In silico indication discovery using deep learning for noise reduction

Dominik Hartl
Idorsia Pharmaceuticals Ltd, Drug Discovery Department, Translational Science, Allschwil, Switzerland

Abstract:

In order to explore the full potential of a drug candidate it is crucial to consider its effectiveness in all diseases. Traditional approaches to match a drug candidate with possible diseases, such as literature research and wet-lab experiments, are however resource intensive. In this novel approach we established an *in silico* framework to discover target diseases for a drug candidate based on matching its molecular effect with patient data based on gene expression. The approach relies on data originating from different sources which introduces considerable noise. To reduce this noise we took advantage of machine learning. We adapted a strategy originating from image analysis where autoencoders decrease random noise. Autoencoders are neural networks that encode a data set into low dimensional space and then decode it again. By forcing the data through a bottleneck in dimensionality, random noise is reduced as only higher order features can be considered. In this application we can imagine that higher order features represent *de novo* generated pathways.

Strategy:

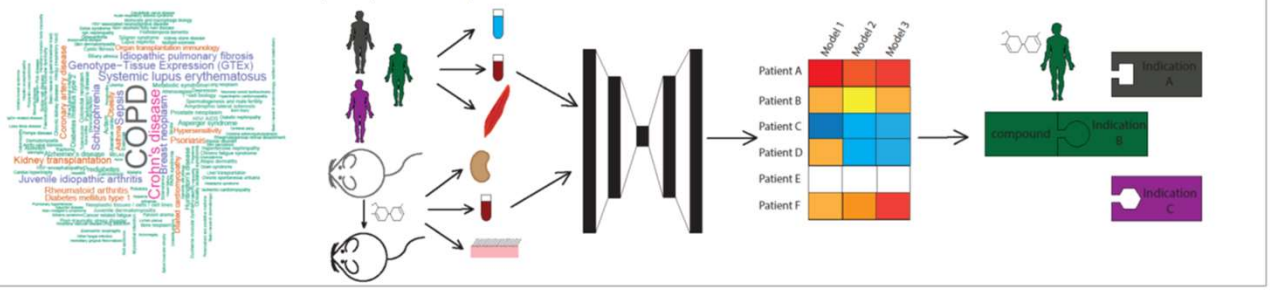
Dataset containing ~140 Diseases from ~12200 Samples

Generation of compound profiles in model organisms in different tissues and patient profiles to compare with.

Noise reduction using autoencoder

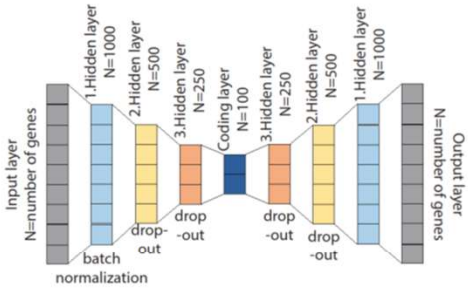
correlation of compound with patient profiles to assess complementarity

Match compounds and patients/diseases based on highest anticorrelation

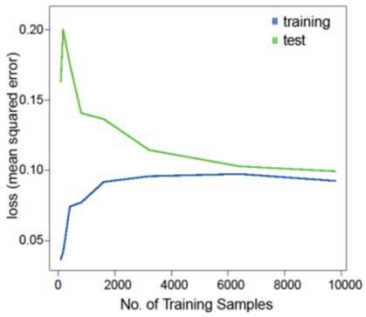


Sufficient data to train the autoencoder:

Autoencoder design:
In the first hidden layer in the encoder batch normalization is applied, hidden layers 2 and 3 were trained with 5% dropout to avoid overfitting. Exponential linear unit was used as activation function in all nodes.

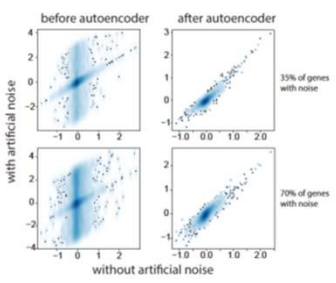


Lineplot showing model loss for training and test set (80/20 split) if it is trained on different number of training samples for 25 epochs.
> 6000 samples are sufficient to train the model.

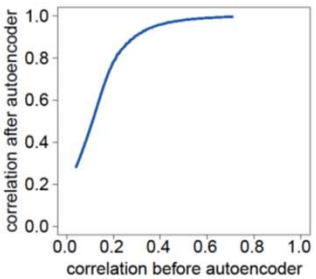


Autoencoder efficiently reduces noise and increases signal:

Reduction of artificial noise in one sample. Random noise was added to randomly selected genes in the dataset (magnitude of noise: -3 to 3), scatterplots show the sample with and without noise before the autoencoder (left panel) and after the autoencoder (right panel).



Lineplot displaying correlation of original samples with samples with random noise introduced before versus after the autoencoder. Correlations were calculated on all samples in the data, the line represents a loess curve fitted to all the data points.



Density plots of the correlations of all samples per disease including the mouse model data before and after the autoencoder processing. The flanks of the distribution are extended indicating an increase in signal.