Leveraging Interpretable Machine Learning for Multiomic and Clinical Data Integration in Biomarker Discovery for Precision Medicine

Dissertation der Fakultät für Biologie
Ludwig-Maximilians-Universität München

Phong Ba Hung Nguyen

Munich, 2025

Leveraging Interpretable Machine Learning for Multiomic and Clinical Data Integration in Biomarker Discovery for Precision Medicine

Dissertation der Fakultät für Biologie

der Ludwig-Maximilians-Universität München

Phong Ba Hung Nguyen

München, 2025

Diese Dissertation wurde angefertigt
unter der Leitung von A.Prof. Michael Menden
im Bereich von Fakultät für Biologie
an der Ludwig-Maximilians-Universität München

Erstgutachter/in:	A.Prof. Michael Menden
Zweitgutachter/in:	Prof. Dr. Wolfgang Enard
Tag der Abgabe:	11.09.2024
Tag der mündlichen Prüfung:	12.03.2025

Erklärung

Ich versichere hiermit an Eides statt, dass meine Dissertation selbstständig und ohne unerlaubte Hilfsmittel angefertigt worden ist.
Die vorliegende Dissertation wurde weder ganz noch teilweise bei einer anderen Prüfungskommission vorgelegt.
Ich habe noch zu keine früheren Zeitpunkt versucht, eine Dissertation einzureichen oder an einer Doktorprüfung teilzunehmen.

München, den June 18, 2025

Phong Nguyen

Abstract

Precision medicine represents a pivotal advancement in healthcare, offering the potential to tailor treatments based on individual molecular, environmental, and lifestyle factors, thus improving patient outcomes and reducing adverse effects. However, significant challenges remain, including the underutilization of different modalities of data, data sparsity, issues with interpretability, and the need for sophisticated computational methods to analyze high-dimensional datasets. To address these challenges, this thesis introduces multiple advancements in methodology for various aspects of precision medicine, from preclinical to clinical studies. For preclinical studies such as drug high throughput screens of cancer cell lines, the stratification with inferred ancestry information enhanced biomarker discovery, demonstrating improved identification of drug response biomarkers. In addition, for observational clinical studies, integration of multiomic data within a biologically interpretable framework, providing an end-to-end comprehensive and transparent machine learning approach to biomarker discovery for complex metabolic diseases. Furthermore, for more complex data such as clinical longitudinal electronic health records, the utilization of pretrained large language models to develop an interpretable prognostic model for type 2 diabetes offered valuable insights into disease progression and patient management. Together, these proposed methods highlight the transformative potential of integrating advanced machine learning techniques and diverse data types to advance biomarker discovery in multiple aspects of precision medicine. Insights into disease mechanisms and actionable biomarkers discovered from these studies serve as valuable resources to help translate both biomedical research and healthcare practice and eventually benefit patients.

Acknowledgements

I would like to express my deep gratitude to several people who have helped me through this long journey.

First and foremost, I would like to thank my supervisor, A/Prof. Dr. Michael Menden, who has given me the opportunity to do a PhD, guided me through the earliest time of my research journey and supported me throughout my development.

In addition, I am grateful for my mentors and collaborators, Prof. Dr. Wolfgang Enard, Prof. Dr. Christian Herder, Prof. Dr. med. Reinhard Holl, Prof. Dr. Annette Peters, Dr. Sebastian Vosberg, Dr. Francesco Paolo Casale, for their invaluable feedback and support to make the projects happen.

Furthermore, my PhD journey would not be fulfilled without the presence and assistance from the fellow researchers in my lab, including Alex Ohnmacht, Ana Galhoz, Ginte Kutkaite, Dr. Ali Farnoud, Dr. Ines Assum, Christina Hillig, Dr. med. Daniel Garger, Martin Meinel, Dr. Diyuan Lu, Nikita Makarov, Maria Bordukova, Clara Meijs, Goeksu Avar, Anantharamanan Rajamani, and many other colleagues from the institute. All the professional meetings and social gatherings have made this a pleasant journey to me.

Last but not least, I am especially thankful for my dear friends Anubhab and Sayantini who are always there for me through good and tough times. I am also extremely grateful for my parents, Tung and Tu, for their unconditional care, and finally my sisters, Uyen and Nhung for their endless mental support throughout this time.

Publications

The publications associated with this work are the following:

Section 3.1: Nguyen, P.B.H.; Ohnmacht, A.J.; Sharifli, S.; Garnett, M.J.; Menden, M.P. Inferred Ancestral Origin of Cancer Cell Lines Associates with Differential Drug Response. *International Journal of Molecular Science*. 2021; 22(18):10135. https://doi.org/10.3390/ijms221810135

Section 3.2: Nguyen, P.B.H. et al. The Interpretable Multimodal Machine Learning (IMML) framework reveals pathological signatures of distal sensorimotor polyneuropathy. Accepted to *Communications Medicine*. Included here is the preprint version: bioRxiv 2024.01.04.574164; doi: https://doi.org/10.1101/2024.01.04.574164

Section 3.3: Nguyen, P.B.H. et al. Leveraging pretrained large language model for prognosis of type 2 diabetes using longitudinal medical records. To-be-submitted manuscript.

Declaration of contributions

Apart from chapter 3, the remaining work in chapters 1, 2 and 4 was not previously presented and has not been published. All used sources are acknowledged as references. The author list for chapter 3 is stated below. The shared-first authors are indicated with asterisks and confirmed their respective contributions.

- Phong B.H. Nguyen (P.B.H.N, author of this thesis)
- Alexander J. Ohnmacht (A.J.O.)*
- Samir Sharifli (S.S.)
- Mathew J. Garnett (M.J.G.)
- Michael P. Menden (M.P.M.)
- Daniel Garger (D.G.)
- Diyuan Lu (D.L.)
- Haifa Maalmi (H.M.)
- Holger Prokisch (H.P.)
- Barbara Thorand (B.T.)
- Jerzy Adamski (J.A.)
- Gabi Karstenmuller (G.K.)
- Melanie Waldenberger (M.W.)
- Christian Gieger (C.G.)
- Annette Peters (A.P.)
- Karsten Suhre (K.S.)
- Gidon J. Bonhof (G.J.B.)
- Wolfgang Rathmann (W.R.)
- Michael Roden (M.R.)
- Harald Grallert (H.G.)
- Dan Ziegler (D.Z.)
- Christian Herder (C.H.)
- Andreas Hungele (A.H.)
- Reinhard W. Holl (R.W.H.)

Section 3.1: Conceptualization, M.P.M.; Data curation, P.B.H.N. and A.J.O.; Formal analysis, P.B.H.N., A.J.O. and S.S.; Methodology, P.B.H.N., A.J.O., S.S. and M.P.M.; Supervision, M.J.G. and M.P.M.; Visualization, P.B.H.N. and A.J.O.; Writing—original draft, P.B.H.N.; Writing—review and editing, A.J.O., M.J.G. and M.P.M. All authors have read and agreed to the published version of the manuscript.

Section 3.2: C.H. and M.P.M. conceptualised the project. H.P., B.T., J.A., G.K., M.W., C.G., A.P., K.S., G.J.B., W.R., H.G. and D.Z. acquired and pre-processed the raw data. P.B.H.N. performed exploratory data analysis, developed the machine learning framework and visualised results. P.B.H.N., D.G., H.M., C.H. and M.P.M. derived biological interpretation. P.B.H.N. wrote the manuscript. M.P.M., C.H., M.R. and D.L. revised the manuscript.

Section 3.3: M.P.M. and P.B.H.N. conceptualised the project. A.H. and R.W.H. acquired and preprocessed the data. P.B.H.N. performed data analysis and developed the computational framework. P.B.H.N., M.P.M., A.H. and R.W.H. derived biological interpretation. P.B.H.N. wrote the manuscript. M.P.M. and R.W.H. revised the manuscript.

Melbourne, 05.09.2024

Michael P. Menden

Munich, 05.09.2024

Phong B. H. Nguyen

Penzberg, 05.09.2024

Alexander J. Ohnmacht

Contents

1 Introduction

The advent of precision medicine has revolutionized the approach to diagnosing, treating, and preventing diseases, offering personalized healthcare solutions tailored to the genetic, environmental, and lifestyle factors of individual patients (1). This paradigm shift is particularly transformative in the fields of oncology and endocrinology, where precision cancer medicine and precision diabetes medicine are at the forefront of clinical innovation. In cancer medicine, the focus on precision approaches has led to significant advancements in identifying genetic mutations and molecular alterations that drive tumor growth and progression. These insights have facilitated the development of targeted therapies and immunotherapies, improving treatment efficacy and patient outcomes (2). Similarly, in diabetes care, precision medicine is redefining disease management by elucidating the genetic and molecular underpinnings of different diabetes subtypes. This enables the creation of more effective, individualized treatment plans that address the specific pathophysiological mechanisms involved (3).

The availability and integration of multiomic and clinical data is pivotal in this context, providing a comprehensive view of disease biology that goes beyond single-dimension analyses (4). Besides, machine learning plays a crucial role in this integrative process, offering sophisticated tools to analyze and interpret complex datasets, identify potential biomarkers, and predict disease outcomes (5). These technologies are instrumental in bridging the gap between preclinical research and clinical application, ensuring that discoveries made in the lab translate into tangible health benefits for patients.

This chapter delves into the current status of precision cancer medicine and precision diabetes medicine, tracing the journey from preclinical studies to clinical applications. It explores the integration of multiomic and clinical data, the role of machine learning in biomarker discovery, and the challenges and opportunities that lie ahead in making precision medicine a reality. Through a detailed examination of these themes, the chapters highlights the profound impact of precision medicine on improving patient care and advancing medical science.

1.1 Current status of precision medicine

The idea of precision medicine could be dated back to the 19th century. Sir William Osler, one of the founders of modern medicine, said "It is much more important to know what sort of a patient has disease than what sort of a disease a patient has" (6), implying the importance of understanding the patient's physiological state (and all other information) in order to treat the disease effectively. Many times throughout nineteenth and twentieth centuries, medical research and practice had taken ideas that could fall under precision medicine paradigm. For example, Werner Kalow's 1962 textbook "Pharmacogenetics" had pointed out that therapeutic responses were associated not only to the biochemical properties of the agents but also to the genetic makeup of the patients (7). In 2015, the Precision Medicine Initiative was launched (1), emphasizing the need of utilizing relevant information of patients to guide personalized treatments and prevention strategies in complex noncommunicable diseases such as cancer and diabetes. Amount of research in the field has been increased significantly and, together with advancement of multi-omics profiling, big data analysis and computational power, benefited researchers, clinical practitioners, patients and their families significantly, paving a way towards translational medicine.

1.1.1 From "one-size-fits-all" to precision medicine

Medicine is a broad field of science with the primary goal of promoting human health through advancing diagnosis, prognosis, treatment and prevention of diseases (8). The practice of medicine could be dated back to the start of human history. Since ancient time, Chinese, Greek and Egyptian medical practitioners had been treating patients through empirical experience and observations of relevant symptoms. Decisions on treatments were usually based on the characteristics of the diseases and universal guideline applied to the whole population (9, 10). Fast forward to nineteenth and twentieth centuries, modern medicine was highly standardized in which medical education and practice strictly followed standard guideline which associated diseases' characteristics with intervention options (11). Most of the knowledge harnessed in these guidelines was driven by clinical trials and population studies which often yielded conclusions by averaging observations from a large heterogeneous group of patients (11). This approach had resulted in multiple achievements in medicine in the past century, especially in the context of infectious diseases and acute conditions (12). For example, successful vaccination programs in general population had led to the global eradication of smallpox (13).

Although this "one-size-fits-all" approach had undeniably contributed to the progress of medicine throughout history, it is not without limitations. In the context of chronic and complex diseases, it becomes more and more evident that a particular drug can only benefit a subset of patients, the others either do not exhibit any positive progression or more dangerously, show side effects (6). As we understand more about human physiology and pathology of the diseases, it is generally accepted that the development and treatment response of these diseases are associated with a multitude of factors including genetic variation, life style, social-environmental factors and the interactions of all these factors, all of which makes patient profiles vary among each other (6, 7, 14). Therefore, when these individual variations are neglected, treatments yield little success. Consequently, this leads to prolonged treatment time, increased cost and poor overall patient and doctor satisfaction. To tackle these obstacles, a more precise approach that tailors medical intervention to groups of patients that exhibit specific characteristics is employed, which is often referred to as precision medicine.

In contrast to the one-size-fits-all approach, precision medicine aims to adapt any medical decision to stratified groups of patients that differ from each other in certain characteristics (6). These characteristics could range from social-environmental factors such as age, sex, ethnicity, physical activity level and use of certain medications to genetic variations and altered protein signaling in the body. The emerging focus on precision medicine is relatively recent, however, some medical practices have existed for many years to some extent. For instance, people who need blood transfusion need to have their and their donor's blood type determined in order to find a suitable match (14). In this case, blood type serves as a genetic marker to stratify patients and have a crucially precise blood transfuse. By adapting medical intervention to patients' characteristics, precision medicine has improved the patients' life and overall satisfaction to the healthcare.

Several advancements in molecular biology and technology have empowered development of the field. Progress in this field has been accelerated recently and aligned with the understanding of our genome as well as the central dogma of biology (14). From the discovery of DNA structure in 1953 (15) to the first generation sequencing technology in 1977 (16) and the completion of Human Genome Project in 2003 (17), we have come a long way to understand our genetic code and the sophisticated link between genes and observed traits. It is now widely accepted that not all diseases follow Mendelian model in which one trait is only affected by one gene, but instead they are often caused by variations of multiple genes and gene (product) interactions, as well as a multitude of extrinsic factors (6). In addition, precision medicine is further enabled by technology advancement, which allows profiling of molecular data such as millions of genetic variations and tens of thousands of gene transcripts and proteins for a patient at a relatively short time and low cost. The deep molecular characterization, coupled with the patient's medical history and social-environmental factors, makes up a complete patient profile that helps discover biomarkers in research, design clinical trials to test drug efficacy, determine the risks of getting diseases, shapes decisions of interventions and disease management in clinical practice (6). All of this is nonetheless empowered by the ability to generate, store and manage big data, as well as interdisciplinary expertise of scientists from biomedicine, bioinformatics, computer science and data science (6). In sum, big efforts are being made to strengthen precision medicine, making it a truly translational paradigm with the ultimate goal of improving human health. Many of these efforts lie in the fields of oncology and diabetology.

1.1.2 Advancements in precision cancer medicine

Cancer is a multifaceted and challenging disease requiring diverse treatment approaches. It is defined by the uncontrolled growth, proliferation, and evasion of normal regulatory mechanisms in cells (18). Traditionally, cancer classification relies on tissue type, histology, and metastatic behavior, which form the basis for diagnosis and treatment (19). However, this limited information often results in therapies that target both healthy and cancerous cells indiscriminately, highlighting a critical need for more targeted approaches.

Traditional cancer treatments, while effective, often come with significant side effects. For decades, the first line of treatment has included cytotoxic chemotherapy, radiotherapy, and surgery, methods that either destroy or remove cancer cells. Unfortunately, these treatments can also damage healthy cells, leading to adverse effects that can be as severe as the clinical benefits (20, 21). This raises concerns about the balance between efficacy and harm, suggesting that these approaches, while useful, are not without significant limitations.

Lung cancer treatment, for example, exemplifies the limitations of traditional therapies. In 2004, the WHO classified lung cancer into small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC), further sub-classifying NSCLC into squamous cell carcinoma and adenocarcinoma (22). Standard treatment for advanced NSCLC has historically involved platinum-based chemotherapy, such as cisplatin and carboplatin, which effectively kill cancer cells by inducing DNA damage (22). However, these treatments also cause severe side effects, including nausea, fatigue, and immunosuppression, emphasizing the need for more refined therapeutic strategies (20, 21).

The evolution of cancer research marks a significant shift towards more targeted and personalized therapies. Advances in molecular biology have paved the way for next-generation cancer management approaches that move beyond the one-size-fits-all model.

These innovations, while promising, are still in development and highlight the potential of personalized medicine to reduce the harmful side effects of traditional treatments while improving overall patient outcomes.

Knowledge of molecular biology have been leveraged to advance cancer treatment for many years. One of the earliest examples is the 1960s' discovery of the association of chronic myeloid leukemia (CML) to the Philadelphia chromosome, which is a modified chromosome 22 caused by a fusion of ABL gene in chromosome 9 to BCR gene in chromosome 22 to create a fused ABL-BCR version of a tyrosine kinase that causes uncontrolled growth of white blood cells (23). This has led to the development of effective targeted therapies such as imatinib, which is a tyrosine kinase inhibitor (TKI) which targets the pathogenic ABL-BCR tyrosine kinase (24). This initial breakthrough has led to a series of relevant achievements in using TKIs to treat many types of cancers. Multiple other pathogenic mutations in NSCLC and melanoma have been discovered and the corresponding TKIs have been developed to target these genes and achieved remarkable clinical accomplishments (24). These successes reforms our definition and characterization of cancer and how we treat the disease leveraging molecular biology.

Cancer can nowadays be thought of as a complex genetic disease where one or more genetic and/or epigenetic alterations that together disrupts the gene expression patterns and subsequent processes, which ultimately leads to the oncogenesis phenotypes (25). In a healthy cell, some proteins work in promoting cell survival, growth and division (oncogenes) while others take care of DNA damage repair, cell cycle arrest and apoptosis (tumor suppressor genes), among others (24). All these proteins work together in highly regulated protein networks to ensure a healthy development of the cell. Cancer happens when genetic and/or epigenetic alterations cause either overexpression of oncogenes or under-expression of tumor suppressor genes, disrupting the regulated protein networks and predisposing the cell to cancer fate (24). In this context, associated mutations that truly drive the development of cancer are called driver mutations, the rest are called passenger mutations. That being said, the distinction between driver and passenger mutations is relatively dynamic due to the lack of necessary information (25).

Understanding the links between these molecular dysregulation and disease phenotype helps drive medical decision making and opens a new opportunity to tackle the disease. For instance, HER2 is a growth receptor that is often overexpressed in breast cancer, triggering downstream signaling pathways that leads to cancer cell proliferation (26). Genomic testing allows identification of HER2-positive patients who could be treated with therapies that specifically block the activity of this oncogene, such as trastuzumab, the first FDA-approved drug for treating HER2-positive breast cancer (26). In the case of NCSLC, an EGFR mutation is a strong biomarker that helps physicians determine whether they should treat the patient with TKIs such as erlotinib or standard chemotherapy (22).

Efforts have been made to advance molecular profiling technologies and further expand the biological understanding of cancer. Since the completion of the human genome (17) and the cancer genome atlas (TCGA) (27), the increasingly comprehensive molecular landscape of cancer has been revealed, with the discovery of many genes and their driver mutations involving in specific cancer development. These genes serve as potential targets for preclinical and clinical research for cancer biomarker discovery. Examples of such research are the Genomics of Drug Sensitivity in Cancer (GDSC) (28) and Cancer Therapeutics Response Portal (CTRP) (29) in which hundreds of compounds were screened against more than 1000 of cancer cell lines which were characterized for the common cancer mutations with the aim of identifying significant pharmacogenetic interactions between drugs and gene mutations. Findings from these projects help designing clinical trials which could potentially leads to new therapy for cancer treatment.

That being said, precision oncology is met with multiple challenges when being translated to clinical practice. Treatment resistance is a major drawback of many targeted therapies. Mechanisms of resistance vary from the heterogeneity of the tumor, tumor evolution to failure to account for the driver mutations in the complex protein network (30). A classic example is the EGFR T790M mutation, which is acquired in approximately 50% of patients previously treated with TKIs (22). Non-invasive liquid biopsy for monitoring the evolution of tumors and detecting resistance clones presents a potential solution to this (24). Furthermore, many germline and environmental-behavioral factors are often neglected in cancer research, leading to failure in clinical trials and treatments that are not generalizable to different groups of patients. Efforts have been made to account for these challenges and further advance precision oncology, benefiting patients.

1.1.3 Advancements in precision diabetes medicine

Diabetes has become a major burden to public health worldwide. In 2019, around 1.5 million people died of diabetes and another 460 000 people died of kidney complications of diabetes. Many of these patients were younger than 70 years (31). The mortality rates are not only high but also increasing at alarming speed. Between 2000 and 2019, the rates increased 3% on average, and up to 13% in lower-middle-income countries (31). The complexity of the disease together with many associated life-threatening complications has made diabetes one of the most urgent epidemics to tackle in the 21st century.

Diabetes is a chronic condition characterized by either lack of production of insulin or the produced insulin cannot be used effectively, leading to hyperglycemia (32). Since insulin is an essential hormone in glucose metabolism affecting virtually all organs in the body, the insulin deficiency causes many microvascular and macrovascular dysfunctions in the end organs (32). Common microvascular conditions associated with diabetes during the course of the disease are nerve damage (neuropathy), kidney damage (nephropathy) and eye damage (retinopathy), while involved macrovascular conditions include heart disease, stroke and peripheral vascular disease (33). The complexity and heterogeneity of these complications make management challenging.

When the body cannot produce insulin due to an autoimmune disease that targets insulin-producing beta cells in pancreas, the diabetes is called type 1 diabetes (T1D), a serious condition in which neither cause nor prevention strategy are fully comprehended (33, 32). On the other hand, type 2 diabetes (T2D) is a more common type of diabetes which is characterized by the inability of the body to use insulin properly, causing increased level of blood glucose and several downstream consequences to the system (33, 32). Since the mechanisms and risk factors of T2D are better understood and its prevalence are much more dominant than the former type, it is often the target for research in precision diabetes medicine.

Precision diabetes medicine provides a transformative approach to managing diabetes by tailoring treatments to individual patient profiles. By leveraging multidimensional data, it helps in making informed decisions for diagnosis, prevention, prognosis, and treatment (32, 34, 35). This method acknowledges the disease's inherent heterogeneity, which traditional, generalized approaches might overlook. A critical challenge in implementing precision diabetes medicine lies in the complexity of integrating such vast and varied data sources into actionable insights.

The heterogeneity of diabetes can be understood at various levels. At social-environmental level, patients may be categorized based on demographic details like age, sex, and ethnicity, as well as behavioral factors such as diet, physical activity, and alcohol consumption. For instance, older individuals with lower physical activity levels show a significantly higher risk of developing type 2 diabetes (T2D) (36). Although these social-environmental factors are helpful for broad categorization, they are often insufficient for accurately predicting disease trajectories due to their general nature. At a deeper level, molecular heterogeneity in diabetes presents an additional layer of complexity. Factors like genetic predispositions, gene expression levels, protein markers, and metabolic regulation also influence disease progression (34). For example, higher levels of serum inflammatory cytokines increase the risk of diabetic neuropathy (37). This molecular diversity necessitates a more nuanced understanding of the disease, though the challenge lies in making these biomarkers accessible and actionable in routine clinical settings.

The growing availability of large-scale data and advancements in medical science have enabled more precise stratification of patients. For instance, a recent study has used clinically relevant pathophysiological variables to segregate pre-diabetes patients into six distinct groups, each with different diabetes progression patterns (38). While this development marks a significant step forward in precision medicine, the effective translation of these insights into widespread clinical practice remains a major hurdle, requiring further refinement in data interpretation and integration into healthcare systems.

Precision approach to diabetes has gained attention in recent years. In 2018, for the first time the American Diabetes Association (ADA), in partnership with the European Association for the Study of Diabetes (EASD), launched the Precision Medicine in Diabetes Initiative (PMDI). Later the PMDI has also partnered with U.S. National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) and Juvenile Diabetes Research Foundation (JDRF). The aim of PMDI is to establish a consensus for the implementation of precision diabetes medicine in aspects including diagnosis, prognosis, prevention and treatment, leveraging expert opinions around the globe (35). The first report of PMDI in 2020 and the second in 2023 have provided detailed reviews of evidence and guideline to support the translation of precision medicine into practice in these four areas (32, 39).

One key area in focus is the precision diagnosis of diabetes, which is a set of methods to classify patients into distinct groups of different disease progression and treatment prognosis. It is conceptualized as a sequential process instead of a single diagnosis step: from assessing the epidemiology aspect (age, sex and ancestry) to observing clinical features and finally a diagnostic test to lead to a conclusion. This pathway helps involve multiple data modality in the diagnostics process instead of a single glycated hemoglobin (HbA1c) value. It is important to involve age, sex, ethnicity, pregnancy status and molecular profile into the diagnosis process (32). Insights gained from precision diagnosis drive the subsequent medical decision making. Many studies have stratified patients at diagnosis into categories of distinct progression and treatment prognosis either using supervised modelling and unsupervised clustering of clinical features, or clustering of genetic variants in case of T2D (40). Other studies utilized many data modalities in studying diabetes and its associated complications (41). Though having some limitations, these studies bring diabetes care closer to a translational precision care.

1.2 Multimodal data integration in precision medicine

The idea of employing multimodal data in studying complex diseases relies on the fact that these conditions interact with multiple factors in the system and environment. In the context of precision medicine, multimodal data analysis, which involves integrating multiple types of data in studying diseases, offers significant advantages over single modality approaches as it would increase the chance of pinpointing specific biomarkers valuable for diagnosis, prognosis and treatment of individual patients. With the current advancement in data generating and analysis technology, the availability of diverse types of data has been increasing, which has revolutionized the way researchers understand and address complex biological and medical problems.

1.2.1 Advantages of multimodal data analysis

While the central dogma of molecular biology often implies that there are certain levels of correlation between molecular entities in complex systems, there is a significant amount of discrepancy. For example, it has been known that the gene expression and corresponding protein abundance levels in multiple cell and tissue types are poorly correlated (42), making the sole use gene expression data to infer protein expression and functions challenging. The involvement of other molecular entities such as epigenomics and metabolomics further increases the complexity of the problem, as each of these entities are not only affected by intrinsic factors, but also independently interact with environmental and behavioral factors (43, 44). In clinical practice, patient stratification for differential treatment prognosis is important as patients who share similar clinical phenotypes might have different pathological mechanisms leading to differential response to interventions (45). Involving diverse types of data increase the insights into these mechanisms to identify the key biomarkers of the disease. Thus, multimodal data analysis offers a holistic understanding of complex biological processes and assist in biomarker discovery in precision medicine.

1.2.2 Types of data

Advanced omics technology has enabled the generation of large quantity of biological and clinical data. Together with the development of computational resources and analysis ability, these data has revolutionized biomedical practice. Each of these data plays a different role in how we understand complex diseases.

1.2.2.1 Genomic data

Often referred to as the blueprint of life, the genome harbours the DNA sequence of an organism, which encodes for information of all cellular activities (46). Variations in genome such as single nucleotide polymorphisms (SNPs), insertions, deletions, and larger structural variations might disrupt these activities, leading to dysfunction in metabolisms and ultimately pathogenicity. Thus, genomic data is a great source to understand the pathological signatures of diseases. For example, in cancer research, genomic sequencing of tumor samples can reveal somatic mutations in oncogenes and tumor suppressor genes, providing insights into the genetic alterations of oncogenicity and evolution, providing actionable targets for therapy.

With the decreasing cost and increasing throughput, next-generation sequencing (NGS) is becoming the standard genomic profiling technology in biological and clinical research (47). NGS technologies, such as short-read Illumina sequencing and single molecule Oxford Nanopore sequencing, have revolutionized genomics by enabling rapid and cost-effective sequencing of entire genomes, exomes or targeted panels (48).

1.2.2.2 Transcriptomic data

While genomic data provides the whole genetic information of the individuals, transcriptomic data captures the specific expression profile of each cell or tissue type, bridging the genetic codes and functions (49). Transcriptomics involves the study of RNA transcripts transcribed using the information encoded in the genome and the specific transcription machinery of the cell, reflecting which genes are actively being expressed at any given time, space and condition (49). Thus, transcriptomic data is invaluable for understanding how metabolisms changes in response to diseases, treatments, or environmental factors. For instance, in autoimmune diseases like rheumatoid arthritis, transcriptomic profiling can identify dysregulated immune response genes, shedding light on disease mechanisms and potential therapeutic targets (50).

RNA sequencing (RNA-seq) is the predominant technology used to generate transcriptomic data (49). RNA-seq provides a high-resolution view of the transcriptome, allowing researchers to quantify gene expression levels and link to physiological states. Modern RNA-seq technology such as single cell transcriptomics and spatial transcriptomics allow studying the heterogeneity of diseases and physiological states at high resolution.

1.2.2.3 Proteomic data

Proteomic data delves into the functional landscape of the cell by profiling proteins, the workhorses performing virtually all activities in biological systems. Proteomics encompasses the study of protein expression, modifications, interactions, and functions. In clinical settings, proteomic data can be instrumental in identifying biomarkers for disease diagnosis and prognosis (51). Mass spectrometry (MS) is the cornerstone technology for proteomics, enabling the simultaneous identification and quantification of thousands of proteins in biological samples. Another popular technology includes proximity extension assay which is especially useful when researchers want focus on a targeted panel of proteins for their research question.

1.2.2.4 Metabolomic data

Metabolomic data captures the chemical fingerprint of biological systems by profiling small molecules and metabolites, thus metabolome provides a snapshot of the metabolic state of a cell or organism, reflecting the interactions between the genome, transcriptome, proteome, and environment (44). In the realm of personalized medicine, metabolomic data can provide insights into individual metabolic profiles, enabling tailored therapeutic interventions. For instance, in diabetes research, metabolomic profiling can identify metabolic changes associated with insulin resistance, facilitating the development of personalized treatment strategies (52). Techniques such as nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS) are commonly used to generate metabolomic data.

1.2.2.5 Clinical data

Clinical data encompasses a wide range of information derived from patient records, including medical history, laboratory measurements, demographic data, treatment outcomes, and lifestyle factors. Electronic health records (EHRs) and clinical databases are primary sources of clinical data. This data type provides critical context for interpreting molecular and omics data, linking biological findings to patient health outcomes. Integrating clinical data with other omics data types can enhance the understanding of disease mechanisms and improve clinical decision-making. For example, in oncology, combining genomic and clinical data can help predict patient responses to specific therapies, paving the way for precision medicine.

1.2.3 Challenges in multimodal data analysis

While multimodal data comes with great potential in precision medicine, there are a few challenges that need to be addressed before we could harness these values. These challenges could stem from the underlying nature of these data to the practice of data generation and clinical usage. Awareness of these obstacles is an important step toward an effective integration strategy to leverage the power of these data.

Multimodal data encompasses diverse types of biological information, each with unique characteristics and formats. This heterogeneity complicates the integration process. Genomic data might be in the form of DNA sequences or SNP genotypes, while proteomic data could be in mass spectrometry readouts, and imaging data in pixel intensities. These differing formats and scales require standardization and normalization to ensure compatibility. Furthermore, the biological meaning of data points varies across modalities. For example, a genetic variant might have a different biological implication compared to an increase in the corresponding protein abundance. Thus, a framework that is aware of the heterogeneity and is able to extract the common biological signal associated with the phenotype of interest would be of great help.

Another challenge stems from the dimensionality. Different data types often have varying dimensions. Genomic data can involve millions of observed genetic variants while metabolomic data might only involve a few hundreds metabolites. This imbalance could lead to models becoming too tailored to the features of one data type and fail to attend all relevant features. Accounting for these discrepancies and select for the most important variables are important for an effective integration strategy. In addition, in many studies, the number of samples available for each data type can vary greatly. For instance, a study might have extensive transcriptomic data from thousands of patients but only limited proteomic data from a subset of these patients. This disparity can lead to biased analyses if not properly addressed.

Last but not least, missing data is a pervasive issue not only in multimodal data, but also in general biomedical datasets, arising from various sources such as incomplete sample collection, technical limitations, or data processing errors. Particularly, in longitudinal studies, patients may drop out, leading to missing data points at certain time intervals. Additionally, not all data types may be collected for every patient, resulting in gaps in the dataset. Deep understanding of the missingness patterns and proper data processing techniques are necessary to mitigate the issue.

1.2.4 Data integration strategies

While it is challenging when dealing with multimodal datasets, several strategies have been proposed to address these challenges and achieve the research goals.

Data pre-processing is undoubtedly one of the most important steps to guarantee the quality of the data and the reliability of downstream analysis result. Ensuring that data from different modalities are standardized and normalized is crucial for effective integration. Techniques such as z-score normalization, quantile normalization, and batch effect correction help make data from different sources comparable (53). Furthermore, to mitigate the impact of missing data, robust imputation methods (e.g., k-nearest neighbors, single value decomposition) (54) to estimate missing values could be employed. It is worth noting that the identification of the missing patterns and their significance is essential to determine the proper data processing technique (55).

To manage the high dimensionality and feature imbalance of multimodal data, dimensionality reduction techniques such as principal component analysis, t-SNE and matrix factorization (56) are frequently employed. Particularly, these techniques aim to map data points to a common latent space which is smaller than the original feature space and might contain meaningful biological signal of interest. In addition, embedded feature selection in machine learning algorithms such as LASSO and random forests are also leverage to select for the relevant features to the phenotype of interest which greatly reduce the feature size to assist downstream predictive modeling.

Data fusion is also a useful set of techniques that could be leveraged to deal with heterogeneous multimodal data. Data fusion could be in the early stage of modeling (concatenating raw features before model training), intermediate stage (combining processed features), and late stage (merging trained models). Each of these methods comes with advantages and disadvantages (57, 58). Thus, it is important to understand the data and the research hypothesis to choose an appropriate method.

Last but not least, ontologies and knowledge graphs are potentially great tools to facilitate data interpretation and interoperability. Particularly, ontologies enable the integration of data from different sources by providing a common framework for annotation and analysis of the data (59). For example, genomic and metabolomic data could be mapped to the same biological annotation using the biological signaling pathway ontology to help identify genetic variants and changes in metabolite levels associated with specific biological process. Similarly, knowledge graphs enable the integration of diverse data types, such as genomic, proteomic, clinical, and drug response data, by representing them as interconnected entities in a graph. This holistic view facilitates the analysis of complex biological questions (60).

In summary, leveraging combinations of these methods, researchers can fully leverage the power of multimodal data to gain comprehensive insights into biological systems and improve patient outcomes.

1.3 From preclinical to clinical research

So far we have talked about how precision medicine approach has transformed biological and medical research and the importance of leveraging multiple data sources for a holistic view of biological systems. That being said, the journey from bench to bed side encompasses several phases that are time consuming and require a wide range of expertise. From preclinical study to clinical trial, each phase is essentially important and consists of several challenges that need to be addressed before the application could benefit patients.

1.3.1 Evolution of clinical trials

Clinical trials, the systematic investigations to assess the efficacy and safety of medical interventions, have a long and fascinating history that dates back to ancient times. The journey of clinical trials from rudimentary experiments to the rigorous, controlled studies of today highlights the evolution of scientific thought and methodology in medicine.

The origins of clinical trials can be traced back to the earliest civilizations, where rudimentary forms of medical experimentation were practiced. Although these early trials lacked the formal structure and ethical considerations of modern studies, they laid the groundwork for the scientific method in modern medicine. For example, ancient biblical text recorded around 500 BC in Babylon indicates experiments and observations with different types of diets to promote overall health (61). While not clinical trials in the modern sense, these writings indicate an early understanding of the need to observe and record the outcomes of medical interventions. Fast forward to the Islamic Golden Age (8th to 14th centuries), Avicenna in his work “The Canon of Medicine” advocated for controlled experiments to test the efficacy of drugs, a concept that foreshadowed modern clinical trials (61).

The modern framework of clinical trials began to take shape in the 18th and 19th centuries. One of the earliest documented clinical trials was conducted by Scottish surgeon

James Lind in 1747, who demonstrated the effectiveness of citrus fruits in preventing scurvy among sailors. Particularly, he selected 12 sailors with similar symptoms and divided them into six groups, each receiving different treatments. Lind's careful documentation led to the identification of vitamin C as the cure for scurvy (61). This trial laid the groundwork for the scientific method in medical research. The 20th century saw significant advancements in clinical trial design and regulation. The establishment of randomized controlled trials (RCTs) by British statistician Sir Austin Bradford Hill in the 1940s revolutionized medical research by introducing randomization and control groups to eliminate bias and improve the validity of results, an example of which is Hill's trial evaluating the effectiveness of streptomycin in treating pulmonary tuberculosis in 1948 (61). The subsequent development of ethical and regulatory framework further enhanced the reliability of findings by preventing both researchers and participants from knowing who received the treatment or placebo (61).

Today, clinical trials are the cornerstone of medicine and health care, providing the critical evidence needed to translate scientific discoveries into effective treatments. These are highly regulated and meticulously designed to ensure the safety and efficacy of medical interventions. Regulatory agencies, such as the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA), oversee the approval and conduct of clinical trials. Modern trials often incorporate advanced technologies, such as electronic health records, genomics, and big data analytics, to enhance their precision and efficiency.

1.3.2 Challenges of clinical trials

One of the foremost challenges in clinical trials, particularly in the context of precision medicine, is patient stratification (62). Precision medicine aims to tailor treatments to individual genetic, molecular, and clinical profiles, which requires identifying and recruiting subgroups of patients who are most likely to benefit from specific interventions. This necessitates extensive genetic and molecular profiling, which can be both costly and time-consuming. Moreover, accurately classifying patients into these subgroups requires computational approaches and a deep understanding of disease heterogeneity (62).

Recruiting a sufficient number of participants who meet the inclusion criteria for clinical trials is another major hurdle (63). Trials often require large and diverse cohorts to ensure statistically significant results and to account for variability in treatment responses across different populations. Recruitment is further complicated by stringent eligibility criteria, which can limit the pool of potential participants. Additionally, retaining participants throughout the trial duration is challenging, as dropouts can skew results and reduce the statistical power of the study. In addition, clinical trials must adhere to strict ethical and regulatory standards to protect participants' rights and safety. This includes obtaining informed consent, ensuring participant confidentiality, and adhering to protocols approved by regulatory bodies such as the FDA and EMA. These requirements are essential but can also lengthen the time required to initiate and conduct trials.

Furthermore, the sheer volume and complexity of data generated in clinical trials pose challenges for data management and analysis (64, 65). Integrating diverse data types, such as molecular and clinical data, requires advanced bioinformatic infrastructure and expertise. Ensuring data accuracy, consistency, and security is critical, as errors or breaches can compromise the validity of the trial results.

With regard to the above challenges, the financial and temporal costs of conducting clinical trials are substantial. Trials can span several years and require significant funding to cover expenses related to recruitment, data collection, monitoring, and analysis (66). The high cost and lengthy timelines can be particularly burdensome for small biotech companies and academic researchers, potentially limiting the scope and scale of the trials they can undertake.

1.3.3 Roles of preclinical studies

Usually situated in the earlier phase of a therapy development, preclinical studies play a crucial role in assisting clinical trials by addressing many of the challenges faced by them. These models provide critical insights into disease mechanisms, drug efficacy, and safety before advancing to human clinical trials (67). By bridging the gap between basic research and clinical application, preclinical studies are essential for the successful development of precision medicine therapies. To integrate preclinical and clinical research, collaboration across disciplines, including genetics, molecular biology, pharmacology, and clinical medicine, is crucial (67). Advanced technologies such as high-throughput sequencing, bioinformatics and machine learning are also instrumental in analyzing complex data and identifying actionable insights (67).

Preclinical studies, using models such as cell lines, animals, and organoids, provide valuable mechanistic insights into how diseases develop and progress. These studies help identify molecular targets for new therapies, validate these targets, and understand the pathways involved (68). For example, in cancer research, high throughput drug screens of cancer cell lines coupled with deep molecular characterisation can reveal specific mechanisms driving tumor sensitivity and resistance, guiding the development of targeted therapies (28). In diabetes and cardiovascular disease research, mice have been widely used as in vivo models to study the effect of metabolic interventions to the systems in treating obesity, diabetes and other conditions (69). By establishing a solid scientific basis for potential treatments, preclinical studies reduce the risk of failure in subsequent clinical trials.

Gaining from the mechanistic insights, preclinical research aids in the identification of biomarkers that can be used for patient stratification and monitoring treatment responses. Biomarkers are measurable indicators of biological processes or responses to treatment, such as specific proteins, genes, or metabolites (70). By validating biomarkers in preclinical studies, researchers can develop companion diagnostics to identify patients who are most likely to benefit from a particular therapy. For example, identifying a genetic mutation that predicts response to a targeted cancer therapy can improve patient selection in clinical trials, enhancing the likelihood of success.

In addition, insights gained from preclinical studies can inform the design of clinical trials, optimizing their efficiency and effectiveness. For instance, preclinical data can guide the selection of endpoints, such as specific clinical outcomes or biomarker changes, that are most relevant for assessing treatment efficacy. Preclinical research can also help identify potential confounding factors and adverse effects that should be monitored during clinical trials. By refining trial design based on robust preclinical evidence, researchers can increase the likelihood of obtaining meaningful and actionable results.

Often time, before new treatments can be tested in humans, their safety and efficacy must be rigorously evaluated in preclinical models. Animal studies, for instance, are used to assess the pharmacokinetics (how a drug is absorbed, distributed, metabolized, and excreted) and pharmacodynamics (the effects of the drug on the body) of potential therapies (68). These studies help determine appropriate dosing, identify potential side effects, and establish initial efficacy. This preclinical evidence is critical for designing safe and effective clinical trial protocols, including dose escalation and safety monitoring strategies.

Lastly, by providing preliminary evidence of safety and efficacy, preclinical studies help reduce their costs and timelines. Early identification of ineffective or unsafe treatments can prevent costly late-stage trial failures. Additionally, preclinical studies can support the use of adaptive trial designs, which allow for modifications based on interim results, thereby increasing trial efficiency and reducing the time required to bring new therapies to market.

1.3.4 Limitations of preclinical research

While preclinical studies are crucial in the early phases of therapy development, they are not without significant limitations. One major challenge is that these models often fail to fully replicate the complexity of human biology. For instance, cell lines and animal models cannot entirely mimic the human tumor microenvironment, immune responses, or genetic diversity. As a result, the translation of findings from animal models to humans is often problematic due to species-specific differences in drug metabolism and disease progression (71). Moreover, the predictive power of preclinical studies is limited by their simplified nature, which may overlook critical aspects of disease mechanisms and drug interactions, resulting in overestimation of a drug's efficacy and underestimation of its toxicity, contributing to high failure rates in subsequent clinical trials (72). The lack of standardized protocols across different laboratories further complicates the reproducibility and reliability of preclinical data (73). These limitations underscore the need for more sophisticated models, such as patient-derived organoids or advanced computational simulations, to better capture human disease complexity and improve the translatability of preclinical findings to clinical settings.

1.3.5 Commonly used preclinical models

Preclinical models enable scientists to explore the biological underpinnings of diseases and evaluate potential therapies. These models allow for controlled experiments that would be unethical or impractical in human subjects. Among the various preclinical models, cell lines, mice, and organoids are the most commonly used (74). Each model has its own unique history, advantages, and limitations, contributing distinctively to the research landscape. As exemplified in Section 3.1, preclinical models such as cancer cell lines offer a simple yet useful means to evaluate drug response mechanisms and identify putative biomarkers. Furthermore, alternatives such as organoids and mouse models serve as complementary options with their own advantages and limitations. Understanding their characteristics is crucial for leveraging their potential effectively.

1.3.5.1 Cell lines

Cell lines have been a cornerstone of biomedical research since the mid-20$^{\text{th}}$ century (75). The first immortal human cell line, known as HeLa, was derived from the cervical cancer cells of Henrietta Lacks in 1951 (75). This groundbreaking development provided researchers with a stable, reproducible model for studying cellular processes and disease mechanisms. Since then, many more cell lines have been derived, representing various tissues and diseases of complex systems (76). Large projects such as The Cancer Cell Line Encyclopedia (CCLE) characterized and obtained mutation, gene expression, protein abundance and metabolite abundance of circa 1,000 cancer cell lines, providing invaluable resources for subsequent exploratory research (77). For example, given insights from CCLE, high throughput drug screens such as the Genomics of Drug Sensitivity in Cancer (GDSC) further assisted biomarker discovery by screening hundreds of compounds on the cancer cell lines and conducted hundreds of thousands of drug-gen association tests to identify putative biomarkers for further validation (28).

Cell lines offer several advantages. They are relatively easy to culture and maintain, allowing for high-throughput screening of drugs and genetic manipulations (78). The ability to generate large quantities of homogeneous cells makes them ideal for biochemical assays and molecular biology studies (79). Additionally, cell lines can be genetically modified to mimic specific disease conditions or to express particular genes, providing a versatile platform for research (80).

Despite their utility, cell lines have notable limitations. They often lack the complexity and organization of tissues in a living organism, which can result in oversimplified models of disease. For example, extra cellular matrix and immune cell interactions can usually not be accounted for in cell culture (81). In addition, prolonged culture can lead to genetic drift and phenotypic changes, potentially affecting the reproducibility and relevance of findings.

1.3.5.2 Organoids

Compared to cell lines, organoids are a relatively recent development in preclinical research. The concept emerged in the 2000s, with significant advancements made in the following decade (74). Organoids are three-dimensional structures derived from primary tissues or stem cells that self-organize into miniaturized versions of in vivo organs (74).

Organoids offer unparalleled advantages in replicating the architecture and functionality of human organs. They provide a more physiologically relevant model compared to traditional two-dimensional cell cultures, allowing for the study of organ development, disease mechanisms, and drug responses in a context that closely mimics human physiology (74). Organoids can be derived from patient-specific cells, enabling personalized medicine approaches and the study of genetic disorders (74). Moreover, organoids would be a more ethical options compared to animal models.

Despite their promise, organoids face several challenges. Their development and maintenance require specialized techniques and expertise, making them less accessible to all research labs. Organoids often lack the full complexity of in vivo organs, such as interactions with the immune system and vascularization. Additionally, the variability between organoid cultures can pose reproducibility issues.

1.3.5.3 Mouse models

Another commonly used preclinical model is mice. The use of mice in research dates back to the early 20th century. C.C. Little established the first inbred mouse strains in the 1900s, providing a genetically consistent model for studying hereditary traits and diseases (82). The development of transgenic and knockout mice in the 1980s further revolutionized the field, allowing for precise genetic manipulations.

Mouse models offer several advantages. Their genetic and physiological similarities to humans make them valuable for studying complex biological processes and disease mechanisms (83). The ability to manipulate their genome enables the creation of models that mimic human genetic disorders (84). For example, transgenic mice expressing the human amyloid precursor protein (APP) gene are used in Alzheimer's disease research to study amyloid plaque formation and test potential treatments (85). On the other hand, knockout mice lacking the p53 tumor suppressor gene provide insights into cancer development and the role of p53 in cell cycle regulation (86). Mice also have a short reproductive cycle and lifespan, facilitating the study of genetic inheritance and disease progression over generations (87).

That being said, there are limitations to using mouse models. Differences in metabolism, immune responses, and lifespan between mice and humans can affect the translation of findings. The ethical considerations of using animals in research also necessitate stringent regulations and oversight. Additionally, maintaining genetically modified mouse strains can be resource-intensive and costly.

1.3.6 Observational clinical studies

Beside clinical trials, another type of clinical research is observational clinical study, which is more practical than randomized control trials (RCTs) in many situations where RCTs are either unethical or studying rare conditions (88). These studies provide invaluable insights into the health and disease patterns of large groups of individuals, offering a real-world perspective that complements the controlled environments of preclinical and clinical studies. Observational clinical studies are research investigations where data is collected on individuals or groups without manipulating the study environment or interventions (88). Unlike RCTs, which introduce specific treatments or conditions to assess their effects, observational studies observe natural variations in exposures and outcomes among populations. These studies can be cohorts, which study incidence, cross-sectional studies, which study prevalence, or case-control studies, which compare groups retrospectively (88).

1.3.6.1 Roles of observational clinical studies

One of the most renowned observational studies is the Framingham Heart Study, initiated in 1948. This longitudinal cohort study has followed multiple generations of residents in Framingham, Massachusetts, to identify risk factors for cardiovascular disease. Key findings from this study have established the roles of hypertension, high cholesterol, smoking, obesity, and diabetes as major risk factors for heart disease, fundamentally shaping public health policies and preventive strategies (89). The Nurses' Health Study, which began in 1976, is another landmark observational study. It has tracked the health and lifestyle behaviors of over 120,000 registered nurses in the United States to investigate the long-term effects of diet, hormones, environment, and lifestyle on women's health. This study has provided critical insights into the links between diet and cancer, the effects of hormone replacement therapy, and risk factors for chronic diseases such as diabetes and cardiovascular disease (90). Today, the UK Biobank is a large-scale biomedical database and research resource containing in-depth genetic and health information from half a million UK participants. Launched in 2006, it aims to improve the prevention, diagnosis, and treatment of a wide range of serious and life-threatening illnesses. The vast amount of data collected, including genetic information, lifestyle factors, and health records, allows researchers to explore complex interactions between genetics and environment in disease development (91).

Observational studies reflect the complexities and variabilities of real-world settings, providing insights that are highly generalizable to broader populations. This contrasts with the often highly controlled environments of clinical trials, which may not fully capture the diversity of everyday clinical practice. In addition, these studies can investigate exposures that would be unethical to assign deliberately, such as smoking or environmental pollutants. They allow researchers to study the effects of such exposures on health outcomes without intervening in harmful ways (88). Furthermore, compared to RCTs where a relatively smaller group of homogeneous participants are recruited, observational studies often involve large sample sizes and diverse populations, enhancing the statistical power to detect associations and the ability to generalize findings across different demographic groups (92). Last but not least, prospective observational studies can follow individuals over long periods, providing valuable longitudinal data that can reveal trends and causal relationships over time. This is particularly useful for studying chronic diseases and long-term health outcomes where it would be challenging for clinical trials to assess (93).

1.3.6.2 Data types of observational clinical studies

Observational studies utilize a broad array of data types and sources to comprehensively capture health-related information from diverse populations. The richness and diversity of these data types allow researchers to explore various aspects of health and disease, leading to a more nuanced understanding of the factors that influence outcomes.

Among the types of data, demographic data forms the foundational layer of these studies, providing essential information on participants' age, gender, ethnicity, and socioeconomic status. These variables are crucial for understanding population diversity and stratifying analyses to identify disparities in health outcomes across different demographic groups (93). Clinical data is another vital component, encompassing medical records, diagnoses, treatment histories, and clinical outcomes. This data is typically obtained from healthcare providers and includes detailed information about patients' interactions with the healthcare system, such as hospital admissions, surgical procedures, and medication prescriptions. Clinical data allows researchers to track the progression of diseases and evaluate the effectiveness of various treatments (94).

Molecular data, including -omic data, which is increasingly available through advances in high throughput technologies, provides insights into the genetic predispositions as well as changes in the molecular landscape of individuals (4). Depending on the scale of the studies, this data type often includes information on single nucleotide polymorphisms (SNPs), gene expression profiles, epigenetic modifications as well as protein and metabolite abundance. This type of data is essential for studying the molecular basis of diseases and identifying potential biomarkers for risk stratification and personalized medicine.

Observational studies often include behavioral data, which captures information on lifestyle factors that influence health, such as diet, physical activity, smoking, and alcohol consumption (95). This data is often collected through self-reported surveys or digital health tools like wearable devices. Behavioral data helps researchers understand how lifestyle choices impact health outcomes and identify potential targets for public health interventions. Another important data type is environmental data, which includes information about external exposures that affect health, such as air pollution, geographic location, climate conditions, and occupational hazards (96). Environmental data is typically gathered from governmental or environmental monitoring agencies and can be linked to individual health records to assess the impact of environmental factors on disease incidence and progression (96).

1.3.6.3 Data sources of observational clinical studies

Electronic Health Records (EHRs) are a primary source of clinical data in observational studies. EHRs are digital versions of patients' medical histories maintained by healthcare providers, containing comprehensive information about patients' diagnoses, treatments, and outcomes (97). The widespread adoption of EHRs has significantly enhanced the ability to conduct large-scale observational studies by providing accessible, longitudinal clinical data.

Another common data source are biobanks, which are repositories that store biological samples, such as blood, saliva, or tissue, along with associated health information (98). These samples can be used for -omic analyses, providing a rich source of molecular data. Biobanks enable researchers to study the molecular basis of diseases and investigate the interactions between genes and environmental factors.

Surveys and questionnaires are traditional tools for collecting behavioral and demographic data directly from study participants. These instruments can be administered in person, by phone, or online, and they gather detailed self-reported information about participants' lifestyles, behaviors, and personal histories (99). Surveys are particularly useful for capturing data on behaviors and exposures that are not typically recorded in medical records. Administrative databases, maintained by health agencies, insurance companies, or other organizations, offer extensive data on healthcare utilization, costs, and outcomes. These databases include information on hospital admissions, outpatient visits, prescription drug use, and healthcare expenditures. Administrative data can be linked with clinical and demographic data to study patterns of healthcare access, quality, and outcomes on a large scale (100).

Last but not least, wearable devices and mobile apps represent a growing source of real-time behavioral and physiological data. These technologies track various health metrics, such as physical activity, heart rate, sleep patterns, and glucose levels (101). The continuous monitoring provided by wearables offers a granular view of health behaviors and physiological responses, enabling researchers to study the dynamic aspects of health and disease in everyday settings.

1.3.6.4 Limitations of observational clinical studies

Despite their many advantages, observational clinical studies face several limitations. Particularly, observational studies are susceptible to confounding, where external factors may influence both the exposure and outcome, leading to spurious associations (102). For example, a study might find an association between coffee consumption and reduced risk of a certain disease, but this could be confounded by other factors such as physical activity and life style (103). In addition, various types of bias can affect the validity of observational studies. Selection bias occurs when the study population is not representative of the general population, while recall bias arises when participants do not accurately remember or report past exposures or behaviors. Measurement bias can also occur if the methods of data collection are inconsistent or inaccurate (88). Furthermore, establishing causality in observational studies is challenging because these studies only observe associations, not direct cause-and-effect relationships. While longitudinal data can suggest temporal sequences, it cannot definitively prove causation without ruling out all potential confounding factors (88). Moreover, the quality and completeness of data in observational studies can vary. Missing data, inaccuracies in self-reported information, and inconsistencies in medical records can compromise the reliability of findings. Ensuring high data quality requires robust data collection methods and thorough validation procedures. Understanding these limitations and implementing mitigation strategies are keys to fully leverage the power of observational clinical data.

1.4 Machine learning for biomarker discovery

Biomarkers are measurable indicators of biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention (70). Biomarkers can be classified into different types based on their applications in medical practice: diagnostic, prognostic, and predictive biomarkers. Diagnostic biomarkers are used to detect or confirm the presence of a disease, enabling early and accurate diagnosis. Prognostic biomarkers provide information about the likely course or outcome of a disease, helping to predict disease progression and patient survival. Predictive biomarkers, on the other hand, are used to predict the response of a patient to a particular treatment, facilitating personalized therapy decisions (70). They are crucial for early disease detection, prognostication, and the development of targeted therapies, thereby improving patient clinical outcomes. The ultimate goal of leveraging multimodal datasets in preclinical and clinical studies is to discover novel biomarkers and facilitate therapy development in the age of precision medicine. However, the complexity of biological systems and the vast amount of data generated by modern high-throughput technologies pose significant challenges to traditional biomarker discovery methods, as discussed in sections 1.2 and 1.3. Meanwhile, machine learning has emerged as a powerful tool to address these challenges by leveraging data-driven algorithms that enables the inference of subtle patterns and relationships within complex biological datasets. That being said, challenges remain to be addressed before we could utilize the full potential of machine learning in biomarker discovery.

1.4.1 Introduction to machine learning

Machine learning is a subset of artificial intelligence (AI) that focuses on developing algorithms that enable computers to learn from and make inferences based on data. The core idea is to allow the machine to learn general rules and relationships, which is then used to make decisions or predictions without being explicitly programmed for specific tasks (104). For example, a machine learning model could learn the relationship between genetic variants to the risk of developing T2D, thereby assist clinical decision making during diagnosis and intervention (105). Among the many ways of classifying machine learning algorithms, one could considers the amount and type of human supervision these algorithms get during training, which results in four categories: supervised, unsupervised, semi-supervised and reinforcement learning (104).

Supervised learning involves using a known outcomes (labels) in the training dataset to develop models that can predict these outcomes (104). Depending on the nature of the label, supervised learning can be categorized into classification tasks, which predict categorical outcomes, and regression tasks, which predict continuous numeric values (104). Common supervised learning algorithms include linear regression, logistic regression, k-nearest neighbors, support vector machines, decision trees, and random forests. In contrast, unsupervised learning deals with unlabeled data, where the system identifies patterns without prior knowledge of outcomes, commonly performing tasks such as clustering (e.g. k-means, hierarchical clustering), dimensionality reduction (e.g. PCA, t-SNE, UMAP), and anomaly detection (104). Semi-supervised learning combines aspects of both supervised and unsupervised learning, handling datasets where only some samples are labeled. Typically, this involves clustering samples based on similarities (unsupervised) and then using the labeled data to inform predictions (supervised) (104). Finally, reinforcement learning differs from the other forms significantly, where an ”agent” learns optimal strategies (policies) to achieve goals by receiving “rewards” or “penalties” based on its actions during each iteration of action (104).

A popular subset of machine learning is the deep learning, relying on a model architecture called deep neural network (DNN), which mimics the structure and function of a biological neural network (106). The general architecture of a DNN consists of an input layer for feature inputs, an output layer for predicted outcomes, and several hidden layers in between to propagate and transform input signals (106). One significant advantage of DNNs is their ability to discern complex patterns across various data types, leading to high accuracy, especially when provided with large datasets (106). However, the nonlinear modeling of data and the complexity of the hidden layers make the interpretation of DNN predictions more challenging (107).

1.4.2 Basics of supervised machine learning in biomedical research

Supervised learning is the most commonly used machine learning approach in biomedical research due to its effectiveness in prediction tasks (104). In a general case, we have a data matrix $X \in \mathbb{R}^{m \times n}$ of n variables and m examples and a target variable $y \in \mathbb{R}^m$ representing the output to be predicted. The training dataset includes a set of input-output pairs (X_i, y_i) where X_i represents the features and y_i represents the labels of the sample i (104). An example of X and y are gene expression profiles and cancer types of patients, respectively.

A supervised model is mathematical representation that aims to map input features to the output labels, by learning the function $y = f(X; \theta) + \varepsilon$, where θ is the model parameters and ϵ is the noise, that describes the relationship between X and y (104). The model training is the process of learning the above function by minimizing the loss function $L(y, \hat{y})$, where $\hat{y} = f(X; \theta)$ is the prediction output of true outcome y (104). The common forms of L are mean squared error for regression task (y is a continuous variable) and cross entropy for classification task (y is a categorical variable). L is minimized through an optimization algorithm to estimate the parameters θ in which L is minimum.

The model $f(X;\theta)$ could come in many forms depending on the underlying relationships between the variables and the form of target variable, ranging from more simple models describing linear relationships between features and target such as linear regression and logistic regression to ones that could deal with more complex relationships such as support vector machine, decision trees and deep neural network (104). The choice of model type to use depends on many factors such as the size and complexity of the data, the computational resources and the research questions of interest. Oftentimes, one should experiment with different model conditions, or utilize an ensemble of models to achieve a desirable result.

1.4.3 Training a supervised machine learning model for biomarker discovery

Supervised models, trained on labeled datasets to predict sample outcome using its input features, is an effective way to learn the relationship between samples' features and their outcome. Through insights obtained from these relationships, we could pinpoint which features are most likely associated with the outcome, making them potential biomarkers of the study. The process from data to biomarker discovery is long and involves many aspects of machine learning that need to be carefully executed.

1.4.3.1 Data preprocessing

In biomarker discovery, data could be sourced from various sources. Common sources are omics technologies, including genomics, epigenomics, transcriptomics, proteomics and metabolomics. Genomic, epigenomic and transcriptomic data might come from variants of next-generation sequencing (NGS) technologies, providing information on gene variants and mutations, epigenomic processes and gene expression levels, respectively. Proteomic data, derived from mass spectrometry, offers insights into protein expression levels and modifications. Metabolomic data, obtained through techniques like nuclear magnetic resonance (NMR) spectroscopy and liquid chromatography-mass spectrometry (LC-MS), reveals small molecule metabolites in biological samples (108). Data could also be in form of clinical information. Clinical data from electronic health records (EHRs), population surveys and patient registries are also integral, offering context on patient demographics, disease states, and treatment outcomes.

The foundation of any supervised learning model is high-quality data. Once the data is collected, preprocessing is essential to ensure its quality and suitability for analysis. This step involves several tasks. First, missing data must be addressed, either through imputation techniques or by excluding incomplete records if appropriate. In this step, the pattern of data missingness must be taken into account in order to implement an appropriate measure (55). Next, data normalization or standardization is performed to scale the features, ensuring that variables with different units or magnitudes do not disproportionately influence the model. For example, gene expression data might be normalized using methods like log transformation or z-score normalization (53). Feature selection is another critical aspect, reducing the dimensionality of the dataset by retaining only the most relevant features. This is especially important due to the fact that the feature space is usually much larger than the sample space in most biological datasets. Techniques such as principal component analysis (PCA) or the least absolute shrinkage and selection operator (LASSO) can be employed to identify and retain significant variables. Furthermore, biology-driven methods such as gene set enrichment analysis could also be leveraged to select features that are both predictive and biologically relevant.

During data preprocessing, data visualization is a pivotal step in understanding the underlying patterns and distributions within the dataset. It helps in identifying potential outliers, trends, and relationships between variables. Visualization techniques such as PCA, t-SNE and UMAP can be used for dimensionality reduction, allowing the representation of high-dimensional data in two or three dimensions. This aids in visualizing the clustering of samples and the separation between different classes. Additionally, heatmaps can be employed to illustrate the expression levels of selected biomarkers across different samples, providing a clear visual representation of potential patterns.

1.4.3.2 Model training and evaluation

The training process of a supervised model involves several steps. Initially, the dataset is split into training and testing subsets, typically in a ratio of 80:20. The training set is used to build the model, while the testing set evaluates its performance. During model training on the training set, one common approach is cross-validation (109), which is employed to evaluate the generalizability and prevent overfitting of our model. This involves dividing the training dataset into k subsets and training the model k times, each time using k - 1 subsets as the training set and the remaining data as the validation set. The performance metric is averaged over all k trials and the final model, which is trained using the whole training set, is then tested with the testing set. For model training, the model architecture is defined based on the problem at hand. For biomarker discovery, common models include linear regression, logistic regression, support vector machines (SVMs), decision trees, random forests, and deep neural networks. Each method corresponds to specific learning algorithm and set of hyperparameters that need to be optimized during training. Common algorithms for hyperparameter optimization include grid search, where the model exhaustively searches through a manually specified subset of the hyperparameter space, random search, where the model randomly samples the hyperparameter space and Bayesian optimization, where the model uses a probabilistic model to select the most promising hyperparameters based on past evaluations.

During training process, the model parameters are optimized to minimize the loss function, which measures the difference between the predicted and actual outcomes. A commonly used optimization algorithm is gradient descent. The core idea behind gradient descent is to use the gradient (or derivative) of the loss function with respect to the model parameters to guide the direction of parameter updates. The gradient points in the direction of the steepest increase of the loss function, so moving in the opposite direction (down the gradient) will decrease the loss (110). A common variant of gradient descent are stochastic gradient descent, where the algorithm uses a small number of training examples to compute the gradients and update the parameters. Besides gradient descent, in the context of linear regression, ordinary least squares is the most common method for parameter optimization. It minimizes the sum of the squared differences between the observed and predicted values. Particularly, for the linear model $y = f(X; \theta) + \varepsilon$, the parameters θ can be estimated by $\hat{\theta} = (X^T X)^{-1} X^T y$ where $X^T X$ has to be invertible.

For biological data, dimensionality imbalance makes a significant challenge for building a good performing model. To mitigate the problem, regularization techniques, such as LASSO or ridge regression, may be applied to prevent overfitting, which occurs when the model captures very well signals from the training dataset but fails to generalize to the testing dataset. Other popular techniques to enhance the learning process include bagging, boosting and ensemble, which should be considered given specific research circumstances.

After training, the model's performance is evaluated using the testing dataset. Several metrics are used to assess the accuracy, precision, recall, and F1-score for classification tasks, or mean squared error (MSE) and root mean squared error (RMSE) for regression tasks. Cross-validation can provide a robust measure of performance. This helps in ensuring that the model generalizes well to unseen data and is not overfitted to the training set.

1.4.3.3 Biomarker discovery and validation

With a trained and evaluated model, the next important step is to interpret the model and identify potential biomarkers. In the context of supervised learning, this usually involves analyzing the feature importance of the model. For instance, in a random forest model, the importance of each feature is determined based on the reduction in impurity it provides (111). For example, presence of certain genetic mutations could stratify individuals into higher and lower risks of developing type 2 diabetes, in other words these mutational features segregate the dataset into two more homogeneous groups than the original one, making them important features of the model. Features with high importance scores are considered potential biomarkers, as they significantly contribute to the model's predictions. Similarly, in linear models, the magnitude of the coefficients can indicate the relevance of each feature. In general, lower complexity models such as random forest and linear models make it more straight forward to interpret and extract feature importance. On the other hand, deep neural networks are more powerful in learning complex patterns within the data at the expense of interpretation as it is more challenging to extract important features and biomarkers out of these models. In most cases, a post hoc interpretation procedure must be employed to analyse the prediction of such models. It is important to take these into account when choosing a suitable model for your research question.

Identified biomarkers must undergo rigorous validation to confirm their biological relevance and clinical utility. Validation involves both *in silico* and experimental approaches. *In silico* validation includes assessing the reproducibility of the findings in independent datasets and performing statistical tests to confirm the significance of the biomarkers. Experimental validation involves laboratory experiments such as quantitative PCR for gene expression, ELISA for protein levels, or targeted mass spectrometry for metabolites. Furthermore, validation studies may extend to clinical trials to evaluate the effectiveness of the biomarkers in predicting disease outcomes or treatment responses in a real-world setting.

It is essential to distinguish between association and causality when interpreting biomarkers. An observed association between a biomarker and a disease does not necessarily imply that the biomarker causes the disease. Establishing causality requires rigorous testing and validation. This often involves longitudinal studies, randomized controlled trials, and techniques such as Mendelian randomization, which uses genetic variants as instrumental variables to infer causal relationships (112). By carefully validating biomarkers through these methods, researchers can ensure that identified biomarkers are not only associated with but also causally linked to the biological processes of interest, enhancing their utility in clinical practice.

1.4.4 From interpretability to translatability

The integration of machine learning into biomedical research has ushered in a new era of precision medicine, particularly through the discovery of biomarkers. These biomarkers, which serve as measurable indicators of biological states or conditions, are pivotal in diagnosing diseases, predicting their progression, and tailoring therapeutic strategies to individual patients. However, the journey from biomarker discovery to clinical application is fraught with challenges, primarily related to the interpretability and translatability of machine learning models. It is crucial to find the balance between predictive power and interpretability in machine learning algorithms, thereby making precision medicine more translational.

Among machine learning algorithms, “shallow” learning models, such as linear regression, logistic regression, decision trees, and support vector machines, are typically more interpretable. These models allow researchers to easily understand the relationship between input features and the predicted outcomes. However, the simplicity of shallow learning models often comes at the cost of reduced predictive power, especially when dealing with complex, high-dimensional data typical of biomedical research (106).

Deep learning models, such as neural networks, convolutional neural networks (CNNs), recurrent neural networks (RNNs), transformer and recent advancements of large language models (LLMs) offer superior predictive performance by capturing intricate patterns and relationships within the data (113), at the expense of interpretability. These models are particularly effective in handling large, multi-dimensional datasets, making them ideal for omics and imaging data. Despite their predictive power, deep learning models are often criticized for being ”black boxes,” where the internal workings and the basis for their predictions are not easily explained (114). This lack of transparency can hinder the acceptance and application of these models in clinical settings, where understanding the rationale behind a decision is crucial. In addition, deep learning models are demanding in terms of amount of data and computational resources required to train, making them less accessible in many biomedical applications.

Despite the high complexity, deep learning models are not entirely black boxes, as several post hoc strategies have been proposed to enhance the interpretability of these models and bridge the gap between high predictive accuracy and practical clinical utility. One common approach is to compute feature importance scores, which provide insights into which features (or biomarkers) significantly influence the model’s predictions. Methods such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) are commonly used to calculate these importance scores (115). SHAP values, for example, quantify the contribution of each feature to the prediction by considering all possible feature combinations, offering a comprehensive view of feature impact (115). Besides deep learning models, SHAP values are also commonly used to interpret non-differentiable machine learning models such as boosted trees and random forest. For most deep learning models that are differentiable (ie. trained with gradient descent), another widely used technique is integrated gradient, which works by integrating the gradients of the model's output with respect to the inputs along a path from a baseline (typically all zeros) to the actual input. The integrated gradients for a feature indicate how much that feature contributes to the prediction, offering a more nuanced view of feature importance compared to standard gradients. By using integrated gradients, researchers can attribute the output of the model to the features in a manner that is consistent with the gradients, but more stable and less noisy (116). In models such as CNNs, RNNs and Transformers, attention mechanisms enable the model to focus on specific parts of the input data that are most relevant to the prediction task (113). By visualizing attention weights, researchers can identify which features or regions of the data are most critical for the model's decisions. Besides these post hoc analysis, simplifying complex models by using fewer layers or neurons can also improve interpretability. Although this may slightly reduce predictive accuracy, it can provide a better balance between performance and transparency. In summary, enhancing the interpretability of deep learning models through techniques like feature importance scores, attention mechanisms, and visualization tools is essential for their acceptance in clinical settings.

Ensuring the translatability of these models through robust validation and interdisciplinary collaboration is crucial for transforming research discoveries into tangible health benefits. By addressing these challenges, we can harness the full power of machine learning to make precision medicine more effective, personalized, and broadly applicable.

2 Aims of research

So far we have discussed many aspects of precision cancer medicine and precision diabetes medicine. Though representing a paradigm shift in healthcare, the field faces significant challenges, including the underutilization of multimodal data, data sparsity, issues with interpretability, and the need for advanced computational methods to analyze high-dimensional datasets. Overcoming these hurdles is essential to fully realize the potential of precision medicine and improve patient outcomes. This thesis introduces novel data analytic and methodological approaches with the aim to address these challenges:

- Aim 1: Leveraging inferred ancestral information to improve biomarker discovery in drug high throughput screens of cancer cell lines.
- Aim 2: Integration of multiomic and clinical data in an interpretable machine learning framework to identify prognostic markers of complex diseases such as distal sensorimotor polyneuropathy.
- Aim 3: Adaptation of pretrained large language models to sparse electronic health records for prognosis and prevention of type 2 diabetes and its complications.

First, to address the challenge of improving biomarker discovery in drug high-throughput screening using cancer cell lines, I derived a computational pipeline to infer the ancestry of the cancer cell lines then leveraged inferred ancestry information to enhance the accuracy and relevance of biomarker identification. By incorporating ancestry data, the genetic diversity was accounted for, which is often overlooked in traditional biomarker discovery approaches. This method not only improves the identification of biomarkers that predict drug response but also highlights the importance of considering genetic ancestry in precision medicine research. The findings underscore the potential for more personalized and effective cancer treatments by acknowledging and utilizing genetic diversity.

In addition, integrating multiomic data collected from sparse observational clinical data is essential to study complex processes in the field of precision diabetes medicine. Multiomic data provides a comprehensive view of biological processes. However, the integration of these heterogeneous data types poses significant analytical challenges. Here I introduce a novel machine learning framework that leverages gene set enrichment analysis to select biologically relevant features then combines these features in a robust integration approach. By doing so, the feature space is shrunk and mapped to the same biological space, facilitating prediction and interpretation. This integrative approach not only enhances the predictive power of the model but also provides insights into the underlying biological mechanisms, making the findings more applicable to clinical settings. The framework demonstrates that combining multiple data types in a biologically meaningful way can lead to more robust and interpretable biomarker discovery.

Last but not least, leveraging clinical electronic health records (EHRs) for pretrained large language models (LLMs) to build a prognostic and interpretable model for type 2 diabetes is an important yet challenging task. EHRs contain a wealth of information that can be leveraged for predictive modeling, but their unstructured nature, variability and sparsity pose significant challenges. Meanwhile, LLMs have demonstrated huge success in the language process field, posing a great potential for adaptation to other domains. Here I explored the utility of pretrained LLMs when adapting to EHRs for predictive modeling. I connected the two domains by customized data processing, prepending a learnable embedding to the LLM and fine-tuning both components using the EHR data. The model not only provides accurate prognostic insights for several important metabolic indicators but also maintains interpretability, allowing clinicians to understand the factors driving the predictions. This approach bridges the gap between complex data analysis and practical clinical application, demonstrating the potential of advanced machine learning techniques in enhancing patient care and management.

In summary, the approaches mentioned above collectively address critical challenges in precision medicine by leveraging advanced machine learning techniques and diverse data types in several biomedical research settings. Together, these approaches illustrate the transformative potential of precision medicine, paving the way for more personalized, effective, and interpretable healthcare solutions.

3 Results

This chapter shows the results of the three projects constituting this cumulative thesis and that were outlined in the previous chapter. Each project takes up one separate section in this chapter. Section 3.1 and 3.2 are peer-reviewed papers and section 3.3 is a preprint.

3.1 Inferred Ancestral Origin of Cancer Cell Lines Associates with Differential Drug Response

International Journal of
Molecular Sciences

MDPI

Article

Inferred Ancestral Origin of Cancer Cell Lines Associates with Differential Drug Response

Phong B. H. Nguyen [1,2,†], Alexander J. Ohnmacht [1,2,†], Samir Sharifli [1,3], Mathew J. Garnett [4] and Michael P. Menden [1,2,5,*]

[1] Helmholtz Center Munich, Institute of Computational Biology, 85764 Neuherberg, Germany; phong.nguyen@helmholtz-muenchen.de (P.B.H.N.); alexander.ohnmacht@helmholtz-muenchen.de (A.J.O.); samir.sharifli@tum.de (S.S.)
[2] Department of Biology, Ludwig-Maximilians University Munich, 82152 Martinsried, Germany
[3] Department of Mathematics, Technical University Munich, 85748 Garching, Germany
[4] Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton CB10 1SA, UK; mathew.garnett@sanger.ac.uk
[5] German Center for Diabetes Research (DZD e.V.), 85764 Neuherberg, Germany
* Correspondence: michael.menden@helmholtz-muenchen.de
† Equal contribution.

check for updates

Citation: Nguyen, P.B.H.; Ohnmacht, A.J.; Sharifli, S.; Garnett, M.J.; Menden, M.P. Inferred Ancestral Origin of Cancer Cell Lines Associates with Differential Drug Response. *Int. J. Mol. Sci.* **2021**, *22*, 10135. https://doi.org/10.3390/ijms221810135

Academic Editor: Ramón Cacabelos

Received: 23 August 2021
Accepted: 17 September 2021
Published: 20 September 2021

Abstract: Disparities between risk, treatment outcomes and survival rates in cancer patients across the world may be attributed to socioeconomic factors. In addition, the role of ancestry is frequently discussed. In preclinical studies, high-throughput drug screens in cancer cell lines have empowered the identification of clinically relevant molecular biomarkers of drug sensitivity; however, the genetic ancestry from tissue donors has been largely neglected in this setting. In order to address this, here, we show that the inferred ancestry of cancer cell lines is conserved and may impact drug response in patients as a predictive covariate in high-throughput drug screens. We found that there are differential drug responses between European and East Asian ancestries, especially when treated with PI3K/mTOR inhibitors. Our finding emphasizes a new angle in precision medicine, as cancer intervention strategies should consider the germline landscape, thereby reducing the failure rate of clinical trials.

Keywords: cancer; ancestry; high-throughput drug screen; biomarkers

1. Introduction

Pre-clinical studies in drug development can help to refine the target population and thus increase the success of clinical trials [1]. To this end, cancer cell lines are simplified and scalable models of human tumours, and enable the high-throughput exploration of pharmacogenetic interactions [2–4]. Among the largest efforts are the Genomics of Drug Sensitivity in Cancer (GDSC) project [2], Cancer Cell Line Encyclopedia (CCLE) [3] and Cancer Therapeutics Response Portal (CTRP) [4]. These efforts have screened hundreds of compounds across >1000 cancer cell lines in order to identify molecular biomarkers of drug response, thereby paving the way for precision oncology.

In the last two decades, the Cancer Genome Atlas (TCGA) [5] and International Cancer Gene Consortium (ICGC) [6] have pioneered the molecular characterisation of cancer patients. These efforts have revealed core cancer genes and their driver mutations, which are conserved in cancer cell lines (CCL) [7], and focusing on these somatic mutations assisted the identification of potential biomarkers [2]. On the other hand, germline variants have been known to influence the somatic mutational landscape of cancer tumours by changing the structures of genes and amino acid sequences, affecting the distribution of somatic mutations and causing global enrichment of mutations [8]. So far, many studies have revealed the direct effect of germline variants or their interaction with somatic mutations in predicting the outcome of patient treatments or sensitivity of cancer cell lines in high-throughput drug screens (HTSs) [9–11].

In contrast, genetic ancestry is mostly neglected in HTSs, although it is an established factor of risk, progression and response to treatment in several cancer types in the clinic [12,13]. Incorporating ancestry as an independent factor or covariate in drug response modelling may result in a more refined discovery of novel biomarkers, and enable us to model interactions with patient demographics.

In this study, we leveraged the drug sensitivity profiles of >1000 molecularly characterised CCLs across >400 drugs obtained from the GDSC project for revealing ancestry-dependent pharmacogenetic interactions. First, we inferred the ancestry of the cell line panel using a Bayesian method, using a list of 100 predictive single-nucleotide polymorphisms (referred to as ancestral SNPs) [14] for inference (Figure 1a), validated in an independent HTS (CTRP). In addition, by using the cell line ancestry, we subsequently inferred the HLA genotypes of the CCLs (Figure 1a). Lastly, we identified cell line ancestries that confer drug sensitivity, ultimately revealing patient subgroups stratified by their ancestry which may show differential responses to treatments in clinical trials (Figure 1b).

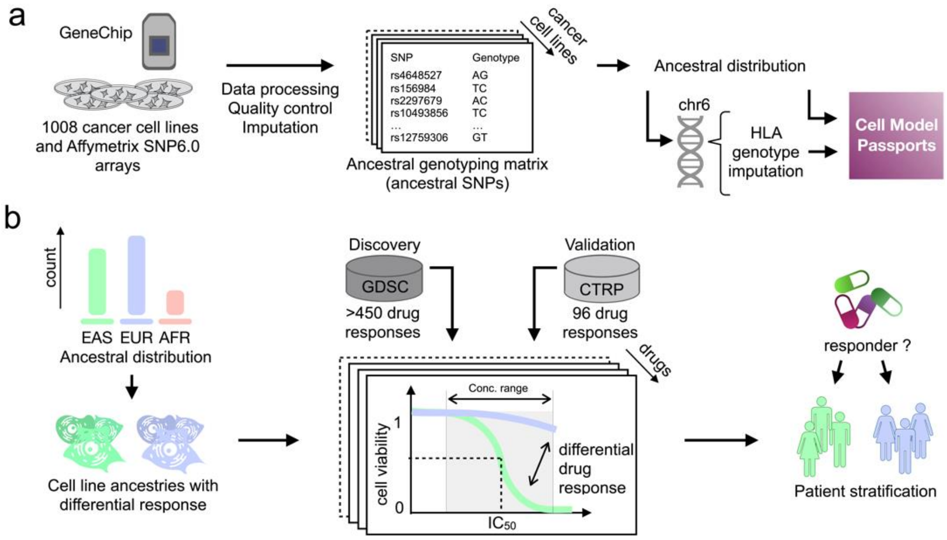


Figure 1. Ancestry inference pipeline and downstream analysis for differential drug response. (**a**) The analysis pipeline for retrieving an ancestral genotyping matrix by using Affymetrix SNP6.0 arrays from the GDSC CCLs, including data processing, quality control and imputation. The inferred ancestry and HLA genotypes are deployed on Cell Model Passports. (**b**) Workflow of pairwise ancestral comparisons of drug response in GDSC, validated with CTRP, generating hypotheses for patient stratifications in clinical trials based on demographics.

2. Results

2.1. Inferred Ancestry of Cancer Cell Lines Is Conserved

Using our processing pipeline (Figure 1a), the ancestral SNPs from the GDSC genotyping data were retrieved and imputed, which was then used to calculate the ancestral probabilities of these CCLs and reveal their ancestry origin (details in Methods). Our ancestry inference pipeline is publicly available and applicable for genetically characterised human cancer models of unknown origin.

The CCLs were classified in 25 subpopulations (Figure 2a), which stemmed from the ancestral origins defined by the 1000 Genomes Project (1000G) [15]. Since this level of stratification resulted in relatively small sample sizes, we summarised the subpopulations into five ancestries, namely European (EUR), East Asian (EAS), African (AFR), American (AMR) and South Asian (SAS) ancestries (Figure 2a). In this context, most cell lines were classified as EUR, followed by EAS ancestry. Particularly, EUR and EAS were assigned to 633 cells (63.1%) and 248 cells (25.0%), respectively, which together constituted the vast majority of the dataset

(88.1%). The other ancestries, AFR, AMR and SAS, only accounted for 56 cells (5.7%), 47 cells (5.4%) and 10 cells (0.8%), respectively (Figure 2a). The distribution of ancestries across cancer types was conserved in many cancer types (Figure 2b). For example, 69.69% of small cell lung cancer (SCLC) were EUR (46/66 cells), in contrast with only 7.58% AMR (5/66 cells). There were a few exceptions, however, in which the vast majority of cells in a cancer type were EAS. For example, 35 CCLs were derived from oesophageal carcinoma (ESCA) samples, of which 24 CCLs were classified into EAS (68.57%). CCLs were labelled according to TCGA classification, in which a significant number of cells could not be classified into any cancer type (183/994 cells, 18.41%) (Figure 2b).

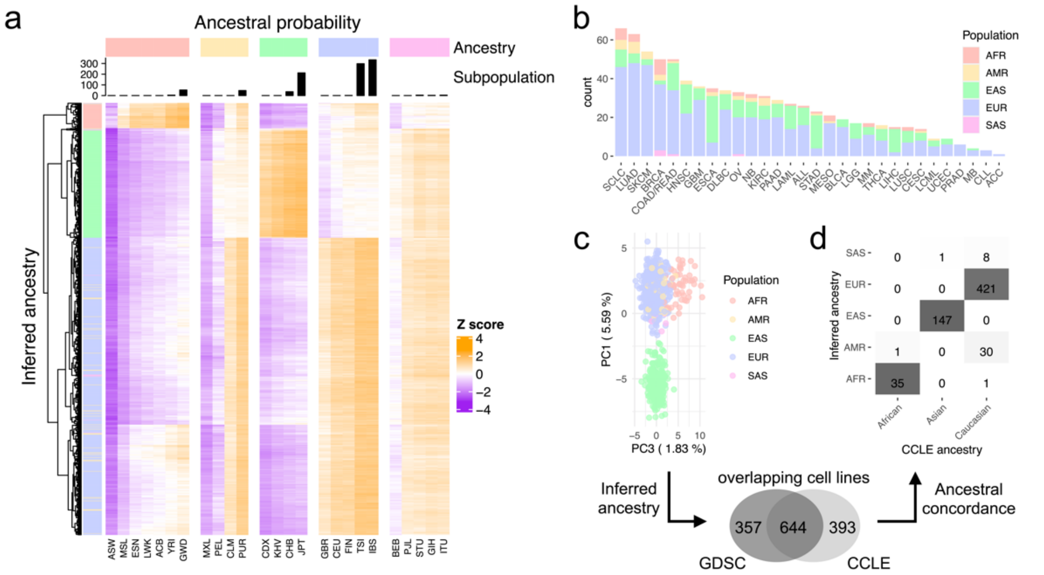


Figure 2. Inference of the ancestral distribution in cancer cell lines. (**a**) Heatmap of the *z*-scores for the inference of ancestry by estimating ancestral probabilities using a Bayesian method. (**b**) Distribution of inferred ancestries across cancer types. (**c**) PCA on the ancestral genotyping matrix. (**d**) Ancestral concordance between inferred ancestries of the GDSC and annotated CCLE CCLs.

A principal component analysis using the genotype matrix showed that the cells were clustered into three distinct groups, which correspond to AFR, EAS and EUR/AMR ancestries (Figure 2c), thus highlighting a stable prediction of ancestry in CCLs.

In order to examine the validity of our method, we compared the results with a similar dataset from the CCLE project. We found 644 cell lines overlapping between the two datasets, using the shared COSMIC and DepMap cell IDs. To make our data compatible with the CCLE data, we summarised both AMR and EUR cell lines as Caucasian and EAS and SAS cell lines as Asian in the annotated GDSC set. As expected, the inferred ancestry showed very high concordance with the referenced ancestry by the CCLE. Particularly, 634 out of 644 cell lines were consistently annotated (98.4%). From the remaining ten cells, nine were CCLE Caucasian that were classified as SAS (eight cells) or AFR (1 cell) and one was an African cell line that was classified as AMR in GDSC, set by our pipeline (Figure 2d). Although based only on a subset of the data, the extremely high concordance of our results with CCLE data proved the accuracy of our analysis pipeline.

Furthermore, using the inferred ancestry information, we managed to predict the HLA genotypes of the CCLs using the HiBAG method [16] with high accuracy (Supplementary Table S1). We validated the result with 56 NCI60 cell lines; the accuracy was up to 85.7% (allowing one mismatched allele for the haplotype consisting of six loci in MHC class I and II: A, B, C, DRB1, DQB1 and DPB1). Together, ancestries and HLA genotypes of CCLs will be imported to Cell Model Passports, a catalogue of CCL annotations that serve as potential features in pharmacogenomic studies (Figure 1a).

2.2. *Differential Drug Responses between Asian and Caucasian Cancer Cell Lines*

Next, we investigated CCL ancestries which confer drug sensitivity, which may be leveraged for selecting target cohorts based on demographics. The drug response data from the GDSC project served as our discovery cohort, and CTRP was used for validation. We performed one-way ANOVA tests for investigating imbalances in drug responses across two pairwise ancestries. Comparing the largest two populations, i.e., Asian and Caucasian (Figure 1b; Methods), revealed 59 significant associations between ancestry and drug response in total across nine cancer types (Figure 3a & Supplementary Table S2, <20% FDR). Comparisons of Asian and African, as well as African and Caucasian ancestries are included in the Supplementary Materials (Supplementary Tables S3 and S4).

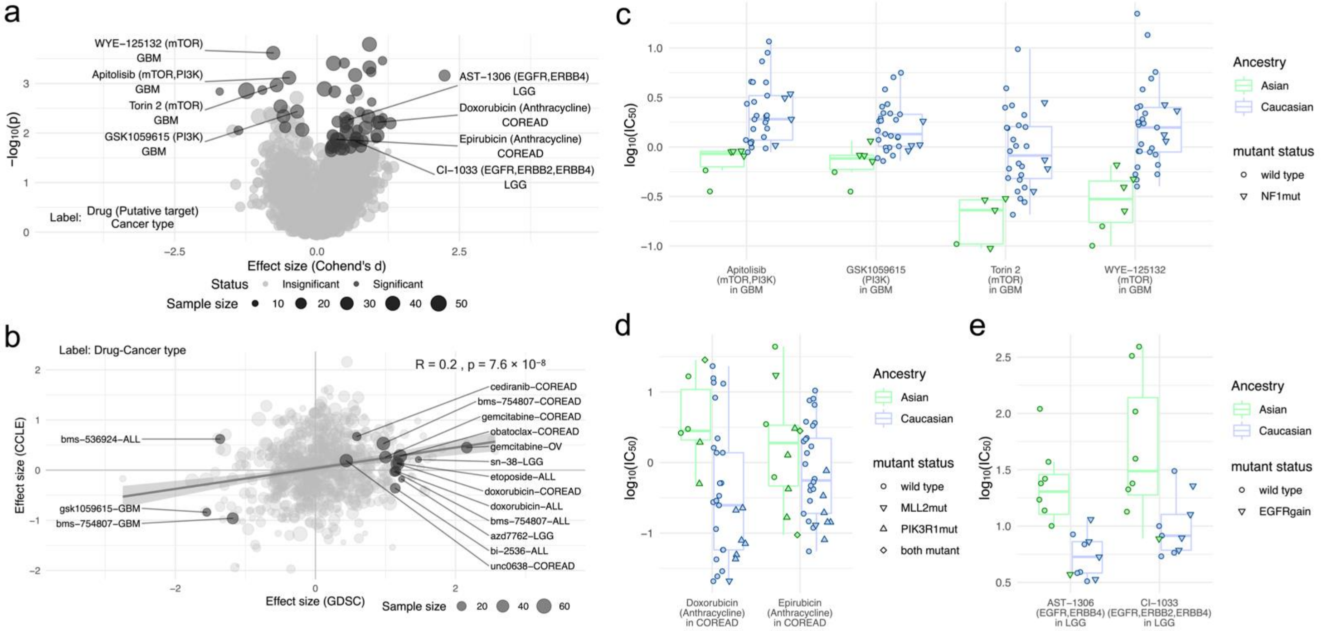


Figure 3. Differential drug responses between ancestral origins. (**a**) Associations between Caucasian and Asian ancestry and drug response across cancer types (<20% FDR). (**b**) Comparison of effect sizes between GDSC and CRTP. Highlighted are significant associations in GDSC, which were overlapping with CTRP. Here, the exemplified associations are higher drug sensitivity of (**c**) Asian CCLs to PI3K/mTOR inhibitors in GBM. (**d**) Caucasian CCLs to anthracyclines in COREAD and (**e**) Caucasian CCLs to TKIs in LGG.

For validation, we investigated the overlapping screens between GDSC and CTRP (Figure 1b; Methods). Focusing on the fraction of associations between ancestry and drug response that were significant in GDSC (<20% FDR), we observed that 11 out of 16 associations from our discovery cohort displayed a consistent sign of effect size (Figure 3b). Four out of the five inconsistent validation experiments were found in acute lymphoblastic leukaemias (ALL). Across all performed tests, the Pearson correlation for effect sizes of the overlapping compounds between GDSC and CCLE was R = 0.2 (Figure 3b), suggesting reproducible associations in independent experiments.

Among the top significant associations, Asian cell lines showed higher sensitivity to PI3K/mTOR inhibitors, especially in glioblastoma (GBM). Specifically, out of all 12 associations in GBM, the cell lines were significantly sensitive to apitolisib (Cohen's $d = -1.73$, adj. $p = 0.06$), GSK1059615 (Cohen's $d = -1.55$, adj. $p = 0.11$), torin 2 (Cohen's $d = -1.87$, adj. $p = 0.06$) and WYE-125132 (Cohen's $d = -1.91$, adj. $p = 0.04$) (Figure 3c). Furthermore, the targets of those drugs (PI3K/mTOR) were enriched among Asian-sensitive associations in GBM (mTOR: adj. $p = 0.0004$, PI3K: adj. $p = 0.01$; Methods). Noticeably, the majority of these inhibitors (three out of four) target only mTOR or a combination of mTOR and PI3K, whereas GSK1059615 only targets PI3K.

Associations in which Caucasian CCLs were found to be more sensitive accounted for 47 out of 59 total significant associations (Supplementary Table S2). In fact, 16 out of 59 significant associations were found in colorectal adenocarcinoma (COREAD), all of which were found to be sensitive in Caucasian CCLs. We found that Caucasian CCLs are more sensitive to the two anthracyclines doxorubicin and epirubicin (Figure 3d). This type of drug is enriched among all significant associations in COREAD (adj. $p = 0.09$).

Furthermore, we identified that Caucasian CCLs in low-grade glioma (LGG) were more sensitive to irreversible tyrosine kinase inhibitors (TKI) targeting EGFR, ERBB2 or ERBB4, such as AST-1306 (Cohen's $d = 1.71$, adj. $p = 0.14$) and CI-1033 (Cohen's $d = 1.43$, adj. $p = 0.17$) (Figure 3e). Remarkably, other screened TKIs such as pelitinib (Cohen's $d = 1.17$, adj. $p = 0.25$) and PF-00299804 (Cohen's $d = 0.98$, adj. $p = 0.25$) showed similar trends but did not pass our set FDR threshold of 20%. Interestingly, all CCLs with a copy number gain of EGFR showed sensitivity to TKIs independent of ancestry, but here we can reveal that other Caucasian CCLs with wild-type EGFR respond better than their Asian CCL wild-type counterparts.

Somatic driver mutations are commonly investigated as potential drug response biomarkers. It is likely that the frequency by which somatic mutations are observed in patients can be dependent on their ancestry. Consequently, we screened for enrichments of high-confidence cancer genes in CCLs in either Asian or Caucasian ancestry CCLs (Methods). We only found a handful of enriched cancer genes, namely for Asian CCLs; NF1 mutations are more abundant for GBM (adj. $p = 0.05$; Figure 3c) and mutations in MLL2 or PIK3R1 are more prevalent for COREAD (adj. $p = 0.26$ and adj. $p = 0.21$, respectively; Figure 3d). However, none of these mutations explained the ancestry-dependent variability of drug responses.

3. Discussion

Drug approval agencies are bound by demographics, e.g., the EMA in Europe or the FDA in the USA, and have an undeniable impact on pharmacology [17]. In order to estimate its impact on drug response, we predicted ancestry in cancer cell lines and showed a differential drug response in high-throughput drug screens. We implemented an efficient Bayesian ancestral inference which utilized ancestral genotype frequencies of SNPs and population weights in the 1000G project to successfully classify cell lines into ancestral populations, demonstrating the possibility to infer missing ancestral information in published data for both patients and cancer cell lines, even with sparse input data.

In general, there is consistency in the distributions of ancestries across cancer types, with a few exceptions. The distribution of ancestries in CCLs may reflect the demographic differences in incidence and prevalence of cancer among ethnic groups, thereby influencing the selection of CCL models. For example, a few studies have shown that Asian populations have significantly higher incidence and prevalence of ESCA and STAD as compared to Caucasian [18–21], which is consistent with our findings.

Strikingly high concordance with published data from the CCLE project supported the validity of our ancestral results and the whole analysis pipeline, building a platform for subsequent analysis and future studies.

A univariate ANOVA analysis was conducted to assess whether ancestry can affect drug sensitivity, especially between Asian and Caucasian CCLs. The results reveal some drugs for which Asian CCLs showed higher sensitivity than Caucasian CCLs. Among the most significant associations were inhibitors targeting PI3K/mTOR signalling in GBM, especially those targeting mTOR. This was consistent with a past study involving clinical trials in solid tumours, concluding that Asian patients suffered from more severe toxicity when treated with PI3K/mTOR inhibitors than European patients who were given similar doses in solid tumours [22]. However, the direct impact of ancestry on the sensitivity of PI3K/mTOR inhibitors and the molecular mechanisms that drive the observed differential drug sensitivity have been hardly studied so far. The response of CCLs with Asian ancestry is consistent with previous studies reporting a high susceptibility of Asian patient-derived cell lines to combination treatment with nimotuzumab (EGFR inhibitor) and rapamycin (mTOR inhibitor) in GBM, which was found to be independent of their EGFR status [23].

In addition, previous studies have shown that East Asian patients are more sensitive to EGFR inhibitors, due to an observation that Asians have a higher mutational frequency of EGFR compared to Caucasians [24]. In contrast, amplifications in EGFR have been found to be more prevalent in Caucasians for cancer types such as non-small cell lung cancer [25]. Accordingly, we found that TKIs targeting EGFR, ERBB2 or ERBB4 conferred sensitivity in Caucasian CCLs in LGG. However, sensitivity was also found for Caucasian CCLs with wild-type EGFR, which is not reported thus far.

We further observed a significant number of Caucasian CCLs in the cancer types COREAD, LGG and ALL, which were more sensitive to various drugs compared to Asian CCLs, but these associations showed no enrichment of specific signalling pathways and putative drug targets. Many studies have shown the ethnic differences in the incidence and survival rates of these cancers, especially between Asian and Caucasian patients [12,13,26,27], and a few studies reported a significantly higher toxicity response to chemotherapy in COREAD Caucasian patients than Asian ones [28,29], but the cause of the differential response to targeted therapies which lies under differences in molecular profiles is still yet to be discovered by more comprehensive molecular investigations.

A limitation of this study lies in the fact that human CCLs remain simplified models which do not capture the full complexity observed in tumours, e.g., the tumour microenvironment, clonal heterogeneity or immune responses. Despite the conserved ancestral origins of CCLs, biological processes within the tumour microenvironment may differ in vivo [7], thus often hampering the generalisation of the results to patients. In addition, CCL models lack patient environments and lifestyle factors, which also can influence the sensitivity to cancer treatments [30]. It would be desirable to explore differences within subpopulations; however, we lack the statistical power due to reduced sample sizes. Thus, differential drug sensitivity analyses of subpopulations may become feasible with additional data releases in the future. Nevertheless, our findings suggest that the impact of ancestry can be partially modelled in vitro. We present a resource for ancestry and HLA subtypes of CCLs, which are shared via Cell Model Passports. This enables in vitro pharmacogenomics analyses considering demographics. In addition, we anticipate that the HLA subtype definition will become an important feature in upcoming CCLs and lymphocyte co-culturing HTS, which are currently pursued for novel immunotherapies. In summary, this study successfully elucidated the distribution of ancestries in the selection of cancer cell lines using an efficient inference pipeline and subsequent differential drug responses to PI3K/mTOR inhibitors and TKIs in GBM and LGG, respectively. We believe that this resource and subsequent findings may shape the next generation of algorithms to identify biomarkers in HTSs.

4. Materials and Methods

4.1. Data Availability

The Affymetrix SNP6.0 arrays genotyping dataset contains 1007 cancer cell lines and 884,148 SNPs ranging from chromosome 1 to chromosome 22, which are deposited in the European Genome-Phenome Archive (EGAS00001000978). For inferring ancestry, we leveraged the set of 100 ancestral SNPs from Sampson et al. [14]. Using synonyms data from Ensembl Biomart, we retrieved the genotypes of 26 ancestral SNPs, and imputed the remaining 74 SNPs.

4.2. Quality Control

CSV files from the Genomics of Drug Sensitivity in Cancer (GDSC) database were transformed into text (PED and MAP) and binary file sets (BED, BIM and FAM) for each chromosome using PLINK [31]. First, SNPs that had a missing rate higher than 10% and a minor allele frequency (MAF) less than 0.05 were removed. Next, positions of SNPs were compared to SNP coordinates of a legend file of the same build retrieved from the

1000 Genome Project (1000G) database (https://www.internationalgenome.org/, accessed on 20 February 2020) [15] and mismatched SNPs were removed. Finally, SNPs were checked for potential swapping of the reference strand, alternative alleles and strand flipping, and the mismatched SNPs were removed.

4.3. Phasing and Imputation

The quality-controlled binary file sets were converted to VCF files using PLINK. A reference genome was downloaded from the 1000G database in the form of phased VCF files and converted to BCF using BCFtools [32] and M3VCF using Minimac3 [33]. The VCF file inputs were phased using Eagle2 [34], provided the BCF reference files and genetic map from the 1000G. The phased VCF files were then imputed using Minimac4 [33] to fill the missing SNPs in each chromosome, provided the M3VCF reference genome files.

4.4. Inference of Ancestry

The imputed genotypes were then combined with the typed set to assemble a complete ancestral list to infer the ancestral origin of the cell lines. We calculated the probability that a cell belongs to a population given its observed ancestral genotype, the population's genotype frequency and the population weight in the 1000G, based on Bayesian inference. The population genotype frequencies were obtained using the Ensembl API. Particularly, for a cell Y_i that has the genotype G_i of the ancestral SNPs independently occurring, the probability $P(G_i)$ that Y_i belongs to a population k (k corresponds to one of 25 subpopulations in the 1000 G) is:

$$P\left(\vec{G_i}\right) = \prod_{j\,=\,1}^{100} \frac{\hat{P}(Y_i \,=\, k) \times (Y_i \,=\, k)}{\hat{P}(G_{ij})}$$

Then, the subpopulation was assigned to the cell line which had the highest corresponding probability: $arg\ max_k\{P(\vec{G_i})\}$. Labelling of the populations was based on the 1000G classification.

4.5. HLA Prediction

The imputed genotypes and the inferred ancestry of cancer cell lines (CCLs) were used to predict the genotypes of seven human leukocyte antigen (HLA) loci, including HLA-A, HLA-B, HLA-C, HLA-DPB1, HLA-DQA1, HLA-DQB1 and HLA-DRB1, at 4-digit resolution using the HIBAG algorithm [16].

4.6. Ancestry Biomarker Analysis

First, we combined GDSC1 and GDSC2 datasets and generated a unique drug identifier (including drug ID and dataset). Next, we removed the drugs that had extrapolated IC_{50} values (considering the maximum screening concentrations) in more than 50% of the screened CCLs. Then, we performed one-way ANOVAs across the remaining drugs and for each cancer type using ancestry as a predictor and drug response (log10(IC_{50})) as the dependent variable, adjusting for cell characteristics and growth properties as covariates. The unclassified cell lines and the cancer types that contained only one population were filtered out of the analysis. We also removed the drugs that were treated in less than 10 cell lines. To simplify the subsequent analysis and comparison to the Cancer Cell Line Encyclopedia (CCLE) dataset, we re-encoded the ancestry variable as follows: American (AMR) and European (EUR) were combined into Caucasian; East Asian (EAS) and South Asian (SAS) were combined into Asian. ANOVAs were performed in a pairwise manner between two out of the three populations. Populations that had less than three cells were not tested. The effect size was calculated as Cohen's *d*. The *p*-values of the ANOVA tests for each cancer type were adjusted using the Benjamini–Hochberg correction.

4.7. Enrichment of Drug Targets

We identified enriched drug targets in drugs with differential drug response for each cancer type independently. We extracted the putative drug targets from the GDSC manifest files and for each drug target that had at least 2 significant drugs per cancer type. We subsequently tested for enrichment using a hypergeometric test for nine drug targets in total, and we adjusted the enrichment *p*-values for multiplicity using the Bonferroni adjustment method.

4.8. Enrichment of Somatic Driver Genes in Ancestries

We downloaded the Binary Event Matrices (BEM) from the GDSC portal (https://www.cancerrxgene.org/gdsc1000/GDSC1000_WebResources/Home.html, accessed on 13 March 2020), which contains a curated set of cancer somatic driver genes observed in both CCLs and primary tumours, from which the binary mutational status is given. We used a two-sided Fisher's test for performing the enrichment tests. We only tested cancer types with at least one drug displaying a differential drug response, and only mutations with at least six mutated CCLs. In total, we performed 88 statistical tests, and adjusted the enrichment *p*-values for multiplicity with the Bonferroni adjustment method for each cancer type independently.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/ijms221810135/s1.

Author Contributions: Conceptualization, M.P.M.; Data curation, P.B.H.N. and A.J.O.; Formal analysis, P.B.H.N., A.J.O. and S.S.; Methodology, P.B.H.N., A.J.O., S.S. and M.P.M.; Supervision, M.J.G. and M.P.M.; Visualization, P.B.H.N. and A.J.O.; Writing—original draft, P.B.H.N.; Writing—review and editing, A.J.O., M.J.G. and M.P.M. All authors have read and agreed to the published version of the manuscript.

Funding: M.P.M. is supported by the European Union's Horizon 2020 Research and Innovation Programme (Grant agreement No. 950293-COMBAT-RES).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The Affymetrix SNP6.0 arrays genotyping dataset contains 1007 cancer cell lines and 884,148 SNPs ranging from chromosome 1 to chromosome 22, which are deposited in the European Genome-Phenome Archive (EGAS00001000978).

Conflicts of Interest: The authors declare no conflict of interest.

Code Availability: All source code has been made available via https://github.com/phngbh/EIC.

References

1. Nelson, M.R.; Tipney, H.; Painter, J.L.; Shen, J.; Nicoletti, P.; Shen, Y.; Floratos, A.; Sham, P.K.; Li, M.J.; Wang, J.; et al. The support of human genetic evidence for approved drug indications. *Nat. Genet.* **2015**, *47*, 856–860. [CrossRef] [PubMed]
2. Iorio, F.; Knijnenburg, T.A.; Vis, D.J.; Bignell, G.R.; Menden, M.P.; Schubert, M.; Aben, N.; Goncalves, E.; Barthorpe, S.; Lightfoot, H.; et al. A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **2016**, *166*, 740–754. [CrossRef]
3. Ghandi, M.; Huang, F.W.; Jané-Valbuena, J.; Kryukov, G.V.; Lo, C.C.; McDonald, E.R., 3rd; Barreyina, J.; Gelfand, E.T.; Bielski, G.M.; Li, H.; et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **2019**, *569*, 503–508. [CrossRef]
4. Seashore-Ludlow, B.; Rees, M.G.; Cheah, J.H.; Cokol, M.; Price, E.V.; Coletti, M.E.; Jones, V.; Bodycombe, N.E.; Soule, C.K.; Gould, J.; et al. Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. *Cancer Discov.* **2015**, *5*, 1210–1223. [CrossRef] [PubMed]
5. Tamborero, D.; Gonzalez-Perez, A.; Perez-Llamas, C.; Deu-Pons, J.; Kandoth, C.; Reimand, J.; Lawrence, M.S.; Getz, G.; Bader, G.D.; Ding, L.; et al. Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.* **2013**, *3*, 2650. [CrossRef]
6. Nik-Zainal, S.; Davies, H.; Staaf, J.; Ramakrishna, M.; Glodzik, D.; Zou, X.; Martincorena, I.; Alexandrov, L.B.; Martin, S.; Wedge, D.C.; et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **2016**, *534*, 47–54. [CrossRef]
7. Goodspeed, A.; Heiser, L.M.; Gray, J.W.; Costello, J.C. Tumor-Derived Cell Lines as Molecular Models of Cancer Pharmacogenomics. *Mol. Cancer Res.* **2016**, *14*, 3–13. [CrossRef] [PubMed]

8. Chatrath, A.; Ratan, A.; Dutta, A. Germline Variants That Affect Tumor Progression. *Trends Genet.* **2021**, *37*, 433–443. [CrossRef]
9. Chatrath, A.; Przanowska, R.; Kiran, S.; Su, Z.; Saha, S.; Wilson, B.; Tsunematsu, T.; Ahn, J.-H.; Lee, K.Y.; Paulsen, T.; et al. The pan-cancer landscape of prognostic germline variants in 10,582 patients. *Genome Med.* **2020**, *12*, 15. [CrossRef]
10. Qing, T.; Mohsen, H.; Marczyk, M.; Ye, Y.; O'Meara, T.; Zhao, H.; Townsend, J.P.; Gerstein, M.; Hatzis, C.; Kluger, Y.; et al. Germline variant burden in cancer genes correlates with age at diagnosis and somatic mutation burden. *Nat. Commun.* **2020**, *11*, 2438. [CrossRef]
11. Menden, M.P.; Casale, F.P.; Stephan, J.; Bignell, G.R.; Iorio, F.; McDermott, U.; Garnett, M.J.; Saez-Rodriguez, J.; Stegle, O. The germline genetic component of drug sensitivity in cancer cell lines. *Nat. Commun.* **2018**, *9*, 3385. [CrossRef] [PubMed]
12. Özdemir, B.C.; Dotto, G.-P. Racial Differences in Cancer Susceptibility and Survival: More Than the Color of the Skin? *Trends Cancer Res.* **2017**, *3*, 181–197. [CrossRef] [PubMed]
13. Oh, S.S.; Galanter, J.; Thakur, N.; Pino-Yanes, M.; Barcelo, N.E.; White, M.J.; de Bruin, D.M.; Greenblatt, R.M.; Bibbins-Domingo, K.; Wu, A.H.B.; et al. Diversity in Clinical and Biomedical Research: A Promise Yet to Be Fulfilled. *PLoS Med.* **2015**, *12*, e1001918. [CrossRef] [PubMed]
14. Sampson, J.N.; Kidd, K.K.; Kidd, J.R.; Zhao, H. Selecting SNPs to identify ancestry. *Ann. Hum. Genet.* **2011**, *75*, 539–553. [CrossRef] [PubMed]
15. Sudmant, P.H.; Rausch, T.; Gardner, E.J.; Handsaker, R.E.; Abyzov, A.; Huddleston, J.; Zhang, Y.; Ye, K.; Jun, G.; Fritz, M.H.-Y.; et al. An integrated map of structural variation in 2,504 human genomes. *Nature* **2015**, *526*, 75–81. [CrossRef] [PubMed]
16. Zheng, X.; Shen, J.; Cox, C.; Wakefield, J.C.; Ehm, M.G.; Nelson, M.R.; Weir, B.S. HIBAG—HLA genotype imputation with attribute bagging. *Pharm. J.* **2014**, *14*, 192–200. [CrossRef] [PubMed]
17. Ramamoorthy, A.; Knepper, T.C.; Merenda, C.; Mendoza, M.; McLeod, H.L.; Bull, J.; Zhang, L.; Pacanowski, M. Demographic Composition of Select Onco-logic New Molecular Entities Approved by the FDA Between 2008 and 2017. *Clin. Pharmacol. Ther.* **2018**, *104*, 940–948. [CrossRef]
18. Kamineni, A.; Williams, M.A.; Schwartz, S.M.; Cook, L.S.; Weiss, N.S. The incidence of gastric carcinoma in Asian migrants to the United States and their descendants. *Cancer Causes Control.* **1999**, *10*, 77–83. [CrossRef]
19. Chen, S.; Zhou, K.; Yang, L.; Ding, G.; Li, H. Racial Differences in Esophageal Squamous Cell Carcinoma: Incidence and Molecular Features. *Biomed. Res. Int.* **2017**, *2017*, 1204082. [CrossRef]
20. Chen, Z.; Ren, Y.; Du, X.L.; Yang, J.; Shen, Y.; Li, S.; Wu, Y.; Lv, M.; Dong, D.; Li, E.; et al. Incidence and survival differences in esophageal cancer among ethnic groups in the United States. *Oncotarget* **2017**, *8*, 47037–47051. [CrossRef]
21. Shah, S.C.; McKinley, M.; Gupta, S.; Peek, R.M., Jr.; Martinez, M.E.; Gomez, S.L. Population-Based Analysis of Differences in Gastric Cancer Incidence Among Races and Ethnicities in Individuals Age 50 Years and Older. *Gastroenterology* **2020**, *159*, 1705–1714.e2. [CrossRef]
22. Aggarwal, R.; Grabowsky, J.; Strait, N.; Cockerill, A.; Munster, P. Impact of patient ethnicity on the metabolic and immunologic effects of PI3K–mTOR pathway inhibition in patients with solid tumor malignancies. *Cancer Chemother. Pharmacol.* **2014**, *74*, 359–365. [CrossRef]
23. Chong, D.Q.; Toh, X.Y.; Ho, I.A.W.; Sia, K.C.; Newman, J.P.; Yulyana, Y.; Ng, W.-H.; Lai, S.H.; Ho, M.M.F.; Dinesh, N.; et al. Combined treatment of Nimotuzumab and rapamycin is effective against temozolomide-resistant human gliomas regardless of the EGFR mutation status. *BMC Cancer* **2015**, *15*, 255. [CrossRef]
24. O'Donnell, P.H.; Dolan, M.E. Cancer pharmacoethnicity: Ethnic differences in susceptibility to the effects of chemotherapy. *Clin. Cancer Res.* **2009**, *15*, 4806–4814. [CrossRef]
25. Calvo, E.; Baselga, J. Ethnic differences in response to epidermal growth factor receptor tyrosine kinase inhibitors. *J. Clin. Oncol.* **2006**, *24*, 2158–2163. [CrossRef]
26. Zavala, V.A.; Bracci, P.M.; Carethers, J.M.; Carvajal-Carmona, L.; Coggins, N.B.; Cruz-Correa, M.R.; Davis, M.; de Smith, A.J.; Dutil, J.; Figueiredo, J.C.; et al. Cancer health disparities in racial/ethnic minorities in the United States. *Br. J. Cancer* **2021**, *124*, 315–332. [CrossRef] [PubMed]
27. Parker, S.L.; Davis, K.J.; Wingo, P.A.; Ries, L.A.; Heath, C.W., Jr. Cancer statistics by race and ethnicity. *CA Cancer J. Clin.* **1998**, *48*, 31–48. [CrossRef] [PubMed]
28. Haller, D.G.; Cassidy, J.; Clarke, S.J.; Cunningham, D.; Van Cutsem, E.; Hoff, P.M.; Rothenberg, M.L.; Saltz, L.B.; Schmoll, H.-J.; Allegra, C.; et al. Potential regional differences for the tolera-bility profiles of fluoropyrimidines. *J. Clin. Oncol.* **2008**, *26*, 2118–2123. [CrossRef] [PubMed]
29. Loh, M.; Chua, D.; Yao, Y.; Soo, R.A.; Garrett, K.; Zeps, N.; Platell, C.; Minamoto, T.; Kawakami, K.; Iacopetta, B.; et al. Can population differences in chemotherapy outcomes be inferred from differences in pharmacogenetic frequencies? *Pharm. J.* **2013**, *13*, 423–429. [CrossRef]
30. Koual, M.; Tomkiewicz, C.; Cano-Sancho, G.; Antignac, J.-P.; Bats, A.-S.; Coumoul, X. Environmental chemicals, breast cancer progression and drug resistance. *Environ. Health* **2020**, *19*, 117. [CrossRef]
31. Purcell, S.; Neale, B.; Todd-Brown, K.; Thomas, L.; Ferreira, M.A.R.; Bender, D.; Maller, J.; Sklar, P.; de Bakker, P.I.; Daly, M.J.; et al. PLINK: A Tool Set for Whole-Genome Associ-ation and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **2007**, *81*, 559–575. [CrossRef] [PubMed]

32. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **2011**, *27*, 2987–2993. [CrossRef] [PubMed]
33. Das, S.; Forer, L.; Schönherr, S.; Sidore, C.; Locke, A.E.; Kwong, A.; Vrieze, S.I.; Chew, E.Y.; Levy, S.; McGue, M.; et al. Next-generation genotype imputation service and methods. *Nat. Genet.* **2016**, *48*, 1284–1287. [CrossRef]
34. Loh, P.-R.; Danecek, P.; Palamara, P.F.; Fuchsberger, C.; A Reshef, Y.; K Finucane, H.; Schoenherr, S.; Forer, L.; McCarthy, S.; Abecasis, G.R.; et al. Reference-based phasing using the Haplo-type Reference Consortium panel. *Nat. Genet.* **2016**, *48*, 1443–1448. [CrossRef] [PubMed]

3.2 The Interpretable Multimodal Machine Learning (IMML) framework reveals pathological signatures of distal sensorimotor polyneuropathy

The Interpretable Multimodal Machine Learning (IMML) framework reveals pathological signatures of distal sensorimotor polyneuropathy

Phong BH Nguyen[1,2,3], Daniel Garger[1,2], Haifa Maalmi[3,4], Holger Prokisch[6,7], Barbara Thorand[3,8,9], Jerzy Adamski[10,11,12], Gabi Kastenmüller[1,13], Melanie Waldenberger[14], Christian Gieger[14], Annette Peters[8,9], Karsten Suhre[15], Gidon J Bönhof[3,4,5], Wolfgang Rathmann[3,16], Michael Roden[3,4,5], Harald Grallert[8,14], Dan Ziegler[3,4,5], Christian Herder[3,4,5*], Michael P Menden[1,2,3*]

[1]Department of Computational Health, Helmholtz Zentrum Munich, 85764 Neuherberg, Germany

[2]Department of Biology, Ludwig-Maximilians University Munich, 82152 Martinsried, Germany

[3]German Center for Diabetes Research (DZD), 85764 Neuherberg, Germany

[4]Institute for Clinical Diabetology, German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany

[5] Department of Endocrinology and Diabetology, Medical Faculty and University Hospital Düsseldorf, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany

[6]Institute of Neurogenomics, Helmholtz Zentrum Munich, 85764 Neuherberg, Germany

[7]Institute of Human Genetics, Technical University Munich, 80333 Munich, Germany

[8]Institute of Epidemiology, Helmholtz Zentrum Munich, 85764 Neuherberg, Germany

[9]Institute for Medical Information Processing, Biometry and Epidemiology, Ludwig-Maximilians University Munich, 81377 Munich, Germany

[10]Institute of Experimental Genetics, Helmholtz Zentrum Munich, 85764 Neuherberg, Germany

[11]Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117597, Singapore

[12]Institute of Biochemistry, Faculty of Medicine, University of Ljubljana, 1000 Ljubljana, Slovenia

[13]Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum Munich, 85764 Neuherberg, Germany

[14]Research Unit Molecular Epidemiology, Helmholtz Zentrum Munich, 85764 Neuherberg, Germany

[15]Department of Physiology and Biophysics, Weill Cornell Medicine - Qatar, Education City, Doha, 24144 Qatar

[16]Institute of Biometrics and Epidemiology, German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany

* Equal contribution and corresponding authors: Christian Herder (Christian.Herder@ddz.de) & Michael P Menden (michael.menden@hemholtz-muenchen.de)

Abstract

Distal sensorimotor polyneuropathy (DSPN) is a common neurological disorder in elderly adults and people with obesity, prediabetes and diabetes and is associated with high morbidity and premature mortality. DSPN is a multifactorial disease and not fully understood yet. Here, we developed the Interpretable Multimodal Machine Learning (IMML) framework for predicting DSPN prevalence and incidence based on sparse multimodal data. Exploiting IMMLs interpretability further empowered biomarker identification. We leveraged the population-based KORA F4/FF4 cohort including 1,091 participants and their deep multimodal characterisation, i.e. clinical data, genomics, methylomics, transcriptomics, proteomics, inflammatory proteins and metabolomics. Clinical data alone is sufficient to stratify individuals with and without DSPN (AUROC = 0.752), whilst predicting DSPN incidence 6.5±0.2 years later strongly benefits from clinical data complemented with two or more molecular modalities (improved ΔAUROC >0.1, achieved AUROC of 0.714). Important and interpretable features of incident DSPN prediction include up-regulation of proinflammatory cytokines, down-regulation of SUMOylation pathway and essential fatty acids, thus yielding novel insights in the disease pathophysiology. These may become biomarkers for incident DSPN, guide prevention strategies and serve as proof of concept for the utility of IMML in studying complex diseases.

Introduction

Type 2 diabetes (T2D) and its comorbidities have become a global challenge given the increasing case numbers and the enormous cost of diagnosis and treatments, putting burden on the public health management worldwide [1–4]. Distal sensorimotor polyneuropathy (DSPN) is the most common neurological complication in T2D which is characterised by a sensory loss of lower limbs, with or without neuropathic pain, caused by nerve damage [5]. Importantly, recent studies show that DSPN is also prevalent in elderly adults and people with prediabetes and obesity, thus affecting an increasing proportion of the general population [6,7].

DSPN diagnosis is challenging. It is based on evaluating the sensing ability of individuals, observing existing physiological conditions and morphological changes, and finally conducting neurophysiological measurements [7]. However, a large proportion of individuals with DSPN remain undiagnosed [8], and we lack computational methods to reliably predict prevalent (i.e. cross-sectional) and incident (i.e. disease trajectory) DSPN. Furthermore, the complex pathogenesis of DSPN is not fully understood yet, and is anticipated to be multifactorial [9], attributed by the interplay of many intrinsic and extrinsic factors [7], thus rendering predictions challenging.

With the advent of multi-omics technologies, we are now able to conduct high-throughput assays that simultaneously characterise hundreds to millions of biomolecules across large patient cohorts [10–12]. As a result, the number of datasets with deep multi-omic characterisation has been exponentially increasing in recent years, e.g. the population-based KORA (Collaborative Research in the Region of Augsburg) F4/FF4 cohort which includes a subset of 1,091 participants with DSPN label defined by the Michigan Neuropathy Screening Instrument (MNSI) [13]. Each participant in KORA is characterised with clinical data, genomics, methylomics, transcriptomics, proteomics, inflammatory proteins and metabolomics [13]. Rise of these large-scale multimodal datasets and computational integration frameworks are the prerequisite to gain insights in complex multifactorial diseases and comorbidities at multiple molecular levels [14,15], here exemplified with DSPN.

Statistical methods empower biological insights. For instance, to select genes associated with a certain phenotype, e.g. gene expression patterns in DSPN, the conventional method is setting a fixed significance threshold for a certain univariate statistical test and selecting genes that fall under the threshold [16,17]. While it is effective in identifying the most univariately significant genes, it tends to neglect smaller effect sizes which may cumulatively contribute to a multivariate model. In order to address this, gene set enrichment analysis (GSEA) is a powerful tool to prioritise functional relevant genes regardless of their global effect size, as it puts genes into context of biological signalling pathways using prior knowledge. Notably, GSEA generalises to other biomolecules such as proteins and metabolites, representing a potential approach to study complex systemic diseases [18].

There are several multimodality data integration strategies. The simplest approach is to concatenate all available features together before supervised learning [19,20]. This method is simple to implement, however, it requires extensive data processing and normalisation to incorporate heterogenous modalities encompassing vast amounts of features, thereby often neglecting important biological signals [20]. To address this, state-of-the-art integration methods such as ensemble stacking (meta learning) are employed to combine the powers of multiple data modalities and/or learning algorithms, whilst increasing weights of more predictive modalities/algorithms. This empowers to learn complicated structures and relationships of the data [21]. A critical assumption of multi-view learning, however, is that the single-view models should be independent [21]. This assumption is often violated in complex metabolic diseases, as there is a high level of redundancy and correlations amongst feature layers. Nevertheless, multi-view learning has proven to be superior compared to models leveraging concatenated feature space in crowd-sourced computational challenges [22,23].

In this study, we present the Interpretable Multimodal Machine Learning (IMML) framework, and exemplify its capability with DSPN classification and predicting DSPN onset over 6.5-years, i.e. prevalent and incidence predictions, respectively. IMML focuses on deriving predictive, interpretable and translational models leveraging sparse multimodal data. For this, we developed a two-step feature selection and integration machine learning framework. The first step extracts functionally relevant features of each molecular layer in isolation and leverages GSEA, whilst the second step benchmarks all combinations of data modalities based on cross-validated and regularised linear models. We hypothesise that well performing models at the minimum number of data modality will give insights into the disease aetiology of DSPN and its incidence, thus may improve diagnosis and pave the way for prevention strategies.

The derived framework successfully classifies cross-sectional DSPN and predicts future incident DSPN, as well as identify relevant and actionable biomarkers of the disease. In particular, the model achieves the AUROC of 0.752 and 0.714 for cross-sectional DSPN and incident DSPN, respectively. Dissecting the model complexity shows that involving molecular data helps improving the prediction performance for incident DSPN, with ΔAUROC >0.1 compared to the clinical data-only model. Importantly, feature analysis shows multiple important signatures of incident DSPN such as up-regulation of inflammatory cytokines and down-regulation of SUMOylation process and essential fatty acids. These putative biomarkers serve as useful resources for future investigation to identify actionable biomarkers for interventions. These findings do not only help identifying individuals at risk of developing the disease but also shed light into the pathological mechanisms and important biomarkers that would help improve patients' life, further advancing precision medicine.

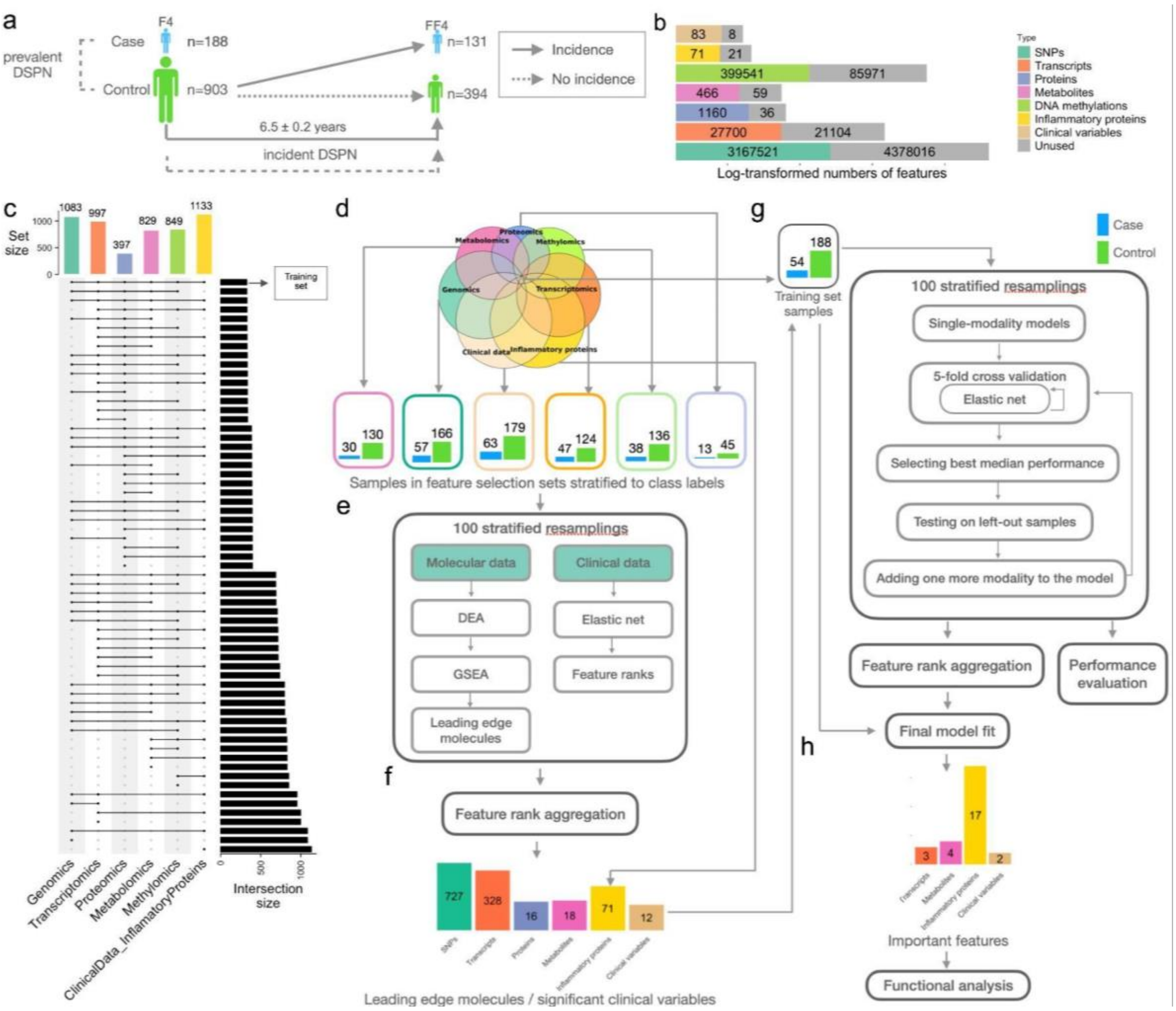


Figure 1: Workflow of interpretable multimodal framework for feature prioritisation, DSPN classification and disease incidence prediction. (a) Distribution of samples across time points (KORA F4 and FF4), disease status (case or control) at baseline (KORA F4) and follow-up (KORA FF4) and prediction tasks. Both models were trained on the same set of F4 features but different labels and a subset of samples. **(b)** Number of features stratified according to data modalities. In grey are removed features after pre-processing. **(c)** Number of samples characterised within each data modality and their overlaps in KORA F4. **(d)** Fully characterised samples in KORA F4 were exclusively leveraged for **(g)** the second and final training step, whilst the remaining sparse samples were used for **(e)** prior feature prioritisation: All molecular features were shortlisted based on differential expression analysis (DEA), gene set enrichment analysis (GSEA) and their leading-edge genes (**Methods**), whilst clinical features were ranked according to feature importance of elastic net models. **(f)** Features for the final training step were selected based on rank aggregation (**Methods**). **(g)** The final training set contained 54 DSPN cases and 188 controls in KORA F4. In the second step, elastic net models determined the optimal number of modalities, features and combination of modalities. These models implemented forward feature selection in a nested cross-validation, using weighted log loss to account for class imbalance, and finally 100 stratified resampling during training and rank aggregation (**Methods**), thus returning **(h)** the refined and final model further subject to functional analysis for gaining insights in DSPN pathophysiology.

Results

We leveraged the population-based KORA study including participants aged 62-81 years with clinical examination of DSPN from the F4 (2006-2008) and FF4 (2013-2014) surveys (**Methods**). The earlier F4 time point surveyed 1,091 individuals of whom 622 were followed up at the later FF4 time point. We used the established Michigan Neuropathy Screening Instrument (MNSI) to assess and define the DSPN status as described in previous studies [24,25]. Using MNSI, we identified 188 DSPN cases and 903 controls at F4, and 131 controls who developed DSPN between F4 and FF4 (**Supp. Table S1, S2**). The first machine learning task was to predict DSPN prevalence at F4 (**Fig. 1a**). The second task was to predict whether controls at F4 will develop incident DSPN during the period from F4 to FF4 (**Fig. 1a**).

Data modalities included in this study were genomics, transcriptomics, proteomics, metabolomics, methylomics, clinical attributes and a panel of inflammatory proteins. The modalities vary greatly in number of features, ranging from 91 clinical attributes to >7.5 million single nucleotide polymorphisms (SNPs; **Fig. 1b**). After data type-specific processing (**Methods**), the number of features was drastically reduced, e.g., only approximately 42% of the assayed SNPs were used for subsequent analyses (**Methods**; **Fig. 1b**). Participants in KORA were sparsely characterised with varying overlaps of data modalities (**Fig. 1c**).

The two-step feature selection and integration machine learning framework

The IMML framework is based on a two-step approach: i) Extensive feature engineering and selection process (**Fig. 1d-f**; **Fig. S1**; **Methods**), and ii) final model training (**Fig. 1g,h**). For enriching biological signals and reducing feature space, we used 849 samples which were lacking at least one data modality, whilst the remaining completely characterised 242 samples were exclusively used for the final model training. Both subsets of data for feature selection and model training and testing were subject to PCA analysis using clinical information to ensure there was no potential bias in sample selection (**Supp. Fig. S2**).

For the first step, i.e. feature engineering and selection processes, we prioritised predictive and biologically relevant biomolecules regardless of their effect sizes (**Methods**). We observed that GSEA-based methods significantly outperformed threshold-based methods (Wilcoxon Rank Sum test, p-value = 3.758e-12; **Supp. Fig. S3**). Therefore, we implemented differential expression analysis (DEA) [17] followed by GSEA [18] to extract a list of molecule sets corresponding to signalling pathways that may be pivotal in DSPN development. This process was repeatedly performed to account for variability (**Methods**) [26]. Finally, we extracted the leading-edge molecules, i.e. those that drive the enrichment of molecule sets [18]. We obtained between zero and 25 significantly enriched molecule sets per data modality (**Supp. Fig. S4, S6; Supp. Table S3, S4**), from which we extracted up to 727 leading-edge features (**Supp. Fig. S1**, **S5**, **S7**). In addition, for clinical feature selection, we trained elastic net models and leveraged rank aggregation to retrieve 13 predictive clinical features (**Methods**; **Fig. 1e,f**; **Supp. Fig. S1**).

For the second step, i.e. final model training and multimodal data integration, we leveraged the short-listed features from the analyses above. The final model was trained with an embedded feature selection whilst balancing number of modalities. We benchmarked three feature integration methods, i.e. forward feature selection (FFS), ensemble and concatenation of all features together (**Supp. Fig. S3a; Methods**), and observed best performance with GSEA-FFS followed by GSEA-ensemble stacking

(**Supp. Fig. S3b**). When comparing the performance of the FFS and ensemble stacking methods using all modalities and with GSEA as the feature selection approach, the FFS algorithm achieved marginally higher predictive performance (**Supp. Fig. S8a**). Both methods retained inflammatory proteins as the most predictive features, however, the GSEA-FFS was further able to detect clinically relevant signals from other modalities (**Supp. Fig. S8b**). Therefore, we implemented an iterative FFS algorithm with resampled cross-validation (**Methods**).

To select the machine learning algorithm for DSPN prediction, we compared the predictive performance of elastic net, random forest and support vector machine, the latter leveraged linear and radial kernels (**Supp. Fig. S9**). For this we performed 100 matched resamples with forward feature selection. Elastic net outperformed the other three machine learning algorithms in both prevalent DSPN (**Supp. Fig. S9a-d**) and incident DSPN predictions (**Supp. Fig. S9e-h**). Best performances in prevalent DSPN (AUROCs of 0.737) and incident DSPN (AUROCs of 0.708) predictions were observed at 1-modality and 3-modality models, respectively. Notably, none of the other machine learning algorithms reached AUROC higher than 0.700 at any number of modalities.

For each iteration of resampling, the most predictive combination of modalities were selected based on cross-validation (**Fig. 1g,h**; **Supp. Fig. S1**). Analysis of the final model returned predictive modality combinations, which became subject to functional analysis for DSPN classification and incidence prediction in the following sections.

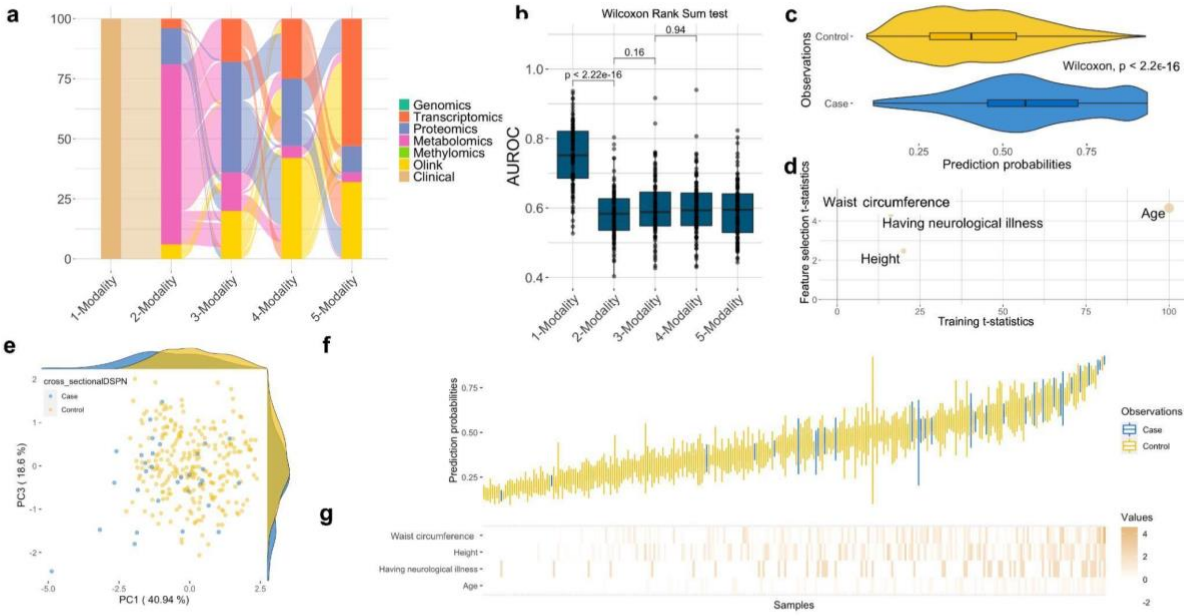


Figure 2: The clinical model can sufficiently stratify DSPN prevalence. (a) Classification of DSPN first leverages clinical attributes, and cumulatively adds molecular modalities with forward feature selection (**Methods**). Here shown for 100 cross-validated models. **(b)** Test set performance of DSPN classification based on leveraging between one to seven data modalities. (**c)** Prediction probabilities of samples in the 100 left-out test sets leveraging clinical features only, stratified into true labels (case and control). (**d)** Feature importance of the final model based on clinical attributes alone applied to training and feature selection set (**Methods**). (**e)** PCA leveraging the four most important clinical features shown in panel **d** to stratify cases from control. **(f)**

Distribution of the test prediction probability of all samples of 100 resampled and cross-validated models. **(g)** Normalised values of the four most important clinical features. The order of samples corresponds to panel **f**.

Clinical data can sufficiently stratify individuals with and without DSPN

In a clinical setting, all suspected DSPN patients are thoroughly clinically characterised and neurologically evaluated. Therefore, our FFS algorithm used KORA clinical attributes as baseline input, and consecutively, evaluated the gained performance by adding more molecular modalities to classify DSPN (**Fig. 2a**). Metabolite and protein features were the most frequently added across 100 iterations, while transcripts were usually added last (**Fig. 2a**). However, the baseline clinical model significantly outperformed any more complex model (Wilcoxon rank sum test, $p<2.22e-16$; **Fig. 2b,c**; **Supp. Fig. S10a, S11**). The clinical model had a median area under the receiver operating curve (AUROC; **Methods**) value of 0.752 with an interquartile range (IQR) of 0.686-0.821 and 95% confidence interval (CI) of 0.733-0.770, whilst the best performing model with molecular data only achieved a median AUROC of 0.583 with IQR of 0.539-0.627. This suggested that clinical variables alone are sufficient to stratify individuals with and without DSPN.

To further dissect the predictive component of clinical attributes, we extracted the most important clinical features from 100 resampled and cross-validated models. For this, we leveraged Robust Rank Aggregation (RRA; FDR < 5%), and used these within the final model (**Methods**). After computing t-statistics of model parameters, four variables had non-zero t-statistics, including age, waist circumference, height and whether the patient had neurological illnesses (self-reported during interview; **Fig. 2d**). The principal component analysis (PCA) of these four clinical variables empowered the segregations of cases and controls (**Fig. 2e**). When we further stratified the prediction probabilities to individual samples and ranked them according to mean probability, most cases were ranked higher than controls, although there were a few outliers (**Fig. 2f**). Values of age, waist circumference and height were significantly higher in cases compared to controls ($p < 0.05$, Wilcoxon Rank Sum test) while having neurological illnesses was significantly enriched in DSPN cases ($p < 0.05$, Fisher's Exact test; **Fig. 2g; Supp. Fig. S12**).

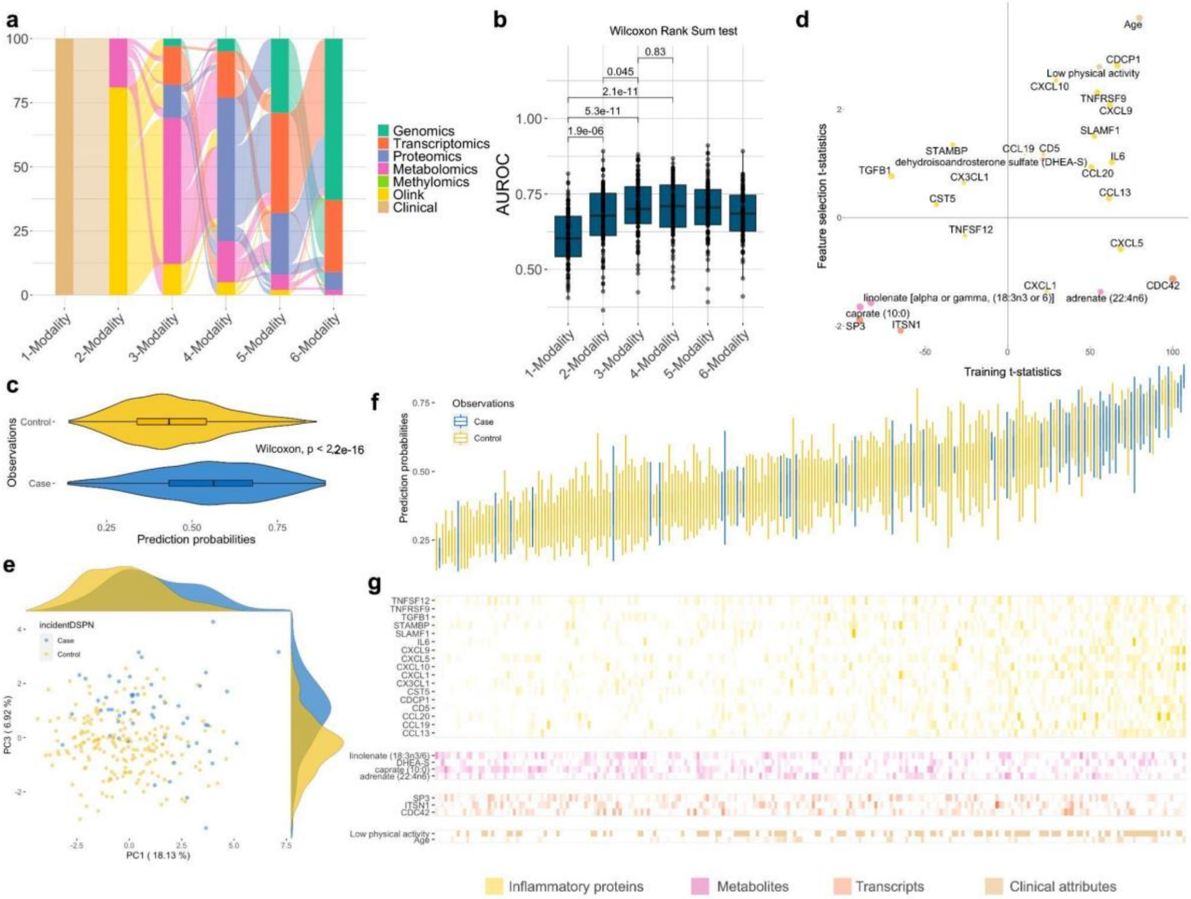


Figure 3: Predicting DSPN incidence benefits from molecular data. (a) Each model starts with clinical attributes at baseline, and consecutively increases the number of modalities by adding the next molecular modality with feed forward selection for 100 cross-validated models **(Methods). (b)** Performance of all model complexities to predict patient trajectories. **(c)** Prediction probabilities of samples in the 100 left-out testing sets using the optimal mode of the corresponding iterations, stratified into true labels (case and control). **(d)** Important features of the final model. X-axis represents the signed model important scores (t-statistics) of the features in the training set, y-axis represents their t-statistics in the feature selection set. **(e)** PCA leveraging the most important features of the final model in panel **d**. (**f**) Waterfall plot of prediction probability of all samples across 100 resampling steps. **(g)** Normalised values of the important features in panel **d** stratified by individual samples and ordered according to panel **f.** Features belonging to the same data modality are grouped together.

Molecular data improves DSPN incidence prediction

DSPN incidence prediction was strongly enhanced by integrating clinical and molecular data. In contrast to clinical baseline models (**Supp. Fig. S13a,b**), we observed a strong benefit in leveraging molecular modalities for predicting whether participants of the KORA F4 cohort will develop DSPN or not within the next 6.5±0.2 years (**Fig. 3a,b; Supp. Fig. S13c**). The baseline DSPN incidence model achieved a median AUROC of 0.603 with an IQR of 0.543-0.676 and 95% CI of 0.588-0.624. This was significantly outperformed by adding either one or two additional molecular data modalities (**Fig. 3a,b**; **Supp. Fig. S10b, S14**), i.e. median AUROC of 0.678 with an IQR of 0.612-0.752 and 95% CI of 0.652-0.692 and AUROC of 0.700 with IQR of 0.651-0.774 and 95% CI of 0.686-0.722, respectively (Wilcoxon rank sum test, p = 1.9e-16 and p = 2.9e-11, respectively). In essence, molecular features significantly enhanced DSPN incidence prediction. We observed that inflammatory proteins were >80% the first picked molecular layer, followed by metabolites, whilst SNPs seemed to carry the least predictive information (**Fig. 3a**). The performance tended to saturate at 3-modality models as adding more modalities did not significantly improve the performance anymore (Wilcoxon rank sum test, p=0.95), i.e. 4-modality models had a median AUROC of 0.714 with an IQR of 0.640-0.774 and 95% CI of 0.684-0.720 (**Fig. 3b**). We observed similar results and a saturation of performance at 3-modalities, when not enforcing clinical attributes as baseline modality, which were still selected in 57% of all 3-modality models (**Supp. Fig. S15**). In essence, prediction of DSPN incidence strongly benefited from adding two or more molecular modalities, and saturated at the 3-modality models.

For feature importance analysis of the final model, we selected 3-modality models (**Methods**). The prediction probabilities of incident cases were significantly higher compared to controls (Wilcoxon rank sum test, p < 2e-16; **Fig. 3c**). We obtained 26 features with non-zero t-statistics including 17 inflammatory proteins, four metabolites, three transcripts and two clinical variables (**Fig. 3d**; **Methods**). Most of the predictive power stemmed from inflammatory proteins, whilst two transcripts (CDC42 and SP3) and two metabolites (caprate and linolenate) displayed the largest t-statistic magnitude **(Supp. Fig. S16**). PCA analysis on these 26 features illustrated that they enable the prediction of DSPN incidence (**Fig. 3e**).

When stratifying the prediction probabilities of the 3-modality models into individual samples and ranking them based on their means, most of the incident DSPN cases were concentrated in higher probability regions, in contrast to controls (**Fig. 3f**). Furthermore, higher prediction probability also corresponded to higher concentrations of many model-important inflammatory proteins and lower concentrations of caprate (**Fig. 3g**). Incident DSPN cases were significantly enriched for low physical activity (Fisher's exact test, p = 0.003; **Supp. Fig. S17**). In essence, levels of inflammatory proteins and metabolites were significantly different between people transitioning to DSPN compared to those who did not (Wilcoxon rank sum tests, p < 0.05; **Supp. Fig. S17**), highlighting the important role of molecular features to predict DSPN incidence.

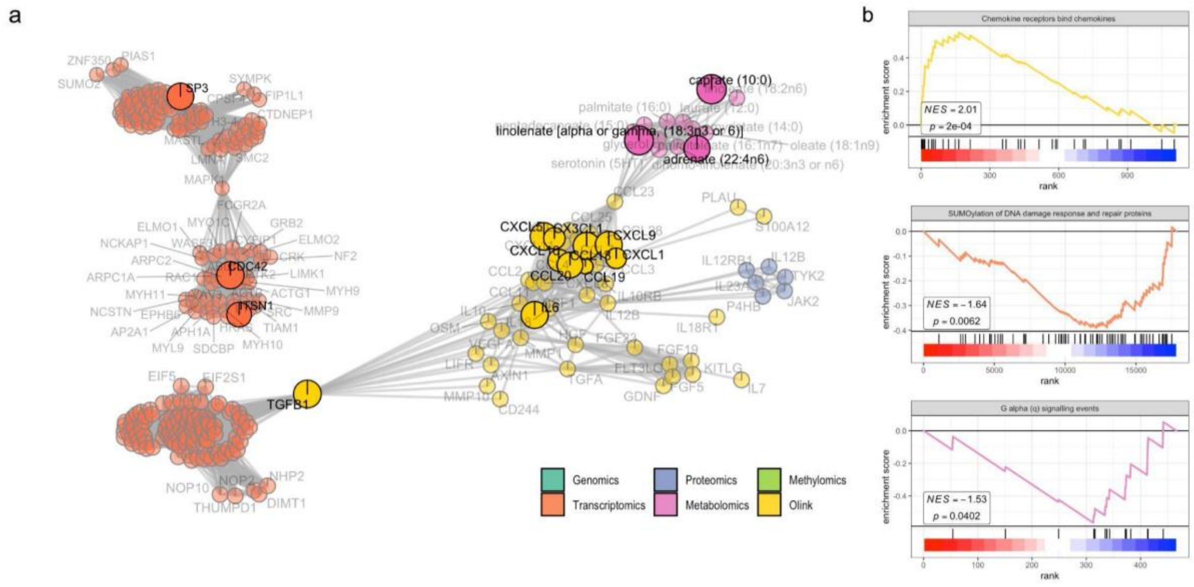


Figure 4. Enrichment of inflammatory cytokines- and essential fatty acids-related pathways as important signatures of DSPN progression. (**a**) Sub-network of important features to predict development of DSPN. Each node is a feature coloured according to its data modality. Edges are the number of shared molecule sets between two nodes. The important features in the final model are highlighted and labelled in black. Below are examples of enriched molecule sets associated with (**b**) inflammation-related proteins, (**c**) transcripts and (**d**) metabolites: (**b**) The up-regulation of "Chemokine receptors bind chemokine" gene set. (**c**) SUMOylation of DNA replication proteins. (**d**) G alpha (q) signalling events. Molecules are ranked in decreasing order of t-statistics, with ticks representing molecules that belong to the examined molecule set.

Increased inflammation, reduced levels of SUMOylation and essential fatty acids as important signatures of incident DSPN

For gaining further insights into the prediction of incident DSPN, we investigated predictive features in the context of the initial GSEA-based feature selection. To this end, we created a network of features connecting all molecular layers by shared signalling pathways (**Methods**; **Fig. 4a**; **Supp. Fig. S7**). For this, all biomolecules were connected to any other leading-edge molecule according to Reactome [27]. We identified two large subnetworks of 15 predictive features containing nine inflammatory proteins, three transcripts (CD42, SP3 and ITSN1) and three metabolites (caprate, linolenate and adrenate; **Fig. 4a**).

Inflammation is an important signature of incident DSPN prediction, which is evident by the increased frequency of functional important inflammatory proteins in the identified large subnetwork (**Fig. 4a**). To gain further understanding of their role, we performed GSEA on the proteomic training data focusing on gene sets involving inflammation. Binding of chemokines to their receptors was significantly upregulated (**Fig. 4b**; adjusted p-value = 0.008), as well as signalling of G protein-coupled receptors (GPCR; **Supp. Fig. S18;** adjusted p-values < 0.2).

Transcriptomic modality encompassed consistently significant gene sets. In particular, down-regulation of SUMOylation-related signalling pathways were consistently observed in both feature selection and training sets (adjusted p-values < 0.2; **Fig. 4c**; **Supp. Fig. S18**). These included SUMOylation of proteins involved in DNA replication and DNA damage response and repair. In addition, the gene set involved in transport of mature RNA from nucleus to cytoplasm was significantly down-regulated (**Fig. 4c**; **Supp. Fig. S18**).

Interestingly, all metabolites in the subnetwork were fatty acids (**Fig. 4a**) and all were significantly down-regulated, i.e. caprate, linolenate and adrenate. As a result, GPCR pathways related to fatty acids activity were significantly down-regulated (**Fig. 4d**; **Supp. Fig. S18**; adjusted p-values < 0.2). Other significant metabolomic pathways included the down-regulation of fatty acid-related signalling and synthesis, secretion, and inactivation of glucagon-like peptide-1 (GLP-1; **Supp. Fig. S18**), and up-regulation of transport of organic anions (**Supp. Fig. S18)**.

Overall, functional analysis of the predictive features revealed molecular signatures of incident DSPN. Particularly, the up-regulation of several inflammatory proteins and down-regulation of SUMOylation-related transcripts and essential fatty acids are the most significant patterns.

Discussion

DSPN is a complex disease attributed to multiple and heterogeneous risk factors [7]. Thus, integration of sparse multimodal data is a prerequisite for a deeper understanding of the disease pathophysiology. In order to address this, here we present the IMML framework, which allows prediction of prevalent and incident DSPN status based on clinical and molecular characterisation. We achieved good performance for both prediction tasks, i.e. AUROC > 0.7. Furthermore, the IMML two-step approach empowered the analysis of sparse clinical and molecular data, which is common in biomedical research. Utilising the modality-specific non-overlapping samples for feature selection increased the number of accessible samples and reproducibility of molecular patterns across different datasets.

The analysis of prevalent DSPN (classification of case-control in the F4 population) suggested that the clinical model (using only clinical variables) outperformed the concatenated models (using clinical + molecular variables) in prediction. Then, feature analysis of the clinical model suggested that age, height, neurological illness, and waist circumference were the most important factors that influence the prediction of prevalent DSPN. Age and height have been reported to be associated with prevalent DSPN [28]. The neurological comorbidity status of patients is not used to classify DSPN yet, however, there might be an intrinsic neurological mechanism that links DSPN to other neurological illnesses. Finally, waist circumference is strongly correlated with BMI, which has been reported to be a risk factor for developing DSPN [29]. From a clinical perspective it is worth mentioning that only waist circumference represents a modifiable risk factor which emphasises the role of obesity prevention and treatment also in the context of DSPN. In summary, for prevalent DSPN, our analysis is confirmatory of previous studies with respect to these clinical variables. However, here we report the clinical variables in the context of a comprehensive multi-modality analysis of DSPN prevalence, thus adding another layer of information to the model.

In the case of incident DSPN prediction, the molecular variables added prediction value as they helped improve the prediction performance (higher AUROC values) compared to the clinical model alone. Feature analysis detected multiple important and potentially actionable biomarkers such as inflammatory proteins, SUMOylation-related transcripts and essential fatty acids. Although the association between inflammatory proteins and incident DSPN has been reported before [24,25], there are as yet no data from population-based studies such as ours implicating SUMOylation-related transcripts and essential fatty acids in the development of DSPN so that these findings are novel and merit further investigation in other cohorts. Additionally, none of these biomarkers and pathways has been reported before in the context of our novel multi-modality analysis of DSPN incidence.

Feature analysis suggested the crucial role of subclinical inflammation in the development of DSPN. We found that 18 out of the 27 most important incident DSPN features were inflammatory proteins. Our finding was consistent with previous studies showing the predictive value of pro-inflammatory cytokines in DSPN [24,25]. One cytokine (IL-6), five chemokines (CXCL9, CXCL10, CCL13, CCL19 and CCL20) and five soluble forms of transmembrane proteins (CDCP1, SLAMF1, TNFRSF9, TNFRSF11B, CD5) were up-regulated at baseline in patients with incident DSPN, suggesting a proinflammatory process which could be observed before DSPN onset [24,25]. CXCL9 and CXCL10 have been shown to directly impact neurotoxic effects [25]. In addition, nerve-derived chemokines may play a role in attracting immune cells to further damage stressed neurons [25]. In accordance with the up-regulation of these proteins, signalling pathways downstream of GPCR signalling, specifically involving chemokine-induced inflammation, were also significantly up-regulated. Subclinical inflammation is an established hallmark of DSPN, as people affected by the disease often have elevated levels of pro-inflammatory cytokines that are associated with nerve damage [24,25,30,31]. It has been hypothesised that a cross-talk of innate and adaptive immune cells contributes to DSPN [25], which is further supported by our study, but requires further mechanistic validation.

Remarkably, inflammatory effects were observed in the blood samples prior to disease onset. Thus, the predictive pro-inflammatory cytokines, chemokines and transmembrane proteins observed in this study could represent modifiable risk factors and therefore therapeutic targets for disease prevention. For example, salicylate was reported in many studies to have inhibitory effects on production of cytokines and chemokines[32]. In addition, novel treatment approaches targeting IL-1beta-related mechanisms have been demonstrated to reduce subclinical inflammation and have beneficial effects on cardiometabolic risk [33,34], and may be generalisable for DSPN. Beyond pharmacological approaches to attenuate subclinical inflammation, it is important to emphasise that subclinical inflammation is triggered by a range of other modifiable risk factors such as high-calorie diet, certain nutrients, physical inactivity, obesity, psychosocial stress and sleep disturbances so that lifestyle changes represent another option for intervention[35].

The transcriptomic layer also gained attention as one of the most important predictors of DSPN. Particularly, significant down-regulation of the small ubiquitin-related modifier (SUMO) pathway was consistent with a recent study [36], which demonstrated SUMO posttranslational modifications are involved in glycolysis. Furthermore, the tricarboxylic acid (TCA) cycle plays a crucial role in maintaining important metabolic processes in sensory neurons, and deficiency of SUMO activity causes damaging effects which may specifically contribute to DSPN pathogenesis [36]. Although this enrichment analysis has to be interpreted with caution due to the small sample size, it is noteworthy that oxidative stress and inflammation have been proposed as mediators linking hyperglycaemia and impaired SUMOylation in diabetic polyneuropathy and that aberrant SUMOylation has also been implicated in the aetiology of neurodegenerative diseases [37]. Thus, this finding extends the aforementioned results on inflammation, corroborates other studies and may point towards another mechanism how DSPN risk could be targeted by addressing modifiable risk factors leading to inflammation and oxidative stress.

Three fatty acids were identified as potential biomarkers of incident DSPN, i.e. caprate, linolenate and adrenate. Capric acid, also known as decanoid acid, binds to the α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) receptor, a glutamate receptor that mediates synaptic transmission in the brain. Capric acid has antioxidative effects in neuronal cells [38] and has been implicated in the amelioration of several neurological diseases [39] so that further studies of the potential role of decreased capric acid levels for the development of DSPN appear promising. Adrenic acid-derived epoxy fatty acids have anti-nociceptive properties and can reduce inflammatory pain [40] so that a link between lower levels and higher risk of DSPN appears biologically plausible. Linolenic acid and adrenic acid, or all-cis-7,10,13,16-docosatetraenoic acid, are also essential polyunsaturated fatty acids (PUFA), which are precursors of more potent derivatives such as arachidonic (omega-6, ARA) and docosahexaenoic (omega-3, DHA) acids, which serve as either building blocks of cell membrane or substrates for the synthesis of inflammation-related compounds and are involved in neural development processes [41,42]. Furthermore, DHA reduces pro-inflammatory cytokines and induces anti-inflammation cytokines, which is consistent with the observed patterns of inflammatory proteins in our datasets [41]. In addition, certain groups of GPCR called free fatty acid receptors (FFAR), such as GPR40/FFAR1 and GPR120/FFAR4, are activated by PUFAs and medium-chain fatty acids (MCFA), such as capric acid, to regulate many cellular processes, i.e. insulin secretion, inflammation, neural cognitive and sensory function [43]. Overall, caprate, linolenate and adrenate have not been linked to DSPN in detailed investigations but nevertheless highlight the possibility that they should be modifiable risk factors that could be modulated by specific dietary interventions or dietary supplements. Importantly, experimental results suggested that PUFA might be a potential agent to treat DSPN [44,45], subject to future studies focussing on high-risk individuals assessing the potential preventive and therapeutic properties of dietary fatty acids in this context.

One strength of our study's design is the utilisation of population-based prospective data from a large cohort (KORA F4). The KORA cohort contains repeated assessment of DSPN status using identical examination methods at two timepoints, which allows studying both prevalent and incident DSPN. The fact that the mean follow-up time was 6.5 years and that we do not have data on DSPN diagnosis between both studies means that our data cannot be extrapolated to considerably shorter or longer time-periods than 6.5 years. It is possible that different variables may be more powerful for short-term or very long-term prediction of DSPN which needs to be addressed in future studies. Furthermore, we presented an innovative machine learning framework to model incidence of DSPN by integrating multi-omic and clinical data. Previous efforts either focused on classifying the disease in a cross-sectional context, lacked multi-omics integration strategies or exhibited limitations of univariate statistical analyses[25,46–50]. In our study, the multi-omic data integration added significant information to boost predictive performance. Strikingly, our findings were observed in blood instead of biopsies containing neuronal cells which would be more tissue-specific for DSPN but are not accessible in large epidemiological cohort studies. Results of this study highlighted the utility of a less invasive blood-based assay to study complex diseases such as DSPN in clinical practice. Although the prediction performance could be improved further by increasing the quantity and quality of data collection and more advanced machine learning development, we believe that using such a model could both be valuable in clinical practice and for the design of future intervention studies. On the one hand, the early identification of people at elevated risk of DSPN could lead to an intensification of (pharmacological and non-pharmacological) risk factor treatment in these people. On the other hand, our model could be used for an enrichment of high-risk individuals in future intervention trials which could reduce required sample size and therefore the costs to assess novel prevention and treatment options. In the long run, our results indicate potentially actionable biomarkers that could be targeted by novel therapy concepts.

One limitation of this study is that our IMML relies substantially on the availability of prior biological knowledge and functional annotation of biomolecules, which thereby reduces the number of evaluated features and may introduce a bias. That being said, incorporating biological knowledge during feature selection increased interpretability, reduced multiple hypothesis testing and utilised cumulatively low-effect size features, overall boosting the model performance. It is our belief that IMML achieves a good balance between predictive power and interpretable DSPN signature, which thereby increases clinical translatability. In addition, validation with external cohorts is currently not feasible due to the uniqueness of the KORA dataset, i.e. deep molecular and longitudinal phenotypic DSPN characterisation, which empowered this study.

One aspect that we were unable to claim are causal relationships due to the inherent limitations of the Kora study design. This is neither addressable in the cross-sectional studies, nor in the prospective segment of our analysis, which concentrates on the occurrence of Diabetic Sensorimotor Polyneuropathy (DSPN), however, the latter sheds light on the temporal associations between risk factors and the onset of DSPN. To delve into the aspect of causality, alternative methodologies are warranted, such as Mendelian randomization studies conducted in human cohorts or investigations utilising animal models and *in vitro* studies employing pertinent cell culture models of neurotoxicity. Our findings concerning incident DSPN offer promising candidates for such inquiries subject to further studies.

In summary, we presented the IMML framework which allows studying multifactorial diseases, here exemplified with DSPN. Leveraging IMML, we were able to stratify individuals according to prevalent DSPN status using only clinical variables. More importantly, IMML showed that molecular data is essential to predict the incidence of DSPN, and pathological signatures are detectable in blood samples 6-7 years before disease onset. IMML is capable of integrating sparse multimodal data, and is generalisable to other cohorts and comorbidities. In essence, IMML simplifies the integration and interpretation, thus giving insights in the disease pathophysiology of DSPN, and may navigate the next generation of diagnostic, prevention and treatment strategies of DSPN.

Methods

Population data

The population-based data in this study was obtained from the “Kooperative Gesundheitsforschung in der Region Augsburg/Cooperative Health Research in the Region of Augsburg” (KORA) platform [24,51]. Specifically, data from the KORA F4 (2006-2008) and the KORA FF4 (2013-2014) studies, both follow-up examinations of the population-based KORA S4 study (1999-2001), were used. All examinations were carried out in accordance with the Declaration of Helsinki, including written informed consent from all participants. The KORA study was approved by the ethics board of the Bavarian Chamber of Physicians (Munich, Germany). The data used in this study was obtained under a data sharing agreement with the Board of Management of KORA and all data owners. Initially there were 1,161 KORA F4 participants aged 62-81 years in the age group with the neuropathy examination module. We excluded 28 individuals with known type 1 diabetes, diabetes forms other than type 2, or unclear glucose tolerance status. In total, we leveraged 1,133 individuals.

DSPN assessment

We used the examination part of the Michigan Neuropathy Screening Instrument (MNSI) score to assess the status of DSPN for all participants of KORA F4 and KORA FF4, as described previously [24]. In the MNSI assessment, we evaluated the appearance of feet (normal or any abnormalities such as dry skin, calluses, infections, fissures, or other irregularities), foot ulceration, ankle reflexes, and vibration perception threshold at the great toes which was assessed with the Rydel-Seiffer graduated C 64 Hz tuning fork [52]. The normal limits for vibration perception threshold, adjusted for age, were determined based on the method outlined by Martina et al. [53]. The MNSI score also included the bilateral examination of touch/pressure sensation using a 10-g monofilament (Neuropen) [54]. Therefore, the total

MNSI score ranged from 0 (indicating normal in all aspects) to a maximum of 10 points. Considering the advanced age of the participants and the inclusion of the monofilament examination, we defined distal sensorimotor polyneuropathy (DSPN) as a score of equal or higher than 3 points [25]. Thus, participants with an MNSI score ≥3 in KORA F4 were considered as prevalent DSPN cases, whereas participants without DSPN in KORA F4 (MNSI <3) but MNSI ≥3 in KORA FF4 were considered as incident cases. This definition meets the minimal diagnostic criteria for possible DSPN, as outlined by the Toronto Diabetic Neuropathy Expert Group [55].

Using this criterion, for prevalent DSPN analysis, among 1,091 out of 1,133 individuals having MNSI scoring records, there were 188 cases and 903 controls. For incident DSPN analysis, we only considered the 903 controls in the KORA F4 and examined their progression of DSPN status in the KORA FF4. Among these, we excluded 378 individuals that either did not participate or lacked MNSI scoring records in the KORA FF4. For the incident DSPN analysis, the remaining 521 participants were split into 131 DSPN cases and 394 controls. For both predictions of prevalent and incident DSPN, we only leveraged clinical and molecular features collected at the early time point of KORA F4.

Data pre-processing

From the KORA F4 study we obtained six types of molecular data, including genomic (Affymetrix Axiom), transcriptomic (Illumina HumanHT 12v3 Expression BeadChip), proteomic (SOMAscan), metabolomic (Metabolon), methylomic (Illumina Methylation 450k) and a small panel of inflammatory proteins (OLINK) data, besides clinical records. Each molecular layer was standardised before downstream analysis by computing the z-score, which accounts for different distributions and numerical scales of features. Our analysis pipeline pre-processed the data in a modality-specific manner, as shown below.

Processing of genomic data

Following microarray assay and initial imputation using the Haplotype Reference Consortium (HRC) as reference genome, the genomic dataset had 3,788 samples and 7,545,537 SNPs. We used PLINK v1.07 [56] for quality control of the genotype data. In particular, we removed SNPs that had equal or higher than 1% missing rate, less than 1% minor allele frequency (MAF) and significant deviation from Hardy Weinberg Equilibrium (HWE, p<1e-10). We used the –annotate function in the MAGMA software [57] to annotate the SNPs to their associated genes, based on the gene location information from the human genome GRCh37, considering SNPs that locate 2Mb upstream and 500b downstream of the genes. Following that we discarded SNPs that could not be annotated to a gene. We removed samples that had heterozygosity rates deviating more than three standard deviations from the mean across all samples. Finally, we filtered samples that had clinical records in the KORA F4 study. Eventually, the pre-processed genomic dataset included 1,083 samples and 3,167,521 SNPs. We transformed the categorical SNP data into continuous alternative allele copy numbers (0, 1 or 2).

Processing of transcriptomic data

The initial data generation, quality control and transformation were performed by the KORA study [58,59]. Specifically, the annotation of probes sequences to known transcripts was based on an annotation file provided by Illumina for HumanHT 12v3 BeadChip (using genome location of hg19). Only probes with the label "good" during mapping (probe sequence mapped uniquely to UCSC transcript) were included in this study. Furthermore, samples with less than 6,000 detected probes were removed using Illumina's GenomeStudio. The data was log2 transformed and quantile-normalised using Bioconductor package *lumi* [60]. The samples were clustered using R and the outliers were removed. We obtained 993 samples and 48,804 transcripts for our analysis. The technical variables including amplification plate, RIN number and sample storage time were regressed out using the R package *limma* [17].

Processing of proteomic data

The SOMAscan proteomic data was obtained from the KORA F4 study, including 1,000 individuals and 1,129 protein probes. One individual and 34 probes were removed due to low quality in accordance with the SomaLogic pipeline. Many probes mapped to multiple proteins/UniProt IDs so we transformed probe annotation into protein annotation. We also filtered for samples that had clinical records. In total, the dataset included 397 individuals and 1,160 proteins.

Processing of metabolomic data

The Metabolon metabolomic data obtained from the KORA F4 study included 1,768 individuals and 525 metabolites, after initial quality control and transformation. Particularly, the data was $\log_{10}$ transformed and values that lied more than four standard deviations from the mean were set to missing. We additionally discarded metabolites that had more than 70% missing values. For the remaining metabolites we imputed missing values using k nearest neighbour algorithm. Furthermore, we discarded samples that had standardised Mahalanobis distance larger than 4 and samples that did not have clinical records. Finally, we leveraged 829 samples and 466 metabolites.

Processing of methylomic data

The Illumina 450k Methylation M-value data was obtained from the KORA F4, which had already undergone filtering for detection rate and data normalisation. The original data had 1,727 samples and 485,512 methylation probes. We leveraged the ChAMP pipeline for methylation data processing [61]. Specifically, we excluded probes spanning SNP regions and probes not associated to genes based on the Illumina annotation file. Then, we imputed missing data using k nearest neighbour algorithm. Finally, technical effects were regressed out using the ComBat function in the *sva* package [62]. Only samples with clinical records were included in this study. In total, we had 849 samples and 399,541 methylation probes for our analysis.

Processing of inflammatory protein data

The OLINK inflammation panel included 92 inflammatory proteins which were measured in 1,133 samples. We additionally removed 21 proteins due to low detection quality, as reported in our previous study [25]. In summary, we used 71 proteins for our analysis.

Processing of clinical data

Clinical data obtained from the KORA F4 study included background information, diabetes and comorbidity status, lifestyle, blood biochemistry and medication usage for 1,161 individuals [24,51]. Together with the filtering mentioned in the “Population data” section, there were 1,133 samples remaining. Categorical variables were transformed using one hot encoding. Subsequently, variables having >10% missing values were discarded. In total, we leveraged 1,133 individuals and 83 variables.

Data partitioning for modality-specific feature selection

Each sample that was lacking at least one data modality was leveraged for modality-specific feature selection. For prevalent DSPN prediction included 710 genomic (141 cases and 569 controls), 621 transcriptomic (133 cases and 488 controls), 67 proteomic (9 cases and 58 controls), 476 metabolomic (76 cases and 400 controls), 495 methylomic (82 cases and 413 controls) and 720 clinical (142 cases and 578 controls) samples. The incident DSPN prediction leveraged 223 genomic (57 cases and 166 controls), 171 transcriptomic (47 cases and 124 controls), 58 proteomic (13 cases and 45 controls), 160 metabolomic (30 cases and 130 controls), 174 methylomic (38 cases and 136 controls) and 242 clinical (63 cases and 179 controls) samples. During gene set enrichment analysis, 100 stratified resampled splits were created for each of the modality-specific dataset, except proteomics due to limited sample size. We used 80% and 20% of samples for feature selection / training and testing, respectively.

Data partitioning for final model training

Fully multi-modal characterised samples were used for final model training. For prevalent DSPN prediction, this was 285 samples (31 cases and 254 controls), whilst for incident DSPN prediction, it was 242 samples (54 cases and 188 controls). We created 100 stratified splits which leveraged 80% samples for feature integration / training, and the remaining 20% for model testing. We further partitioned the 80% training samples into stratified five folds for cross-validation. The cross-validation performance was used as a criterion for the FFS algorithm to select the optimal model. We never used any test data for neither model training nor tuning of model parameters.

Gene set enrichment analysis

For the gene set enrichment analysis (GSEA), we leveraged the Bioconductor *fgsea* R-package [63], which is a more computationally efficient implementation compared to the original method [18]. For ranking genes, we used the t-statistics of the differential expression analysis from the Bioconductor *limma* R-package [17], which estimated the univariate association of the genes to the phenotype using a linear model. For calculating the enrichment score (ES), we used gene sets from the Reactome database [27]. Finally, the p-values were adjusted for multiple hypothesis testing with false discovery rate (FDR) < 20%, which is a lenient threshold allowing the selection of features with lower effect size, which may add predictive value in multivariate models in later integration steps.

The mapping of biomolecules to Reactome was customised for each data modality. For transcriptomic and proteomic data, we used the Reactome gene set annotation with Entrez IDs. For metabolomic data, we used the Reactome metabolite set annotation based on ChEBI IDs.

For genomic data, we leveraged the MAGMA software [57] to estimate the gene effect and subsequently perform gene set analysis. First, we annotated SNPs according to nearby genes (2 kb upstream and 0.5 kb downstream), and consecutively used MAGMA to estimate the gene effect on the phenotype, taking into account the SNPs that were mapped to this gene. MAGMA estimated the gene effect by first conducting principal component analysis (PCA) using all SNPs linked to this gene, and afterwards used PCs to train a linear regression model predicting the phenotype. Finally, MAGMA computed the gene's p-value with F-test, and converted these to Z-values for the gene set analysis leveraging a linear regression model [57].

For methylomic data, we used the methylRRA method[64] to perform gene set enrichment analysis (GSEA) on the CpG probes. First, this required a differential expression analysis on the probes using the R package limma, followed by using the ranked list of p-values as input for methylRRA. To this end, methylRRA computed a p-value for each gene leveraging the ranking of all CpGs annotated to that gene by implementing Robust Rank Aggregation algorithm[26]. Consequently, the p-values were transformed into z-scores and were used for the GSEA to extract significant gene sets[64].

In all cases, we included the full set of Reactome signaling pathways at the lowest levels of pathway hierarchy to avoid redundancy, and at the time, ensure full unbiased coverage (**Supplementary Data 1**). Furthermore, except for the proteomic data, the GSEA was performed across the 100 stratified splits accounting for heterogeneity.

Robust Rank Aggregation

We leveraged the implementation of Robust Rank Aggregation of Kolde et al.[26]. The molecules/molecule sets were ranked according to p-values, leading to a different ranked list per cross-validation/resampling run. Then, the rank distribution of each molecule/molecule set across all lists was tested against the random ranking distribution generated by permutation with the null hypothesis that there was no difference between the two distributions. The p-values of the test were adjusted for multiple hypothesis testing by multiplying the number of tested lists and additionally adjusted for the number of tested molecules/molecule sets by Benjamini-Hochberg method.

Extraction of leading-edge genes

Leading-edge genes in upregulated gene sets are all genes from the beginning of the ranked gene list until the enrichment score (ES). In contrast, in case of down-regulated gene sets, leading-edge genes are from the ES to the end of the ranked gene list. Here, we leveraged 80% of all data for each of the 100 stratified resamples, did GSEA, extracted the leading edge molecules to train an elastic net model, and finally tested the model prediction on the left out remaining 20% of samples. For aggregating results of these 100 stratified resamples, we only considered predictive models (AUROC > 0.5), and leveraged a Robust Rank Aggregation (RRA) algorithm[26] with a false discovery rate (FDR) cutoff of 5%, which delivered a union of leading-edge gene sets. Afterwards, a GSEA was conducted on the union of leading-edge gene sets to extract the final consensus significant gene sets and leading edge molecules, which were subject to final model training.

Clinical feature selection

For select clinical variables, we leveraged elastic nets using the R package *caret* with an 80% and 20% split for training and testing, respectively. We used weighted log-loss as the performance metric for hyper parameter tuning. The feature importance of the models was evaluated using the magnitude of t-statistics. Features with zero t-statistic were omitted. Finally, we used RRA with FDR cutoff of 5% to aggregate the important features across the 100 bootstraps.

Iterative forward feature selection

The iterative forward feature selection (FFS) integrates multiple data modalities. It is based on 100 independent runs of five-fold cross-validation. We tuned elastic net's hyperparameters alpha and lambda by grid search of 20 alphas and lambdas in range [0,1], resulting in 400 parameter sets. The chosen hyperparameter combination was the one having best mean performance across 5-fold cross-validations. In each run, we randomly sampled 80% of the dataset to perform five-fold cross-validation and the performance was tested with the remaining 20% data. For each fold of model training, elastic net models with weighted log-loss function to overcome class imbalances were implemented. Within the inner loop, a 5-fold cross-validation selected the best data modality to add next. The prediction performance of the model was tested by predicting on the outer test set (20% samples). The prediction probabilities were calibrated using the Platt scaling method. In each step, the model adds the next best data modality based on increased performance until all data modalities are included.

Extracting feature importance

We selected the optimal number of modalities based on their testing AUROC distribution. For this, the chosen number of modalities had a significant improvement in testing AUROC distribution compared to the previous number, and no significant improvement could be observed in the later complexities

(Wilcoxon rank sum test, p<0.05). After choosing the optimal number of modalities, the important features in the 100 models of that number of modalities were aggregated using RRA. Consecutively, we used the selected features to train a final elastic net model on the whole training dataset using the same settings as previous steps. The feature importance of the final model was accessed by t-statistics of model parameters.

Benchmarking of feature selection and integration methods

For feature selection, we compared GSEA with the conventional thresholding methods. For feature integration, we benchmarked FFS, data concatenation and ensemble stacking approaches. Thus, in total there were six combinations of methods to compare. For the thresholding method, we implemented differential expression analysis using *limma* and selected features having p-values<0.05.

Regarding feature integration, we benchmarked the FFS with ensemble stacking and feature concatenation. The latter concatenated all features into a single matrix before training the model. The ensemble stacking approach leveraged 100 independent runs with stratified resampling. This is, we generated 100 sets of stratified resamples, each consisting of 80% training and 20% test set (i.e. outer loop). Within each 80% training set, we further divided the data into 5-fold cross validation sets (i.e. inner loop). For each iteration of the inner loop, we trained an elastic net model on four out of five validation sets and made predictions on the remaining validation set. After five iterations, we obtained the predictions of all samples for that inner loop (corresponding to the 80% training set from the outer loop). We used this together with the ground truth (80% outer train) to train an elastic net meta model in the outer loop, and consecutively tested the predictive performance on the remaining 20% test set. Importantly, the test sets were never used for any parameter optimisation nor training, and only leveraged for unbiased performance evaluation. This process was repeated for each data modality. For the feature analysis we implemented Robust Rank Aggregation on the meta models of the ensemble stacking across 100 resamplings, then extracted the individual feature importance.

Statistics and Reproducibility

The multiomic datasets were preprocessed by the KORA study using customised software mentioned above. The development of the computational framework and the statistical analyses were conducted using the R packages and independent software detailed above. To reproduced the analysis results, one can obtain the data from https://www.helmholtz-munich.de/en/epi/cohort/kora/kora-studienzentrum, following a data sharing agreement with the KORA study. Details about sample sizes, types of data and code availability could be found in **Methods, Data availability** and **Code availability**.

Data availability

The KORA F4/FF4 cohort dataset is not publicly available due to data protection agreement to protect patient confidentiality. However, data access for research purposes could be requested via the KORA platform at https://helmholtz-muenchen.managed-otrs.com/external/. Data of the results shown in the main figures could be found in **Supplementary Data 3**.

Code availability

The analysis code of this paper is available at https://github.com/phngbh/DSPN [65]. The developed IMML framework is available at https://github.com/phngbh/IMML [66].

Author contributions

C.H. and M.P.M conceptualised the project. H.P., B.T., J.A., G.K., M.W., C.G., A.P., K.S., G.J.B., W.R., H.G. and D.Z. acquired and pre-processed the raw data. P.B.H.N. performed exploratory data analysis, developed the machine learning framework and visualised results. P.B.H.N., D.G., H.M., C.H. and M.P.M. derived biological interpretation. P.B.H.N. wrote the manuscript. M.P.M., C.H. and M.R. revised the manuscript.

Acknowledgements

The KORA study was initiated and financed by the Helmholtz Zentrum München – German Research Center for Environmental Health, which is funded by the German Federal Ministry of Education and Research and by the State of Bavaria. Furthermore, KORA research was supported within the Munich Center of Health Sciences (MC-Health), Ludwig-Maximilians-Universität, as part of LMUinnovativ. The German Diabetes Center is supported by the Ministry of Culture and Science of the State of North Rhine-Westphalia and the German Federal Ministry of Health. This study was supported in part by a grant from the German Federal Ministry of Education and Research to the German Center for Diabetes Research (DZD). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Health Organization. Global health estimates 2016: deaths by cause, age, sex, by country and by region, 2000–2016. *Geneva: World Health Organization* (2018).
2. Shaw, J. E., Sicree, R. A. & Zimmet, P. Z. Global estimates of the prevalence of diabetes for 2010 and 2030. *Diabetes Res. Clin. Pract.* **87**, 4–14 (2010).
3. Bayliss, E. A., Steiner, J. F., Fernald, D. H., Crane, L. A. & Main, D. S. Descriptions of barriers to self-care by persons with comorbid chronic diseases. *Ann. Fam. Med.* **1**, 15–21 (2003).
4. Kerr, E. A. *et al.* Beyond comorbidity counts: how do comorbidity type and severity influence diabetes patients' treatment priorities and self-management? *J. Gen. Intern. Med.* **22**, 1635–1640

(2007).

5. Pop-Busui, R. *et al.* Diabetic Neuropathy: A Position Statement by the American Diabetes Association. *Diabetes Care* **40**, 136–154 (2017).
6. Herder, C., Roden, M. & Ziegler, D. Novel Insights into Sensorimotor and Cardiovascular Autonomic Neuropathy from Recent-Onset Diabetes and Population-Based Cohorts. *Trends Endocrinol. Metab.* **30**, 286–298 (2019).
7. Bönhof, G. J. *et al.* Emerging Biomarkers, Tools, and Treatments for Diabetic Polyneuropathy. *Endocr. Rev.* **40**, 153–192 (2019).
8. Bongaerts, B. W. C. *et al.* Older subjects with diabetes and prediabetes are frequently unaware of having distal sensorimotor polyneuropathy: the KORA F4 study. *Diabetes Care* **36**, 1141–1146 (2013).
9. Albers, J. W. & Pop-Busui, R. Diabetic neuropathy: mechanisms, emerging treatments, and subtypes. *Curr. Neurol. Neurosci. Rep.* **14**, 473 (2014).
10. Dai, X. & Shen, L. Advances and Trends in Omics Technology Development. *Front. Med.* **9**, 911861 (2022).
11. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biol.* **18**, 83 (2017).
12. Zhou, Z.-H., Chawla, N. V., Jin, Y. & Williams, G. J. Big Data Opportunities and Challenges: Discussions from Data Analytics Perspectives [Discussion Forum]. *IEEE Comput. Intell. Mag.* **9**, 62–74 (2014).
13. Holle, R., Happich, M., Löwel, H., Wichmann, H. E. & MONICA/KORA Study Group. KORA--a research platform for population based health research. *Gesundheitswesen* **67 Suppl 1**, S19–25 (2005).
14. Conesa, A. & Beck, S. Making multi-omics data accessible to researchers. *Sci Data* **6**, 251 (2019).
15. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.* **19**, A68–77 (2015).
16. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for

RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

17. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
18. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).
19. Li, Y., Wu, F.-X. & Ngom, A. A review on machine learning principles for multi-view biological data integration. *Brief. Bioinform.* **19**, 325–340 (2018).
20. Picard, M., Scott-Boyer, M.-P., Bodein, A., Périn, O. & Droit, A. Integration strategies of multi-omics data for machine learning analysis. *Comput. Struct. Biotechnol. J.* **19**, 3735–3746 (2021).
21. Sewell, M. Ensemble Learning. http://www.cs.ucl.ac.uk/fileadmin/UCL-CS/research/Research_Notes/RN_11_02.pdf (2011).
22. Menden, M. P. *et al.* Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nat. Commun.* **10**, 2674 (2019).
23. Costello, J. C. *et al.* A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.* **32**, 1202–1212 (2014).
24. Herder, C. *et al.* Proinflammatory Cytokines Predict the Incidence and Progression of Distal Sensorimotor Polyneuropathy: KORA F4/FF4 Study. *Diabetes Care* **40**, 569–576 (2017).
25. Herder, C. *et al.* A Systemic Inflammatory Signature Reflecting Cross Talk Between Innate and Adaptive Immunity Is Associated With Incident Polyneuropathy: KORA F4/FF4 Study. *Diabetes* **67**, 2434–2442 (2018).
26. Kolde, R., Laur, S., Adler, P. & Vilo, J. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* **28**, 573–580 (2012).
27. Fabregat, A. *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* **46**, D649–D655 (2018).
28. Ziegler, D. *et al.* Screening, diagnosis and management of diabetic sensorimotor polyneuropathy in clinical practice: International expert consensus recommendations. *Diabetes Res. Clin. Pract.* **186**, 109063 (2022).

29. Fakkel, T. M. *et al.* Risk Factors for Developing Diabetic Peripheral Neuropathy: a Meta-analysis. *SN Comprehensive Clinical Medicine* **2**, 1853–1864 (2020).

30. Román-Pintos, L. M., Villegas-Rivera, G., Rodríguez-Carrizalez, A. D., Miranda-Díaz, A. G. & Cardona-Muñoz, E. G. Diabetic Polyneuropathy in Type 2 Diabetes Mellitus: Inflammation, Oxidative Stress, and Mitochondrial Function. *J Diabetes Res* **2016**, 3425617 (2016).

31. Vikram, A., Tripathi, D. N., Kumar, A. & Singh, S. Oxidative stress and inflammation in diabetic complications. *Int. J. Endocrinol.* **2014**, 679754 (2014).

32. Pop-Busui, R., Ang, L., Holmes, C., Gallagher, K. & Feldman, E. L. Inflammation as a Therapeutic Target for Diabetic Neuropathies. *Curr. Diab. Rep.* **16**, 29 (2016).

33. Herder, C., Dalmas, E., Böni-Schnetzler, M. & Donath, M. Y. The IL-1 Pathway in Type 2 Diabetes and Cardiovascular Complications. *Trends Endocrinol. Metab.* **26**, 551–563 (2015).

34. Rohm, T. V., Meier, D. T., Olefsky, J. M. & Donath, M. Y. Inflammation in obesity, diabetes, and related disorders. *Immunity* **55**, 31–55 (2022).

35. Furman, D. *et al.* Chronic inflammation in the etiology of disease across the life span. *Nat. Med.* **25**, 1822–1832 (2019).

36. Agarwal, N. *et al.* SUMOylation of Enzymes and Ion Channels in Sensory Neurons Protects against Metabolic Dysfunction, Neuropathy, and Sensory Loss in Diabetes. *Neuron* **107**, 1141–1159.e7 (2020).

37. Mandel, N. & Agarwal, N. Role of SUMOylation in Neurodegenerative Diseases. *Cells* (2022).

38. Mett, J. & Müller, U. The medium-chain fatty acid decanoic acid reduces oxidative stress levels in neuroblastoma cells. *Sci. Rep.* **11**, 6135 (2021).

39. Shekhar, N., Tyagi, S., Rani, S. & Thakur, A. K. Potential of Capric Acid in Neurological Disorders: An Overview. *Neurochem. Res.* **48**, 697–712 (2023).

40. Singh, N. *et al.* Adrenic Acid-Derived Epoxy Fatty Acids Are Naturally Occurring Lipids and Their Methyl Ester Prodrug Reduces Endoplasmic Reticulum Stress and Inflammatory Pain. *ACS Omega* **6**, 7165–7174 (2021).

41. Falomir-Lockhart, L. J., Cavazzutti, G. F., Giménez, E. & Toscani, A. M. Fatty Acid Signaling Mechanisms in Neural Cells: Fatty Acid Receptors. *Front. Cell. Neurosci.* **13**, 162 (2019).

42. Tracey, T. J., Steyn, F. J., Wolvetang, E. J. & Ngo, S. T. Neuronal Lipid Metabolism: Multiple Pathways Driving Functional Outcomes in Health and Disease. *Front. Mol. Neurosci.* **11**, 10 (2018).

43. Kimura, I., Ichimura, A., Ohue-Kitano, R. & Igarashi, M. Free Fatty Acid Receptors in Health and Disease. *Physiol. Rev.* **100**, 171–210 (2020).

44. Tao, M., McDowell, M. A., Saydah, S. H. & Eberhardt, M. S. Relationship of Polyunsaturated Fatty Acid Intake to Peripheral Neuropathy Among Adults With Diabetes in the National Health and Nutrition Examination Survey (NHANES) 1999–2004. *Diabetes Care* vol. 31 93–95 Preprint at https://doi.org/10.2337/dc07-0931 (2008).

45. Durán, A. M., Lawrence Beeson, W., Firek, A., Cordero-MacIntyre, Z. & De León, M. Dietary Omega-3 Polyunsaturated Fatty-Acid Supplementation Upregulates Protective Cellular Pathways in Patients with Type 2 Diabetes Exhibiting Improvement in Painful Diabetic Neuropathy. *Nutrients* vol. 14 761 Preprint at https://doi.org/10.3390/nu14040761 (2022).

46. Haque, F. *et al.* Performance Analysis of Conventional Machine Learning Algorithms for Diabetic Sensorimotor Polyneuropathy Severity Classification. *Diagnostics (Basel, Switzerland)* **11**, (2021).

47. Shin, D. Y., Lee, B., Yoo, W. S., Park, J. W. & Hyun, J. K. Prediction of Diabetic Sensorimotor Polyneuropathy Using Machine Learning Techniques. *J. Clin. Med. Res.* **10**, (2021).

48. Kazemi, M., Moghimbeigi, A., Kiani, J., Mahjub, H. & Faradmal, J. Diabetic peripheral neuropathy class prediction by multicategory support vector machine model: a cross-sectional study. *Epidemiol. Health* (2016).

49. Jian, Y., Pasquier, M., Sagahyroon, A. & Aloul, F. A Machine Learning Approach to Predicting Diabetes Complications. *Healthc. Pap.* **9**, (2021).

50. Dagliati, A. *et al.* Machine Learning Methods to Predict Diabetes Complications. *J. Diabetes Sci. Technol.* **12**, (2018).

51. Rathmann, W. *et al.* Incidence of Type 2 diabetes in the elderly German population and the effect of clinical and lifestyle risk factors: KORA S4/F4 cohort study. *Diabet. Med.* **26**, 1212–1219 (2009).

52. Feldman, E. L. *et al.* A practical two-step quantitative clinical and electrophysiological assessment for the diagnosis and staging of diabetic neuropathy. *Diabetes Care* **17**, 1281–1289 (1994).

53. Martina, I. S., van Koningsveld, R., Schmitz, P. I., van der Meché, F. G. & van Doorn, P. A. Measuring vibration threshold with a graduated tuning fork in normal aging and in patients with polyneuropathy. European Inflammatory Neuropathy Cause and Treatment (INCAT) group. *J. Neurol. Neurosurg. Psychiatry* **65**, 743–747 (1998).

54. Boyraz, O. & Saracoglu, M. The effect of obesity on the assessment of diabetic peripheral neuropathy: a comparison of Michigan patient version test and Michigan physical assessment. *Diabetes Res. Clin. Pract.* **90**, 256–260 (2010).

55. Tesfaye, S. *et al.* Diabetic neuropathies: update on definitions, diagnostic criteria, estimation of severity, and treatments. *Diabetes Care* **33**, 2285–2293 (2010).

56. Purcell. PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.*

57. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).

58. Schurmann, C. *et al.* Analyzing Illumina Gene Expression Microarray Data from Different Tissues: Methodological Aspects of Data Analysis in the MetaXpress Consortium. *PLoS One* **7**, e50938 (2012).

59. Mehta, D. *et al.* Impact of common regulatory single-nucleotide variants on gene expression profiles in whole blood. *Eur. J. Hum. Genet.* **21**, 48–54 (2013).

60. Du, P., Kibbe, W. A. & Lin, S. M. lumi: a pipeline for processing Illumina microarray. *Bioinformatics* **24**, 1547–1548 (2008).

61. Morris, T. J. *et al.* ChAMP: 450k Chip Analysis Methylation Pipeline. *Bioinformatics* **30**, 428–430 (2014).

62. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).

63. Korotkevich, G. *et al.* Fast gene set enrichment analysis. *bioRxiv* 060012 (2021) doi:10.1101/060012.

64. Ren, X. & Kuan, P. F. methylGSA: a Bioconductor package and Shiny app for DNA methylation data length bias adjustment in gene set testing. *Bioinformatics* **35**, 1958–1959 (2019).

65. Nguyen, P. B. H. *phngbh/DSPN: DSPN Analysis*. (Zenodo, 2024). doi:10.5281/ZENODO.13646753.

66. wglaas, UlrichAsemann & Nguyen, P. B. H. *phngbh/IMML: IMML*. (Zenodo, 2024). doi:10.5281/ZENODO.13646779.

3.3 Leveraging pretrained large language model for prognosis of type 2 diabetes using longitudinal medical records

Leveraging pretrained large language model for prognosis of type 2 diabetes with longitudinal medical records

Phong B.H. Nguyen[1,2], Andreas Hungele[3,5], Reinhard W. Holl[3,5+], Michael P. Menden[1,2,4+]

[1]Computational Health Center, Helmholtz Munich, Munich, Germany
[2]Faculty of Biology, Ludwig-Maximilian University Munich, Munich, Germany
[3]Institute for Epidemiology and Medical Biometry, Ulm University, Ulm, Germany
[4]Department of Biochemistry and Pharmacology, University of Melbourne, Melbourne, Australia
[5]German Center for Diabetes Research (DZD), Munich-Neuherberg

[+]Corresponding authors

Abstract

Timely prognosis of type 2 diabetes-associated complications becomes crucial to drive intervention strategies, save time and cost and increase overall treatment satisfaction. The emergence of AI-driven large language models (LLM) provides an opportunity to gain insights into patients' journey and infer significant clinical features. However, adapting LLMs to the healthcare domain remains challenging due to the sparse numerical nature and large feature space of the longitudinal medical records. In this work we demonstrated the utility of LLM in medical time series prediction by proposing a novel approach. Particularly, we first preprocessed and concatenated a missing mask to the data, then constructed and prepended an embedding layer to a pretrained LLM and fine-tuned both components leveraging the preprocessed data. To demonstrate the model performance, we leveraged the DPV registry dataset consisting of real world medical records of 449 185 T2D patients. Results showed that the fine-tuned LLM model outperformed baseline models in predicting HbA1c and LDL levels in the next doctor visit, achieving Pearson's correlation coefficients of 0.749 and 0.754, respectively. In addition, the model showed robust long-term prediction when predicting HbA1c in the next 554.3 days (95% CI: [547.0, 561.5]), outperforming the prediction by last observation. Furthermore, integrated gradient feature analyses revealed the significance of specific visits and clinical features contributing to the prediction of both variables, pinpointing potential biomarkers for early interventions. Overall, the results showed the possibility to leverage the prediction power of LLM in T2D prognosis using sparse medical time series, assisting clinical prognosis and biomarker discovery, ultimately advancing precision medicine.

Introduction

Type 2 diabetes (T2D) is a chronic metabolic disorder characterized by insulin resistance and impaired insulin secretion, leading to elevated blood glucose levels [1]. This condition is associated with a myriad of micro- and macrovascular dysfunctions in the end organs such as cardiovascular diseases, neuropathy, nephropathy and retinopathy, which collectively contribute to significant morbidity and mortality worldwide [2]. The global prevalence of T2D has been increasing steadily, highlighting the urgent need for effective management strategies to mitigate its complications. Early prognosis of these complications is crucial as it enables timely interventions that can slow or prevent the progression of the disease, thereby improving patient outcomes and reducing healthcare costs.

Longitudinal medical records, which contain comprehensive and continuous patient data over time [3], are invaluable for the early prognosis of T2D. These records typically include a wide range of information such as demographics, medical history, laboratory test results, medication prescriptions and lifestyle factors, providing a holistic view of the patient’s health journey. For instance, the Framingham Heart Study has demonstrated the utility of longitudinal data in predicting the risk of cardiovascular diseases [4]. Similarly, the Diabetes Control and Complications Trial (DCCT) has shown that long-term monitoring of blood glucose levels can predict the onset of diabetic complications [5]. The availability of these datasets allows for the identification of patterns that may not be apparent in cross-sectional data, facilitating more accurate and individualized prognostic assessments. That being said, the substantial volume and complexity of such datasets complicate the effort to analyze and extract useful insights.

The advent of machine learning, particularly deep learning, has revolutionized the field of medical prognosis, especially in the analysis of complex and voluminous longitudinal medical data. Several variants of deep learning architectures have been employed to analyze and model such datasets such as feed-forward neural networks, convolutional neural networks and recurrent neural networks. One of the earlier works is Deepr, which encoded medical data as sentences of medical codes and mapped them using word embedding, before modeling using a convolutional neural network in an end-to-end prediction for hospital readmission [6]. In another work, the authors introduced RETAIN which used a learnable word embedding approach and a recurrent neural network (RNN) with an attention mechanism to model the temporal medical data and assist interpretation [7]. Several more recent deep learning methods have been developed using advances such as transformers. One notable method is Med-BERT, which applies a transformer-based architecture to encode and model electronic health records (EHR) data. Med-BERT effectively captures complex patterns and temporal dependencies within the data, showing significant improvements in various prediction tasks, including disease progression and hospital readmission [8]. Another study introduced BEHRT, which was inspired by the pretraining of the BERT language model. Particularly, BEHRT introduces positional embeddings to account for the time intervals between medical events and pretrained the model using mask language model on a large dataset, thus providing a more nuanced representation of patient history and improving prediction performance in multiple clinical scenarios [9].

While these studies showed superior performance of deep learning models on various clinical and healthcare tasks, they are not without limitations. Most of these studies demonstrated the prediction performance on specific datasets with relatively low number of variables as predictors. This does not reflect real world scenarios where healthcare datasets are usually filled with hundreds of variables coming from diverse sources such as diagnostic codes, laboratory tests, medication usage, lifestyle and demographic information. The models should be able to deal with this amount of features or go through intensive preprocessing. Another limitation lies in the fact that in order to train such complex models, a vast amount of data and computational resources are required. For example, models such as transformers could go up to millions of parameters, making it challenging to train and account for the important complex relationships.

Furthermore, sparsity of the data possesses another substantial challenge that was not exclusively tackled in most of the studies. Missingness is very common in healthcare datasets and could harbor important information for the prediction, thus needs to be handled carefully. A few methods have been proposed over the years to address the missingness in medical time series, by modeling the spatiotemporal relationships while reducing inflated errors due to long range sequences. For example, GRU-D utilizes gated recurrent units (GRUs) and introduces a decay mechanism to model the influence of missing values over time, enhancing the imputation accuracy for time series data [10]. On the other hand, BRITS employs a bidirectional RNN approach, incorporating both forward and backward temporal dynamics to handle irregular time series data more effectively [11]. Last but not least, NAOMI uses a non-autoregressive model that conditions on both past and future values, applying a multi-resolution strategy to iteratively refine the imputed values across different time scales [12]. While these methods demonstrated success in some cases, the complexity of such models makes it challenging to scale up to more sophisticated datasets. Furthermore, some inherent aspects of autoregressive models such as memory loss, exploding/vanishing gradients and irregular time intervals were not fully addressed, making them less feasible to implement.

Recently, pretrained large language models (LLMs) such as GPT-3 and BERT have shown promise in various natural language processing tasks, including those in the biomedical domain. These models, which are trained on vast amounts of text data, can be fine-tuned for specific tasks, including the analysis of longitudinal health records for prognostic purposes. For instance, Huang et al. fine-tuned the BERT model on clinical notes from EHRs to predict patient outcomes such as hospital readmission and mortality [13]. Similarly, GatorTron, a transformer model pretrained on large-scale clinical data, has shown efficacy in predicting disease progression and identifying high-risk patients by analyzing both structured and unstructured health data [14]. Another study introduced MedGPT, a model fine-tuned on longitudinal EHR data to predict the onset of complications in diabetic patients, outperforming traditional machine learning approaches [15]. Most of these studies pretrained a LLM from scratch using clinical data, while did not explore the possibility to fine tune pretrained ones. Furthermore, these studies mostly focused on unstructured clinical text, which is inline with the NLP nature of LLMs, they did not attempt to model the structured clinical data directly. In addition, the intepretability of these models was also overlooked.

In this study, we aimed to explore the utility of pretrained LLM on prognosis tasks for a large longitudinal medical dataset. To this end, we developed a novel approach to deal with the sparse structured numeric nature of the data and adapt it directly to a pretrained LLM without training from scratch. In particular, we implemented a minimal data preprocessing and transformation to deal with missingness, constructed and prepended a learnable embedding layer to the LLM architecture and finally trained both components simultaneously to predict prognosis endpoints. Furthermore, we also demonstrated the interpretability of such a model in clinical settings, making it more translational and assisting precise and personalized healthcare.

Results

This study utilized the data from the DPV Initiative, a project initiated by Ulm University and supported by the German Diabetes Center (DZD) and its partners. The large cohort of electronic health records consisted of 649 331 patients from over 400 treatment facilities in four European countries. From these, a subset of 449 185 type 2 diabetes (T2D) patients were selected, from which 50 183 samples that had at least 10 clinical visits recorded were further selected (**Fig. 1a**). This ensures that the dataset includes patients with sufficient historical data to inform the predictions. From this initial dataset, we performed intensive and specific data and feature processing for both of the target variables HbA1c and LDL (**Methods**). At the end, there were 36 733 and 15 139 samples for the prediction of HbA1c and LDL, respectively (**Fig. 1a**). Both target variables showed a relatively normal distribution of values across the patient population (**Fig. 1a**).

To prepare the data for the prediction, we defined important points along the patients' medical records. Each patient record is a timeline from the first visit to the target point, where there is the value to be predicted, in this study the patient records consisted of 10 to 50 visits. Along this timeline, prediction point is the time point before which the data is used to make predictions. The period from the first visit up until the prediction point is called feature window. During this time, 321 features were collected from patient records, such as diagnoses, demographics, physical measurements, laboratory measurements, medication usage and hospitalization, which were not recorded for all patients in all visits. The prediction window is defined as the time period from the prediction point till target point, indicating how far in the future the prediction is made (**Fig. 1b**). The means of record lengths for HbA1c and LDL samples are 2651 (95% CI [2633,2669]) and 2661 (95% CI [2642,2679]), respectively (**Fig 1c,d**).

The machine learning framework to adapt sparse multivariate medical time series to pretrained large language model

This machine learning framework was designed to handle sparse medical time series data and adapt it to a pretrained large language model to leverage the prediction power of such models in the specific domain of medical data. The framework aims to mitigate issues related to missing data, which is common in medical records, and transforming the data into a suitable format for pretrained language models.

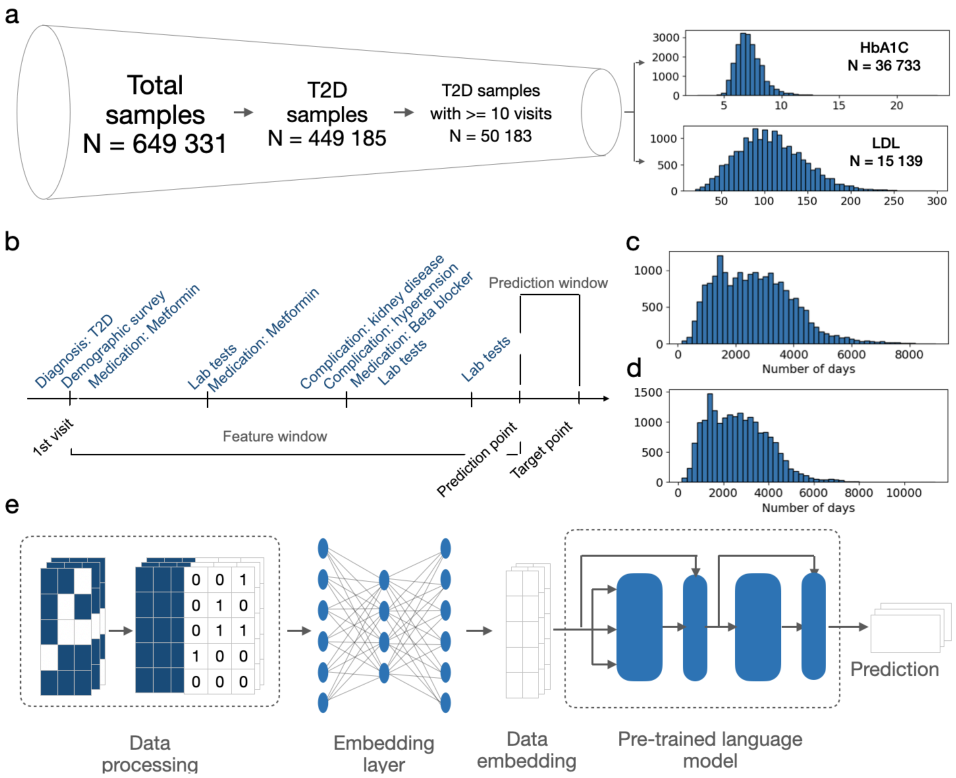


Figure 1. Summary of the data and machine learning framework to adapt sparse medical time series to pretrained large language models. (a) Samples were gone through a filtering process to select relevant samples for each prediction task. The histograms show the distribution of selected samples for HbA1c and LDL prediction. (b) Example of a patient medical record used in this study, from the first visit to the target point, feature window and prediction window are also defined. (c,d) Distributions of the lengths of the time series for HbA1c and LDL, respectively. (e) The machine learning framework started with processing the sparse medical data, which was then fed into an embedding layer that outputs a reduced representation of the data, which in turn was used as input for the pre-trained language model, which outputs the prediction (e).

The framework is a multi-step process to make prediction from raw medical data. The process begins with data processing, which includes the generation of a missing mask containing missing information of the variables and appending the missing mask to the original data matrix, retaining the missing information of the variables. The missing values were then filled with the means of the corresponding variables across time points. The processed data is then passed through an embedding layer, which reduces the dimensionality of the data while preserving its essential features. The embedding layer is crucial for handling the complexity and sparsity of the medical data, as it transforms the input into a more compact and meaningful representation, and at the same time makes it compatible for the pretrained language model. The embedded data is subsequently fed into a pretrained large language model. This model, which has been trained on vast amounts of textual data, is adapted to understand and predict outcomes based on the structured medical data provided (**Methods, Fig. 1e**). In this study we leveraged the framework to predict HbA1c and LDL levels either in the next visit or some visit in the future.

The LLM-based model trained on medical time series data is useful in case of longer prediction windows

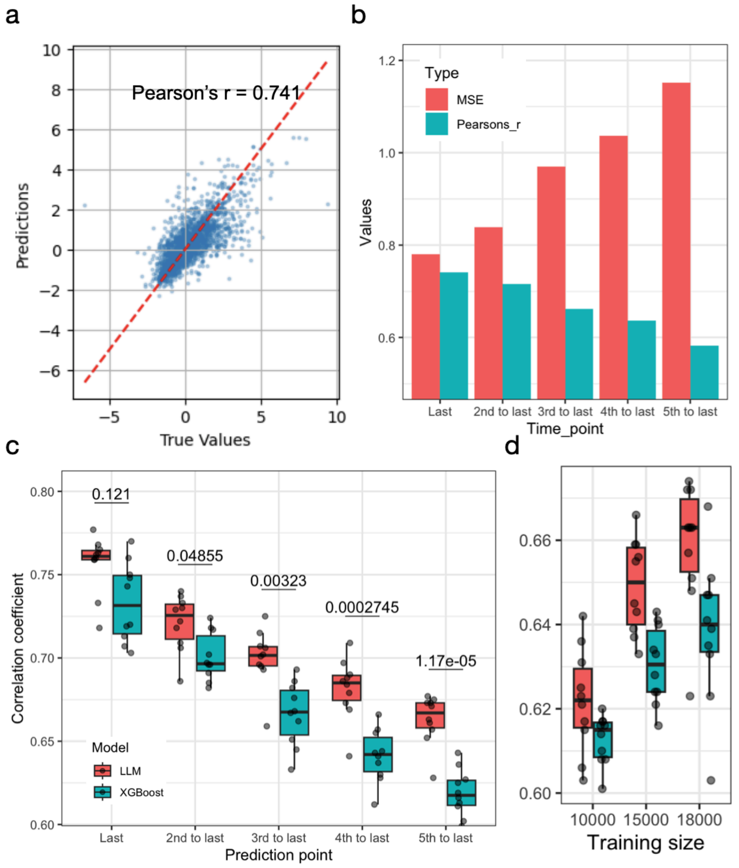


Figure 2. LLM-based model trained on time series outperformed the XGBoost trained on snapshot data. a) Correlation between observed data and predicted data using XGBoost trained on snapshot data of 17 500 samples at prediction point. Data points are results from a test set of 2500. b) Performance of the XGBoost model at different prediction points ("last" means last time point before the target point), in terms of mean squared error (MSE) and Pearson's correlation coefficient. c) Performance of the XGBoost model compared to the LLM-based model trained on time series at different prediction points, with 10 resamples. P-values are from Wilcoxon Rank Sum test. d) Performance of both models at different training size within a subset of 20 000 samples, at the prediction point of "4th to last", with 10 resamples.

Before training a deep learning model on time series data, we wanted to validate the utility of such model and data in clinical settings. Using a random training sample of 18 000, we observed that to predict the HbA1c level in the next doctor visit, an XGBoost model trained only on the data at the prediction point is sufficiently well performed. In particular, the model achieved a Pearson's correlation coefficient of 0.741 on the test set, indicating the sufficiency of using a shallow learner on the snapshot data at the prediction point (**Fig. 2a**). However, when we widened the prediction window by moving the prediction point further away from the target point, the prediction performance deteriorated drastically, in terms of mean squared error (MSE) and Pearson's correlation coefficient. Specifically, the MSE increased from 0.780 to 1.152, and the correlation coefficient reduced from 0.741 to 0.582, when shifting the prediction point four visits to the past (**Fig. 2b**). From this, we hypothesized that using deep learning trained on longitudinal data would improve the prediction in such cases by learning the complex patterns of historical data.

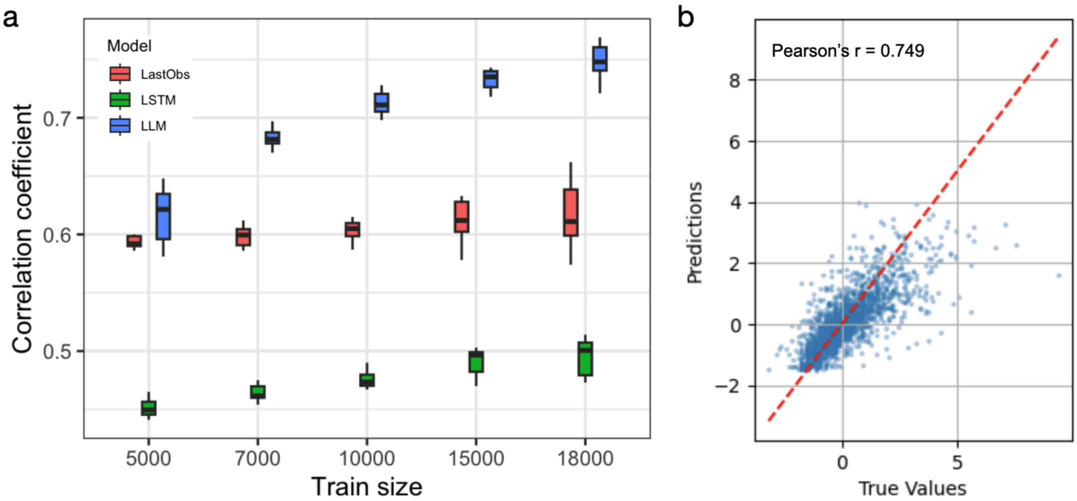


Figure 3. The LLM-based framework effectively predicted HbA1c level in the next visit. (a) Performance of the LLM-based framework was compared to the LSTM-based method and random forest model using only the last observed HbA1c value in the patient's record (LastObs), in terms of Pearson's correlation coefficient. (b) Predicted values of the HbA1c were compared against the true values in a testing dataset when using 18000 samples for the train set and 2000 samples for the test set.

To prove our hypothesis, we benchmarked the XGBoost model as described above with our LLM-based model trained using the medical time series data, using the same data set but at multiple different prediction points. In particular, we compared the prediction performance in predicting HbA1c of the two methods at different lengths of prediction window, ranging from the mean of 124.5 days (95% CI: [121.0, 126.8]) for predicting the next visit to 554.3 days (95% CI: [547.0, 561.5]) for predicting the fifth visit in the future (**Supp. Fig. S1**). Results showed that the LLM-based model showed significant improvement in prediction performance compared to the XGBoost model, especially in longer prediction windows. Specifically, in predicting HbA1c in the next visit, the LLM-based model showed marginally higher correlation coefficient compared to the XGBoost model, although the difference was not significant (Wilcoxon's Rank Sum Test, adjusted p-value = 0.121). However, when gradually increasing the prediction window, the prediction benefited more from the prediction power of LLM and the time series data, showing the significantly higher performance compared to the XGBoost model, with the lowest p-value at the widest prediction window (Wilcoxon's Rank Sum Test, adjusted p-value = 1.17e-05) (**Fig. 2c**). When predicting the HbA1c at the fifth visit in the future at different training sizes ranging from 10 000 to 18 000, we also observed the significantly higher performance of LLM-based models compared to the XGBoost models (**Fig.2d**). Overall, the results indicated the essentiality of our LLM-based model trained on time series data, especially in predicting further in the future.

The model could effectively predict HbA1c and LDL levels in the next doctor visit

To illustrate the performance of the derived framework, we used it to make predictions of HbA1c and LDL levels in the next doctor visit. We then compared the prediction performance with an LSTM-based model, where the data processing was similar but the pretrained LLM was replaced by an LSTM trained from scratch. In addition, we also compared the LLM-based model against a random forest model trained on the last observed value of the target variable and the length of the prediction window (LastObs). It is worth mentioning that for LLM- and LSTM-based models, we removed all target-related variables during training, including the target variables themselves, so that although the prediction task would be harder, we would be able to identify the contribution of important biomarkers for the specific targets. For all three methods, the training and testing were performed at different ratios of train and test sets in a subset of 20 000 samples. These comparisons would give detailed evaluation of the model performance in many aspects.

Using Pearson's correlation coefficient as the performance metrics, the LLM-based model consistently outperformed both the LSTM and LastObs models across all training sizes from 5000 to 18 000 samples out of the 20 000 dataset, for both HbA1c and LDL. In particular, for HbA1c prediction, the LLM-based models achieved the means of correlation coefficient from 0.618 (95%CI [0.601,0.653]) at 5000 training size to 0.749 (95%CI [0.738,0.759]) at 18 000 training size. These results significantly outperformed the LastObs models where the last observation of HbA1c and the prediction window were used in random forest training, as their performance did not show much improvement when increasing the sample size, ranging from 0.593 (95%CI [0.590,0.597]) to 0.619 (95% CI [0.598,0.640]). On the other hands, LSTM-based models where only variables that were not known to be related to the target variable were used to train did not learn very well the complex patterns of the time series and showed poor performance, achieving the mean correlations from 0.451 (95% CI [0.445,0.457]) to 0.495 (95% CI [0.484,0.507]) (**Fig. 3a**). Noticeably, increasing the number of samples used for training improved the performance of the LLM-based models more drastically, compared to other methods, especially at the lower sample sizes. At the sample size of 18 000, the predicted HbA1c values showed strong correlation to the true values (**Fig. 3b**).

In an additional analysis, we evaluated the performance of the LLM-based model in a post hoc setting. Particularly, we transformed the raw predicted values of HbA1c into its original range using the known mean and standard deviation of HbA1c in our dataset, then divided the samples into healthy (<7.5%) and unhealthy (>= 7.5%), then evaluated the classification with the observed HbA1c values. Fisher's Exact Test showed strong enrichment between the classes of the true and predicted values (p-value = 4.90e-170) (**Supp. Fig. S2**). The result further emphasized the superior performance and clinical relevance of our LLM-based model.

The benchmarking of LDL prediction models showed similar patterns as in the case of the HbA1c prediction. In particular, the mean correlation coefficients of the LLM-based models ranged from 0.650 (95% CI [0.639, 0.660]) to 0.749 (95% CI [0.738, 0.759]), from 5000 to 18 000 training size, respectively, which outperformed both LSTM-based models (**Supp. Fig. S3a**). At the training size of 18 000, the LLM-based models showed high correlation between the predicted and true values (**Supp. Fig. S3b**). The LastObs models, however, outperformed the LLM-based models at training size of 5000 with correlation coefficient of 0.692 (95% CI [0.690, 0.694]). Nevertheless, when increasing the sample size, the LLM-based showed significantly higher prediction performance than the LastObs models, especially from the sizes of 10 000 (Wilcoxon Rank Sum test, p-values < 0.05) (**Supp. Fig. S3a**).

The model gives valuable insights for prognosis of T2D

In order to interpret the model prediction and gain clinically relevant insights regarding T2D developments, we computed the integrated gradients, a robust algorithm to assess the feature importance in association with the target variables, for the prediction of HbA1c and LDL (**Methods**). Specifically, we computed the integrated gradients for each feature for a subset of 500 samples that went to 50 doctor visits, then aggregated these gradients across all samples and timepoints to identify the most important features, i.e. ones that have highest absolute gradients (**Methods**).

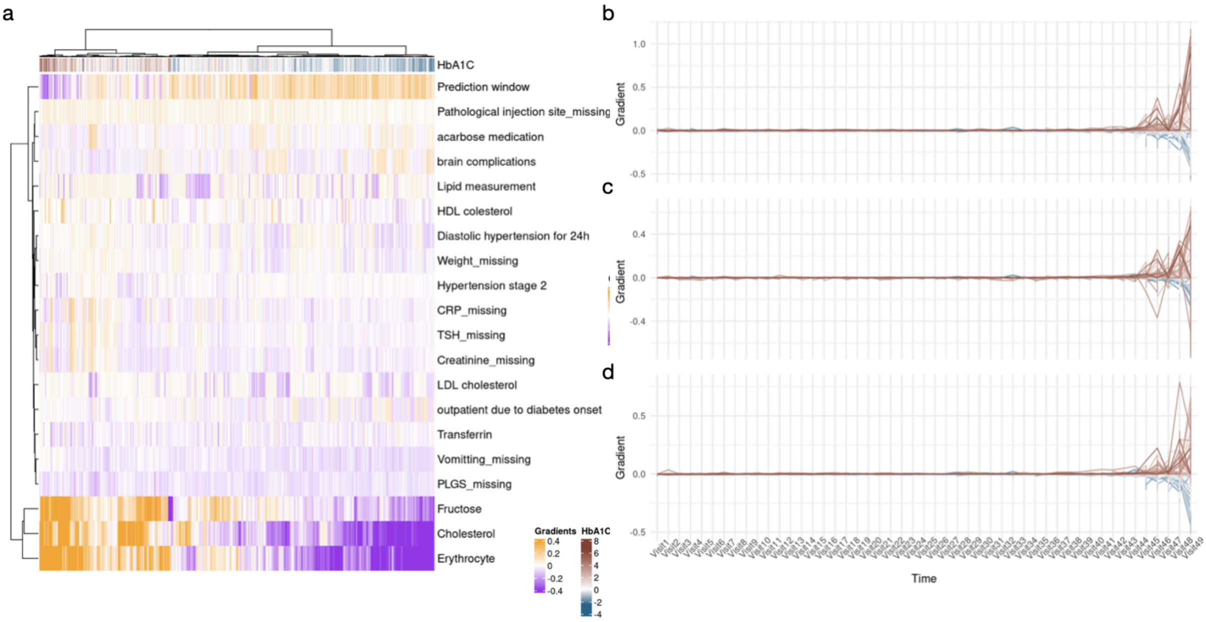


Figure 4. Most important features of HbA1c prediction. a) Heat map of feature importance score represented by integrated gradients of top 20 most important features. Values are the mean absolute gradients of the features across samples and time points. b-d) Trajectory of erythrocyte count, cholesterol and fructose levels across visits in the feature window, respectively. Each line represents one of 500 samples, colours represent the observed HbA1c level at the target point.

For HbA1c prediction, the integrated gradients gave insights into the importance of specific visits and individual features. In particular, most of the gradients concentrated at the last five visits before the target point, indicating the essentiality of these visits in providing necessary information for making the prediction at the target point (**Supp. Fig. S4a**). When analyzing the most important features in terms of absolute average gradient across samples and time points, the top features included erythrocyte level, cholesterol level, fructose level and prediction window (**Supp. Fig. S4b**). Specifically, the levels of erythrocyte, cholesterol and fructose were positively associated to the predicted HbA1c level (**Fig. 4a**), but mostly influence the prediction in the last five visits of the feature window (**Fig. 4b-d**). Besides, several features exhibited the association with HbA1c in their missing patterns. For instance, the missing information of thyroid stimulating hormone and C-reactive protein levels were among the important ones.

The integrated gradients also gave insights into important visits and features for the LDL prediction. Specifically, most of the gradients concentrated at the last ten visits before the target point, showing the importance of these visits for LDL prediction(**Supp. Fig. S5a**). When analyzing the most important features, the top features were thyroid stimulating hormone level, triglyceride level and missing number of prandial insulin injections (**Supp. Fig. S5b**). All of these features are positively associated with LDL level, especially in the last ten visits of the feature window (**Supp. Fig. S6**)

Discussion

In this work, we successfully introduced a novel approach to adapt sparse numerical data from a medical time series dataset to pretrained large language model (LLM), utilizing the prediction power of such models. Specifically, understanding that LLMs were specifically designed for language processing task, we alleviated the text prompt engineering step by first processing the sparse data to impute the missingness then added a learnable embedding layer after the input layer and before the pretrained LLM to feed the input data directly into the transformer architecture of the LLM. By using the learnable embedding layer, we successfully adapted the numerical input data and avoided the prompt engineering process, while effectively leveraging the prediction power of the LLM. Results of this paper show positive performance of this approach, in terms of prediction performance as well as clinical utility.

The LLM-based approach showed strong prediction performance in various aspects, exemplified by the forecasting of HbA1c and LDL. In particular, the model outperformed both LSTM-based and LastObs models in various training sizes, indicating the robust performance of the model even at lower number of training samples. This could be due to the fact that the LLM was already pretrained using much larger sample size, thereby reducing the number of samples, computational resources and time needed to fine-tune such model. Nevertheless, it’s worth noticing that increasing the number of training samples increased the performance drastically for the LLM-based models, especially at lower sample sizes, indicating that such models still benefit from the larger sample sizes to account for variability, especially in the case of clinical data where the levels of variation is high. In addition, the LLM-based models showed robust performance even for larger prediction windows. Particularly, they showed relatively high correlation coefficients when predicting HbA1c level in the next 18 months (554.3 days (95% CI: [547.0, 561.5])). This also highlighted the utility of historical data when predicting future values, as the LLM-based leveraging time series data outperformed the XGBoost model using only data at the prediction point. This demonstrates an important implication to clinical practice as historical data is relatively important to make decision and influence patient trajectory. Altogether, our LLM-based approach showed strong clinical utility in terms of prognostics for type 2 diabetes patients.

One important aspect of our model is the ability to give mechanistic insights into the pathological trajectory of the patients. For the model to be more accessible to patients and healthcare practitioners, we should be able to interpret its predictions. In this study, we interpreted the model prediction in two aspects, including important visits along the patient's medical history and individual influential features. To this end, we leveraged integrated gradients, a robust algorithm particularly useful for interpreting highly complex models such as LLMs (**Methods**). When averaging the gradients across all samples, we observed that not all visits were influential for prognosis of both HbA1c and LDL. It was interesting that only the last few visits (five for HbA1c and ten for LDL) of the feature window determined the predictions. From the clinical point of view, this implies that although historical data is important, only the doctor visits that are closer to the present are essential and the events that are further away are less informative.

Regardless of the visits, for HbA1c the most important features in terms of absolute gradients were erythrocyte, cholesterol and fructose levels. Erythrocytes (or red blood cells) are abundant in the blood and are greatly affected by blood glucose levels, particularly by the binding of glucose to hemoglobin A1C to form HbA1c [16]. It is interesting that the erythrocyte count seemed to be inversely associated with HbA1c levels, since the relationship between the two entities has not been straightforward, depending on many conditions [17]. More functional investigations are needed to validate this finding. On the other hand, cholesterol and fructose are well known markers that indirectly associate with HbA1c. Particularly, high cholesterol level (especially LDL) and high consumption of fructose might lead to insulin resistance, worsen glycemic control and finally lead to higher HbA1c level [18,19]. These detected important features showed potential prognostic and prevention biomarkers for T2D.

For LDL prediction, the most important features are thyroid stimulating hormone (TSH), triglycerides and the missing number of prandial insulin injections, all of which showed positive association to LDL. In conditions such as hypothyroidism, the level of TSH increases due to reduced function of thyroid, which leads to the reduced levels of multiple metabolic processes, including clearance of LDL cholesterol, leading to the increased level of LDL [20]. Triglycerides is a different class of lipids that is often associated with many metabolic disorders, which is also positively associated with LDL level [21]. Lastly, prandial insulin is usually injected to patients with diabetes to control blood glucose level, missing of which would lead to diabetes and dyslipidemia, which also increases the LDL level [22]. In summary, these important features represent potential diagnostic and prognostic biomarkers for LDL, as well as conditions such as dyslipidemia.

Despite valuable findings, there is still room for future research to improve the impact of this study. The putative biomarkers detected in this study need to be validated using an external dataset to validate the association with the target variables. In the future, molecular data could be integrated to further elucidate the mechanistic insights. In addition, the high level of missingness also limited the number of conditions we could predict. As more patients are being recruited, more conditions could be studied and the utility of the method could be validated more. Nevertheless, superior performance in HbA1c and LDL prediction served as a proof of concept for the method.

In conclusion, the study demonstrated a novel method to utilize the prediction power of pretrained LLMs in the healthcare domain, specifically in the prognosis of T2D patients using time series electronic health records. The LLM-based method exhibited outstanding performance in predicting HbA1c and LDL in the future doctor visit, as well as pinpointing important visits in the past and influential features as putative biomarkers for these conditions. These findings have strong clinical utility in diagnosis, prognosis and prevention strategy, further advancing precision medicine.

Methods

Study dataset, data processing and experimental design

The study was conducted under a data sharing agreement with the DPV Initiative, a project initiated by the German Diabetes Center (DZD) and executed and maintained by Ulm University, Germany, from which we obtained the data. The DPV Initiative has been collecting patient medical records since the 1990s from over 500 treatment facilities across Germany, Austria, Switzerland and Luxembourg. At the time of acquisition, the data consisted of 649 331 patients, of which 449 185 were of T2D. The number of doctor visits in T2D patients ranged from 1 to 241. To ensure sufficient information for model training, we selected patients who had at least 10 visits, which consisted of 50 183 patients. After accounting for the missing data in the target variables, the number of patients used for HbA1c and LDL predictions were 36 733 and 15 139, respectively. To reduce variability in time sequence lengths and increase the training sample size, for all the patients that had more than 50 visits, we extracted the sliding windows of 50 visits each, 10 visits per sliding step from the time sequences of these patients. Therefore, the number of visits of the final dataset ranged from 10 to 50. The final numbers of samples were 52 122 for HbA1c and 18 960 for LDL.

The variables were gone through intensive processing. Originally, there were 321 variables in the dataset, including demographics, lifestyle, physiological measurements, diagnosis, treatments, medications and other administrative records. All uniform, completely missing variables and ones with less than 5% data were removed. Inconsistent encodings were adjusted. Highly correlated and redundant variables were dropped. Categorical variables were transformed using one hot encoding. To deal with missing data, for numerical variables we filled the missing data points with the global mean of the corresponding variables, for categorical variables we encoded the missing values as "missing" before one hot encoding. Additionally, to retain the missing patterns which might be useful information during model training, we created a missing mask which had the same dimensions as the original data table but encoded '1' for missing data and '0' for present data. The mask was appended to the processed data table. At the end, the numbers of processed variables were 760 for HbA1c and 749 for LDL. All variables were standardized to normal distribution.

Benchmarking with different training sizes

To reduce the training time and computational resources required, if not indicated otherwise, we selected a random subset of 20 000 samples for HbA1c and all samples for LDL during our benchmarking experiments. Within the selected samples, we evaluated the prediction performance at different sizes of training set, including 5000, 7000, 10 000, 15 000 and 18 000 (only for HbA1c). The remaining samples were used for testing. For the deep learning models, 2000 samples within the training dataset were utilized for validation. Using the validation set, we implemented early stopping with a patience parameter of 30 epochs to avoid overfitting. To ensure the robustness, for each train-test ratio, we trained models in 10 resamples.

Benchmarking with different prediction windows

The prediction windows were created by moving the prediction point. In particular, we moved the prediction points backward from the visit preceding the target point to five visits away from the target point, creating the corresponding prediction windows. For each prediction window, we implemented the same training and evaluation scheme as with the benchmarking of training sizes above. However, since we moved the target point maximum five visits to the past, the sequence length was reduced by five, therefore the range of sequence lengths was now 5 to 45.

Adaptation of the pretrained LLM

To adapt the numerical input data to the pretrained LLM without having to engineer text prompts, we constructed a learnable embedding layer prepended to the pretrained LLM. The embedding layer was used to learn the suitable embeddings of the raw input data to be used as input for the LLM. In theory one can use any architecture that could be customized to output appropriate embeddings for the specific LLM. In our study, to simplify the training process, we implemented an multilayer perceptron (MLP) of two hidden layers, each with 50 nodes. The dimensions of input and output layers of the MLP could be customized depending on the input data and the internal architecture of the LLM. Input dimensions were 760 for HbA1c and 746 for LDL, output dimension was 768 for embedding dimension of BERT. Here we used BERT as the pretrained LLM for simplicity, but in theory one can use any pretrained LLM for their specific purpose. We plugged the embedding layer directly into the input encoder layer of BERT, bypassing the tokenizing and embedding steps of BERT. Both the embedding layer and BERT were trained simultaneously during model training.

The LLM was initiated with the pretrained parameters. During forward pass, to avoid overfitting, we implemented a dropout mechanism for both the embedding and BERT, in which 70% of the parameters were dropped out of the feedforward, only the remaining parameters were used to compute the output. We trained the models for a maximum 150 epochs, but with an early stopping mechanism. Particularly, the training would stop if there was not improvement in prediction performance on the validation set after 30 epochs. We used Adam optimizer with mean square error as loss function, learning rate of 2e-05 and weight decay of 5e-03. Pearson's correlation coefficient was used as prediction performance on the test set.

Implementation of long short term memory (LSTM)

The LSTM included two hidden layers of 100 dimensions. The training process was similar to the LLM-based method as the processed input data was fed directly into the LSTM. However, the LSTM model was initiated from scratch. The input and output dimensions, training hyperparameters and regularization mechanisms were the same as with the LLM-based method.

Implementation of XGBoost

The XGBoost models were trained with 100 boosted trees and mean squared error as the loss function. All other parameters were set as default from the XGBRegressor function of the xgboost library in Python. Here we only used the variables in the prediction time point as training features, therefore no time series was used. The rest of the training and testing scheme was the same as with LLM- and LSTM-based methods.

Integrated gradients and analysis of feature importance

Here we leveraged integrated gradients to interpret the model prediction. Integrated gradients is a feature attribution method that quantifies the contribution of each input feature to a model's prediction. It works by calculating the path integral of the gradients from a baseline input (usually zero) to the actual input. The method computes the gradients of the model's output with respect to the input at points along this path and averages them. By integrating the gradients, it captures how the prediction changes as the input moves from the baseline to its actual value, providing a more stable and accurate attribution of feature importance compared to simple gradient methods. Details could be found in the original publication [23].

To interpret the model prediction, we computed the integrated gradients for a subset of 500 samples that had 50 visits, using the trained model. We computed both total gradients and average absolute gradients across all samples. The average absolute gradients were used to analyze the influential visits, from there we identify most important individual features by summing up the gradients across time points.

Statistical and computational analysis of results

The analysis framework was developed using pytorch. Data processing, model training and evaluation of the results were conducted using various libraries from Python. Hypothesis testing for two continuous variables was done with the Wilcoxon Rank Sum test, p-values were adjusted for FDR. Hypothesis testing for two categorical variables was done with Fisher's Exact test. Data visualization was done using ggplot2 in R.

Code availability

The code for the analysis and computational framework is located here: https://github.com/phngbh/DPV.

Data availability

The data is not publicly accessible due to the patient's confidentiality. Researchers who would like to reproduce the results could contact the responsible person in the DPV Initiative in Ulm University. Detailed information can be found here: https://buster.zibmt.uni-ulm.de/.

References

1. Diabetes. https://www.who.int/news-room/fact-sheets/detail/diabetes.

2. Farmaki, P. *et al.* Complications of the type 2 diabetes mellitus. *Curr. Cardiol. Rev.* **16**, 249–251 (2020).

3. Lee, S. *et al.* Unlocking the Potential of Electronic Health Records for Health Research. *Int J Popul Data Sci* **5**, 1123 (2020).

4. Kannel, W. B. Some lessons in cardiovascular epidemiology from Framingham. *Am. J. Cardiol.* **37**, 269–282 (1976).

5. Nathan, D. M. & for the DCCT/EDIC Research Group. The Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and complications study at 30 years: Overview. *Diabetes Care* **37**, 9–16 (2014).

6. Nguyen, P., Tran, T., Wickramasinghe, N. & Venkatesh, S. Deepr: A Convolutional Net for Medical Records. *arXiv [stat.ML]* (2016).

7. Choi, E. *et al.* RETAIN: An interpretable predictive model for healthcare using REverse Time AttentIoN mechanism. *arXiv [cs.LG]* (2016).

8. Rasmy, L., Xiang, Y., Xie, Z., Tao, C. & Zhi, D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit. Med.* **4**, (2021).

9. Li, Y. *et al.* BEHRT: Transformer for electronic health records. *Sci. Rep.* **10**, 7155 (2020).

10. Che, Z., Purushotham, S., Cho, K., Sontag, D. & Liu, Y. Recurrent neural networks for multivariate time series with missing values. *Sci. Rep.* **8**, (2018).

11. Cao, W. *et al.* BRITS: Bidirectional recurrent imputation for time series. *arXiv [cs.LG]* (2018).

12. Liu, Y., Yu, R., Zheng, S., Zhan, E. & Yue, Y. NAOMI: Non-Autoregressive Multiresolution Sequence Imputation. *arXiv [cs.LG]* (2019).

13. Huang, K., Altosaar, J. & Ranganath, R. ClinicalBERT: Modeling clinical notes and predicting hospital readmission. *arXiv [cs.CL]* (2019).

14. Yang, X. *et al.* GatorTron: A large clinical language model to unlock patient information from

unstructured electronic health records. *bioRxiv* (2022) doi:10.1101/2022.02.27.22271257.

15. Kraljevic, Z. *et al.* MedGPT: Medical concept prediction from clinical narratives. *arXiv [cs.CL]* (2021).
16. Wang, Y. *et al.* The relationship between erythrocytes and diabetes mellitus. *J. Diabetes Res.* **2021**, 6656062 (2021).
17. English, E. *et al.* The effect of anaemia and abnormalities of erythrocyte indices on HbA1c analysis: a systematic review. *Diabetologia* **58**, 1409–1421 (2015).
18. American Diabetes Association. Dyslipidemia management in adults with diabetes. *Diabetes Care* **27**, s68–s71 (2004).
19. Stanhope, K. L. Role of fructose-containing sugars in the epidemics of obesity and metabolic syndrome. *Annu. Rev. Med.* **63**, 329–343 (2012).
20. Duntas, L. H. & Brenta, G. A renewed focus on the association between thyroid hormones and lipid metabolism. *Front. Endocrinol. (Lausanne)* **9**, 511 (2018).
21. Parhofer, K. G. & Laufs, U. The diagnosis and treatment of hypertriglyceridemia. *Dtsch. Arztebl. Int.* **116**, 825–832 (2019).
22. Slattery, D., Amiel, S. A. & Choudhary, P. Optimal prandial timing of bolus insulin in diabetes management: a review. *Diabet. Med.* **35**, 306–316 (2018).
23. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic Attribution for Deep Networks. *arXiv [cs.LG]* (2017).

4 Discussion

This thesis represent key research findings supporting biomarker discovery for complex multifactorial diseases such as cancer and diabetes, in both aspects of biomedical insights and methodological development. While these findings potentially serve as the basis to translate biomedical research and advance towards precision medicine, there remain challenges and limitations for future research to tackle. This chapter summarizes key messages that have been elaborated through in the previous chapter and consolidates final remarks for this thesis.

4.1 Importance of involving comprehensive data in biomarker discovery for complex diseases

The integration of multiomic and clinical data is essential for biomarker discovery in the context of complex diseases, as it provides a comprehensive and nuanced understanding of the multifaceted nature of these conditions. Complex diseases, such as cancer, diabetes, and cardiovascular diseases, arise from intricate interactions between molecular, environmental and lifestyle factors. Multiomic data offers detailed insights into the various biological layers that contribute to disease pathology. By integrating these diverse datasets, researchers can capture the complexity of biological systems and identify intricate interactions and pathways that may not be evident when examining a single omic layer alone. On the other hand, clinical data adds another critical dimension by providing patient-specific information, such as demographics, medical histories, treatment responses, and outcomes. This rich, real-world context is crucial for validating the clinical relevance and utility of potential biomarkers.

Combining data from various modality enables the identification of biomarkers that are both biologically meaningful and clinically actionable. This holistic approach enhances the predictive power and specificity of biomarkers, facilitating the development of personalized therapeutic strategies tailored to individual patient profiles. In the realm of complex diseases, such integrated approaches are particularly valuable, as they help unravel the underlying heterogeneity and pave the way for targeted interventions that can improve patient outcomes. Ultimately, the convergence of multiomic and clinical data fosters a deeper understanding of disease mechanisms, driving more effective and personalized healthcare solutions.

4.1.1 Demographics, ancestry in particular, associates with cancer drug response

Demographics, particularly ancestry information, play a crucial role in cancer research, especially in understanding and predicting cancer drug responses. Genetic diversity across different populations can significantly influence the efficacy and safety of cancer therapies. Ancestry-specific genetic variations can affect drug metabolism, efficacy, and toxicity, leading to variations in treatment outcomes. For instance, certain genetic mutations that influence drug response may be more prevalent in specific ancestral groups. Incorporating ancestry information in cancer research allows for the identification of these population-specific genetic markers and the development of tailored therapeutic strategies.

In this thesis, I explore the distribution of genetic ancestry among various cancer cell lines and its implications for cancer research. In particular, I utilized a Bayesian method to infer the ancestry of over 1000 cancer cell lines, classifying them primarily into five main groups: European, East Asian, African, American and South Asian. The study highlighted the discrepancy in ancestral distribution across cancer types, in which European and East Asian accounted for a vast majority of cells, 63.1% and 25%, respectively, with smaller proportions from African, American, and South Asian ancestries. The distribution of ancestries in cancer cell lines likely mirrors ethnic variations in cancer incidence and prevalence, influencing the choice of cell line models for research. This finding highlights the need for diverse representation in cancer research to better understand drug responses across different populations.

More importantly, the study found notable differences in how cancer cell lines from European and East Asian ancestries responded to PI3K/mTOR inhibitors. In particular, Asian cell lines appeared to be significantly more sensitive to PI3K/mTOR then Caucasian ones, especially in glioblastoma. This aligns with clinical trial findings that Asian patients experience greater toxicity from PI3K/mTOR inhibitors compared to European patients. This suggests that genetic ancestry can influence drug efficacy and underscores the importance of considering ancestral background in preclinical drug testing. Despite this, the influence of ancestry on drug sensitivity and the underlying molecular mechanisms remain underexplored. The findings emphasize the potential of integrating genetic ancestry into precision medicine approaches, which could lead to more effective and personalized cancer treatments, reducing the failure rates of clinical trials by tailoring interventions to genetic backgrounds.

4.1.2 Multiomic data is essential for prognosis of complex diseases

Type 2 diabetes and its complications as well as other metabolic conditions are heterogeneous diseases characterized by complex underlying molecular mechanisms. Thus, it is hypothesized that incorporating molecular data together with clinical information facilitates the identification of novel biomarkers, elucidates disease mechanisms, and enhances the development of targeted therapies, ultimately leading to more precise and effective interventions for such diseases.

In this thesis I demonstrated how molecular data was leveraged to study complex metabolic diseases. The subject of this study is distal sensorimotor polyneuropathy (DSPN), a neuropathy condition occuring in both diabetic and non-diabetic individuals. Leveraging a comprehensive machine learning framework on a prospective population dataset, the study found that addition of multiomic data on top of regular clinical data can significantly improve the prognosis of the participants. In particular, the comprehensive model outperformed the clinical model in predicting the future incident of DSPN for healthy individuals (6.5 ± 0.2 years), achieving an AUROC of 7.1. The findings imply the detectable changes in metabolic mechanisms before the onset of observable clinical symptoms.

The study essentially revealed important putative biomarkers of DSPN incidence. Specifically, inflammatory proteins and metabolites were significant contributors to the model's predictive power. The analysis identified 26 important features, including 17 inflammatory proteins, four metabolites, three transcripts, and two clinical variables. These findings highlight the critical role of inflammation and metabolic alterations in the development of DSPN and underscore the potential of using molecular biomarkers for early prediction and prevention strategies. Ultimately, the findings highlight the need of multiomic data and its integration in biomarker discovery.

4.2 Interpretable machine learning frameworks help translate biomarker research

Interpretable machine learning plays a crucial role in biomarker discovery by providing transparency and insight into the predictive models used in biomedical research. Rid of black-box models, interpretable machine learning allows researchers, clinicians and patients to understand how input features contribute to the prediction of disease outcomes. This transparency is essential for validating the relevance and reliability of identified biomarkers, ensuring that they are not only statistically significant but also biologically and clinically meaningful. By elucidating the relationships between multiomic data and clinical phenotypes, interpretable models facilitate the identification of novel biomarkers that can improve early diagnosis, prognostic assessments, and personalized treatment strategies. Furthermore, the ability to interpret machine learning models helps build trust among stakeholders, including clinicians and patients, and supports regulatory approval processes by demonstrating the scientific rigor behind biomarker discoveries. Ultimately, the integration of interpretability in machine learning enhances the translation of theoretical findings into practical medical applications, advancing the field of precision medicine.

4.2.1 Linear model coupled with biology-driven feature selection for biomarker discovery

Biological datasets are usually enriched with noise on top of relevant signal of interest. This issue is further worsened by the extremely imbalanced in data dimensions, where the number of samples is incomparable to the huge feature space. Thus, an efficient method to filter through the feature space and select relevant features for biomarker analysis is much needed, either serving as the basis for exploratory analysis or part of an end to end predictive modeling.

Here I introduce IMML - a robust two-step end to end machine learning framework for predictive modeling and biomarker discovery in complex diseases, showcased by the analysis of distal sensorimotor polyneuropathy (DSPN). The first step of this framework emphasizes the importance of integrating prior biological knowledge during feature selection, through the use of gene set enrichment analysis (GSEA). The GSEA feature aims to map heterogeneous molecular variables to an interpretable biological space, by identifying the molecule sets that are enriched and associate with the phenotype of interest. By leveraging GSEA, the framework enriches the feature selection process with insights from existing biological knowledge, which improves the interpretability and relevance of the selected features. This method allows for a more targeted approach in identifying potential biomarkers that are intricately linked to the underlying biology of DSPN.

The use of interpretable models ensures that the relationships between the features (biomarkers) and the disease outcomes are clearly understood, facilitating the validation and application of these biomarkers in clinical settings. The impact of this framework is significant, as it not only enhances the predictive performance of the model but also leads to the discovery of actionable biomarkers that can be directly applied in precision medicine. The discovery of these biomarkers, such as the up-regulation of inflammatory proteins, down-regulation of SUMOylation pathways and essential fatty acids, provides new avenues for early diagnosis, prognosis, and personalized treatment strategies for DSPN, ultimately advancing the field of precision medicine with reliable and interpretable findings.

4.2.2 Post-hoc interpretability for fine-tuned large language model

Deep learning (DL) models have demonstrated remarkable predictive power, particularly in complex and high-dimensional datasets typical of healthcare. However, the interpretability of these models poses significant challenges due to their intricate architectures, often seen as ”black boxes.” Interpretability in DL is crucial as it fosters trust and transparency, essential for clinical decision-making and regulatory approval. Understanding how DL models make predictions enables clinicians to validate and leverage these models effectively, ensuring patient safety and adherence to medical standards. As discussed in section 1.4.4, to enhance interpretability, several approaches have been developed. Techniques such as Shapley additive explanations (SHAP) and local interpretable model-agnostic explanations (LIME) provide insights into feature importance and model decision processes. Additionally, methods like attention mechanisms and integrated gradients offer a more nuanced understanding of the model's focus and the contribution of individual features to predictions.

For instance, the machine learning framework introduced in section 3.3 painted out an elegant way to access the large language model's internal work, with integrated gradients. In particular, the computed gradients of the features revealed significant variables essential for the prediction, at specific time points. As an example, erythrocyte concentration, fructose and cholesterol level in the last five visits (approximately two years) before the target time point strongly influenced the model prediction for HbA1C percentage. This information does not only help researchers and physicians understand the model's behaviors and validate its prediction in clinical perspective, but also drive decision making for early intervention to improve patients' prognosis. By making DL models more interpretable, we can bridge the gap between advanced computational methods and practical healthcare applications, ultimately making these powerful tools more translatable and actionable in clinical settings. This not only enhances the reliability of the models but also ensures that they can be seamlessly integrated into patient care, driving forward the field of precision medicine.

4.3 Novel method contributions

Chapter 3 introduces several contributions in novel methodology for advancing biomarker discovery research in multiple settings. Here is a summary of a few key methods.

4.3.1 The Bayesian ancestry inference framework

In section 3.1, we derived a novel Bayesian method to infer the ancestry of cancer cell lines (CCLs). To this end, this method utilizes genotype data of the ancestry-associated variants to estimate the probability that each cell line belongs to various ancestral populations. Essentially, the process involved associating the genetic profiles of the CCLs to reference populations with known ancestries. By calculating the likelihood of each cell line belonging to specific ancestral groups, we effectively assigned ancestry labels such as European, East Asian, African, American, and South Asian. This probabilistic framework allowed for a nuanced classification that could capture the complex admixture patterns present in many cell lines. The inferred ancestries were validated against known ancestries from an independent dataset and showed high accuracy, providing a reliable basis for further analysis.

The accurate inference of ancestry in CCLs has significant implications for cell line characterization and biomarker analysis. Understanding the ancestral background of CCLs enhances the ability to select appropriate models for studying cancer biology and drug responses. It helps in identifying population-specific genetic variants and molecular pathways that might influence disease progression and treatment efficacy. For instance, the study found that Asian CCLs exhibited higher sensitivity to certain PI3K/mTOR inhibitors compared to Caucasian CCLs. This suggests that ancestry-specific biomarkers can be critical in predicting drug response and tailoring treatments to diverse patient populations. In sum, this methodological advancement underscores the importance of incorporating genetic diversity into cancer research to develop more effective and inclusive therapeutic strategies.

4.3.2 The interpretable and comprehensive machine learning framework for predictive modeling

An end-to-end, easy-to-implement predictive modeling for multiomic integration and biomarker discovery is essential in precision medicine. To this end, the IMML (Interpretable Multimodal Machine Learning) method proposed in section 3.2 integrates multiomic and clinical data to enhance the prediction of disease incidence and prevalence, specifically distal sensorimotor polyneuropathy (DSPN). It employs a two-step approach: first, using gene set enrichment analysis (GSEA) for feature selection to incorporate prior biological knowledge, and second, selective integration of the features in a robust model training framework. This method improves predictive performance with ΔAUROC > 0.1 for the prediction of incident DSPN and ensures that identified biomarkers are biologically relevant and clinically actionable.

The IMML method has significant implications for understanding disease mechanisms and discovering prognostic and predictive biomarkers. By integrating multiomic and clinical data, it provides a comprehensive view of the biological processes involved in complex diseases like DSPN. The use of GSEA during feature selection enriches the analysis with biological insights, enabling the identification of key pathways and molecular interactions. This method not only enhances the accuracy of disease prediction but also aids in identifying actionable biomarkers that can inform personalized treatment strategies, ultimately advancing precision medicine.

4.3.3 Healthcare domain adaptation framework for finetuning pretrained large language model

As elaborated in section 3.3, domain adaptation is essential to leverage the prediction power of large language models (LLMs) for sparse clinical data in healthcare prognosis. Specifically for type 2 diabetes prognosis, we first preprocessed the data to address the missingness issue, then leveraged a fine-tuned BERT (bidirectional encoder representations from transformers) LLM model prepended with a learnable linear embedding layer for nummerical data adaptation. In particular, each data point was transformed into a fixed-size embedding through the linear embedding layer, effectively reducing the dimensionality while preserving the essential information and omitting the need for text prompt construction and engineering. Both the embedding layer and the LLM model were then fine-tuned on the transformed time series data. The LLM's self-attention mechanism captures the complex temporal dependencies and interactions between different clinical measurements. Additionally, dropout and weight decay are employed during training to prevent overfitting, ensuring the model generalizes well to unseen data.

Applying the novel machine learning framework for type 2 diabetes prognosis represents a significant contribution in disease prediction and management. By capturing intricate temporal patterns and relationships within the patient data, this model can provide highly accurate predictions about disease progression. Particularly, the model achieved the mean Pearson's correlation efficient of 0.749 (95%CI [0.738,0.759]) for HbA1C, outperforming conventional methods. Such precise prognostic tools are invaluable in clinical settings, enabling healthcare providers to identify high-risk patients early and tailor interventions accordingly. Moreover, the interpretability of the model can offer insights into which clinical features and time periods are most influential in predicting disease progression, providing valuable feedback for both medical practitioners and patients in understanding and managing the disease.

4.4 Limitations and outlook

In general, this thesis represents significant scientific contributions in predictive modeling and biomarker discovery for complex diseases, further advancing precision medicine. While the studies highlight innovative approaches, they bring to light several limitations and areas for future development. Addressing these challenges and further development of research in this area would transform precision medicine into a truly translational field and benefit patients.

4.4.1 Enrichment of data annotation in multiple modalities to assist biomarker discovery

Enriching the availability of data in precision medicine is crucial for advancing the field, yet several challenges hinder progress. From the preclinical perspective, one significant challenge is the lack of annotated demographic information, sparsity in molecular characterization and and other pathological information in cancer cell lines. Comprehensive databases such as Cell Model Passport encompass molecular characterization, pathological information and drug response information of cancer cell models which serve as useful resources for preclinical study (117). That being said, efforts are being made to complete the annotation for these cell line models. This limitation is particularly critical as precision medicine relies on tailoring therapies to individual patient profiles, which requires comprehensive datasets that include detailed annotations on ancestry, lifestyle factors, and clinical histories. In particular, findings from Section 3.1. suggest that ancestral information plays a crucial role in differential drug response, especially between East Asian and Caucasian populations. Therefore, leveraging this information in preclinical study would greatly enhance the predictive performance.

Another challenge is the incomplete functional annotation of biological molecules in various multiomic datasets. Despite the growing use of multiomics to explore complex biological systems, many molecules remain poorly characterized, which hampers the identification of meaningful biomarkers. The vast amount of data generated by high-throughput technologies often outpaces our ability to accurately annotate and interpret these molecular features, leading to a gap between data generation and actionable insights. For instance, while databases like Reactome and KEGG provide extensive information on molecular pathways and interactions, they do not cover the full spectrum of molecules identified in high-throughput studies (118, 119). As pointed out in Section 3.2., the lack of functional annotation in multiple modalities such as genomics and metabolomics reduced the number of analyzed features of the framework significantly, complicating the model interpretation and subsequent biomarker discovery. In genomics, not all genetic variants have known functions or clinical relevance, leading to challenges in interpreting how specific mutations might contribute to disease. Similarly, in metabolomics, a substantial proportion of detected metabolites remain unannotated or poorly characterized, limiting our understanding of their roles in cellular processes and disease mechanisms.

To address these challenges, future research must focus on several key areas. Firstly, there is a need for standardized protocols for data collection, annotation, and sharing across laboratories and institutions. This would ensure that demographic, molecular, and clinical data are consistently recorded and made available for research. Initiatives like global biobanks and consortiums that aggregate data from diverse populations could play a pivotal role in enriching datasets. Secondly, advances in bioinformatics and computational biology are essential for improving the annotation of biological molecules in multiomic studies. Developing more sophisticated algorithms and machine learning models to predict the function and significance of uncharacterized molecules could bridge the gap between data generation and biological understanding. Additionally, collaborative efforts to expand databases like Reactome and KEGG with new experimental data are essential. Furthermore, initiatives to create new databases specifically focused on underrepresented molecules or to integrate existing databases for a more holistic view of molecular interactions could significantly advance the field. By addressing these challenges, the availability of robust, annotated data will significantly improve, paving the way for more accurate biomarker discovery and the advancement of precision medicine, which will ultimately lead to more effective, personalized treatments that consider the full spectrum of individual patient characteristics.

4.4.2 Completion of sample characterisation to increase prediction power

Integrating clinical data more effectively into precision medicine research is a complex but crucial endeavor. Currently, clinical datasets are often fragmented, with inconsistent formats, missing data, or variable accuracy, which hinders comprehensive analyses. This leads to the lack of depth and breadth required to capture the complexity of diseases, particularly when it comes to correlating molecular profiles with clinical outcomes. This sparsity is compounded by the challenges of integrating data from different sources, such as electronic health records, biobanks, and research studies, which may have inconsistent formats, missing information, or varying degrees of accuracy. These challenges were evidenced by the clinical datasets introduced in Section 3.2 and 3.3, where the data modality were distributed unevenly across samples and missingness was at higher rate. These data incompleteness complicated the analysis process and reduced prediction power of the models.

To overcome these challenges, the development of interoperable platforms that facilitate seamless data sharing and analysis across different healthcare systems is essential. For instance, creating standardized data repositories that can aggregate and harmonize data from diverse sources—such as electronic health records, biobanks, and research studies—would vastly improve the depth and quality of available clinical information. Additionally, incorporating more comprehensive molecular profiling of patients during routine care could significantly enrich clinical datasets. For example, routinely collecting genomic, proteomic, and metabolomic data alongside traditional clinical measures would provide a more detailed and actionable dataset for researchers. This would not only enhance the precision of biomarker discovery but also enable more personalized treatment strategies that are based on a deeper understanding of the patient's unique molecular and clinical profile.

Efforts to standardize protocols for data generation and annotation across laboratories and institutions are also critical for advancing precision medicine. Currently, the lack of standardized protocols means that data generated in different studies or at different institutions may not be directly comparable, which limits the ability to integrate and analyze data at a broader scale. To address this, international collaborations and initiatives aimed at establishing universal standards for data handling and sharing should be prioritized. These efforts would not only improve the reliability and reproducibility of research findings but also enhance the ability to identify and validate biomarkers that are relevant across different populations. In turn, this would help bridge the gap between research and clinical application, making precision medicine more effective and accessible to a broader range of patients.

4.4.3 Functional validation of putative biomarkers to translate biomedicine

Functional validation of putative biomarkers is a crucial step in translating biomedical research findings into clinical applications. For example in Section 3.2, where biomarkers for distal sensorimotor polyneuropathy (DSPN) were identified, including inflammatory cytokines, metabolic pathways, and genetic markers, and in section 3.3, where erythrocyte, fructose and cholesterol were identified as putative biomarkers of future HbA1C level, functional validation ensures that these biomarkers are not only statistically significant but also biologically relevant and clinically actionable.

The process of functional validation typically involves several key steps. First, replicating the findings in independent cohorts can confirm the reliability of the biomarkers across different populations and settings. This helps ensure that the biomarkers are not artifacts of a specific dataset or analytical method. For instance, in the DSPN study, the validation of biomarkers such as proinflammatory cytokines would require further studies to confirm their role in the disease process and their predictive power for DSPN onset. In addition to replication, functional validation often involves experimental studies designed to test the biological roles of the biomarkers. This could include in vitro experiments, such as using cell cultures to examine the effects of specific cytokines on neuronal cells, or in vivo studies, where animal models are used to explore how manipulating these biomarkers affects disease progression. For example, we could use genetically modified mice to overexpress or knock out specific inflammatory cytokines identified as biomarkers in the DSPN study, observing the resultant impact on nerve function and disease development. In the context of cholesterol, fructose, and erythrocyte levels identified as biomarkers for future HbA1C level (a measure of long-term blood glucose levels), functional validation would involve exploring the mechanistic links between these biomarkers and glucose metabolism. For instance, we might investigate how elevated cholesterol levels influence insulin resistance or how fructose metabolism affects HbA1C levels. Additionally, understanding the role of erythrocytes in glucose regulation could provide insights into how changes in red blood cell function impact HbA1C predictions.

These experimental approaches are complemented by computational methods that can model the interactions between different biomarkers and their impact on disease pathways, broadening the understanding of the disease's mechanisms. By integrating multiomic data, we can build network models that predict how changes in one biomarker might influence others and contribute to disease progression. These models can then be tested and refined through further experimental studies, ultimately leading to a deeper understanding of the disease mechanisms.

Functional validation also plays a crucial role in determining the clinical utility of biomarkers. This involves assessing whether the biomarkers can be reliably measured in a clinical setting, whether they provide additional information beyond existing diagnostic tools, and whether they can be used to guide treatment decisions. For the DSPN biomarkers, this could mean developing assays that can measure the levels of specific cytokines in patient blood samples and testing whether these levels correlate with disease progression or response to treatment.

In summary, functional validation of putative biomarkers is an essential process that bridges the gap between discovery and application. By confirming the biological relevance, experimental reproducibility, and clinical utility of biomarkers, researchers can ensure that their findings lead to meaningful advances in patient care. The examples of inflammatory cytokines in DSPN and cholesterol, fructose, and erythrocyte levels in HbA1C prediction highlight the importance of this process in translating research discoveries into tools for precision medicine.

A Abbreviations

A.1 List of Acronyms

WHO	World Health Organization
NSCLC	Non-small cell lung cancer
SCLC	Small cell lung cancer
CML	Chronic myeloid leukemia
TKI	Tyrosine kinase inhibitor
FDA	U.S. Food and Drug Administration
TCGA	The Cancer Genome Atlas
GDSC	Genomics of Drug Sensitivity in Cancer
CTRP	Cancer Therapeutics Response Portal
T790M	Gatekeeper mutation of the epidermal growth factor receptor
T1D	Type 1 diabetes
T2D	Type 2 diabetes
ADA	American Diabetes Association
EASD	European Association for the Study of Diabetes
PMDI	Precision Medicine in Diabetes Initiative
NIDDK	U.S. National Institute of Diabetes and Digestive and Kidney Diseases
JDRF	Juvenile Diabetes Research Foundation
HbA1c	Glycated hemoglobin
SNP	Single nucleotide polymorphism
NGS	Next generation sequencing
RNA-seq	RNA sequencing
MS	Mass spectrometry
NMR	Nuclear magnetic resonance
EHR	Electronic health record
t-SNE	t-distributed stochastic neighbor embedding
LASSO	Least absolute shrinkage and selection operator
RCT	Randomized control trial
EMA	European Medicine Agency
CCLE	The Cancer Cell Line Encyclopedia
AI	Artificial Intelligence
PCA	Principal Component Analysis

UMAP	Uniform Manifold Approximation and Projection
DNN	Deep neural network
LC-MS	Liquid chromatography mass spectrometry
SVM	Support vector machine
MSE	Mean squared error
RMSE	Root mean squared error
PCR	Polymerase chain reaction
ELISA	Enzyme-linked immunosorbent assay
CNN	Convolutional neural network
RNN	Recurrent neural network
LLM	Large language model
SHAP	Shapley Additive Explanation
LIME	Local Interpretable Model-agnostic Explanation
CCL	Cancer cell line
ICGC	International Cancer Genome Consortium
HTS	High throughput drug screen of cancer cell lines
HLA	Human Leukocyte Antigen
MHC	Major Histocompatibility Complex
COSMIC	Catalogue of Somatic Mutation in Cancer
DepMap	The Cancer Dependency Map Project
NCI60	A group of 60 human cancer cell lines used by the NCI for screening of compounds
ANOVA	Analysis of variance hypothesis testing
FDR	False discovery rate
GBM	Glioblastoma
COREAD	Colon and rectum adenocarcinoma
LGG	Brain Lower Grade Glioma
PI3K/mTOR	PI3K/AKT/mTOR intracellular signaling pathway
ALL	Acute lymphoblastic leukemia
1000G	1000 Genome Project
ESCA	Esophageal carcinoma
STAD	Stomach adenocarcinoma
DSPN	Distal sensorimotor polyneuropathy
KORA	Collaborative Health Research in the Region of Augsburg
AUROC	Area under the ROC curve
MNSI	Michigan Neuropathy Screening Instrument
GSEA	Gene set enrichment analysis

DEA	Differential expression analysis
FFS	Forward feature selection
IQR	Interquartile range
CI	Confidence interval
RRA	Robust Rank Aggregation
BMI	Body mass index
SUMO	Small ubiquitin-related modifier
TCA cycle	Tricarboxylic acid cycle
OLINK	Protein profiling platform using proximity extension assay
LDL	Low-density lipoprotein
DCCT	Diabetes Control and Complications Trial
NLP	Natural language processing
BERT	Bidirectional Encoder Representations from Transformers
RETAIN	Reverse Time Attention model
BEHRT	Implementation of BERT for EHR data
GRU	Gated Recurrent Unit
LSTM	Long Short-Term Memory
BRITS	Bidirectional Recurrent Imputation of Time Series data
NAOMI	Non-Autoregressive Multiresolution Sequence Imputation
GPT-3	Generative Pre-trained Transformer 3
XGBoost	Extreme Gradient Boost
TSH	Thyroid stimulating hormone

A.2 List of Proteins

ABL	Tyrosine-protein kinase ABL1
BCR	Breakpoint cluster region protein
HER2	Receptor tyrosine-protein kinase erbB-2
EGFR	Epidermal growth factor receptor
p53	Tumor protein p53
PI3K	Phosphoinositide 3-kinase
mTOR	Mammalian target of Rapamycin
ERBB2	HER2
ERBB4	Erb-B2 Receptor Tyrosine Kinase 4
GPCR	G protein-coupled receptors

GLP-1	Glucagon-like peptide-1
IL-6	Interleukin 6
CXCL9	Chemokine (C-X-C motif) ligand 9
CXCL10	Chemokine (C-X-C motif) ligand 10
CCL13	Chemokine (C-C motif) ligand 13
CCL19	Chemokine (C-C motif) ligand 19
CCL20	Chemokine (C-C motif) ligand 20
CDCP1	CUB domain containing protein 1
SLAMF1	Signaling lymphocytic activation molecule 1
TNFRSF9	TNF receptor superfamily member 9
TNFRSF11B	TNF receptor superfamily member 11B
CD5	Cluster of differentiation on the surface of T cells
FFAR	Free fatty acid receptors

A.3 List of Genes

NF1	Neurofibromin 1
MLL2	Lysine methyltransferase 2D
PIK3R1	Phosphoinositide-3-kinase regulatory subunit 1
CDC42	Cell division control protein 42 homolog
SP3	Sp3 Transcription Factor
ITSN1	Intersectin 1

A.4 List of Metabolites

AMPA	alpha-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid
PUFA	Polyunsaturated fatty acid
ARA	Arachidonic (omega-6) acid
DHA	Docosahexaenoic (omega-3) acid
MCFA	Medium-chain fatty acid

B Supplementary information to section 3.1

Inferred Ancestral Origin of Cancer Cell Lines Associates with Differential Drug Response

Phong B. H. Nguyen, Alexander J. Ohnmacht, Samir Sharifli, Mathew J. Garnett and Michael P. Menden

Corresponding author: Menden MP (michael.menden@helmholtz-munich.de)

The Supplementary Data for this manuscript include Supplementary Table S1 to S3 could be found on the online version of this manuscript

C Supplementary information to section 3.2

The Interpretable Multimodal Machine Learning (IMML) framework reveals pathological signatures of distal sensorimotor polyneuropathy

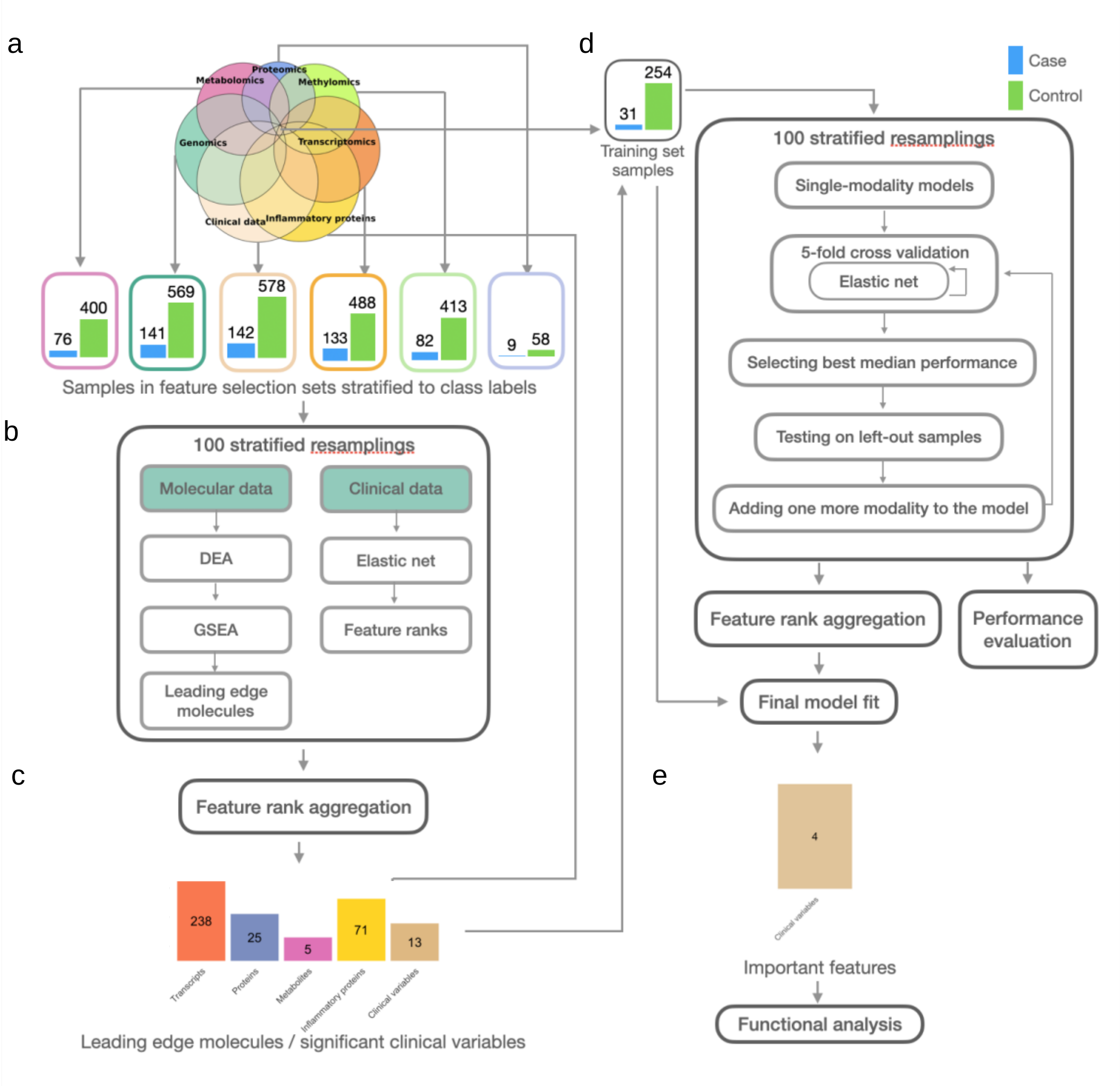
Phong BH Nguyen, Daniel Garger, Diyuan Lu, Haifa Maalmi, Holger Prokisch, Barbara Thorand, Jerzy Adamski, Gabi Kastenmueller, Melanie Waldenberger, Christian Gieger, Annette Peters, Karsten Suhre, Gidon J Boenhof, Wolfgang Rathmann, Michael Roden, Harald Grallert, Dan Ziegler, Christian Herder, Michael Menden

Corresponding author: Menden MP (michael.menden@helmholtz-munich.de) and Herder C (christian.herder@ddz.de)

This section consists of the Supplementary Information for this manuscript including Supplementary Figures S1 to S18 and Supplementary Table S1 to S3.

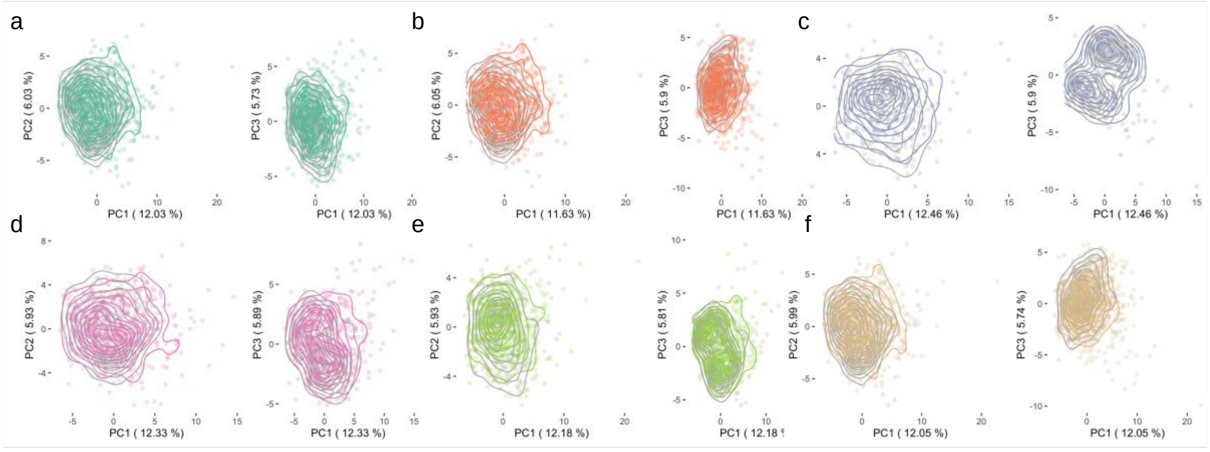
The Supplementary Data 1 to 3 could be found on the online version of this manuscript.

Supplementary Information

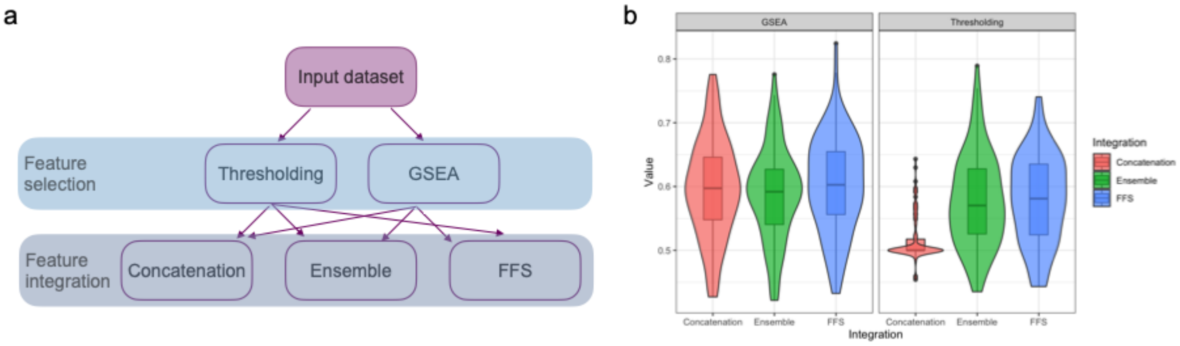


Supplementary Figure S1. Analysis pipeline and results for cross-sectional DSPN

a. Features in modality-specific datasets were selected independently using non-overlapping modality-specific samples. The selected features stratified into case and control are shown in the barplots.
b. Molecular data went through differential expression analysis (DEA) which generated a molecule list sorted by t-statistics which was then used as input for gene set enrichment analysis (GSEA). GSEA output leading edge genes which drive the enrichment of their respective gene sets. Clinical features was selected by training elastic net models and extracting important features. The process was repeated using 100 stratified resamplings.
c. The final significant list of molecules and clinical variables were selected using a rank aggregation algorithm.
d. After feature selection step, the selected features were then integrated to train models to predict DSPN, using the left-out overlapping dataset (training set). The training aimed to determine the optimal complexity and composition of the models by implementing elastic net with forward feature selection in a nested cross-validation manner, using weighted log loss as performance metric to account for class imbalance. We used 100 stratified resamplings during training and the rank aggregation at the end to select the most stable model.

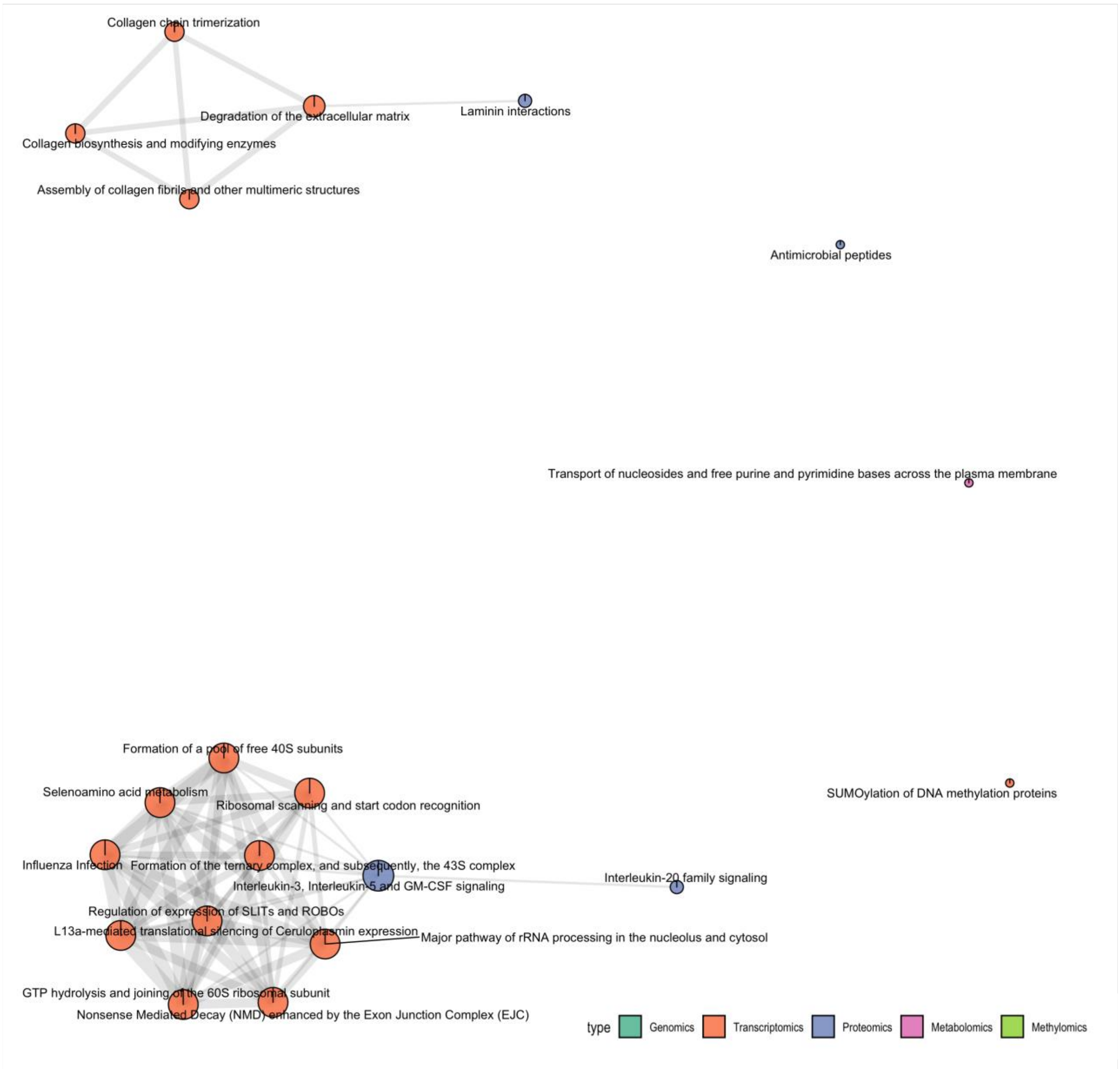


Supplementary Figure S2. PCA of clinical features for feature selection and model training datasets. Grey contour plots highlight the model training sets, whilst other colors indicate the feature selection set of the different modalities: (a) Genomics, (b) Transcriptomics, c) Proteomics, (d) Metabolomics, (e) Methylomics and (f) Clinical data.



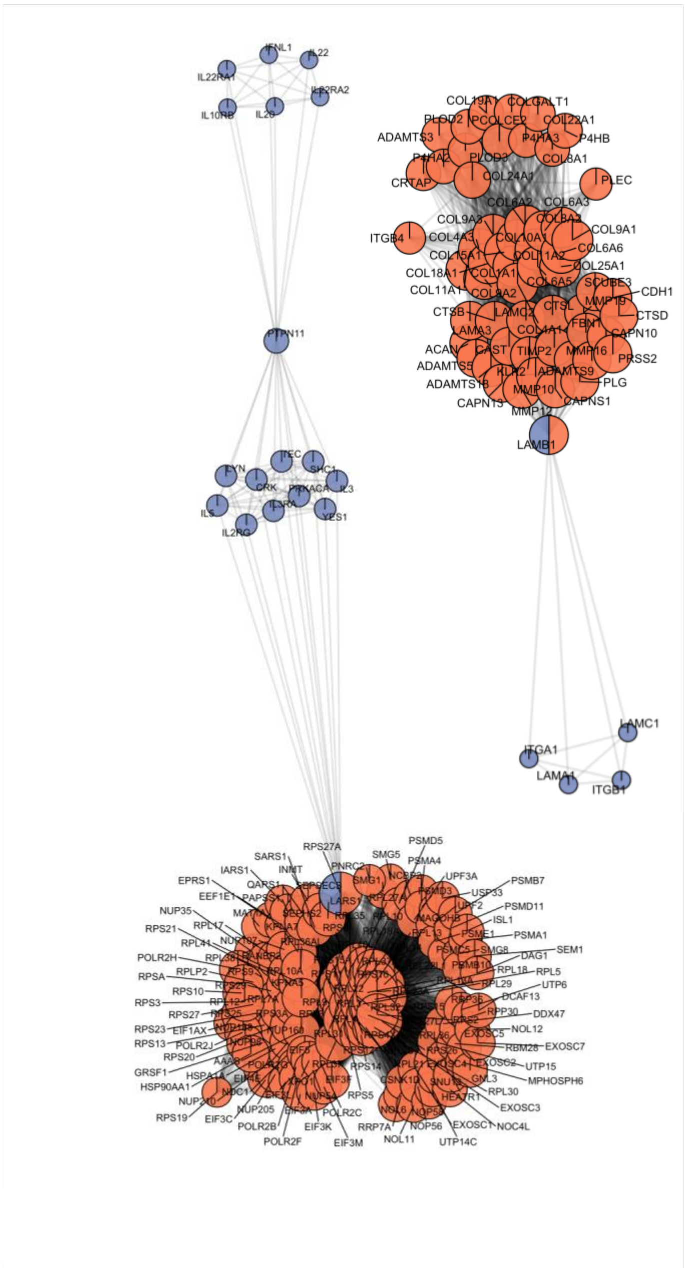
Supplementary Figure S3. Benchmarking of feature selection and integration methods

a. Illustration of methods for feature selection (thresholding and GSEA) and feature integration (concatenation, ensemble and our FFS algorithm) in a conventional multi-modal machine learning process. Arrows show the possible trajectory of the process in which different combinations of these methods could be used.
b. Benchmarking result showing prediction performance on the test set of different selection-integration methods for incident DSPN prediction using transcriptomic, proteomic, metabolomic and clinical data. Distributions of AUROC for the matched 100 stratified resamplings are shown in the y-axis and different methods are shown on the x-axis.


Collagen chain trimerization
Degradation of the extracellular matrix
Laminin interactions
Collagen biosynthesis and modifying enzymes
Assembly of collagen fibrils and other multimeric structures
Antimicrobial peptides
Transport of nucleosides and free purine and pyrimidine bases across the plasma membrane
Formation of a pool of free 40S subunits
Selenoamino acid metabolism
Ribosomal scanning and start codon recognition
SUMOylation of DNA methylation proteins
Influenza Infection
Formation of the ternary complex, and subsequently, the 43S complex
Interleukin-3, Interleukin-5 and GM-CSF signaling
Interleukin-20 family signaling
Regulation of expression of SLITs and ROBOs
L13a-mediated translational silencing of Ceruloplasmin expression
Major pathway of rRNA processing in the nucleolus and cytosol
GTP hydrolysis and joining of the 60S ribosomal subunit
Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC)
type
Genomics
Transcriptomics
Proteomics
Metabolomics
Methylomics

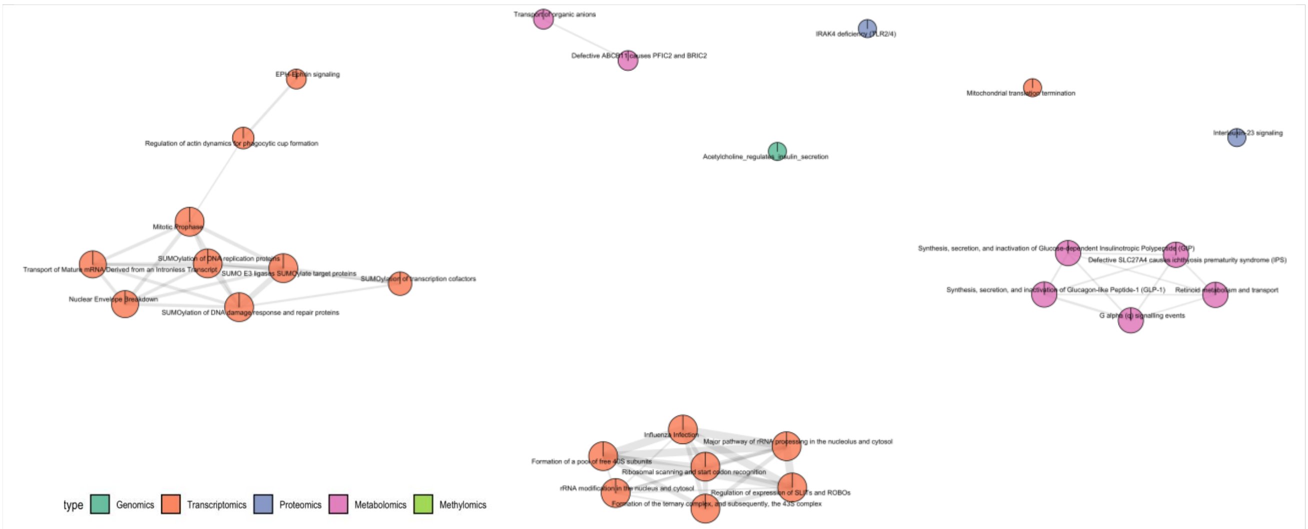
Supplementary Figure S4. Network of enriched gene sets in cross-sectional DSPN

Network of enriched gene sets from which the predictive features were selected, for cross-sectional DSPN prediction. Nodes are the gene sets coloured with their corresponding data modality. Size of the nodes reflects their centrality with respect to the network. Edges are the number of shared leading-edge molecules between two nodes.

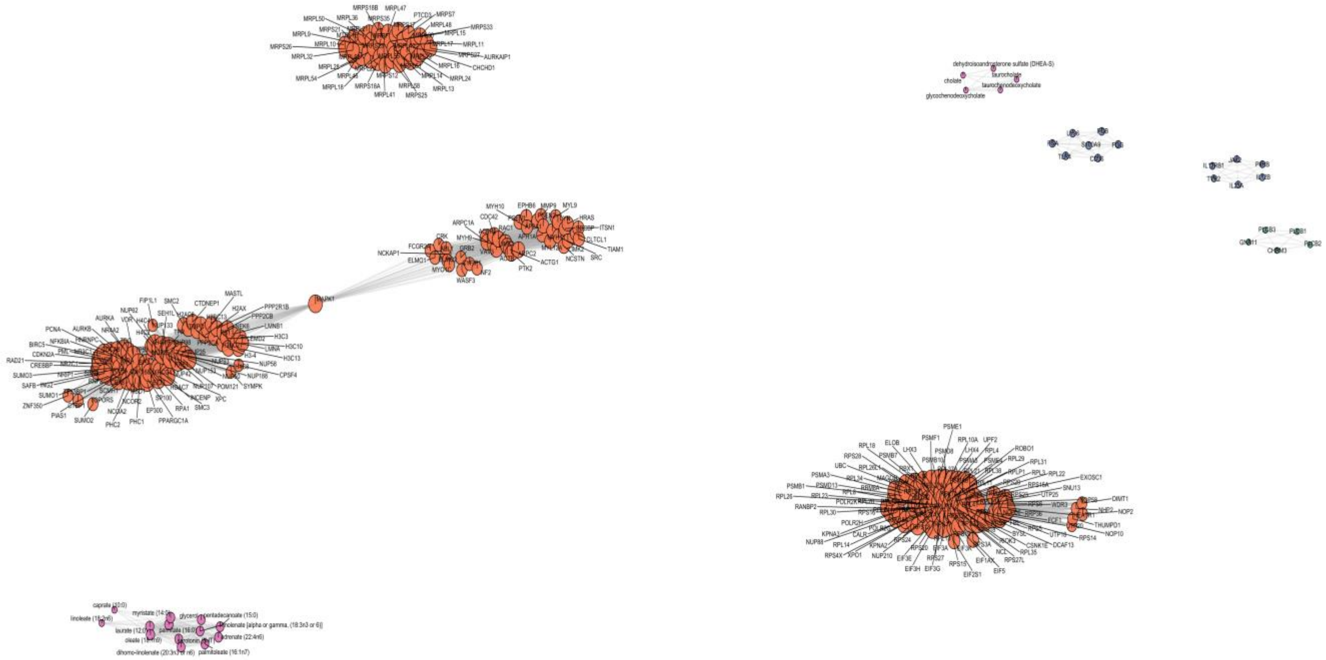


Supplementary Figure S5. Network of enriched features in cross-sectional DSPN

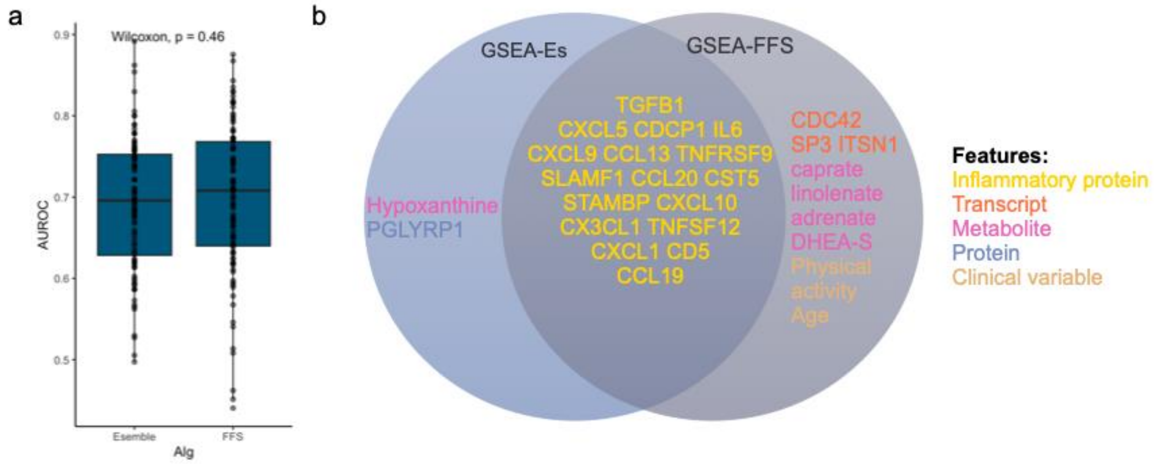
Network of all selected features for training cross-sectional DSPN models. Nodes are the features coloured with their corresponding data modality. Edges are the number of shared gene sets between two nodes.


Transport of organic anions
Defective ABCB11 causes PFIC2 and BRIC2
IRAK4 deficiency (TLR2/4)
EPH-Ephrin signaling
Mitochondrial translation termination
Regulation of actin dynamics for phagocytic cup formation
Acetylcholine_regulates_insulin_secretion
Interleukin-23 signaling
Mitotic Prophase
SUMOylation of DNA replication proteins
Transport of Mature mRNA Derived from an Intronless Transcript
SUMO E3 ligases SUMOylate target proteins
SUMOylation of transcription cofactors
Nuclear Envelope Breakdown
SUMOylation of DNA damage response and repair proteins
Synthesis, secretion, and inactivation of Glucose-dependent Insulinotropic Polypeptide (GIP)
Defective SLC27A4 causes ichthyosis prematurity syndrome (IPS)
Synthesis, secretion, and inactivation of Glucagon-like Peptide-1 (GLP-1)
Retinoid metabolism and transport
G alpha (q) signalling events
Influenza Infection
Major pathway of rRNA processing in the nucleolus and cytosol
Formation of a pool of free 40S subunits
Ribosomal scanning and start codon recognition
rRNA modification in the nucleus and cytosol
Regulation of expression of SLITs and ROBOs
Formation of the ternary complex, and subsequently, the 43S complex
type
Genomics
Transcriptomics
Proteomics
Metabolomics
Methylomics

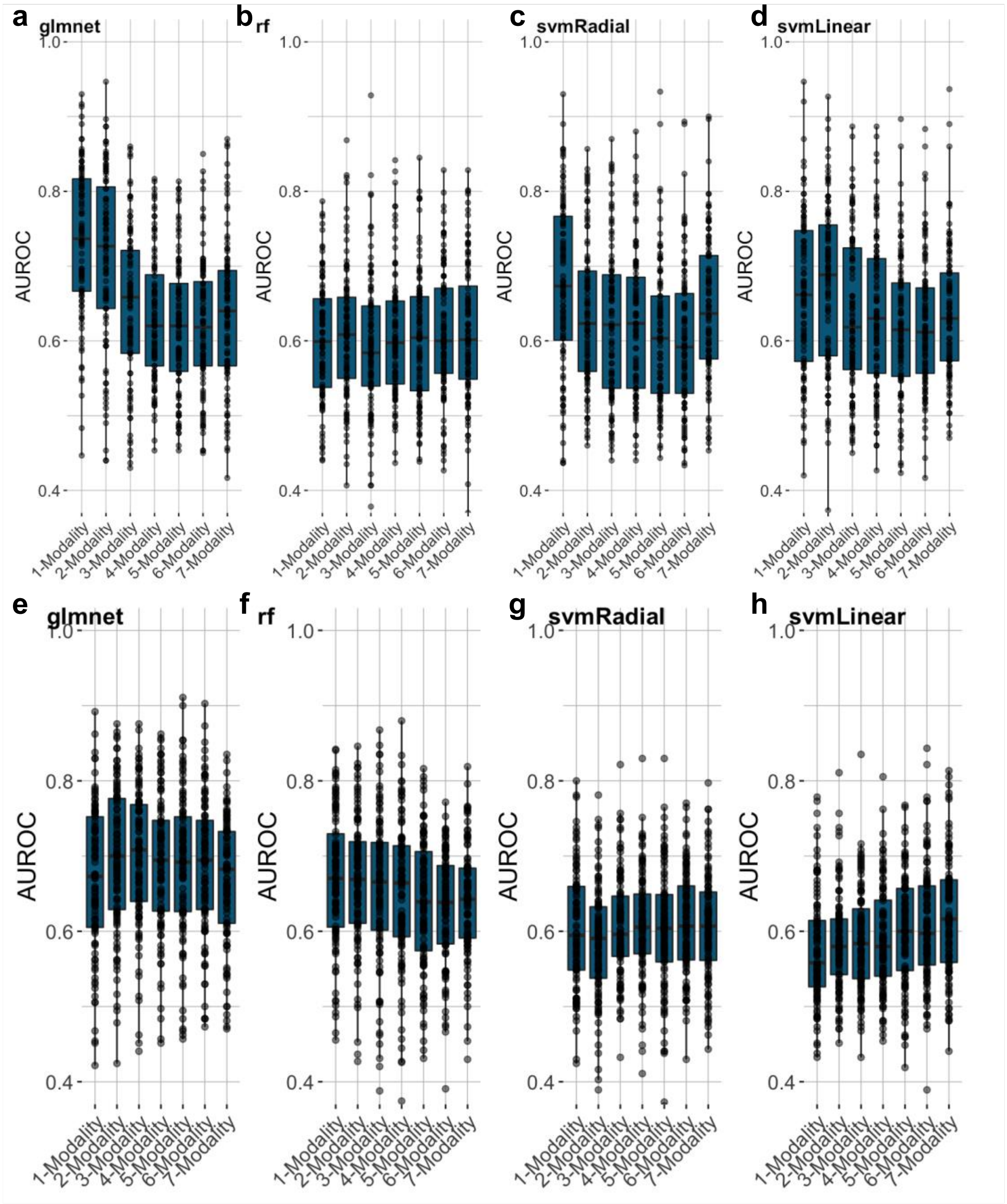
Supplementary Figure S6. Network of enriched gene sets in incident DSPN
Network of enriched gene sets from which the predictive features were selected, for incident DSPN prediction. Nodes are the gene sets coloured with their corresponding data modality. Size of the nodes reflects their centrality with respect to the network. Edges are the number of shared leading-edge molecules between two nodes.



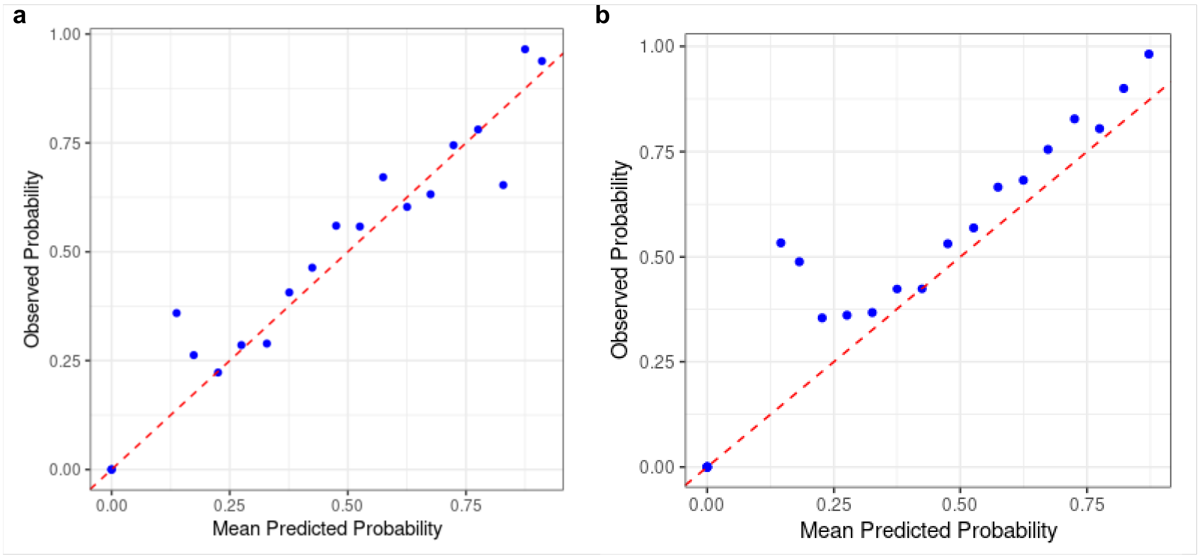
Supplementary Figure S7. Network of enriched features in incident DSPN
Network of all selected features for training incident DSPN models. Nodes are the features coloured with their corresponding data modality. Edges are the number of shared gene sets between two nodes.



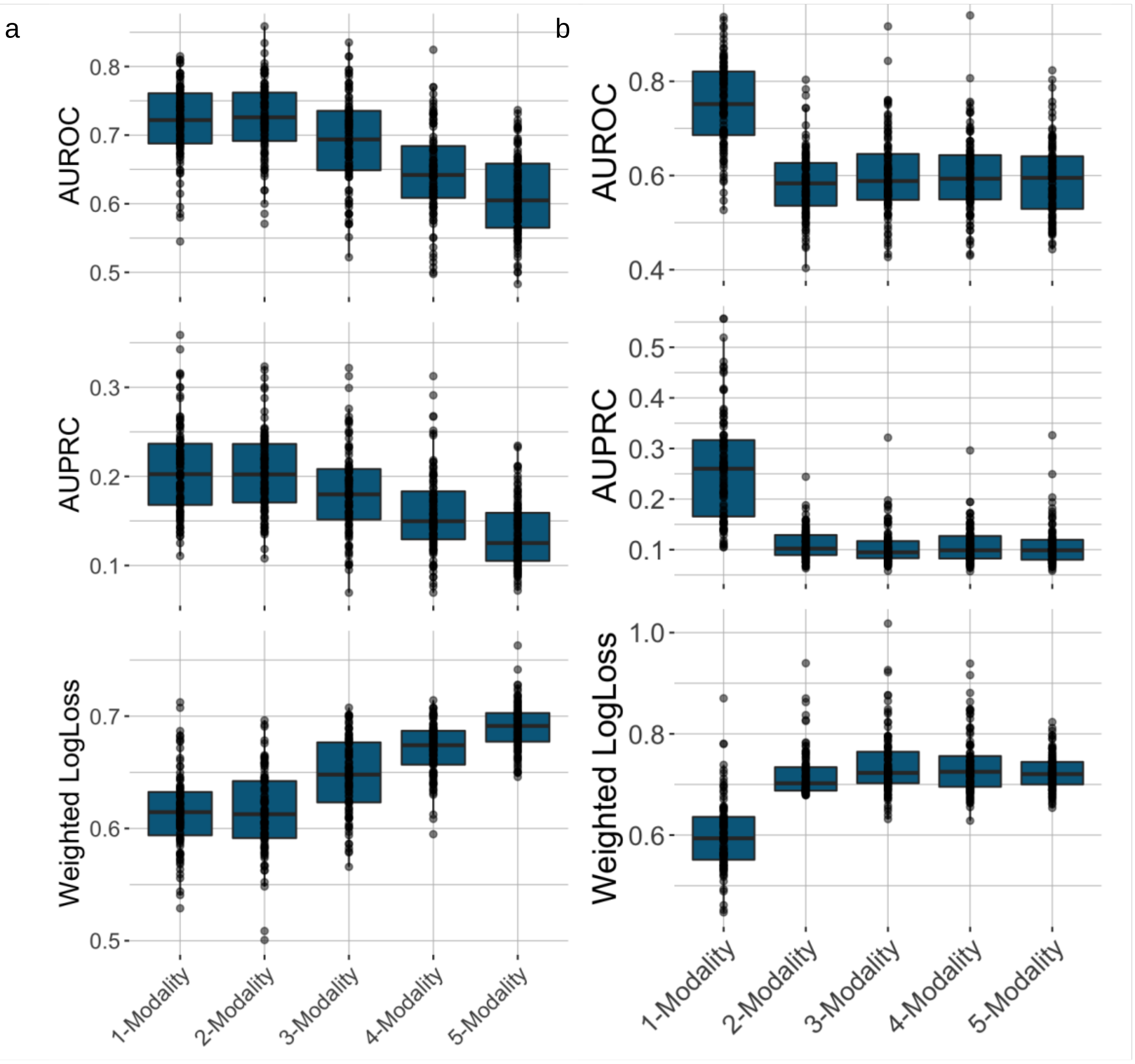
Supplementary Figure S8. Performance of forward feature selection (FFS) and ensemble stacking feature integration methods across 100 stratified resamples. (a) AUROC of the testing prediction of the two algorithms. P-value of Wilcoxon rank sum test is shown.(b) Important features selected by the GSEA-ensemble stacking (GSEA-Es) and GSEA-FFS methods and their overlapping.



Supplementary Figure S9: Prediction performance of four different machine learning algorithms. Here we compare the predictive power of (a-d) prevalent DSPN and (e-h) incident DSPN. We benchmarked (a,e) elastic net (glmnet), (b,f) random forest (rf), and support vector machine with (c,g) radial (svmRadial) and (d,h) linear kernel (svmLinear).



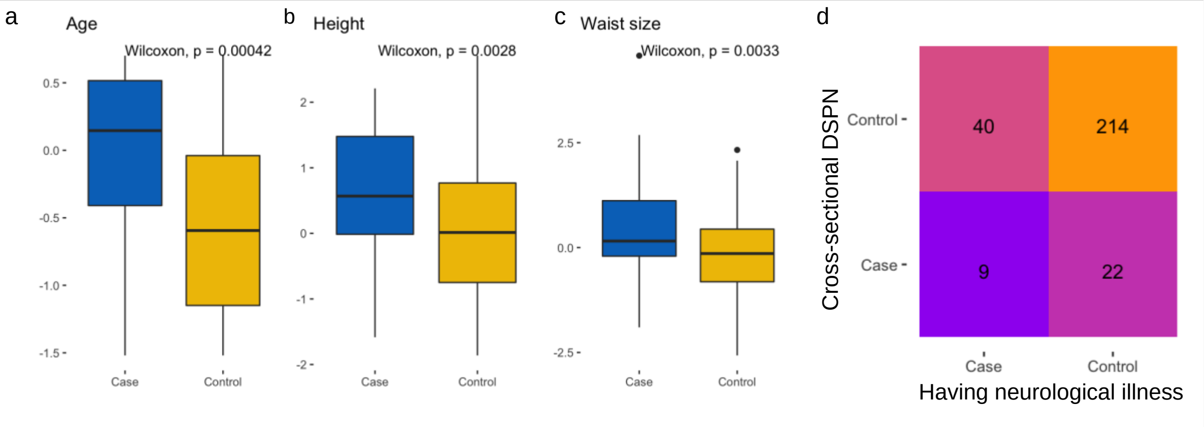
Supplementary Figure S10: Calibration plots of predicted probabilities for prevalent DSPN (a) and incident DSPN (b). The predicted probabilities were calibrated using the Platt scaling method.



Supplementary Figure S11. Prediction performance of prevalent DSPN models, when forcing the FFS algorithm to choose clinical model at the beginning.

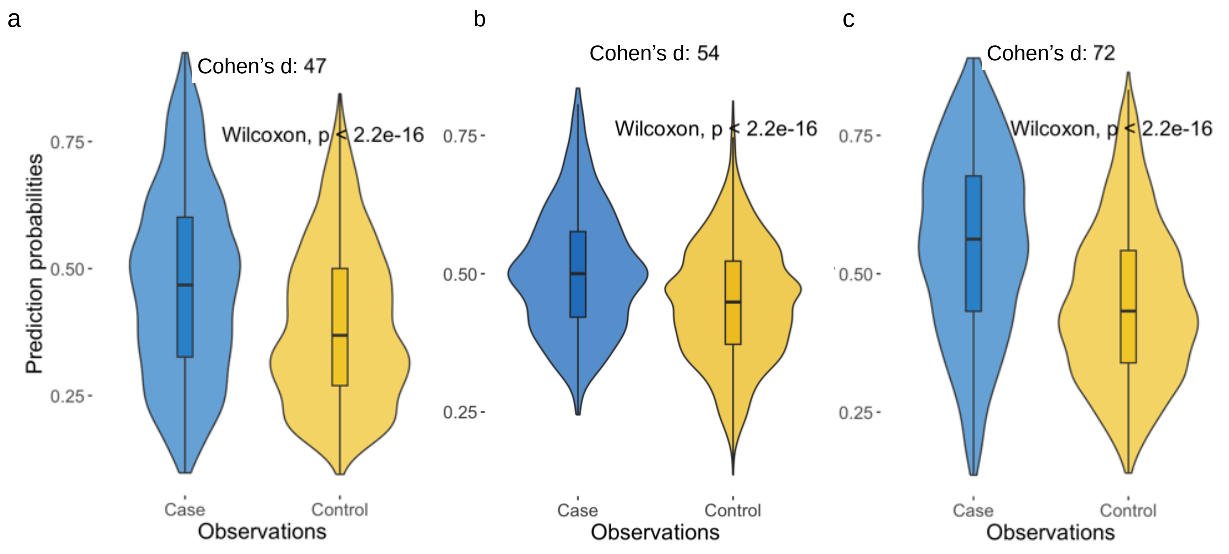
a. Prediction performance during cross-validation. X-axis shows the increasing model complexity. Y-axis shows the median of performance values across 5-fold cross-validation for AUROC, AUPRC and weighted log-loss

b. Prediction performance on the testing sets. X-axis shows the increasing model complexity. Y-axis shows the performance values on the testing sets for AUROC, AUPRC and weighted log-loss

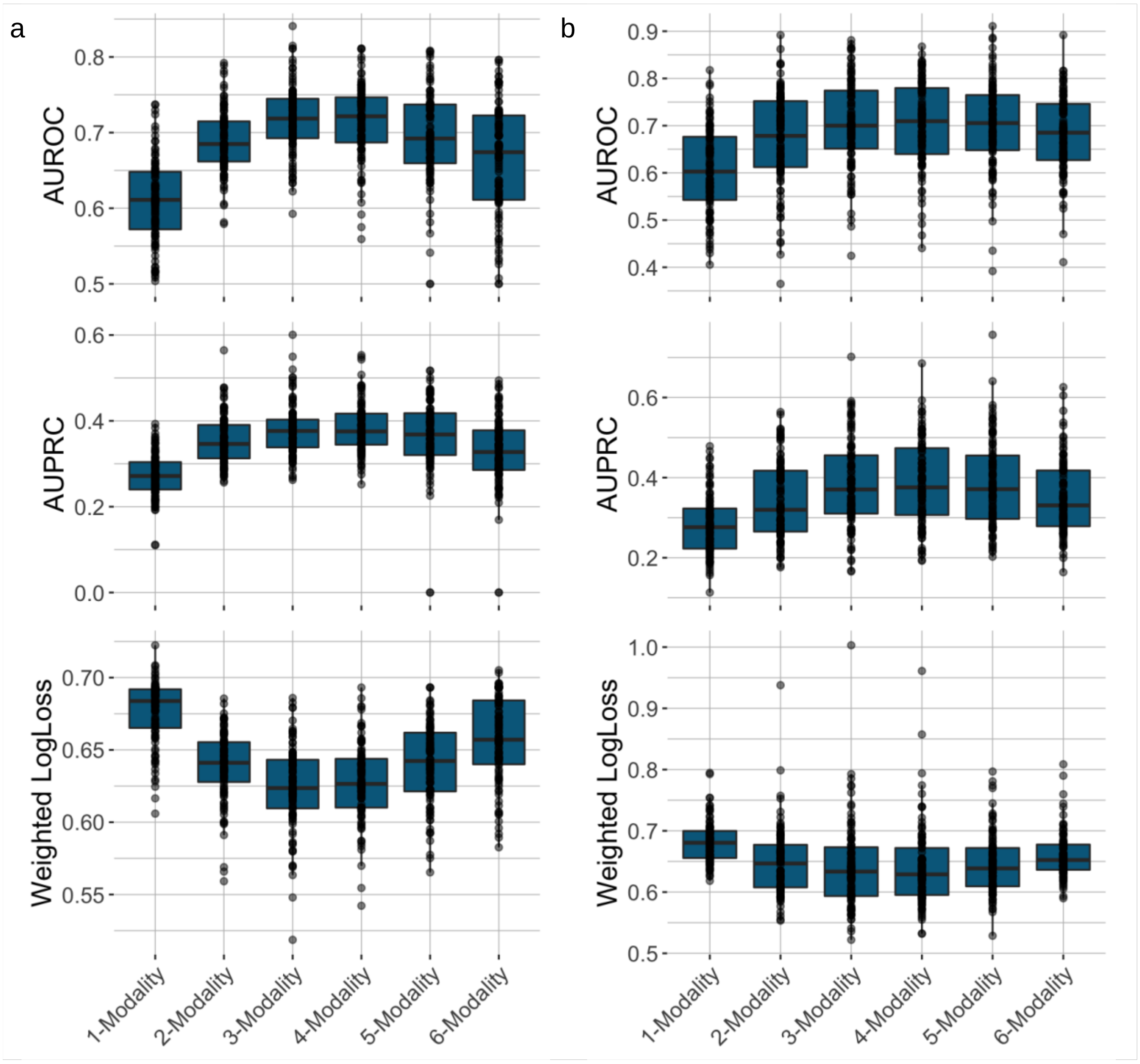


Supplementary Figure S12. Distribution of important clinical variables for cross-sectional DSPN model

Distribution of age, height and waist size in the training set stratified into case and control (panel a, b and c respectively). Panel d shows association of patients who have neurological illness in general and cross-sectional DSPN. P-values for Wilcoxon rank sum test and Fisher's exact test are shown.



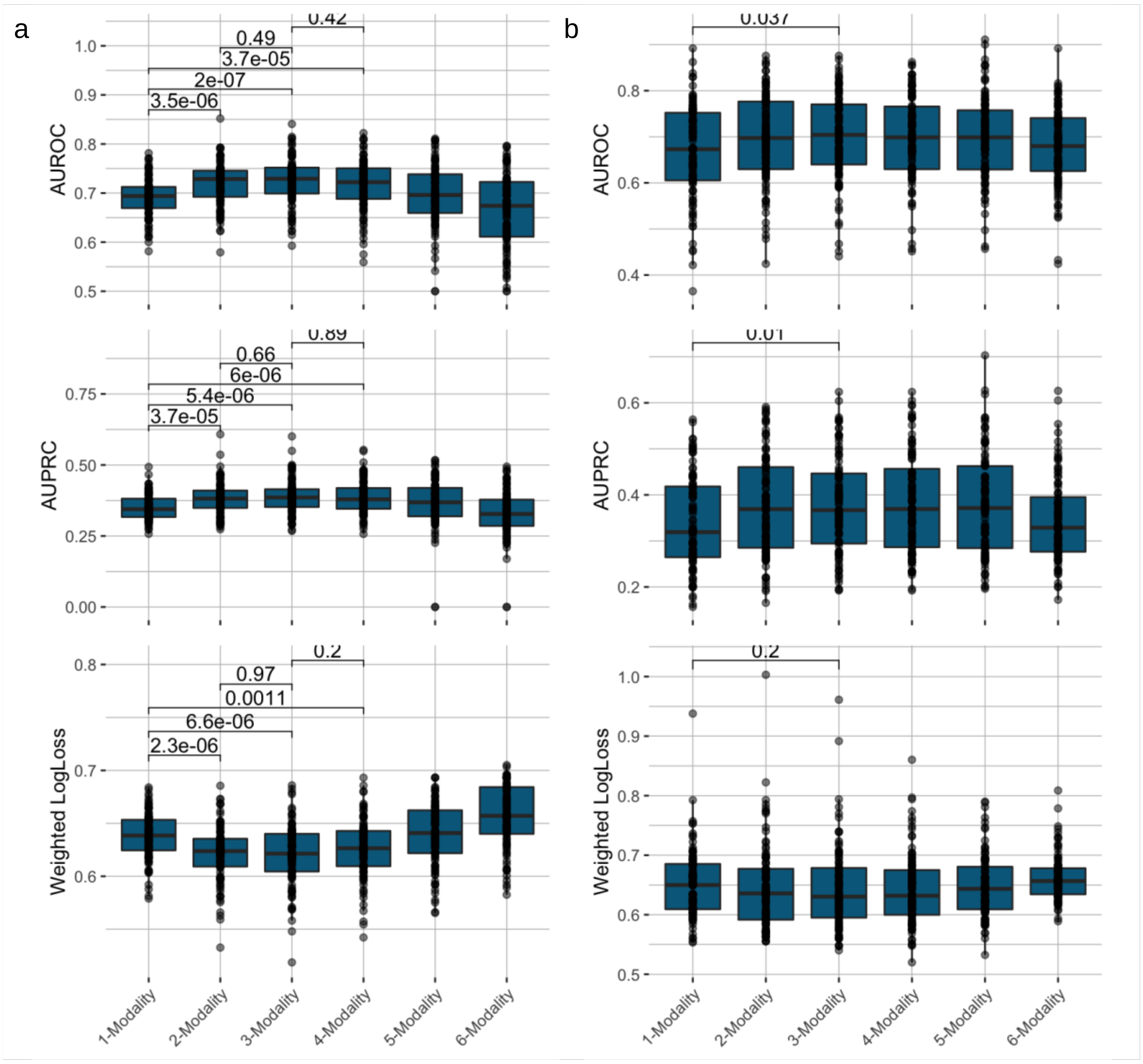
Supplemental Figure S13: Baseline models to predict DSPN incidence. Prediction probabilities during testing of negative samples using a) the prevalent DSPN model trained on clinical data alone at F4, b) baseline incidence model trained only on clinical variables at F4 and incidence label at FF4 and c) the full incidence model trained on clinical + molecular variables at F4 and incidence label at FF4. Cases are samples developing DSPN from F4 to FF4, and controls are ones remaining negative. For each comparison, Cohen's d was used as the measure of the difference between groups.



Supplementary Figure S14. Prediction performance of incident DSPN models, when forcing the FFS algorithm to choose clinical model at the beginning.

a. Prediction performance during cross-validation. X-axis shows the increasing model complexity. Y-axis shows the median of performance values across 5-fold cross-validation for AUROC, AUPRC and weighted log-loss

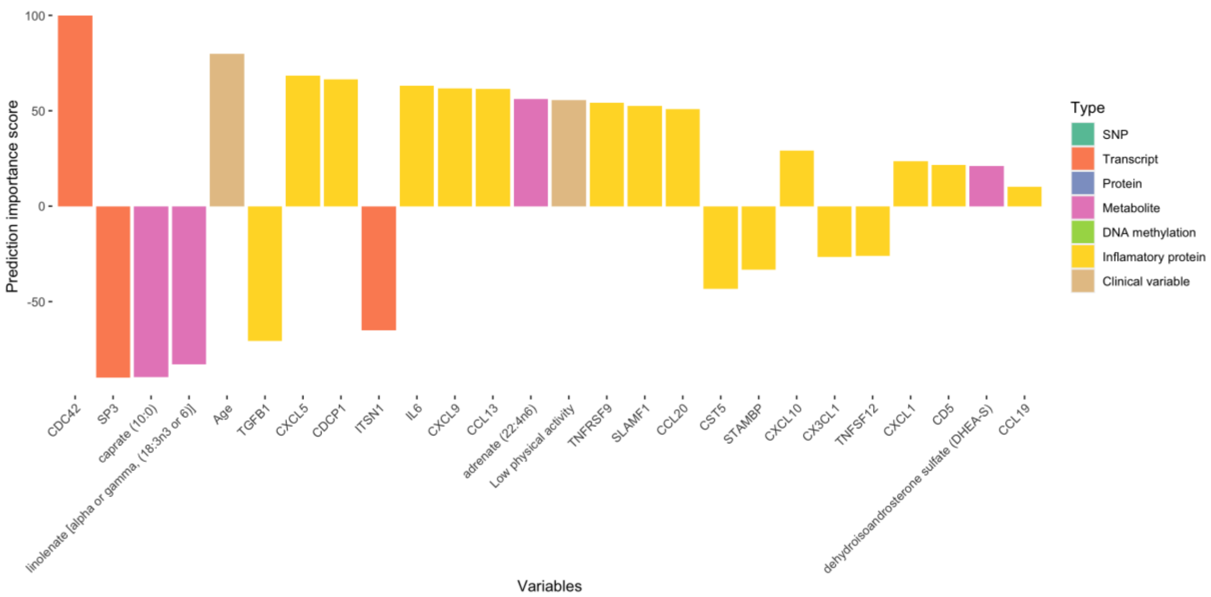
b. Prediction performance on the testing sets. X-axis shows the increasing model complexity. Y-axis shows the performance values on the testing sets for AUROC, AUPRC and weighted log-loss



Supplementary Figure S15. Prediction performance of incident DSPN models, when allowing the FFS algorithm to choose starting model based on cross-validation.

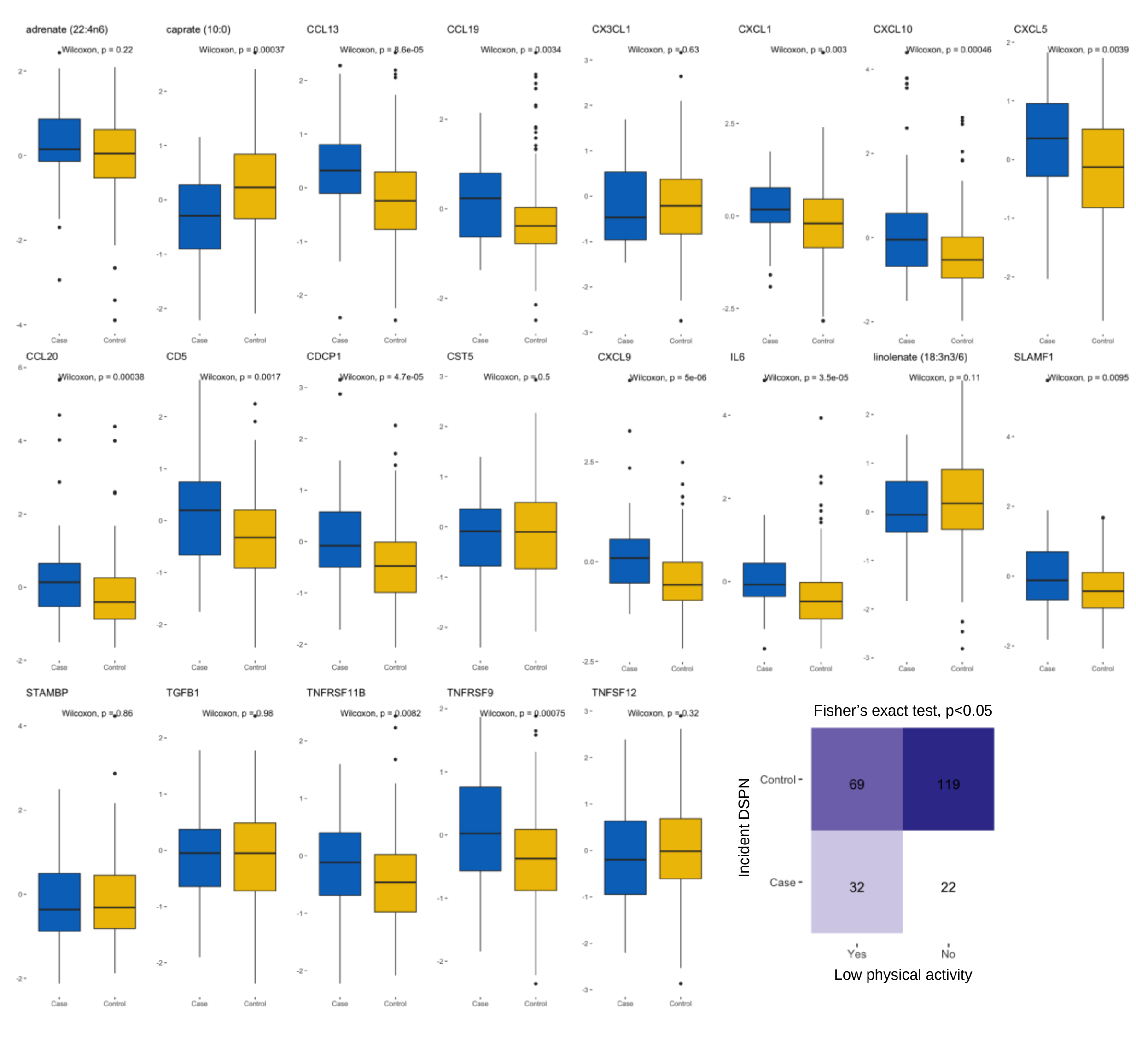
a. Prediction performance during cross-validation. X-axis shows the increasing model complexity. Y-axis shows the median of performance values across 5-fold cross-validation for AUROC, AUPRC and weighted log-loss

b. Prediction performance on the testing sets. X-axis shows the increasing model complexity. Y-axis shows the performance values on the testing sets for AUROC, AUPRC and weighted log-loss



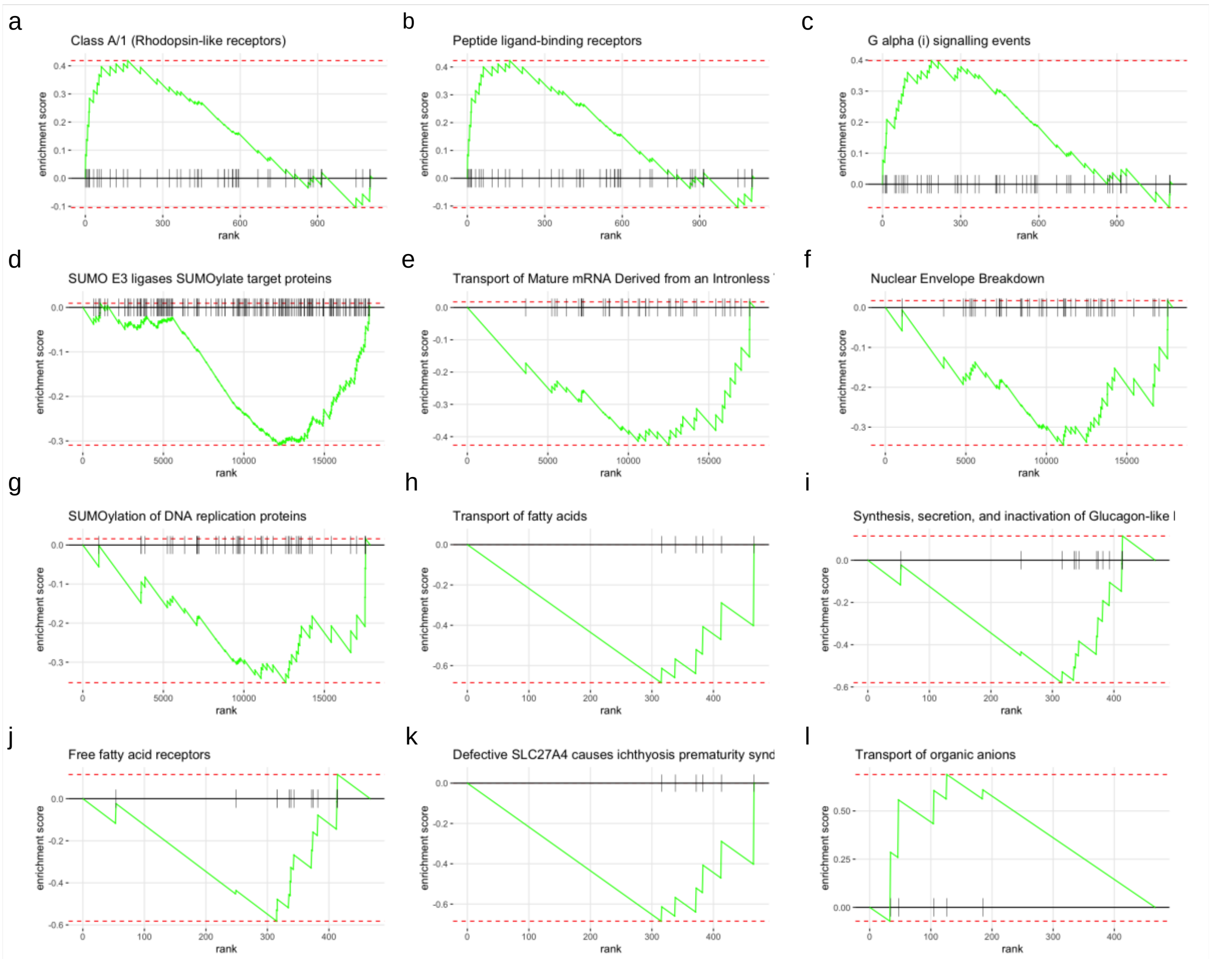
Supplementary Figure S16. Feature importance score of the important features of the final incident DSPN model

X-axis shows the features in decreasing magnitude of the t-statistics in the final model. Y-axis shows the t-statistics (signed importance scores) of the features. Colors represent the data modality.



Supplementary Figure S17. Distribution of important clinical variables for incident DSPN model

Distribution of the features in the training set stratified into case and control. P-values for Wilcoxon rank sum test and Fisher's exact test are shown.



Supplementary Figure S18. Examples of consistently enriched signalling pathways that are predictive of incident DSPN

X-axis represents all evaluated genes ranked in decreasing order of t-statistics, with ticks represent genes that belong to the examined gene set. Y-axis represent the enrichment score. Panels a-c are inflammation protein pathways, d-g are transcriptomic pathways and h-l are metabolomic pathways.

Supplementary Table S1. Clinical characteristics of the dataset for prevalent DSPN prediction

Variable	Control (MNSI <3)	Case (MNSI >= 3)	P
N	903	188	
Age, years	69.7 ± 5.2	72.5 ± 5.2	1.09e-10
Sex, % male	49.4	60.6	0.005
Height, cm	165.3 ± 8.8	167.9 ± 9.6	0.00071
BMI, kg/m2	28.4 ± 4.2	30.2 ± 5.2	1.30e-05
Waist circumference, cm	97.2 ± 11.7	103.7 ± 12.9	8.11e-10
Systolic blood pressure, mmHg	128.8 ± 20	128.6 ± 20	0.873
Diastolic blood pressure, mmHg	74.4 ± 10.1	72.4 ± 9.8	0.007
Hypertension, %	62.0	64.4	0.561
Smoking, %, never/former/current	51.6/40.7/7.7	44.9/48.1/7.0	0.233
High alcohol consumption, %	29.1	33.7	0.220
Low physical activity, %	36.8	51.9	0.014
Previous myocardial infarction, %	5.9	9.1	0.104
Previous stroke, %	3.2	8.0	0.006
Presence of neurological diseases, %	16.2	31.0	4.33e-06
Absent ankle reflexes, %	5	72.3	6.63e-112
Foot ulcer present, %	0	2.1	0.001
MNSI score	1.7 ± 1	4.3 ± 0.9	2.34e-107
Use of NSAIDs, %	3.4	7.4	0.024
NGT, %	53.7	45.7	0.054
i-IFG, %	5.3	3.7	0.464
i-IGT, %	16.7	12.2	0.154
IFG/IGT, %	4.3	6.9	0.133
Newly diagnosed diabetes, %	6.4	4.8	0.504
Known diabetes, %	13.5	26.6	1.25e-05
Diabetes duration, years*	8.1 ± 6.4	15 ± 10.6	1.58e-15
Metabolic parameters			
Fasting glucose, mg/dL⁺	103.6 ± 21.2	110.4 ± 29.9	0.015
2-h glucose, mg/dL⁺	128.0 ± 41.9	127.2 ± 38.6	0.945
HbA1c, %	5.7 ± 0.7	6.0 ± 0.8	3.06e-06
Total cholesterol, mg/dL	222.7 ± 41.0	210.8 ± 37.9	0.00014
LDL cholesterol, mg/dL	140.7 ± 36.2	131.7 ± 33.4	0.001
HDL cholesterol, mg/dL	56.0 ± 14.3	53.4 ± 12.2	0.075
Creatinine, mg/dL	0.95 ± 0.3	1.02 ± 0.3	0.001
Uric acid, mg/dL	5.5 ± 1.4	5.8 ± 1.5	0.015

* Only applicable to people with diabetes

+ Only applicable to people without known diabetes

Supplementary Table S2. Clinical characteristics of the dataset for incident DSPN prediction

Variable	Control (no incident F4->FF4)	Case (incident F4-> FF4)	P
N	394	131	
Age, years	68.0 ± 4.6	70.1 ± 4.9	2.46e-05
Sex, % male	49.2	56.5	0.159
Height, cm	165.9 ± 8.5	167.6 ± 9.4	0.064
BMI, kg/m2	27.7 ± 3.8	29.1 ± 4.0	0.00054
Waist circumference, cm	94.8 ± 11.2	99.9 ± 11.4	1.34e-05
Systolic blood pressure, mmHg	128.4 ± 19.2	131.3 ± 19.9	0.217
Diastolic blood pressure, mmHg	75.5 ± 10.1	75.5 ± 9.2	0.950
Hypertension, %	56.3	65.6	0.066
Smoking, %, never/former/current	52.0/42.4/5.6	55.0/33.6/11.4	0.054
High alcohol consumption, %	29.4	35.9	0.191
Low physical activity, %	26.4	42.7	0.00064
Previous myocardial infarction, %	4.8	6.9	0.373
Previous stroke, %	1.0	0.8	1
Presence of neurological diseases, %	14.7	21.4	0.102
Absent ankle reflexes, %	3.8	6.1	0.323
Foot ulcer present, %	0	0	1
MNSI score	1.5 ± 1.0	1.9 ± 0.9	2.65e-05
Use of NSAIDs, %	1.0	2.3	0.374
NGT, %	62.9	50.4	0.013
i-IFG, %	3.0	7.6	0.040
i-IGT, %	14.5	16.8	0.573
IFG/IGT, %	4.6	4.6	1
Newly diagnosed diabetes, %	5.6	5.3	1
Known diabetes, %	9.4	15.3	0.074
Diabetes duration, years*	6.9 ± 5.5	8.9 ± 5.2	0.116
Metabolic parameters			
Fasting glucose, mg/dL+	101.0 ± 16.4	103.8 ± 17.2	0.078
2-h glucose, mg/dL+	123.9 ± 38.6	127.4 ± 38.4	0.371
HbA1c, %	5.7 ± 0.5	5.8 ± 0.7	0.027
Total cholesterol, mg/dL	226.3 ± 40.5	216.1 ± 42.7	0.009
LDL cholesterol, mg/dL	142.5 ± 36.3	136.6 ± 37.5	0.069
HDL cholesterol, mg/dL	57.3 ± 14.2	52.5 ± 12.3	0.00025
Creatinine, mg/dL	0.9 ± 0.2	1.0 ± 0.3	0.071
Uric acid, mg/dL	5.5 ± 1.3	5.6 ± 1.4	0.692

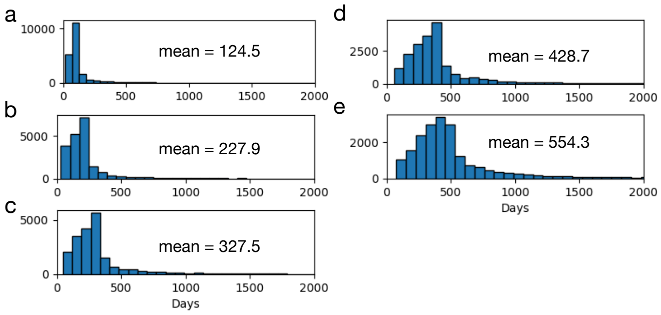
* Only applicable to people with diabetes

+ Only applicable to people without known diabetes

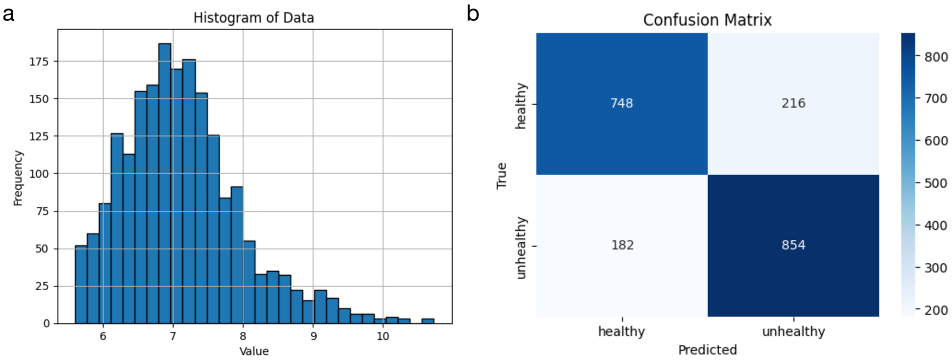
Supplementary Table S3. Significantly enriched signalling pathways during feature selection for prevalent DSPN prediction

pathway	pval	padj	ES	NES	size	Type
Formation of a pool of free 40S subunits	1.874e-06	2.024e-05	-0.445	-2.140	95	Transcriptomics
GTP hydrolysis and joining of the 60S ribosomal subunit	3.308e-05	8.132e-05	-0.405	-1.986	104	Transcriptomics
L13a-mediated translational silencing of Ceruloplasmin expression	3.555e-05	8.132e-05	-0.404	-1.972	103	Transcriptomics
Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC)	6.2356e-05	0.0001	-0.395	-1.944	106	Transcriptomics
Collagen chain trimerization	0.0001	0.0002	0.522	2.029	39	Transcriptomics
Influenza Infection	0.0002	0.0004	-0.341	-1.762	143	Transcriptomics
Collagen biosynthesis and modifying enzymes	0.0003	0.0005	0.441	1.921	61	Transcriptomics
Assembly of collagen fibrils and other multimeric structures	0.0003	0.0005	0.448	1.928	58	Transcriptomics
Degradation of the extracellular matrix	0.0004	0.0006	0.338	1.678	131	Transcriptomics
Formation of the ternary complex and subsequently the 43S complex	0.0005	0.0007	-0.460	-1.941	47	Transcriptomics
SUMOylation of DNA methylation proteins	0.0015	0.002	-0.635	-2.006	15	Transcriptomics
Selenoamino acid metabolism	0.0018	0.002	-0.333	-1.642	110	Transcriptomics
Major pathway of rRNA processing in the nucleolus and cytosol	0.0019	0.002	-0.293	-1.550	170	Transcriptomics
Ribosomal scanning and start codon recognition	0.002	0.003	-0.413	-1.774	53	Transcriptomics
Regulation of expression of SLITs and ROBOs	0.005	0.005	-0.291	-1.526	159	Transcriptomics
Laminin interactions	0.0001	0.0713	0.829	2.017	9	Proteomics
Antimicrobial peptides	0.0004	0.0713	-0.676	-2.111	16	Proteomics
Interleukin-20 family signaling	0.0004	0.0713	0.745	2.009	13	Proteomics
Interleukin-3 Interleukin-5 and GM-CSF signaling	0.0006	0.082	0.618	1.922	23	Proteomics
Transport of nucleosides and free purine and pyrimidine bases across the plasma membrane	0.0002	0.0002	-0.931	-2.083	5	Metabolomics

D Supplementary information to section 3.3



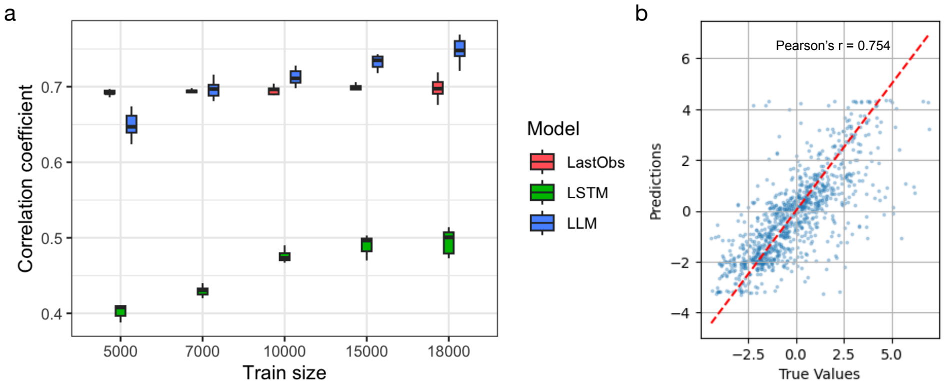
Supplementary Figure S1. Distribution of prediction windows for the benchmarking of LLM-based and XGBoost models. a-e) Prediction windows in case of last visit, second to last, third to last, fourth to last and fifth to last visits, respectively. The mean numbers of days are also shown.



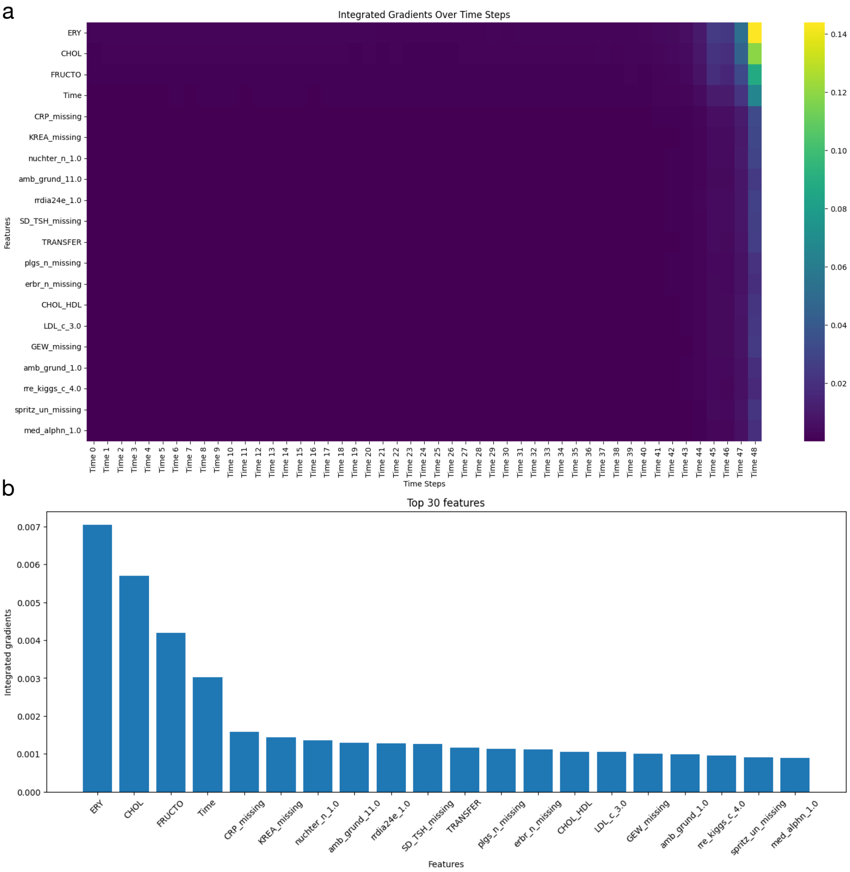
(Raw) predictions

Fisher’s exact test, p-value=4.9e-170

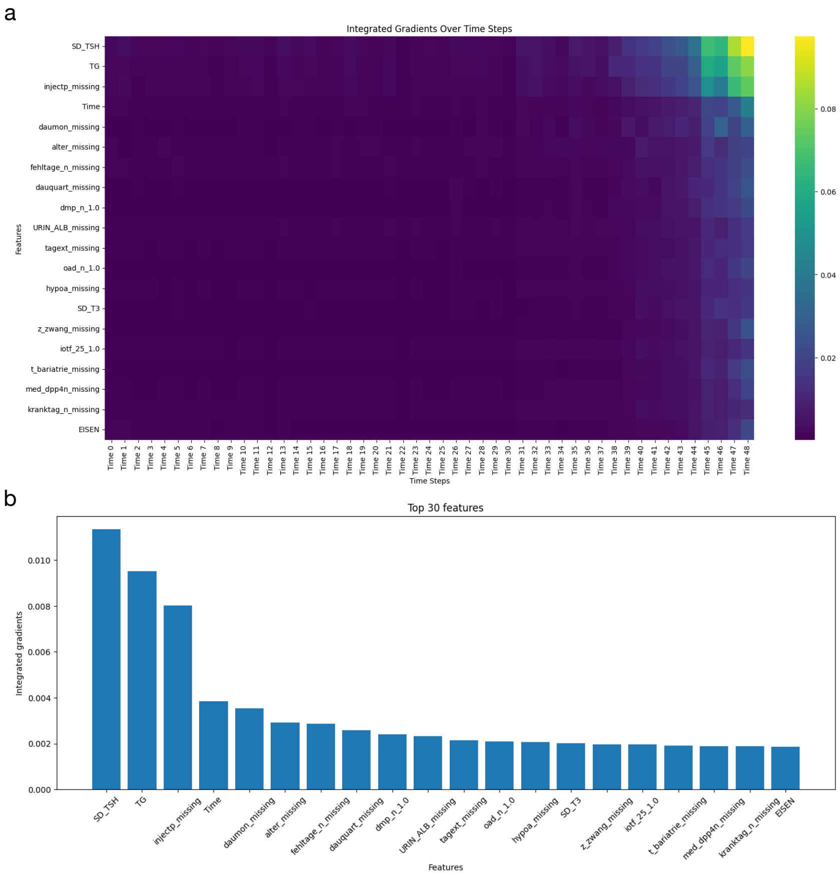
Supplementary Figure S2. Performance of the model in stratifying healthy and unhealthy patients. (a) Distribution of predicted HbA1C when put back in the normal range. b) Confusion matrix of predictions and observations for HbA1C prediction.



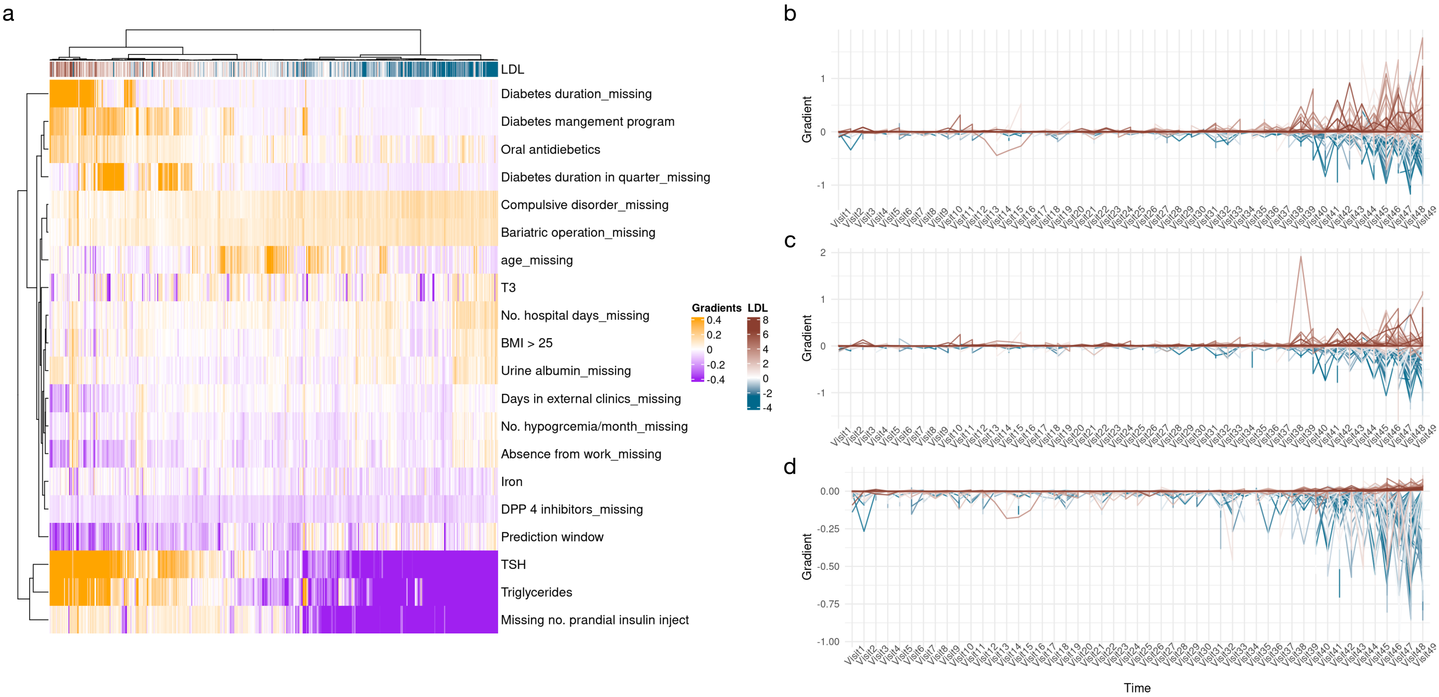
Supplementary Figure S3. The LLM-based framework effectively predicted LDL level in the next visit. (a) Performance of the LLM-based framework was compared to the LSTM-based method and random forest model using only the last observed LDL value in the patient's record (LastObs), in terms of Pearson's correlation coefficient. (b) Predicted values of the LDL was compared against the true values in a testing dataset when using 18000 samples for train set and 2000 samples for test set.



Supplementary Figure S4. Important features in HbA1C prediction. (a) Heatmap of top 20 important features across timepoints, colours represent average gradient across samples. b) Top 20 important features ordered according to average absolute gradient across timepoints and samples.



Supplementary Figure S5. Important features in LDL prediction. (a) Heatmap of top 20 important features across timepoints, colours represent average gradient across samples. b) Top 20 important features ordered according to average absolute gradient across timepoints and samples.



Supplementary Figure S6. Most important features of LDL prediction. a) Heat map of feature importance score represented by integrated gradients of top 20 most important features. Values are the mean absolute gradients of the features across samples and time points. b-d) Trajectory of erythrocyte count, cholesterol and fructose levels across visits in the feature window, respectively. Each line represents one of 500 samples, colours represent the observed LDL level at the target point.

References

[1] Francis S Collins and Harold Varmus. A new initiative on precision medicine. *N. Engl. J. Med.*, 372(9):793, February 2015.

[2] Rodrigo Dienstmann, Louis Vermeulen, Justin Guinney, Scott Kopetz, Sabine Tejpar, and Josep Tabernero. Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. *Nat. Rev. Cancer*, 17(2):79–92, February 2017.

[3] Domenico Accili. Insulin action research and the future of diabetes treatment: The 2017 banting medal for scientific achievement lecture. *Diabetes*, 67(9):1701–1709, September 2018.

[4] Yehudit Hasin, Marcus Seldin, and Aldons Lusis. Multi-omics approaches to disease. *Genome Biol.*, 18(1):83, May 2017.

[5] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.*, 25(1):44–56, January 2019.

[6] Nardeep Naithani, Sharmila Sinha, Pratibha Misra, Biju Vasudevan, and Rajesh Sahu. Precision medicine: Concept and tools. *Armed Forces Med. J. India*, 77(3):249–257, July 2021.

[7] Christopher J Phillips. Precision medicine and its imprecise history. *Harvard Data Science Review*, January 2020.

[8] John D Firth, Christopher P Conlon, and Timothy Cox. *Oxford Textbook of Medicine.* Oxford University Press, 2020.

[9] Roy Porter. *The Greatest Benefit to Mankind: A Medical History of Humanity from Antiquity to the Present.* HarperCollins, 1997.

[10] George Rosen. *A History of Public Health.* JHU Press, April 2015.

[11] Abraham Flexner and Carnegie Foundation for the Advancement. *Medical Education in the United States and Canada: A Report to the Carnegie Foundation for the Advancement of Teaching, Issues 1-3... - Primary Source.* BiblioLife, December 2013.

[12] Institute of Medicine and Committee on Emerging Microbial Threats to Health. *Emerging Infections: Microbial Threats to Health in the United States.* National Academies Press, February 1992.

[13] Frank Fenner. *Smallpox and Its Eradication.* World Health Organization, 1988.

[14] Yi-Fan Lu, David B Goldstein, Misha Angrist, and Gianpiero Cavalleri. Personalized medicine and human genetic diversity. *Cold Spring Harb. Perspect. Med.*, 4(9), September 2014.

[15] J D Watson and F H C Crick. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, April 1953.

[16] F Sanger and A R Coulson. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.*, 94(3):441–448, May 1975.

[17] M V Olson. The human genome project. *Proc. Natl. Acad. Sci. U. S. A.*, 90(10):4338–4344, May 1993.

[18] National Cancer Institute. What is cancer? `https://www.cancer.gov/about-cancer/understanding/what-is-cancer`, September 2007. Accessed: 2024-6-21.

[19] Antonino Carbone. Cancer classification at the crossroads. *Cancers*, 12(4):980, April 2020.

[20] Bruce A Chabner and Thomas G Roberts, Jr. Timeline: Chemotherapy and the war on cancer. *Nat. Rev. Cancer*, 5(1):65–72, January 2005.

[21] M Baum, M A Chaplain, A R Anderson, M Douek, and J S Vaidya. Does breast cancer exist in a state of chaos? *Eur. J. Cancer*, 35(6):886–891, June 1999.

[22] Ailbhe C O'Neill, Jyothi P Jagannathan, and Nikhil H Ramaiya. Evolving cancer classification in the era of personalized medicine: A primer for radiologists. *Korean J. Radiol.*, 18(1):6–17, January 2017.

[23] J D Rowley. Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and giemsa staining. *Nature*, 243(5405):290–293, June 1973.

[24] Lee Schwartzberg, Edward S Kim, David Liu, and Deborah Schrag. Precision oncology: Who, how, what, when, and when not? *Am. Soc. Clin. Oncol. Educ. Book*, 37:160–169, 2017.

[25] Michael R Stratton, Peter J Campbell, and P Andrew Futreal. The cancer genome. *Nature*, 458(7239):719–724, April 2009.

[26] Viviana Masoud and Gilles Pagès. Targeted therapies in breast cancer: New challenges to fight against resistance. *World J. Clin. Oncol.*, 8(2):120–134, April 2017.

[27] Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, October 2008.

[28] Wanjuan Yang, Jorge Soares, Patricia Greninger, Elena J Edelman, Howard Lightfoot, Simon Forbes, Nidhi Bindal, Dave Beare, James A Smith, I Richard Thompson, Sridhar Ramaswamy, P Andrew Futreal, Daniel A Haber, Michael R Stratton, Cyril Benes, Ultan McDermott, and Mathew J Garnett. Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.*, 41(Database issue):D955–61, January 2013.

[29] Amrita Basu, Nicole E Bodycombe, Jaime H Cheah, Edmund V Price, Ke Liu, Giannina I Schaefer, Richard Y Ebright, Michelle L Stewart, Daisuke Ito, Stephanie Wang, Abigail L Bracha, Ted Liefeld, Mathias Wawer, Joshua C Gilbert, Andrew J Wilson, Nicolas Stransky, Gregory V Kryukov, Vlado Dancik, Jordi Barretina, Levi A Garraway, C Suk-Yee Hon, Benito Munoz, Joshua A Bittker, Brent R Stockwell, Dineo Khabele, Andrew M Stern, Paul A Clemons, Alykhan F Shamji, and Stuart L Schreiber. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell*, 154(5):1151–1161, August 2013.

[30] Caitriona Holohan, Sandra Van Schaeybroeck, Daniel B Longley, and Patrick G Johnston. Cancer drug resistance: an evolving paradigm. *Nat. Rev. Cancer*, 13(10):714–726, October 2013.

[31] World Health Organisation. Diabetes. `https://www.who.int/news-room/fact-sheets/detail/diabetes`. Accessed: 2024-6-21.

[32] Wendy K Chung, Karel Erion, Jose C Florez, Andrew T Hattersley, Marie-France Hivert, Christine G Lee, Mark I McCarthy, John J Nolan, Jill M Norris, Ewan R Pearson, Louis Philipson, Allison T McElvaine, William T Cefalu, Stephen S Rich, and Paul W Franks. Precision medicine in diabetes: A consensus report from the american diabetes association (ADA) and the european association for the study of diabetes (EASD). *Diabetes Care*, 43(7):1617–1635, July 2020.

[33] Anjali D Deshpande, Marcie Harris-Hayes, and Mario Schootman. Epidemiology of diabetes and diabetes-related complications. *Phys. Ther.*, 88(11):1254, November 2008.

[34] Richard David Leslie, Ronald Ching Wan Ma, Paul W Franks, Kristen J Nadeau, Ewan R Pearson, and Maria Jose Redondo. Understanding diabetes heterogeneity: key steps towards precision medicine in diabetes. *Lancet Diabetes Endocrinol*, 11(11):848–860, November 2023.

[35] John J Nolan, Anna R Kahkoska, Zhila Semnani-Azad, Marie-France Hivert, Linong Ji, Viswanathan Mohan, Robert H Eckel, Louis H Philipson, Stephen S Rich, Chandra Gruber, and Paul W Franks. ADA/EASD precision medicine in diabetes initiative: An international perspective and future vision for precision medicine in diabetes. *Diabetes Care*, 45(2):261–266, February 2022.

[36] F B Hu, J E Manson, M J Stampfer, G Colditz, S Liu, C G Solomon, and W C Willett. Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. *N. Engl. J. Med.*, 345(11):790–797, September 2001.

[37] Christian Herder, Julia M Kannenberg, Cornelia Huth, Maren Carstensen-Kirberg, Wolfgang Rathmann, Wolfgang Koenig, Margit Heier, Sonja Püttgen, Barbara Thorand, Annette Peters, Michael Roden, Christa Meisinger, and Dan Ziegler. Proinflammatory cytokines predict the incidence and progression of distal sensorimotor polyneuropathy: KORA F4/FF4 study. *Diabetes Care*, 40(4):569–576, April 2017.

[38] Robert Wagner, Martin Heni, Adam G Tabák, Jürgen Machann, Fritz Schick, Elko Randrianarisoa, Martin Hrabě de Angelis, Andreas L Birkenfeld, Norbert Stefan, Andreas Peter, Hans-Ulrich Häring, and Andreas Fritsche. Pathophysiology-based subphenotyping of individuals at elevated risk for type 2 diabetes. *Nat. Med.*, 27(1):49–57, January 2021.

[39] Deirdre K Tobias, Jordi Merino, Abrar Ahmad, Catherine Aiken, Jamie L Benham, Dhanasekaran Bodhini, Amy L Clark, Kevin Colclough, Rosa Corcoy, Sara J Cromer, Daisy Duan, Jamie L Felton, Ellen C Francis, Pieter Gillard, Véronique Gingras, Romy Gaillard, Eram Haider, Alice Hughes, Jennifer M Ikle, Laura M Jacobsen, Anna R Kahkoska, Jarno L T Kettunen, Raymond J Kreienkamp, Lee-Ling Lim, Jonna M E Männistö, Robert Massey, Niamh-Maire Mclennan, Rachel G Miller, Mario Luca Morieri, Jasper Most, Rochelle N Naylor, Bige Ozkan, Kashyap Amratlal Patel, Scott J Pilla, Katsiaryna Prystupa, Sridharan Raghavan, Mary R Rooney, Martin Schön, Zhila Semnani-Azad, Magdalena Sevilla-Gonzalez, Pernille Svalastoga, Wubet Worku Takele, Claudia Ha-Ting Tam, Anne Cathrine B Thuesen, Mustafa Tosur, Amelia S Wallace, Caroline C Wang, Jessie J Wong, Jennifer M Yamamoto, Katherine Young, Chloé Amouyal, Mette K Andersen, Maxine P Bonham, Mingling Chen, Feifei Cheng, Tinashe Chikowore, Sian C Chivers, Christoffer Clemmensen, Dana Dabelea, Adem Y Dawed, Aaron J Deutsch, Laura T Dickens, Linda A DiMeglio, Monika Dudenhöffer-Pfeifer, Carmella Evans-Molina, María Mercè Fernández-Balsells, Hugo Fitipaldi, Stephanie L Fitzpatrick, Stephen E Gitelman, Mark O Goodarzi, Jessica A Grieger, Marta Guasch-Ferré, Nahal Habibi, Torben Hansen, Chuiguo Huang, Arianna Harris-Kawano, Heba M Ismail, Benjamin Hoag, Randi K Johnson, Angus G Jones, Robert W Koivula, Aaron Leong, Gloria K W Leung, Ingrid M Libman, Kai Liu, S Alice Long, William L Lowe, Jr, Robert W Morton, Ayesha A Motala, Suna Onengut-Gumuscu, James S Pankow, Maleesa Pathirana, Sofia Pazmino, Dianna Perez, John R Petrie, Camille E Powe, Alejandra Quinteros, Rashmi Jain, Debashree Ray, Mathias Ried-Larsen, Zeb Saeed, Vanessa Santhakumar, Sarah Kanbour, Sudipa Sarkar, Gabriela S F Monaco, Denise M Scholtens, Elizabeth Selvin, Wayne Huey-Herng Sheu, Cate Speake, Mag-

gie A Stanislawski, Nele Steenackers, Andrea K Steck, Norbert Stefan, Julie Støy, Rachael Taylor, Sok Cin Tye, Gebresilasea Gendisha Ukke, Marzhan Urazbayeva, Bart Van der Schueren, Camille Vatier, John M Wentworth, Wesley Hannah, Sara L White, Gechang Yu, Yingchai Zhang, Shao J Zhou, Jacques Beltrand, Michel Polak, Ingvild Aukrust, Elisa de Franco, Sarah E Flanagan, Kristin A Maloney, Andrew McGovern, Janne Molnes, Mariam Nakabuye, Pål Rasmus Njølstad, Hugo Pomares-Millan, Michele Provenzano, Cécile Saint-Martin, Cuilin Zhang, Yeyi Zhu, Sungyoung Auh, Russell de Souza, Andrea J Fawcett, Chandra Gruber, Eskedar Getie Mekonnen, Emily Mixter, Diana Sherifali, Robert H Eckel, John J Nolan, Louis H Philipson, Rebecca J Brown, Liana K Billings, Kristen Boyle, Tina Costacou, John M Dennis, Jose C Florez, Anna L Gloyn, Maria F Gomez, Peter A Gottlieb, Siri Atma W Greeley, Kurt Griffin, Andrew T Hattersley, Irl B Hirsch, Marie-France Hivert, Korey K Hood, Jami L Josefson, Soo Heon Kwak, Lori M Laffel, Siew S Lim, Ruth J F Loos, Ronald C W Ma, Chantal Mathieu, Nestoras Mathioudakis, James B Meigs, Shivani Misra, Viswanathan Mohan, Rinki Murphy, Richard Oram, Katharine R Owen, Susan E Ozanne, Ewan R Pearson, Wei Perng, Toni I Pollin, Rodica Pop-Busui, Richard E Pratley, Leanne M Redman, Maria J Redondo, Rebecca M Reynolds, Robert K Semple, Jennifer L Sherr, Emily K Sims, Arianne Sweeting, Tiinamaija Tuomi, Miriam S Udler, Kimberly K Vesco, Tina Vilsbøll, Robert Wagner, Stephen S Rich, and Paul W Franks. Second international consensus report on gaps and opportunities for the clinical translation of precision diabetes medicine. *Nat. Med.*, 29(10):2438–2457, October 2023.

[40] Emma Ahlqvist, Petter Storm, Annemari Käräjämäki, Mats Martinell, Mozhgan Dorkhan, Annelie Carlsson, Petter Vikman, Rashmi B Prasad, Dina Mansour Aly, Peter Almgren, Ylva Wessman, Nael Shaat, Peter Spégel, Hindrik Mulder, Eero Lindholm, Olle Melander, Ola Hansson, Ulf Malmqvist, Åke Lernmark, Kaj Lahti, Tom Forsén, Tiinamaija Tuomi, Anders H Rosengren, and Leif Groop. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *Lancet Diabetes Endocrinol*, 6(5):361–369, May 2018.

[41] Masato Akiyama. Multi-omics study for interpretation of genome-wide association study. *J. Hum. Genet.*, 66(1):3–10, January 2021.

[42] Christine Vogel and Edward M Marcotte. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.*, 13(4):227–232, March 2012.

[43] Robert Feil and Mario F Fraga. Epigenetics and the environment: emerging patterns and implications. *Nat. Rev. Genet.*, 13(2):97–109, January 2012.

[44] Jeremy K Nicholson and John C Lindon. Systems biology: Metabonomics. *Nature*, 455(7216):1054–1056, October 2008.

[45] J Larry Jameson and Dan L Longo. Precision medicine–personalized, problematic, and promising. *N. Engl. J. Med.*, 372(23):2229–2234, June 2015.

[46] Bruce Alberts. *Molecular Biology of the Cell.* W.W. Norton & Company, 6th edition edition, August 2017.

[47] Elaine R Mardis. Next-generation sequencing platforms. *Annu. Rev. Anal. Chem.*, 6:287–303, 2013.

[48] Sara Goodwin, John D McPherson, and W Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.*, 17(6):333–351, May 2016.

[49] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10(1):57–63, January 2009.

[50] Fan Zhang, Kevin Wei, Kamil Slowikowski, Chamith Y Fonseka, Deepak A Rao, Stephen Kelly, Susan M Goodman, Darren Tabechian, Laura B Hughes, Karen Salomon-Escoto, Gerald F M Watts, A Helena Jonsson, Javier Rangel-Moreno, Nida Meednu, Cristina Rozo, William Apruzzese, Thomas M Eisenhaure, David J Lieb, David L Boyle, Arthur M Mandelin, 2nd, Accelerating Medicines Partnership Rheumatoid Arthritis and Systemic Lupus Erythematosus (AMP RA/SLE) Consortium, Brendan F Boyce, Edward DiCarlo, Ellen M Gravallese, Peter K Gregersen, Larry Moreland, Gary S Firestein, Nir Hacohen, Chad Nusbaum, James A Lederer, Harris Perlman, Costantino Pitzalis, Andrew Filer, V Michael Holers, Vivian P Bykerk, Laura T Donlin, Jennifer H Anolik, Michael B Brenner, and Soumya Raychaudhuri. Defining inflammatory cell states in rheumatoid arthritis joint synovial tissues by integrating single-cell transcriptomics and mass cytometry. *Nat. Immunol.*, 20(7):928–942, July 2019.

[51] Ruedi Aebersold and Matthias Mann. Mass-spectrometric exploration of proteome structure and function. *Nature*, 537(7620):347–355, September 2016.

[52] Yen Chin Koay, Adelle C F Coster, Daniel L Chen, Brad Milner, Amani Batarseh, John F O'Sullivan, Jerry R Greenfield, and Dorit Samocha-Bonet. Metabolomics and lipidomics signatures of insulin resistance and abdominal fat depots in people living with obesity. *Metabolites*, 12(12), December 2022.

[53] Yaxing Zhao, Limsoon Wong, and Wilson Wen Bin Goh. How to do quantile normalization correctly for gene expression data analyses. *Sci. Rep.*, 10(1):15534, September 2020.

[54] O Troyanskaya, M Cantor, G Sherlock, P Brown, T Hastie, R Tibshirani, D Botstein, and R B Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, June 2001.

[55] Peter C Austin, Ian R White, Douglas S Lee, and Stef van Buuren. Missing data in clinical research: A tutorial on multiple imputation. *Can. J. Cardiol.*, 37(9):1322–1331, September 2021.

[56] Haonan Wang. Nonlinear dimensionality reduction by LEE, J. a. and VERLEYSEN, M. *Biometrics*, 65(2):665–665, May 2009.

[57] Milan Picard, Marie-Pier Scott-Boyer, Antoine Bodein, Olivier Périn, and Arnaud Droit. Integration strategies of multi-omics data for machine learning analysis. *Comput. Struct. Biotechnol. J.*, 19:3735–3746, June 2021.

[58] Yifeng Li, Fang-Xiang Wu, and Alioune Ngom. A review on machine learning principles for multi-view biological data integration. *Brief. Bioinform.*, 19(2):325–340, March 2018.

[59] Olivier Bodenreider and Robert Stevens. Bio-ontologies: current trends and future directions. *Brief. Bioinform.*, 7(3):256–274, September 2006.

[60] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, José Emilio Labra Gayo, Sabrina Kirrane, Sebastian Neumaier, Axel Polleres, Roberto Navigli, Axel-Cyrille Ngonga Ngomo, Sabbir M Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. Knowledge graphs. *arXiv [cs.AI]*, March 2020.

[61] Arun Bhatt. Evolution of clinical research: a history before and beyond james lind. *Perspect. Clin. Res.*, 1(1):6–10, January 2010.

[62] Carla Abdelnour, Federica Agosta, Marco Bozzali, Bertrand Fougère, Atsushi Iwata, Ramin Nilforooshan, Leonel T Takada, Félix Viñuela, and Martin Traber. Perspectives and challenges in patient stratification in alzheimer's disease. *Alzheimers. Res. Ther.*, 14(1):112, August 2022.

[63] Achilleas Thoma, Forough Farrokhyar, Leslie McKnight, and Mohit Bhandari. Practical tips for surgical research: how to optimize patient recruitment. *Can. J. Surg.*, 53(3):205–210, June 2010.

[64] Sarah Asad, Kathryn Kananen, Kurt R Mueller, W Fraser Symmans, Yujia Wen, Charles M Perou, James S Blachly, James Chen, Benjamin G Vincent, and Daniel G Stover. Challenges and gaps in clinical trial genomic data management. *JCO Clin Cancer Inform*, 6(1):e2100193, March 2022.

[65] Binny Krishnankutty, Shantala Bellary, Naveen B R Kumar, and Latha S Moodahadu. Data management in clinical research: An overview. *Indian J. Pharmacol.*, 44(2):168–172, March 2012.

[66] David B Fogel. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: A review. *Contemp Clin Trials Commun*, 11:156–164, September 2018.

[67] Isabella Wy Mak, Nathan Evaniew, and Michelle Ghert. Lost in translation: animal models and clinical trials in cancer treatment. *Am. J. Transl. Res.*, 6(2):114–118, January 2014.

[68] Kimberly Scearce-Levie, Pascal E Sanchez, and Joseph W Lewcock. Leveraging preclinical models for the development of alzheimer disease therapeutics. *Nat. Rev. Drug Discov.*, 19(7):447–462, July 2020.

[69] Maximilian Kleinert, Christoffer Clemmensen, Susanna M Hofmann, Mary C Moore, Simone Renner, Stephen C Woods, Peter Huypens, Johannes Beckers, Martin Hrabe de Angelis, Annette Schürmann, Mostafa Bakhti, Martin Klingenspor, Mark Heiman, Alan D Cherrington, Michael Ristow, Heiko Lickert, Eckhard Wolf, Peter J Havel, Timo D Müller, and Matthias H Tschöp. Animal models of obesity and diabetes mellitus. *Nat. Rev. Endocrinol.*, 14(3):140–162, March 2018.

[70] Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin. Pharmacol. Ther.*, 69(3):89–95, March 2001.

[71] H Bart van der Worp, David W Howells, Emily S Sena, Michelle J Porritt, Sarah Rewell, Victoria O'Collins, and Malcolm R Macleod. Can animal models of disease reliably inform human studies? *PLoS Med.*, 7(3):e1000245, March 2010.

[72] Ismail Kola and John Landis. Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.*, 3(8):711–715, August 2004.

[73] Florian Prinz, Thomas Schlange, and Khusru Asadullah. Believe it or not: how much can we rely on published data on potential drug targets? *Nat. Rev. Drug Discov.*, 10(9):712, August 2011.

[74] Aliya Fatehullah, Si Hui Tan, and Nick Barker. Organoids as an in vitro model of human development and disease. *Nat. Cell Biol.*, 18(3):246–254, March 2016.

[75] W F Scherer, J T Syverton, and G O Gey. Studies on the propagation in vitro of poliomyelitis viruses. IV. viral multiplication in a stable strain of human malignant epithelial cells (strain HeLa) derived from an epidermoid carcinoma of the cervix. *J. Exp. Med.*, 97(5):695–710, May 1953.

[76] John R Masters. HeLa cells 50 years on: the good, the bad and the ugly. *Nat. Rev. Cancer*, 2(4):315–319, April 2002.

[77] Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A Margolin, Sungjoon Kim, Christopher J Wilson, Joseph Lehár, Gregory V Kryukov, Dmitriy Sonkin, Anupama Reddy, Manway Liu, Lauren Murray, Michael F Berger, John E Monahan, Paula Morais, Jodi Meltzer, Adam Korejwa, Judit Jané-Valbuena, Felipa A Mapa, Joseph Thibault, Eva Bric-Furlong, Pichai Raman, Aaron Shipway, Ingo H Engels, Jill Cheng, Guoying K Yu, Jianjun Yu, Peter Aspesi, Jr, Melanie de Silva, Kalpana Jagtap, Michael D Jones, Li Wang, Charles Hatton, Emanuele Palescandolo, Supriya Gupta, Scott Mahan, Carrie Sougnez, Robert C Onofrio, Ted Liefeld, Laura MacConaill, Wendy Winckler, Michael Reich, Nanxin Li, Jill P Mesirov, Stacey B Gabriel, Gad Getz, Kristin Ardlie, Vivien Chan, Vic E Myer, Barbara L Weber, Jeff Porter, Markus Warmuth, Peter Finan, Jennifer L Harris, Matthew Meyerson, Todd R Golub, Michael P Morrissey, William R Sellers, Robert Schlegel, and Levi A Garraway. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, March 2012.

[78] Peter Horvath, Nathalie Aulner, Marc Bickle, Anthony M Davies, Elaine Del Nery, Daniel Ebner, Maria C Montoya, Päivi Östling, Vilja Pietiäinen, Leo S Price, Spencer L Shorte, Gerardo Turcatti, Carina von Schantz, and Neil O Carragher. Screening out irrelevant cell-based models of disease. *Nat. Rev. Drug Discov.*, 15(11):751–769, November 2016.

[79] J R Masters. Human cancer cell lines: fact and fantasy. *Nat. Rev. Mol. Cell Biol.*, 1(3):233–236, December 2000.

[80] Sylvia Merkert and Ulrich Martin. Targeted gene editing in human pluripotent stem cells using site-specific nucleases. *Adv. Biochem. Eng. Biotechnol.*, 163:169–186, 2018.

[81] Susan Breslin and Lorraine O'Driscoll. Three-dimensional cell culture: the missing link in drug discovery. *Drug Discov. Today*, 18(5-6):240–249, March 2013.

[82] Kenneth Paigen. One hundred years of mouse genetics: an intellectual history. I. the classical period (1902-1980). *Genetics*, 163(1):1–7, January 2003.

[83] Nadia Rosenthal and Steve Brown. The mouse ascending: perspectives for human-disease models. *Nat. Cell Biol.*, 9(9):993–999, September 2007.

[84] Monica J Justice and Paraminder Dhillon. Using the mouse to model human disease: increasing validity and reproducibility. *Dis. Model. Mech.*, 9(2):101–103, February 2016.

[85] D Games, D Adams, R Alessandrini, R Barbour, P Berthelette, C Blackwell, T Carr, J Clemens, T Donaldson, and F Gillespie. Alzheimer-type neuropathology in transgenic mice overexpressing V717F beta-amyloid precursor protein. *Nature*, 373(6514):523–527, February 1995.

[86] L A Donehower, M Harvey, B L Slagle, M J McArthur, C A Montgomery, Jr, J S Butel, and A Bradley. Mice deficient for p53 are developmentally normal but susceptible to spontaneous tumours. *Nature*, 356(6366):215–221, March 1992.

[87] Lee M Silver. *Mouse Genetics Concepts and Applications.* Oxford University Press, April 1995.

[88] C J Mann. Observational research methods. research design II: cohort, cross sectional, and case-control studies. *Emerg. Med. J.*, 20(1):54–60, January 2003.

[89] W B Kannel. Some lessons in cardiovascular epidemiology from framingham. *Am. J. Cardiol.*, 37(2):269–282, February 1976.

[90] G A Colditz, J E Manson, and S E Hankinson. The nurses' health study: 20-year contribution to the understanding of health among women. *J. Womens. Health*, 6(1):49–62, February 1997.

[91] Nicole Rusk. The UK biobank. *Nat. Methods*, 15(12):1001, December 2018.

[92] J Concato, N Shah, and R I Horwitz. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N. Engl. J. Med.*, 342(25):1887–1892, June 2000.

[93] Morten Schmidt, Sigrun Alba Johannesdottir Schmidt, Kasper Adelborg, Jens Sundbøll, Kristina Laugesen, Vera Ehrenstein, and Henrik Toft Sørensen. The danish health care system and epidemiological research: from health care contacts to database records. *Clin. Epidemiol.*, 11:563–591, July 2019.

[94] William R Hersh. Medical informatics: improving health care through information. *JAMA*, 288(16):1955–1958, 2002.

[95] Gemmae M Fix, Bo Kim, Mollie Ruben, and Megan B McCullough. Direct observation methods: a practical guide for health researchers. *PEC Innov*, 1, December 2022.

[96] Cathryn Tonne and Paul Wilkinson. Long-term exposure to air pollution is associated with survival following acute coronary syndrome. *Eur. Heart J.*, 34(17):1306–1311, May 2013.

[97] S Lee, Y Xu, A G D Apos Souza, E A Martin, C Doktorchik, Z Zhang, and H Quan. Unlocking the potential of electronic health records for health research. *Int J Popul Data Sci*, 5(1):1123, January 2020.

[98] Laura Annaratone, Giuseppe De Palma, Giuseppina Bonizzi, Anna Sapino, Gerardo Botti, Enrico Berrino, Chiara Mannelli, Pamela Arcella, Simona Di Martino, Agostino Steffan, Maria Grazia Daidone, Vincenzo Canzonieri, Barbara Parodi, Angelo Virgilio Paradiso, Massimo Barberis, Caterina Marchiò, and Alleanza Contro il Cancro (ACC) Pathology and Biobanking Working Group. Basic principles of biobanking: from biological samples to precision medicine for patients. *Virchows Arch.*, 479(2):233–246, August 2021.

[99] R Holle, M Happich, H Löwel, H E Wichmann, and MONICA/KORA Study Group. KORA–a research platform for population based health research. *Gesundheitswesen*, 67 Suppl 1:S19–25, August 2005.

[100] Mathieu Ravaut, Hamed Sadeghi, Kin Kwan Leung, Maksims Volkovs, Kathy Kornas, Vinyas Harish, Tristan Watson, Gary F Lewis, Alanna Weisman, Tomi Poutanen, and Laura Rosella. Predicting adverse outcomes due to diabetes complications with machine learning using administrative health data. *NPJ Digit Med*, 4(1):24, February 2021.

[101] Lukasz Piwek, David A Ellis, Sally Andrews, and Adam Joinson. The rise of consumer health wearables: Promises and barriers. *PLoS Med.*, 13(2):e1001953, February 2016.

[102] Mohamad Amin Pourhoseingholi, Ahmad Reza Baghestani, and Mohsen Vahedi. How to control confounding effects by statistical analysis. *Gastroenterol Hepatol Bed Bench*, 5(2):79–83, 2012.

[103] Neal D Freedman, Yikyung Park, Christian C Abnet, Albert R Hollenbeck, and Rashmi Sinha. Association of coffee drinking with total and cause-specific mortality. *N. Engl. J. Med.*, 366(20):1891–1904, May 2012.

[104] M I Jordan and T M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, July 2015.

[105] Carla Márquez-Luna, Po-Ru Loh, South Asian Type 2 Diabetes (SAT2D) Consortium, SIGMA Type 2 Diabetes Consortium, and Alkes L Price. Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genet. Epidemiol.*, 41(8):811–823, December 2017.

[106] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.

[107] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queueing Syst.*, 16(3):31–57, June 2018.

[108] Meng Song, Jonathan Greenbaum, Joseph Luttrell, 4th, Weihua Zhou, Chong Wu, Hui Shen, Ping Gong, Chaoyang Zhang, and Hong-Wen Deng. A review of integrative imputation for multi-omics datasets. *Front. Genet.*, 11:570255, October 2020.

[109] Payam Refaeilzadeh, Lei Tang, and Huan Liu. Cross-validation. In Ling Liu and M Tamer Özsu, editors, *Encyclopedia of Database Systems*, pages 532–538. Springer US, Boston, MA, 2009.

[110] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv [cs.LG]*, September 2016.

[111] Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, October 2001.

[112] George Davey Smith and Gibran Hemani. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.*, 23(R1):R89–98, September 2014.

[113] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv [cs.CL]*, June 2017.

[114] Xuhong Li, Haoyi Xiong, Xingjian Li, Xuanyu Wu, Xiao Zhang, Ji Liu, Jiang Bian, and Dejing Dou. Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. *arXiv [cs.LG]*, March 2021.

[115] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *arXiv [cs.AI]*, May 2017.

[116] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *arXiv [cs.LG]*, March 2017.

[117] Dieudonne van der Meer, Syd Barthorpe, Wanjuan Yang, Howard Lightfoot, Caitlin Hall, James Gilbert, Hayley E Francies, and Mathew J Garnett. Cell model passports-a hub for clinical, genetic and functional datasets of preclinical cancer models. *Nucleic Acids Res.*, 47(D1):D923–D929, January 2019.

[118] Antonio Fabregat, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, Bijay Jassal, Florian Korninger, Bruce May, Marija Milacic, Corina Duenas Roca, Karen Rothfels, Cristoffer Sevilla, Veronica Shamovsky, Solomon Shorser, Thawfeek Varusai, Guilherme Viteri, Joel Weiser, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The reactome pathway knowledgebase. *Nucleic Acids Res.*, 46(D1):D649–D655, January 2018.

[119] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, 45(D1):D353–D361, January 2017.

Phong BH Nguyen

LinkedIn: www.linkedin.com/in/phngbh-2019

Github: https://github.com/phngbh

Google Scholar: Phong BH Nguyen

EDUCATION & QUALIFICATION

Jun 2020 till now	**PhD Candidate** Computational Health Center – Helmholtz Munich Interpretable integration of multimodal datasets to build prognostic models and understand the pathology of complex diseases
Oct 2017 – Mar 2020	**Master of Science in Biology** Ludwig-Maximilian University of Munich Thesis: Development and benchmarking of a highly sensitive and efficient RNA-seq method Current GPA: 1.3 (Very good)
Sep 2011 – Jun 2015	**Bachelor of Science in Biotechnology** International University – Vietnam National University in HCMC Thesis: Optimization of High Resolution Melt Analysis for Genotyping SNP rs895819 in association with breast cancer in Vietnamese Population GPA: 3.2/4 (Good)

RESEARCH/TRAINING/INTERNSHIP

Oct 2019-Mar 2020	**INSTITUTE OF COMPUTATIONAL BIOLOGY-HELMHOLTZ CENTER MUNICH** **Intern** Investigating impact of ancestry information on differential drug response in high throughput screen of cancer cell lines.
Oct 2018-Dec 2018	**BIOINFORMATICS CORE FACILITY-BIOMEDICAL CENTER, LMU MUNICH** **Intern** Investigation of dosage compensation in gene expression using a published scRNA-seq data in Drosophila melanogaster.

PERSONAL SKILLS:

Languages	Vietnamese (mother tongue), English (C1), German (B1)
Programming	R, Python, Perl, UNIX, shell scripting

Computation High Performance Computing, version control, containerization, data visualization, bioinformatics pipelines

PUBLICATIONS:

July 2024 Phong BH Nguyen, Daniel Garger, Diyuan Lu, Haifa Maalmi, ProfileHolger Prokisch, Barbara Thorand, Jerzy Adamski, Gabi Kastenmüller, Melanie Waldenberger, Christian Gieger, Annette Peters, Karsten Suhre, Gidon J Bönhof, Wolfgang Rathmann, Michael Roden, Harald Grallert, Dan Ziegler, Christian Herder, Michael P Menden, *The interpretable multimodal machine learning framework (IMML) revealed pathological signatures of distal sensorimotor polyneuropathy*, accepted to Communications Medicine

Mar 2022 Aleksandar Janjic, Lucas E Wange, Johannes W Bagnoli, Johanna Geuder, Phong BH Nguyen, Daniel Richter, Beate Vieth, Binje Vick, Irmela Jeremias, Christoph Ziegenhain, Ines Hellmann, Wolfgang Enard, *Prime-seq, efficient and powerful bulk RNA sequencing*, Genome Biol volume 23, Article number: 88 (2022). https://doi.org/10.1186/s13059-022-02660-8

Sep 2021 Phong BH Nguyen, Alexander J Ohnmacht, Samir Sharifli, Mathew J Garnett, Michael P Menden. *Inferred Ancestral Origin of Cancer Cell Lines Associates with Differential Drug Response*, Int. J. Mol. Sci. 2021, 22(18), 10135; https://doi.org/10.3390/ijms221810135

CONFERENCE ATTENDANCES:

Oct 2022 European Association for the Study of Diabetes (EASD) 58th Annual Meeting, *Artificial Intelligence and Machine Learning in Diabetes Research*, Keynote talk

July 2023 Intelligent Systems for Molecular Biology (ISMB) 31st Conference, *The Interpretable Multimodal Machine Learning (IMML) framework reveals pathological signatures of distal sensorimotor polyneuropathy*, Poster presentation

Sep 2024 European Conference in Computational Biology (ECCB) 23rd Conference, *Leveraging pretrained large language model and longitudinal medical records for prognosis of type 3 diabetes,* Poster presentation

REFERENCES:

Prof. Dr. Wolfgang Enard
Chair of Department Biology II
Anthropology & Human Genomics Dept.
Contact: enard@bio.lmu.de

Dr. Tobias Straub
Head of Bioinformatics Core Facility
Biomedical Center - LMU Munich
Contact: tstraub@med.uni-muenchen.de

Dr. Michael Menden
Group Leader
Institute of Computational Biology, HZM
Contact: michael.menden@helmholtz-muenchen.de