Part A – Applicant

A.1 Applicant

Name student (initials, first name, last name, student number): DPMV, Dylan, Maassen-Veeters, 2808196

Affiliation (university/institute + department): Utrecht University, MScs Bioinformatics and Biocomplexity

Name first examiner: Ernest Diez Benavente

Affiliation (university/institute + department): Utrecht University, UMC, Heart and Lungs Division, Experimental Cardiology

Name second examiner: Michal Mokry

Affiliation (university/institute + department): Utrecht University, UMC, Heart and Lungs Division, Central Diagnostics Laboratory

Part B – Scientific proposal

B.1 BASIC DETAILS

B.1.1 Title

Deep Learning-Driven Coronary Artery Disease Diagnostics leveraging Cell-Free DNA Methylation

B.1.2 Abstract

Coronary artery disease (CAD), caused by atherosclerotic plaque buildup within coronary arteries, usually remains undiagnosed until severe symptoms occur, such as heart attacks (1-4). CAD affects approximately 200 million people globally, a leading cause of death in every population (1-3). The current 'gold standard' for CAD diagnostics is cardiac catheterization and angiogram, a costly and invasive procedure (2). Collectively coronary heart diseases (CHDs) cost approximately 11% of the EU healthcare budget over 77 billion euros annually. Due to the high prevalence and costs associated with CAD, novel diagnostic and predictive tests are ever more necessary. Here we propose the introduction of cell-free DNA (cfDNA) diagnostics, already leveraged in cancer and prenatal diagnostics, within the cardiac diagnostic field (5,6,8,9,13-15). CfDNA is released into blood from dying cells, it contains sequence and methylation information from their tissue of origin throughout the body. While varying cfDNA compositions have been described for cardiovascular diseases, reliable biomarkers and diagnostic tests have yet to be determined (10-12). Therefore, we propose the incorporation of natural language processing to create novel deep learning approaches using cfDNA for a classification model to differentiate CAD between patients. Using the novel human methylation atlas as a pre-training data corpus and creating novel tokenizers we aim to pre-train and fine tune a novel methylation language model capable of incorporating DNA methylation within the model pre-training (17). Furthermore, leveraging language model's unique capabilities to track model attention mechanisms we can calculate which parts of the sequence or sequences help differentiate CAD patients from healthy, allowing us to further speculate alternative biomarkers (16). Given the success of such a model we can significantly impact the clinical setting of CAD diagnostics as well as create a novel language model which can be fine-tuned for various DNA methylation tasks. Furthermore, this approach could be leveraged to identify additional CHDs in the future.

B.1.3 Layman's summary

Cardiovascular diseases are the largest contributors to death globally. Coronary artery disease (CAD), a common cardiovascular disease (CVD) is a result of plaque buildup in the coronary arteries, which typically begins with basic symptoms such as chest pain. CAD usually goes unnoticed until life-threatening symptoms occur such as heart attacks. During the CAD diagnosis patients must undergo multiple cardiovascular tests, the final and most conclusive tests are coronary catheterization

and angiograms, which are extremely invasive and expensive. Due to the prevalence of this disease and the increasing need for innovative diagnostic tools, we propose the use of cell-free DNA (cfDNA). Recent advancements in sequencing technologies have initiated the research of cfDNA, which can be sequenced from a blood sample and currently can be used as a diagnostic tool for cancer, organ transplant compatibility, and pregnancy complications. More recently it has been shown that patients with CVDs have different cfDNA patterns than healthy patients, and we suspect this can be used as a novel diagnostic tool for CAD patients. Currently cfDNA research is being propelled using artificial intelligence, but traditional models are not yet sufficient for clinical implementation. State-of-the-art artificial intelligence models, Language Models, have recently entered the scene of cfDNA research for cancer diagnosis. These models excel at capturing information, patterns, and elements from DNA sequences, and in cfDNA cancer research are able to identify tumorous sequences and estimate tumour presence in a patient. We suspect using this model and incorporating DNA methylation into a language model we can create an accurate diagnostic alternative for CAD disease. Once trained, language models can be fine-tuned to perform a wide variety of tasks depending on the type of data. We aim to fine-tune our language model to take the cfDNA sequences of a person and predict whether someone has CAD. Building off this first task, we want to predict how at risk a patient is for CAD. Once we've accomplished these goals, we will perform a study to identify what is the minimum amount of data to perform an accurate prediction of CAD. This can allow us to cut down of the amount of data required. Separately, we want to use language models to research the biological mechanisms of CAD. Language models allow us to identify which parts of the sequence is the model paying attention to. If we can determine these specific parts of the genome from the cfDNA, we can create a list of potentially relevant biological markers for research.

B.1.4 Keywords

Language Model, Cell-free DNA, DNA Methylation, Coronary Artery Disease Diagnoses, Deep Learning

B.2 SCIENTIFIC PROPOSAL

B.2.1 Research Topic

Coronary Heart Disease (CHD) is one of the leading causes of death globally in both developing and developed countries (1). In 2021, CHD accounted for approximately 34% of deaths in Europe, and cost 11% of the EU healthcare costs, over 77 billion euros annually (3). One of the leading underlying causes of CHD is coronary artery disease (CAD) caused by atherosclerotic plaques in the lumen of coronary arteries. These plaques can grow, rupture, or erode impeding the normal blood flow (2). Untreated CAD eventually leads to myocardial infarctions (heart attacks) in which complete artery blockage prevents oxygen and blood from reaching the heart. CAD typically develops over decades and **minor symptoms usually go unnoticed until severe symptoms arise, such as myocardial infarctions**. CAD has been linked to several genetic variants (5). However, studies claim lifestyle and environmental factors play a much larger role (1,2,4). CAD has been shown to affect all ethnicities. Furthermore, men are more at risk than women, however after the age of 65 these differences gradually diminish. The general symptoms of CAD, chest pain and troubled breathing, are shared with numerous other differential diagnoses. Thus, **patients suspected of CAD are required to undergo numerous tests** to confirm CAD as the cause of their symptoms. **The current 'gold standard' diagnostics are cardiovascular catheterization and angiograms, being particularly invasive and costly procedures**. Furthermore, not all patients are eligible for such procedures, due to potential kidney damage, allergic reactions, or additional problems (2). As human life expectancy continues to grow, and sedentary lifestyles become more prevalent the need for alternative CHD diagnostic and predictive tools becomes ever more necessary. Given the severity and prevalence of CHDs globally, predictive assessments could provide opportunity for countermeasures before the onset of the debilitating and life-threatening CHD symptoms.

Recent advances in high-throughput methylation sequencing have initiated the emergence of cell-free DNA (cfDNA) sequencing as a diagnostic tool in cancer, organ transplant rejection, preeclampsia, and fetal chromosome abnormalities (6-10, 14, 15). During cell apoptosis and necrosis chromosomes are degraded by nuclease enzymes, however, nucleosome-wound DNA and their respective methylations remain intact. These sequences, averaging 165 bps, are released into the bloodstream where they remain for typically six hours until cleared by the liver (11). From a blood sample, these small cfDNA sequences contain information including DNA sequence, methylation modifications, and fragmentation patterns which can be leveraged to give real-time insight into cell and tissue-specific injuries. **Current research claims patients with CHDs have significantly increased cfDNA levels for specific tissues** when compared to healthy controls, as well as between different CHDs pathophysiology (12,13). Despite this, consistent CHD-specific biomarkers for cfDNA have failed to be determined and implemented.

Atherosclerosis formation and progression is a complex biological process that develops while ageing and is affected by a wide range of factors (namely dyslipidaemia, inflammation, blood pressure and others). This complexity is likely behind the difficulty in identifying biological biomarkers that can be linked to the presence of CAD, more specifically high-risk CAD. Therefore, **a more nuanced approach is required that can account for such complexity**. We propose the use of deep learning technologies. Deep learning is a division of machine learning which uses neural networks to learn and adapt to large datasets. Deep learning models for cfDNA cardiovascular diagnoses are currently focused on feed-forward and convolutional neural networks (14,15). While convolutional neural networks (CNNs) typically out-perform traditional feed-forward neural networks (NNs) they are significantly more computationally exhaustive. Both require extensive high-quality data to perform well and are conditionally outperformed by less demanding models such as random forests (RFs) and logistic regressions (LRs) (6,13). In cancer cfDNA deep learning models leverage the disparities between fragment lengths of healthy tissue and tumour cfDNA which does not occur in the normal cells of CHD patients (7,16). Therefore, **we propose the use natural language processing models to overcome traditional deep learning insufficiencies.** Language models (LMs), such as DNABERT, have emerged as state-of-the-art deep learning models which excel at capturing global contextual information between DNA sequences, with better results and a fraction of the computation cost of a convolutional neural network (17). **LMs outperform alternative deep learning models in data scarce environments** and are proficient at capturing patterns, subregions, and elemental aspects of DNA sequences entirely unsupervised (16,17)**.** Making it an optimal approach to understand cfDNA discrepancies towards CAD patients. Additionally, LM's multi-headed attention mechanisms can be leveraged to identify which elements of a sequence the model focuses attention. This aids in understanding the model's decision making and depending on the model's performance could give insight into unexplored biological elements. The leading methylation LM to date, MethylBERT, was pre-trained using the entire human genome which leveraged cfDNA for tumour sequences and yielded impressive capabilities for tissue deconvolution and tumour cfDNA classification. The developers have gone so far as to accurately predict tumour purity based on cfDNA inputs alone (16). The authors did not however, include DNA methylation within the pre-training of their language model, only leveraging DNA methylation for cfDNA deconvolution. Here we propose the creation of a novel language model, tokenizing methylated DNA all while pretraining on the first complete DNA-methylation atlas (18). These adjustments aim to further optimize deconvolutional capabilities while better leveraging specific DNA methylation patterns. Given enough high-quality CAD patient cfDNA data we will train a LM to classify and predict CAD, utilizing unique mechanisms of language models to observe sequence and methylation patterns the model identifies relevant for CAD diagnosis.

Project Goals:

1. **Train a general language model Incorporating DNA methylation.**
2. **Fine tune model to classify and predict CAD from coronary angiography cohort.**
3. **Use language model attention mechanisms to determine potential biomarkers through CAD classification.**

B.2.2 Approach

We propose a CAD diagnostic tool leveraging deep learning through the creation of a novel language model. This language model will employ patient cfDNA sequences and create embeddings, which will later be fine-tuned for CAD classification between patients. After fine-tuning optimization, we can begin to speculate patient discrepancies leveraged by the model in its predictions. Language models possess unique capabilities unlike other deep learning models, allowing observation of the model's attention, via its attention mechanisms. If the model focuses on a specific sequences or methylation patterns for a given diagnosis, we can begin speculating potential CHD biomarkers from biologically relevant sequences. After discovering specific methylations and sequences this information can be further researched to disentangle the biology of CHDs. We split the project into three tasks, pre-training, fine-tuning, and biomarker detection and model optimization as shown in Figure 1.

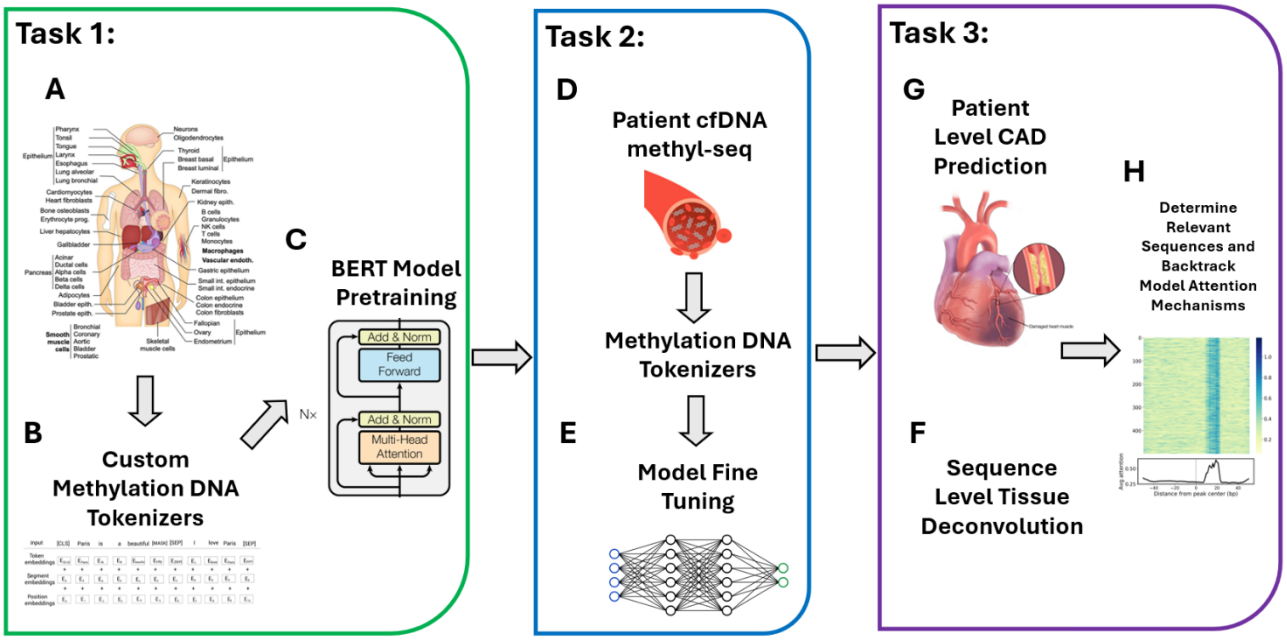

Figure 1: Experimental Workflow for creation of novel LM for CAD diagnosis. (A) Leverage normal cell methylation atlas as foundational training corpus for LM. (B) Create custom tokenizer, which takes methylated sequences and assigns values for each depending on sequence and methylation. This value is combined with positional value to make unique tokens. (C) Pretrain BERT model using tokenized methylation atlas whole genome sequences. (D) Run unseen patient cfDNA sequences through tokenizer and then pretrained BERT model to obtain model embeddings. (E) Fine tune model using patient cfDNA embeddings for two tasks. (F) The first fine-tuned task is tissue deconvolution per unique sequence, to estimate patient compositions. (G) Separately fine tune model for patient healthy and CAD classifications. (H) Use results from F and G to find unique and relevant sequences determining differentiation. Backtrack BERT model's attention mechanisms to determine methylations and sequences used for classifications.

Task 1. Creating a LM which incorporates methylation data from the human DNA methylation atlas:

We plan to utilize the methylation sequencing from the whole genome methylation atlas (18) within the LM's pre-training, where the model learns cfDNA context and methylation patterns. This is based on two fundamental concepts, the pre-training data corpus and the tokenization of sequences. LMs begin with a tokenization process in which n-kmer length sequences are converted to numeric representations for each unique segment, referred to as tokens as shown in Figure 2A-C. These tokens are further differentiated with positional values per token (Figure 2D-E). The tokens can then be input into the LM for pretraining or data analysis and embedding extraction. Previous methylation language models make a non-optimal use of methylation information, only incorporating it in fine-tuning. We aim to incorporate sequence methylation into the pre-training phase during the tokenization process by creating a custom tokenizer which incorporates DNA methylations, more specifically, methylated and unmethylated CpG sites as unique tokens (Figure 2C-D). Following the tokenization, the LM will be pretrained on methylated sequencing data unlike any DNA LM to date. Integrating the methylation data creates an additional layer of information within the pre-training from different tissues of the atlas. The model will be able to incorporate not only sequence-specific information, but also, tissue- and methylation-specific information, enhancing its capacity for comprehension and analysis.

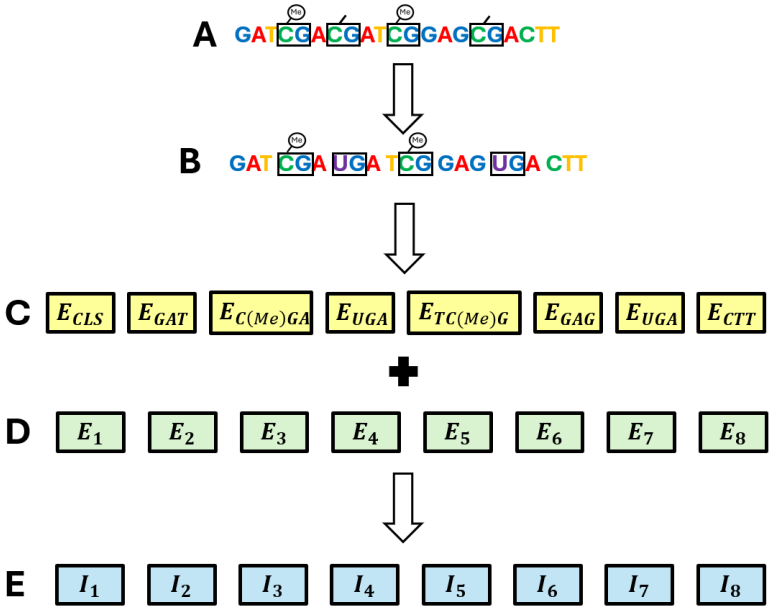

Figure 2: Workflow for custom methylation tokenizer. (A) Example methylation DNA sequence highlighting methylated and unmethylated CpG sites. (B) Methylation DNA sequences will initially be split into n length kmers for tokenization process. Unmethylated Cytosine from CpG sites are labelled as Uracil, while non-CpG Cytosines remain Cytosines. (C) Each sequence is converted to unique token value for that specific sequence, repeated sequences share the same numerical representation for sequence. We will include novel tokens for unique CpG sites, methylated and unmethylated. (D) The numerical representations for sequences are combined with positional values. (E) The combined sequence and positional values create the unique input embeddings to be leveraged by the pretrained BERT model.

We adopt the use of the complete normal cell methylation atlas as the foundation for the model pretraining opposed to the complete genome traditionally used in DNA language models (18). Initially, these whole epigenomes will be randomly broken up into a range of varying length sequences which are tokenized then input as the pretraining data corpus. **The model learns the contextual information of the methylated sequences here allowing it to differentiate cell types, tissue, and DNA activation through the process of Masked Language Modelling.** Throughout this process the model randomly masks tokens and learns to predict masked tokens. Similarly to MethylBERT we will follow DNABERT's optimized DNA pre-training parameters (17).

Task 2. Fine-tuning of the foundational model to predict CAD in patients cfDNA:

After pre-training, we can implement cfDNA sequences from healthy and CVD patients within the language model to extract embeddings and fine tune our model (Figure 1D). To classify patient's cfDNA embeddings they will be grouped per patient and trained on multiple ML classifiers, such as a neural network and random forest classifiers, to determine the most appropriate classifier (Figure 1E). Post fine-tuning the classifiers performance will be validated using unseen data. To verify the model's performance, the same sequences will be identically run through DNABERT, without methylation information, as well as one-hot encoded and run through the fine-tuning classifiers opposed to using the language model embeddings. MethylBERT demonstrated accurate prediction of individuals with tumours and tumour purity from 14 healthy controls, 40 colorectal cancer patients, and 44 pancreatic ductal adenocarcinoma patients. We currently have cfDNA sequences from 70 donors, 10 healthy controls, 40 CAD donors, and 20 alternative CHDs. To practically achieve our goals, we will need to increase our healthy controls by 10 donors, CAD patients by 10, costing approximately 20 thousand euros together. Additionally, DNABERT used an entire human genome as the pre-training data corpus and pre-trained their model for 25 days on 8 GPUs. Given that we plan to use the entire human methylation atlas for the entire genome, the computational costs for training computing capacity are estimated at 4 months with 8 GPUs.

We plan to fine tune cfDNA embeddings with two distinct approaches, cfDNA tissue specific deconvolution and patient specific CAD classification. First, we will fine tune the language model to deconvolute individual sequences within patients. Giving a real-time view on patient's cell death compositions (Figure 1F). This approach requires modification of the final layers

of the language model, specifying it for this task. Secondly, we can leverage the embeddings on a patient level, creating fine tuning models to classify patients with CAD (Figure 1G).

Task 3. CAD Biomarker Detection and Model Optimization:
While the combined patient cfDNA embeddings are to be used for CAD diagnoses, when combined with the CAD labels we can begin to contrast the disease with the controls. Leveraging cell-tissue compositions and tools such as UMAP we will further assess disparities between patient groups. Moreover, **leveraging BERT's attention mechanisms, we can calculate attention scores for various tokens from the input sequences.** This can help us further speculate relevant sequences and methylation patterns by determining which sequences the model focuses its attention during the fine-tuning processes.

Given the high costs of cfDNA methylation sequencing, it is essential we minimize the sequencing requirements to increase accessibility and feasibility for a clinical setting. New sequencing technologies such as Oxford Nanopore Sequencing can allow real-time sequencing, in a clinical setting for diagnosis. These can be equally as costly as modern diagnostic methods, while much less invasive, there is yet a great need to minimize the requirements needed for accurate diagnosis. We propose, beginning with a process of randomly sub-setting cfDNA sequences per patient to establish the minimum data requirements for accurate results from patient cfDNA sequencing.

B.2.3 Feasibility / Risk assessment

CfDNA has already proven to be a multifaceted diagnostic tool (6-10). However, there is still no feasible and straightforward methodology to bridge the gap between these cfDNA techniques and CAD diagnostics within a clinical setting. Traditional deep learning models have shown sufficient for other sects of cfDNA research such as transplant rejection and pregnancy complications (8,15). Nonetheless, creating alternative CHD diagnostic tools is beyond the scope of traditional deep learning models. Current models such as feed forward NNs require significant data contributions beyond what is currently available for CHDs. Alternatively, language models have proven contextual understanding of DNA sequences from the human genome alone. Therefore, the data prerequisites for construction are entirely feasible (Figure 3A). Furthermore, language models have sequence input length maximums ranging from 500-2000 bps allowing cfDNA sequences to optimally utilize LMs. The novel whole genome methylation atlas provides an ideal infrastructure for a novel methylation language model training on sequence methylation unlike MethylBERT. **This language model will be constructed to understand any human DNA methylation patterns, which can be leveraged in a range of methylated DNA applications across fields.**

Recent studies have elaborated on the significant differences in cfDNA profiles for CHD patients compared to healthy individuals, such as increased neutrophil and CD4+ cfDNA. (13). Naturally, language models for circulating DNA have already begun to emerge, MethylBERT, which when fine-tuned on tumour DNA, can accurately differentiate tumour DNA from normal cfDNA as well as use leverage DNA methylation post embedding to accurately estimate tumour purity per individual (16, 17). We aim to take this novel deep learning model and incorporate methylation data within the pre-training of the language model, unlike MethylBERT. Incorporating an additional layer of information for the model to learn from would potentially allow discovery of unforeseen context between patient and cfDNA features to their methylation and sequences. Post sequence specific fine-tuning, we believe this will allow for more accurate tissue deconvolution from cfDNA samples. As shown for DNABERT, the foundational DNA language model (17).

For any machine learning model, the demands for high quality data are infinite. While language models surpass other models when working with limited data quantities, the fine-tuning still preforms proportionally to the quantity of data. This is a major limitation in terms of maximizing model accuracy and performance for CAD diagnosis and prediction. Therefor the feasibility for an accurate CAD classification and tissue deconvolution is yet limited (Figure 3A). Liquid biopsies are recent sequencing innovations and are currently costly. However, as cfDNA exploratory research and practical applications continue to expand this will improve in time. The majority of publicly available cfDNA data is from cancer research. Although the controls samples from the experiments are cancer free, we do not know the cardiovascular health of these individuals. However, cohorts from similar CHD cfDNA research with similar phenotypes and selection processes would be useful contributions to our own datasets, especially regarding healthy CHD controls. Lastly, given inaccurate results from our CAD classifier, we can still leverage our embeddings in alternative analyses, fine tuning for more specific biological features within the data. The language model's unique attention mechanisms can show unbeknownst relevance between specific biological features of our methylation sequences. While LMs are currently state-of-the-art AI models, there remains risk for the models to overfit and underfit on irrelevant sequence elements rather than potential biomarkers (Figure 3A).

This project is not without risks, and a comprehensive assessment of the risks has been presented in Figure 3:

Project Goals	A Risk	B Impact
Language Model Methylation Tokenizer	Low Risk	Insignificant
Language Model DNA Methylation Atlas Pretraining	Low Risk	Significant
Model Fine Tuning for cfDNA CAD Diagnosis	High Risk	Very Significant
Model Fine Tuning for cfDNA Tissue Deconvolution	Medium Risk	Significant
CAD Biomarker Detection	Medium Risk	Very Significant

Figure 3: Risk assessment table for the proposal goals. (A) Feasibility for project goals. (B) Risk assessment for proposed goals, in respect to required work and costs. (C) Impact Potential for Clinical and Research Impact per project goal.

B.2.4 Scientific (a) and societal (b) impact

While the CAD diagnosis and predictive tools are the overarching goal of the project, the creation of a state-of-the-art language model fully incorporating methylation DNA data has major implications for the field of cfDNA research (Figure 3B). Pre-training on the novel complete human methylation atlas allows for any type of model fine-tuning for various DNA methylation tasks, not limited to cfDNA or CHD research. Given the unprecedented capabilities of DNABERT, which will share the same foundation as our own language model, we can expect to predict various features from methylated data and additionally accurately deconvolute cfDNA to the tissue level. While, inevitably, newer models will precede our own with larger data corpuses and computational resources, this model will lay the foundation for such advancements with little to no risk.

If the fine-tuning portion of our project were to be successful, it would significantly advance CAD diagnostics (Figure 3B). Given a high sensitivity, an initial test of patients presenting symptoms within the risk groups for CAD would provide enough proof to warrant invasive diagnostics, preventing the need for a series of increasingly invasive and costly tests. Contrarily, a high specificity from such tests, would better ensure patients undergoing cardiac catheterization and angiograms were not doing so unnecessarily. Given the severity of CAD symptoms an initial cfDNA blood test would be quick, given the convenience of on-site nanopore sequencing, minimally invasive predictions could be given the same day, warranting immediate preventive action. Treatment for CAD and other CHDs typically involves open heart surgery and such procedures do not occur without irrefutable assurance of diagnosis, which can only be provided via a cardiac angiogram. However, in time, as data availability and model capabilities increase, this will change given the widespread application of such models, as shown in various diagnostic application today.

CAD, and other CHD, diagnoses patients must spend considerable funds and undergo series of increasingly invasive tests to be confidently diagnosed. If successful, an AI based diagnostic tool could not only enhance the patients experience, doing a simple blood test in place of angiograms, but in the future become cheaper than the current diagnosis. Although the costs of cfDNA sequencing is currently high, approximately 1200 euros per patient, with the increase demand for onsite nanopore sequencing, and given the historical decrease in sequencing costs the likelihood of sequencing within a clinical setting is rapidly approaching. Furthermore, the costs of the sequencing only needs to be less than the collective costs of diagnostic tests per patient, which can range from electrocardiograms (EKGs), echocardiography tests, stress tests, chest x-rays, blood work, and cardiac catheterization and angiograms. Given the slow onset of CAD and the multi-faceted learning approach from LMs, we suspect embeddings will additionally highlight disparities between CAD patients with regards to CAD severity. This can later be further leveraged into novel prediction models for CAD onset and severity. Predictive tests and diagnostics for aging diseases, such as CHDs, could delay serious risk of myocardial infarctions by years in patients, preventing avoidable medical care costs and improving patients' quality of life.

B.2.5 Ethical considerations

It is important to note that while the data will be publicly available, only patient data given with consent will be used. Their anonymity will be maintained throughout this process using anonymised data in secure servers only within the UMC Utrecht.

B.2.6 Literature/references

1. Malakar AK, Choudhury D, Halder B, Paul P, Uddin A, Chakraborty S. A review on coronary artery disease, its risk factors, and therapeutics. *J Cell Physiol*. 2019; 234: 16812–16823. https://doi.org/10.1002/jcp.28350

2. Ergashov Behro'zjon Komilovich. (2023). Coronary Artery Disease. *EUROPEAN JOURNAL OF MODERN MEDICINE AND PRACTICE*, *3*(12), 81–87. Retrieved from https://www.inovatus.es/index.php/ejmmp/article/view/2186

3. Luengo-Fernandez R, Walli-Attaei M, Gray A, Torbica A, Maggioni AP, Huculeci R, Bairami F, Aboyans V, Timmis AD, Vardas P, Leal J. Economic burden of cardiovascular diseases in the European Union: a population-based cost study. Eur Heart J. 2023 Dec 1;44(45):4752-4767. doi: 10.1093/eurheartj/ehad583. PMID: 37632363; PMCID: PMC10691195.

4. Brown JC, Gerhardt TE, Kwon E. Risk Factors for Coronary Artery Disease. 2023 Jan 23. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan–. PMID: 32119297.

5. Aragam, K. G., Jiang, T., Goel, A., Kanoni, S., Wolford, B. N., Atri, D. S., Weeks, E. M., Wang, M., Hindy, G., Zhou, W., Grace, C., Roselli, C., Marston, N. A., Kamanu, F. K., Surakka, I., Venegas, L. M., Sherliker, P., Koyama, S., Ishigaki, K., Åsvold, B. O., … CARDIoGRAMplusC4D Consortium (2022). Discovery and systematic characterization of risk variants and genes for coronary artery disease in over a million participants. Nature genetics, 54(12), 1803–1815. https://doi.org/10.1038/s41588-022-01233-6

6. Cristiano, S., Leal, A., Phallen, J. *et al.* Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature* **570**, 385–389 (2019). https://doi.org/10.1038/s41586-019-1272-6

7. Chung DC, Gray DM 2nd, Singh H, Issaka RB, Raymond VM, Eagle C, Hu S, Chudova DI, Talasaz A, Greenson JK, Sinicrope FA, Gupta S, Grady WM. A Cell-free DNA Blood-Based Test for Colorectal Cancer Screening. N Engl J Med. 2024 Mar 14;390(11):973-983. doi: 10.1056/NEJMoa2304714. PMID: 38477985.

8. Agbor-Enoh S, Shah P, Tunc I, Hsu S, Russell S, Feller E, Shah K, Rodrigo ME, Najjar SS, Kong H, Pirooznia M, Fideli U, Bikineyeva A, Marishta A, Bhatti K, Yang Y, Mutebi C, Yu K, Kyoo Jang M, Marboe C, Berry GJ, Valantine HA; GRAfT Investigators. Cell-Free DNA to Detect Heart Allograft Acute Rejection. Circulation. 2021 Mar 23;143(12):1184-1197. doi: 10.1161/CIRCULATIONAHA.120.049098. Epub 2021 Jan 13. PMID: 33435695; PMCID: PMC8221834.

9. De Borre, M., Che, H., Yu, Q. *et al.* Cell-free DNA methylome analysis for early preeclampsia prediction. *Nat Med* **29**, 2206–2215 (2023). https://doi.org/10.1038/s41591-023-02510-5

10. Scarff KL, Flowers N, Love CJ, Archibald AD, Hunt CE, Giouzeppos O, Elliott J, Delatycki MB, Pertile MD. Performance of a cell-free DNA prenatal screening test, choice of prenatal procedure, and chromosome conditions identified during pregnancy after low-risk cell-free DNA screening. Prenat Diagn. 2023 Feb;43(2):213-225. doi: 10.1002/pd.6307. Epub 2023 Jan 21. PMID: 36617980.

11. Zemmour, H., Planer, D., Magenheim, J. *et al.* Non-invasive detection of human cardiomyocyte death using methylation patterns of circulating DNA. *Nat Commun* **9**, 1443 (2018). https://doi.org/10.1038/s41467-018-03961-y

12. Brusca SB, Elinoff JM, Zou Y, Jang MK, Kong H, Demirkale CY, Sun J, Seifuddin F, Pirooznia M, Valantine HA, Tanba C, Chaturvedi A, Graninger GM, Harper B, Chen LY, Cole J, Kanwar M, Benza RL, Preston IR, Agbor-Enoh S, Solomon MA. Plasma Cell-Free DNA Predicts Survival and Maps Specific Sources of Injury in Pulmonary Arterial Hypertension. Circulation. 2022 Oct 4;146(14):1033-1045. doi: 10.1161/CIRCULATIONAHA.121.056719. Epub 2022 Aug 25. PMID: 36004627; PMCID: PMC9529801.

13. Rafael R C Cuadrat, Adelheid Kratzer, Hector Giral Arnal, Anja C Rathgeber, Katarzyna Wreczycka, Alexander Blume, Irem B Gündüz, Veronika Ebenal, Tiina Mauno, Brendan Osberg, Minoo Moobed, Johannes Hartung, Kai Jakobs, Claudio Seppelt, Denitsa Meteva, Arash Haghikia, David M Leistner, Ulf Landmesser, Altuna Akalin, Cardiovascular disease biomarkers derived from circulating cell-free DNA methylation, *NAR Genomics and Bioinformatics*, Volume 5, Issue 2, June 2023, lqad061, https://doi.org/10.1093/nargab/lqad061

14. Khalil A, Bellesia G, Norton ME, Jacobsson B, Haeri S, Egbert M, Malone FD, Wapner RJ, Roman A, Faro R, Madankumar R, Strong N, Silver RM, Vohra N, Hyett J, Macpherson C, Prigmore B, Ahmed E, Demko Z, Ortiz JB, Souter V, Dar P. The Role of cfDNA Biomarkers and Patient Data in the Early Prediction of Preeclampsia: Artificial Intelligence Model. Am J Obstet Gynecol. 2024 Mar 1:S0002-9378(24)00380-6. doi: 10.1016/j.ajog.2024.02.299. Epub ahead of print. PMID: 38432413.

15. Bahado-Singh R, Friedman P, Talbot C, Aydas B, Southekal S, Mishra NK, Guda C, Yilmaz A, Radhakrishna U, Vishweswaraiah S. Cell-free DNA in maternal blood and artificial intelligence: accurate prenatal detection of fetal congenital heart defects. Am J Obstet Gynecol. 2023 Jan;228(1):76.e1-76.e10. doi: 10.1016/j.ajog.2022.07.062. Epub 2022 Aug 7. PMID: 35948071.

16. Jeong Y., Rohr K., Lutsik P.. MethylBERT: A Transformer-based model for read-level DNA methylation pattern identification and tumour deconvolution. bioRxiv 2023.10.29.564590; doi: https://doi.org/10/1101/2023.10.29.564590

17. Ji Y, Zhou Z, Liu H, Davuluri RV. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics. 2021;**37**: 2112–2120. doi: 10.1093/bioinformatics/btab083

18. Loyfer, N., Magenheim, J., Peretz, A. *et al.* A DNA methylation atlas of normal human cell types. *Nature* **613**, 355–364 (2023). https://doi.org/10.1038/s41586-022-05580-6