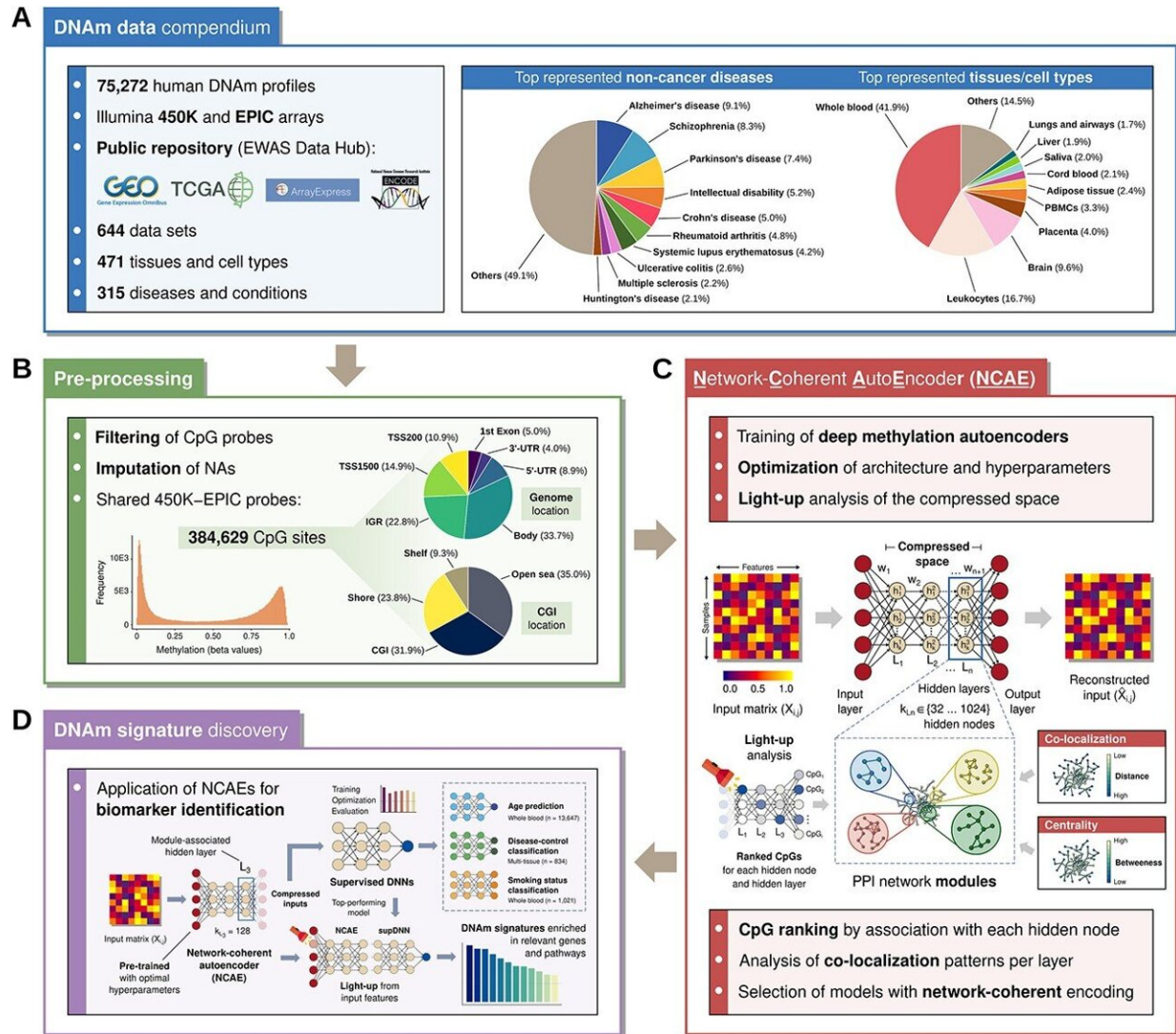# A step towards AI-based precision medicine

October 11 2023, by Karin Söderlund Leifler

Schematic of the workflow for training and functionalization of network-coherent autoencoders (NCAEs). (A) Summary of DNAm datasets included in the compendium, top represented non-cancer diseases, and tissue and cell types. (B) Pre-processing steps and description of probe set genomic and CpG island

(CGI) locations. (C) Training and selection of deep autoencoders based on the network coherence of their latent space. (D) Functionalization of deep NCAE compressed representations for the identification of DNAm signatures using concatenated task-specific deep supervised neural networks (DNNs). Credit: *Briefings in Bioinformatics* (2023). DOI: 10.1093/bib/bbad293

Artificial intelligence, AI, which finds patterns in complex biological data could eventually contribute to the development of individually tailored health care. Researchers at Linköping University, Sweden, have developed an AI-based method applicable to various medical and biological issues. Their models can for instance accurately estimate people's chronological age and determine whether they have been smokers or not.

There are many factors that can affect which out of all our genes are used at any given point in time. Smoking, dietary habits and environmental pollution are some such factors. This regulation of gene activity can be likened to a power switch determining which genes are switched on or off, without altering the actual genes, and is called epigenetics.

Researchers at Linköping University have used data with epigenetic information from more than 75,000 human samples to train a large number of AI neural network models. They hope that such AI-based models could eventually be used in precision medicine to develop treatments and preventive strategies tailored to the individual. Their models are of the autoencoder type, which self-organizes the information and finds interrelation patterns in the large amount of data.

To test their model, the LiU researchers compared it with existing models. There are already existing models of the effects of smoking on

the body, building on the fact that specific epigenetic changes reflect the effect of smoking on the functioning of the lungs.

These traces remain in the DNA long after a person has quit smoking, and this type of model can identify whether someone is a current, former or never smoker. Other models can, based on epigenetic markers, estimate the chronological age of an individual, or group individuals according to whether they have a disease or are healthy.

The LiU researchers trained their autoencoder and then used the result to answer three different queries: age determination, smoker status and diagnosing the disease systemic lupus erythematosus, SLE. Although the existing models rely on selected epigenetic markers known to be associated with the condition they aim to classify. However, it turned out that the LiU researchers' autoencoders functioned better or equally well.

"Our models not only enable us to classify individuals based on their epigenetic data. We found that our models can identify previously known epigenetic markers used in other models, but also new markers associated with the condition we're examining. One example of this is that our model for smoking identifies markers associated with respiratory diseases, such as lung cancer, and DNA damage," says David Martínez, Ph.D. student at Linköping University.

The objective of the autoencoder models is to enable compression of extremely complex biological data into a representation of the most relevant characteristics and patterns in data.

"We didn't steer the model and had no hypotheses based on existing biological knowledge, but let the data speak for itself. When subsequently looking at what was happening in the autoencoder, we saw that data self-organized in a way similar to how it works in the body," says Mika Gustafsson, professor of translational bioinformatics at

Linköping University, who led the study now published in *Briefings in Bioinformatics*.

In the next step, the researchers can use the most important characteristics found by the autoencoder to create models able to classify for a large amount of environment-related, individual-specific factors where there is not enough training data to train more complex AI models on.

Certain types of AI are sometimes likened to a black box that provides answers, but humans cannot see how the AI arrived at the answer. Mika Gustafsson and his colleagues however strive to create interpretable AI models that, so to speak, let the researchers peek under the lid of the "black box" to understand what is going on inside.

"We want to be able to understand what the model shows us about the biology behind disease and other conditions. Then we'll see not only whether someone is ill or not, but, by interpreting data, we'll also have a chance to learn why," says Mika Gustafsson.

  More information: David Martínez-Enguita et al, NCAE: data-driven representations using a deep network-coherent DNA methylation autoencoder identify robust disease and risk factor signatures, *Briefings in Bioinformatics* (2023). DOI: 10.1093/bib/bbad293

Provided by Linköping University

study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.