Interpretable Deep Learning for Single-Cell Data Analysis

Introduction to Interpretable AI

Maren Hackenberg, Sara Al-Rawi, Martin Treppner, Moritz Hess
Institute of Medical Biometry and Statistics, Medical Center – University of Freiburg

GCB 2022 – September 5th, 2022

Why Explainable AI?

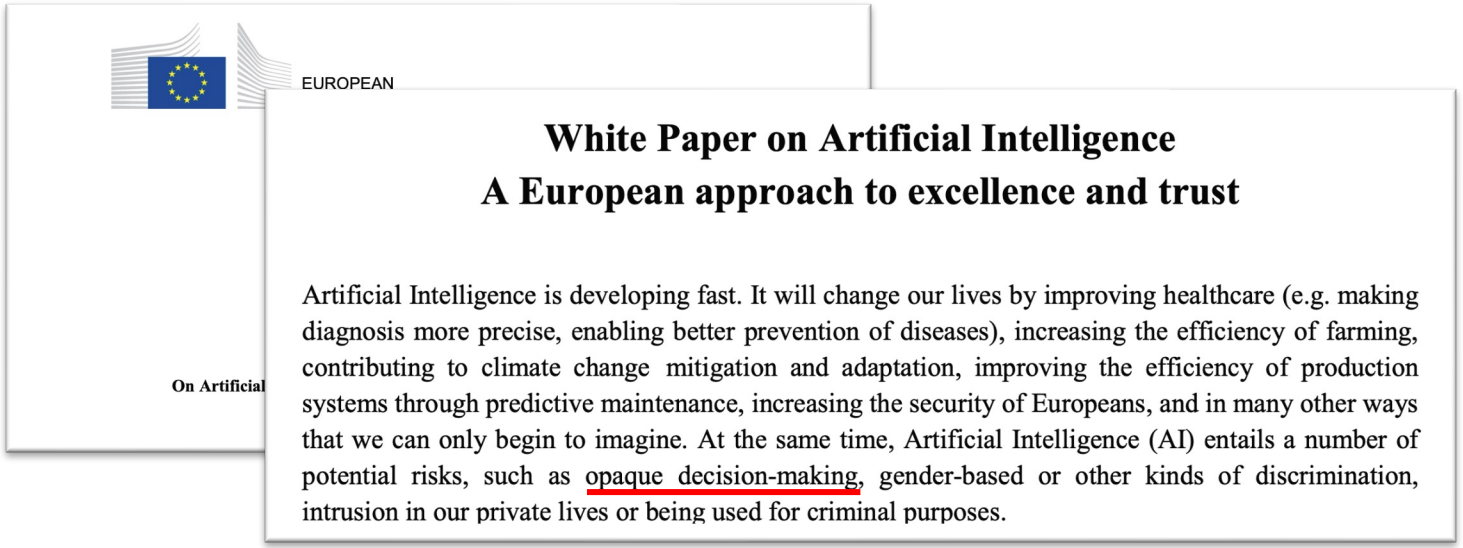
EUROPEAN
COMMISSION

Brussels, 19.2.2020
COM(2020) 65 final

WHITE PAPER

On Artificial Intelligence - A European approach to excellence and trust

Why Explainable AI?


EUROPEAN

On Artificial

White Paper on Artificial Intelligence
A European approach to excellence and trust

Artificial Intelligence is developing fast. It will change our lives by improving healthcare (e.g. making diagnosis more precise, enabling better prevention of diseases), increasing the efficiency of farming, contributing to climate change mitigation and adaptation, improving the efficiency of production systems through predictive maintenance, increasing the security of Europeans, and in many other ways that we can only begin to imagine. At the same time, Artificial Intelligence (AI) entails a number of potential risks, such as opaque decision-making, gender-based or other kinds of discrimination, intrusion in our private lives or being used for criminal purposes.

Explainable AI Taxanomy:

Linardatos, P. et al. (2020). Explainable AI: A review of machine learning interpretability methods. *Entropy*.

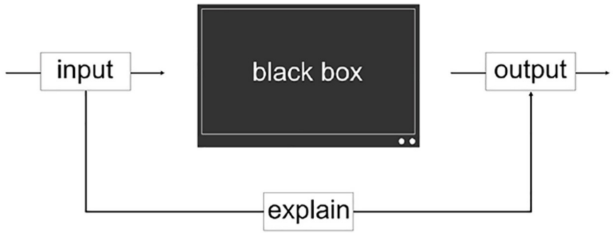
Post-hoc vs. Model-based interpretability

Treppner, M. et al. (2022) Interpretable generative deep learning: an illustration with single cell gene expression data. *Hum Gen*.

Post-hoc vs. Model-based interpretability

Post-hoc:

- Infer feature importance by approximating the non-interpretable model with simpler models

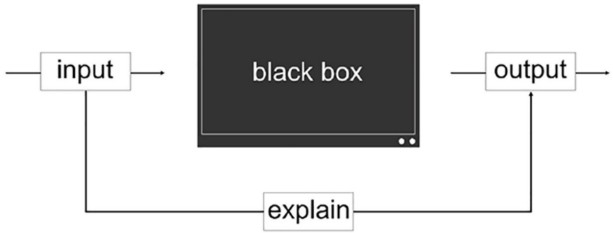


Treppner, M. et al. (2022) Interpretable generative deep learning: an illustration with single cell gene expression data. *Hum Gen*.

Post-hoc vs. Model-based interpretability

Post-hoc:

- Infer feature importance by approximating the non-interpretable model with simpler models



Model-based:

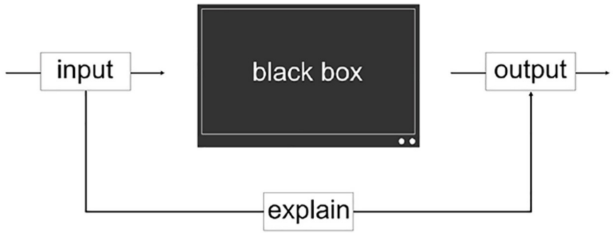
- Refers to models that incorporate mechanisms that allow direct interpretation of learned relationships

Treppner, M. et al. (2022) Interpretable generative deep learning: an illustration with single cell gene expression data. *Hum Gen*.

Post-hoc vs. Model-based interpretability

Post-hoc:

- Infer feature importance by approximating the non-interpretable model with simpler models



Model-based:

- Refers to models that incorporate mechanisms that allow direct interpretation of learned relationships

Treppner, M. et al. (2022) Interpretable generative deep learning: an illustration with single cell gene expression data. *Hum Gen*.

Post-hoc explainability methods

Post-hoc explainability methods

Local explainability:

- Local interpretability relates to the determination of features that explain the predictions made for a given observational unit, e.g., a patient

Post-hoc explainability methods

Local explainability:

- Local interpretability relates to the determination of features that explain the predictions made for a given observational unit, e.g., a patient

Global explainability:

- Global interpretability relates to the features (variables) that are overall important

Local explainability

Gradient-based or perturbation-based

Ancona, M. et al. (2017) Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv*.

Local explainability

Gradient-based or perturbation-based

- **Gradient-based methods** compute the attributions for all input features in a single forward and backward pass through the network

Ancona, M. et al. (2017) Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv*.

Local explainability

Gradient-based or perturbation-based

- **Gradient-based methods** compute the attributions for all input features in a single forward and backward pass through the network

- **Perturbation-based methods** directly compute the attribution of an input feature (or set of features) by removing, masking or altering them, and running a forward pass on the new input, measuring the difference with the original output.

Ancona, M. et al. (2017) Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv*.

Gradient-based methods – Integrated Gradients

Importance value for the ith feature:

$$\phi_i^{IG}(f, x, x') = \overbrace{(x_i - x'_i)}^{\text{Difference from baseline}} \times \underbrace{\int_{\alpha=0}^{1}}_{\text{From baseline to input...}} \overbrace{\frac{\delta f(x' + \alpha(x - x'))}{\delta x_i} d\alpha}^{\text{...accumulate local gradients}}$$

Sundararajan et al. (2017) Axiomatic attribution for deep networks. *PMLR*.

Importance value for the ith feature:

$$\phi_i^{IG}(f, x, x') = \overbrace{(x_i - x'_i)}^{\text{Difference from baseline}} \times \underbrace{\int_{\alpha=0}^{1}}_{\text{From baseline to input...}} \overbrace{\frac{\delta f(x' + \alpha(x - x'))}{\delta x_i} d\alpha}^{\text{...accumulate local gradients}}$$

where x is the current input, f is the model function and x' is some baseline input that is meant to represent "absence" of feature input. α is the interpolation constant.

Sundararajan et al. (2017) Axiomatic attribution for deep networks. *PMLR*.

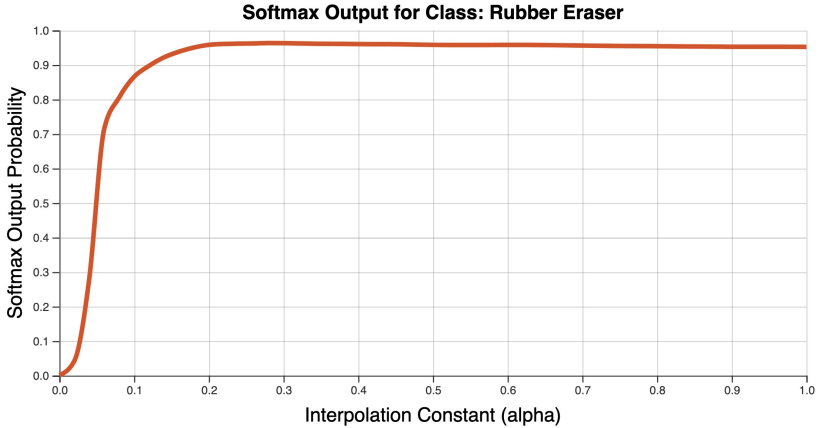
Problem with taking gradients

Saturation

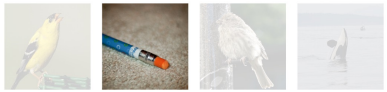
Gradients of input features may have small magnitudes around a sample even if the network depends heavily on those features.

This can happen if the network function flattens after those features reach a certain magnitude.

By integrating over a path, integrated gradients avoids problems with local gradients being saturated.



Click to select a different image:



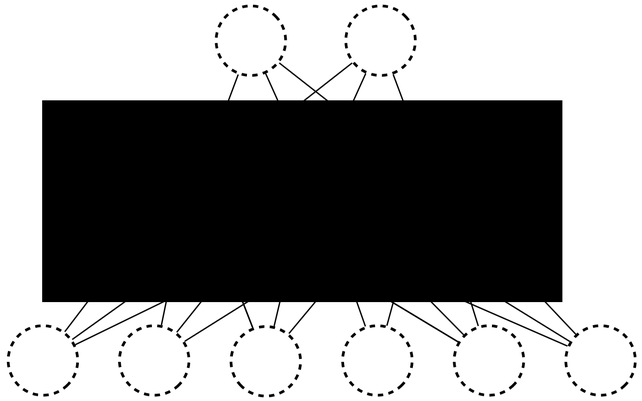
A plot of network outputs at $x' + \alpha(x - x')$. Notice that the network output saturates the correct class at small values of α. By the time $\alpha = 1$, the network output barely changes.

https://distill.pub/2020/attribution-baselines/

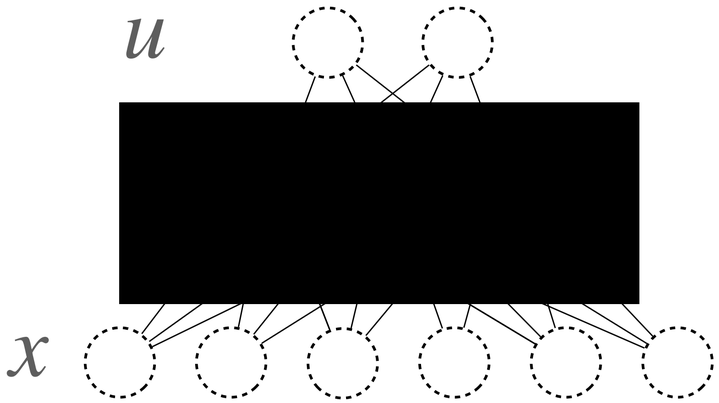
Global methods – Pattern extractor

Investigating the joint distribution of discretized synthetic observations

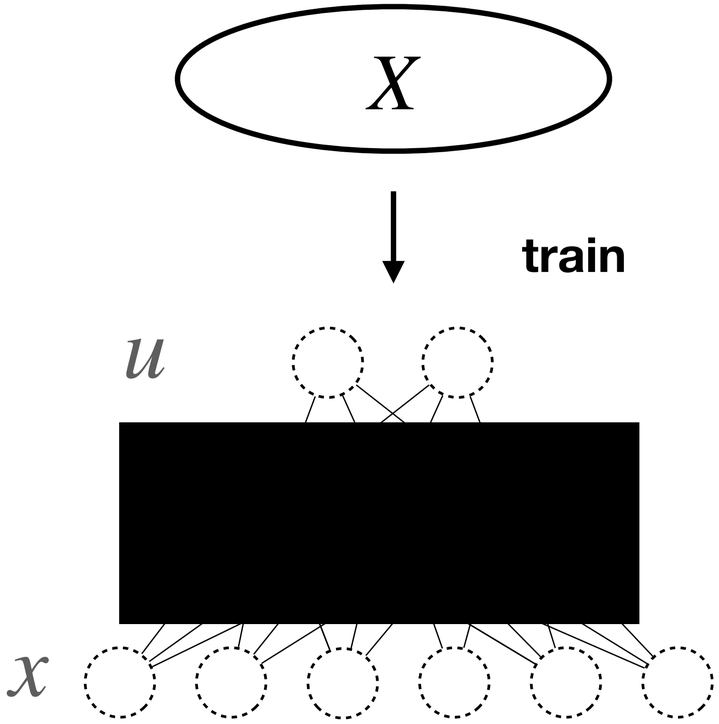
Investigating the joint distribution of discretized synthetic observations



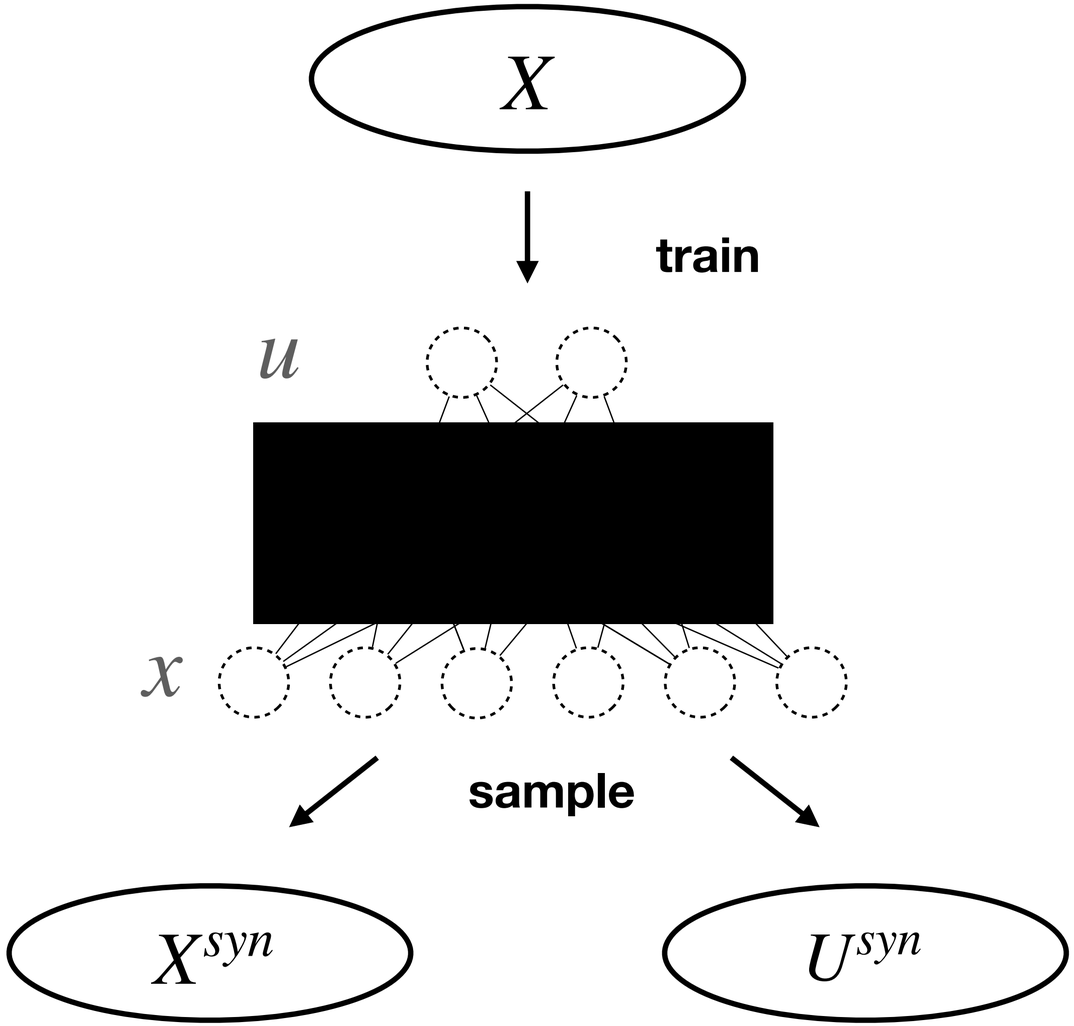
Investigating the joint distribution of discretized synthetic observations



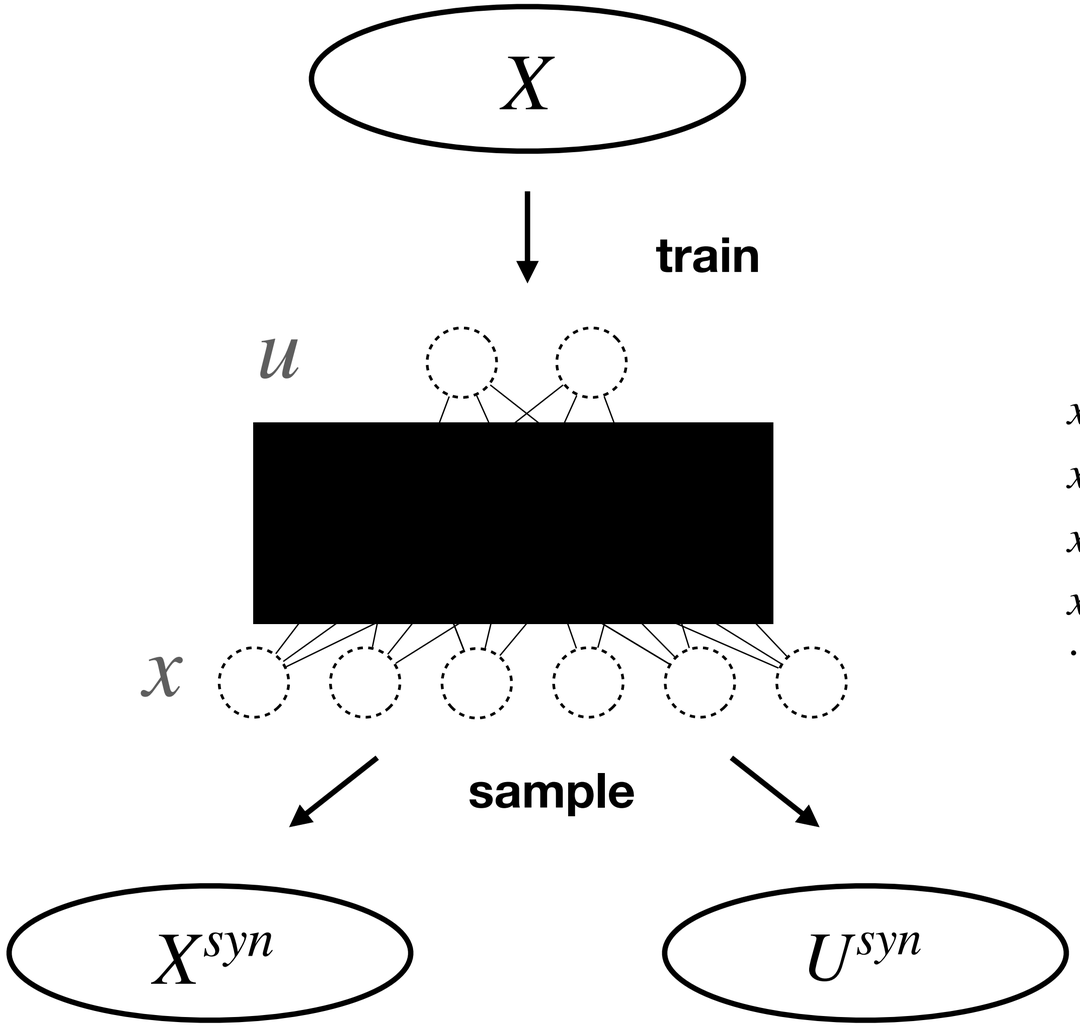
Investigating the joint distribution of discretized synthetic observations



Investigating the joint distribution of discretized synthetic observations

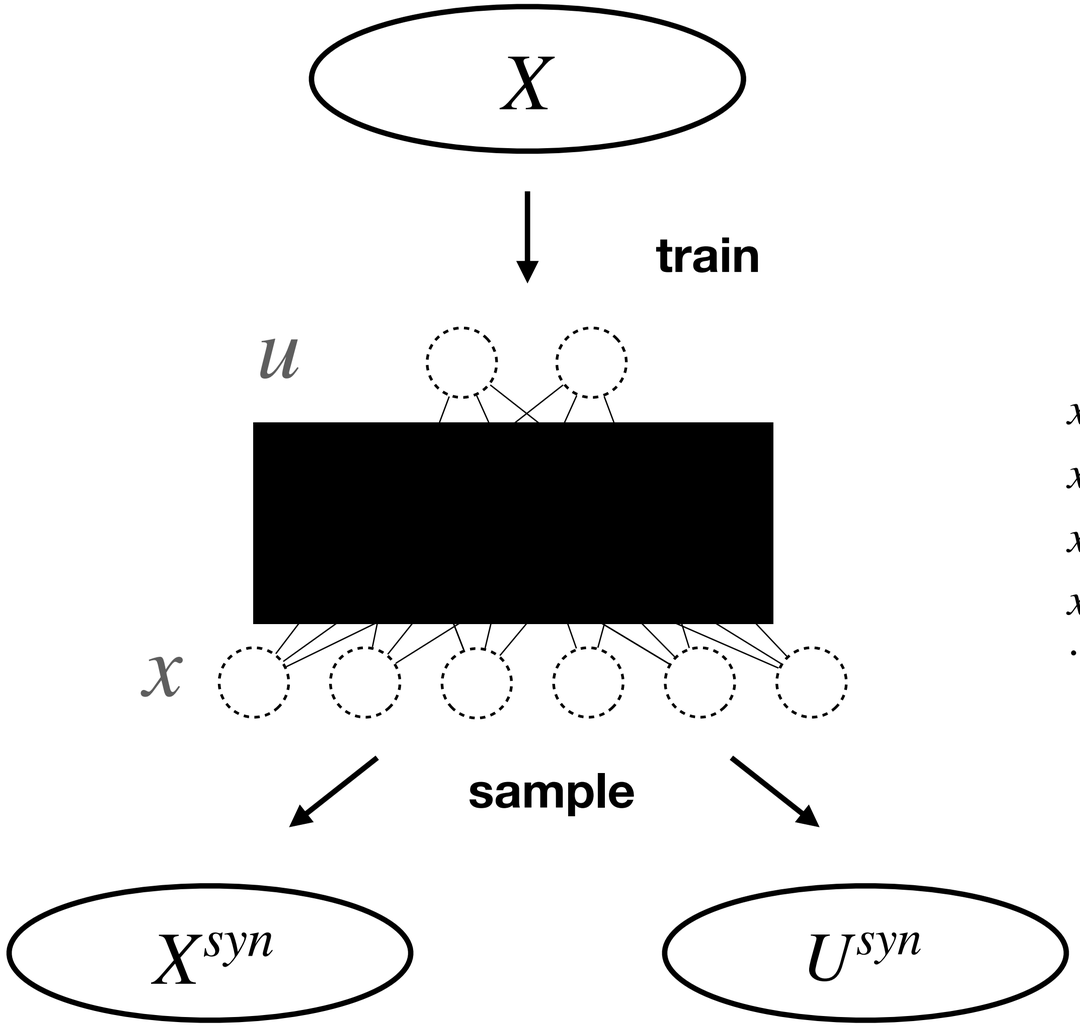


Investigating the joint distribution of discretized synthetic observations



$$x^{syn}_{1.} = 12\ 32\ 4\ 5\ 13\ 5 \qquad u^{syn}_{1.} = -3.2\ 2.8$$
$$x^{syn}_{2.} = 68\ 90\ 8\ 44\ 31\ 9 \qquad u^{syn}_{2.} = -1.2\ 1.4$$
$$x^{syn}_{3.} = 6\ 5\ 78\ 12\ 24\ 36 \qquad u^{syn}_{3.} = 1.2\ -1.2$$
$$x^{syn}_{4.} = 98\ 5\ 6\ 76\ 47\ 67 \qquad u^{syn}_{4.} = 4.2\ -2.0$$
$$\cdots \qquad \cdots$$

Investigating the joint distribution of discretized synthetic observations



$$x^{syn}_{1.} = 0\ 1\ 0\ 0\ 1\ 0 \qquad u^{syn}_{1.} = 0\ 1$$
$$x^{syn}_{2.} = 1\ 1\ 0\ 1\ 1\ 0 \qquad u^{syn}_{2.} = 0\ 1$$
$$x^{syn}_{3.} = 0\ 0\ 1\ 0\ 1\ 1 \qquad u^{syn}_{3.} = 1\ 0$$
$$x^{syn}_{4.} = 1\ 0\ 0\ 1\ 1\ 1 \qquad u^{syn}_{4.} = 1\ 0$$
$$\cdots \qquad \cdots$$

Interactions between observed and latent variables are identified via likelihood ratio statistics

Interactions between observed and latent variables are identified via likelihood ratio statistics

$$
\begin{aligned}
x^{syn}_{1.} &= 0\ 1\ 0\ 0\ 1\ 0 & u^{syn}_{1.} &= 0\\
x^{syn}_{2.} &= 1\ 1\ 0\ 1\ 1\ 0 & u^{syn}_{2.} &= 0\\
x^{syn}_{3.} &= 0\ 0\ 1\ 0\ 1\ 1 & u^{syn}_{3.} &= 1\\
x^{syn}_{4.} &= 1\ 0\ 0\ 1\ 1\ 1 & u^{syn}_{4.} &= 1\\
&\cdots & &\cdots
\end{aligned}
$$

Interactions between observed and latent variables are identified via likelihood ratio statistics

$$
\begin{aligned}
x_{1.}^{syn} &= 0\ 1\ 0\ 0\ 1\ 0 \qquad & u_{1.}^{syn} &= 0 \\
x_{2.}^{syn} &= 1\ 1\ 0\ 1\ 1\ 0 \qquad & u_{2.}^{syn} &= 0 \\
x_{3.}^{syn} &= 0\ 0\ 1\ 0\ 1\ 1 \qquad & u_{3.}^{syn} &= 1 \\
x_{4.}^{syn} &= 1\ 0\ 0\ 1\ 1\ 1 \qquad & u_{4.}^{syn} &= 1 \\
&\cdots & &\cdots
\end{aligned}
$$

	$X_*^{syn}=0$	$X_*^{syn}=1$
$U_1^{syn}=0$		
$U_1^{syn}=1$		

Interactions between observed and latent variables are identified via likelihood ratio statistics

$$
\begin{array}{llllllll}
x_{1.}^{syn} = & 0 & 1 & 0 & 0 & 1 & 0 & \\
x_{2.}^{syn} = & 1 & 1 & 0 & 1 & 1 & 0 & \\
x_{3.}^{syn} = & 0 & 0 & 1 & 0 & 1 & 1 & \\
x_{4.}^{syn} = & 1 & 0 & 0 & 1 & 1 & 1 & \\
\cdots & & & & & & & \\
\end{array}
\qquad
\begin{array}{ll}
u_{1.}^{syn} = & 0 \\
u_{2.}^{syn} = & 0 \\
u_{3.}^{syn} = & 1 \\
u_{4.}^{syn} = & 1 \\
\cdots &
\end{array}
$$

	$X_*^{syn} = 0$	$X_*^{syn} = 1$
$U_1^{syn} = 0$		
$U_1^{syn} = 1$		

$$log(\mu_{i,j,k}) = \lambda + \lambda_i^{X_2} + \lambda_j^{X_6} + \lambda_k^{U} + \lambda_{ij}^{X_2X_6} + \lambda_{ik}^{X_2U} + \lambda_{jk}^{X_6U}$$

$$i, j, k \in \{1,2\}$$

Interactions between observed and latent variables are identified via likelihood ratio statistics

$$
\begin{aligned}
x^{syn}_{1.} &= 0\ 1\ 0\ 0\ 1\ 0 \\
x^{syn}_{2.} &= 1\ 1\ 0\ 1\ 1\ 0 \\
x^{syn}_{3.} &= 0\ 0\ 1\ 0\ 1\ 1 \\
x^{syn}_{4.} &= 1\ 0\ 0\ 1\ 1\ 1 \\
&\cdots
\end{aligned}
\qquad
\begin{aligned}
u^{syn}_{1.} &= 0 \\
u^{syn}_{2.} &= 0 \\
u^{syn}_{3.} &= 1 \\
u^{syn}_{4.} &= 1 \\
&\cdots
\end{aligned}
$$

	$X^{syn}_{*}=0$	$X^{syn}_{*}=1$
$U^{syn}_{1}=0$		
$U^{syn}_{1}=1$		

$$log(\mu_{i,j,k}) = \lambda + \lambda_i^{X_2} + \lambda_j^{X_6} + \lambda_k^{U} + \lambda_{ij}^{X_2X_6} + \lambda_{ik}^{X_2U} + \lambda_{jk}^{X_6U}$$

$$i, j, k \in \{1,2\}$$

Interactions between observed and latent variables are identified via likelihood ratio statistics

$$
\begin{array}{llll}
x^{syn}_{1.} = 0\ 1\ 0\ 0\ 1\ 0 & & u^{syn}_{1.} = 0 \\
x^{syn}_{2.} = 1\ 1\ 0\ 1\ 1\ 0 & & u^{syn}_{2.} = 0 \\
x^{syn}_{3.} = 0\ 0\ 1\ 0\ 1\ 1 & & u^{syn}_{3.} = 1 \\
x^{syn}_{4.} = 1\ 0\ 0\ 1\ 1\ 1 & & u^{syn}_{4.} = 1 \\
\cdots & & \cdots
\end{array}
$$

	$X^{syn}_{*} = 0$	$X^{syn}_{*} = 1$
$U^{syn}_{1} = 0$		
$U^{syn}_{1} = 1$		

$$log(\mu_{i,j,k}) = \lambda + \lambda_i^{X_2} + \lambda_j^{X_6} + \lambda_k^{U} + \lambda_{ij}^{X_2X_6} + \lambda_{ik}^{X_2U}$$

$$i, j, k \in \{1,2\}$$

Interactions between observed and latent variables are identified via likelihood ratio statistics

$$
\begin{array}{llll}
x_{1.}^{syn} = & 0\ 1\ 0\ 0\ 1\ 0 & u_{1.}^{syn} = & 0 \\
x_{2.}^{syn} = & 1\ 1\ 0\ 1\ 1\ 0 & u_{2.}^{syn} = & 0 \\
x_{3.}^{syn} = & 0\ 0\ 1\ 0\ 1\ 1 & u_{3.}^{syn} = & 1 \\
x_{4.}^{syn} = & 1\ 0\ 0\ 1\ 1\ 1 & u_{4.}^{syn} = & 1 \\
\cdots & & \cdots &
\end{array}
$$

	$X_*^{syn} = 0$	$X_*^{syn} = 1$
$U_1^{syn} = 0$		
$U_1^{syn} = 1$		

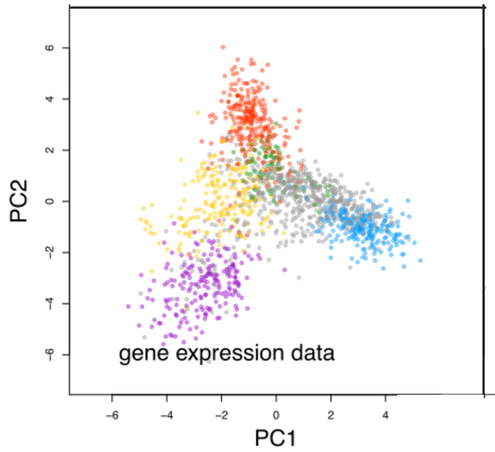
$$log(\mu_{i,j,k}) = \lambda + \lambda_i^{X_2} + \lambda_j^{X_6} + \lambda_k^{U} + \lambda_{ij}^{X_2X_6} + \lambda_{ik}^{X_2U}$$

$$log(\mu_{i,j,k}) = \lambda + \lambda_i^{X_2} + \lambda_j^{X_6} + \lambda_k^{U} + \lambda_{ij}^{X_2X_6} + \lambda_{ik}^{X_2U} + \lambda_{jk}^{X_6U}$$

$$i, j, k \in \{1,2\}$$

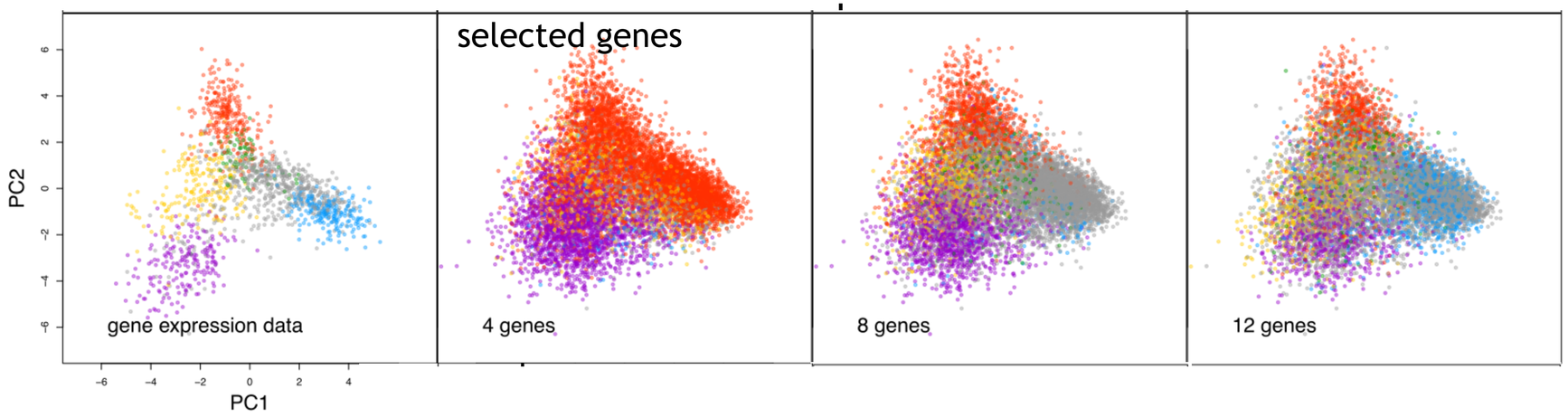
Patterns in selected genes allow to assign synthetic data to their overall most similar training example

Synthetic data matched to training data based on patterns in ...



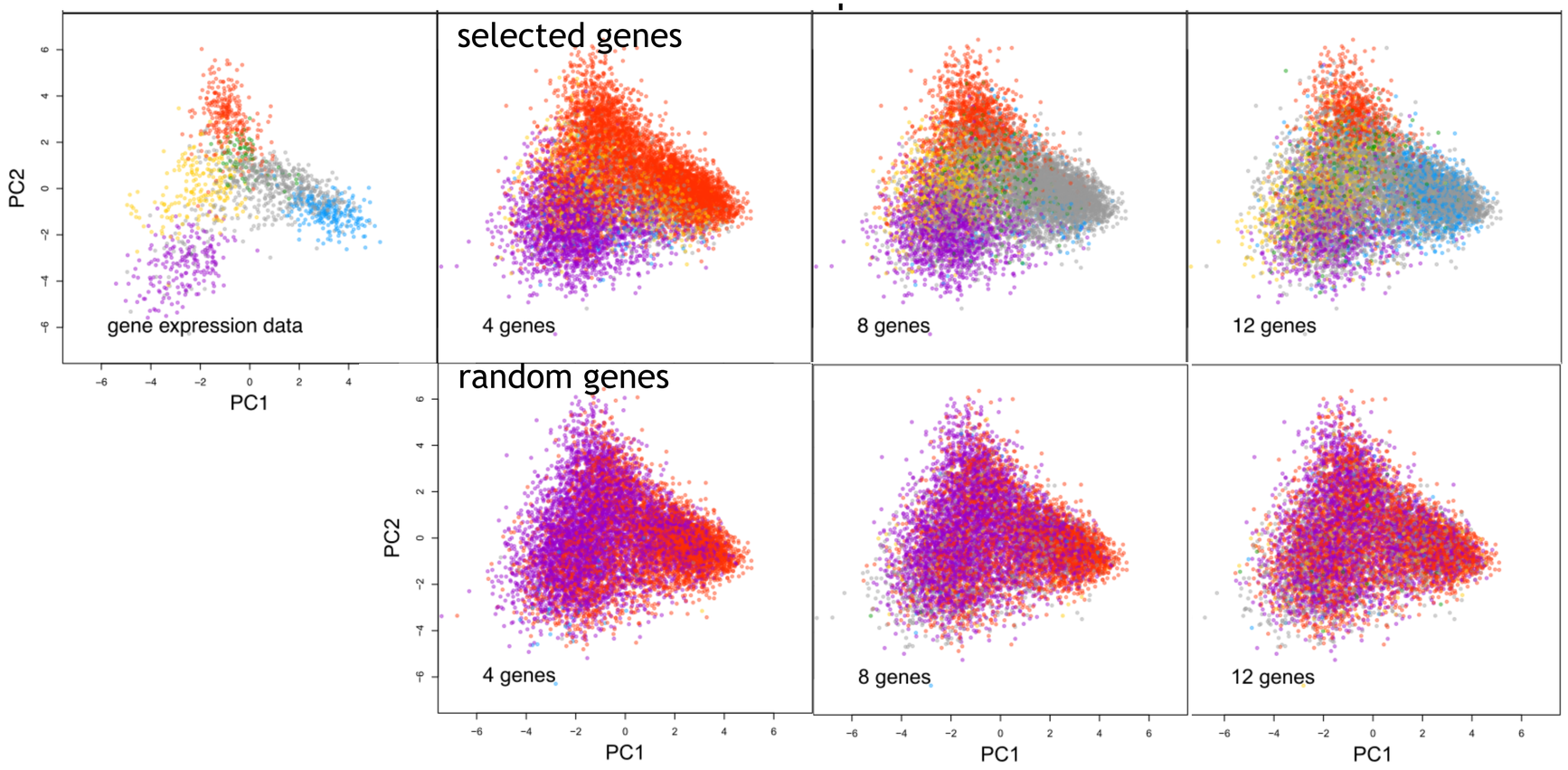
Patterns in selected genes allow to assign synthetic data to their overall most similar training example

Synthetic data matched to training data based on patterns in ...



Patterns in selected genes allow to assign synthetic data to their overall most similar training example

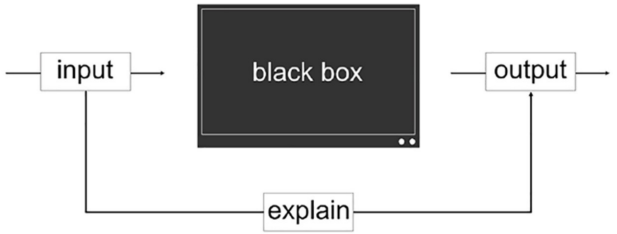
Synthetic data matched to training data based on patterns in ...



Post-hoc vs. Model-based interpretability

Post-hoc:

- Infer feature importance by approximating the non-interpretable model with simpler models



Model-based:

- Refers to models that incorporate mechanisms that allow direct interpretation of learned relationships

Treppner, M. et al. (2022) Interpretable generative deep learning: an illustration with single cell gene expression data. *Hum Gen*.

Model-based interpretability

Model-based interpretability

- Refers to models that incorporate mechanisms that allow direct interpretation of learned relationships

Model-based interpretability

- Refers to models that incorporate mechanisms that allow direct interpretation of learned relationships

- Gain in interpretability is often accompanied by a loss in predictive accuracy or a lower support for the modeled data, i.e., lower likelihoods

Model-based interpretability

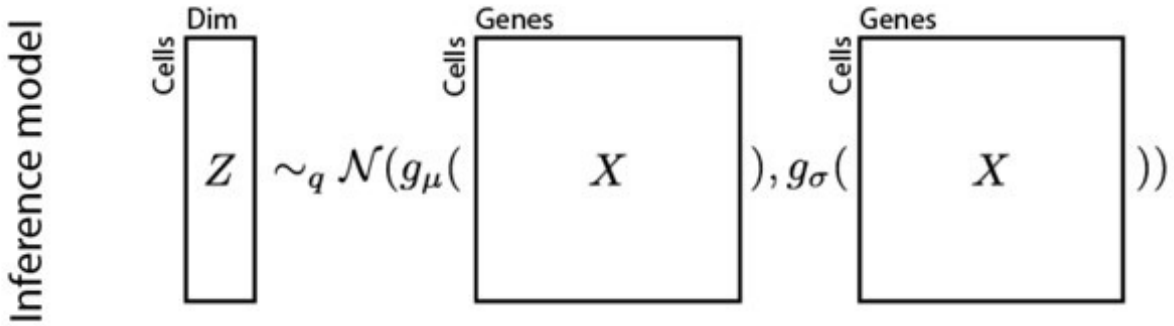
- Refers to models that incorporate mechanisms that allow direct interpretation of learned relationships

- Gain in interpretability is often accompanied by a loss in predictive accuracy or a lower support for the modeled data, i.e., lower likelihoods

- Disentanglement strategies, deep learning approaches are modifed to enforce, e.g., linearity or monotonicity

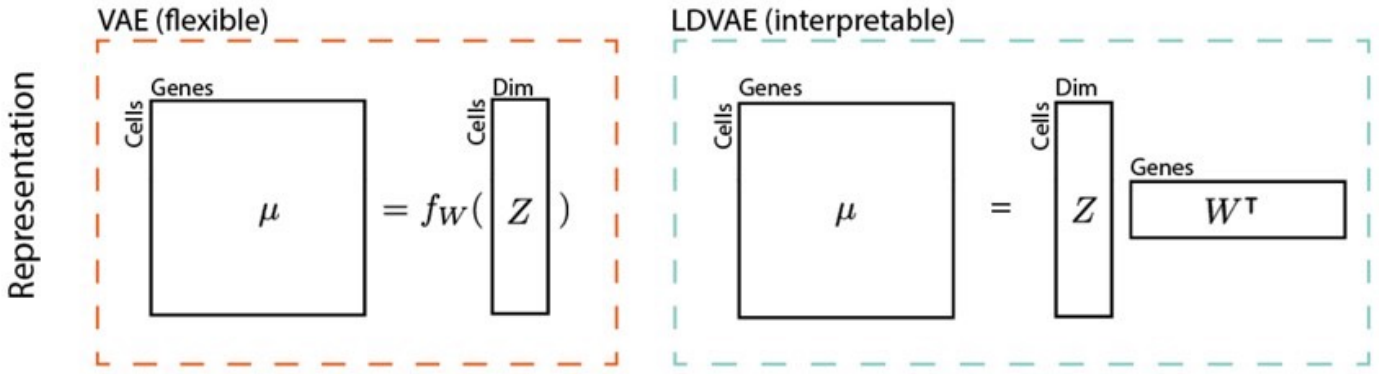
Linearization: LDVAE

Linearly Decoded Variational Autoencoder



Linearization: LDVAE

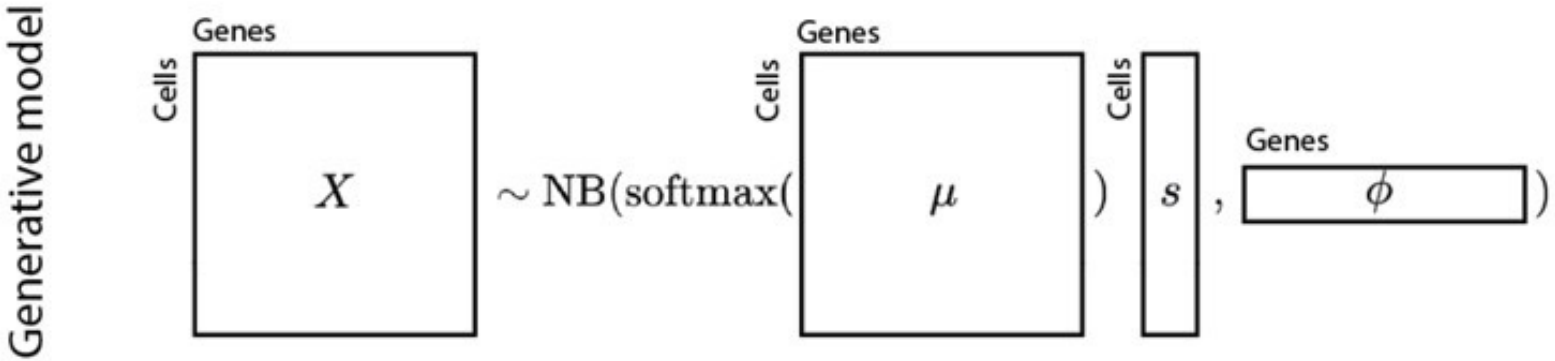
Linearly Decoded Variational Autoencoder



Svensson et al. (2020) Interpretable factor models of single-cell RNA-seq via variational autoencoders. *Bioinformatics*.

Linearization: LDVAE

Linearly Decoded Variational Autoencoder



Svensson et al. (2020) Interpretable factor models of single-cell RNA-seq via variational autoencoders. *Bioinformatics*.

Pros and cons – Post-hoc vs. model-based

- Post-hoc approaches more flexible than model-based approaches

- Post-hoc approaches more flexible than model-based approaches
- Possible to employ the best performing approach and only in the second step care about the interpretations

- Post-hoc approaches more flexible than model-based approaches
- Possible to employ the best performing approach and only in the second step care about the interpretations
- Risk in post-hoc approaches is that the model, due to uncontrollable noise, focuses on artifacts

- Post-hoc approaches more flexible than model-based approaches
- Possible to employ the best performing approach and only in the second step care about the interpretations
- Risk in post-hoc approaches is that the model, due to uncontrollable noise, focuses on artifacts
- Model-based interpretability approaches are adapted to a specific model class and thus cannot be used flexibly

- Post-hoc approaches more flexible than model-based approaches
- Possible to employ the best performing approach and only in the second step care about the interpretations
- Risk in post-hoc approaches is that the model, due to uncontrollable noise, focuses on artifacts
- Model-based interpretability approaches are adapted to a specific model class and thus cannot be used flexibly
- Additionally, model-based interpretability often leads to reduced predictive performance or increased reconstruction errors

Thanks to:

UNIVERSITÄTS KLINIKUM FREIBURG

fdm freiburger zentrum für datenanalyse und modellbildung

DFG Deutsche Forschungsgemeinschaft

MEDIZIN INFORMATIK BIOLOGIE

NEPH GENOS

Special thanks to:

Harald Binder and the AG Machine Learning

Contact:
Martin Treppner
Institute of Medical Biometry and Statistics
Faculty of Medicine and Medical Center - University of Freiburg, Germany
treppner@imbi.uni-freiburg.de

@IMBI_fr

UNIVERSITÄTS KLINIKUM FREIBURG

Sources:

Nice overview: https://christophm.github.io/interpretable-ml-book/index.html

Linardatos, P. et al. (2020). Explainable AI: A review of machine learning interpretability methods. Entropy.

Treppner, M. et al. (2022) Interpretable generative deep learning: an illustration with single cell gene expression data. Hum Gen.

Ancona, M. et al. (2017) Towards better understanding of gradient-based attribution methods for deep neural networks. arXiv.

Sundararajan et al. (2017) Axiomatic attribution for deep networks. PMLR.

https://distill.pub/2020/attribution-baselines/

Hess, M. et al. (2020). Exploring generative deep learning for omics data using log-linear models. Bioinformatics .