When, Where, and How: Adaptive Binning for Tabular Self-Supervised Learning

Daehwan Kim[1][0009−0006−7856−7850], Haejun Chung[1†][0000−0001−8959−237X], and Ikbeom Jang[2†][0000−0002−6901−983X]

[1] Hanyang University, Seoul, Republic of Korea
[2] Hankuk University of Foreign Studies, Yongin, Republic of Korea
{officialhwan, haejun}@hanyang.ac.kr, ijang@hufs.ac.kr

Abstract. Medical tabular data are ubiquitous in clinical research, but deep learning for tables remains underexplored because reliable labels often require costly expert adjudication, even though structured clinical variables are routinely available in tabular form. Self-supervised learning can leverage these unlabeled tables, and recent binning-based pretexts offer a promising inductive bias, but existing objectives fix a single global quantile discretization and apply feature-agnostic supervision. We propose Adaptive Binning, a training-adaptive discretization pretext for tabular SSL that couples discretization to learning through a feature-wise coarse-to-fine curriculum. Motivated by the spectral bias of neural networks and the principles of curriculum learning, our method progressively refines discretization per feature upon plateau detection and selects representation-aware splits to jointly improve value-space concentration and representation-space coherence. A heterogeneity-aware objective unifies categorical reconstruction with ordinal supervision for numerical features, and experiments on public medical tabular datasets under unified evaluation protocols show consistent gains for linear probing and fine-tuning without dataset-specific discretization tuning. We further introduce a medical tabular SSL benchmark with standardized protocols to support reproducible progress in this underexplored domain. Our code is available at https://github.com/labhai/Adaptive-Binning.

Keywords: Medical Tabular Data · Self-Supervised Learning · Adaptive Binning · Tabular SSL

1 Introduction

Clinical trials, registries, and epidemiological studies routinely tabulate baseline characteristics, laboratory panels, graded findings, and outcomes; in a pilot review of comparative clinical trials, 99% of articles reported baseline or outcome measures in at least one table, and 85% reported both [13]. This reliance on tables reflects a broader pattern where tabular data dominates structured decision systems but remains underexplored in deep learning [4]. Primary challenges are

† Corresponding authors.

that tables mix categorical and numerical variables, exhibit non-smooth interactions, and lack spatial or sequential structure [11,12]. These properties favor tree ensembles such as XGBoost [6] and CatBoost [18], whose recursive partitioning yields piecewise-constant functions for mixed types [22,15]. While tabular neural architectures narrow this gap [11,2,26], deep models are most compelling when they exploit self-supervised representation learning from unlabeled data, a setting that aligns with healthcare, where labels require expert adjudication [9,28]. Nonetheless, medical self-supervision has focused on imaging and language, leaving clinical tabular data underserved.

Recent tabular SSL recasts binning as a pretext task [14], reconstructing bin indices to inject a tree-like continuous-to-discrete inductive bias and harmonize supervision across heterogeneous features. However, fixed global discretization uses a single bin count T and static quantile boundaries during training, fitting integer-index targets by squared-error regression. This feature-agnostic design neither adapts resolution to feature saturation nor uses representations to localize refinement, and provides limited support for jointly modeling ordinal numerical targets and categorical reconstruction. Beyond direct heterogeneous reconstruction objectives [8,23] and learnable discretization or tokenization schemes [27], these limitations call for label-free ordinal targets whose resolution is training-coupled and evolves during pretraining.

Motivated by these gaps, we replace globally fixed discretization with a training-adaptive coarse-to-fine curriculum. This design mirrors the clinical progression from broad stratification to finer severity grading encoded in diagnostic criteria [16,1] and aligns with neural network learning dynamics. We leverage the synergy between curriculum learning [3] and spectral bias [19]–the tendency of networks to fit coarse structures before fine details–to implement an adaptive binning strategy that gradually increases task complexity. To realize this curriculum, we develop an autoencoding-based tabular SSL framework that refines discretization feature-wise during pretraining, coupled with type-aware reconstruction. It specifies *when/where/how* to refine discretization via feature-wise saturation triggers, representation-aware split selection, and a type-aware reconstruction objective for mixed categorical and ordinal numerical targets, so discretization targets evolve online during pretraining. Our contributions are:

1. We propose Adaptive Binning (Fig. 1), a training-adaptive discretization pretext task for tabular SSL that replaces fixed global binning with feature-wise, coarse-to-fine refinement and explicitly specifies *when*, *where*, and *how* discretization evolves during pretraining.
2. We evaluate this pretext across medical tabular datasets spanning binary, nominal, and ordinal multiclass classification, and regression, using linear probing (Tab. 2) and fine-tuning with multiple tabular encoders (Tab. 4), with a single default configuration that avoids dataset-specific tuning (Fig. 2).
3. We establish a medical tabular SSL benchmark with unified evaluation protocols (Tab. 1), providing a reproducible foundation for progress in self-supervised learning for clinical tabular data.

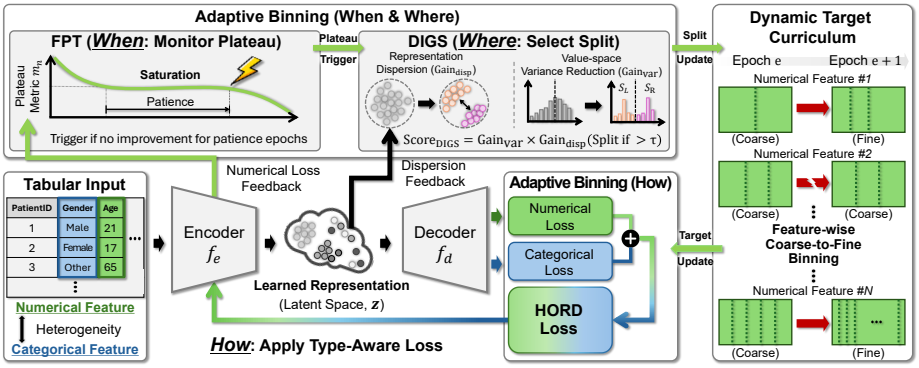

Fig. 1. Overview of our proposed Adaptive Binning framework: HORD provides type-aware mixed-type reconstruction, while FPT and DIGS refine numerical binning via plateau-triggered, representation-aware splitting to form a feature-wise coarse-to-fine target curriculum; see Section 2.2 for details.

2 Method

We first formalize the masking–reconstruction framework for tabular SSL and the fixed quantile-binning objective [14] as our baseline. We then describe Adaptive Binning, which refines discretization during pretraining and couples it with type-aware ordinal supervision for mixed categorical–numerical schemas.

2.1 Preliminaries: Masking and Fixed Binning

We follow the autoencoding-based tabular SSL setup of [14] with inputs $\mathbf{x} = [\mathbf{x}^{\text{cat}}, \mathbf{x}^{\text{num}}]$. During pretraining, we optionally apply feature-wise masking with probability p_m and impute masked entries with a fixed constant (*Const*) [28] or an in-batch value (*Random*) [14]; setting $p_m = 0$ corresponds to *NoMask*. An encoder–decoder maps the corrupted input to a representation and reconstructs pretext targets, subsuming denoising/value reconstruction [25] and mask detection [28]. For numerical feature n, the fixed-binning baseline maps x_n^{num} to a quantile index $y^{(n)} \in \{0, \ldots, T-1\}$ using a single global bin count T with static boundaries, and learns **BinRecon** by squared-error regression on $y^{(n)}$ [14]. We consider three baseline objectives: **ValueRecon** (raw-value reconstruction), **MaskXent** (mask prediction), and **BinRecon** (fixed-T quantile-index prediction).

2.2 Proposed Method

For each numerical feature n, we maintain an adaptive discretizer $B^{(n)}$ with a feature-specific bin count T_n, initialized at T_{init} and capped at $T_{\max}$. The

bin-index target is $y^{(n)} = B^{(n)}(x_n^{\text{num}}) \in \{0, \ldots, T_n - 1\}$, and we refer to the per-feature schedule $\{T_n\}$ as *Feature-Wise Adaptation* (FWA).

***When: Feature-Wise Plateau Trigger* (FPT).** Numerical features vary in complexity and convergence speed, making a globally synchronized refinement schedule inefficient. We therefore propose *Feature-Wise Plateau Trigger* (FPT) to monitor each feature independently. At the end of each epoch, we compute a feature-specific plateau metric m_n from the numerical reconstruction loss, defined as the normalized weighted sum of the components in Eq. (2) over the epoch; we then track the running best value best_n and update a patience counter cnt_n. If $m_n < \text{best}_n - \delta$, we set $\text{best}_n \leftarrow m_n$ and reset cnt_n; otherwise, we increment cnt_n. Feature n is eligible for splitting when $\text{cnt}_n \geq \text{patience}$ (set to 5; see Fig. 2) and $T_n < T_{\max}$, triggering refinement upon saturation.

***Where: Dispersion-Informed Gain-based Splitting* (DIGS).** When FPT marks feature n as ready to refine, we propose DIGS to choose *which* bin to split and *where* to place a new boundary by coupling variance reduction in the feature-value space with dispersion reduction in the encoder-induced representation space. For each bin $B_t^{(n)}$ of the flagged feature, we use the within-bin median as the candidate split point, yielding a near-balanced partition that preserves equal-frequency standardization and avoids sparsity-induced target instability. Let S denote the samples in $B_t^{(n)}$, split by the median into S_L and S_R with weights $w_L = |S_L|/|S|$ and $w_R = |S_R|/|S|$. For any statistic $g(\cdot)$, define the split-induced reduction as $\Delta_g(S \to S_L, S_R) = g(S) - w_L g(S_L) - w_R g(S_R)$, and set the value-space gain to $\text{Gain}_{\text{var}} = \Delta_{\text{Var}}(S \to S_L, S_R)$ [5]. Since variance reduction alone is agnostic to the encoder-induced representation space, we additionally quantify within-subset coherence [21] from embeddings of uncorrupted inputs, $\mathbf{z}_i = f_\theta(\mathbf{x}_i)$ with $\hat{\mathbf{z}}_i = \mathbf{z}_i/|\mathbf{z}_i|$, and define $\text{Disp}(S) = \left|\log\!\left(\epsilon + \left\|\frac{1}{|S|}\sum_{i\in S}\hat{\mathbf{z}}_i\right\|^2\right)\right|$ with $\epsilon > 0$, yielding $\text{Gain}_{\text{disp}} = \Delta_{\text{Disp}}(S \to S_L, S_R)$. We define the split score as

$$\text{Score}_{\text{DIGS}}(S \to S_L, S_R) = \text{Gain}_{\text{var}} \cdot \text{Gain}_{\text{disp}}. \tag{1}$$

We split only if $\text{Gain}_{\text{var}} > 0$, $\text{Gain}_{\text{disp}} > 0$, and $\text{Score}_{\text{DIGS}} > \tau$ ($\tau = 10^{-4}$; see Fig. 2). At each refinement event for feature n, we score all bins and apply all qualifying splits in parallel, potentially inserting multiple boundaries per trigger. After refinement, we reset the FPT statistics for feature n.

***How: Heterogeneity-aware ORDinal Loss* (HORD).** We propose HORD, a type-aware reconstruction objective that unifies nominal supervision for categorical features with ordinal, distribution-aware supervision for numerical features. Mapping a continuous value to an ordered bin index yields an ordinal surrogate that preserves ordering and local proximity; thus, numerical reconstruction should penalize errors by bin distance rather than treat bins as nominal classes [14]. We supervise categorical features with cross entropy, $\mathcal{L}_{\text{cat}}^{(c)} = \text{CE}(\boldsymbol{\ell}^{(c)}, y^{(c)})$. For numerical feature n, we reconstruct a distribution over T_n ordered bins with logits $\boldsymbol{\ell}^{(n)} \in \mathbb{R}^{T_n}$ and probabilities $\mathbf{p}^{(n)} = \text{softmax}(\boldsymbol{\ell}^{(n)})$ for target index $y^{(n)}$. We use soft ordinal targets [7] $q_t = \frac{\exp\left(-(t-y^{(n)})^2\right)}{\sum_{k=0}^{T_n-1}\exp\left(-(k-y^{(n)})^2\right)}$ and supervise numerical reconstruction with soft-target cross entropy (SORD). We

Table 1. Dataset summary with full names, abbreviations, task type, number of classes, presence of missing values (Missing), instances (#Inst.), features (#Feat.), and dataset-specific evaluation configuration (batch size and MLP width/depth).

Dataset (Abbr.)	Task (#Classes)	#Inst.	#Feat.	Missing	Batch size	MLP Width	MLP Depth
Indian Liver Patient Dataset (ILPD)	BC (2)	579	10	Yes	64	512	1
Heart Failure Clinical Records (HFC)	BC (2)	299	12	No	64	512	5
Cardiotocography (CTG)	NMC (10)	2126	21	No	128	256	2
Epileptic Seizure Recognition (ESR)	NMC (5)	11500	178	No	256	512	4
Estimation of Obesity Levels (EOL)	OMC (7)	2111	16	No	128	128	2
Maternal Health Risk (MHR)	OMC (3)	1013	6	No	64	1024	4
Parkinsons Telemonitoring (PT)	Reg (-)	5875	19	No	128	1024	2
Body Fat Prediction (BFP)	Reg (-)	252	13	No	64	512	5

augment SORD with mean–variance regularization [17] on $\mathbf{p}^{(n)}$ using $\mu^{(n)} = \sum_t p_t^{(n)}\, t$ and $\sigma^{2(n)} = \max\!\big(0, \sum_t p_t^{(n)}\, t^2 - (\mu^{(n)})^2\big)$:

$$\mathcal{L}_{\text{num}}^{(n)} = w_{\text{SORD}}\Big(-\sum_t q_t \log p_t^{(n)}\Big) \;+\; w_{\text{mse}}\big(\mu^{(n)} - y^{(n)}\big)^2 \;+\; w_{\text{var}}\,\sigma^{2(n)}. \quad (2)$$

We fix $w_{\text{SORD}} = 10$, $w_{\text{mse}} = 0.1$, and $w_{\text{var}} = 0.001$ in all experiments (see Fig. 2 for sensitivity). Finally, we average losses by feature type and weight by feature counts, yielding uniform feature-wise weighting under varying compositions:

$$\mathcal{L}_{\text{HORD}} = \frac{C}{C+N}\,\frac{1}{C}\sum_{c=1}^{C}\mathcal{L}_{\text{cat}}^{(c)} \;+\; \frac{N}{C+N}\,\frac{1}{N}\sum_{n=1}^{N}\mathcal{L}_{\text{num}}^{(n)}. \quad (3)$$

Adaptive Binning as a Pretext Task. Fig. 1 and Alg. 1 summarize the method.

Adaptive Binning replaces a fixed global T with feature-wise resolutions $\{T_n\}$ refined online. HORD defines the numerical loss underlying m_n (*How*); FPT flags saturated features $\mathcal{R}$ (*When*); and DIGS selects representation-aware qualifying splits (*Where*). Together, they form a label-free coarse-to-fine curriculum.

Algorithm 1 Adaptive Binning Pretraining.

```
Initialize B^(n) with T_init bins; T_n ← T_init
for epoch e = 1, ..., E do
  HORD: optimize Eq. (3); compute m_n
  FPT: update states; set R
  for n ∈ R with T_n < T_max do
    Encode uncorrupted inputs z_i = f_θ(x_i)
    DIGS: score median splits by Eq. (1)
    Split bins; update B^(n), T_n; reset FPT
```

3 Experiments and Results

Datasets. We curate a benchmark of publicly available medical tabular datasets spanning binary classification (BC), multiclass classification (MC; NMC for nom-

Table 2. Linear evaluation with an MLP encoder across diverse datasets and tasks. We vary numerical binning (B: none (-), fixed binning (FIX), Ours), masking (M: none (-), constant replacement (C), random replacement (R)), and the pretext objective (O: ValueRecon (VR), MaskXent (MX), MaskXent+ValueRecon (MR), BinRecon (BR), Ours). We report Average Rank (Avg. Rank) aggregated from per-dataset rankings. Results are mean_{std}; metrics are denoted as $\text{Task}_{\text{metric}}$; **best** and second-best.

Datasets			ILPD	HFC	CTG	ESR	EOL	MHR	PT	BFP	Avg.
B	M	O	$\text{BC}_{\text{AUC(\%)}}\uparrow$		$\text{NMC}_{\text{Acc.(\%)}}\uparrow$		$\text{OMC}_{\text{QWK(\%)}}\uparrow$		$\text{Reg}_{\text{RMSE}}\downarrow$		Rank
-	-	VR	$75.89_{0.3}$	$86.08_{2.4}$	$85.66_{0.5}$	$53.70_{0.2}$	$86.38_{0.4}$	$30.57_{1.2}$	$15.98_{0.1}$	$5.50_{0.0}$	10.88
-	C	VR	$76.28_{0.1}$	$85.09_{2.2}$	$86.71_{0.3}$	$57.00_{0.2}$	$91.67_{0.4}$	$41.97_{1.4}$	$17.74_{0.2}$	$5.53_{0.0}$	9.25
-	C	MX	$76.52_{0.3}$	$90.11_{0.5}$	$83.52_{0.6}$	$61.81_{0.2}$	$87.05_{0.4}$	$55.35_{0.8}$	$19.57_{0.3}$	$5.13_{0.1}$	8.19
-	C	MR	$76.40_{0.2}$	$84.85_{2.0}$	$86.53_{0.3}$	$58.72_{0.1}$	$89.99_{0.3}$	$38.91_{2.9}$	$17.60_{0.2}$	$5.64_{0.0}$	9.56
-	R	VR	$76.39_{0.3}$	$88.08_{2.1}$	$86.46_{0.3}$	$55.54_{0.2}$	$91.91_{0.4}$	$40.67_{2.1}$	$17.40_{0.1}$	$5.48_{0.0}$	8.38
-	R	MX	$76.53_{0.3}$	$89.19_{0.4}$	$84.84_{0.3}$	$62.67_{0.1}$	$86.27_{0.3}$	$64.06_{1.0}$	$15.29_{0.3}$	$5.45_{0.1}$	7.31
-	R	MR	$76.48_{0.2}$	$85.82_{2.1}$	$86.53_{0.2}$	$57.39_{0.2}$	$90.66_{0.6}$	$43.47_{1.7}$	$15.27_{0.1}$	$5.37_{0.1}$	7.31
FIX	-	BR	$75.54_{0.2}$	$86.76_{0.3}$	$84.86_{0.5}$	$62.00_{0.1}$	$86.83_{0.4}$	$60.45_{0.5}$	$16.67_{0.1}$	$5.32_{0.1}$	8.88
FIX	C	BR	$76.25_{0.3}$	$90.11_{0.4}$	$86.76_{0.2}$	$62.94_{0.2}$	$90.52_{0.4}$	$61.03_{0.5}$	$17.66_{0.2}$	$5.30_{0.0}$	6.31
FIX	R	BR	$76.10_{0.3}$	$86.34_{0.5}$	$85.89_{0.3}$	$63.83_{0.2}$	$88.13_{0.8}$	$62.41_{0.4}$	$15.71_{0.4}$	$5.33_{0.0}$	7.38
Ours	-	**Ours**	$76.53_{0.1}$	$93.25_{0.6}$	$84.91_{0.2}$	$64.10_{0.2}$	$\mathbf{93.78}_{0.4}$	$64.13_{0.9}$	$\underline{14.27}_{0.1}$	$4.65_{0.0}$	3.56
Ours	C	**Ours**	$\mathbf{77.80}_{0.2}$	$\underline{95.00}_{0.3}$	$\mathbf{87.70}_{0.5}$	$\mathbf{66.86}_{0.1}$	$89.71_{0.5}$	$\underline{66.91}_{0.8}$	$14.98_{0.1}$	$\underline{4.57}_{0.0}$	$\underline{2.50}$
Ours	R	**Ours**	$\underline{77.25}_{0.2}$	$\mathbf{96.88}_{0.3}$	$\underline{87.61}_{0.4}$	$\underline{65.60}_{0.1}$	$\underline{93.17}_{0.4}$	$\mathbf{70.51}_{1.2}$	$\mathbf{11.32}_{0.1}$	$\mathbf{4.56}_{0.0}$	**1.50**

inal, OMC for ordinal), and regression (Reg), with diverse clinical tasks, heterogeneous schemas, and varying categorical–numerical compositions (see Table 1).

Implementation Details. To ensure rigorous comparability, we adopt the 1000 epoch pretraining protocol of the fixed-binning baseline [14]. As architectural complexity yields diminishing returns on tabular data [11,12,10], we employ a standard MLP encoder f_θ with a symmetric decoder f_d. Dataset-specific depth $\in \{1, 2, 3, 4, 5\}$ and width $\in \{128, 256, 512, 1024\}$ are selected via supervised validation (see Table 1); all remaining configurations inherit from [14] [3]. Adaptive refinement is confined to pretraining and introduces no downstream overhead.

Evaluation. To isolate and quantify representation quality [14,20], we employ two 100 epoch protocols: linear probing of frozen embeddings and fine-tuning. The fine-tuning phase pairs our encoder with MLPs and tabular architectures (ResNet, TabNet [2], SAINT [23], FT-Transformer [11], T2G-Former [26]), explicitly retaining their default configurations to minimize orthogonal influences from hyperparameter tuning. We report AUC for BC, Accuracy (Acc.) for NMC, QWK for OMC, and RMSE for Reg, as 10-seed averages on one RTX 4090.

Linear Evaluation. Table 2 supports a clear takeaway on medical tabular tasks: adaptive discretization achieves the best average rank across masking options and pretext objectives. Under masking, we tune $p_m \in \{0.1, 0.2, 0.3\}$ and choose the initial resolution (fixed T for BinRecon, T_{init} for ours) from $\{2, 10\}$. The margin over fixed-binning BinRecon [14] persists even at its best masked setting, and our no-mask variant still surpasses masked BinRecon. This indicates that

[3] $\text{lr} = 10^{-4}, 10^{-2}, 10^{-3}$ for pretraining, linear probing, and fine-tuning, respectively.

Table 3. Linear-evaluation ablations of our method across datasets. Each variant removes one component from the full model (w/o FWA/HORD: direct removal; w/o FPT: DIGS with fixed-epoch refinement; w/o DIGS: FPT with variance-only splitting); all other settings follow Table 2. Mean_{std}; metrics as $\text{Task}_{\text{metric}}$; best in **bold**.

Datasets	ILPD	HFC	CTG	ESR	EOL	MHR	PT	BFP
Method	$\text{BC}_{\text{AUC}(\%)}\uparrow$		$\text{NMC}_{\text{Acc.}(\%)}\uparrow$		$\text{OMC}_{\text{QWK}(\%)}\uparrow$		$\text{Reg}_{\text{RMSE}}\downarrow$	
Ours	$\mathbf{77.80}_{0.2}$	$\mathbf{96.88}_{0.3}$	$\mathbf{87.70}_{0.5}$	$\mathbf{66.86}_{0.1}$	$\mathbf{93.78}_{0.4}$	$\mathbf{70.51}_{1.2}$	$11.32_{0.1}$	$\mathbf{4.56}_{0.0}$
w/o FWA	$76.73_{0.2}$	$\mathbf{96.88}_{0.3}$	$85.02_{0.4}$	$63.97_{0.1}$	$89.15_{0.3}$	$67.07_{0.6}$	$\mathbf{11.23}_{0.1}$	$4.97_{0.0}$
w/o FPT	$76.48_{0.3}$	$\mathbf{96.88}_{0.3}$	$86.50_{0.3}$	$64.31_{0.2}$	$84.08_{0.5}$	$66.23_{1.3}$	$14.20_{0.2}$	$4.91_{0.1}$
w/o DIGS	$74.07_{0.3}$	$\mathbf{96.88}_{0.3}$	$86.29_{0.4}$	$65.97_{0.1}$	$85.59_{0.4}$	$67.29_{0.8}$	$13.55_{0.1}$	$5.44_{0.1}$
w/o HORD	$76.51_{0.2}$	$88.41_{0.3}$	$86.71_{0.5}$	$62.80_{0.1}$	$91.39_{0.6}$	$63.31_{1.1}$	$15.84_{0.2}$	$5.12_{0.0}$

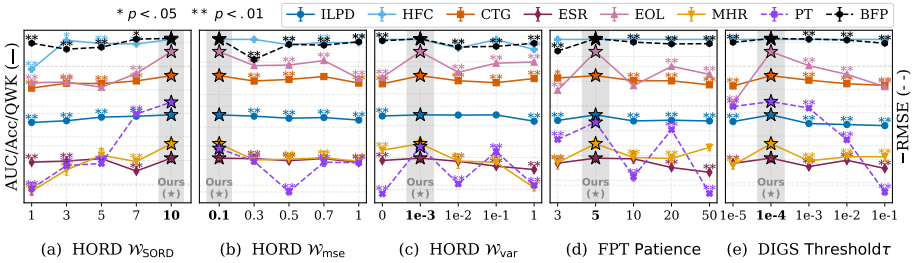

Fig. 2. Linear-probing hyperparameter sweeps under the same setup as Table 2. We sweep (a) $w_{\text{SORD}} \in \{1, 3, 5, 7, \mathbf{10}\}$, (b) $w_{\text{MSE}} \in \{\mathbf{0.1}, 0.3, 0.5, 0.7, 1\}$, (c) $w_{\text{Var}} \in \{0, \mathbf{10^{-3}}, 10^{-2}, 10^{-1}, 1\}$, (d) FPT patience $\in \{3, \mathbf{5}, 10, 20, 50\}$, and (e) DIGS threshold $\tau \in \{10^{-5}, \mathbf{10^{-4}}, 10^{-3}, 10^{-2}, 10^{-1}\}$. The gray shaded line denotes the default configuration; statistical significance markers are shown above each point.

improvements are driven primarily by training-adaptive, feature-wise refinement rather than input corruption, with masking as a complementary regularizer. The *When–Where–How* coupling turns discretization into a pretext that adaptively sharpens supervision and yields stronger representations for clinical tables.

Table 3 shows that ablating individual components generally degrades linear probing, revealing complementary effects not attributable to any single module. The HFC setting is especially instructive. Because FPT never triggers, supervision remains effectively fixed-binned, yet removing HORD still induces a marked drop, demonstrating the value of type-aware ordinal supervision even without refinement. HFC is the sole no-refinement case, whereas refinement occurs in the other seven datasets, with the median per-dataset mean bin count rising from 5.54 to 17.44 over pretraining. Overall, the ablations indicate that performance is driven by the integrated Adaptive Binning pipeline.

Finally, Figure 2 demonstrates the method's robustness to hyperparameter choice. A single default configuration provides a reliable starting point across

Table 4. Fine-tuning with tabular-specific encoders: supervised from scratch vs. SSL-pretrained initialization. MaskXent+ValueRecon (MR) and fixed-binning BinRecon (BR) serve as SSL baselines based on their strong linear-probe results (Table 2). Results are mean_{std} across runs; metrics are denoted as $\text{Task}_{\text{metric}}$; **best** and second-best.

Encoder	Datasets	HFC	ESR	EOL	BFP
	Method	$\text{BC}_{\text{AUC}(\%)}\uparrow$	$\text{NMC}_{\text{Acc.}(\%)}\uparrow$	$\text{OMC}_{\text{QWK}(\%)}\uparrow$	$\text{Reg}_{\text{RMSE}}\downarrow$
MLP	Supervised	$\underline{90.74}_{1.6}$	$76.24_{0.5}$	$94.06_{0.6}$	$5.50_{0.3}$
	SSL MR	$89.90_{1.2}$	$76.66_{0.6}$	$\underline{94.64}_{0.6}$	$\underline{5.40}_{0.2}$
	SSL BR	$90.54_{0.7}$	$\mathbf{77.02}_{0.6}$	$\mathbf{94.74}_{0.6}$	$5.65_{0.3}$
	SSL **Ours**	$\mathbf{90.87}_{0.9}$	$\underline{76.87}_{0.6}$	$93.89_{0.9}$	$\mathbf{5.09}_{0.2}$
ResNet	Supervised	$\mathbf{91.12}_{1.6}$	$\underline{76.33}_{0.7}$	$88.22_{1.3}$	$5.43_{0.4}$
	SSL MR	$\underline{90.97}_{0.6}$	$76.27_{0.3}$	$88.81_{0.8}$	$5.18_{0.2}$
	SSL BR	$90.96_{0.9}$	$75.95_{0.5}$	$\underline{88.89}_{1.3}$	$\underline{4.82}_{0.2}$
	SSL **Ours**	$90.52_{1.6}$	$\mathbf{76.70}_{0.3}$	$\mathbf{89.17}_{1.0}$	$\mathbf{4.77}_{0.1}$
TabNet [2]	Supervised	$85.07_{7.3}$	$47.89_{1.8}$	$64.51_{6.8}$	$10.22_{8.0}$
	SSL MR	$87.22_{4.3}$	$48.82_{2.4}$	$62.69_{5.1}$	$9.40_{1.7}$
	SSL BR	$\underline{87.39}_{2.7}$	$\underline{50.88}_{1.8}$	$\underline{74.26}_{4.8}$	$\underline{8.42}_{1.6}$
	SSL **Ours**	$\mathbf{89.31}_{2.5}$	$\mathbf{52.32}_{1.6}$	$\mathbf{75.41}_{4.3}$	$\mathbf{8.23}_{1.4}$
SAINT [23]	Supervised	$92.34_{2.0}$	$\mathbf{71.80}_{0.5}$	$\underline{95.52}_{0.7}$	$4.41_{0.1}$
	SSL MR	$\underline{92.59}_{1.7}$	$\underline{71.54}_{0.5}$	$94.65_{0.9}$	$\underline{4.40}_{0.1}$
	SSL BR	$92.42_{2.7}$	$70.70_{0.5}$	$93.75_{1.2}$	$4.58_{0.1}$
	SSL **Ours**	$\mathbf{92.82}_{2.2}$	$71.06_{0.3}$	$\mathbf{95.77}_{1.0}$	$\mathbf{4.39}_{0.1}$
FT-Transformer [11]	Supervised	$89.43_{3.5}$	$67.16_{1.0}$	$\mathbf{95.78}_{1.0}$	$5.52_{0.1}$
	SSL MR	$92.21_{2.4}$	$70.39_{0.6}$	$\underline{94.37}_{1.1}$	$\underline{5.27}_{0.2}$
	SSL BR	$\underline{92.47}_{2.1}$	$\underline{74.14}_{0.3}$	$93.62_{1.2}$	$5.70_{0.3}$
	SSL **Ours**	$\mathbf{93.43}_{1.5}$	$\mathbf{75.97}_{0.7}$	$93.53_{1.4}$	$\mathbf{5.05}_{0.3}$
T2G-Former [26]	Supervised	$91.77_{3.7}$	$67.93_{0.0}$	$96.67_{1.4}$	$\mathbf{5.01}_{0.1}$
	SSL MR	$92.90_{1.5}$	$\underline{72.62}_{0.9}$	$96.75_{0.9}$	$5.32_{0.2}$
	SSL BR	$\underline{93.17}_{2.0}$	$71.90_{0.0}$	$\mathbf{97.09}_{1.3}$	$5.27_{0.3}$
	SSL **Ours**	$\mathbf{94.22}_{1.7}$	$\mathbf{73.12}_{0.0}$	$\mathbf{97.09}_{0.9}$	$\underline{5.02}_{0.3}$

tasks and datasets, reducing the need for per-dataset tuning, where extensive tuning can amplify the risk in clinical deployment [24]. Across broad sweeps of loss weights and refinement controls, deviations from the default consistently reduce performance, reinforcing it as a robust choice.

Fine-tuning Evaluation. Table 4 shows that the benefits of our pretraining persist under end-to-end fine-tuning on one representative dataset per task type, rather than being confined to linear probing. Among SSL objectives, adaptive discretization provides the most reliable initialization and typically improves over MR and fixed-binning BR after fine-tuning. While purely supervised training can be optimal for particular model–task pairs, the overall trend indicates a robustness advantage: our pretraining reaches competitive or superior optima with reduced sensitivity to the downstream model choice, which is valuable given diverse encoder choices in medical tabular modeling. Overall, these results suggest that learning-driven discretization acts as a transferable inductive bias that persists under downstream optimization, rather than a probe-specific artifact.

4 Conclusion

We introduce Adaptive Binning, a tabular self-supervised pretext for medical data that elevates discretization from a fixed design choice to a learning-coupled, feature-wise coarse-to-fine curriculum. By designing plateau-triggered refinement, representation-aware split selection, and heterogeneity-aware ordinal supervision, our approach yields stronger representations across tasks and datasets. We further establish a benchmark of medical tabular datasets with unified evaluation protocols, enabling reproducible comparisons for self-supervised learning on clinical tables. Our study is limited to in-dataset transfer and a small set of downstream protocols; future work will extend evaluation to broader clinical endpoints and cross-dataset pretraining with adaptation to new targets.

Acknowledgments. This work was supported by NRF (RS-2024-00338048, RS-2024-00455720, RS-2026-25487796), National Institute of Health (2026-ER0901-00, 2026-ER0904-00), High-Performance Computing Support project (RQT-25-070083), Hankuk University of Foreign Studies Research Fund of 2025, KOCCA R&D Program (RS-2024-00332210), IITP AI Graduate School Program (RS-2020-II201373, Hanyang University), and IITP AI Semiconductor Support Program (RS-2023-00253914).

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Amin, M.B., Greene, F.L., Edge, S.B., Compton, C.C., Gershenwald, J.E., Brookland, R.K., Meyer, L., Gress, D.M., Byrd, D.R., Winchester, D.P.: The eighth edition ajcc cancer staging manual: continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. CA: a cancer journal for clinicians **67**(2), 93–99 (2017)
2. Arik, S.Ö., Pfister, T.: Tabnet: Attentive interpretable tabular learning. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 6679–6687 (2021)
3. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Proceedings of the 26th annual international conference on machine learning. pp. 41–48 (2009)
4. Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., Kasneci, G.: Deep neural networks and tabular data: A survey. IEEE transactions on neural networks and learning systems **35**(6), 7499–7519 (2022)
5. Breiman, L., Friedman, J., Olshen, R.A., Stone, C.J.: Classification and regression trees. Chapman and Hall/CRC (2017)
6. Chen, T., Guestrin, C.: 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)
7. Diaz, R., Marathe, A.: Soft labels for ordinal regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4738–4747 (2019)
8. Du, S., Zheng, S., Wang, Y., Bai, W., O'Regan, D.P., Qin, C.: Tip: Tabular-image pre-training for multimodal classification with incomplete data. In: European Conference on Computer Vision. pp. 478–496. Springer (2024)

9. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., Dean, J.: A guide to deep learning in healthcare. Nature medicine **25**(1), 24–29 (2019)
10. Gorishniy, Y., Rubachev, I., Babenko, A.: On embeddings for numerical features in tabular deep learning. Advances in Neural Information Processing Systems **35**, 24991–25004 (2022)
11. Gorishniy, Y., Rubachev, I., Khrulkov, V., Babenko, A.: Revisiting deep learning models for tabular data. Advances in neural information processing systems **34**, 18932–18943 (2021)
12. Grinsztajn, L., Oyallon, E., Varoquaux, G.: Why do tree-based models still outperform deep learning on tabular data? (2022), `https://arxiv.org/abs/2207.08815`
13. Holub, K., Hardy, N., Kallmes, K.: Toward automated data extraction according to tabular data structure: Cross-sectional pilot survey of the comparative clinical literature. JMIR Formative Research **5**(11), e33124 (2021)
14. Lee, K., Sim, Y., Cho, H.S., Eo, M., Yoon, S., Yoon, S., Lim, W.: Binning as a pretext task: Improving self-supervised learning in tabular domains. In: International Conference on Machine Learning. PMLR (2024)
15. McElfresh, D., Khandagale, S., Valverde, J., Prasad C, V., Ramakrishnan, G., Goldblum, M., White, C.: When do neural nets outperform boosted trees on tabular data? Advances in Neural Information Processing Systems **36**, 76336–76369 (2023)
16. McGorry, P.D., Hickie, I.B., Yung, A.R., Pantelis, C., Jackson, H.J.: Clinical staging of psychiatric disorders: a heuristic framework for choosing earlier, safer and more effective interventions. Australian & New Zealand Journal of Psychiatry **40**(8), 616–622 (2006)
17. Pan, H., Han, H., Shan, S., Chen, X.: Mean-variance loss for deep age estimation from a face. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5285–5294 (2018)
18. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A.: Catboost: unbiased boosting with categorical features. Advances in neural information processing systems **31** (2018)
19. Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F., Bengio, Y., Courville, A.: On the spectral bias of neural networks. In: International conference on machine learning. pp. 5301–5310. PMLR (2019)
20. Rubachev, I., Alekberov, A., Gorishniy, Y., Babenko, A.: Revisiting pretraining objectives for tabular deep learning. arXiv preprint arXiv:2207.03208 (2022)
21. de Sá, C.R., Soares, C., Knobbe, A.: Entropy-based discretization methods for ranking data. Information Sciences **329**, 921–936 (2016)
22. Shwartz-Ziv, R., Armon, A.: Tabular data: Deep learning is not all you need. Information fusion **81**, 84–90 (2022)
23. Somepalli, G., Goldblum, M., Schwarzschild, A., Bruss, C.B., Goldstein, T.: SAINT: Improved neural networks for tabular data via row attention and contrastive pre-training. arXiv preprint arXiv:2106.01342 (2021)
24. Varoquaux, G., Cheplygina, V.: Machine learning for medical imaging: methodological failures and recommendations for the future. NPJ digital medicine **5**(1), 48 (2022)
25. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.: Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on Machine learning. pp. 1096–1103 (2008)

26. Yan, J., Chen, J., Wu, Y., Chen, D.Z., Wu, J.: T2g-former: organizing tabular features into relation graphs promotes heterogeneous feature interaction. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 10720–10728 (2023)
27. Yan, J., Zheng, B., Xu, H., Zhu, Y., Chen, D., Sun, J., Wu, J., Chen, J.: Making pre-trained language models great on tabular prediction. In: International Conference on Learning Representations. vol. 2024, pp. 23570–23593 (2024)
28. Yoon, J., Zhang, Y., Jordon, J., van der Schaar, M.: VIME: Extending the success of self- and semi-supervised learning to tabular domain. In: Advances in Neural Information Processing Systems. vol. 33, pp. 11033–11043 (2020)