XBANK: CROSS-BANK TRANSFER OF TRANSACTION FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised learning enables banks to derive reusable representations from customer transaction histories. However, the large datasets needed for pretraining typically remain proprietary, limiting opportunities to study how these representations transfer across institutions. We introduce XBANK, a dataset comprising around 91.5 million daily-aggregated transaction records from 378,305 corporate clients, with 13 categorical features, two numerical features, and three monthly product-propensity targets. Together with the Multimodal Banking Dataset (MBD), which contains transaction histories for approximately 1.5 million corporate clients, XBANK supports a benchmark for cross-bank transfer of frozen representations. We evaluate five models spanning contrastive, autoregressive, masked-modelling, and temporal point-process objectives. By evaluating MBD-pretrained encoders on raw MBD transactions, their daily aggregates, and XBANK, we isolate the effect of aggregation within the source institution and assess cross-bank transfer at the same aggregation level under each dataset's native evaluation protocol. To handle unknown field correspondences, we propose a dictionary-free schema-alignment method based on semi-relaxed Fused Gromov-Wasserstein optimization. It combines statistical column profiles with within-bank feature dependencies, accommodates schemas of different sizes, and determines a fixed mapping before evaluation without using downstream labels. Together, these contributions provide a framework for studying the transfer of transaction representations across banks with heterogeneous data schemas.

1 INTRODUCTION

Customer transaction histories provide a basis for predicting financial behaviour and demand for banking products. These histories form event sequences in which each transaction combines numerical and categorical attributes with a timestamp, and intervals between events are irregular. To model such sequences, self-supervised learning methods are widely adopted: contrastive learning (CoLES), masked event modeling (MLM), next-event prediction (NEP), as well as temporal point process models (THP, COTIC) that rely on inter-event intervals as their primary signal. Chronos-2 serves as an external baseline in a zero-shot regime, consuming only the transaction-amount channel. The trained model produces vector representations of client histories—embeddings—which serve as the basis for downstream tasks such as forecasting demand for banking products.

Applying a pretrained encoder at another bank introduces several changes at once. First, the institutions may differ in their client populations, transaction distributions, and category vocabularies. Second, their data pipelines may represent financial activity at different levels of aggregation: one may retain individual transactions, while another stores daily totals for each combination of categorical attributes. Third, the correspondence between their feature schemas may be unknown, particularly when field names and category codes are anonymized. Each change can affect the representations produced by a frozen encoder. Evaluating transfer therefore requires examining both the conditions under which the encoder is applied and the alignment of the inputs it receives.

The public Multimodal Banking Dataset (MBD) Mollaev et al. (2025) stands as the first open, large-scale source of corporate transaction (B2B) histories for pretraining. To study transfer to another institution, we introduce XBANK, an anonymized corpus comprising 91.5 million daily-aggregated transaction records from 378,305 corporate clients over 2022–2024, with 13 categorical features,

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.

two numerical features, and three binary targets indicating monthly banking product purchases. Together, the two datasets support a benchmark for cross-bank transfer of frozen transaction representations. We pretrain five models on MBD, covering contrastive, autoregressive, masked-modelling, and temporal point-process objectives. Each encoder is then kept frozen, while downstream classifiers are trained on its embeddings to predict each dataset's own product-propensity targets. The two datasets allow us to examine aggregation and institution changes through three evaluation conditions. Whereas XBANK contains only daily aggregates, MBD retains individual transactions with precise timestamps. We construct a aggregated version of MBD by grouping transactions by client and day. Each encoder pretrained on raw MBD is evaluated on the original transactions, their daily aggregates, and XBANK. Comparing the two MBD versions isolates the effect of aggregation while holding the client population, prediction targets, and evaluation protocol fixed. Evaluation on XBANK then examines cross-bank transfer with daily-aggregated inputs.

To support transfer without a field dictionary, we propose a schema-alignment method based on semi-relaxed Fused Gromov–Wasserstein optimization. The method combines statistical profiles of individual columns with the dependency structure among features within each bank. It accommodates schemas of different sizes by allowing an XBANK column to supply more than one input field of the pretrained encoder and leaving fields unmatched when no admissible candidate exists. We first reconstruct the field correspondences within MBD after masking column identities and evaluate recovery against the known schema. We then apply the same method to map XBANK columns to the input fields of the MBD-pretrained encoder. The resulting mapping is fixed before downstream evaluation without using prediction labels. This procedure addresses column correspondence; aligning the meanings of individual category codes across banks remains outside its scope.

Our contributions are:

- **XBANK**, a public corpus of 91.5M daily-aggregated corporate-client transaction rows over $\sim$378K clients, released with three monthly product-propensity targets, forming with MBD the first public cross-institution banking pair with genuinely distinct schemas.
- **A nested decomposition of the transfer gap**, separating degradation caused by a change in temporal aggregation from the incremental degradation caused by a change of institution, across five architectures spanning self-supervised objective families.
- **A dictionary-free schema-alignment protocol** based on semi-relaxed Fused Gromov–Wasserstein optimization combining column profiles and feature dependencies to obtain a fixed input mapping without downstream labels.

2 Related Work

Self-supervised models of event sequences. Foundation models for transaction histories differ primarily in the self-supervised objective used to learn from unlabeled client histories. Contrastive methods, of which CoLES (Babaev et al., 2022) is the canonical instance in the transactional domain, form positive pairs from disjoint sub-sequences of the same client's history and negatives from other clients, yielding an embedding space in which proximity encodes behavioural similarity. MLM adapts the BERT objective (Devlin et al., 2019) to event sequences: a fraction of events is masked and reconstructed from bidirectional context, which forces the encoder to model co-occurrence structure among categorical attributes. Autoregressive next-event prediction instead factorizes the sequence causally, predicting each event from its prefix. These three families treat the sequence as an ordered list of events and model timestamps, if at all, as an auxiliary feature.

A distinct family models client histories as realizations of a marked temporal point process, where the learning signal is the conditional intensity governing *when* the next event occurs, not only *what* it is. The Transformer Hawkes Process (Zuo et al., 2020) parameterizes this intensity with self-attention over event history, while COTIC (Zhuzhel et al., 2025) uses a continuous-time convolutional backbone. This distinction matters directly for our study: point-process models depend on inter-arrival times as their primary signal, and are therefore the family most exposed to a change in temporal aggregation, which is precisely one of the shifts we isolate.

Public banking transaction datasets. Public transaction corpora have overwhelmingly originated from data-science competitions (Alfa-Bank, 2020; Sberbank, 2021; Rosbank, 2021; VTB Bank,

2022; American Express, 2022). These share three properties that limit their use for transfer research. First, nearly all cover retail (B2C) clients, whose behavioural dynamics differ substantially from corporate accounts. Second, each defines its own schema, encoding, and target, so cross-dataset evaluation confounds representation quality with incompatible task definitions. Third, and most importantly, no two of them are matched closely enough in client segment and task structure to serve as source and target of a controlled transfer experiment. MBD (Mollaev et al., 2025) is the first public corpus at industrial scale covering corporate clients, with multiple modalities and a well-specified five-fold client-disjoint benchmark protocol; prior work using it has evaluated architectures in-domain, on those folds, rather than as one half of an institutional pair.

Dataset shift and transfer in structured domains. Dataset-shift research typically asks how source and target differ once the feature space itself is held fixed. This framing does not extend to a setting where the feature space itself differs across source and target—where the two disagree not on the distribution of a feature, but on what a feature *is*. Two shifts specific to structured event data fall into exactly this gap. The first is *aggregation shift*: the same underlying events summarized at a different temporal granularity, so that a model trained on raw events encounters pre-summed rows at deployment, losing intra-period ordering and inter-arrival structure while preserving totals. A recent tabular benchmark studies a related feature-shift setting—columns added, removed, or replaced within a single schema (Cheng et al., 2025)—but assumes source and target are drawn from the same data-preparation pipeline; aggregation shift as we define it changes the pipeline itself, not the column set it emits.

The second is *schema shift*: the same semantic fields presented under an unknown permutation of column identities, with no dictionary linking them. This is the setting studied by *heterogeneous transfer learning*, the branch of transfer learning concerned with source and target occupying genuinely different feature spaces, which most standard transfer methods rule out by assumption (Bao et al., 2026). A recent survey of the field further separates methods into symmetric approaches, which project both feature spaces into a shared learned subspace, and asymmetric approaches, which map one feature space into the other without retraining it (Bao et al., 2026); a frozen, already-pre-trained encoder forces the asymmetric case, since the source-side representation cannot be touched. The same survey notes that most heterogeneous transfer methods, symmetric or asymmetric, still assume some labeled target data or a pivot set of instances known to correspond across domains—an assumption a receiving institution with an unlabeled, dictionary-free corpus cannot satisfy. Correspondence-free alignment methods built on Gromov-Wasserstein optimal transport, which match domains by comparing intra-domain distance structure rather than requiring paired points, relax exactly this assumption (Alvarez-Melis & Jaakkola, 2018), and have been used to align feature spaces with no shared instances or dimensionality at all (Demetci et al., 2022). We adopt semi-relaxed Fused Gromov–Wasserstein optimization (Vincent-Cuaz et al., 2022) for this setting, pairing it with column-profile and within-bank structural distances designed for cross-institution schema matching, and fix the resulting correspondence before downstream evaluation.

Both aggregation shift and schema shift arise routinely when a representation crosses an organizational boundary, and both are invisible in benchmarks where source and target share a data-preparation pipeline and a field dictionary. Our benchmark examines aggregation effects through controlled transformations of MBD and addresses unknown field correspondences through a fixed, dictionary-free alignment procedure.

3 THE XBANK DATASET

We introduce XBANK, an anonymized dataset of corporate banking transactions released with this work. It contains 91,547,558 daily-aggregated records from 378,305 clients of a major commercial bank, together with three monthly product-propensity targets. Table 1 summarizes the dataset.

Row semantics and aggregation. Each row represents a unique combination of client, calendar day, and categorical feature values. The numeric fields carry amounts already summed over whatever raw transactions share that day and that combination of categorical values; the corpus contains no raw per-transaction records. Because the date field carries day resolution with no time-of-day component, XBANK flattens the sub-day inter-arrival structure that continuous-time point-process models are designed to exploit. This is not a defect of the release but a property of the underlying

Table 1: XBANK transactions summary statistics.

Clients	∼378K
Rows	91,547,558
Row granularity	one row per (client, day, category-combination)
Transaction period	2022-01-01 – 2024-01-31
Target period	2023-01-01 – 2024-02-29
Categorical features	13
Numeric features	2 (summed amounts)
Targets	3 binary, monthly product-propensity flags

warehouse format, and it is one of the conditions our benchmark measures rather than controls away. It is also the reason the aggregation-level variation in our benchmark is constructed on the MBD side (Section 4): XBANK's granularity cannot be un-aggregated after the fact. XBANK is scoped to small and medium-sized enterprises (SMEs): clients whose total transaction volume over the collection period exceeds the 99th percentile of the originating institution's full client base are excluded, since firms at that scale are treated as large corporates rather than SMEs and fall outside this study's target population.

Targets. The target table contains three binary targets, each indicating whether a client purchases a given banking product in a given month, aligned to a monthly grid. Target months do not cover the transaction period uniformly—a small number of months are absent from the target file entirely—and our evaluation code surfaces such gaps explicitly rather than silently producing empty splits. Per-target base rates are reported in Appendix A.1.

Anonymization and release. Client identifiers are irreversibly hashed prior to release: the `id` field is a salted SHA-256 digest of the original client identifier — the salt is fixed within this release (so a client's `id` is stable across all their rows here) but distinct per release, so the digest cannot be reversed to recover the original identifier or linked to the same client's `id` in any other release or system. Categorical fields (`col_2`–`col_10`, `col_13`–`col_16`) are re-coded to institution-specific integer or small-cardinality codes that do not correspond to any external standard classification, so that no field can be joined against outside data sources by value alone. Numeric fields (`col_11`, `col_12`) are min-max scaled using statistics computed on the released 378K-client population itself, not on the originating institution's full internal client base. The anonymization procedure was reviewed and approved by the bank's legal and compliance team before release.

4 BENCHMARK DESIGN: CROSS-INSTITUTION TRANSFER

4.1 ARCHITECTURES

We pretrain five models spanning self-supervised objective families under a common data preparation and evaluation protocol to compare their transfer behaviour.

CoLES (contrastive) encodes client histories with a recurrent sequence encoder and optimizes a contrastive loss over sub-sequence pairs drawn from the same client. **NEP** (autoregressive) predicts the next event from a causal transformer encoding of the prefix. **MLM** (masked modelling) reconstructs masked events from a bidirectional transformer encoding of the surrounding context. **THP** and **COTIC** (continuous-time point processes) model the conditional intensity of the next event in continuous time, over transformer and convolutional backbones respectively; both consume event time and a single categorical mark rather than the full feature set, a real limitation of the marked-TPP formulation that we note rather than work around.

Client histories are truncated to their 500 most recent events. All five models are trained for up to 100 epochs, with early stopping (patience 5) on a client-disjoint validation set comprising 5% of clients. CoLES samples shorter subsequences during training.

All architectures are pretrained on MBD, so that released checkpoints and the benchmark protocol remain reproducible from public data alone. XBANK is reserved entirely as a transfer target and is

Table 2: Pre-training hyperparameters. All models: `max_epochs` 100

Model	Architecture parameters	Batch size
CoLES	hidden 128, embedding 32, 2 layers	256
NEP	d_{model} 128, 1 layer, lr 10^{-3}	256
MLM	d_{model} 128, 1 layer, lr 10^{-3}	256
THP	hidden 128, 4 layers, lr 10^{-3}	256
COTIC	128 filters, 128 channels, 6 layers	32

never seen during pretraining. Section 5 describes how XBANK columns are mapped to the input fields of the pretrained encoders.

4.2 Rendering MBD at two aggregation levels

MBD ships as raw per-transaction records. We additionally construct a daily-aggregated rendering by grouping records by (client, day, full category-combination) and summing the transacted amount within each group—exactly the convention under which XBANK is natively stored. Amount normalization is applied *after* aggregation, on the resulting per-day sums.

This yields two renderings of the same client population, differing only in temporal granularity: MBD-raw, with ∼947M rows, and MBD-daily, with ∼557M rows—a 41.2% reduction. Both cover the same number of distinct clients. The reduction is smaller than daily aggregation typically yields on XBANK-style data because several of MBD's sender- and receiver-type fields are high-cardinality, so two same-day transactions rarely share the exact category combination that an aggregation group requires.

4.3 A nested decomposition of the transfer gap

Because XBANK is natively daily, a model pre-trained on MBD-raw and deployed on XBANK undergoes two simultaneous shifts: the data format changes from raw events to daily aggregates, and the institution changes. Evaluating the *same* pre-trained models on three successive conditions separates these (Table 3).

Table 3: Three evaluation conditions for a single set of MBD-raw pre-trained models, tracing a nested path in which the aggregation level changes first and the institution changes second.

	Evaluated on	Aggregation	Institution
(A) Source domain	MBD-raw	raw (matched)	MBD (matched)
(B) Aggregation shift	MBD-daily	daily (shifted)	MBD (matched)
(C) Full transfer	XBANK	daily (shifted)	XBANK (shifted)

From these we report two quantities. The *aggregation-shift component*, $\Delta_{\text{agg}} = M(A) - M(B)$, measures degradation caused purely by presenting pre-summed daily rows to a model trained on raw events, with the institution and client population held fixed. The *incremental institution-shift component*, $\Delta_{\text{inst}|\text{agg}} = M(B) - M(C)$, measures the additional degradation from changing institution once the format has already changed. Their sum is the total transfer gap $M(A) - M(C)$ by construction.

We emphasize what this design identifies: the aggregation effect cleanly, and the institution effect *conditional on the aggregation level already having shifted*. It does not identify an institution effect at matched aggregation, nor an interaction term between the two shifts. We regard that as the natural extension of this benchmark and leave it to future work.

4.4 Evaluation protocol

Representations are frozen at evaluation: no fine-tuning, and no re-fitting of any preprocessor or category vocabulary on target data, which would defeat the purpose of a zero-shot transfer test. For

each client and each target date, an embedding is computed from that client's trailing history window up to—but never past—that date.

On MBD (conditions A and B), we follow the dataset's own published protocol: five client-disjoint folds, each rotated in turn as the held-out test set with the remaining four used for training, reporting mean $\pm$ standard deviation across the five rotations rather than a single arbitrarily chosen fold. Since fold assignment is per client, no client straddles the train/test boundary in any rotation; a further client-grouped split carves validation data out of the training folds. On XBANK (condition C), we hold out a later calendar period—training on 2023 and testing on 2024-01/02—so that the probe is never evaluated on a period it could have trained on, and split train from validation by unique client identifier so that a client's several monthly rows cannot straddle the two. These two protocols test different notions of generalization: MBD's folds test generalization across clients, while XBANK's calendar cut tests generalization across clients and time. We inherit each dataset's native convention rather than forcing them into artificial alignment, and flag the asymmetry explicitly wherever the two are compared.

For each target column we train independent probe on the frozen embeddings with gradient-boosted trees (LightGBM) and report ROC-AUC, PR-AUC, and precision and recall at 5%. Targets in both corpora are severely imbalanced (see A.1 for details), so all probes apply class weighting derived from the training labels.

5 Schema Alignment Without a Dictionary

The benchmark of Section 4 presupposes a known correspondence between MBD's fields and XBANK's column slots. In practice, an institution receiving a frozen representation trained elsewhere frequently has no such dictionary. We formulate recovering this correspondence as an optimal transport problem: for each MBD field Y_j, choose which XBANK column X_i supplies its data, using only the statistical shape of each column — never its name — as evidence. A single XBANK column may be selected as the source for more than one MBD field, since the two schemas need not have the same number of columns.

Column passports. Each column is summarized by a *passport*: a small set of statistics computed over its valid (non-missing) values, plus a missingness rate r_{null} computed over all rows. Categorical passports carry cardinality U, normalized entropy H_{norm}, and the frequency of the most common value $top1$; numeric passports carry a 21-point quantile vector $z(p)$, $p \in \{0.01, 0.05, 0.10, \ldots, 0.95, 0.99\}$, normalized by a robust scale $b = Q(0.95) - Q(0.05)$ (falling back to $Q(0.99) - Q(0.01)$, then the full range, if this is degenerate), together with a zero-rate r_{zero}. This normalization is invariant to shifting and positively rescaling a column's values — exactly the invariance needed when two institutions encode the same underlying quantity on different scales.

Passport distance. For a pair of matching-type columns (X_i, Y_j) (categorical–categorical or numeric–numeric), we compare their *passports* and define a passport distance $D_{ij} \in [0, 1]$ between the two summaries, smaller meaning the two columns' statistical shapes are more alike. For categorical pairs,

$$D_{ij} = 0.35\, d_U + 0.30\, d_H + 0.20\, d_{top1} + 0.15\, d_{\text{null}},$$

where d_H, d_{top1}, d_{null} are absolute differences of the corresponding passport statistics and d_U is a sample-size-corrected difference in log-cardinality (a pair with $d_U > 1$ is excluded outright). Exact definitions are provided in Appendix A.2. For numeric pairs,

$$D_{ij} = 0.70\, d_Q + 0.15\, d_{\text{zero}} + 0.15\, d_{\text{null}}, \qquad d_Q = \frac{\ell_Q}{1 + \ell_Q}, \quad \ell_Q = \frac{1}{21} \sum_p \left| z_{X_i}(p) - z_{Y_j}(p) \right|.$$

The combination weights above are hand-tuned, not learned or validated — a limitation we carry forward rather than obscure. A pair is excluded from consideration entirely if the two columns' types or numeric subtypes disagree, if either has no valid values, or if D_{ij} exceeds a fixed threshold (0.70 by default). Let $\mathcal{A}_j$ denote the set of XBANK columns surviving these filters for field Y_j; if $\mathcal{A}_j = \emptyset$, Y_j is left `unfilled` rather than forced to an unjustified assignment.

Column passports alone discard a signal that matters: how columns relate to *each other* within one institution's schema. We additionally compute, separately within XBANK and within MBD, a

pairwise structural distance between columns based on normalized mutual information over jointly-valid rows (numeric columns discretized into decile bins for this purpose only):

$$C^X_{ik} = 1 - \mathrm{NMI}(X_i, X_k), \qquad C^Y_{jl} = 1 - \mathrm{NMI}(Y_j, Y_l).$$

A correct schema mapping should not just match columns whose individual distributions look alike; it should preserve the relational structure among columns, so that two XBANK columns that are strongly associated map onto two MBD fields that are similarly associated.

Fused Gromov-Wasserstein assignment. Both signals are combined into a single transport problem over an assignment matrix $T_{ij} \geq 0$, interpreted as the mass of MBD field j's data attributed to XBANK source i:

$$F(T) = (1-\alpha)\sum_{i,j} D_{ij}T_{ij} + \alpha \sum_{i,j,k,l} \left(C^X_{ik} - C^Y_{jl}\right)^2 T_{ij}T_{kl}, \qquad T^* \in \arg\min_T F(T),$$

subject to $T_{ij} = 0$ for $i \notin \mathcal{A}_j$ and $\sum_i T_{ij} = 1/m$ for each of the m active MBD fields. Critically, only the MBD-side mass is constrained to sum to a fixed value; XBANK columns' row sums are left free. This is what makes the problem *semi-relaxed*, as opposed to balanced Gromov-Wasserstein, which fixes both sides — and it is what permits one XBANK column to be the chosen source for several MBD fields, the many-to-one case a bijective dictionary cannot express. The weight $\alpha \in [0, 1]$ trades off passport similarity against structural consistency; at $\alpha = 0$ only column shape matters, at $\alpha = 1$ only relational structure does (subject to the same admissibility filters throughout). F is non-convex in T; we solve it with Frank–Wolfe exact line search (Titouan et al., 2019) from five initializations (uniform-over-candidates, greedy nearest-passport, and three random restarts) and keep the lowest-loss result. A run's `CONVERGED` status certifies that the stopping criterion was met, not that a global optimum was found.

From transport plan to a mapping. The solved plan is normalized per MBD field, $P_{ij} = T_{ij}/\sum_k T_{kj}$, and each field is finally assigned its highest-weight source, $i^*(j) \in \arg\max_{i \in \mathcal{A}_j} P_{ij}$, with ties broken first by smaller D_{ij} and then lexicographically. A weight of $P_{i^*(j),j} = 1$ does not mean the assignment is correct with probability 1 — it means the optimizer's found (possibly non-optimal, possibly wrong) plan placed all of field j's mass on one source. The size of the margin between the best and second-best candidate is not, by itself, evidence of reliability: a degenerate zero/one plan can produce a margin of 1 even when the underlying assignment is wrong. This is precisely why the mapping's accuracy is validated separately, against a testbed with a known ground-truth correspondence (masked MBD–MBD recovery), rather than read off the optimizer's own loss or confidence.

6 Experiments

6.1 Setup

All pre-training was performed on 2×NVIDIA RTX A5000 (24GB) GPUs. The category vocabulary is fit on the first chunk and reused, unchanged, for all subsequent chunks, exactly as it is later reused unchanged at inference time. COTIC uses a smaller batch size than the other architectures: at MBD's sequence lengths, which sit at the 500-event cap for most clients, the common setting exceeded device memory. We report this as a genuine constraint on point-process architectures at this scale, not a tuning choice.

6.2 Results

Aggregation shift does not uniformly degrade performance. Tables 4, 5 report mean PR-AUC and ROC-AUC across all four MBD targets and five folds, for the same pre-trained encoders evaluated on MBD-raw (condition A) and MBD-daily (condition B). Contrary to the expectation that temporal aggregation should degrade representation quality — and should do so most for the point-process architectures, which depend on inter-arrival timing (Section 2) — four of five encoders score *higher* on daily-aggregated inputs than on raw ones, on both metrics. Only CoLES shows the expected-direction (and small) degradation. Δ_{agg} is therefore negative — an improvement, not a

cost — for CoTIC, THP, NEP, and MLM. We report this as a genuine, unexpected finding rather than adjust the hypothesis after the fact; per-target detail (Appendix A.4 (Tables 10–11) show the pattern holds across most individual targets, not just in aggregate, though fold-to-fold variance is large for the rarest targets, so individual-target rankings should be read cautiously).

Table 4: PR-AUC across four MBD targets and five folds, condition A (MBD-raw) vs. condition B (MBD-daily). $\Delta_{\text{agg}} = M(A) - M(B)$; negative means daily outperforms raw.

Encoder	PR-AUC (A)	PR-AUC (B)	Δ_{agg} (pp)
CoLES	0.819%	0.808%	$+0.011$
COTIC	1.339%	1.378%	-0.039
THP	1.511%	1.527%	-0.016
NEP	1.768%	1.791%	-0.023
MLM	1.638%	1.733%	-0.095

Table 5: ROC-AUC, same conditions as Table 4.

Encoder	ROC-AUC (A)	ROC-AUC (B)	Δ_{agg}
CoLES	0.6795	0.6786	$+0.0009$
COTIC	0.7820	0.7849	-0.0029
THP	0.7940	0.7961	-0.0021
NEP	0.8113	0.8141	-0.0028
MLM	0.8045	0.8099	-0.0054

Full transfer under the recovered schema mapping. On XBANK (condition C), we compare downstream performance under a simpler Hungarian-assignment column matcher (one-to-one, cosine-similarity cutoff) against the mapping recovered by the semi-relaxed FGW procedure of Section 5, matched pairwise within each encoder on the same evaluation cohort.

Table 6: XBANK (condition C), old vs. recovered FGW mapping: mean effect per encoder, averaged over its 3 targets (full per-target breakdown in Appendix A.4, Tables 12–13). Δ = New $-$ Old.

Encoder	ΔPR-AUC (pp)	ΔROC-AUC	ΔP@5% (pp)	ΔR@5% (pp)
CoLES	$+0.393$	$+0.0059$	$+0.02$	-0.18
COTIC	-0.628	$+0.0269$	-3.54	-2.30
THP	$+0.098$	$+0.0378$	-1.66	$+0.18$
NEP	-0.031	$+0.0196$	-0.54	-0.03
MLM	$+0.147$	$+0.0265$	-0.61	-0.12

Across the 15 encoder×target combinations, the recovered mapping improves ROC-AUC in 13/15 cases and PR-AUC in 10/15 — real, if not universal, improvements. P@5% improves in only 9/15 rows, which undersells the direction of the effect: averaged per encoder (Table 6), mean P@5% decreases for four of five encoders, since a few large losses on the highest-prevalence target outweigh several small gains elsewhere. ROC-AUC and top-of-list precision and recall thus move in different aggregate directions across the benchmark. The clearest case is COTIC on its highest-prevalence target: ROC-AUC improves substantially ($0.650 \rightarrow 0.709$), while P@5% and R@5% nearly halve ($21.7\% \rightarrow 10.9\%$ and $15.2\% \rightarrow 7.6\%$) — a mapping can improve overall ranking while worsening top-of-list predictions for a specific pair. It bears directly on Section 5's point that a transport plan's confidence is not evidence of downstream safety.

7 Discussion and Limitations

The decomposition is nested, not factorial. We do not report a single $\Delta_{\text{inst}|\text{agg}}$ number derived by directly subtracting condition C's metrics from condition B's. MBD and XBANK predict different targets with substantially different base rates, so raw PR-AUC is not comparable across the

two without normalizing for prevalence (Section 4.4). A naive $M(B) - M(C)$ would conflate the institution-shift effect with this base-rate confound. We leave a prevalence-normalized version of this comparison to future work.

Asymmetric evaluation protocols. MBD's client-disjoint folds and XBANK's calendar cut test different notions of generalization. Comparisons across the two corpora should be read as comparisons of transfer behaviour under each dataset's own native protocol, not as a single metric on a common axis.

Aggregation is confounded with rescaling. Section 4 states that MBD-raw and MBD-daily differ only in temporal granularity, but amount normalization is refit separately on each rendering's post-aggregation sums, placing the two on measurably different effective scales. We cannot rule out that part of the aggregation-shift result in Section 6 — daily-aggregated inputs outperforming raw for four of five encoders — reflects this rescaling rather than aggregation itself. Disentangling the two would require holding the amount channel's scale fixed across both renderings.

Schema-alignment gains are metric-dependent, not just non-uniform. The recovered mapping's benefit in Section 6 is real on ROC-AUC but frequently absent or negative on P@5%, the metric closest to how such a mapping would actually be used downstream. Evaluating a schema-alignment method by a single summary statistic risks missing exactly the failure mode that matters operationally.

8 Conclusion

We introduced XBANK, an anonymized corpus of 91.5M daily-aggregated corporate-client transaction rows, which together with MBD forms the first public pair of industrial-scale banking datasets with genuinely distinct schemas. Using this pair, we built a nested benchmark examining aggregation changes and cross-bank transfer, and a dictionary-free schema-alignment protocol based on semi-relaxed Fused Gromov-Wasserstein optimization.

The experiments revealed two main findings. Under the evaluated preprocessing pipeline, daily-aggregated inputs yielded slightly higher mean ROC-AUC and average precision for four of five encoders, including both point-process models; only CoLES showed a small decline. This comparison reflects both aggregation and the associated amount scaling. We also do not collapse the institution-shift component into a single number, since MBD and XBANK's differing target base rates would conflate a real institutional effect with an unadjusted prevalence confound — a boundary of what our nested design can isolate, left to future work.

The downstream effects of schema alignment are mixed: the FGW mapping improves ROC-AUC in 13 of 15 encoder–target combinations, while mean precision (P@5%) and mean recall (R@5%) each decrease for four of five encoders. A method's aggregate discriminative quality does not automatically speak to its behavior on an operational top-k metric, which is worth checking directly before deployment. Dictionary-free transfer is feasible and, in aggregate, competitive with a similarity-based baseline; we release XBANK, our checkpoints, and the alignment protocol to support further study of this gap.

Reproducibility Statement

The XBANK dataset and benchmark code will be released publicly upon acceptance; MBD itself is already publicly available (Mollaev et al., 2025). Exact splits, hyperparameters and compute budget are specified in (Table 9), and Appendix A.3.

AI use statement

Generative AI assistants were used in literature search, data inspection and computational audits. Reported numerical results come from executed experiments and retained output artifacts. The authors are responsible for the submitted text, claims, references, and artifacts.

REFERENCES

Alfa-Bank. Alfa battle 2.0: Machine learning competition. https://boosters.pro/championship/alfabattle2/overview, 2020. Accessed: 2024-06-07.

David Alvarez-Melis and Tommi Jaakkola. Gromov-Wasserstein alignment of word embedding spaces. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1881–1890, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1214.

American Express. American express - default prediction. https://www.kaggle.com/competitions/amex-default-prediction, 2022. Accessed: 2024-06-07.

Dmitrii Babaev, Nikita Ovsov, Ivan Kireev, Maria Ivanova, Gleb Gusev, Ivan Nazarov, and Alexander Tuzhilin. Coles: Contrastive learning for event sequences with self-supervision. In *Proceedings of the 2022 International Conference on Management of Data*, SIGMOD '22, pp. 1190–1199, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392495. doi: 10.1145/3514221.3526129. URL https://doi.org/10.1145/3514221.3526129.

Runxue Bao, Yiming Sun, Yuhe Gao, Jiabao Li, Ran Fang, Jindong Wang, Qiang Yang, Zhi-Hong Mao, Chenxi Song, Ai Ren, and Ye Ye. A recent survey of heterogeneous transfer learning. *Knowledge-Based Systems*, 350:116577, 2026. ISSN 0950-7051.

Zi-Jian Cheng, Zi-Yi Jia, Zhi Zhou, Yu-Feng Li, and Lan-Zhe Guo. Tabfsbench: Tabular benchmark for feature shifts in open environments, 2025. URL https://arxiv.org/abs/2501.18935.

Pinar Demetci, Rebecca Santorella, Björn Sandstede, William Stafford Noble, and Ritambhara Singh. Scot: Single-cell multi-omics alignment with optimal transport. *Journal of Computational Biology*, 29(1):3–18, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

Dzhambulat Mollaev, Alexander Kostin, Maria Postnova, Ivan Karpukhin, Ivan Kireev, Gleb Gusev, and Andrey Savchenko. Multimodal banking dataset: Understanding client needs through event sequences. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2025. arXiv:2409.17587.

Rosbank. Rosbank ml competition: Client churn prediction. https://boosters.pro/championship/rosbank1/overview, 2021. Accessed: 2024-06-07.

Sberbank. Age prediction dataset (sberbank sirius lesson competition). https://ods.ai/competitions/sberbank-sirius-lesson, 2021. Accessed: 2024-06-07.

Vayer Titouan, Nicolas Courty, Romain Tavenard, Chapel Laetitia, and Rémi Flamary. Optimal transport for structured data with application on graphs. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6275–6284. PMLR, 09–15 Jun 2019.

Cédric Vincent-Cuaz, Rémi Flamary, Marco Corneli, Titouan Vayer, and Nicolas Courty. Semi-relaxed gromov-wasserstein divergence with applications on graphs, 2022. URL https://arxiv.org/abs/2110.02753.

VTB Bank. Data fusion contest 2022: Matching and educational status prediction. https://ods.ai/tracks/data-fusion-2022-competitions, 2022. Accessed: 2024-06-07.

Vladislav Zhuzhel, Vsevolod Grabar, Galina Boeva, Alexander Stepikin, Artem Zabolotnyi, Vladimir Zholobov, Maria Ivanova, Mikhail Orlov, Ivan Kireev, Evgeny Burnaev, Rodrigo Rivera-Castro, and Alexey Zaytsev. Cotic: Embracing non-uniformity in event sequence data via multilayer continuous convolution. *IEEE Access*, 13:210741–210757, 01 2025. doi: 10.1109/ACCESS.2025.3641785.

Simiao Zuo, Haoming Jiang, Zichong Li, Tuo Zhao, and Hongyuan Zha. Transformer Hawkes process. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 11692–11702. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/zuo20a.html.

A Appendix

A.1 Per-target statistics

XBANK. Table 7 reports, for each of the three binary product-propensity targets, the positive rate pooled over the full target period. Base rates are severely imbalanced across all four targets (Section 4.4), which motivates the class-weighting and undersampling used by both probes.

Table 7: XBANK per-target positive rate pooled over all rows ($n_{\text{rows}} \sim 5.0M$).

	target_3	target_4	target_5
Target rate, %	1.3	7.8	0.9

MBD. The corresponding per-fold positive rates for MBD's native targets are reported in Table 8, following the dataset's published five-fold client-disjoint protocol (Section 4.4).

Table 8: MBD per-target positive rate by fold.

Target	Pooled rate,%	Per-fold range,%
col_2	0.448	0.444-0.452
col_3	0.050	0.047-0.053
col_4	0.375	0.369-0.378
col_5	0.253	0.245-0.260

A.2 Schema-alignment details

Cardinality correction. For a categorical column A, let N_A denote the number of valid observations and U_A the number of distinct valid values.

$$\delta_U = \log_{10} \frac{U_{X_i}}{U_{Y_j}}, \qquad \delta_N = \log_{10} \frac{N_{X_i}}{N_{Y_j}},$$

$$h = \begin{cases} |\delta_N|, & \delta_U \delta_N > 0, \\ 0, & \text{otherwise}, \end{cases} \qquad d_U = \frac{\max(0, |\delta_U| - h)}{0.75}.$$

This heuristic allows for higher observed cardinality in larger samples when both differences have the same direction. Pairs with $d_U > 1$ are excluded. Columns with no valid observations are excluded before computing these quantities.

Quantile normalization. For a numeric column A, let $Q_A(p)$ denote its empirical quantile over valid finite observations. We compute

$$z_A(p) = \frac{Q_A(p) - Q_A(0.50)}{b_A}, \qquad p \in \{0.01, 0.05, 0.10, \ldots, 0.95, 0.99\}.$$

The scale b_A is the first positive value among $Q_A(0.95) - Q_A(0.05)$, $Q_A(0.99) - Q_A(0.01)$, and $\max(A) - \min(A)$. For a constant column, all normalized quantiles are set to zero. The quantile representation is invariant to additive shifts and positive rescaling; the zero fraction used separately in the numeric distance is not invariant to shifts.

Structural dissimilarity. After discretization, we use arithmetic-normalized mutual information:

$$\mathrm{NMI}(A, B) = \frac{2I(A; B)}{H(A) + H(B)}, \qquad C(A, B) = 1 - \mathrm{NMI}(A, B),$$

where entropies and mutual information are estimated on jointly valid observations. Diagonal entries are zero. Off-diagonal entries are set to one when fewer than 60 jointly valid observations are available or either column is constant on those observations. This fallback represents an unavailable dependency estimate, not established independence.

A.3 Reproducibility details

Software and hardware. All experiments used PyTorch 2.5.4, pytorch-lifestream 0.7.0, and DuckDB 1.5.5 for data loading. Pre-training ran on 2×NVIDIA RTX A5000 (24GB) GPUs; COTIC and THP were run one-per-GPU rather than data-parallel across both, since ptls-lifestream's data pipeline is single-process per run. Full container/environment specification, pinned by digest, is released alongside the code.

Computational audit All five architectures use seed 0, `valid_frac`=0.05, and a client-disjoint train/validation split decided once on the full client-identifier list before any chunked loading begins. Models converged under early stopping on MBD-raw. Table 9 reports wall-clock training time per architecture.

Table 9: Pre-training compute budget on MBD-raw.

Model	Validation metric	Wall-clock (h)	Epochs to convergence
CoLES	recall@k	2.6	7
NEP	loss	34.9	82
MLM	loss	26.7	56
THP	NLL	9.7	10
COTIC	log-likelihood	5.4	10

Schema-alignment solver parameters. The mapping used to produce the XBANK results in Section 6 was computed with $\alpha = 0.5$, five initializations ($n_{\text{init}} = 5$), a 300-iteration cap per run, stopping tolerance 10^{-9}, and seed 42. Column passports were built from a 2,000,000-row sample per institution; structural (mutual-information) distances were estimated on a 200,000-row subsample with numeric columns discretized into 10 bins and categorical columns reduced to at most 64 groups, including a pooled group for rare categories. Admissible pairs were required to fall under a maximum passport distance of 0.70, and structural-distance estimates required at least 60 jointly-valid rows per column pair.

A.4 Full downstream results by target

Table 10: MBD PR-AUC by encoder and target, raw vs. daily (mean over 5 folds; SD omitted for space, see released CSVs). Test-row counts differ by ∼3.5K out of ∼11M between raw and daily due to aggregation.

Encoder	Target	Raw PR-AUC	Daily PR-AUC	ΔPR (pp)
CoLES	col_2	0.695%	0.705%	+0.010
CoLES	col_3	0.591%	0.543%	−0.048
CoLES	col_4	0.735%	0.728%	−0.006
CoLES	col_5	1.256%	1.254%	−0.002
COTIC	col_2	1.473%	1.537%	+0.064
COTIC	col_3	0.710%	0.743%	+0.033
COTIC	col_4	1.897%	1.977%	+0.080
COTIC	col_5	1.277%	1.255%	−0.021
THP	col_2	1.535%	1.613%	+0.079
THP	col_3	1.044%	0.967%	−0.076
THP	col_4	1.967%	2.047%	+0.080
THP	col_5	1.499%	1.481%	−0.018
NEP	col_2	1.602%	1.664%	+0.062
NEP	col_3	1.436%	1.409%	−0.027
NEP	col_4	1.789%	1.854%	+0.064
NEP	col_5	2.243%	2.238%	−0.005
MLM	col_2	1.477%	1.530%	+0.053
MLM	col_3	1.121%	1.414%	+0.293
MLM	col_4	1.692%	1.751%	+0.059
MLM	col_5	2.261%	2.238%	−0.023

Table 11: MBD ROC-AUC, same encoders/targets as Table 10.

Encoder	Target	Raw ROC-AUC	Daily ROC-AUC	ΔROC
CoLES	col_2	0.6339	0.6340	+0.0001
CoLES	col_3	0.6874	0.6882	+0.0008
CoLES	col_4	0.6399	0.6390	−0.0009
CoLES	col_5	0.7567	0.7533	−0.0034
COTIC	col_2	0.7437	0.7466	+0.0029
COTIC	col_3	0.8064	0.8090	+0.0026
COTIC	col_4	0.7944	0.7982	+0.0038
COTIC	col_5	0.7835	0.7856	+0.0021
THP	col_2	0.7549	0.7575	+0.0027
THP	col_3	0.8236	0.8257	+0.0022
THP	col_4	0.8035	0.8061	+0.0026
THP	col_5	0.7942	0.7950	+0.0008
NEP	col_2	0.7608	0.7636	+0.0028
NEP	col_3	0.8366	0.8417	+0.0051
NEP	col_4	0.8005	0.8039	+0.0034
NEP	col_5	0.8473	0.8474	+0.0001
MLM	col_2	0.7555	0.7575	+0.0020
MLM	col_3	0.8211	0.8384	+0.0174
MLM	col_4	0.7969	0.7994	+0.0025
MLM	col_5	0.8445	0.8442	−0.0003

Table 12: XBANK (condition C), old vs. recovered FGW mapping: PR-AUC and ROC-AUC. Single test period (2024-01/02), paired within encoder on the same cohort.

Encoder	Target	Old PR-AUC	New PR-AUC	ΔPR (pp)	Old ROC	New ROC	ΔROC
CoLES	col_3	2.777%	2.664%	−0.112	0.6621	0.6497	−0.0123
CoLES	col_4	14.564%	15.811%	+1.247	0.6342	0.6661	+0.0319
CoLES	col_5	2.085%	2.130%	+0.045	0.6384	0.6368	−0.0016
COTIC	col_3	2.780%	2.846%	+0.067	0.6527	0.6717	+0.0191
COTIC	col_4	14.441%	12.503%	−1.938	0.6496	0.7091	+0.0596
COTIC	col_5	2.241%	2.231%	−0.011	0.6410	0.6431	+0.0022
THP	col_3	2.793%	3.330%	+0.537	0.6552	0.6953	+0.0400
THP	col_4	13.760%	13.306%	−0.454	0.6719	0.7310	+0.0591
THP	col_5	2.209%	2.420%	+0.211	0.6398	0.6539	+0.0141
NEP	col_3	2.655%	2.806%	+0.150	0.6491	0.6632	+0.0140
NEP	col_4	15.123%	14.845%	−0.278	0.6485	0.6858	+0.0373
NEP	col_5	2.026%	2.060%	+0.034	0.6314	0.6389	+0.0075
MLM	col_3	3.050%	3.071%	+0.021	0.6676	0.6801	+0.0125
MLM	col_4	15.332%	15.601%	+0.269	0.6454	0.6994	+0.0539
MLM	col_5	2.114%	2.265%	+0.152	0.6428	0.6558	+0.0130

Table 13: XBANK (condition C), old vs. recovered FGW mapping: P@5% and R@5%. Same encoders/targets as Table 12.

Encoder	Target	Old P@5%	New P@5%	Old R@5%	New R@5%
CoLES	col_3	3.50%	3.23%	10.99%	10.13%
CoLES	col_4	22.44%	22.76%	15.40%	15.62%
CoLES	col_5	2.82%	2.84%	11.38%	11.48%
COTIC	col_3	3.30%	3.45%	10.27%	10.74%
COTIC	col_4	21.70%	10.87%	15.18%	7.60%
COTIC	col_5	2.97%	3.02%	12.00%	12.21%
THP	col_3	3.26%	4.41%	10.17%	13.75%
THP	col_4	18.13%	11.62%	12.62%	8.08%
THP	col_5	3.06%	3.43%	12.34%	13.84%
NEP	col_3	3.03%	3.54%	9.52%	11.10%
NEP	col_4	22.64%	20.57%	15.53%	14.12%
NEP	col_5	2.63%	2.57%	10.62%	10.37%
MLM	col_3	3.65%	3.72%	11.46%	11.67%
MLM	col_4	23.29%	21.17%	15.98%	14.52%
MLM	col_5	2.84%	3.06%	11.48%	12.36%