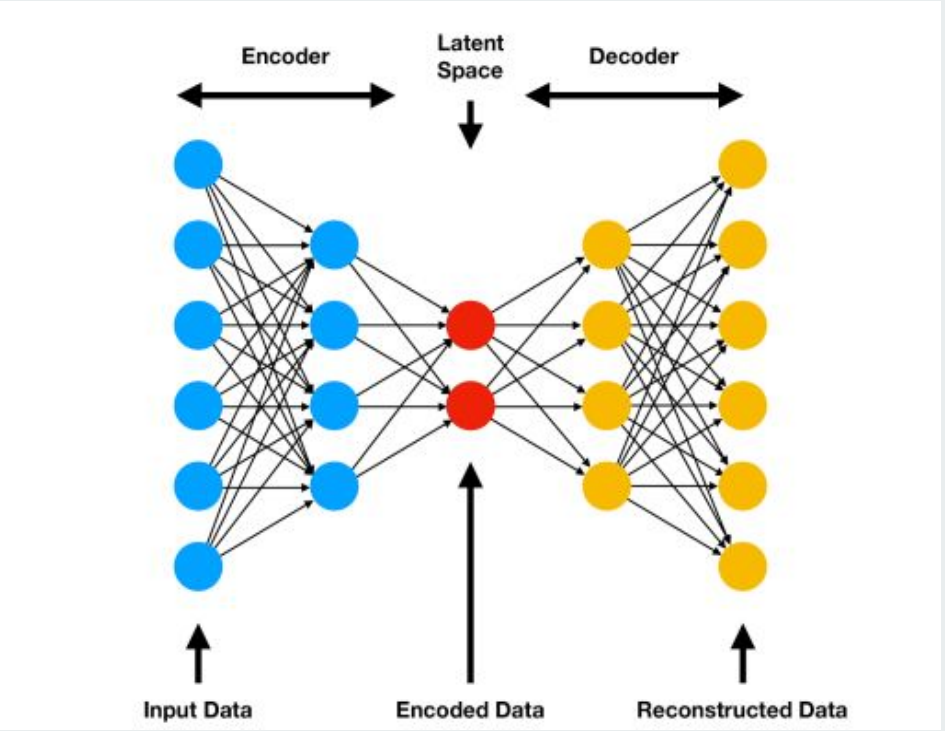
XA4C: eXplainable representation learning via Autoencoders revealing Critical genes



https://neptune.ai/blog/representation-learning-with-autoencoder

Εισαγωγή/Υπόβαθρο

1. Autoencoders are able to learn the hidden representations of input data despite the input being noisy and heterogeneous, leading to “latent variables” that are cleaner and more orthogonal for next stages of analysis.
2. Learned representations, e.g., the “latent variables” in an autoencoder (AE), are difficult to interpret, not to mention prioritizing essential genes for functional follow-up.
3. XA4C = eXplainable Autoencoder for Critical genes
4. Critical genes = genes that contribute highly to learned representations (e.g., latent variables in an autoencoder).
5. XA4C disentangles the black box of the neural network of an autoencoder by providing each gene’s contribution to the latent variables in the autoencoder, based on which Critical genes are prioritized.

Μέθοδοι/Αλγόριθμοι

These methods and algorithms collectively form the basis of the XA4C approach, enabling the identification of Critical genes and their impact on biological pathways:

A. **Autoencoder**: to learn representations (i.e., latent variables) of input gene expression profiles.

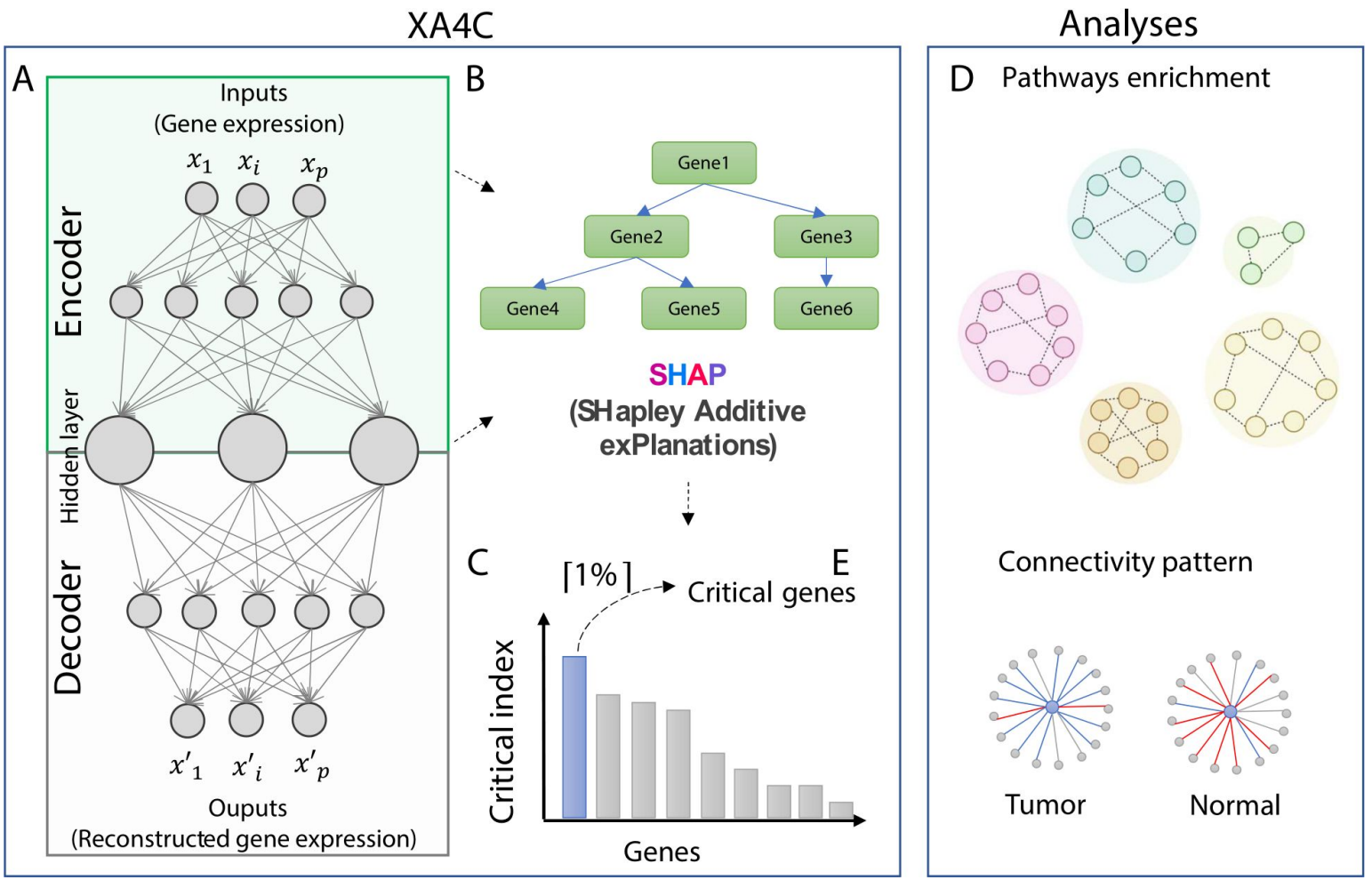
B. **XGBoost** to quantify each gene's contribution to the latent variables and **Tree SHAP explanation** to assess the contribution of inputs to representations. XGBoost and TreeSHAP are utilized to evaluate SHAP values and Critical indexes for all genes.

   a. **SHapley Additive exPlanations (SHAP) values**: a method (inspired by the popular economic concept of "Shapley Value") quantifying the contribution of a player in a game.

      XA4C outputs SHAP values for input gene expression individually and quantifies the contribution of each input to each representation (i.e., latent variable) and aggregate them to form "Critical index" for each gene. These Critical indexes will be used to prioritize Critical genes based on user specified cutoff, e.g., 1%.

C. KEGG Pathway is a collection of manually drawn pathway maps representing our knowledge of the molecular interaction, reaction, and relation networks. **KEGG Pathway Enrichment analysis** is a computational method used to identify which KEGG pathways are overrepresented (meaning these pathways are more represented in the input list than would be expected by chance) by prioritized genes with SHAP values. (https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0126492)

D. **Connectivity analysis** discloses interaction patterns among genes centered by Critical genes in pathways. Comparison between cancerous (tumor) cells and normal cells.


XA4C
Analyses
A
Inputs
(Gene expression)
x_1 x_i x_p
Encoder
Hidden layer
Decoder
x'_1 x'_i x'_p
Ouputs
(Reconstructed gene expression)
B
Gene1
Gene2
Gene3
Gene4
Gene5
Gene6
SHAP
(SHapley Additive
exPlanations)
C
[1%]
Critical index
Genes
E
Critical genes
D
Pathways enrichment
Connectivity pattern
Tumor
Normal

Αποτελέσματα

1. Applying XA4C to gene expression data in six cancers showed that Critical genes capture essential pathways underlying cancers.

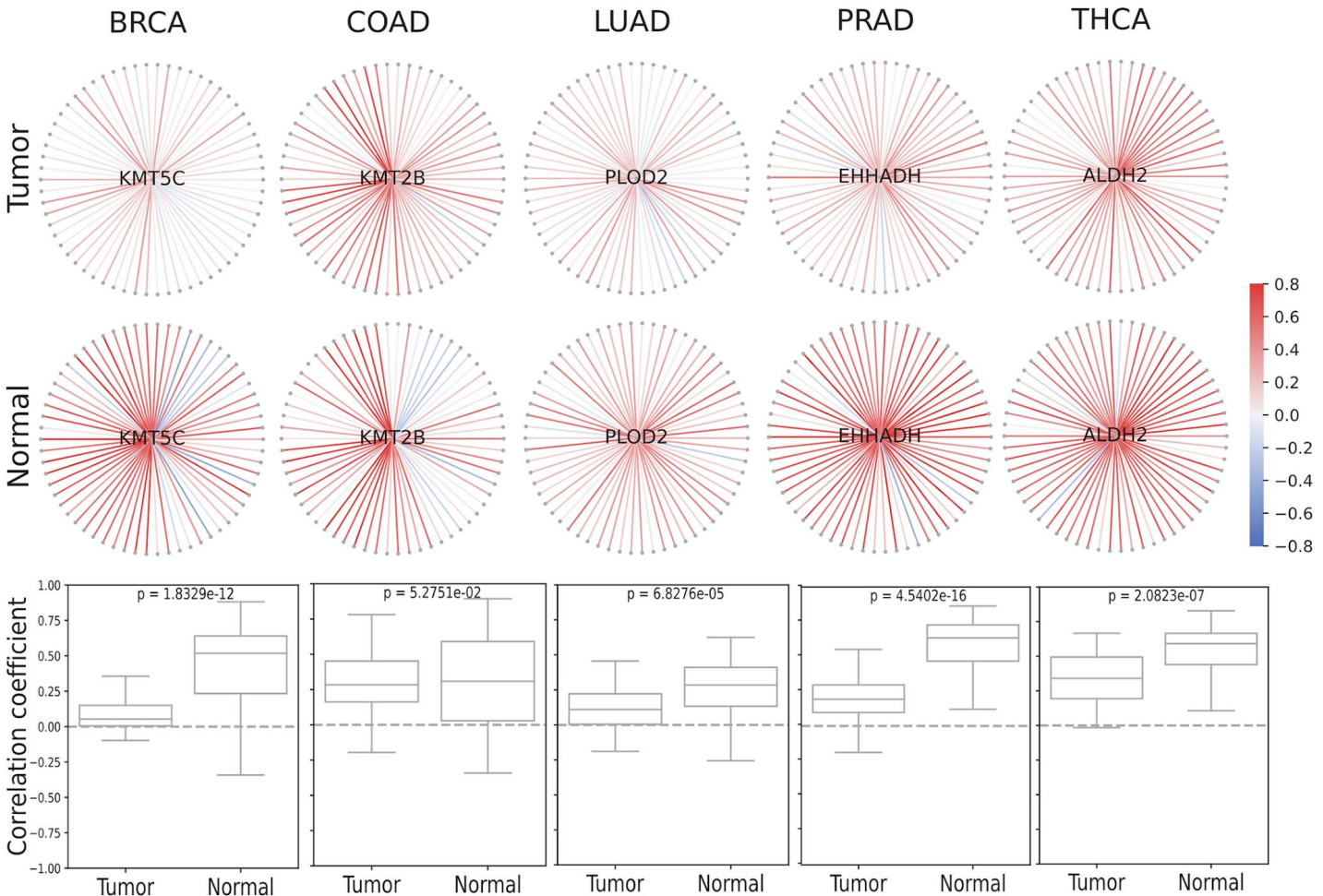
2. Remarkably, Critical genes has little overlap with Hub or DiffEx genes, however, has a higher enrichment in a comprehensive disease gene database (DisGeNET) and a cancer-specific database (COSMIC), evidencing its potential to disclose massive unknown biology. As an example, we discovered five Critical genes sitting in the center of Lysine degradation (hsa00310) pathway, displaying distinct interaction patterns in tumor and normal tissues. In conclusion, XA4C facilitates explainable analysis using RL and Critical genes discovered by explainable RL empowers the study of complex interactions.

3. Notably, we discovered that Critical genes enjoy two properties: (1) Their overlap with traditional differentially expressed genes and hub genes are poor, suggesting that they indeed brought novel insights into transcriptome data that cannot be captured by traditional analysis. (2) The enrichment of Critical genes in a comprehensive disease gene database (DisGeNET) and cancer-specific database (COSMIC) are higher than differentially expressed or hub genes, evidencing their strong relevance to disease pathology and its potential to disclose massive unknown biology. Therefore, we conclude that XA4C can reveal an additional landscape of gene expression data.

4. As an example, we discovered five Critical genes sitting in the center of Lysine degradation (hsa00310) pathway, displaying distinct interaction patterns in tumor and normal tissues. In conclusion, XA4C facilitates explainable analysis using RL and Critical genes discovered by explainable RL empowers the study of complex interactions.

Αποτελέσματα (συνέχεια)

The Lysine degradation pathway (I00310) is used. Critical genes (light blue) are located at the core of the network, surrounded by additional genes from the same pathway (gray). The boundaries of Pearson's correlation coefficients range from +0.8 (red) to -0.8 (blue). Boxplots show the distributions of two sets of correlations (tumor vs. normal) together with the P-value of the Kolmogorov-Smirnov test, with the null hypothesis being that the two samples were chosen from the same distribution. Critical genes shown in this figure are novel as they have not been identified by traditional analysis search for Hub nor DiffEx genes.



The End!