AI FOUNDATION MODELS FOR PERSONALIZED HEALTH MONITORING: LEARNING MEANINGFUL REPRESENTATIONS OF METABOLIC PROFILES

Raphael Kozlovsky, Tobias Ameismeier & Roland Geyer
lifespin GmbH
Am Biopark 13
93053 Regensburg, Germany
raphael.kozlovsky@lifespin.health

INTRODUCTION

Metabolism plays a central role in health, with most diseases directly or indirectly linked to metabolic processes. A comprehensive understanding of these connections forms the basis for more precise diagnoses and individualized therapies. Nuclear Magnetic Resonance (NMR) spectroscopy of blood offers a remarkably broad and detailed data foundation, providing direct insights into a person's metabolic state. However, conventional methods fall short of fully exploiting this potential due to the inherent complexity and richness of NMR spectral data.

Advances in artificial intelligence (AI)—particularly in the areas of deep learning, Transformer architectures, and Foundation Models—offer a promising pathway to overcome these challenges. By leveraging AI's capacity for recognizing patterns and structures within large datasets, NMR metabolomics could achieve a significant leap forward, ultimately establishing new standards in personalized medicine. To this end, we are developing an AI foundation model that operates directly on 1D high-resolution NMR-spectra from blood samples, providing a versatile platform that can be fine-tuned to extract clinically relevant insights across a wide range of medical applications.

CURRENT STATE

We maintain a substantial database of digitized blood samples (over 200k samples), spanning more than 600 distinct diseases. Currently, concentrations of specific substances (around 200 parameters) are extracted from blood NMR spectra through rule-based algorithms, which are inherently complex and computationally intensive. These parameters support health status predictions for donors, but they tap only a fraction of the total information available in NMR spectra (which can contain around 100k data points). The proposed approach aims to harness the full spectral data to substantially increase the scope of insights drawn from each sample. Although still in an early development stage, initial results from Transformer-based model architectures are promising.

TECHNICAL APPROACH

We are developing a Transformer-based AI foundation model capable of learning the intricate patterns within NMR spectra. The approach involves two key phases:

1. **Pre-training (Masked-Data Modeling):** The model will be exposed to a large corpus of unlabeled NMR spectra (both real and artificial), without associated metadata or clinical labels. During this unsupervised learning phase, it will autonomously discern fundamental patterns and dependencies within the spectral data.
2. **Fine-Tuning:** The pre-trained model is then fine-tuned on a smaller set of labeled spectra, allowing it to specialize in specific tasks facilitating targeted applications, including disease diagnosis and the assessment of metabolic health for use in precision medicine and preventative care.

As we develop an AI Foundation Model tailored to NMR-based metabolomics, several key challenges and questions guide our efforts:

- **Spectrum Tokenization:** How can NMR signals be partitioned into meaningful tokens to facilitate efficient learning in Transformer-based models?
- **Hybrid Architectures:** Can convolutional neural networks (CNNs) effectively extract local features before passing data to a Transformer? By integrating CNN-driven local feature extraction with the Transformer's global attention, we seek to reduce the data demands of attention mechanisms while retaining critical spectral information.
- **Biomarker Discovery via Attention:** Attention mechanisms can highlight spectral regions most relevant to disease states. We plan to investigate whether these patterns can serve as a pathway to identify novel biomarkers, thereby advancing both diagnostic capabilities and clinical research.

CONTRIBUTION AND IMPACT

This work introduces an AI foundation model that directly processes NMR spectra, rather than relying on predefined metabolite panels. Key advantages of this approach include:

- **Comprehensive Spectral Utilization:** The model leverages the information from the entire NMR-spectrum, including unresolved signals from larger molecules and other spectral features that traditional methods disregard.
- **Efficient Use of Data:** Unsupervised pre-training enables the model to learn structural patterns from large unlabeled datasets, minimizing the need for extensive labeled data and facilitating adaptation to new tasks, such as rare disease detection.
- **Improved Inference Efficiency:** The model enables rapid spectral interpretation, providing a computationally efficient alternative to rule-based methods for real-time clinical and research applications.
- **Adaptability to Future Applications:** As a foundation model, it can be fine-tuned for diverse metabolomics tasks, maintaining relevance in evolving research and clinical settings. The model encodes high-dimensional spectral data into latent representations, which serve as the core processing units, capturing key biochemical patterns of the donor's blood. These representations enable clustering, classification, and anomaly detection while integrating seamlessly into broader AI-driven workflows. By fine-tuning the foundation model on labeled spectra (e.g., spectra annotated with disease labels), it processes these representations to deliver comprehensive health evaluations.

By developing and training an AI foundation model to process the intricate structures in NMR-spectra, this project aims improve diagnostic precision and contribute to the next generation of healthcare solutions.

MEANINGFULNESS STATEMENT

Metabolism is essential for life, transforming nutrients into the energy our bodies need to function. Blood serves as the central medium for these metabolic processes, continuously reflecting changes in our biochemical state. NMR spectroscopy captures this biochemical activity in detail, providing a snapshot of metabolic processes and thus life itself. By training a foundation model on NMR spectra, we aim to extract the underlying information and distill them into meaningful representations. These representations serve as a bridge between raw spectral data and fundamental biological insights, helping us better understand how metabolism influences health and disease.