ECOLE DOCTORALE PIERRE LOUIS DE SANTE PUBLIQUE A PARIS
EPIDEMIOLOGIE ET SCIENCES DE L'INFORMATION BIOMEDICALE

Directeur : Pierre-Yves Boëlle
Responsable pour l'Université Paris Cité : Isabelle Boutron

PROPOSITION DE SUJET DE THESE

SIGLE ET NOM DU LABORATOIRE : UR 7537 — BIOSTM BIOSTATISTIQUE, TRAITEMENT ET MODELISATION DES DONNEES BIOLOGIQUES
NOM DE L'EQUIPE : UR 7537 — BIOSTM BIOSTATISTIQUE, TRAITEMENT ET MODELISATION DES DONNEES BIOLOGIQUES
DIRECTEUR DE THESE : CHANTAL GUIHENNEUC
ADRESSE : 4 AVENUE DE L'OBSERVATOIRE
75006 PARIS

TITRE DE LA THESE : FROM SEQUENCE TO CONSEQUENCE: A BAYESIAN NEURAL NETWORK APPROACH TO ELUCIDATE THE CAUSAL PATH FROM GENETIC VARIANT TO COMPLEX DISEASE

CO-ENCADRANT : MARIE VERBANCK
EQUIPE DU CO-ENCADRANT : UR 7537 — BIOSTM BIOSTATISTIQUE, TRAITEMENT ET MODELISATION DES DONNEES BIOLOGIQUES
LABORATOIRE : UR 7537 — BIOSTM BIOSTATISTIQUE, TRAITEMENT ET MODELISATION DES DONNEES BIOLOGIQUES

PRESENTATION DU SUJET

1 Scientific context: a massive paradigm shift in human genetics

With the initial sequencing and analysis of the human genome in 2001 [1] a new era began in the field of human genetics driven by Genome-wide association studies (GWASs). These new and unprecedentedly massive genetics data were to revolutionize the study of heritable complex diseases (*i.e.* caused by more than 1 gene) both in terms of elucidation of disease etiology and novel diagnostics and therapeutics options, up to the ultimate goal of personalized medicine. GWASs consist in testing the effect of genetic variants in a genome-wide manner on a single trait of interest. According to the GWAS catalog inventory [2], **5,975 publications and 421,875 unique genetic variant-trait associations** have been reported as of September 2022. Although GWASs still are successful in identifying robust associations between genetic variants and diseases, these associations have small effect sizes and the path from genetic variant to gene and ultimately to disease mostly remains unsolved. Instead, the field has experienced several paradigm shifts leading to the emergence of novel concepts and hypotheses.

The many faces of genetic regulation. Most identified genetic variants from GWAS are intergenic and were since shown to be involved in the regulation of the expression. The majority of the genome is transcribed into non-coding RNAs and the role of epigenetics through DNA and histone modifications which control chromatin structure and DNA accessibility is major.

Pervasive pleiotropy. The vast majority of variants are associated with more than one complex trait or disease, a phenomenon called pleiotropy which is widespread in the human genome and crucial to our understanding of the architecture of complex diseases [3].

Omnigenic architecture. Most associations with complex traits comprise a large number of variants distributed widely across the genome, conveying the highly polygenic nature of complex traits. The omnigenic model hypothesizes that complex traits are influenced by an infinitely large number of core and peripheral genes making small to infinitely small contributions.

As Research uncovers the genetic architecture of complex traits and apprehends the etiology of heritable diseases, new paradigms keep emerging revealing more of the complexity of biological models. Consequently, the field of genetics needs global **computational approaches** to **encompass the staggering complexity of variant effects on disease** based on a **substantial understanding of the complex biological structure of the genetics data**.

2 Research questions

This PhD project aims at providing methodologies to elucidate the link between genetic variants and complex diseases. The approach is based on three innovative features:

1. Joint modeling of a multicomponent disease. Metabolic syndrome (MetS) is defined by WHO as characterized by abdominal obesity, insulin resistance, hypertension, and hyperlipidemia which largely increases the risk of coronary heart disease, stroke, and type 2 diabetes [4]. There has been controversy as to whether MetS is a unique syndrome because it is defined as a cluster of interconnected factors that exhibit comorbidities and genetic correlation [5]. What is certain is that components of MetS are highly heritable diseases [6] and the leading causes of death worldwide.

2. Comprehensive omics regulation mechanisms. Our knowledge of genetic regulation mechanisms has considerably increased and has brought multiple levels of corresponding omics data which could be integrated to develop the most comprehensible approach.

3. Prediction task using a deep learning approach. The effect of the genetic variants on disease as well as on multiple intermediate omics phenotypes is available and has not been addressed yet as a prediction problem. Therefore, if our current knowledge of genetic regulation mechanisms is comprehensive enough, we could build a deep learning prediction model to predict the effect size of any genetic variant on a set of related diseases from omics data.

3 Methods

Multiple levels of omics data as well as biological knowledge can be used at several levels to predict the effect of genetic variants on metS.

3.1 Modular neural network guided by biological knowledge

Although the causal link between genetic variant and complex disease is not elucidated in a majority of cases, the estimation of effect sizes is highly consistent between studies (within similar ancestry). In the majority of elucidated variant-disease causal relationships, the effect of the variant goes through genetic or epigenetic control of gene regulation [7,8]. This could indicate that effect sizes could be predicted from regulatory omic features. Recent approaches have successfully predicted one body of omic data (e.g. transcription factor motifs, histone modifications, or chromatin opening) from sequence data using machine learning [10,11,9].

In this PhD project, we propose a higher-scale modular strategy, in the form of a modular neural network [12] (MNN) where biological knowledge will allow to build omics modules. A relevant module will be built to predict the "following" omic outputs (Figure 3.1). Each module will be trained sequentially and assessed in terms of overall disease effect prediction which is the final outcome. The interpretability of the model is guaranteed by the modular approach oriented by biological knowledge: the per-variant predictions will be interpreted in light of the modules to elucidate the variant-to-disease causal path. Tissue specificity will definitely be a challenge and several strategies could be considered such as greedy algorithms to prioritize relevant candidate tissues or biological knowledge to inform disease–tissue maps.

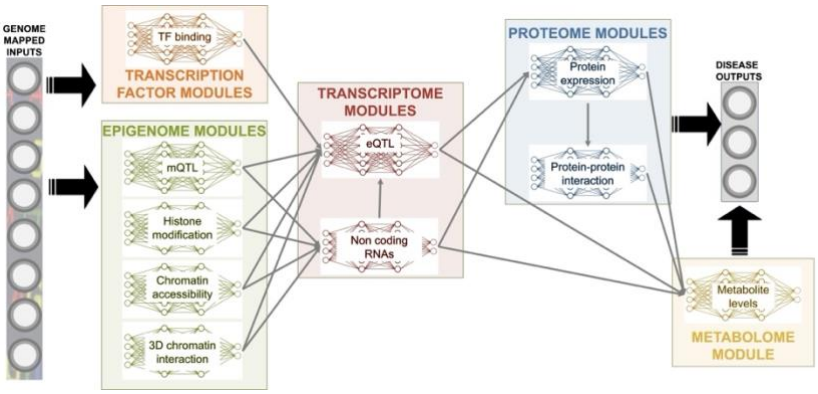


Figure 3.1: Biological-knowledge-oriented structure of the MNN

3.2 Bayesian inference and interpretable neural networks

Traditional neural network approaches approximate a function by learning a set of parameters (network weights) by minimizing a loss function regarding training samples; this often leads to overfitting as the number of model parameters can quickly explode. In Bayesian inference, instead of learning the parameter values, we seek to compute the conditional distribution of the weights given the training data. The Bayesian inference framework does not only mitigate the risk of overfitting but also enables us to gauge the level of uncertainty of our model regarding its parameters. To further enhance the interpretability of neural networks, methods have been developed to guide their architecture by restricting the connections between neurons to plausible connections only based on biological knowledge [13]. We will therefore explore the possibility of guiding the architecture of our modules using biological knowledge such as annotations. Finally, we could adopt a Bayesian approach by considering a given observed module as a realization from a parametric family of random modules to incorporate uncertainty in the neural network architecture [14].

4 Data used

This PhD project proposes to reuse publicly available data. This list is not final and will be evolving with the PhD project and the availability of novel data. GWAS summary statistics are the basis for the developed methods and will come from 2 main cohorts for replication purposes: UK Biobank [15] and FINNGEN [16]. In addition, different levels of information on top of **the genome** information will be mapped from various databases: **the transcriptome** including gene expression, splicing, and other non-coding RNAs such as miRNA, lncRNA or circRNA [18,17], potential human transcription factor genes [19], and phased allelic expression data [20]; **the epigenome** comprising DNA and histone modifications [22,21,23] as well as 3D chromatin interactions [24]; **the proteome** including protein levels and protein interactions [27,25,26]; **the metabolome** [30,28,29,31]; and **the phenome** [15].

5 PhD timeline

S1: Literature review & data collection from public databases.
S2: development and test of the basic modular neural network for variant prediction.
S3: writing & submission of article 1 / development and test of the Bayesian prediction framework.
S4: Bayesian prediction framework / writing & submission of article 2.
S5: article's revisions / start of thesis writing.
S6: articles' revisions, thesis writing & thesis review by the committee to prepare the defense.

6 Theme of the articles

Article 1: A Modular neural network integrating several levels of omics data to predict the effect of genetic variants on metabolic syndrome.
Article 2: Interpretable AI: a Bayesian approach to guarantee the interpretation of a multiomics neural network to predict the effect of genetic variants on metabolic syndrome.

References

[1] Lander, E. S.; Linton, L. M.; Birren, B.; et al. *Nature* **2001**, *409* (6822), 860–921.
[2] MacArthur, J.; Bowler, E.; Cerezo, M.; et al. *Nucleic Acids Research* **2017**, *45* (D1), D896–D901.
[3] Watanabe, K.; Stringer, S.; Frei, O.; et al. *Nature Genetics* **2019**, *51* (9), 1339–1348.
[4] Saklayen, M. G. *Current Hypertension Reports* **2018**, *20* (2), 12.
[5] Bulik-Sullivan, B.; Finucane, H. K.; Anttila, V.; et al. *Nature Genetics* **2015**, *47* (11), 1236–1241.
[6] van Dongen, J.; Willemsen, G.; Chen, W.-M.; et al. *Journal of Lipid Research* **2013**, *54* (10), 2914–2923.
[7] Bernstein, B. E.; Meissner, A.; Lander, E. S. *Cell* **2007**, *128* (4), 669–681.
[8] Nicolae, D. L.; Gamazon, E.; Zhang, W.; et al. *PLOS Genetics* **2010**, *6* (4), e1000888.
[9] Zhou, J.; Theesfeld, C. L.; Yao, K.; et al. *Nature Genetics* **2018**, *50* (8), 1171–1179.
[10] Agarwal, V.; Shendure, J. *Cell Reports* **2020**, *31* (7).
[11] Kelley, D. R.; Reshef, Y. A.; Bileschi, M.; et al. *Genome Research* **2018**, *28* (5), 739–750.
[12] Happel, B. L. M.; Murre, J. M. J. *Neural Networks* **1994**, *7* (6), 985–1004.
[13] van Hilten, A.; Kushner, S. A.; Kayser, M.; et al. *Communications Biology* **2021**, *4* (1), 1094.
[14] Zhang, Y.; Pal, S.; Coates, M.; et al. Bayesian graph convolutional neural networks for semi-supervised classification, 2018.
[15] Sudlow, C.; Gallacher, J.; Allen, N.; et al. *PLOS Medicine* **2015**, *12* (3), e1001779.

[16] Kurki, M. I.; Karjalainen, J.; Palta, P.; et al. FinnGen: Unique genetic insights from combining isolated population and national health register data, 2022, 2022.03.03.22271360.
[17] Consortium, T. G. *Science* **2020**, *369* (6509), 1318–1330.
[18] Abugessaisa, I.; Ramilowski, J. A.; Lizio, M.; et al. *Nucleic Acids Research* **2021**, *49* (D1), D892–D898.
[19] Lambert, S. A.; Jolma, A.; Campitelli, L. F.; et al. *Cell* **2018**, *172* (4), 650–665.
[20] Castel, S. E.; Aguet, F.; Mohammadi, P.; et al. *Genome Biology* **2020**, *21* (1), 234.
[21] Rosenbloom, K. R.; Dreszer, T. R.; Long, J. C.; et al. *Nucleic Acids Research* **2012**, *40* (Database issue), D912–D917.
[22] Kundaje, A.; Meuleman, W.; Ernst, J.; et al. *Nature* **2015**, *518* (7539), 317–330.
[23] Stunnenberg, H. G.; Abrignani, S.; Adams, D.; et al. *Cell* **2016**, *167* (5), 1145–1149.
[24] Wang, Y.; Song, F.; Zhang, B.; et al. *Genome Biology* **2018**, *19* (1), 151.
[25] Szklarczyk, D.; Gable, A. L.; Nastou, K. C.; et al. *Nucleic Acids Research* **2021**, *49* (D1), D605–D612.
[26] Uhlén, M.; Fagerberg, L.; Hallström, B. M.; et al. *Science* **2015**, *347* (6220), 1260419.
[27] Deming, Y.; Xia, J.; Cai, Y.; et al. *Scientific Reports* **2016**, *6* (1), 18092.
[28] Kettunen, J.; Tukiainen, T.; Sarin, A.-P.; et al. *Nature Genetics* **2012**, *44* (3), 269–276.
[29] Shin, S.-Y.; Fauman, E. B.; Petersen, A.-K.; et al. *Nature Genetics* **2014**, *46* (6), 543–550.
[30] Demirkan, A.; Henneman, P.; Verhoeven, A.; et al. *PLOS Genetics* **2015**, *11* (1), e1004835.
[31] Yin, X.; Chan, L. S.; Bose, D.; et al. *Nature Communications* **2022**, *13* (1), 1644.

PREREQUIS, FORMATION : DATA SCIENTIST WITH A BACKGROUND IN DEEP LEARNING & BAYESIAN STATISTICS AND KNOWLEDGE OF BIOMEDICAL APPLICATION.

CONTACT POUR CE SUJET : MARIE VERBANCK
EMAIL : MARIE.VERBANCK@U-PARIS.FR
TELEPHONE : 06 48 89 62 85

SPECIALITE DE LA THESE

Santé publique - Epidémiologie	☐
Santé publique - Epidémiologie clinique	☐
Santé publique - Epidémiologie sociale	☐
Santé publique - Epidémiologie génétique	☒
Santé publique - Biostatistique	☐
Santé publique - Biomathématiques	☐
Santé publique - Biostatistique et Biomathématiques	☒
Santé publique - Informatique médicale	☐
Santé publique - Imagerie biomédicale	☐
Santé publique - Bioinformatique	☐
Santé publique - Recherches sur les services de santé	☐
Santé publique - Economie de la santé	☐

ECOLE DOCTORALE PIERRE LOUIS DE SANTE PUBLIQUE A PARIS
EPIDEMIOLOGIE ET SCIENCES DE L'INFORMATION BIOMEDICALE

Directeur : Pierre-Yves Boëlle
Responsable pour l'Université Paris Cité : Isabelle Boutron

Santé publique - Science des données ☒

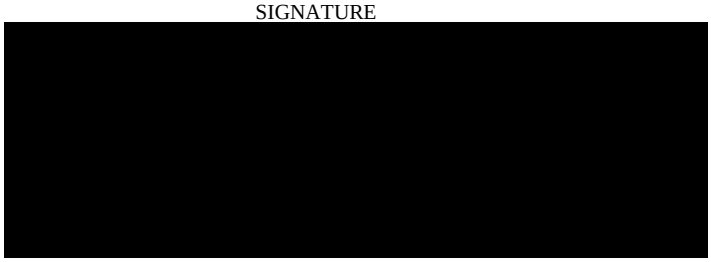
Santé publique – Prévention et promotion de la santé ☐

SIGNATURE DU. DE LA DIRECTEUR.TRICE DE THESE

VISA DU.DE LA DIRECTEUR.TRICE DU LABORATOIRE
(DEROGATION DE SIGNATURE NON ACCEPTEE)

AVIS FAVORABLE ☒

SIGNATURE



Ecole Doctorale 393
Centre Biomédical des Cordeliers
15, rue de l'Ecole de Médecine 75006 Paris
https://ed393.sorbonne-universite.fr/

Contact : ed393@sorbonne-universite.fr/Téléphone : 01.44.27.24.35