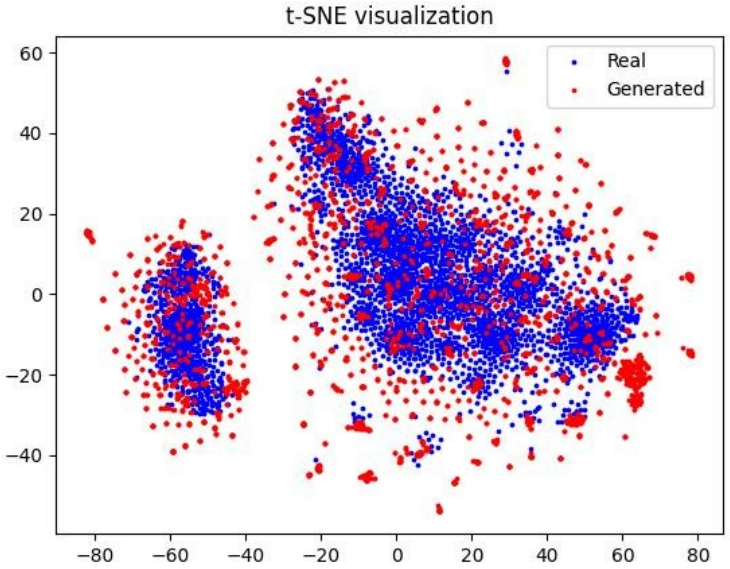


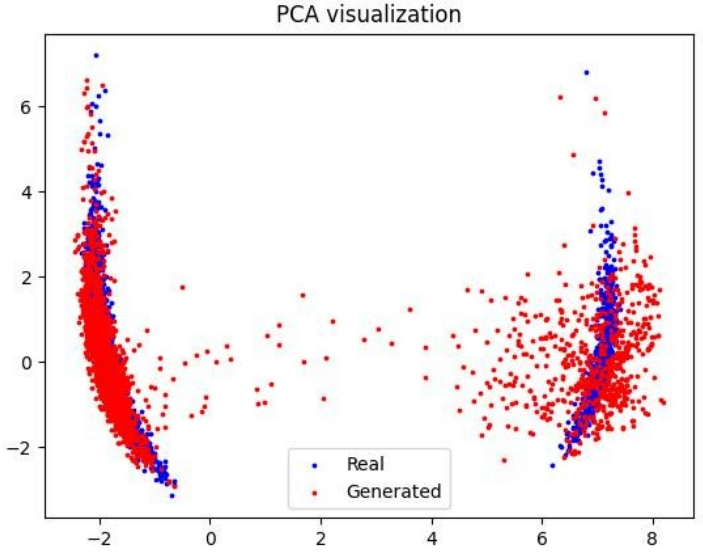
VAE





Resultados GANFORMER datos simulados





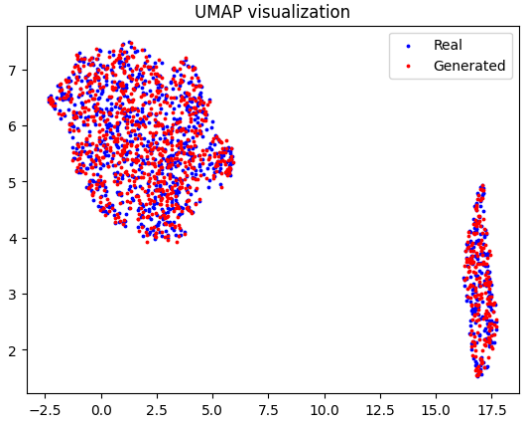
Métricas usadas para conversión de datos ómicos

- MSE
- RMSE
- PCC
- t-SNE
- UMAP
- PCA

Resultados Expr. génica → Metilación datos simulados

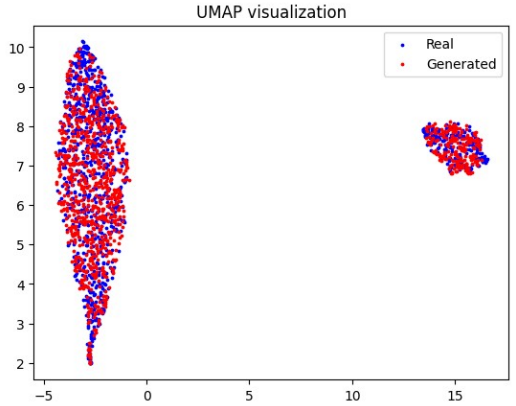
GAN

MSE: 0.0039
RMSE: 0.0627
PCC: 0.9833



TRANSFORMER

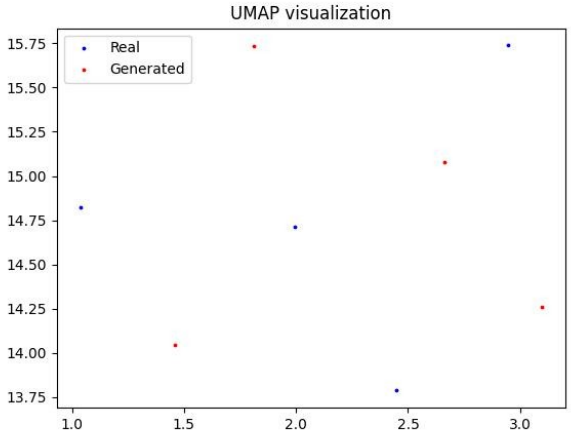
MSE: 0.0114
RMSE: 0.1071
PCC: 0.9505



Resultados Expr. génica → Metilación datos reales

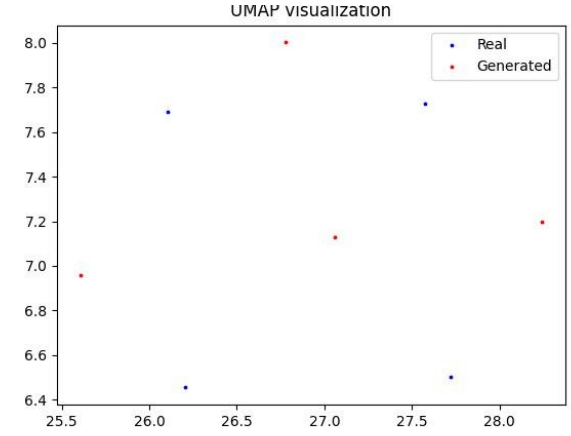
GAN

MSE: 0.0888
RMSE: 0.2980
PCC: 0.2114



TRANSFORMER

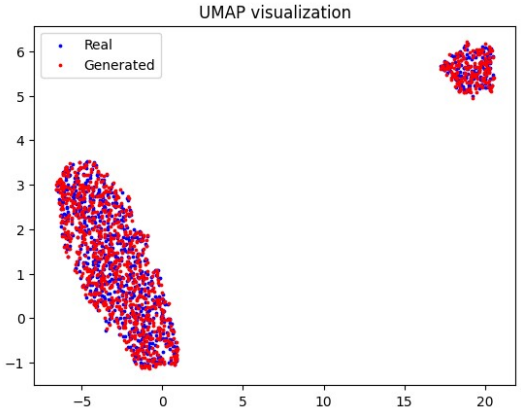
MSE: 0.0938
RMSE: 0.3063
PCC: 0.1178



Resultados Metilación → Expr. Génica datos simulados

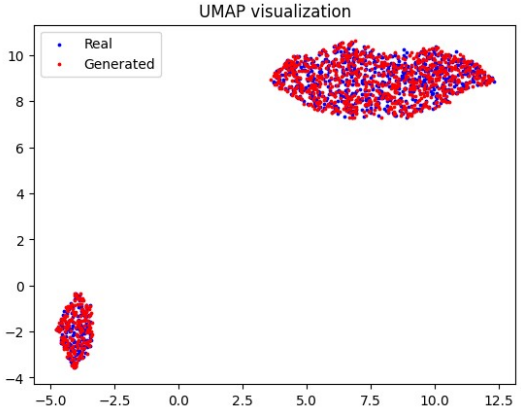
GAN

MSE : 0.020
RMSE: 0.1432
PCC: 0.7032



TRANSFORMER

MSE: 0.0126
RMSE: 0.1125
PCC: 0.8075



Conclusiones

Generación datos sintéticos

- GAN +calidad, VAE +diversidad
- GAN ≈ VAE datos simulados
- GAN > VAE datos reales con pocas muestras
- Rendimiento GANFORMER pobre

Conversión de datos ómicos

- GAN > Transformer conversión E → M
- Transformer > GAN conversión M → E
- Datos reales muestras insuficientes

Fortalezas y debilidades

Fortalezas

- Buen rendimiento con datos simulados
- No hay necesidad de grandes recursos computacionales

Debilidades

- Volumen pequeño de datos reales
- Limitación en tiempo y/o recursos computacionales
- No hay validación de significación biológica
- Comprensión superficial de la arquitectura Transformer
- Limitación en la generalización de los modelos

Futuras lineas de trabajo

- Exploración de modelos VAE con datos reales y espacio latente ampliado
- Comprensión más profunda de la arquitectura Transformer
- Mejora y exploración modelo GANFORMER
- Entrenamiento modelos con mayor volúmen de datos reales
- Exploración Cycle GAN
- Mejora en la exploración de modelos e hiperparámetros

Universitat Oberta
de Catalunya

UOC.universitat
@UOCuniversidad
UOCuniversitat