



DOCTORAL THESIS

Learning More from Less: Accurate and Trustworthy Foundation Models for Patient Trajectories

Ali Amirahmadi



Learning More from Less: Accurate and Trustworthy Foundation Models for Patient Trajectories

Ali Amirahmadi

Learning More from Less: Accurate and Trustworthy Foundation Models for Patient Trajectories

©Ali Amirahmadi

Halmstad University Dissertation No. 141

ISBN 978-91-90123-03-4 (printed)


ISBN 978-91-90123-04-1 (pdf)

Publisher: Halmstad University Press, 2026 | www.hh.se/hup

Printer: Media-Tryck, Lund



Media-Tryck is a Nordic Swan Ecolabel certified provider of printed material. Read more about our environmental work at www.mediatryck.lu.se

MADE IN SWEDEN 

To my wife

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Mattias Ohlsson and Farzaneh Etmnani, for their continuous guidance, encouragement, and scientific insight throughout my PhD journey. Your support, critical feedback, and high standards have shaped my thinking and helped me grow as a researcher. I am deeply thankful for your patience and for always making time for discussion and direction when it mattered most.

I am especially grateful to Jonas Björk and Olle Melander for steady advice, perspective, and encouragement along the way.

I also wish to thank Alessandro Tibo for the guidance and collaboration during my visiting period at AstraZeneca. That experience broadened my perspective. I truly appreciate your generosity with time and expertise.

My thanks also go to all my colleagues and peers. Your collaboration, discussions, and friendship have made this journey much richer and more enjoyable. A supportive research environment makes a tremendous difference, and I am grateful to have shared this time with you.

I would also like to acknowledge the faculty and administrative staff for their support behind the scenes. Your work enables everything we do as researchers.

Finally, and most importantly, I am forever grateful to my family for their unconditional support and belief in me. A very special thank you to my wife for her patience, encouragement, and constant presence throughout this journey. This work would not have been possible without you.

Abstract

Electronic health records (EHRs) contain longitudinal traces of patients’ interactions with the healthcare system. These *patient trajectories*—sequences of diagnoses, medications, and other events over time—offer opportunities to predict adverse outcomes early to intervene. In practice, however, EHR data are heterogeneous, temporally complex, and often available only in limited-sized cohorts with scarce labels. This thesis, *Learning More from Less: Accurate and Trustworthy Foundation Models for Patient Trajectories*, investigates how to build foundation-style models for such data.

The work is guided by the question: *How can we improve prediction and provide trustworthy explanations for adverse health outcomes by modeling longitudinal EHR trajectories?* It follows two tracks: (i) robust EHR-specific representation learning, and (ii) trustworthy modeling.

First, the thesis enriches self-supervised pretraining for structured EHR. A trajectory-order objective (TOO-BERT) teaches models to distinguish true temporal order from plausible permutations, while a source-masked objective model cross-sources dependencies. These objectives exploit the structure already present in trajectories, yielding stronger representations and improved prediction of incident outcomes.

Second, the thesis targets robust adaptation under label scarcity. Adaptive Noise-Augmented Attention (ANAA) perturbs and smoothly augments attention scores during fine-tuning, broadening overly sharp attention patterns and improving performance.

Third, the thesis develops explanation methods tailored to multimodal transformers EHR telemetry models. A manifold-aware baseline for Integrated Gradients keeps attribution paths in high-density regions of the representation space, improving faithfulness. Group-Sparse IG further adjusts the path schedule to produce sparse, token-level explanations that are more concise. Building on these methods, the thesis also proposes an approach to aggregate individual-level attributions into population-level insights for greater actionability, and applies it to identify key drivers of longevity and early mortality in the Malmö Diet and Cancer cohort.

Finally, the thesis explores uncertainty estimation in small, sequence-based datasets through a Gaussian process model with a decoupled global alignment kernel for peptide permeability prediction. This demonstrates how structured sequence

kernels can provide better accuracy and calibrated uncertainty when data are limited.

Overall, the thesis shows that in complex, data-scarce EHR settings, “learning more from less” requires making the pretraining, fine-tuning, and explanation stages explicitly reflect the structure of patient trajectories, leading to more accurate and trustworthy models for clinical risk prediction.

List of Papers

The following papers, referred to in the text by their Roman numerals, are included in this thesis.

PAPER I: Deep learning prediction models based on EHR trajectories: A systematic review

Ali Amirahmadi, Mattias Ohlsson, and Farzaneh Etminani. Journal of biomedical informatics, 2023.

PAPER II: Trajectory-Ordered Objectives for Self-Supervised Representation Learning of Temporal Healthcare Data Using Transformers: Model Development and Evaluation Study

Ali Amirahmadi, Farzaneh Etminani, Olle Melander, Jonas Björk, and Mattias Ohlsson. JMIR Medical Informatics 13.1, 2025.

PAPER III: A Masked language model for multi-source EHR trajectories contextual representation learning

Ali Amirahmadi, Farzaneh Etminani, Olle Melander, Jonas Björk, and Mattias Ohlsson. 33rd Medical Informatics Europe Conference (MIE2023), short paper, 2023.

PAPER IV: Adaptive noise-augmented attention for enhancing Transformer fine-tuning on longitudinal medical data.

Ali Amirahmadi, Farzaneh Etminani, and Mattias Ohlsson. Frontiers in Artificial Intelligence 8, 2025.

PAPER V: Group-Sparse Manifold-Aware Integrated Gradients for Multimodal Transformers on EHR Trajectories

Ali Amirahmadi, Farzaneh Etminani, and Mattias Ohlsson. Proceedings of the 5th Machine Learning for Health Symposium, PMLR, 2025.

PAPER VI: From Individual Attributions to Population Risk: Identifying Key Drivers

of Longevity and Early Mortality from Longitudinal EHR Data

Ali Amirahmadi, Farzaneh Etminani, Olle Melander, Jonas Björk, and Mattias Ohlsson. manuscript

PAPER VII: A decoupled alignment kernel for peptide membrane permeability predictions

Ali Amirahmadi, Gökçe Geylan, Leonardo De Maria, Farzaneh Etminani, Mattias Ohlsson, Alessandro Tibo. submitted.

The following papers are also contributed but not included in this thesis.

PAPER 1: Graph neural networks for clinical risk prediction based on electronic health records: A survey.

Heloísa Oss Boll, Ali Amirahmadi, Mirfarid Musavian Ghazani, Wagner Ourique de Moraes, Edison Pignaton de Freitas, Amira Soliman, Farzaneh Etminani, Stefan Byttner, Mariana Recamonde-Mendoza. *Journal of Biomedical Informatics*, 2024.

PAPER 2: Graph Neural Networks for Heart Failure Prediction on an EHR-Based Patient Similarity Graph

Heloisa Oss Boll, Ali Amirahmadi, Amira Soliman, Stefan Byttner, Mariana Recamonde-Mendoza. arXiv preprint arXiv:2411.19742. 2024.

PAPER 3: Developing a novel prediction model in opioid overdose using machine learning; a pilot analytical study

Sakhaee, Ehsan, Ali Amirahmadi, Morteza Mahdiani, Maziar Shojaei, Hossein Hassanian-Moghaddam, Roman Bauer, Nasim Zamani, Hossein Pakdaman, and Kourosh Gharagozli. *Health science reports*, 5(5), 2022.

Table of Contents

Acknowledgements	iii
Abstract	iv
List of Papers	vii
List of Figures	xiii
List of Tables	xv
1 Introduction	1
1.1 Introduction	1
1.2 Challenges	3
1.3 Research Questions	4
1.4 Contributions	6
2 Background	9
2.1 Foundation models	9
2.2 Self-supervised learning	10
2.2.1 Pretraining objectives	10
2.2.2 Adapting models to downstream tasks	11
2.3 Transformer and self-attention	13
2.4 Explainability in foundation models	16
2.5 Uncertainty estimation in foundation models	17
3 Summary of Included Papers	21
3.1 PAPER I: Deep learning prediction models based on EHR trajectories	21
3.1.1 Aim and Scope	21
3.1.2 Search Strategy and Inclusion	21
3.1.3 Methods Landscape	21
3.1.4 Trends and Gaps	22

3.2	PAPER II: Trajectory-Ordered Objectives (TOO-BERT) for Self-Supervised Representation Learning of Temporal Healthcare Data	23
3.2.1	Aim and Contribution	23
3.2.2	Method Overview	23
3.2.3	Key Results	24
3.3	PAPER III: A Masked language model for multi-source EHR trajectories	25
3.3.1	Aim and Contribution	25
3.3.2	Method overview	25
3.3.3	Key Results	26
3.4	PAPER IV: Adaptive noise-augmented attention (ANAA) for enhancing Transformer fine-tuning	26
3.4.1	Aim and Contribution	26
3.4.2	Method Overview	26
3.4.3	Key Results	26
3.5	PAPER V: Group-Sparse Manifold-Aware Integrated Gradients	28
3.5.1	Aim and Contribution	28
3.5.2	Method Overview	29
3.5.3	Key Results	30
3.6	PAPER VI: From Individual Attributions to Population Risk: Identifying Key Drivers of Longevity and Early Mortality from Longitudinal EHR Data	30
3.6.1	Aim and Contribution	30
3.6.2	Method Overview	31
3.6.3	Key Results	32
3.7	PAPER VII: A decoupled alignment kernel for peptide membrane permeability predictions	33
3.7.1	Aim and Contribution	33
3.7.2	Method Overview	33
3.7.3	Key Results	34
4	Concluding Remarks and Future Work	35
4.1	Concluding remarks	35
	References	41
	Appendix	47
A	PAPER I	47
B	PAPER II	61
C	PAPER III	87
D	PAPER IV	93

E	PAPER V	115
F	PAPER VI	137
G	PAPER VII	163

List of Figures

1.1	Thesis roadmap linking goals, challenges, and methods.	5
2.1	The Transformer model architecture from [1].	14
3.1	Trajectory order objective-BERT (TOO-BERT) architecture for an example patient trajectory.	24
3.2	Learning the effective representation of multi-source patient trajectories by a two-step Masking MLM and a multi-head transformer encoder	25
3.3	Comparison of the impact of ANAA on self-attention score distributions in fine-tuned models. Attention scores from each head are individually scaled to the [0, 1] range before plotting their distributions.	28
3.4	Qualitative comparison for a held-out patient from the MDC cohort predicted as early death. Top: IG with [MASK]-all baseline; Middle: IG with (manifold-aware) baseline; Bottom: GS-IG (manifold-aware + group sparsity). Each panel shows medical-code attributions over time and reports the number of <i>active</i> codes (non-zero attribution). Here, GS-IG reduces active codes from 35 to 19 (vs. the middle panel), yielding a sparser, more readable list of decisive factors. Red indicates contributions toward early death; blue indicates contributions toward long life.	31
3.5	Patient-level Integrated Gradients (IG) explanation for long-life prediction (2-year window): correctly predicted long life. Bars show token attributions for the longitudinal ICD/ATC trajectory in temporal order (positive vs. negative contributions toward the long-life class). Dashed vertical lines separate visits. The green curve shows the net sum of attributions within each visit.	32

List of Tables

1. Introduction

1.1 Introduction

The digital transformation of healthcare has driven the widespread adoption of electronic health records (EHRs), creating rich, longitudinal data that document patients’ medical histories over time. These records contain heterogeneous and temporally structured information—including diagnosis codes, prescribed medications, laboratory results, vital signs, procedures, demographic attributes, and clinical notes. When linked across encounters for the same individual, these data form an EHR trajectory: a time-ordered sequence of visits in which each visit aggregates the clinical events recorded at that point in time. EHR trajectories capture how diseases emerge and evolve, how treatments are initiated and adjusted, and how risks accumulate or resolve. As such, they have become a valuable source for machine learning, enabling predictive modeling, early risk identification, and data-informed treatment planning [2; 3].

Over the last few years, deep learning has achieved impressive results on EHR-based prediction tasks (e.g., incident disease, readmission, length of stay) [2–4]. Recent models learn directly from full trajectories, using recurrent networks, e.g., LSTM [5], attention mechanisms, and especially Transformer architectures that can capture both local and long-range dependencies via self-attention, enabling rich context aggregation[1]. In parallel, self-supervised learning has emerged as a practical response to label scarcity, where models are first pretrained on unlabeled data via proxy objectives—most commonly masked language modeling (MLM), which masks a subset of tokens and trains the model to reconstruct them from context [6]—and then fine-tuned on downstream task with comparatively few labeled examples, yielding strong performance even in label-scarce settings. *Foundation models* formalize this paradigm as models trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks [7].

In this thesis, we use *patient trajectories* to denote time-ordered sequences of clinical events—such as diagnoses (ICD¹) and medications (ATC²)—recorded across visits. For many clinically important endpoints (e.g., incident disease, ad-

¹<https://www.cdc.gov/nchs/icd/icd-10-cm/index.html>

²<https://www.who.int/tools/atc-ddd-toolkit/atc-classification>

verse events), high-quality labels are scarce and cohort-specific, which limits purely supervised learning. The foundation-model paradigm addresses this by pretraining on large collections of unlabeled EHR trajectories to learn general-purpose backbones that can then be fine-tuned on downstream tasks with limited labeled data, as demonstrated for structured EHR in models such as BEHRT and Med-BERT [8; 9].

Despite this progress, core challenges remain. EHR data are sparse, irregularly sampled, and often incomplete; events are encoded in diverse systems whose semantics are complex and evolving; temporal dependencies span variable intervals and can involve delayed effects; and labels for important outcomes are rare, noisy, and cohort-specific. Privacy and governance constraints further restrict broad data sharing, limiting opportunities for large-scale model development and reproducible evaluation. Moreover, while Transformers are well-suited to model both local (within-visit) and global (across-visit) relations, they are data-hungry: in small- to medium-sized EHR cohorts—typical for many registries—naïve fine-tuning can fail to learn the nuanced disease–intervention dependencies that drive clinically meaningful predictions. These realities motivate representation-learning strategies that (i) respect heterogeneity and temporal structure, (ii) are robust and data-efficient, and (iii) transfer across settings [2].

Beyond technical performance, societal impact and clinical adoption hinge on transparency and trust. Accurate trajectory-based risk models can enable earlier interventions, more equitable resource allocation, and better outcomes. But safe deployment requires that clinicians understand why a prediction was made and how confident the model is. Local (case-level) explanations help clinicians audit and act on individual predictions, while global (population-level) summaries support cohort stratification, guideline development, and policy decisions. Faithful, concise explanations and reliable uncertainty estimates are therefore essential for actionable, equitable, and trustworthy use in practice.

Scope and focus. This thesis lies at the intersection of *predictive modeling* and *model interpretability* for longitudinal, EHR trajectories. It targets foundation-model settings typical of registry data—where access to vast training corpora (as in general-domain NLP) is uncommon—and thus emphasizes *efficient* pretraining and robust adaptation under data constraints. Concretely, the work develops methods to (i) tailor self-supervised pretraining to EHR structure so that models capture cross-source and temporal dependencies, (ii) stabilize fine-tuning in low-label regimes without altering the backbone architecture, and (iii) generate concise, faithful, clinically useful explanations at both the individual and population levels.

Central question.

How can we improve prediction and provide trustworthy explanations for adverse health outcomes by modeling longitudinal EHR trajectories?

To answer this question, the thesis articulates a set of research questions that align with two overarching objectives: (1) improving the modeling of health trajectories through representation learning and robust adaptation, and (2) improving the interpretability, uncertainty quantification, and usability of the resulting models. These directions are pursued through a sequence of studies that advance the state of the art in self-supervised objectives, attention-space augmentations, explanation methods for structured EHR, and aggregation of case-level attributions into cohort-level insights. The next sections summarize the key challenges (Section 1.2) and detail the research questions that guide the contributions (Section 1.3).

1.2 Challenges

Modern machine learning on EHR *trajectories* must operate across messy, multi-source clinical data, respect irregular and long-range temporal structure, learn effectively under privacy and label constraints, and produce outputs clinicians can trust and act upon. The challenges below distill these requirements into concrete design constraints that motivate the research questions and methods in this thesis.

Patient trajectories combine *diverse* sources—structured medical codes (ICD diagnoses, ATC medications, procedure codes), laboratory results, vitals, demographics, medical imaging, signals, and often free text—recorded with non-uniform granularity across institutions. Turning these into a coherent learning signal requires (i) *harmonization* and semantic mapping across coding systems and data models, (ii) *aggregation* within visits and across visits, and (iii) *multimodal* representation learning that exploits cross-source relations instead of naïve concatenation [2; 10].

Accurate prediction depends on capturing both *inner-encounter dependencies* (interactions among codes, labs, and medications *within* a visit) and the *longitudinal flow* of disease and intervention across visits [2; 11; 12].

EHR trajectories are event-driven and therefore *irregularly sampled*: variable time gaps carry signal but complicate alignment, interpolation, and evaluation. Missingness is often structured (not missing at random), so simplistic masking or imputation can bias both training and metrics [2; 12; 13].

Since EHR contains sensitive personal information, broad data sharing is restricted and public trajectory datasets are scarce. Even within institutions, cohort sizes for specific endpoints can be limited, and long-tail code distributions reduce the effective sample size for many concepts. Training modern deep models, which benefit from large-scale pretraining and robust validation, is therefore challenging without careful data governance, augmentation, or proxy-task design [2; 3]. High-quality labels for incident outcomes are costly to curate, imbalanced (rare diseases), and often site-specific. [2; 14].

Shifts in coding practice, case-mix, measurement regimes, and care pathways

degrade out-of-domain performance. Robustness, therefore, hinges on domain adaptation, careful evaluation across institutions, and designs that separate signal from site-specific artifacts[2; 15].

Multi-institutional training and external validation are essential for developing generalizable models, but are hindered by legal, ethical, and technical constraints. Practical deployments must contend with heterogeneity, trust, auditing, and communication efficiency [16–18].

For models to be adopted in clinical settings, they must be reliable and *trustworthy*. Providing information about prediction uncertainty and explaining the key factors behind each decision enables developers, clinicians, and practitioners to better understand, inspect closely, and verify model outputs. Explanations must be both *faithful*—accurately reflecting the model’s internal reasoning—and *actionable*, highlighting the features that genuinely drive predictions in a way that supports clinical decision-making. Reliable uncertainty estimates help clinicians gauge the level of confidence to place in each prediction, thereby reducing the risk of overreliance on uncertain outputs. In addition to case-level explanations, population-level summaries and fairness-aware evaluations are increasingly emphasized [2; 19].

Scientific progress in EHR trajectory modeling is limited by the scarcity of publicly available datasets and the field’s heavy reliance on a few benchmark cohorts—most notably MIMIC—which focus on ICU populations and are specific to U.S. healthcare standards. This narrow focus can lead to models that are overfitted to particular settings and not generalizable to broader or more diverse populations. Recent initiatives such as EHRSHOT [20] and other benchmarking efforts advocate for the development and use of more diverse datasets, richer and clinically relevant prediction tasks, and few-shot evaluation settings that better reflect real-world clinical challenges. Equally important is the transparent sharing of code, pretrained models, and evaluation protocols tailored to the structure of EHR data. These practices are essential for improving reproducibility, enabling fair comparisons across methods, and accelerating meaningful progress in the field [2; 20; 21].

1.3 Research Questions

This thesis is guided by a central aim: to improve predictive modeling and uncover new risk patterns that can help prevent adverse health outcomes, using longitudinal EHR data. To achieve this, the thesis pursues two tightly coupled research directions.

First, it aims to improve the predictive performance of deep learning models on heterogeneous, temporally complex, and data-scarce EHR trajectories. Second, it seeks to increase their transparency and trustworthiness by developing techniques to explain predictions, so that risk estimates are not only accurate but also accompanied by insight into *why* they were made. Fig. 1.1 shows this roadmap.

Accordingly, the research questions addressed in this thesis are as follows:

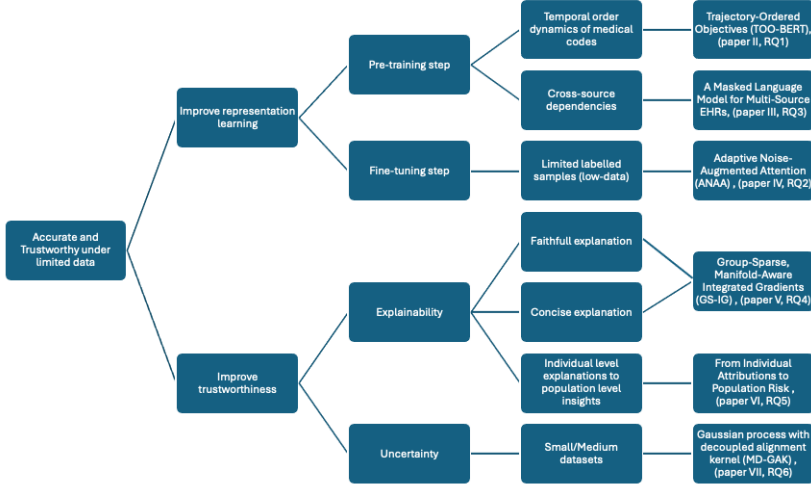


Figure 1.1: Thesis roadmap linking goals, challenges, and methods.

RQ1. *How can temporal order dynamics of medical codes be leveraged to improve the learned representations in foundational EHR models beyond what is captured by standard MLM?*

Standard MLM, given limitations in available EHR sources, captures contextual relationships but often neglects the temporal sequence in which clinical events occur. This question investigates whether explicitly modeling the temporal order of visits and codes enhances a model’s ability to learn meaningful disease–intervention patterns and long-range dependencies.

RQ2. *How can we improve the performance of fine-tuning foundational EHR models, particularly in low-data regimes, without requiring large labeled datasets?*

Transformers pretrained on large-scale, unlabeled EHR can struggle when fine-tuned with limited labeled data. This question seeks augmentation techniques at the fine-tuning stage—especially in attention space—that stabilize training and preserve performance under data scarcity.

RQ3. *How can self-supervised pretraining be adapted to learn cross-source dependencies in multi-source EHR trajectories?*

Many clinical features (e.g., diagnoses, medications, and questionnaire responses) interact across modalities. This question explores how pretraining objectives—particularly masked language modeling (MLM)—can be ex-

tended with source-aware strategies to better capture inter-source relationships and contextual dependencies in patient trajectories.

RQ4. *How can we produce concise and faithful token-level explanations for predictions made by multimodal Transformers on EHR trajectories?*

Local interpretability is essential for clinician trust. This question explores how to develop explanation methods that are sparse (highlighting only the most relevant tokens), faithful (aligned with the model’s true decision process), and grounded in realistic patient representations (i.e., staying on the learned manifold).

RQ5. *How can individual-level explanations be aggregated to generate population-level insights for cohort-wide analysis?*

While case-level explanations support individual decisions, system-level interventions require population summaries. This question addresses methods for aggregating individual-level attributions across cohorts to identify features that are consistently important at the population level.

RQ6. *How can we provide reliable uncertainty quantification for sequence prediction models, especially in small- to medium-sized datasets?*

Uncertainty is critical for safety and model calibration. This question investigates techniques to quantify predictive uncertainty in data-limited settings using molecular sequences.

1.4 Contributions

The main contributions of the thesis can be summarized as follows:

Paper I: Deep learning on EHR trajectories—a systematic review. This paper presents a systematic review of 63 studies (published between 2016 and 2022) that investigated deep learning in EHR trajectories. Crucially, the review identifies the several persistent challenges in modeling EHR trajectories. These include: (i) handling heterogeneous, multi-source data and aligning semantics across structured modalities; (ii) capturing temporal coherence and long-range dependencies within and across clinical encounters; (iii) managing irregular sampling, missing data, and measurement bias; and (iv) addressing data insufficiency, while ensuring trustworthiness and transparency in deep learning models. These findings collectively motivate the research directions pursued in the rest of the thesis.

Paper II: Trajectory-Ordered Objectives (TOO-BERT). This work introduces a novel self-supervised auxiliary objective—called the *trajectory order objective*—to complement masked language modeling (MLM) during pretraining on EHR tra-

jectories. The goal is to enrich temporal representations by explicitly teaching the model to understand the natural ordering of medical codes across visits (RQ1).

The method operates by generating contrastive pairs of sequences: *ordered* sequences that preserve the true temporal flow of events, and *permuted* sequences that are synthetically reordered. These permutations are generated under constraints that preserve clinical plausibility by conditioning on medical code co-occurrence statistics. Specifically, code-swapping is applied to pairs of codes where one frequently follows the other, encouraging the model to learn from meaningful disruptions in temporal order. This auxiliary task improves representation quality during the self-supervised pretraining phase.

Paper III: A Masked Language Model for Multi-Source EHR Trajectories.

This paper develops a two-step masking strategy for learning representations from multi-source EHR data. In the first step, we mask individual codes along a patient’s trajectory randomly. In the second step, we mask codes from only one source (for example, only diagnoses) and train the model to reconstruct them from the remaining sources (e.g., medication codes). This encourages the model to capture the interaction between diseases and treatments, and improves downstream prediction performance (RQ3).

Paper IV: Adaptive Noise-Augmented Attention (ANAA) for robust fine-tuning.

Foundational EHR models, usually pretrained with self-supervised objectives (e.g., MLM) often face label scarcity at fine-tuning time. In such cases, fine-tuned models tend to under-explore the complex dependencies within patient histories, often converging to narrow or oversimplified attention patterns, resulting in polarized distributions (values concentrated near 0 or 1) (RQ2).

We propose Adaptive Noise-Augmented Attention (ANAA), a lightweight augmentation applied during fine-tuning. ANAA operates in two steps on each head’s attention scores (pre-softmax): (i) inject adaptive Gaussian noise whose scale is calibrated to the head/layer statistics, promoting exploration of diverse dependency paths; and (ii) apply a 2D Gaussian smoothing kernel over the score matrix, restoring structural coherence and preventing the noise from devolving into unstructured attention. This simple perturb-then-smooth procedure yields more informative self-attention maps, improves robustness in low-data settings, and enhances downstream predictive performance—while maintaining unchanged inference and minimizing computational overhead.

Paper V: Group-Sparse, Manifold-Aware Integrated Gradients (GS-IG).

Integrated Gradients (IG) is a widely used attribution method for deep networks because it satisfies key axioms such as sensitivity and implementation invariance. However, for categorical, sequential inputs—e.g., medical codes in a patient trajectory or words in a text—the choice of baseline is ambiguous and produces dense, hard-to-interpret attribution maps (RQ4).

We address both issues with Group-Sparse, Manifold-Aware IG (GS-IG). First, we introduce a manifold-aware baseline defined as the mean of the embeddings, ensuring that the IG path originates in a high-density region of the representation space. Second, we reparameterize the IG path with a schedule per instance by optimizing a group-sparsity objective. Together, these design choices keep attribution paths close to the data manifold and transform dense token heatmaps into concise, clinician-readable rationales that better reflect the model’s decision-making process.

Paper VI: From Individual Attributions to Population Risk: Identifying Key Drivers of Longevity and Early Mortality.

Individual-level explanations help case review, but do not directly reveal cohort-level drivers of risk (RQ5). This paper presents a framework that turns Integrated Gradients—computed for longitudinal medical codes (diagnoses/medications) and baseline phenotype/lifestyle variables—into population-level insights for mortality prediction by aggregating attributions using (i) magnitude-based importance and (ii) direction-aware positive/negative importance (toward vs. against the outcome). To reduce sensitivity to modeling and aggregation choices, it also performs a consensus analysis across multiple trained models and aggregation schemes, producing a stable set of cohort-level features whose attributions consistently push toward versus against the outcome.

Paper VII: Gaussian Processes with a Novel Molecular Decoupled Alignment Kernel for peptide membrane permeability.

In small to medium-sized molecular sequence datasets, foundation models often struggle to achieve high accuracy, reliable calibration, or effectively quantify predictive uncertainty (RQ6).

To address this, we develop a kernel-based probabilistic framework using Gaussian Processes (GPs) tailored for peptide sequence modeling. Recognizing that peptides can be represented as sequences of monomers, we design the *Monomer Decoupled Global Alignment Kernel* (MD-GAK), which computes similarity between sequences through alignment costs. Unlike traditional global-alignment kernels, MD-GAK decouples chemical similarity from gap penalties, enabling the model to capture sequence-sensitive chemical relationships more faithfully. We also provide a constructive proof that MD-GAK is a positive semidefinite (PSD) kernel, guaranteeing it defines a valid GP covariance. When used as the covariance function in a GP classifier, MD-GAK produces accurate, well-calibrated predictions and provides trustworthy uncertainty estimates, demonstrating strong performance even in small- to medium-sized datasets.

2. Background

2.1 Foundation models

Foundation models are large models trained on broad, often heterogeneous data—typically via self-supervision—that can be adapted to a wide range of downstream tasks with minimal task-specific supervision [7]. This paradigm unifies earlier pretrain–fine-tune workflows from Natural Language Processing (NLP) and vision and has recently been proposed as a pathway to medical AI, where a single model family supports many clinical tasks with limited labels [22].

Two landmark families—BERT [6] and the GPT line of models [23; 24]—were among the first to show convincingly that **(i)** large-scale *self-supervised pretraining* and **(ii)** the *Transformer* architecture [1] together form a powerful recipe for building general-purpose models.

BERT (Bidirectional Encoder Representations from Transformers) is a stack of transformer encoder layers pretrains a bidirectional stack of Transformer *encoder* layers on two proxy tasks: (i) masked language modeling (MLM), where a subset of tokens is masked and reconstructed from context, and (ii) next sentence prediction (NSP), which encourages modeling of inter-sentence relations [6]. After this self-supervised stage on large unlabeled corpora (BooksCorpus and English Wikipedia), BERT is adapted to a target task by simply adding a shallow task head (a shallow MLP) and fine-tuning all parameters for a few epochs (pretraining and fine-tuning). This minimal adaptation step delivered state-of-the-art performance on GLUE [25], SQuAD [26], and several other benchmarks, mostly because (a) MLM produces rich, transferable semantic representations; (b) bidirectional self-attention captures both left and right context in a single pass; and (c) the Transformer encoder makes it easy to reuse the same backbone across tasks with very little labeled data [27].

GPT took the complementary route: it used a causal (left-to-right) Transformer *decoder* trained with a single, simple objective—next-token prediction—on a large, diverse WebText corpus [28]. Instead of task-specific fine-tuning, GPT-2 demonstrated strong *zero-shot* and *few-shot* performance by prompting: the same pretrained model could perform summarization, translation, and question answering when the task was described in the input. GPT-3 scaled the very same idea—same causal objective, same decoder-style Transformer—to 175B parameters, trained on a much larger and more heterogeneous corpus, and extended the context window

from 1,024 to 2,048 tokens [24]. At that scale, GPT-3 showed that in-context learning (zero/one/few-shot) can approach the performance of task-specific fine-tuning on many benchmarks, without updating model weights. Its success rested on three ingredients: (i) a *simple, scalable* self-supervised objective (next-token prediction); (ii) *broad* pretraining data; and (iii) *sufficient model capacity* to internalize general linguistic and world knowledge.

Self-supervised learning paradigm lets models exploit large, unlabeled corpora and thus alleviates label scarcity, and *Transformers* (encoder-, decoder-, or encoder–decoder-style) provide a flexible sequence model that can learn both short- and long-range dependencies and be reused across tasks with very light adaptation.

2.2 Self-supervised learning

Supervised learning assumes access to input–output pairs $\{(x_i, y_i)\}$ and learns a function $f_\theta : x \mapsto y$ by minimizing a task loss (e.g., cross-entropy). Its performance and generalization are therefore tightly linked to the size and quality of the labeled set [29–31]. In EHR, high-quality labels (e.g., incident disease, mortality, complications) are costly, rare, and often site-specific, which makes purely supervised training fragile. Self-supervised learning (SSL) addresses this by defining proxy tasks on unlabeled data, pretraining a model on these tasks, and then fine-tuning it on the small labeled dataset.

2.2.1 Pretraining objectives

Most EHR foundation–inspired models follow the NLP recipe: a primary generative objective that learns broad contextual structure, plus one or more auxiliary objectives to improve the learned representation.

Generative/reconstruction objectives. The primary objective for structured EHR is masked language modeling (MLM): mask a subset of medical tokens (ICD, ATC, procedure) and predict them from context, learning bidirectional dependencies across codes and visits in a context. MLM objectives have found widespread application in EHR trajectory prediction tasks, largely owing to the capabilities of BERT models to learn the context [8; 9; 11; 32–35]. A complementary family uses an *autoregressive* objective [24] to predict upcoming medical events (e.g., codes for the next day or visit) from the patient’s history, aligning the pretext task with the temporal direction of care [36–38].

Knowledge-injection objectives. To explicitly couple diagnoses and interventions, source-conditioned prediction heads can be added—predicting medications

from diagnoses and vice versa—to force the representation to encode disease–treatment relations [35; 39]. Other works inject external domain knowledge via auxiliary heads, such as length-of-stay prediction (patient acuity) or visit type classification, which helps mitigate sparsity and improves learned representation [9; 32; 38]. Ada-Diag [40] added a domain classifier to distinguish data from different institutes and enhance the generalizability and robustness of the learned representation against dataset shifts.

Contrastive / discrimination objectives. Contrastive SSL encourages agreement between different “views” of the same trajectory while separating different patients. Hi-BEHT employs a BYOL-style [41] agreement loss over augmented EHR sequences [33]. RAPT trains the encoder to discriminate patient trajectories and detect mixed (stitched) sequences, strengthening identity and temporal coherence [42]. GRACE augments MLM with a real-versus-GAN-generated contrastive signal to cope with data insufficiency [43]. Plain MLM can under-encode global temporal structure. To address this, trajectory-ordered objectives require the model to distinguish correctly ordered visit/code sequences from clinically plausible permutations, directly teaching temporal coherence beyond local context [11]. This echoes the motivation behind sentence-order prediction in NLP [44].

2.2.2 Adapting models to downstream tasks

Once a foundation model has been pretrained and acquired general knowledge, there are several strategies to adapt it to a new prediction task that differ in how many parameters are updated, how much storage/compute they require, and how well they preserve the general knowledge encoded during pretraining.

Full model fine-tuning. The classical approach updates all parameters of the pretrained network on the target task while adding a small task-specific head (e.g., a classifier). Popularized in language modeling by ULMFiT and BERT, end-to-end fine-tuning delivers strong transfer with minimal architectural change [6; 45]. In practice, effective tuning typically relies on small learning rates with warmup and linear decay [6], layer-wise learning-rate decay or discriminative rates so early layers move less than higher layers [45], gradual unfreezing [45], proximity regularization around the pretrained point (e.g., AdamW) [46], and standard stabilization (dropout, gradient clipping, early stopping) to mitigate optimization fragility [47]. While it often maximizes single-task performance, full fine-tuning is parameter- and storage-intensive (one full copy per task) and can be susceptible to catastrophic forgetting in sequential training [48].

Parameter-efficient fine-tuning (PEFT). PEFT updates only a small fraction of parameters while keeping the backbone frozen. Representative families include: Adapters (small bottleneck modules inserted between Transformer layers) that add only a few percent task-specific parameters yet approach full fine-tuning accuracy [49; 50]. LoRA (low-rank adaptation), which injects trainable low-rank matrices into attention/Feed Forward Network (FFN) projections; it can reduce trainable parameters by orders of magnitude and adds negligible inference latency [51]. BitFit, which tunes only bias terms and is competitive in small/medium data [52]. $(IA)^3$, which learns per-vector multiplicative scales on key/value/FFN activations and is effective in few-shot regimes [53]. PEFT is attractive when many tasks must be served (one small adapter per task), when storage/latency are constrained, or when we wish to avoid overwriting pretrained knowledge.

Linear probing (feature reuse). Here the backbone is frozen and only a linear classifier is trained on top of the final representations. Linear probes were introduced to assess representational quality in deep networks [54] and are widely used as a low-cost adaptation/baseline (e.g., “linear evaluation protocol” in self-supervised vision) [55]. Probing is fast, avoids overfitting in low-label regimes, and reveals how separable the task is in the pretrained feature space, though it usually underperforms parameter-updating methods when substantial task-specific adaptation is needed.

In-context learning (prompting). Large decoder-only models can perform new tasks from instructions and a few demonstrations *without any weight updates*—the model conditions on a prompt that contains task description and k exemplars (zero/one/few-shot) [24]. This is appealing when labeled data are scarce, rapid iteration is needed, or model weights cannot be modified. Prompt-based *PEFT* variants bridge prompting and fine-tuning: *prefix/prompt-tuning* learn small continuous “soft prompts” while keeping the backbone frozen, often matching full fine-tuning in low-data settings and improving with scale [56].

Choosing an adaptation strategy. In practice, the choice depends on data, compute, and how far the downstream task departs from the pretraining signal. Full fine-tuning maximizes single-task targets when labels and compute permit, because all layers can specialize to the new endpoint [6; 45]. Parameter-efficient methods (adapters, LoRA, BitFit, $(IA)^3$) retain most of the backbone while updating a small set of weights, offering an excellent performance–efficiency trade-off for multi-task deployment or tight memory budgets [49; 51–53]. Linear probing is a fast, low-variance baseline and a diagnostic of representation quality, though it usually demonstrates a weaker performance [54; 55]. Prompting/in-context learning adapts

decoder LMs without requiring weight updates and is particularly attractive when labels are scarce and rapid iteration is needed [24].

Most structured-EHR studies—including our work—use end-to-end fine-tuning because: (i) outcome tasks (e.g., incident of Heart Failure, Alzheimer Disease, Early Death) and cohort/coding shifts differ substantially from the pretext objectives, so updating all layers improves discrimination and calibration; (ii) encoder-only Transformers over categorical codes are modest in size, making full fine-tuning feasible in clinical settings; Accordingly, fine-tuning remains the default for EHR outcome models [8; 9; 11; 32; 34; 35; 57].

2.3 Transformer and self-attention

Overview. Transformers model a sequence $x_{1:n}$ by repeatedly applying self-attention and position-wise feed-forward layers with residual connections and normalization (fig 2.1) [1]. Let $X \in \mathbb{R}^{n \times d_{\text{model}}}$ denote the token embeddings for a length- n sequence. A Transformer block computes

$$Y_1 = \text{LN}(X + \text{MHA}(X)), \quad Y_2 = \text{LN}(Y_1 + \text{FFN}(Y_1)),$$

where LN is LayerNorm, MHA is multi-head self-attention, and FFN is a two-layer MLP applied identically at each position.

Tokens and embedding layer. Transformers operate on sequences of discrete *tokens* (e.g., words in text, or medical codes in EHR). Each token is represented by an integer index in a vocabulary \mathcal{V} , and an *embedding layer* maps these indices to continuous vectors. Concretely, a learned embedding matrix $E \in \mathbb{R}^{|\mathcal{V}| \times d_{\text{model}}}$ stores one d_{model} -dimensional vector per token; given a sequence of token indices (t_1, \dots, t_n) , the corresponding input matrix $X \in \mathbb{R}^{n \times d_{\text{model}}}$ is formed by looking up the rows E_{t_1}, \dots, E_{t_n} and (optionally) adding type, visit, or time embeddings. In language models, tokens are typically subwords or words [1], whereas in EHR models such as BEHRT and Med-BERT they correspond to clinical codes (e.g., ICD diagnoses, ATC medications) and related event types [8; 9]. These token embeddings are then combined with positional encodings (below) and passed through the stack of Transformer layers.

Scaled dot-product attention. For one head h with per-head width d_k , queries, keys, and values are linear projections

$$Q = XW_Q^{(h)}, \quad K = XW_K^{(h)}, \quad V = XW_V^{(h)}, \quad W_Q^{(h)}, W_K^{(h)}, W_V^{(h)} \in \mathbb{R}^{d_{\text{model}} \times d_k}.$$

Multiple heads allow the model to attend to heterogeneous patterns (e.g., short-range co-occurrence and long-range dependencies) in parallel subspaces. Dropout is commonly applied to A or to the output projection Y_1 or Y_2 to regularize training.

Position-wise feed-forward network. The MLP applies the same transformation at each position:

$$\text{FFN}(x) = W_2 \sigma(W_1 x + b_1) + b_2, \quad W_1 \in \mathbb{R}^{d_{\text{model}} \times d_{\text{ff}}}, W_2 \in \mathbb{R}^{d_{\text{ff}} \times d_{\text{model}}},$$

with a nonlinearity σ (ReLU or GELU). Residual connections and layer normalization (pre-LN or post-LN variants) improve optimization stability [58].

Positional information. Because self-attention is permutation-invariant over positions, Transformers add position encodings $P \in \mathbb{R}^{n \times d_{\text{model}}}$ to X (absolute sinusoidal encodings, learned embeddings, or relative position biases). A common absolute scheme uses

$$\text{PE}(\text{pos}, 2i) = \sin\left(\frac{\text{pos}}{10000^{2i/d_{\text{model}}}}\right), \quad \text{PE}(\text{pos}, 2i+1) = \cos\left(\frac{\text{pos}}{10000^{2i/d_{\text{model}}}}\right),$$

while relative variants add content- and distance-dependent terms directly to S .

Masks (causal, padding, structure). Masks are injected additively in S via M :

$$M_{ij} = \begin{cases} 0, & \text{if } j \text{ is visible to } i, \\ -\infty, & \text{otherwise,} \end{cases}$$

so that $\exp(S_{ij} + M_{ij}) = 0$ for disallowed positions. *Padding masks* prevent attending to pad tokens. *Causal masks* enforce $j \leq i$ for auto regressive decoding.

Complexity and long-sequence variants. Self-attention forms QK^\top with $O(n^2 d_k)$ time and $O(n^2)$ memory, which can be limiting for long trajectories. Approaches such as sparse attention (Longformer [59], BigBird [60]), low-rank projections (Linformer [61]), or kernelized attention (Performer) [62] reduce cost while approximating full attention.

In trajectory modeling, tokens represent clinical events (ICD diagnoses, ATC medications, procedures, labs). Self-attention mixes information within and across visits, enabling the model to learn short-range co-occurrence structure (within-visit code interactions) and long-range disease–intervention dynamics across visits.

2.4 Explainability in foundation models

Why explanations matter in EHR modeling. Foundation models promise strong transfer and data efficiency, but clinical deployment requires that predictions be *understandable*, *auditable*, and *trustworthy*. In healthcare, explanations support model debugging, bias assessment, and ultimately clinician acceptance; Recent works consistently emphasize that explanation goals (safety, accountability, fairness) should shape what is explained and how it is evaluated [19].

What counts as a good explanation. Two key criteria recur in the NLP/health literature: faithfulness (does the explanation reflect the model’s actual reasoning process) and usefulness/plausibility (is it clinically meaningful) [63]. Perturbation-based metrics such as comprehensiveness and sufficiency operationalize faithfulness by measuring what happens when highlighted tokens are removed/retained [64–66]. However, explanation metrics can be gamed or misaligned with human intuition, so careful validation is essential [67].

Attention as explanation. Because BERT-style models produce attention weights, a natural idea is to treat attention as importance. Yet multiple studies show that raw attention often correlates weakly with causal influence: one can perturb attention without changing predictions, and gradient-based importance can diverge from attention magnitudes [68; 69].

Gradient-based attribution. For bidirectional encoders, gradient backpropagation provides token-level attributions. Integrated Gradients (IG) [70] attributes a trained neural network ($F(x)$), to input features by integrating the gradient along a path from a baseline x' to the input:

$$\text{IG}_i(x; x') = (x_i - x'_i) \int_0^1 \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha,$$

and is justified by axioms of sensitivity and implementation invariance [70] and passed model-parameter randomization sanity check [71]. Two design choices are critical when applying IG to embedding-based models (e.g., text or EHR codes). Baseline selection: straight-line IG presumes an input representing “absence of evidence.” While a black image (or zero vector) is sensible for pixels, categorical sequences lack a natural null; heuristics such as using a special token (e.g., [MASK]) or explaining one token at a time (Sequential IG) keep interpolants syntactic and can push paths off the empirical data distribution, affecting attribution quality [66; 72]. Path choice: linear interpolation in embedding space may traverse low-density

regions, amplifying gradient noise and reducing faithfulness; moreover, gradient-based methods usually return dense saliency scores that are difficult to read and even harder to act upon [72; 73].

Perturbation-based explanations. These methods estimate feature importance by directly altering the input and measuring the change in the model’s output. In NLP, common variants include word/token erasure or leave-one-out, where a token is removed or masked and the resulting confidence drop is measured [74]. Another example is *iterative input reduction*, which repeatedly deletes the least-salient tokens until the prediction flips; this often reveals overconfident, nonsensical rationales [75]. A third family are local surrogate approaches like LIME and SHAP, which approximate feature contributions by fitting simple models to many randomized perturbations of the input [76; 77]. In structured EHR Transformers, the same idea is implemented by masking ICD/ATC codes, entire visits, or modality blocks and measuring the change in predicted risk [78].

Pros. (i) These methods are model-agnostic. (ii) Directly tied to causal influence on the model’s output and supports natural “what-if” questions (e.g., remove a diagnosis/medication and measure the risk drop). (iii) Comes with practical fidelity checks such as deletions/insertions to validate that highlighted features truly matter [79]. (iv) Within the perturbation family, Shapley-based SHAP offers axiomatic guarantees (completeness, consistency, missingness) when its assumptions hold [77]. **Cons.** (i) Most simple perturbation rules (token/feature erasure, occlusion, input-reduction, LIME/RISE-style masking) lack formal attribution axioms. (ii) Naive perturbations can produce out-of-distribution inputs in language and EHR sequences (e.g., breaking grammar or clinical plausibility). (iii) Replacement/baseline bias: importance can depend strongly on how features are “removed” (e.g., [MASK], zeroed embeddings, or synthetic codes), and different baselines yield different scores. (iv) Non-additivity: interactions among codes/visits mean single-feature deletions can understate synergistic effects; attribution sums need not equal the model’s output change. (v) *Computational cost and instability*: reliable perturbation estimates require many forward passes. Input reduction can also expose overconfident but unfaithful rationales; therefore, these scores are best interpreted together with fidelity checks such as comprehensiveness and sufficiency. [64; 75].

2.5 Uncertainty estimation in foundation models

Why uncertainty matters. For clinical deployment, a model should not only be accurate but also well-calibrated and able to express when it is unsure. It is useful to distinguish epistemic (model) uncertainty, which stems from limited data or misspecified parameters and can shrink with more data, from aleatoric

(data) noise that is irreducible, and distributional (shift) uncertainty under out-of-distribution (OOD) inputs [80]. In practice, we monitor predictive uncertainty and its calibration via reliability diagrams and Expected Calibration Error (ECE) or Brier score [81]. Below, we summarize the main families used with Transformer-based NLP foundation models and with structured-EHR Transformers (e.g., BEHRT/Med-BERT-style encoders).

Sampling (approximate Bayesian) methods. *MC Dropout* keeps dropout active at test time and averages many stochastic forward passes[82]. It is simple to add to the trained model in fine-tuning loops and is widely reported in healthcare as a practical baseline for uncertainty [83]. *Deep ensembles* train K independently-seeded models and average their outputs; they yield strong accuracy and calibrated uncertainties and are robust under shift, at a considerable cost [84]. **Pros:** black-box, minimal code changes (MC dropout), strong OOD detection and calibration for ensembles. **Cons:** multiple forward passes (MC dropout) or multiple models (ensembles) raise latency/cost; variance may still be miscalibrated under strong correlations or heavy label noise.

Distribution-aware single-model variants. Gaussian processes (GPs) provide a classical Bayesian framework where predictions come with a full predictive distribution: a mean (point prediction) and a variance that directly quantifies uncertainty. The quality of this approach, however, hinges on the choice of kernel, which encodes the similarity structure of the inputs. For structured inputs such as molecular or sequence data, well-designed kernels are crucial [85]. When an appropriate kernel matches the data’s inductive biases, uncertainty can be both principled and reliable—but designing such kernels for high-dimensional, heterogeneous inputs (images, text, EHR trajectories) is challenging.

To bridge deep representations and GP-style uncertainty, several *deep kernel* and distance-aware architectures combine neural feature extractors with probabilistic output layers. SNGP (spectral-normalized neural Gaussian Process) adds spectral normalization to the encoder to enforce approximate Lipschitz continuity and replaces the final linear layer with a GP. This yields a single deterministic network that is more distance-aware in feature space and improves calibration and OOD detection on modern architectures such as BERT.[86] DUE (Deterministic Uncertainty Estimation) follows a related deep-kernel idea: it couples a bi-Lipschitz feature extractor with an inducing-point GP head, producing calibrated, distance-aware uncertainty with a single forward pass [87]. Evidential deep learning instead predicts evidence parameters of a prior (e.g., a Dirichlet) so that low evidence corresponds to high uncertainty; it yields rich uncertainty structure with a single deterministic network but requires bespoke losses, priors, and regularization.[88]

Pros: single forward pass; better OOD awareness (SNGP, PriorNets); no sampling at test time (evidential). **Cons:** extra architectural pieces and hyperparameters; dependence assumptions (e.g., GP feature maps) and training stability matter; may still need post-hoc calibration.

Post-hoc calibration (confidence, not uncertainty decomposition). *Temperature scaling* rescales logits with a single scalar fit on a validation set; it reliably reduces ECE without changing predictions and is common for LLMs and EHR risk models [81; 89]. *Beta calibration* improves logistic/Platt scaling in binary settings by modeling score distributions with a beta family [90]; **Pros:** trivial to apply; cheap; preserves accuracy. **Cons:** no decomposition; can fail under shift; needs a clean validation set.

Conformal prediction (simple, coverage-guaranteed outputs). Conformal prediction (CP) is a wrapper around any trained model that turns point predictions into prediction sets (or intervals) with a guaranteed error rate. Given a held-out calibration set and a user-chosen error level α (e.g., 0.1), CP chooses a threshold so that, on the calibration data, the true label lies inside the set at least $1 - \alpha$ of the time. Under a mild assumption called exchangeability (data points are i.i.d. or, more generally, order-invariant), this coverage guarantee transfers to new data [91]. In clinical time-series/EHR deployment, CP is often used to abstain: if the set is large (or empty), the system says “indeterminate” instead of issuing an over-confident alert. **Pros:** explicit, finite-sample coverage guarantees; model-agnostic wrapper; naturally supports abstention and risk-tiering.

Cons: outputs are sets (not calibrated probabilities); there is a trade-off between set size and coverage; standard guarantees assume exchangeability and can degrade under covariate shift.

3. Summary of Included Papers

3.1 PAPER I: Deep learning prediction models based on EHR trajectories

3.1.1 Aim and Scope

This paper presents a systematic review of Deep Learning (DL) methods for predicting patient outcomes from longitudinal electronic health record (EHR) trajectories, with the goal of identifying challenges, knowledge gaps, and active research directions in reliability, data efficiency, and explainability.

3.1.2 Search Strategy and Inclusion

Titles, abstracts, and keywords were searched across Scopus, PubMed, IEEE Xplore, and the ACM Digital Library (January 2016–April 2022) using terms centered on EHR trajectories, DL, and disease prediction. After removing duplications and screening against three eligibility criteria (use of DL; longitudinal/trajectory aspect; prediction of patient health outcomes), 63 studies were included.

The most frequent targets were “all diagnoses at the next visit” and cardiovascular outcomes; diabetes and kidney disease were also common. Mortality and readmission were the dominant non-disease tasks.

3.1.3 Methods Landscape

We synthesize five recurring stages in EHR–DL pipelines and the main solutions reported:

1. **Preprocessing.** Standardization (e.g., FHIR/OpenEHR), bucketing visits, handling missingness (e.g., Bi-GANs, dropout), and reducing code granularity are widely used to address heterogeneity, sparsity, and irregular sampling.
2. **Data aggregation.** The dominant pattern is to (i) reduce the granularity of medical codes or convert codes to their text names, then merge them with features extracted from clinical notes; (ii) represent codes/text as multi-hot or

one-hot sequences (treating a visit as a “sentence”); and (iii) aggregate structured signals (labs, vitals, demographics) directly and include time-between-visits as an explicit feature. Most papers either concatenated all inputs into a single embedding layer or used modality-specific embeddings followed by late fusion/projection into a shared subspace; recent work also trains cross-modal alignment models. Feature extraction from text typically used NER, topic models (LDA/NMF), or pre-trained embeddings. Beyond that fusion, some studies encode EHRs as graphs (e.g., diagnosis–medication–test dependencies, patient–diagnosis, and diagnosis–patient–lab), enabling inner- and cross-modality relations to be leveraged in downstream models.

3. **Representation learning.** Two concrete strategies dominate. *Fully supervised, end-to-end* models learn representations during task training (often relying on co-occurrence), and were frequently instantiated with RNN variants for temporal coherence (bi-GRU, LSTM, T-LSTM), and CNNs/GNNs to capture intra-visit structure. *Pre-train & fine-tune* methods build representations via self-/unsupervised tasks (e.g., Skip-gram, NMF/autoencoders/VAEs, GNNs, BERT-inspired transformers) before downstream adaptation; these better address data insufficiency.
4. **Data insufficiency.** Two main Approaches were: (i) *Transfer learning*—multi-task setups and weight transfer across similar datasets/tasks—to initialize from richer sources. (ii) *External medical knowledge*—injecting ontologies/graphs (ICD/ATC trees, knowledge graphs), posterior regularization, or learning with privileged information—to guide representation learning and upweight rare conditions (e.g., ancestors in ICD). Synthetic data is noted as a complementary option.
5. **Explainability.** Three families are used, each with limits. (i) *Time-attention* highlights which visits drive predictions but are sensitive to preprocessing/visit embeddings and cannot localize intra-visit drivers. (ii) *Self-attention visualizations* reveal interactions among diseases/interventions across visits but do not directly attribute outcomes. (iii) *Embedding-space plots* show clustering/representation quality rather than direct contribution.

3.1.4 Trends and Gaps

Recent work increasingly adopts self-supervised and language-model-inspired methods, multi-modal learning, and GNNs for both performance and transparency. At the same time, benchmarking remains limited by the scarcity of public trajectory datasets (with a heavy reliance on ICU-focused MIMIC), and few models jointly

address heterogeneity, temporal irregularity, intra-visit structure, label scarcity, and interpretability within a single framework.

3.2 PAPER II: Trajectory-Ordered Objectives (TOO-BERT) for Self-Supervised Representation Learning of Temporal Healthcare Data

3.2.1 Aim and Contribution

This paper introduces *TOO-BERT*, a transformer-based pretraining framework that augments masked language modeling (MLM) with a *trajectory-order objective* (TOO) designed to explicitly encode temporal coherence in electronic health record (EHR) trajectories. TOO-BERT is pretrained to discriminate whether a sequence is in its true temporal order or a permutation, at both the code and visit levels. This objective complements MLM by injecting explicit temporal structure and encouraging the encoder to model order-sensitive dependencies.

3.2.2 Method Overview

Objectives. We pretrain a transformer with two complementary self-supervised objectives: MLM loss to learn context, and a contrastive trajectory–order objective (TOO) to learn order-sensitive dependencies (Fig. 3.1).

We apply permutations at two granularities (i) code swapping within a patient across visits, and (ii) visit swapping that exchanges entire visits.

Conditional permutations. To bias learning toward more plausible, asymmetric transitions, we introduce conditional variants that prioritize frequently observed temporal relations. For codes c_i, c_j observed in different visits, the *conditional code-swapping* score $\text{CCS}(c_i, c_j)$ upweights pairs where c_i tends to *follow* c_j :

$$\text{CCS}(c_i, c_j) \propto \max(\text{CCnt}(c_i, c_j) - \text{CCnt}(c_j, c_i), 0) + \varepsilon,$$

where $\text{CCnt}(c_i, c_j)$ counts occurrences with c_i *after* c_j across the corpus, and ε ensures exploration. Pairs with larger CCS are sampled more often during swapping. For visits v_x, v_y , *conditional visit-swapping* aggregates code-level relations:

$$\text{CVS}(v_x, v_y) = \sum_{c_i \in v_x} \sum_{c_j \in v_y} \text{CCS}(c_i, c_j),$$

so visit pairs with higher CVS are preferentially swapped.

A compact transformer encoder receives token embeddings augmented with visit-index embeddings. Two classifier heads are used during pretraining. For downstream tasks, a Bi-GRU classifier is stacked on top of the pre-trained encoder.

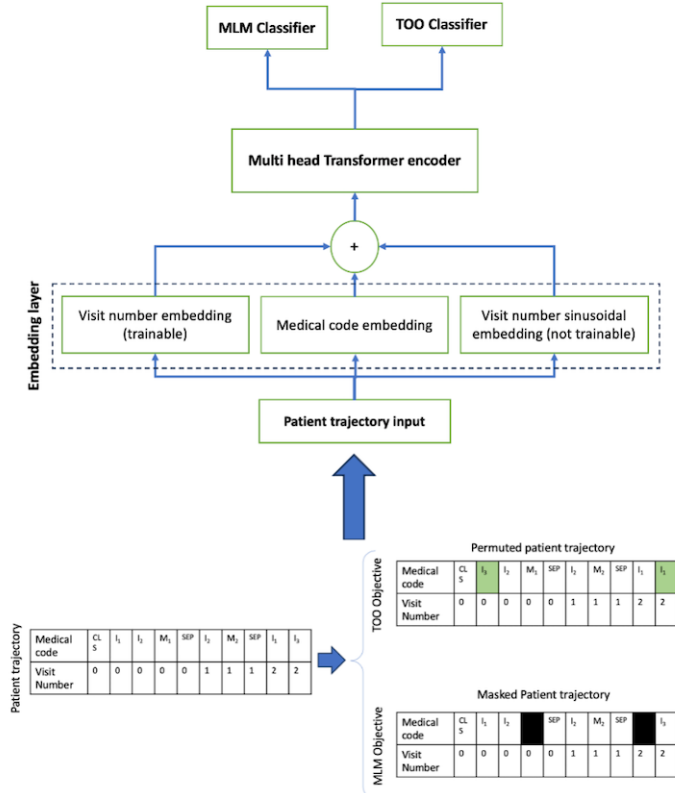


Figure 3.1: Trajectory order objective-BERT (TOO-BERT) architecture for an example patient trajectory.

Data and Tasks (i) MIMIC-IV (hosp): ~173 k patients, and ~10.6 M medical codes; (ii) Malmö Diet and Cancer (MDC): ~30 k individuals, and ~7.6 M medical codes. Downstream tasks included prediction of Heart Failure (HF) incidence, Alzheimer’s Disease (AD) incidence (MDC), and prolonged length of stay (PLS).

3.2.3 Key Results

Downstream performance. TOO-BERT improves over strong baselines across datasets and tasks. On MDC (longer trajectories), *visit-level* ordering was most beneficial: MLM + TOO_{CVS} achieved AUCs of **0.739** (HF) and **0.719** (AD), outperforming MLM-only (HF 0.677; AD 0.695). On MIMIC, *code-level* ordering yielded the best HF result: MLM + TOO_{CCS} reached AUC **0.898** vs. 0.862 (MLM-only). PLS prediction remained challenging.

Attention analysis. Visualization showed that MLM-only models concentrated attention near the most recent visits, whereas TOO -BERT exhibited richer patterns.

3.3 PAPER III: A Masked language model for multi-source EHR trajectories

3.3.1 Aim and Contribution

The paper proposes a self-supervised representation learning scheme for longitudinal, *multi-source* EHR (diagnoses, medications), targeting two core challenges: (i) capturing long/short-term temporal dependencies and (ii) modeling interactions across sources (e.g., how medications relate to diagnoses) to improve downstream clinical prediction.

3.3.2 Method overview

The approach pretrains a transformer encoder with a two-step masking strategy (Fig.3.2). First, a standard MLM step randomly masks a proportion of tokens along a patient trajectory to learn contextual dependencies. Second, a cross-source masking step masks some medical codes only from one source (e.g., ICD10) and trains the model to reconstruct it from the remaining sources (e.g., ATC). It encourages the encoder to learn inter-source relationships. After pretraining, the model is fine-tuned to predict the incidence of Heart Failure (HF) in the Malmö Diet and Cancer (MDC) cohort. The learned representations are evaluated by fine-tuning to predict heart failure (HF) at the next visit.

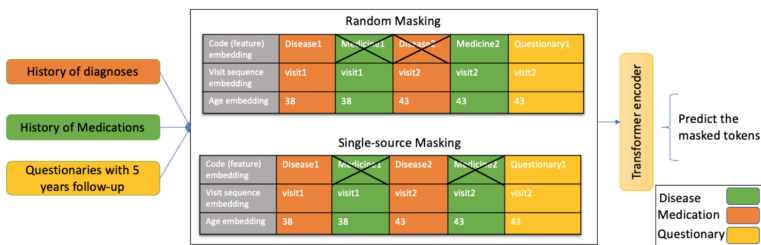


Figure 1. Learning the effective representation of multi-source patient trajectories by a two-step Masking MLM and a multi-head transformer encoder

Figure 3.2: Learning the effective representation of multi-source patient trajectories by a two-step Masking MLM and a multi-head transformer encoder

3.3.3 Key Results

The two-step masked LM achieves the best performance AUC 0.919 (± 0.006), surpassing the simple MLM+Bi-GRU baseline (0.909 ± 0.002)

3.4 PAPER IV: Adaptive noise-augmented attention (ANAA) for enhancing Transformer fine-tuning

3.4.1 Aim and Contribution

This paper proposes **Adaptive Noise-Augmented Attention (ANAA)**, a fine-tuning augmentation technique for pretrained Transformer encoders on longitudinal EHR. ANAA perturbs self-attention scores with adaptive Gaussian noise and then smooths them by a 2D Gaussian kernel, encouraging exploration of alternative dependencies while preserving coherent structure. Across multiple clinical prediction tasks (HF, AD, PLS), ANAA yields consistent AUC gains, particularly under label scarcity, and reshapes attention from near-binary weights to richer, more informative distributions.

3.4.2 Method Overview

Let $A^h = \text{softmax}(Q^h K^{h\top} / \sqrt{d_k})$ be the attention matrix for head h . ANAA forms an augmented attention

$$\tilde{A}^h = (A^h + \mathcal{N}(\mu, \sigma_{GN}^2)) * G_{\sigma_{eh}},$$

where μ and σ_{GN} are the mean and standard deviation of A^h (computed per head), $*$ denotes 2D convolution, and $G_{\sigma_{eh}}$ is a Gaussian kernel with bandwidth (“event horizon”) σ_{eh} . The resulting self-attention output is $\tilde{H}^h = \tilde{A}^h V^h$, and multi-head outputs are concatenated as usual. During inference, stochasticity is removed by replacing noise with its expectation μ :

$$\tilde{A}_{\text{test}}^h = (A^h + \mu) * G_{\sigma_{eh}}.$$

Algorithm 1 details the fine-tuning loop.

3.4.3 Key Results

ANAA improves pretrained Transformers across cohorts and tasks: on MDC, HF from 72.2 ± 2.5 to 74.5 ± 2.9 , AD from 72.2 ± 1.1 to 73.2 ± 0.3 on MIMIC, HF from 85.2 ± 1.1 to **87.2 ± 0.4** .

Algorithm 1 Fine-tuning Transformer Encoder with ANAA

Input: $D_{\text{fine-tuning}} = \{(X_i, y_i)\}_1^N$ tokenized dataset, embedding layer $\text{emb}(\cdot)$, attention score matrix A_h , normal noise $\mathcal{N}(\mu, \sigma_{\text{GN}}^2)$, two-dimensional Gaussian noise $n_{\sigma_{\text{eh}}}$, rest of the model $f(\cdot)$

Parameter: Normal noise $\mu, \sigma_{\text{GN}}^2$ calculated from A_h , event horizon hyperparameter σ_{eh} (adjust based on data characteristics)

- 1: Initialize θ from a pre-trained model
- 2: **repeat**
- 3: Sample $(X_i, y_i) \sim D_{\text{fine-tuning}}$
- 4: $X_{\text{emb}} \leftarrow \text{emb}(X_i)$
- 5: **for** each Attention Head A_h in Transformer Block **do**
- 6: $A_h(X_{\text{attn}}) \leftarrow A_h(X_{\text{emb}}) + \mathcal{N}(\mu, \sigma_{\text{GN}}^2)$
- 7: $A_h(X_{\text{attn}}) \leftarrow \text{Convolve}(A_h(X_{\text{attn}}), n_{\sigma_{\text{eh}}})$
- 8: $H_h(X_{\text{attn}}) \leftarrow A_h(X_{\text{attn}})V$
- 9: **end for**
- 10: MultiHead(H) $\leftarrow \text{concat}(H_0(X_{\text{attn}}), \dots, H_h(X_{\text{attn}}))$
- 11: $\hat{y}_i \leftarrow f(\text{MultiHead}(H))$
- 12: $\theta \leftarrow \text{opt}(\theta, \text{loss}(\hat{y}_i, y_i))$
- 13: **until** Stopping criteria met or maximum iterations reached

Few-shot robustness. With only 50%, 20%, or 10% of fine-tuning data, **HF** prediction gains persist (about 3% points on MIMIC across all sizes); on MDC, gains hold to 50% and diminish at very small sizes due to few positives.

Where to perturb? Against embedding/feed-forward layers perturbations (NEF-Tune; HyPe-style), ANAA achieves the best AUC on both MDC and MIMIC, suggesting that operating in attention space is more effective for EHR sequences.

Attention Behavior Histograms over heads (Fig. 3.3) show vanilla fine-tuning produces near-binary attention (mass near 0 or 1). Noise injected Attention (RNA) broadens the distribution; ANAA broadens and stabilizes it, yielding smoother, overlapping attention patterns. ANAA first perturbs each head with variance-scaled Gaussian noise (broadening discrete spikes into continuous modes), then applies Gaussian convolution as a data-adaptive low-pass filter that suppresses high-frequency artifacts and links neighboring tokens. Analytically, this acts like a *structured, variance-scaled drop-connect* regularizer directly on attention scores, unlike DroAttention regularizer, driving exploration without breaking essential dependencies.

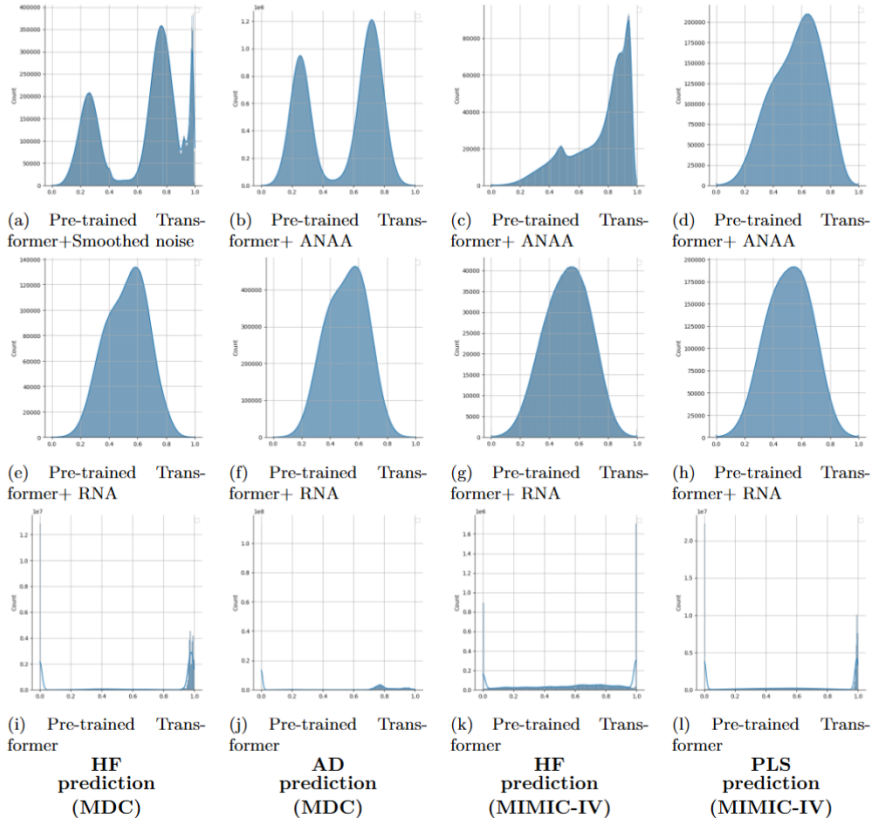


Figure 3.3: Comparison of the impact of ANAA on self-attention score distributions in fine-tuned models. Attention scores from each head are individually scaled to the $[0, 1]$ range before plotting their distributions.

3.5 PAPER V: Group-Sparse Manifold-Aware Integrated Gradients

3.5.1 Aim and Contribution

This paper revisits Integrated Gradients (IG) for multimodal Transformers operating in embedding space and introduces two modifications: (i) a **manifold-aware baseline** built from empirical mean embeddings (to keep IG paths near high-support regions), and thus produce more faithful explanations; (ii) **Group-Sparse IG (GS-IG)**, which re-parameterizes the IG path schedule to generate token-level sparse attributions which more concise. On MIMIC-IV (incident heart failure) and MDC (early mortality), the manifold-aware baseline improves faithfulness (Comprehensiveness \uparrow , Sufficiency \downarrow), and GS-IG reduces token-level $\ell_{2,1}$ sparsity

by ~5–18% with negligible change in those metrics.

3.5.2 Method Overview

IG Let $F : \mathbb{R}^d \rightarrow \mathbb{R}$ denote a trained neural network whose scalar output we wish to explain; Given a baseline x' and a path $\gamma(t)$ from x' to x , IG attributes feature i as

$$\text{IG}_i(x; x', \gamma) = \int_0^1 \frac{\partial F(\gamma(t))}{\partial \gamma_i(t)} \frac{\partial \gamma_i(t)}{\partial t} dt,$$

In practice, the integral is approximated by a Riemann sum over K points.

Manifold-aware baseline. As the IG baseline, we use the *position-wise mean input embedding* computed on the validation set, which keeps the IG path near high-support regions of the embedding space. For modality m , let $E_n^{(m)}$ denote the embeddings of the validation sequence, EHR trajectory, n (obtained by passing the sequence through the encoder). We define the empirical mean

$$\mu^{(m)} = \frac{1}{N} \sum_{n=1}^N E_n^{(m)},$$

and use $\mu^{(m)}$ as the baseline for that modality; for tabular features we use feature-wise means. This semantically grounded choice keeps the IG interpolation points close to typical sequences, reducing noisy gradient accumulation. Empirically, the mean-embedding baseline lies closer to the validation embedding cloud (smaller Euclidean/Mahalanobis distances and higher KDE log-density) than other heuristics.

Group-Sparse IG (GS-IG). GS-IG produces compact, token-level explanations by re-parameterizing the IG path’s speed/schedule and selecting its schedule to encourage group sparsity across each token’s embedding coordinates. Keep the straight geometry, we replace the straight-line schedule $\alpha(t) = t$ with

$$\gamma_\theta(t) = x' + \alpha_\theta(t) (x - x'), \alpha_\theta(t) = t^\theta, \quad \theta \in [0.1, 5.0],$$

where the exponent θ smoothly interpolates between baseline-biased ($\theta > 1$) and input-biased ($\theta < 1$) trajectories.

Let $\widehat{\text{IG}}_{\text{emb}}(x, x'; \theta) \in \mathbb{R}^{L \times d_e}$ be the embedding-space IG matrix; define token saliency $s_j = \|\widehat{\text{IG}}_{\text{emb}}[j, :]\|_2$ and select θ^* per input by minimizing an $\ell_{2,1}$ Group Lasso

$$\mathcal{L}(\theta) = \lambda_{\text{grp}} \sum_{j=1}^L m_j \|\widehat{\text{IG}}_{\text{emb}}[j, :]\|_2,$$

promoting *token-level* sparsity (entire rows shrink to zero). $m_j \in \{0, 1\}$ masks out padding/special tokens. A small Bayesian search over θ (e.g., $T=10$) used in practice.

We used A multimodal Transformer pre-trained with TOO-BERT and fine-tuned per task; faithfulness is evaluated with Comprehensiveness (Comp; removal test) and Sufficiency (Suff; keep-only test).

3.5.3 Key Results

Manifold-aware baseline improves faithfulness. On sequential data, IG(mean-embedding) achieves higher Comp and lower Suff than common heuristics on both MDC and MIMIC-IV (e.g., MDC: Comp 0.244 vs. 0.205 for MASK-all; Suff 0.022 vs. 0.030). More broadly, the manifold-aware baseline outperforms alternatives across $k \in \{10, \dots, 50\}$ and across modalities.

GS-IG yields concise token lists with minimal cost to fidelity. Optimizing θ reduces token-level group sparsity by $\approx 5\text{--}18\%$ across baselines while leaving Comp/Suff essentially unchanged. Qualitative panel (Fig. 3.4) shows fewer active codes and crisper token sets with GS-IG.

3.6 PAPER VI: From Individual Attributions to Population Risk: Identifying Key Drivers of Longevity and Early Mortality from Longitudinal EHR Data

3.6.1 Aim and Contribution

Individual-level explanations are useful for case review, but do not directly answer cohort-level questions about which factors consistently drive risk across a population (RQ5). This paper introduces a population-level explanation framework for longitudinal EHR models by aggregating individual IG attributions across patients to identify robust risk and protective factors for early mortality and exceptional longevity. The key contribution is a set of complementary global attribution definitions that capture both (i) magnitude-based importance (overall sensitivity) and (ii) direction-aware importance (risk- vs. protective effects), together with a consensus procedure that stabilizes feature rankings across multiple trained model instances and aggregation schemes, yielding cohort-level insights suitable for system-level analysis.

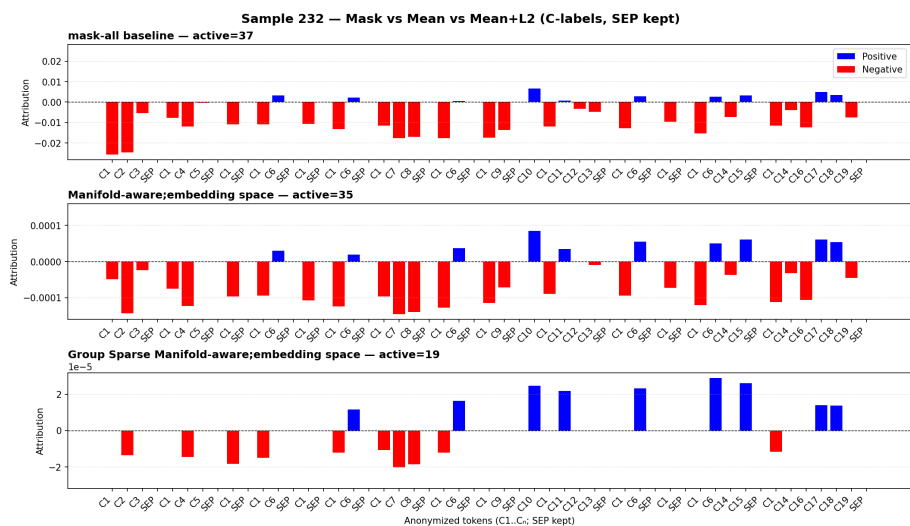


Figure 3.4: Qualitative comparison for a held-out patient from the MDC cohort predicted as early death. **Top:** IG with [MASK]-all baseline; **Middle:** IG with (manifold-aware) baseline; **Bottom:** GS-IG (manifold-aware + group sparsity). Each panel shows medical-code attributions over time and reports the number of *active* codes (non-zero attribution). Here, GS-IG reduces active codes from 35 to 19 (vs. the middle panel), yielding a sparser, more readable list of decisive factors. Red indicates contributions toward early death; blue indicates contributions toward long life.

3.6.2 Method Overview

Signed IG at the patient level. For each patient trajectory, we compute IG attributions and map them to (i) longitudinal medical codes (e.g., diagnoses and medications) and (ii) baseline phenotype/lifestyle features. Attributions are *signed* so that positive vs. negative contributions indicate whether a feature pushes the prediction toward vs. away from the target outcome (e.g., early mortality). Figure 3.5 illustrates a representative long-life case, showing explanations for the longitudinal trajectory.

Global importance via cohort aggregation. To translate individual explanations into population summaries, we aggregate patient-level attributions into global scores using multiple definitions: (i) *magnitude-based* importance (absolute attribution) to measure overall sensitivity, and (ii) *direction-aware* importance that separates positive and negative contributions to distinguish features that consistently push toward versus against the target outcome.

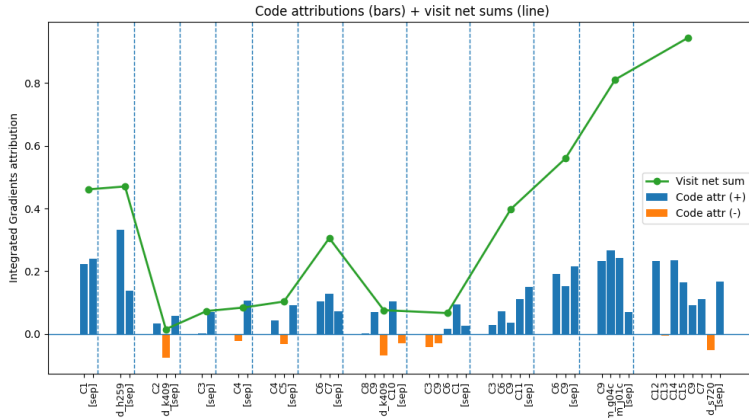


Figure 3.5: Patient-level Integrated Gradients (IG) explanation for long-life prediction (2-year window): correctly predicted long life. Bars show token attributions for the longitudinal ICD/ATC trajectory in temporal order (positive vs. negative contributions toward the long-life class). Dashed vertical lines separate visits. The green curve shows the net sum of attributions within each visit.

Robust consensus across model and aggregation variability. Because global rankings can vary with model initialization and aggregation choice, we perform a consensus analysis by combining rankings obtained from multiple trained model instances and aggregation schemes. Features that repeatedly emerge as important are prioritized, producing a stability-focused set of cohort-level drivers.

3.6.3 Key Results

Stable cohort-level drivers from individual explanations. Across outcomes, the proposed global attribution definitions recover coherent, direction-aware population-level feature sets, demonstrating that individual attributions can be translated into meaningful cohort-level summaries.

Consensus reduces sensitivity to modeling and aggregation choices. Combining rankings across multiple trained model instances and aggregation schemes yields more robust, reproducible feature importance lists than any single aggregation alone, supporting reliable cohort-wide interpretation beyond single-patient explanations.

3.7 PAPER VII: A decoupled alignment kernel for peptide membrane permeability predictions

3.7.1 Aim and Contribution

In small-to-medium datasets of molecular sequences, we need models that deliver *both* strong predictive performance and trustworthy uncertainty. This paper addresses the need for membrane permeability of cyclic peptides by introducing a **monomer-decoupled global alignment kernel (MD-GAK)** for Gaussian processes (GPs). We represent each peptide as a sequence of monomers (residues) and introduce the MD-GAK to compare two peptides. Unlike the global alignment kernel, it decouples the penalty for warping (gaps or insertions) from the contribution of chemical matches. Concretely, local matches are scored with a chemistry-aware kernel (Tanimoto on Morgan fingerprints), while gap propagation is handled separately in the dynamic program. Combined with GP classification, this yields calibrated probabilistic predictions and improved discrimination under leakage-controlled evaluation on CycPeptMPDB—providing reliable “what the model knows” estimates in the data-constrained regime. We proved that MD-GAK is a valid kernel.

Motivation: limits of pretrain-fine-tune for peptides. While the pretrain fine-tune paradigm (e.g., chemical language models trained on large unlabeled corpora and adapted to downstream property prediction) has shown promise, it can underperform in small/medium peptide sequences datasets due to (i) **distribution mismatch** (pretraining often targets small molecules or linear peptides, whereas cyclic peptides have distinct topology, stereochemistry, and monomer vocabularies), (ii) **objective mismatch** (self-supervised text/SMILES objectives do not align with permeability mechanisms), and (iii) **data scarcity during adaptation**, which increases overfitting and degrades calibration. Our kernel-GP approach addresses these issues by operating directly on monomer sequences with a chemistry-aware local similarity and an alignment mechanism tailored to sequence structure, yielding strong discrimination with calibrated uncertainties without relying on large-scale pretraining.

3.7.2 Method Overview

Monomer representation and local similarity. Each cyclic peptide is represented as an ordered sequence of monomer SMILES; every monomer is encoded by Morgan fingerprint, and local chemistry between monomers is compared with the (PSD) Tanimoto similarity kernel κ_0 :

$$\kappa_0(\phi(s), \phi(t)) = \frac{\langle \phi(s), \phi(t) \rangle}{\|\phi(s)\|_1 + \|\phi(t)\|_1 - \langle \phi(s), \phi(t) \rangle}.$$

Decoupled global alignment (MD-GAK). Let $A = (s_1, \dots, s_n)$ and $B = (t_1, \dots, t_m)$. MD-GAK sums over all monotone alignments but *decouples* match contributions from gap propagation:

$$M_{0,0}=1, \quad M_{i,0}=M_{0,j}=0, \quad M_{i,j} = \kappa_0(\phi(s_i), \phi(t_j)) M_{i-1,j-1} + M_{i-1,j} + M_{i,j-1},$$

and $K_{\text{MD-GAK}}(A, B) = M_{n,m}$ (optionally cosine-normalized). Intuitively, chemistry κ_0 affects only diagonal (match) steps, while gaps accumulate independently; this avoids extinguishing paths at poor local matches and reduces sensitivity to isolated mismatches. We proved $K_{\text{MD-GAK}}$ is positive semidefinite if κ_0 is PSD.

Gaussian-process classification. MD-GAK serves as the GP covariance; binary permeability (PAMPA threshold $P \geq -6$) is modeled with a logistic likelihood and Laplace approximation, yielding calibrated posteriors. We compared MD-GAK against the global alignment kernel, tree-based models, the chemical language model ChemBERTa, and strong graph-, string-, and image-based benchmark models.

Data and Evaluation Protocols Experiments use CycPeptMPDB (PAMPA assay) with careful near duplicate control and two complementary settings: **(A)** applicability-domain-*aware* with (i) label-stratified and (ii) canonical-group-stratified folds to curb leakage; **(B)** a length-focused subset (6/7/10-mers) under random vs. scaffold splits, mirroring recent benchmarks. Metrics include ACC, F1, ROC-AUC, Brier score, and Estimated Calibration Error (ECE).

3.7.3 Key Results

Setting A: label-stratified nested CV. MD-GAK GP attains the strongest *threshold* metrics among GPs and vector baselines: ACC **83.0** \pm 0.5, F1 **73.7** \pm 0.9, ROC-AUC 87.8 \pm 0.7, with competitive calibration (Brier 14.30 \pm 0.46, ECE 12.31 \pm 1.61). Tree ensembles have decent calibration but lower AUC.

Setting A: canonical-group-stratified nested CV (harder). Removing near-duplicates reduces performance across models; MD-GAK remains strong (ACC 80.1 \pm 1.4, F1 67.8 \pm 4.4, ROC-AUC 84.7 \pm 1.0).

Setting B: length-focused 6/7/10-mers. On both random and scaffold splits, the MD-GAK GP achieves top ROC-AUC (random: 88.8 \pm 0.2; scaffold: 79.8 \pm 0.0), exceeding strong graph baselines (e.g., AttentiveFP random 86.2 \pm 1.8, MPNN scaffold 73.4 \pm 8.7) and showing a smaller degradation from random \rightarrow scaffold, suggesting better robustness to distribution shift.

4. Concluding Remarks and Future Work

4.1 Concluding remarks

In this thesis, we set out to address a central question:

How can we improve prediction and provide trustworthy explanations for adverse health outcomes by modeling longitudinal EHR trajectories?

Starting from a systematic review of deep learning on EHR trajectories, the work has progressed along two tightly connected tracks: (i) improving representation learning and fine-tuning of foundation-style models on longitudinal, multi-source EHR data under realistic data constraints; and (ii) developing methods for local explanations and uncertainty quantification that make such models more transparent and trustworthy.

Summary of answers to the research questions

RQ1: Temporal order. TOO-BERT (Paper II) augments MLM with a trajectory-order objective that discriminates between correctly ordered and permuted trajectories at the code and visit level. By using more plausible permutations, the model learns order-sensitive dependencies and complements the learned context by MLM. This improves downstream AUCs on both MIMIC-IV and MDC and yields richer attention patterns across the full trajectory. Temporal order thus provides an effective auxiliary pretraining signal.

RQ2: Robust fine-tuning. To improve fine-tuning under realistic issue of label scarcity, ANAA (Paper IV) perturbs attention scores with adaptive Gaussian noise and then smooths them with a 2D Gaussian kernel. This broadens overly sharp attention distributions while preserving structure, leading to more informative attention maps and higher AUCs. Operating directly in attention space provides a straightforward, architecture-preserving method to stabilize fine-tuning without incurring additional inference costs.

RQ3: Cross-source dependencies. For multi-source trajectories, Paper III extends MLM with a source-masked prediction step, where masked contents from one source (e.g., ICD) must be reconstructed from other sources (e.g., ATC). This forces the model to encode disease–intervention and cross-source relations rather than treating modalities independently. On next-visit heart failure prediction in MDC, this two-step masking improves the model’s performance.

RQ4. Concise and faithful token-level explanations. Paper V revisits Integrated Gradients for embedding-based EHR trajectories models. A manifold-aware baseline, defined from empirical mean embeddings, keeps IG paths in high-density regions and improves faithfulness. Group-Sparse IG (GS-IG) then adjusts the path schedule to induce token-level sparsity, yielding concise attributions. These sparse, manifold-aware explanations are well-suited for aggregation across patients to reveal common risk patterns.

RQ5: From individual attributions to population-level explanations. While feature-level explanations support case-based interpretation, they do not directly provide cohort-level insight. Paper VI introduces a population-level explanation framework that aggregates Integrated Gradients attributions across patients to identify stable cohort-level features whose attributions consistently push toward versus against early mortality and exceptional longevity. By combining magnitude-based and direction-aware global attribution measures with a consensus ranking across multiple trained model instances and aggregation schemes, the method produces robust cohort-level drivers and reduces sensitivity to modeling and aggregation choices.

RQ6: Uncertainty in small/medium sequence datasets. Paper VII studies membrane permeability of peptides and proposes the Monomer Decoupled Global Alignment Kernel (MD-GAK) within a Gaussian process classifier. By decoupling chemical similarity from gap penalties in the alignment dynamic program, MD-GAK captures sequence-sensitive structure while remaining robust to local mismatches. The resulting GP achieves competitive accuracy and well-calibrated uncertainty on multiple evaluation protocols, illustrating that domain-informed kernel–GP models are a strong alternative to pretrain–fine-tune pipelines in data-limited sequence settings.

Broader contributions and lessons learned

Beyond the individual research questions, the thesis offers a broader perspective on foundation-style modeling for longitudinal EHR:

- **EHR-specific pretraining matters.** In longitudinal EHR, data are complex and scarce, so naive MLM-style pretraining is often insufficient. By designing self-supervised objectives that exploit the structure of trajectories—such as temporal-order objectives (TOO-BERT) and cross-source masking— we can enrich the pretraining stage, learn more robust representations, and enhance downstream performance.
- **Attention is both a modeling tool and a regularizer.** Self-attention is central for capturing dependencies in patient trajectories, but naive fine-tuning can yield brittle, near-binary attention maps. Augmenting attention scores directly, as in ANAA, provides a simple way to regularize these patterns and expose the model to more diverse views of the data.
- **Explanation quality depends on geometry.** Integrated Gradients is a powerful, axiomatically grounded method for explaining deep models, but for embedding-based architectures, its behavior is highly sensitive to how attribution paths traverse representation space. In this thesis, we treat the choice of baseline and path as degrees of freedom that can be optimized so that IG attains the desired properties for the specific task: by using sparse manifold-aware baselines and path schedules that better follow the data distribution, we substantially improve both the fidelity and the practical interpretability of token-level attributions, bringing them closer to what is needed in real clinical settings.
- **Small data and uncertainty warrant dedicated methods.** While foundation models excel in high-data regimes, this thesis shows that in small-to medium-sized sequence datasets, kernel-based probabilistic models with domain-informed similarity functions can deliver not only competitive accuracy but also calibrated uncertainty estimates that are crucial for risk-aware decision-making.
- **Systematic understanding of the field is a prerequisite.** The systematic review in Paper I provided the conceptual “map” for this thesis, identifying recurring challenges—heterogeneity, temporal irregularity, data scarcity, and limited explainability—that guided the subsequent methodological contributions. It also underscored gaps in benchmarking and evaluation that remain obstacles for community-wide progress.

Taken together, the thesis demonstrates that substantial gains in both performance and trustworthiness can be achieved not by ever larger models, but by aligning objectives, fine-tuning strategies, and explanation methods with the specific structure and constraints of EHR trajectories.

Limitations

Several limitations of this work should be acknowledged. First, most empirical evaluations are based on a small number of cohorts (MIMIC-IV and a Swedish regional registry), with specific healthcare practices and coding standards. Although these datasets are diverse in population and setting, the generalizability of the proposed methods to other countries, health systems, and coding schemes remains to be tested systematically.

Second, the focus has been on structured, code-based trajectories (ICD, ATC), with only limited integration of other modalities. While the methods are, in principle, extensible to richer multimodal inputs (e.g., such as free text, imaging, or high-frequency physiological signals), their behavior and benefits need further investigation.

Third, the evaluation of explanation methods relies primarily on proxy metrics (comprehensiveness, sufficiency) and qualitative case studies. These are necessary but not sufficient proxies for clinical usefulness. User studies with clinicians, as well as prospective assessments of whether explanations improve decision-making, are outside the scope of this thesis.

Fourth, the uncertainty quantification work is demonstrated on molecular rather than EHR data. The conceptual insights carry over, but a direct integration of kernel-GP heads or conformal wrappers with EHR Transformers would be needed to fully realize the potential of calibrated uncertainty in the clinical trajectory setting.

Finally, most methods are evaluated in retrospective, offline prediction scenarios. Issues such as deployment in real-time systems, updating models over time, and handling changing data distributions are not addressed here and remain important challenges for future work.

Future directions

Building on the contributions of this thesis, several directions appear particularly promising:

- **Richer multimodal trajectories.** Extending order-aware and source-masked pretraining to settings that include free-text notes, imaging features, and sensor data could enable models that better reflect the full information available in clinical practice.
- **Federated and privacy-preserving pretraining.** Combining the proposed objectives with federated or distributed learning frameworks would allow leveraging multi-institutional data while respecting privacy and governance

constraints, and would directly address challenges around dataset shift and external validation.

- **Integrated uncertainty and explanation.** Bringing together GS-IG style explanations with calibrated uncertainty (e.g., via GP heads or conformal prediction) could support interfaces that present not only “why” a prediction was made, but also “how sure” the model is, in a way that is actionable for clinicians.
- **Human-in-the-loop evaluation.** Finally, embedding these methods in interactive tools and studying how clinicians use them in practice—what they find trustworthy, confusing, or useful—will be essential for translating methodological advances into real-world impact.

In conclusion, this thesis demonstrates that modeling longitudinal EHR trajectories with foundation-style architectures is not only a question of scale, but also a question of design. By tailoring pretraining objectives, fine-tuning strategies, and explanation methods to the particularities of clinical data, we can move towards models that are more accurate, more data-efficient, and more transparent. Such models bring us a step closer to the overarching goal: using EHR trajectories to anticipate adverse outcomes in time to intervene, while providing clinicians with explanations and confidence estimates they can understand, inspect, and act upon.

References

- [1] ASHISH VASWANI, NOAM SHAZEER, NIKI PARMAR, JAKOB USZKOREIT, LLION JONES, AIDAN N GOMEZ, ŁUKASZ KAISER, AND ILLIA POLOSUKHIN. **Attention is all you need.** *Advances in neural information processing systems*, **30**, 2017. xiii, 1, 9, 13, 14
- [2] ALI AMIRAHMADI, MATTIAS OHLSSON, AND KOBRA ETMINANI. **Deep learning prediction models based on EHR trajectories: A systematic review.** *Journal of biomedical informatics*, **144**:104430, 2023. 1, 2, 3, 4
- [3] YUQI SI, JINGCHENG DU, ZHAO LI, XIAOQIAN JIANG, TIMOTHY MILLER, FEI WANG, W JIM ZHENG, AND KIRK ROBERTS. **Deep representation learning of patient data from Electronic Health Records (EHR): A systematic review.** *Journal of biomedical informatics*, **115**:103671, 2021. 1, 3
- [4] CAO XIAO, EDWARD CHOI, AND JIMENG SUN. **Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review.** *Journal of the American Medical Informatics Association*, **25**(10):1419–1428, 2018. 1
- [5] SEPP HOCHREITER AND JÜRGEN SCHMIDHUBER. **Long short-term memory.** *Neural computation*, **9**(8):1735–1780, 1997. 1
- [6] JACOB DEVLIN, MING-WEI CHANG, KENTON LEE, AND KRISTINA TOUTANOVA. **Bert: Pre-training of deep bidirectional transformers for language understanding.** In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 1, 9, 11, 12
- [7] RISHI BOMMASANI. **On the opportunities and risks of foundation models.** *arXiv preprint arXiv:2108.07258*, 2021. 1, 9
- [8] YIKUAN LI, SHISHIR RAO, JOSÉ ROBERTO AYALA SOLARES, ABDELAALI HASSAINE, REMA RAMAKRISHNAN, DEXTER CANOY, YAJIE ZHU, KAZEM RAHIMI, AND GHOLAMREZA SALIMI-KHORSHIDI. **BEHRT: transformer for electronic health records.** *Scientific reports*, **10**(1):7155, 2020. 2, 10, 13
- [9] NING LIU, QIAN HU, HUAYUN XU, XING XU, AND MENGXIN CHEN. **Med-BERT: A pretraining framework for medical records named entity recognition.** *IEEE Transactions on Industrial Informatics*, **18**(8):5600–5608, 2021. 2, 10, 11, 13
- [10] MUHAMMAD AYAZ, MUHAMMAD F PASHA, MOHAMMED Y ALZAHIRANI, RAHMAT BUDIARTO, AND DERIS STIAWAN. **The Fast Health Interoperability Resources (FHIR) standard: systematic literature review of implementations, applications, challenges and opportunities.** *JMIR medical informatics*, **9**(7):e21929, 2021. 3
- [11] ALI AMIRAHMADI, FARZANEH ETMINANI, JONAS BJÖRK, OLLE MELANDER, AND MATTIAS OHLSSON. **Trajectory-Ordered Objectives for Self-Supervised Representation Learning of Temporal Healthcare Data Using Transformers: Model Development and Evaluation Study.** *JMIR Medical Informatics*, **13**(1):e68138, 2025. 3, 10, 11, 13
- [12] ZIYANG SONG, QINGCHENG LU, HE ZHU, DAVID BUCKERIDGE, AND YUE LI. **TrajGPT: Irregular Time-Series Representation Learning for Health Trajectory Analysis.** *arXiv preprint arXiv:2410.02133*, 2024. 3

- [13] LINGLONG QIAN, YIYUAN YANG, WENJIE DU, JUN WANG, RICHARD DOBSONI, AND ZINA IBRAHIM. **Beyond Random Missingness: Clinically Rethinking for Healthcare Time Series Imputation.** *arXiv preprint arXiv:2405.17508*, 2024. 3
- [14] YUANYUAN ZHENG, ADEL BENSATLA, MINA BJELOGRLIC, JAMIL ZAGHIR, HUGUES TURBE, LYDIE BEDNARCZYK, CHRISTOPHE GAUDET-BLAVIGNAC, JULIEN EHRSAM, STÉPHANE MARCHAND-MAILLET, AND CHRISTIAN LOVIS. **A scoping review of self-supervised representation learning for clinical decision making using EHR categorical data.** *npj Digital Medicine*, 8(1):362, 2025. 3
- [15] JOHANNA DRIEVER, MANUEL LENTZEN, SUMIT MADAN, AND HOLGER FRÖHLICH. **Domain Adaptation Strategies for Transformer-Based Disease Prediction using Electronic Health Records.** *medRxiv*, pages 2025–06, 2025. 4
- [16] HUIYA ZHAO, DEHAO SUI, YASHA WANG, LIANTAO MA, AND LING WANG. **Privacy-Preserving Federated Learning Framework for Multi-Source Electronic Health Records Prognosis Prediction.** *Sensors*, 25(8):2374, 2025. 4
- [17] ASGHAR ALI, VÁCLAV SNÁŠEL, AND JAN PLATOŠ. **Health-FedNet: A Privacy-Preserving Federated Learning Framework for Scalable and Secure Healthcare Analytics.** *Results in Engineering*, page 106484, 2025.
- [18] PUSHPENDRA SINGH ET AL. **CrypTen-FL: A Secure Federated Learning Framework for Multi-Disease Prediction from MIMIC-IV Using Encrypted EHRs.** *International Journal of Advanced Computer Science & Applications*, 16(9), 2025. 4
- [19] ANIEK F MARKUS, JAN A KORS, AND PETER R RIJNBEEK. **The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies.** *Journal of biomedical informatics*, 113:103655, 2021. 4, 16
- [20] MICHAEL WORNOW, RAHUL THAPA, ETHAN STEINBERG, JASON FRIES, AND NIGAM SHAH. **Ehrshot: An ehr benchmark for few-shot evaluation of foundation models.** *Advances in Neural Information Processing Systems*, 36:67125–67137, 2023. 4
- [21] ALISTAIR EW JOHNSON, LUCAS BULGARELLI, LU SHEN, ALVIN GAYLES, AYAD SHAMMOUT, STEVEN HORNG, TOM J POLLARD, SICHENG HAO, BENJAMIN MOODY, BRIAN GOW, ET AL. **MIMIC-IV, a freely accessible electronic health record dataset.** *Scientific data*, 10(1):1, 2023. 4
- [22] MICHAEL MOOR, OISHI BANERJEE, ZAHRA SHAKERI HOSSEIN ABAD, HARLAN M KRUMHOLZ, JURE LESKOVEC, ERIC J TOPOL, AND PRANAV RAJPURKAR. **Foundation models for generalist medical artificial intelligence.** *Nature*, 616(7956):259–265, 2023. 9
- [23] ALEC RADFORD, KARTHIK NARASIMHAN, TIM SALIMANS, ILYA SUTSKEVER, ET AL. **Improving language understanding by generative pre-training.** *OpenAI blog*, 2018. 9
- [24] TOM BROWN, BENJAMIN MANN, NICK RYDER, MELANIE SUBBIAH, JARED D KAPLAN, PRAFULLA DHARIWAL, ARVIND NEELAKANTAN, PRANAV SHYAM, GIRISH SASTRY, AMANDA ASKELL, ET AL. **Language models are few-shot learners.** *Advances in neural information processing systems*, 33:1877–1901, 2020. 9, 10, 12, 13
- [25] ALEX WANG, AMANPREET SINGH, JULIAN MICHAEL, FELIX HILL, OMER LEVY, AND SAMUEL BOWMAN. **GLUE: A multi-task benchmark and analysis platform for natural language understanding.** In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pages 353–355, 2018. 9
- [26] PRANAV RAJPURKAR, JIAN ZHANG, KONSTANTIN LOPYREV, AND PERCY LIANG. **Squad: 100,000+ questions for machine comprehension of text.** *arXiv preprint arXiv:1606.05250*, 2016. 9
- [27] ANNA ROGERS, OLGA KOVALEVA, AND ANNA RUMSHISKY. **A primer in BERTology: What we know about how BERT works.** *Transactions of the association for computational linguistics*, 8:842–866, 2021. 9

- [28] ALEC RADFORD, JEFFREY WU, REWON CHILD, DAVID LUAN, DARIO AMODEI, ILYA SUTSKEVER, ET AL. **Language models are unsupervised multitask learners.** *OpenAI blog*, 1(8):9, 2019. 9
- [29] YOSHUA BENGIO, IAN GOODFELLOW, AARON COURVILLE, ET AL. *Deep learning*, 1. MIT press Cambridge, MA, USA, 2017. 10
- [30] CURTIS G NORTHCUIT, ANISH ATHALYE, AND JONAS MUELLER. **Pervasive label errors in test sets destabilize machine learning benchmarks.** *arXiv preprint arXiv:2103.14749*, 2021.
- [31] ANTONIO TORRALBA AND ALEXEI A EFROS. **Unbiased look at dataset bias.** In *CVPR 2011*, pages 1521–1528. IEEE, 2011. 10
- [32] CHAO PANG, XINZHUO JIANG, KRISHNA S KALLURI, MATTHEW SPOTNITZ, RUIJUN CHEN, ADLER PEROTTE, AND KARTHIK NATARAJAN. **CEHR-BERT: Incorporating temporal information from structured EHR data to improve prediction tasks.** In *Machine Learning for Health*, pages 239–260. PMLR, 2021. 10, 11, 13
- [33] YIKUAN LI, MOHAMMAD MAMOUEI, GHOLAMREZA SALIMI-KHORSHIDI, SHISHIR RAO, ABDELAALI HASSAINE, DEXTER CANOY, THOMAS LUKASIEWICZ, AND KAZEM RAHIMI. **Hi-BEHT: hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records.** *IEEE journal of biomedical and health informatics*, 27(2):1106–1117, 2022. 11
- [34] PKS PRAKASH, SRINIVAS CHILUKURI, NIKHIL RANADE, AND SHANKAR VISWANATHAN. **RareBERT: transformer architecture for rare disease patient identification using administrative claims.** In *Proceedings of the AAAI conference on artificial intelligence*, 35, pages 453–460, 2021. 13
- [35] ALI AMIRAHMADI, MATTIAS OHLSSON, KOBRA ETMINANI, OLLE MELANDER, AND JONAS BJÖRK. **A Masked language model for multi-source EHR trajectories contextual representation learning.** *arXiv preprint arXiv:2402.06675*, 2024. 10, 11, 13
- [36] LIN LAWRENCE GUO, ETHAN STEINBERG, SCOTT LANYON FLEMING, JOSE POSADA, JOSHUA LEMMON, STEPHEN R PFOHL, NIGAM SHAH, JASON FRIES, AND LILLIAN SUNG. **EHR foundation models improve robustness in the presence of temporal distribution shift.** *Scientific Reports*, 13(1):3767, 2023. 10
- [37] ETHAN STEINBERG, KEN JUNG, JASON A FRIES, CONOR K CORBIN, STEPHEN R PFOHL, AND NIGAM H SHAH. **Language models are an effective representation learning technique for electronic health record data.** *Journal of biomedical informatics*, 113:103637, 2021.
- [38] XIANLONG ZENG, SIMON L LINWOOD, AND CHANG LIU. **Pretrained transformer framework on pediatric claims data for population specific tasks.** *Scientific Reports*, 12(1):3651, 2022. 10, 11
- [39] JUNYUAN SHANG, TENGFEI MA, CAO XIAO, AND JIMENG SUN. **Pre-training of graph augmented transformers for medication recommendation.** *arXiv preprint arXiv:1906.00346*, 2019. 11
- [40] TIANRAN ZHANG, MUHAO CHEN, AND ALEX AT BUI. **AdaDiag: Adversarial domain adaptation of diagnostic prediction with clinical event sequences.** *Journal of biomedical informatics*, 134:104168, 2022. 11
- [41] JEAN-BASTIEN GRILL, FLORIAN STRUB, FLORENT ALTCHÉ, CORENTIN TALLEC, PIERRE RICHEMOND, ELENA BUCHATSKAYA, CARL DOERSCH, BERNARDO AVILA PIRES, ZHAOHAN GUO, MOHAMMAD GHESLAGHI AZAR, ET AL. **Bootstrap your own latent-a new approach to self-supervised learning.** *Advances in neural information processing systems*, 33:21271–21284, 2020. 11
- [42] HOUXING REN, JINGYUAN WANG, WAYNE XIN ZHAO, AND NING WU. **Rapt: Pre-training of time-aware transformer for learning robust healthcare representation.** In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 3503–3511, 2021. 11
- [43] HOUXING REN, JINGYUAN WANG, AND WAYNE XIN ZHAO. **Generative adversarial networks enhanced pre-training for insufficient electronic health records modeling.** In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3810–3818, 2022. 11

- [44] ZHENZHONG LAN, MINGDA CHEN, SEBASTIAN GOODMAN, KEVIN GIMPEL, PIYUSH SHARMA, AND RADU SORICUT. **Albert: A lite bert for self-supervised learning of language representations.** *arXiv preprint arXiv:1909.11942*, 2019. 11
- [45] JEREMY HOWARD AND SEBASTIAN RUDER. **Universal language model fine-tuning for text classification.** *arXiv preprint arXiv:1801.06146*, 2018. 11, 12
- [46] ILYA LOSHCHEV AND FRANK HUTTER. **Decoupled weight decay regularization.** *arXiv preprint arXiv:1711.05101*, 2017. 11
- [47] MARIUS MOSBACH, MAKSYM ANDRIUSHCHENKO, AND DIETRICH KLAKOW. **On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines.** *arXiv preprint arXiv:2006.04884*, 2020. 11
- [48] SEYED IMAN MIRZADEH, MEHRDAD FARAJTABAR, DILAN GORUR, RAZVAN PASCANU, AND HASSAN GHASEMZADEH. **Linear mode connectivity in multitask and continual learning.** *arXiv preprint arXiv:2010.04495*, 2020. 11
- [49] NEIL HOULSBY, ANDREI GIURGIU, STANISLAW JASTRZEBSKI, BRUNA MORRONE, QUENTIN DE LAROUSSILHE, ANDREA GESMUNDO, MONA ATTARIYAN, AND SYLVAIN GELLY. **Parameter-efficient transfer learning for NLP.** In *International conference on machine learning*, pages 2790–2799. PMLR, 2019. 12
- [50] JONAS PFEIFFER, AISHWARYA KAMATH, ANDREAS RÜCKLÉ, KYUNGHYUN CHO, AND IRYNA GUREVYCH. **Adapterfusion: Non-destructive task composition for transfer learning.** *arXiv preprint arXiv:2005.00247*, 2020. 12
- [51] EDWARD J HU, YELONG SHEN, PHILLIP WALLIS, ZEYUAN ALLEN-ZHU, YUANZHI LI, SHEAN WANG, LU WANG, WEIZHU CHEN, ET AL. **Lora: Low-rank adaptation of large language models.** *ICLR*, 1(2):3, 2022. 12
- [52] ELAD BEN ZAKEN, SHAULI RAVFOGEL, AND YOAV GOLDBERG. **Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models.** *arXiv preprint arXiv:2106.10199*, 2021. 12
- [53] HAOKUN LIU, DEREK TAM, MOHAMMED MUQEETH, JAY MOHTA, TENGHAO HUANG, MOHIT BANSAL, AND COLIN A RAFFEL. **Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning.** *Advances in Neural Information Processing Systems*, 35:1950–1965, 2022. 12
- [54] GUILLAUME ALAIN AND YOSHUA BENGIO. **Understanding intermediate layers using linear classifier probes.** *arXiv preprint arXiv:1610.01644*, 2016. 12
- [55] TING CHEN, SIMON KORNBLITH, MOHAMMAD NOROUZI, AND GEOFFREY HINTON. **A simple framework for contrastive learning of visual representations.** In *International conference on machine learning*, pages 1597–1607. PmLR, 2020. 12
- [56] XIANG LISA LI AND PERCY LIANG. **Prefix-tuning: Optimizing continuous prompts for generation.** *arXiv preprint arXiv:2101.00190*, 2021. 12
- [57] ALI AMIRAHMADI, FARZANEH ETMINANI, AND MATTIAS OHLSSON. **Adaptive noise-augmented attention for enhancing Transformer fine-tuning on longitudinal medical data.** *Frontiers in Artificial Intelligence*, 8:1663484, 2025. 13
- [58] JIMMY LEI BA, JAMIE RYAN KIROS, AND GEOFFREY E HINTON. **Layer normalization.** *arXiv preprint arXiv:1607.06450*, 2016. 15
- [59] IZ BELTAGY, MATTHEW E PETERS, AND ARMAN COHAN. **Longformer: The long-document transformer.** *arXiv preprint arXiv:2004.05150*, 2020. 15
- [60] MANZIL ZAHAEER, GURU GURUGANESH, KUMAR AVINAVA DUBEY, JOSHUA AINSLIE, CHRIS ALBERTI, SANTIAGO ONTANON, PHILIP PHAM, ANIRUDH RAVULA, QIFAN WANG, LI YANG, ET AL. **Big bird: Transformers for longer sequences.** *Advances in neural information processing systems*, 33:17283–17297, 2020. 15

- [61] SINONG WANG, BELINDA Z LI, MADIAN KHABSA, HAN FANG, AND HAO MA. **Linformer: Self-attention with linear complexity.** *arXiv preprint arXiv:2006.04768*, 2020. 15
- [62] KRZYSZTOF CHOROMANSKI, VALERII LIKHOSHERSTOV, DAVID DOHAN, XINGYOU SONG, ANDREEA GANE, TAMAS SARLOS, PETER HAWKINS, JARED DAVIS, AFROZ MOHIUDDIN, LUKASZ KAISER, ET AL. **Rethinking attention with performers.** *arXiv preprint arXiv:2009.14794*, 2020. 15
- [63] QING LYU, MARIANNA APIDIANAKI, AND CHRIS CALLISON-BURCH. **Towards faithful model explanation in nlp: A survey.** *Computational Linguistics*, **50**(2):657–723, 2024. 16
- [64] JAY DEYOUNG, SARTHAK JAIN, NAZNEEN FATEMA RAJANI, ERIC LEHMAN, CAIMING XIONG, RICHARD SOCHER, AND BYRON C WALLACE. **ERASER: A benchmark to evaluate rationalized NLP models.** *arXiv preprint arXiv:1911.03429*, 2019. 16, 17
- [65] SOUMYA SANYAL AND XIANG REN. **Discretized integrated gradients for explaining language models.** *arXiv preprint arXiv:2108.13654*, 2021.
- [66] JOSEPH ENGUEHARD. **Sequential integrated gradients: a simple but effective method for explaining language models.** *arXiv preprint arXiv:2305.15853*, 2023. 16
- [67] JENNIFER HSIA, DANISH PRUTHI, AARTI SINGH, AND ZACHARY C LIPTON. **Goodhart’s law applies to nlp’s explanation benchmarks.** *arXiv preprint arXiv:2308.14272*, 2023. 16
- [68] SARTHAK JAIN AND BYRON C WALLACE. **Attention is not explanation.** *arXiv preprint arXiv:1902.10186*, 2019. 16
- [69] SOFIA SERRANO AND NOAH A SMITH. **Is attention interpretable?** *arXiv preprint arXiv:1906.03731*, 2019. 16
- [70] MUKUND SUNDARARAJAN, ANKUR TALY, AND QIQI YAN. **Axiomatic attribution for deep networks.** In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. 16
- [71] JULIUS ADEBAYO, JUSTIN GILMER, MICHAEL MUELLY, IAN GOODFELLOW, MORITZ HARDT, AND BEEN KIM. **Sanity checks for saliency maps.** *Advances in neural information processing systems*, **31**, 2018. 16
- [72] ALI AMIRAHMADI, FARZANEH ETMINANI, AND MATTIAS OHLSSON. **Group-Sparse Manifold-Aware Integrated Gradients for Multimodal Transformers on EHR Trajectories.** In *Machine Learning for Health 2025*, 2025. 16, 17
- [73] JINGWEI ZHANG AND FARZAN FARNIA. **Moreaugrad: Sparse and robust interpretation of neural networks via moreau envelope.** In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2021–2030, 2023. 17
- [74] JIWEI LI, WILL MONROE, AND DAN JURAFSKY. **Understanding neural networks through representation erasure.** *arXiv preprint arXiv:1612.08220*, 2016. 17
- [75] SHI FENG, ERIC WALLACE, ALVIN GRISSOM II, MOHIT IYER, PEDRO RODRIGUEZ, AND JORDAN BOYD-GRABER. **Pathologies of neural models make interpretations difficult.** *arXiv preprint arXiv:1804.07781*, 2018. 17
- [76] MARCO TULLIO RIBEIRO, SAMEER SINGH, AND CARLOS GUESTRIN. **" Why should i trust you?" Explaining the predictions of any classifier.** In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016. 17
- [77] SCOTT M LUNDBERG AND SU-IN LEE. **A unified approach to interpreting model predictions.** *Advances in neural information processing systems*, **30**, 2017. 17

- [78] SHISHIR RAO, YIKUAN LI, REMA RAMAKRISHNAN, ABDELAALI HASSAINE, DEXTER CANOY, JOHN CLELAND, THOMAS LUKASIEWICZ, GHOLAMREZA SALIMI-KHORSHIDI, AND KAZEM RAHIMI. **An explainable transformer-based deep learning model for the prediction of incident heart failure.** *IEEE journal of biomedical and health informatics*, **26**(7):3362–3372, 2022. 17
- [79] SARA HOOKER, DUMITRU ERHAN, PIETER-JAN KINDERMANS, AND BEEN KIM. **A benchmark for interpretability methods in deep neural networks.** *Advances in neural information processing systems*, **32**, 2019. 17
- [80] EYKE HÜLLERMEIER AND WILLEM WAEGEMAN. **Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods.** *Machine learning*, **110**(3):457–506, 2021. 18
- [81] CHUAN GUO, GEOFF PLEISS, YU SUN, AND KILIAN Q WEINBERGER. **On calibration of modern neural networks.** In *International conference on machine learning*, pages 1321–1330. PMLR, 2017. 18, 19
- [82] YARIN GAL AND ZOUBIN GHAHRAMANI. **Dropout as a bayesian approximation: Representing model uncertainty in deep learning.** In *international conference on machine learning*, pages 1050–1059. PMLR, 2016. 18
- [83] TYLER J LOFTUS, BENJAMIN SHICKEL, MATTHEW M RUPPERT, JEREMY A BALCH, TEZCAN OZRAZGAT-BASLANTI, PATRICK J TIGHE, PHILIP A EFRON, WILLIAM R HOGAN, PARISA RASHIDI, GILBERT R UP-CHURCH JR, ET AL. **Uncertainty-aware deep learning in healthcare: a scoping review.** *PLOS digital health*, **1**(8):e0000085, 2022. 18
- [84] BALAJI LAKSHMINARAYANAN, ALEXANDER PRITZEL, AND CHARLES BLUNDELL. **Simple and scalable predictive uncertainty estimation using deep ensembles.** *Advances in neural information processing systems*, **30**, 2017. 18
- [85] ALI AMIRAHMADI, GÖKÇE GEYLAN, LEONARDO DE MARIA, FARZANEH ETMINANI, MATTIAS OHLSSON, AND ALESSANDRO TIBO. **A decoupled alignment kernel for peptide membrane permeability predictions.** *arXiv preprint arXiv:2511.21566*, 2025. 18
- [86] JEREMIAH LIU, ZI LIN, SHREYAS PADHY, DUSTIN TRAN, TANIA BEDRAX WEISS, AND BALAJI LAKSHMINARAYANAN. **Simple and principled uncertainty estimation with deterministic deep learning via distance awareness.** *Advances in neural information processing systems*, **33**:7498–7512, 2020. 18
- [87] JOOST VAN AMERSFOORT, LEWIS SMITH, YEE WHYI TEH, AND YARIN GAL. **Uncertainty estimation using a single deep deterministic neural network.** In *International conference on machine learning*, pages 9690–9700. PMLR, 2020. 18
- [88] MURAT SENSOY, LANCE KAPLAN, AND MELIH KANDEMIR. **Evidential deep learning to quantify classification uncertainty.** *Advances in neural information processing systems*, **31**, 2018. 18
- [89] LINGKAI KONG, HAOMING JIANG, YUCHEN ZHUANG, JIE LYU, TUO ZHAO, AND CHAO ZHANG. **Calibrated language model fine-tuning for in-and out-of-distribution data.** *arXiv preprint arXiv:2010.11506*, 2020. 19
- [90] MEELIS KULL, TELMO SILVA FILHO, AND PETER FLACH. **Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers.** In *Artificial intelligence and statistics*, pages 623–631. PMLR, 2017. 19
- [91] ANASTASIOS N ANGELOPOULOS, STEPHEN BATES, ET AL. **Conformal prediction: A gentle introduction.** *Foundations and trends® in machine learning*, **16**(4):494–591, 2023. 19



School of Information Technology

ISBN: 978-91-90123-03-4 (printed)
Halmstad University Dissertations, 2026

