novo nordisk® Radcliffe Department of Medicine UNIVERSITY OF OXFORD

Integrating genetics and deep neural networks to identify future drug targets for cardiometabolic disease

Oxford supervisor: Associate Prof Alejo Nevado-Holgado[1]
Novo Nordisk supervisors: Dr Joanna Howson[2], Dr Sile Hu[2] and Dr Robert Kitchen[2]

Department: 1. Department of Psychiatry
2. Novo Nordisk Research Centre Oxford

Project outline

Background

In the last 10 years, GWAS has revolutionised our understanding of the inherited basis of disease and we can now use human genetics for drug target discovery. The current challenge is to identify the underlying causal genes, pathways and mechanisms. AI techniques, more particularly Neural Networks (NNs), are transforming multiple engineering industries by revolutionizing our capacity to analyse complex data [1,2]. They are also starting to be used in biology [3,4], medicine [5,6], and genomics [7,8], where they hold considerable promise. The sensitivity of NNs becomes especially acute where large amounts of information-rich data is available. This is particularly the case in genomics, where deep genotyping data (e.g. Whole Genome Sequencing - WGS) is being made available for ever-larger datasets of hundreds of thousands of individuals (e.g. UK Biobank). We have started successfully applying neural networks to genetic data to capture non-linear effects and interactions more readily than conventional statistical modelling. This success has been achieved with the early NN architectures that emerged during 2010s – e.g. fully connected NNs, LSTMs, and transformers. More recent and sophisticated algorithms, such as contrastive learning or convolutional-transformer methods, have very recently been proposed and have achieved significant breakthroughs in image recognition and language understanding. There are good reasons to explore the applications of these newer architectures to genetics, as well as expanding the more traditional ones that we are continuing to develop, because they could lead to new understanding of the genetic architecture of traits of interest.

Aims

Here we propose leveraging (1) Artificial Intelligence (AI) techniques, (2) the wealth of large human genomic datasets now available, (3) and our experience of applying AI to genetic data to identify potential causal genes for cardiometabolic traits. The proteins encoded by these causal genes are potential pharmacological targets for the treatment of cardiometabolic diseases, such as atherosclerosis, diabetes, insulin resistance and non-alcoholic fatty liver disease.

Work to be undertaken

Building on this prior experience, we will further refine our fully connected NNs, recurrent NNs and transformers, and adopt new architectures that have been transforming the field [9-11]

The NNs will use whole exome and whole genome sequencing from UK Biobank [9], while other comparable datasets will be used for validation. The fellow will be expected to lead a thorough investigation of the optimal way to select variants for inclusion in the NN models. Approaches could include, all 95% credible sets from fine-mapping approaches, polygenic risk scores, variants based on functional annotations, or variants within a window containing the gene of interest . Fully NN methods will also be explored to select the most relevant variants and genetic regions. Regions of interest will then be fed to a NN with the objective of further decoding the genetic context and more complex features, such as motifs associated with protein conformation, which NNs are proficient at detecting (e.g. AlphaFold architecture) As example, this NN could be trained to model the structure of the genome in the same fashion as natural language models (e.g. BERT, GPT) are trained to model the structure of natural human language. Then this model will be used and fine-tuned in downstream tasks, such as in finding which genes associated with disease or protein concentration. We encourage the fellow to investigate and explore further alternatives and methods.

We will apply the approach across cardiometabolic conditions including (the following outcomes) diagnosis of cardiovascular disease (e.g. atherosclerosis), type 2 diabetes, chronic kidney disease, obesity and adiposity related traits. We will harness information on family history of diseases when available (e.g. family history of diabetes), and endophenotypic markers of disease (e.g. insulin resistance).

Recent work, from Langenberg and colleagues in Science and from Stefannson and colleagues in Nature Genetics, have shown the utility of proteins in relation to disease understanding, therefore we will also experiment with including protein concentrations in the NNs. For each such input we will train our NNs on on >3,000 proteins from >50,000 volunteers. After cross-validation, we will record the accuracy that our NNs achieved on predicting the outcome variable (e.g. diagnosis of the given cardiometabolic disease).

We will prioritise the genes that are identified as significant for follow-up using a suite of bioinformatic approaches and assess their viability as potential therapeutic targets. We will collaborate with *in vitro* scientists at Novo Nordisk to validate the hypothesised targets in phenotypic assays.

Bibliography

1. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
2. Rusk, N. Deep learning. *Nat. Methods* **13**, 35–35 (2016).
3. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).
4. Zhavoronkov, A. *et al.* Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* **37**, 1038–1040 (2019).
5. Esteva, A. *et al.* A guide to deep learning in healthcare. *Nat. Med.* **25**, 24–29 (2019).
6. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
7. Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).
8. Wong, A. K., Sealfon, R. S. G., Theesfeld, C. L. & Troyanskaya, O. G. Decoding disease: from genomes to networks to phenotypes. *Nat. Rev. Genet.* (2021) doi:10.1038/s41576-021-00389-x.

9. Poplin, R., Chang, PC., Alexander, D. et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol* (2018) doi:10.1038/nbt.4235
10. Liu, Zhuang, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell and Saining Xie. A ConvNet for the 2020s. *ArXiv* (2022).
11. He, Kaiming, Haoqi Fan, Yuxin Wu, Saining Xie and Ross B. Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition CVPR* (2020)
12. Palmer, L. J. UK Biobank: bank on it. *The Lancet* **369**, 1980–1982 (2007).

Contributions of Oxford and NNRCO supervisors:

Oxford supervisor, Alejo Nevado-Holgado will provide expertise on NNs methodology and application. NNRCO supervisor will provide expertise on cardiometabolic disease genetics, statistical methodology and target identification.

Supervisor’s recent relevant publications (5 max per supervisor):

Prof Alejo Nevado-Holgado, Oxford

1. *Replication study of plasma proteins relating to Alzheimer’s pathology.* L Shi, LM Winchester, S Westwood, […] AJ Nevado-Holgado. 10.1002/alz.12322 Alzheimer's & Dementia, 2021. **IF 14.4**
2. *High Blood Pressure and Risk of Dementia: a Two-Sample Mendelian Randomization study in the UK Biobank.* W Sproviero, L Winchester […] AJ Nevado-Holgado. 10.1016/j.biopsych.2020.12.015 Biological Psychiatry. 2021. **IF 12.1**
3. *Med7: A transferable clinical natural language processing model for electronic health records.* A Kormilitzina, N Vaci, Q Liu, AJ Nevado-Holgado. 10.1016/j.artmed.2021.102086 Artificial Intelligence in Medicine, 2021. **IF 5.3**
4. *Discovery and validation of plasma proteomic biomarkers relating to brain amyloid burden by SOMAscan assay.* L Shi, S Westwood, AL Baird, L Winchester, […] A Nevado-Holgado. Alzheimer's & Dementia, 2019. **IF 14.4**
5. *A metabolite-based machine learning approach to diagnose Alzheimer’s type dementia in blood: Results from the European Medical Information Framework for Alzheimer's Disease biomarker discovery cohort.* D Stamate, M Kim, P Proitsi, S Westwood, A Baird, AJ Nevado-Holgado, [...] C Legido-Quigley. Alzheimer's & Dementia, 2019. **IF 14.4**

Dr Joanna Howson, NNRCO

1. An atlas of mitochondrial DNA genotype-phenotype associations in the UK Biobank; Yonova Doing E, Calabrese C, Gomez-Duran A, Schon K, Wei W, Karthikeyan S, Chinnery P*, Howson JMM*; Nature Genetics, 2021; 53:982-993
2. Discovery of rare variants associated with Blood pressure regulation through meta-analysis of 1.3 million individuals; Surendran P, +200 authors, Howson JMM; Nature Genetics, 2020 52;1314-1332
3. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes; Malik R, + 80 authors, Howson JMM*, Kamatani Y*, Debette S*, Dichgans M*; Nature Genetics, 2018, 50; 524-537

4. Fifteen new risk loci for coronary artery disease highlight arterial-wall specific mechanisms; Howson JMM et al.; *Nature Genetics*, 2017, 49; 1113-1119
5. Robust and efficient method for Mendelian Randomisation with hundreds of genetics variants; Burgess S, Foley CN, Allara E, Staley JR, Howson JMM; *Nat Communications*, 2020 11;376

* Authors contributed equally