REBASE: RELATIONAL BASIS LEARNING FOR HETEROGENEOUS MULTI-SOURCE TABULAR DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-source learning for tabular data is fundamentally challenged by *feature heterogeneity* across distributed sources, arising from both *structural heterogeneity* (differences in feature dimensionality and composition) and *semantic inconsistency* (variation in feature definitions, naming, and units). This challenge is particularly pronounced in healthcare, where clinical variables are collected under institution-specific protocols. While such surface-level discrepancies exist, the underlying disease mechanisms are largely shared across populations, manifesting through complex interactions among clinical features. To address this challenge, we propose REBASE, a framework that learns *reusable* relational bases that can be shared across heterogeneous tabular sources. REBASE represents each sample as a graph, where nodes and edges are grounded in large language model embeddings, mapping heterogeneous features into a unified semantic space. Over this space, our model learns a set of relational bases, each capturing a distinct relational pattern. It aligns each sample graph to these bases using Fused Gromov–Wasserstein distance, which jointly accounts for semantic similarity and structural correspondence. This alignment produces sample-specific coordinates that compose predictions from the bases. Experiments on three real-world benchmarks demonstrate that REBASE consistently outperforms existing multi-source tabular pretraining methods, in both robust generalization and few-shot adaptation to unseen, data-scarce domains, with its learned relational bases exhibiting clinically meaningful patterns across domains. Code is available here.

1 INTRODUCTION

Tabular data remains the most prevalent format in many real-world machine learning applications, including healthcare, finance, and manufacturing (Joseph et al., 2022; Johnson et al., 2016; Afonso et al., 2019). While deep learning has shown increasing promise for tabular prediction, it typically requires large amounts of labeled data to learn effective representations. In clinical settings, however, a given disease or clinical outcome is often studied across multiple institutions, each following its own (and possibly different) data collection protocol and contributing only a limited number of labeled samples (Wang et al., 2024). Jointly leveraging multiple sources has the potential to improve predictive performance beyond what any individual source can achieve, as disease progression is often governed by mechanisms that recur across patient populations, such as shared risk factors and their complex interactions (Yusuf et al., 2004).

In practice, however, integrating such tabular data is fundamentally challenged by *feature heterogeneity*, which manifests along two axes: *structural heterogeneity* arises when sources differ in feature dimensionality and composition, while *semantic inconsistency* occurs when the same underlying concept is represented using different definitions, names, or units. These discrepancies make direct feature-level alignment difficult and remain a fundamental obstacle to effectively transferring knowledge across heterogeneous tabular sources. Realizing the potential of multi-source learning therefore requires a representation that maps heterogeneous features into a shared, unified semantic space while preserving the relational structure among them. Such a representation can capture recurring relational patterns (Radenovic et al., 2022) across sources without being tied to any particular domain-specific feature set, providing a basis for transferring knowledge to new domains that share similar underlying mechanisms.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Contribution. To this end, we propose REBASE, a relational basis learning framework for heterogeneous multi-source tabular data that learns *reusable relational bases* that can be shared across domains, capable of resolving both structural heterogeneity and semantic inconsistency. Specifically, REBASE learns a set of *relational basis functions*, each representing a distinct relational pattern. These bases serve as *building blocks* for prediction: each sample is characterized by a set of sample-specific coordinates that govern how the bases are composed to provide its prediction. To enable this decomposition across heterogeneous feature spaces, REBASE represents each sample as a graph, with nodes corresponding to features and node and edge representations grounded in language model embeddings. This representation maps heterogeneous features into a unified semantic space while preserving the relational structure among them, allowing semantically related features to be meaningfully compared across sources. We then align each sample graph with the shared relational bases using Fused Gromov–Wasserstein (FGW) distance (Vayer et al., 2019), which jointly captures semantic similarity at the feature level and structural correspondence at the interaction level. This alignment provides the sample-specific coordinates to compose predictions from the bases. Since the learned relational knowledge is encoded in shared bases rather than tied to domain-specific feature sets, REBASE enables transfer across heterogeneous sources and adaptation to unseen target domains with limited labeled data. Experiments on three real-world datasets, each comprising multiple heterogeneous sources, show that REBASE generally outperforms existing multi-source tabular methods in both multi-source learning and few-shot adaptation to unseen target domains. Also, the learned relational bases exhibit clinically meaningful patterns across domains.

2 Related Works

Tabular Deep Learning. Gradient-boosted decision trees (GBDT) such as XGBoost (Chen & Guestrin, 2016), LightGBM (Ke et al., 2017), and CatBoost (Prokhorenkova et al., 2018) remain the predominant approach for tabular prediction, owing to their data efficiency and robustness (Grinsztajn et al., 2022; Shwartz-Ziv & Armon, 2022). Deep tabular models such as FT-Transformer (Gorishniy et al., 2021), SAINT (Somepalli et al., 2021), and TabNet (Arik & Pfister, 2021) have narrowed this gap but require substantial data to match GBDT performance. Moreover, they fix the input to a predefined feature schema, so any added or missing feature requires re-learning the embedding layer. Levin et al. (2022) relax this constraint by imputing pseudo-values for features missing in the target dataset, but this requires substantial feature overlap between the source and target datasets and a separate imputation model for each such pair.

Multi-Source Pretraining for Heterogeneous Tabular Data. Several approaches pretrain neural networks across tabular datasets with heterogeneous feature spaces. However, they do not explicitly represent relational patterns over features in a form reusable across sources with different feature compositions. TransTab (Wang & Sun, 2022) concatenates each feature name with its value and processes the resulting tokens as a flat sequence, making it difficult to explicitly capture relational structures among features. CARTE (Kim et al., 2024) represents each sample as a graph with language model embeddings, but connects each feature node only to a central node, limiting its ability to model richer relational structures among features. XTFormer (Chen et al., 2024) takes a complementary approach by composing predictions from shared basis functions; however, its basis coefficients are determined by a dataset-level context vector learned during training, rather than by the relational structure of individual samples. Consequently, the coefficients remain fixed across samples within a dataset, while a new dataset requires labeled data to learn its own context vector. In contrast, our framework explicitly captures relational patterns in a shared representation space and derives sample-specific coordinates over reusable relational bases, allowing these patterns to be recomposed across heterogeneous feature compositions and unseen domains.

While several existing frameworks capture generalizable tabular knowledge, they rely on target-domain labeled data to adapt to unseen datasets. XTab (Zhu et al., 2023) routes each table into a shared backbone using a dataset-specific featurizer, which must be explicitly trained on new labels prior to inference. In-context learning approaches like TabPFN-2.5 (Grinsztajn et al., 2025) and TabICLv2 (Qu et al., 2026) extract priors from large synthetic corpora but adapt to a new dataset by ingesting its labeled samples as context, making zero-shot prediction impossible. In contrast, REBASE composes predictions from shared relational bases using sample-specific coordinates. Because these coordinates are computed purely through graph alignment, REBASE requires neither dataset-specific featurizers nor labeled context, supporting zero-shot generalization to new domains.

3 PROBLEM FORMULATION

Problem Setup. We consider M heterogeneous domains $\{\mathcal{D}^{(m)}\}_{m=1}^{M}$, where each domain $\mathcal{D}^{(m)}$ is characterized by a feature schema $\mathbf{s}^{(m)} = (s_1^{(m)}, \ldots, s_{p_m}^{(m)})$ and comprises N_m labeled samples, $\{(\mathbf{x}_i^{(m)}, y_i^{(m)})\}_{i=1}^{N_m}$, with input features $\mathbf{x}_i^{(m)} \in \mathcal{X}^{(m)}$ and target label $y_i^{(m)} \in \mathcal{Y}$.[1] Here, p_m denotes the feature dimension of domain m, and each descriptor $s_j^{(m)}$ provides textual metadata – such as the name, description, and data type of the j-th feature – enabling semantic grounding via pre-trained language models. Each sample $\mathbf{x}_i^{(m)} = (x_{i,1}^{(m)}, \ldots, x_{i,p_m}^{(m)})$ lies in a mixed-type product space $\mathcal{X}^{(m)} = \mathcal{X}_1^{(m)} \times \cdots \times \mathcal{X}_{p_m}^{(m)}$ of numerical or categorical features. We use a shared label space $\mathcal{Y}$ across domains for a common clinical outcome, although the specific clinical endpoints may differ. Hereafter, we omit the sample index i and domain index (m) when clear from the context.

Basis-Function Decomposition. Our objective is to learn a set of reusable relational bases that can be shared across heterogeneous domains and support robust generalization and adaptation. A model trained on a single domain observes relational patterns only through its available feature set, providing a potentially incomplete view of the underlying input-output relationships. Particularly in data-scarce settings, the model may rely heavily on domain-specific patterns that do not transfer beyond the source domain. Multi-source learning can alleviate this issue by jointly exposing the model to diverse views of the underlying mechanisms, enabling it to learn relational patterns.

To realize the benefits of multi-source learning in a modular and transferable form, we conceptualize the input-output relationship through a basis-function decomposition. Formally, for any domain m, we represent the prediction model $f : \mathcal{X}^{(m)} \to \Delta^{C-1}$ as a convex combination of K bases:

$$f(\mathbf{x}) = \textstyle\sum_{k=1}^{K} c_k(\mathbf{x})\phi_k, \tag{1}$$

where $\Phi = \{\phi_k\}_{k=1}^{K}$ denotes the set of bases shared across all domains. Each basis $\phi_k \in \Delta^{C-1}$ captures a distinct relational pattern that recurs across domains, and $\mathbf{c}(\mathbf{x}) = (c_1(\mathbf{x}), \ldots, c_K(\mathbf{x})) \in \Delta^{K-1}$ denotes the sample-specific coordinate vector. Thus, $c_k(\mathbf{x}) \in [0, 1]$ determines the contribution of the k-th relational basis to the prediction for $\mathbf{x}$. This formulation is driven by the intuition that a given target label often stems from multiple distinct relational patterns (Radenovic et al., 2022). By extracting a set of relational bases across domains, we preserve these patterns in a reusable form, providing the flexibility to *recompose* them via sample-specific coefficients for any given instance.

Challenges. Realizing the basis-function decomposition in Eq. (1) is non-trivial in heterogeneous tabular settings due to fundamental discrepancies across sources. Specifically, for any two domains m and n, these discrepancies arise along two primary axes:

- **Structural Heterogeneity** (i.e., feature-space mismatch): Sources typically vary in dimensionality ($p_m \neq p_n$) and feature composition ($\mathcal{X}^{(m)} \neq \mathcal{X}^{(n)}$), often with minimal or even no direct overlap. Consequently, inputs from different sources cannot be projected into a shared functional space using standard architectural components.
- **Semantic Inconsistency** (i.e., feature-semantic mismatch): Even when features are conceptually related, their representations are often inconsistent due to differing naming conventions, definitions, or units. Furthermore, a single underlying concept may be encoded via entirely different sets of features, obscuring multi-source correspondence.

4 METHOD: REBASE

Overview. To address the challenges above, we propose REBASE, a relational basis learning framework for heterogeneous multi-source tabular data. REBASE jointly leverages multiple sources to learn a set of *reusable* relational bases that can be shared across domains, enabling it to recover relational patterns that may be only partially observed within any individual domain. REBASE consists of three key components:

- **Semantic Grounding via Language Models.** Feature schemas from different domains are mapped into a unified semantic space using language models. This enables semantically related features to be aligned despite potential differences in definition, name, and type.

[1] For concreteness, we focus on C-class classification with $\mathcal{Y} = \{1, \ldots, C\}$, though the formulation naturally extends to other common tabular tasks, such as regression.

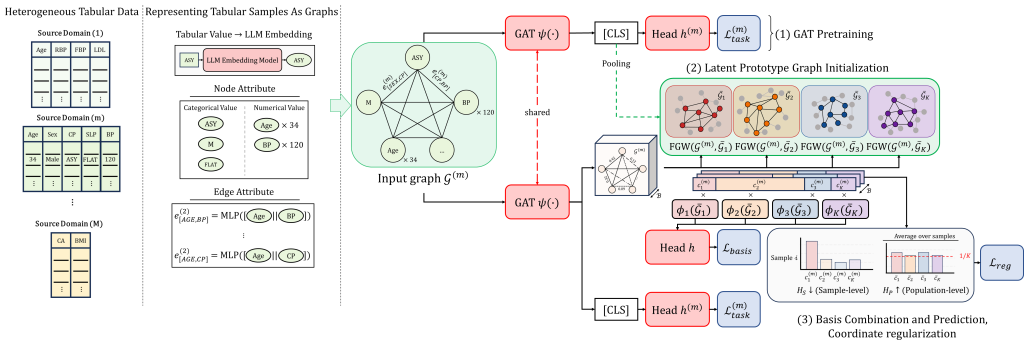

Figure 1: **Overall architecture of REBASE.** (1) GAT pretraining on all source domains with graph construction grounded in language model embeddings. (2) Latent prototype graph initialization. (3) FGW alignment and joint optimization of the relational bases.

- **Graph-based Encoding.** Each sample is represented as a graph over its features within this semantic space. A shared graph attention network (GAT) encoder processes these instance graphs and maps them into a common representation space, producing comparable representations despite heterogeneous feature compositions across sources.
- **Relational Basis Learning.** Within this shared representation space, REBASE learns K relational bases that capture distinct relational patterns observed across the source domains. These bases serve as reusable building blocks for prediction, with each instance graph aligned to the learned bases via FGW distance.

All three components are jointly optimized across multiple sources, enabling the basis-function decomposition in Eq. (1) to be realized across heterogeneous feature spaces.

4.1 Representing Tabular Samples as Graphs

To capture complex relational patterns arising from interactions among features, we represent each tabular instance as a fully connected graph over its constituent features. We ground both node and edge attributes in language model embeddings, enabling semantically related features to be meaningfully aligned despite differences across domains. This graph-based representation also naturally accommodates the varying feature dimensions and compositions inherent to heterogeneous multi-source data. Below, we describe the graph construction focusing on a single source domain and omit the domain index for notational simplicity.

Given a sample $\mathbf{x}$, we construct a graph whose nodes correspond to its p features. For each feature j, we obtain a semantic embedding $\mathbf{e}_j \in \mathbb{R}^d$ by encoding its descriptor s_j with a language model. For categorical variables, the node is initialized with the language model embedding of its specific realized value (e.g., encoding the string "Male" for a gender feature). For numerical variables, the node is initialized as $x_j \cdot \mathbf{e}_j$, where x_j is the normalized scalar value; scaling the semantic embedding by the continuous value preserves relative magnitudes across samples while retaining semantic identity (Wang & Sun, 2022; Kim et al., 2024). To capture structural relationships, the edge attribute between any pair of nodes j and j' is defined as $\mathbf{e}_{j,j'} = \texttt{MLP}_{\text{edge}}([\mathbf{e}_j \,\|\, \mathbf{e}_{j'}])$, which encodes their pairwise semantic affinity and supplies the feature-name context to categorical nodes.

To aggregate these feature-level representations into a sample-level embedding, regardless of the varying feature dimension p, we append a global [CLS] node to the graph. A shared GAT encoder processes this augmented graph using an asymmetric attention mechanism, where the [CLS] node attends to all feature nodes, but not vice versa. For each input $\mathbf{x}$, the GAT produces three outputs: a refined node embedding matrix $F = [\mathbf{f}_1, \ldots, \mathbf{f}_p]^\top \in \mathbb{R}^{p \times d}$, an attention-derived affinity matrix $A \in \mathbb{R}^{p \times p}$ capturing the interaction topology among the feature nodes, and a [CLS] embedding $\mathbf{z} \in \mathbb{R}^d$. These representations serve two distinct roles. The structural instance graph $\mathcal{G} = (F, A)$ is routed to the FGW alignment module to compute the basis coordinates (Section 4.2), while the global embedding $\mathbf{z}$ is passed to a source-specific prediction head $h^{(m)} : \mathbb{R}^d \to \Delta^{C-1}$ (Section 4.3).

4.2 Latent Prototype Graphs and FGW Alignment

Now, we realize K relational bases $\{\phi_k\}_{k=1}^{K}$ through *latent prototype graphs*, where each prototype graph represents one relational basis and captures a distinct relational pattern. Specifically, the k-th prototype graph consists of T latent feature nodes, $\bar{F}_k = [\bar{\mathbf{f}}_1, \ldots, \bar{\mathbf{f}}_T]^\top \in \mathbb{R}^{T\times d}$, initialized via K-means clustering over the source representations (detailed in Section 4.3), together with a learnable affinity matrix $\bar{A}_k \in \mathbb{R}^{T\times T}$ that captures the structural relationships among these latent nodes. The resulting prototype graphs, $\bar{\mathcal{G}}_1, \ldots, \bar{\mathcal{G}}_K$ where $\bar{\mathcal{G}}_k = (\bar{F}_k, \bar{A}_k)$, provide a compact and shared representation of the relational bases learned from the source domains.

To compose a prediction for a given sample according to Eq. (1), we need to determine how strongly each relational basis corresponds to the sample, such that prototypes more similar to the sample receive greater contributions to the prediction. Specifically, the coordinate vector $\mathbf{c}$ can be computed based on the FGW distance between $\mathcal{G}$ and each prototype graph $\bar{\mathcal{G}}_k$ as

$$\mathbf{c} = \mathrm{softmax}(-\mathtt{dist}/\tau) \in \mathbb{R}^K, \quad \text{where} \quad \mathtt{dist}_k = \mathrm{FGW}\big(\mathcal{G},\ \bar{\mathcal{G}}_k\big)\,. \tag{2}$$

Here, τ is a temperature hyperparameter. Thus, latent prototypes with smaller FGW distances receive larger coordinates and contribute more strongly to the prediction.

Utilizing the FGW distance is particularly suitable for our setting because instance and prototype graphs may differ in sizes, with p feature nodes and T latent nodes, respectively, while their correspondence depends jointly on node-level semantic similarity and edge-level structural correspondence. Its formal definition is provided in Appendix A. More importantly, the two components of FGW provide complementary signals for instance–prototype matching. The Gromov-Wasserstein component captures correspondence in relational structure but may align semantically different features that occupy similar structural roles. Conversely, the Wasserstein component leverages node attributes to identify semantically related features but does not account for their interactions. Combining the two therefore allows the alignment to jointly reflect feature semantics and relational structure. We further validate the importance of this joint formulation empirically in Appendix B.6.

4.3 Multi-Source Pretraining

With the graph representation (Section 4.1) and FGW-based alignment (Section 4.2) in place, we train the framework across heterogeneous source domains in two stages. We first pretrain the shared GAT encoder jointly across all source domains using source-specific prediction heads. We then use the resulting representations to initialize the latent prototype graphs and jointly optimize the encoder, prototypes, and prediction heads to learn the shared relational bases.

GAT Pretraining. We pretrain the shared GAT encoder ψ on all source domains with source-specific prediction heads $h^{(m)}$. For each domain m, the task loss is defined as:

$$\mathcal{L}_{\text{task}}^{(m)} = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}^{(m)}}\left[\ell\Big(h^{(m)}(\psi(\mathbf{x})),\ y\Big)\right]. \tag{3}$$

Once the shared GAT encoder is pretrained via Eq. (3), we extract and aggregate the `[CLS]` representations from all samples across all source domains into a unified global pool, $\mathcal{H} = \{\mathbf{z}_i^{(m)} \mid m \in \{1,\ldots,M\}, i \in \{1,\ldots,N_m\}\}$. Collecting these embeddings into a single cross-domain manifold ensures that the subsequent K-means clustering – used to initialize the latent prototype nodes – is performed jointly over the aggregated multi-source distribution, rather than being restricted to or biased by any individual domain.

Latent Prototype Graph Initialization. Each prototype graph represents one relational basis and is intended to capture a distinct relational pattern. A flat clustering of $\mathcal{H}$ into $K \times T$ clusters does not explicitly enforce this structure, as centroids corresponding to different relational patterns may be arbitrarily assigned to the same prototype graph. We therefore adopt a hierarchical clustering strategy (Ye et al., 2024; Shin et al., 2020; 2024). We first apply K-means to $\mathcal{H}$ to obtain K coarse clusters, each corresponding to a prototype graph, and then independently perform T-means within each coarse cluster to obtain T sub-centroids. These sub-centroids initialize the latent feature nodes $\bar{F}_k$ of the corresponding prototype graph. We initialize the learnable affinity matrix $\bar{A}_k$ based on the pairwise distances between the sub-centroids, assigning higher initial affinity to more closely located nodes. Implementation details are provided in Appendix C.1.

Basis Combination and Prediction. To derive the output for each relational basis ϕ_k, we process the corresponding prototype graph $\bar{\mathcal{G}}_k$ through a lightweight two-layer Graph Neural Network (GNN). Thus, we can rewrite the basis-function decomposition from Eq. (1) as follows:

$$f(\mathbf{x}) = \sum_{k=1}^{K} c_k(\mathbf{x})\, \phi_k(\bar{\mathcal{G}}_k). \tag{4}$$

Once the prototype graphs are initialized, we optimize the prototype graphs and the GNN parameters using $\mathcal{L}_{\text{basis}} = \sum_{m=1}^{M} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}^{(m)}}[\ell(f(\mathbf{x}), y)]$. Since the sample-specific coordinates depend on the instance graphs produced by the shared GAT encoder ψ, we also continue to update ψ during this stage using the task loss $\mathcal{L}_{\text{task}}^{(m)}$ in Eq. (3) across the domains, but with a smaller learning rate. This preserves the pretrained feature representations while allowing the GAT encoder to adapt to the learned prototype graphs.

Coordinate Regularization. Training with the basis-combination loss alone may lead to *coordinate collapse*, where a small subset of relational bases receives most of the coordinate mass across samples, leaving the remaining bases underutilized. To encourage diverse and meaningful use of the relational bases, we regularize the coordinate entropy at both the sample and population levels:

$$H_S = \mathbb{E}_{\mathbf{x}}\left[\frac{-\sum_{k=1}^{K} c_k(\mathbf{x}) \log c_k(\mathbf{x})}{\log K}\right], \quad H_P = \frac{-\sum_{k=1}^{K} \bar{c}_k \log \bar{c}_k}{\log K} \quad \text{where } \bar{c}_k = \mathbb{E}_{\mathbf{x}}\left[c_k(\mathbf{x})\right]. \tag{5}$$

Minimizing $\mathcal{L}_{\text{reg}} = H_S + (1 - H_P)^2$ encourages the coordinates of each sample to concentrate on a few bases, i.e., low H_S, while keeping every basis in use across the population, i.e., low $(1 - H_P)^2$.

Overall, REBASE is jointly trained across all source domains using the following objective:

$$\mathcal{L}_{\text{pre}} = \sum_{m=1}^{M} \mathcal{L}_{\text{task}}^{(m)} + \mathcal{L}_{\text{basis}} + \alpha_{\text{reg}} \mathcal{L}_{\text{reg}}. \tag{6}$$

4.4 Inference and Adaptation

Since the relational bases are shared across domains, they are kept fixed after pretraining. At inference, a sample is first represented as a graph as in Section 4.1, and its coordinates are computed by FGW alignment to these bases; the prediction is then composed from the bases as in Eq. (4). For adaptation to an unseen target domain, a target head is attached to this prediction and trained.

5 Experiments

To evaluate REBASE, we examine three questions aligned with our formulation: (i) *does learning across heterogeneous sources improve predictive performance on each domain*; (ii) *do the learned relational bases capture clinically meaningful patterns consistently across domains*; (iii) *do they transfer to domains unseen during pretraining*. Accordingly, we consider the following scenarios:

- **Scenario 1. Multi-source integration** (Section 5.1, 5.2): We pretrain each model on all source domains and fine-tune on each one. We additionally analyze the learned relational bases of REBASE to examine the patterns captured through multi-source training.
- **Scenario 2. Few-shot adaptation** (Section 5.3, 5.4): We hold out each domain in turn as an unseen target domain and evaluate adaptation using only a few labeled samples. We also compare REBASE with tabular foundation models under the same setting.

Baselines. We compare REBASE against classical models (Logistic Regression, Random Forest, XGBoost (Chen & Guestrin, 2016), CatBoost (Prokhorenkova et al., 2018)), deep tabular models (MLP, TabM (Gorishniy et al., 2025)), an LLM-based method (TabLLM (Hegselmann et al., 2023)), and multi-source pretrained models (TransTab (Wang & Sun, 2022), XTab (Zhu et al., 2023), XTFormer (Chen et al., 2024), CARTE (Kim et al., 2024)). We implement all baselines using their official codebases; implementation details are provided in Appendix C.2.

Datasets. We construct three multi-source benchmarks from publicly available clinical datasets to evaluate multi-source tabular learning under both *structural heterogeneity* and *semantic inconsistency*. Detailed preprocessing and benchmark construction for each dataset are provided in Appendices D.1, D.2, and D.3.

Table 1: Multi-source performance across all source domains. Best in **bold**, second best underlined.

	Cardiovascular						ICU						Breast Cancer			
Model	MED	CDD	HD	ERB	SAH	HF	MIMIC	eICU	HiRiD	SUPPORT	SicDB	Zigong	C1	C2	C3	C4
LR	89.9±5.7	89.4±3.7	91.9±3.3	93.8±3.3	75.6±3.6	85.1±3.0	86.0±0.1	80.4±0.2	85.7±0.4	83.9±0.3	88.3±0.8	72.9±2.6	71.0±1.4	69.5±6.7	75.1±6.3	69.5±3.2
RF	94.6±3.3	90.8±2.7	88.9±2.7	94.6±1.2	71.8±4.7	84.2±3.1	86.1±0.2	80.4±0.1	86.4±0.5	85.9±0.2	88.7±0.5	73.8±3.0	69.6±3.9	69.1±3.6	74.7±4.1	62.9±5.8
XGB	94.3±4.2	92.6±2.3	91.8±3.2	96.6±2.0	75.5±3.7	88.3±2.8	86.9±0.1	82.2±0.2	85.6±0.5	85.2±0.3	87.5±0.5	72.3±3.1	64.7±4.0	62.3±6.5	66.8±4.7	62.5±4.8
CatBoost	95.3±2.7	93.3±2.1	90.4±3.0	95.5±3.0	75.1±3.8	88.2±2.9	87.4±0.1	82.6±0.2	86.7±0.5	86.0±1.5	89.3±0.3	76.0±5.3	69.8±2.6	67.8±4.4	75.2±4.6	68.6±5.8
MLP	92.2±1.9	90.9±4.1	87.5±3.2	93.1±3.9	72.9±4.8	83.3±3.0	83.0±0.2	79.3±0.1	81.7±1.3	82.2±0.7	82.7±0.8	68.0±3.6	61.9±6.1	67.4±6.9	73.3±5.4	68.4±5.0
TabM	96.2±1.4	93.9±3.0	89.2±3.3	95.3±1.7	76.8±3.1	88.1±2.9	87.0±0.3	82.9±0.3	87.1±0.9	86.4±1.4	87.8±0.8	77.3±3.6	71.4±2.2	70.5±3.4	76.3±5.1	68.2±4.7
CARTE	96.5±2.2	97.4±2.7	95.3±2.1	97.6±2.2	74.2±4.1	91.5±2.9	88.4±0.3	82.6±0.4	88.9±0.7	87.5±1.1	**91.8**±0.6	79.2±4.1	72.2±2.9	72.8±3.2	76.5±4.3	70.5±1.9
XTFormer	94.3±2.5	93.9±1.8	91.9±2.4	93.9±2.6	75.3±4.4	91.6±1.9	87.6±0.4	84.6±0.3	88.3±0.8	86.4±1.6	91.2±0.6	78.5±4.0	73.9±1.6	68.9±6.2	73.5±5.1	67.4±3.2
TabLLM	92.3±2.4	90.4±5.2	88.7±2.7	96.2±1.7	**79.3**±4.0	86.8±3.0	87.0±0.5	83.5±0.3	88.7±0.9	85.1±2.5	86.5±1.5	77.1±4.8	70.9±3.5	68.9±5.7	74.2±5.2	68.3±4.1
XTab	90.8±2.6	92.4±3.3	89.3±2.2	95.3±2.3	72.5±4.2	89.3±2.4	88.0±0.4	82.1±0.6	86.6±1.7	84.8±1.4	88.8±1.1	69.4±6.0	71.8±2.1	71.0±3.3	76.1±4.9	68.9±3.8
TransTab	89.8±2.8	93.4±5.6	87.3±2.6	95.1±3.4	71.5±4.0	87.9±3.0	87.6±0.5	82.2±0.3	85.7±0.7	86.2±1.1	87.2±0.7	81.3±2.5	71.2±2.2	69.7±1.5	77.3±5.9	69.1±5.7
Ours	**97.5**±2.8	**98.7**±1.6	**96.7**±1.4	**97.7**±1.6	76.2±3.2	**92.2**±2.7	**91.1**±0.3	**86.0**±0.2	**90.7**±0.8	**90.8**±1.2	89.4±0.7	**83.7**±2.7	**74.8**±2.4	**75.2**±2.8	**78.9**±4.0	**71.3**±2.1

- **Cardiovascular Benchmark:** We curate six publicly available cardiovascular cohorts collected from diverse institutions and repositories. Following clinical guidelines (Virani et al., 2023), we organize these cohorts into four clinical categories: (i) coronary artery disease (CAD), including **MED** (*Medicaldataset*), **CDD** (*Cardiovascular Disease Dataset*), and **HD** (*Heart*); (ii) cardiovascular disease (CVD), including **ERB** (*Erbil Cardiovascular*); (iii) coronary heart disease (CHD), including **SAH** (*Cardio SAheart*); and (iv) heart failure (HF), including **HF** (*Heart Failure Clinical Records*).
- **ICU Benchmark:** We construct cohorts from six ICU databases – **MIMIC-IV** (Johnson et al., 2020), **eICU** (Pollard et al., 2018), **HiRID** (Yèche et al., 2021), **SICdb** (Rodemund et al., 2023), **SUPPORT** (Knaus et al., 1995), and **Zigong** (Xu et al., 2022a). These datasets are used for two distinct tasks: mortality prediction (classification) and kidney function estimation (regression).
- **Breast Cancer Benchmark:** We construct cohorts from four breast cancer adjuvant clinical trials in Project Data Sphere: **C1** (*NCT00041119*), **C2** (*NCT00174655*), **C3** (*NCT00312208*), and **C4** (*NCT00688740*). All four cohorts share disease-free survival as a common endpoint but differ in their recorded features.

5.1 Performance on Multi-Source Tabular Learning

We first evaluate whether multi-source pretraining improves predictive performance on each source domain. Each domain is divided into 60% training, 20% validation, and 20% test sets using stratified random splits; REBASE is pretrained jointly on the training splits of all source domains and then fine-tuned and evaluated on each one individually. During pretraining, source domains are iterated sequentially so that each receives the same number of updates regardless of its size. All implementation details are provided in Appendix C.1.

Table 1 summarizes the classification results on the three benchmarks, with full results, including AUPRC, reported in Appendix E. Regression results on the ICU benchmark are provided in Table 8. While several existing multi-source methods are competitive with tree-based ensembles, REBASE achieves the best average rank on all three benchmarks.

A complexity analysis with measured runtimes is given in Appendix B.1, where REBASE remains within the same order of magnitude as the baselines.

5.2 Interpreting the Shared Relational Bases

Coordinate Similarity across Source Domains. We first examine how consistently each basis is associated with positive outcomes across source domains. For each basis k in domain m, we compute a weighted positive rate, $\rho_k^{(m)} = \sum_{i=1}^{N_m} c_{i,k} \cdot y_i / \sum_{i=1}^{N_m} c_{i,k}$. We then evaluate the alignment of the resulting domain-level vectors, $\boldsymbol{\rho}^{(m)} = [\rho_1^{(m)}, \ldots, \rho_K^{(m)}]$, using cosine similarity across all source domain pairs in Figure 2. The five intensive-care databases exhibit high similarity, indicating that the learned bases show consistent outcome associations across these domains.

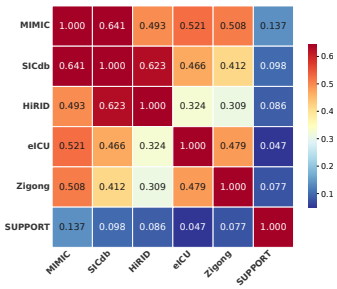

Figure 2: Pairwise similarity among source domains (ICU).

Residual variation within this group is consistent with differences in the variables recorded by each source beyond those shared across all databases. For example, ventilation and hemodynamic monitoring are available only in HiRID and SICdb, and source pairs that do not share these measurements exhibit lower similarity. Conversely, SUPPORT shows the lowest similarity, likely reflecting its distinct cohort composition: patients were enrolled with advanced stages of one of nine illnesses rather than at intensive-care admission (The SUPPORT Principal Investigators, 1995). Consistent with these similarities, source-exclusion analysis shows that removing a given source domain degrades performance on those most similar to it (Appendix B.3).

Clinical Interpretation of the Learned Bases. We next examine what each basis captures. We assign each sample to the basis with its largest coordinate and, within each source domain, compare the samples assigned to it (high-risk, for bases associated with positive outcomes) against the remaining samples (rest). Each variable is tested with the Mann-Whitney U test, and those separating the two groups with a non-negligible effect size are grouped by clinical category under the organ dysfunction and acute physiology scores (Vincent et al., 1996; Knaus et al., 1985) (Table 18). Figure 3 shows the t-SNE of all ICU test coordinates, with one representative variable per category for four bases; full profiles are in Appendix D.2. In each of these bases, the variables of a clinical category differ across sources in name or definition, yet move together toward high-risk:

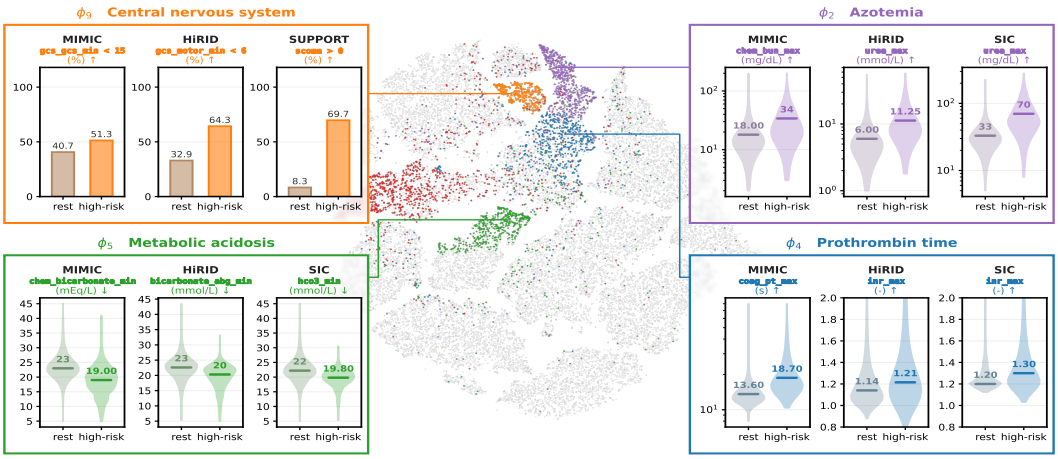

Figure 3: t-SNE of ICU test coordinates. Positive samples are colored by their dominant basis and negative samples are shown in gray; arrows mark the clinically adverse direction.

- **Cardiorenal Hypoperfusion (ϕ_2):** ϕ_2 captures variables consistent with a pathway in which arterial hypotension lowers renal perfusion and glomerular filtration, driving the accumulation of nitrogenous waste and impairing tissue perfusion (Prescott et al., 2026; Kidney Disease: Improving Global Outcomes (KDIGO) Acute Kidney Injury Work Group, 2012). MIMIC-IV reports the nitrogen fraction as blood urea nitrogen, whereas HiRID and SICdb report the whole molecule as urea, recorded in mg/dL, mmol/L and mg/dL respectively.
- **Hemostatic Consumption (ϕ_4):** ϕ_4 captures variables consistent with the depletion of platelets and clotting factors under systemic activation of coagulation, which prolongs the time the blood takes to clot (Taylor et al., 2001; Vincent et al., 1996; Singer et al., 2016). MIMIC-IV records this time directly in seconds, whereas HiRID and SICdb record it as a ratio normalized against a laboratory standard.
- **Metabolic Acidosis (ϕ_5):** ϕ_5 captures variables consistent with the buffering of an acid load, in which bicarbonate is consumed, arterial pH falls, and ventilation rises to clear carbon dioxide in compensation (Knaus et al., 1985). Bicarbonate is recorded under three different names: as a serum chemistry result in MIMIC-IV, as a blood-gas measurement in HiRID, and by its chemical formula in SICdb.
- **Neurologic Depression (ϕ_9):** ϕ_9 captures variables consistent with impaired consciousness and the diminished physiologic reserve of older patients (Vincent et al., 1996; Knaus et al., 1985).

Consciousness is assessed on a different scale in each source: MIMIC-IV records the total Glasgow Coma Scale, HiRID the motor subscale, and SUPPORT a separate coma score.

Table 2: Few-shot performance on the three benchmarks. Best in **bold**, second best underlined.

	Cardiovascular						ICU						Breast Cancer					
Model	0	4	8	16	32	64	0	4	8	16	32	64	0	4	8	16	32	64
LR	–	64.6	67.1	74.6	78.5	83.5	–	63.4	65.3	69.0	72.4	73.4	–	59.8	63.4	65.7	69.0	70.0
RF	–	68.5	74.4	78.1	80.8	83.9	–	62.9	65.2	69.7	72.1	72.9	–	60.5	63.2	65.3	67.9	68.9
XGB	–	55.9	65.8	77.3	81.9	84.4	–	59.3	65.4	70.2	73.6	74.9	–	53.6	60.7	62.3	64.2	63.8
CatBoost	–	69.8	76.6	81.1	84.2	86.7	–	64.5	69.8	71.2	75.0	76.8	–	58.9	61.6	63.8	66.8	68.8
MLP	–	65.3	68.5	74.6	79.0	81.8	–	62.2	63.0	69.0	71.7	73.7	–	58.9	62.1	64.2	66.4	66.4
TabM	–	75.1	77.2	80.5	83.7	86.4	–	66.0	70.0	72.9	75.5	77.1	–	62.8	65.2	67.4	69.3	71.5
CARTE	–	76.4	79.3	82.6	86.1	88.6	–	67.5	69.3	73.2	77.0	79.6	–	66.1	67.5	68.4	69.8	71.2
XTFormer	–	75.1	77.6	80.7	84.1	86.7	–	67.0	69.9	74.6	77.4	78.7	–	63.4	64.8	66.8	68.3	70.3
TabLLM	71.4	73.5	76.3	79.9	82.2	84.3	63.0	63.3	66.9	70.1	73.9	75.0	62.2	64.1	65.8	67.2	68.2	69.5
XTab	–	71.9	76.1	78.8	82.2	85.0	–	65.2	68.0	71.8	74.2	75.8	–	61.9	64.0	67.4	69.8	71.4
TransTab	68.9	71.6	75.6	78.0	81.5	83.7	62.2	62.3	67.1	71.2	74.6	75.2	57.1	60.1	63.9	65.7	68.8	70.8
Ours	**76.8**	**78.0**	**81.2**	**84.4**	**87.3**	**89.7**	**70.9**	**71.7**	**73.5**	**75.9**	**78.9**	**81.1**	**63.1**	**67.1**	**70.5**	**71.9**	**72.8**	**73.0**

5.3 Few-Shot Adaptation to Unseen Target Domains

We evaluate whether the learned relational bases transfer to unseen target domains with only a few labels. We adopt a leave-one-domain-out protocol on each benchmark: each domain is treated in turn as the target, while the model is pretrained on the remaining source domains, allocating 85% of each source for training and 15% for validation. For each target, 20% of the data is held out as a fixed test set, and from the remaining split we sample k labels per class for adaptation, without a target validation set. All results are averaged over 25 support-set resamples, with 5 random seeds. (Appendix C.1). Since the alignment to the fixed bases needs no target labels, adaptation extends to the zero-shot case (Section 4.4). Table 2 reports, for each benchmark, the mean AUROC across target domains: REBASE achieves the highest AUROC in nearly all settings, including the zero-shot case where most baselines are not applicable. Critical difference diagrams and full results are provided in Appendix E.

5.4 Comparison with Tabular Foundation Models

We compare REBASE against tabular foundation models on the ICU benchmark, reporting AUPRC alongside AUROC since mortality is a minority outcome (Saito & Rehmsmeier, 2015). Unlike TabPFN-2.5 (Grinsztajn et al., 2025) and TabICLv2 (Qu et al., 2025), which adapt by ingesting labeled samples as context, REBASE predicts zero-shot and exceeds their 4-shot results in both metrics (Table 3). As more labels become available, these models overtake it in AUROC, whereas REBASE remains higher in AUPRC at every shot size. This is consistent with Section 5.2: the bases are already associated with high-risk outcomes, while the labeled context of these models contains only a few positive cases.

Table 3: Few-shot performance on the ICU benchmark against tabular foundation models.

Model	0	4	8	16	32	64
AUROC						
TabPFN-2.5	–	68.4	72.73	76.31	**79.61**	**83.75**
TabICLv2	–	68.56	71.89	**76.67**	79.27	82.03
Ours	**70.89**	**71.65**	**73.52**	75.90	78.85	81.07
AUPRC						
TabPFN-v2.5	–	21.26	23.73	28.68	33.93	37.55
TabICLv2	–	20.87	22.33	27.64	31.34	38.32
Ours	**23.20**	**24.42**	**26.62**	**30.90**	**34.88**	**39.00**

6 Conclusion

We present REBASE, a relational basis learning framework for heterogeneous multi-source tabular data. REBASE learns relational bases shared across domains and composes each prediction from these bases through sample-specific coordinates, obtained by Fused Gromov–Wasserstein alignment between a sample graph and the bases. Across three clinical benchmarks, it generally outperforms existing multi-source methods in both multi-source learning and few-shot adaptation, and its learned bases exhibit clinically meaningful patterns across domains.

AI USE STATEMENT

In this work, we used generative AI tools to edit the manuscript for readability and to proofread it for consistency across the text, tables, and figures. We did not use generative AI tools for designing the methodology or experiments, implementing the method, or interpreting the results, and the remaining disclosure items, such as generating synthetic data, formulating mathematical claims, and assisting with proofs, are not applicable to this work. All AI-suggested edits were reviewed by the authors. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only publicly available, de-identified clinical datasets. The intensive care databases and the breast cancer trial data were accessed under the terms of their respective data use agreements (PhysioNet and Project Data Sphere), and no attempt was made to re-identify individuals. The proposed framework is intended for research purposes and has not been validated for clinical deployment.

REPRODUCIBILITY STATEMENT

Code is available at the anonymous link given in Appendix C.1, together with implementation details and hyperparameter search spaces (Table 12). Baseline configurations are in Appendix C.2, and the cohort construction, preprocessing, and exclusion criteria for all benchmarks in Appendix D.

REFERENCES

Bruno Afonso, Luckeciano Melo, Willian Oliveira, Samuel Sousa, and Lilian Berton. Housing prices prediction with a deep learning and random forest ensemble. In *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pp. 389–400. SBC, 2019.

Sercan Ö Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 6679–6687, 2021.

Donna K Arnett, Roger S Blumenthal, Michelle A Albert, et al. 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease: A report of the American College of Cardiology/American Heart Association task force on clinical practice guidelines. *Circulation*, 140(11): e596–e646, 2019.

Hajira Basit, Beenish S. Bhutta, and Martin R. Huecker. Myocardial infarction serum markers. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2023. Updated 2023 Aug 14. Available from: https://www.ncbi.nlm.nih.gov/books/NBK532966/.

Lukas Biewald. Experiment tracking with weights and biases, 2020. URL https://www.wandb.com/. Software available from wandb.com.

Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? In *International Conference on Learning Representations (ICLR)*, 2022.

Jintai Chen, Zhen Lin, Qiyuan Chen, and Jimeng Sun. Cross-table pretraining towards a universal function space for heterogeneous tabular data. *arXiv preprint arXiv:2406.00281*, 2024.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Davide Chicco and Giuseppe Jurman. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Medical Informatics and Decision Making*, 20(16), 2020. doi: 10.1186/s12911-020-1023-5.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

Bhanu Prakash Doppala, Debnath Bhattacharyya, Midhunchakkaravarthy Janarthanan, and Namkyun Baik. A reliable machine intelligence model for accurate identification of cardiovascular diseases using ensemble techniques. *Journal of Healthcare Engineering*, 2022:2585235, 2022. doi: 10.1155/2022/2585235.

Anca Florian, Massimo Slavich, Pier Giorgio Masci, Stefan Janssens, and Jan Bogaert. Electrocardiographic Q-wave "remodeling" in reperfused ST-segment elevation myocardial infarction: Validation study with CMR. *JACC: Cardiovascular Imaging*, 5(10):1003–1013, 2012. doi: 10.1016/j.jcmg.2012.02.018.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Yury Gorishniy, Akim Kotelnikov, and Artem Babenko. TabM: Advancing tabular deep learning with parameter-efficient ensembling. In *International Conference on Learning Representations*, 2025.

L Grinsztajn, K Flöge, O Key, F Birkel, P Jund, B Roof, B Jäger, D Safaric, S Alessi, A Hayler, et al. Tabpfn v2.5: Advancing the state of the art in tabular foundation models. *arXiv preprint arXiv:2511.08667*, 2025.

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*, volume 35, 2022.

Ibrahim Hamarash, Shwan Amen, Banan Rasool, and Hangaw Ahmed. Erbil cardiovascular health dataset (ECHD), 2024.

Trevor Hastie and Robert Tibshirani. Non-parametric logistic and proportional odds regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 36(3):260–276, 1987.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. Tabllm: Few-shot classification of tabular data with large language models. In *International conference on artificial intelligence and statistics*, pp. 5549–5581. PMLR, 2023.

Paul A Heidenreich, Biykem Bozkurt, David Aguilar, et al. 2022 AHA/ACC/HFSA guideline for the management of heart failure: A report of the American College of Cardiology/American Heart Association joint committee on clinical practice guidelines. *Circulation*, 145(18):e895–e1032, 2022.

Andras Janosi, William Steinbrunn, Matthias Pfisterer, and Robert Detrano. Heart disease. UCI Machine Learning Repository, 1988.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV, 2020. PhysioNet. Available from: https://physionet.org/content/mimiciv/1.0/.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Lionel P Joseph, Erica A Joseph, and Ramendra Prasad. Explainable diabetes classification using hybrid bayesian-optimized tabnet architecture. *Computers in Biology and Medicine*, 151:106178, 2022.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

Kidney Disease: Improving Global Outcomes (KDIGO) Acute Kidney Injury Work Group. KDIGO clinical practice guideline for acute kidney injury. *Kidney International Supplements*, 2(1):1–138, 2012.

Myung Jun Kim, Léo Grinsztajn, and Gaël Varoquaux. Carte: pretraining and transfer for tabular learning. *arXiv preprint arXiv:2402.16785*, 2024.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

William A. Knaus, Elizabeth A. Draper, Douglas P. Wagner, and Jack E. Zimmerman. APACHE II: a severity of disease classification system. *Critical Care Medicine*, 13(10):818–829, 1985.

William A Knaus, Frank E Harrell, Joanne Lynn, Lee Goldman, Russell S Phillips, Alfred F Connors, Neal V Dawson, William J Fulkerson, Robert M Califf, Norman Desbiens, et al. The support prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Annals of internal medicine*, 122(3):191–203, 1995.

Jennifer S Lawton, Jacqueline E Tamis-Holland, Sripal Bangalore, et al. 2021 ACC/AHA/SCAI guideline for coronary artery revascularization: A report of the American College of Cardiology/American Heart Association joint committee on clinical practice guidelines. *Circulation*, 145(3):e18–e114, 2022.

Roman Levin, Valeriia Cherepanova, Avi Schwarzschild, Arpit Bansal, C Bayan Bruss, Tom Goldstein, Andrew Gordon Wilson, and Micah Goldblum. Transfer learning with deep tabular models. *arXiv preprint arXiv:2206.15306*, 2022.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965, 2022.

Theresa A McDonagh, Marco Metra, Marianna Adamo, et al. 2021 ESC guidelines for the diagnosis and treatment of acute and chronic heart failure. *European Heart Journal*, 42(36):3599–3726, 2021.

Facundo Mémoli. Gromov-Wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 11(4):417–487, 2011.

Latha P. Palaniappan, Norrina B. Allen, Zaid I. Almarzooq, Cheryl A.M. Anderson, Pankaj Arora, Christy L. Avery, Carissa M. Baker-Smith, Nisha Bansal, Maria E. Currie, Rebecca S. Earlie, Wenjun Fan, Jessica L. Fetterman, Bethany Barone Gibbs, Debra G. Heard, Swapnil Hiremath, Haoyun Hong, Hyacinth I. Hyacinth, Chinwe Ibeh, Tian Jiang, Michelle C. Johansen, Dhruv S. Kazi, Darae Ko, Tak W. Kwan, Michelle H. Leppert, Yilun Li, Jared W. Magnani, Karlyn A. Martin, Seth S. Martin, Erin D. Michos, Michael E. Mussolino, Oluwabunmi Ogungbe, Nisha I. Parikh, Marco V. Perez, Sarah M. Perman, Ashish Sarraju, Nilay S. Shah, Mellanie V. Springer, Marie-Pierre St-Onge, Evan L. Thacker, Seda Tierney, Sarah M. Urbut, Harriette G.C. Van Spall, Jenifer H. Voeks, Seamus P. Whelton, Sally S. Wong, Juan Zhao, Sadiya S. Khan, on behalf of the American Heart Association Council on Epidemiology, Prevention Statistics Committee, and Stroke Statistics Committee. 2026 heart disease and stroke statistics: A report of us and global data from the american heart association. *Circulation*, 153(9):e275–e906, 2026. doi: 10.1161/CIR.0000000000001412. URL https://www.ahajournals.org/doi/abs/10.1161/CIR.0000000000001412.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-Wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pp. 2664–2672. PMLR, 2016.

Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):180178, 2018.

Hallie C. Prescott, Massimo Antonelli, Waleed Alhazzani, Morten Hylander Møller, Fayez Alshamsi, Luciano C. P. Azevedo, Emilie Belley-Cote, Jan De Waele, Lennie Derde, Joanna C. Dionne, Laura Evans, Hayley B. Gershengorn, Carol L. Hodgson, Kimia Honarmand, Jozef Kesecioglu, Lauralyn McIntyre, Mervyn Mer, Mark E. Nunnally, Simon J. W. Oczkowski, Bram Rochwerg, Andrew Rhodes, and Craig M. Coopersmith. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2026. *Intensive Care Medicine*, 52 (5):863–936, 2026. doi: 10.1007/s00134-026-08361-1.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31, 2018.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. Tabicl: A tabular foundation model for in-context learning on large data. *arXiv preprint arXiv:2502.05564*, 2025.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICLv2: A better, faster, scalable, and open tabular foundation model. *arXiv preprint arXiv:2602.11139*, 2026.

Filip Radenovic, Abhimanyu Dubey, and Dhruv Mahajan. Neural basis models for interpretability. *Advances in Neural Information Processing Systems*, 35:8414–8426, 2022.

Tarik A. Rashid and Bryar Hassan. Heart attack dataset, 2022.

Niklas Rodemund, Bernhard Wernly, Christian Jung, Crispiana Cozowicz, and Andreas Koköfer. The salzburg intensive care database (sicdb): an openly available critical care dataset. *Intensive care medicine*, 49(6):700–702, 2023.

J. E. Rossouw, J. P. du Plessis, A. J. S. Benadé, P. C. J. Jordaan, J. P. Kotzé, P. L. Jooste, and J. J. Ferreira. Coronary risk factor screening in three rural communities: The CORIS baseline study. *South African Medical Journal*, 64:430–436, 1983.

Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, 10(3):e0118432, 2015.

Rai Dilawar Shahjehan, Sanjeev Sharma, Ehab Dababneh, and Beenish S. Bhutta. Coronary artery disease. In *StatPearls [Internet]*. StatPearls Publishing, Treasure Island (FL), 2024. Updated 2024 Oct 9. Available from: https://www.ncbi.nlm.nih.gov/books/NBK564304/.

Yunsheng Shi, Zhengjie Huang, Shikun Feng, Hui Zhong, Wenjin Wang, and Yu Sun. Masked label prediction: Unified message passing model for semi-supervised classification. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

Su-Jin Shin, Kyungwoo Song, and Il-Chul Moon. Hierarchically clustered representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5776–5783, 2020.

Yong-Min Shin, Sun-Woo Kim, and Won-Yong Shin. Page: Prototype-based model-level explanations for graph neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 46(10):6559–6576, 2024.

Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.

Mervyn Singer, Clifford S. Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R. Bernard, Jean-Daniel Chiche, Craig M. Coopersmith, Richard S. Hotchkiss, Mitchell M. Levy, John C. Marshall, Greg S. Martin, Steven M. Opal, Gordon D. Rubenfeld, Tom van der Poll, Jean-Louis Vincent, and Derek C. Angus. The third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA*, 315(8):801–810, 2016.

Gowthami Somepalli, Micah Goldblum, Avi Schwarzschild, C Bayan Bruss, and Tom Goldstein. Saint: Improved neural networks for tabular data via row attention and contrastive pre-training. *arXiv preprint arXiv:2106.01342*, 2021.

Fletcher B. Taylor, Cheng-Hock Toh, W. Keith Hoots, Hideo Wada, and Marcel Levi. Towards definition, clinical and laboratory criteria, and a scoring system for disseminated intravascular coagulation. *Thrombosis and Haemostasis*, 86(5):1327–1330, 2001.

The SUPPORT Principal Investigators. A controlled trial to improve care for seriously ill hospitalized patients: The study to understand prognoses and preferences for outcomes and risks of treatments (SUPPORT). *JAMA*, 274(20):1591–1598, 1995.

Kristian Thygesen, Joseph S. Alpert, Allan S. Jaffe, et al. Fourth universal definition of myocardial infarction (2018). *Circulation*, 138(20):e618–e651, 2018. doi: 10.1161/CIR.0000000000000617.

Robin Van De Water, Hendrik Schmidt, Paul Elbers, Patrick Thoral, Bert Arnrich, and Patrick Rockenschaub. Yet another icu benchmark: A flexible multi-center framework for clinical ml. *arXiv preprint arXiv:2306.05109*, 2023.

Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. Fused Gromov-Wasserstein distance for structured objects: theoretical foundations and mathematical properties. *arXiv preprint arXiv:1811.02834*, 2018.

Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. Optimal transport for structured data with application on graphs. In *International Conference on Machine Learning*, pp. 6275–6284, 2019.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, et al. Embeddinggemma: Powerful and lightweight text representations. *arXiv preprint arXiv:2509.20354*, 2025.

Jean-Louis Vincent, Rui Moreno, Jukka Takala, Sheila Willatts, Arnaldo De Mendonça, Hajo Bruining, C. K. Reinhart, Peter M. Suter, and Lambertius G. Thijs. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. *Intensive Care Medicine*, 22(7): 707–710, 1996.

Salim S Virani, L Kristin Newby, Suzanne V Arnold, et al. 2023 AHA/ACC/ACCP/ASPC/NLA/PCNA guideline for the management of patients with chronic coronary disease: A report of the American Heart Association/American College of Cardiology joint committee on clinical practice guidelines. *Circulation*, 148(9):e9–e119, 2023.

Zifeng Wang and Jimeng Sun. Transtab: Learning transferable tabular transformers across tables. *Advances in Neural Information Processing Systems*, 35:2902–2915, 2022.

Zifeng Wang, Chufan Gao, Cao Xiao, and Jimeng Sun. Meditab: Scaling medical tabular data predictors via data consolidation, enrichment, and refinement. In *International Joint Conference on Artificial Intelligence*, 2024.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations (ICLR)*, 2019.

Ping Xu, Lin Chen, and Zhongheng Zhang. Critical care database comprising patients with infection at Zigong Fourth People's Hospital. *PhysioNet*, June 2022a. doi: 10.13026/xpt9-z726. URL https://doi.org/10.13026/xpt9-z726. Version 1.1.

Ping Xu, Lin Chen, Yongchao Zhu, Shuling Yu, Ronglin Chen, Wenlong Huang, Feng Wu, and Zhongheng Zhang. Critical care database comprising patients with infection. *Frontiers in Public Health*, 10:852410, 2022b. doi: 10.3389/fpubh.2022.852410.

Hangting Ye, Wei Fan, Xiaozhuang Song, Shun Zheng, He Zhao, Dandan Guo, and Yi Chang. Ptarl: Prototype-based tabular representation learning via space calibration. *arXiv preprint arXiv:2407.05364*, 2024.

Hugo Yèche, Rita Kuznetsova, Marc Zimmermann, Matthias Hüser, Xinrui Lyu, Martin Faltys, and Gunnar Rätsch. Hirid-icu-benchmark–a comprehensive machine learning benchmark on high-resolution icu data. *arXiv preprint arXiv:2111.08536*, 2021.

Salim Yusuf, Steven Hawken, Stephanie Ôunpuu, Tony Dans, Alvaro Avezum, Fernando Lanas, Matthew McQueen, Andrzej Budaj, Prem Pais, John Varigos, and Liu Lisheng. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. *The Lancet*, 364(9438):937–952, 2004.

Bingzhao Zhu, Xingjian Shi, Nick Erickson, Mu Li, George Karypis, and Mahsa Shoaran. Xtab: Cross-table pretraining for tabular transformers. *arXiv preprint arXiv:2305.06090*, 2023.

APPENDIX

A Preliminaries **17**

A.1 Graph represented as metric measure spaces. 17

A.2 Gromov-Wasserstein (GW) Distance. 17

A.3 Fused Gromov-Wasserstein (FGW) Distance. 17

B Further Results **17**

B.1 Time Complexity Analysis 17

B.2 Full Results for Multi-Source Pretraining across Sampling Ratios 19

B.3 Source Exclusion Analysis 20

B.4 Graph Size without Imputation 21

B.5 Performance results on Kidney Function regression 22

B.6 Sensitivity to model components 22

B.7 Necessity of the FGW Alignment 23

B.8 Contribution of Individual Bases 24

C Experimental Setup - Details **24**

C.1 Implementation Details 24

C.2 Baseline Implementation Details 25

D Dataset details **27**

D.1 Cardiovascular datasets 27

D.2 ICU datasets 32

D.3 Breast Cancer datasets 37

E Full Results **38**

E.1 Full Basis Profiles 38

E.2 Critical Difference diagram 41

E.3 Full Finetuning Results 42

E.4 Full Cardiovascular Few-shot Results. 43

E.5 Full ICU Few-shot Results 46

E.6 Full Breast Cancer Few-shot Results 49

A PRELIMINARIES

A.1 GRAPH REPRESENTED AS METRIC MEASURE SPACES.

An attributed graph is represented as a triplet $\mathcal{G} = (C, F, h)$, where $C \in [0,1]^{n\times n}$ is a symmetric matrix encoding pairwise structural relations between n nodes (e.g., adjacency or affinity matrix), $F = [F(i)]_{i=1}^{n} \in \mathbb{R}^{n\times d}$ is the node feature matrix whose rows correspond to d-dimensional feature vectors, and $h = n^{-1}\mathbf{1}_n$ represents a uniform discrete probability distribution over nodes. Each vertex i is therefore characterized by its structural role within C and its feature embedding $F(i)$. When only structural information is considered, a graph is represented as a pair $\mathcal{G} = (C, h)$.

A.2 GROMOV-WASSERSTEIN (GW) DISTANCE.

The Gromov-Wasserstein distance (Mémoli, 2011) compares two graphs based solely on their internal structural relations. Given two graphs $\mathcal{G}_1 = (C_1, n_1^{-1}\mathbf{1}_{n_1})$ and $\mathcal{G}_2 = (C_2, n_2^{-1}\mathbf{1}_{n_2})$, the GW distance is defined as:

$$\mathrm{GW}(\mathcal{G}_1, \mathcal{G}_2) = \min_{\pi\in\Pi_{n_1,n_2}} \sum_{i,k,j,l} \left(C_1(i,k) - C_2(j,l)\right)^2 \pi_{i,j}\pi_{k,l},$$

where $\Pi_{n_1,n_2} = \{\pi \in \mathbb{R}_+^{n_1\times n_2} \mid \pi\mathbf{1}_{n_2} = n_1^{-1}\mathbf{1}_{n_1},\ \pi^\top\mathbf{1}_{n_1} = n_2^{-1}\mathbf{1}_{n_2}\}$ denotes the set of admissible transport plans. The optimal coupling π^* aligns nodes across graphs by preserving pairwise structural relations: if nodes i and k are similarly related in $\mathcal{G}_1$, they are mapped to nodes j and l that are similarly related in $\mathcal{G}_2$. However, GW considers only structural geometry and ignores node attributes entirely, making it unable to distinguish nodes that occupy similar structural roles but carry different semantics.

A.3 FUSED GROMOV-WASSERSTEIN (FGW) DISTANCE.

The FGW distance (Vayer et al., 2018) extends GW to attributed graphs by jointly considering structural relations and node attributes. Given two attributed graphs $\mathcal{G}_1 = (C_1, F_1, n_1^{-1}\mathbf{1}_{n_1})$ and $\mathcal{G}_2 = (C_2, F_2, n_2^{-1}\mathbf{1}_{n_2})$, the FGW distance is defined as:

$$\mathrm{FGW}(\mathcal{G}_1, \mathcal{G}_2) = \min_{\pi\in\Pi_{n_1,n_2}} \sum_{i,k,j,l} \Big[(1-\beta)\left\|F_1(i) - F_2(j)\right\|_{\mathbb{R}^d}^2 + \beta\left(C_1(i,k) - C_2(j,l)\right)^2\Big]\pi_{i,j}\pi_{k,l},$$

where Π_{n_1,n_2} is as defined above and $\beta \in [0,1]$ controls the trade-off between preserving the structural consistency (C_1, C_2) and matching node features (F_1, F_2). FGW defines an optimal transport–based distance between metric measure graphs by aligning pairwise structural relations and node attributes, enabling comparison between graphs with different numbers of nodes and heterogeneous domains. The optimal coupling π^* yields a soft alignment that minimizes discrepancies across structural and feature domains. In practice, solving the FGW optimization exactly is computationally expensive. We adopt entropic regularization, which adds a negative entropy term $-\epsilon H(\pi)$ to the objective, making the optimization differentiable and efficiently solvable via the Sinkhorn algorithm (Cuturi, 2013).

B FURTHER RESULTS

B.1 TIME COMPLEXITY ANALYSIS

We analyze the time complexity of REBASE against existing tabular models, with particular attention to the Fused Gromov–Wasserstein alignment. Runtime is measured on the ICU benchmark, which includes the widest schema in our study (HiRID, $p_m = 283$).

Graph Encoding. A fully-connected instance graph over p_m feature nodes has p_m^2 edges, so one GAT layer costs $O(p_m^2 d + p_m d^2)$: the first term covers attention coefficients and message aggregation over all edges, the second covers node projections. This matches the per-layer cost of self-attention over p_m tokens, the form used by transformer-based tabular models.

FGW Alignment. Let $\pi \in \mathbb{R}^{p_m \times T}$ denote the transport plan between the instance graph $\mathcal{G}$ and a prototype graph $\bar{\mathcal{G}}_k$. Each solver iteration requires the tensor product $\mathcal{L} \otimes \pi$ with $\mathcal{L}_{iujv} = L\big(A(i,u), \bar{A}_k(j,v)\big)$, which costs $O(p_m^2 T^2)$ if $\mathcal{L}$ is formed explicitly. Our implementation adopts the squared Euclidean loss, for which $L(a,b) = a^2 + b^2 - 2ab$ separates and the product admits the factorization (Peyré et al., 2016)

$$\mathcal{L} \otimes \pi \;=\; c \;-\; h_1(A)\, \pi\, h_2(\bar{A}_k)^\top, \tag{7}$$

where $h_1(a) = a$ and $h_2(b) = 2b$ are applied element-wise, and c does not depend on π and is computed once per alignment. Each iteration thus reduces to two matrix products of cost $O(p_m^2 T)$ and $O(p_m T^2)$. Over K prototype graphs and n_{it} iterations, alignment costs $O\big(K\, n_{\text{it}}(p_m^2 T + p_m T^2)\big)$, whose dependence on p_m matches that of self-attention. Since $T \ll p_m$, the first term dominates: the cost is linear in T but carries a factor of K, so enlarging the prototype graphs is far cheaper than adding more of them. Two further properties keep this term small. The operands are compact ($K, T \leq 12$ in all our experiments), and the solver is warm-started with the converged plan from the previous step rather than a uniform plan. Both this warm start and the squared Euclidean loss are options of the POT (Python Optimal Transport) package.[2] Table 4 shows that REBASE remains comparable in magnitude to transformer-based baselines in per-step cost and memory, despite the fully-connected representation and the added alignment.

Table 4: Per-step time and peak memory on the ICU benchmark, averaged over the six sources, measured as in Table 6.

Model	Batch	s/step	Peak mem.
XTab	128	0.048	1.96 GB
CARTE	128	0.067	1.77 GB
TransTab	128	0.035	1.56 GB
XTFormer	128	0.297	10.03 GB
REBASE	128	0.110	3.17 GB

Table 5: Scaling with the number of prototype graphs K and prototype nodes T on HiRID, batch size 128. The other quantity is held at 12. AUROC is the zero-shot result on HiRID, averaged over five runs.

	K (with T=12)			T (with K=12)		
Value	AUROC	s/step	GB	AUROC	s/step	GB
4	71.49 ± 6.6	0.093	1.74	76.15 ± 4.6	0.173	3.51
8	76.88 ± 8.0	0.152	2.62	79.75 ± 4.1	0.181	3.52
12	80.42 ± 4.2	0.190	3.54	80.42 ± 4.2	0.190	3.54
16	79.58 ± 3.9	0.233	4.46	81.23 ± 3.3	0.195	3.56
32	79.18 ± 2.8	0.394	8.13	82.17 ± 2.5	0.203	3.72

Effect of the Number of Prototype Graphs and Nodes. Enlarging K raises the per-step cost almost linearly, whereas enlarging T leaves it nearly unchanged, so the two quantities may differ in how much accuracy they can buy. We fix one at 12 and vary the other over $\{4, 8, 12, 16, 32\}$ on HiRID. As shown in Table 5, both are lowest at 4 and approach their level at 12 from 8 onward. With too few prototype graphs, the coordinate space may be too low-dimensional to represent diverse relational patterns; with too few nodes, each prototype graph may lack sufficient structure for FGW alignment to establish informative correspondences. Above 12, accuracy tends to decline with larger K, possibly because the supervisory signal is spread across more bases, while it tends to increase with larger T; in both directions, in both directions, however, the differences fall within one standard deviation across seeds. Enlarging K beyond this range therefore adds cost without a clear gain in accuracy, whereas T can be enlarged at almost no additional cost. The configurations selected on the source validation split lie in $K, T \in \{8, 12\}$ (Table 12).

[2] https://pythonot.github.io/gen_modules/ot.batch.html#ot.batch.solve_gromov_batch

Table 6: Per-step time and peak memory on HiRID ($p_m = 283$, the widest schema in our benchmark) as a function of batch size B.

	B=4		B=8		B=16		B=32		B=64		B=128		B=256	
Model	s/step	GB	s/step	GB	s/step	GB	s/step	GB	s/step	GB	s/step	GB	s/step	GB
CARTE	0.031	1.23	0.034	1.28	0.034	1.37	0.035	1.55	0.053	1.91	0.084	2.62	0.166	4.05
XTFormer	0.032	0.58	0.034	1.10	0.062	2.13	0.121	4.17	0.239	8.27	0.486	16.46	0.988	32.84
XTab	0.031	0.12	0.032	0.22	0.035	0.42	0.039	0.82	0.045	1.61	0.066	3.21	0.107	6.39
TransTab	0.015	0.16	0.016	0.25	0.015	0.43	0.021	0.79	0.032	1.51	0.056	2.95	0.104	5.83
REBASE	0.117	0.22	0.147	0.33	0.146	0.54	0.149	0.97	0.142	1.83	0.190	3.54	0.298	9.51

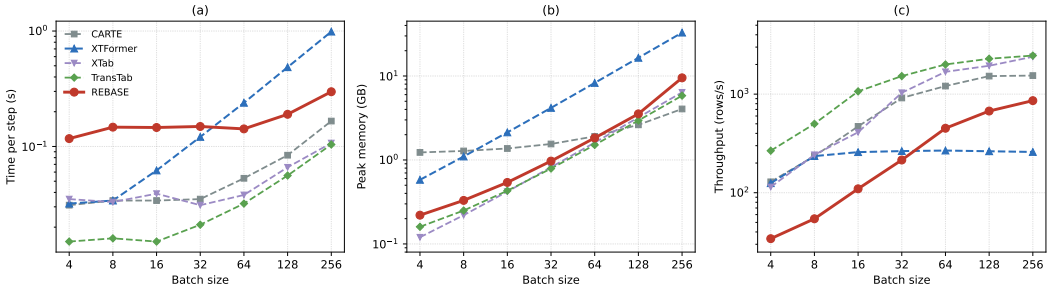

Figure 4: Batch-size scaling on HiRID ($p_m = 283$): (a) per-step training time, (b) peak GPU memory, and (c) training throughput, on log scales. The per-step cost of REBASE grows slowly with the batch size, while the throughput of XTFormer saturates beyond B=16.

Batch-Size Scaling. Table 6 and Figure 4 vary the batch size B from 4 to 256 on HiRID, the widest schema in our benchmark. REBASE aligns each sample against a fixed set of K prototype graphs with a fixed number of solver iterations, so the alignment acts as a per-step cost that grows slowly with B: from 0.117 to 0.298 s/step over a 64× larger batch, with throughput rising from 34 to 859 rows/s. The other baselines instead scale more steeply with B, so the gap narrows as the batch grows; at B=256, REBASE is within 2.9× of TransTab and 1.8× of CARTE. The overhead of alignment is therefore most visible at small batch sizes.

XTFormer. XTFormer is the only baseline whose cost exceeds that of REBASE beyond B=32, reaching 0.486 s/step and 16.46 GB at B=128, and its cost grows 31× across the sweep. Since its coordinates are assigned to each feature rather than shared across the table (Section 2), every term of its cost scales with B, and its throughput saturates near 260 rows/s beyond B=16 (Figure 4c).

B.2 Full Results for Multi-Source Pretraining across Sampling Ratios

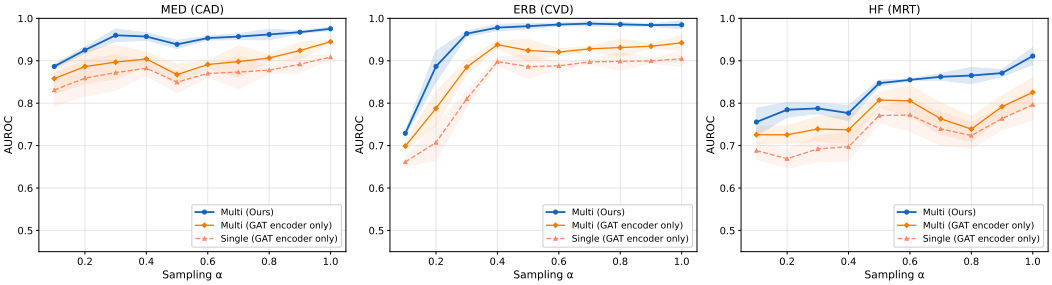

Figure 5: Model performance vs sampling ratio α (MED (CAD), ERB (CVD), HF (MRT)).

Does architectural decomposition drive performance? To determine whether the gains from multi-source learning stem from our basis-function decomposition rather than a simple increase in training data volume, we conduct an ablation study across three configurations: (i) **Single (GAT)**,

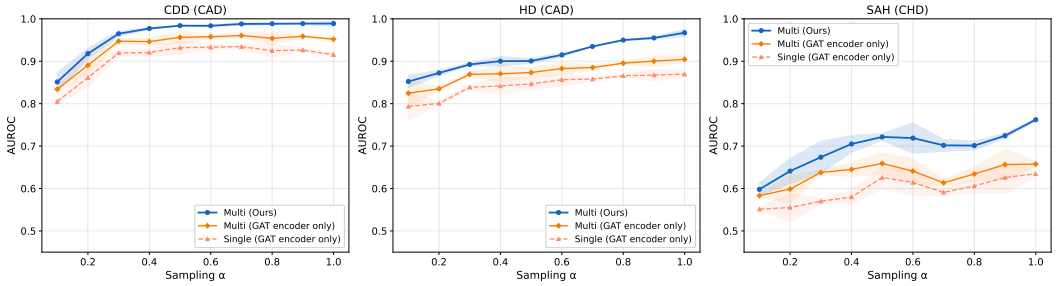

Figure 6: Model performance vs sampling ratio α (CDD (CAD), HD (CAD), SAH (CHD)).

a GAT trained on each source independently; (ii) **Multi (GAT)**, a GAT trained on the union of all sources without basis decomposition; and (iii) REBASE (Ours). We evaluate these configurations by varying the sampling ratio $\alpha \in \{0.1, 0.2, \ldots, 1.0\}$ of the total available training pool, where each domain contributes $\lfloor \alpha \cdot N_m \rfloor$ samples. As shown in Figure 5 and Figure 6, Multi (GAT) yields only marginal improvements over Single (GAT), suggesting that merely aggregating heterogeneous data is insufficient for effective knowledge transfer. In contrast, REBASE consistently outperforms both baselines by a substantial margin across all data scales. This gap demonstrates that our relational basis-function decomposition is essential for extracting and utilizing the latent mechanisms shared across heterogeneous feature spaces.

B.3 Source Exclusion Analysis

To verify that the coordinate similarities reflect genuine sharing of relational patterns, we perform a leave-one-source-out analysis: we remove one source from pretraining and measure the resulting change in fine-tuning performance on the remaining domains.

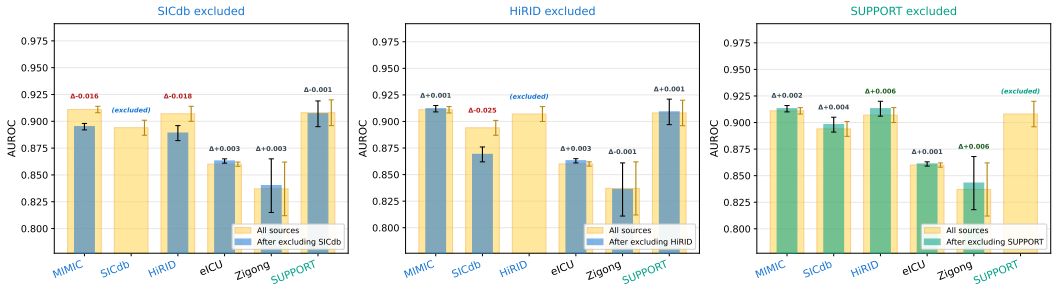

Figure 7: Source exclusion analysis on the ICU benchmark. Each panel removes one source from pretraining and reports fine-tuning AUROC on the rest, averaged over five seeds with standard deviation.

On the intensive care benchmark, all sources share the same outcome (mortality), so the coordinate similarities in Figure 2 follow from which clinical categories each source records (Table 18). In Figure 7, labels are colored by these similarities: blue for MIMIC-IV, SICdb and HiRID, which are mutually similar, green for SUPPORT, which is least similar to all others, and black for the remaining sources. Removing SICdb degrades MIMIC-IV and HiRID, the two sources it is most similar to, and removing HiRID degrades SICdb; these three sources are the ones that record arterial blood gases, coagulation assays and ventilation. Removing SUPPORT leaves all domains unaffected.

On the cardiovascular benchmark, the sources differ further in their clinical endpoints (Table 14) as well as in their recorded variables (Table 15), and the coordinate similarities in Figure 8 follow this separation. Figure 9 shows the same pattern. When MED (CAD) is excluded, domains with similar basis usage – CDD (CAD) and HD (CAD) – experience noticeable performance degradation, while domains with dissimilar basis usage remain largely unaffected. Excluding SAH (CHD) or HF (MRT), which exhibit distinct basis usage patterns, leads to negligible performance changes to other domains.

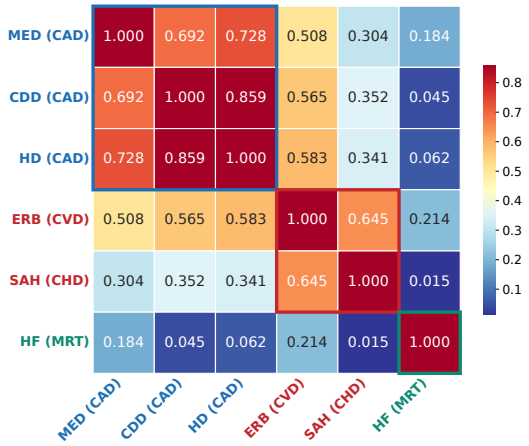

Figure 8: Pairwise similarity among source domains (Cardiovascular).

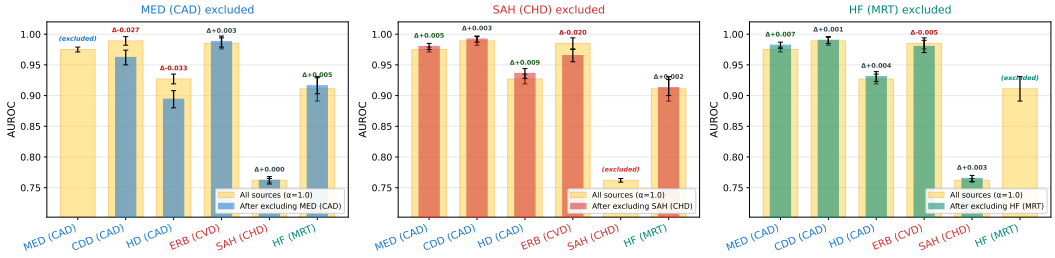

Figure 9: Source exclusion analysis. Each panel removes one source from pretraining and reports fine-tuning AUROC on the rest. Bar colors reflect coordinate similarity from Figure 8.

B.4 Graph Size without Imputation

Table 7: Effect of dropping imputed nodes on the ICU benchmark. AUROC is the fine-tuning result after multi-source pretraining, averaged over five runs. Nodes denotes the number of feature nodes per sample, averaged over samples when imputed nodes are dropped.

	w imputed nodes		w/o imputed nodes	
Source	Nodes	AUROC	Nodes	AUROC
MIMIC-IV	236	91.1 ± 0.34	184.6	92.5 ± 0.29
eICU	144	86.0 ± 0.20	107.6	87.1 ± 0.24
HiRID	283	90.7 ± 0.70	219.9	90.2 ± 0.63
SICdb	246	89.4 ± 0.69	213.5	89.2 ± 0.78
Zigong	236	83.7 ± 2.55	179.5	84.2 ± 2.13
Average		88.2		88.64

The graph representation does not require a fixed column set: only the items a patient actually has become nodes, so the graph size follows the number of observed values rather than the size of the schema. Our reported experiments nonetheless impute missing values, so that every sample carries the same node set and all methods receive the same input format. Dropping the imputed nodes instead reduces the average graph size by 13–25% across the intensive care sources, while predictive performance is unchanged or slightly higher (Table 7). SUPPORT is excluded from this comparison: it reports admission observations rather than serial measurements, so almost no value is missing.

B.5 Performance results on Kidney Function regression

Task Definition. We adopt the Kidney Function (KF) regression task from YAIB (Van De Water et al., 2023). The target variable is the median serum creatinine (mg/dL) measured during the 24–48 h window after ICU admission. Patients with length of stay shorter than 48 h or lacking at least one creatinine measurement in the target window are excluded. To prevent information leakage, all creatinine- and eGFR-derived input columns, AKI and sepsis labels computed within the first 24 h, and target-related metadata are removed from the feature set. Target values are clipped to $[0.1,\ 15.0]$ mg/dL to suppress physiologically implausible outliers. SUPPORT is excluded from this task, as it provides only a single baseline creatinine at admission without longitudinal ICU measurements, making it incompatible with the 24–48 h target window definition.

Metrics. We report R^2 as the primary metric, supplemented by the mean absolute error (MAE) in mg/dL. All metrics are evaluated on a held-out 20% test split stratified by target quartile across five random seeds.

Baselines. We compare against three single-source ML models—Ridge, Random Forest, and XGBoost—trained independently on each target cohort, and four multi-source DL baselines—XTab, TransTab, CARTE, and XTFormer—pretrained on all five source cohorts before per-target fine-tuning, following the same protocol as in the classification experiments (Section 5.1). For TransTab, we use the official `TransTabRegressor` (v0.0.7).

Table 8: Kidney Function regression results across five ICU cohorts, averaged over five seeds. The best result is in **bold** and the second best is underlined.

	Test R^2 (↑)						*Test MAE (↓, mg/dL)*					
Model	MIMIC	eICU	HiRID	SICdb	Zigong	AVG	MIMIC	eICU	HiRID	SICdb	Zigong	AVG
Ridge	0.495±0.01	0.386±0.02	0.436±0.02	0.458±0.03	0.451±0.02	0.445±0.01	0.443±0.01	0.419±0.01	0.461±0.01	0.343±0.01	0.607±0.02	0.455±0.01
RF	0.585±0.02	0.416±0.02	0.563±0.02	0.510±0.03	0.530±0.02	0.521±0.02	0.378±0.01	0.400±0.01	0.390±0.02	0.316±0.02	0.462±0.02	0.389±0.01
XGB	0.646±0.02	0.444±0.03	0.618±0.03	0.564±0.04	0.541±0.03	0.563±0.02	0.356±0.01	0.392±0.02	0.373±0.02	0.301±0.03	0.479±0.03	0.380±0.01
XTab	0.652±0.02	0.448±0.04	0.595±0.03	0.553±0.05	0.526±0.04	0.555±0.03	0.357±0.02	0.391±0.02	0.385±0.02	0.310±0.02	0.471±0.03	0.383±0.01
TransTab	0.640±0.03	0.443±0.04	0.583±0.04	0.561±0.05	0.518±0.05	0.549±0.03	0.359±0.02	0.393±0.02	0.391±0.02	0.305±0.03	0.476±0.03	0.385±0.02
CARTE	**0.671**±0.02	0.461±0.03	0.642±0.02	0.598±0.03	0.580±0.04	0.590±0.02	**0.345**±0.02	0.380±0.02	0.358±0.02	0.291±0.03	0.453±0.02	0.365±0.01
XTFormer	0.668±0.02	0.463±0.03	0.628±0.03	0.587±0.04	0.588±0.04	0.587±0.02	0.347±0.02	0.378±0.02	0.365±0.02	0.294±0.02	0.448±0.02	0.366±0.01
Ours	0.660±0.02	**0.470**±0.02	**0.658**±0.03	**0.612**±0.04	**0.601**±0.02	**0.600**±0.02	0.350±0.01	**0.374**±0.02	**0.351**±0.02	**0.288**±0.03	**0.441**±0.02	**0.361**±0.01

B.6 Sensitivity to model components

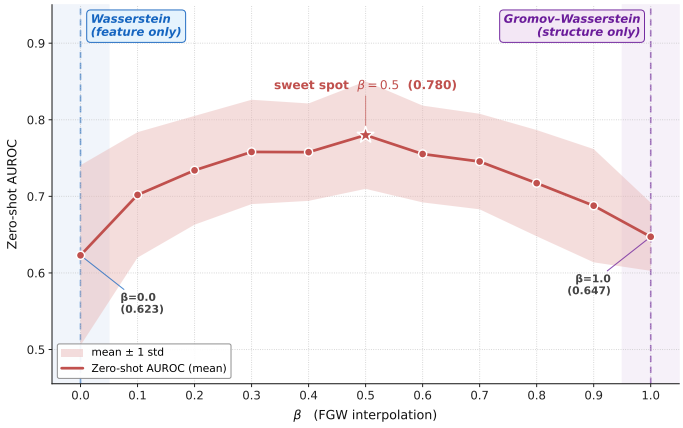

Figure 10: Zero-shot AUROC on the HD target dataset as a function of the FGW interpolation coefficient β. The two extremes correspond to pure Wasserstein (β=0, feature-only) and pure Gromov–Wasserstein (β=1, structure-only), while intermediate values balance the two signals. Shaded region denotes mean $\pm$ 1 std.

Fused Gromov Wasserstein Coefficients Variation We conduct an ablation on the interpolation coefficient β of the Fused Gromov–Wasserstein objective. The parameter $\beta \in [0, 1]$ controls the

balance between the Wasserstein term (node-attribute alignment) and the Gromov–Wasserstein term (structural alignment), where $\beta = 0$ reduces to the Wasserstein distance and $\beta = 1$ to the Gromov–Wasserstein distance. As shown in Figure 10, zero-shot performance on HD at the two extremes is noticeably lower: β=0 relies solely on node attributes and fails to leverage any relational information, while β=1 depends entirely on structural geometry and becomes unable to distinguish nodes whose structural roles coincide despite carrying different semantics. As β moves toward intermediate values, both sources of information are jointly exploited and zero-shot AUROC rises substantially, peaking at 0.780 at β=0.5, with values between β=0.3 and 0.7 remaining within one standard deviation of this peak. This inverted-U trajectory indicates that neither attributes nor structure alone are sufficient, and that the FGW framework benefits from a balanced contribution of the two.

Table 9: Sensitivity to the graph encoder on the cardiovascular benchmark (AUROC, %).

Encoder	0	4	8	16	32	64
GAT (Veličković et al., 2017)	76.83	**78.01**	**81.17**	**84.38**	**87.34**	**89.65**
GATv2 (Brody et al., 2022)	**77.23**	77.60	80.10	83.00	86.10	88.90
Graph Transformer (Shi et al., 2021)	76.46	77.30	79.80	82.80	85.90	88.70
GCN (Kipf & Welling, 2017)	65.23	68.10	71.20	74.80	78.30	81.50
GIN (Xu et al., 2019)	67.11	69.80	72.60	75.90	79.00	81.90

Sensitivity to the Graph Encoder. We examine how the choice of graph encoder affects performance, since REBASE uses the encoder's attention matrix as the relational structure for FGW alignment. We compare three attention-based encoders—GAT (Veličković et al., 2017), GATv2 (Brody et al., 2022), and a Graph Transformer (Shi et al., 2021)—with two non-attention encoders, GCN (Kipf & Welling, 2017) and GIN (Xu et al., 2019). GCN and GIN aggregate over a fixed adjacency matrix, so on a fully-connected instance graph their affinity becomes uniform; the structure term of the FGW objective then carries no information and the alignment reduces to attribute-only matching. As shown in Table 9, the three attention-based encoders perform within a narrow range (92.77–93.18 at full supervision), so the choice among them is not decisive, whereas GCN and GIN drop by more than eight points, with the gap widening further in the zero-shot regime. Notably, their zero-shot performance (65.23 and 67.11) closely matches the β=0 ablation above, where the structure term is removed and alignment relies solely on node attributes. This correspondence confirms that the structure term depends on a learned affinity: once the affinity is uniform, it contributes no information and performance falls to the level of attribute-only alignment.

Table 10: Few-shot AUROC on the cardiovascular benchmark under different coordinate computations, with all other components fixed.

Alignment	0	4	8	16	32	64
Pooled `[CLS]` cosine	61.2	64.6	70.6	72.4	73.4	81.1
Cross-attention	55.9	67.6	67.3	72.0	74.9	80.0
FGW (ours)	**76.8**	**78.0**	**81.2**	**84.4**	**87.3**	**89.7**

Table 11: Per-epoch cost of each coordinate computation on the cardiovascular benchmark.

Alignment	s/epoch
Pooled `[CLS]` cosine	13.94
Cross-attention	14.22
FGW (ours)	16.08

B.7 Necessity of the FGW Alignment

We ask whether the FGW alignment itself is necessary, by replacing only the coordinate computation while keeping the encoder, the prototype graphs and the basis GNN unchanged. Two alternatives are

considered: the cosine similarity between the [CLS] embedding and a pooled prototype representation, and the maximum cross-attention score between the instance nodes and the prototype nodes. In both cases only the coordinates are obtained differently; the basis outputs and their combination remain the same.

Neither alternative meets the requirement stated in Section 4.2. The cosine similarity does not use the instance affinity A, and pooling the prototype nodes discards $\bar{A}_k$, so each graph is compared as a single vector. The cross-attention score is computed at the node level but uses only the node embeddings, with no place for either affinity. Neither therefore represents the relations between pairs of variables explicitly in the alignment.

The results follow accordingly. Both alternatives degrade most severely in the zero-shot setting, where the coordinates must be determined without any target label, and remain behind at every shot size (Table 10). The FGW alignment costs about 15% more per epoch than the cheaper of the two (Table 11). Without a correspondence between individual nodes, the alternatives also do not support the analysis in Section 5.2, where each basis is interpreted through the variables that separate its high-risk group from the rest.

B.8 Contribution of Individual Bases

To examine whether each basis is used where its clinical category is recorded, we remove one basis from the combination in Eq. (4) at inference and renormalize the remaining coordinates, leaving every learned component unchanged. Figure 11 reports the resulting change in AUROC for three of the bases interpreted in Section 5.2, with N the number of the 43,472 test samples assigned to each basis by the largest coordinate, which indicates how often each basis is used.

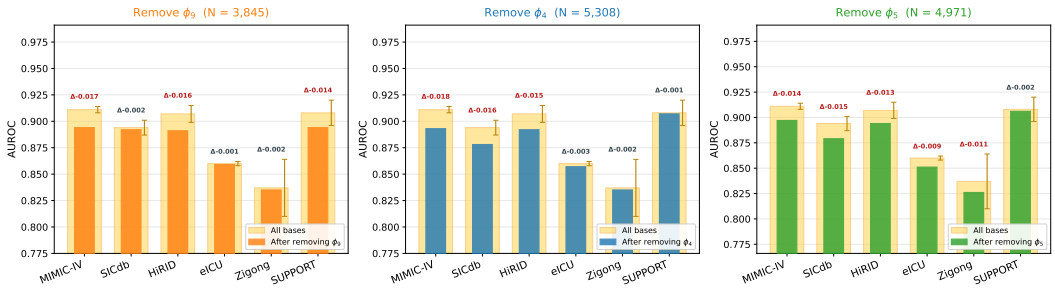

Figure 11: Change in AUROC when a single basis is removed from the combination and the remaining coordinates are renormalized. Colors follow the bases in Figure 3.

The degradation follows the clinical categories each source domain records (Table 18). Removing ϕ_9, which captures impaired consciousness, degrades MIMIC-IV, HiRID and SUPPORT, the three sources that record a consciousness score, and leaves the others largely unaffected. Removing ϕ_4 degrades MIMIC-IV, SICdb and HiRID, the three sources that record prothrombin time. Removing ϕ_5 degrades MIMIC-IV, SICdb and HiRID, which record both bicarbonate and arterial carbon dioxide; SUPPORT, which records no bicarbonate, is unaffected.

C Experimental Setup - Details

C.1 Implementation Details

The code for reproducing all experiments is available.[3] All experiments were conducted on NVIDIA RTX A6000 GPUs (48GB VRAM).

We use EmbeddingGemma-300M (Vera et al., 2025), a lightweight text embedding model based on the Gemma architecture, as the default LLM encoder for computing feature name and value embeddings, using its medical-domain fine-tuned variant[4].

[3] https://anonymous.4open.science/r/REBASE-E3B2/

[4] https://huggingface.co/sentence-transformers/embeddinggemma-300m-medical

The GAT encoder follows the GATv1 architecture (Veličković et al., 2017) with 2 layers and edge attributes included in the attention computation. The FGW distance is computed using the POT (Python Optimal Transport) package.[5] For source pretraining, we use the Adam optimizer with a batch size of 128 and a linear learning rate decay schedule. A warmup period of 60 epochs is applied, during which early stopping is deactivated. After warmup, early stopping is based on the mean validation AUROC across source domains with a patience of 20 epochs. For few-shot target adaptation, a linear learning rate decay is applied, and the final results are averaged over $R = 25$ support-set resampling episodes with a fixed test split.

The prototype affinity is parameterized as $\bar{A}_k = \sigma(\theta_k)$ and initialized as $\theta_{k,tt'} = \mathrm{logit}(1 - \delta_{k,tt'})$, where $\delta_{k,tt'}$ is the squared Euclidean distance between sub-centroids t and t', divided by the 90th percentile of these distances over all prototype graphs and capped at 1. The resulting $1 - \delta_{k,tt'}$ is clipped to $[0.01, 0.99]$ before the logit.

All hyperparameters are selected by mean AUROC on the source-domain validation split of each pretraining run; target performance plays no role, and no target validation set is used in the zero-shot or few-shot setting (Section 5.3). We first fix the number of prototype graphs K and nodes per graph T, since they determine the size of the transport plan and thus the cost of every subsequent run, then select the remaining hyperparameters with Weights & Biases sweeps (Biewald, 2020). The search spaces are summarized in Table 12; across all pretraining runs, the selected values lie in $K, T \in \{8, 12\}$ and $\beta \in [0.5, 0.7]$.

Table 12: Hyperparameters of REBASE.

Block	**Set of Hyperparameters**
Initialization	Xavier
Optimizer	Adam
Language model encoder	EmbeddingGemma-300M
No. of GAT layers	2
No. of basis GNN layers	2
Hidden dimension	192
Batch size	{32, 64, 128}
LR schedule	Linear decay
Search Space	
Source learning rate (η_{src})	{0.0005, 0.001, 0.002}
Target learning rate (η_{tgt})	{0.00005, 0.0001, 0.0003}
Dropout rate	{0.1, 0.2, 0.3}
No. of prototype graphs (K)	{4, 8, 12, 16, 32}
No. of nodes per graph (T)	{4, 8, 12, 16, 32}
FGW β	{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}
Temperature τ	{0.01, 0.02, 0.05, 0.1}
Reg. coefficient α_{reg}	{0.1, 0.25, 0.5, 0.75, 1.0, 1.5, 2.0}

C.2 Baseline Implementation Details

For deep tabular baselines, we use official implementations with the recommended hyperparameter settings from their original works. Classical and tree-based baselines are tuned via grid search as described below. All experiments are evaluated on classification tasks using stratified 80/20 train/test splits, and results are averaged over 5 random seeds.

Classical Machine Learning Methods. For Logistic Regression and Random Forest, we determine hyperparameters via grid search with k-shot cross-validation, following (Hegselmann et al., 2023). MLP uses a three-layer architecture with hidden dimensions $[256, 256, 1]$ and dropout rate 0.1, trained with early stopping (patience 5, maximum 100 epochs) and tuned over batch size and learning rate. The search spaces for all classical methods are specified in Table 13.

[5] https://pythonot.github.io/gen_modules/ot.batch.html#ot.batch.solve_gromov_batch

Table 13: Hyperparameter search spaces for classical, tree-based, and MLP baselines.

Model	Parameter	Search space
LogReg	penalty	{l1, l2}
	C	{100, 10, 1, 1e-1, 1e-2, 1e-3, 1e-4, 1e-5}
RF	bootstrap	{True, False}
	max_depth	{2, 4, 6, 8, 10, 12}
	n_estimators	{2, 4, 8, 16, 32, 64, 128, 256}
XGBoost	n_estimators	{50, 100, 200}
	max_depth	{2, 4, 6, 8, 10, 12}
	alpha	{1e-8, 1e-7, ..., 1e-1, 1}
	lambda	{1e-8, 1e-7, ..., 1e-1, 1}
	eta	{0.01, 0.03, 0.1, 0.3}
CatBoost	depth	{3, 4, 5, 6}
	iterations	{100, 200, 300}
MLP	architecture	3-layer, hidden [256, 256, 1] (fixed)
	dropout	0.1 (fixed)
	batch size	{16, 32, 64, 128}
	learning rate	{5e-5, 1e-4, 5e-4, 1e-3}
	early stopping	patience 5, max 100 epochs (fixed)

Tree-Based Methods. For XGBoost (Chen & Guestrin, 2016), hyperparameters are selected via grid search with k-shot cross-validation following (Hegselmann et al., 2023). For CatBoost (Prokhorenkova et al., 2018), we use grid search over depth and iterations with k-shot cross-validation; categorical features are handled via native encoding. The search spaces are specified in Table 13.

Cross-Table Pretrained Methods. For CARTE (Kim et al., 2024), we use version 0.0.26 of the official `carte-ai` package[6] with `CARTEClassifier` under default parameters. For TransTab (Wang & Sun, 2022), we use the official implementation[7] with 2 layers of gated transformers, embedding dimension 128, hidden dimension 256, and 8 attention heads, trained with Adam optimizer and early stopping with patience 10 over a maximum of 100 supervised training epochs. For XTab (Zhu et al., 2023), we use the publicly available pretrained FT-Transformer backbone[8] with 3 transformer blocks, embedding size 192, and 8 attention heads, finetuned with AdamW (learning rate 1×10^{-4}, weight decay 1×10^{-5}, batch size 128). For XTFormer (Chen et al., 2024), we use the official implementation[9] with the CaLinear-based architecture and the coordinate positioning protocol as described in the original work. To ensure fair comparison across datasets of varying scale, we adjust the batch size and the number of calibration and refinement epochs for XTFormer following the training configurations recommended by the original authors. More broadly, across all pretrained baselines, hyperparameters such as batch size and finetuning epochs are adjusted per dataset scale under the training configurations of each original work.

Deep Tabular and LLM-Based Methods. For TabM (Gorishniy et al., 2025), we use `TabM_D_Classifier`[10] with default settings. For TabLLM (Hegselmann et al., 2023), we use the official implementation[11] with the T0 model (`bigscience/T0_3B`) as the backbone. We adopt the T-Few fine-tuning method (Liu et al., 2022) with default hyperparameters, training for 30 epochs following the original protocol.

Tabular Foundation Models. For TabPFN-2.5 (Grinsztajn et al., 2025), we use version 6.4.1 of the official `tabpfn` package with the `tabpfn-v2.5-classifier-v2.5_default` check-

[6] https://github.com/soda-inria/carte
[7] https://github.com/ryanwangzf/transtab
[8] https://github.com/bingzhaozhu/xtab
[9] https://openreview.net/forum?id=iy31dsnsHz
[10] https://github.com/yandex-research/tabm
[11] https://github.com/clinicalml/TabLLM

Table 14: Clinical categories of the cardiovascular datasets. Abbreviations follow the 2023 AHA/ACC Guideline for Chronic Coronary Disease (Virani et al., 2023) (Section 2, Abbreviations); definitions are drawn from the cited sources.

Abbr.	Clinical Category	Definition	Datasets
CAD	Coronary Artery Disease	Atherosclerotic plaque buildup in the coronary arteries (Virani et al., 2023; Shahjehan et al., 2024)	MED, CDD, HD
CVD	Cardiovascular Disease	Any condition involving the heart or blood vessels (Virani et al., 2023)	ERB
CHD	Coronary Heart Disease	Encompasses stable angina, ACS, and silent myocardial ischemia (Shahjehan et al., 2024)	SAH
HF	Heart Failure	Clinical syndrome resulting from structural or functional impairment of ventricular filling or ejection of blood (Heidenreich et al., 2022); included in the CVD category (Palaniappan et al., 2026)	HF

point and `TabPFNClassifier` under default parameters (an ensemble of 8 estimators). For training sets exceeding its native limit of 50,000 samples, which occurs only in the full-data setting of MIMIC-IV and eICU, we draw a stratified random subset of 50,000 samples per seed; no dataset exceeds its feature limit. For TabICL v2 (Qu et al., 2026), we use version 2.0.3 of the official `tabicl` package with the `tabicl-classifier-v2-20260212` checkpoint and `TabICLClassifier` under default parameters (an ensemble of 8 estimators), using the full training pool in all settings. Both models handle missing values and categorical features natively, and receive the raw inputs without additional preprocessing.

D Dataset Details

D.1 Cardiovascular Datasets

Task Definition. All six cardiovascular datasets are posed as binary classification, where the positive class denotes a clinically adverse cardiovascular outcome. Their specific endpoints differ, however: five datasets target the presence of cardiovascular disease, whereas HF targets mortality among patients already diagnosed with heart failure (Chicco & Jurman, 2020). This benchmark therefore relaxes the shared-label assumption further than the intensive care and breast cancer benchmarks, whose sources share the same type of endpoint (mortality and disease-free survival, respectively). We observe no negative transfer from this relaxation in either direction: excluding HF from pretraining leaves the other domains largely unaffected (Figure 9), and multi-source training improves performance on HF over single-source training, both with and without the basis decomposition (Figure 5). To interpret the learned bases, we further categorize the cohorts into four groups by their target definition and feature scope, using the abbreviations of the 2023 AHA/ACC guideline (Virani et al., 2023) (Table 14). **MED**, **CDD**, and **HD** are classified as **CAD** because their features include diagnostic modalities that the 2023 AHA/ACC Chronic Coronary Disease Guideline associates with established chronic coronary disease: cardiac biomarkers (Troponin, CK-MB) in MED, and electrocardiography and exercise stress testing in CDD and HD (Table 5 of Virani et al. (2023); see Table 15 for the variable-level mapping). **ERB** is classified as **CVD** because its target denotes a general heart problem (Hamarash et al., 2024) and its features span both diagnostic and risk-factor categories. **SAH** targets coronary heart disease (**CHD**) (Rossouw et al., 1983)—a term frequently used interchangeably with CAD (Shahjehan et al., 2024); we separate it because its features consist exclusively of risk factors with no diagnostic test variables, corresponding to a primary prevention scope (Arnett et al., 2019). Finally, **HF** is separated as it is managed under a distinct guideline (Heidenreich et al., 2022).



Figure 12: Radar plots of basis-wise clinical profiles for all six cardiovascular datasets. [H] and [L] indicate whether each basis is associated with high- or low-risk outcomes.

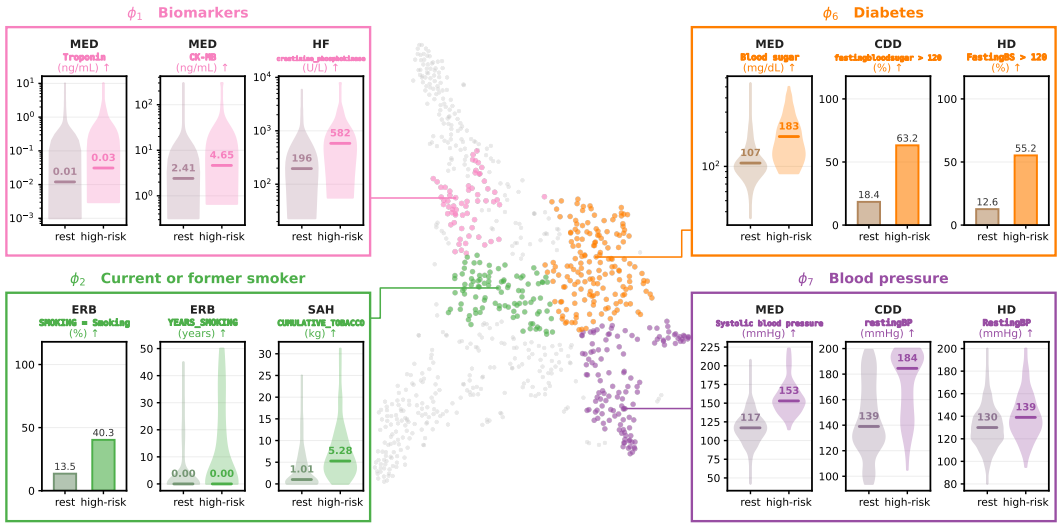

Figure 13: t-SNE of cardiovascular test coordinates with four clinically interpretable bases. Each panel shows a representative clinical category of that basis, comparing samples assigned to the basis (high-risk) against the rest across sources; variables differ in name or definition across sources yet shift in the same clinical direction. Full profiles are reported in Appendix E.

Basis-Wise Clinical Profile Analysis. In this section, we examine the radar profiles (Figure 12) for all six cardiovascular datasets to assess whether each relational basis captures a clinically meaningful relational pattern across sources. For each dataset, we display the top three high-risk bases and their low-risk counterparts. For ERB, 15 of 20 features with the strongest basis-wise significance are displayed; the full feature set is listed in Table 15. Variable colors correspond to the clinical categories defined in Virani et al. (Virani et al., 2023) (Table 15).

Ischemic Cascade (ϕ_1): Myocardial ischemia occurs when coronary blood flow is insufficient to meet the oxygen demand of the heart muscle (Shahjehan et al., 2024; Thygesen et al., 2018); prolonged ischemia leads to myocardial necrosis (infarction), releasing intracellular proteins into the bloodstream (Basit et al., 2023; Thygesen et al., 2018). Across the CAD sources, ϕ_1 consistently captures variables associated with this process, despite substantial differences in how each source records it. In **MED**, ϕ_1 responds to cardiac biomarkers (orange: Troponin, CK-MB)—proteins released upon cardiomyocyte death, listed under "Biomarkers" in Virani et al. (Virani et al., 2023) (Table 5)—and elevated heart rate (blue), consistent with compensatory tachycardia when damaged myocardium cannot maintain adequate cardiac output. In **CDD** and **HD**, the same basis shifts to exercise stress test variables (blue: slope, oldpeak, exerciseangia): ST-segment depression during exercise indicates that narrowed coronary arteries cannot supply sufficient oxygen under increased demand (Virani et al., 2023) (§3.1, Recommendation 4), provoking angina and reduced functional capacity. In **ERB**, whose feature set shares no overlap with the CAD sources, ϕ_1 captures resting electrocardiographic findings (blue: ECG_PATTERN with ST Depression, Q_WAVE) together with prior cardiac event history (yellow: HEART_SURGERY_HISTORY). Pathological Q waves on the ECG are an established marker of prior myocardial necrosis per ESC/ACCF/AHA/WHF consensus criteria (Florian et al., 2012), linking them to the same ischemic pathway that Troponin and CK-MB reflect in MED. In **SAH**, ϕ_1 shows no clear ischemic signature, as the dataset contains no diagnostic test variables. In **HF**, ϕ_1 responds to ejection fraction (blue)—listed under "Echocardiography: reduced left ventricular ejection fraction" in Virani et al. (Virani et al., 2023) (Table 5)—and creatinine phosphokinase (orange), both indirectly related to myocardial damage, but shows no distinct ischemic signature, as variables that specify the cause of injury (e.g., troponin, ECG findings) are not recorded.

Metabolic Synergy (ϕ_6): Dyslipidemia and diabetes are established risk factors for coronary artery disease (Shahjehan et al., 2024); elevated cholesterol contributes to atherosclerotic plaque buildup in the coronary arteries, while chronic hyperglycemia damages vascular endothelium over time. Both are listed under "Past or Concurrent Medical, Mental Health Conditions" in Virani et al. (Virani et al., 2023) (Table 5). ϕ_6 consistently isolates these metabolic variables across sources. In **CDD** and **HD**, it responds to cholesterol and fasting blood sugar (yellow), with minimal activation on stress test variables (blue), consistent with a metabolic rather than an ischemic risk profile. In **MED**, ϕ_6 captures elevated blood sugar (yellow) with moderately raised blood pressure (red), consistent with metabolic derangement increasing vascular resistance. In **SAH**, ϕ_6 captures LDL cholesterol, adiposity, and obesity (yellow), forming a coherent dyslipidemia–obesity profile aligned with the risk factors emphasized in the 2019 ACC/AHA Primary Prevention Guideline (Arnett et al., 2019). In **ERB**, ϕ_6 is largely inactive; although DIABETIC is present, the dataset lacks complementary metabolic variables (e.g., cholesterol, fasting glucose) that would allow this basis to form a meaningful metabolic pattern. In **HF**, ϕ_6 responds to diabetes and serum creatinine (yellow), the latter being a renal function marker in the standard diagnostic evaluation for HF (Heidenreich et al., 2022) that is not recorded in the other sources. It shows no distinct metabolic profile, as diabetes has near-zero label association in this domain (41.7% among positives vs. 41.8% overall).

Hypertensive-Structural Axis (ϕ_7): Hypertension is one of the most prevalent comorbidities among patients with CCD, co-occurring in over 60% of cases (Virani et al., 2023) (§4.2.7); sustained elevated blood pressure increases afterload on the heart, leading to left ventricular hypertrophy (LVH) and elevated risk of major adverse cardiovascular events. ϕ_7 selectively captures blood pressure variables (red) across all sources where they are available. In **MED**, it responds to systolic and diastolic blood pressure with elevated heart rate. In **CDD** and **HD**, it isolates resting blood pressure (red) with minimal response to stress test or metabolic variables, consistent with a hypertensive profile distinct from ischemic or metabolic pathways. In **ERB**, ϕ_7 captures systolic and diastolic blood pressure (red) together with IVS_DIASTOLE (blue)—interventricular septum thickening measured by echocardiography, listed under "Echocardiography: left ventricular hypertrophy" in Virani et al. (Virani et al., 2023) (Table 5)—long-standing hypertension. In **SAH**, ϕ_7 responds to systolic blood pressure (red) and age, consistent with age-related chronic hypertension. In **HF**, high blood pressure is recorded only as a categorical variable (red), and ϕ_7 shows no distinct hypertensive profile in this domain.

Behavioral Risk Profile (ϕ_2): Tobacco use, physical inactivity, and excessive alcohol consumption are modifiable behavioral risk factors addressed in dedicated management sections of the guideline (Virani et al., 2023) (§4.2.3, §4.2.4, §4.2.11). ϕ_2 captures these behavioral variables, but only in

sources that contain them. **MED**, **CDD**, and **HD** lack lifestyle variables entirely, so ϕ_2 has no signal in these domains. In **ERB**, ϕ_2 responds to smoking status, years of smoking, physical inactivity, and body weight (green, yellow). In **SAH**, it captures cumulative tobacco consumption, alcohol intake, and Type A behavior (green). These behavioral factors contribute to cardiovascular risk through long-term cumulative exposure; their effects typically manifest over decades, eventually overlapping with the pathways captured by ϕ_1 (ischemia), ϕ_6 (metabolic), and ϕ_7 (hypertension). In **HF**, lifestyle variables show near-zero label association (smoking: 31.2% among positives vs. 32.1% overall), and ϕ_2 shows no distinct behavioral profile in this domain.

Summary. Across the cardiovascular sources, each basis responds to variables of the same clinical category even when they are recorded under different names, modalities, or measurement scales, as with ϕ_1 capturing cardiac biomarkers in MED, exercise stress test variables in CDD and HD, and resting ECG findings in ERB. Where a source does not record the variables of a category, or where they show little association with the label, the corresponding basis shows no distinct profile, as with ϕ_1 in SAH, ϕ_2 in MED, CDD, and HD, and ϕ_6, ϕ_7, and ϕ_2 in HF. The clinical role of each basis thus depends on which variables a source provides (Table 15), while remaining consistent across the sources that record them.

Table 15: Clinical variable composition of the cardiovascular datasets. Variable categories follow the risk stratification taxonomy of Virani et al. (Virani et al., 2023) (Table 5, "Potential Features Associated With a Higher Risk of MACE Among Patients With Established CCD"). Variables not listed in Table 5 are mapped to supplementary sections of the same guideline, indicated by superscripts.

	CAD			**CVD**	**CHD**	**HF**
Category	MED (8)	CDD (12)	HD (11)	ERB (20)	SAH (9)	HF (11)
Biomarkers						
High-sensitivity troponin	Troponin CK-MB	—	—	—	—	creatinine_phosphokinase
Ancillary Cardiac Testing or Imaging						
ECG	Heart rate	restingrelectro	RestingECG	ECG_PATTERN Q_WAVE HEART_RATE	—	—
EST	—	slope oldpeak exerciseangia maxheartrate	ST_Slope Oldpeak ExerciseAngina MaxHR	—	—	—
Echocardiography	—	—	—	IVS_DIASTOLE	—	ejection_fraction
Past or Concurrent Medical, Mental Health Conditions						
Elevated body mass index	—	—	—	WEIGHT HEIGHT	ADIPOSITY OBESITY	—
Diabetes	Blood sugar	fastingBS	FastingBS	DIABETIC	—	diabetes
Dyslipidemia	—	cholesterol	Cholesterol	LDL_CHOLESTEROL	LDL_CHOLESTEROL	—
Current or former smoker	—	—	—	SMOKING YEARS_SMOKING	CUMULATIVE_TOBACCO	smoking
Chronic kidney disease	—	—	—	—	—	serum_creatinine
Previous MI, PCI, or CABG	—	—	—	HEART_SURGERY_HISTORY	—	—
Demographics and Socioeconomic Status						
Age	Age	age	Age	AGE	AGE	age
Male sex	Gender	gender	Sex	SEX	—	sex
Additional clinical variables[a–h]						
Coronary angiography[a]	—	noofmajorvessels	—	—	—	—
Symptom assessment[a]	—	chestpain	ChestPainType	CHEST_PAIN	—	—
Blood pressure[b]	Systolic BP Diastolic BP	restingBP	RestingBP	BLOOD_PRESSURE_SYSTOLIC BLOOD_PRESSURE_DIASTOLIC HYPERTENSION	SYSTOLIC_BLOOD_PRESSURE	high_blood_pressure
Alcohol[c]	—	—	—	—	ALCOHOL_CONSUMPTION	—
Physical activity[d]	—	—	—	PHYSICALLY_ACTIVE	—	—
Family history[e]	—	—	—	FAMILY_HISTORY	FAMILY_HISTORY	—
Behavioral[f]	—	—	—	—	TYPE_A_BEHAVIOR	—
Social determinants of health[g]	—	—	—	LIFESTYLE	—	—
Dataset-specific markers[h]	—	—	—	—	—	serum_sodium platelets anaemia

[a–h]Not listed in Table 5 of Virani et al. (Virani et al., 2023); mapped to the indicated sections of the same guideline where applicable. [a]§3.1 (Diagnostic Evaluation), [b]§4.2.7 (Blood Pressure Management), [c]§4.2.4 (Alcohol and Substance Use), [d]§4.2.11 (Physical Activity), [e]§6.3 (Young Adults), [f]§4.2.2 (Mental Health Conditions), [g]§4.1.4 (Social Determinants of Health), [h]Dataset-original clinical features; no corresponding entry in the guideline.

Cardiovascular Datasets. We collect six publicly available cohorts from different institutions and repositories (Kaggle, Mendeley Data, and the UCI ML Repository), all formulated as binary classification. These datasets target distinct cardiovascular endpoints; Table 14 summarizes the clinical categories and their mapping to each dataset, following standard abbreviations from AHA/ACC guidelines (Virani et al., 2023; Arnett et al., 2019; Lawton et al., 2022; Heidenreich et al., 2022; McDonagh et al., 2021).

Table 16: Statistics of the cardiovascular datasets.

Dataset	Samples	Features	Cat.	Num.	Pos. rate
MED (CAD)	1319	8	1	7	0.614
CDD (CAD)	1000	12	7	5	0.580
HD (CAD)	918	11	6	5	0.553
ERB (CVD)	287	20	12	8	0.355
SAH (CHD)	462	9	1	8	0.346
HF (HF)	299	11	5	6	0.321

MED (Medicaldataset)[12] was collected at Zheen Hospital in Erbil, Iraq, between January and May 2019 (Rashid & Hassan, 2022). It contains 1319 patient records with 9 features including heart rate, systolic and diastolic blood pressure, blood sugar, CK-MB, and troponin, labeled as CAD positive or negative. After excluding the target label, 8 input features remain.

CDD (Cardiovascular Disease Dataset)[13] was acquired from a multispecialty hospital in India (Doppala et al., 2022). It consists of 1000 subjects described by 14 features covering demographic, examination, and diagnostic attributes for CVD prediction. After excluding the target label and the patient identifier, which carries no clinical information, 12 input features remain.

HD (Heart)[14] combines five independently published heart disease datasets — Cleveland, Hungarian, Switzerland, Long Beach VA, and Statlog — over 11 common features (Janosi et al., 1988). After removing 272 duplicate records from the combined 1190 observations, the final dataset contains 918 samples (6 categorical, 5 numerical) for CAD prediction.

ERB (Erbil Cardiovascular Health Dataset)[15] consists of 287 visitors to a private clinic in Erbil City, Kurdistan Region of Iraq, collected between November 2018 and April 2019 (Hamarash et al., 2024). The 21 features span four categories: demographic, physical examination and symptomatic, medical lab tests, and diagnostic features including ECG and echocardiography results, for CVD prediction. After excluding the target label, 20 input features remain.

SAH (SAheart)[16] is a retrospective sample of 462 males from a CHD high-risk region of the Western Cape, South Africa (Rossouw et al., 1983), selected by Hastie and Tibshirani (Hastie & Tibshirani, 1987) from the larger CORIS baseline study. It contains 9 input features including systolic blood pressure, cumulative tobacco consumption, LDL cholesterol, and family history.

HF (Heart Failure Clinical Records)[17] contains 299 heart failure patients from the Allied Hospital and Faisalabad Institute of Cardiology, Pakistan, collected between April and December 2015 (Chicco & Jurman, 2020). Each patient is described by 13 clinical features including ejection fraction, serum creatinine, and serum sodium, with the target indicating death during the follow-up period. After excluding the target label and the follow-up time variable, which constitutes label leakage, 11 input features remain.

Cardiovascular Dataset Preprocessing Since our framework relies on LLM embeddings to encode feature semantics, categorical features must be represented as textual descriptions rather than numerical codes. For datasets where categorical values were already provided as ordinal integers (e.g., `sex`: 0/1), we map them back to their original descriptions (e.g., "Female"/"Male"), following the convention of (Wang & Sun, 2022). Missing numerical values are imputed with the training-set mean of the corresponding feature.

[12] https://data.mendeley.com/datasets/wmhctcrt5v/1
[13] https://www.kaggle.com/datasets/jocelyndumlao/cardiovascular-disease-dataset
[14] https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction
[15] https://data.mendeley.com/datasets/396fd9yp6m/2
[16] https://www.kaggle.com/datasets/yassinehamdaoui1/cardiovascular-disease/data
[17] https://archive.ics.uci.edu/dataset/519/heart+failure+clinical+records

D.2 ICU DATASETS

We use six publicly available ICU databases as source domains for multi-source pretraining, all posed as binary classification tasks predicting in-hospital mortality. Cohort construction and preprocessing follow the YAIB framework (Van De Water et al., 2023), whose task definition and exclusion criteria are detailed below.

Task Definition. We define the mortality prediction task as a binary classification of in-hospital death based on clinical observations recorded during the first 24 hours of ICU admission. Following (Van De Water et al., 2023), the feature extraction window spans from ICU admission time to 24 hours thereafter; the prediction is issued at the end of this window. To ensure that the prediction is clinically meaningful, we exclude stays shorter than 30 hours — this prevents the inclusion of patients who died or were discharged within the first 6 hours after the prediction time, which would otherwise conflate the sickest patients (imminent death) and the healthiest patients (early discharge) at the time of prediction. For SICdb, only the ICU discharge status is available, so mortality is defined as death at ICU discharge; for HiRID, we use the discharge status as provided.

Exclusion Criteria. Following the protocol of (Van De Water et al., 2023), we include ICU stays of adult patients and exclude stays with an invalid admission or discharge time (missing or negative length of stay), stays shorter than 6 hours in the ICU, stays with fewer than 4 separate hours of recorded measurements, and stays containing any gap of 12 or more consecutive hours without a measurement. For the mortality task, we further exclude stays with a total length of stay shorter than 30 hours. After applying these criteria, features are extracted from the first 24 hours of each remaining stay, aggregated into five summary statistics (mean, standard deviation, minimum, maximum, count) per variable, and columns with more than 50% missing values are removed. This aggregation yields one tabular sample per stay, in line with our focus on feature heterogeneity across sources rather than temporal dynamics within a stay; all methods receive the same inputs. Each database retains its own feature schema without cross-database harmonization. SUPPORT, which lacks timestamped measurements, follows the cohort construction described in its dataset description instead.

Table 17: Statistics of the ICU source datasets (mortality task).

Dataset	Stays	Features	Mortality prev.	Origin
MIMIC-IV	65,745	236	12.70%	USA
eICU	114,069	144	9.76%	USA (multi-center)
HiRID	13,886	283	8.52%	Switzerland
SICdb	14,189	246	4.54%	Austria
SUPPORT	6,842	76	19.20%	USA
Zigong	2,616	236	5.20%	China

MIMIC-IV[18] (v3.1) was collected at the Beth Israel Deaconess Medical Center, USA (Johnson et al., 2016). Features are extracted from 10 derived concept tables provided by the MIT-LCP mimic-code repository, covering vital signs, chemistry, blood gas, complete blood count, coagulation, enzymes, inflammation markers, cardiac markers, GCS, and urine output. Mortality is determined by the `hospital_expire_flag` in the admissions table. The final cohort contains 65,745 stays with 236 features and a mortality prevalence of 12.70%.

eICU[19] (v2.0) is a multi-center dataset collected from 200+ hospitals across the USA using Philips ICU monitoring systems (Pollard et al., 2018). Features are extracted from the vitalPeriodic, vitalAperiodic, and lab tables. Laboratory values are stored as VARCHAR and require safe numeric casting. Mortality is determined when either `unitdischargestatus` or `hospitaldischargestatus` equals "Expired". The final cohort contains 114,069 stays with 144 features and a mortality prevalence of 9.76%.

HiRID[20] (v1.1.1) was collected at Bern University Hospital, Switzerland (Yèche et al., 2021). Observations are stored in long format (variableid × value × datetime), from which we extract 58

[18] `https://physionet.org/content/mimiciv/3.1/`
[19] `https://physionet.org/content/eicu-crd/2.0/`
[20] `https://physionet.org/content/hirid/1.1.1/`

feature groups covering vitals, GCS, blood gas, electrolytes, renal, liver, coagulation, cardiac markers, inflammation, and hematology. Mortality is determined by `discharge_status` = "dead". The final cohort contains 13,886 stays with 283 features and a mortality prevalence of 8.52%.

SICdb[21] (v1.0.8) was collected at the University Hospital Salzburg, Austria. Time offsets are recorded in seconds. Features are extracted from vital signs (data_float_h table) and laboratory results (laboratory table), mapped via the official d_references vocabulary. Mortality is determined by `DischargeState` = 2215. The final cohort contains 14,189 stays with 246 features and a mortality prevalence of 4.54%.

SUPPORT[22] is a static tabular dataset from the Vanderbilt SUPPORT study covering 9,105 seriously ill hospitalized adults (Knaus et al., 1995). As it contains no time-stamped measurements, the temporal window and exclusion criteria above do not apply; we instead retain adults enrolled within three days of admission whose stay extended at least one day beyond enrollment. Mortality is defined as in-hospital death (`hospdead`), and columns determined only after the outcome (e.g., survival estimates, physician prognoses, DNR orders, charges, and length-of-stay variables) are removed.We compute clinically derived indicators from the existing measurements (e.g., shock index, SIRS criteria), one-hot encode categorical variables, and median-impute missing values with missingness indicators. The final cohort contains 6,842 patients with 76 input features and a mortality prevalence of 19.20%.

Zigong[23] refers to the Critical Care Database Comprising Patients with Infection (Xu et al., 2022b), collected at Zigong Fourth People's Hospital, Sichuan, China, between January 2019 and December 2020. The database consists exclusively of infection-related ICU admissions. Features are extracted from nursing chart vitals and laboratory results, with time offsets recorded in hours. Mortality is determined by `Status_Discharge` = "Dead". The final cohort contains 2,616 stays with 236 features and a mortality prevalence of 5.20%. Table 17 summarizes the source dataset statistics.

Basis-Wise Clinical Profile Analysis (Intensive Care). In this section, we examine the variable-level profiles (Figure 16) for all six intensive care datasets to assess whether each relational basis captures a clinically meaningful relational pattern across sources. For each basis, we display the clinical categories it spans, with one representative variable per category and source. Variable names are given exactly as each cohort records them, and category assignments follow the organ dysfunction and acute physiology scores (Vincent et al., 1996; Knaus et al., 1985) (Table 18). A source appears in a category only when the variable it contributes separates the two groups in the clinically adverse direction; variables that are recorded but do not meet this criterion are noted where relevant.

Cardiorenal Hypoperfusion (ϕ_2): Sepsis-induced hypoperfusion is identified by hypotension or elevated lactate (Prescott et al., 2026), and a sustained fall in renal perfusion reduces glomerular filtration, which in turn raises serum creatinine and allows nitrogenous waste to accumulate (Kidney Disease: Improving Global Outcomes (KDIGO) Acute Kidney Injury Work Group, 2012). ϕ_2 captures variables consistent with this sequence across all six sources, despite substantial differences in how each records it. Renal function is the one category to which every cohort contributes, and it is also where the naming and unit differences are most visible: the same measurand appears as `chem_creatinine`, `lab_creatinine`, `crea` and `creatinine` in mg/dL, and as `creatinine_blood` and `creatinine` in μmol/L. The arterial pressure arm is likewise recorded as a mean pressure in five cohorts and as a non-invasive systolic pressure in **Zigong**, and nitrogenous waste as the nitrogen fraction (`bun`, mg/dL) in two sources and as the whole molecule (`urea`) in mmol/L and mg/dL in the other two. Across these formats the values move together: mean arterial pressure falls below the threshold at which organ perfusion is compromised (Prescott et al., 2026), creatinine rises past the first scoring band of the organ dysfunction score (Vincent et al., 1996), and lactate crosses the threshold used in the septic shock criteria (Singer et al., 2016). **SUPPORT** records blood urea nitrogen, but it does not separate the two groups, and lactate is contributed only by the three cohorts that record an arterial blood gas panel.

Hemostatic Consumption (ϕ_4): Systemic activation of coagulation consumes platelets and clotting factors, and the resulting depletion is assessed by the platelet count together with the prothrombin

[21] `https://physionet.org/content/sicdb/1.0.8/`
[22] `https://hbiostat.org/data/repo/support2csv.zip`
[23] `https://physionet.org/content/icu-infection-zigong-fourth/1.1/`

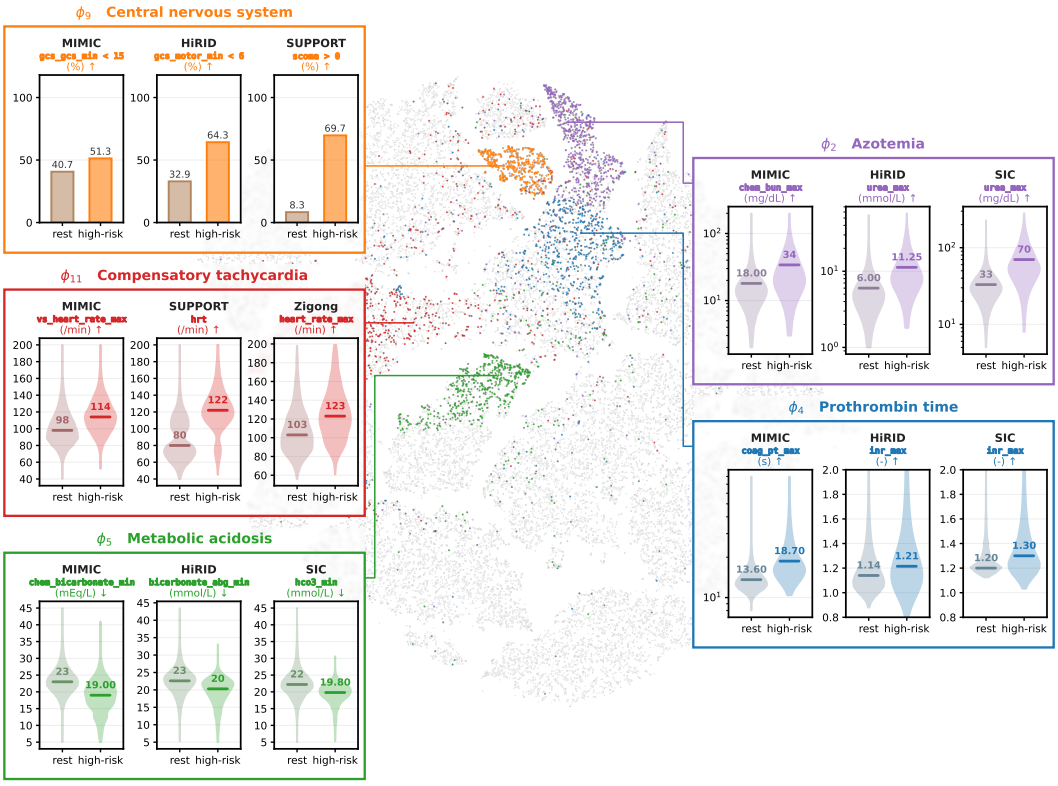

Figure 14: t-SNE of ICU test coordinates with five clinically interpretable bases. Each panel shows a representative clinical category of that basis, comparing samples assigned to the basis (high-risk) against the rest across sources; variables differ in name or definition across sources yet shift in the same clinical direction. Full profiles are reported in Appendix E.

time (Taylor et al., 2001); the platelet count alone defines coagulation failure in the organ dysfunction score (Vincent et al., 1996; Singer et al., 2016). ϕ_4 captures variables consistent with this consumption. In **MIMIC-IV**, **eICU** and **SICdb** the platelet count falls toward the first scoring band, reported in K/μL in two sources and 10^9/L in the third. The clotting time is recorded in three incompatible formats: directly in seconds (`coag_pt`), as a ratio normalized against a laboratory standard (`inr`), and as a percentage of normal activity (`prothrombin_time`), the last of which falls as the other two rise. ϕ_4 moves with all three, so a single basis reads the same prolongation signal from quantities whose scales and directions differ. In **SUPPORT**, which records no coagulation assay, ϕ_4 instead responds to the white blood cell count, an inflammatory rather than a hemostatic marker.

Metabolic Acidosis (ϕ_5): An accumulating acid load consumes bicarbonate, lowers arterial pH, and elicits a respiratory response in which ventilation rises to clear carbon dioxide; the acute physiology score records arterial pH and, where blood gases are unavailable, venous bicarbonate in its place (Knaus et al., 1985). ϕ_5 captures variables consistent with this compensated acidosis. Bicarbonate falls in five cohorts and is recorded under three names – `chem_bicarbonate` and `lab_bicarbonate` from serum chemistry, `bicarbonate_abg` from the blood gas panel, and `hco3` as the bare chemical formula – and arterial pH falls in four. The respiratory response appears through two different variables reporting the same compensation: a rising respiratory rate in **MIMIC-IV** and a falling arterial carbon dioxide tension in **HiRID** and **SICdb**. In **SUPPORT**, only arterial pH is recorded, and neither it nor the respiratory rate separates the two groups, so ϕ_5 forms no compensated pattern in this cohort.

Neurologic Depression (ϕ_9): Impaired consciousness is graded by the Glasgow Coma Scale in the organ dysfunction score (Vincent et al., 1996), and chronological age enters the acute physiology score as a separate term reflecting diminished physiologic reserve (Knaus et al., 1985). ϕ_9 captures variables consistent with both. In **MIMIC-IV**, **HiRID** and **SUPPORT** the proportion of patients

with impaired consciousness rises toward high-risk, although each source assesses it on a different instrument: the total scale, the motor subscale alone, and a separate coma score whose polarity is inverted relative to the Glasgow Coma Scale. The three instruments differ in range, granularity and sign, yet all move in the same clinical direction. Age rises in the three cohorts that contribute it. ϕ_9 also responds to admission glucose in four sources, reported in mg/dL in three and mmol/L in the fourth; the acute physiology score evaluated serum glucose and excluded it for contributing little explanatory power (Knaus et al., 1985), so this arm has no counterpart in the reference standards. **eICU**, **SICdb** and **Zigong** record no Glasgow Coma Scale, and contribute to ϕ_9 through the remaining categories.

Oxygen Delivery Failure (ϕ_{11}): Oxygen delivery is the product of cardiac output and arterial oxygen content, so a fall in either the oxygen-carrying capacity of the blood or its saturation elicits a compensatory rise in heart rate; the acute physiology score records the hematocrit, the arterial oxygen tension and the heart rate as separate terms (Knaus et al., 1985). ϕ_{11} captures variables consistent with this compensation. Arterial oxygenation is contributed by four cohorts and recorded as a partial pressure in two of them and as a saturation percentage in the other two – quantities on different scales describing the same state, although saturation is bounded above and separates less sharply. The compensatory rise in heart rate appears in four cohorts, and the oxygen-carrying capacity arm in **MIMIC-IV** and **eICU**, as hemoglobin in g/dL and as a hematocrit percentage. The remaining cohorts also record a red cell index, but it does not fall toward high-risk in these populations and therefore does not enter this category.

Table 18: Clinical taxonomy of the ICU benchmark. Each cell lists the variables a cohort contributes to that category, under that cohort's own naming convention. Categories follow the organ dysfunction score of Singer et al. (2016) and the acute physiology score of Knaus et al. (1985); variables defined by neither are grouped below and footnoted to the standard that defines them.

	MIMIC-IV	**eICU**	**HiRID**	**SUPPORT**	**SICdb**	**Zigong**
Category	(48)	(31)	(57)	(42)	(49)	(50)
Organ Dysfunction						
Respiration	bg_po2 vs_spo2	vp_sao2	po2_arterial sao2_arterial spo2	pafi	po2 sao2 spo2	spo2
Coagulation	cbc_platelet	lab_platelets	platelets	—	platelets	platelets
Liver	—	—	—	bili	—	—
Cardiovascular	vs_mbp vs_mbp_ni	va_noninvasivemean	mbp_invasive mbp_noninvasive	meanbp	mbp_invasive	—
Central nervous system	gcs_gcs gcs_gcs_motor gcs_gcs_verbal gcs_gcs_eyes	—	gcs_motor gcs_verbal gcs_eyes	scoma[a]	—	—
Renal	chem_creatinine uo_total_24h	lab_creatinine	creatinine_blood urine_output	crea urine	creatinine urine_output	creatinine urine_output
Acute Physiology						
Temperature	vs_temperature	—	temperature	temp	temperature	temperature
Heart rate	vs_heart_rate	vp_heartrate	heart_rate	hrt	heart_rate	heart_rate
Respiratory rate	vs_resp_rate	vp_respiration	resp_rate	resp	resp_rate	resp_rate
Arterial pH	bg_ph	—	ph_arterial	ph	ph	ph
Serum sodium	chem_sodium	lab_sodium	sodium	sod	sodium	sodium
Serum potassium	chem_potassium	lab_potassium	potassium	—	potassium	potassium
Hematocrit	cbc_hematocrit	lab_hct	—	—	hematocrit	hematocrit
White blood count	cbc_wbc	lab_wbc	wbc	wblc	wbc	wbc
Serum bicarbonate	chem_bicarbonate bg_totalco2	lab_bicarbonate	bicarbonate_abg	—	hco3	hco3 co2
Age	age	age	age	age	age	age
Additional physiologic variables[c]						
Azotemia[d]	chem_bun	lab_bun	urea	bun	urea	urea

Category	MIMIC-IV	eICU	HiRID	SUPPORT	SICdb	Zigong
Tissue hypoperfusion[e]	bg_lactate	—	lactate_arterial	—	lactate	—
Prothrombin time[f]	coag_pt coag_inr	—	inr	—	prothrombin_time inr	—
Other coagulation[f]	coag_ptt	—	—	—	aptt fibrinogen	—
Arterial PCO_2	bg_pco2	—	pco2_arterial	—	pco2	—
Acid–base (other)	bg_baseexcess chem_aniongap	lab_aniongap	base_excess	—	base_excess anion_gap	base_excess anion_gap
Dyshemoglobins	—	—	carboxyhemoglobin methemoglobin	—	cohgb methgb	o2hb
Red cell indices	cbc_hemoglobin cbc_rbc cbc_mcv cbc_mch cbc_mchc cbc_rdw	lab_hgb lab_rbc lab_mcv lab_mch lab_mchc lab_rdw	hemoglobin hemoglobin_abg mcv mch mchc	—	hemoglobin rbc mcv mch mchc	hemoglobin rbc mcv mch mchc rdwsd mpv pct[b]
Leukocyte differential	—	—	—	—	—	neutrophils_count neutrophil_pct lymphocytes_count lymphocyte_pct monocytes_count monocyte_pct
Acute-phase response	—	—	crp	—	crp	—
Serum glucose	vs_glucose	lab_glucose	glucose	glucose	glucose	blood_sugar glucose
Blood pressure (S/D)	vs_sbp vs_dbp vs_sbp_ni vs_dbp_ni	va_noninvasivesystolic va_noninvasivediastolic	sbp_invasive dbp_invasive sbp_noninvasive dbp_noninvasive	—	sbp_invasive dbp_invasive	sbp_noninvasive dbp_noninvasive
Hemodynamic monitoring	—	—	cvp	—	cvp	—
Ventilation	—	—	fio2 peep tidal_volume pip plateau_pressure etco2	—	etco2 tidal_volume pip	—
Hepatic enzymes	—	—	ast alt	—	—	ast
Cardiac markers	—	—	ck	—	—	—
Protein and nutrition	—	—	—	alb	—	—
Electrolytes and minerals	chem_chloride chem_calcium	lab_chloride lab_calcium lab_magnesium	chloride calcium_ionized magnesium phosphate	—	chloride calcium_ionized calcium_total magnesium phosphate	chloride calcium
Renal function (other)	—	—	—	—	—	uric_acid egfr cystatin_c
Neurologic (other)	—	—	—	—	—	pupil_left pupil_right
Anthropometry	—	—	weight height	—	—	—
Cohort-specific and non-physiologic variables[g]						
Demographics and socioeconomic status	gender	gender unittype admission_source	gender	sex race edu income	gender	gender infection_site admit_dept
Comorbidity and diagnosis	—	—	—	num.co diabetes dementia ca dzgroup dzclass	—	—

Category	MIMIC-IV	eICU	HiRID	SUPPORT	SICdb	Zigong
Severity indices	—	—	—	aps sps sirs_score shock_index temp_abnormal hr_abnormal resp_abnormal wbc_abnormal aki_indicator hypotension hypoxemia tachycardia fever	—	—
Functional status	—	—	—	adlp adls adlsc	—	—
Dataset-specific markers	—	—	—	—	—	sputum_volume gastric_juice
Derived indicators[h]	sepsis_24h aki_24h aki_stage_max_24h uo_count_24h	sepsis_24h aki_24h	sepsis_24h aki_24h	—	sepsis_24h aki_24h	—

[a] SUPPORT reports a coma score whose polarity is inverted relative to the Glasgow Coma Scale (higher indicates worse neurological function). [b] Zigong pct denotes plateletcrit, a red cell analyser index, not procalcitonin. [c] Measured physiology not listed in Table 1 of Singer et al. (2016); mapped to the indicated standard where applicable. [d] Retention of nitrogenous waste, defined in the KDIGO Clinical Practice Guideline for Acute Kidney Injury (2012). [e] Threshold of $>$ 2 mmol/L in the septic shock criteria of Singer et al. (2016). [f] Platelet count, prothrombin time and fibrinogen constitute the global coagulation tests of the overt disseminated intravascular coagulation algorithm of Taylor et al. (2001); the activated partial thromboplastin time is not part of that score. [g] Fields that are not measured physiology. [h] Binary indicators of sepsis and acute kidney injury within the first 24 hours after ICU admission, derived from the same observation window as the other features using each cohort's labeling criteria. For mortality prediction, all time-series cohorts include only stays of at least 30 hours, so these indicators precede the outcome.

D.3 Breast Cancer Datasets

Task Definition. We define the outcome for each breast cancer cohort as a binary classification of disease-free survival (DFS), where an event is the earliest of local or distant recurrence, second primary cancer, or death, and the positive class denotes an adverse prognosis. Each cohort is derived from an adjuvant clinical trial in Project Data Sphere, retaining the baseline clinical variables recorded at enrollment as features and excluding treatment-arm assignments and post-treatment measurements. The DFS endpoint is obtained from each trial's recurrence and death records; where a precomputed DFS status is unavailable, we reconstruct it from the recorded dates, harmonizing their units to days. Since raw survival times are right-censored, we binarize DFS at a landmark time $T = 24$ months: a patient is labeled positive if an event occurs within T, and negative if event-free follow-up extends beyond T.

Exclusion Criteria. Patients censored before the landmark are excluded, as their status at T is undetermined; retaining them as negatives would treat unobserved events as event-free outcomes. Wang & Sun (2022) instead record censored patients as alive and use mortality rather than DFS as the endpoint, so our cohorts differ from theirs in both size and target for the trials in common. Per-cohort statistics after this exclusion are reported in Table 19.

Table 19: Statistics of the breast cancer datasets.

Dataset	NCT ID	Samples	Features	Cat.	Num.	Pos. rate
C1	NCT00041119	3646	23	23	0	0.039
C2	NCT00174655	438	23	19	4	0.096
C3	NCT00312208	1602	16	12	4	0.077
C4	NCT00688740	739	10	7	3	0.164

The cohorts differ in how the disease-free survival (DFS) endpoint is obtained:

- **C1** (NCT00041119)[24]: the trial provides a precomputed DFS status, where an event is local or distant recurrence, or death without recurrence.
- **C2** (NCT00174655)[25]: no precomputed status is available, so we reconstruct the endpoint from the recurrence and death records, harmonizing their units to days. Every recorded death follows a prior recurrence, so the events are recurrences.
- **C3** (NCT00312208)[26]: the endpoint is reconstructed from the recurrence and death records, taking the earliest of the two as the event.
- **C4** (NCT00688740)[27]: the trial provides a survival record, where an event is breast cancer relapse, second primary malignancy, or death without relapse.

E Full Results

E.1 Full Basis Profiles

[24] https://data.projectdatasphere.org/projectdatasphere/html/content/596a13c5-62a3-4f01-a53e-32966ef6bfe8

[25] https://data.projectdatasphere.org/projectdatasphere/html/content/40d9ee0c-a9ed-46bc-b09e-a937c1530223

[26] https://data.projectdatasphere.org/projectdatasphere/html/content/45260b6c-8df0-4249-a5a0-956cb8a26c16

[27] https://data.projectdatasphere.org/projectdatasphere/html/content/c28e5975-8936-47cc-b81b-78cca652c2f3



Figure 15: Full variable-level profiles of the high-risk bases in the cardiovascular benchmark. Each row shows one basis and the clinical categories it spans; high-risk denotes samples assigned to the basis by the highest coordinate. Arrows indicate the clinically adverse direction.

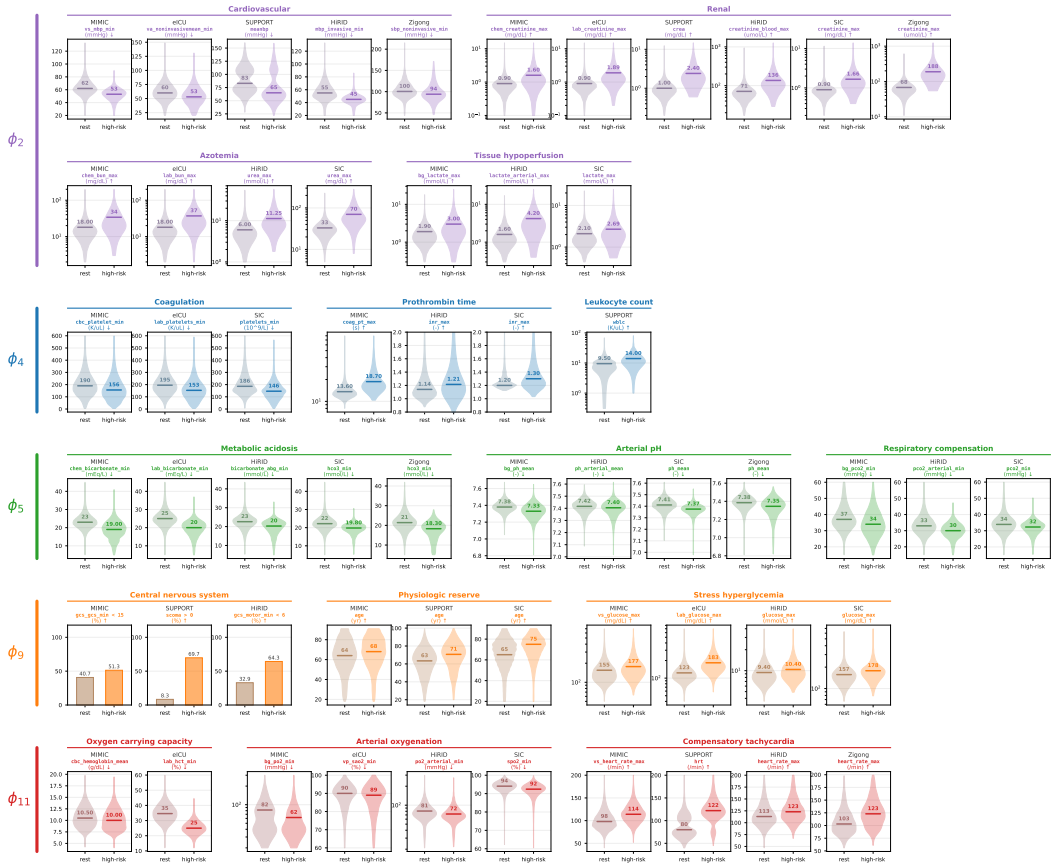

Figure 16: Full variable-level profiles of the high-risk bases in the intensive care benchmark. Each row shows one basis and the clinical categories it spans; high-risk denotes samples assigned to the basis by the highest coordinate. Arrows indicate the clinically adverse direction.

E.2 CRITICAL DIFFERENCE DIAGRAM

We compare the models by their average rank across the few-shot settings of each benchmark, following the procedure of Demšar (2006). For each (target domain, shot count) pair we rank the twelve models by their performance, and average these ranks over all pairs; AUROC and AUPRC are ranked separately, and the full-data setting is excluded. This gives 30 pairs for the cardiovascular and intensive care benchmarks (six domains, five shot counts) and 20 for the breast cancer benchmark (four domains). Models whose average ranks differ by less than the critical difference of the Nemenyi test ($\alpha = 0.05$) are connected by a horizontal bar; the critical difference is 3.04 for the former two benchmarks and 3.72 for the latter. Figures 17, 18 and 19 show the resulting diagrams.

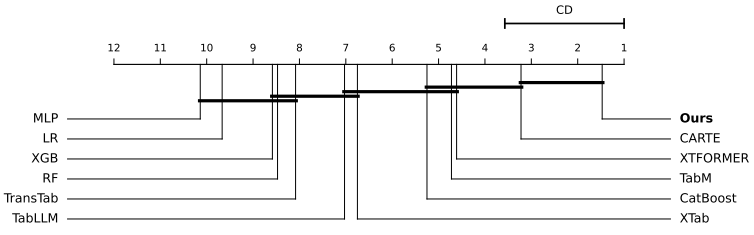

Figure 17: Critical difference diagram over cardiovascular datasets.

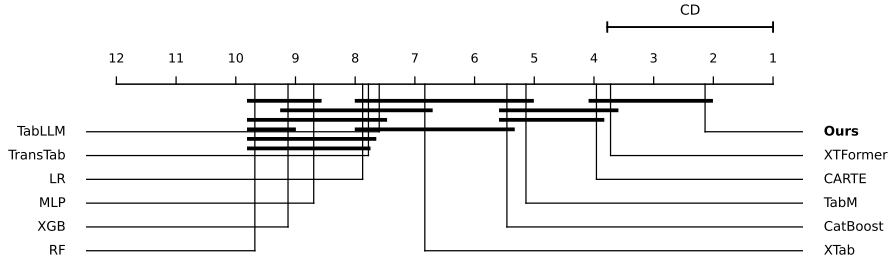

Figure 18: Critical difference diagram over ICU datasets.

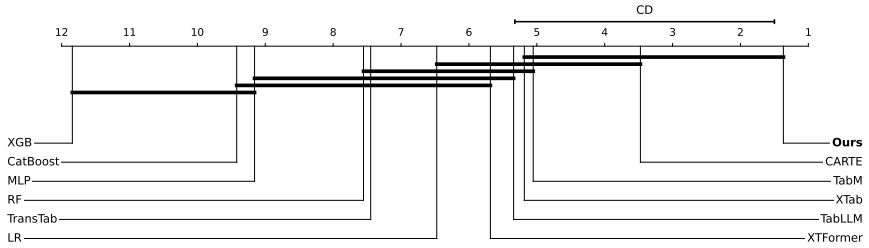

Figure 19: Critical difference diagram over breast cancer datasets.

E.3 FULL FINETUNING RESULTS

Table 20: Full evaluation results on the cardiovascular and ICU datasets. AUROC and AUPRC are reported for each dataset, with the best result in **bold** and the second best underlined.

		Cardiovascular						ICU					
Model	Metric	MED	CDD	HD	ERB	SAH	HF	MIMIC	eICU	HiRiD	SUPPORT	SicDB	Zigong
LR	AUROC	89.9±5.68	89.4±3.67	91.9±3.34	93.8±3.29	75.6±3.57	85.1±3.03	86.0±0.07	80.4±0.24	85.7±0.43	83.9±0.35	88.3±0.77	72.9±2.59
	AUPRC	92.7±5.23	90.7±2.01	85.3±2.45	94.6±3.00	64.8±4.37	73.7±3.53	48.9±0.33	33.0±0.26	41.4±0.95	64.3±1.17	34.2±2.42	21.5±4.03
RF	AUROC	94.6±3.27	90.8±2.71	88.9±2.70	94.6±1.17	71.8±4.71	84.2±3.15	86.1±0.20	80.4±0.14	86.4±0.51	85.9±0.23	88.7±0.49	73.8±2.96
	AUPRC	91.7±2.19	92.1±1.63	89.8±1.97	95.8±2.30	62.0±4.90	72.8±3.80	48.7±0.40	30.2±0.22	41.3±1.51	62.6±0.67	33.0±1.97	25.0±2.81
XGB	AUROC	94.3±4.25	92.6±2.26	91.8±3.16	96.6±2.00	75.5±3.73	88.3±2.79	86.9±0.12	82.2±0.18	85.6±0.50	85.2±0.32	87.5±0.45	72.3±3.13
	AUPRC	94.0±4.18	95.2±2.51	92.5±1.69	94.8±2.53	63.1±4.37	82.1±5.54	52.4±0.35	36.4±0.58	45.2±1.08	62.6±0.76	35.2±0.68	21.9±2.66
CatBoost	AUROC	95.3±2.72	93.3±2.13	90.4±2.98	95.5±3.04	75.1±3.75	88.2±2.89	87.4±0.09	82.6±0.25	86.7±0.49	86.0±1.49	89.3±0.28	76.0±5.29
	AUPRC	95.4±1.91	95.1±0.93	91.2±1.02	94.7±2.27	63.2±5.07	82.7±4.65	53.9±0.21	37.1±0.73	47.3±1.52	64.0±1.11	38.6±1.31	27.7±3.56
MLP	AUROC	92.2±1.94	90.9±4.12	87.5±3.25	93.1±3.86	72.9±4.80	83.3±2.97	83.0±0.25	79.3±0.13	81.7±1.29	82.2±0.70	82.7±0.82	68.0±3.60
	AUPRC	92.1±3.13	92.9±3.64	84.1±3.11	92.1±3.49	66.4±5.33	73.1±5.15	46.1±0.32	32.4±0.30	39.8±1.50	57.4±1.80	29.8±0.54	20.1±4.14
TabM	AUROC	96.2±1.42	93.9±2.96	89.2±3.33	95.3±1.69	76.8±3.10	88.1±2.93	87.0±0.35	82.9±0.28	87.1±0.88	86.4±1.37	87.8±0.75	77.3±3.65
	AUPRC	96.2±1.84	95.3±2.52	89.8±3.05	**96.7**±1.72	66.7±4.33	81.5±3.88	54.3±1.14	37.7±0.74	52.1±1.60	63.4±2.79	42.4±1.69	45.1±7.52
CARTE	AUROC	96.5±2.19	97.4±2.72	95.3±2.08	97.6±2.22	74.2±4.08	91.5±2.85	88.4±0.33	82.6±0.40	88.9±0.70	87.5±1.05	**91.8**±0.55	79.2±4.09
	AUPRC	96.5±1.83	96.8±1.77	95.7±2.06	93.9±2.27	65.9±4.15	85.2±4.41	56.6±1.70	37.1±1.15	54.6±1.55	66.6±2.81	44.9±2.23	47.4±5.87
XTFormer	AUROC	94.3±2.46	93.9±1.84	91.9±2.43	93.9±2.61	75.3±4.43	91.6±1.85	87.6±0.39	84.6±0.33	88.3±0.81	86.4±1.58	91.2±0.59	78.5±3.96
	AUPRC	93.0±3.41	94.8±3.15	92.7±2.35	94.3±2.45	65.1±4.67	85.7±4.23	55.2±1.00	37.9±0.66	53.9±1.33	66.5±2.22	**45.2**±1.72	47.0±6.70
TabLLM	AUROC	92.3±2.40	90.4±5.17	88.7±2.73	96.2±1.69	**79.3**±3.96	86.8±2.98	87.0±0.50	83.5±0.26	88.7±0.95	85.1±2.48	86.5±1.48	77.1±4.79
	AUPRC	92.1±2.06	90.8±3.37	88.2±3.05	95.1±1.84	**72.9**±6.18	80.4±5.36	52.6±1.45	37.2±0.63	47.0±2.30	64.0±3.95	34.1±4.75	28.3±9.30
XTab	AUROC	90.8±2.62	92.4±3.26	89.3±2.25	95.3±2.34	72.5±4.25	89.3±2.43	88.0±0.39	82.1±0.56	86.6±1.67	84.8±1.45	88.8±1.11	69.4±5.99
	AUPRC	91.0±2.46	93.8±2.92	90.3±1.39	94.3±2.56	64.0±5.43	82.7±4.44	54.4±0.99	37.4±0.97	45.3±1.65	61.2±2.33	33.4±4.43	23.1±8.43
TransTab	AUROC	89.8±2.84	93.4±5.60	87.3±2.65	95.1±3.40	71.5±4.01	87.9±3.00	87.6±0.45	82.2±0.34	85.7±0.70	86.2±1.15	87.2±0.72	81.3±2.55
	AUPRC	90.0±2.94	89.8±4.57	88.3±2.62	93.3±2.62	66.0±4.60	81.2±4.42	52.4±1.70	37.2±0.45	43.5±2.40	63.2±3.52	36.2±2.63	26.8±6.75
Ours	AUROC	**97.5**±2.78	**98.7**±1.59	**96.7**±1.35	**97.7**±1.56	76.2±3.22	**92.2**±2.71	**91.1**±0.34	**86.0**±0.20	**90.7**±0.83	**90.8**±1.19	89.4±0.69	**83.7**±2.71
	AUPRC	**97.5**±2.48	**97.8**±0.89	**97.1**±1.11	96.3±2.46	67.1±4.41	**86.1**±4.82	**57.4**±1.84	**41.2**±0.52	**55.7**±1.68	**68.7**±1.16	44.2±2.48	**50.1**±8.47

Table 21: Full evaluation results on the breast cancer datasets. AUROC and AUPRC are reported for each dataset, with the best result in **bold** and the second best underlined.

Model	Metric	c1	c2	c3	c4
LR	AUROC	70.97±1.42	69.50±6.69	75.07±6.35	69.52±3.24
	AUPRC	10.68±1.87	26.05±8.23	22.13±3.24	32.27±3.93
RF	AUROC	69.56±3.85	69.12±3.59	74.67±4.10	62.91±5.83
	AUPRC	10.19±2.39	18.49±2.92	21.23±6.10	25.99±4.99
XGB	AUROC	64.67±3.98	62.31±6.51	66.79±4.72	62.47±4.85
	AUPRC	8.06±2.54	15.15±3.64	16.47±3.31	25.14±1.94
CatBoost	AUROC	69.82±2.61	67.84±4.40	75.23±4.63	68.57±5.84
	AUPRC	8.62±1.31	24.56±7.79	21.71±4.60	30.64±6.41
MLP	AUROC	61.87±6.10	67.38±6.87	73.26±5.39	68.42±4.96
	AUPRC	6.23±2.28	20.71±11.00	22.47±7.49	30.09±5.70
TabM	AUROC	71.37±2.23	70.48±3.41	76.28±5.07	68.16±4.70
	AUPRC	10.78±1.06	26.47±8.46	24.32±5.53	28.14±4.98
CARTE	AUROC	72.15±2.92	72.84±3.16	76.48±4.27	70.48±1.94
	AUPRC	12.37±1.93	30.54±6.58	23.78±4.31	33.92±2.76
XTFormer	AUROC	73.86±1.56	68.92±6.24	73.45±5.08	67.38±3.17
	AUPRC	10.15±1.38	23.14±5.17	19.68±5.48	34.16±7.21
TabLLM	AUROC	70.89±3.47	68.93±5.68	74.23±5.23	68.32±4.13
	AUPRC	10.42±2.73	24.38±7.62	21.89±5.17	29.47±6.23
XTab	AUROC	71.83±2.06	71.04±3.32	76.14±4.91	68.93±3.78
	AUPRC	12.68±2.87	28.94±8.47	23.47±4.59	29.67±3.24
TransTab	AUROC	71.24±2.25	69.65±1.48	77.26±5.90	69.12±5.72
	AUPRC	11.94±3.31	32.25±9.73	23.61±4.83	33.58±8.76
Ours	AUROC	**74.82**±2.37	**75.17**±2.81	**78.93**±3.96	**71.26**±2.13
	AUPRC	**13.82**±1.76	**32.87**±5.74	**25.86**±4.02	**35.48**±2.43

E.4 FULL CARDIOVASCULAR FEW-SHOT RESULTS.

Table 22: Few-shot results on the six cardiovascular datasets. AUROC and AUPRC are reported for each shot count, with the best result in **bold** and the second best underlined.

Model	Metric	Few-shot					
		0	4	8	16	32	64
Medicaldataset (MED)							
LR	AUROC	–	68.6±14.39	74.3±9.05	78.7±6.41	82.2±2.76	86.3±3.29
	AUPRC	–	68.0±16.58	77.4±6.41	79.3±8.12	85.3±3.87	90.8±3.97
RF	AUROC	–	70.2±12.51	81.6±11.52	85.3±4.47	88.9±3.43	91.1±2.3
	AUPRC	–	70.5±9.02	80.9±6.25	**90.2**±4.04	91.9±3.04	91.2±3.03
XGB	AUROC	–	57.3±9.7	68.3±4.6	83.3±6.01	88.4±4.57	91.6±5.37
	AUPRC	–	62.4±9.98	71.9±7.13	82.3±3.19	90.3±3.52	90.6±4.35
CatBoost	AUROC	–	70.4±14.93	83.6±10.53	88.0±6.3	91.1±5.22	93.7±2.96
	AUPRC	–	71.3±12.34	83.0±8.86	**90.2**±8.17	92.0±5.99	93.2±3.38
MLP	AUROC	–	74.1±13.06	76.2±8.06	78.8±7.41	82.0±6.04	85.3±3.45
	AUPRC	–	73.4±11.46	79.6±5.28	80.7±4.83	84.7±6.99	89.0±4.83
TabM	AUROC	–	81.2±6.38	84.6±7.26	87.9±6.79	91.3±6.59	93.9±4.91
	AUPRC	–	81.8±5.44	85.0±9.52	89.9±3.88	92.3±6.51	**96.8**±3.63
CARTE	AUROC	–	79.0±8.73	83.2±10.0	86.0±5.26	88.2±6.09	91.6±2.87
	AUPRC	–	77.7±9.89	84.5±9.53	90.1±3.88	89.5±5.97	92.2±3.71
XTFormer	AUROC	–	72.2±11.98	78.3±6.97	82.2±4.78	89.2±3.81	92.2±5.26
	AUPRC	–	77.6±9.59	78.5±7.08	85.6±4.33	92.1±5.09	95.3±4.02
TabLLM	AUROC	74.4±8.59	77.2±11.08	81.2±4.21	84.6±7.77	88.3±4.76	91.1±4.34
	AUPRC	77.8±7.41	80.4±9.23	82.0±6.68	87.8±5.2	90.3±4.61	93.4±3.69
XTab	AUROC	–	75.3±6.04	79.3±11.4	82.1±3.34	85.2±2.55	89.0±3.33
	AUPRC	–	79.3±9.13	83.8±8.79	83.0±4.74	88.7±4.56	93.0±2.83
TransTab	AUROC	75.3±4.55	76.4±8.87	78.2±5.32	83.1±5.32	84.2±6.97	87.9±2.43
	AUPRC	77.8±7.5	79.0±8.53	82.7±4.92	85.6±5.85	85.0±3.98	89.2±3.13
Ours	AUROC	**80.7**±4.48	**81.3**±5.78	**85.9**±4.38	**88.8**±4.44	**91.5**±2.78	**95.1**±2.67
	AUPRC	**82.4**±8.33	**82.6**±9.68	**87.9**±6.82	88.1±5.88	**93.8**±3.38	94.9±2.78
Cardiovascular Disease (CDD)							
LR	AUROC	–	72.0±10.75	71.3±9.07	82.1±4.15	84.7±4.21	88.3±2.98
	AUPRC	–	72.2±9.34	71.9±4.98	83.7±5.08	86.2±6.65	92.5±5.3
RF	AUROC	–	79.0±7.0	84.3±11.65	83.8±8.48	84.4±2.81	86.7±3.12
	AUPRC	–	77.4±8.69	86.8±9.23	87.9±8.11	89.0±3.53	90.8±5.06
XGB	AUROC	–	51.0±12.64	57.3±4.81	80.3±5.2	86.2±6.84	90.1±5.25
	AUPRC	–	60.6±11.26	61.0±6.21	84.1±4.24	87.6±4.7	92.4±5.22
CatBoost	AUROC	–	81.7±6.92	**85.2**±8.3	85.8±6.2	87.7±5.35	90.4±5.48
	AUPRC	–	86.6±6.53	87.4±7.27	88.6±7.09	88.8±4.25	91.9±6.62
MLP	AUROC	–	68.3±10.97	76.3±11.11	81.3±3.26	83.0±5.38	86.2±2.08
	AUPRC	–	68.3±8.13	74.0±8.45	87.2±4.88	88.6±4.11	91.7±1.82
TabM	AUROC	–	82.3±11.3	83.4±10.69	84.6±3.62	87.0±4.44	88.2±5.18
	AUPRC	–	**86.9**±9.17	88.2±7.21	88.6±4.15	87.4±6.34	93.2±3.51
CARTE	AUROC	–	82.4±12.31	84.8±7.65	**86.9**±8.73	**93.2**±5.64	**95.9**±2.55
	AUPRC	–	84.6±9.81	83.9±7.77	85.3±5.67	**92.2**±3.61	**97.0**±2.24
XTFormer	AUROC	–	81.6±8.85	84.8±10.58	86.0±4.16	88.0±4.14	90.6±3.13
	AUPRC	–	85.7±8.01	86.2±7.68	**90.6**±4.16	**92.2**±4.45	91.7±5.08
TabLLM	AUROC	75.9±8.41	80.3±9.71	81.2±6.25	81.8±4.06	83.1±2.74	86.6±2.63
	AUPRC	79.1±3.85	79.7±6.65	80.8±9.13	85.9±4.08	87.4±2.65	90.9±3.17
XTab	AUROC	–	75.3±6.49	81.8±7.43	82.9±4.81	85.4±4.62	87.8±2.04
	AUPRC	–	78.0±9.51	80.5±4.32	84.4±8.23	87.1±3.94	90.0±3.38
TransTab	AUROC	73.1±8.87	75.0±11.32	82.3±5.91	82.6±5.31	84.3±6.24	86.7±4.31

Continued on next page

Table 22 – continued from previous page

Model	Metric	Few-shot					
		0	4	8	16	32	64
	AUPRC	72.4±5.64	74.3±8.0	86.4±6.29	87.5±8.69	88.8±6.04	87.9±4.01
Ours	AUROC	**81.4**±5.11	**84.0**±6.27	**85.2**±5.36	86.7±4.38	89.3±3.25	92.9±2.92
	AUPRC	**79.5**±7.23	84.9±8.17	**89.0**±7.67	88.6±6.65	89.7±3.5	94.2±3.49
Heart (HD)							
LR	AUROC	–	70.2±9.52	73.9±10.63	78.1±6.99	81.7±3.45	86.1±2.01
	AUPRC	–	73.5±7.47	73.9±8.44	80.4±6.89	82.3±2.77	89.6±2.74
RF	AUROC	–	69.4±13.62	70.4±6.3	78.3±7.09	83.5±4.05	87.1±5.15
	AUPRC	–	67.2±8.84	69.5±4.29	75.9±5.21	85.1±6.01	87.4±3.75
XGB	AUROC	–	57.0±7.14	71.3±9.52	78.5±3.04	86.8±3.75	87.7±2.4
	AUPRC	–	60.0±8.44	75.2±8.81	77.2±4.15	85.1±2.9	85.7±2.82
CatBoost	AUROC	–	70.4±5.88	72.6±9.94	80.7±5.38	84.3±6.03	87.0±2.22
	AUPRC	–	71.0±8.41	77.2±8.94	84.7±6.7	84.7±4.28	89.7±2.92
MLP	AUROC	–	69.9±10.04	71.4±8.38	78.3±7.19	85.4±3.25	86.4±5.24
	AUPRC	–	73.6±6.92	74.4±6.09	79.7±7.09	88.1±5.59	89.5±6.05
TabM	AUROC	–	70.3±10.88	75.3±8.57	80.4±3.07	84.3±4.81	86.5±2.86
	AUPRC	–	74.3±8.46	77.4±6.84	80.2±2.67	86.5±5.62	85.6±4.75
CARTE	AUROC	–	78.6±13.12	81.7±5.3	84.7±5.87	87.1±6.58	90.9±4.82
	AUPRC	–	77.6±7.14	**83.9**±7.39	84.5±5.18	85.6±6.8	90.0±4.46
XTFormer	AUROC	–	74.4±14.58	77.7±5.37	83.2±4.79	86.4±5.23	88.5±4.96
	AUPRC	–	72.3±9.94	78.4±9.38	**85.8**±8.25	85.2±5.65	89.0±5.43
TabLLM	AUROC	69.2±6.73	70.3±8.76	79.2±8.75	84.3±8.11	86.1±3.83	85.2±5.49
	AUPRC	70.4±4.67	71.8±7.07	82.1±9.26	83.4±5.96	**89.1**±5.69	88.7±3.29
XTab	AUROC	–	68.4±10.07	79.4±9.37	81.3±7.42	85.2±5.18	86.2±3.48
	AUPRC	–	72.6±9.37	81.0±8.5	84.3±7.41	84.2±3.73	85.6±3.91
TransTab	AUROC	67.1±7.33	68.0±10.04	77.3±11.21	79.2±7.08	83.3±6.72	84.2±4.76
	AUPRC	66.1±7.38	70.9±8.99	75.1±9.87	81.1±7.37	84.1±6.11	83.4±5.45
Ours	AUROC	**78.9**±3.74	**79.1**±7.31	**82.8**±5.92	**85.3**±3.9	**89.4**±3.25	**91.3**±2.37
	AUPRC	**73.5**±4.64	**75.8**±9.06	80.1±5.02	84.7±4.13	86.5±2.92	**90.1**±2.75
Erbil Cardio (ERB)							
LR	AUROC	–	58.3±12.55	61.4±10.94	73.3±3.64	79.9±5.15	89.8±2.55
	AUPRC	–	46.4±7.93	54.6±12.19	65.4±5.51	68.5±6.04	78.8±2.55
RF	AUROC	–	67.2±10.68	80.8±9.85	83.3±5.91	85.8±5.87	89.9±5.04
	AUPRC	–	60.7±7.68	71.3±8.67	77.8±8.91	79.5±4.85	78.5±4.21
XGB	AUROC	–	57.1±6.36	81.4±6.49	85.5±4.47	86.2±2.66	90.6±3.14
	AUPRC	–	39.5±3.48	72.9±6.82	74.9±3.73	80.2±2.97	83.5±4.0
CatBoost	AUROC	–	67.9±13.2	79.1±7.79	88.7±7.84	91.9±5.64	93.6±2.66
	AUPRC	–	58.7±7.35	67.7±8.06	75.8±5.18	85.9±5.16	87.5±2.71
MLP	AUROC	–	59.2±12.53	63.4±10.46	74.3±8.1	81.4±3.13	83.4±4.22
	AUPRC	–	52.1±9.83	50.8±9.57	68.5±8.43	68.9±3.49	78.3±5.32
TabM	AUROC	–	81.3±11.28	83.0±4.21	85.7±7.98	88.1±3.2	92.8±2.8
	AUPRC	–	70.5±6.42	72.5±3.59	80.0±7.08	80.2±3.24	84.0±3.22
CARTE	AUROC	–	78.8±8.42	85.2±8.22	91.8±8.02	93.3±6.01	96.2±4.46
	AUPRC	–	69.8±8.56	**78.8**±8.3	79.4±5.66	83.5±3.6	85.6±5.35
XTFormer	AUROC	–	**82.4**±10.26	84.3±4.81	86.3±7.56	89.0±4.11	91.8±5.03
	AUPRC	–	**75.0**±9.33	75.6±5.14	75.4±5.03	78.8±3.69	85.9±5.26
TabLLM	AUROC	77.1±7.35	79.0±14.05	80.3±11.1	87.2±5.59	88.3±5.67	91.2±4.75
	AUPRC	**68.7**±5.02	73.7±8.72	73.4±6.88	80.7±5.13	80.5±5.92	80.8±5.13
XTab	AUROC	–	75.2±5.65	78.2±10.83	81.6±3.78	84.6±6.15	89.6±3.75
	AUPRC	–	62.8±6.29	72.4±8.43	71.7±4.24	74.8±4.79	76.9±2.87
TransTab	AUROC	72.3±5.87	75.3±9.66	77.1±11.64	80.1±5.7	84.1±6.97	89.2±3.18

Continued on next page

Table 22 – continued from previous page

Model	Metric	Few-shot					
		0	4	8	16	32	64
	AUPRC	64.6±3.48	65.6±6.82	66.5±6.83	68.7±4.21	78.6±5.92	82.8±4.39
Ours	AUROC	78.7±5.82	79.3±7.01	**86.1**±5.92	**93.3**±4.99	**96.1**±3.62	**96.4**±3.28
	AUPRC	67.6±3.72	68.0±4.28	76.1±9.55	**84.9**±4.67	**87.9**±4.18	**91.1**±2.48
Cardio SAheart (SAH)							
LR	AUROC	–	57.1±9.12	56.9±4.78	63.9±7.93	65.4±5.21	67.4±5.16
	AUPRC	–	47.9±8.87	48.8±7.6	55.3±7.61	54.7±6.39	54.2±4.57
RF	AUROC	–	60.0±10.86	60.6±6.2	65.3±5.1	64.2±4.67	69.5±2.52
	AUPRC	–	49.6±5.84	49.9±4.12	55.1±4.74	55.0±5.16	56.9±3.52
XGB	AUROC	–	57.9±9.41	58.2±7.41	63.8±4.75	64.9±4.91	66.0±4.74
	AUPRC	–	38.6±10.2	50.2±9.63	57.7±6.6	59.1±3.01	60.3±5.63
CatBoost	AUROC	–	56.4±6.66	65.9±6.24	67.4±6.29	68.8±4.99	70.2±3.3
	AUPRC	–	48.4±8.8	57.0±5.21	59.1±7.75	60.3±2.85	62.3±2.85
MLP	AUROC	–	60.5±5.27	63.1±6.89	65.1±7.26	67.0±6.83	71.5±3.91
	AUPRC	–	50.4±7.26	52.1±8.73	55.1±4.84	60.3±5.65	59.4±3.8
TabM	AUROC	–	69.4±6.56	68.7±8.94	68.7±3.46	69.7±4.24	72.2±4.19
	AUPRC	–	56.8±8.98	56.3±7.48	57.0±5.9	57.9±5.77	60.7±2.84
CARTE	AUROC	–	66.2±9.34	65.5±10.73	67.2±5.35	69.9±3.13	70.2±5.31
	AUPRC	–	57.9±9.06	56.4±5.98	56.0±6.64	60.2±3.42	**63.6**±4.95
XTFormer	AUROC	–	65.2±7.61	64.1±9.43	65.3±4.37	67.2±3.62	68.3±5.46
	AUPRC	–	57.9±8.1	59.7±9.62	58.4±6.22	61.0±3.28	61.9±5.05
TabLLM	AUROC	64.3±6.18	65.3±10.04	64.2±4.4	65.4±3.12	67.4±4.39	68.2±3.12
	AUPRC	**56.6**±5.12	58.5±8.08	58.4±4.04	57.6±5.52	58.0±2.66	61.8±5.05
XTab	AUROC	–	66.9±14.91	67.2±9.41	67.0±5.14	69.6±5.45	72.4±3.03
	AUPRC	–	56.8±9.2	57.7±5.8	57.1±3.92	60.3±6.36	60.9±3.2
TransTab	AUROC	61.3±5.15	65.9±6.87	67.2±6.03	66.0±5.65	71.6±3.78	70.3±4.91
	AUPRC	54.6±3.98	56.4±4.72	58.2±9.73	57.5±7.27	60.7±6.72	61.2±3.24
Ours	AUROC	**69.5**±5.41	**69.8**±7.82	**70.7**±4.81	**71.9**±4.68	**72.2**±3.24	**73.2**±2.62
	AUPRC	53.9±4.21	**62.8**±5.73	**61.7**±6.39	**62.5**±5.12	**61.5**±5.27	62.5±4.54
Heart Failure (HF)							
LR	AUROC	–	61.4±6.15	64.7±8.71	71.2±3.85	77.0±3.41	82.8±4.84
	AUPRC	–	52.4±6.26	52.2±9.93	54.7±3.24	62.3±3.26	74.8±4.98
RF	AUROC	–	64.5±12.27	68.8±10.86	72.4±4.67	77.8±5.54	79.3±2.96
	AUPRC	–	54.8±9.37	54.5±9.13	60.8±5.29	67.7±6.66	67.3±5.0
XGB	AUROC	–	55.3±12.62	58.8±4.93	72.4±8.84	79.2±5.3	83.5±2.85
	AUPRC	–	33.9±10.38	42.6±7.74	57.3±7.96	66.6±5.27	67.0±4.24
CatBoost	AUROC	–	72.4±9.64	73.5±4.56	76.1±5.7	81.9±4.58	85.0±2.65
	AUPRC	–	61.9±7.18	64.4±6.9	63.5±5.68	70.7±5.58	72.1±3.74
MLP	AUROC	–	59.1±8.42	65.8±8.16	69.9±4.42	75.4±4.44	78.1±3.51
	AUPRC	–	45.6±7.71	57.5±8.08	60.1±5.17	66.4±4.54	66.8±3.41
TabM	AUROC	–	65.6±7.27	70.2±7.8	76.8±5.73	81.5±4.85	84.4±2.93
	AUPRC	–	52.4±8.24	60.0±5.05	62.4±5.58	72.5±6.51	72.3±4.68
CARTE	AUROC	–	73.7±6.11	75.3±4.68	79.4±7.59	84.6±4.21	87.0±2.97
	AUPRC	–	**63.8**±8.71	64.1±7.22	63.1±8.05	68.6±3.76	78.6±4.86
XTFormer	AUROC	–	**74.7**±12.25	**76.9**±11.99	**80.9**±7.1	85.0±5.52	88.8±2.31
	AUPRC	–	58.4±12.92	61.9±6.5	63.1±5.15	74.7±4.42	77.0±2.52
TabLLM	AUROC	67.1±4.32	69.2±5.23	71.6±7.37	76.3±8.95	79.8±5.9	83.2±2.94
	AUPRC	**55.5**±3.1	58.7±5.2	62.5±9.37	**67.0**±7.08	70.8±3.49	70.8±2.6
XTab	AUROC	–	70.3±11.73	70.9±6.89	78.1±5.86	82.9±3.19	85.2±2.42
	AUPRC	–	55.7±9.01	54.7±7.47	61.5±8.47	72.6±3.43	76.0±4.18
TransTab	AUROC	64.2±5.01	68.9±6.17	71.3±6.34	76.5±8.36	81.0±5.94	83.8±2.4

Continued on next page

Table 22 – continued from previous page

Model	Metric	Few-shot					
		0	4	8	16	32	64
	AUPRC	46.3±5.51	60.2±8.58	58.8±5.57	66.1±7.38	65.5±5.22	72.7±2.84
Ours	AUROC	**71.8**±6.38	74.1±7.31	76.3±5.3	80.3±3.93	**85.5**±3.89	**89.1**±2.25
	AUPRC	54.4±5.62	63.6±9.56	**64.8**±8.81	63.9±4.75	**75.8**±6.09	**80.7**±2.62

E.5 Full ICU Few-shot Results

Table 23: Few-shot results on the six ICU datasets. AUROC and AUPRC are reported for each shot count, with the best result in **bold** and the second best underlined.

Model	Metric	Few-shot					
		0	4	8	16	32	64
MIMIC-IV (MIMIC)							
LR	AUROC	–	61.8±8.23	64.3±6.16	69.6±3.56	73.3±2.21	73.8±2.66
	AUPRC	–	17.4±5.35	25.0±5.10	27.5±3.86	28.6±2.61	35.0±2.94
RF	AUROC	–	59.9±5.28	63.2±4.86	68.9±3.89	72.3±2.77	72.8±1.97
	AUPRC	–	20.7±6.44	24.3±4.43	26.6±4.04	32.1±2.80	34.9±2.50
XGB	AUROC	–	57.0±4.35	63.4±4.54	69.2±3.02	73.3±2.82	75.0±1.72
	AUPRC	–	13.7±1.12	20.4±2.93	25.4±2.44	30.5±2.95	35.3±2.85
CatBoost	AUROC	–	61.8±7.24	69.1±4.98	70.1±3.38	74.7±1.67	76.2±1.04
	AUPRC	–	20.7±5.20	24.6±4.63	28.6±4.19	33.8±2.91	36.8±1.72
MLP	AUROC	–	61.6±7.18	62.8±5.80	70.2±2.60	72.6±1.66	74.0±1.59
	AUPRC	–	18.3±4.85	23.7±5.05	28.2±3.38	29.8±2.47	33.2±2.35
TabM	AUROC	–	63.6±8.03	68.1±5.54	73.0±3.52	76.4±2.08	77.4±1.75
	AUPRC	–	20.6±5.33	24.4±4.79	27.6±3.61	33.2±2.60	36.6±2.83
CARTE	AUROC	–	69.2±6.38	72.5±2.33	76.4±1.77	79.0±1.41	80.9±1.02
	AUPRC	–	24.5±4.73	**28.7**±3.23	**31.8**±2.57	34.9±2.36	37.2±1.75
XTFormer	AUROC	–	65.2±7.45	70.0±4.84	75.7±2.87	78.5±2.13	79.5±1.54
	AUPRC	–	21.8±5.45	25.7±4.99	30.4±3.69	33.9±3.28	37.0±2.23
TabLLM	AUROC	64.4±7.15	64.6±5.57	69.1±4.83	73.2±3.87	76.6±2.76	77.5±1.86
	AUPRC	19.0±6.40	19.6±6.81	23.1±6.46	26.8±3.59	28.5±3.11	32.0±1.65
XTab	AUROC	–	66.0±5.05	68.9±4.91	71.6±3.39	74.3±2.90	76.2±2.57
	AUPRC	–	22.9±5.68	26.1±5.82	27.8±3.27	29.7±2.87	34.5±2.36
TransTab	AUROC	61.9±8.42	62.1±6.59	67.8±4.13	72.2±2.45	75.2±1.89	75.4±1.16
	AUPRC	18.1±6.87	19.3±5.68	24.8±4.61	29.7±3.21	32.8±2.94	35.4±2.50
Ours	AUROC	**73.6**±5.38	**74.8**±4.17	**77.0**±3.41	**78.9**±2.62	**81.2**±1.89	**83.4**±1.25
	AUPRC	**24.8**±5.14	**26.4**±5.28	28.7±4.36	31.6±3.05	**34.9**±2.28	**38.2**±1.47
eICU							
LR	AUROC	–	60.7±7.35	61.8±5.21	65.7±3.14	67.7±3.42	70.0±2.80
	AUPRC	–	13.3±3.67	18.0±2.56	18.7±2.69	18.5±2.87	21.6±2.74
RF	AUROC	–	59.3±4.46	60.4±3.67	64.5±3.03	65.7±2.95	66.6±2.30
	AUPRC	–	15.0±3.20	17.2±2.65	17.8±2.89	20.1±2.45	22.8±2.27
XGB	AUROC	–	57.6±4.83	61.0±3.53	65.0±2.95	66.9±2.99	67.7±2.57
	AUPRC	–	10.8±1.14	16.1±2.01	19.3±2.11	22.3±2.17	25.1±2.18
CatBoost	AUROC	–	61.6±6.40	65.9±4.89	65.3±3.37	68.7±2.29	71.5±1.72
	AUPRC	–	16.3±3.06	17.7±2.54	20.1±2.52	21.1±2.22	22.4±2.01
MLP	AUROC	–	59.7±5.63	61.1±5.64	65.3±3.81	67.4±3.13	70.7±1.70
	AUPRC	–	13.2±2.95	17.7±2.94	20.1±2.59	18.6±2.60	21.1±2.06
TabM	AUROC	–	64.1±5.66	65.5±6.40	69.3±3.36	70.1±3.03	70.2±1.62

Continued on next page

Table 23 – continued from previous page

Model	Metric	Few-shot					
		0	4	8	16	32	64
	AUPRC	–	16.4±3.14	19.3±2.97	19.6±2.66	23.0±2.09	25.2±1.96
CARTE	AUROC	–	68.3±2.99	67.4±4.10	70.8±2.32	74.9±1.67	77.3±1.17
	AUPRC	–	18.5±2.51	20.3±3.09	20.8±2.61	24.5±2.23	26.7±1.79
XTFormer	AUROC	–	65.2±5.30	67.3±4.36	72.3±3.90	74.7±2.38	75.6±1.49
	AUPRC	–	17.6±3.24	20.4±2.79	23.8±3.14	26.9±2.80	29.0±2.09
TabLLM	AUROC	59.8±5.90	60.0±4.87	63.6±4.23	67.0±3.54	70.8±2.76	71.7±1.94
	AUPRC	12.7±5.60	13.8±4.62	15.5±4.08	20.1±3.37	22.2±2.53	24.6±1.86
XTab	AUROC	–	63.4±4.42	65.8±3.32	68.6±2.98	71.2±2.64	72.5±2.31
	AUPRC	–	16.5±3.40	18.9±2.55	19.9±2.62	21.3±1.77	25.4±1.97
TransTab	AUROC	60.5±10.86	59.4±8.57	61.7±5.19	65.1±3.34	67.3±2.26	70.4±1.38
	AUPRC	12.6±4.92	13.6±4.18	15.1±2.99	16.9±2.66	18.9±2.72	22.4±1.93
Ours	AUROC	**68.3**±4.45	**69.5**±3.94	**71.6**±3.28	**74.0**±2.71	**76.5**±2.13	**78.4**±1.34
	AUPRC	**18.2**±4.28	**20.5**±3.65	**23.5**±3.14	**27.0**±2.53	**31.5**±1.97	**36.5**±1.34
HiRID							
LR	AUROC	–	66.5±7.68	67.1±6.95	71.5±3.55	75.6±2.40	75.8±1.27
	AUPRC	–	13.9±3.38	18.9±3.18	19.3±3.62	22.7±2.56	27.7±2.73
RF	AUROC	–	68.7±4.98	68.9±4.13	73.3±3.42	76.7±2.76	77.7±1.94
	AUPRC	–	20.0±5.72	21.9±5.38	23.7±4.67	31.9±4.83	34.9±3.28
XGB	AUROC	–	59.1±4.54	66.9±5.30	71.5±4.31	76.2±2.87	77.9±2.23
	AUPRC	–	10.1±2.20	17.1±3.07	21.5±4.52	27.5±5.38	32.8±4.09
CatBoost	AUROC	–	69.5±9.50	72.9±5.89	74.7±3.13	78.6±2.68	79.2±1.35
	AUPRC	–	20.4±6.90	22.1±5.45	26.8±4.91	33.9±4.68	36.5±3.35
MLP	AUROC	–	67.8±6.48	64.6±5.89	72.3±3.14	74.9±2.48	75.7±1.56
	AUPRC	–	16.0±4.19	20.4±4.35	24.5±3.80	27.6±4.15	31.8±3.04
TabM	AUROC	–	69.7±7.84	72.5±5.09	74.6±3.89	77.1±2.93	80.0±2.13
	AUPRC	–	19.8±5.16	22.3±3.74	22.9±3.90	28.0±3.79	33.0±3.90
CARTE	AUROC	–	71.8±5.60	73.9±5.10	77.7±4.98	82.2±1.94	84.4±0.92
	AUPRC	–	23.3±6.11	27.9±6.26	29.6±5.49	34.3±2.87	36.7±1.97
XTFormer	AUROC	–	71.8±6.96	73.7±6.58	78.3±1.77	81.3±1.44	82.6±1.05
	AUPRC	–	21.8±5.74	24.0±4.99	27.8±3.39	32.9±3.44	35.6±2.46
TabLLM	AUROC	67.4±8.19	68.4±6.83	72.4±5.42	75.4±4.17	78.6±3.06	80.0±2.24
	AUPRC	19.0±5.93	20.3±5.38	24.2±4.76	27.5±3.87	30.1±2.84	32.4±2.13
XTab	AUROC	–	70.2±4.80	72.5±4.26	74.7±3.84	74.9±3.53	78.5±2.81
	AUPRC	–	18.3±4.83	21.2±4.90	22.6±4.42	20.0±2.36	27.3±1.93
TransTab	AUROC	68.0±7.73	67.1±5.81	72.0±6.57	77.0±3.80	79.8±2.01	78.6±1.71
	AUPRC	23.0±9.86	24.3±8.19	26.0±7.68	29.4±5.65	31.0±4.76	34.3±5.24
Ours	AUROC	**80.4**±5.98	**80.8**±4.83	**82.2**±4.51	**83.7**±3.02	**84.7**±2.16	**86.2**±1.46
	AUPRC	**27.0**±5.87	**28.4**±5.42	**30.6**±4.68	**33.8**±3.29	**37.5**±2.47	**42.2**±1.68
SUPPORT							
LR	AUROC	–	64.4±5.37	65.6±7.97	70.3±3.23	72.9±2.18	74.1±2.07
	AUPRC	–	24.2±4.74	34.7±7.86	38.5±4.83	37.5±3.62	49.0±3.14
RF	AUROC	–	64.5±4.30	66.6±3.64	72.0±2.90	73.8±2.73	74.5±2.29
	AUPRC	–	30.8±4.50	**36.4**±7.44	37.8±5.22	44.5±4.74	46.8±2.75
XGB	AUROC	–	63.9±3.45	68.0±3.69	74.4±3.08	76.3±2.87	76.7±2.44
	AUPRC	–	24.2±4.97	31.9±6.07	37.7±4.60	40.7±4.14	46.3±3.77
CatBoost	AUROC	–	**66.3**±7.52	**70.7**±8.64	73.9±3.87	76.9±2.52	77.5±1.87
	AUPRC	–	30.6±5.81	35.4±8.22	41.5±5.68	46.2±5.13	**50.8**±3.18
MLP	AUROC	–	60.9±4.85	62.1±8.49	68.4±3.70	71.7±3.14	73.6±2.29
	AUPRC	–	24.7±3.66	30.6±7.07	34.9±4.26	40.0±4.66	45.7±3.80
TabM	AUROC	–	65.7±6.30	70.0±7.26	74.5±4.96	77.7±3.75	77.9±2.67

Continued on next page

Table 23 – continued from previous page

Model	Metric	Few-shot					
		0	4	8	16	32	64
	AUPRC	–	28.8±4.96	32.9±7.24	39.5±6.70	44.8±5.45	47.7±4.01
CARTE	AUROC	–	61.6±6.69	61.3±7.86	66.2±5.15	70.3±3.57	73.2±4.00
	AUPRC	–	26.7±5.74	29.8±6.43	32.0±4.61	34.1±4.14	36.2±4.42
XTFormer	AUROC	–	63.6±5.07	65.7±6.25	71.0±3.53	74.0±2.67	75.2±2.21
	AUPRC	–	26.1±4.13	28.7±5.41	31.7±4.31	34.9±3.52	38.7±2.14
TabLLM	AUROC	61.8±7.86	61.8±6.53	64.9±5.84	67.8±4.57	72.7±3.26	73.5±2.47
	AUPRC	26.1±6.24	28.5±5.43	30.9±6.18	33.7±4.72	36.2±3.87	40.0±2.83
XTab	AUROC	–	63.9±4.13	66.3±2.89	**75.3**±3.41	**78.4**±2.92	**78.6**±2.65
	AUPRC	–	26.9±3.66	31.0±5.97	**42.5**±4.43	**46.7**±3.64	44.4±3.20
TransTab	AUROC	59.2±7.20	60.6±5.90	63.6±7.42	67.9±4.29	72.0±3.19	73.5±1.87
	AUPRC	19.8±6.70	22.7±5.99	28.0±7.25	35.5±6.36	41.2±4.42	45.3±2.68
Ours	AUROC	**65.6**±6.42	66.2±4.76	67.6±4.29	71.5±3.11	75.3±2.34	77.2±1.82
	AUPRC	**32.1**±5.42	**32.2**±4.68	33.3±4.03	39.8±3.06	42.5±2.29	47.7±1.64
SICdb							
LR	AUROC	–	68.8±9.74	72.7±5.20	75.6±3.72	77.7±2.65	77.1±2.24
	AUPRC	–	12.8±7.34	22.8±5.23	21.9±5.19	23.0±3.27	27.4±3.19
RF	AUROC	–	70.4±4.05	75.9±3.24	79.4±2.25	80.1±1.70	80.2±1.42
	AUPRC	–	15.7±5.86	23.5±4.96	24.0±3.65	26.4±4.18	28.5±3.04
XGB	AUROC	–	60.8±2.90	73.0±4.11	79.5±3.35	81.3±2.57	82.9±1.76
	AUPRC	–	6.0±1.46	15.3±3.73	21.9±5.24	25.1±5.04	28.8±2.84
CatBoost	AUROC	–	70.0±8.73	76.8±5.32	79.4±2.65	80.7±2.55	80.8±1.70
	AUPRC	–	16.2±6.26	20.5±5.51	23.9±4.24	25.8±4.30	27.5±2.66
MLP	AUROC	–	65.0±8.90	67.2±4.02	76.2±3.26	76.8±2.01	77.0±1.59
	AUPRC	–	10.2±4.97	18.4±3.47	20.7±3.62	20.0±2.83	22.2±3.06
TabM	AUROC	–	71.5±6.88	78.4±3.08	79.4±3.65	81.9±2.01	84.4±1.69
	AUPRC	–	16.9±7.29	22.8±3.45	22.3±4.56	25.9±3.67	28.3±2.82
CARTE	AUROC	–	74.7±6.17	**80.8**±2.51	81.8±2.55	84.0±2.31	85.9±1.11
	AUPRC	–	18.9±6.04	**26.7**±4.67	27.4±4.85	32.2±4.20	35.7±2.39
XTFormer	AUROC	–	75.4±6.77	79.9±3.88	**82.6**±2.43	**84.4**±1.81	86.0±1.21
	AUPRC	–	20.0±7.48	24.6±4.94	28.2±5.11	30.5±4.90	32.8±4.04
TabLLM	AUROC	67.9±7.53	68.0±6.24	70.4±4.87	73.7±3.68	77.5±2.63	78.9±1.72
	AUPRC	14.1±6.32	15.6±5.78	19.6±4.76	21.7±3.84	24.8±2.87	27.5±1.74
XTab	AUROC	–	72.0±3.62	75.1±3.87	79.0±2.92	80.5±2.45	82.9±1.97
	AUPRC	–	16.4±6.75	19.4±4.17	23.4±4.95	24.1±4.89	30.4±3.30
TransTab	AUROC	65.2±6.03	65.4±4.98	72.9±3.31	77.6±1.83	80.8±1.59	81.2±1.17
	AUPRC	16.0±8.53	17.4±7.19	23.9±4.32	25.3±4.45	28.7±3.15	31.5±2.70
Ours	AUROC	**76.7**±5.14	**76.9**±4.37	79.5±3.19	80.5±2.35	83.5±1.66	**86.2**±1.15
	AUPRC	**20.7**±6.28	**22.1**±5.17	24.5±4.02	**29.3**±2.94	**34.5**±2.21	**38.5**±1.55
Zigong							
LR	AUROC	–	58.2±6.50	60.2±7.99	61.5±5.66	67.3±5.09	69.7±2.94
	AUPRC	–	8.4±3.57	12.3±3.94	10.6±4.29	14.7±4.42	16.6±3.15
RF	AUROC	–	54.3±3.83	56.5±3.49	60.8±5.27	64.2±5.10	65.4±4.13
	AUPRC	–	8.7±5.42	11.2±6.18	12.3±4.71	16.8±3.89	23.9±2.94
XGB	AUROC	–	57.5±0.92	60.4±2.61	61.5±5.28	67.7±5.65	69.0±4.61
	AUPRC	–	6.2±1.87	11.9±4.35	15.7±5.92	21.0±4.68	24.9±3.24
CatBoost	AUROC	–	57.9±6.76	63.7±6.76	63.9±6.59	70.5±6.04	75.7±3.30
	AUPRC	–	11.3±3.11	13.5±4.26	16.2±6.08	19.3±6.27	25.7±5.44
MLP	AUROC	–	58.1±5.54	60.3±7.92	61.6±4.87	67.0±5.52	71.5±3.31
	AUPRC	–	8.6±3.12	13.2±3.70	15.0±5.98	16.1±6.24	20.0±4.61
TabM	AUROC	–	61.3±7.60	**65.6**±7.69	66.6±5.49	69.9±5.45	72.9±3.59

Continued on next page

Table 23 – continued from previous page

Model	Metric	Few-shot					
		0	4	8	16	32	64
	AUPRC	–	11.6±5.73	15.7±4.08	13.3±4.32	19.9±5.19	23.9±5.83
CARTE	AUROC	–	59.2±7.47	60.1±7.82	66.2±4.61	71.6±3.13	**75.8**±2.96
	AUPRC	–	9.4±1.42	12.7±3.43	14.4±2.67	20.0±4.17	23.5±2.84
XTFormer	AUROC	–	60.8±7.28	62.7±7.57	**67.9**±4.15	71.3±4.67	73.5±3.98
	AUPRC	–	13.0±4.52	15.3±4.47	19.9±6.16	23.9±6.18	26.9±5.05
TabLLM	AUROC	57.0±9.43	57.2±7.36	61.1±6.72	63.7±5.48	67.5±4.57	68.5±3.62
	AUPRC	9.3±4.82	10.1±4.65	13.6±4.94	17.0±4.13	19.1±3.76	22.1±2.94
XTab	AUROC	–	55.7±4.00	59.4±3.68	61.6±4.71	65.9±4.37	66.1±3.84
	AUPRC	–	9.9±5.83	12.0±6.24	12.2±5.16	13.9±3.72	17.9±3.05
TransTab	AUROC	58.6±9.11	59.4±7.08	64.5±6.20	67.4±5.04	**72.7**±3.69	73.8±3.04
	AUPRC	10.9±5.26	11.4±4.83	14.7±4.84	14.5±4.76	20.2±5.60	22.4±3.61
Ours	AUROC	**60.8**±6.39	**61.7**±5.47	63.2±4.68	66.8±3.62	71.9±2.87	75.0±2.24
	AUPRC	**16.4**±4.87	**16.9**±5.24	**19.1**±4.53	**23.9**±4.15	**28.4**±3.29	**30.9**±2.76

E.6 Full Breast Cancer Few-shot Results

Table 24: Few-shot results on the four Breast Cancer datasets. AUROC and AUPRC are reported for each shot count, with the best result in **bold** and the second best underlined. For C2, the 64-shot setting is omitted, as the adaptation split contains fewer than 64 positive samples and we do not resample with replacement.

Model	Metric	Few-shot					
		0	4	8	16	32	64
NCT00041119 (C1)							
LR	AUROC	–	59.5±2.07	62.1±2.09	64.6±1.08	67.9±1.76	69.6±1.61
	AUPRC	–	6.8±0.80	7.5±0.55	8.0±0.34	8.9±0.45	9.8±0.70
RF	AUROC	–	60.0±1.83	62.8±2.59	63.8±0.63	67.2±2.15	69.2±2.08
	AUPRC	–	6.9±0.53	7.4±0.35	7.7±0.28	8.3±0.82	9.1±0.83
XGB	AUROC	–	55.1±1.47	60.7±2.37	61.4±0.98	63.0±1.81	62.8±1.67
	AUPRC	–	4.6±0.26	6.4±0.39	6.7±0.21	7.2±0.57	7.1±0.98
CatBoost	AUROC	–	56.7±1.13	61.2±2.71	62.0±0.80	65.8±2.01	67.9±2.07
	AUPRC	–	6.1±0.36	6.8±0.59	7.1±0.43	7.9±0.59	8.6±0.66
MLP	AUROC	–	57.3±1.08	60.3±1.75	61.9±0.58	64.4±1.73	64.5±1.56
	AUPRC	–	6.5±0.38	7.0±0.29	7.3±0.20	7.7±0.59	8.0±0.52
TabM	AUROC	–	61.4±4.76	62.6±4.15	64.1±3.18	66.9±2.23	69.6±2.02
	AUPRC	–	7.7±2.16	8.1±1.46	8.7±1.78	9.5±1.10	10.5±1.00
CARTE	AUROC	–	**68.7**±4.16	69.5±3.11	70.5±2.49	70.9±2.49	73.9±1.76
	AUPRC	–	10.5±3.03	9.7±2.03	11.1±2.51	11.7±1.49	12.2±1.05
XTFormer	AUROC	–	63.6±4.25	64.3±3.54	65.6±2.79	67.5±2.94	69.5±2.19
	AUPRC	–	7.3±2.58	7.7±1.79	8.3±1.68	8.8±1.83	10.4±1.25
TabLLM	AUROC	63.4±7.49	64.4±4.73	65.7±3.65	67.1±4.32	68.8±2.23	71.2±1.77
	AUPRC	8.1±3.74	8.4±2.63	7.8±2.51	9.2±1.66	9.9±1.70	10.6±1.85
XTab	AUROC	–	61.4±5.72	63.2±3.17	65.8±3.58	67.7±2.45	70.7±2.28
	AUPRC	–	7.7±1.99	8.1±2.09	8.8±2.60	9.5±2.20	10.6±1.12
TransTab	AUROC	58.3±5.64	59.9±6.04	61.9±3.46	63.8±4.16	68.0±2.28	72.8±2.41
	AUPRC	7.7±3.20	7.2±2.24	8.4±2.33	9.0±1.85	9.6±1.62	10.4±1.93
Ours	AUROC	**64.4**±4.37	67.9±5.09	**71.5**±3.90	**73.0**±2.94	**74.3**±2.59	**74.5**±1.39
	AUPRC	**10.1**±2.31	**10.7**±2.44	**11.4**±1.63	**11.9**±1.50	**12.3**±1.44	**12.6**±1.51
NCT00174655 (C2)							

Table 24 – continued

Model	Metric	Few-shot					
		0	4	8	16	32	64
LR	AUROC	–	61.1±5.94	67.0±2.90	68.7±4.00	71.2±4.29	–
	AUPRC	–	21.1±5.74	23.4±4.07	**26.0**±5.98	**27.5**±7.55	–
RF	AUROC	–	62.6±2.71	67.0±3.85	69.8±3.88	71.0±5.16	–
	AUPRC	–	20.1±3.57	**23.5**±4.07	24.5±4.26	23.9±4.91	–
XGB	AUROC	–	53.2±1.58	63.1±3.60	64.9±3.67	66.4±5.59	–
	AUPRC	–	10.6±0.85	20.0±3.05	21.0±4.29	21.0±4.98	–
CatBoost	AUROC	–	61.4±4.05	64.1±2.11	66.5±3.48	68.2±5.56	–
	AUPRC	–	20.4±2.37	21.4±2.53	23.1±1.44	23.3±5.94	–
MLP	AUROC	–	61.3±4.41	66.2±2.88	68.0±3.48	69.4±3.92	–
	AUPRC	–	20.8±4.50	22.0±3.24	25.3±3.01	26.1±5.80	–
TabM	AUROC	–	62.5±5.62	64.7±7.29	66.7±4.63	69.1±5.42	–
	AUPRC	–	15.4±5.56	16.3±4.81	20.4±4.06	21.9±3.52	–
CARTE	AUROC	–	**68.0**±6.55	69.4±5.16	70.3±4.05	72.5±3.05	–
	AUPRC	–	**22.3**±4.67	23.5±4.85	24.4±4.20	24.9±2.39	–
XTFormer	AUROC	–	61.7±8.43	61.2±7.66	65.4±5.74	66.6±5.31	–
	AUPRC	–	14.2±5.30	16.7±5.45	19.0±3.30	19.4±4.40	–
TabLLM	AUROC	62.8±8.65	65.3±6.88	67.9±7.01	68.1±4.96	69.6±6.00	–
	AUPRC	15.9±5.63	16.2±4.29	17.9±4.73	19.5±3.71	20.1±4.68	–
XTab	AUROC	–	61.3±6.31	61.9±8.08	67.2±6.01	71.2±5.64	–
	AUPRC	–	15.3±6.13	18.4±5.73	20.7±6.38	22.1±4.65	–
TransTab	AUROC	57.9±7.34	59.8±6.33	65.4±6.09	67.1±6.45	70.8±4.41	–
	AUPRC	15.4±8.53	18.1±7.80	20.0±4.68	20.7±4.95	22.2±2.88	–
Ours	AUROC	**63.1**±6.13	66.1±6.05	**70.5**±8.39	**71.8**±7.37	**72.8**±4.12	–
	AUPRC	**18.6**±5.60	19.8±5.26	21.8±4.69	22.4±4.62	22.9±3.87	–
NCT00312208 (C3)							
LR	AUROC	–	62.3±2.62	65.1±3.62	68.5±2.57	72.1±3.78	73.8±4.37
	AUPRC	–	15.4±1.25	16.9±2.25	17.9±1.17	19.4±2.23	20.9±2.50
RF	AUROC	–	62.2±2.25	64.0±2.82	67.0±2.40	70.0±3.17	71.3±4.02
	AUPRC	–	14.5±1.31	16.2±3.19	17.6±1.98	18.5±4.27	19.8±4.31
XGB	AUROC	–	53.4±1.83	62.0±1.55	64.3±1.86	66.1±2.36	66.7±3.02
	AUPRC	–	8.7±0.43	13.4±1.17	15.2±1.51	15.6±2.19	15.9±3.02
CatBoost	AUROC	–	60.6±1.70	63.9±2.34	66.9±2.50	70.0±3.89	71.8±4.04
	AUPRC	–	13.8±0.53	15.1±2.24	17.0±2.57	18.2±3.88	19.0±3.40
MLP	AUROC	–	61.3±2.48	64.7±4.08	68.3±3.01	70.4±3.32	71.4±2.99
	AUPRC	–	14.6±1.55	16.4±3.00	17.7±1.84	18.6±2.82	18.6±2.39
TabM	AUROC	–	65.8±4.33	68.5±3.86	70.7±3.65	72.4±2.99	74.6±2.30
	AUPRC	–	17.0±4.20	17.5±3.45	20.9±3.67	21.4±2.82	23.0±3.30
CARTE	AUROC	–	66.1±2.52	68.2±2.84	69.5±2.86	70.3±1.51	72.5±1.38
	AUPRC	–	21.5±4.15	22.5±4.44	23.7±4.75	24.4±3.78	25.0±2.63
XTFormer	AUROC	–	65.5±5.34	68.0±4.03	69.1±3.74	70.2±2.28	71.6±2.06
	AUPRC	–	15.6±2.30	17.7±3.34	19.8±2.90	21.1±3.31	22.2±2.79
TabLLM	AUROC	61.5±6.22	62.9±4.59	64.2±3.72	66.8±4.07	68.0±2.18	69.0±2.08
	AUPRC	16.5±4.87	18.1±3.90	19.0±3.71	20.0±2.28	21.0±2.12	21.6±1.94
XTab	AUROC	–	62.7±4.81	66.1±3.26	68.2±2.58	70.0±2.49	72.1±2.27
	AUPRC	–	16.0±6.75	18.6±4.17	20.1±4.95	21.3±4.89	22.1±3.30
TransTab	AUROC	57.2±6.11	61.6±4.91	64.7±3.72	65.8±3.76	69.2±3.40	71.2±1.86
	AUPRC	15.8±6.03	17.4±5.47	18.8±3.65	19.2±3.48	20.7±3.96	21.4±2.70
Ours	AUROC	**65.0**±3.90	**69.1**±2.74	**72.5**±2.13	**74.3**±1.87	**74.8**±1.75	**75.2**±1.01
	AUPRC	**20.6**±5.68	**22.9**±3.60	**24.3**±3.82	**25.0**±3.40	**25.5**±2.83	**25.8**±2.13
NCT00688740 (C4)							
LR	AUROC	–	56.3±2.04	59.3±2.89	61.1±3.14	64.9±3.11	66.7±4.16

Table 24 – continued

Model	Metric	Few-shot					
		0	4	8	16	32	64
	AUPRC	–	24.4±2.32	26.0±2.49	27.5±1.91	29.5±2.02	31.6±3.48
RF	AUROC	–	57.3±1.81	58.9±2.09	60.4±2.88	63.3±2.92	66.1±4.44
	AUPRC	–	23.6±1.72	24.1±1.40	25.6±2.08	26.8±2.06	29.9±3.80
XGB	AUROC	–	52.6±1.26	57.1±1.47	58.4±2.31	61.2±2.41	62.0±3.92
	AUPRC	–	17.4±0.62	22.5±0.90	23.5±1.00	25.4±1.82	26.3±2.97
CatBoost	AUROC	–	56.8±2.33	57.3±0.88	59.7±3.26	63.3±3.23	66.7±4.36
	AUPRC	–	22.8±1.36	22.9±0.98	24.7±2.23	27.0±2.13	29.7±3.74
MLP	AUROC	–	55.7±1.95	57.2±2.98	58.5±2.01	61.3±1.28	63.4±2.80
	AUPRC	–	22.7±1.89	23.7±1.77	24.2±1.46	25.2±1.38	27.0±2.14
TabM	AUROC	–	61.3±3.42	64.9±2.66	68.0±2.29	68.6±1.58	70.2±1.72
	AUPRC	–	22.4±3.33	25.1±3.26	27.9±3.57	29.3±3.03	30.4±1.71
CARTE	AUROC	–	61.5±2.29	62.9±1.83	63.1±2.43	65.3±1.57	67.2±1.20
	AUPRC	–	27.6±2.95	29.7±4.23	30.9±2.62	31.7±1.96	32.1±1.31
XTFormer	AUROC	–	62.6±2.91	65.5±2.96	67.1±2.70	69.0±2.61	69.9±1.99
	AUPRC	–	20.8±4.28	25.3±5.73	27.1±4.98	28.9±5.19	29.6±3.76
TabLLM	AUROC	**60.9**±5.48	63.6±3.62	65.4±4.04	66.6±2.64	66.2±2.73	68.4±2.20
	AUPRC	22.2±5.05	23.9±3.90	25.3±4.24	26.8±3.75	28.4±2.87	29.1±2.35
XTab	AUROC	–	62.0±4.12	64.6±3.10	**68.4**±2.52	**70.3**±1.70	**71.4**±1.90
	AUPRC	–	21.6±6.54	24.7±4.20	27.0±2.83	29.0±3.36	30.0±2.37
TransTab	AUROC	54.9±4.70	59.1±3.69	63.4±2.94	66.1±3.08	67.0±1.87	68.4±1.99
	AUPRC	20.6±6.03	23.1±5.88	24.2±5.94	26.8±3.59	27.3±3.28	27.9±2.70
Ours	AUROC	60.0±4.39	**65.3**±4.40	**67.5**±2.77	68.3±2.07	69.3±1.79	69.4±2.15
	AUPRC	**27.3**±5.32	**30.1**±5.10	**32.1**±3.56	**32.8**±3.33	**33.7**±3.28	**34.1**±1.78