ID N. 7





UNIVERSITÀ CAMPUS BIO-MEDICO DI ROMA

DEPARTMENT OF ENGINEERING

UNIVERSITÀ DEL PIEMONTE ORIENTALE

DEPARTMENT OF SCIENCE AND TECHNOLOGICAL INNOVATION

Italian National Ph.D. in Artificial Intelligence
Health and Life Sciences
XXXVIII Cycle

From Graphs to Reasoning in Medicine: The Role of Structured Knowledge in Deep Learning Systems

Supervisors

Portinale Luigi

Piovesan Luca

Chiocchetti Annalisa

Candidate

Dossena Marco

April, 2026

Abstract

Deep learning has become a central methodology in biomedical research, yet its expansion into medicine is shaped by a fundamental tension: the most powerful neural architectures presuppose large, clean, well-annotated datasets, while clinical practice, constrained by privacy regulations, the cost of expert annotation, and the rarity of many conditions, can rarely supply them. Medicine is, however, a domain of extraordinary opportunity in another respect: decades of clinical research have produced an exceptionally rich corpus of structured knowledge, encoded as ontologies, knowledge graphs, and decision pathways that articulate the very reasoning patterns that human clinicians follow. This thesis develops a research program centred on a single argument: that structured knowledge constitutes a principled and increasingly necessary inductive bias for deep learning systems applied to medicine, and that its role evolves along a coherent trajectory of increasing structural commitment. At the first stage, knowledge is implicit, and the burden of generalization falls entirely on data; here, data-centric techniques can extend what neural models achieve in scarce-label regimes, but they simultaneously reveal a structural ceiling that no amount of statistical refinement can transcend. At the second stage, knowledge becomes an explicit inductive bias, embedded into neural architectures so as to align their representational space with the functional organization of the domain; the resulting models are not merely more accurate, but more clinically coherent, in that their predictions respect the priorities encoded in established medical protocols. At the third stage, knowledge becomes the primary object of learning: when deep models are trained directly on biomedical knowledge graphs with appropriately chosen geometric inductive biases, their latent spaces inherit, by virtue of algebraic closure, a compositional inferential capacity that exceeds the training distribution and supports multi-step queries that no single training example could justify in isolation. At the fourth stage, knowledge becomes a normative yardstick: structured decision pathways drawn from clinical sources are turned into a controlled benchmark against which the reasoning capabilities of state-of-the-art Large Language Models are measured, revealing that the structural, and not merely the informational, character of graph-based knowledge is what closes the gap between parametric recall and stepwise clinical inference. Read in sequence, the four stages describe

a single trajectory in which structured knowledge plays a progressively more central role, and in which a recurring principle emerges across very different methodological settings: that structure, more than scale, is the operative variable in domains characterised by limited data, strong prior organization, and high requirements of robustness, interpretability, and reasoning fidelity. By tracing this trajectory from small clinical cohorts to large-scale biomedical knowledge graphs, and ultimately to the evaluation of how generative models reason when confronted with the very structured pathways that human clinicians follow, the thesis offers a unified perspective on how the next generation of biomedical artificial intelligence can be designed to be simultaneously data-efficient, robust, interpretable, and capable of reasoning in the structured terms that medicine itself demands.

Contents

List of Acronyms **17**

1 Introduction **20**
1.1 Motivation . . . 20
1.2 Research questions . . . 24
1.3 Thesis contribution . . . 27
1.4 Thesis outline . . . 29
1.5 List of publications . . . 32
1.5.1 Peer-reviewed publications . . . 32
1.5.2 Preprints under review . . . 33
1.5.3 Work in progress . . . 33
1.5.4 Other publications . . . 33

2 Machine Learning in Biomedicine: The Data Scarcity Problem **35**
2.1 Reasons for data scarcity and related issues . . . 36
2.1.1 Data Scarcity in the Biomedical Context: Beyond Volume . . . 36
2.1.2 Consequences for Deep Learning Models . . . 37
2.1.3 The Problem-Dependent Nature of Data Scarcity . . . 38
2.2 Predicting Long-COVID Sequelae: Data Scarcity in Clinical Data . . . 42
2.2.1 Problem definition . . . 42
2.2.2 The Multi-Label Classification Framework . . . 45
2.2.3 Dataset Analysis and Augmentation Strategies . . . 49
2.2.4 Experimental framework . . . 52
2.2.5 Classification results and considerations . . . 53
2.2.6 Interpretability Analysis . . . 56
2.2.7 Conclusions on data-driven approach experiment . . . 59
2.3 Beyond Labels: A Self-Supervised Approach in Histopathology . . . 61
2.3.1 Problem definition . . . 62

2.3.2 HMAE framework . . . 63
2.3.3 Experimental framework & results . . . 66
2.3.4 Generalization capabilities & qualitative evaluation . . . 70
2.3.5 Conclusions on self-supervised approach experiment . . . 74
2.4 Lessons from the Two Case Studies . . . 75

3 Knowledge Integration in Deep Learning Models 78
3.1 Inductive Bias and the Necessity of Prior Knowledge . . . 79
3.1.1 Architectural and Relational Inductive Biases in Neural Networks . . 80
3.2 Inductive Bias in Transformers . . . 84
3.2.1 The Inductive Bias Profile of Transformers . . . 87
3.3 Knowledge Integration in Transformers for Biomedical Applications . . . 88
3.3.1 The Spectrum of Integration Approaches . . . 90
3.3.2 Beyond Transformers: Alternative Paradigms of Informed Machine Learning . . . 92
3.3.3 Positioning of the Proposed Approach and Motivation for an Additional Inductive Bias . . . 94
3.4 Structural Positional Encoding for Knowledge Integration in Medical Process Monitoring . . . 97
3.4.1 Medical Process Monitoring and Trace Classification . . . 98
3.4.2 Structural Positional Encoding: Embedding Structured Knowledge . 100
3.4.3 Integration into the Transformer . . . 103
3.5 Experimental Validation: Stroke Management . . . 106
3.5.1 Next Activity Prediction Results . . . 109
3.5.2 Trace Classification Results . . . 112
3.5.3 External Validation on a Non-Clinical Process: BPI Challenge 2012 . 114
3.6 Conclusion: Why Structured Knowledge Matters . . . 116

4 Learning on Knowledge Graphs 121
4.1 Foundations of Graph-Based Knowledge Representation . . . 123
4.1.1 From Homogeneous to Heterogeneous Networks . . . 123
4.1.2 Knowledge Graphs . . . 125
4.1.3 Relation to Clinical Knowledge Representations . . . 130
4.2 Deep Learning Paradigms for Knowledge Graphs . . . 132
4.3 Knowledge Graph Embeddings: From TransE to Modern Approaches . . . 133
4.4 Graph Neural Networks for Relational Learning . . . 139
4.4.1 Message Passing Neural Networks: The General Framework . . . 140

4.4.2 Relational Graph Convolutional Networks . . . 141
4.5 Open-World vs. Closed-World Assumptions: Implications for Learning . . . 143
4.5.1 False Negative Contamination in Training Data . . . 143
4.5.2 Strategies for Mitigating False Negatives . . . 145
4.6 Experimental Validation: Knowledge Graph Completion . . . 146
4.6.1 Experimental Setup and Datasets . . . 146
4.6.2 Knowledge Graph Completion Results . . . 151
4.7 Compositional Query Answering in Latent Space . . . 157
4.8 From Knowledge Representation to Emergent Compositional Behavior . . . 167
4.8.1 Summary of Empirical Findings . . . 167
4.8.2 Theoretical Implications . . . 168
4.8.3 Biomedical Significance . . . 170
4.8.4 Limitations and Future Directions . . . 170

5 Measuring LLM Reasoning through Clinically Grounded Decision Pathways 173
5.1 Existing benchmarks . . . 174
5.2 Proposed Methodology . . . 176
5.3 Evaluation Protocol . . . 179
5.4 Experiments and Findings . . . 181
5.4.1 Dataset Characteristics . . . 181
5.4.2 Experiments on Dataset Creation . . . 182
5.4.3 Benchmark Results and Reasoning Analysis . . . 183
5.5 Conclusion . . . 186

6 Conclusions and Future Directions 188
6.1 The trajectory of structured knowledge: a synthesis . . . 188
6.2 Methodological reflections . . . 192
6.3 Limitations . . . 194
6.4 Future directions . . . 195
6.5 Closing remarks . . . 196

7 Computational Cost and Reproducibility 198
7.1 Reporting Conventions . . . 198
7.2 Chapter 2 (i): Long-COVID Multi-Label Classification . . . 199
7.3 Chapter 2 (ii): HMAE (Histopathological Masked Autoencoder) . . . 199
7.4 Chapter 3: SPE-Augmented Transformer for Stroke Process Monitoring . . . 201

7.5 Chapter 4: UMLS Knowledge Graph Embedding Models 201
7.6 Chapter 5: HealthBranches LLM Benchmark 203

A HealthBranches: Supplementary Material 226
A.1 Dataset Details 226
A.2 Extract Option Rule Set 229
A.3 Q&A Validation 230
A.4 Quiz Inference 233
A.5 Open Answer Inference 235
A.6 Inference Times Details 236
A.7 Results for Different Modalities 236
A.8 LLM-as-a-Judge Implementation 236
A.9 Experimental and Hardware Setup 238
A.10 Medical Validation Protocol 239
A.11 Web App Review Examples and Final Results 241

List of Figures

2.1 Dataset distribution 49
2.2 Concurrence plots: `orig` (left) and `k3U` (right) 51
2.3 SHAP values for the original dataset using a Random Forest model 57
2.4 SHAP values for arthalgia at follow-up. 58
2.5 SHAP values for diarrhea at follow-up. 58
2.6 Comparison of SHAP values between the original dataset and different augmentations. The SHAP values of each dataset have been normalized independently. 59
2.7 **HMAE architecture**. (top) Tissue regions are randomly sampled from the original image (WSI). Subsequently, a random mask is applied, occluding 75% of the image data. (down) The autoencoder architecture receives the masked image as input and aims to reconstruct the hidden regions. 65
2.8 Class taxonomy of BRACS dataset. 66
2.9 t-SNE visualization of the region of interest (RoI) embeddings from the BRACS dataset. Each point represents a sample, colored according to its class: benign (blue), malignant (red), and atypical (green). This visualization highlights the clustering of different histopathological categories based on their learned feature representations. 73
2.10 Self-attention heatmaps over histopathology images. The left column in each pair shows the original histopathology image, while the right column presents the corresponding heatmap, where warmer colors (red, yellow) indicate regions of high attention. 73

3.1 Excerpt of the domain ontology. 102
3.2 The model architecture. The left side of the figure represents the transformer-decoder model, the right side shows the Structural Positional Encoding module 104
3.3 Example of a trace. The activities $ and # represent the start and the end of the trace, respectively. 106

3.4 Trace length histogram 107
3.5 Trace given as a query to the transformer without using domain knowledge, with the predicted next activity colored in red. 112
3.6 Trace given as a query to the transformer using SPE, with the predicted next activity colored in red. 112

4.1 Structure of a q_{2p1} conjunctive query. Two anchor entities (e_1, e_2) are projected through their respective relations (r_1, r_2) to identify an intermediate entity v that satisfies both constraints simultaneously. A third relation r_3 is then applied to v to reach the final answer a. The biomedical example illustrates a query that composes the *treats* and *has_symptom* relations with *associated_with* to perform multi-hop inference across the UMLS knowledge graph. 159
4.2 Geometric interpretation of q_{2p1} query solving in RotatE's complex latent space. Anchor entities $\mathbf{e}_1$ and $\mathbf{e}_2$ are rotated by their respective relation embeddings $\mathbf{r}_1$ and $\mathbf{r}_2$ to produce projected points $\mathbf{v}_1$ and $\mathbf{v}_2$. The intersection is approximated by the centroid $\mathbf{v}^*$, which is then rotated by $\mathbf{r}_3$ to yield the estimated answer location $\mathbf{a}^*$. Candidate entities are ranked by proximity to $\mathbf{a}^*$. 161
4.3 MRR comparison between single-hop knowledge graph completion and compositional q_{2p1} query answering, both using RotatE on UMLS. The 58.5% degradation reflects the accumulated approximation error across three sequential geometric operations in latent space. 163
4.4 Hits@k comparison between single-hop KGC and compositional q_{2p1} query answering. While the gap is most pronounced at $k=1$ (0.659 absolute difference), it narrows rapidly: by $k=25$, the compositional model recovers to within 8.1 percentage points of single-hop performance. 164
4.5 Performance degradation analysis from single-hop to compositional queries. The absolute gap (bars) and relative degradation percentage (line) both decrease monotonically with k. The disproportionate drop at $k=1$ (87.5% relative degradation) compared to $k=25$ (8.5%) reveals that compositional error primarily affects top-rank precision while preserving broader retrieval quality. 165

4.6 Summary of experimental findings. (a) Performance comparison between single-hop knowledge graph completion and compositional q_{2p1} query answering using RotatE. MRR exhibits a 58.5% relative degradation, while Hits@10 degrades by only 10.4%, indicating that compositional error primarily affects top-rank precision. (b) Model comparison on single-hop KGC, with RotatE achieving the highest MRR among all six models. GNN-based augmentation yields divergent effects depending on the scoring function. 169

5.1 A step-by-step workflow for building the HealthBranches Q&A dataset. *Parsing the knowledge source*: textual and graphical streams are extracted from medical books and parsed. *Path extraction and refinement*: graph-based representations are processed to extract and refine relevant semantic paths. *Q&A generation*: an LLM is prompted with textual and path information to generate initial question–answer pairs. *Q&A refinement*: the dataset is further refined using another LLM to ensure consistency and quality. 177

5.2 Detailed example of the Q&A generation process, starting from the extraction of a reasoning path. The left side shows an example graph outlining treatment steps for dyspnea, while the right side displays the extracted reasoning path and the resulting Q&A pair. The correct answer (option 5, Pulmonary function testing) is derived directly from the structured path, demonstrating how graph-based clinical knowledge is translated into a benchmark item. . . 178

5.3 Improved accuracy across models after the Q&A refinement process. 184

5.4 Model performance on HealthBranches. Top: quiz accuracy across settings. Middle: G-eval scores for open-ended answers. Bottom: cosine similarity between predictions and ground truth. The bottom plot's y-axis is truncated for visual clarity, as cosine similarity measures relative semantic alignment rather than absolute distance from zero. Error bars show 95% confidence intervals across all plots. 185

A.1 Main types of corrections after the Q&A refinement process. 233

A.2 Examples of reviews conducted by a medical specialist and a medical student receiving at least one score of 3 or lower. 241

List of Tables

2.1 Instantiation of the four dimensions of data scarcity on the two case studies of this chapter. . . . 41
2.2 Summary of the input feature space of the long-COVID predictive model. Features are grouped by semantic block; representative examples are reported for each group, the complete variable list being available in the clinical companion studies [17, 19]. . . . 43
2.3 Main features of considered datasets; n: number of instances, #ls: cardinality of label-set, #pl: number of label-set with more than one label . . . 50
2.4 Accuracy (Jaccard Index) . . . 55
2.5 Exact Match . . . 55
2.6 Hamming Score . . . 55
2.7 Area under ROC (macro-averaged) . . . 56
2.8 Cancer Classification (3 classes): AUC and Weighted F1-score . . . 68
2.9 Cancer Classification (3 classes): statistical difference (95% confidence) . . . 68
2.10 Sub-type cancer classification: weighted F1-score . . . 70
2.11 Sub-cancer Classification (7 classes): p-values at 95% confidence . . . 70
2.12 Single class classification F1-score (based on top-3 models in Table 2.10). . . . 70
2.13 Single-class performance: p-values between HMAE and top-2 performers . . . 71
2.14 Testing Results on the BACH dataset (weighted F1-score). . . . 71
2.15 BRACS to BACH test: p-value . . . 71

3.1 Positioning of Structural Positional Encoding within the space of Informed Machine Learning approaches that inject prior knowledge during learning, following the integration-point axis of von Rueden et al. [156]. . . . 97
3.2 Parameter search space used to find optimal hyperparameters and optimal configuration, by Optuna. Search spaces included between square brackets are discrete sets, while for continuous search spaces the values are sampled by Optuna space in that range. . . . 108

3.3 Test accuracy-at-k and standard deviation on 10 random initialization and train-val-test splits. Model performance with different positional encoding methods at different embedding sizes are reported 109
3.4 Test accuracy-at-k and standard deviation on 10 random initialization and train-val-test splits. Model performance with 2 different noisy datasets. The models are trained using the SPE method. 111
3.5 Results of the trace classification task. We compare the transformer model with SPE embedding to the CNN-based model, the LSTM-based model and the random forest model. All the metrics are weighted in order to account for dataset imbalance 113
3.6 Next activity prediction on the BPI Challenge 2012 event log, with and without Structural Positional Encoding. Model size fixed at 32, experimental protocol identical to the stroke-management experiments of Table 3.3. 115

4.1 Relational patterns captured by different knowledge graph embedding models. (✓✓ indicates particular strength, ✓ indicates capability, × indicates limitation.) 139
4.2 Knowledge graph completion results on UMLS (2M filtered ranking queries). Best results per column are shown in **bold**. 152
4.3 Effect of R-GCN aggregation: difference (Δ) between hybrid and standalone models. Positive values (↑) indicate improvement; negative values (↓) indicate degradation. 155
4.4 Comparison of single-hop knowledge graph completion and compositional q_{2p1} query answering performance using RotatE on the UMLS knowledge graph. . 163

5.1 (✓= present, ✗= absent.) Summary of existing medical Q&A datasets and comparison with **HealthBranches**, considering four qualitative dimensions: (1) **Knowledge**: whether the dataset tests knowledge tied to a particular domain; (2) **Qualitative Reasoning**: whether the dataset tests logical or conceptual reasoning rather than quantitative computation; (3) **No Comput.**: whether the dataset does not require quantitative calculations, formulas, or numeric estimation; (4) **Explanation**: whether the dataset includes a step-by-step justification or reasoning trace. While these criteria are inherently qualitative, the classification follows the precedent set in MedCalc-BENCH [80]. 175
5.2 Distribution of questions and number of leaf nodes across clinical categories in the HealthBranches dataset. 182

7.1 Hardware platforms. The GPU nodes are high-performance nodes equipped with an Intel Xeon Gold CPU, 192 GB of system RAM and Ubuntu Linux. All experiments were run on a single accelerator; no distributed or multi-GPU training was employed. 199
7.2 Computational profile of the long-COVID multi-label classification study. . . 199
7.3 Structural complexity of the long-COVID MLC pipeline. Each base ensemble is a 200-tree Random Forest. CDN performs inference through Gibbs sampling over $I = 500$ iterations, so a single test instance requires $500 \times 7 = 3{,}500$ binary Random Forest evaluations, making it by two orders of magnitude the most expensive method at inference time despite an identical training footprint to BR or CC. 200
7.4 Computational profile of the HMAE study. 200
7.5 Trainable parameter budget of HMAE. Positional encodings are fixed sinusoidal buffers and are excluded. The decoder accounts for 54.5% of the parameters trained and is discarded after pre-training: only the 21.6 M-parameter encoder is retained to produce ROI embeddings, and the supervised stage updates solely the MLP head on top of those frozen embeddings. 201
7.6 Computational profile of the SPE study. 202
7.7 Trainable parameter counts for the next-activity prediction models, computed from the Optuna-selected configuration (feed-forward size 128, 4 heads, 4 layers, $k = 32$, vocabulary 85). Sinusoidal encodings are non-trainable, so the None, PE and SPE variants at a given d differ only by the $32d + d$ parameters of the structural projection: at $d = 128$ this is 1.0% of the model. 202
7.8 Computational profile of the UMLS knowledge graph completion study. . . . 203
7.9 Trainable parameter counts for the UMLS models at embedding dimension 512. Training a 1.07B-parameter embedding table with Adam requires approximately 4GB. 203
7.10 Measured training throughput for the UMLS models on the Quadro RTX 6000 node. The rate is governed by memory traffic over the 4 GiB embedding table and by the rejection sampling of type-constrained negatives, rather than by arithmetic: a single step performs only $O(|\mathcal{B}| \cdot d)$ arithmetic operations, which on this hardware would otherwise complete in a small fraction of a second. . 204
7.11 Computational profile of the HealthBranches benchmark. 204

7.12 Frozen parameter counts, quantized memory footprints and measured average per-response latency for the HealthBranches benchmark. Latency is largely independent of parameter count, as expected in a memory-bandwidth-bound autoregressive decoding regime; the outlier is DeepSeek-R1 8B, whose latency reflects the generation of reasoning tokens that are suppressed from the visible output but still produced. 205

A.1 Average inference time per response (in seconds) for quiz and open-ended answer settings. 237
A.2 Model performance for each modality (Zero-shot, RAG, etc.) in terms of Accuracy, Judge score, and Similarity. Scores highlighted in red denote the best-performing model (rank 1), while those in blue indicate the second-best performance (rank 2) for the corresponding metric. 237
A.3 Reviewer performance and feedback statistics by expertise level for randomly selected questions. 242
A.4 Coverage across HealthBranches categories. 242

List of Algorithms

1 Type-Constrained Negative Sampling for UMLS 150
2 Compositional Query Solver for q_{2p1} Queries in RotatE Latent Space 162
3 ExtractOption(answer) 229

List of Acronyms

ADH	Atypical Ductal Hyperplasia
AI	Artificial Intelligence
AUC	Area Under the ROC Curve
BACH	BreAst Cancer Histology
BCC	Bayesian Classifier Chain
BMI	Body Mass Index
BPI	Business Process Intelligence
BR	Binary Relevance
BRACS	BReAst Carcinoma Subtyping
CC	Classifier Chain
CDN	Conditional Dependency Network
CNN	Convolutional Neural Network
CT	Computed Tomography
CUI	Concept Unique Identifier
CVIR	Coefficient of Variation of the Imbalance Ratio per Label
CXR	Chest X-Ray
DCIS	Ductal Carcinoma In Situ
DL	Deep Learning
ECG	Electrocardiogram
EHR	Electronic Health Record
FEA	Flat Epithelial Atypia
FFN	Feedforward Network
FL	Federated Learning
GDPR	General Data Protection Regulation
GNN	Graph Neural Network
GRU	Gated Recurrent Unit

H&E	Hematoxylin and Eosin
HIN	Heterogeneous Information Network
HIPAA	Health Insurance Portability and Accountability Act
HMAE	Histopathological Masked AutoEncoder
ICD	International Classification of Diseases
ICU	Intensive Care Unit
IID	Independent and Identically Distributed
IRLbl	Imbalance Ratio per Label
KG	Knowledge Graph
KGC	Knowledge Graph Completion
KIML	Knowledge-Informed Machine Learning
LC	Label Combination
LLM	Large Language Model
LSTM	Long Short-Term Memory
MAE	Masked Autoencoder
MANN	Memory Augmented Neural Network
MAP	Mean Average Precision
MeanIR	Mean Imbalance Ratio
MIL	Multiple Instance Learning
MLC	Multi-Label Classification
MLP	Multi-Layer Perceptron
MLSMOTE	Multi-Label Synthetic Minority Over-sampling Technique
mMRC	modified Medical Research Council
MPNN	Message Passing Neural Network
MRI	Magnetic Resonance Imaging
MRR	Mean Reciprocal Rank
MS	Multi-Similarity (loss)
MSE	Mean Squared Error
NFL	No Free Lunch (theorem)
NLP	Natural Language Processing
OWL	Web Ontology Language
PCA	Principal Component Analysis
PE	Positional Encoding
PFT	Pulmonary Function Test

PINN	Physics-Informed Neural Network
PS	Pruned Set
Q&A	Question Answering
RAG	Retrieval-Augmented Generation
RAkEL	RAndom k-labEL Pruned Set
RDF	Resource Description Framework
R-GCN	Relational Graph Convolutional Network
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
ROI	Region of Interest
SC	Stroke Center
SCUMBLE	Score of ConcUrrence among iMBalanced LabEls
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Over-sampling Technique
SNOMED CT	Systematized Nomenclature of Medicine Clinical Terms
SPARQL	SPARQL Protocol and RDF Query Language
SPE	Structural Positional Encoding
SSL	Self-Supervised Learning
SU	Stroke Unit
SVM	Support Vector Machine
t-SNE	t-distributed Stochastic Neighbor Embedding
UDH	Usual Ductal Hyperplasia
UMLS	Unified Medical Language System
ViT	Vision Transformer
WF1	Weighted F1-score
WSI	Whole-Slide Image

Chapter 1

Introduction

1.1 Motivation

The intersection of artificial intelligence and biomedicine has, over the past decade, evolved from a frontier of academic curiosity into one of the most consequential research programs in modern science. What began as a series of isolated proofs-of-concept, such as a convolutional network trained to recognise diabetic retinopathy or a recurrent model predicting hospital readmission, has matured into an established sub-discipline with its own conferences, benchmarks, regulatory frameworks, and clinical trials. Deep learning, in particular, has shown a remarkable capacity to extract complex, non-linear patterns from high-dimensional biological signals, opening unprecedented opportunities for diagnosis, prognosis, and clinical decision support. From convolutional architectures that recognise malignancies in histopathological slides, to Transformers that interpret clinical narratives and structured electronic health records, to Large Language Models capable of generating differential diagnoses or summarising medical literature, the breadth of applications continues to expand at a pace that few other domains can match. The promise is no longer hypothetical: in narrow, well-circumscribed tasks such as image-based screening, signal-derived risk scoring, and automated documentation, neural systems already approach or, in some published studies, exceed the performance of trained human specialists.

Yet the very premise on which these advances rest, namely the availability of large, clean, well-annotated datasets, sits in direct tension with the reality of biomedical practice. The textbook recipe for state-of-the-art deep learning calls for millions of independent samples, dense and reliable labels, balanced class distributions, and a degree of homogeneity in acquisition protocols that is the norm in consumer datasets but the exception in medicine. Patient privacy regulations such as GDPR and HIPAA, the institutional fragmentation of clinical

records, the legal and ethical requirements surrounding informed consent, the inherent rarity of many conditions, and the cost of expert annotation combine to produce an environment in which data is simultaneously abundant in volume and effectively scarce in usable form. Hospitals possess massive archives of imaging studies, laboratory results, and free-text reports, yet the fraction of those archives that can be released, harmonised, labelled by qualified specialists, and curated to a research-grade standard remains modest. Furthermore, the labels themselves are expensive in a way that has no direct analogue in general-domain machine learning: while object recognition in natural images can be crowdsourced to thousands of non-expert annotators, delineating a tumour boundary on a magnetic-resonance volume or assigning a histopathological subtype to a Whole-Slide Image requires hours of attention from board-certified clinicians whose time is structurally limited and inherently scarce. Inter-observer variability further complicates matters: medicine is rarely binary, and even highly trained experts routinely disagree on borderline cases, producing a soft "noise floor" that no amount of statistical regularisation can eliminate.

The most powerful deep-learning models, designed under the implicit assumption of asymptotic data regimes, are therefore confronted in medicine with the opposite condition: small cohorts, imbalanced labels, heterogeneous acquisition devices, and an additional, strict requirement for interpretability and reliability that few other application domains impose. A model that misclassifies a frame in a video-recommendation pipeline produces an annoyance; a model that fails on a clinical decision can contribute to misdiagnosis, inappropriate treatment, or harm. This raises the bar from "performs well on average" to "performs reliably across patient subpopulations, hospitals, and acquisition devices, and supplies an explanation that a clinician can scrutinise and, if necessary, override". Each of these requirements is, on its own, demanding; their conjunction is one of the defining methodological challenges of contemporary biomedical AI. That this bar is rarely cleared is not a rhetorical claim but an empirically documented one, and the gap between what the literature produces and what obtains regulatory authorisation constitutes an independent motivation for the present work. Tens of thousands of papers proposing medical machine-learning algorithms have been published over the past decade, yet the number of systems that have completed a regulatory pathway remains comparatively modest: a systematic review identified 1,016 authorisations of AI/ML-enabled devices between 1995 and 2024, corresponding to only 736 distinct devices, of which 84.4% operate on images and merely three take tabular electronic health records as input [127]. The translational bottleneck, in other words, is not primarily one of algorithmic novelty. It is one of evidence: of models validated on a single site, on retrospective cohorts, and on populations that do not span the heterogeneity of those on which they would be deployed. These are precisely the failure modes that data scarcity produces, and they mo-

tivate the search for mechanisms that constrain a model's behaviour by means other than the accumulation of yet more training examples. This tension has motivated a fundamental reconsideration of what it means to learn in the biomedical setting. A first response, dominant for several years, has been to apply increasingly sophisticated data-centric techniques, such as augmentation, transfer learning from natural-image backbones, multi-task learning, and semi-supervised and self-supervised pre-training, in the hope of extracting more signal from the limited samples available. These techniques have produced genuine advances, and Chapter 2 of this thesis will argue, with empirical support, that they remain a necessary first line of defence.

A distinct and particularly influential branch of this data-centric response deserves separate mention, since it attacks the problem from the side of data access rather than data efficiency. Federated Learning (FL) inverts the conventional pipeline: instead of moving patient data to a central repository where a model is trained, the model is moved to the data, trained locally within each institution, and only parameter updates are exchanged and aggregated. In the biomedical setting this is an attractive proposition, because the barrier to large-scale training is frequently not the non-existence of records but their confinement in institutional and jurisdictional zones [120]. The paradigm has moved well beyond proof-of-concept: the EXAM study, for instance, trained a model predicting oxygen requirements in COVID-19 patients across twenty institutions on four continents without any exchange of patient records [39]. FL is therefore, alongside augmentation and self-supervision, one of the most relevant paradigms currently available for confronting biomedical data scarcity, and it is complementary to, rather than in competition with, the approaches developed in this thesis. It is nevertheless important to be precise about what FL does and does not solve, since this delimits its role in the present argument. FL addresses the *accessibility* dimension of scarcity: it enlarges the effective sample size available to a model without violating the privacy constraints that make centralization impossible. It does not address the *annotation* dimension, because labels remain as expensive to produce at each participating site as they were at one; nor does it address the *rarity* dimension, since a condition that is rare is rare in every silo, and federating a hundred hospitals multiplies rather than resolves the problem of long-tailed distributions.

Nevertheless, every such method shares the same fundamental property: its effectiveness is bounded by the information content of the data itself. When that information is genuinely insufficient, whether because the cohort is small, because rare disease subtypes are absent, or because annotation noise dominates weak signals, no amount of architectural cleverness can manufacture knowledge that the data does not contain. The structural ceiling here is not an engineering inconvenience but a theoretical limit, formalised in machine learning by

the No Free Lunch theorem: in the absence of assumptions about the structure of the problem space, no learning algorithm can be expected to generalize. Effective learning requires inductive bias, that is, a set of prior assumptions about which hypotheses are plausible. The question, in biomedicine, is therefore not whether to use prior knowledge, but where that knowledge will come from and how it will be made operational. Medicine is, in this respect, a domain of extraordinary opportunity. Decades, and in some specialties centuries, of clinical research have produced an exceptionally rich corpus of structured knowledge: ontologies such as SNOMED CT and the Unified Medical Language System describing millions of medical concepts and their relationships; clinical guidelines and decision pathways that codify the consensus reasoning of expert panels; pharmacological databases capturing drug-drug interactions and contraindications; anatomical hierarchies; disease taxonomies; and care protocols refined through countless iterations of clinical practice. Unlike many other fields where machine learning is applied, the question is not whether such structure exists, but how to render it intelligible to neural systems whose default mode of operation is statistical association over raw data. A Transformer trained from scratch on a small clinical cohort has no built-in understanding that "myocardial infarction" and "heart attack" denote the same concept, that electrocardiography typically precedes thrombolysis in acute stroke management, or that a drug contraindicated for renal impairment should not be recommended for a patient with stage-IV kidney disease. All of this knowledge is, however, available, encoded in standardised vocabularies, decision graphs, and practice guidelines that were produced precisely so that human clinicians could share, transfer, and refine it. The methodological challenge is to construct learning architectures that can leverage these external resources, not as post-hoc rationalisations of their predictions, but as constituent elements of the learning process itself. This challenge has acquired renewed urgency with the arrival of Large Language Models. The latest generation of generative models exhibits, on certain medical benchmarks, performance levels that would have been considered science-fictional a decade ago: passing scores on licensing-style examinations, plausible differential diagnoses generated from free-text vignettes, and an apparent fluency with medical terminology that rivals that of trained clinicians. These results have led to widespread enthusiasm and to an equally widespread discomfort: the same models that produce confident, articulate, and sometimes correct clinical reasoning also hallucinate, fabricate citations, and exhibit failure modes that are difficult to characterise systematically. The community has begun to recognise that what is being measured by current medical benchmarks is, in large part, factual recall against a slightly modified version of the medical literature on which the model was trained, and that genuine clinical reasoning, namely the structured, multi-step navigation of a decision pathway from presenting symptom to recommended action, is far harder to evaluate, and almost certainly

far less developed in current models than aggregate accuracy figures suggest. Distinguishing memorisation from reasoning, and quantifying the gap between them, has become one of the open problems of medical AI.

This thesis develops a research program grounded in the convergence of these three observations: that purely data-driven solutions face a structural ceiling in biomedicine; that the field is uniquely rich in structured prior knowledge that has yet to be fully exploited by deep-learning systems; and that the rise of LLMs has made the question of what counts as reasoning a practical and measurable concern. Its central argument is that structured knowledge, encoded as ontologies, knowledge graphs, or decision pathways, constitutes a principled and increasingly necessary inductive bias for deep learning systems applied to medicine, and that its role evolves along a coherent trajectory: from a remedy against scarcity, to an architectural constraint, to an object of learning, and finally to a yardstick against which the reasoning capabilities of generative models are measured. The work begins by characterising the limits of purely data-driven solutions in two complementary biomedical scenarios, and then progressively builds toward architectures and benchmarks in which structural knowledge is embedded directly into the learning process. In doing so, it traces an arc from small clinical cohorts to large-scale biomedical knowledge graphs, and ultimately to the evaluation of how Large Language Models reason, or fail to reason, when confronted with the very structured pathways that human clinicians follow.

1.2 Research questions

The progression from data-driven mitigation to knowledge-grounded reasoning described in the previous section is articulated through four research questions, each addressed in a dedicated chapter of this dissertation. The questions are deliberately ordered along a trajectory of increasing structural commitment: the first treats knowledge as implicit and seeks to compensate for its absence through statistical means; the second injects structured knowledge as an inductive bias inside an otherwise standard architecture; the third makes the knowledge graph itself the primary object of learning; and the fourth turns the structured pathways extracted from medical sources into a diagnostic instrument for measuring the reasoning capabilities of modern LLMs.

RQ1. *Limits of data-driven solutions in biomedical machine learning.* To what extent can purely data-driven techniques mitigate the data-scarcity problem in biomedical applications, and what are the structural limits of such approaches? This first question grounds the thesis in the empirical reality that motivates the rest of the program. Data scarcity in biomedicine is not a single phenomenon but a family of related conditions: a small absolute number of

samples; a small number of labelled samples in a sea of unlabelled raw data; severe class imbalance, with rare conditions dominated by common ones; and limited representational diversity, where a dataset that is large in volume nevertheless fails to span the heterogeneity of the underlying patient population. Each variant suggests different methodological responses, and the question therefore subdivides naturally into two complementary subquestions, each tackled through an original case study. The first asks whether oversampling and dimensionality-reduction techniques, combined with careful interpretability analysis, can produce a robust multi-label predictor from a small, high-dimensional clinical cohort, and is instantiated by the task of predicting long-COVID sequelae from acute-phase clinical data. The second asks whether self-supervised representation learning can decouple feature extraction from expert annotation in a setting where labels are the binding constraint, and is instantiated by the task of breast-cancer subtype classification from histopathological Whole-Slide Images. Taken together, the two studies delineate both the practical reach and the theoretical ceiling of strictly data-driven mitigation, and provide the empirical motivation for the turn toward structured knowledge that defines the remainder of the thesis.

RQ2*. Knowledge as inductive bias in Transformer architectures.* Can structured domain knowledge, encoded as a graph or an ontology, be integrated into Transformer architectures so as to provide a useful inductive bias in data-limited biomedical scenarios? This second question is the natural successor of RQ1: once it has been established that data-driven techniques have a structural ceiling, the question becomes how to introduce non-data sources of generalization pressure into the learning process. The Transformer architecture is a particularly interesting target. Its near-complete absence of architectural inductive bias, in that every token can attend to every other token with no a priori notion of locality, hierarchy, or graph structure, is what enables its scalability and broad applicability, but it is also what forces it to re-derive, from data alone, the structural regularities of any specialised domain. In medicine, where many such regularities are already explicitly encoded in ontologies and decision graphs, this is doubly wasteful: the model spends its capacity learning patterns that domain experts could have specified directly. The question therefore concerns how, mechanically, graph-structured knowledge can be injected into the attention machinery of a Transformer, what kind of knowledge yields measurable improvements, and under what conditions the resulting architecture remains compatible with standard training pipelines. The investigation is concretized in the development of Structural Positional Encoding for medical process monitoring, in which ontological knowledge about clinical activities is embedded into the positional representations consumed by the attention mechanism, and is empirically validated on stroke management traces under both clean and noise-perturbed conditions, and over both next-activity prediction and trace-classification tasks.

RQ3. *Knowledge as the primary object of learning.* Do deep-learning models trained directly on biomedical knowledge graphs learn latent spaces whose geometry supports inference patterns beyond the single triples observed during training? The third question marks a shift in perspective. In RQ2, the knowledge graph was an auxiliary signal, computed offline and used to enrich a model whose inputs were sequences of clinical activities. In RQ3, the knowledge graph itself becomes the primary object of study: the model is trained to predict missing relationships, infer new connections, and ultimately answer multi-hop queries that no single triple in the training set could justify in isolation. This shift introduces theoretical questions that have no counterpart in the sequence-modelling setting: how do different geometric inductive biases, namely translations (TransE), bilinear products (DistMult, ComplEx), and rotations (RotatE), capture different relational patterns, and how does their geometry interact with neighbourhood-aggregation mechanisms such as those of Graph Neural Networks? Most importantly, it raises an empirical question whose answer has implications well beyond knowledge graphs: when a model is trained exclusively on single-hop facts, does its latent space nevertheless acquire properties that emergently support multi-hop, compositional inference? The investigation is concretised in a large-scale study on the UMLS knowledge graph, comprising more than two million biomedical concepts and 3.5 million relational assertions, in which single-hop knowledge-graph completion is systematically compared across embedding methods and GNN-based variants, and a closed-form compositional q2p1 query solver is constructed to probe the multi-hop behaviour of the learned latent space.

RQ4. *Structured guidance and the reasoning gap in Large Language Models.* Are current Large Language Models genuinely capable of clinical reasoning, and to what extent does the explicit availability of graph-structured decision pathways close the reasoning gap that persists in their parametric knowledge? The fourth and final question addresses the most prominent recent development in medical AI: the rise of LLMs as candidate clinical reasoners. Their headline-grabbing performance on multiple-choice medical examinations has fuelled both enthusiasm and concern; what it has not produced, until recently, is a principled methodology for separating what the model remembers from what the model can derive. Existing benchmarks, dominated by closed-form questions whose answers can be matched against the pre-training corpus, are systematically biased toward measuring factual recall. The question therefore is twofold: how can clinical reasoning be operationalised so that it can be measured rather than asserted, and to what extent does the structured, graph-shaped knowledge advocated in the preceding chapters, when made available to the model at inference time, close the gap between what LLMs know and what they can reliably reason about. The investigation is concretised in the construction of HealthBranches, a benchmark of 4,063 clinically grounded question-answer items derived from explicit decision pathways extracted

from medical textbooks across 17 domains. Crucially, every benchmark item is paired with the full reasoning chain from which it was derived, enabling a controlled comparison among four conditions: zero-shot answering, retrieval-augmented generation, supplementation with the unstructured textual description of the relevant clinical scenario, and supplementation with the structured graph-derived path itself. The asymmetry between the latter two conditions provides a direct empirical test of the central thesis claim, namely that the structural character of graph-based knowledge, and not merely the information it contains, is what unlocks reasoning capabilities in modern generative models.

These four questions are, deliberately, not independent. RQ1 establishes the limits of relying on data alone, motivating RQ2's introduction of structured knowledge as inductive bias. RQ2's success in turning a knowledge graph into a positional signal for a Transformer raises the question, addressed by RQ3, of whether such graphs can be learned directly by neural models, with inferential properties that go beyond their training distribution. RQ3's finding that suitably structured latent spaces support emergent compositional inference, in turn, sharpens the question, addressed by RQ4, of whether LLMs, the current state of the art in reasoning over natural language, exhibit comparable structural sensitivity when supplied with graph-derived pathways. Together, the four questions describe a single trajectory in which structured knowledge plays a progressively more central role: from a remedy against data scarcity, to an explicit inductive bias in neural architectures, to a learning target in its own right, and finally to a yardstick against which the reasoning capabilities of state-of-the-art LLMs can be measured.

1.3 Thesis contribution

The contributions of this thesis are both methodological and empirical, and span four complementary axes that map directly onto the research questions introduced above. To these four, a fifth, transversal contribution is added: the integrative perspective that organises the work into a single research program rather than a collection of independent studies.

A characterisation of data scarcity in biomedical machine learning, with two original case studies. The thesis argues, and supports through empirical evidence, that scarcity in biomedicine cannot be reduced to a deficit of samples: it is a multidimensional condition that includes annotation cost, demographic representativeness, and distributional coverage. This view is operationalised in two complementary case studies. The first proposes a multi-label classification framework for predicting long-COVID sequelae from a real-world clinical databank of over 300 patients, systematically comparing six MLSmote upsampling policies against supervised (RELIEF) and unsupervised (PCA) dimensionality-

reduction strategies, and assessing the impact of each on model performance via several MLC algorithms. A SHAP-based interpretability analysis demonstrates that augmentation does not distort the qualitative explanation of the resulting models, a non-trivial property given the reluctance with which clinicians adopt opaque predictive systems. The second case study introduces HMAE, a self-supervised Histopathological Masked AutoEncoder that learns Vision-Transformer representations directly from randomly cropped Whole-Slide Image regions, without recourse to manual annotation during the representation-learning stage. Evaluated on the BRACS dataset, HMAE reaches performance statistically comparable to fully supervised state-of-the-art pipelines, while its competitive cross-dataset transfer to BACH demonstrates the robustness of the learned representations under shifts of acquisition protocol and class taxonomy. Together, the two studies establish, with concrete numbers, both the reach and the structural ceiling of strictly data-driven mitigation strategies.

Structural Positional Encoding for knowledge-enhanced Transformers. The thesis develops SPE, a lightweight mechanism that encodes ontological proximity between clinical activities through Laplacian eigenvectors of an ontology graph and injects this signal into Transformer positional representations through a single learnable projection. The mechanism is deliberately minimalist (its entire contribution to the forward pass is one matrix multiplication followed by an element-wise addition), yet experimental validation on a stroke management dataset shows that SPE outperforms a standard sinusoidal-positional-encoding Transformer as well as CNN-based, LSTM-based, and Random Forest baselines on both next-activity prediction and trace classification. The advantage persists under input perturbation (5% and 10% noise on the test traces), and the resulting predictions align with clinical priorities as judged by domain experts. A revealing negative result accompanies these findings: sinusoidal encoding, designed for rigidly sequential data, occasionally degrades performance on medical processes, exposing a misalignment between the default inductive bias of Transformers and the functional, hierarchical structure of clinical pathways, and thereby justifying, beyond a marginal improvement in accuracy, the principled need for domain-aware positional schemes.

A large-scale empirical study of deep learning on UMLS, with emergent compositional inference. The thesis presents a comparative evaluation of TransE, DistMult, ComplEx, RotatE, and their R-GCN-augmented variants on the Unified Medical Language System, comprising more than two million biomedical concepts and 3.5 million relational assertions. Beyond standard knowledge-graph completion, the study introduces a closed-form compositional query solver for q2p1 conjunctive queries in RotatE's complex latent space. A central empirical finding is that RotatE, trained exclusively on single-hop triples, exhibits a non-trivial form of emergent compositional capacity: on 88,669 multi-hop queries it correctly

retrieves the answer entity within the top-10 predictions for over 83% of queries, despite never having been trained to compose relations. Further analysis identifies the centroid-based intersection approximation as the principal bottleneck, and the chapter draws explicit theoretical conclusions about the alignment between geometric inductive bias and relational structure, explaining, in particular, why GNN-based augmentation helps TransE but harms RotatE, and why this asymmetry has implications well beyond UMLS.

HealthBranches: a path-grounded benchmark for medical reasoning in LLMs. The thesis introduces HealthBranches, a benchmark of 4,063 question-answer items derived from explicit decision pathways extracted from medical textbooks across 17 clinical domains. The benchmark supports both multiple-choice and open-ended formats, and pairs every item with the full reasoning chain from which it was generated. A controlled evaluation of eleven LLMs under zero-shot, retrieval-augmented, topline-text, and topline-path conditions provides a quantitative measure of the reasoning gap separating current LLMs from structured clinical inference: providing the explicit graph-derived path lifts the best model from 75.8% to 92.7% accuracy, whereas providing the same information as unstructured text yields only marginal gains. This asymmetry constitutes direct empirical evidence that the structural, rather than merely informational, character of graph-based knowledge is what drives reasoning improvements in modern LLMs. The benchmark is released together with an extensible template that allows the community to add new sources, from textbooks to society guidelines to institutional care pathways, provided that they can be represented in the same node-edge format.

A unifying conceptual contribution. Beyond these four axes, the thesis articulates an integrative perspective in which data-driven mitigation, knowledge-as-inductive-bias, knowledge-as-learning-target, and reasoning-under-structured-guidance are not disjoint research themes but successive layers of a single program. In each layer, structured knowledge acquires a more prominent role: first as a remedy, then as an architectural constraint, then as an object of learning, and finally as a benchmark against which generative models are measured. The chapters can therefore be read either as four self-contained studies, each with its own methodology, dataset, and evaluation, or as a coherent argument about the role that graph-shaped knowledge can, and increasingly must, play in the next generation of biomedical AI systems.

1.4 Thesis outline

The remainder of this dissertation is organised as follows.

Chapter 2. Machine Learning in Biomedicine: The Data Scarcity Problem. This chapter examines the data-scarcity problem in biomedicine and develops two complementary case studies. After a discussion that disentangles the several distinct conditions usually conflated under the label "scarcity" (limited volume, limited labels, limited representativeness, and limited distributional coverage), and reviews their consequences for deep learning models (overfitting, poor generalization, algorithmic bias), the chapter turns to its first case study: a multi-label classification framework for predicting long-COVID sequelae from a small, high-dimensional clinical cohort, augmented through MLSmote, contracted through PCA and RELIEF, and analyzed via SHAP. The second case study proposes the HMAE framework, a self-supervised pipeline based on Vision Transformers and Masked Autoencoders that learns histopathological representations directly from unlabelled Whole-Slide Images, evaluated on breast-cancer subtype classification on BRACS and on cross-dataset transfer to BACH. Together, the two studies establish both the practical impact and the structural limits of strictly data-driven solutions, motivating the turn toward knowledge-informed approaches developed in the subsequent chapters.

Chapter 3. Knowledge Integration in Deep Learning Models. This chapter introduces the perspective of Knowledge-Informed Machine Learning. It first reframes the problem of generalization through the lens of inductive bias, drawing on the No Free Lunch theorem and on the relational and architectural biases of contemporary deep-learning models. It then characterises the inductive bias profile of Transformers, and in particular the sense in which their near-absence of architectural prior is both their greatest strength and their principal liability in specialised domains. A focused survey of representative knowledge-integration strategies (SapBERT, KeBioLM, and K-BERT) positions these along a spectrum that ranges from input-level injection to deep architectural restructuring. The chapter culminates in the development and experimental validation of Structural Positional Encoding for medical process monitoring, with an extended evaluation on a stroke management dataset covering both next-activity prediction and trace classification, and a qualitative analysis of the clinical coherence of the resulting predictions.

Chapter 4. Learning on Knowledge Graphs. This chapter shifts perspective from knowledge as an auxiliary signal to knowledge as the primary object of learning. It develops the formal foundations of graph-based knowledge representation, from homogeneous graphs to heterogeneous information networks and ontology-grounded knowledge graphs, and surveys the two main paradigms of deep learning over such structures: Knowledge Graph Embeddings (TransE, DistMult, ComplEx, RotatE) and Graph Neural Networks (with particular attention to R-GCN). After a dedicated discussion of the open-world assumption and its consequences for negative sampling, including the false-negative contamination problem and

the type-constrained sampling strategy adopted in the experiments, the chapter presents a systematic empirical study on UMLS, followed by a closed-form analysis of compositional q2p1 query answering in RotatE's complex latent space. The chapter concludes by drawing theoretical and practical implications from the emergent compositional behaviour observed in the experiments, and by discussing its biomedical significance for tasks such as drug-disease association discovery and hypothesis-guided literature exploration.

Chapter 5. Measuring LLM Reasoning through Clinically Grounded Decision Pathways. This chapter turns to the most articulated form of reasoning examined in this thesis: the linguistically explicit, step-by-step deductions expected from Large Language Models. It positions the proposed HealthBranches benchmark relative to the existing landscape of medical question-answering datasets, including HeadQA, MedQA, MeDiaQA, PubMedQA, MedCalc-BENCH, MedXpertQA, and PathVQA, describes the semi-automated pipeline through which decision pathways are extracted from medical textbooks and refined into 4,063 Q&A items, and reports a comparative evaluation of eleven LLMs under zero-shot, retrieval-augmented, topline-text, and topline-path conditions. The asymmetry between structured and unstructured context is analyzed as a quantitative measure of the reasoning gap that, in different forms, motivates the entire thesis.

Chapter 6. Conclusions and Future Directions. This chapter draws the four investigations together into a single argument. It first retraces the trajectory of structured knowledge across the dissertation, from the structural ceiling encountered by purely data-centric mitigation in Chapter 2, through the correction of a misaligned architectural prior in Chapter 3 and the elevation of the knowledge graph to primary object of learning in Chapter 4, to the use of decision pathways as a normative criterion in Chapter 5. It then makes explicit three methodological reflections that apply with equal force to all four research questions: that structure has repeatedly proved a more efficient lever than scale within the regimes considered; that interpretability has been treated throughout as a property designed into the system rather than recovered after the fact; and that the design of evaluation protocols capable of capturing the relevant dimensions of reliability is itself a contribution of comparable importance to the design of the models. The chapter closes by stating the limitations of each study, together with the concrete mitigation strategies that would address them, and by identifying four directions for future work, among them the joint optimisation of ontology-derived positional representations with knowledge-graph embeddings, and a systematic characterization of the trade-off between structural priors and scale.

Chapter 7. Computational Cost and Reproducibility. This chapter consolidates the computational profile of every model trained or evaluated in the experimental chapters, so that the hardware budget required to replicate any of the studies can be sized in advance. Each

study is described through a homogeneous set of entries covering the hardware platform, the software framework, the scale of the data, the model configuration, the number of trainable parameters, the training workload and time, and the per-sample inference cost, with the conventions governing these quantities stated at the outset.

A final note on terminology is in order before proceeding. In the title of this thesis, and throughout its chapters, the word reasoning is used in its broadest sense, encompassing both the inferential operations carried out by neural models over graph-structured knowledge (Chapters 3 and 4) and the step-by-step, linguistically articulated reasoning probed in Large Language Models through the benchmark presented in Chapter 5. Throughout the technical development, the two uses are kept terminologically distinct: inference is preferred when referring to mechanical operations in latent spaces, and reasoning is reserved for cases, primarily in the LLM context, where the model is expected to produce or follow an explicit chain of deductions in natural language.

1.5 List of publications

The research carried out during the PhD has led to the publications listed below. The first group collects the works on which this thesis is directly based, with an indication of the chapter in which each contribution is developed and of its current publication status; the second group collects further work produced during the program but not included in the present dissertation.

1.5.1 Peer-reviewed publications

1. **M. Dossena**, C. Irwin, L. Piovesan, L. Portinale.
   Predicting Long-COVID Sequelae: a Multi-Label Classification Approach.
   Intelligenza Artificiale, 2025.
   doi.org/10.1177/17248035251317937 **[Chapter 2, Section 2.2]**

2. **M. Dossena**, A. Chiocchetti, C. Irwin, L. Portinale.
   Beyond Labels: A Self-Supervised Framework with Masked Autoencoders and Random Cropping for Breast Cancer Subtype Classification.
   Proceedings of the AAAI Symposium Series, vol. 5(1), pp. 10–17, 2025.
   ojs.aaai.org/index.php/AAAI-SS/article/view/35546 **[Chapter 2, Section 2.3]**

3. **M. Dossena**, C. Irwin, G. Leonardi, S. Montani.
   Structural Positional Encoding for Knowledge Integration in Transformer-based Medical Process Monitoring and Trace Classification.
   Progress in Artificial Intelligence, Springer, 2024.
   doi.org/10.1007/s13748-024-00356-7 **[Chapter 3]**

1.5.2 Preprints under review

1. C. Cosentino, A. Defilippo, **M. Dossena**, C. Irwin, S. Joubbi, P. Liò.
   HealthBranches: Synthesizing Clinically-Grounded Question Answering Datasets via Decision Pathways.
   arXiv:2508.07308, 2025.
   arxiv.org/abs/2508.07308 **[Chapter 5]**

1.5.3 Work in progress

1. *Compositional Query Answering in Latent Space*
   [Chapter 4]

Publication status of the contributions. The studies presented in Chapters 2 and 3 have appeared in peer-reviewed venues. The HealthBranches benchmark of Chapter 5 has been released as a preprint. The knowledge-graph study of Chapter 4 is the most recent of the four investigations and its results are being prepared for submission; the material is therefore presented here in its partial form for the first time. This asymmetry reflects the chronological order in which the studies were carried out rather than a difference in the maturity of the underlying results.

1.5.4 Other publications

The following works were produced during the doctoral program and, while not part of the argument developed in this thesis, contributed to the background in graph representation learning on which Chapters 3 and 4 build.

1. **M. Dossena**, C. Irwin, L. Portinale.
   Graph-based Recommendation using Graph Neural Networks.
   21st IEEE International Conference on Machine Learning and Applications (ICMLA), 2022.
   ieeexplore.ieee.org/document/10068865

2. L. Aceto, **M. Dossena**, P. Grassi, C. Irwin, S. Montani, L. Portinale.
   Graph Neural Diffusion: an Approach Based on Diffusion PDE.
   AIxIA Workshop on Machine Learning and Data Mining, 2023.
   iris.uniupo.it/handle/11579/166643

3. **M. Dossena**, C. Irwin.
   Autoencoder with Dual Loss Function for Follow-up Estimation.
   AIxIA Doctoral Consortium, CEUR-WS vol. 3670, 2023.
   ceur-ws.org/Vol-3670/paper80.pdf

Chapter 2

Machine Learning in Biomedicine: The Data Scarcity Problem

Deep Learning (DL) has emerged as a cornerstone of modern biomedical research, offering unprecedented capabilities for analysis, interpretation, and prediction from complex, high-dimensional biological data. Its distinctive power lies in the ability to automatically identify intricate, non-linear patterns that often elude traditional statistical methods. This allows researchers to reframe complex biological and clinical challenges into well-defined, data-driven computational problems, such as classification, regression or segmentation. However, the remarkable performance of these sophisticated, data-hungry models is fundamentally tethered to the availability of vast, high-quality, and annotated datasets. This strict requirement poses a significant bottleneck in the biomedical domain, which is pervasively afflicted by data scarcity. This scarcity arises from unavoidable practical challenges, including the prohibitive costs of data acquisition and expert annotation, stringent patient privacy regulations, and the inherent rarity of specific diseases, which makes gathering large cohorts exceptionally difficult. The consequence of this data gap is a critical risk: models may fail to generalize to new patients, perpetuate underlying data biases, and ultimately hinder their translational potential from the research bench to the clinical bedside.

This chapter investigates this critical data scarcity problem, arguing for the necessity of data-driven solutions designed specifically to enhance model robustness and reliability in these common data-limited scenarios. Two complementary strategies will be analyzed and discussed that confront this challenge from two distinct but synergistic angles. The first explores a data-level enhancement approach, utilizing multi-label upsampling (MLSmote) [32] and feature analysis to improve the prediction of Long-COVID sequelae [18]. This method

addresses the challenge of building a reliable classifier from a novel, small, and inherently imbalanced clinical trial dataset. The second investigates a feature-level solution, presenting a self-supervised framework based on Masked Autoencoders and random cropping [61]. This approach aims to learn powerful and generalizable data representations for breast cancer subtype classification directly from unlabeled images. By capturing the underlying tissue morphology, it bypass the need for large-scale, manually annotated datasets.

2.1 Reasons for data scarcity and related issues

The rise of deep learning has sparked a significant shift across various scientific fields, with biomedical research emerging as a particularly promising area for its application. The most effective machine learning models are extremely data-hungry, necessitating vast amounts of high-quality, annotated data for effective training. Yet, the biomedical field is inherently data-constrained by its nature and design [149]. The tension between the demands of algorithms and the realities of the domain results in the issue of data scarcity, a challenge that extends beyond simply having insufficient data points. It encompasses a complex array of limitations related to data volume, variety, quality, and accessibility, representing a significant barrier to translating artificial intelligence from promising research prototypes into robust, reliable clinical tools. Models are likely to suffer from severe failures, including **low predictive accuracy**, **biased training**, and difficulties in **generalizing to new patient** populations.

2.1.1 Data Scarcity in the Biomedical Context: Beyond Volume

At its core, data scarcity in machine learning is a condition where the available data is insufficient to train a model that can achieve the desired level of performance and, crucially, generalize its learned patterns to unseen, real-world data. This insufficiency can manifest in several ways: a **limited absolute volume of data**, a complete **lack of labeled examples** for supervised learning, or a **severe class imbalance** where data for a given label is vastly underrepresented compared to others. A more nuanced understanding reveals that raw data volume is often a misleading metric of data availability. Many healthcare institutions possess massive archives of electronic health records and medical images, suggesting a state of data abundance. Yet, the vast majority of this data is unstructured, uncurated, non-standardized, and closed behind privacy and institutional barriers [83]. The process of transforming this raw data into a clean, labeled, and analysis-ready format is so resource-intensive that the volume of *usable* data remains critically low. This gap between the apparent volume of

raw data and the actual volume of usable data creates a condition of *effective scarcity*. Consequently, the problem is not merely that data does not exist, but that **the data that exists cannot be effectively utilized for machine learning**.
Therefore, a biomedical-specific definition of data scarcity must include:

- **Lack of High-Quality Annotations**: Beyond the simple presence of labels, scarcity in the biomedical domain is often a scarcity of reliable, expert-derived annotations. Labeling medical data, such as segmenting a tumor in an MRI scan, requires domain expertise from certified clinicians and is a costly, time-consuming bottleneck.

- **Lack of Representative Diversity**: A dataset may be large in volume but scarce in diversity. If it fails to represent the heterogeneity of the human population across demographics (age, gender, ethnicity), disease subtypes, stages of progression, and co-morbidities, it is functionally scarce. Models trained on such data will fail to generalize and may perpetuate health inequities [7].

- **Lack of Data Representing the Full Distribution**: Ultimately, the performance of a machine learning model depends less on the absolute size of the dataset and more on how well that dataset represents the true underlying distribution of the problem [5]. A small but highly representative dataset can yield a better model than a large but biased or incomplete one.

It is also critical to distinguish data **scarcity** from the related concept of data **sparsity**. Data scarcity refers to an insufficient volume of data points. Data sparsity, in contrast, refers to the distributional properties of the data, particularly in high-dimensional feature spaces. A dataset can be large (not scarce) but sparse, meaning the data points are widely distributed across the vast feature space, with many possible feature combinations having no examples. While sparsity presents its own set of challenges related to the "curse of dimensionality" [20], data scarcity is a more foundational problem that impacts the very ability to undertake a machine learning algorithm.

2.1.2 Consequences for Deep Learning Models

The consequences of training data-hungry deep learning models in a state of data scarcity are not subtle; they manifest as a cascade of failures that undermine a model's performance, reliability, and trustworthiness. As the complexity of a model (i.e., its degrees of freedom or number of parameters) increases, so too does its requirement for data to learn a meaningful and generalizable function. When this requirement is not met, the following critical issues arise:

- **Overfitting**: This is the principal and most immediate consequence of data scarcity [13]. Overfitting occurs when a model, especially high-capacity models such as a deep neural networks, learns the training data too well. Instead of capturing the underlying biological signal, it begins to memorize the noise, outliers, and artifacts present in the limited training samples. This results in a model with deceptively high accuracy on the data it was trained on but an inability to perform on any new data it encounters, rendering it clinically useless.

- **Poor Generalization**: Generalization is the ability of a model to maintain its performance on new, unseen data drawn from the same underlying distribution. A model that has overfit a small dataset will exhibit poor generalization [12]. In a clinical context, this means a model trained at one hospital may fail when deployed at another, or a model trained on images from one MRI scanner may not work on images from a different manufacturer. This lack of robustness is a critical barrier to the widespread adoption of AI in healthcare, as models must be reliable across diverse clinical settings and patient populations.

- **Algorithmic Bias and Health Inequity**: Data scarcity significantly amplifies algorithmic bias. When a training dataset is not only small but also unrepresentative of the broader population, the resulting model will inevitably be biased. It will learn patterns specific to the over-represented group and perform poorly, or even make systematically incorrect predictions, for underrepresented demographic groups (e.g., based on race, gender, or socioeconomic status) [7]. Deploying these modes in a clinical setting may perpetuate existing health disparities, resulting in ethically troubling and adverse outcomes. [21].

2.1.3 The Problem-Dependent Nature of Data Scarcity

The discussion so far has characterised data scarcity in terms that apply broadly to the biomedical domain. Such a characterization, however, risks obscuring a fact that is central to the methodological choices made in the remainder of this chapter: scarcity is not an absolute property of a dataset, but a **relation between the available data, the task to be solved, and the hypothesis class chosen to solve it**. A cohort of three hundred patients may be entirely adequate for estimating a logistic regression over five clinically curated covariates, and hopelessly insufficient for training a deep architecture over several hundred features; neither statement about "three hundred patients" carries meaning in isolation. Any claim about the sufficiency of data must therefore be qualified along four interacting dimensions, namely the application domain, the type of data, the formulation of the prediction

task, and the inductive bias of the model.

The application domain fixes the acquisition regime and, with it, the structural origin of the bottleneck. Clinical studies on emerging syndromes are bounded by the number of patients a single centre can recruit within the observation window, which typically yields cohorts in the order of 10^2–10^3 instances, so that scarcity is a scarcity of subjects. Digital pathology is not bounded by raw signal at all: a single Whole-Slide Image yields tens of thousands of tissue patches, and the constraint lies entirely in the expert annotation required to make that signal usable. Genomic and transcriptomic studies operate instead in the $p \gg n$ regime, where the number of measured variables exceeds the number of subjects by orders of magnitude and the difficulty is one of statistical identifiability rather than of volume. The same label thus denotes three qualitatively different conditions, each calling for a different remedy.

The modality of the data determines, in turn, which remedies are even admissible. For images, a rich repertoire of label-preserving transformations is available, and the relevant invariances can often be stated exactly: rotation and reflection are semantically neutral in histopathology, where tissue has no canonical orientation, but are not neutral in chest radiography, where laterality carries diagnostic meaning. For the mixed binary and categorical vectors typical of clinical records, no comparable canonical invariance exists. Synthetic instances must be produced by interpolation in a feature space whose geometry has no intrinsic clinical semantics, which is precisely why the plausibility of the resulting virtual patients cannot be taken for granted and must be assessed separately. Multimodal settings add a further variant, in which scarcity manifests not as a reduced number of subjects but as missing modalities for a subset of them.

Equally decisive, and more easily ignored, is the formulation of the prediction task, since it is this that governs the amount of supervision each instance actually carries. A binary classification problem extracts one bit of supervision per subject; a multi-label problem over Q strongly imbalanced targets is governed instead by the rarest positive class, so that a cohort of several hundred patients may supply only a few dozen positive examples for the labels of clinical interest, and the effective sample size is a small fraction of the nominal one. Annotation cost scales along the same axis: image-level classification requires a single expert judgement per slide, whereas semantic segmentation requires pixel-level delineation, raising the marginal cost of each labelled example by orders of magnitude and shifting the binding constraint from acquisition to expert time.

The fourth dimension concerns the **learner** itself, and is the one on which the argument of this thesis ultimately turns. The data requirement of a model is a function of the priors it encodes: hypothesis classes whose structure already matches that of the problem are, by

construction, more sample-efficient. Gradient-boosted trees remain competitive with, and frequently superior to, deep architectures on small heterogeneous tabular data [54], precisely because axis-aligned partitioning is well matched to the structure of such data and because their capacity can be controlled with far fewer effective degrees of freedom. Within deep learning the same gradient reappears: convolutional architectures encode locality, translation equivariance and hierarchical composition, and therefore learn from comparatively modest image corpora, whereas Vision Transformers dispense with these priors and must recover them from data, requiring substantially larger training regimes before they match convolutional performance [43]. The consequence is that the earlier statement, that model complexity increases the demand for data, holds only at a fixed inductive bias, and that scarcity admits two distinct lines of attack: one may enlarge or restructure the data, or one may strengthen the prior. The former is the subject of the present chapter; the latter is developed systematically from Chapter 3 onwards.

Taken together, these four dependencies suggest that scarcity is better diagnosed empirically than declared by threshold. A learning problem may be considered data-scarce when generalization performance has not saturated as a function of training set size, that is, when the learning curve retains a non-negligible slope at the maximum available sample size; empirical studies across vision, language and speech show that such curves follow regular power-law regimes whose exponent depends jointly on the domain and on the architecture, and which flatten only once the model class has extracted the information the data can supply [63]. This criterion is intrinsically relative to the model, and is consistent with the observation of Section 2.1.1 that **representativeness matters more than raw volume**: a dataset stops to be scarce not when it becomes large, but when additional samples stop to be informative for the hypothesis class under consideration.

This relational view also organises the space of countermeasures. Where the lack is one of instances and is compounded by class imbalance, the natural response is data-level, acting on the training set through augmentation, minority oversampling or synthetic generation. Where the shortage is one of annotations while raw signal is adequate, the response is representation-level, whether through transfer from related corpora or through self-supervised pre-training that learns structure from unlabelled data before any label is consumed; where the difficulty is dimensional, the same level offers feature selection and dimensionality reduction. A third family acts on the hypothesis class itself, through regularization, capacity control or the deliberate adoption of architectures with stronger priors. A fourth addresses scarcity that is institutional rather than intrinsic: Federated Learning leaves data within the centres that hold it and aggregates only model updates, so that the effective training population spans multiple institutions without any transfer of patient records [119];

although it lies outside the scope of this thesis, it remains one of the principal paradigms in the field and is complementary to, rather than in competition with, the approaches developed here. A fifth and final family injects structured domain knowledge into the learning process, reducing the quantity of information that must be extracted from data alone, and constitutes the central subject of the remainder of this dissertation.
The two studies presented in the following sections were selected because they occupy distinct positions along the dimensions just introduced, and consequently call for strategies drawn from different families; their contrast is methodological rather than merely illustrative. Table 2.1 makes this positioning explicit.

	Case study I	**Case study II**
Domain	Clinical cohort study	Digital histopathology
Data type	Tabular, mixed binary/categorical	Whole-Slide Images
Scale	324 patients, high dimensionality	Abundant images, scarce labels
Task	Multi-label classification, $Q = 7$	Multi-class subtype classification
Scarcity type	Few *instances*; label imbalance	Few *annotations*
Model class	Classical MLC learners	Vision Transformer (weak prior)
Strategy	Data-level augmentation and feature reduction	Representation-level self-supervision

Table 2.1: Instantiation of the four dimensions of data scarcity on the two case studies of this chapter.

Read jointly, the two studies delimit the reach of purely data-driven mitigation. The first shows that instance-level scarcity can be partially compensated without sacrificing the interpretability on which clinical adoption depends; the second, that annotation scarcity can be circumvented whenever the raw signal is abundant. Neither, however, supplies the model with information that is absent from the data itself, and it is this common limit, rather than the shortcomings of any individual technique, that motivates the knowledge-informed turn taken in the chapters that follow.

2.2 Predicting Long-COVID Sequelae: Data Scarcity in Clinical Data

In recent years, the intelligent analysis of clinical data has become a cornerstone of modern biomedical research. An increasing number of projects are focusing on the creation of biomedical databases fed by diverse sources, such as electronic health records, clinical trial data, and laboratory analysis results, which are usually updated over time. In such contexts, providing practitioners with direct and easy access to sophisticated, AI-driven data analysis is highly beneficial. This work moves a step forward in that direction by studying how machine learning techniques can be adopted to work with a real-world databank. This work is part of the **TECNOMED-HUB** research project [1], which aims to build a research platform supporting the collection and intelligent analysis of long-COVID data, with the long-term goal of being applicable to a wide range of other diseases.
The specific goal of this study is to define a framework for predicting long-COVID sequelae using patients' clinical data collected during the acute phase of their COVID infection. **Post(long)-COVID-19** [103] syndrome is characterized by signs and symptoms (sequelae) that persist beyond 12 weeks from the onset of the acute infection and are not ascribable to other causes. From a machine learning perspective, this task is characterized as a **Multi-Label Classification** (MLC) problem, where patients' data collected at hospitalization are the instances and the long-COVID symptoms persisting at follow-up are the labels. Coping with this kind of individualized predictive modeling [72] in a real-world setting means working in a low-data regime with unbalanced labels. Therefore, this work also evaluates solutions like *upsampling techniques* (MLSmote [32]) and *feature reduction* (RELIEF and PCA) to find a valid trade-off between creating a more stable dataset and keeping the data close to the real-world clinical evidence.

2.2.1 Problem definition

The primary clinical objective of this study is to predict the long-term persistence of specific sequelae following an acute COVID-19 infection. The work is based on a real-world dataset from a long-COVID study conducted at **Ospedale Maggiore di Novara (IT)**, which collected data on 324 patients hospitalized during the first three pandemic waves. The predictive goal is to determine, at a one-year follow-up, which symptoms will persist. **Long-COVID** [103] *syndrome is clinically defined as symptoms consistent with the initial infection that are present beyond 12 weeks from the acute onset and are not ascribable to alternative causes.* For the purpose of this study, a patient is considered to have the syndrome if they

exhibit **at least** one such sequela. To make this prediction, the model is provided with two categories of patient data collected at the time of the acute illness:

- **Baseline Data**: This includes demographic information (e.g., sex, age) and a history of prior comorbidities (e.g., obesity, hypertension, anxiety, depression).
- **Hospitalization Data**: This encompasses the patient's symptoms at the onset of acute COVID-19 (e.g., fever, cough), drugs administered during their stay (e.g., glucocorticoids, antivirals), and other hospitalization details (e.g., duration of stay, oxygen administration, or ICU intubation).

The specific prediction targets are the presence or absence of 7 distinct sequelae: *arthromyalgia, asthenia, cough, diarrhea, dysgeusia, anosmia, and severe dyspnea* (derived from the mMRC respiratory test).

Block	**Group**	**Type**	**Representative variables**
Baseline	Demographics	numeric, binary	sex, age, BMI, smoking attitude, number of comorbidities
	Prior comorbidities	binary	obesity, chronic liver disease, hypertension, anxiety and depression, number of comorbidities
Hospitalization	Symptoms at onset	binary	fever, cough, dyspnea, arthralgia, anosmia, dysgeusia, diarrhea
	Administered drugs	binary	hydroxychloroquine, monoclonal antibodies, glucocorticoids, antivirals
	Hospitalization course	numeric, binary	length of stay, oxygen administration, ICU intubation, severity of the acute episode
Total: 57 features (47 binary, 10 numeric); no missing values.			

Table 2.2: Summary of the input feature space of the long-COVID predictive model. Features are grouped by semantic block; representative examples are reported for each group, the complete variable list being available in the clinical companion studies [17, 19].

The selection of these 7 sequelae as target labels was not trivial, given that a consolidated clinical definition of long-COVID was still emerging at the time the study was designed. The target set was therefore identified through a two-step process: a preliminary survey of the available medical literature on post-acute COVID-19 syndrome, followed by validation and refinement in close collaboration with the clinical team at Ospedale Maggiore di Novara, who contributed their direct experience with the cohort under study. On the literature side, the earliest systematic follow-up studies of hospitalised patients converged on a recurrent core of persistent manifestations. **Fatigue** and **dyspnea** were reported as the two most

prevalent sequelae at six months in the Wuhan cohort of Huang et al. [71], and appeared with comparable frequency, together with **arthralgia**, in the Italian cohort described by Carfì et al. [31]; chemosensory dysfunction, namely **anosmia** and **dysgeusia**, together with persistent **cough** and **gastrointestinal** manifestations, emerged consistently in large-scale symptom-tracking studies [132, 38] and in the clinical characterisation of the syndrome by Nalbandian et al. [103]. On the clinical side, the label set was constrained to those manifestations for which the Novara cohort provided reliable and systematically collected follow-up measurements, as documented in the companion clinical studies of the same population [19, 17]; the choice of the mMRC questionnaire as a proxy for **severe dyspnea**, in particular, follows the protocol adopted in those studies. This cross-validation between published evidence and expert clinical judgement was essential to ensure that the resulting label set was simultaneously clinically meaningful, supported by the literature, and consistent with the data actually collected during hospitalization and follow-up.

From a machine learning standpoint, this clinical objective is formulated as a **Multi-Label Classification problem**. This formulation is essential because a single patient may present with any combination of the 7 target sequelae, rather than only one. The formal goal is to learn a *function* $h : \mathcal{X} \to 2^{\mathcal{L}}$ that maps a patient's feature vector to a subset of possible labels $\mathcal{L}$. To this end, the problem is usually tackled by considering the set of labels represented as a binary vector of size Q, where 0 means that the label is absent (or not predicted) and 1 means the label is present (or predicted). The problem can be generalized to label ranking by learning a function $h : \mathcal{X} \times \mathcal{L} \to \mathcal{R}$ such that $h(x_i, \lambda_j) = r$ is the prediction score for label λ_j in the example x_i. Usually, the score is the probability of label λ_j given x_i, and the presence of the label can be predicted if it exceeds a given threshold (usually 0.5). Given an instance $(x_i, \mathcal{Y}_i) \in E$ we define $\mathcal{Y}_i$ as the *label-set* of x_i. Usually, the number of label-sets in a given instance space (dataset) is much less than $2^{|\mathcal{L}|}$, making the label-set space very sparse.
The dataset's structure defines the learning problem's parameters:

- **Input Space ($\mathcal{X}$)**: The instance space consists of n=324 patients. Each patient is represented by a feature vector of D=57 features, which are a mix of 47 binary and 10 numeric attributes.
- **Label Space ($\mathcal{L}$)**: The label space consists of $\mathcal{L}$=7 binary labels, one for each of the target symptoms.

This real-world medical dataset presents two significant technical challenges that define the problem:

- **Low-Data Regime**: The study operates with a limited number of instances (n=324) relative to the feature-space dimensionality (D=57). This "high dimensionality vs low sample size" scenario makes models prone to high variance and overfitting, complicating the task of finding generalizable patterns. The scarcity of available patients for the study was a primary motivator for investigating augmentation.
- **Label Imbalance and Sparsity**: The dataset exhibits "significant" label imbalance , with a Mean Imbalance Ratio of 2.303. Furthermore, the SCUMBLE metric is very low (S=0.012) , indicating a low level of concurrence, or interaction, between minority and majority labels (see next section about metrics).

A small, imbalanced dataset with low label concurrence is critical. They directly motivate the two main analytical strategies explored in this work: *data augmentation* (specifically via MLSmote) to synthetically increase the number of representative instances and feature reduction (via RELIEF and PCA) to mitigate the *"curse of dimensionality"*

2.2.2 The Multi-Label Classification Framework

MLC can be approached in two different ways: *problem transformation* and *algorithm transformation* [24]. Methods in the former category transform the multi-label dataset into one or more datasets that are then targeted using traditional single-label classification algorithms; they finally build one or multiple single-label models. Methods in the second category adapt traditional single-label algorithms to the multi-label setting such as decision trees, functional models (SVM or NN), instance-based models and probabilistic models. In the present study we concentrate on the **problem transformation approaches**, which is the most widely adopted in MLC; in fact, most of the algorithm transformation methods actually rely on an internal problem transformation in order to solve the MLC task. Moreover, this allows in principle to experiment with a larger number of base classifiers. Concerning problem transformation methods, they can further be divided into **binary**, **multi-class** or **ensemble methods** [24]. In **binary methods**, each pair of labels is considered to produce a set of single-target binary datasets following a one-vs-all strategy; the result is the construction of L binary classifiers, one for each possible label. Methods belonging to this category are Binary Relevance (BR) [50] and different versions of classifier chain such as CC [114] or BCC [168]. **Multi-class** methods, on the contrary, build one single multi-class classifier: the target class has a set of possible values equal to the cardinality of the label-sets. The Label Powerset or Label Combination (LC) algorithm [50] and the Pruned Set (PS) method [115] are examples of this kind of approach. Finally, **ensemble methods** adopt ensemble-like techniques in order to train multiple single-label classifiers; important representatives of

such methods are the Conditional Dependency Network (CDN) [58] and the RAKEL [148] algorithms.

Beyond the classification methods, **MLC dataset characterization** is a very important aspect, especially in terms of label distribution and balancing. Different measures can be used to characterize a dataset for MLC [139] starting from basic measures such as number of instances, number of attributes, number of labels, number of distinct label-sets. Another characterization can be given in terms of label distribution through measures such as **label cardinality** (average number of labels in the examples) and **label density** (the cardinality divided by the number of labels), as reported in 2.1:

$$Card(E) = \frac{1}{n}\sum_{i=1}^{n}|\mathcal{Y}_i| \quad Dens(E) = \frac{1}{|\mathcal{L}|}Card(E) \tag{2.1}$$

Imbalance level measures are also important in order to characterize how frequent or rare are certain labels. The most relevant measures are the **Imbalance Ratio per Label** (IRLbl) and the **Mean Imbalance Ratio** (MeanIR) which are reported in equations 2.2 and 2.4 respectively:

$$IRLbl(\lambda) = \frac{\max_{\lambda' \in \mathcal{L}} \sum_{i=1}^{n} h(\lambda', \mathcal{Y}_i)}{\sum_{i=1}^{n} h(\lambda, \mathcal{Y}_i)}, \tag{2.2}$$

$$h(\lambda, \mathcal{Y}_i) = \begin{cases} 1 & \text{if } \lambda \in \mathcal{Y}_i \\ 0 & \text{if } \lambda \notin \mathcal{Y}_i \end{cases}; \tag{2.3}$$

$$MeanIR = \frac{1}{|\mathcal{L}|}\sum_{\lambda \in \mathcal{L}} IRLbl(\lambda). \tag{2.4}$$

IRLbl(λ) is the ratio between the occurrence of the majority label and the current label λ; IRLbl is ≥ 1 (1 only for majority labels), and the larger the value the greater the imbalance of the label. MeanIR characterises the level imbalance of the whole dataset. In addition, the standard coefficient of variation of IRL (**CVIR**, the ratio between standard deviation and mean) can be useful to measure if labels experience a similar level of imbalance or if there are large differences, in terms of imbalance, among them.
Finally, the **SCUMBLE** metric [33] provides a way to understand the level of concurrence between minority and majority labels; values are in the [0, 1] range, and the higher the value the more instances sharing minority and majority labels exist in the dataset. It is based on the computation of the **Atkinson index** [8] over the IRLbl of the labels occurring in each instance; the final score is the average over all the instances of the dataset (see [33] for details). A small value of SCUMBLE (usually < 0.1) denotes a low concurrency between

minority and majority labels; this is a proxy for the possibility of adopting data augmentation techniques without the risk of increasing in a significant way the label imbalance. In particular, upsampling techniques such as **MLSmote** [32] are well justified and applicable when the dataset SCUMBLE is low [33, 113].

Concerning **evaluation measures** for MLC, several metrics have been proposed with different aims. They can be categorized as *bipartition-based* and *ranking-based* [50]. In this setting, bipartition-based are more significant, since the emphasis is on predicting the right set of long-COVID symptoms, rather than a correct ranking. In the present study, we then consider some of the most popular and natural bipartition-based metrics. Let $\mathcal{Y}_i$ be the ground-truth label-set of instance x_i, $\mathcal{Z}_i$ the predicted label-set and Δ the symmetric set difference operator:

- **HammingScore**: This is 1 - HammingLoss. **Hamming Loss** evaluates how many times, on average, an example-label pair is misclassified:

$$HammingLoss = \frac{1}{n}\sum_{i=1}^{n}\frac{|\mathcal{Y}_i \Delta \mathcal{Z}_i|}{|\mathcal{L}|} \tag{2.5}$$

- **Accuracy** or **Jaccard Index**: This evaluates the average proportion of labels correctly classified on the total number (predicted and actual) of labels, and averaged over all instances:

$$Accuracy = \frac{1}{n}\sum_{i=1}^{n}\frac{|\mathcal{Y}_i \cap \mathcal{Z}_i|}{|\mathcal{Y}_i \cup \mathcal{Z}_i|} \tag{2.6}$$

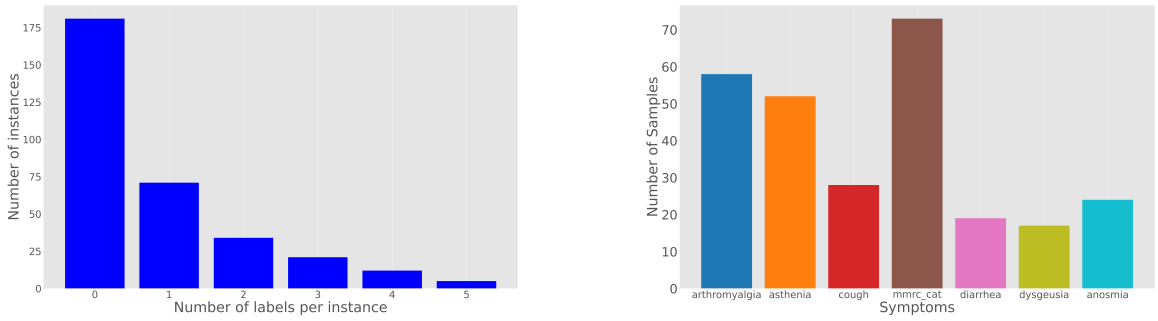
- **ExactMatch**: evaluates the percentage of label-sets that are correctly predicted as a whole:

$$ExactMatch = \frac{1}{n}\sum_{i=1}^{n} 1(\mathcal{Y}_i = \mathcal{Z}_i). \tag{2.7}$$

Notice that the Hamming Score is somewhat lenient, since when several labels are absent in a label-set, the score may find several correct matches; however, it is an important metric since the correct prediction of an absent label should be relevant in several applications, as in this case study. On the contrary, Exact Match is a very strict measure, since partial predictions are completely ruled out. A somewhat intermediate metric is the Accuracy/Jaccard Index, where however, due to the nature of an MLC problem, one can hardly expect to get results close to 1 (as can be in the case of single-label classification). Jaccard Index of about 0.65/0.7 can be regarded as good results (see [94]).
Finally, it interesting consider the **AUC** (Area Under Roc Curve) in the macro-averaged version (the metric is computed independently for each label and then averaged). Since all labels have the same importance (they should be treated equally as one of the symptoms of long-COVID), macro-average is preferred to micro-averaging (where the specific measures of each class are combined together).

2.2.3 Dataset Analysis and Augmentation Strategies

The dataset produced in the long-COVID study (named **orig** in the following) resulted in $n = 324$ instances (324 different patients under study) with $D = 57$ features (47 binary features and 10 numeric features). No missing value was reported. The cardinality of the label-set is $|\mathcal{L}| = 7$, representing the 7 long-COVID sequelae. Some distributions concerning labels and instances are reported in Figure 2.1. The first row of Table 1 shows the main characterization measures of the original dataset that was obtained. It can be noticed that the imbalance level (as measured by MeanIR) is significant, even if a low CVIR shows that there are similar levels of imbalance among the labels. Moreover, the SCUMBLE metric S=0.012 is relatively low, showing a low level of interactions (concurrence) between minority and majority labels. This analysis highlights a key challenge: the "high dimensionality vs low sample size" problem. The low SCUMBLE value suggests that data augmentation techniques through resampling can be attempted to get more significant data. This is very important in the present context, since the study was able to process just a few hundred real patients, and the possibility of increasing the dataset size is considered fundamental. An alternative to upsampling is to reduce the number of features through some feature reduction methods. The section discusses both alternatives, starting with the upsampling approach.



(a) Label number by instance number

(b) Label distribution

Figure 2.1: Dataset distribution

The first approach tested in the study was to resort to **MLSmote** [32], an *upsampling technique* suited to the multi-label framework. It produces new synthetic data by first identifying minority labels (using IRLbl measures), followed by a kNN search and a feature generation from that search, as in the standard **SMOTE** technique [34]. Finally, label-sets for the synthetic instances are generated using 3 possible criteria: *Intersection* (I): Labels appearing in the reference sample and in all the neighbors are added to the new sample. *Union* (U): Labels appearing in either the reference sample or in any of the neighbors are

added to the new sample. *Ranking* (R): Labels present in more than half of the reference and neighbor samples are added to the new sample (majority voting). The study began with the original dataset (orig) and produced six augmented datasets using MLSmote with k = 3, 5 (the number of considered neighbors). For each k value, three generation strategies were applied (intersection, union, and ranking) resulting in datasets k3I, k3U, k3R, k5I, k5U, and k5R, respectively. The number of synthetic instances to introduce was set to **100**, incrementing the available examples in a reasonable way.

Table 2.3 shows a summary of the characterization measures for each of these datasets. It can be noticed that the more aggressive the augmentation (U more aggressive than R more aggressive than I), the larger the number of label-sets and the presence of label-sets with several labels, as reflected by cardinality and density as well. This also allows to reduce the label imbalance, by also keeping the SCUMBLE index under control. To this end, considering a smaller number of neighbours in the instance generation (k=3) seems to take the label imbalance more under control.

To give an intuition of how MLSmote influences the resulting labelsets, Figure 2.2 reports the concurrence plots among the labels for datasets orig and k3U respectively. The plot concerning k3U shows more uniform bands than the plot for orig, meaning that the balance of the label-sets is more noticeable in the former.

Dataset	n	#ls	#pl	Card	Dens	MeanIR	CVIR	SCUMBLE
orig	324	36	54	0.670	0.096	2.303	0.417	0.012
k3I	424	36	54	0.592	0.085	2.173	0.415	0.009
k3R	424	38	79	0.814	0.116	1.762	0.373	0.007
k3U	424	48	154	1.752	0.250	1.566	0.239	0.008
k5I	424	36	54	0.512	0.073	2.303	0.417	0.009
k5R	424	36	54	0.599	0.086	2.089	0.466	0.008
k5U	424	48	154	1.745	0.249	1.602	0.311	0.009

Table 2.3: Main features of considered datasets; n: number of instances, #ls: cardinality of label-set, #pl: number of label-set with more than one label

It is worth stressing, however, that while the low SCUMBLE value justifies the applicability of MLSmote from the point of view of multi-label imbalance (i.e., the augmentation is not expected to worsen the concurrence between minority and majority labels), this technical condition does not guarantee per se that the synthetic patients generated by the algorithm are clinically plausible. MLSmote, being an extension of SMOTE, produces new instances through linear interpolation in the feature space of the nearest neighbours: this procedure can, in principle, give rise to feature combinations (e.g., binary comorbidities and hospi-

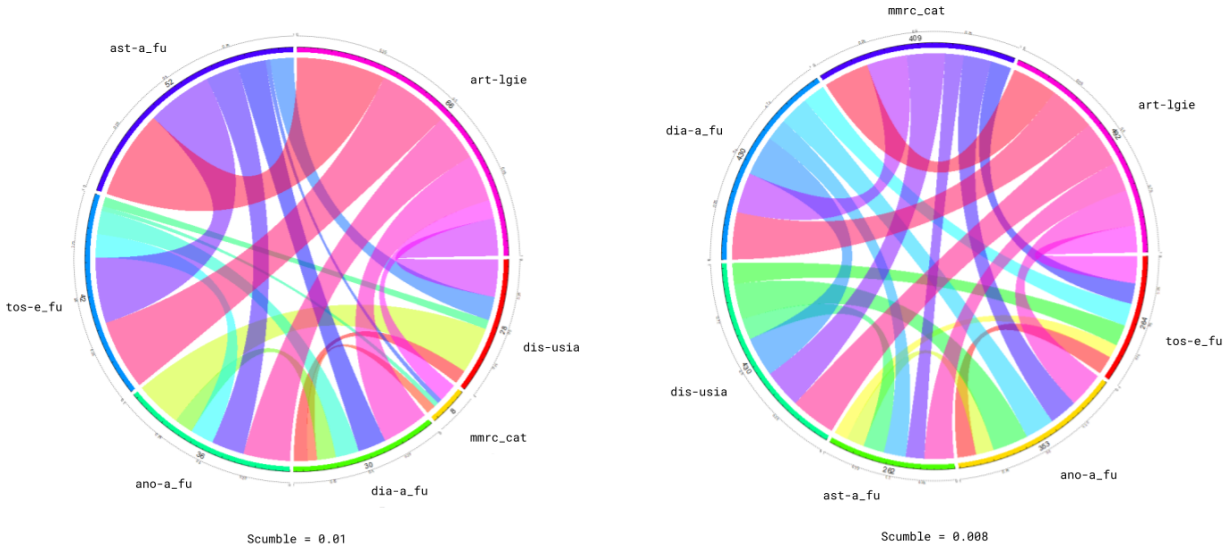


Figure 2.2: Concurrence plots: `orig` (left) and `k3U` (right)

talization drugs) that are statistically consistent with the training distribution but are not fully realistic from a clinical standpoint. The generation of realistic synthetic patient data is, in fact, still recognized as an open research problem in the medical informatics community [108, 101, 47]. In the present work, we therefore adopt MLSmote as a principled and computationally lightweight baseline that is appropriate given the low SCUMBLE value, while acknowledging that more sophisticated generative approaches (e.g., CTGAN, VAE-based models, or LLM-assisted tabular synthesis) may in the future provide a more faithful reproduction of the underlying clinical phenomenology.

The second approach to mitigate the problem of a reduced sample size was to consider methods of *feature reduction*. Taking the number of attributes under control (by avoiding noisy ones or those with low information) may be a way of dealing with a reduced number of examples when evaluating predictive results [111]. To this end, the authors considered two different methods based on feature ranking: a supervised approach called **RELIEF** [81], and a standard unsupervised approach, namely **Principal Component Analysis** (PCA).

Relief computes a score for each feature which can then be applied to rank and select them. The score is based on identifying feature value differences between nearest neighbor instance pairs. A "hit" is registered when a feature value difference is observed in a neighboring instance pair with the same class, and a "miss" when the difference is in a pair with different classes. The score is suitably increased in case of a miss and decreased in case of a hit. Since the method relies on a single label class, it was adapted to the multi-label setting by considering a single class label obtained from the possible label-sets occurring in the

dataset (as in the LC multi-label classification method) . The study adopted the RELIEF implementation from WEKA [49] with $k = 10$ nearest neighbors. By adopting the widely adopted "rule of thumb" that considers as significant a correlation between two variables when their correlation score is above the threshold $\sigma = \frac{2}{\sqrt{n}}$ (where n is the sample size), we finally selected the first top 8 features, those having a score exceeding the above threshold (in this case $n = 324$ and $\sigma = 0.11$). The resulting dataset has been indicated as `Relief` in the following (see Section Results).
Concerning the unsupervised feature reduction, the attributes related to the sequelae were removed and a standard PCA was performed (again using WEKA). By considering an explanation score of 95% of the data variance, 44 new features were obtained which are linear combinations of the 57 original ones. The resulting dataset has been labeled as PCA. Notice that, apart from the supervised vs unsupervised approach, the two methods have a significant difference in the total reduction of features (86% reduction for `Relief` and 23% for PCA).

2.2.4 Experimental framework

The experimental analysis has been performed using Meka [116], a multi-label extension to Weka. The following problem transformation approaches to MLC were considered:

- BR (Binary Relevance): a set of 7 independent binary classification models (one for each possible label) has been built and results have been merged; label correlation information is completely neglected in this case.

- CC (Classifier Chain): labels are processed in a random order by a set of binary classifiers; each classifier predicts the presence of the corresponding label, by considering the classification produced by the previous ones, i.e., classifier C_i $(2 \leq i \leq 7)$ uses problem features augmented with the label values predicted by classifiers C_j $(1 \leq j \leq i-1)$. Classifier C_1 uses only the problem features.

- BCC (Bayesian Classifier Chain): as the CC method, but the order in which labels are processed is not random; we tested two possible versions of BCC corresponding to different label ordering: BCC(I), where the label order is induced by the mutual information among labels, and BCC(C) where label ordering is determined by label co-occurrence counts.

- LC (Label Combination): the problem features are augmented with a class attribute taking discrete values in the range $[1, \#ls]$ where #ls is the cardinality of the label-sets; the problem is then solved as a standard single-label multi-class classification.

- PS (Pruned Set): first it prunes all examples having label-sets that occur less than p times in the training set (set $p = 2$); then it subsamples the label-sets of these examples for label subsets that occur more frequently in the training data. It then attaches these label sets to the example, creating new examples and reintroducing them into the training set; after these steps, it trains a standard LC classifier.

- CDN (Conditional Dependency Network): it builds a fully-connected network where nodes are the labels, then it builds a set of binary classifiers (7 in this experiment) one for each label λ_j, predicting $p(\lambda_j|x_i, \lambda_1, \ldots \lambda_{j-1}, \lambda_{j+1}, \ldots \lambda_Q)$; inference is performed through Gibbs sampling [121] over a set of I iterations, and by collecting results from last I_c iterations ($I = 500$ and $I_c = 100$).

- RAkEL (RAndom k-labEL Pruned Sets): it randomly draws M subsets of labels, each with k labels, from the set of labels, and trains PS upon each one (we set PS as indicated above, $M = 10$ and $k = 3$).

The tested MLC methods are 4 binary methods (BR, CC and the two versions of BCC), 2 multi-class methods (LC and PS), and 2 ensemble methods (CDN and RAkEL). We tested the above methods with several base classifiers (from lazy classifiers, to neural nets, SVM and tree-based classifiers), and we have finally found (through an inner cross-validated grid-search) that a 200-trees Random Forest provided the most interesting results[1]. In the following, we will report results concerning the above transformation-based methods using this base classifier.
It is worth to remark that multi-class methods rely on the label-sets which are actually present in the data. A suitable data augmentation strategy could be really important to consider potential label-sets which are not occurring in the original dataset.

2.2.5 Classification results and considerations

This section presents the comparative results obtained by the MLC algorithms described in previous section, evaluated on the various datasets generated. All methods were tested using 5 runs of **10-fold outer cross-validation**, and the results were subsequently averaged. In the following tables, the best results obtained for each specific dataset (the original, the ones from MLSmote upsampling, and those from RELIEF and PCA feature reduction) are highlighted in bold. The last row of each table reports the average performance of each MLC method across all datasets, with the best overall average also shown in bold.

[1]The other relevant hyperparameters we have adopted are: a split factor (number of attributes to split on) equal to $\log_2(\#\text{features}) + 1$, and an unconstrained depth of the trees.

Table 2.4 shows the Accuracy (Jaccard Index) for the various datasets and tested methods. It can be noticed that multi-class methods (LC, PS) generally achieve better performance, with chain classifiers (BCC, CC) showing very close results, independent of label ordering. The use of the RAkEL ensemble does not appear to actually improve upon the basic PS version. As outlined in Section 2.2.2, these reported scores can be regarded as quite satisfactory. However, analysis of the feature-reduced datasets shows that the reduction obtained through RELIEF **is too drastic**, significantly reducing the score compared to other datasets. Moreover, PCA does not show an actual improvement in performance over using the original dataset with all original features. Similar considerations apply to the Exact Match score, presented in Table 2.5. For this metric, the **PS method is noted as definitively the best performer**. Given that Exact Match is a much stricter metric than the Jaccard Index, the obtained results are considered very satisfactory. It is also noted that the augmented datasets produced with MLSmote (U) exhibit a decrease in Exact Match; this is attributed to their larger cardinality and density, which decreases the probability of achieving a perfect label-set match.

The final bipartition-based metric considered was the Hamming Score, with results shown in Table 2.6. The obtained scores are in general very good. Again, comparable performances are observed between the multi-class methods and the classifier chains. As seen with the previous metrics, the feature reduction approaches show that PCA usually provides the same results as using every available feature, while the reduction proposed by RELIEF appears to lose important attributes required for the final predictions. An additional observation is that, in general, the dataset obtained through PCA shows a **more stable behavior** across all tested methods. While for other datasets there are significant differences in the metrics depending on the adopted multi-label classification algorithm, this difference is not as evident in the case of PCA.

Finally, the macro-averaged AUC was computed, as shown in Table 2.7. In this case, the results of all methods are comparable over all datasets. A slightly better performance can be noticed for the basic BR method and for the ensemble algorithm CDN. This can be explained by the fact that in these approaches, the contribution of labels that are **more represented dominates the aggregated** ROC, disregarding the contribution of some minority labels (which are usually better captured by multi-class methods). It is also noted here that, in terms of AUC, the use of datasets obtained via feature reduction (Relief, PCA) does not allow the tested multi-label classification algorithms to reach performances close to those obtained on some upsampled datasets, particularly those using the union (U) method of MLSmote.

Datasets	BCC(C)	BCC(I)	BR	CC	CDN	LC	PS	RAkEL
orig	0.604	**0.605**	0.420	0.604	0.415	0.604	**0.605**	0.604
k3I	0.646	0.647	0.507	0.649	0.503	0.653	**0.673**	0.651
k3R	0.615	0.615	0.546	0.613	0.527	**0.660**	0.643	0.652
k3U	0.637	0.639	0.577	0.638	0.537	0.637	0.624	**0.651**
k5I	0.666	0.666	0.488	0.669	0.466	0.646	**0.680**	0.641
k5R	0.645	0.648	0.499	0.645	0.496	0.648	**0.668**	0.647
k5U	0.651	0.649	0.582	0.649	0.552	0.648	0.637	**0.666**
Relief	0.556	0.558	0.471	0.559	0.425	0.552	**0.566**	0.539
PCA	**0.606**	**0.606**	0.445	**0.606**	0.432	0.605	0.605	0.603
Avg	0.625	0.626	0.504	0.626	0.484	0.628	**0.633**	0.628

Table 2.4: Accuracy (Jaccard Index)

Datasets	BCC(I)	BCC(C)	BR	CC	CDN	LC	PS	RAkEL
orig	0.601	0.603	0.383	0.601	0.377	0.604	**0.605**	0.601
k3I	0.641	0.643	0.475	0.645	0.471	0.644	**0.668**	0.642
k3R	0.590	0.593	0.479	0.590	0.461	0.620	**0.623**	0.614
k3U	0.559	0.564	0.460	0.568	0.431	0.589	**0.594**	0.586
k5I	0.666	0.666	0.473	0.669	0.449	0.646	**0.680**	0.641
k5R	0.641	0.645	0.472	0.640	0.468	0.639	**0.664**	0.640
k5U	0.571	0.571	0.472	0.580	0.435	0.586	**0.594**	0.593
Relief	0.546	0.548	0.435	0.548	0.388	0.544	**0.561**	0.523
PCA	**0.605**	**0.605**	0.406	**0.605**	0.391	**0.605**	**0.605**	0.603
Avg	0.602	0.604	0.451	0.605	0.430	0.609	**0.622**	0.605

Table 2.5: Exact Match

Datasets	BCC(I)	BCC(C)	BR	CC	CDN	LC	PS	RAkEL
orig	**0.904**	**0.904**	0.847	0.903	0.844	**0.904**	**0.904**	**0.904**
k3I	0.915	0.915	0.874	0.915	0.875	0.911	**0.922**	0.913
k3R	0.903	0.903	0.874	0.903	0.871	0.908	**0.910**	0.908
k3U	0.873	0.876	0.850	0.876	0.839	0.859	0.856	**0.876**
k5I	0.918	0.918	0.868	0.918	0.869	0.908	**0.922**	0.909
k5R	0.913	0.914	0.872	0.913	0.874	0.909	**0.922**	0.911
k5U	0.881	0.880	0.859	0.882	0.842	0.861	0.858	**0.883**
Relief	0.885	0.886	0.847	0.885	0.846	0.882	**0.891**	0.877
PCA	**0.905**	**0.905**	0.851	**0.905**	0.847	0.904	0.904	0.904
Avg	**0.900**	**0.900**	0.860	**0.900**	0.856	0.899	0.898	0.900

Table 2.6: Hamming Score

Datasets	BCC(C)	BCC(I)	BR	CC	CDN	LC	PS	RAkEL
orig	0.504	0.504	0.543	0.503	**0.547**	0.501	0.500	0.501
k3I	0.535	0.537	0.574	0.533	**0.579**	0.563	0.560	0.550
k3R	0.621	0.618	0.727	0.622	**0.731**	0.673	0.639	0.666
k3U	0.763	0.765	**0.847**	0.768	0.836	0.746	0.724	0.778
k5I	0.496	0.496	0.452	0.496	0.459	0.490	**0.497**	0.491
k5R	0.529	0.531	0.544	0.532	**0.560**	0.553	0.551	0.550
k5U	0.785	0.784	**0.870**	0.787	0.846	0.760	0.737	0.797
Relief	0.503	0.504	0.518	0.504	**0.547**	0.495	0.495	0.502
PCA	0.501	0.501	0.588	0.501	**0.589**	0.500	0.500	0.500
Avg	0.582	0.582	0.629	0.583	**0.636**	0.587	0.578	0.593

Table 2.7: Area under ROC (macro-averaged)

2.2.6 Interpretability Analysis

To make the approach truly usable in the medical domain, an essential aspect is to provide physicians with techniques for interpreting the model's predictions. To this end, the problem was approached using the Shapley Additive Explanations (SHAP) framework [93]. SHAP assigns an importance value to each feature for a given prediction, mathematically representing its contribution. This is based on cooperative game theory, ensuring each feature's impact is fairly calculated. Frameworks like SHAP are **particularly useful for black-box models**, such as the Random Forest used in this study, where direct interpretation is not straightforward. All interpretability experiments were conducted using the Random Forest model with the PS (Pruned Set) approach, which was one of the best-performing methods on average.

The first objective was to develop tools that physicians could inspect to understand why the model made specific predictions. Two types of SHAP plots were used. The first is an a**ggregate bar plot** (shown in Figure 2.3) which sums the absolute SHAP values across all samples. This assesses the overall contribution of each feature to the model's output, with segments in each bar showing the feature's contribution to each specific label (symptom). The second type, **beeswarm plots** (shown in Figures 2.5 and 2.4), provides a more detailed, sample-level insight, showing how the specific value of a feature (low values in blue, high in red) influences the prediction for a particular class.

From the SHAP analysis and evaluations by physicians, several findings were identified that are consistent with existing medical literature:

- The most significant features influencing the model are **sex, BMI, and hospitalization duration** (see Figure 2.3), which aligns with previous studies on long-COVID

[126, 154, 98].

- These features, along with **anxious symptoms, depression, and sleep changes**, were found to be particularly important predictors of arthromyalgia, as also reported in [110] (see Figure 2.5).
- The **administration of glucocorticoids** at the onset of COVID-19 was associated with the absence of **diarrhea** at follow-up (see Figure 2.4).
- The number of comorbidities and the severity of the initial case also appear to be linked to long COVID symptoms [98].
- A strong predictive pattern was observed where **anosmia** at the time of hospitalization is a strong predictor of anosmia at follow-up, a pattern not seen for other symptoms.

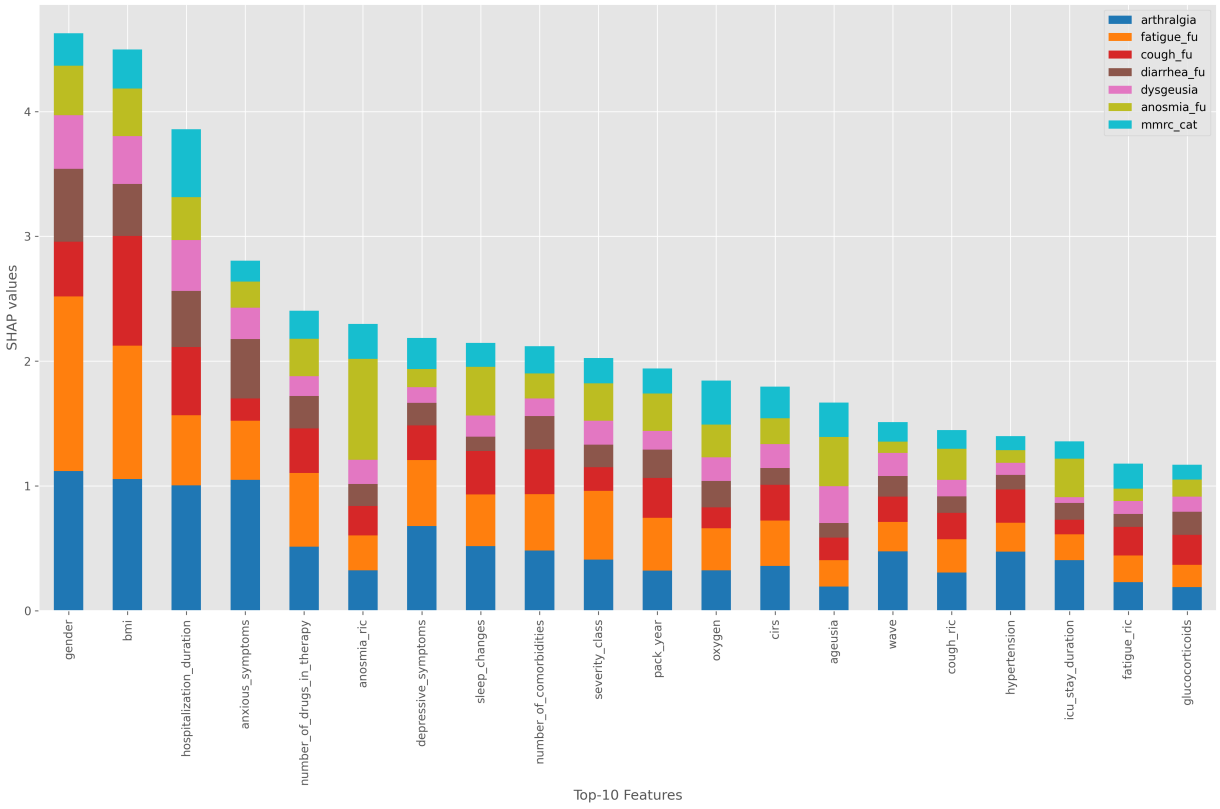


Figure 2.3: SHAP values for the original dataset using a Random Forest model

The second part of the analysis focused on **evaluating the impact of upsampling methods** on the model's interpretability. Given the data scarcity, augmentation is necessary, but it was important to validate the hypothesis that these augmentations maintain

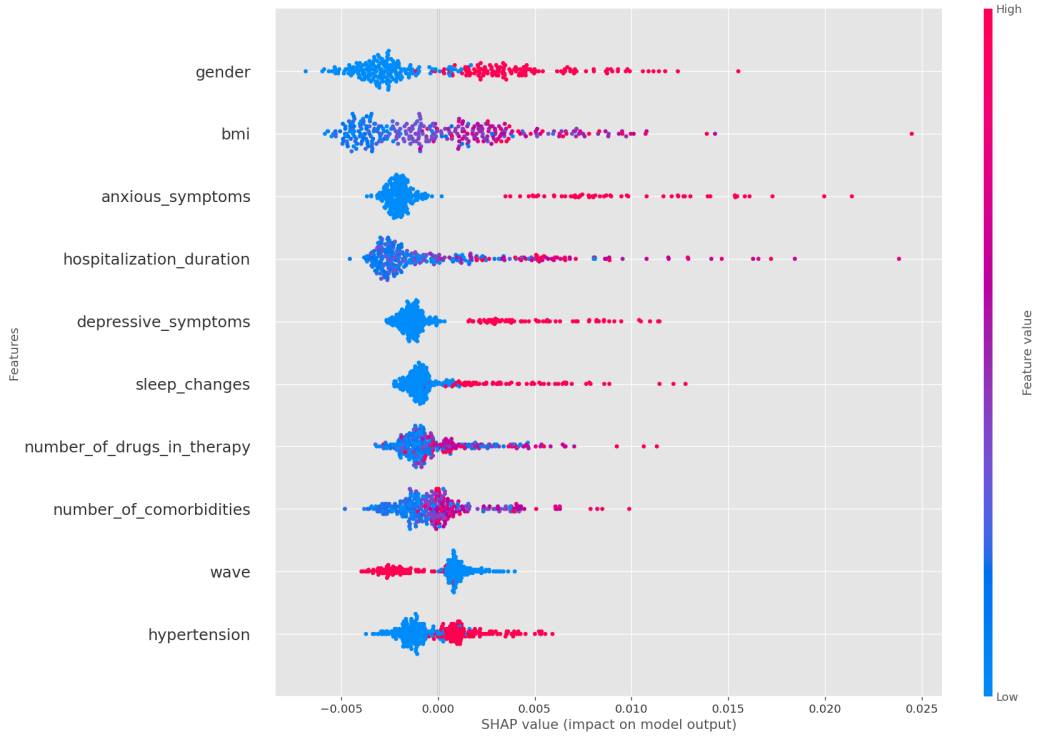


Figure 2.4: SHAP values for arthalgia at follow-up.

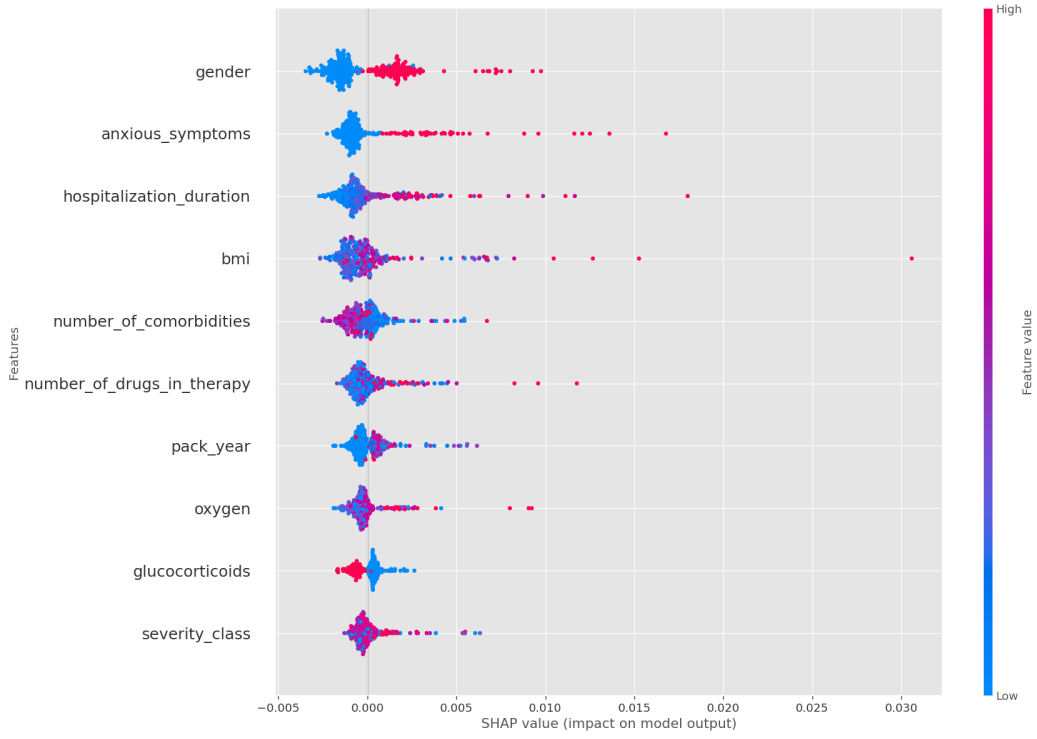


Figure 2.5: SHAP values for diarrhea at follow-up.

consistency in terms of **interpretability**. To achieve this, the SHAP values of each dataset were normalized, and the top 15 features identified in the original dataset were examined to see if they retained their importance (high SHAP values) in the models trained on the augmented datasets. The results of this comparison are presented in Figure 2.6.
Overall, the analysis shows that the top 15 features identified in the base model consistently achieve high SHAP values across most upsampled models. This consistency was particularly strong in the k3I, k5I, and k5R methods. In contrast, the union-based upsampling methods (k3U, k5U) showed lower SHAP values for the top features. This is likely due to the more aggressive nature of union-based upsampling, which may significantly alter the dataset's statistical properties and thus the model's feature dependencies.

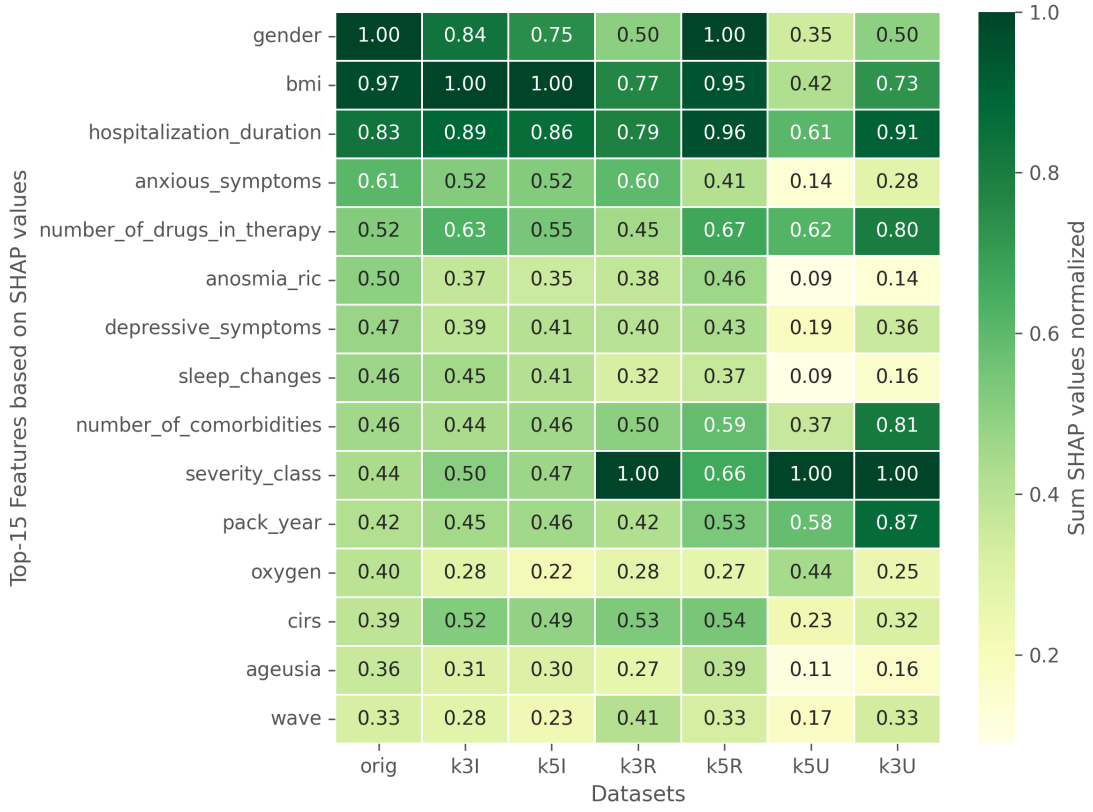


Figure 2.6: Comparison of SHAP values between the original dataset and different augmentations. The SHAP values of each dataset have been normalized independently.

2.2.7 Conclusions on data-driven approach experiment

This section presented a study that directly confronts a fundamental challenge in medical informatics: the pervasive issue of **data scarcity** when applying data-driven methods to

novel clinical syndromes. The long-COVID study, involving data from over 300 patients, served as a representative case study for this "*low sample size, high dimensionality*" problem, where the original dataset was deemed potentially under-representative.
The core methodological contribution of this work was a systematic investigation into data-driven strategies to mitigate this scarcity. Two primary, and often competing, strategies were explored:

- **Data Augmentation**: The dataset was synthetically expanded using the MLSmote algorithm under six different policies, testing various heuristics for generating new, plausible patient profiles.
- **Dimensionality Reduction**: The feature space was contracted using both supervised (RELIEF) and unsupervised (PCA) methods to reduce complexity.

A representative set of MLC models was subsequently trained on all resulting datasets to evaluate which scarcity-mitigation strategy yielded the most effective predictive models. The results demonstrated that the ability to predict long-COVID sequelae was highly dependent on these initial data-driven choices. The findings indicated that data augmentation via MLSmote was a successful strategy, often improving model performance. Conversely, aggressive dimensionality reduction (particularly the supervised RELIEF method) appeared to discard salient information, leading to degraded performance.
Crucially for the topic of applying data-driven solutions in a clinical setting, the impact of these synthetic augmentations on model trustworthiness was assessed. An interpretability analysis using the SHAP framework was conducted. A significant finding of this work is that the use of data augmentation **did not harm or distort the qualitative interpretability** of the models. The SHAP analysis of models trained on augmented data still highlighted clinically relevant features, confirming that this data-driven approach can enhance model robustness without sacrificing the clinical actionability and trustworthiness essential for medical adoption.
Finally, it is important to frame the scope of these results within a broader methodological reflection. While MLSmote proved effective in this setting the question of whether the generated instances correspond to truly plausible virtual patients remains open. Interpolation-based oversampling techniques are, by construction, agnostic with respect to the underlying clinical semantics, and the risk of producing synthetic profiles that include unrealistic combinations of comorbidities, symptoms and treatments cannot be formally excluded. This is consistent with a growing body of recent literature that identifies realistic synthetic patient generation as an open problem, still actively investigated through generative adversarial networks, variational autoencoders, diffusion models and, more recently, large language models.

Integrating such approaches with explicit clinical constraints represents a promising direction for future work on long-COVID and, more generally, for data-scarce clinical prediction problems.

2.3 Beyond Labels: A Self-Supervised Approach in Histopathology

The preceding section examined data augmentation as a mechanism to artificially expand dataset sizes, addressing scenarios where the absolute number of clinical samples is insufficient for robust model training. However, expanding the volume of raw inputs addresses only one dimension of the data scarcity problem. As the focus shifts from the collection of raw data to its utilization, a more pervasive economic and logistical barrier emerges: **the scarcity of annotated labels**.

While the digitization of healthcare has led to the accumulation of massive repositories of electronic health records and medical imaging, the vast majority of this information remains unstructured and unlabeled. In this context, the primary bottleneck is not the absence of data, but rather the absence of supervision. This distinction is critical because standard supervised learning algorithms, which currently dominate the landscape of medical AI, require large sets of examples paired with definitive ground truths to function effectively.

The scarcity of labels in medicine stems from the unique constraints of **clinical expertise**. Unlike computer vision tasks in general domains, where image annotation can often be crowdsourced to non-experts, medical annotation requires deep domain knowledge. Validating a dataset of MRI scans or histopathological slides demands the attention of board-certified specialists, such as radiologists or pathologists. These experts are a finite resource, and their time is exceptionally valuable, making the creation of large-scale, fully annotated datasets prohibitively expensive and slow. Furthermore, the annotation process itself is often far more complex than simple classification. Tasks such as semantic segmentation require pixel-perfect delineation of lesions or anatomical structures in three-dimensional space, a process that can take hours for a single patient volume. This difficulty is compounded by **inter-observer variability**; medicine is rarely binary, and experts often disagree on difficult cases, leading to "noisy" labels that complicate the definition of a true gold standard.

To address the theoretical and practical constraints imposed by the scarcity of annotated medical data, the following section introduces a case study that moves beyond traditional supervised paradigms. While data augmentation addresses the quantity of raw samples, the bottleneck of expert supervision remains a critical hurdle. In this context, we propose

a **Self-Supervised Learning** (SSL) framework applied to histopathological analysis [44], specifically designed to extract robust diagnostic features without the need for extensive labeled datasets. This work serves as a paradigmatic example of how algorithmic innovation, specifically the use of **Masked Autoencoders** (MAEs), can mitigate the costly acquisition of ground-truth labels in complex medical domains, such as breast cancer subtype classification.

2.3.1 Problem definition

Histopathological image analysis constitutes a fundamental pillar in clinical applications, particularly for cancer diagnosis. Pathologists rely on **Whole-Slide Images** (WSIs), which provide high-resolution views of entire tissue sections, enabling a comprehensive evaluation of the patient's condition. However, the manual analysis of these massive images is inherently time-consuming and prone to inter-observer variability. Consequently, Deep Learning models have emerged as critical tools for automating this process, offering the potential for faster, more consistent, and potentially more accurate diagnoses [150].

The primary data modality in this domain consists of high-resolution WSIs. Due to their massive size, analyzing these images necessitates a localized approach. The standard methodology involves extracting smaller image patches from the WSI. This extraction process reduces the data to a manageable size, facilitating the application of computer vision architectures such as Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) for tasks like tissue classification and cell segmentation. Once extracted, these patches are often treated as a "bag of tissue regions," making the **Multiple Instance Learning** (MIL) framework a natural choice for classification tasks [157]. While **CNNs** have historically been the dominant architecture due to their ability to capture spatial relationships [130, 69, 70], they often struggle to capture long-range dependencies or overall features within complex tissues, particularly as network depth increases. Alternatively, **Graph Neural Networks** can model tissue structures by representing cells as nodes, though this approach requires heavy pre-processing to construct the required graph structure [173, 9]. **ViTs** offer a compelling alternative by processing patches directly to learn global dependencies, yet they require substantial data to generalize effectively [56, 158].

The primary problem addressed in this work is the accurate classification of breast cancer sub-types using histopathological data without relying on extensive manual annotations. The objective is to distinguish between specific cancer types, such as benign, malignant, and atypical tumors, as well as their granular sub-types. To achieve this, the work exploits the incredibly high dimensionality of WSIs to overcome data scarcity. Instead of relying

on limited labeled regions, the strategy involves automatically generating a large dataset of unlabeled image regions through random cropping of WSIs. This process captures a diverse pool of tissue regions, encompassing both tumor and non-tumor areas, effectively expanding the available training data. By generating this **massive dataset of Regions of Interest** (ROIs), the work applies a **Self-Supervised Learning** method. This approach aims to learn informative latent representations (embeddings) of the tissues **solely through reconstruction tasks**, such as **masked autoencoding**, without the need for class labels during the representation learning phase The ultimate goal is to utilize these self-supervised embeddings to enhance performance on downstream **multi-class classification tasks** with simple classifiers, effectively decoupling feature learning from the challenges of annotation.

2.3.2 HMAE framework

This section outlines the theoretical and practical framework employed to address histopathological image analysis. The approach integrates Vision Transformers with a Masked Autoencoding strategy to perform image reconstruction, thereby learning robust feature representations.

Vision Transformers

The architecture leverages Vision Transformers, which represent a significant departure from traditional Convolutional Neural Networks. In this framework, an input image is divided into **fixed-size patches**, which are subsequently embedded into **vector representations**. These vectors serve as input to an encoder block that employs a **self-attention mechanism** [152]. This mechanism allows the model to learn long-range dependencies across different image regions, capturing the global context essential for classification tasks.

Masked encoder/decoder

The encoding process begins by dividing the image into equally sized, non-overlapping patches. A **masking strategy** is then applied where a specific number of patches are randomly selected to remain visible, while the remainder are hidden [61]. To prevent the model from biasing toward the image center, patch selection is randomized across the entire image. The framework retains only a small percentage of the original image (specifically 25% of the patches) forcing the model to infer missing information from sparse data rather than relying on local neighbors.
Following encoding, the system processes the encoded data from visible patches alongside "*mask tokens*", which represent the missing areas to be predicted. **Positional encodings**

are added to these tokens to retain spatial information. This combined input is fed into the decoder, which comprises a series of transformer blocks. The decoder is utilized exclusively during the training phase to learn how to reconstruct the missing patches by predicting their actual pixel values. The reconstruction accuracy is evaluated by comparing the output against the original image using Mean Squared Error (MSE) .

WSI sampling pipeline

To learn informative tissue representations that distinguish between tumor and non-tumor areas, the methodology employs a pipeline to extract a large volume of unlabeled image regions from WSIs:

- **Region Selection**: Square-shaped image patches are selected from the WSI. The dimensions of these regions are determined by sampling from a normal distribution, with parameters derived from the size statistics of ROIs in the BRACS dataset.
- **Region Quality Control**: To ensure data quality, the average variance of pixel intensities is calculated for each extracted patch. If the variance exceeds a predefined threshold, the region is deemed informative; otherwise, it is excluded. This step filters out uninformative areas, such as background or borders, which typically display low variability.

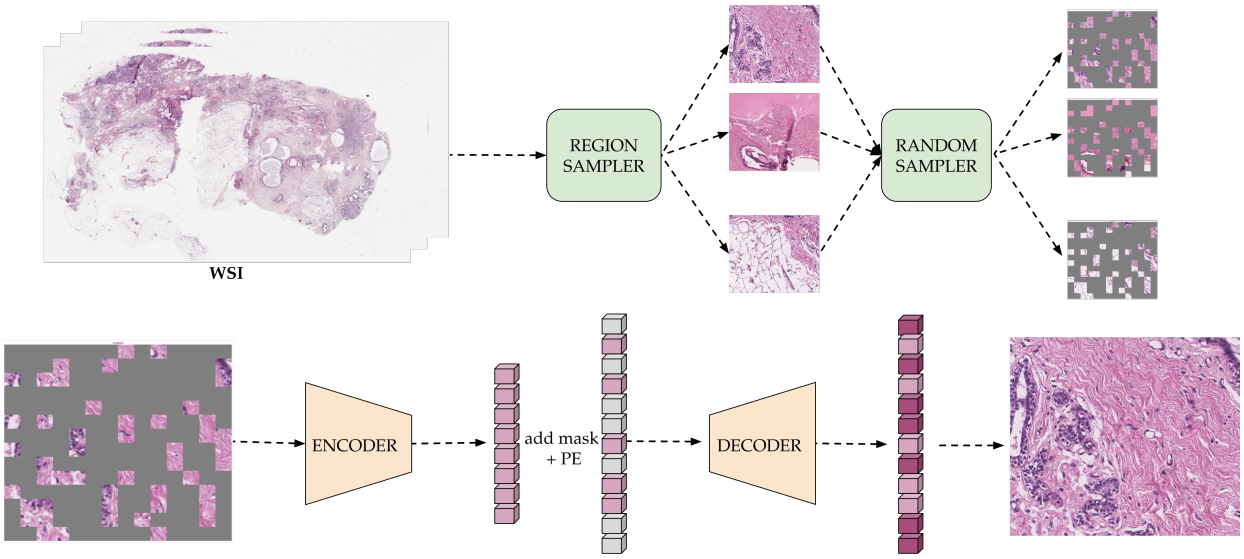


Figure 2.7: **HMAE architecture**. (top) Tissue regions are randomly sampled from the original image (WSI). Subsequently, a random mask is applied, occluding 75% of the image data. (down) The autoencoder architecture receives the masked image as input and aims to reconstruct the hidden regions.

Annotated RoI classification strategy

The final classification capability is achieved through a two-step process utilizing the pre-trained HMAE. First, annotated ROIs are processed by the model to generate embeddings, effectively capturing the essential characteristics of the data. A mean aggregation of patch-level embeddings is performed to produce a single vector per ROI. Second, these embeddings serve as input features for a Multi-Layer Perceptron (MLP) classifier with one hidden layer. This approach allows the complex representations to be learned via self-supervision without extensive labeling, **requiring labels only for the training of the simple MLP classifier**.

Datasets

After defining the learning architecture, the study then considered several reference datasets to evaluate classification performance on histopathological breast cancer images. The datasets examined in this work are described below.

BRACS. The BReAst Carcinoma Subtyping (BRACS) dataset [26] is a collection of digital images used to study breast lesions. It includes 547 WSIs, which are high-resolution scans of entire tissue samples. Additionally, 4539 smaller, more specific areas called regions of in-

terest (RoIs) are extracted from these whole-slide images. Each WSI and its corresponding RoIs are carefully examined and labeled by three pathologists. These categories encompass three main lesion types, that is three different classes: *benign* (healthy tissue), *malignant* (cancerous tissue), and *atypical*. Further details are provided by seven subcategories within these main types, allowing for a more precise understanding of the specific lesion (see Figure 2.8).

BACH. The BreAst Cancer Histology (BACH) dataset [6] is a significant resource for researchers developing computer algorithms to automatically diagnose breast cancer. It consists again of a collection of digitized WSIs from breast biopsies. The RoIs extracted from each WSI are labeled according to four different classes: *normal tissue*, *benign tumors*, *in situ carcinoma* (precancerous cells), and *invasive carcinoma* (cancerous cells). The dataset contains a total of 400 RoIs (100 RoIs for each class).

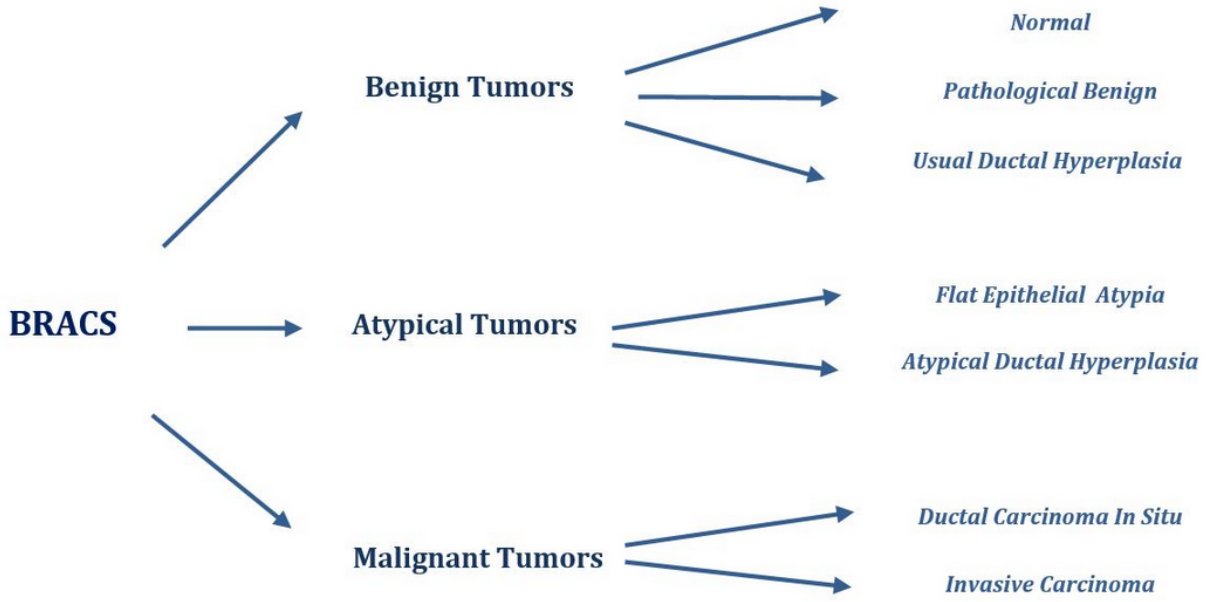


Figure 2.8: Class taxonomy of BRACS dataset.

2.3.3 Experimental framework & results

The primary evaluation was conducted using the BRACS dataset, utilizing Hematoxylin and Eosin (H&E) stained tissue images. The dataset includes 547 WSIs and 4539 annotated Regions of Interest.

To ensure robust evaluation, in all the experiments the corresponding dataset has been split into training, validation, and test sets using a 70/10/20 split. To account for potential

variability, each classification experiment was run 100 times, and the average value for each chosen metric was reported. The HMAE training was performed on a single Nvidia A40 GPU, taking a total time of 32 hours to complete.

Cancer Classification (3-Class Task)

The first evaluation concerned the task of classifying tissues from BRACS images as **cancerous, non-cancerous or atypical**. We have benchmarked HMAE with respect to several state-of-the-art approaches to this problem and the results are shown in Table 2.8 (for a 3-class classification task). For a correct comparison, all models employed a **ViT-S/16 encoder** architecture. This facilitated a comprehensive evaluation against various baseline models, leveraging the groundwork established in [171].

These models use either an attention mechanism or a ViT to correlate different region of a tissue image. They then classify the tissue by considering the relationships between these regions (the first two in the table perform the maximum and the average between the image patches embedding respectively).

In table, the best performing approach is shown in bold and the HMAE results are colored. In light of HMAE not achieving the best performance, a statistical evaluation was conducted to determine whether significant differences exist in the measured performance scores compared to the top-performing approach. The p-values computed in a T-test at the 95% confidence level are reported in Table 2.9. Since p-values are greater than 0.05 we can conclude that, with a 95% confidence level, there is no significant difference between HMAE and ACMIL in terms of both WF1-score and AUC.

It is worth emphasizing at this point what this result actually means in practice. Whereas the top-performing approaches listed in Table 2.8 (ACMIL, IBMIL, DTFD-MIL, ABMIL and so on) are fully supervised methods that depend on pixel or patch-level annotations produced by trained pathologists, HMAE learns its representations in a **purely self-supervised** fashion and uses labels only for training a lightweight MLP classifier on top of the frozen embeddings. Matching the best supervised baseline without a statistically significant gap, while requiring a drastically smaller annotation budget, is the actual measure of success in this setting: in a real clinical pipeline, the production of even a few hundred annotated ROIs is a tangible bottleneck, given that each annotation must be generated by board-certified pathologists whose time is the scarcest resource of the whole workflow.

Model	**AUC**	**WF1-score**
Max-pooling	0.823±0.033	0.596±0.029
Mean-pooling	0.739±0.007	0.522±0.038
Clam-SB [92]	0.863±0.005	0.631±0.034
TransMIL [125]	0.841±0.006	0.631±0.030
DSMIL [87]	0.816±0.028	0.577±0.028
DTFD-MIL [169]	0.870±0.022	0.612±0.080
IBMIL [88]	0.871±0.014	0.645±0.041
MHIM-MIL [138]	0.865±0.017	0.625±0.060
ABMIL [73]	0.866±0.029	0.680±0.051
ACMIL [171]	**0.888±0.010**	**0.722±0.030**
HMAE	0.866±0.003	0.704±0.009

Table 2.8: Cancer Classification (3 classes): AUC and Weighted F1-score

	ACMIL	**HMAE**	p-value
F1-score	0.722	0.704	0.2598
AUC	0.888	0.866	0.1728

Table 2.9: Cancer Classification (3 classes): statistical difference (95% confidence)

Sub-type Cancer Classification (7-Class Task)

A second evaluation has concerned the sub-type cancer classification in the BRACS dataset. This task represents an even greater challenge than previous efforts due to the often subtle morphological distinctions between some cancer types. Furthermore, the inherent heterogeneity of the training data, which includes a significant percentage of non-tumor tissue, introduces an inherent class imbalance within the latent representation space. The performance of the model was evaluated against the benchmark reported in [131] under identical experimental conditions (7-class classification task, weighted F1-score for evaluation).
Most of the models used for this comparison differ from those employed in cancer classification (reported on Table 2.8) since benchmarks for cancer and sub-type cancer classification are different; in fact, models for cancer classification (discussed in [171]) would require training from scratch specifically for sub-type classification, whereas the model does not. Results are presented in Table 2.10 (best performance in bold and HMAE results colored). In this case, statistical significance analysis between HMAE and the top-2 performer approaches reveals a significant difference (at 95% confidence level) in the computed WF1-scores (see

Table 2.11). In order to get a deeper understanding of such differences, we investigated the behavior, among the top-3 performers (ScoreNet, HACT-Net and HMAE), in terms of F1-score on each label (i.e., normal tissue and sub-cancer categories). Results are reported on Table 2.12 (best results shown in bold as usual).

Previous exposure to a larger proportion of non-cancerous tissue during the training phase appears to have influenced the HMAE prediction distribution. This is reflected by the best F1-score achieved by HMAE for the "normal" class compared to both the benchmark models and other classes within this investigation. However, performance on the remaining classes, particularly the "Invasive" class with the highest F1-score, remains comparable to previously tested models. In order to be more detailed about this aspect, we have performed a statistical significance test also in this case, and results are reported in Table 2.13. Bold values represent situations where HMAE performs either as significantly the best ("normal" class) or with no significant difference with respect to the other approaches.

The symmetric observation concerns the ADH class, on which the gap with respect to the supervised competitors is widest (0.301 against 0.474 of ScoreNet). It should be noted, however, that ADH is the *minimum of the per-class profile of all three models* in Table 2.12: even the best approach misclassifies more than half of these samples, which indicates a difficulty intrinsic to the class rather than a failure specific to the proposed architecture. This is consistent with the curation process of BRACS itself, whose authors report that some ROIs could not be resolved by the annotating pathologists because morphologically similar subtypes:

> *Indeed, for some ROIs, pathologists were in difficulty on deciding among different lesion classes, as different subtypes of lesions have similar morphological features (e.g. ADH and DCIS) and cannot be easily classified disregarding the WSI-level context. Thus, the pathologists examined together the corresponding WSI, to jointly assign a final label to the ROI.*

The discriminative evidence for ADH is therefore only partially contained in the input that a ROI-level model receives, and the residual error incorporates a component of irreducible annotation uncertainty.

Based on these results it can be concluded that, despite a potential bias towards non-cancerous tissue introduced by the training data and a residual weakness on the atypical branch, the model retains in general the ability to discriminate effectively between different cancer sub-types, quite often in a comparable way with respect to the best performing approaches.

This observation must be weighed against the fundamental operational difference between

HMAE and the competing architectures: ScoreNet, HACT-Net, CGC-Net and the other top-ranking models are supervised (or heavily supervision-dependent) pipelines that rely on expert-curated annotations for every training sample, often including fine-grained cell or region-level graph construction. HMAE, in contrast, produces its ROI embeddings without accessing any class label at all, and only the final MLP head is trained with supervision.

Model	WF1-score
CLAM-MB/B [92]	0.548±0.010
CGC-Net [173]	0.436±0.005
Patch-GNN [9]	0.521±0.006
TG-GNN [106]	0.559±0.001
CG-GNN [106]	0.566±0.013
HACT-Net [106]	0.615±0.009
TransPath [159]	0.567±0.02
TransMIL [125]	0.575±0.007
ScoreNet [131]	**0.644±0.009**
HMAE	0.578±0.015

Table 2.10: Sub-type cancer classification: weighted F1-score

	ScoreNet	HACT-Net
HMAE	2.21E−13	3.58E−05

Table 2.11: Sub-cancer Classification (7 classes): p-values at 95% confidence

Label	ScoreNet	HACT-Net	HMAE
Normal	0.646±0.022	0.616±0.021	**0.683±0.022**
Benign	**0.540±0.022**	0.475±0.029	0.485±0.020
UDH	**0.484±0.022**	0.436±0.019	0.445±0.070
ADH	**0.474±0.024**	0.404±0.025	0.301±0.015
FEA	**0.779±0.007**	0.742±0.014	0.702±0.018
DCIS	0.629±0.020	**0.664±0.026**	0.633±0.019
Invasive	**0.910±0.014**	0.884±0.002	0.893±0.015

Table 2.12: Single class classification F1-score (based on top-3 models in Table 2.10).

2.3.4 Generalization capabilities & qualitative evaluation

To assess the model's representational capacity, its ability to generalize to data derived from entirely unseen WSIs was investigated. This has been achieved by assessing its performance on a classification task using a dataset (BACH) **not included in the training phase**

Label	ScoreNet	HACT-Net
normal	**0.0198**	**0.0198**
benign	3.00E−04	**0.5779**
UDH	**0.2976**	**0.8078**
ADH	7.41E−32	6.00E−12
FEA	1.00E−14	6.00E−4
DCIS	**0.0594**	**0.0594**
Invasive	**0.1047**	**0.2447**

Table 2.13: Single-class performance: p-values between HMAE and top-2 performers

(tumor and non-tumor tissues, further categorized into four distinct classes). In particular, the latent representation learning is performed using BRACS, then during inference each image from BACH is used as input to the HMAE architecture, the latent representation is produced and the classifier (now trained on a subset of BACH for a 4-class classification task) can finally predict the class. As in the previous experiments, the corresponding dataset has been split into training, validation, and test sets using an 70/10/20 split. Results (in terms of weighted F1-score) are shown on Table 2.14 (bold shows best result and HMAE results are background colored). As usual, we tested the statistical significance of the difference

Model	BRACS → BACH
HACT-Net	0.402±.028
TransPath	0.618±0.048
TransMIL	0.465±0.100
CLAM-SB/B	0.575±0.036
ScoreNet	**0.734±0.035**
HMAE	0.687±0.032

Table 2.14: Testing Results on the BACH dataset (weighted F1-score).

in performance between HMAE and the top performer (ScoreNet); Table 2.15 shows that there is no significant difference between the approaches, in terms of weighted F1-score (on 4 classes). We can then conclude that also in this case, HMAE performs at the top level.

	Scorenet	HMAE	p-value
WF1-score	0.734	0.687	0.0522

Table 2.15: BRACS to BACH test: p-value

This cross-dataset finding is especially meaningful for clinical deployment: embeddings learned via self-supervision on BRACS generalize to BACH, a dataset acquired under different protocols and defined by a distinct class taxonomy, while maintaining competitive

performance. In a real hospital setting this is precisely the scenario where supervised models tend to fail, since each new dataset would normally require a fresh round of expert annotation. The fact that HMAE only needs to retrain a small MLP classifier on the target labels, without re-annotating the images used for representation learning, makes it a far more sustainable option when the deployment context changes (e.g., a different hospital, a different scanner, or a newly defined diagnostic category).

Qualitative Evaluation

In order to have a better understanding of the quality of the learned representations in the HMAE model, t-SNE dimensionality reduction was also applied to the BRACS reference dataset using the higher level of the class hierarchy (benign, cancer and atypical tissue). Figure 2.9 provides a visualization of the embeddings in a 2D latent space. Each embedding is plotted and colored according to its corresponding class label. Interestingly, even though the model was trained unsupervised, a good degree of class separation is already evident.
Finally, have been analyzed the attention maps, a crucial component of the transformer architecture. Attention maps reveal which parts of the input image the model focuses on the most. By visualizing them, it is possible to understand which image regions are most relevant for the model's predictions. Figure 2.10 showcases four original images alongside the attention maps generated by the final layer of the MAE in the HMAE architecture. Notably, even during unsupervised training, the model appears to differentiate between connective and glandular tissue. This distinction likely arises because glandular tissue is structurally more complex, requiring the model to retain more information for reconstruction.

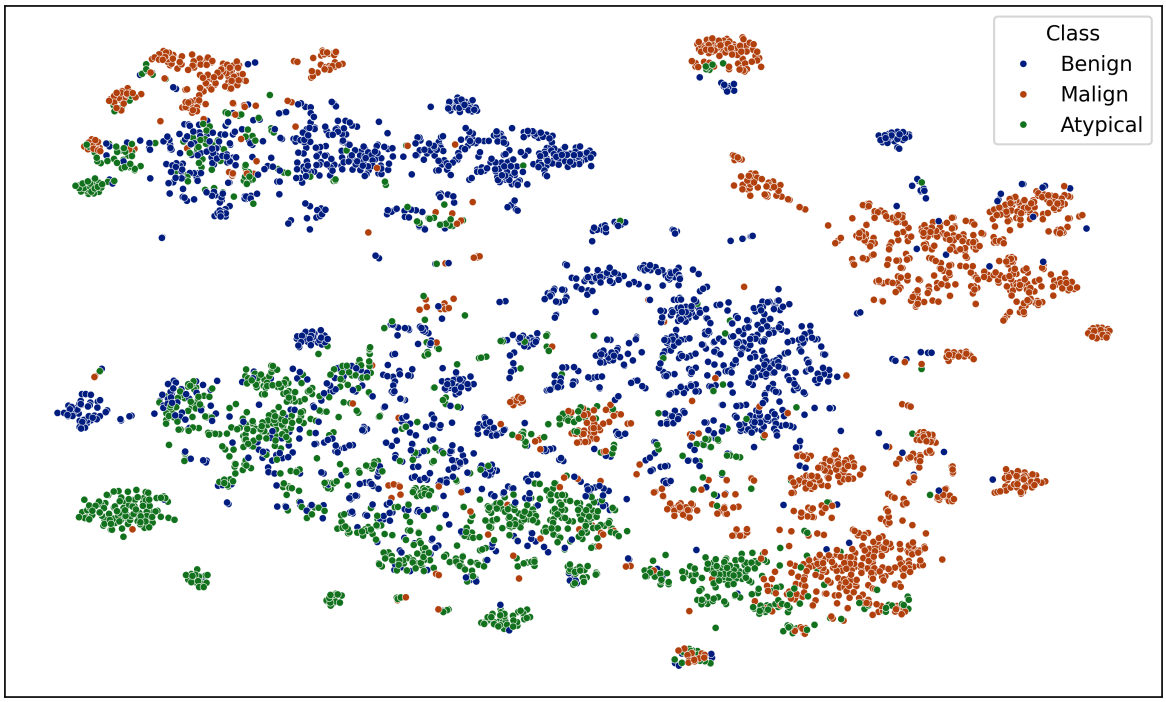


Figure 2.9: t-SNE visualization of the region of interest (RoI) embeddings from the BRACS dataset. Each point represents a sample, colored according to its class: benign (blue), malignant (red), and atypical (green). This visualization highlights the clustering of different histopathological categories based on their learned feature representations.

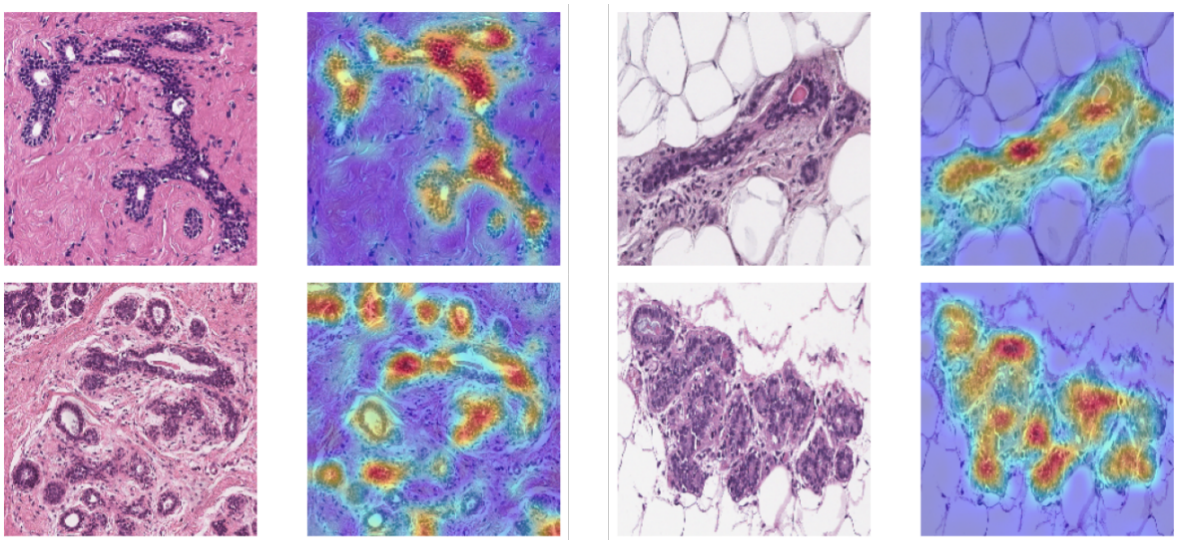


Figure 2.10: Self-attention heatmaps over histopathology images. The left column in each pair shows the original histopathology image, while the right column presents the corresponding heatmap, where warmer colors (red, yellow) indicate regions of high attention.

2.3.5 Conclusions on self-supervised approach experiment

This research has presented a Self-Supervised Learning framework designed to address the challenges of breast cancer sub-type classification in histopathological images. The proposed Histopathological Masked AutoEncoder (HMAE) leverages the architectural strengths of Vision Transformers (ViT) combined with a masked autoencoding strategy to generate robust feature representations without relying on extensive labeled datasets. The core contribution of this work lies in the effective utilization of high-resolution WSIs. By implementing a random cropping pipeline, the methodology successfully generated a massive, diverse dataset of unlabeled tissue regions, enabling the model to capture intricate patterns and long-range dependencies through the pretext task of image reconstruction. This approach effectively mitigates the constraints imposed by the scarcity of annotated medical data and the limitations of transfer learning from general-domain datasets.

Experimental validation on the BRACS dataset demonstrated the efficacy of the learned embeddings. In the task of 3-class cancer classification (benign, malignant, and atypical), the HMAE framework achieved performance levels statistically comparable to state-of-the-art supervised methods, such as ACMIL, with no significant difference in AUC or weighted F1-score. Furthermore, in the more granular 7-class sub-type classification, the model exhibited strong discriminative capabilities, particularly in identifying normal and invasive carcinoma tissues, despite the inherent challenges of class imbalance and morphological subtlety. Although on the most granular 7-class task HMAE does not outperform the best supervised competitors (ScoreNet, HACT-Net), and on the BACH transfer experiment it trails ScoreNet by a small (non-significant) margin, these results should not be read in isolation from the supervision regime under which each model operates.

The single most important property of the proposed framework and justifies its adoption even when it does not top the leaderboard, is that HMAE is self-supervised at the representation-learning stage. In concrete terms, this means that the hundreds to thousands of annotated ROIs that competing models require during training are simply not needed by HMAE to build its embedding space. Labels are used only to fit a shallow MLP classifier on top of frozen features, which typically requires one to two orders of magnitude fewer annotated examples to reach competitive performance. In a real hospital deployment scenario, this is not a marginal implementation detail but a decisive operational advantage: histopathological annotation is performed by board-certified pathologists whose time is both expensive and structurally limited, inter-observer variability makes consensus annotation even more costly, and every change in scanner, staining protocol, tissue type or diagnostic taxonomy would, under a fully supervised paradigm, trigger a new annotation campaign. By decou-

pling representation learning from label availability, HMAE provides a foundation on which new downstream tasks can be instantiated at the cost of a lightweight fine-tuning, rather than of a full re-annotation effort.

It is precisely in this light that the numerical results acquire their full meaning. Matching or approaching the accuracy of fully supervised state-of-the-art pipelines while requiring only a small fraction of their annotation budget is, from the perspective of clinical adoption, a more valuable outcome than a marginal improvement in F1-score obtained at the price of an unsustainable labeling effort. Critically, these results confirm that a latent space optimized solely for the unsupervised objective of pixel-level reconstruction can successfully encode diagnostically relevant features: a property further corroborated by the t-SNE visualization in Figure 2.9, which shows that class-coherent clusters emerge even though the model never observed the class labels during training, and by the attention maps of Figure 2.10, which highlight structurally informative tissue regions without any explicit supervisory signal. The findings validate the hypothesis that masked image modeling is a scalable and powerful mechanism for histopathological analysis, offering a viable pathway to automate cancer diagnosis while significantly reducing the dependency on manual pathological annotations: reduction that, in the day-to-day reality of a hospital pathology department, is arguably the binding constraint on the adoption of deep-learning-based diagnostic tools.

2.4 Lessons from the Two Case Studies

The two studies presented in this chapter approached the same underlying obstacle from opposite ends of the taxonomy introduced in Section 2.1.3, and it is worth drawing together what each established, where each fell short, and what conclusions survive the transition from the specific to the general.

The long-COVID study addressed scarcity in its most classical form, that of a **limited number of instances**, aggravated by label imbalance and by a feature space whose dimensionality was substantial relative to the cohort size. Its principal finding is that the two data-level solutions available in this regime are not symmetric. Synthetic oversampling through MLSmote improved predictive performance across most configurations, whereas aggressive dimensionality reduction, and particularly the supervised RELIEF criterion, degraded it, suggesting that in a multi-label setting with correlated targets, the features discarded as individually uninformative carry predictive signal jointly. Equally important, and less expected, is the interpretability result: the SHAP analysis showed that models trained on augmented data continued to attribute importance to the same clinically salient variables identified on the original cohort, so that the gain in robustness was not paid for with a loss

of clinical actionability. In a domain where a model that cannot be interrogated cannot be adopted, this is a substantive rather than an incidental finding.

The histopathology study addressed the complementary condition, in which the raw signal is abundant and only **the annotation** is scarce. The HMAE framework showed that this asymmetry can be exploited directly: by learning representations from unlabelled tissue regions through masked reconstruction, supervision is confined to a small classifier trained on the target labels, and the annotation budget is decoupled from the representation-learning budget. The resulting embeddings proved statistically indistinguishable from fully supervised state-of-the-art methods on the coarse three-class task, and the cross-dataset transfer from BRACS to BACH, involving different acquisition protocols and a different class taxonomy, was sustained with only a retrained classification head. This last result is the one with the clearest translational reading, since it is precisely the re-deployment scenario, a new hospital or a new scanner, in which supervised pipelines ordinarily require a fresh annotation campaign.

However, both results have clear limits that need to be pointed out. The long-COVID cohort was collected at a single centre during the early pandemic waves, and no external validation on an independent population was possible, so that the generalization of the learned associations beyond that setting remains unverified. More fundamentally, interpolation-based oversampling is by construction agnostic to clinical semantics: MLSmote guarantees that synthetic instances lie within the convex hull of observed ones, not that they correspond to physiologically coherent patients, and the possibility that some generated profiles combine comorbidities, treatments and symptoms in clinically implausible ways cannot be formally excluded. The augmentation therefore purchased statistical stability at the cost of a guarantee that no purely data-driven procedure is in a position to provide. The histopathology study faces a limitation of the same family from the opposite direction. Its pretext task rewards the reconstruction of tissue appearance, which is a proxy for diagnostic relevance rather than a specification of it; the model is never told that nuclear atypia matters more than stromal texture. This is a plausible explanation for the fact that HMAE matched supervised competitors on the coarse classification while trailing them on the finest-grained subtypes, where the distinguishing evidence is morphologically subtle and semantically defined. Neither study, moreover, achieved its result without cost, whether in the computational expense of self-supervised pre-training over large patch collections or in the sensitivity of the augmented results to the choice of resampling policy.

Read together, the two studies support a single conclusion that generalizes beyond either. Data-driven mitigation is genuinely effective at redistributing the information a dataset already contains: it can rebalance it across classes, compress it into a lower-dimensional space,

or re-organize it into representations from which a small number of labels suffices to learn. What it cannot do is introduce information that the data never carried. The convex hull of a small cohort does not extend to the patient profiles the cohort failed to sample, and a reconstruction objective cannot discover which morphological features a pathologist regards as diagnostic. This limit explains why the two studies succeeded most clearly where the missing information was implicitly available elsewhere, in the abundant unlabelled pixels of the second case, and least clearly where it was genuinely absent, in the finest-grained distinctions of the same case and in the clinical plausibility of synthetic patients in the first.

The consequence is methodological rather than merely diagnostic. If the binding constraint in these settings is the quantity of information available to the learner, and if that quantity cannot be increased from the data alone, then the remaining source of information is the **structured knowledge** that the biomedical domain, unusually among application areas, possesses in codified and machine-readable form: ontologies, clinical guidelines, care pathways, and curated knowledge graphs. Section 2.1.3 observed that scarcity admits two lines of attack, one acting on the data and one on the inductive bias of the model; this chapter has pursued the first to its limits. The chapters that follow pursue the second.

Chapter 3

Knowledge Integration in Deep Learning Models

In the previous chapter, two data-driven approaches were examined to partially address the limitations imposed by restricted sample availability and the reliance on handcrafted labels. As the term *data-driven* suggests, these methods derive their effectiveness exclusively from the information contained within the data itself. However, data is not the sole source of insight; **prior knowledge**, representing established domain understanding, offers an additional and often more powerful foundation for model development. Recognizing the limitations of relying solely on empirical observations has prompted a fundamental epistemological shift toward **Knowledge-Informed Machine Learning** (KIML) [155]. This paradigm seeks to combine the representational strength and scalability of deep learning with the interpretability, robustness, and data efficiency characteristic of symbolic knowledge. Rather than tasking neural models with rediscovering known principles from data alone, knowledge integration strategies enable the explicit incorporation of domain expertise, such as structured knowledge graphs, ontological hierarchies, logical constraints, or physical laws, directly within the learning process. The integration of external knowledge into deep learning models represents more than an engineering convenience; it constitutes a fundamental reconsideration of how artificial intelligence systems should be constructed. This chapter provides a comprehensive theoretical and practical examination of knowledge integration in deep learning, with particular emphasis on *Transformer-based architectures in biomedical applications*. It first establishes the theoretical underpinnings through the lens of inductive bias, illustrating that knowledge integration can be interpreted as the deliberate engineering of learning biases to constrain the hypothesis space of neural models. It then introduces a taxonomy of knowledge integration techniques, detailing how knowledge can be sourced, represented, and incorporated at various stages of the machine learning pipeline.

Finally, sections 3.4–3.6 present the original methodological contribution of this thesis, first published in [74]: Structural Positional Encoding, a mechanism that injects ontology-derived structural information into the positional representations consumed by self-attention, together with its experimental validation on a real stroke-management dataset. The first part is not intended as a systematic review, but as the minimal conceptual apparatus required to introduce and identify the specific gap that SPE addresses, which Section 3.3.3 states explicitly.

3.1 Inductive Bias and the Necessity of Prior Knowledge

To understand why knowledge integration is not merely beneficial but theoretically necessary, its important begin with the foundations of machine learning itself. At its core, machine learning addresses the problem of **generalization**: *given a finite set of observed examples, how can an algorithm make reliable predictions about unseen data?* This seemingly straightforward question conceals a profound logical challenge known as the **problem of induction**, which has occupied philosophers and epistemologists for centuries.

The computational manifestation of this philosophical problem is formalized in the N**o Free Lunch (NFL)** theorem for machine learning [161]. In its most general form, the NFL theorem states that when averaged over all possible data-generating distributions, every learning algorithm **has identical** expected error on previously unobserved points. This counterintuitive result implies that no algorithm is universally superior; an algorithm that performs exceptionally well on one class of problems will necessarily perform poorly on others. The theorem's profound implication is this: *effective learning requires assumptions about the structure of the problem space.* These assumptions, known as **inductive biases** [15], represent the set of presuppositions that a learning algorithm utilizes to prefer certain hypotheses over others when multiple hypotheses are consistent with the training data. Without inductive bias, a learner faces an intractable infinity of equally plausible functions that fit the observed data but diverge arbitrarily on new examples. The inductive bias thus acts as a "*prior probability*" in the Bayesian sense, constraining the hypothesis space to a tractable subset of functions that the algorithm deems probable a priori.

3.1.1 Architectural and Relational Inductive Biases in Neural Networks

In deep learning, inductive biases manifest in two primary forms: implicit architectural biases and explicit knowledge-based biases. This section examines the former, tracing a progression from simple structural constraints to sophisticated relational reasoning capabilities.
Relational inference capabilities mean the ability of a neural architecture to compute a function that depends explicitly on the relational structure of the input, i.e., on the set of pairwise (or higher-order) connections among the entities that constitute it, rather than on the entities taken in isolation. Operationally, this is realized through mechanisms such as message-passing updates between connected nodes in a Graph Neural Network, attention weights computed over pairs of tokens in a Transformer, or scoring functions that operate jointly on triples (head, relation, tail) in knowledge graph embedding models. In all these cases, the "reasoning" label is used in an engineering rather than cognitive sense: no symbolic manipulation or step-wise deduction is performed; rather, the model computes, in one or more forward passes, a function whose inductive bias respects the relational structure of the domain.
Classical neural architectures embed generic inductive biases through their structural design, imposing domain-agnostic assumptions about data organization and information flow:

- **Convolutional Neural Networks** impose a strong bias of translation invariance and locality. By sharing weights across spatial dimensions and restricting connectivity to local receptive fields, CNNs assume that visual features useful in one image region (e.g., edges, textures) will be equally useful in another, and that nearby pixels exhibit stronger correlations than distant ones. This bias is domain-generic—it applies to natural images, medical scans, and even audio spectrograms when converted to time-frequency representations. The hierarchical composition of convolutional layers further embeds an assumption that complex features emerge from combinations of simpler local patterns.

- **Recurrent Neural Networks** encode a bias of temporal sequentiality and temporal invariance. They assume that the current hidden state is a function of immediate temporal history and that the transformation rule remains constant across time steps. **Long Short-Term Memory** (LSTM) networks extend this with gating mechanisms that introduce an additional bias: the assumption that certain information should persist over long temporal horizons while other information should be rapidly forgotten. This architectural choice reflects an implicit theory about the nature of sequential

dependencies: relevant context may need to be maintained across varying time scales.

- **Fully Connected Networks** possess the weakest structural bias, allowing arbitrary relationships between all input and output neurons. While this maximal flexibility provides *universal approximation* capacity [68], it results in poor sample efficiency for structured problems, as the model must learn every relationship from scratch without any prior assumptions about which connections are more likely to be meaningful.

These architectural biases are fundamentally domain-agnostic: they encode general principles about the nature of images, sequences, or tabular data, but they lack specificity about the particular problem. A CNN applied to medical imaging possesses no inherent understanding that certain anatomical structures must maintain spatial relationships or that certain pixel intensity patterns correspond to pathological conditions. This limitation motivates the need for more sophisticated forms of inductive bias.

Relational Inductive Biases and Compositional Structure

A particularly important evolution in architectural design concerns **relational inductive biases** [14], structural constraints specifically designed to encourage learning about entities, relations, and the compositional rules governing their interactions. This concept [14], provides a unifying framework for understanding how different neural architectures impose assumptions about the relational structure of data. This structure brokes one of the fundamental assumptions underpinning classical machine learning approaches: **the independent and identically distributed (IID) nature of the dataset**. When relational dependencies or structured knowledge are incorporated, data samples can no longer be considered mutually independent nor identically distributed, as their representation becomes influenced by contextual relationships and domain-specific priors. From the perspective of inductive bias, this shift introduces a relational inductive bias, deliberately embedding structured assumptions into the learning process to guide the model toward capturing meaningful interactions rather than treating observations as isolated instances. The key insight is that many real-world problems involve reasoning about discrete entities and their interactions, rather than merely processing raw sensory inputs. In a medical process monitoring context, the entities might be clinical activities (CT scan, blood test, surgical procedure), and the relations might be temporal ordering, causal dependencies, or functional similarities. In molecular biology, entities are atoms or amino acids, and relations are chemical bonds or spatial proximities. Standard feedforward networks possess minimal relational bias they treat inputs as unstructured vectors and outputs as independent predictions, with no explicit

representation of entities or their relationships. Graph Neural Networks (GNNs) [123], in contrast, impose a strong relational inductive bias by assuming that:

- **Graph-based representation**: The data consists of discrete entities (graph nodes) and relations (graph edges), with each entity possessing its own feature representation.
- **Neighborhood aggregation**: The properties and representations of an entity are determined by its interactions with related entities: its **neighborhood** in the graph. Information propagates through the graph structure via iterative message-passing operations.
- **Permutation invariance**: The rules governing these interactions exhibit invariance to the ordering in which neighbors are considered. The aggregated information from a node's neighborhood should not depend on an arbitrary enumeration of that neighborhood.
- **Shared transformation rules**: The same relational processing rules apply across all nodes and edges, enabling parameter sharing and generalization across different graph structures.

This architectural prior enables **combinatorial generalization**: the ability to correctly reason about novel combinations of known entities and relations, even when such combinations were never observed during training. This capacity mirrors a fundamental aspect of human cognition: the ability to understand that "the red ball is on the table" generalizes to "the blue ball is under the chair" by composing known concepts (objects, properties, spatial relations) in novel configurations. Similarly, a medical reasoning system that understands the general principle that "diagnostic procedures precede treatments" can apply this knowledge to novel disease-treatment combinations not seen during training. The progression from unstructured neural networks to relationally-structured architectures represents a shift in the granularity of inductive bias:

- **MLPs**: No structural assumptions beyond continuity and smoothness.
- **CNNs/RNNs**: Generic assumptions about locality (spatial or temporal).
- **GNNs**: Explicit assumptions about entity-relation structure and compositional inference

Each level of specificity trades flexibility for sample efficiency. GNNs excel on graph-structured problems but require that data be represented as graphs. When the relational structure is known a priori, as is the case with medical ontologies, molecular structures, or social networks, this constraint becomes a powerful advantage rather than a limitation.

Explicit Knowledge-Based Biases

In contrast to architectural biases, knowledge-based inductive biases are domain-specific constraints derived from external sources such as scientific principles, expert knowledge, or structured databases. When we integrate a medical ontology into a neural network e.g for clinical process monitoring, we are imposing an explicit prior: not all sequences of medical activities are equally probable. The ontology encodes the constraint that activities serving similar clinical goals (e.g., diagnostic imaging procedures) are more likely to occur in temporal proximity and are more strongly related than activities from disparate categories (e.g., administrative procedures versus surgical interventions). This explicit encoding serves multiple critical functions:

- **Hypothesis Space Reduction**: By constraining the set of permissible functions, knowledge integration reduces the effective dimensionality of the learning problem, in other words the number of examples required to achieve a given level of generalization performance.
- **Extrapolation Guidance**: In regions of the input space poorly covered by training data, explicit knowledge provides principled guidance, steering the model toward physically plausible or clinically coherent predictions rather than arbitrary extrapolations.
- **Interpretability Enhancement**: Models that respect known domain structures are inherently more interpretable, as their behavior can be partially explained by reference to the incorporated knowledge rather than being entirely opaque.
- **Robustness to Distribution Shift**: When test data differs from training data (e.g., new patient populations, rare disease presentations), models grounded in domain knowledge maintain performance by falling back on structural constraints rather than spurious correlations.

The process of integrating knowledge can be framed within the principles of **statistical learning theory**. In particular, knowledge integration may be interpreted as the introduction of structured inductive biases to constrain the hypothesis space and guide the learning process toward solutions that are consistent with established domain knowledge. Let $\mathcal{H}$ denote the hypothesis space, the set of all functions that a learning algorithm might consider. In standard empirical risk minimization, the learner seeks to find:

$$\hat{h} = argmin_{h \in \mathcal{H}} \mathcal{L}(h, \mathcal{D}) \tag{3.1}$$

where $\mathcal{L}$ is the empirical loss on training data $\mathcal{D}$.
Without additional constraints, $\mathcal{H}$ may be vast (e.g., the set of all neural networks with a given architecture), and many hypotheses $\mathcal{H}$ will achieve low training loss but fail to generalize. Inductive bias corresponds to a restriction or weighting of $\mathcal{H}$:

$$\mathcal{H}_{biased} = h \in \mathcal{H} : h^* \tag{3.2}$$

where h^* are the hypotheses that satisfy the bias constraints. Equivalently, the previous formula can be written as:

$$\hat{h} = argmin_{h \in \mathcal{H}}[\mathcal{L}(h, \mathcal{D}) + \lambda \cdot \Omega(h, \mathcal{K})] \tag{3.3}$$

where $\Omega(h, \mathcal{K})$ is a penalty term quantifying how much hypothesis violates constraints derived from external knowledge $\mathcal{K}$, and λ controls the strength of this constraint.

3.2 Inductive Bias in Transformers

The Transformer architecture [152] represents a paradigm shift in sequence modeling and constitutes a interesting case study in inductive bias design. Unlike its predecessors (RNNs and CNNs) the Transformer avoid both recurrence and convolution in favor of a pure attention-based mechanism. This architectural choice fundamentally alters the inductive biases encoded in the model, trading strong structural assumptions for maximal representational flexibility. Understanding the Transformer's mathematical foundations and its position in the bias-flexibility spectrum is essential for comprehending how knowledge integration strategies can enhance its capabilities, particularly in specialized domains like biomedicine.
The Transformer processes sequences through a composition of identical layers, each containing two primary sub-modules: a multi-head self-attention mechanism and a position-wise feedforward network. These components are connected via residual connections [137] and layer normalization. This section provides a mathematical description of these building blocks.

Self-Attention Mechanism

The fundamental operation in a Transformer is scaled dot-product attention, which computes a weighted sum of value vectors based on the compatibility between query and key vectors. Given an input sequence represented as a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, where n is the sequence length and d is the embedding dimension, the attention mechanism first computes three

representations:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}^Q, \quad \mathbf{K} = \mathbf{X}\mathbf{W}^K, \quad \mathbf{V} = \mathbf{X}\mathbf{W}^V \tag{3.4}$$

where $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V \in \mathbb{R}^{d \times d_k}$ are learned projection matrices, and d_k is the dimension of the key/query space (typically $d_k = d/h$ where h is the number of attention heads). The attention operation then computes:

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}})\mathbf{V} \tag{3.5}$$

The scaling factor $1/\sqrt{d_k}$ prevents the dot products from growing too large in magnitude, which would push the softmax function into regions with vanishingly small gradients. The softmax operation, applied row-wise, produces an attention distribution $\mathbf{A} \in \mathbb{R}^{n \times n}$ where $\mathbf{A}_{ij}$ represents the weight assigned to position j when computing the representation of position i:

$$\mathbf{A}_{ij} = \frac{exp(q_i^\top k_j / \sqrt{d_k})}{\sum_{j'=1}^{n} exp(q_i^\top k_{j'} / \sqrt{d_k})} \tag{3.6}$$

This formulation reveals a critical architectural property: every position attends to every other position with learned, content-dependent weights. There is no hard-coded notion of locality (as in CNNs with fixed receptive fields) or sequential ordering (as in RNNs with their temporal dependencies). The attention pattern emerges entirely from the learned compatibility between queries and keys.

Multi-Head Attention

Rather than performing a single attention operation, Transformers employ multi-head attention, which allows the model to jointly attend to information from different representation subspaces at different positions. This is accomplished by applying h parallel attention operations with different projection matrices. **Parallelizing attention heads** has been key to scaling these models to their current state:

$$head_i = Attention(\mathbf{X}\mathbf{W}_i^Q, \mathbf{X}\mathbf{W}_i^K, \mathbf{X}\mathbf{W}_i^V) \tag{3.7}$$

where $\mathbf{X}\mathbf{W}_i^Q, \mathbf{X}\mathbf{W}_i^K, \mathbf{X}\mathbf{W}_i^V \in \mathbb{R}^{d \times d_k}$ are the projection matrices for the i-th head. The outputs of all heads are concatenated and linearly transformed:

$$MultiHead(\mathbf{X}) = Concat(head_1, ..., head_h)\mathbf{W}^O \tag{3.8}$$

where $\mathbf{W}^O \in \mathbb{R}^{hd_k \times d}$ is an output projection matrix. Multi-head attention can be interpreted as learning multiple distinct notions of relevance simultaneously; one head might focus on syntactic relationships while another captures semantic associations.

Position-wise Feedforward Networks

Following the attention sub-layer, each Transformer layer applies a position-wise feedforward network (FFN) independently to each position:

$$FFN(\mathbf{x}) = ReLU(\mathbf{x}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2 \tag{3.9}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d \times d_{ff}}, \mathbf{W}_2 \in \mathbb{R}^{d_{ff} \times d}$ are weight matrices. This two-layer network with ReLU activation provides additional non-linear transformation capacity and can be viewed as learning position-specific feature combinations after the global information aggregation performed by attention.

Residual Connections and Layer Normalization

Both the attention and FFN sub-layers are wrapped in residual connections [137] with layer normalization [10]:

$$\mathbf{X}' = LayerNorm(\mathbf{X} + MultiHead(\mathbf{X}), \tag{3.10}$$

$$\mathbf{X}'' = LayerNorm(\mathbf{X}' + FFN(\mathbf{X}') \tag{3.11}$$

Layer normalization, applied across the feature dimension, normalizes activations to have zero mean and unit variance:

$$LayerNorm(\mathbf{x}) = \gamma \odot \frac{\mathbf{x} - \mu}{\sqrt{\alpha^2 + \epsilon}} + \beta \tag{3.12}$$

where μ and σ^2 are the mean and variance computed across the feature dimension, and γ, β are learned scale and shift parameters. This normalization stabilizes training in deep networks and has been shown to be particularly effective in the "Pre-LN" configuration, where normalization is applied before rather than after the sub-layer operations.

Positional Encoding

A critical limitation of the basic attention mechanism is its **permutation invariance**: without positional information, the model treats input sequences as unordered sets. To

address this, Transformers incorporate positional encodings [152] that are added to the input embeddings before the first layer. The original formulation uses sinusoidal functions:

$$\mathbf{PE}_{pos,2i} = sin(\frac{pos}{10000^{\frac{2i}{d}}}), \tag{3.13}$$

$$\mathbf{PE}_{pos,2i+1} = cos(\frac{pos}{10000^{\frac{2i}{d}}}) \tag{3.14}$$

where *pos* is the position index and i is the dimension index. This encoding allows the model to learn to attend by relative positions, as any fixed offset k corresponds to a linear transformation of the positional encoding. The complete input to the first Transformer layer is thus:

$$\mathbf{H}^{(0)} = \mathbf{E} + \mathbf{PE} \tag{3.15}$$

where $\mathbf{E}$ contains the token embeddings and $\mathbf{PE}$ contains the positional encodings.

3.2.1 The Inductive Bias Profile of Transformers

Having the mathematical foundations, one can now examine more holistically the inductive biases (or the relative absence thereof) embedded in the Transformer architecture, and how these biases influence what the model can and cannot learn.

At its core, the self-attention mechanism in a Transformer confers a form of permutation equivariance: in the absence of positional encoding, permuting the order of the input tokens results in an identically permuted output. This means that the architecture does not, by default, commit to any inherent sequential or spatial ordering, a sharp contrast to recurrent networks, which encode sequential dependencies, or convolutional networks, which impose a fixed spatial structure. This weak ordering bias yields a dual effect. On one side, it affords the **model flexibility**: it can discover long-range dependencies without being constrained by a fixed, pre-imposed notion of ordering.This is a powerful asset for tasks such as language modeling, where dependencies (e.g. subject–verb agreement, long-distance syntactic relations) may span arbitrary distances. On the other side, the **lack of a strong structural prior** implies that the model must learn even basic sequential or positional patterns, which typically demands large amounts of data; only when positional encodings (whether learned or fixed) are introduced can the model recover order information in a controlled way.

Closely related is the fact that self-attention in Transformers **induces all-to-all connectivity**: each token can directly attend to every other token, giving rise to a fully connected computational graph with on the order of $\mathcal{O}(n^2)$ pairwise interactions when processing a

sequence of length n. This stands in stark contrast with convolutional architectures where connectivity is local and sparse limited by kernel size and receptive field and with recurrent architectures, where information flows sequentially, so distant positions must be bridged through a chain of intermediate states. In practice, this dense connectivity gives the model maximal representational flexibility: in principle, any token can influence any other, regardless of distance. However, this also entails computational and statistical costs: complexity grows quadratically with sequence length, and the model often depends on very large amounts of data to learn meaningful attention patterns from scratch.

So, standard Transformers lack built-in relational or structural priors. Unlike architectures explicitly designed for graphs or relational data, where nodes (entities) and edges (relations) are defined a priori, the Transformer must infer from data alone which positions (tokens) are "related" and how. The attention matrix learned for each input instance can be viewed as a dynamic, instance-specific adjacency matrix, but it is ephemeral: it encodes relationships only for that particular input and does not reflect stable, domain-inherent relational structure. This design makes the architecture highly general and domain-agnostic; yet, in domains where relational structure is known and relevant, such as scientific data, biomedical processes, or time series with fixed dependencies, the absence of structural bias can make learning less efficient or less robust.

Altogether, this properties: **permutation equivariance, all-to-all connectivity, weak structural priors and emergent hierarchy**, helps to explain why Transformers achieve remarkable success in data-rich, high-variability domains such as large-scale language modeling or general-purpose vision. When abundant and diverse data is available, the flexibility of the architecture allows it to discover and represent complex, long-range, non-local dependencies without being constrained by human-imposed structural biases. However, the same flexibility becomes a limitation in low-resource or structure-rich domains: when data is scarce, or when one expects regular, domain-inherent relational or structural patterns (e.g. fixed dependencies in biomedical monitoring, time-series with clear temporal dynamics, scientific data with known relationships), the absence of strong inductive bias can hinder the model's ability to learn effectively or generalize robustly.

3.3 Knowledge Integration in Transformers for Biomedical Applications

The application of Transformer architectures to biomedical domains has revealed a fundamental tension between the models' remarkable representational capacity and the specialized

requirements of medical language understanding. While Transformers trained on general corpora (e.g., BERT, GPT) have demonstrated unprecedented performance on standard NLP benchmarks, their direct application to biomedical text processing exposes critical limitations. **Medical language** is characterized by extreme terminological density along with pervasive synonymy, where a single clinical concept may be expressed through dozens of lexically distinct surface forms. For instance, "myocardial infarction," "heart attack," "acute MI," and "AMI" all refer to the same pathological condition, yet standard language models treat these as distinct tokens whose semantic equivalence must be inferred purely from distributional co-occurrence patterns. This reliance on statistical correlation rather than explicit semantic grounding becomes particularly problematic in specialized subdomains where training data is scarce and the consequences of misunderstanding are severe. Early attempts to address this challenge focused on **domain-adaptive pre-training**, where models like BioBERT [86] and PubMedBERT [55] were initialized from general-domain BERT checkpoints and then continued pre-training on large biomedical corpora (PubMed abstracts, PMC full-text articles, clinical notes). These approaches yielded substantial improvements over general-domain baselines, demonstrating that exposure to in-domain text helps the model learn field-specific terminology and discourse patterns. However, domain adaptation alone proved insufficient for tasks requiring precise entity disambiguation and relational reasoning. A model that has seen "diabetes mellitus type 1" and "type 1 diabetes" in similar contexts may learn to place them nearby in embedding space, but it lacks explicit knowledge that these are synonymous expressions for the same ICD-10 code (E10). Moreover, domain-adapted models remain vulnerable to the "long tail" problem: rare medical concepts, specialized procedures, and emerging terminology (e.g., COVID-19 variants) appear infrequently even in massive biomedical corpora, resulting in poor representations for precisely those cases where clinical decision support is most needed. This recognition catalyzed a paradigm shift toward knowledge-enhanced Transformers: architectures that explicitly **integrate structured biomedical knowledge** from curated resources such as UMLS, SNOMED-CT, Gene Ontology, and domain-specific ontologies. The fundamental insight driving this research direction is that medical knowledge exists in two complementary forms: *the implicit*, distributional knowledge captured by language models through statistical learning over text, and *the explicit*, symbolic knowledge encoded in terminological systems, ontologies, and knowledge graphs constructed through decades of expert curation. Effective biomedical NLP systems must leverage both forms, using symbolic knowledge to ground and constrain the statistical patterns learned from text. The evolution of knowledge-enhanced biomedical Transformers can be traced through increasingly sophisticated integration strategies, progressing from simple entity linking to deep architectural modifications. Early approaches adopted a pipeline

architecture, where a standard Transformer first generated contextual representations, which were then post-processed using knowledge base lookups for entity disambiguation. For example, the entity "Apple" in the sentence "Patient reports eating an apple daily" would first be encoded by BERT, then disambiguated against UMLS to distinguish the fruit (C0003625) from the technology company. While effective for explicit entity recognition tasks, this pipeline approach suffered from error propagation (mistakes in entity recognition cascade to downstream components) and semantic disconnect (the language model's internal representations remain unaware of the structured knowledge used in post-processing).

3.3.1 The Spectrum of Integration Approaches

Knowledge-enhanced biomedical Transformers can be categorized along several dimensions: the **depth of integration** (from training objective modifications to fundamental architectural changes), the type of **constraints imposed** (soft guidance versus hard restrictions), and the **stage of the pipeline** where knowledge enters the system. This section examines representative approaches spanning this design space, progressing from shallow to deep integration strategies.

SapBERT: Metric Learning for Entity Alignment

SapBERT(Self-Alignment Pretraining for BERT) [89] addresses the medical synonym problem through specialized pre-training rather than architectural modification. The core innovation lies in reframing knowledge integration as a **metric learning task**: instead of predicting masked tokens, SapBERT trains the model to produce embeddings where synonymous medical terms cluster tightly together while semantically distinct concepts are pushed apart. The training data comes directly from UMLS, where each Concept Unique Identifier (CUI) is associated with multiple lexical variants forming a synonym set.
The metric learning objective is the **Multi-Similarity** (MS) loss:

$$\mathcal{L}_{\text{MS}}(i) = \frac{1}{\alpha} \log \left[1 + \sum_{k \in P_i} e^{-\alpha(S_{ik} - \lambda)} \right] + \frac{1}{\beta} \log \left[1 + \sum_{k \in N_i} e^{\beta(S_{ik} - \lambda)} \right] \tag{3.16}$$

where $S_{ik} = \cos(\mathbf{e}_i, \mathbf{e}_k)$ is the cosine similarity between entity embeddings, P_i represents positive pairs (synonyms), N_i represents negative pairs (distinct concepts), and α, β, λ are hyperparameters. The loss simultaneously pulls synonyms together and pushes distinct concepts apart, with emphasis on hard examples near the decision boundary.
A critical component is **online hard negative mining**: rather than training on random negative pairs (which are trivially distinguishable), SapBERT dynamically identifies seman-

tically related but distinct concepts that are currently positioned close in embedding space (e.g: "Type 1 diabetes" versus "Type 2 diabetes") forcing the model to learn fine-grained clinical distinctions. SapBERT achieves SOTA performance on biomedical entity linking while using the standard BERT architecture without modifications, demonstrating that knowledge can be integrated through training objectives alone. However, its integration remains entity-centric, optimizing isolated entity similarities without explicitly modeling relationships between entities.

KeBioLM: Architectural Fusion of Text and Entities

KeBioLM (Knowledge-enhanced Biomedical Language Model) [167] takes a more structurally explicit approach by modifying the Transformer architecture to maintain dual awareness of linguistic context and ontological identity. The key innovation is a **Text-Entity Fusion Layer** that processes words and entities through separate pathways before merging them via learned gating.

The processing begins with entity recognition and linking: entity mentions are identified and linked to UMLS Concept Unique Identifiers. Each entity is then represented through pre-trained knowledge graph embeddings (using TransE on the full UMLS graph), capturing its relationships to diseases, symptoms, treatments, and anatomical structures. The fusion mechanism combines contextual and ontological information:

$$g_{ij} = \sigma\Big(\mathbf{W}_g\big[\mathbf{h}^{(i)}_{\text{text}}; \mathbf{e}_j\big] + \mathbf{b}_g\Big), \tag{3.17}$$

$$\mathbf{h}^{(i)}_{\text{fused}} = g_{ij} \odot \mathbf{e}_j + (1 - g_{ij}) \odot \mathbf{h}^{(i)}_{\text{text}} \tag{3.18}$$

where $\mathbf{h}^{(i)}_{\text{text}}$ is the contextualized token representation, $\mathbf{e}_j$ is the knowledge graph embedding, and the gate $\mathbf{g}_{ij}$ learns to balance between linguistic context and structured knowledge.

Training employs multiple objectives beyond standard masked language modeling:

$$\mathcal{L}_{total} = \mathcal{L}_{MLM} + \lambda_1\mathcal{L}_{entity-detect} + \lambda_2\mathcal{L}_{entity-link} + \lambda_3\mathcal{L}_{relation} \tag{3.19}$$

encompassing entity detection (identifying boundaries), entity linking (predicting CUIs), and relation extraction (predicting relationships between entity pairs). This multi-task approach forces the model to maintain explicit entity awareness throughout processing. KeBioLM excels at tasks requiring precise entity disambiguation but requires accurate entity linking systems and introduces significant computational overhead.

K-BERT: Structural Injection via Attention Masking

K-BERT (Knowledge-Enabled BERT) [90] represents the most architecturally ambitious approach, restructuring how the Transformer's attention mechanism operates by injecting knowledge graph structure directly into attention patterns. K-BERT addresses the "knowledge noise" problem: when knowledge is naively appended to text (e.g., adding "Apple is located in California" to "Cook applied to Apple"), it can disrupt sentence understanding and create spurious associations. K-BERT's solution involves sentence tree construction: starting with a flat input sentence, relevant knowledge triples are recursively injected to create a hierarchical structure. To handle this non-sequential structure, K-BERT uses **soft positional encoding** where original tokens maintain sequential positions while injected knowledge tokens receive positions tied to their anchors with small offsets. The core innovation is the visible matrix $\mathbf{M}_{\text{vis}}$ that acts as a hard constraint on attention:

$$\mathbf{M}_{\text{vis}}[i,j] = \begin{cases} 0 & \text{if token } i \text{ should attend to token } j \\ -\infty & \text{otherwise} \end{cases} \tag{3.20}$$

The visibility rules encode the tree structure: original sentence tokens freely attend to each other, knowledge tokens attend to their anchors and ancestors, but knowledge tokens are masked from unrelated sentence parts. For the example above, "Cupertino" can attend to "Apple" but is masked from "Tim Cook", preventing incorrect associations. The modified attention becomes:

$$Attention(\mathbf{Q},\mathbf{K},\mathbf{V}) = softmax(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} + \mathbf{M}_{vis})\mathbf{V} \tag{3.21}$$

where $-\infty$ entries become zero after softmax, effectively blocking information flow. This transforms the Transformer's weak relational bias (all-to-all attention) into a strong, structured bias defined by external knowledge. K-BERT demonstrates substantial improvements on knowledge-intensive tasks but requires high-quality knowledge graphs, accurate entity linking, and adds preprocessing overhead.

3.3.2 Beyond Transformers: Alternative Paradigms of Informed Machine Learning

The three systems examined above share a common premise: knowledge is injected into a Transformer, whether through its training objective, its intermediate representations, or its attention mask. This premise is, however, only one region of a substantially broader design space. Informed Machine Learning, in the systematic sense given to the term by von Rue-

den et al. [156], is characterized along three orthogonal axes: the **source** of the knowledge (scientific laws, expert judgement, formal ontologies, world knowledge), its **representation** (differential equations, logic rules, graphs, probabilistic relations, invariances), and the **integration point** at which it enters the pipeline (training data, hypothesis space, learning algorithm, final hypothesis). Under this taxonomy, SapBERT integrates expert-curated synonymy at the level of the learning algorithm, KeBioLM at the level of the hypothesis space, and K-BERT at the level of the hypothesis space through a hard architectural constraint. What remains unexamined are the families that occupy the same conceptual position as the approach proposed in this chapter, namely knowledge injected *during* learning, but that realise it on entirely different architectural substrates. Three such families are particularly relevant to the present argument.

Knowledge in the Objective Function: Physics-Informed and Neuro-Symbolic Learning

The most widespread alternative dispenses with architectural modification completely and encodes knowledge as an additional differentiable penalty, instantiating directly the regularised objective of Equation 3.4. **Physics-Informed Neural Networks** (PINNs) [112, 78] are the canonical case: the residual of a governing partial differential equation is evaluated at collocation points and added to the data-fitting loss, so that the network is penalized for every violation of the underlying conservation law. The paradigm has proven remarkably effective in low-data regimes precisely because the physical law supplies supervision where measurements are absent, and it has since been generalized from differential equations to symbolic constraints. **Semantic Loss** [162] compiles a propositional formula into an arithmetic circuit and penalises the probability mass that the network assigns to assignments falsifying it; **DeepProbLog** [96] embeds neural predicates inside a probabilistic logic program and back-propagates through exact weighted model counting; **Logic Tensor Networks** [11] interpret first-order formulas in a real-valued fuzzy semantics, so that satisfiability itself becomes the training signal.

The strength of this family is its architecture-agnosticism and the formal guarantee, in the limit of perfect optimization, of consistency with the stated axioms. Its limitations, however, are structural rather than incidental. First, the knowledge must admit a *truth-functional* formulation: a constraint is either satisfied or violated, and a differentiable relaxation of that dichotomy must exist. The knowledge available in medical process monitoring is of a different nature. A clinical taxonomy does not assert that a thrombolytic therapy must never follow a CT scan; it asserts that thrombolysis and thrombectomy are *closer* to one another, in terms of therapeutic goal, than either is to a discharge-administration activity.

This is a statement about graded semantic proximity, not about logical admissibility, and forcing it into a satisfaction-based semantics either discards the gradation or requires an arbitrary thresholding of the ontology into hard rules. Second, a loss penalty operates on the *output distribution* and leaves the internal representation space unchanged: the model is discouraged from producing inconsistent predictions, but it is never told that two activities are related, and must therefore still rediscover that relatedness from the event log. Third, the penalty is active only at training time and its effect degrades silently under distribution shift, with no residual structural guarantee at inference.

Knowledge in a Different Architecture: Graph-Based and Hierarchy-Aware Models

A second family accepts the premise that structure should be architectural but rejects the Transformer as the substrate, reformulating the task so that a Graph Neural Network can operate on it directly [123, 14]. Where the domain is natively relational this is unquestionably the stronger inductive bias, and Chapter 4 adopts exactly this stance. Applied to process monitoring, however, the reformulation is costly: a patient trace is a *sequence* of events, and the ontology is an *exogenous graph* over the activity vocabulary, not over the trace instances. Encoding both simultaneously requires either constructing a heterogeneous graph per trace, which fragments the sequential signal that next-activity prediction depends upon, or a two-tower design in which a GNN encodes the ontology and a sequence model encodes the trace, which reintroduces the very fusion problem that KeBioLM illustrates. The literature on *graph* Transformers is instructive here for a different reason. Graphormer [166] and the spectral attention networks of Kreuzer et al. [82], together with the learnable structural and positional representations of Dwivedi et al. [46], established that graph structure can be conveyed to a Transformer not by rewiring attention but by *encoding the structure into the positional representation*, with Laplacian eigenvectors serving as a canonical spectral coordinate system for the nodes. The construction developed in this chapter inherits that insight while inverting its premise: in graph Transformers the input itself is the graph, whereas here the input is a sequence and the graph is external domain knowledge over the token alphabet.

3.3.3 Positioning of the Proposed Approach and Motivation for an Additional Inductive Bias

The preceding survey permits a precise formulation of the question that motivates the remainder of this chapter: given that the Transformer is a universal sequence approximator

and that several mature mechanisms already exist for supplying it with domain knowledge, why should an *additional* architectural bias be introduced in this particular setting? Three arguments, one theoretical, one empirical, and one concerning the nature of the available knowledge, converge on an affirmative answer.

The sample regime invalidates the amortization argument. The Transformer's weak architectural prior is not a defect in general; it is a deliberate trade of inductive bias for representational flexibility, and it is repaid whenever the missing prior can be amortised over a sufficiently large corpus. This is exactly what occurs in natural language processing, where the absence of a syntactic prior is compensated by pre-training over billions of tokens. Medical event logs afford no such compensation. The stroke-management log analyzed in Section 3.4 comprises a few hundred traces over a vocabulary of clinical activities of comparable cardinality, and no general-purpose corpus of clinical process traces exists from which a prior could be transferred. The consequence is not that learning is impossible but that the burden of generalization must be shifted from data to assumptions. In this regime the marginal value of an architectural prior is at its maximum, and the flexibility that justifies the Transformer's design elsewhere becomes a liability.

The default prior is not merely absent but misaligned. A second and stronger argument emerges from the experimental evidence reported in Section 3.4: sinusoidal positional encoding, the standard prior of the original Transformer, yields no benefit in this setting and in several configurations degrades performance relative to a model with no positional information at all. This is not an artefact. Sinusoidal encoding assumes a metric ordering in which absolute and relative offsets carry semantic weight, an assumption well matched to natural language and poorly matched to clinical pathways, where activities cluster by therapeutic goal and the ordering within a functional group is frequently clinically immaterial. The Transformer therefore does not begin from a neutral position that additional knowledge might merely improve; it begins from an actively inappropriate one. Removing a misaligned bias and replacing it with a domain-aligned one is a qualitatively different intervention from adding knowledge to a neutral baseline, and none of the existing approaches surveyed above performs it: SapBERT, KeBioLM and K-BERT all leave the positional prior untouched, while loss-based and filtering-based methods operate outside the representation entirely.

The available knowledge is taxonomic, not deontic. The knowledge that clinical practice actually makes available in this domain is a goal-oriented taxonomy of activities, an artefact that is inexpensive to elicit from clinicians and frequently already present in

institutional care pathways. It expresses graded functional proximity rather than obligation or prohibition. Logic-based integration requires the latter and must therefore discretise the former; conformance-based filtering requires a control-flow specification that the taxonomy does not provide; entity-fusion architectures such as KeBioLM presuppose an entity-linking stage that is vacuous here, since every token in a trace is already an ontology node by construction. Structural Positional Encoding is designed for precisely the representation that is available: the Laplacian spectrum of the taxonomy furnishes a continuous coordinate system in which distance is monotone in ontological proximity, and the projection of those coordinates into the embedding space makes that proximity directly legible to the attention mechanism, before any attention weight is computed.

Three properties follow from this design and constitute the advantage claimed over the alternatives. The bias is *soft*: unlike K-BERT's visibility matrix, which excludes attention paths outright, SPE biases the representation without prohibiting any interaction, so that an incomplete or partially incorrect ontology degrades performance gracefully rather than catastrophically, a property corroborated by the robustness of the reported gains under five and ten per cent input perturbation. The bias is *architecture-agnostic*: because it acts on the input representation rather than on the attention computation, it transfers to recurrent architectures with comparable benefit, as the LSTM experiments confirm, which neither K-BERT's masking nor graph-Transformer structural encodings permit. And the bias is *inference-free*: the eigendecomposition is computed once offline and the only added parameters are those of a single learnable projection matrix, so no decoding-time overhead is incurred, in contrast to conformance-checking pipelines whose cost scales with the candidate set.

Table 3.1 summarises this positioning. The claim advanced is not that SPE dominates the alternatives in general; each of the surveyed families is preferable under the conditions for which it was designed. The claim is the narrower and more defensible one that, for a task in which the data are scarce, the default positional prior is actively misaligned, the available knowledge is taxonomic rather than deontic, and the deployment context penalises decoding-time overhead, the existing mechanisms are either inapplicable or address the wrong integration point, and a representational bias derived from the spectrum of the domain ontology occupies the gap they leave.

Family	**Integration point**	**Knowledge form required**	**Constraint type**	**Inference cost**
PINN / Semantic Loss / LTN / DeepProbLog	Objective function	Equations or logic formulas	Soft, output-level	None
GNN / graph Transformers	Architecture	Graph over the instances	Hard, structural	Encoding of the graph
K-BERT	Attention mask	Triples + entity linking	Hard, path-level	Preprocessing
KeBioLM	Hidden states	KG embeddings + entity linking	Soft, fusion	Entity linking
SPE	**Input representation**	**Activity taxonomy**	**Soft, representational**	**None (offline)**

Table 3.1: Positioning of Structural Positional Encoding within the space of Informed Machine Learning approaches that inject prior knowledge during learning, following the integration-point axis of von Rueden et al. [156].

3.4 Structural Positional Encoding for Knowledge Integration in Medical Process Monitoring

The integration of domain knowledge into Transformer architectures has demonstrated significant potential across various biomedical applications. While approaches like SapBERT, KeBioLM, and K-BERT have primarily focused on natural language processing tasks, the principles of knowledge-enhanced learning extend naturally to other structured domains. This section presents a case study in medical process monitoring, where the temporal sequence of clinical activities can be enriched with ontological knowledge to improve predictive accuracy and trace classification performance.

Medical process monitoring represents a compelling domain for knowledge integration. Clinical pathways are not arbitrary sequences of activities; they are structured by well-established medical knowledge about diagnostic procedures, treatment protocols, and care coordination. An ontology that captures these relationships by grouping activities according to clinical goals, temporal dependencies, or functional similarities offers valuable structural priors to guide a Transformer's learning process. The work presented here demonstrates how such knowledge can be incorporated through SPE, a technique that embeds graph-structured domain knowledge directly into the positional representations used by the attention mechanism.

3.4.1 Medical Process Monitoring and Trace Classification

Medical process monitoring encompasses two critical predictive tasks that have gained attention in the process mining and healthcare AI communities. The first (and most interesting), **next activity prediction**, aims to forecast which clinical activity should be executed next given a partially observed patient trace. This capability provides valuable decision support in complex or atypical situations where clinicians must navigate resource constraints, patient comorbidities, or uncertainty about guideline applicability. The second task, **trace classification**, seeks to categorize complete patient traces according to specific properties, such as distinguishing between simple and complex patient cases, or identifying the appropriate level of care facility. Classification enables quality assessment by verifying whether different patient types receive appropriately differentiated care protocols.
Formally, a **patient trace** is represented as a sequence $S = \{a_1, ..., a_i, ..., a_n\}$ where each a_i belongs to a set A of possible clinical activities. In next activity prediction, the model learns a function that maps a trace prefix to a probability distribution over the next activity. In trace classification, the model assigns the entire trace to one of several predefined classes. Notably, these two tasks can be **unified within a single Transformer architecture** by treating the class label as a special token to be predicted at the end of the sequence, an approach analogous to the classification strategy employed in BERT [40].
About the **knowledge integration**, Medical process monitoring faces several distinctive challenges that justify the use of this technique:

- **Limited Training Data**: Clinical datasets are often small due to privacy constraints, the rarity of certain conditions, and the cost of data collection. A dataset of several hundred patient traces, while valuable, may be insufficient for a data-driven Transformer to learn robust representations.

- **Activity Relationships**: Medical activities are not independent tokens. Diagnostic procedures (CT scans, blood tests, ECGs) share functional similarities and are more likely to occur in temporal proximity than, say, administrative activities. These relationships are encoded in medical ontologies but are not explicitly represented in raw activity sequences.

- **Weak Sequential Bias**: While clinical pathways do exhibit temporal ordering (diagnosis typically precedes treatment), the precise order of certain activities within a functional group may be less critical than their collective presence. A purely sequential positional encoding may impose rigid assumptions about ordering.

- **Domain Generalization**: Models trained on data from one hospital or patient population must generalize to others. Grounding the model in stable domain knowledge rather than spurious correlations in limited training data, enhances robustness to distribution shift.

Evolution of Predictive Process Monitoring Approaches

The application of machine learning to predictive process monitoring has evolved through several generations of increasingly sophisticated techniques. Early approaches relied on classical methods such as Transition Systems [85], Hidden Markov Models [84], and Support Vector Machines [29], which provided baseline capabilities but struggled to capture the complex temporal dependencies inherent in clinical pathways. The advent of deep learning ushered in a new era of performance, with various architectures being adapted to the sequential nature of process traces.

Autoencoders [100, 65], were among the first deep learning approaches applied to this domain, offering the advantage of dimensionality reduction for high-dimensional activity representations. However, their fundamental limitation lies in their inability to model long-range temporal dependencies, a critical requirement given that clinical decisions may depend on activities performed hours or days earlier in a patient's care trajectory.

Convolutional Neural Networks [4, 99] introduced a different perspective, treating process traces as one-dimensional grids and applying convolutional filters to detect local patterns. While CNNs can identify recurring subsequences of activities, their reliance on fixed-size kernels inherently limits their receptive field. An activity's relevance may depend on events occurring beyond the kernel's reach, and this architectural constraint becomes particularly problematic in medical processes where the temporal distance between related activities can vary dramatically across patients.

The sequential nature of process traces makes **Recurrent Neural Networks** [105] and their more sophisticated variants a more natural architectural choice. **Long Short-Term Memory** networks [66], in particular, emerged as a dominant approach due to their ability to learn complex temporal dynamics through gating mechanisms that selectively retain or forget information over extended sequences. The works of [48, 140, 30, 141] demonstrated that LSTMs could effectively predict next activities by maintaining a hidden state that accumulates relevant context from the trace prefix. Similarly, **Gated Recurrent Units** (GRUs) [36], which simplify the LSTM architecture with fewer parameters, have been successfully applied to both next activity prediction and general trace classification tasks [64]. For exceptionally long traces or processes with extensive repeated cycles, **Memory Augmented Neural Networks** (MANNs) [79] offer enhanced capacity to maintain dependencies that

might cause standard RNNs to forget early-sequence information. However, this additional memory capacity comes at the cost of significantly increased training complexity, making MANNs most appropriate for domains with particularly challenging temporal structure.
The transition from recurrent architectures to **Transformer-based** models represents the current state of the art in predictive process monitoring. Works [109, 28] pioneered the application of Transformers to this domain, adopting a decoder-only architecture with multi-head self-attention that processes trace prefixes of varying lengths during training. This architectural choice eliminates the sequential processing bottleneck of RNNs, enabling parallel computation across all positions in a trace while maintaining the capacity to capture long-range dependencies through the attention mechanism. The flexibility of attention-based models has led to their widespread adoption, exploring their potential in process monitoring contexts [42].
A few approaches have addressed the idea of incorporating **domain knowledge** for improving next activity prediction. The work in [41] adopts an LSTM architecture to predict all possible suffixes of a process traces (given the prefix), ranks the suffixes on the basis of their likelihood, and then keeps the first suffix which is compliant with domain/background knowledge (expressed by means of Linear Temporal Logic constraints). The work in [42] is inspired by such a contribution, but moves towards the adoption of a transformer. Moreover, conformance with domain knowledge is tested while the suffix is still being constructed, and not just at the end. Domain knowledge is represented as a Petri Net, and compliance is obtained using token-based replay [22].

3.4.2 Structural Positional Encoding: Embedding Structured Knowledge

The core innovation of this work lies in the incorporation of domain knowledge through **Structural Positional Encoding (SPE)**. Rather than using only the standard sinusoidal positional encoding that captures sequential order, SPE augments each activity's representation with information about its position within a domain ontology encoded as a graph.
The medical ontology is structured as a graph $G = (V, E)$ where:

- Nodes (V): Include both activity nodes (one for each clinical activity in A) and activity-type nodes (representing higher-level categories such as "diagnostic procedures," "pharmacological therapy," or "patient management").
- Edges (E): Connect activity nodes to their corresponding activity-type nodes, and connect activity-type nodes to each other to form a hierarchical taxonomy. The graph

is designed to be connected, ensuring that every activity has a path to every other activity, with path length reflecting semantic similarity.

The application could rely on an available ontology (see Figure 3.1 for an excerpt) that groups together activities involved in similar diagnostic or treatment objectives. Grouping-by-goal is achieved through a taxonomic representation, where activities involved in similar goals are placed in a close relationship (having a shorter path to reach one from the other). According to the goal semantics, two main goals (sons of the root) can be identified: the **emergency phase** patient care goal, and the **hospitalization phase** patient care goal. Within both phases, it is possible to distinguish three sub-goals: **patient management**, **patient assessment** and **therapy**. Patient management includes as leaves some administrative ground activities, such as arrival or discharge. Patient assessment includes diagnostic activities, such as Computer Tomography (CT). Therapy includes therapeutic activities, such as pharmacological ones, or life style improvement ones. For instance, ECG_H (electrocardiogram), RX_H (X rays) and Angiography_H (angiography) are three leaves, which represent diagnostic activities, sharing the goal of assessing the patient's condition at the time of hospitalization.

It is important to clarify upfront the nature of this contribution: SPE is not proposed as a general-purpose, domain-agnostic replacement for the sinusoidal positional encoding introduced in the original Transformer formulation. The standard PE encodes the only piece of positional information that is universally available in any token sequence and remains the appropriate default when no additional structural information about the domain is known. SPE, in contrast, is a domain-informed positional encoding: it is applicable only in those settings where an ontology (or, more generally, a graph-structured description of the semantic relationships among the tokens) is available, and its purpose is to expose this structural information directly to the attention mechanism. Rather than using only the standard sinusoidal positional encoding that captures sequential order, SPE augments each activity's representation with information about its position within a domain ontology encoded as a graph. The comparison between SPE and PE carried out in the following sections is therefore not intended to adjudicate which of the two is methodologically superior in the abstract, but to quantify the gain achievable by replacing (or complementing) a generic sequential prior with a domain-specific structural prior, whenever such a prior is available.

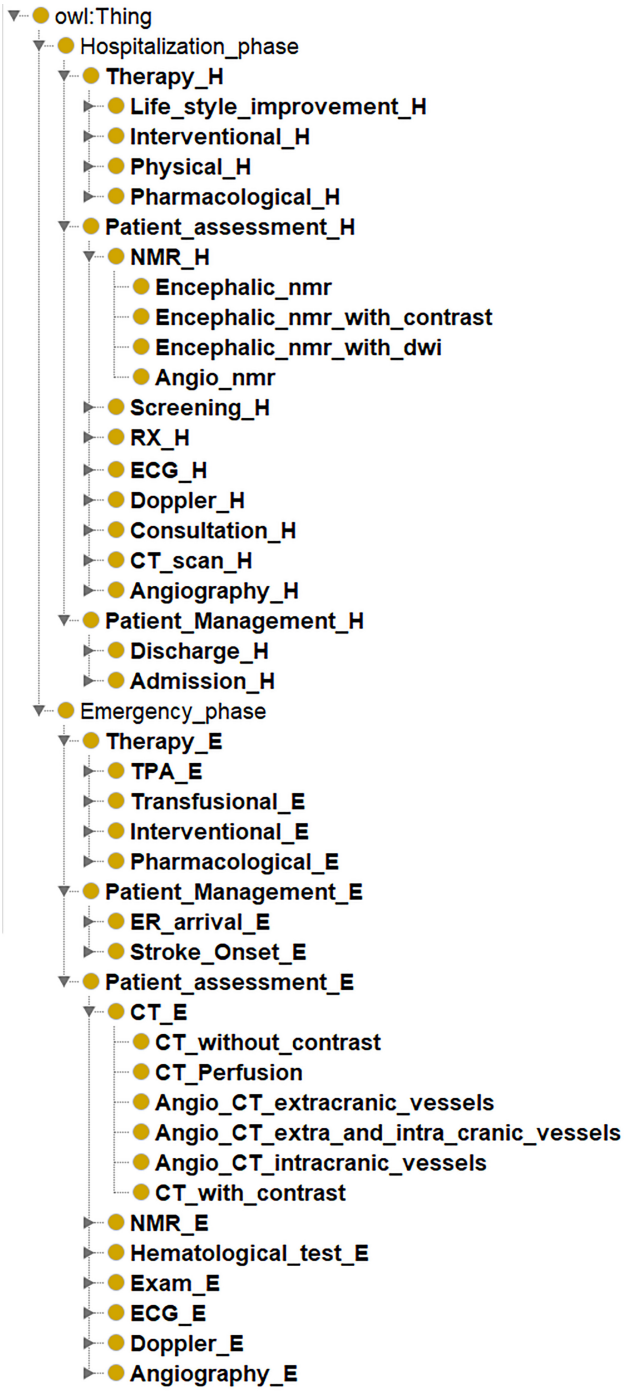


Figure 3.1: Excerpt of the domain ontology.

Laplacian Eigenvector Encoding

To translate this graph structure into vector representations suitable for integration with the Transformer, SPE employs the Laplacian eigenvector technique [16]. This method computes node embeddings that preserve the graph's geometric structure: nodes that are close in the

graph topology receive similar embeddings.
The process begins by constructing the graph Laplacian matrix:

$$\Delta = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}} = U^T \Lambda U \tag{3.22}$$

where I is the identity matrix, $A \in \mathbb{R}^{n \times n}$ assuming $n = |V|$ is the adjacency matrix where $A_{ij} = 1$ if node i and node j are connected by an edge (i.e. the graph connectivity) and $D \in R^{n \times n}$ where $D_{ii} = \sum_{i=1}^{n} A_{ij}$ is a diagonal matrix representing the degree of every node in G; Λ, U denote the eigenvalues and eigenvectors respectively. The embedding for a node i is a vector $\lambda_i \in \mathbb{R}^k$ defined as the k smallest nontrivial eigenvector components of that node [45]. Finally, to incorporate the node embedding $\lambda_i \in \mathbb{R}^k$ into the corresponding activity embedding $X_i \in \mathbb{R}^d$, we employ a fully-connected layer $\Theta \in \mathbb{R}^{k \times d}$ to ensure compatibility between the embedding sizes. This encoding has several desirable properties:

- **Structural Smoothness**: Nodes with similar eigenvector values tend to belong to the same cluster or community within the graph, meaning they are structurally "close". Intuitively, if two activities share many common neighbors or belong to the same densely connected subgraph, their eigenvector components will be similar.
- **Global-Local Balance**: Laplacian eigenvectors, particularly the **Fiedler vector** (the eigenvector corresponding to the second-smallest eigenvalue), capture both global graph structure (major divisions and communities) and local neighborhood information (fine-grained connectivity patterns), providing a multi-scale representation of each node's position.
- **Computational Efficiency**: The encoding is computed once offline through matrix factorization, requiring no additional learning during training. This stands in contrast to graph neural network embeddings, which would need to be learned jointly with the Transformer parameters, potentially introducing optimization challenges and computational overhead.

3.4.3 Integration into the Transformer

The node embeddings are integrated into the Transformer architecture at the positional encoding stage. Given an input trace $S = \{a_1, ..., a_n\}$, the model first computes standard token embeddings $X \in \mathbb{R}^{n \times d}$ where d is the embedding dimension. Rather than only adding standard positional encodings, the model also adds the structural encodings:

$$\mathbf{X}' = \mathbf{X} + SPE(\mathbf{X}, G) \tag{3.23}$$

Specifically, for each activity a_i in the sequence, the corresponding Laplacian eigenvector embedding $\lambda_i \in \mathbb{R}^k$ is projected to the model's embedding dimension using a learned linear transformation $\Theta \in \mathbb{R}^{k \times d}$:

$$SPE(a_i) = \lambda_i \Theta \tag{3.24}$$

This structural encoding is then added to the token embedding before it enters the multi-head attention layer. As a result, the augmented representation jointly encodes three forms of information: the **semantic content** captured by the learned token embedding, which represents what the activity is; the **sequential position** captured by standard positional encoding, which represents where the activity appears within the trace; and the **ontological position** captured by the structural encoding, which represents the activity's role and its relationships within the domain knowledge graph.

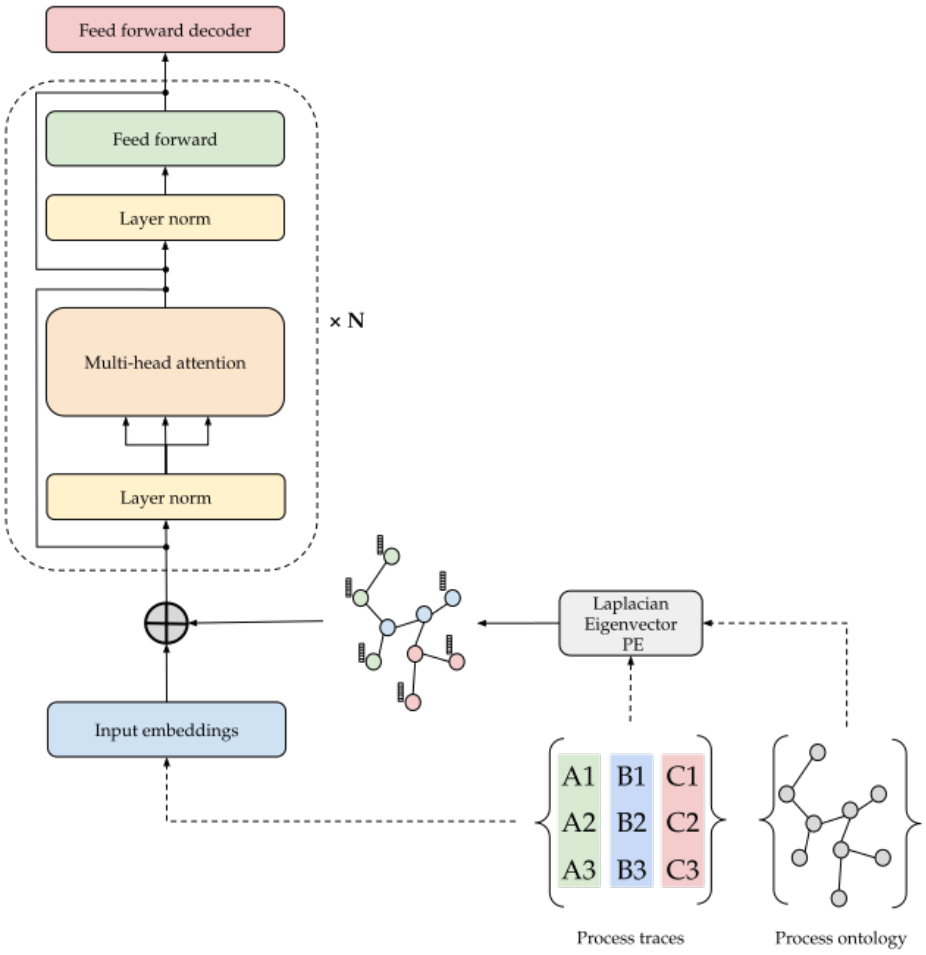


Figure 3.2: The model architecture. The left side of the figure represents the transformer-decoder model, the right side shows the Structural Positional Encoding module

This design enables the attention mechanism to learn interactions not only based on learned token compatibility and sequential proximity, but also guided by ontologically-defined relatedness. Two activities that are semantically similar according to the ontology will have similar structural encodings, increasing the likelihood that the attention mechanism will learn to relate them even if they appear distant in the sequence or were rarely co-occurring in the limited training data. in the standard Transformer, the attention weights between positions i and j are computed based solely on the compatibility of their query and key representations (see eq. 3.5). With SPE, the query and key representations now incorporate structural information:

$$\mathbf{Q} = (\mathbf{X} + SPE(\mathbf{X}, G))\mathbf{W}_Q, \quad \mathbf{K} = (\mathbf{X} + SPE(\mathbf{X}, G))\mathbf{W}_K \tag{3.25}$$

This means that the attention score between activity a_i at position i and activity a_j at position j is influenced not only by what these activities are and where they occur in the sequence, but also by their ontological proximity in the domain knowledge graph. Activities that are ontologically related (e.g., two different imaging modalities) will have similar structural encodings, making their query-key dot products larger and increasing the attention weights between them, even if they never co-occur in the training data.

Trace Classification via Predictive Modeling

The **trace classification** task is integrated at no extra cost into the same Transformer architecture. Rather than training a separate classification model, the approach treats the class label as a special token appended to the end of the trace sequence. During training, the model learns to predict this class token given the complete activity sequence as context. To prioritize the classification objective, the loss function is modified to assign higher weight to the prediction of the class token compared to intermediate activity predictions. The weighted cross-entropy loss is defined as:

$$\mathcal{L}_{wCE} = -\sum_{i=0}^{C} w_i \cdot y_i \cdot log(p_i) \tag{3.26}$$

where C is the number of classes, w_i is the weight for class i (empirically determined to be higher than the weight for activity tokens), y_i is the true label, and p_i is the predicted probability. This formulation allows the same model and training procedure to address both next activity prediction and trace classification within a unified framework.
This approach offers several advantages. First, it eliminates the need for task-specific architectural modifications: the same encoder that learns representations for next activity

prediction automatically produces representations suitable for classification. Second, the multi-task learning regime, where the model simultaneously optimizes for activity prediction and trace classification, acts as a form of regularization that can improve generalization.

3.5 Experimental Validation: Stroke Management

The approach was validated on a **real-world dataset from stroke management**, a domain where predictive process monitoring offers significant clinical value. Stroke care involves time-critical decisions about thrombolytic therapy, imaging modalities, and intervention procedures. The ability to predict the next appropriate activity, or to correctly classify patients as simple versus complex case, directly supports better resource allocation and clinical decision-making.

Dataset Characteristics

The dataset contained 5,341 stroke patient traces, including:

- 905 **complex patients**: Individuals with rare stroke types or significant comorbidities requiring advanced care.
- 4436 **simple patients**: Standard stroke cases, of which 1724 were treated at Stroke Centers (SCs) and 2712 at Stroke Units (SUs).

Traces exhibited an **average length of 15 activities**, drawn from a vocabulary of 82 possible clinical activities. The activities span categories such as emergency arrival, diagnostic imaging (CT, MRI, angiography), laboratory tests (hematological, biochemical), neurological examinations, pharmacological interventions (thrombolysis, anticoagulants), and administrative procedures (discharge, transfer). The standard deviation of trace length was 3 activities, with a minimum length of 2 and maximum of 25 activities, reflecting the heterogeneity of stroke presentations and care trajectories.

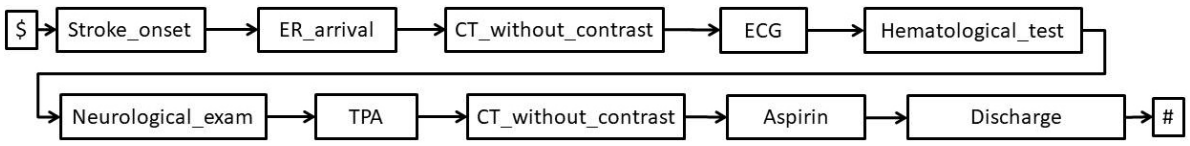


Figure 3.3: Example of a trace. The activities $ and # represent the start and the end of the trace, respectively.

The associated ontology was structured as a graph with 110 nodes and 111 edges, organized in a five-level hierarchy. At the top level, activities were grouped into “emergency

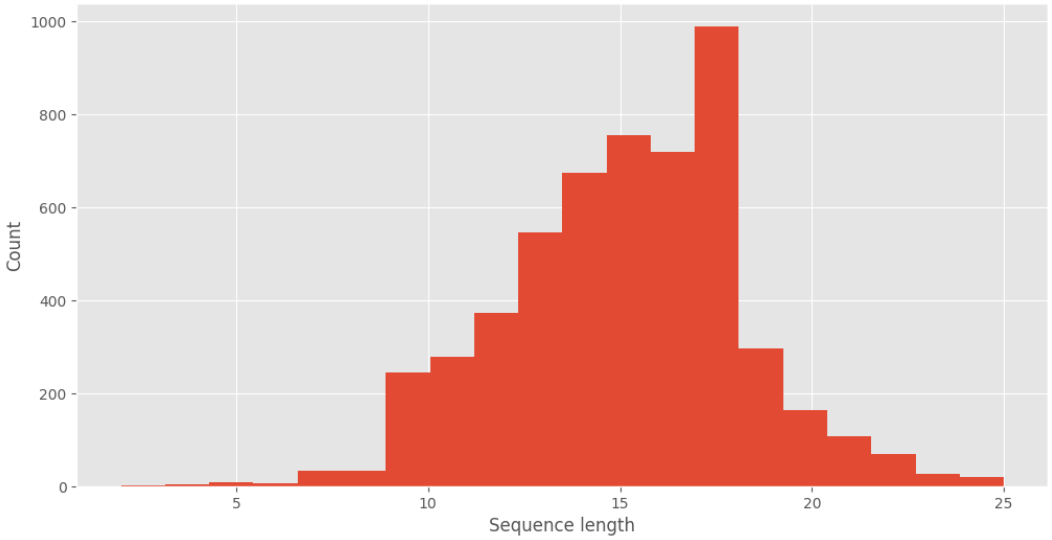


Figure 3.4: Trace length histogram

phase care" and "hospitalization phase care," reflecting the two major temporal phases of stroke management. Within each phase, three sub-goals were defined: patient management (administrative activities such as ER arrival and discharge), patient assessment (diagnostic procedures including all imaging and laboratory tests), and therapy (treatment interventions including both pharmacological and surgical procedures).

For example, within the hospitalization phase, all neuroimaging procedures (CT without contrast, CT with contrast, MRI, angiography) were children of a common "neuroimaging" category under "patient assessment during hospitalization," reflecting their shared diagnostic goal of visualizing brain structure and vascular anatomy. Similarly, pharmacological interventions were organized hierarchically: thrombolytic agents (TPA, tenecteplase) were grouped under "thrombolysis," anticoagulants (aspirin, heparin, warfarin) under "anticoagulation therapy," and both categories were children of "pharmacological therapy" under "therapy during hospitalization." This hierarchical organization encodes clinical knowledge at multiple levels of granularity. At the finest level, it distinguishes specific medications or imaging modalities. At intermediate levels, it groups activities by therapeutic mechanism (e.g., anticoagulation) or diagnostic target (e.g., neuroimaging). At the coarsest level, it separates the fundamental clinical functions of assessment, management, and treatment. This multi-level structure allows SPE to capture both subtle distinctions between closely related activities and broad functional similarities between activities serving the same overarching clinical purpose.

Experimental setup

All traces were preprocessed with special tokens marking the start and end of each sequence. Padding was applied to ensure uniform length, with padded positions represented by a special "pad" token. During training, padded positions and start-of-sequence tokens were excluded from loss calculation to prevent the model from learning patterns associated with these artificial markers, while end-of-sequence tokens were retained to enable the model to learn sequence termination. The dataset was split 80/10/10 into training, validation, and test sets. Hyperparameters were optimized using the Optuna framework [3], which performed a grid search over the parameter space to identify the configuration maximizing validation set accuracy. Table 3.2 reports the search space considered and the best parameters found.

Parameter	**Search space**	**Best configuration**
Embedding size	[16, 32, 64, 128, 256]	64
Hidden size	[16, 32, 64, 128, 256]	128
Number of heads	[1, 2, 4, 8]	4
Number of layers	[1, 2, 3, 4, 5]	4
Dropout rate	0.1 - 0.5	0.216375
Optimizer	AdamW	AdamW
Scheduler	StepLR	StepLR
Gamma	0.85 - 0.99	0.989695
Learning rate	1×10^{-2} - 3×10^{-2}	0.002836
Ontology embededing size	[8, 16, 32]	32

Table 3.2: Parameter search space used to find optimal hyperparameters and optimal configuration, by Optuna. Search spaces included between square brackets are discrete sets, while for continuous search spaces the values are sampled by Optuna space in that range.

Three model variants were compared:

- **No Positional Encoding** (None): Baseline Transformer without any positional information beyond learned token embeddings.
- **Standard Positional Encoding** (PE): Transformer with sinusoidal positional encoding, using the formulation:

$$PE_{pos,2i} = \sin(pos/10000^{2i/d}), \quad PE_{pos,2i+1} = \cos(pos/10000^{2i/d})$$

- **Structural Positional Encoding** (SPE): Transformer augmented with Laplacian eigenvector embeddings from the domain ontology.

Additionally, an **LSTM baseline** with 128 hidden units and dropout of 0.39 was trained to assess whether classical recurrent architectures could also benefit from structural encoding.

This comparison is particularly important given the extensive prior work demonstrating LSTM effectiveness for process monitoring tasks.

3.5.1 Next Activity Prediction Results

Model performance was evaluated using accuracy-at-k metrics, measuring whether the true next activity appeared among the model's top-k predictions. Results were averaged over 10 random initializations and data splits to ensure statistical reliability.

SPE Substantially Outperforms Baselines: Across all model sizes, the SPE variant achieved significantly higher accuracy than both the baseline (no encoding) and the standard PE approach, see table 3.3.

Architecture	Method	Model size	Accuracy@1	Accuracy@3	Accuracy@5
Transformer	None	16	45.6±0.3	67.1±0.3	75.8±0.3
	None	32	46.9±0.2	69.4±0.3	78.8±0.3
	None	64	47.4±0.4	70.4±0.5	79.8±0.5
	None	128	47.3±0.2	70.6±0.2	80.1±0.3
Transformer	PE	16	45.1±0.4	65.8±0.4	74.7±0.4
	PE	32	46.9±0.2	68.6±0.3	78.4±0.3
	PE	64	47.5±0.1	70.4±0.3	79.9±0.2
	PE	128	47.0±0.2	70.1±0.5	79.8±0.3
Transformer	SPE	16	48.2±0.5	69.7±0.3	77.6±0.4
	SPE	32	52.1±0.4	75.3±0.4	83.4±0.2
	SPE	64	55.1±0.3	79.4±0.6	87.8±0.5
	SPE	128	**56.4±0.5**	**81.5±0.4**	**89.8±0.4**
LSTM	None	128 (best)	46.8±0.2	69.5±0.2	78.8±0.3
	PE	128 (best)	46.7±0.4	69.5±0.4	78.7±0.3
	SPE	128 (best)	52.2±0.3	75.7±0.4	84.1±0.4

Table 3.3: Test accuracy-at-k and standard deviation on 10 random initialization and train-val-test splits. Model performance with different positional encoding methods at different embedding sizes are reported

These results demonstrate that grounding the model in domain knowledge yields improvements of approximately 9 percentage points at top-1 accuracy and over 10 percentage points at top-3 accuracy, a substantial gain in a clinical decision support context where the top few suggestions guide physician decision-making. The standard deviations across runs were relatively small (±0.5% for SPE at 128 dimensions), indicating that the performance gains are robust to random initialization and data splitting.

Notably, the PE approach **did not improve** over the baseline without positional encoding, and in some configurations even slightly underperformed it. This observation must be read

carefully, because it is the key to understanding the nature of the improvement brought by SPE. The failure of the standard PE in this setting is not a symptom of a general methodological weakness of sinusoidal positional encodings but is instead a symptom of a domain mismatch. Medical process traces are not **strictly sequential objects**: once a patient has reached the emergency department, the ordering of CT scan, ECG and blood test within the initial assessment phase is in most cases clinically immaterial, whereas their collective completion before thrombolysis is not. The quantity that carries predictive signal is therefore not the absolute sequential index of each activity, but the membership of each activity in a functional group defined by the clinical ontology, together with the set of groups that have been visited so far. A positional encoding that only conveys sequential order is, in this sense, injecting a prior that is either useless (when the trace happens to respect a canonical order) or actively misleading (when the same set of activities is performed in a different but clinically equivalent order).

SPE improved performance consistently across embedding sizes from 16 to 128 dimensions. Interestingly, performance saturated at 64-dimensional embeddings, with the improvement from 64 to 128 dimensions being marginal (less than 1.5 percentage points). This suggests that the model does not require excessive capacity when guided by structural knowledge, a desirable property for deployment in resource-constrained clinical environments where model size and inference speed are practical concerns.

Finally, the LSTM baseline, when augmented with SPE, achieved 52.2% accuracy@1, compared to 46.8% without SPE. While still underperforming the Transformer+SPE combination, this result indicates that structural positional encoding is a general technique applicable beyond Transformers, providing value to any architecture capable of incorporating auxiliary embeddings. The LSTM's sequential processing bottleneck and limited capacity for long-range dependencies prevent it from fully exploiting the structural information, but the 5.4 percentage point improvement demonstrates that even architecturally constrained models benefit from explicit knowledge integration.

Robustness to Noise

To simulate real-world logging errors and data quality issues common in clinical environments, the models were tested on datasets where 5% and 10% of traces had activities randomly inserted or removed.

While performance declined with increasing noise, the model retained substantial predictive capability even under significant data corruption (see Table 3.4). The fact that accuracy@1 on 10% noisy data (48.3%) remains higher than the clean-data baseline without SPE (47.3%) is particularly striking: a knowledge-enhanced model operating on corrupted

Setup		5% noise dataset			10% noise dataset		
Architecture	**Model size**	**Accuracy@1**	**Accuracy@3**	**Accuracy@5**	**Accuracy@1**	**Accuracy@3**	**Accuracy@5**
Transformer	16	43.9±0.4	65.1±0.4	74.0±0.4	39.0±0.7	60.2±0.5	69.5±0.5
	32	48.1±0.3	71.5±0.4	80.4±0.4	43.5±2.1	66.3±2.9	75.7±2.9
	64	51.6±0.5	76.6±0.6	84.8±0.5	46.1±2.9	69.7±4.7	79.2±5.1
	128	**53.4±0.6**	**78.3±0.4**	**87.2±0.5**	**48.3±2.3**	**73.9±3.1**	**83.8±3.2**
LSTM	16	37.0±0.4	58.1±0.5	69.0±0.4	33.3±0.6	53.6±0.6	64.7±0.5
	32	42.1±0.1	63.7±0.8	73.0±0.6	36.5±1.2	57.6±1.2	68.1±0.7
	64	46.7±0.5	69.4±0.9	78.0±0.7	42.3±0.6	64.4±0.8	73.4±0.8
	128	49.6±0.4	73.3±0.6	81.9±0.6	45.3±0.9	68.5±1.1	77.8±1.0

Table 3.4: Test accuracy-at-k and standard deviation on 10 random initialization and train-val-test splits. Model performance with 2 different noisy datasets. The models are trained using the SPE method.

data still outperforms a purely data-driven model operating on clean data. This suggests practical applicability in clinical settings where electronic health record data may contain errors, omissions, or inconsistencies due to delayed documentation, manual entry mistakes, or system integration issues.

The LSTM baseline exhibited similar degradation patterns under noise, though starting from a lower baseline, while the standard PE approach showed no particular robustness advantage, confirming that the improved noise tolerance is attributable to the structural knowledge encoded in SPE rather than the Transformer architecture itself.

Qualitative Analysis: Clinical Coherence

Beyond quantitative metrics, a qualitative evaluation with domain experts examined the clinical relevance of the model's predictions. Consider a trace where a patient, after emergency arrival, undergoes CT scan, hematological testing, and neurological evaluation, followed by mechanical thrombectomy (surgical clot removal). The question posed to the model: what should be done next?

Without SPE: The model predicted an encephalic MRI (see Figure 3.5), a neuroimaging procedure to assess the surgical outcome. While clinically sensible as a follow-up assessment to verify the success of the intervention and check for procedural complications, this suggestion does not align with immediate post-surgical guidelines established by stroke care protocols.

With SPE: The model predicted a trans-chest ECG (electrocardiogram, see Figure 3.6). This prediction is more clinically appropriate according to established guidelines. The **SPREAD** stroke clinical guidelines [129] specify that after mechanical thrombectomy, monitoring the patient's blood pressure and cardiac function within the first 24 hours is mandatory to ensure cardiovascular parameters remain below 180/105 mm Hg. ECG monitoring, which typically includes blood pressure measurement as a standard component of cardiac

monitoring, directly addresses this guideline recommendation.
Referring to the ontology structure, both suggestions fall under "patient assessment during hospitalization," demonstrating that both models correctly identified the need for post-procedural assessment. However, they differ in the specificity and appropriateness of that assessment. Encephalic MRI is a child of "neuroimaging," focused on brain structure evaluation, a reasonable but non-urgent follow-up. Trans-chest ECG is a child of "cardiological monitoring," directly relevant to the post-intervention cardiovascular stability requirement specified in clinical guidelines as an immediate priority. The SPE model's prediction, informed by the ontology's encoding of functional relationships and clinical priorities, produced a recommendation more closely aligned with established clinical protocols. This alignment suggests that the model has learned not merely to predict statistically frequent next activities, but to reason about the clinical goals and priorities encoded in the domain ontology. The domain experts evaluating the predictions confirmed that the SPE model's suggestion represented better adherence to standard-of-care protocols.

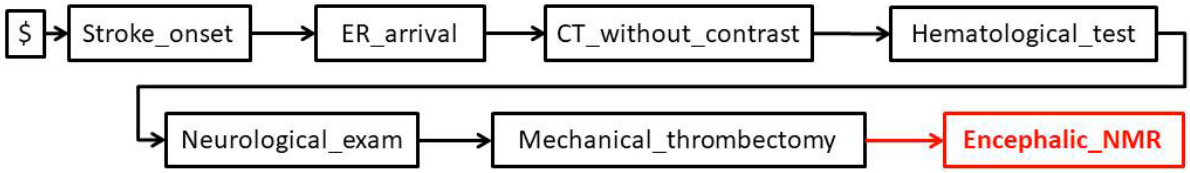


Figure 3.5: Trace given as a query to the transformer without using domain knowledge, with the predicted next activity colored in red.

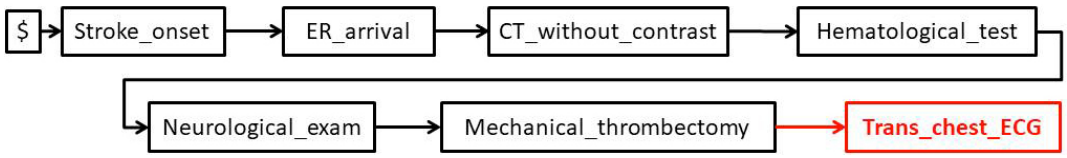


Figure 3.6: Trace given as a query to the transformer using SPE, with the predicted next activity colored in red.

3.5.2 Trace Classification Results

The second set of experiments focuses on evaluating the model's performance in trace classification. As anticipated, a connection has been established between a trace and its corresponding label by appending a class token to the trace when inputted into the transformer model. Then, by increasing the weights in the cross-entropy loss associated with the class token, the model was directed to prioritize the prediction of these tokens.
Two different sets of class labels were assigned to the traces:

- **Simple vs. Complex Patients**: A binary classification distinguishing patients with standard stroke presentations from those with rare variants or comorbidities (80-20 class distribution). This task supports quality assessment by verifying whether complex cases are identified and receive appropriately intensive monitoring.

- **Stroke Center vs. Stroke Unit**: A binary classification identifying the appropriate level of care facility based on case complexity (53-47 class distribution). This task supports resource planning by determining whether patients are being treated at facilities with the appropriate capabilities for their clinical needs.

The SPE-enhanced Transformer was compared against established baselines representing diverse architectural paradigms: a CNN-based classifier [97] using three one-dimensional convolution layers, an LSTM-based classifier with two LSTM layers, and a Random Forest model. All experiments were repeated 100 times with random initialization to ensure statistical reliability.

Setup		**Original Dataset**			**5% noise Dataset**			**10% noise Dataset**		
Dataset	**Model**	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Simple/complex patients	CNN	0.78	0.78	0.78	0.75	0.75	0.72	0.71	0.68	0.69
	LSTM	0.43	0.65	0.52	0.43	0.65	0.52	0.43	0.65	0.52
	Random Forest	0.75	0.75	0.74	0.72	0.73	0.71	0.70	**0.70**	0.68
	Transformer	**0.83**	**0.83**	**0.83**	**0.77**	**0.78**	**0.77**	**0.80**	0.69	**0.72**
SC/SU hospital	CNN	0.70	0.70	0.70	0.68	0.66	0.66	0.64	0.64	0.64
	LSTM	0.43	0.52	0.36	0.28	0.52	0.36	0.28	0.52	0.36
	Random Forest	0.66	0.66	0.66	0.66	0.66	0.66	0.63	0.63	0.63
	Transformer	**0.72**	**0.72**	**0.72**	**0.69**	**0.70**	**0.69**	**0.68**	**0.65**	**0.65**

Table 3.5: Results of the trace classification task. We compare the transformer model with SPE embedding to the CNN-based model, the LSTM-based model and the random forest model. All the metrics are weighted in order to account for dataset imbalance

The Transformer+SPE outperformed all baselines (see Table 3.5), with the improvement being particularly pronounced for the more challenging simple/complex classification where the class imbalance (80-20) and the subtle clinical distinctions between patient types pose significant difficulties for standard classifiers.

The LSTM's poor performance on trace classification (F1-scores of 0.52 and 0.36) is particularly noteworthy given its strong performance on next activity prediction in prior literature. This suggests that the hidden state representations learned by LSTMs, while effective for sequential prediction, **may not capture the global trace properties** necessary for classification. The Transformer's ability to attend to all positions simultaneously, combined with SPE's encoding of functional relationships, provides a more suitable representation for classification tasks. Notably, the advantage of the knowledge-enhanced approach persisted even when noise was introduced into the test traces, with the Transformer+SPE maintaining an

F1-score of 0.77 on 5% noisy data and 0.72 on 10% noisy data for simple/complex classification. This robustness advantage reinforces the principle that models grounded in stable domain knowledge are less susceptible to perturbations in surface-level input patterns.
These results validate that the learned representations, enriched with ontological structure, capture clinically meaningful distinctions between patient types, distinctions that enable automated quality assessment and resource planning in healthcare organizations. The ability to accurately classify patient complexity can inform resource allocation decisions (ensuring complex cases are directed to appropriately equipped facilities), quality monitoring (verifying that treatment protocols match patient needs), and operational planning (forecasting capacity requirements based on expected case mix).

3.5.3 External Validation on a Non-Clinical Process: BPI Challenge 2012

The experiments reported so far were conducted on a single stroke-management dataset, and the ontology exploited by SPE was a curated clinical taxonomy designed by domain experts. Two legitimate concerns follow from this setting. The first is whether the observed gain is an artefact of the specific care pathway considered, or whether it survives a change of domain. The second, and more subtle, is whether the gain depends on the availability of a **handcrafted, deep taxonomy**, which would considerably narrow the practical applicability of the method: most event logs, in business as well as in clinical settings, are not accompanied by a formal ontology. To address both concerns, the approach was replicated on a publicly available process-mining benchmark that is neither clinical nor equipped with an expert-designed ontology.

Dataset

The BPI Challenge 2012 event log [151] is one of the most widely adopted benchmarks in predictive process monitoring. It records the execution of a loan and overdraft application process at a Dutch financial institution, and it was released for the 8th International Business Process Intelligence Challenge. The log contains approximately 13,000 cases and 262,000 events, distributed over 36 event classes, and it is organised around three intertwined subprocesses, distinguishable from the event-name prefixes:

- **A** (*Application*): events describing the state of the loan application itself, from submission to final approval, cancellation or rejection.

- **O** (*Offer*): events describing the state of the offers sent to the customer, including their creation, dispatch and acceptance.
- **W** (*Work item*): events describing the tasks performed by the bank's employees on the case, each associated with a lifecycle transition (`SCHEDULE`, `START`, `COMPLETE`).

The choice of this benchmark is deliberate and, in a sense, adversarial with respect to the hypothesis under test. Unlike the stroke pathway, a loan application process is highly **standardised and strongly sequential**: the order in which activities occur is largely prescribed by the administrative procedure, there is little clinician-style discretionary reordering, and the number of distinct activities is smaller. This is precisely the regime in which a purely sequential prior is expected to be adequate and in which, therefore, the marginal value of a structural prior should be at its lowest.
Since no formal ontology is distributed with the log, the graph $G = (V, E)$ required by SPE was derived from the intrinsic structure of the event names, following the same construction principle described in Section 3.4.2: leaf nodes correspond to the individual event classes, intermediate nodes correspond to the sub-process to which each event belongs (A, O, W) and to its lifecycle transition, and a single root connects the sub-process nodes, guaranteeing connectivity. The resulting taxonomy is therefore *shallow and automatically induced*, rather than deep and expert-curated. All the remaining experimental conditions (80/10/10 split, loss masking of padding and start-of-sequence tokens, accuracy-at-k averaged over repeated runs) were kept identical to those of the stroke experiments, and the model size was fixed at 32 to allow a direct comparison with the corresponding row of Table 3.3.

Results

The replication confirms the direction of the effect: SPE improves next activity prediction on a domain that is neither clinical nor equipped with a curated ontology, raising top-1 accuracy from 0.8140 to 0.8601. Two observations are worth making about the *magnitude* of this improvement.

Metric	No SPE	SPE
Accuracy@1	0.8140	**0.8601**
Accuracy@3	0.9742	**0.9800**
Accuracy@5	0.9925	**0.9939**

Table 3.6: Next activity prediction on the BPI Challenge 2012 event log, with and without Structural Positional Encoding. Model size fixed at 32, experimental protocol identical to the stroke-management experiments of Table 3.3.

First, the gain is concentrated at $k = 1$ and becomes negligible at higher k. This is primarily an arithmetic consequence of the baseline's starting point: with accuracy@3 and accuracy@5 already at 0.9742 and 0.9925, the available headroom is 2.6 and 0.8 percentage points respectively, so the observed improvements of 0.6 and 0.1 points correspond to error reductions comparable to the one measured at $k = 1$. The behaviour at higher k is therefore uninformative in this setting, and the comparison with the stroke dataset, where accuracy@5 left more than 20 points of headroom, should not be read as a direct measure of how much structural knowledge contributes in the two domains. What can be said is that the top-1 decision is the only regime in which the two datasets are both far enough from saturation for the comparison to carry information, and it is the regime in which SPE improves on the baseline in both.
Second, and most relevant to the claim of general applicability, the ontology used here was **induced automatically from event-name conventions** and not designed by an expert. That a two-level, mechanically derived taxonomy is already sufficient to produce a measurable gain suggests that the benefit of SPE does not hinge on the availability of a rich formal ontology, but on the weaker and far more common condition that activities admit *some* meaningful grouping. This is a good property for deployment, and it is consistent with the interpretation given in Section 3.5.1: what the attention mechanism exploits is the membership of each activity in a functional group, and even a coarse notion of group carries signal.

Taken together, the two experiments delimit the empirical claim of this chapter more precisely than either could alone. SPE is not a clinical technique that happens to work on process data, nor a generic improvement to positional encoding: it is a mechanism for injecting a grouping structure over the token vocabulary, whose benefit scales with the amount of predictive signal that such grouping carries and that a purely sequential prior cannot capture.

3.6 Conclusion: Why Structured Knowledge Matters

The experimental validation of Structural Positional Encoding in stroke management demonstrates a fundamental principle: **when domain knowledge exists in structured form, its explicit integration into neural architectures can enhance performance, robustness, and clinical coherence**. The results across next activity prediction and trace classification tasks reveal that this integration addresses not merely an engineering challenge but a theoretical gap in how Transformers are applied to specialized domains.
The comparison between the stroke pathway and BPI2012 makes it possible to state, in

reasonably precise terms, the conditions under which the structural prior is expected to pay off, and those under which it is expected to be neutral or even counterproductive.

Depth and degree of hierarchy. SPE derives its embeddings from the spectrum of the graph Laplacian, and the eigenvectors associated with the smallest non-zero eigenvalues encode the coarsest partitions of the graph. A **deep taxonomy** therefore yields embeddings that are informative at several scales simultaneously: the leading components separate broad clinical functions (assessment, management, therapy), while later components discriminate between closely related leaves. The stroke ontology, with its four levels, exploits this property fully. A flat ontology, in which all activities hang from a single root, collapses this multi-scale structure into a one-hot-like partition and reduces SPE to little more than a learned type embedding; the BPI2012 taxonomy sits close to this degenerate end of the spectrum, which is consistent with its smaller absolute gain.

Branching factor. A high branching factor at the leaf level means that many activities share the same parent and therefore receive nearly identical structural embeddings, which weakens the model's ability to discriminate within a group. Conversely, a very low branching factor produces a chain-like graph whose Laplacian eigenvectors approximate a sequential ordering, making SPE asymptotically equivalent to a standard positional encoding. The regime in which SPE is most effective is therefore an **intermediate branching factor**, where groups are large enough to constitute a useful abstraction but small enough to remain informative.

Density and connectivity. The construction described in Section 3.4.2 assumes a connected graph in which path length reflects semantic similarity. Dense graphs, in which most activities are linked to most others, have a Laplacian spectrum with a large eigenvalue gap and produce embeddings that are close to uniform: the prior carries little information because the ontology asserts that everything is related to everything. Sparse, tree-structured ontologies preserve the correspondence between graph distance and semantic distance on which the method relies. This is an important practical caveat, since general-purpose biomedical terminologies are often far denser than a purpose-built care-pathway taxonomy, and a pruning or subsetting step is likely to be required before they can be used as input to SPE.

Data regime. Finally, the value of any inductive bias is inversely related to the amount of data available to learn the same regularity from scratch. The stroke cohort comprises a few hundred traces, the BPI2012 log several thousand, and the ordering of the observed relative

gains is consistent with the expectation that structural priors matter most in the small-data regime that characterises most clinical applications. This is not an incidental observation but a restatement of the thesis of this chapter: structure is a substitute for data, and its marginal value declines as data becomes abundant.

Perhaps more revealing than the magnitude of improvements is the pattern of what does and does not help. Standard sinusoidal positional encoding, the foundation of the original Transformer architecture, provides **no benefit and occasionally even degrades performance** in medical process monitoring. This negative result brings out a fundamental misalignment between architectural assumptions and domain structure. **Medical processes are not rigidly sequential chains** where position 3 differs fundamentally from position 5. Rather, they are functionally organized hierarchies where activities cluster by clinical goal (assessment, treatment, management), and the precise ordering within functional groups is often clinically immaterial. A patient might receive an ECG before or after a blood test during initial assessment, but both activities serve the same diagnostic purpose and their relative position carries little semantic weight. SPE's graph-based encoding captures this functional organization through ontological proximity rather than sequential position, providing an inductive bias that actually matches the problem structure rather than imposing generic assumptions about temporal order. The qualitative evaluation with domain experts reinforces this alignment with medical reasoning. When predicting the next activity after mechanical thrombectomy, the SPE model recommended trans-chest ECG monitoring, directly addressing the guideline mandate for cardiovascular monitoring within 24 hours post-procedure. The baseline model suggested encephalic MRI, clinically right but non-urgent, demonstrating that knowledge-enhanced predictions better respect **the priorities encoded in clinical protocols**. This is not merely about accuracy metrics; it reflects a model whose internal r**epresentations are structured by the same functional relationships** that guide expert clinical reasoning. When a system's predictions can be explained by reference to the ontological structure it was designed to **respect**, clinicians are more likely to understand its reasoning, identify when it might be incorrect, and trust its suggestions in practice.

The **generalizability of the approach** extends along multiple dimensions. Architecturally, even LSTM models benefit from structural encoding (52.2% vs 46.8% accuracy@1), indicating that SPE is not Transformer-specific but applicable to any model capable of incorporating **auxiliary embeddings**. Task-wise, the same knowledge-enhanced representations support both next activity prediction and trace classification without architectural modifications, with the class label treated simply as a special token to predict. This **unified framework** exemplifies how well-designed representations can serve multiple downstream objectives simultaneously, reducing the engineering complexity of deploying AI systems that must sup-

port diverse clinical workflows. These empirical findings illuminate broader principles about the relationship between knowledge and learning in specialized domains. Domain knowledge, encoded as graph structure and embedded through positional representations, serves as an inductive bias that constrains the hypothesis space to functions respecting known relationships. This constraint is particularly valuable in data-limited regimes characteristic of medical applications, enabling generalization from fewer examples by ruling out hypotheses that could fit training data but **violate ontological structure**. Critically, this knowledge integration is complementary rather than competitive with data-driven learning. SPE does not replace the attention mechanism's learned patterns; it structures the representation space within which those patterns are discovered. The model still learns which activity sequences are predictive, but it does so in a space where ontologically related activities have similar encodings, making it easier to discover regularities while **harder to overfit** to spurious correlations.

The **simplicity of the implementation** is noteworthy. Unlike approaches requiring complex attention mechanism modifications (K-BERT's visible matrices) or dual encoder pathways (KeBioLM's text-entity fusion), SPE integrates knowledge through a single matrix multiplication projecting Laplacian eigenvectors to embedding dimension, followed by element-wise addition to token embeddings. This architectural minimalism facilitates adoption, maintains compatibility with standard Transformers, and adds negligible computational overhead—the entire knowledge integration mechanism can be implemented in a few lines of code, with the graph eigendecomposition computed once offline. This combination of effectiveness and simplicity positions SPE as a practical tool rather than merely a research demonstration.

Looking beyond medical process monitoring, **SPE is applicable to any domain where sequences of discrete events constitute the primary data and where graphs or ontologies capture event relationships**: business process mining, software execution analysis, scientific experiment protocols, or time-series forecasting with known causal dependencies. Each possesses structured knowledge that could be encoded as ontologies and integrated to improve predictive models operating on sequential data. The ultimate goal of medical AI is not to replace clinical expertise but to augment it, providing decision support that respects established knowledge while adapting to individual patients and local contexts. The Structural Positional Encoding approach demonstrates that this integration need not be architecturally complex or computationally expensive, but can be achieved through thoughtful design aligning the inductive biases of neural models with domain structure. As Transformers continue their expansion across scientific and clinical domains, techniques like SPE offer a pathway for encoding expertise into powerful but inherently knowledge-agnostic

architectures, producing systems that are simultaneously **accurate and interpretable**, essential properties for AI systems operating in high-stakes environments where human lives depend on getting the answer right.

Chapter 4

Learning on Knowledge Graphs

The previous chapter demonstrated how structured domain knowledge, encoded through ontologies and embedded via positional encodings, can enhance Transformer-based models for sequential prediction tasks in medical process monitoring. However, medical knowledge extends far beyond process ontologies: it encompasses vast, **interconnected networks of entities** (diseases, symptoms, treatments, anatomical structures, genes) and their **semantic relationships** (causes, treats, located_in, part_of). These rich, multi-relational knowledge structures are naturally represented as **Knowledge Graphs** (KGs), where entities are nodes and typed relationships are edges, forming heterogeneous networks that capture the complexity of biomedical knowledge at scale.

While the Structural Positional Encoding approach leveraged ontological structure to guide attention mechanisms, it treated the knowledge graph primarily as a **static source** of node embeddings, computed once offline through spectral methods. This chapter explores a more direct paradigm: **learning directly on knowledge graphs**, where neural models are trained to perform inference over the graph structure itself. Rather than using knowledge as an auxiliary signal to enhance sequence modeling, the knowledge graph becomes the primary object of study. The model must learn to predict missing relationships, infer new connections, and answer complex multi-hop queries by navigating the graph's relational structure.

This shift from knowledge-enhanced sequence modeling to knowledge-centric graph learning introduces new theoretical and practical challenges. Knowledge graphs exhibit distinct properties that differentiate them from the graphs encountered in social networks or molecular structures: they are **heterogeneous** (containing multiple entity and relation types with distinct semantics), **incomplete** (the open-world assumption means absence of a fact does not imply its falsehood), and often **massive** in scale (biomedical KGs like UMLS contain millions of entities and relations). These characteristics demand specialized learning paradigms that can capture semantic richness, handle incompleteness, and scale efficiently.

This chapter provides a comprehensive treatment of deep learning on knowledge graphs, progressing from foundational concepts to state-of-the-art methods and culminating in experimental validation on biomedical knowledge inference tasks. The presentation is structured in three interconnected parts that together address the full pipeline from representation to inference over the graph:

Part I: Foundations of Graph-Based Knowledge Representation establishes the theoretical groundwork by formalizing knowledge graphs as semantic, multi-relational structures. It begins with the progression from simple homogeneous graphs (where all nodes and edges are of the same type) to heterogeneous information networks that capture the diversity of entity and relation types characteristic of real-world knowledge. The discussion then focuses on the specific structure of knowledge graphs, triples of the form (head entity, relation, tail entity), and examines how ontological foundations and schema design constrain and enrich these representations. A critical theoretical distinction is introduced: the contrast between open-world and closed-world assumptions, which fundamentally influences how models are trained, evaluated, and interpreted in the presence of incomplete knowledge.

Part II: Deep Learning Paradigms for Knowledge Graphs surveys the landscape of neural approaches for learning on KGs, organized by their underlying geometric or relational inductive biases. The first class of methods, **knowledge graph embeddings**, represents entities and relations as vectors in continuous spaces, where relational patterns (e.g., "X causes Y" or "A is_part_of B") are encoded through geometric operations like translations, rotations, or tensor decompositions applied to these embeddings. Beginning with the TransE model and progressing through its descendants (DIstMult, RotatE, ComplEx), this section examines how different geometric constraints capture distinct relational semantics. The second class of methods, **Graph Neural Networks** adapted for knowledge graphs, takes a complementary approach: rather than purely geometric scoring, these models perform iterative message passing over the graph structure, aggregating information from multi-hop neighborhoods while respecting relation types through attention mechanisms or relation-specific transformations.

Part III: Experimental Validation grounds these theoretical concepts in a comprehensive empirical study using the Unified Medical Language System (UMLS), one of the largest and most widely used biomedical knowledge graphs. UMLS integrates over 200 source vocabularies, covering millions of biomedical concepts and their relationships, making it an ideal testbed for evaluating knowledge graph learning methods in a real-world domain.

The transition from knowledge-enhanced Transformers (Chapter 3) to direct knowledge graph learning represents a shift in the field of knowledge integration. In SPE, knowledge guided the learning of activity sequences; here, the knowledge graph itself is the object to

be learned and reasoned over. This shift reveals both the power and the challenges of graph-based knowledge representation: while KGs provide semantic expressiveness and support for complex reasoning, they also introduce fundamental difficulties related to incompleteness and the evaluation of learned models under uncertainty. The experimental findings on UMLS will demonstrate that biomedical knowledge reasoning remains an open challenge. Neither purely geometric embeddings nor GNN-based approaches alone achieve human-level performance on complex queries but there are interesting signs that, under certain conditions, the latent space of these models can encode more complex reasoning than simply completing a triplet query.

4.1 Foundations of Graph-Based Knowledge Representation

The representation of structured knowledge has evolved through progressively more expressive formalisms, each designed to capture increasingly complex aspects of relational information. This section traces this evolution from simple homogeneous graphs to knowledge graphs enriched with semantic types and ontological constraints, establishing the formal foundations necessary for understanding deep learning methods on these structures.

4.1.1 From Homogeneous to Heterogeneous Networks

At this foundational level, a graph $G = (V, E)$ consists of a set of nodes V and a set of edges $E \subseteq V \times V$, where all nodes represent entities of a single type and all edges represent relationships of a single type. This uniform structure called **Homogeneous Graph**, provides a natural representation only for domains characterized by a single kind of entity and a single kind of relationship. A social network where nodes represent users and edges represent "friendship" exemplifies this structure: every node is a person, and every edge indicates the same symmetric relationship of mutual connection.

The elegance of homogeneous graphs, however, provide several limitations: they are fundamentally limited in their capacity to represent **the semantic diversity** inherent in real-world knowledge. Consider the challenge of representing biomedical information, where knowledge spans multiple entity categories (diseases, symptoms, genes, drugs) and multiple relationship types (treats, causes, located_in, part_of). A homogeneous representation could model only a single relation at a time. One might construct a "disease-symptom" graph where edges connect diseases to their symptoms, capturing statements like "Diabetes has_symptom Polyuria." However, this representation would be unable to simultaneously

capture "Metformin treats Diabetes" or "Mutation_in_Gene_X causes Diabetes" without creating entirely separate, disconnected graphs. This fragmentation of knowledge into multiple isolated graphs creates problems for inference. The therapeutic efficacy of a drug for a particular disease often depends critically on the symptoms that disease produces and the underlying genetic or molecular mechanisms that cause it. By representing these different facets of knowledge in separate graphs, we lose the ability to reason about their interconnections. A physician making a treatment decision needs to consider not just that "Drug A treats Disease B," but also whether the patient exhibits the symptoms characteristic of Disease B, whether they have genetic risk factors for Disease B, whether Drug A is contraindicated with other medications the patient is taking, and how Disease B interacts with other conditions the patient may have. This **holistic view of interconnected medical knowledge** cannot be captured by a collection of homogeneous graphs operating in isolation.

Heterogeneous Information Networks (HINs) [133] emerge as a response to these representational limitations, providing a formalism that explicitly captures multiple entity types and multiple relation types within a unified graph structure. The transition from homogeneous to **heterogeneous graphs** represents more than a simple technical extension, it reflects a fundamental shift in how we conceptualize knowledge representation, moving from a purely structural view (graphs as topology) to a semantic view (graphs as typed, meaningful relationships between typed entities). Formally, a heterogeneous information network (or Heterogeneus Graph) is defined as a graph $G = (V, E, \tau, \phi)$ where the structure is augmented with explicit type functions:

- V is the set of nodes representing entities in the knowledge domain.
- $E \subseteq V \times V$ is the set of edges representing relationships between entities.
- $\tau : V \to T_V$ is a node type mapping function that assigns each node to an entity type from a finite set T_V of entity types.
- $\phi : E \to T_E$ is an edge type mapping function that assigns each edge to a relation type from a finite set T_E of relation types.

The introduction of **explicit type mapping** functions fundamentally transforms the semantics of the graph structure. In a homogeneous graph, all edges are ontologically equivalent: they represent instances of the same relationship, differing only in the specific entities they connect. In a heterogeneous graph, the type of an edge carries semantic information that must be respected during inference and learning. An edge of type "treats" connecting a Drug entity to a Disease entity has a completely different meaning than an edge of type

"part_of" connecting two Anatomical Structure entities, even though both are simply edges in the graph topology.

This semantic differentiation has significant consequences for the analysis and learning processes applied to such graphs. Many algorithms designed for homogeneous graphs implicitly assume that all paths are equally meaningful, regardless of the types of nodes or edges involved. In the context of heterogeneous information networks, however, this assumption no longer holds.

Consider a path from *Drug_A* to *Disease_C* that passes through *Drug_B* (*Drug_A* $\rightarrow$ *Drug_B* $\rightarrow$ *Disease_C*). While such a path may suggest a potential relationship between the drug and the disease, its interpretation is entirely determined by the semantics of the intermediate relations. For instance, a path composed of (*Drug_A*, *similar_to*, *Drug_B*) followed by (*Drug_B*, *treats*, *Disease_C*) may indicate that *Drug_A* could also be effective in treating *Disease_C* due to shared mechanisms of action. In contrast, a path defined by (*Drug_A*, *contraindicated_with*, *Drug_B*) and (*Drug_B*, *treats*, *Disease_C*) conveys a fundamentally different implication, suggesting that *Drug_A* may be unsuitable or harmful for patients with *Disease_C*.

4.1.2 Knowledge Graphs

While heterogeneous information networks provide a general framework for representing typed entities and relationships, **knowledge graphs** [172] emerge as a specialized formalism optimized for representing **factual**, **assertional knowledge** about the world. The knowledge graph paradigm, rooted in the Semantic Web tradition and standardized through frameworks like RDF (Resource Description Framework) and OWL (Web Ontology Language), has become the dominant representation for large-scale structured knowledge in both academic research and industrial applications. At the heart of the knowledge graph formalism lies a simple representational primitive: the **(subject, predicate, object)** triple, more commonly written in knowledge graph literature as **(h, r, t)** where h is the **head** entity, r is the **relation**, and t is the **tail** entity. This triple-based representation encodes a single atomic fact or assertion about the world. For instance, the triple (*Aspirin*, *treats*, *Pain*) asserts that the entity "Aspirin" stands in the "treats" relationship to the entity "Pain"—in natural language, "Aspirin treats pain." Despite its apparent simplicity, this triple structure possesses several powerful properties that have made it the foundation for knowledge representation at massive scale.

Atomicity and Modularity: Each triple represents a single, indivisible assertion. This atomic granularity provides exceptional modularity: individual facts can be added, removed,

or updated independently without affecting other triples. In a traditional relational database, adding information about a new drug-disease relationship might require modifying table schemas, inserting rows in multiple linked tables, and maintaining integrity constraints across the database. In a knowledge graph, the same information is simply a new triple: (*New_Drug*, *treats*, *New_Disease*). This modularity dramatically simplifies knowledge acquisition, integration, and maintenance, particularly in domains like biomedicine where knowledge is constantly evolving and growing.

Compositionality and Emergent Structure: While individual triples are atomic, complex knowledge emerges through the composition of simple assertions. The statement "Aspirin treats headache, and headache is a symptom of migraine" is represented as two independent triples: (*Aspirin*, *treats*, *Headache*) and (*Headache*, *symptom_of*, *Migraine*). Although these triples can exist independently, their composition enables inference. If Aspirin treats a symptom of Migraine, then Aspirin may be useful for Migraine sufferers. This compositional structure supports inference by chaining relationships, a capability fundamental to question answering and inference in knowledge graphs. Consider a more complex example from biomedical knowledge. Suppose we have the following triples in a knowledge graph:

- (*Metformin*, *treats*, *Type_2_Diabetes*)
- (*Type_2_Diabetes*, *has_symptom*, *Hyperglycemia*)
- (*Type_2_Diabetes*, *has_symptom*, *Insulin_Resistance*)
- (*Type_2_Diabetes*, *associated_with*, *Obesity*)
- (*Obesity*, *risk_factor_for*, *Cardiovascular_Disease*)

From these simple, atomic assertions, we can compose more sophisticated understanding. A question like "What symptoms might improve with Metformin?" can be answered by composing (*Metformin*, *treats*, *Type_2_Diabetes*) with (*Type_2_Diabetes*, *has_symptom*, **X**), yielding **Hyperglycemia** and **Insulin_Resistance** as answers. A more complex query, "What cardiovascular risks might be reduced by treating Type 2 Diabetes?", requires chaining multiple relationships: *Type_2_Diabetes* → *associated_with* → **Obesity** → *risk_factor_for* → *Cardiovascula_Disease*. The knowledge graph structure makes these compositional inferences explicit and computationally tractable.

Standardization and Interoperability: The triple-based formalism aligns precisely with

established Semantic Web standards, particularly the RDF data model. This standardization facilitates knowledge integration and interoperability across systems. Different research groups, institutions, or countries can represent biomedical knowledge using the same triple format, and these knowledge graphs can be merged, aligned, and queried using standardized protocols (SPARQL query language, RDF serialization formats). This interoperability is particularly crucial in biomedicine, where knowledge is fragmented across numerous specialized databases, clinical systems, and research publications.

Formally, a knowledge graph is defined as a structure $K = (E, R, T)$ consisting of three fundamental components:

- $\mathbf{E}$ is the set of entities (nodes in the graph), representing real-world or abstract objects, concepts, or instances in the domain.
- $\mathbf{R}$ is the set of relations (edge types), representing the types of relationships that can hold between entities.
- $\mathbf{T} \subseteq \mathbf{E} \times \mathbf{R} \times \mathbf{E}$ is the set of true triples, representing the facts that are known to hold in the domain

Each triple $(h, r, t) \in \mathbf{T}$ makes a specific assertion: the relation $r \in \mathbf{R}$ holds between head entity $h \in \mathbf{E}$ and tail entity $t \in \mathbf{E}$.
A crucial aspect of knowledge graph semantics, often overlooked in technical treatments but essential for correct inference over relations, concerns the directionality and asymmetry of relations. Unlike edges in undirected graphs, which simply indicate the presence of some relationship between nodes without specifying direction or semantics, relations in knowledge graphs are inherently directional and carry specific semantic interpretations that depend critically on this directionality. The triple (*Diabetes*, *has_symptom*, *Polyuria*) makes a meaningful biomedical assertion: the disease diabetes manifests with the symptom polyuria. The reversed triple (*Polyuria*, *has_symptom*, *Diabetes*) is not merely false but it is semantically incoherent. Symptoms do not "**have**" diseases in any meaningful sense; symptoms are manifestations or signs of diseases, but the relation cannot be inverted while preserving its meaning. This directional semantics is not merely a matter of convention or labeling; it reflects the underlying causal and ontological structure of the domain. This directionality introduces fundamental asymmetry into the learning problem that deep learning methods must address. When training a model to predict missing relationships, the model must learn not only that certain entities are related, but the specific directional semantics of that relationship and how those semantics constrain inference. A model that learns that (*Drug_A*,

treats, *Disease_B*) is plausible must simultaneously learn that (*Disease_B*, *treats*, *Drug_A*) is implausible: diseases do not treat drugs. This asymmetry cannot be captured by models that treat relationships as symmetric similarity measures; it requires architectures that explicitly represent directional transformations or asymmetric scoring functions. The semantics of relations in knowledge graphs often exhibit specific logical properties that further structure the space of valid inferences and constrain learning:

Symmetry: A relation r is symmetric if whenever $(h, r, t) \in \mathbf{T}$, the reverse $(t, r, h) \in \mathbf{T}$ must also hold. In formal logic: $\forall h, t : r(h, t) \rightarrow r(t, h)$. Few biomedical relations are truly symmetric, but some examples include:

- "is_similar_to" in drug similarity networks: if *Drug_A is_similar_to Drug_B* (based on chemical structure or pharmacological profile), then *Drug_B is_similar_to Drug_A*
- "co-occurs_with" in symptom networks: if *Symptom_X co_occurs_with Symptom_Y* in clinical presentations, then *Symptom_Y co_occurs_with Symptom_X*
- "synonym_of" in terminology systems: if *Concept_A synonym_of Concept_B*, then *Concept_B synonym_of Concept_A*

Antisymmetry: A relation r is antisymmetric if whenever both (h, r, t) and (t, r, h) hold, it must be the case that $h = t$; that is, the only way for the relation to hold in both directions is if the entities are **identical**. In formal logic: $\forall h, t : [r(h, t) \wedge r(t, h)] \rightarrow h = t$. Many hierarchical or compositional relations are antisymmetric:

- "part_of" in anatomical hierarchies: if *Structure_A* is *part_of Structure_B* and *Structure_B* is *part_of Structure_A*, then *Structure_A* and *Structure_B* must be the same structure (you cannot have two distinct structures each being part of the other).
- "is_a" in taxonomic hierarchies: if *Disease_A is_a Disease_B* and *Disease_B is_a Disease_A*, then they must be the same disease concept
- "causes" in causal networks: if *Event_A causes Event_B* and *Event_B causes Event_A*, this indicates a simultaneous or identity relationship, not distinct causal events

Transitivity: A relation r is transitive if whenever (h, r, e) and (e, r, t) both hold, it follows that (h, r, t) must also hold. In formal logic: $\forall h, e, t : [r(h, e) \wedge r(e, t)] \rightarrow r(h, t)$. Transitivity enables multi-hop inference and is characteristic of many important biomedical relations:

- "is_a" in taxonomic hierarchies: if *Type_1_Diabetes is_a Diabetes* and *Diabetes is_a Metabolic_Disorder*, then *Type_1_Diabetes is_a Metabolic_Disorder*.

- "part_of" in anatomical composition: if *Mitochondria part_of Cell* and *Cell part_of Tissue*, then *Mitochondria part_of Tissue*.
- "located_in" in spatial hierarchies: if *Gene_X located_in Chromosome_7* and *Chromosome_7 located_in Nucleus*, then *Gene_X located_in Nucleus*.

Transitive relations are particularly important for knowledge graph completion because they enable the inference of new facts through path composition. If a model learns that a relation is transitive, it can apply this logical constraint to predict missing edges: whenever a path of length 2 exists through a transitive relation, a direct edge should also exist.

Inversion: Some relation pairs exist as explicit inverses of each other, where $(h, r1, t)$ if and only if $(t, r2, h)$. Common examples include:

- "treats" and "treated_by": if (*Aspirin*, *treats*, *Pain*) then (*Pain*, *treated_by*, *Aspirin*)
- "causes" and "caused_by": if (*Smoking*, *causes*, *Lung_Cancer*) then (*Lung_Cancer*, *caused_by, Smoking*)
- "part_of" and "has_part": if (*Mitochondria*, *part_of*, *Cell*) then (*Cell*, *has_part*, *Mitochondria*)

Inverse relations can be explicitly represented in the knowledge graph (both relation types exist as separate entries in R), or they can be implicit (only one direction is stored, and the inverse is computed on demand). Embedding models that respect inverse relations should learn that $r2 \approx -r1$ (for translational models) or that the scoring functions satisfy complementary properties that enforce inverse semantics.

These logical properties are not merely abstract mathematical curiosities. They represent deep structural regularities in how real-world knowledge is organized, particularly in domains like biomedicine where knowledge is grounded in physical, causal, and ontological structures. These properties provide powerful inductive biases that can be exploited during learning: a model that understands transitivity can infer exponentially many new facts through multi-hop inference; a model that respects antisymmetry will not generate logically impossible predictions; a model aware of inverse relations can use bidirectional evidence to strengthen inferences.

4.1.3 Relation to Clinical Knowledge Representations

Before examining how deep learning models operate on knowledge graphs, it is necessary to clarify the relationship between the triple-based formalism adopted in this chapter and the knowledge representations that are already established in clinical practice. Medicine is, in this respect, an atypical application domain: it does not merely possess unstructured expertise waiting to be extracted from data, but a long tradition of *explicitly codified* and partially machine-readable knowledge. Failing to position a knowledge-graph methodology against that tradition would leave the impression that the two are competing accounts of the same object, which they are not.

A taxonomy of codified clinical knowledge. Four families of representation are relevant here, and they differ in formality rather than in content. *Narrative clinical guidelines* are consensus documents in which recommendations are graded by strength and quality of the underlying evidence; they are authoritative but expressed in natural language and therefore not directly executable. *Diagnostic and therapeutic flowcharts*, of the kind found in the algorithmic textbooks adopted in Chapter 5, are the graphical distillation of those documents into decision trees whose nodes are diagnostic or therapeutic steps. *Institutional care protocols and clinical pathways* localize guideline content to the resources and organization of a specific setting, adding temporal ordering and resource constraints. Finally, *computer-interpretable guideline languages* and *rule-based decision support systems* render this knowledge formally executable. A comparative account of these formalisms and of the modelling choices that distinguish them is given in [107].

Procedural versus assertional knowledge. What unites these four families is that they encode **procedural, prescriptive** knowledge: they specify what a clinician *should do* for a patient in a given state. Their inference is deterministic, their execution auditable, and their semantics rests on an implicit closed-world assumption, whereby a condition not asserted about the patient is treated as absent. Their strength is normative authority combined with complete traceability: every recommendation can be referred back to a body of evidence and, in the executable case, every fired rule to an identifiable premise. Their weakness is equally structural. Coverage is restricted to predefined clinical scenarios, rendering system behavior outside this scope undefined rather than degraded. Additionally, maintenance relies on manual, expert-intensive curation, driving costs up in step with the growth of biomedical evidence rather than the operational capacity of the maintaining institutions [136].
A biomedical knowledge graph occupies a different position. It encodes **assertional, descriptive** knowledge: it states what *is the case* about biomedical entities, that a drug treats

a disorder or that a finding is associated with a condition, without prescribing any course of action, without temporal ordering, and without any notion of patient eligibility. Crucially, it does so under an explicit open-world assumption: the absence of a triple is read as absence of evidence rather than as evidence of absence. This is not a defect of the formalism but the condition of possibility for learning over it, since it is precisely the space of unasserted-yet-plausible facts that embedding models are trained to rank. Hogan et al. [67] frame this contrast as the coexistence of *deductive* knowledge, made explicit through ontologies and rules, and *inductive* knowledge, extracted from the graph through numerical representations, and observe that mature knowledge-graph systems are characterized by the interaction of the two rather than by the dominance of either.

Consequences for the scope of the present contribution. It follows that the methodology investigated in this chapter neither competes with nor substitutes for a clinical guideline, and that no experimental result reported iin the following sections should be read as evidence of clinical decision-making capability. Learning embeddings of UMLS does not yield recommendations; it yields a ranked plausibility over unasserted relational facts. The two representations operate on distinct layers of the same knowledge stack, and their interaction can be characterised along three directions, each of which corresponds to an experimental finding of this chapter.

(i) Knowledge graph learning as an upstream curation instrument. Guidelines, protocols and rule bases are authored on top of an assertional substrate of accepted biomedical facts, and the quality of the former is bounded by the completeness of the latter. The single-hop completion results of Section 4.6, with Hits@10 above 0.91 for every model considered, indicate that missing links in this substrate can be recovered with high recall. This positions the learned models as a triage mechanism for ontology maintenance: candidate assertions are surfaced automatically and ranked, and are subsequently submitted to expert adjudication before entering any prescriptive artefact. The model acts as a *hypothesis generator* feeding a human-authored and human-validated normative layer, and never as a bypass of it. This addresses directly the maintenance bottleneck identified in [136], without displacing the human judgement that gives the normative layer its authority.

(ii) Knowledge graph learning as a fallback beyond guideline coverage. Prescriptive artefacts are by construction silent on cases their authors did not anticipate: unusual comorbidity combinations, off-pathway presentations, or associations spanning semantic categories that no single guideline addresses. In these regions a task-network guideline model has no applicable plan and a rule engine no premise to satisfy, and the system returns an empty answer set. The compositional queries of Section 4.7 target precisely this regime:

a conjunctive query composing `treats` and `has_symptom` with `associated_with` crosses semantic boundaries that no individual clinical algorithm spans. The Hits@10 of 0.8339 obtained there is far below what an autonomous clinical recommendation would require, but it is obtained where the symbolic alternative offers nothing at all. The trade-off is explicit and must be stated as such: graceful degradation is purchased at the cost of the provenance and the formal guarantees that rule-based systems provide by design.

(iii) Symbolic constraints as a prior on sub-symbolic learning. The relationship is not unidirectional. The type-constrained negative sampling procedure adopted in Section 4.6.1 is itself a symbolic constraint, derived from the UMLS Semantic Network, injected into a sub-symbolic training objective: candidate corruptions violating the semantic type signature of a relation are excluded a priori, so that the model is never asked to discriminate against clinically absurd alternatives. Schema-level knowledge of the same kind that underpins rule-based systems is therefore already shaping the geometry of the learned space.

Placement within the thesis. Viewed across the dissertation as a whole, the two families of representation are addressed in complementary chapters rather than set in opposition. Chapter 3 operates on procedural knowledge, encoding a clinical process ontology through Structural Positional Encoding and thereby constraining a Transformer with the ordering information that a care pathway supplies. Chapter 5 operates on flowcharts in the strict sense, extracting root-to-leaf traversals from clinically validated decision algorithms and using them as ground-truth reasoning chains. The present chapter addresses the assertional counterpart, where no ordering and no prescription are available and the graph must be learned as an object in its own right. The methodological claim advanced here is consequently a modest and precise one: latent representations of a biomedical knowledge graph extend the reach of codified clinical knowledge into regions that its prescriptive forms do not cover, and support the maintenance of the substrate on which those forms are built, while remaining subordinate to expert validation wherever a decision affecting a patient is at stake.

4.2 Deep Learning Paradigms for Knowledge Graphs

The challenge of learning from knowledge graphs can be framed as follows: given an incomplete set of observed triples $\mathbf{T}$, learn a model that can (**1**) predict missing triples (link prediction/kg completition), (**2**) answer complex multi-hop queries (query answering), and (**3**) generalize to new entities or relations not seen during training (inductive inference). Deep learning approaches to this challenge fall into two broad paradigms that reflect different inductive biases: **geometric embedding methods** that represent entities and relations as

vectors in continuous spaces and score triples through geometric operations, and **graph neural network methods** that learn entity representations through iterative message passing over the graph structure. This section examines both paradigms, their theoretical foundations, and their relative strengths and limitations.

4.3 Knowledge Graph Embeddings: From TransE to Modern Approaches

The fundamental insight underlying knowledge graph embedding methods is that the symbolic structure of a KG (discrete entities connected by discrete relations) can be projected into a continuous vector space where relational patterns are preserved through geometric constraints. Once entities and relations are represented as vectors, the plausibility of any triple (h, r, t) can be assessed by computing a scoring function $f(h, r, t)$ that measures how well the triple conforms to the learned geometric pattern. This transformation from discrete symbolic structures to continuous representations offers several critical advantages: it enables the application of **gradient-based** optimization methods, facilitates scalability to large knowledge graphs through efficient **vector operations**, and most importantly, provides a natural framework for **generalization**: entities with similar roles in the graph should have similar embeddings, allowing the model to make reasonable predictions about unseen relationships.

The evolution of knowledge graph embedding methods can be understood as an exploration of increasingly sophisticated geometric and algebraic structures for capturing relational semantics. Each model makes different assumptions about how relations transform entity representations, leading to different capabilities in capturing various relational patterns that characterize real-world knowledge.

The TransE Model: Translation as Relation

The **TransE** model [25] established the foundational paradigm for geometric knowledge graph embeddings through an elegantly principle: representing **relations as translations** in embedding space. The core intuition is deeply geometric: if a relation r connects entity h to entity t, then in a well-structured embedding space, the relation should act as a displacement vector that, when added to h's position, yields approximately t's position.

Formally, given entity embeddings $h, t \in \mathbb{R}^d$ and relation embedding $r \in \mathbb{R}^d$, TransE encodes the principle that if (h, r, t) is a true triple, then $h + r \approx t$. The relation vector r acts as a translation from the head entity to the tail entity in the d-dimensional embedding space.

This geometric relationship is enforced through a distance-based scoring function:

$$f_{TransE}(h, r, t) = -||h + r - t||_p \tag{4.1}$$

where $|| \cdot ||_p$ denotes either the L1 norm (Manhattan distance) or L2 norm (Euclidean distance), and lower scores indicate higher plausibility. The negative sign converts the distance measure into a score that should be maximized for true triples. Triples where the translation holds should have $h + r$ close to t, resulting in small distance and thus high score (less negative).
The model (and all models described in the section) is trained using a **margin-based ranking loss** [160] that encourages true triples to score higher than corrupted triples:

$$\mathcal{L} = \sum_{(h,r,t)\in \mathbf{T}} \sum_{(h',r,t')\in \mathbf{T'}} max(0, \gamma + f(h, r, t) - f(h', r, t')) \tag{4.2}$$

where **T** is the set of observed triples, **T'** is a set of negative samples generated by corrupting true triples (typically by replacing either the head or tail entity with a random entity), and γ is a margin hyperparameter that controls the desired separation between positive and negative examples. This loss function pushes true triples to score at least γ points higher than corrupted triples, creating a margin of separation in the learned embedding space.
The translational principle captures several important relational patterns that are fundamental to knowledge representation:

- **Inversion**: If r_1 and r_2 are inverse relations (e.g., "treats" and "treated_by"), the model can learn $r_1 \approx -r_2$. This means applying r_1 followed by r_2 (or vice versa) returns to the starting point: $h + r_1 + r_2 \approx h + r_1 - r_1 = h$. This encoding of inverse relations emerges naturally from the translational geometry without requiring explicit supervision about which relations are inverses.

- **Transitivity**: A relation r is transitive if whenever (h, r, e) and (e, r, t) both hold, it follows that (h, r, t) must also hold. In formal logic: $\forall h, e, t : [r(h, e) \wedge r(e, t)] \rightarrow r(h, t)$. Transitivity enables multi-hop inference and is characteristic of many important biomedical relations. **Multi-hop** relational patterns can be captured through vector addition. If there exists a path (h, r_1, e) and (e, r_2, t), and if there's also a direct relation r_3 connecting h to t that is semantically equivalent to the two-hop path, the model can learn $r_3 \approx r_1 + r_2$. This enables inference through relation chains: "if Gene_X causes Disease_Y, and Disease_Y manifests_as Symptom_Z, then Gene_X indirectly_causes Symptom_Z" can be captured if indirectly_causes $\approx$ causes + mani-

fests_as.

Despite its elegance and empirical success on many benchmarks, TransE faces fundamental geometric limitations that arise from the strict translational constraint.
The most severe limitation concerns **symmetric** relations. If a relation r is symmetric (e.g., "is_similar_to"), the model must satisfy both $h + r \approx t$ and $t + r \approx h$ for the same pair of entities. Adding these two constraints:

- $h + r \approx t$ implies $r \approx t - h$
- $t + r \approx h$ implies $r \approx h - t = -(t - h)$

These two conditions can only be satisfied simultaneously if $r \approx 0$, forcing all symmetric relations to collapse to the zero vector. This means symmetric relations lose their distinctiveness: he model cannot distinguish between different symmetric relations like "is_similar_to," "co-occurs_with," or "synonym_of" because they all must be embedded near zero.
The second problem are **One-to-Many**, **Many-to-One**, and **Many-to-Many** Relations. TransE also struggles with relations where a single entity participates in multiple instances of the same relation type. Consider the relation "has_symptom": a disease like diabetes has many symptoms (polyuria, polydipsia, polyphagia, fatigue, etc.). If Diabetes + has_symptom should equal Polyuria, Polydipsia, Polyphagia, etc., simultaneously, this is geometrically impossible unless all these symptoms are embedded at the same point, which would make them indistinguishable. The translation Diabetes + has_symptom can only point to one location in space, yet it must point to many different symptoms. This creates a fundamental tension in the embedding geometry that TransE cannot resolve without sacrificing the distinctiveness of entity representations.

DistMult: Symmetric Bilinear Scoring

To address TransE's limitations with symmetric and complex relations, **DistMult** (Bilinear-Diagonal Model) [164] takes a fundamentally different approach, abandoning the geometric metaphor of translation in favor of an algebraic formulation based on bilinear scoring. Rather than modeling relations as vectors that translate entities, DistMult represents relations as diagonal matrices that modulate similarity between entities.
The core idea is to score triples based on a **weighted similarity** between head and tail embeddings, where the weights are relation-specific. Given entity embeddings $h, t \in \mathbb{R}^d$ and a relation represented as a diagonal matrix with diagonal entries $r \in \mathbb{R}^d$, the scoring function is:

$$f_{DistMult}(h, r, t) = \sum_{i=1}^{d} h_i r_i t_i = \mathbf{h}^T diag(\mathbf{r})\mathbf{t} \tag{4.3}$$

where $diag(r)$ creates a diagonal matrix with r on the diagonal. This can be more compactly written as:

$$f_{DistMult}(h, r, t) = \langle \mathbf{h}, \mathbf{r}, \mathbf{t} \rangle = \sum_{i=1}^{d} h_i r_i t_i \tag{4.4}$$

This trilinear form computes a weighted dot product ($\langle \cdot \rangle$ represent the generalized dot product) between head and tail embeddings, where each dimension's contribution is modulated by the corresponding component of the relation vector. High scores indicate that h and t are similar in the dimensions that relation r considers important. The DistMult scoring function can be understood as measuring relation-specific similarity. Each relation defines which dimensions of entity embeddings are **relevant** for that relationship.
The critical advantage of DistMult is its natural handling of **symmetric** relations. The scoring function is inherently symmetric in h and t:

$$f_{DistMult}(h, r, t) = \sum h_i r_i t_i = \sum t_i r_i h_i = f_{DistMult}(t, r, h) \tag{4.5}$$

This means DistMult automatically assigns the same score to (h, r, t) and (t, r, h), making it ideal for symmetric relations like "is_similar_to" or "co-occurs_with." Unlike TransE, which must collapse symmetric relations to zero, DistMult can learn rich, distinctive representations for different symmetric relations by varying the diagonal elements of r.
The symmetry that makes DistMult excel at symmetric relations becomes a liability for **asymmetric** relations. This model cannot properly represent antisymmetric relations like "part_of" or "causes" because it always assigns identical scores to (h, r, t) and (t, r, h). If Hand is part_of Arm, DistMult will incorrectly assign the same plausibility to Arm being part_of Hand.
Another problem emerge in case of **inversion**. DistMult cannot distinguish between a relation and its inverse because both directions receive the same score. This is acceptable if inverse relations are explicitly represented as different relation types (r_1 and r_2), but it means DistMult cannot learn the inverse structure implicitly.
Despite these limitations, DistMult has proven highly effective in practice, particularly for knowledge graphs with many symmetric relations or where the symmetric component of asymmetric relations carries important signal. Its simplicity, requiring only $O(d)$ parameters per relation compared to $O(d^2)$ for full matrix models, also makes it highly scalable.

RotatE: Rotation in Complex Space

RotatE [134] represents a sophisticated synthesis of translational and symmetric approaches by moving from real-valued vector spaces to complex-valued representations. The key insight

is that relations can be modeled as rotations in complex space, with each relation defining an element-wise rotation applied to entity embeddings.

Given complex-valued entity embeddings $h, t \in \mathbb{C}^d$ and relation embeddings $r \in \mathbb{C}^d$, where each complex number can be written in polar form as $re^{i\theta}$, RotatE constrains relation embeddings to lie on the unit circle in each dimension: $|r_i| = 1$ for all dimensions i. This constraint means each component of r is a pure rotation without any scaling: $r_i = e^{i\theta_i}$ for some angle θ_i.

The RotatE scoring function models the principle that tail embeddings should be obtainable by rotating head embeddings:

$$t = h \circ r \tag{4.6}$$

where $\circ$ denotes element-wise (Hadamard) product in complex space. The corresponding distance-based score is:

$$f_{Rotate} = -\|h \circ r - t\| \tag{4.7}$$

where $\|\cdot\|$ is the modulus (magnitude) in complex space.

Geometric Intuition: In complex space, multiplication by $e^{i\theta}$ rotates a vector by angle θ. Element-wise multiplication means each dimension of the entity embedding is rotated independently by the corresponding angle in the relation embedding. This provides enormous flexibility: different dimensions can be rotated by different amounts, allowing relations to **affect different aspects** of entity semantics differently.

RotatE elegantly handles all the challenging relational patterns described 4.1.2:

- **Symmetry**: If r is symmetric, RotatE can learn $r = [1, 1, ..., 1]$ (the identity rotation with all angles equal to 0 or 2π), satisfying $h \circ r = h$ and $t \circ r = t$, thus making $f(h, r, t) = f(t, r, h)$. Unlike TransE, this doesn't collapse the relation to zero and r remains a well-defined vector.

- **Antisymmetry**: Antisymmetric relations are captured by non-trivial rotations. If $h \circ r = t$, then for the reverse to hold ($t \circ r = h$), we would need $h \circ r \circ r = h$, which requires $r \circ r = [1, 1, ..., 1]$. This only occurs if r is the identity or has order 2 (180° rotations), allowing RotatE to distinguish most antisymmetric relations from symmetric ones.

- **Inversion**: If r_1 and r_2 are inverse relations, RotatE can learn $r_2 = r_1^{-1}$, where the inverse is computed element-wise as the complex conjugate (negating all angles). This means applying r_1 then r_2 returns to the original position: $h \circ r_1 \circ r_2 = h \circ r_1 \circ r_1^{-1} = h$.

- **Transitivity**: Transitive relations like "is_a" or "part_of" benefit particularly from rotation-based composition. If r is transitive and we observe paths of length k through

relation r, the composition $r \circ r \circ ... \circ r$ (k times) should approximate r^k. In complex space, this is simply rotation by k times the angle, which RotatE can learn naturally.

ComplEx: Bilinear Models with Complex Embeddings

While TransE and RotatE are fundamentally distance-based or translation-based models, **ComplEx** [147] takes a different approach by introducing a bilinear scoring function operating on complex-valued entity and relation embeddings. This represents a generalization of DistMult into the complex domain, combining the relation-specific similarity interpretation of DistMult with the expressiveness advantages of complex representations.
Given complex-valued embeddings $h, r, t \in \mathbb{C}^d$, ComplEx defines the scoring function as:

$$f_{ComplEx}(h, r, t) = Re(\sum_{i=1}^{d} h_i r_i \bar{t}_i) = Re(\mathbf{h}^T diag(\mathbf{r}) \mathbf{t}) \tag{4.8}$$

where $\bar{t}$ denotes the complex conjugate of t (negating the imaginary components), and $Re(\cdot)$ extracts the real part of the complex result. This can be expanded as:

$$f_{ComplEx}(h, r, t) = Re(\langle \mathbf{h}, \mathbf{r}, \bar{\mathbf{t}} \rangle) \tag{4.9}$$

The use of complex conjugates is crucial for introducing **asymmetry** into the scoring function. Without conjugation, the function would be symmetric in h and t like DistMult. With conjugation, $f(h, r, t) \neq f(t, r, h)$ in general because:

$$f_{ComplEx}(h, r, t) = Re(\sum h_i r_i \bar{t}_i), \tag{4.10}$$

$$f_{ComplEx}(t, r, h) = Re(\sum t_i r_i \bar{h}_i) \tag{4.11}$$

These are not equal unless the embeddings are real-valued (in which case ComplEx reduces to DistMult). The asymmetry introduced by conjugation enables ComplEx to distinguish between head and tail positions and thus model directed relations.
ComplEx provides high representational flexibility:

- **Antisymmetry**: The complex conjugate in the scoring function naturally handles antisymmetric relations. The model can learn embeddings where (h, r, t) scores highly but (t, r, h) scores poorly by exploiting the phase information in complex space.
- **Symmetry**: For symmetric relations, ComplEx can learn embeddings where the imaginary components cancel out, reducing the scoring function to effectively real-valued

similarity (like DistMult). This allows symmetric relations to be properly represented without collapsing to trivial solutions.

- **Inversion**: ComplEx can distinguish inverse relation pairs by learning appropriate complex structures. If r_1 and r_2 are inverses, the model learns r_2 patterns that undo the scoring effects of r_1.

ComplEx uses a scoring function based on complex vector products to assign high scores to true triples (h, r, t). This function is good at modeling asymmetry and symmetry, but it does not enforce relational composition. It doesn't guarantee that combining two facts like (h_1, r, h_2) and (h_2, r, h_3) will yield a high score for (h_1, r, h_3) just because of its embedding structure. Representing transitivity requires the relation representation to be **closed under composition**, meaning you can algebraically "chain" relation embeddings so composing r with itself implies r again and ComplEx's scoring formulation simply doesn't provide such a mechanism (also DistMult by definition).

Pattern	**TransE**	**DistMult**	**RotatE**	**ComplEx**
Symmetry	×	✓	✓	✓
Antisymmetry	✓	×	✓	✓
Inversion	✓	×	✓	✓
Transitivity	✓	×	✓✓	×

Table 4.1: Relational patterns captured by different knowledge graph embedding models. (✓✓ indicates particular strength, ✓ indicates capability, × indicates limitation.)

4.4 Graph Neural Networks for Relational Learning

While embedding methods treat the knowledge graph as a collection of independent triples to be scored through geometric operations, **Graph Neural Networks** take a fundamentally different approach rooted in the principles of local neighborhood aggregation and iterative refinement. Rather than learning static entity representations through global optimization of a scoring function across all triples, GNNs compute entity representations dynamically by propagating and aggregating information along graph edges, allowing each entity's representation to be informed by the structural context of its multi-hop neighborhood.

This paradigm shift introduces a powerful relational **inductive bias**: the assumption that an entity's representation should be influenced by the entities it is connected to, the relations through which they are connected, and the broader topological context in which it is embedded. This principle aligns naturally with how knowledge is actually structured: the

meaning and properties of a medical concept are tied to its relationships with other concepts. Understanding "Type 1 Diabetes" requires knowing not just its isolated properties, but that it is a subtype of "Diabetes," that it is caused by autoimmune destruction of pancreatic beta cells, that it is treated with insulin, that it manifests with specific symptoms, and that it affects primarily younger populations. These relationships collectively define what Type 1 Diabetes is, and GNNs provide a computational framework for learning representations that respect this relational structure.

4.4.1 Message Passing Neural Networks: The General Framework

The **message passing** [27] framework provides a unified mathematical formulation for a broad family of graph neural network architectures. This framework abstracts the common computational pattern underlying most GNN variants: nodes iteratively update their representations by receiving, aggregating, and processing information from their neighbors. At each layer l of a GNN, every node v updates its representation $h_v^{(l)}$ through a two-stage process. First, the node collects messages from all its neighbors. A message from neighbor u to node v is computed by a message function that considers the representations of both nodes and potentially the edge connecting them:

$$msg_{u \to v}^{(l)} = MESSAGE(h_u^{(l-1)}, h_v^{(l-1)}, e_{uv}) \tag{4.12}$$

where e_{uv} represents edge features or, in knowledge graphs, the relation type connecting u to v. The **message** function transforms the neighbor's representation into information relevant for updating v's representation.
Second, all incoming messages are aggregated through a **permutation-invariant** aggregation function, yielding a neighborhood summary:

$$m_v^{(l)} = AGGREGATE(msg_{u \to v}^{(l)} : u \in \mathcal{N}(v)) \tag{4.13}$$

where $\mathcal{N}(v)$ denotes the set of neighbors of node v. The permutation invariance of the aggregation function is essential because graph neighborhoods have no natural ordering. Common choices for **AGGREGATE** include:

- Sum aggregation: $m_v^{(l)} = \sum_{u \in N(v)} msg_{u \to v}^{(l)}$
- Mean aggregation: $m_v^{(l)} = \frac{1}{\mathcal{N}(v)|} \sum_{u \in N(v)} msg_{u \to v}^{(l)}$
- Max aggregation: $m_v^{(l)} = max_{u \in \mathcal{N}(v)} msg_{u \to v}^{(l)}$ (element-wise maximum)

- Attention-weighted aggregation: $m_v^{(l)} = \sum_{u \in \mathcal{N}(v)} \alpha_{uv} msg_{u \to v}^{(l)}$, where α_{uv} are learned attention weights

Finally, the node updates its own representation by combining its previous representation with the aggregated neighborhood information:

$$h_v^{(l)} = UPDATE(h_v^{(l-1)}, m_v^{(l)}) \tag{4.14}$$

The **UPDATE** function typically applies a non-linear transformation, often implemented as a feed-forward neural network with activation functions. A common pattern is:

$$h_v^{(l)} = \sigma(W^{(l)}[h_v^{(l-1)} \| m_v^{(l)}] + b^{(l)}) \tag{4.15}$$

where $\|$ denotes concatenation, $W^{(l)}$ is a learnable weight matrix, $b^{(l)}$ is a bias vector, and σ is a non-linear activation function (ReLU, tanh, etc.).

The power of this iterative scheme lies in its ability to capture multi-hop dependencies. After K layers of message passing, each node's representation has aggregated information from its K-hop neighborhood (all nodes reachable within K steps). This enables the model to capture complex relational patterns that extend beyond immediate connections. This multi-hop aggregation provides a natural mechanism for compositional inference (in the sense defined in Section 4.7). If we want to infer that "Drug_A treat Disease_C" based on the path "Drug_A treats Disease_B $\leftarrow$ similar_to $\rightarrow$ Disease_C," a 2-layer GNN can aggregate information along this path, allowing Drug_A's representation to be influenced by Disease_C through the intermediate Disease_B.

However, this multi-hop aggregation also introduces challenges. As the number of layers increases, information from distant nodes gets mixed and diluted through repeated aggregation, leading to **over-smoothing** [122], a phenomenon where node representations become increasingly similar and lose their distinctiveness. In knowledge graphs with high-degree hub entities (concepts that appear in thousands of triples), over-smoothing can be particularly severe, as these hubs dominate neighborhood aggregations and cause many entities to receive similar signals.

4.4.2 Relational Graph Convolutional Networks

For knowledge graphs, the standard GNN message passing framework must be adapted to handle the critical aspect that distinguishes knowledge graphs from generic graphs: multiple relation types. In a social network, all edges represent the same relationship type (friendship, follows, etc.), allowing a single weight matrix to transform all messages uniformly. In

a knowledge graph, edges of different types, "treats," "causes," "part_of," "has_symptom", carry fundamentally different semantics and should be processed differently.

Relational Graph Convolutional Networks (R-GCN) [124] address this through relation-specific transformations: rather than using a single weight matrix shared across all edge types, R-GCN maintains separate weight matrices for each relation type, allowing the model to learn distinct aggregation patterns for different relationships. The R-GCN update rule for entity e at layer l is:

$$h_e^{(l)} = \sigma\left(\sum_{r\in\mathbb{R}} \sum_{t:(e,r,t)\in T} \frac{1}{c_{e,r}} W_r^{(l)} h_t^{(l-1)} + W_0^{(l)} h_e^{(l-1)}\right) \tag{4.16}$$

Let's unpack this equation component by component:

Relation-Specific Aggregation: The outer sum $\sum_{r\in\mathbb{R}}$ iterates over all relation types in the knowledge graph. For each relation type r, the inner sum $\sum_{t:(e,r,t)\in T}$ aggregates over all tail entities t that are connected to entity e through relation r. This means the model processes each relation type separately before combining their contributions. Each relation r has its own learnable weight matrix $W_r^{(l)} \in \mathbb{R}^{d^{(l)}\times d^{(l-1)}}$, where $d^{(l-1)}$ is the dimension of representations in layer $l-1$ and $d^{(l)}$ is the dimension in layer l. When aggregating information from a neighbor t through relation r, the neighbor's representation $h_t^{(l-1)}$ is first transformed by $W_r^{(l)}$. This transformation is specific to the relation type, allowing the model to extract relation-dependent features.

Normalization: The term $\frac{1}{c_{e,r}}$ is a normalization constant, typically set to $c_{e,r} = |t : (e, r, t) \in T|$, the in-degree of entity e with respect to relation r (the number of incoming edges of type r). This normalization prevents entities with many incoming connections from receiving disproportionately large aggregated signals, which would cause their representations to dominate during training and lead to numerical instability.

Self-Loop: The term $W_0^{(l)} h_e^{(l-1)}$ adds a transformed version of the entity's own previous representation. This self-loop ensures that an entity's representation is not entirely determined by its neighbors: it retains information from previous layers and can maintain stable features even as neighborhood information is integrated. The self-loop is essential for preventing information loss during aggregation.

The relation-specific transformation captures a crucial intuition: **the meaning of a neighbor depends on the relationship type**. Consider an entity "Type 1 Diabetes." Its

neighbors include:

- "Diabetes" (connected via is_a relation)
- "Insulin" (connected via treated_by relation)
- "Pancreatic Beta Cells" (connected via affects relation)
- "Polyuria" (connected via has_symptom relation)

These neighbors provide different kinds of semantic information. The "is_a" neighbor tells us about taxonomic categorization, the "treated_by" neighbor tells us about therapeutic interventions, the affects neighbor tells us about pathological mechanisms, and the "has_symptom" neighbor tells us about clinical manifestations. R-GCN's relation-specific weight matrices allow the model to process each type of information appropriately: extracting taxonomic features from is_a neighbors, therapeutic features from treated_by neighbors, and so on.

With this comprehensive overview of geometric and neural approaches to knowledge graph learning, we now turn to the critical challenge that distinguishes real-world knowledge systems from idealized benchmarks: the open-world assumption and its implications for training, evaluation, and inference.

4.5 Open-World vs. Closed-World Assumptions: Implications for Learning

The distinction between **open-world** and **closed-world** assumptions has profound practical implications for how knowledge graph learning systems are trained and evaluated. This section examines these implications in detail, focusing on the specific challenges they create for deep learning methods and the strategies developed to mitigate them.

4.5.1 False Negative Contamination in Training Data

The standard training procedure for knowledge graph embedding models relies on a contrastive objective: given a true triple $(h, r, t) \in T$, generate negative samples by corrupting the triple, typically by replacing the head or tail entity with a random entity from E, and train the model to assign higher scores to true triples than to corrupted ones. This procedure implicitly **assumes that corrupted triples are false**, which holds under the

closed-world assumption but fails under the open-world assumption characteristic of real knowledge graphs.

Consider a biomedical knowledge graph where (Aspirin, treats, Pain) is a known fact. If we corrupt this triple by replacing "Pain" with a randomly selected entity, there is a non-negligible probability of selecting another condition that Aspirin actually treats but which is not yet recorded in the knowledge graph, for instance, (Aspirin, treats, Inflammation). If this triple happens to be true but unobserved, using it as a negative sample sends an incorrect training signal to the model: it is told to assign low scores to a relationship that actually holds. The severity of false negative contamination depends on several factors:

- **Knowledge Graph Density**: Sparser graphs have lower collision probability. If only 0.1% of possible triples are true, random corruption has a 99.9% chance of generating a true negative. However, many real-world KGs are not uniformly sparse; certain relations are dense (e.g., "has_synonym" in UMLS might connect many entities), while others are sparse (e.g., specific drug-disease treatment relationships).

- **Entity Frequency Distribution**: High-frequency entities (those appearing in many triples) are more likely to be sampled during corruption, increasing collision probability. In UMLS, common concepts like "Disease" or "Symptom" appear in thousands of triples, making them frequent corruption targets.

- **Relation Semantics**: Some relations naturally have many true instances (e.g., "related_to" in UMLS is extremely broad and has many valid instances), while others are highly specific (e.g., "is_biochemical_structure_of" has few valid instances). False negative rates vary dramatically across relation types.

These challenges are particularly acute in biomedical knowledge graphs. Medical knowledge is characterized by several distinct features. First, it undergoes rapid evolution as new drugs, diseases, and treatment relationships are discovered continuously, causing knowledge graphs to quickly become outdated. Second, there is heterogeneous coverage across different medical domains: well-studied diseases such as cancer and heart disease have extensive documentation, while rare diseases may have minimal information. Third, source integration presents significant challenges; UMLS integrates over 200 source vocabularies with overlapping but inconsistent content, meaning a fact may be present in one source but absent in another, creating apparent incompleteness. Finally, publication bias affects the availability of data, as negative results (for example, findings that Drug X does not treat Disease Y) are rarely published or recorded, making it impossible to definitively identify true negatives. As a result, any deep learning model trained on Biomedical KGs operates in a profoundly open-world setting.

4.5.2 Strategies for Mitigating False Negatives

Several strategies have been developed to reduce false negative contamination, each with tradeoffs between effectiveness and computational cost:

Filtered Negative Sampling: The most widely adopted strategy is filtered evaluation: during testing, when ranking a true triple (h, r, t) against corruptions, any corrupted triple that appears in the training, validation, or test sets is excluded from consideration. This ensures that the model is not penalized for assigning high scores to known facts that were held out for testing. However, filtered evaluation addresses only the evaluation phase: it does not prevent false negatives from contaminating training. To address this, some approaches apply filtering during training as well: when generating negative samples, check each candidate against all known triples (training, validation, test) and reject any matches. This significantly increases computational cost, as it requires maintaining a hash table of all observed triples and checking every negative sample.

Type-Constrained Negatives: Rather than corrupting triples with completely random entities, type-constrained negative sampling replaces entities only with others of the same semantic type. For instance, when corrupting (Aspirin, treats, Diabetes), the tail entity "Diabetes" would be replaced only with other diseases, not with drugs, symptoms, or genes. This strategy reduces false negative probability by exploiting domain constraints: if we know a priori that "treats" can only connect drugs to diseases, then replacing "Diabetes" with "Gene_XYZ" generates a triple that is semantically invalid (a type violation) rather than false. Type-constrained sampling relies on the availability of type information, which is present in ontology-based KGs like UMLS but may not be available in general-domain KGs.

Adversarial Negative Sampling: Rather than random sampling, adversarial methods train a generator network to produce hard negatives, triples that are plausible (high score from the current model) but false [95]. The intuition is that hard negatives force the model to refine its decision boundaries, while random easy negatives (which are trivially false) provide little training signal. However, adversarial approaches face a critical risk in open-world settings: if the generator produces a triple that looks highly plausible to the model, there is an increased risk that the triple is actually true but unobserved. The generator may inadvertently discover valid knowledge that is not in the training set, only for that knowledge to be labeled as negative. This represents a form of learned false negative contamination, where the very act of generating hard negatives increases the probability of collision with true facts.

Semi-Supervised and Self-Training Approaches: Some recent methods attempt to explicitly model the uncertainty in negative labels. Rather than treating all negative samples as definitively false, these approaches assign soft labels or confidence weights to negatives based on their plausibility. Triples that score highly according to the current model (and thus might be true but unobserved) are down-weighted during training, reducing their influence on the loss. Self-training approaches iteratively re-label high-confidence predictions as positive examples, effectively expanding the training set with inferred knowledge. This can help recover from false negative contamination if the model has correctly identified a true but unobserved fact and can now learn from it as a positive example. However, this strategy risks compounding errors if the model's high-confidence predictions are incorrect.

4.6 Experimental Validation: Knowledge Graph Completion

This section presents the experimental contribution of this thesis to knowledge graph learning on biomedical data. The evaluation reported hereafter was designed and carried out as an original part of the present work. It includes the construction of the benchmark dataset from the 2025AA release of UMLS, the choice and configuration of the six models under comparison, and the type-constrained negative sampling procedure. It also covers the analysis of results across single-hop and compositional inference tasks. This evaluation does not report pre-existing benchmarks drawn from the literature. The aim of this experimental campaign is twofold: first, to assess the behavior of representative geometric, neural, and hybrid approaches on a current, large-scale, real-world biomedical knowledge graph rather than on the small legacy benchmarks (such as WN18 or FB15k-237) typically used in the knowledge graph embedding literature; and second, to investigate the extent to which the learned latent spaces capture compositional inference patterns that can be exploited for downstream biomedical applications. The testbed adopted for this study is the UMLS, one of the largest and most comprehensive biomedical knowledge graphs currently available.

4.6.1 Experimental Setup and Datasets

UMLS Knowledge Graph Construction

The **Unified Medical Language System** in its 2025AA [23] release represents one of the most ambitious efforts in biomedical knowledge integration, containing over 200 source

vocabularies that span the entire landscape of medical terminology: from clinical diagnoses (ICD-10, SNOMED CT) to medications (RxNorm), from anatomical concepts (FMA) to genetic information (Gene Ontology), from procedures (CPT) to research classifications (MeSH). However, this comprehensiveness comes with significant heterogeneity in both coverage and quality across different semantic domains. For this experimental study, the complete UMLS Metathesaurus was processed to construct a knowledge graph suitable for deep learning experiments. The resulting UMLS knowledge graph contains:

- **2,096,327** unique medical concepts CUIs
- **438** relation types (219 base relation types plus their inverses)
- **3,504,739** entity-relation-entity assertions

The raw UMLS contains numerous relation types with extremely low cardinality. Some relations have a cardinality that falls below 10% of the mean cardinality calculated across all relations ($mean(|rels|) - 10\% \approx 350$) in the knowledge graph. These ultra-rare relations cannot provide enough examples for models to learn meaningful transformation patterns or scoring functions. Additionally, many entities participate in only a single relation, with only a small subset of entities having more than 2 relations (common entities). One last thing to add: many entities were translations or synonyms. All "duplicate occurrences" were removed and only one entity was retained.

Inverse Relation Augmentation: To ensure bidirectional information flow in graph neural network models and to enable inference in both directions through asymmetric relations, the knowledge graph was augmented with explicit inverse relations. For every triple (h, r, t) in the original UMLS, a corresponding inverse triple $(t, r^{(-1)}, h)$ was added, where $r^{(-1)}$ is the inverse of relation r. This augmentation is standard practice in knowledge graph learning and effectively doubles the number of relation types while providing symmetric access to relational information regardless of the direction in which relationships were originally encoded.

Data Split Strategy: The knowledge graph was partitioned into training (70%), validation (10%), and test (20%) sets using a **stratified random split** to ensure balanced representation of all relation types. Stratification ensures that each relation type appears in all three splits proportionally to its overall frequency, preventing scenarios where rare relations might be entirely absent from validation or test sets. This is critical for evaluating model performance across the full diversity of UMLS relation types rather than only on frequent relations.

Models Evaluated

The experimental study evaluated representative methods from both major paradigms of knowledge graph learning: geometric embedding models and graph neural networks, along with hybrid approaches that combine both. Regarding the hyperparameters, the objective of this work is not to achieve state-of-the-art performance on a specific dataset, but rather to demonstrate that the latent space encodes meaningful relational structure and the potential application in bio-medicine. Therefore, the model dimension is fixed at **512** across all experiments, and the number of negative samples is set to **1**. The choice of a single negative sample is motivated by computational efficiency and conceptual clarity: it provides sufficient contrastive signal for the model to learn discriminative features while minimizing training overhead. Since the goal is to validate the interpretability of the learned representations rather than maximize performance metrics, this minimal contrastive setup allows us to isolate and analyze the relational patterns captured in the latent space without introducing unnecessary complexity from multiple negative samples.

Geomtric embedding models:

- **TransE**: The foundational translational model
- **DistMult**: Symmetric bilinear model
- **RotatE**: Rotation-based model in complex space, expected to excel on hierarchical and transitive relations prevalent in medical ontologies
- **ComplEx**: Complex-valued bilinear model providing high expressiveness for diverse relation types

Graph Neural Network models:

- **R-GCN + TransE**: Hybrid architecture using 2-layer R-GCN to compute contextualized entity representations, then applying TransE scoring function
- **R-GCN + RotatE**: Hybrid architecture using 2-layer R-GCN representations with RotatE scoring, combining neighborhood aggregation with compositional rotation-based scoring

The inclusion of hybrid models tests the hypothesis that GNN-based neighborhood aggregation can provide richer entity representations that, when combined with strong geometric scoring functions, might outperform either approach alone.

Training Configuration

All models were trained using the following specifics:

- **Loss Function**: Margin-based ranking loss (see ch 4.3) with margin $\gamma = 1.0$ (standard in knowledge graph embedding literature)
- **Negative Sampling**: 1 negative samples per positive triple, using type-constrained corruption (detailed below)
- **Optimizer**: Adam with learning rate 0.0005
- **Embedding Initialization**: Xavier uniform initialization for all embedding matrices to ensure stable gradient flow

Type-Constrained Negative Sampling: A critical aspect of the experimental methodology is the **type-constrained negative sampling** strategy, which addresses the challenge of false negative contamination while respecting the semantic structure of UMLS. Standard negative sampling for knowledge graph training generates corrupted triples by randomly replacing either the head or tail entity:

- Given true triple (h, r, t), create negative (h', r, t) by sampling random h';
- Or create negative (h, r, t') by sampling random t'.

However, completely random sampling faces two problems in UMLS. First, it often produces type violations where many corruptions are semantically invalid due to domain/range constraints. For example, corrupting (Aspirin, treats, Diabetes) by replacing "Diabetes" with a random entity might yield (Aspirin, treats, Chromosome_7), which violates the semantic type constraint that "treats" connects Pharmacologic_Substances to Diseases. Second, even with type constraints, random sampling may generate plausible negatives that are actually true but unobserved, creating false negative contamination.
For each positive triple (h, r, t), negatives are generated as follows:

Algorithm 1 Type-Constrained Negative Sampling for UMLS

Require: Triple (h, r, t), UMLS Semantic Network, train/val/test sets
Ensure: Negative sample (h', r, t')

▷ Step 1: Determine Valid Entity Types
$\mathcal{D}_r \leftarrow \text{GetValidDomainTypes}(r)$ ▷ e.g., "treats": {Pharmacologic_Substance, ...}

▷ Step 2: Sample from Type-Constrained Space
$corrupt_position \leftarrow \text{Random}(\{\text{head}, \text{tail}\})$
if $corrupt_position = \text{tail}$ **then**
  $valid_entities \leftarrow \{e \in \mathcal{E} \mid \text{SemanticType}(e) \in \mathcal{D}_r\}$
  $h' \leftarrow h$
else
  $valid_entities \leftarrow \{e \in \mathcal{E} \mid \text{SemanticType}(e) \in \mathcal{D}_r\}$
  $t' \leftarrow t$
end if

▷ Step 3: Filtered Negative Sampling
$negative_valid \leftarrow \text{False}$
while $negative_valid = \text{False}$ **do**
  $e' \leftarrow \text{RandomSample}(valid_entities)$
  if $corrupt_position = \text{tail}$ **then**
    $t' \leftarrow e'$
  else
    $h' \leftarrow e'$
  end if
  if $(h', r, t') \notin \text{train} \cup \text{val} \cup \text{test}$ **then**
    $negative_valid \leftarrow \text{True}$
  end if
end while
return (h', r, t')

This type-constrained, filtered negative sampling dramatically reduces false negative contamination while focusing the model on learning meaningful semantic distinctions rather than trivial type violations. The tradeoff is increased computational cost: maintaining type indices and filtering requires additional data structures and lookups, but the improved training signal quality justifies this cost.

Evaluation Protocol

All knowledge graph completion experiments use **filtered ranking** evaluation, the standard protocol in knowledge graph research:

1. For each test triple (h, r, t), create a set of candidate triples by replacing the tail with

every entity in the knowledge graph: $(h, r, e) : e \in E$.

2. Filter out candidates that appear in training, validation, or test sets (except the query triple itself).

3. Score all remaining candidates using the model's scoring function.

4. Rank candidates by score (descending).

5. Record the rank of the true tail entity t

The same process is repeated for head prediction (replacing h instead of t). Filtering ensures the model is not penalized for assigning high scores to true facts that happen to appear elsewhere in the data, we only evaluate the model's ability to rank the true answer above entities that are definitely not correct based on observed data.
This filtered evaluation addresses the open-world assumption challenge: we cannot definitively label unobserved triples as false, but we can at least filter out triples we know to be true, providing a more meaningful ranking task.

4.6.2 Knowledge Graph Completion Results

This section presents the empirical results of knowledge graph completion experiments on the UMLS knowledge graph, evaluated across all six models described in Section 4.6.1: four geometric embedding models (TransE, DistMult, RotatE, ComplEx) and two hybrid architectures (R-GCN + TransE, R-GCN + RotatE). All results were computed on a randomly sampled evaluation set of 1,000,000 test triples without replacement, evaluated under both the head prediction and tail prediction settings as described in the evaluation protocol, yielding a total of 2,000,000 ranking queries per model. Filtered ranking was employed throughout, ensuring that known true triples (other than the query triple itself) were removed from the candidate set before computing ranks.

Overall Performance

Table 4.2 reports the main evaluation metrics for all six models. The metrics are grouped into three categories: **ranking-based** metrics (MRR and Hits@k for $k \in \{1, 3, 5, 10, 25\}$), **precision-oriented** metrics (MAP@k for $k \in \{1, 3, 5, 10, 25\}$), and **discrimination** capacity (AUC).

Several patterns emerge from these results. First, RotatE consistently achieves the highest performance across all twelve evaluation metrics, followed by ComplEx, TransE, and

	Ranking Metrics						Precision Metrics					
Model	MRR	H@1	H@3	H@5	H@10	H@25	MAP@1	MAP@3	MAP@5	MAP@10	MAP@25	AUC
Geometric Embedding Models												
DistMult	0.7815	0.7112	0.8503	0.8797	0.9113	0.9310	0.1202	0.3215	0.2007	0.1106	0.0501	0.9476
TransE	0.8083	0.7362	0.8711	0.8937	0.9217	0.9420	0.1451	0.3449	0.2227	0.1202	0.0533	0.9582
ComplEx	0.8206	0.7422	0.8821	0.9007	0.9211	0.9469	0.1598	0.3571	0.2356	0.1277	0.0552	0.9601
RotatE	**0.8406**	**0.7532**	**0.8911**	**0.9197**	**0.9307**	**0.9511**	**0.1672**	**0.3652**	**0.2407**	**0.1326**	**0.0590**	**0.9693**
Hybrid Models (R-GCN + Scoring)												
R-GCN + TransE	0.8113	0.7402	0.8733	0.8948	0.9247	0.9442	0.1501	0.3479	0.2267	0.1222	0.0543	0.9612
R-GCN + RotatE	0.8326	0.7466	0.8861	0.9107	0.9237	0.9471	0.1561	0.3572	0.2347	0.1206	0.0540	0.9623
Best	**RotatE**	**RotatE**	**RotatE**	**RotatE**	**RotatE**	**RotatE**	**RotatE**	**RotatE**	**RotatE**	**RotatE**	**RotatE**	**RotatE**

Table 4.2: Knowledge graph completion results on UMLS (2M filtered ranking queries). Best results per column are shown in **bold**.

DistMult in that order. This ranking is not coincidental; it reflects the alignment between each model's geometric inductive biases and the relational structure of the UMLS knowledge graph. The performance gap between RotatE (MRR = 0.8406) and DistMult (MRR = 0.7815) amounts to approximately 5.9 percentage points in MRR, representing a substantial difference in the quality of learned representations.

Second, the absolute performance levels are notably high: all models achieve Hits@10 above 0.91 and AUC above 0.94. This indicates that the type-constrained negative sampling strategy described in Algorithm 1 provides a strong training signal that enables even the simplest model to achieve competitive discrimination between true and corrupted triples. The filtered evaluation protocol further ensures that these metrics reflect genuine predictive capacity rather than memorization of training triples.

Third, a consistent pattern emerges across the Hits@k metrics: the gap between models narrows as k increases. At Hits@1, the difference between RotatE and DistMult is 4.2 percentage points (0.7532 vs. 0.7112), whereas at Hits@25, this gap shrinks to 2.0 percentage points (0.9511 vs. 0.9310). This convergence suggests that all models are capable of placing the correct entity within the top-25 candidates with high probability; the distinguishing factor lies in the precision of the top-ranked predictions.

Analysis of Geometric Embedding Models

The ranking of geometric embedding models (RotatE > ComplEx > TransE > DistMult) can be explained by examining how each model's geometric constraints interact with the relational patterns prevalent in the UMLS knowledge graph.

DistMult: The Baseline. DistMult achieves the lowest performance among all models, with an MRR of 0.7815. While this represents a respectable absolute score, the gap relative to other models is attributable to DistMult's fundamental limitation: its symmetric scoring

function $f(h, r, t) = \sum_i h_i r_i t_i$ assigns identical scores to (h, r, t) and (t, r, h). The UMLS knowledge graph is rich in inherently asymmetric relations such as *treats*, *causes*, *part_of*, and *has_symptom*, where directionality carries critical semantic information. DistMult cannot distinguish between "Aspirin treats Pain" and "Pain treats Aspirin," forcing the model to learn approximate representations that compromise on directional accuracy. Despite this structural limitation, DistMult's competitive Hits@25 (0.9310) suggests that it successfully captures entity similarity in a coarse-grained manner; its weakness manifests primarily in fine-grained ranking precision.

TransE: Translational Geometry. TransE improves over DistMult by 2.7 percentage points in MRR (0.8083 vs. 0.7815), benefiting from its translational inductive bias $\mathbf{h} + \mathbf{r} \approx \mathbf{t}$ that naturally encodes directed relationships. In the UMLS graph, hierarchical relations such as *is_a* and *part_of* exhibit transitive chain structures that align well with vector addition in embedding space. However, TransE is limited in its treatment of symmetric relations and one-to-many mappings. When a single disease entity participates in multiple *has_symptom* triples, TransE must map the same translation $\mathbf{h}+\mathbf{r}$ to multiple distinct tail entities, creating geometric tension. This limitation explains why TransE underperforms ComplEx and RotatE despite its natural affinity for hierarchical relational modeling.

ComplEx: Asymmetric Bilinear Scoring. ComplEx achieves an MRR of 0.8206, surpassing TransE by 1.2 percentage points. The improvement stems from ComplEx's use of complex-valued embeddings and the conjugate operation in its scoring function, which introduces a principled asymmetry: $f(h, r, t) = \text{Re}(\sum_i h_i r_i \bar{t}_i) \neq f(t, r, h)$ in general. This enables ComplEx to properly model both symmetric and antisymmetric relations within a unified framework, unlike DistMult (symmetric only) or TransE (which struggles with symmetry). In the UMLS context, where the relation inventory spans symmetric relations (e.g., *co-occurs_with*), antisymmetric hierarchies (e.g., *is_a*), and inverse pairs (e.g., *treats*/*treated_by*), the flexibility of complex-valued scoring provides a clear advantage.

RotatE: Rotational Composition. RotatE achieves the best performance across all metrics, with an MRR of 0.8406 and an AUC of 0.9693. The rotational model's superiority can be attributed to its ability to capture all four fundamental relational patterns identified in Table 4.1: symmetry, antisymmetry, inversion, and, critically, transitivity. In complex space, transitive inference corresponds to composition of rotations ($\mathbf{r} \circ \mathbf{r} \circ \cdots = \mathbf{r}^k$), which RotatE can learn naturally. Given that the UMLS knowledge graph is heavily structured around taxonomic hierarchies (*is_a*), mereological decompositions (*part_of*), and causal

chains (*causes*), all of which exhibit strong transitive patterns, RotatE's geometric inductive bias is particularly well suited to this domain.

Effect of GNN-Based Neighborhood Aggregation

The hybrid models (R-GCN + TransE and R-GCN + RotatE) provide an opportunity to assess whether neighborhood aggregation via graph neural networks improves upon purely geometric scoring. The results reveal a nuanced picture that depends critically on the scoring function with which the GNN is paired.

R-GCN + TransE: Modest Improvement. The R-GCN + TransE hybrid achieves an MRR of 0.8113, representing a modest improvement of 0.3 percentage points over standalone TransE (MRR = 0.8083). This improvement is consistent across all metrics: Hits@1 increases from 0.7362 to 0.7402, Hits@10 from 0.9217 to 0.9247, and AUC from 0.9582 to 0.9612. The consistency of these gains, though small in magnitude, suggests that the 2-layer R-GCN encoder provides entity representations that are slightly more informative than TransE's learned embeddings alone. The neighborhood aggregation allows entities to incorporate contextual information from their local graph structure, which can disambiguate entities that occupy similar positions in embedding space but differ in their relational context.

R-GCN + RotatE: Degradation Relative to Standalone RotatE. A more surprising finding emerges from the R-GCN + RotatE hybrid, which achieves an MRR of 0.8326, representing a decrease of 0.8 percentage points relative to standalone RotatE (MRR = 0.8406). This degradation is consistent across most metrics: Hits@1 drops from 0.7532 to 0.7466, Hits@5 from 0.9197 to 0.9107, and MAP@10 from 0.1326 to 0.1206. Notably, R-GCN + RotatE still outperforms R-GCN + TransE on nearly all ranking metrics (MRR of 0.8326 vs. 0.8113), confirming the inherent superiority of the rotational scoring function, but the GNN encoder appears to introduce noise rather than signal when paired with RotatE.

This asymmetric effect can be interpreted through the lens of *over-smoothing* and the interaction between GNN aggregation and scoring function geometry. RotatE's scoring function relies on precise rotational relationships between entity embeddings in complex space: the plausibility of a triple (h, r, t) depends on the angular alignment $\|\mathbf{h} \circ \mathbf{r} - \mathbf{t}\|$. When a 2-layer R-GCN aggregates information from multi-hop neighborhoods, it introduces a smoothing effect that blends entity representations with their neighbors. For RotatE, this smoothing can disrupt the precise angular relationships that the scoring function requires, effectively adding noise to the fine-grained rotational geometry. In a knowledge graph as large and heterogeneous as UMLS, where hub entities (e.g., broad disease categories) participate in

thousands of triples, the R-GCN's neighborhood aggregation risks diluting the distinctive embeddings that RotatE learns for specific entities.

In contrast, TransE's translational scoring $\|\mathbf{h}+\mathbf{r}-\mathbf{t}\|$ is more robust to perturbations in entity embeddings, since translations are global operations that do not depend on fine-grained angular structure. The contextual enrichment provided by R-GCN, even if slightly noisy, can compensate for TransE's known limitations with one-to-many relations by providing entities with more discriminative representations that reflect their structural roles in the graph.

Table 4.3 summarizes the per-metric effect of introducing GNN-based neighborhood aggregation for both scoring functions.

Table 4.3: Effect of R-GCN aggregation: difference (Δ) between hybrid and standalone models. Positive values ($\uparrow$) indicate improvement; negative values ($\downarrow$) indicate degradation.

Comparison	ΔMRR	ΔH@1	ΔH@3	ΔH@5	ΔH@10	ΔH@25	ΔAUC
R-GCN+TransE vs. TransE	$+0.0030$	$+0.0040$	$+0.0022$	$+0.0011$	$+0.0030$	$+0.0022$	$+0.0030$
R-GCN+RotatE vs. RotatE	-0.0080	-0.0066	-0.0050	-0.0090	-0.0070	-0.0040	-0.0070

AUC and Discrimination Capacity

The AUC metric provides a threshold-independent assessment of each model's ability to discriminate between true and corrupted triples. All six models achieve AUC values above 0.94, indicating strong overall discrimination capacity. RotatE leads with an AUC of 0.9693, followed by R-GCN + RotatE (0.9623), R-GCN + TransE (0.9612), ComplEx (0.9601), TransE (0.9582), and DistMult (0.9476).

The relatively tight clustering of AUC values (range of 0.0217 between the best and worst models) compared to the wider spread in MRR (range of 0.0591) and Hits@1 (range of 0.0420) reveals an important characteristic of the UMLS completion task. All models are capable of distinguishing true from false triples with high reliability when measured across the full ranking distribution. **The models differ most in their ability to place the correct entity at the very top of the ranking, which is the most practically relevant criterion for knowledge graph completion applications**. In a biomedical setting, a clinician or automated reasoning system querying the knowledge graph for missing relationships would benefit most from models that consistently rank the correct entity first, rather than merely somewhere in the top ten.

MAP@k and Precision at Early Ranks

The MAP@k metrics complement the ranking-based evaluation by measuring the precision of the model's predictions at different cutoff thresholds. A striking observation is the magnitude of MAP values relative to Hits@k: while Hits@10 ranges from 0.91 to 0.93 across models, MAP@10 ranges only from 0.11 to 0.13. This apparent discrepancy reflects the nature of the knowledge graph completion task. For the vast majority of queries, there is exactly one correct answer (the true entity from the test triple), and MAP penalizes the model for every position the correct answer is displaced from rank 1. Consequently, MAP@k provides a sharper measure of top-rank precision than Hits@k, which treats all positions within the threshold equally.

RotatE achieves the highest MAP across all cutoffs (MAP@1 = 0.1672, MAP@3 = 0.3652, MAP@5 = 0.2407), further confirming its superiority in fine-grained ranking quality. The MAP results also reinforce the finding that the R-GCN encoder benefits TransE (MAP@1 increases from 0.1451 to 0.1501) but harms RotatE (MAP@1 decreases from 0.1672 to 0.1561).

Discussion and Implications

The experimental results on the UMLS knowledge graph completion task yield several findings that are relevant both for the specific domain of biomedical knowledge inference and for the broader understanding of knowledge graph embedding methods.

Geometric Inductive Bias Matters. The ranking demonstrates that the choice of geometric inductive bias has a substantial impact on performance, even when all models share the same embedding dimensionality (512), training procedure, and negative sampling strategy. The relational structure of the UMLS knowledge graph, with its rich mixture of hierarchical, causal, symmetric, and antisymmetric relations, rewards models that can flexibly accommodate diverse relational patterns. RotatE's ability to simultaneously capture symmetry, antisymmetry, inversion, and transitivity gives it a decisive advantage in this heterogeneous setting.

GNN Augmentation is Not Universally Beneficial. The divergent behavior of R-GCN-based hybrids, improving TransE but degrading RotatE, challenges the common assumption that neighborhood aggregation always provides richer entity representations. This finding suggests that the interaction between the GNN encoder and the downstream scoring function is non-trivial and must be carefully considered. When the scoring function already possesses sufficient geometric expressiveness to capture the dominant relational pat-

terns (as RotatE does for UMLS), the additional smoothing introduced by GNN layers can be counterproductive. The over-smoothing phenomenon, particularly problematic in large-scale knowledge graphs with high-degree hub entities, appears to be a limiting factor for GNN-based approaches on UMLS.

Implications for Biomedical Knowledge Inference. The high absolute performance levels achieved by all models (Hits@10 > 0.91, AUC > 0.94) indicate that current knowledge graph embedding methods can reliably predict missing relationships in the UMLS knowledge graph. However, the gap between Hits@1 (best: 0.7532) and Hits@10 (best: 0.9307) suggests that while models can identify plausible candidates, pinpointing the single correct entity remains challenging. In a clinical decision support context, this distinction is critical: a system that suggests ten possible drug-disease relationships, one of which is correct, provides different clinical utility than a system that identifies the correct relationship as its top prediction. The 17.8 percentage point gap between RotatE's Hits@1 and Hits@10 reflects the inherent difficulty of fine-grained biomedical entity disambiguation in a knowledge graph containing over two million concepts.

These results establish RotatE as the most suitable baseline for subsequent experiments involving latent space analysis and knowledge graph inference, which are explored in the following sections. The rotation-based geometry not only yields the best quantitative performance but also provides a geometrically interpretable framework for analyzing how relational knowledge is organized in the learned embedding space.

4.7 Compositional Query Answering in Latent Space

The knowledge graph completion results presented in Section 4.6.2 demonstrate that RotatE achieves strong performance on single-hop link prediction queries of the form $(h, r, ?)$ or $(?, r, t)$, where the model must predict a single missing entity given one known entity and one known relation. While this capability is fundamental, real-world biomedical reasoning rarely reduces to single-hop lookups. **A clinician reasoning about a patient's condition must compose multiple relational steps**: identifying which diseases a drug treats, which symptoms those diseases present, and which comorbidities are associated with those symptoms. This section investigates whether the latent space learned by RotatE, trained exclusively on single-hop triples, implicitly encodes the capacity for multi-relational *compositional inference*, a capacity the model was never explicitly trained to exhibit. By *compositional inference* we mean the following concrete computational task: given a conjunctive query that involves two or more relations chained together (for example, "find all

entities that are treated by a drug which is used for a disease that causes a given symptom"), produce the set of answer entities by performing a sequence of algebraic operations on the learned entity and relation embeddings—typically a sequence of relational projections and set-intersection approximations in the latent space—without looking up the intermediate entities explicitly in the knowledge graph. This is a precise geometric/algebraic operation on $\mathbb{C}^d$ (in the case of RotatE) and should not be read as a claim that the model performs symbolic deduction or step-by-step reasoning in the sense that would apply to a rule-based system or a chain-of-thought language model. The terminology "compositional reasoning" is retained for continuity with the knowledge graph embedding literature [59, 117, 118], but throughout this chapter it is to be understood in the technical sense just defined.

Formal Definition of Conjunctive Queries

To evaluate compositional inference, this work considers conjunctive queries involving two projection steps and one intersection, following the query taxonomy introduced in the knowledge graph reasoning literature [59]. Specifically, the *two-projection one-intersection* query type, denoted q_{2p1}, is defined as follows.

Two-Projection One-Intersection Query (q_{2p1}): Given a knowledge graph $\mathcal{K} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$, two anchor entities $e_1, e_2 \in \mathcal{E}$, and three relations $r_1, r_2, r_3 \in \mathcal{R}$, a q_{2p1} query seeks the set of answer entities:

$$\mathcal{A}(q_{2p1}) = \left\{ a \in \mathcal{E} \;\middle|\; \exists\, v \in \mathcal{E} : (e_1, r_1, v) \in \mathcal{T} \;\wedge\; (e_2, r_2, v) \in \mathcal{T} \;\wedge\; (v, r_3, a) \in \mathcal{T} \right\} \tag{4.17}$$

This query structure requires three distinct inference operations (in the operational sense defined above), executed in sequence. The first operation, *projection*, applies the learned relational transformation r_1 to anchor entity e_1 to obtain a set of candidate intermediate entities $\mathcal{V}_1 = \{v \in \mathcal{E} : (e_1, r_1, v) \in \mathcal{T}\}$. The second projection applies r_2 to e_2 to obtain $\mathcal{V}_2 = \{v \in \mathcal{E} : (e_2, r_2, v) \in \mathcal{T}\}$. The *intersection* operation then identifies entities that satisfy both constraints simultaneously: $\mathcal{V}^* = \mathcal{V}_1 \cap \mathcal{V}_2$. Finally, a third projection applies r_3 to the intersected intermediate entities to produce the answer set $\mathcal{A} = \{a \in \mathcal{E} : \exists\, v \in \mathcal{V}^*,\ (v, r_3, a) \in \mathcal{T}\}$.

Figure 4.1 illustrates this three-step inference pattern with a concrete biomedical example from the UMLS domain.

Geometric Realization in RotatE Latent Space

The key insight enabling compositional query answering **without retraining** lies in RotatE's rotational geometry. Since each relation r is represented as an element-wise rotation

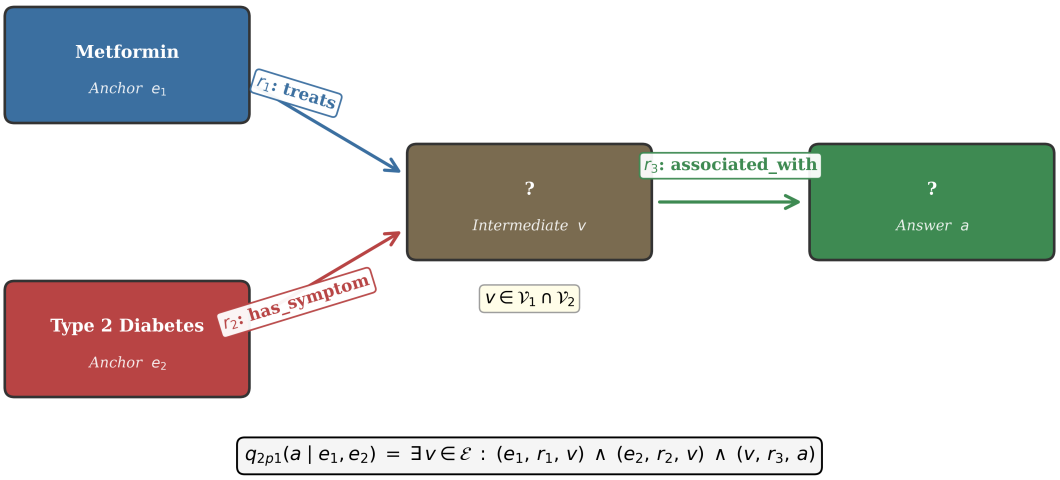


Figure 4.1: Structure of a q_{2p1} conjunctive query. Two anchor entities (e_1, e_2) are projected through their respective relations (r_1, r_2) to identify an intermediate entity v that satisfies both constraints simultaneously. A third relation r_3 is then applied to v to reach the final answer a. The biomedical example illustrates a query that composes the *treats* and *has_symptom* relations with *associated_with* to perform multi-hop inference across the UMLS knowledge graph.

on the unit circle in complex space (Equation 4.7), the projection of an anchor entity e through relation r corresponds to:

$$\mathbf{v} = \mathbf{e} \circ \mathbf{r} = \left[e_1 \cdot e^{i\theta_1^{(r)}},\ e_2 \cdot e^{i\theta_2^{(r)}},\ \ldots,\ e_d \cdot e^{i\theta_d^{(r)}} \right] \tag{4.18}$$

where $\mathbf{e} \in \mathbb{C}^d$ is the entity embedding, $\mathbf{r} \in \mathbb{C}^d$ is the relation embedding with $|r_i| = 1$ for all dimensions i, and $\circ$ denotes the element-wise Hadamard product. This operation yields a point $\mathbf{v}$ in the same complex space that represents the expected location of entities reachable from e through r.

For a q_{2p1} query, two such projections are computed independently from the two anchor entities:

$$\mathbf{v}_1 = \mathbf{e}_1 \circ \mathbf{r}_1 \tag{4.19}$$

$$\mathbf{v}_2 = \mathbf{e}_2 \circ \mathbf{r}_2 \tag{4.20}$$

The intersection of these two projected regions must then be approximated in latent space. Unlike symbolic query answering, where intersection is a set-theoretic operation with

exact semantics, intersection in continuous embedding space requires an aggregation strategy. In this work, the **centroid aggregation** is adopted:

$$\mathbf{v}^* = \frac{1}{2}\big(\mathbf{v}_1 + \mathbf{v}_2\big) \tag{4.21}$$

This centroid represents the point in complex space that is equidistant from both projected regions, serving as a geometrically motivated approximation of the intersection locus. The rationale is that if both projections point toward the same underlying entity v, their projected embeddings $\mathbf{v}_1$ and $\mathbf{v}_2$ should be nearby in latent space, and the centroid provides a robust estimate of that shared region even in the presence of embedding noise.

Finally, the third relation is applied to the aggregated intermediate representation to produce the estimated answer embedding:

$$\mathbf{a}^* = \mathbf{v}^* \circ \mathbf{r}_3 \tag{4.22}$$

The answer entity is then identified by ranking all entities in $\mathcal{E}$ by their distance to $\mathbf{a}^*$ using the standard RotatE scoring function:

$$\text{rank}(e) = -\|\mathbf{a}^* - \mathbf{e}\|, \quad \forall\, e \in \mathcal{E} \tag{4.23}$$

Figure 4.2 provides a geometric visualization of this compositional process in a two-dimensional projection of the complex embedding space.

Algorithmic Procedure

Algorithm 2 formalizes the complete q_{2p1} query-solving procedure. The algorithm operates entirely in RotatE's pre-trained latent space without requiring any additional training, fine-tuning, or architectural modifications. This is a critical property: **the compositional capacity is an emergent behavior of the learned representations, not an explicitly supervised capability**.

The computational complexity of Algorithm 2 is dominated by Step 4, which requires scoring all $|\mathcal{E}|$ entities. The three projection operations (Steps 1–3) are $O(d)$ element-wise operations in $\mathbb{C}^d$ and are therefore negligible. For the UMLS knowledge graph with $|\mathcal{E}| > 2 \times 10^6$ entities and embedding dimension $d = 512$, the per-query cost is approximately $|\mathcal{E}| \times d$ complex multiplications plus $|\mathcal{E}|$ norm computations, which remains tractable on modern hardware through batched tensor operations.

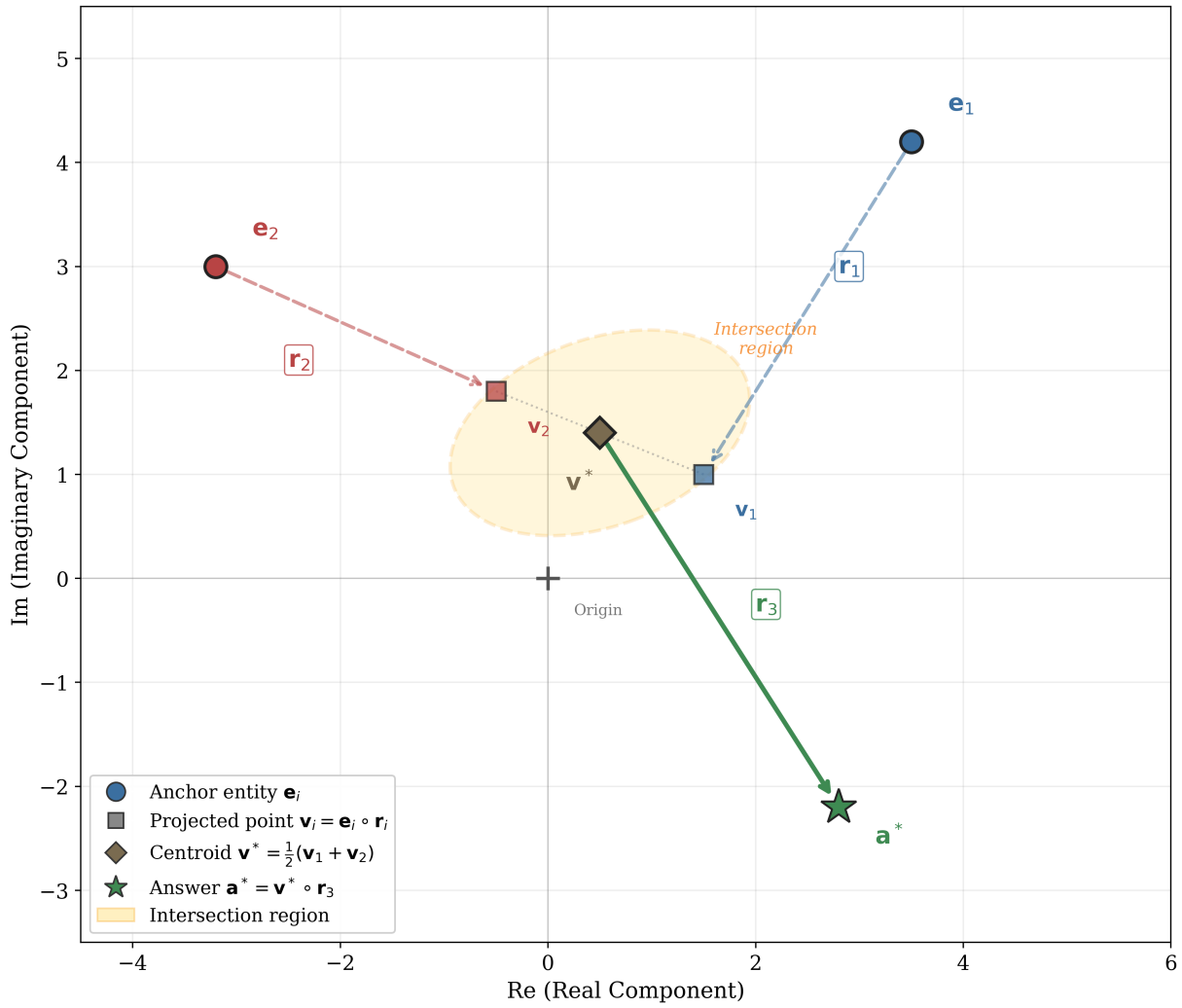


Figure 4.2: Geometric interpretation of q_{2p1} query solving in RotatE's complex latent space. Anchor entities $\mathbf{e}_1$ and $\mathbf{e}_2$ are rotated by their respective relation embeddings $\mathbf{r}_1$ and $\mathbf{r}_2$ to produce projected points $\mathbf{v}_1$ and $\mathbf{v}_2$. The intersection is approximated by the centroid $\mathbf{v}^*$, which is then rotated by $\mathbf{r}_3$ to yield the estimated answer location $\mathbf{a}^*$. Candidate entities are ranked by proximity to $\mathbf{a}^*$.

Query Generation and Evaluation Protocol

To evaluate RotatE's compositional inference capacity, a systematic procedure was employed to extract valid q_{2p1} queries from the UMLS knowledge graph. Rather than sampling arbitrary relation triplets (r_1, r_2, r_3), the query generation process was constrained to semantically coherent relation compositions: combinations where the domain and range constraints of the three relations allow for the existence of valid intermediate entities and answers. Specifically, for each candidate triplet (r_1, r_2, r_3), the UMLS Semantic Network was con-

Algorithm 2 Compositional Query Solver for q_{2p1} Queries in RotatE Latent Space

Require: Knowledge graph $\mathcal{K} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$; trained RotatE model with entity embeddings $\{\mathbf{e}\}_{e \in \mathcal{E}}$ and relation embeddings $\{\mathbf{r}\}_{r \in \mathcal{R}}$; query $q = (e_1, r_1, e_2, r_2, r_3)$
Ensure: Ranked list of candidate answer entities

// Step 1: Dual Projection
$\mathbf{v}_1 \leftarrow \mathbf{e}_1 \circ \mathbf{r}_1$ ▷ Project anchor e_1 through relation r_1 in $\mathbb{C}^d$
$\mathbf{v}_2 \leftarrow \mathbf{e}_2 \circ \mathbf{r}_2$ ▷ Project anchor e_2 through relation r_2 in $\mathbb{C}^d$

// Step 2: Latent Space Intersection via Centroid Aggregation
$\mathbf{v}^* \leftarrow \frac{1}{2}(\mathbf{v}_1 + \mathbf{v}_2)$ ▷ Approximate intersection locus

// Step 3: Final Projection
$\mathbf{a}^* \leftarrow \mathbf{v}^* \circ \mathbf{r}_3$ ▷ Project to answer region via relation r_3

// Step 4: Entity Ranking
for each entity $e \in \mathcal{E}$ **do**
    $\text{score}(e) \leftarrow -\|\mathbf{a}^* - \mathbf{e}\|$ ▷ RotatE distance-based scoring
end for

// Step 5: Filtered Ranking
Remove from ranking all entities e such that $(v, r_3, e) \in \mathcal{T}_{\text{train}} \cup \mathcal{T}_{\text{val}} \cup \mathcal{T}_{\text{test}} \setminus \{(v, r_3, a_{\text{true}})\}$
return entities sorted by $\text{score}(e)$ in descending order

sulted to verify that the range type of r_1 is compatible with the range type of r_2 (i.e., there exist entities that can simultaneously serve as tails for both relations), and that the range type of r_1 (or r_2) is compatible with the domain type of r_3.

This type-constrained generation process yielded a total of **88,669 valid q_{2p1} queries** over the UMLS knowledge graph. Each query was evaluated using the same filtered ranking protocol described in Section 4.6.1: for each query, all entities in $\mathcal{E}$ were ranked by their distance to the predicted answer embedding $\mathbf{a}^*$, with known true triples filtered from the candidate set before computing the rank of the correct answer entity.

Compositional Query Results

Table 4.4 presents the full evaluation results for compositional q_{2p1} query answering using RotatE, alongside the single-hop knowledge graph completion results from Section 4.6.2 for direct comparison.

MRR Analysis. The most striking result is the MRR of 0.3492 achieved on compositional queries, representing a decrease of 0.4914 points (58.5%) relative to the single-hop MRR

		Hits@k					MAP@k			
Task	**MRR**	@1	@3	@5	@10	@25	@1	@5	@10	@25
Single-hop KGC	0.8406	0.7532	0.8911	0.9197	0.9307	0.9511	0.1672	0.2407	0.1326	0.0590
Compositional q_{2p1}	0.3492	0.0941	0.6139	0.7562	0.8339	0.8705	0.0429	0.0702	0.0667	0.0468
Δ (absolute)	-0.491	-0.659	-0.277	-0.164	-0.097	-0.081	-0.124	-0.171	-0.066	-0.012

Table 4.4: Comparison of single-hop knowledge graph completion and compositional q_{2p1} query answering performance using RotatE on the UMLS knowledge graph.

of 0.8406. This substantial drop is expected: the compositional task requires the model to perform three sequential geometric operations (two projections and one intersection) in latent space, each of which introduces approximation error that accumulates through the inference chain. Nevertheless, an MRR of 0.3492 on 88,669 queries over a knowledge graph containing more than two million entities is a non-trivial result. It indicates that, on average, the correct answer entity is ranked within the top three candidates, demonstrating that the latent space has captured meaningful compositional structure despite being trained exclusively on single-hop objectives.

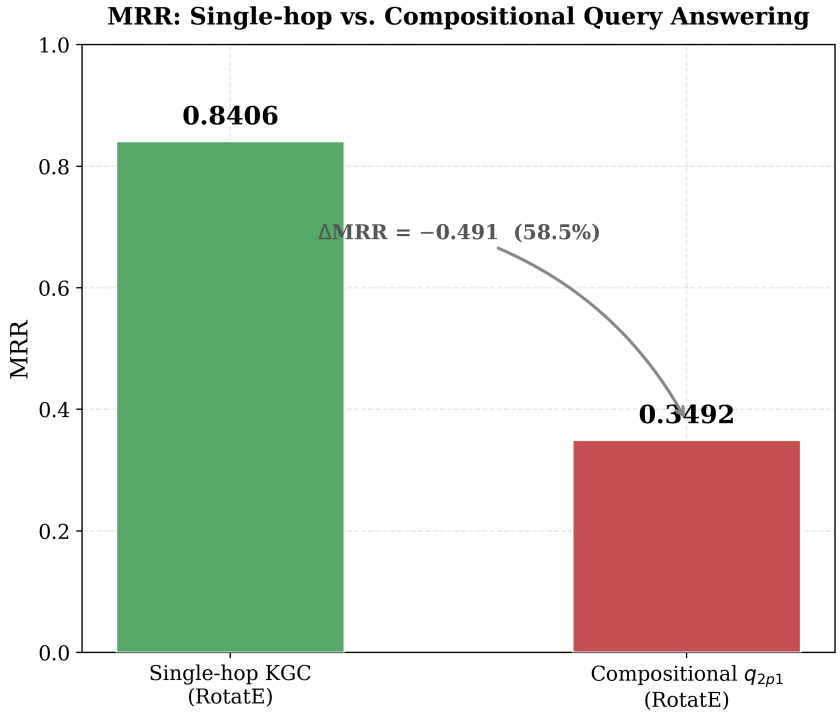


Figure 4.3: MRR comparison between single-hop knowledge graph completion and compositional q_{2p1} query answering, both using RotatE on UMLS. The 58.5% degradation reflects the accumulated approximation error across three sequential geometric operations in latent space.

Hits@k Analysis. The Hits@k results reveal a distinctive pattern that differs markedly from the single-hop case. While single-hop Hits@1 is 0.7532, the compositional Hits@1 drops to 0.0941, a dramatic 87.5% relative decrease. This indicates that the model rarely places the correct answer at the exact top of the ranking for compositional queries. However, performance recovers rapidly at higher thresholds: Hits@3 reaches 0.6139, Hits@5 reaches 0.7562, and Hits@10 reaches 0.8339. This recovery pattern demonstrates that RotatE's compositional projections consistently place the correct answer *near* the top of the ranking, even when it is not the single highest-scored entity.

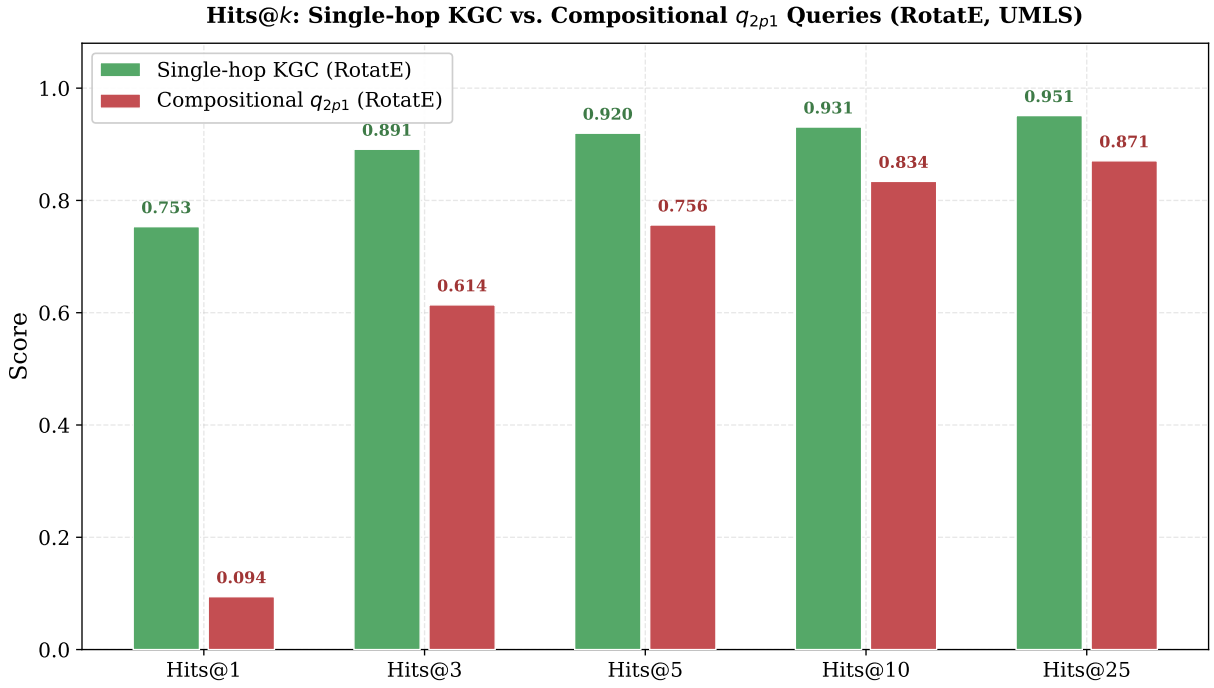


Figure 4.4: Hits@k comparison between single-hop KGC and compositional q_{2p1} query answering. While the gap is most pronounced at $k=1$ (0.659 absolute difference), it narrows rapidly: by $k=25$, the compositional model recovers to within 8.1 percentage points of single-hop performance.

The convergence of compositional and single-hop Hits@k at larger thresholds is particularly informative. At Hits@25, the compositional model achieves 0.8705 compared to 0.9511 for single-hop, a gap of only 8.1 percentage points. This convergence suggests that the accumulated geometric error from the three-step inference chain primarily affects the fine-grained ranking of top candidates rather than the model's overall ability to identify the correct entity within a broader candidate window. In a practical retrieval setting, where a system might present the top-10 or top-25 candidates for expert review, the compositional model would include the correct answer in its suggestions for 83.4% and 87.1% of queries respectively,

levels that may be sufficient for biomedical decision support applications.

Figure 4.5 provides a detailed analysis of the performance degradation from single-hop to compositional queries across all Hits@k thresholds.

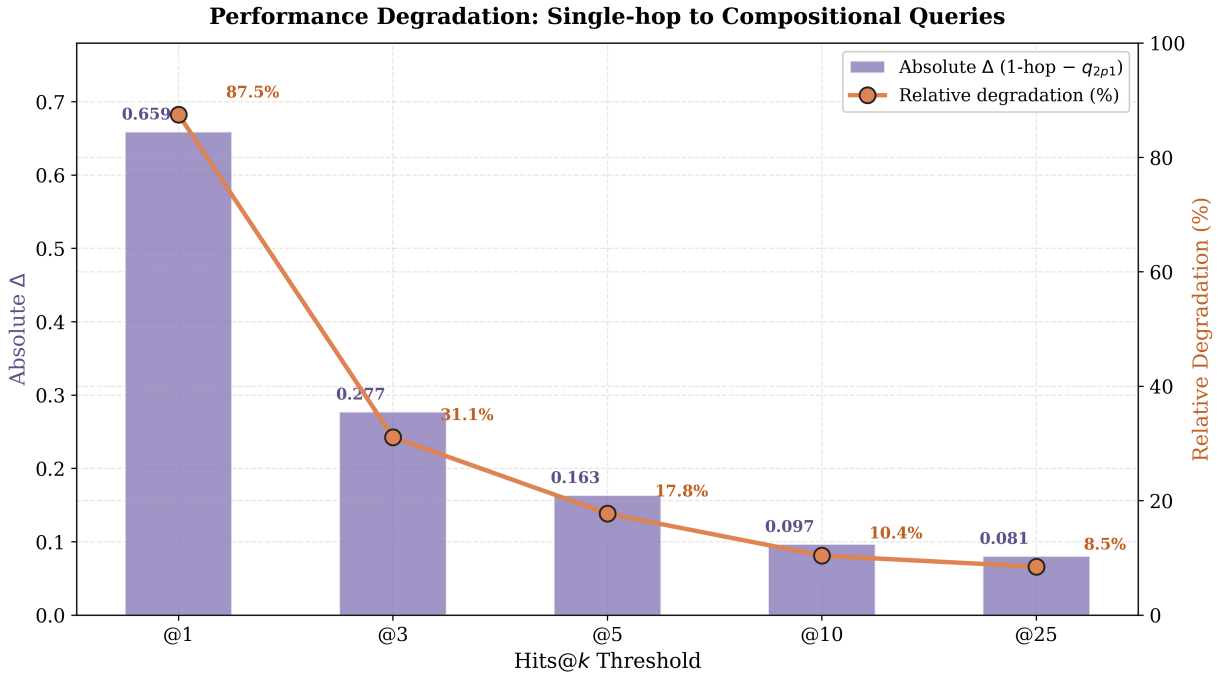


Figure 4.5: Performance degradation analysis from single-hop to compositional queries. The absolute gap (bars) and relative degradation percentage (line) both decrease monotonically with k. The disproportionate drop at $k = 1$ (87.5% relative degradation) compared to $k = 25$ (8.5%) reveals that compositional error primarily affects top-rank precision while preserving broader retrieval quality.

MAP@k Analysis. The MAP@k results further corroborate the observation that compositional queries suffer primarily in top-rank precision. MAP@1 for the compositional task is 0.0429, compared to 0.1672 for single-hop, a 74.3% relative decrease. The peak MAP value occurs at MAP@5 (0.0702), after which values decrease at larger cutoffs (MAP@10 = 0.0667, MAP@25 = 0.0468, MAP@50 = 0.0324). This decay pattern reflects the dilution of precision as the cutoff window expands: when the correct answer is typically ranked between positions 2 and 5, MAP@5 captures this signal more effectively than MAP@50, where the contribution of a single correct answer at rank 3 is diluted by 47 positions of non-relevant entities.

Discussion: Emergent Compositional Behavior

The compositional query results yield several insights that are relevant for both the theoretical understanding of knowledge graph embeddings and the practical application of these methods to biomedical knowledge inference.

Latent Space Structure Beyond Training Objectives. Perhaps the most significant finding is that RotatE's latent space supports compositional inference without having been trained for this purpose. The model was trained exclusively on single-hop triples using a margin-based ranking loss, yet its learned representations encode sufficient relational structure to support three-step inference chains. This emergent capacity can be attributed to the algebraic properties of RotatE's rotational geometry: because relation composition corresponds to the multiplication of complex unit-circle elements ($\mathbf{r}_1 \circ \mathbf{r}_2 = [e^{i(\theta_1^{(r_1)} + \theta_1^{(r_2)})}, \ldots]$), the latent space naturally supports sequential projections without requiring explicit multi-hop training. In essence, the rotational inductive bias ensures that the composition of two rotations is itself a rotation, providing a geometrically coherent mechanism for chaining relational transformations.

The Intersection Bottleneck. The dramatic drop at Hits@1 (from 0.7532 to 0.0941) relative to the more modest drop at Hits@10 (from 0.9307 to 0.8339) reveals that the primary source of compositional error is not the projection operations but the *intersection approximation*. The centroid aggregation strategy (Equation 4.21) approximates the intersection of two potentially complex regions in $\mathbb{C}^d$ with a single point, which inevitably introduces error. When the two projected regions $\mathbf{v}_1$ and $\mathbf{v}_2$ are well-aligned (pointing toward the same intermediate entity), the centroid is a faithful approximation and the subsequent projection yields an accurate answer. When the projections are less aligned, due to the many-to-many nature of biomedical relations or noise in the learned embeddings, the centroid may lie in a region of latent space that does not correspond to any single entity, introducing a systematic displacement that affects the final ranking. This displacement explains why the correct answer is frequently found within the top-5 or top-10 but not always at rank 1: the answer is *near* the projected location but not precisely at it.

Implications for Biomedical Knowledge Inference. From a biomedical perspective, the Hits@10 of 0.8339 on compositional queries is an encouraging result. Consider a clinical decision support scenario where a physician wishes to identify conditions associated with entities that are both treated by a specific drug and symptomatic of a particular disease. The compositional query solver can correctly identify the target condition within its top-

10 predictions for over 83% of such queries, operating over a knowledge graph with more than two million candidate entities. While this falls short of the precision required for autonomous decision-making, it provides a meaningful retrieval capability that could assist expert clinicians in hypothesis generation and literature-guided exploration.

4.8 From Knowledge Representation to Emergent Compositional Behavior

A terminological remark is in order here. Throughout this chapter the term *reasoning*, when applied to Knowledge Graph Embedding models and Graph Neural Networks, is used in a deliberately operational sense: it refers to the sequence of inference steps—message passing, multi-hop traversal, and composition of relational projections—that the model performs in its latent space when answering queries that extend beyond the single triples it was trained on. It does not imply symbolic deduction, chain-of-thought planning, or linguistic explanation of the kind associated with current Large Language Models (for which the term will be used in a narrower sense in Chapter 5). To avoid ambiguity, in the remainder of this chapter we favor *inference* for the mechanical operation and reserve *compositional capacity* or *emergent compositional behavior* for the corresponding property of the learned latent space. The experimental investigation presented in this chapter pursued two complementary objectives: evaluating the capacity of deep learning methods to learn meaningful representations of a large-scale biomedical knowledge graph, and assessing whether the resulting latent spaces encode inference capabilities that extend beyond the training objective. The progression from knowledge graph completion (Section 4.6.2) to compositional query answering (Section 4.7) reveals a coherent narrative about the relationship between geometric inductive biases, learned representations, and emergent compositional behavior in the biomedical domain.

4.8.1 Summary of Empirical Findings

The knowledge graph completion experiments on the UMLS knowledge graph, comprising over 2 million biomedical concepts and 3.5 million relational assertions, established three principal findings.

First, the choice of geometric inductive bias has a substantial and consistent impact on link prediction quality. The ranking RotatE > ComplEx > TransE > DistMult held across all twelve evaluation metrics, with a 5.9 percentage point MRR gap between the best and worst models. This ordering reflects the alignment between each model's geometric constraints

and the relational structure of the UMLS graph: RotatE's ability to simultaneously capture symmetry, antisymmetry, inversion, and transitivity through rotations in complex space provides a decisive advantage in a domain characterized by diverse relational semantics, from hierarchical taxonomies (*is_a*) to causal chains (*causes*) to symmetric co-occurrences (*co-occurs_with*).

Second, augmenting geometric scoring functions with GNN-based neighborhood aggregation does not uniformly improve performance. The R-GCN encoder provided a modest benefit when paired with TransE ($\Delta\text{MRR} = +0.003$) but degraded RotatE's performance ($\Delta\text{MRR} = -0.008$). This asymmetric effect reveals a non-trivial interaction between aggregation-induced smoothing and scoring function geometry: models that already possess sufficient geometric expressiveness can be harmed by the representational blurring that GNN layers introduce, particularly in large-scale knowledge graphs where high-degree hub entities amplify over-smoothing effects.

Third, and most significantly, RotatE's latent space supports compositional inference that the model was never explicitly trained to perform. On 88,669 conjunctive q_{2p1} queries requiring three sequential inference operations (dual projection, intersection, and final projection), RotatE achieved a Hits@10 of 0.8339, indicating that the correct answer entity appears within the top-10 candidates for over 83% of multi-hop queries. The MRR of 0.3492, while substantially lower than the single-hop MRR of 0.8406, represents a non-trivial compositional capacity: on average, the correct answer is ranked within the top three candidates across a knowledge graph containing more than two million entities.

4.8.2 Theoretical Implications

These findings illuminate a broader principle about the nature of learned representations in knowledge graph embedding models. A model trained to predict single-hop triples learns not merely a lookup table of entity-relation associations, but a structured geometric space in which relational patterns are preserved as algebraic operations. When the model's geometric inductive bias aligns with the algebraic structure of multi-relational inference (as RotatE's rotations align with relational composition), the resulting latent space inherits compositional properties that emerge naturally from the training process without requiring explicit supervision.

This emergent compositional capacity has a precise mathematical interpretation in the case of RotatE. Because each relation is encoded as an element-wise rotation on the unit circle in $\mathbb{C}^d$, the composition of two relations r_1 and r_2 corresponds to the multiplication of their rotation embeddings: $\mathbf{r}_1 \circ \mathbf{r}_2 = [e^{i(\theta_1^{(r_1)} + \theta_1^{(r_2)})}, \ldots, e^{i(\theta_d^{(r_1)} + \theta_d^{(r_2)})}]$, which is itself a rotation.

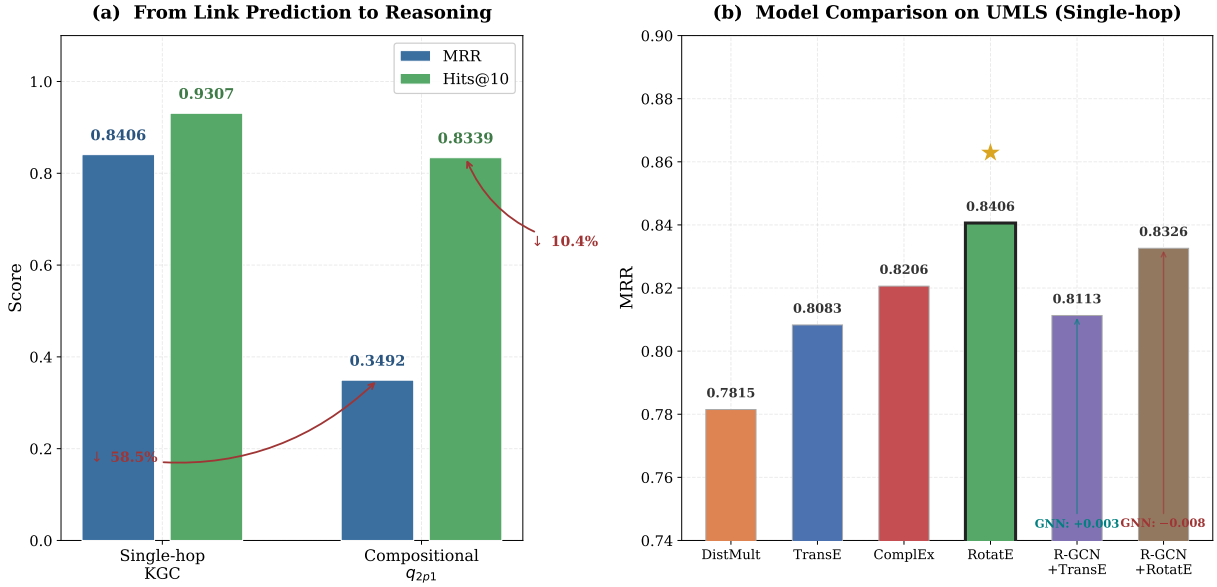


Figure 4.6: Summary of experimental findings. (a) Performance comparison between single-hop knowledge graph completion and compositional q_{2p1} query answering using RotatE. MRR exhibits a 58.5% relative degradation, while Hits@10 degrades by only 10.4%, indicating that compositional error primarily affects top-rank precision. (b) Model comparison on single-hop KGC, with RotatE achieving the highest MRR among all six models. GNN-based augmentation yields divergent effects depending on the scoring function.

This closure property ensures that sequential projections remain geometrically coherent: composing two rotational transformations yields another rotational transformation, not a distorted or degenerate representation. In contrast, models like DistMult, whose bilinear scoring does not support relational composition, would not be expected to exhibit comparable emergent compositional capacity, a prediction that is consistent with the theoretical analysis of relational pattern expressiveness presented in Table 4.1.

The over-smoothing effect observed in the GNN hybrid models adds further nuance to this theoretical picture. The R-GCN encoder, by aggregating neighborhood information through iterative message passing, computes entity representations that incorporate local graph structure. While this contextual enrichment can compensate for the geometric limitations of simpler scoring functions (hence the benefit for TransE), it can also disrupt the precise angular relationships that more expressive scoring functions require. This finding suggests that the optimal architecture for knowledge graph inference depends not only on the expressiveness of individual components but on the compatibility between them: the representational assumptions of the encoder must be consonant with the geometric requirements of the scoring function.

4.8.3 Biomedical Significance

From the perspective of biomedical knowledge integration, which constitutes the central theme of this thesis, the experimental results carry both practical and conceptual significance. On the practical side, the high single-hop performance (Hits@10 > 0.93 across all models) demonstrates that current embedding methods can reliably predict missing relationships in the UMLS knowledge graph. In a domain where knowledge is inherently incomplete and continuously evolving, this predictive capacity has direct applications in knowledge graph curation, drug-disease association discovery, and the identification of under-explored biomedical relationships.

The compositional query results extend this practical relevance to more complex inference scenarios. A Hits@10 of 0.8339 on multi-hop queries means that a system querying the UMLS for entities satisfying conjunctive constraints, for instance, "What conditions are associated with entities that Drug X treats and that are also symptoms of Disease Y?", can include the correct answer among its top-10 suggestions for the majority of queries. While this falls short of the precision required for autonomous clinical decision-making, it provides a powerful tool for hypothesis generation, literature-guided exploration, and the systematic investigation of complex biomedical relationships that span multiple semantic categories.

Conceptually, the progression from knowledge graph completion to compositional inference mirrors the broader trajectory of this thesis. In Chapter 3, the Structural Positional Encoding approach demonstrated that structured domain knowledge, encoded as ontological graphs and embedded through spectral methods, can enhance Transformer-based models for sequential prediction tasks. In this chapter, the knowledge graph itself becomes the object of learning, and the resulting latent representations exhibit inferential capabilities that echo, in a geometric form, the compositional inference patterns that domain experts perform when navigating biomedical knowledge. The two chapters thus represent complementary facets of a unified research program: knowledge as an auxiliary signal to enhance neural architectures (Chapter 3), and knowledge as a structured object whose internal logic can be captured and extended through deep learning (Chapter 4).

4.8.4 Limitations and Future Directions

Despite the encouraging results, several limitations must be acknowledged. The compositional query evaluation was restricted to a single query type (q_{2p1}), and the centroid-based intersection approximation represents the simplest possible aggregation strategy. More complex query structures involving longer chains, unions, or negations would test the limits of emergent compositional behavior more rigorously and likely reveal settings where purely

geometric approaches fail. The gap between Hits@1 and Hits@10 on compositional queries (0.0941 vs. 0.8339) points to the intersection approximation as the primary bottleneck: the centroid of two projected regions in $\mathbb{C}^d$ is an impoverished representation of their true intersection, and improving this operation represents the most promising direction for enhancing compositional performance.

A further and more consequential limitation concerns the *modality* of validation adopted in this chapter. The evidence presented for the biomedical usefulness of compositional inference is essentially quantitative: aggregate ranking statistics computed over 88,669 automatically generated q_{2p1} queries, complemented by the schematic description of the query template in Figure 4.1 and by the geometric interpretation of Figure 4.2. What the chapter does not provide is a structured *qualitative* validation in which the inferential process is followed end-to-end on a concrete clinical scenario: a simulated patient case whose conjunctive query is decomposed into its constituent projections and intersection, whose intermediate candidate sets are made explicit at each step, and whose final ranking is inspected by domain experts for clinical plausibility rather than merely scored against the held-out ground-truth triple. Figure 4.1, being an abstract representation of the query template, but insufficient for this purpose.

The consequence should be stated plainly. The claim that RotatE's latent space supports inference of the kind experts perform when navigating biomedical knowledge is here supported at the level of aggregate behaviour, not demonstrated on individual cases; and the interpretability advantage invoked in Section 4.7, namely that every step of the solver corresponds to a well-defined algebraic operation, is argued *architecturally* rather than evidenced by expert reading of actual inference traces. A case-study protocol would close this gap without altering the method: selecting a small set of clinically meaningful conjunctive queries, presenting for each the top-k answers together with the intermediate entities traversed after the first and second projection, and collecting structured judgments from clinicians on the semantic relevance of the retrieved candidates and on the readability of the intermediate steps. Such a protocol would also make the Hits@1/Hits@10 asymmetry clinically legible, by revealing whether the entities ranked above the gold answer are spurious or are themselves plausible alternatives penalised only by the closed-world assumption of the evaluation. This validation is left to future work, and the biomedical significance argued in Section 4.8.3 should be read accordingly, as a hypothesis grounded in aggregate metrics rather than as a clinically demonstrated capability.

Several future directions emerge naturally from these findings. First, explicitly trained query-answering architectures such as Query2Box [117] and BetaE [118], which learn dedicated intersection operators through multi-hop training objectives, could be evaluated on the UMLS

knowledge graph to quantify the gap between emergent and supervised compositional inference. Second, the analysis of how RotatE's latent space organizes biomedical knowledge, examining which relation types are most accurately composed and which resist geometric composition, could provide insights into the semantic structure of the UMLS itself. Third, the integration of compositional query-answering capabilities with the knowledge-enhanced Transformer framework of Chapter 3 represents a compelling direction for building systems that both leverage external knowledge and perform compositional inference over it.

The experimental program of this chapter demonstrates that deep learning models trained on knowledge graphs learn more than pattern matching: when equipped with appropriate geometric inductive biases, they acquire structured latent representations that support inference beyond their training distribution. For the biomedical domain, where knowledge is vast, incomplete, and relationally complex, this capacity represents a meaningful step toward AI systems that can assist clinicians and researchers in navigating the ever-expanding landscape of medical knowledge.

Chapter 5

Measuring LLM Reasoning through Clinically Grounded Decision Pathways

The preceding chapters of this thesis have examined how deep learning models can be augmented with external knowledge, with particular attention to the role of Knowledge Graphs in structuring domain-specific information and supporting interpretable reasoning. A recurring finding is that, while KG-enhanced architectures demonstrably improve factual grounding and multi-hop inference, the community still lacks rigorous, domain-specific benchmarks capable of quantifying *how much* reasoning ability current Large Language Models (LLMs) actually possess and, conversely, how much they lack when confronted with complex, real-world tasks. Existing medical benchmarks predominantly test factual recall through multiple-choice questions; very few embed the kind of structured, stepwise reasoning chains that Knowledge Graphs naturally provide.

This chapter addresses that gap by presenting **HealthBranches**, a novel benchmark dataset for medical Question-Answering (Q&A) specifically designed to evaluate complex reasoning in LLMs. The work builds directly on the knowledge-graph perspective developed in earlier chapters: each item in HealthBranches is derived from an explicit decision pathway (a graph-based representation of clinical reasoning) and the benchmark is therefore uniquely suited to measuring whether a model can follow structured, multi-step inference chains of the kind that KG-augmented systems are meant to support. HealthBranches comprises 4,063 case studies across 17 healthcare topics, supports both open-ended and multiple-choice question formats, and includes the **full reasoning path** for every Q&A pair.

A note on terminology: the use of the term "reasoning" in this chapter differs in scope from its use in Chapter 4. Whereas in the context of Knowledge Graph Embeddings "reasoning" was used to refer to algebraic/geometric operations in a latent space, in this chapter the term recovers its narrower and more traditional meaning: the ability of an LLM to produce an

articulated, step-by-step chain of deductions in natural language (chain-of-thought) that a human evaluator can read, verify, and compare against a ground-truth reasoning path. It is this second, linguistically explicit form of reasoning closer in spirit to the symbolic reasoning tradition in AI, that HealthBranches is designed to probe.

From the standpoint of the thesis argument, HealthBranches serves a dual purpose. First, it provides an empirical measurement instrument: by comparing LLM performance with and without access to the underlying reasoning paths, the benchmark yields a quantitative proxy for the "reasoning gap" that persists in current models. The baseline results reported in this chapter confirm that models benefit dramatically from explicit graph-structured reasoning paths yet gain little from unstructured textual descriptions alone, thereby underscoring both the promise and the present limitations of KG-based enhancement strategies. Second, the dataset itself embodies the kind of structured, path-aware knowledge resource whose construction and exploitation have been a central concern of this thesis. In this sense, HealthBranches is not merely a benchmark *about* reasoning; it is a benchmark *built from* the same graph-structured knowledge that earlier chapters have advocated as external support for deep learning models.

The remainder of this chapter is organised as follows. Section 5.1 surveys the landscape of medical Q&A benchmarks and positions HealthBranches relative to existing work. Section 5.2 describes the semi-automated pipeline used to generate the dataset, elucidating how structured reasoning paths are extracted and why they constitute a principled basis for evaluating LLM reasoning. Section 5.3 presents the evaluation protocol, metrics, and the set of models under study. Section 5.4 reports the experimental results, covering both the dataset-creation phase and the benchmark evaluation. Finally, Section 5.5 discusses implications, limitations, and connections to future directions.

5.1 Existing benchmarks

Medical Q&A benchmarks vary widely in format and scope. Closed-ended multiple-choice datasets such as **HeadQA**, **MedQA**, **MeDiaQA**, and **PubMedQA** focus on fixed-answer formats for dialogue or biomedical text analysis [153, 135, 76, 77], while open-ended datasets like **MedCalc-BENCH** and **PathVQA** emphasise free-form responses but require manual verification [80, 62]. More recently, multi-task and multi-modal benchmarks have integrated textual and visual inputs to test reasoning under zero or few-shot conditions [174], and query-based datasets have emerged for structured data such as Electronic Health Records (EHRs) [104]. Other efforts include curated biomedical resources like **UltraMedical**, which aggregates manual and synthetic datasets for LLM fine-tuning [170], and SM3-Text-to-Query, the

Dataset (Year)	# questions	Type of questions	Knowledge	Qual. Reasoning	No-Comput.	Explanation
HeadQA (2019) [153]	6,765	Multiple choice	✓	✓	✓	✗
MEDCALC-BENCH (2024) [80]	1,000	Open-ended	✓	✓	✗	✓
MeDiaQA (2021) [135]	23,048	Multiple choice	✓	✗	✓	✗
MedQA (2020) [76]	61,097	Multiple choice	✓	✓	✓	✗
MedXpertQA (2025) [174]	4,460	Multi-modal questions	✓	✓	✓	✗
MIMIC-SPARQL (2020) [104]	10,000	Table- and graph-based queries	✓	✗	✗	✗
PathVQA (2020) [62]	32,799	Open-ended	✓	✗	✓	✗
PubMedQA (2019) [77]	273,500	Multiple choice	✓	✓	✓	✗
HealthBranches (2025)	4,063	Multiple choice & open-ended	✓	✓	✓	✓

Table 5.1: (✓= present, ✗= absent.) Summary of existing medical Q&A datasets and comparison with **HealthBranches**, considering four qualitative dimensions: (1) **Knowledge**: whether the dataset tests knowledge tied to a particular domain; (2) **Qualitative Reasoning**: whether the dataset tests logical or conceptual reasoning rather than quantitative computation; (3) **No Comput.**: whether the dataset does not require quantitative calculations, formulas, or numeric estimation; (4) **Explanation**: whether the dataset includes a step-by-step justification or reasoning trace. While these criteria are inherently qualitative, the classification follows the precedent set in MedCalc-BENCH [80].

first multi-model benchmark for SNOMED-CT–based queries [128].

As summarised in Table 5.1, most existing datasets either assess knowledge recall or offer limited explanation structures. MedQA and PubMedQA test factual knowledge without reasoning paths, while MedCalc-BENCH provides explanations but focuses on quantitative calculations. Similarly, **ACI-Bench** targets ambient clinical intelligence by benchmarking AI-assisted note generation from doctor–patient dialogues, emphasising narrative documentation rather than stepwise decision-making [165]. In contrast, HealthBranches uniquely combines multiple-choice and open-ended formats with explicit reasoning paths from clinical decision graphs, allowing for the evaluation of non-computational, interpretable, and step-by-step reasoning essential for trustworthy medical LLMs. As presented in Table 5.1, benchmarks that include richer supervision signals (e.g., calculus, images, graphs, and explicit reasoning paths) are typically smaller than simple QA corpora (e.g., MedQA) due to the substantially higher curation effort.

This distinction is especially relevant to the thesis argument. The structured reasoning paths embedded in HealthBranches are analogous to the traversals of a Knowledge Graph: each path encodes a sequence of clinical decisions from a root symptom to a leaf diagnosis or treatment, mirroring the node-edge traversals that KG-enhanced architectures perform at inference time. The benchmark is therefore a natural testbed for assessing whether the KG-enhancement strategies discussed in the previous chapter translate into measurable reasoning improvements and, when they do not, for diagnosing exactly where the reasoning chain breaks down.

5.2 Proposed Methodology

The central aim of the proposed methodology is to construct a benchmark that can measure the *reasoning* capabilities of LLMs, not merely their ability to recall medical facts, by grounding every question–answer pair **in a structured, graph-derived decision pathway**. The intuition is straightforward: if a clinical decision can be decomposed into a sequence of nodes (symptoms, tests, findings, treatments) connected by directed edges (decision steps), then an LLM that truly "reasons" should be able to follow that path when it is provided as context, and should perform markedly worse when it is forced to answer from parametric knowledge alone. Conversely, supplying the same clinical information as unstructured text rather than as a structured path, should yield a smaller benefit if the model cannot internally reconstruct the reasoning chain. This asymmetry between structured and unstructured context is the core signal that HealthBranches is designed to capture, and it connects directly to the thesis-level argument that graph-structured external knowledge is qualitatively different from, and superior to, flat textual context for supporting deep-learning-based reasoning.

The pipeline consists of three main stages, illustrated in Figure 5.1: (1) parsing medical decision pathways from authoritative textbooks, (2) generating Q&A pairs from the extracted reasoning paths, and (3) refining outputs through automated and human-in-the-loop validation. Each stage is described in the paragraphs that follow.

Parsing of the Knowledge Source. The knowledge source consists of medical textbooks [102, 60, 51] divided into sections, each detailing the management of a patient with a specific symptom or condition through a clinically validated algorithmic framework. These frameworks are typically graph-based representations (often decision trees), where each node corresponds to a diagnostic or therapeutic step. This structure encapsulates the clinical reasoning and domain knowledge for one of the 17 covered domains, while remaining applicable to other medical areas, ensuring broader generalizability. Each section provides two complementary information streams: the *Textual stream*, a narrative of clinical rationale and procedural steps, and the *Graphical stream*, a decision tree organising the reasoning process. Both are automatically extracted using **Gemini-flash 2.0** [142], producing plain-text files for content and graph files encoding the decision structure. While the primary sources are textbooks, the same pipeline applies to any resource that provides an equivalent structured pathway: society guidelines, institutional care protocols, or purpose-built Knowledge Graphs. This reinforces the thesis-level claim that the methodology is agnostic to the specific provenance of graph-structured knowledge and depends only on its availability in a node-edge format.

Path Extraction and Refinement. From each graph, root-to-leaf traversals are enumer-

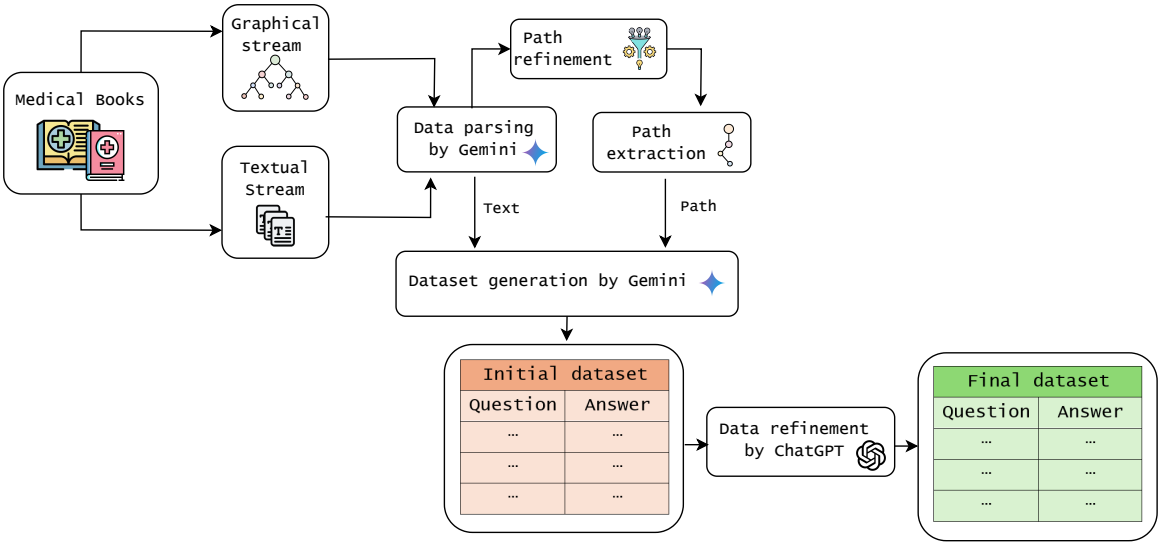


Figure 5.1: A step-by-step workflow for building the HealthBranches Q&A dataset. *Parsing the knowledge source*: textual and graphical streams are extracted from medical books and parsed. *Path extraction and refinement*: graph-based representations are processed to extract and refine relevant semantic paths. *Q&A generation*: an LLM is prompted with textual and path information to generate initial question–answer pairs. *Q&A refinement*: the dataset is further refined using another LLM to ensure consistency and quality.

ated, retaining up to two distinct paths per leaf to balance coverage and dataset size. Each path is then refined with Gemini-flash 2.0 (Appendix A.1), which normalizes terminology and removes formatting artefacts while preserving clinical semantics. The resulting paths are the structured reasoning chains that constitute the backbone of the benchmark: they define what a "correct reasoning process" looks like for each clinical scenario, and they serve as the reference against which LLM-generated answers are later judged. Figure 5.2 illustrates the complete transformation from a clinical decision graph to a refined reasoning path and its associated Q&A pair.

Q&A Generation. The dataset's Q&A content was generated by prompting Gemini-flash 2.0 with the text for a specific condition and its reasoning path. First, Gemini produced an open-ended answer based on the patient scenario, resulting in a detailed response. This answer was then used as the correct option in a multiple-choice format. To construct challenging distractors, Gemini also generated four alternative answers that were clinically plausible within the broader context of the condition yet *invalid* according to the specific reasoning path. This design is deliberate: it ensures that a model cannot answer correctly by pattern-matching surface-level medical terms, but must instead follow the logic of the decision pathway. Each reasoning path is thus linked to a Q&A pair and its distractors,

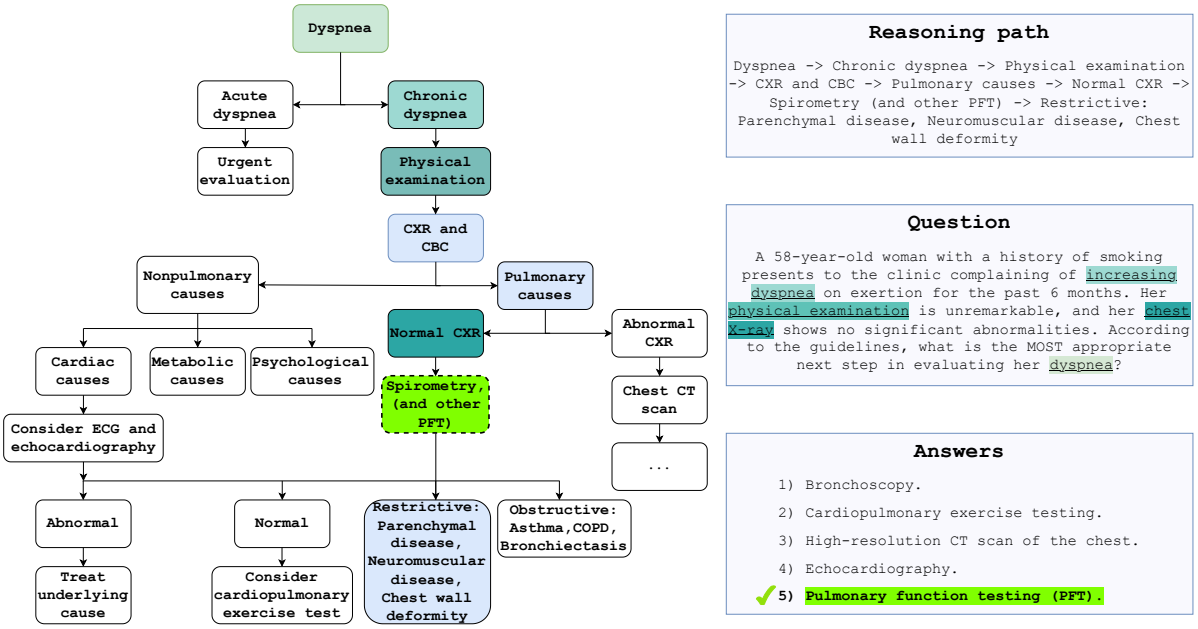


Figure 5.2: Detailed example of the Q&A generation process, starting from the extraction of a reasoning path. The left side shows an example graph outlining treatment steps for dyspnea, while the right side displays the extracted reasoning path and the resulting Q&A pair. The correct answer (option 5, Pulmonary function testing) is derived directly from the structured path, demonstrating how graph-based clinical knowledge is translated into a benchmark item.

enabling benchmarking in both open-ended and multiple-choice formats.

Q&A Refinement. The generation process may occasionally produce flawed examples, such as unanswerable questions, inconsistent answers, or details misaligned with clinical guidelines. To examine the impact of model scale on question difficulty and robustness, **Llama 3.1 (405B)** [52] was incorporated in the refinement phase, as it exceeds the capacity of the benchmark models and helps identify intrinsically flawed items. A two-stage procedure was implemented: first, questions misanswered by both **Llama 3.3 (70B)** and Llama 3.1 were collected; from this subset, only those that at least half of the benchmark models answered incorrectly were retained, resulting in 1,203 items individually proofread by five reviewers. Each item was then submitted to **GPT-4o** with web search and reasoning enabled, refined according to a defined protocol, and subsequently reviewed by a human evaluator. The review protocol, detailed in Appendix A.3, includes a standardised prompt and a Q&A template for submitting the item to be corrected. The results were compared to determine whether to modify the question, adjust the correct option, or discard the item if fundamentally flawed. The motivation for combining ChatGPT and Gemini throughout the pipeline (generation, flawed-path identification, and refinement) was to reduce systematic

biases from relying on a single model, thereby enhancing robustness and generalisability. This multi-model strategy is itself a form of the "external knowledge diversification" principle that has been a recurring theme in this thesis: just as KG-enhanced architectures benefit from drawing on multiple knowledge sources, the dataset construction pipeline benefits from the complementary strengths (and complementary failure modes) of different LLMs.

The Dataset as a Reasoning Probe. It is worth emphasising the conceptual link between the HealthBranches pipeline and the broader thesis. The decision pathways extracted from medical textbooks are, in essence, small Knowledge Graphs: directed acyclic graphs in which nodes represent clinical states or actions and edges represent decision steps. The Q&A pairs generated from these graphs test whether an LLM can reproduce the graph traversal—i.e., can it follow a structured reasoning chain? The topline evaluation (Section 5.3) then measures the *gap* between what a model achieves when the reasoning graph is explicitly provided and what it achieves from parametric knowledge alone. This gap is a direct, quantitative indicator of the "reasoning deficit" that KG-augmented systems are designed to address. A large gap indicates that the model *can* use structured reasoning when given it, but *cannot* reconstruct it internally the scenario in which external knowledge from a KG is most valuable. A small gap, by contrast, would suggest either that the model already internalizes sufficient structure, or that the benchmark items are too easy to discriminate. As the results in Section 5.4 will show, the observed gaps are large for all models tested, confirming the continued need for external, graph-structured reasoning support.

5.3 Evaluation Protocol

The evaluation strategies were applied to both the quiz-based and open-ended versions of the dataset. Models were tested in two complementary settings designed to isolate the contribution of structured reasoning versus unstructured context. In the **Topline** setting, each model receives, alongside the question, one of three forms of additional context: (a) only the reasoning path, (b) only the textual description, or (c) both. This three-way comparison is essential for the thesis argument: condition (a) simulates a KG-enhanced scenario in which the model has access to a graph-derived reasoning chain; condition (b) provides the same information in unstructured prose; and condition (c) combines the two. Any performance difference between (a) and (b) is attributable to the *structure* of the knowledge representation rather than to its informational content, thereby testing the central claim that graph-structured knowledge has a qualitative advantage over flat text.

In the **Benchmark** setting, models answer with the standard zero-shot Q&A prompt (question only), with and without Retrieval-Augmented Generation (RAG) context. The RAG

configuration employed the following parameters:

- Embedder Model: `mxbai-embed-large`
- Chunk Size: `500`
- Chunk Overlap: `150`
- Split Strategy: `RecursiveCharacterTextSplitter (LangChain)`

These values were not chosen arbitrarily. A grid search was performed over:

$$\text{chunk_size} = 500 + 100k, \quad k \in \{0, 1, 2, 3, 4, 5\}$$
$$\text{chunk_overlap} = 50 + 50j, \quad j \in \{0, 1, 2, 3\}$$

holding the embedder fixed, and the configuration reported above was the one maximizing retrieval quality on a testing subset of the benchmark. The retrieval strategy is deliberately kept simple, i.e. dense semantic search over a single vector index with no reranking, query expansion, or hybrid sparse-dense scoring. This choice is methodological rather than incidental: the purpose of the RAG condition in this chapter is *not* to build a competitive retrieval system, but to establish a reference point for what unstructured, retrieved context contributes relative to structured reasoning paths. A heavily engineered retrieval pipeline would confound the comparison by making the RAG condition a study of retrieval engineering rather than of knowledge representation.

All evaluation prompts are reported in Appendix A.4 (quiz) and Appendix A.5 (open-ended). Ten open, decoder-only LLMs that are freely available and deployable via the Ollama platform were selected for benchmarking: **Mistral 7B** [75], **Llama 2 7B** [146], **Llama 3.1 8B** [53], **Gemma 7B** [144], **Gemma 2 9B** [145], **Gemma 3 4B** [143], **Qwen 2.5 7B** [163], **DeepSeek-R1 8B** [57], **Phi-4 14B** [2], and **Mistral NeMo 12B**. To ensure a consistent evaluation framework, all models were evaluated in **chat mode under zero-shot settings without chain-of-thought prompting**. Two larger models were also assessed via the Together AI platform: **Llama 3.3 70B** and **Llama 3.1 405B** [53], the latter used only during the question control phase.

Three complementary metrics were employed, each targeting a different facet of answer quality:

- **Exact Match** was used in the quiz evaluation: models were instructed via the system prompt to respond with a single letter (A–E), a regular expression extracted the letter from each response, and accuracy was computed as the proportion of extracted letters matching the ground-truth key (the extraction algorithm is detailed in Appendix A.2).

- **LLM-as-a-Judge** scoring was used to assess the quality of open-ended answers: an LLM-based judge (*Gemini-flash 2.0*) employing in-context learning scored each response on a 0–10 scale using the ground-truth answer and reasoning path as reference. This approach aligns with recent work showing that LLMs can effectively serve as automated evaluators in clinical settings, reducing the cost and variability of expert annotation while maintaining strong agreement with human judgements [37]. Specifically, the G-Eval metric [91] was adopted, which operates in two phases: first, the judge is introduced to the task and evaluation criteria to generate assessment steps; second, these steps are combined with the criteria and task description to produce the final score. The same judge was used throughout all evaluations; implementation details and prompts are provided in Appendix A.8.

- **Semantic Similarity** metric based on the BGE-M3 embedding model [35] was computed, comparing the ground-truth answer with the generated open response. BGE-M3 can produce dense, sparse, and ColBERT-style embeddings, allowing for comparison between semantic and lexical similarity. This metric serves two purposes: validating the reliability of the LLM-as-a-judge scores and providing insight into the strategies models use to answer open-ended Q&A.

5.4 Experiments and Findings

This section presents the main findings from pipeline development and model evaluation. The discussion proceeds in three stages: first, the characteristics of the resulting dataset are described; second, the quality-assurance experiments conducted during dataset creation are reported; third, the benchmark evaluation of LLMs on the final dataset is analyzed, with particular attention to the reasoning implications.
Details on experimental and hardware setup are available in Appendix A.9.

5.4.1 Dataset Characteristics

After filtering, the final HealthBranches dataset contains 4,063 question–answer pairs across various medical specialties (Table 5.2). This diversity ensures that the dataset reflects both common and specialised domains encountered in clinical practice. Categories with a similar number of sub-conditions (e.g., *Hematology/Oncology* vs. *Rheumatology*) differ in question count due to variations in the complexity of their clinical reasoning graphs: more complex domains produce a wider variety of reasoning paths and, consequently, more questions, while simpler categories yield fewer. A useful proxy for this complexity is the number of

Category	Questions	Conditions	Leafs
Hematology/Oncology	531	22	147
General Medicine	457	15	228
Infectious Diseases	441	15	241
Neurology	373	19	258
Women's Health	350	18	152
Gastrointestinal	338	21	239
Nephrology	279	17	138
Cardiology	278	18	153
Rheumatology	217	22	197
Behavioral Medicine	190	8	79
Endocrinology	183	16	117
Pharmacology	146	8	68
Pulmonary Disease	125	12	97
Emergency Medicine	54	7	36
Urology	45	5	30
Dermatology	35	5	25
Ocular	21	4	22

Table 5.2: Distribution of questions and number of leaf nodes across clinical categories in the HealthBranches dataset.

leaf nodes in the reasoning graph for each category, as each leaf node can contribute up to two sampled reasoning paths (to avoid skewing overrepresented classes). Simpler categories were intentionally not inflated artificially in order to maintain semantic diversity and reduce redundancy.

Each question is aligned with a clear and structured reasoning path, similar to a clinical decision tree, that integrates domain-specific knowledge checks with qualitative reasoning steps and interpretable explanations. Combined with its support for both multiple-choice and open-ended formats, this design makes HealthBranches well-suited for training and evaluating trustworthy, explanation-aware medical LLMs.

5.4.2 Experiments on Dataset Creation

As noted in Section 5.2, the automatic generation process may introduce flawed or imprecise examples. A two-pronged quality-assurance campaign was therefore conducted, combining automated model-based validation with expert human review.

Figure 5.3 shows the accuracy gains after applying the refinement protocol, with significant improvements in both zero-shot and RAG settings, particularly for larger models. The

improvements confirm that the refinement step is not merely cosmetic: it removes items whose ground-truth labels were incorrect or whose question stems were ambiguous, thereby tightening the benchmark signal. For additional information regarding the primary types of corrections, including adjusted clinical phrasing, fixed inconsistencies, and replacement of incorrect answers, please refer to Appendix A.3.

An initial expert validation of the dataset was also conducted with practising physicians and senior medical students, following the structured protocol in Appendix A.10. Each item comprising the question, multiple-choice answers, and reasoning, was evaluated through written feedback and a structured scoring system. Each subcomponent (question formulation, answer accuracy, and reasoning quality) received a score ranging from 0 to 5, yielding a maximum total score of 15 per item. A total of 145 reviews were submitted by 22 reviewers, yielding high average scores across all dimensions: 4.68 for question accuracy, 4.68 for answer accuracy, and 4.61 for reasoning path quality. Physicians gave more conservative ratings, while students scored higher (full results in Appendix A.11, Table A.3). Only 19 questions received a 3 or below in any category, and just 6 fell below 9/15 overall. These results suggest that the pipeline produces clinically reliable Q&A content, though broader expert review could further enhance consistency.

5.4.3 Benchmark Results and Reasoning Analysis

Models were evaluated under all topline and benchmark settings to assess the impact of reasoning paths on performance. Figure 5.4 presents the results across three metrics: quiz accuracy (top), G-Eval judge scores for open-ended answers (middle), and cosine similarity between predictions and ground truth (bottom). Numerical values for all conditions are available in Appendix A.7.

The findings reveal a consistent pattern across all three metrics and all eleven models. When models receive the explicit reasoning path (*topline-path*), performance improves substantially. For instance, Mistral 7B jumps from 51.9% accuracy in the zero-shot setting to 87.7% when given the path, while even the best-performing model, Llama 3.3 70B, gains nearly 17 percentage points (from 75.8% to 92.7%). However, when models receive only the textual description of the same clinical scenario (*topline-text*), gains are marginal: Llama 3.3 70B improves by less than 1 percentage point, and several smaller models show no improvement at all. This asymmetry, large gains from structured paths, negligible gains from unstructured text containing the same information, demonstrates that current LLMs cannot reliably extract reasoning steps from natural-language descriptions, even when the relevant knowledge is present in the context window.

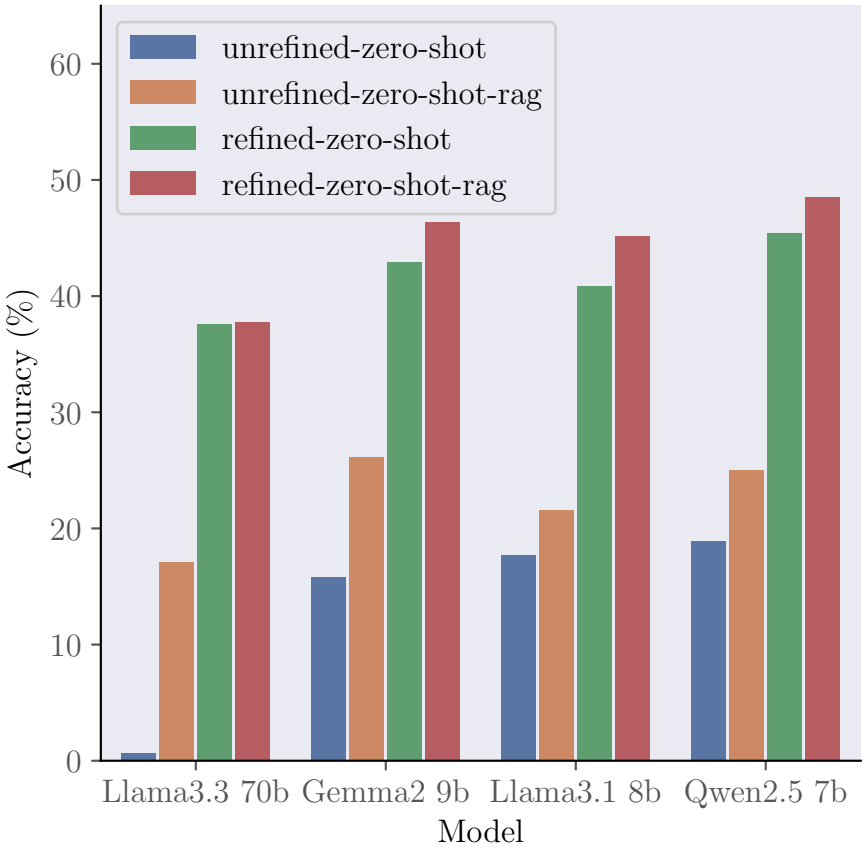


Figure 5.3: Improved accuracy across models after the Q&A refinement process.

The introduction of RAG information similarly yielded only modest improvements over zero-shot performance, particularly for larger models. Notably, smaller models such as Llama 2 exhibited greater enhancements from RAG, suggesting that the contextual information retrieved by RAG is already embedded in the parametric knowledge of newer, larger models. Results from open-ended questions align with those from the quiz setting: G-Eval judge scores and BGE-M3 semantic similarity both track the same pattern, reinforcing the use of LLM-as-a-judge as a dependable quality metric and confirming that the reasoning gap is not an artefact of the multiple-choice format.

These results carry a clear implication for the broader thesis. The topline condition with the explicit reasoning path simulates a scenario in which a KG-derived reasoning chain is made available to the model at inference time, precisely the architecture advocated in the preceding chapters. The large performance gap between this condition and the zero-shot baseline constitutes direct empirical evidence that the *structured* character of knowledge-graph-style representations is what drives reasoning improvements, not merely the presence of additional context. The baseline (zero-shot) performance of even the best-performing models, which

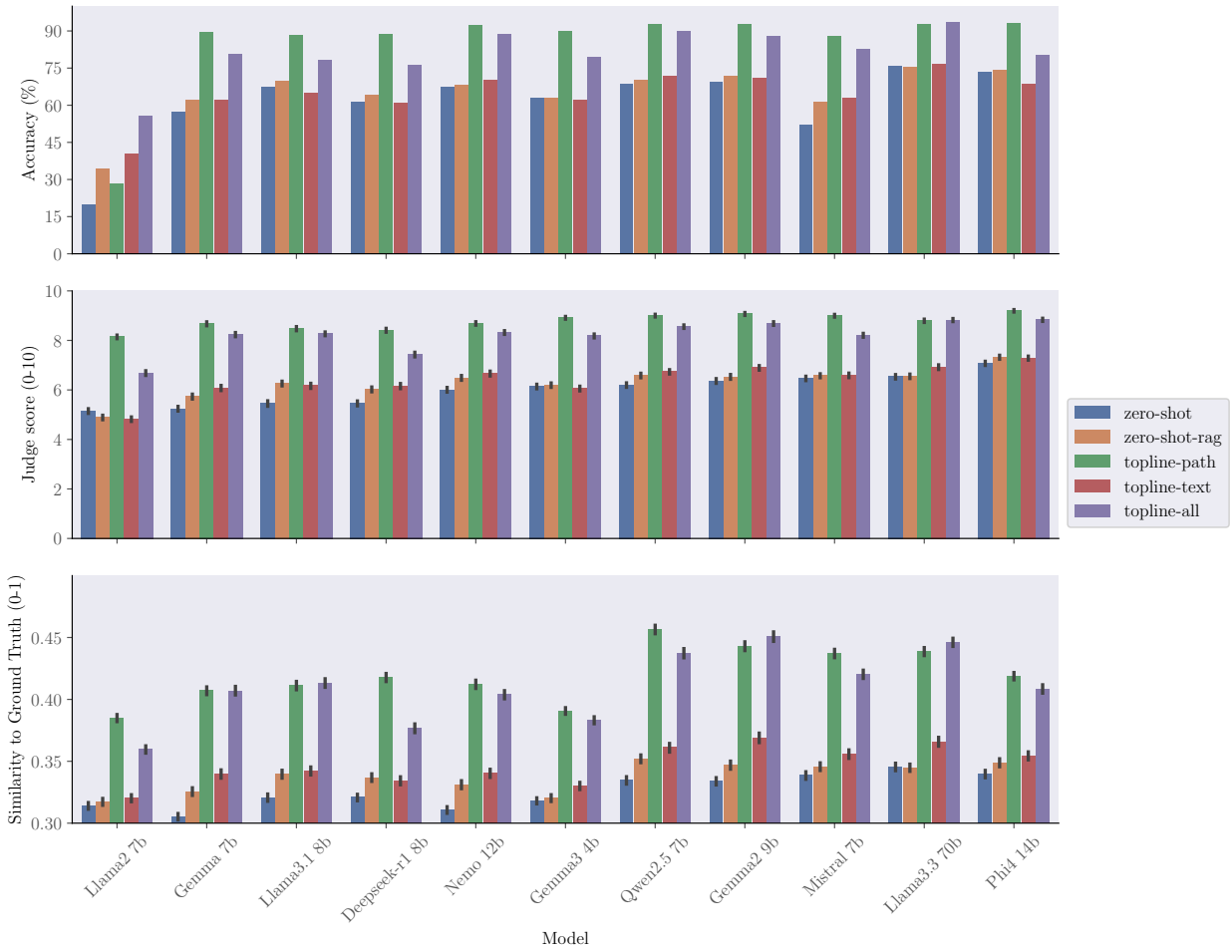


Figure 5.4: Model performance on HealthBranches. Top: quiz accuracy across settings. Middle: G-eval scores for open-ended answers. Bottom: cosine similarity between predictions and ground truth. The bottom plot's y-axis is truncated for visual clarity, as cosine similarity measures relative semantic alignment rather than absolute distance from zero. Error bars show 95% confidence intervals across all plots.

plateaus around 75% accuracy on a benchmark where the reasoning path alone pushes accuracy above 92%, provides a quantitative measure of the remaining reasoning gap that future KG-enhanced architectures must close. Conversely, the minimal benefit of unstructured text in the topline-text condition confirms that the challenge is not one of information availability but of *structural representation*: models need knowledge to be pre-organized into reasoning chains, which is exactly what Knowledge Graphs provide.

5.5 Conclusion

This chapter has presented HealthBranches, a benchmark dataset for advancing medical Q&A through clinically grounded reasoning. By extracting and refining decision pathways from medical literature, a resource was created that supports quiz-style and open-ended formats while embedding reasoning chains for multi-step inference. The semi-automated pipeline combines structured knowledge extraction, LLM generation, and expert refinement, producing high-quality Q&A pairs across 17 clinical domains. Evaluations across 11 LLMs under zero-shot and RAG conditions show that models benefit significantly from structured reasoning paths, highlighting the limits of relying only on text. In addition, LLM-as-a-judge and semantic similarity metrics align closely with human assessments, reinforcing the dataset's validity as a rigorous benchmark.

The primary aim of HealthBranches is to provide a reasoning-focused benchmark that assesses the stepwise decision quality of LLMs rather than their ability to memorise facts. To facilitate community growth, a public, extensible template is released so that researchers can add new sources such as textbooks, society guidelines, or institutional care pathways, provided they can be represented in the same node-edge format. Path-aware evaluation promotes explainability and helps curb hallucinations by requiring models to justify answers via explicit steps.

The dataset generation pipeline has certain limitations that must be acknowledged. While it utilises advanced language models, these models can introduce inherent biases, and the pipeline may inadvertently propagate inaccuracies or outdated information from source texts. Additional caution is warranted with sensitive medical content, as expert review is necessary to prevent errors that could impact clinical understanding. The dataset is also imbalanced across medical specialties, as more complex domains naturally generate a greater number of reasoning paths and questions, while simpler categories yield fewer. This skew may limit generalizability, and future updates will address it by incorporating additional Q&A pairs from dedicated textbooks to strengthen underrepresented areas. Expert validation was conducted on a limited scale (17 clinical topics), reflecting the ongoing refinement of the dataset which requires substantial time and domain-specific effort; future iterations will aim to expand expert involvement to enhance clinical fidelity and robustness across topics.

It should nonetheless be acknowledged that the RAG baseline explores a narrow region of the design space. Whether advanced retrieval strategies, such as hybrid sparse-dense scoring, cross-encoder reranking, or graph-based retrieval over the decision pathways themselves, can narrow the gap towards the topline-path condition remains an open question, and one that HealthBranches is explicitly designed to support.

Within the broader arc of this dissertation, HealthBranches demonstrates that the gap between LLMs' parametric knowledge and the reasoning demanded by complex real-world tasks remains substantial and that graph-structured external knowledge offers a principled means of closing it. The benchmark results provide both a baseline against which future KG-enhanced models can be measured and a concrete illustration of how structured decision pathways, the very artefacts produced and exploited by Knowledge Graph pipelines, unlock reasoning capabilities that unstructured text alone cannot. HealthBranches represents a significant advancement in the development of interpretable, safe, and trustworthy LLMs for medical applications. By aligning evaluations with clinical decision-making processes, it serves as a valuable resource for both research and education.

Chapter 6

Conclusions and Future Directions

This dissertation has pursued, across four interconnected research questions, a single guiding hypothesis: that the deep learning systems on which contemporary biomedical artificial intelligence is increasingly being built cannot reach the levels of robustness, generalization, and reasoning that their applications demand by relying on data alone, and that the principled integration of structured prior knowledge encoded as ontologies, knowledge graphs, or decision pathways, provides the inductive bias on which such progress depends. The argument has been developed neither as a critique of data-driven methods nor as a defence of symbolic representations, but as an attempt to characterise, with empirical precision and methodological rigour, the roles that structured knowledge can play along a trajectory of increasing depth: from a remedy against scarcity, to an architectural constraint, to an object of learning, and finally to a criterion against which the reasoning capabilities of state-of-the-art generative models are measured. The four chapters that constitute the technical body of this thesis correspond, one for one, to the four stations of this trajectory, and each has produced concrete contributions whose synthesis is the subject of the present chapter.

6.1 The trajectory of structured knowledge: a synthesis

The starting point of the dissertation, articulated in Chapter 2, is the empirical reality that motivates the rest of the program. Biomedical machine learning operates under conditions that violate, in often-aggravated combination, every assumption on which the most powerful deep learning models rely. Cohorts are small; labels are scarce, expensive, and noisy; class distributions are severely imbalanced; and representational diversity is bounded by the institutional and demographic homogeneity of the centres from which the data originate. The

two case studies presented in that chapter were chosen so as to instantiate two complementary facets of this multidimensional condition. The first, on the prediction of long-COVID sequelae from a real-world clinical databank of over three hundred patients, addressed the prototypical low-sample, high-dimensional regime, and showed that a careful combination of multi-label upsampling through MLSmote and dimensionality control can yield predictive performance of clinical interest while preserving, under SHAP-based interpretability analysis, the qualitative explanations on which clinical adoption ultimately rests. The second, on breast-cancer subtype classification from histopathological Whole-Slide Images, confronted the parallel problem of label scarcity, and introduced the Histopathological Masked AutoEncoder (HMAE) as a self-supervised pipeline capable of learning Vision-Transformer representations from unlabelled tissue regions; the resulting embeddings reached performance statistically comparable to fully supervised state-of-the-art methods on BRACS, transferred competitively to BACH, and produced t-SNE projections and self-attention maps in which class-coherent structure emerged without the model ever having been exposed to a single class label during representation learning.

Taken together, these two studies establish, with concrete numbers, that data-centric techniques retain genuine value in the biomedical setting, and that they remain a necessary first line of defence against scarcity. They also, however, reveal the structural ceiling that the rest of the dissertation sets out to address. Augmentation cannot manufacture the rare disease subtypes that are absent from the cohort; self-supervision can decouple representation learning from annotation, but it cannot encode a clinical guideline that the training images do not exemplify; and no purely statistical regularisation can ensure that a model, however well-fit to its training distribution, respects the functional, hierarchical, and causal structure that decades of medical research have already articulated. The No Free Lunch theorem, recalled at the opening of Chapter 3, formalises this limit: in the absence of assumptions about the structure of the problem space, no learning algorithm can be expected to generalize. The question, in biomedicine, is therefore not whether to use prior knowledge, but how to make it operational within learning architectures whose default mode of operation is statistical association over raw signals.

Chapter 3 took the first step along this path by reframing the problem as one of inductive bias engineering. The Transformer was singled out as a particularly revealing target: its near-complete absence of architectural prior, the very property responsible for its scalability across domains, is also what forces it to rediscover, from data alone, the regularities that domain experts have already specified. In a field as densely structured as medicine, this is doubly wasteful, and Structural Positional Encoding (SPE) was developed as a minimalist mechanism to remedy it. By computing Laplacian eigenvectors of an ontology graph,

projecting them through a single learnable matrix, and adding the result to the token embeddings consumed by the attention mechanism, SPE injects ontological proximity between clinical activities directly into the representation space within which the Transformer attends. Experimental validation on a stroke-management dataset showed that this minimal intervention yields gains of approximately nine percentage points at top-1 accuracy and over ten percentage points at top-3 over a Transformer with sinusoidal positional encoding, persists under five and ten per cent input perturbation, and transfers to LSTM architectures with similar magnitude. The qualitative analysis with domain experts further demonstrated that an SPE-augmented model, when asked what should follow mechanical thrombectomy, predicts trans-chest ECG monitoring rather than encephalic MRI, aligning with the SPREAD guidelines on post-procedural cardiovascular surveillance and demonstrating that knowledge integration can shift the model not only toward more accurate predictions, but toward predictions that respect the priorities encoded in clinical protocols.

A revealing negative finding accompanies these positive results, and is in many respects more theoretically eloquent than the gains themselves: standard sinusoidal positional encoding, the foundation of the original Transformer, provides no benefit and occasionally degrades performance in this setting. The chapter argues that this is not an isolated curiosity but a symptom of a structural mismatch: medical processes are not rigidly sequential chains in which position three differs categorically from position five, but functionally organised hierarchies in which activities cluster by clinical goal and the ordering within a functional group is often clinically immaterial. A positional encoding designed for natural-language token sequences therefore introduces a prior that is at best uninformative and at worst actively misleading. The lesson generalizes well beyond the SPE construction itself: the question one must ask of any inductive bias is not whether it is sophisticated, but whether it is the right one for the structure of the domain. Structural knowledge integration, on this reading, is not an add-on to the Transformer but a corrective to a default assumption that does not match the data.

Chapter 4 marked a deeper commitment by elevating the knowledge graph from auxiliary signal to primary object of study. The empirical platform was the Unified Medical Language System, comprising more than two million biomedical concepts and 3.5 million relational assertions, and the methodological aim was twofold: to evaluate, in a controlled comparison, how different geometric inductive biases, namely translations (TransE), bilinear products (DistMult, ComplEx), and rotations (RotatE), capture the relational structure of biomedical knowledge, and to investigate whether the resulting latent spaces support inference patterns that go beyond the single-hop facts on which they were trained. The first investigation produced a clear ranking, RotatE > ComplEx > TransE > DistMult, consistent across all twelve

evaluation metrics and amounting to a 5.9% gap in MRR between the best and worst models, and showed that augmenting these scoring functions with R-GCN-based neighbourhood aggregation yields an improvement when paired with TransE but a degradation when paired with RotatE. The asymmetry was interpreted theoretically: aggregation-induced smoothing helps geometric priors that lack the expressiveness to capture rich relational patterns, but harms those whose precise angular structure can be disrupted by representational blurring, and over-smoothing in particular is amplified at the high-degree hub entities that dominate biomedical graphs.

The second investigation produced what the dissertation regards as one of its most theoretically suggestive findings. A closed-form compositional query solver was developed for q2p1 conjunctive queries in RotatE's complex latent space, in which two anchor entities are projected through their respective relations, the intersection is approximated by the centroid of the projected images, and a final relation projects the centroid toward the answer region. Evaluated on 88,669 type-coherent multi-hop queries, the model that had been trained exclusively on single-hop triples nevertheless retrieved the correct answer entity within the top ten predictions for over 83% of queries (Hits@10 = 0.8339), with a Hits@25 of 0.8705, only 8.1 percentage points below the single-hop ceiling. The interpretation offered in the chapter is mathematically precise: because RotatE encodes relations as element-wise rotations on the unit circle in C^d, and because the composition of two rotations is itself a rotation, the latent space inherits, by closure, a compositional structure that the training objective never explicitly demanded. Emergent compositional behaviour is not, on this reading, a fortuitous side-effect; it is the direct geometric consequence of an inductive bias whose algebraic properties match those of relational composition itself. The principle generalizes beyond UMLS: when the chosen geometry is consonant with the algebra of the domain, learning a single-step task can implicitly endow the latent space with the capacity to support multi-step inference, while geometries that lack this consonance, such as the bilinear scoring of DistMult, would not be expected to exhibit the same behaviour.

Chapter 5, finally, turned the structured-knowledge argument outward toward what is currently the most prominent question in medical artificial intelligence: whether modern Large Language Models are genuinely capable of clinical reasoning, and whether the gap between what they remember and what they can derive can be closed by the very graph-shaped representations advocated in the preceding chapters. The proposed contribution, HealthBranches, is a benchmark of 4,063 question-answer items derived, through a semi-automated pipeline of structured knowledge extraction, language-model generation, and expert refinement, from explicit decision pathways covering 17 clinical domains. Crucially, every item is paired with the full reasoning chain from which it was generated, enabling a controlled four-way com-

parison among zero-shot answering, retrieval-augmented generation, supplementation with the unstructured textual description of the relevant scenario (topline-text), and supplementation with the structured graph-derived path itself (topline-path). The asymmetry between the latter two conditions was constructed precisely to provide a direct empirical test of the central claim of the dissertation, namely that the structural character of graph-based knowledge, and not merely the information it conveys, is what unlocks reasoning capabilities in modern generative models.

The eleven evaluated models confirmed this prediction unambiguously. Mistral 7B, supplied with the explicit reasoning path, rose from 51.9% to 87.7% accuracy; Llama 3.3 70B, the strongest baseline at 75.8% in the zero-shot setting, rose to 92.7% under topline-path. By contrast, the same information conveyed as unstructured text yielded only marginal gains: less than one percentage point for Llama 3.3 70B, and no improvement at all for several smaller models. RAG context produced similarly modest improvements, with the gains concentrated in older or smaller models, suggesting that retrieval mostly recovers information that newer models already possess parametrically. The G-Eval and BGE-M3 metrics on the open-ended portion of the benchmark tracked the multiple-choice results closely, ruling out the possibility that the asymmetry is an artefact of the quiz format. The conclusion is empirical and direct: current Large Language Models cannot, on their own, reliably extract reasoning steps from natural-language descriptions of clinical scenarios, even when the relevant knowledge is fully present in the context window; what they require, in order to close the gap between parametric recall and stepwise inference, is for that knowledge to be pre-organised into the kind of structured, node-and-edge pathways that knowledge graphs provide.

6.2 Methodological reflections

Beyond the specific contributions, several methodological reflections deserve to be made explicit, as they apply with equal force to all four research questions and may have value for the broader research community.

The first concerns the relationship between scale and structure. A recurrent temptation in contemporary deep learning is to assume that any obstacle to performance can be overcome by larger models, larger datasets, or more compute. The empirical results of this dissertation suggest that, in specialised domains characterised by limited data and strong prior structure, this assumption can be misleading. SPE's 9% gain over a sinusoidal-encoded Transformer was obtained without changing the model size, the optimiser, or the training data; the gain came from aligning the inductive bias of the architecture with the functional organization

of the domain. RotatE's emergent compositional capacity was not produced by additional training on multi-hop queries, but by a geometry whose algebraic closure happens to support relational composition. HealthBranches' topline-path improvements were not produced by larger models, but by re-organising the same information into a structured form. In each case, structure proved to be a more efficient lever than scale, at least within the regimes considered.

The second concerns interpretability, which has surfaced repeatedly in the dissertation under different guises. The SHAP analyzes of Chapter 2 demonstrated that data augmentation, when carried out carefully, need not distort qualitative explanations. The qualitative comparison of Chapter 3 showed that knowledge-enhanced predictions can be inspected against clinical guidelines and judged for their adherence to medical priorities. The geometric construction of Chapter 4 makes the inference process algorithmically transparent: every step of the q2p1 query solver corresponds to a well-defined operation in complex latent space. The path-based evaluation of Chapter 5 confronts a model with the explicit reasoning chain a human would follow. Across the four chapters, interpretability has therefore been treated not as a post-hoc rationalisation but as a property to be designed into the system from the outset, and the consistency with which structured knowledge has supported this property across very different methodological settings is, in the view of this work, one of the strongest practical arguments for the integration paradigm.

The third concerns the criterion of evaluation. Each chapter has, in its own way, complemented standard quantitative metrics with a more demanding form of validation. The long-COVID study juxtaposed predictive accuracy with SHAP explanations vetted for clinical plausibility; HMAE's matched-or-better numbers against fully supervised baselines were re-read in the operational light of annotation cost; SPE's accuracy gains were reinforced by a domain-expert evaluation of the clinical coherence of the predicted activities; the q2p1 results were not reported as a single number but decomposed across Hits@k thresholds whose convergence pattern was itself diagnostic of where compositional error accumulates; HealthBranches was deliberately designed so that the asymmetry between topline-path and topline-text could constitute, on its own, a measurement of the reasoning gap. The cumulative implication is that, in domains as consequential as biomedicine, a benchmark number is rarely a sufficient summary of model behaviour, and the design of evaluation protocols that capture the relevant dimensions of reliability is itself a research contribution of comparable importance to the design of the models.

6.3 Limitations

The dissertation has explicit limitations that should be discussed.

The long-COVID study was conducted on a single-centre cohort recruited at a single Italian hospital, whose representativeness across health-care systems and patient populations cannot be assumed. The consequences of this constraint extend beyond external generalizability: because admission criteria, intensive-care thresholds and follow-up protocols vary systematically between institutions, a multi-centre cohort would in all likelihood alter the very label statistics on which the methodological choices of Section 2.2 were predicated. The low SCUMBLE value ($S = 0.012$) that justified resampling-based augmentation reflects a comparatively homogeneous case-mix; a pooled cohort spanning centres with different severity profiles would plausibly increase the concurrence between minority and majority labels, narrowing the regime in which MLSmote can be applied without distorting the joint label distribution. Three mitigation strategies follow directly. First, leave-one-site-out validation, in place of the pooled cross-validation adopted here, would estimate transportability rather than internal fit, and the difference between the two estimates would itself quantify site-specific overfitting. Second, mixed-effects formulations treating the recruiting centre as a random effect, together with per-centre stratification of the synthetic augmentation, would prevent the resampling procedure from interpolating between clinically incompatible practice patterns. Third, where data-sharing agreements preclude pooling, federated training would allow the estimation of shared parameters without centralising patient records. The expected outcome of such a design is a *lower* nominal performance than the one reported in this chapter, but a less optimistically biased one: the informative question would no longer be whether accuracy is preserved, but whether the SHAP attribution rankings that were shown here to survive augmentation also remain stable across centres, since it is the invariance of the explanations, and not the magnitude of the metrics, that would establish the clinical robustness of the model.

The HMAE evaluation, although it covered both BRACS and a cross-dataset transfer to BACH, does not yet probe the full diversity of staining protocols, scanners, and tissue types encountered in deployed digital-pathology pipelines, and its self-supervised representations have been validated only on a restricted set of downstream classification tasks.

The SPE experiments were conducted on one stroke-management dataset, and while the ontology-based inductive bias is in principle domain-agnostic, the magnitude of the improvement may vary across clinical pathways with different ratios of structured to unstructured variability.

The compositional inference study of Chapter 4 was restricted to a single query type (q2p1)

and a single intersection approximation (centroid aggregation); both choices were deliberate and methodologically motivated, but they leave open the question of how the observed emergent behaviour scales to longer chains, conjunctions of more than two anchors, and queries that involve unions or negations.

The HealthBranches benchmark, finally, is currently imbalanced across medical specialties, has been expert-validated on a subset of its 17 domains, and reflects, by construction, the choices of the source textbooks from which its decision pathways were extracted. The extension of the corpus to society guidelines and institutional care pathways is part of its public template, and it is worth stating explicitly what such an extension would be expected to change, since the two sources would act in opposite directions. Clinical practice guidelines issued by learned societies are already formulated in algorithmic form, with graded recommendations and explicit decision nodes, and their incorporation would therefore reduce the extraction noise introduced by the semi-automated pipeline and supply a principled criterion for weighting pathways by strength of evidence. They are, however, also heavily represented in the pre-training corpora of the evaluated models, so their addition would be expected to raise the zero-shot baseline and thus to *compress* the measured gap between the no-context and topline-path conditions. Institutional care pathways have the converse property: locally negotiated, rarely published in machine-readable form, and probably absent from pre-training data, they would constitute the cleanest available test of genuine path-following as opposed to parametric recall, and would be expected to *widen* that gap. Read jointly, these two predictions suggest that the 75.8% zero-shot accuracy reported here should be regarded as an optimistic upper bound and the 16.9-point improvement under the topline-path condition as a conservative estimate of the reasoning gap. Two further consequences of such an expansion should be anticipated: guidelines carry an explicit temporal validity that textbooks do not, so the benchmark would require versioning and periodic re-validation to avoid rewarding adherence to superseded recommendations; and institutional pathways vary across health-care systems, which would make it necessary to report performance stratified by provenance rather than pooled. The empirical results reported here should therefore be read as a baseline relative to that future expansion.

6.4 Future directions

Several directions for future work emerge naturally from the trajectory developed across the four chapters, and from the limitations just acknowledged.

A first direction concerns the integration of explicitly trained query-answering architectures, such as Query2Box and BetaE, with the biomedical knowledge-graph framework of Chapter

4, in order to quantify the gap between the emergent compositional inference reported here and the supervised compositional inference these architectures are designed to support. A second direction lies in the closer coupling of the knowledge-enhanced Transformer of Chapter 3 with the compositional latent space of Chapter 4: a model in which the positional representations of clinical activities are derived not from offline spectral methods but from a knowledge-graph embedding optimised jointly with the sequence-modelling objective could, in principle, combine the architectural simplicity of SPE with the inferential richness of RotatE's geometry. A third direction, and the one most directly suggested by Chapter 5, is the development of generative architectures in which structured pathways are not simply provided at inference time, but are induced from text and grounded in a knowledge graph as part of the reasoning trace itself; HealthBranches has been designed so that the gap between parametric reasoning and path-augmented reasoning can be measured for any new architecture, and it is the ambition of this dissertation that this gap will narrow over time, not by forcing models to memorise more pathways, but by equipping them with the structural sensitivity that the topline-path condition currently approximates from outside.

A fourth direction, finally, follows from the methodological reflections of Section 6.2. The empirical evidence accumulated here suggests that structure is, in many respects, a more efficient lever than scale within the regimes considered, but the trade-off between the two has not been characterised systematically. Future work could therefore investigate, across the four problem settings considered in this dissertation, how the marginal benefit of structural priors evolves as model capacity, training-data volume, and pre-training regime are scaled up; the prediction that follows from the present results is that structure and scale are complementary rather than substitutive, and that the most capable systems will be those in which both axes are exploited together.

6.5 Closing remarks

The thesis began with the observation that biomedicine is, simultaneously, the domain in which deep learning faces its hardest version of the data-scarcity problem and the domain that is most generously endowed with structured prior knowledge. The four chapters that followed have argued, through a combination of empirical investigation and theoretical analysis, that this asymmetry is not a coincidence but an opportunity: the very richness of medical ontologies, knowledge graphs, and decision pathways that has accumulated through decades of clinical research is precisely the resource that can compensate for the scarcity of usable data, provided that learning architectures are designed so as to operate on it natively rather than treat it as an afterthought. The four contributions, the multidimensional

characterisation of biomedical scarcity and its data-driven mitigation, the development of Structural Positional Encoding for knowledge-enhanced Transformers, the large-scale empirical and theoretical study of learning on UMLS with its emergent compositional inference, and the construction of the HealthBranches benchmark for measuring the reasoning gap in Large Language Models, were each conceived to instantiate one stage of this argument, and together describe a single coherent program.

If a single message is to be drawn from the dissertation as a whole, it is the following. In the next generation of biomedical artificial intelligence, the question will not be whether deep learning models can be trained on enough data to perform well in a single specialised task; it will be whether they can reason, robustly and transparently, over the structured knowledge that defines their domain. The empirical evidence assembled here suggests that this capacity is achievable, that it depends on the careful alignment of architectural inductive biases with the algebra of medical knowledge, and that it can be measured, today, with benchmarks that distinguish reasoning from recall. The road from graph to reasoning, in the title of this thesis, is therefore not a metaphor but a methodological claim, and it is the hope of the author that the contributions presented here may serve as one milestone along a path on which much remains to be travelled.

Chapter 7

Computational Cost and Reproducibility

7.1 Reporting Conventions

This chapter consolidates the computational profile of every model and system trained or evaluated in the experimental chapters of this thesis, so that the hardware budget required to replicate any of the studies can be sized in advance.
Each study is described by a homogeneous set of entries: hardware platform, software framework, dataset scale, model configuration, trainable parameters, training workload, training time, and per-sample inference cost. Two conventions apply throughout.

- **Trainable parameters** are those updated by gradient descent during the training run of the study in question. Fixed sinusoidal positional encodings are excluded, as are the parameters of frozen components.
- **Inference cost per sample** is reported as wall-clock latency where it was measured, and otherwise as the exact arithmetic cost of a single forward pass in multiply-accumulate operations (MAC), which is hardware-independent and derives deterministically from the documented architecture.

A dash (—) denotes a quantity that was not instrumented during the original experimental runs.

Platform	Accelerator	Memory	Used in
CPU workstation	—	System RAM	Chapter 2 (long-COVID)
GPU node A	1× NVIDIA A40	48 GB GDDR6 ECC	Chapter 2 (HMAE)
GPU node B	1× NVIDIA Quadro RTX 6000	24 GB GDDR6	Chapters 3, 4, 5

Table 7.1: Hardware platforms. The GPU nodes are high-performance nodes equipped with an Intel Xeon Gold CPU, 192 GB of system RAM and Ubuntu Linux. All experiments were run on a single accelerator; no distributed or multi-GPU training was employed.

7.2 Chapter 2 (i): Long-COVID Multi-Label Classification

The long-COVID study is the only experimental contribution of this thesis that does not involve a neural model. Eight problem-transformation MLC strategies were instantiated with a 200-tree Random Forest base classifier and executed entirely on CPU.

Hardware	CPU workstation
Framework	Meka (multi-label extension of Weka)
Dataset scale	324 instances (`orig`), 424 after MLSmote; 57 features (`orig`), 8 (`Relief`), 44 (`PCA`); 7 labels
Model configuration	Random Forest, 200 trees, unconstrained depth, split factor $\lfloor \log_2(\#\text{features}) + 1 \rfloor$ (= 6 for 57 and 44 features, = 4 for 8 features)
Trainable parameters	Not applicable (non-parametric); structural complexity in Table 7.3
Training workload	9 datasets × 8 methods × 5 runs of 10-fold cross-validation = 3,600 outer fits, corresponding to 21,150 Random Forest ensembles and 4,230,000 induced decision trees
Training time	—
Inference per sample	—

Table 7.2: Computational profile of the long-COVID multi-label classification study.

7.3 Chapter 2 (ii): HMAE (Histopathological Masked Autoencoder)

HMAE couples a ViT-S/16 encoder with a masked-autoencoder decoder trained self-supervised on randomly cropped tissue regions, with 75% of patches masked. Downstream classification

MLC method	Base ensembles	Trees	Base-model calls / sample
BR	7	1,400	7
CC	7	1,400	7
BCC(I)	7	1,400	7
BCC(C)	7	1,400	7
LC	1	200	1
PS	1	200	1
CDN	7	1,400	3,500
RAkEL	10	2,000	10
Total per fold	**47**	**9,400**	—

Table 7.3: Structural complexity of the long-COVID MLC pipeline. Each base ensemble is a 200-tree Random Forest. CDN performs inference through Gibbs sampling over $I = 500$ iterations, so a single test instance requires $500 \times 7 = 3{,}500$ binary Random Forest evaluations, making it by two orders of magnitude the most expensive method at inference time despite an identical training footprint to BR or CC.

uses a single-hidden-layer MLP over mean-aggregated, frozen ROI embeddings; the decoder is discarded after pre-training.

Hardware	1× NVIDIA A40, 48 GB
Dataset scale	BRACS: 547 WSIs, 4,539 annotated ROIs; BACH: 400 ROIs; 70/10/20 train/validation/test split, 100 repetitions per classification experiment
Model configuration	ViT-S/16 encoder: embedding dimension 384, 12 blocks, 6 heads, MLP ratio 4, patch size 16, masking ratio 75%. Reference MAE decoder: embedding dimension 512, 8 blocks, 16 heads, MLP ratio 4
Trainable parameters	47,401,728 (≈ 47.4 M) during self-supervised pre-training, of which 21,590,016 (≈ 21.6 M) belong to the encoder that is retained afterwards (Table 7.5)
Training workload	Self-supervised reconstruction (MSE) on random crops extracted from the BRACS WSIs
Training time	32 hours (single A40) for pre-training. The downstream stage trains only the MLP head on pre-computed embeddings, at a cost negligible with respect to pre-training
Inference per sample	—

Table 7.4: Computational profile of the HMAE study.

Component	**Configuration**	**Parameters**
Patch embedding	$16 \times 16 \times 3 \to 384$	295,296
`[CLS]` token	384	384
Transformer blocks	12× (d=384, 6 heads, MLP×4)	21,293,568
ViT-S/16 encoder subtotal		**21,590,016**
Decoder input projection	$384 \to 512$	197,120
Mask token	512	512
Decoder blocks	8× (d=512, 16 heads, MLP×4)	25,219,072
Reconstruction head	$512 \to 768$	393,984
Decoder subtotal		**25,811,712**
Pre-training total		**47,401,728**

Table 7.5: Trainable parameter budget of HMAE. Positional encodings are fixed sinusoidal buffers and are excluded. The decoder accounts for 54.5% of the parameters trained and is discarded after pre-training: only the 21.6 M-parameter encoder is retained to produce ROI embeddings, and the supervised stage updates solely the MLP head on top of those frozen embeddings.

7.4 Chapter 3: SPE-Augmented Transformer for Stroke Process Monitoring

A Transformer decoder is trained autoregressively on stroke process traces, with Laplacian eigenvector embeddings of the domain ontology injected through a learned linear projection $\Theta \in \mathbb{R}^{k \times d}$. The eigendecomposition is computed once offline and adds no cost to training.

7.5 Chapter 4: UMLS Knowledge Graph Embedding Models

Six models were trained on a knowledge graph derived from the UMLS 2025AA release. Parameter counts are dominated by the entity embedding table: with over two million entities, the 438 relation embeddings account for less than 0.03% of the total.

All models allocate 512 *real* components per entity, so that the comparison is made at equal representational budget. For the complex-valued models this corresponds to 256 complex components: RotatE stores 256 complex entity coordinates and 256 relation phases, while ComplEx stores 256 complex components for both entities and relations. In the terms of the reference implementation, this is a hidden dimension of 256 with entity (and, for

Hardware	1× NVIDIA Quadro RTX 6000, 24 GB
Dataset scale	5,342 traces; vocabulary of 82 activities plus 3 special tokens; mean trace length 15, maximum 25; ontology of 110 nodes and 111 edges; 80/10/10 split; 10 fits per configuration (100 for the trace-classification task)
Model configuration	Transformer decoder: 4 layers, 4 heads, feed-forward size 128, dropout 0.216, ontology embedding size $k = 32$, AdamW with StepLR ($\gamma = 0.990$), learning rate 2.836×10^{-3}; embedding size $d \in \{16, 32, 64, 128\}$. LSTM baseline: 1 layer, 128 hidden units, dropout 0.39
Trainable parameters	25 K to 441 K depending on d (Table 7.7)
Training workload	Autoregressive next-activity prediction over all prefixes of length 1 to n, plus a weighted cross-entropy classification objective on an appended class token
Training time	—
Inference per sample	—

Table 7.6: Computational profile of the SPE study.

Model	d	**Per layer**	**SPE projection Θ**	**Total**
Transformer	16	5,392	528	25,205
Transformer	32	12,704	1,056	58,517
Transformer	64	33,472	2,112	151,253
Transformer	128	99,584	4,224	441,173
LSTM baseline	128	—	4,224	158,165

Table 7.7: Trainable parameter counts for the next-activity prediction models, computed from the Optuna-selected configuration (feed-forward size 128, 4 heads, 4 layers, $k = 32$, vocabulary 85). Sinusoidal encodings are non-trainable, so the None, PE and SPE variants at a given d differ only by the $32d + d$ parameters of the structural projection: at $d = 128$ this is 1.0% of the model.

ComplEx, relation) embeddings doubled.

Training is organised in *optimization steps* rather than epochs. Each step consumes one batch of 50,000 positive triples together with their type-constrained negatives, and successive steps alternate between the two corruption modes, head prediction and tail prediction. Traversing the training split once therefore requires $\lceil 2{,}453{,}317/50{,}000 \rceil = 50$ steps in a single corruption mode, and ≈ 98 steps for a full epoch in which every training triple has been presented under both modes. Table 7.10 converts the measured throughput into the corresponding wall-clock quantities.

Hardware	1× NVIDIA Quadro RTX 6000, 24 GB
Dataset scale	2,096,327 entities, 438 relation types, 3,504,739 assertions; stratified 70/10/20 split
Model configuration	Embedding dimension 512 real components for all models; margin-based ranking loss with $\gamma = 1.0$; 1 type-constrained negative sample per positive triple; Adam with learning rate 5×10^{-4}; Xavier uniform initialisation; batch size 50,000 triples; hybrid models use a 2-layer R-GCN encoder without basis decomposition
Trainable parameters	$\approx$ 1.07 B per geometric model; +230.2 M for the R-GCN encoder (Table 7.9)
Training workload	2,453,317 training triples; 50 optimization steps per corruption mode to traverse the split once, i.e. $\approx$ 98 steps per full epoch (Table 7.10)
Training time	3.2 optimization steps per minute, i.e. 18.75 s per step and $\approx$ 31 minutes per full epoch
Inference per sample	—

Table 7.8: Computational profile of the UMLS knowledge graph completion study.

Model	**Trainable parameters**
TransE	1,073,543,680
DistMult	1,073,543,680
ComplEx	1,073,543,680
RotatE	1,073,543,680
R-GCN encoder (2 layers, $d = 512$, $R = 438$)	230,162,432
R-GCN + TransE	1,303,706,112
R-GCN + RotatE	1,303,593,984

Table 7.9: Trainable parameter counts for the UMLS models at embedding dimension 512. Training a 1.07B-parameter embedding table with Adam requires approximately 4GB.

7.6 Chapter 5: HealthBranches LLM Benchmark

HealthBranches is an inference-only benchmark: no model was trained or fine-tuned, so the number of trainable parameters is zero throughout. The relevant figures are the frozen parameter counts, the quantised memory footprints and the measured per-response latencies.

Quantity	**Value**
Batch size (positive triples)	50,000
Training triples	2,453,317
Steps per single pass over the split	50
Steps per full epoch (both corruption modes)	≈98
Measured throughput	3.2 steps/min
Time per optimization step	18.75 s
Time per full epoch	≈31 min
Positive triples processed per second	≈2,670
Scored triples per second (positives + negatives)	≈5,330

Table 7.10: Measured training throughput for the UMLS models on the Quadro RTX 6000 node. The rate is governed by memory traffic over the 4 GiB embedding table and by the rejection sampling of type-constrained negatives, rather than by arithmetic: a single step performs only $O(|\mathcal{B}| \cdot d)$ arithmetic operations, which on this hardware would otherwise complete in a small fraction of a second.

Hardware	1× NVIDIA Quadro RTX 6000, 24 GB
Framework	Ollama v0.6.3 with CUDA and cuDNN; Together AI for the two largest models
Dataset scale	4,063 Q&A items across 17 clinical domains, in quiz and open-ended formats, evaluated under five conditions (zero-shot, zero-shot + RAG, topline-path, topline-text, topline-all)
Model configuration	Ten local decoder-only LLMs under two quantisation schemes, Q8_0 (8-bit weights) and Q4_K_M (4-bit k-means); zero-shot chat mode without chain-of-thought prompting. RAG: `mxbai-embed-large` embedder, chunk size 500, overlap 150, LangChain `RecursiveCharacterTextSplitter`
Trainable parameters	0 (no fine-tuning); frozen counts in Table 7.12
Training workload	None
Training time	None
Inference per sample	0.41–7.23 s (quiz), 0.69–6.00 s (open-ended); see Table 7.12

Table 7.11: Computational profile of the HealthBranches benchmark.

Model	**Quantization**	**Params**	**Weights**	**Quiz (s)**	**Open (s)**
Gemma 3 4B	Q4_K_M	4.3 B	2.4 GiB	0.528	0.738
Llama 2 7B	Q8_0	6.7 B	6.7 GiB	0.489	1.567
Mistral 7B	Q8_0	7.2 B	7.2 GiB	0.440	0.906
Qwen 2.5 7B	Q4_K_M	7.6 B	4.3 GiB	0.413	0.691
Llama 3.1 8B	Q4_K_M	8.0 B	4.5 GiB	0.547	1.012
DeepSeek-R1 8B	Q4_K_M	8.0 B	4.5 GiB	7.233	6.000
Gemma 7B	Q8_0	8.5 B	8.5 GiB	0.449	0.834
Gemma 2 9B	Q8_0	9.2 B	9.1 GiB	0.572	0.935
Mistral NeMo 12B	Q8_0	12.2 B	12.1 GiB	0.574	0.746
Phi-4 14B	Q4_K_M	14.7 B	8.3 GiB	1.429	2.240
Llama 3.3 70B	remote (Together AI)	70.6 B	—	—	—
Llama 3.1 405B	remote (Together AI, control phase only)	405 B	—	—	—
`mxbai-embed-large`	—	0.335 B	0.6 GiB	—	—

Table 7.12: Frozen parameter counts, quantized memory footprints and measured average per-response latency for the HealthBranches benchmark. Latency is largely independent of parameter count, as expected in a memory-bandwidth-bound autoregressive decoding regime; the outlier is DeepSeek-R1 8B, whose latency reflects the generation of reasoning tokens that are suppressed from the visible output but still produced.

Bibliography

[1] Tecnomed-hub webpage. https://www.tecnomedhub.it. Accessed: 2023-06-30.

[2] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

[3] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[4] M. Z. Alom, T. M. Taha, C. Yakopcic, S. Westberg, P. Sidike, M. S. Nasrin, M. Hasan, B. C. Van Essen, A. A. S. Awwal, and V. K. Asari. A state-of-the-art survey on deep learning theory and architectures. *Electronics*, 8(3), 2019.

[5] Alhanoof Althnian, Duaa AlSaeed, Heyam Al-Baity, Amani Samha, Alanoud Bin Dris, Najla Alzakari, Afnan Abou Elwafa, and Heba Kurdi. Impact of dataset size on classification performance: an empirical evaluation in the medical domain. *Applied sciences*, 11(2):796, 2021.

[6] Guilherme Aresta, Teresa Araújo, Scotty Kwok, Sai Saketh Chennamsetty, Mohammed Safwan, Varghese Alex, Bahram Marami, Marcel Prastawa, Monica Chan, Michael Donovan, et al. Bach: Grand challenge on breast cancer histology images. *Medical image analysis*, 56:122–139, 2019.

[7] Anmol Arora, Joseph E Alderman, Joanne Palmer, Shaswath Ganapathi, Elinor Laws, Melissa D Mccradden, Lauren Oakden-Rayner, Stephen R Pfohl, Marzyeh Ghassemi, Francis Mckay, et al. The value of standards for health datasets in artificial intelligence-based applications. *Nature medicine*, 29(11):2929–2938, 2023.

[8] A.B. Atkinson. On the measurement of inequality. *Journal of Economic Theory*, 2(3):244—263, 1970.

[9] Bulut Aygüneş, Selim Aksoy, Ramazan Gökberk Cinbiş, Kemal Kösemehmetoğlu, Sevgen Önder, and Ayşegül Üner. Graph convolutional networks for region of interest classification in breast histopathology. In *Medical Imaging 2020: Digital Pathology*, volume 11320, pages 134–141. SPIE, 2020.

[10] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[11] Samy Badreddine, Artur d'Avila Garcez, Luciano Serafini, and Michael Spranger. Logic tensor networks. *Artificial Intelligence*, 303, 2022.

[12] Jinshuai Bai, Laith Alzubaidi, Qingxia Wang, Ellen Kuhl, Mohammed Bennamoun, and Yuantong Gu. Utilising physics-guided deep learning to overcome data scarcity. *arXiv preprint arXiv:2211.15664*, 2022.

[13] Ms Aayushi Bansal, Dr Rewa Sharma, and Dr Mamta Kathuria. A systematic review on data scarcity problem in deep learning: solution and applications. *ACM Computing Surveys (Csur)*, 54(10s):1–29, 2022.

[14] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

[15] Jonathan Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198, 2000.

[16] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[17] Mattia Bellan, Daria Apostolo, Alice Albè, Martina Crevola, Nicolò Errica, Giacomo Ratano, Stelvio Tonello, Rosalba Minisini, Davide D'Onghia, Alessio Baricich, et al. Determinants of long covid among adults hospitalized for sars-cov-2 infection: A prospective cohort study. *Frontiers in Immunology*, 13:1038227, 2022.

[18] Mattia Bellan, Annalisa Chiocchetti, Marco Dossena, Christopher Irwin, Luca Piovesan, and Luigi Portinale. Predicting long-covid sequelae: A multi-label classification approach. *Intelligenza Artificiale*, page 17248035251317937, 2025.

[19] Mattia Bellan, Daniele Soddu, Piero Emilio Balbo, et al. Respiratory and psychophysical sequelae among patients with COVID-19 four months after hospital discharge. *JAMA Network Open*, 4(1):e2036142, 2021.

[20] Richard Bellman. Dynamic programming. *science*, 153(3731):34–37, 1966.

[21] Mourad Benmalek and Abdessamed Seddiki. Bias in federated learning: Factors, effects, mitigations, and open issues. *Revue des Sciences et Technologies de l'Information-Série ISI: Ingénierie des Systèmes d'Information*, 29(6):2137–2160, 2024.

[22] Alessandro Berti and Wil M. P. van der Aalst. Reviving token-based replay: Increasing speed while improving diagnostics. In Wil M. P. van der Aalst, Robin Bergenthum, and Josep Carmona, editors, *Proceedings of the International Workshop on Algorithms & Theories for the Analysis of Event Data 2019 Satellite event of the conferences: 40th International Conference on Application and Theory of Petri Nets and Concurrency Petri Nets 2019 and 19th International Conference on Application of Concurrency to System Design ACSD 2019, ATAED@Petri Nets/ACSD 2019, Aachen, Germany, June 25, 2019*, volume 2371 of *CEUR Workshop Proceedings*, pages 87–103. CEUR-WS.org, 2019.

[23] Olivier Bodenreider. The unified medical language system (umls): Integrating biomedical terminology. *Nucleic acids research*, 32:D267–70, 02 2004.

[24] Jasmin Bogatinovski, Ljupčo Todorovski, Sašo Džeroski, and Dragi Kocev. Comprehensive comparative study of multi-label classification methods. *Expert Systems with Applications*, 203:117215, 2022.

[25] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26, 2013.

[26] Nadia Brancati, Anna Maria Anniciello, Pushpak Pati, Daniel Riccio, Giosuè Scognamiglio, Guillaume Jaume, Giuseppe De Pietro, Maurizio Di Bonito, Antonio Foncubierta, Gerardo Botti, et al. Bracs: A dataset for breast carcinoma subtyping in h&e histology images. *Database*, 2022:baac093, 2022.

[27] Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

[28] Zaharah A. Bukhsh, Aaqib Saeed, and Remco M. Dijkman. Processtransformer: Predictive business process monitoring with transformer network, 2021.

[29] C. Cabanillas, C. DiCiccio, J. Mendling, and A. Baumgrass. Predictive task monitoring for business processes. In S. Wasim Sadiq, P. Soffer, and H. Völzer, editors,

Business Process Management - 12th International Conference, BPM 2014, Haifa, Israel, September 7-11, 2014. Proceedings, volume 8659 of *Lecture Notes in Computer Science*, pages 424–432. Springer, 2014.

[30] M. Camargo, M. Dumas, and O. González Rojas. Learning accurate LSTM models of business processes. In T. T. Hildebrandt, B. F. van Dongen, M. Röglinger, and J. Mendling, editors, *Business Process Management - 17th International Conference, BPM 2019, Vienna, Austria, September 1-6, 2019, Proceedings*, volume 11675 of *Lecture Notes in Computer Science*, pages 286–302. Springer, 2019.

[31] Angelo Carfì, Roberto Bernabei, and Francesco Landi. Persistent symptoms in patients after acute COVID-19. *JAMA*, 324(6):603–605, 2020.

[32] F. Charte, A.J. Rivera, M.J. delJesus, and F. Herrera. MLSMOTE: approaching imbalanced multilabeled learning through synthetic instance generation. *Knowledge Based Systems*, 89:385–397, 2015.

[33] F. Charte, A.J. Rivera, M.J. delJesus, and F. Herrera. Dealing with difficult minority labels in imbalanced mutilabel data sets. *Neurocomputing*, 326-327:39–53, 2019.

[34] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16(1):321–357, 2002.

[35] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 2024.

[36] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation, 2014.

[37] Emma Croxford, Yanjun Gao, Elliot First, Nicholas Pellegrino, Miranda Schnier, John Caskey, Madeline Oguss, Graham Wills, Guanhua Chen, Dmitriy Dligach, et al. Automating evaluation of ai text generation in healthcare with a large language model (llm)-as-a-judge. *medRxiv*, pages 2025–04, 2025.

[38] Hannah E Davis, Gina S Assaf, Lisa McCorkell, et al. Characterizing long COVID in an international cohort: 7 months of symptoms and their impact. *EClinicalMedicine*, 38:101019, 2021.

[39] Ittai Dayan, Holger R. Roth, Aoxiao Zhong, et al. Federated learning for predicting clinical outcomes in patients with COVID-19. *Nature Medicine*, 27:1735–1743, 2021.

[40] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[41] Chiara DiFrancescomarino, Chiara Ghidini, Fabrizio Maria Maggi, Giulio Petrucci, and Anton Yeshchenko. An eye into the future: leveraging a-priori knowledge in predictive business process monitoring. In *Business Process Management: 15th International Conference, BPM 2017, Barcelona, Spain, September 10–15, 2017, Proceedings 15*, pages 252–268. Springer, 2017.

[42] Ivan Donadello, Jonghyeon Ko, Fabrizio Maria Maggi, Jan Mendling, Francesco Riva, and Matthias Weidlich. Knowledge-driven modulation of neural networks with attention mechanism for next activity prediction, 2023.

[43] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[44] Marco Dossena, Christopher Irwin, Annalisa Chiocchetti, and Luigi Portinale. Beyond labels: A self-supervised framework with masked autoencoders and random cropping for breast cancer sub-type classification. *Proceedings of the AAAI Symposium Series*, 5(1):10–17, May 2025.

[45] Vijay Prakash Dwivedi and Xavier Bresson. A generalization of transformer networks to graphs. *AAAI Workshop on Deep Learning on Graphs: Methods and Applications*, 2021.

[46] Vijay Prakash Dwivedi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Graph neural networks with learnable structural and positional representations. In *International Conference on Learning Representations (ICLR)*, 2022.

[47] Jan-Niklas Eckardt, Waldemar Hahn, Christoph Röllig, Sebastian Stasik, Uwe Platzbecker, Carsten Müller-Tidow, Hubert Serve, Claudia D Baldus, Christoph Schliemann, Kerstin Schäfer-Eckart, et al. Mimicking clinical trials with synthetic acute myeloid leukemia patients using generative artificial intelligence. *NPJ digital medicine*, 7(1):76, 2024.

[48] J. Evermann, J.R. Rehse, and P. Fettke. Predicting process behaviour using deep learning. *Decision Support Systems*, 100:129–140, 2017.

[49] E. Frank, M.A. Hall, and I.H. Witten. The WEKA workbench. In *Data Mining: Practical Machine Learning Tools and Techniques, 4th ed. (online Appendix)*. 2016.

[50] Eva Gibaja and Sebastián Ventura. A tutorial on multilabel learning. *ACM Computing Surveys (CSUR)*, 47(3):1–38, 2015.

[51] Alexander Goldfarb-Rumyantzev. *Critical Care Medicine: An Algorithmic Approach E-Book*. Elsevier Health Sciences, 2021.

[52] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[53] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[54] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 507–520, 2022.

[55] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

[56] Ahmet Gokberk Gul, Oezdemir Cetin, Christoph Reich, Nadine Flinner, Tim Prangemeier, and Heinz Koeppl. Histopathological image classification based on self-supervised vision transformer and weak labels. In *Medical Imaging 2022: Digital and Computational Pathology*, volume 12039, pages 366–373. SPIE, 2022.

[57] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[58] Yuhong Guo and Suicheng Gu. Multi-label classification using conditional dependency networks. In *Proc. 22nd Intern. Joint Conference on Artificial Intelligence (IJCAI 11)*, pages 1300–1305, 2011.

[59] William L. Hamilton, Payal Bajaj, Marinka Zitnik, Dan Jurafsky, and Jure Leskovec. Embedding logical queries on knowledge graphs, 2019.

[60] Mark Harrison and Ala Mohammed. *Algorithms for emergency medicine*. Oxford University Press, 2023.

[61] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[62] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

[63] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

[64] M. Hinkka, T. Lehto, K. Heljanko, and A. Jung. Classifying process instances using recurrent neural networks. In F. Daniel, Q. Z. Sheng, and H. Motahari, editors, *Business Process Management Workshops - BPM 2018 International Workshops, Sydney, NSW, Australia, September 9-14, 2018, Revised Papers*, volume 342 of *Lecture Notes in Business Information Processing*, pages 313–324. Springer, 2018.

[65] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality ofdata with neural networks. *Science*, 313:504–507, 2006.

[66] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[67] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, et al. Knowledge graphs. *ACM Computing Surveys*, 54(4), 2021.

[68] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[69] Le Hou, Dimitris Samaras, Tahsin M Kurc, Yi Gao, James E Davis, and Joel H Saltz. Patch-based convolutional neural network for whole slide tissue image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2424–2433, 2016.

[70] Le Hou, Dimitris Samaras, Tahsin M. Kurc, Yi Gao, James E. Davis, and Joel H. Saltz. Patch-based convolutional neural network for whole slide tissue image classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2424–2433, 2016.

[71] Chaolin Huang, Lixue Huang, Yeming Wang, et al. 6-month consequences of covid-19 in patients discharged from hospital: a cohort study. *The Lancet*, 397(10270):220–232, 2021.

[72] Youhe Huang, Rongfeng Zhang, Hongru Li, Yunlong Xia, Xia Yu, Songbai Liu, and Yiheng Yang. A multi-label learning prediction model for heart failure in patients with atrial fibrillation based on expert knowledge of disease duration. *Applied Intelligence*, pages 1–12, 2023.

[73] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.

[74] Christopher Irwin, Marco Dossena, Giorgio Leonardi, and Stefania Montani. Structural positional encoding for knowledge integration in transformer-based medical process monitoring and trace classification. *Progress in Artificial Intelligence*, pages 1–13, 2024.

[75] Fengqing Jiang. Identifying and mitigating vulnerabilities in llm-integrated applications. Master's thesis, University of Washington, 2024.

[76] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

[77] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, 2019.

[78] George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6), 2021.

[79] M. A. Khan, H. Le, K. Do, T. Tran, A. Ghose, K. H. Dam, and R. Sindhgatta. Memory-augmented neural networks for predictive process analytics. *CoRR*, abs/1802.00938, 2018.

[80] Nikhil Khandekar, Qiao Jin, Guangzhi Xiong, Soren Dunn, Serina Applebaum, Zain Anwar, Maame Sarfo-Gyamfi, Conrad Safranek, Abid Anwar, Andrew Zhang, et al. Medcalc-bench: Evaluating large language models for medical calculations. *Advances in Neural Information Processing Systems*, 37:84730–84745, 2024.

[81] Kenji Kira and Larry A. Rendell. A practical approach to feature selection. In *Proceedings of the Ninth International Workshop on Machine Learning (ML 92)*, page 249–256, San Francisco, CA, USA, 1992. Morgan Kaufmann Publishers Inc.

[82] Devin Kreuzer, Dominique Beaini, William L. Hamilton, Vincent Létourneau, and Prudencio Tossou. Rethinking graph transformers with spectral attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

[83] Clemens Scott Kruse, Rishi Goswamy, Yesha Raval, and Sarah Marawi. Challenges and opportunities of big data in health care: A systematic review. *JMIR Med Inform*, 4(4):e38, Nov 2016.

[84] G.T. Lakshmanan, D. Shamsi, Y. N. Doganata, M. Unuvar, and R. Khalaf. Markov prediction model for data-driven semi-structured business processes. *Knowl. Inf. Syst.*, 42 (1):97–126, 2015.

[85] M. Le, B. Gabrys, and D. Nauck. A hybrid model for business process event prediction. In M. Bramer and M. Petridis, editors, *Research and Development in Intelligent Systems XXIX, Incorporating Applications and Innovations in Intelligent Systems XX: Proceedings of AI-2012, The Thirty-second SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence, Cambridge, England, UK, December 11-13, 2012*, pages 179–192. Springer, 2012.

[86] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[87] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2021.

[88] Tiancheng Lin, Zhimiao Yu, Hongyu Hu, Yi Xu, and Chang-Wen Chen. Interventional bag multi-instance learning on whole-slide pathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19830–19839, 2023.

[89] Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 4228–4238, 2021.

[90] Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 2901–2908, 2020.

[91] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore, December 2023. Association for Computational Linguistics.

[92] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.

[93] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. NIPS'17, page 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc.

[94] Gjorgji Madjarov, Dragi Kocev, Dejan Gjorgjevikj, and Sašo Džeroski. An extensive experimental comparison of methods for multi-label learning. *Pattern recognition*, 45(9):3084–3104, 2012.

[95] Tiroshan Madushanka and Ryutaro Ichise. Negative sampling in knowledge graph representation learning: A review. *arXiv preprint arXiv:2402.19195*, 2024.

[96] Robin Manhaeve, Sebastijan Dumančić, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. DeepProbLog: Neural probabilistic logic programming. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

[97] R. Mans, H. Schonenberg, G. Leonardi, S. Panzarasa, A. Cavallini, S. Quaglini, and et al. Process mining techniques: an application to stroke care. In S. Andersen, G. O.

Klein, S. Schulz, and J. Aarts, editors, *Proc. MIE*, volume 136 of *Studies in Health Technology and Informatics*, pages 573–578. IOS Press, Amsterdam, 2008.

[98] Lourdes Mateu, Cristian Tebe, Cora Loste, José Ramón Santos, Gemma Lladós, Cristina López, Sergio España-Cueto, Ruth Toledo, Marta Font, Anna Chamorro, et al. Determinants of the onset and prognosis of the post-covid-19 condition: a 2-year prospective observational cohort study. *The Lancet Regional Health–Europe*, 33, 2023.

[99] N. Di Mauro, A. Appice, and T. M. A. Basile. Activity prediction of business process instances with inception CNN models. In M. Alviano, G. Greco, and F. Scarcello, editors, *AI*IA 2019 - Advances in Artificial Intelligence - XVIIIth International Conference of the Italian Association for Artificial Intelligence, Rende, Italy, November 19-22, 2019, Proceedings*, volume 11946 of *Lecture Notes in Computer Science*, pages 348–361. Springer, 2019.

[100] N. Mehdiyev, J. Evermann, and P. Fettke. A multi-stage deep learning approach for business process event prediction. In P. Loucopoulos, Y. Manolopoulos, O. Pastor, B. Theodoulidis, and J. Zdravkovic, editors, *19th IEEE Conference on Business Informatics, CBI 2017, Thessaloniki, Greece, July 24-27, 2017, Volume 1: Conference Papers*, pages 119–128. IEEE Computer Society, 2017.

[101] Jorge M Mendes, Aziz Barbar, and Marwa Refaie. Synthetic data generation: a privacy-preserving approach to accelerate rare disease research. *Frontiers in Digital Health*, 7:1563991, 2025.

[102] Stuart B Mushlin and Harry L Greene. *Decision making in medicine: an algorithmic approach*. Elsevier Health Sciences, 2009.

[103] A Nalbandian and et al. Post-acute covid-19 syndrome. *Nature Medicine*, 27(4):601–615, 2021.

[104] Junwoo Park, Youngwoo Cho, Haneol Lee, Jaegul Choo, and Edward Choi. Knowledge graph-based question answering with electronic health records. In *Machine Learning for Healthcare Conference*, pages 36–53. PMLR, 2021.

[105] R. Pascanu, Ç. Gülçehre, K. Cho, and Y. Bengio. How to construct deep recurrent neural networks. In Y. Bengio and Y. LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[106] Pushpak Pati, Guillaume Jaume, Lauren Alisha Fernandes, Antonio Foncubierta-Rodríguez, Florinda Feroce, Anna Maria Anniciello, Giosue Scognamiglio, Nadia Brancati, Daniel Riccio, Maurizio Di Bonito, et al. Hact-net: A hierarchical cell-to-tissue graph neural network for histopathological image classification. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis: Second International Workshop, UNSURE 2020, and Third International Workshop, GRAIL 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 8, 2020, Proceedings 2*, pages 208–219. Springer, 2020.

[107] Mor Peleg. Computer-interpretable clinical guidelines: A methodological review. *Journal of Biomedical Informatics*, 46(4), 2013.

[108] Vasileios C Pezoulas, Dimitrios I Zaridis, Eugenia Mylona, Christos Androutsos, Kosmas Apostolidis, Nikolaos S Tachos, and Dimitrios I Fotiadis. Synthetic data generation methods in healthcare: A review on open-source tools and methods. *Computational and structural biotechnology journal*, 23:2892–2910, 2024.

[109] P. Philipp, R. Jacob, S. Robert, and J. Beyerer. Predictive analysis of business processes using neural networks with attention mechanism. In *2020 International Conference on Artificial Intelligence in Information and Communication, ICAIIC 2020, Fukuoka, Japan, February 19-21, 2020*, pages 225–230. IEEE, 2020.

[110] Doan Hoang Phu, Sarunya Maneerattanasak, Shamarina Shohaimi, Le Thanh Thao Trang, Truong Thanh Nam, Muminah Kuning, Aunchalee Like, Hameedah Torpor, and Charuai Suwanbamrung. Prevalence and factors associated with long covid and mental health status among recovered covid-19 patients in southern thailand. *PLoS One*, 18(7):e0289382, 2023.

[111] L. Portinale and L. Saitta. Feature selection. Technical Report D14.1 EU Project MiningMart, University of Dortmund, 2002.

[112] Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378, 2019.

[113] P. Rana, A. Sowmya, E. Meijering, and Y. Song. Imbalanced classification for protein subcellular localization with multilabel oversampling. *Bioinformatics*, 39(1):btac841, 2023.

[114] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. *Machine learning*, 85:333–359, 2011.

[115] Jesse Read, Bernhard Pfahringer, and Geoffrey Holmes. Multi-label classification using ensembles of pruned sets. In *Proc. 8th IEEE International Conference on Data Mining (ICDM 08)*, pages 995–1000, 2008.

[116] Jesse Read, Peter Reutemann, Bernhard Pfahringer, and Geoff Holmes. MEKA: A multi-label/multi-target extension to Weka. *Journal of Machine Learning Research*, 17(21):1–5, 2016.

[117] Hongyu Ren, Weihua Hu, and Jure Leskovec. Query2box: Reasoning over knowledge graphs in vector space using box embeddings, 2020.

[118] Hongyu Ren and Jure Leskovec. Beta embeddings for multi-hop logical reasoning in knowledge graphs, 2020.

[119] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletarì, et al. The future of digital health with federated learning. *npj Digital Medicine*, 3(1):119, 2020.

[120] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletarì, Holger R. Roth, Shadi Albarqouni, et al. The future of digital health with federated learning. *npj Digital Medicine*, 3:119, 2020.

[121] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods (2nd edition)*. Springer-Verlag, 2004.

[122] T Konstantin Rusch, Michael M Bronstein, and Siddhartha Mishra. A survey on oversmoothing in graph neural networks. *arXiv preprint arXiv:2303.10993*, 2023.

[123] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

[124] Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks, 2017.

[125] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021.

[126] Julio Silva, Takehiro Takahashi, Jamie Wood, Peiwen Lu, Alexandra Tabachnikova, Jeff R Gehlhausen, Kerrie Greene, Bornali Bhattacharjee, Valter Silva Monteiro, Carolina Lucas, et al. Sex differences in symptomatology and immune profiles of long covid. *medRxiv*, pages 2024–02, 2024.

[127] Rohan Singh, Monika Bapna, Abdul Rahman Diab, Emily S. Ruiz, and William Lotter. How ai is used in fda-authorized medical devices: a taxonomy across 1,016 authorizations. *npj Digital Medicine*, 2025.

[128] Sithursan Sivasubramaniam, Cedric E Osei-Akoto, Yi Zhang, Kurt Stockinger, and Jonathan Fürst. Sm3-text-to-query: Synthetic multi-model medical text-to-query benchmark. *Advances in Neural Information Processing Systems*, 37:88627–88663, 2024.

[129] http://www.iso-spread.it. The ISO-SPREAD clinical guideline group web page.

[130] Chetan L Srinidhi, Ozan Ciga, and Anne L Martel. Deep neural network models for computational histopathology: A survey. *Medical image analysis*, 67:101813, 2021.

[131] Thomas Stegmüller, Behzad Bozorgtabar, Antoine Spahr, and Jean-Philippe Thiran. Scorenet: Learning non-uniform attention and augmentation for transformer-based histopathological image classification. In *Proceedings of the IEEE/CVF winter Conference on applications of computer vision*, pages 6170–6179, 2023.

[132] Carole H Sudre, Benjamin Murray, Thomas Varsavsky, et al. Attributes and predictors of long COVID. *Nature Medicine*, 27(4):626–631, 2021.

[133] Yizhou Sun and Jiawei Han. *Mining heterogeneous information networks: principles and methodologies*. Morgan & Claypool Publishers, 2012.

[134] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. Rotate: Knowledge graph embedding by relational rotation in complex space. *arXiv preprint arXiv:1902.10197*, 2019.

[135] Huqun Suri, Qi Zhang, Wenhua Huo, Yan Liu, and Chunsheng Guan. Mediaqa: A question answering dataset on medical dialogues. *arXiv preprint arXiv:2108.08074*, 2021.

[136] Reed T. Sutton, David Pincock, Daniel C. Baumgart, Daniel C. Sadowski, Richard N. Fedorak, and Karen I. Kroeker. An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digital Medicine*, 3(17), 2020.

[137] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[138] Wenhao Tang, Sheng Huang, Xiaoxian Zhang, Fengtao Zhou, Yi Zhang, and Bo Liu. Multiple instance learning framework with masked hard instance mining for whole slide image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4078–4087, 2023.

[139] A.N Tarekegn, M. Giacobini, and K. Michalak. A review of methods for imbalanced multi-label classification. *Pattern Recognition*, 118:107965, 2021.

[140] N. Tax, I. Teinemaa, and S. J. van Zelst. An interdisciplinary comparison of sequence modeling methods for next-element prediction. *CoRR*, abs/1811.00062, 2018.

[141] Niek Tax, Ilya Verenich, Marcello La Rosa, and Marlon Dumas. Predictive business process monitoring with LSTM neural networks. In *Proceedings of the 29th International Conference on Advanced Information Systems Engineering (CAiSE)*, pages 477–492. Springer, 2017.

[142] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[143] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[144] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

[145] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[146] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[147] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *International conference on machine learning*, pages 2071–2080. PMLR, 2016.

[148] G. Tsoumakas, I. Katakis, and I. Vlahavas. Random k-labelsets for multi-label classification. *IEEE Transactions on Knowledge and Data Engineering*, 23:1079—1089, 2011.

[149] Alfred Ultsch and Jörn Lötsch. Augmenting small biomedical datasets using generative ai methods based on self-organizing neural networks. *Briefings in Bioinformatics*, 26(1):bbae640, 12 2024.

[150] Jeroen Van der Laak, Geert Litjens, and Francesco Ciompi. Deep learning in histopathology: the path to the clinic. *Nature medicine*, 27(5):775–784, 2021.

[151] Boudewijn van Dongen. Bpi challenge 2012, 2012.

[152] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[153] David Vilares and Carlos Gómez-Rodríguez. Head-qa: A healthcare dataset for complex reasoning. *arXiv preprint arXiv:1906.04701*, 2019.

[154] Luigi Vimercati, Luigi De Maria, Marco Quarato, Antonio Caputi, Loreto Gesualdo, Giovanni Migliore, Domenica Cavone, Stefania Sponselli, Antonella Pipoli, Francesco Inchingolo, et al. Association between long covid and overweight/obesity. *Journal of Clinical Medicine*, 10(18):4143, 2021.

[155] Laura Von Rueden, Sebastian Mayer, Katharina Beckh, Bogdan Georgiev, Sven Giesselbach, Raoul Heese, Birgit Kirsch, Julius Pfrommer, Annika Pick, Rajkumar Ramamurthy, et al. Informed machine learning–a taxonomy and survey of integrating prior knowledge into learning systems. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):614–633, 2021.

[156] Laura von Rueden, Sebastian Mayer, Katharina Beckh, Bogdan Georgiev, Sven Giesselbach, Raoul Heese, Birgit Kirsch, Julius Pfrommer, Annika Pick, Rajkumar Ramamurthy, Michal Walczak, Jochen Garcke, Christian Bauckhage, and Jannis Schuecker. Informed machine learning – a taxonomy and survey of integrating prior knowledge into learning systems. *IEEE Transactions on Knowledge and Data Engineering*, 35(1), 2023.

[157] Jun Wang, Yu Mao, Nan Guan, and Chun Jason Xue. Advances in multiple instance learning for whole slide image analysis: Techniques, challenges, and future directions. *arXiv preprint arXiv:2408.09476*, 2024.

[158] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Junzhou Huang, Wei Yang, and Xiao Han. Transpath: Transformer-based self-supervised learning for histopathological image classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, pages 186–195. Springer, 2021.

[159] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Junzhou Huang, Wei Yang, and Xiao Han. Transpath: Transformer-based self-supervised learning for histopathological image classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, pages 186–195. Springer, 2021.

[160] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research*, 10(2), 2009.

[161] David H Wolpert, William G Macready, et al. No free lunch theorems for search. Technical report, Technical Report SFI-TR-95-02-010, Santa Fe Institute, 1995.

[162] Jingyi Xu, Zilu Zhang, Tal Friedman, Yitao Liang, and Guy Van den Broeck. A semantic loss function for deep learning with symbolic knowledge. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

[163] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[164] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*, 2014.

[165] Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific data*, 10(1):586, 2023.

[166] Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

[167] Zheng Yuan, Yijia Liu, Chuanqi Tan, Songfang Huang, and Fei Huang. Improving biomedical pretrained language models with knowledge. *arXiv preprint arXiv:2104.10344*, 2021.

[168] Julio Zaragoza, Luis Sucar, Eduardo Morales, Concha Bielza, and Pedro Larranaga. Bayesian chain classifiers for multidimensional classification. In *Proc. 22nd International Joint Conference on Artificial Intelligence (IJCAI 11)*, pages 2192–2197, 2011.

[169] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18802–18812, 2022.

[170] Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, et al. Ultramedical: Building specialized generalists in biomedicine. *Advances in Neural Information Processing Systems*, 37:26045–26081, 2024.

[171] Yunlong Zhang, Honglin Li, Yunxuan Sun, Sunyi Zheng, Chenglu Zhu, and Lin Yang. Attention-challenging multiple instance learning for whole slide image classification. In *18th European Conference on Computer Vision*, pages 125–143. Springer, 2024.

[172] Jie Zheng. *Knowledge Graph*, pages 217–230. Springer Nature Singapore, Singapore, 2026.

[173] Yanning Zhou, Simon Graham, Navid Alemi Koohbanani, Muhammad Shaban, Pheng-Ann Heng, and Nasir Rajpoot. Cgc-net: Cell graph convolutional network for grading

of colorectal cancer histology images. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019.

[174] Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*, 2025.

Appendix A

HealthBranches: Supplementary Material

A.1 Dataset Details

In this section, the prompts used for data extraction using Gemini are presented.

Prompt used in the experiments to extract the graphical stream

For each pdf file that i send you will respond in the following way:

1. Put the content of the decision tree/graph in a valid JSON format. It has to be structured as a decision tree. If the branch of a decision is unclear or missing, get it from the text. The JSON must contain these keys:
   (a) node: containing the node name
   (b) content: a brief description of the node
   (c) children: a list of children node (can be optional for leaf nodes)
   (d) make a root node with the name of the condition
2. Make sure that the JSON is valid and well structured. Make sure that you don't get 'Invalid control character at line' errors.

The system prompt used to polish raw text extracted from the PDF (textual stream) is:

Prompt used to refine raw text extracted from a PDF

```
The uploaded PDF contains information and decisions to be made about a
specific disease/condition.
Each PDF consists of (sorted from the top of the first page): a title,
the authors, a brief description, observations (begin with a capital
letter followed by a period), a decision graph, a continuation of an
observation if it did not fit on the previous page, references.
In some documents there may be tables, ignore them.
```

While the single message with the single condition attached (PDF file):

Prompt to extract the condition from the PDF file

```
Analyze the PDF provided and extract the brief description and
observations. The title, authors, decision graph and references should
NOT be extracted. There must be ONLY text contained in the provided pdf
in the output.
```

Additionally, an example of path refinement using a real medical case is provided below.

System Prompt (path refinement)

```
Given a sequence of reasoning and associated condition, you have to
refine it to be uniform with medical terminology.
The meaning of the sequence must NOT change (at most, you can remove
superfluous information or sanitize the text). If the reasoning step is
poorly explained or ambiguous, refine it using the context and medical
knowledge.
The answer must be a sequence of reasoning with -> indicating the
transition between one step and the next.
Do not add any more text or reasoning in your answer, just the sequence.
```

Pleural Effusion Decision Paths before and after refinement process

Old Path

1. Imaging: PA & lateral CXR; consider decubitus film; consider ultrasound
2. Suspect CHF or viral pleurisy
3. No or atypical progression
4. Thoracentesis
5. Light's criteria: TP_f/TP_s > 0.5. LDH_f/LDH_s > 0.6. LDH_{pf} > 2/3 upper limit of normal.
6. Exudate
7. Cell count & differential
8. Mononuclear cell predominant. Viral infection. Chronic causes

New Path

1. Obtain radiological images: PA & lateral CXR; consider decubitus film or ultrasound
2. Suspect CHF or viral pleurisy
3. No or atypical progression
4. Perform thoracentesis
5. Apply Light's criteria: TP_f/TP_s > 0.5.LDH_f/LDH_s > 0.6. LDH_{pf} > 2/3 upper limit of normal.
6. Exudative effusion identified
7. Perform cell count & differential
8. Mononuclear cell predominance
9. Consider viral infection or chronic etiologies

System Prompt (question generation)

Given a sequence of reasoning and a text related to it about how to treat a symptom/condition, generate:
1. a question reflecting the reasoning of the sequence provided.
The question must include a clinical case, e.g: "A 67-year-old man is brought to the physician because of increasing forgetfulness, unsteadiness, and falls over the past year..."
2. A set of 5 possible answers (A,B,C,D,E). Should not be too long and should also reflect the sequence of reasoning. One of them must be the correct answer. The other answers need not be correct for the generated question but must be related to the topic of the question.
The sequence does NOT have to be explicit in both question and answers!
The correct option must be the same as the answer.
The output should be structured in the following format: ["question", "answer", "['Option A', 'Option B', 'Option C', 'Option D', 'Option E']", "letter of correct option"]

Algorithm 3 ExtractOption(answer)

Require: answer (string)
Ensure: Extracted option letter or None

ans ← Trim(answer)

for each rule $R_i \in \{R1, \ldots, R8\}$ **do**
    if ans matches R_i.pattern **then**
        ch ← Uppercase(captured letter)
        return ch
    end if
end for

return None

```
Do not generate any additional texts.
```

Generation question prompt

```
The reasoning sequence is as follows: {path}, the context associated is:
{text} and the symptom/condition to be treated is: {cond}.
```

A.2 Extract Option Rule Set

Below is the algorithm used for extracting an option from a string.

Rule definitions:

R1: `^[A-Ea-e]$` — exact single letter

R2: `^([A-Ea-e])\s` — letter + space at start

R3: `^([A-Ea-e])\` — letter + period at start

R4: `\(([A-Ea-e])\)` or
`\s([A-Ea-e])\)` — letter in parentheses or before ")"

R5: `^[A-E][A-Z]` — leading uppercase + noise

R6: `^[A-Ea-e][^A-Za-z]` — letter + non-letter at start

R7: any of

  - `the correct answer is [:\s]*\(?([A-Ea-e])\)?`
  - `the answer is [:\s]*\(?([A-Ea-e])\)?`
  - `option *\(?([A-Ea-e])\)?`
  - `answer: *\(?([A-Ea-e])\)?`

R8: `[:\.,]\s*([A-Ea-e])$` — trailing letter after punctuation

A.3 Q&A Validation

This section reports the prompt used to validate the questions that were incorrectly answered by both Llama3.3 (70 B) and Llama3.1 (405 B), along with the corresponding answers and path. In addition, the template used to submit the Q&A is presented using an example with the matching answer given by ChatGPT applied with web search and reasoning capabilities.

Prompt submitted to ChatGPT

```
You will now act as a reviewer for some questions and answers to a
quiz on the medical domain.  I will give you a question, some possible
answers, the supposed correct answer and a reasoning path that can
help to respond correctly which is structured exactly like a path with
arrows separating each step of the path.Your task is to check if the
question makes sense, if the reported correct option is the correct
option, if other options may be also correct and if the reasoning path
is correct.  You must check the soundness of the quiz using existing
medical literature and clinical guidelines.  If there are any issues
report them in an order manner and suggest which modification should be
done to the quiz in order to make it right.The modification can be of 4
types:

  1. The question is regarded as complex or ambiguous to answer.

  2. Change the correct answer (i.e.  change the correct option to
     another or adding/correcting an option).

  3. Change the question so that the resulting option is correct (i.e.
     if the question has some imprecise or missing information).
```

4. If these changes are affecting the reasoning path change it accordingly but keep the same schematic/path style (make sure to keep the same action like structure using ->).

After evaluating all of this, suggest the simplest modification between 0, 1, 2, or 3 to make the quiz sound. Give a brief motivation about it. Finally, output the quiz in the same template I sent it to you, but corrected with your best suggestion.

Example of Q&A using our template

Q&A template:

- Question: *A 34-year-old male is newly diagnosed with HIV. What is the correct order of actions to take after the diagnosis?*
- Options: *['Prescribe prophylactic antibiotics.', 'Refer the patient for immediate psychological counseling.', 'Perform a lumbar puncture.', 'Administer tetanus, diphtheria, and acellular pertussis boosters.', 'Initiate antiretroviral therapy immediately.']*
- Correct option: *Administer tetanus, diphtheria, and acellular pertussis boosters.*
- Reasoning path: *Initial steps for newly diagnosed HIV-positive patients -> Immunizations -> Tetanus, diphtheria, and acellular pertussis boosters.*

Reasoned about the consistency of the medical quiz for 27 seconds.

ChatGPT answer

Below is the evaluation and recommendation:
ISSUES IDENTIFIED:

1. Incorrect Correct Option:

- Observation: Current clinical guidelines for newly diagnosed HIV patients prioritize the initiation of antiretroviral therapy (ART) immediately. Immunizations (such as tetanus, diphtheria, and acellular pertussis boosters) are important, but they do not supersede the need for immediate ART initiation.
- Literature/Guidelines: The U.S. Department of Health and Human Services (DHHS) and WHO both recommend that ART should be started as soon as possible after diagnosis.
- Recommendation: Change the correct answer to "Initiate antiretroviral therapy immediately."

2. Question Wording Ambiguity:

- Observation: The phrasing "order of actions" implies a sequential management plan. However, the provided options mix interventions that occur at different stages.
- Recommendation: To avoid ambiguity, modify the question so that it asks for the "most immediate" or "priority" intervention after an HIV diagnosis.

3. Reasoning Path Issues:

- Observation: The current reasoning path directs the reader toward immunizations rather than immediate ART initiation.
- Recommendation: Update the reasoning path to emphasize that immediate ART is the priority, with immunizations coming later in the comprehensive care plan.

SIMPLEST MODIFICATION CHOSEN:

- Modification Type 1: Change the correct answer (and adjust the reasoning path accordingly).This is the simplest modification because it only requires updating the correct answer and its supporting reasoning without having to rework the entire quiz structure.

Figure A.1 shows the main correction types that were obtained in the Q&A refinement process.

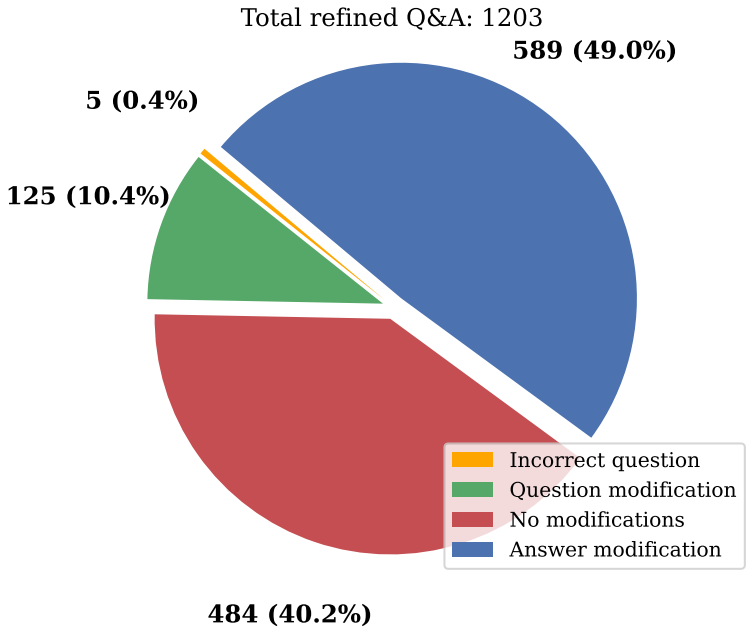


Figure A.1: Main types of corrections after the Q&A refinement process.

A.4 Quiz Inference

The prompts used during the evaluation phase of the quiz dataset, employing various models and setups, are presented below.

Zero shot prompt

```
The following is a multiple-choice question about how to manage a
patient affected by {condition}.
Reply ONLY with the letter (A,B,C,D,E) of the answer you think is
CORRECT, without any additional text!
Question: {question}
Options:
```

- A. {option1}
- B. {option2}
- C. {option3}
- D. {option4}
- E. {option5}

The answer MUST be a single letter.
Correct option:

Zero shot RAG prompt

The following is a multiple-choice question about how to manage a patient affected by {condition}.
Reply ONLY with the letter (A,B,C,D,E) of the answer you think is CORRECT, without any additional text!
Answer the question using the provided context.
Context: {context}
Question: {question}
Options:

- A. {option1}
- B. {option2}
- C. {option3}
- D. {option4}
- E. {option5}

The answer MUST be a single letter.
Correct option:

Zero shot Topline (path only) prompt

The following is a multiple-choice question about how to manage a patient affected by {condition}.
Reply ONLY with the letter (A,B,C,D,E) of the answer you think is CORRECT, without any additional text!
The choice of the correct answer should be based on the following sequence of reasoning: {path}.
Question: {question}

- A. {option1}
- B. {option2}
- C. {option3}

- D. {option4}
- E. {option5}

The answer MUST be a single letter.
Correct option:

System instruction for deepseek

Answer the question directly with the letter of the correct option only (e.g., A, B, C, D, or E).
Do NOT include any internal reasoning or chain-of-thought in your response.

A.5 Open Answer Inference

The prompts used during the evaluation phase of the open-ended dataset, employing various models and setups, are presented below.

Zero shot prompt

Answer the question about how to manage a patient affected by {condition}. The answer must:

- no longer than 100/150 characters;
- must contain a decision made on the decision-making process of a clinical case described in the question.

Question: {question}
Answer:

Zero shot RAG prompt

Answer the question about how to manage a patient affected by {condition}. The answer must:

- no longer than 100/150 characters;
- must contain a decision made on the decision-making process of a clinical case described in the question.

Answer the question using the provided context.

```
Context: {context}
Question: {question}
Answer:
```

Zero shot topline (path only) prompt

```
Answer the question about how to manage a patient affected by
{condition}. The answer must:

  • no longer than 100/150 characters;
  • must contain a decision made on the decision-making process of a
    clinical case described in the question.

The answer must be based on the following reasoning path: {path}.
Question: {question}
Answer:
```

System instruction for deepseek

```
Answer the question directly.
Do NOT include any internal reasoning or chain-of-thought in your
response.
```

A.6 Inference Times Details

The table below presents the average inference times (in seconds) required to generate a single response in both the quiz and open-ended answer settings.

A.7 Results for Different Modalities

Model performance by modality is shown in Table A.2.

A.8 LLM-as-a-Judge Implementation

The prompt used for the evaluation phase with the LLM-as-a-judge is reported below.

Model	Avg. Quiz	Avg. Open
gemma3:4b	0.528	0.738
mistral:7b	0.440	0.906
gemma:7b	0.449	0.834
qwen2.5:7b	0.413	0.691
llama2:7b	0.489	1.567
llama3.1:8b	0.547	1.012
deepseek-r1:8b	7.233	6.000
gemma2:9b	0.572	0.935
mistral-nemo:12b	0.574	0.746
phi4:14b	1.429	2.240

Table A.1: Average inference time per response (in seconds) for quiz and open-ended answer settings.

Model	Zero-Shot			Zero-Shot RAG			Topline Path			Topline Text			Topline All		
	Acc	Judge	Sim	Acc	Judge	Sim	Acc	Judge	Sim	Acc	Judge	Sim	Acc	Judge	Sim
Llama2 7b	19.71	5.14	0.314	34.36	4.89	0.317	28.25	8.15	0.385	40.09	4.82	0.320	55.45	6.68	0.360
Mistral 7b	51.93	6.46	0.339	61.14	6.57	0.346	87.69	9.00	0.437	62.86	6.59	0.356	82.40	8.21	0.420
Gemma 7b	57.27	5.24	0.305	62.22	5.73	0.325	89.37	8.67	0.407	61.90	6.08	0.340	80.56	8.23	0.407
Deepseek-r1 8b	61.33	5.46	0.321	63.97	6.02	0.337	88.48	8.41	0.418	60.94	6.16	0.334	76.15	7.43	0.377
Gemma3 4b	62.79	6.14	0.318	62.79	6.20	0.320	89.74	8.91	0.391	61.92	6.06	0.330	79.30	8.18	0.383
Llama3.1 8b	67.22	5.45	0.321	69.87	6.25	0.340	88.01	8.48	0.411	64.73	6.17	0.342	78.05	8.27	0.413
Nemo 12b	67.29	6.00	0.311	68.23	6.49	0.331	92.35	8.69	0.412	70.17	6.67	0.340	88.43	8.32	0.404
Qwen2.5 7b	68.55	6.20	0.335	70.02	6.59	0.352	92.84	9.00	0.456	71.55	6.73	0.361	89.66	8.55	0.437
Gemma2 9b	69.16	6.36	0.334	71.65	6.53	0.347	92.69	9.07	0.443	71.03	6.89	0.369	87.96	8.68	0.451
Phi4 14b	73.20	7.07	0.340	74.18	7.32	0.349	92.99	9.20	0.419	68.30	7.28	0.354	80.04	8.84	0.408
Llama3.3 70b	75.83	6.53	0.346	75.44	6.55	0.345	92.67	8.80	0.439	76.57	6.91	0.366	93.45	8.82	0.446

Table A.2: Model performance for each modality (Zero-shot, RAG, etc.) in terms of Accuracy, Judge score, and Similarity. Scores highlighted in red denote the best-performing model (rank 1), while those in blue indicate the second-best performance (rank 2) for the corresponding metric.

LLM-as-a-judge prompt

```
*** TASK: Based on the following task description and evaluation
criteria, generate a detailed Chain of Thought (CoT) that outlines
the necessary Evaluation Steps to assess the solution. The CoT should
clarify the reasoning process for each step of evaluation.
*** INPUT:
TASK INTRODUCTION: You are an evaluator for Q&A medical tasks. You will
evaluate the quality of answers to medical questions. You will have
some ground truth to help you evaluate.
You will be given 4 inputs:
```

- The question

- The ground truth answer
- The reasoning path that entails the answer
- The given answer which needs to be evaluated

The inputs will be given in csv format as a batch of 3 samples). The columns will contain:

- question
- reasoning path
- real answer
- predicted answer

The output must be a csv formatted including ALL the examples given in the input. But only of a column indicating the score
*** EVALUATION CRITERIA:
The evaluation criteria must take into account how much the given answer (input 4) is adherent to the reasoning path and to the ground truth answer.
The evaluation MUST not take into account the "style" of the answer but only the content. The score must be an integer between 0 to 10.
FINAL SCORE:
IF THE USER'S SCALE IS DIFFERENT FROM THE 0 TO 10 RANGE, RECALCULATE THE VALUE USING THIS SCALE.
SCORE VALUE MUST BE AN INTEGER.

A.9 Experimental and Hardware Setup

All experiments were run on a high-performance node equipped with an NVIDIA Quadro RTX 6000 GPU, an Intel Xeon Gold CPU, 192 GB RAM and Ubuntu Linux. Inference was performed using the Ollama framework (v0.6.3), with CUDA and cuDNN support. Statistics on the inference times are available in Appendix A.6. Two quantisation schemes were used: `Q_0`: 8-bit weight representation, no further compression (Gemma1/2, Mistral, Llama2, Mistral-nemo); `Q4_K_M`: 4-bit k-means quantisation (Gemma3, Qwen2.5, Deepseek-r1, Llama3.1, Phi4).

The experiments were completed on April 12th, 2025; therefore, models released after this date are not included in the benchmark evaluation.

A.10 Medical Validation Protocol

This section describes the structured protocol used by medical professionals to review and validate the Q&A content in the dataset. The following protocol was administered to students, medical specialists, and doctors using a simple and user-friendly web app to evaluate a sample of Q&A.

1. **Purpose.** To provide a structured framework for medical professionals to assess and ensure the accuracy, clarity, dependability, and identification of potential misinformation or factual inaccuracies in the Q&A content provided to patients and the general public.

2. **Scope.** Applicable to all medical professionals engaged in the evaluation process (physicians, specialists, medical researchers). Focuses on verifying:

   - Content accuracy
   - Adherence to current medical guidelines
   - Suitability for patient education

3. **Reviewer Qualifications.** Individuals must meet at least one of the following criteria:

   - Medical Student
   - Doctor / Physician
   - Medical Specialist

4. **Review Process.**

   (a) **Selection of Q&A Content.** A curated set of Q&A instances is chosen to span a broad range of medical topics.

   (b) **Evaluation Parameters and Scoring.** Each Q&A entry is scored on a 1–5 scale (1 = Poor, 5 = Excellent) according to:

   - *Question Medical Accuracy:* Alignment with current medical knowledge and reliable sources.
   - *Answer Medical Accuracy:* Correctness of the diagnosis or explanation. If incorrect, the reviewer selects the appropriate answer.

- *Path Medical Accuracy:* Correctness and completeness of the management steps outlined.

Total score ranges from 3 to 15. Entries scoring below a predefined threshold (i.e., 9/15) are flagged as incorrect.

(c) **Providing Feedback and Documentation.**

- Entries below threshold of 3/5: reviewers submit detailed feedback highlighting strengths and areas for revision.

(d) **Structured Review Form.** A web-app form is used to collect expert reviews:

- Each Q&A instance is displayed with fields for scores and free-text comments.
- The form auto-calculates total scores and flags entries below threshold.
- Aggregated results are exported for trend analysis and identification of common issues.

5. **Implementation of Revisions.**

- The Q&A development team reviews all feedback.
- Clarifications or further expert input are obtained as needed.
- A final verification pass is performed before publishing revisions.

6. **Ongoing Review and Quality Assurance.**

- Periodic re-assessment of Q&A content to ensure continued accuracy and relevance.
- Analysis of reviewer feedback trends to guide future improvements.
- Protocol updates to incorporate advances in medical science and best practices.

7. **Confidentiality and Ethical Considerations.**

- Reviewers must adhere to confidentiality agreements regarding unpublished material.
- All activities must comply with ethical standards and relevant regulatory requirements for medical information dissemination.

A.11 Web App Review Examples and Final Results

To illustrate the detailed corrections for Q&A that received at least one score of 3 or lower, two examples of the feedback provided by clinicians involved in this phase are displayed in Figure A.2. These cases highlight common sources of ambiguity: in the first example (bradycardia management), the reviewer emphasised the need to consider the possibility of a high potassium level that was not explicitly addressed in the original formulation; in the second example (platelet transfusion threshold), the reviewer noted that the question lacked sufficient clinical specificity (type of leukaemia), which could affect interpretation. The reviews generally highlighted instances where a diagnostic modality was unspecified, key labs were omitted, reasoning was unclear, specificity was insufficient, and terminology and guidelines were outdated. The validation was conducted meticulously, following the standardised protocol described in the previous section. These suggestions incorporate field-specific details to enhance the Q&A.

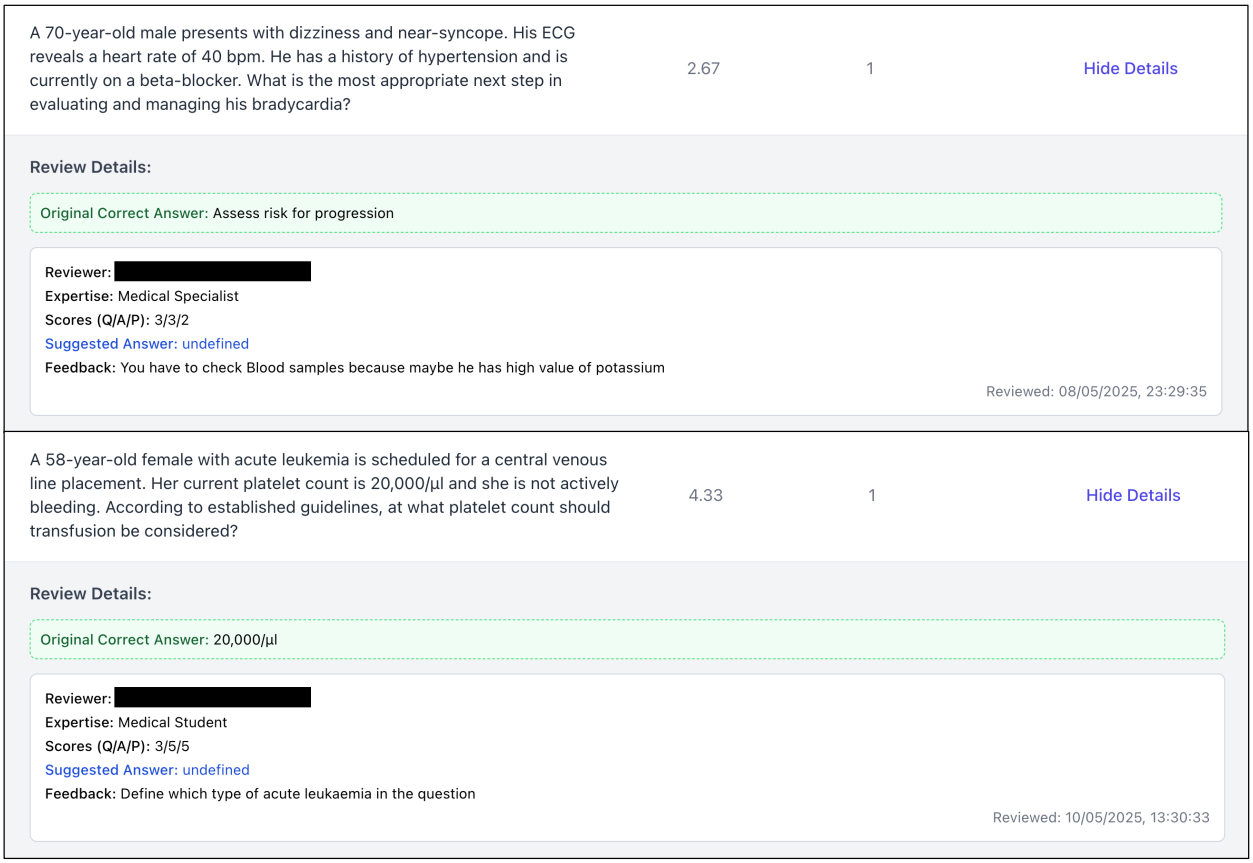


Figure A.2: Examples of reviews conducted by a medical specialist and a medical student receiving at least one score of 3 or lower.

Table A.3 provides an overview of the scores from the different reviewers for the Q&A

dataset.

Expertise	Question	Answer	Path	Mean	Reviewer	Reviews
Doctor/Physician	4.83/5	4.67/5	4.67/5	14.17/15	3	18
Medical Specialist	4.32/5	4.20/5	4.16/5	12.68/15	5	25
Medical Student	4.74/5	4.79/5	4.72/5	14.25/15	12	102
Total	13.89/15	13.66/15	13.55/15	13.97/15	22	145

Table A.3: Reviewer performance and feedback statistics by expertise level for randomly selected questions.

Category	Validated Questions	Total Questions
Hematology/Oncology	12	531
General Medicine	5	457
Infectious Diseases	20	441
Neurology	12	373
Women's Health	5	350
Gastrointestinal	9	338
Nephrology	6	279
Cardiology	23	278
Rheumatology	9	217
Behavioral Medicine	11	190
Endocrinology	10	183
Pharmacology	2	146
Pulmonary Disease	10	125
Emergency Medicine	0	54
Urology	5	45
Dermatology	2	35
Ocular	4	21
Total	145	4063

Table A.4: Coverage across HealthBranches categories.