Combining transfer learning using latent variables learned UK Biobank repository to predict gene-environment interactions for classification of multiple sclerosis and rheumatoid arthritis.

Many complex diseases are **major health challenges** because they are common but difficult to decipher. Intricate interactions between the genetic landscape and environmental risk factors contribute to disease but remain elusive.

To capture the multifactorial complexity in common complex disease, where many genetic variants with small effects contribute to disease, we will here employ multiple **deep artificial neural networks** (DNNs) in combination with other **machine learning** algorithms on genetic and risk factor data from large well-defined cohorts of cases and controls including several national cohorts together with the UK Biobank. For this purpose we will use three large genetics cohorts that we believe can be integrated using transfer learning. First, UK Biobank genotype data from ~500,000 individuals has been used to identify several auto-encoders that compress the ~780,000 single nucleotide polymorphisms ~10-fold. We have seen that these underlying latent variables can efficiently be used for predicting disease phenotypes and new disease associated factors using our previously published methods (Dwivedi et al Nature Comm 2020). In this project we would like to make use of transfer learning and adapt this representation together with 10-20 different life-style factors associated with multiple sclerosis and rheumatoid arthritis in two cohorts of ~20,000 individuals (which we have access to through collaborators). The aim is then to identify synergistic factors between the two, using first simple machine learning methods and also deep neural networks. This could lead to both a generally applicable strategy to dissect other complex genetic and environmental diseases and to better advice of risks in the studied diseases.

Ref.

Dwivedi, S.K., Tjärnberg, A., Tegnér, J. and Gustafsson, M., 2020. Deriving disease modules from the compressed transcriptional space embedded in a deep autoencoder. *Nature communications*, *11*(1), pp.1-10.