ADAPTIVE-DEPTH LAYER GRAFTING: REPRESENTATION-LEVEL TRANSFER ACROSS HETEROGENEOUS TABULAR FEATURE SPACES

Anonymous authors
Paper under double-blind review

ABSTRACT

Transfer learning is especially challenging when a label-scarce target comes from a disparate domain and shares the same prediction endpoint as a source model but uses a different feature space, with neither paired samples nor known feature correspondence. We ask whether knowledge encoded in a pretrained source hierarchy can still be leveraged under this schema mismatch. To that effect, we introduce Adaptive-Depth Layer Grafting (ADLG), which maps target inputs through a lightweight bridge into a selected internal source representation while keeping the retained source tail fixed. ADLG-Frozen trains only the bridge and the interface selection is handled separately, with quantile-aligned CKA used as one optional pre-fit screening method. On Census-Income→Adult, ADLG improves balanced accuracy over a model trained from scratch by 1.52, 2.56, and 2.89 percentage points with 0.5%, 2%, and 10% of the target labels, respectively, and by 2.36 points across the prespecified 0.5–20% label regime. It also gains 1.88 points over a tuned target-only MLP and 3.41 points in minority recall over the model trained from scratch. Additional studies span heterogeneous tabular settings, including structured multimodal and inter-modal representations, as well as multiclass and regression endpoints. Controlled audits show that attachment depth is consequential and that source-label information is most discernible at shallow interfaces. Overall, ADLG provides a compact mechanism for leveraging knowledge in a pretrained source hierarchy when schemas differ and target labels are scarce.

1 INTRODUCTION

Many tabular transfer settings involve a data-rich source and a label-scarce target that describe the same prediction task using disparate features, dimensions, or semantics. The domains may also lack paired observations or reliable feature correspondence, making direct input-level transfer difficult. Existing heterogeneous-transfer methods commonly remap features, construct shared spaces, augment inputs, or align distributions (Duan et al., 2012; Wang & Mahadevan, 2011; Feuz & Cook, 2015; Li et al., 2019; Wang & Sun, 2022; Acharya & Das, 2026).

We instead treat hidden layers of a pretrained source network as *candidate transfer interfaces*. A lightweight bridge maps target features into one of these internal representations, after which prediction proceeds through the retained source tail. This yields our canonical estimator, **ADLG-Frozen**. At attachment depth k, the target bypasses source layers $L_1, \ldots, L_k$, while an affine–ReLU bridge maps the target input to the representation consumed by L_{k+1}. Only the bridge is trained, while the retained source layers and output head remain fixed. ADLG therefore requires neither paired source–target samples nor an explicit feature correspondence.

Attachment depth is selected separately from the graft itself. Fixed depth and inner validation are valid policies, while quantile-aligned Centered Kernel Alignment (CKA) provides one supervised pre-fit compatibility screen (Kornblith et al., 2019). Source and target examples are independently ordered by task score and compared on a common quantile-rank coordinate, without imposing sample-level correspondence. Attachment depth determines how much of the pretrained hierarchy is retained and how demanding the target-side mapping becomes. A shallow attachment preserves more source structure but requires the bridge to match an earlier representation, whereas a deeper attachment shortens the retained hierarchy but changes the approximation burden. We formalize this approximation–estimation tradeoff and analyze CKA as a measure of interface compatibility, not as an estimator of target risk. The proposed method is depicted schematically in Figure 1.

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.

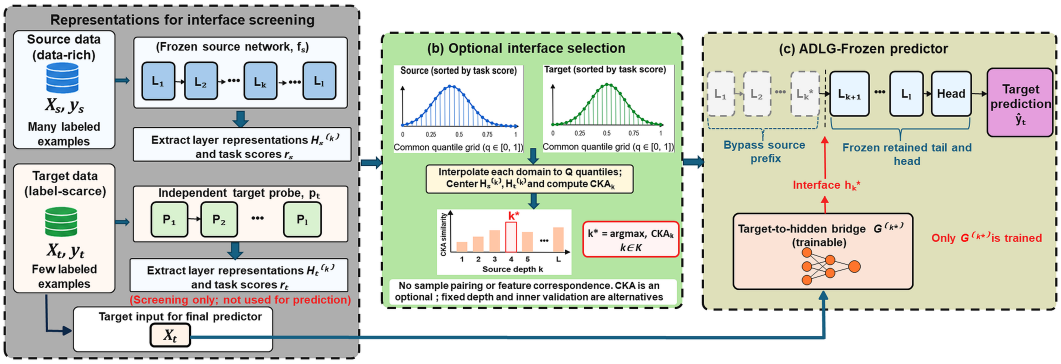

Figure 1: **ADLG-Frozen.** (a) Source and target representations are extracted for interface screening. (b) Quantile-aligned CKA is one optional policy for selecting $k^\star$; no sample pairing is assumed. (c) A trainable target-to-hidden bridge maps the raw target input to the selected interface, while the retained source tail and head remain frozen.

On Census-Income→Adult, ADLG-Frozen improves balanced accuracy by 2.36 points over an architecture-matched model trained from scratch and by 1.88 points over a tuned target-only MLP across the prespecified 0.5–20% label regime. The low-label gain over the trained-from-scratch model is accompanied by a 3.41-point increase in minority recall. Across five heterogeneous transfer pairs, the matched ADLG–MLP effect varies by pair, with supported gains on some pairs and a target-only advantage on WDBC→Mammographic Masses. Additional experiments extend the analysis to multimodal, inter-modal, multiclass, and regression settings. These results delineate where source-hierarchy transfer is beneficial and where direct target learning remains stronger.

ADLG-Frozen is a controlled method for testing whether knowledge in a pretrained source hierarchy can be transferred through a compact target-to-hidden bridge when schemas are incompatible and target labels are scarce. Its benefit is pair-, modality-, and label-budget-dependent; it is not universal.

Contributions. First, we introduce a target-to-hidden graft that maps an incompatible target schema into a pretrained source hierarchy without paired samples or feature correspondence. Second, we analyze how attachment depth balances retained-tail mismatch, bridge approximation, and target-side estimation, and relate CKA to interface compatibility. Third, we characterize where frozen source-hierarchy transfer helps and where well-tuned target-native learning remains preferable, using Census as a motivating low-label case and extending the analysis across heterogeneous tabular, multimodal, inter-modal, multiclass, and regression settings.

2 Related Work

Heterogeneous-transfer methods reconcile mismatched feature spaces through projection, shared latent spaces, feature augmentation, or distribution alignment (Duan et al., 2012; Wang & Mahadevan, 2011; Feuz & Cook, 2015; Li et al., 2019; Wang & Sun, 2022; Acharya & Das, 2026). ORCA and optimal-transport approaches similarly align input-side or relational structure (Shen et al., 2023; Courty et al., 2017; Mémoli, 2011; Vayer et al., 2020). FLAT meta-trains across task collections, whereas PU-HUDA assumes positive-only source labels and an unlabeled target (Zhu et al., 2023b; Mori et al., 2023). ADLG instead transfers knowledge from a single pretrained source hierarchy using labeled target data and no explicit input correspondence.

Layerwise transfer, adapters, and model stitching connect internal representations through learned interfaces (Yosinski et al., 2014; Houlsby et al., 2019; Lenc & Vedaldi, 2015; Bansal et al., 2021; Pan et al., 2023). Stitching typically assumes a common task and compatible activation geometry, often using linear maps between matched levels. ADLG addresses incompatible input schemas by learning a target-to-hidden bridge and selecting among candidate source interfaces from unpaired rank-aligned representations.

Semantic and cross-table methods such as TransTab, CARTE, TAROT, and CM2/OpenTabs exploit feature semantics, metadata, or broader pretraining (Wang & Sun, 2022; Kim et al., 2024; Shi et al., 2026; Ye et al., 2024). LATTLE transfers attention structure, while TabPFN and HyperFast are target-native pretrained predictors (Kowsar et al., 2025; Hollmann et al., 2025; Bonet et al., 2024). These methods operate under different supervision or pretraining assumptions and are not directly interchangeable with transferring internal representations from an already-trained source predictor.

Table 1: Transfer-family scope and the assumptions that distinguish ADLG.

Method family	Heterogeneous inputs	Paired cross-domain examples	Explicit input correspondence	Target input → source-hidden bridge	Data-dependent attachment depth
Model stitching	No	Often	No	No	No
Adapters	Usually no	No	No	No	No
Explicit feature remapping	Yes	Often	Yes	No	No
Latent-subspace HDA (e.g., HFA)	Yes	No	No	No	No
Cross-table pretraining (CM2/OpenTabs)†	Yes	No	Schema mechanism	No	No
Layer transplant (LATTLE)	Yes	No	No	No	No
ADLG	**Yes**	**No**	**No**	**Yes**	**Yes**

†Complementary pretraining setting; not executed in the matched five-pair benchmark.

Representation-similarity measures motivate our optional pre-fit compatibility screen (Raghu et al., 2017; Morcos et al., 2018; Kornblith et al., 2019), although similarity does not in general determine target risk (Davari et al., 2023). We therefore treat CKA as a screening policy and evaluate its cost and regret separately from the grafted estimator. ADLG's central contribution is a trainable target-input-to-source-hidden interface that exposes internal source knowledge when schemas are unmatched, examples are unpaired, and feature correspondence is unavailable. Frozen transfer alone is not claimed as novel.

3 Problem Formulation

Let

$$\mathcal{D}_S = \{(x_i^S, y_i^S)\}_{i=1}^{n_S}, \qquad x_i^S \in \mathbb{R}^{d_S}, \qquad y_i^S \in \mathcal{Y}, \qquad f_S = L_n \circ \cdots \circ L_1 \tag{1}$$

denote the labeled source data and its pretrained predictor. After an outer target split, let

$$\begin{aligned} \mathcal{P}_T &= \{x_i^T : i \in \mathcal{I}_T^{\text{train}}\}, \qquad x_i^T \in \mathbb{R}^{d_T}, \\ \mathcal{S}_T^{(f)} &= \{(x_i^T, y_i^T) : i \in \mathcal{I}_T^{(f)}\}, \quad y_i^T \in \mathcal{Y}, \quad m_T = |\mathcal{I}_T^{(f)}|, \end{aligned} \tag{2}$$

where $\mathcal{I}_T^{(f)} \subseteq \mathcal{I}_T^{\text{train}}$. The held-out target test set is disjoint from both sets.

Source and target may differ in dimensionality, feature identity, semantics, scale, and ordering, so $\mathcal{X}_S \neq \mathcal{X}_T$. Because ADLG retains the source output head, both domains must share the same prediction endpoint and label space, $\mathcal{Y}_S = \mathcal{Y}_T = \mathcal{Y}$, with $\mathcal{Y}$ binary, multiclass, or real-valued in our experiments.

ADLG requires neither paired source–target observations nor a known feature map. It selects an internal source interface and learns a compact mapping from the target input to that representation. Thus, it addresses schema mismatch while assuming compatible prediction endpoints.

4 Adaptive-Depth Layer Grafting

ADLG treats internal source layers as candidate transfer interfaces. At a chosen depth, a lightweight bridge maps the target input into the source hierarchy, while the retained source tail remains frozen. Interface selection is separate from the graft and may be fixed, random, validation-based, or CKA-based.

4.1 Grafted Predictor

For candidate depth $k \in \mathcal{K}$, let

$$F_k = L_n \circ \cdots \circ L_{k+1}, \qquad G^{(k)} : \mathbb{R}^{d_T} \to \mathbb{R}^{d_k}, \qquad f_T^{(k)} = F_k \circ G^{(k)}, \tag{3}$$

where d_k is the input width of L_{k+1}. In the main experiments,

$$G^{(k)}(x) = \text{ReLU}(W_k x + b_k). \tag{4}$$

Thus, the target bypasses $L_1, \ldots, L_k$ and enters the frozen tail through the bridge. We use the same low-capacity bridge family at every depth so that comparisons reflect the attachment interface. Bridge-capacity controls and the deepest-interface equivalence are given in Appendices G.8 and B.

4.2 Interface Selection

For representations $X \in \mathbb{R}^{m \times p}$ and $Y \in \mathbb{R}^{m \times q}$, linear CKA is

$$\text{CKA}(X, Y) = \frac{\|X_c^\top Y_c\|_F^2}{\|X_c^\top X_c\|_F \, \|Y_c^\top Y_c\|_F}, \qquad X_c = H_m X, \quad Y_c = H_m Y, \tag{5}$$

where $H_m = I_m - m^{-1}\mathbf{1}_m\mathbf{1}_m^\top$.

Because source and target rows are unpaired, our optional CKA policy first trains a split-local target probe, independently orders source and target examples by task score, and interpolates their hidden activations onto a common $Q = 128$ quantile grid. It then selects

$$\widehat{C}^{(k)} = \mathrm{CKA}\Big(\widetilde{H}_S^{(k)}, \widetilde{H}_T^{(k)}\Big), \qquad k^\star = \arg\max_{k\in\mathcal{K}} \widehat{C}^{(k)}. \tag{6}$$

Binary tasks use positive-class probability, multiclass tasks use class-conditional probabilities averaged across classes, and regression uses the predicted outcome; endpoint-specific details appear in Appendix A.1. The quantile grid is a within-domain rank coordinate and does not create sample correspondence. CKA is therefore a supervised compatibility screen, not an estimate of target risk. Fixed depth, random choice, and inner validation use the same grafted predictor. Further selector details, tie-breaking, cross-fitting, and labeled-only controls are in Appendices D and G.6.

4.3 Design Rationale

ADLG separates three questions: whether a pretrained hierarchy contains transferable structure, whether a target input can be mapped into that hierarchy, and how much of the pretrained source hierarchy should be retained. Freezing the source tail isolates representation transfer from additional target-side fine-tuning, while the lightweight bridge keeps the interface comparison focused on compatibility while controlling capacity.

4.4 Training and Evaluation

ADLG-Frozen is the canonical estimator: only the Xavier-initialized bridge is optimized, while the retained source tail remains fixed. Primary experiments use the prespecified candidate set $\mathcal{K} = \{1, 2, 3\}$, with CKA as the reported selection policy. Appendix D records the executed configuration for each study.

All compared methods use the same outer target splits and labeled indices. Preprocessing, probe fitting, depth selection, initialization, training, and model selection use only outer-training data; the held-out target test set is used once for evaluation. Full preprocessing provenance, schedules, freezing details, and selector settings appear in Appendices C.1 and D.

5 Theoretical Analysis

Attachment depth affects both the amount of source structure retained and the mapping the target bridge must learn. Shallow grafts preserve more of the source hierarchy but may require a harder interface match; deeper grafts retain less of the hierarchy and shift more of the learning to the target data. The analysis below formalizes this tradeoff for ADLG-Frozen. It does not prescribe a universally optimal depth, and CKA is treated only as a compatibility measure.

Let

$$R_T(f) = \mathbb{E}_{(x,y)\sim T}[\ell(f(x), y)], \qquad R_T^\star = \inf_f R_T(f).$$

For depth k, let $g_k^\dagger$ minimize target risk for the retained tail F_k, and define

$$\delta_k = R_T(F_k \circ g_k^\dagger) - R_T^\star, \qquad \epsilon_k^\star = \inf_{G\in\mathcal{H}_G^{(k)}} \mathbb{E}_{x\sim T}[\rho_k(G(x), g_k^\dagger(x))]. \tag{7}$$

Here, δ_k measures the limitation of the retained source tail and $\epsilon_k^\star$ the bridge approximation error. Full assumptions and proofs appear in Appendix B.

5.1 Fixed-Tail Risk

Theorem 1 (Fixed-depth target excess risk)**.** *Assume bounded inputs and loss, a norm-bounded bridge class, Lipschitz loss, and a Lipschitz retained tail. Then, with probability at least* $1 - \delta$*, an approximate empirical-risk minimizer* $\widehat{G}_k$ *satisfies*

$$R_T(F_k \circ \widehat{G}_k) - R_T^\star \le \delta_k + L_{\mathrm{loss}}\,\mathrm{Lip}(F_k)\epsilon_k^\star + \Gamma_k(m_T, \delta) + \epsilon_{\mathrm{opt},k}. \tag{8}$$

The terms correspond to retained-tail mismatch, bridge approximation, finite-sample estimation, and optimization. For the affine–ReLU bridge with $P_k = d_k(d_T + 1)$ trainable parameters,

$$\Gamma_k(m_T, \delta) = \widetilde{\mathcal{O}}\left(\sqrt{\frac{P_k + \log(1/\delta)}{m_T}}\right).$$

Thus, depth can alter both source compatibility and the amount of target-side estimation required. When bridge width is fixed across depths, the tradeoff is driven mainly by δ_k, $\epsilon_k^\star$, and $\mathrm{Lip}(F_k)$, not parameter count. Detailed complexity and deepest-interface results appear in Appendices B.2 and B.11.

Operationally, these terms correspond to quantities varied in the controlled audits: source-tail condition, bridge capacity, labeled-target sample size, and optimization repeat. The bound organizes those dependencies; it does not make CKA or any other policy a guaranteed risk-optimal selector. The regression study is an empirical extension; the bound applies only under its stated bounded-loss and Lipschitz assumptions.

5.2 Data-Dependent Depth Selection

Because $\mathcal{K}$ is finite, applying the fixed-depth result at confidence $\delta/|\mathcal{K}|$ gives a simultaneous bound over all candidate depths. The guarantee therefore remains valid when depth is selected using permitted training data, including the CKA policy. This controls finite-set selection; it does not imply that CKA minimizes target risk. The formal statement is given in Appendix B.3.

5.3 CKA and Interface Compatibility

For the quantile-indexed source and target representations, linear CKA measures similarity of their centered linear geometry. Under the spectral conditions derived in Appendix B, higher CKA can correspond to smaller linear reconstruction error. In general, however, rank and spectrum vary across depths, so CKA need not order interface compatibility or downstream target risk. We therefore use it only as a pre-fit screen and compare it with fixed, random, and inner-validation policies; oracle regret is reported only as a diagnostic.

6 Experimental Setup

We evaluate ADLG under scarce target labels, across different transfer settings and endpoint types, and through controlled mechanism audits. All primary experiments use ADLG-Frozen with the prespecified candidate set $\mathcal{K} = \{1, 2, 3\}$. Complete dataset construction, architectures, training schedules, preprocessing, and executed configurations are given in Appendices C and D. The accompanying supplementary ZIP contains anonymized code, configurations, frozen outputs, and the available split identifiers. Appendix Table 4 provides a claim-to-evidence ledger; the supplementary reader map links every principal result summarized here to its complete tables, sensitivity analyses, provenance checks, and limitations.

6.1 Benchmarks

Census-Income→Adult is the primary low-label study. We use five paired stratified 80/20 target splits, three optimization repeats per split, and nested label budgets $\{0.5, 1, 2, 5, 10, 20, 40, 60, 80, 100\}\%$; 0.5–20% is the prespecified low-label regime. Optimization repeats are averaged within each outer split, so inference uses five paired split-level differences. Comparisons include an architecture-matched model trained from scratch, a tuned target-only MLP, TransTab-Transfer, and CATTLE-Supervised (Akhter et al., 2026).

The five-pair benchmark additionally includes HCV→ILPD, Student→OULAD, Telco→Iranian, and WDBC→Mammographic Masses. It uses ten matched outer splits and absolute target-label budgets $\{20, 50, 100, 200\}$. Target-native comparisons include TabPFN v2, HyperFast, CARTE, and LATTLE (Hollmann et al., 2025; Bonet et al., 2024; Kim et al., 2024; Kowsar et al., 2025). Telco→Iranian and Taiwan→German also provide matched comparisons with linear HFA and XTab-style source pretraining (Appendix E.8).

We further examine multimodal HAM10000→Derm7pt and METABRIC→TCGA-BRCA, bidirectional clinical–omics transfer, and clinically harmonized eICU→MIMIC. HAM10000→Derm7pt uses repeated training runs on one official target test split, while the METABRIC studies use disjoint source and target patients. Multiclass PAM50 prediction and Washington/DC→Seoul Bike regression extend the same frozen-tail graft to four-class and continuous outcomes. Complete endpoint and cross-modality protocols are in Appendices A.1, H.9, H.10, H.11, and H.12.

6.2 Information Isolation and Statistical Analysis

Matched methods use the same target splits and labeled subsets. Target-test data never enter preprocessing, probe fitting, CKA, bridge training, depth selection, or model and threshold selection. Target preprocessing is fitted within the corresponding outer-training partition and then applied unchanged

to held-out data. CKA may use permitted unlabeled target-training covariates, but never target-test observations. Split construction and integrity checks are reported in Appendices D.1 and K.1.
Balanced accuracy (BA) is primary for classification and RMSE for regression; minority recall and other secondary metrics are reported in the supplement. Matched effects use paired 95% intervals over the stated outer-split or run-level units. Optimization repeats are averaged before split-level inference, preventing training repeats from being treated as independent datasets. Inferential claims are pair-specific; cross-pair aggregates in the supplement are descriptive. For studies using one fixed test partition, such as HAM10000→Derm7pt, repeated runs quantify optimization and initialization variability, not uncertainty over new target populations.

6.3 Mechanism Audits

Census and WDBC→Mammographic Masses controls examine attachment depth, source condition, bridge capacity, and interface selection under matched target splits and budgets. Fixed-depth controls separate attachment effects from the CKA policy, while source-condition audits compare correctly pretrained, shuffled-label, and random-frozen source tails. The test-based oracle is used only as a retrospective diagnostic and never for model selection. Complete audits and executed settings are reported in Appendices I and D. Candidate depths use the same bridge family. Their depth-specific trainable parameter counts are reported with the audits, and a constant-width fixed-depth control separates attachment location from dimensionality variation.

7 Results

All main-text results use canonical ADLG-Frozen; exploratory fine-tuning variants are intentionally excluded from the main evidence chain. Full pair–budget tables and unfavorable outcomes are retained in the supplement.

7.1 Five-Pair Benchmark and Census Case Study

The five-pair benchmark tests the portability of frozen source-hierarchy transfer; it is not designed to assert universal target-accuracy superiority. Effects vary across transfer pairs: ADLG is favored on Census-Income→Adult and HCV→ILPD, inconclusive on Student→OULAD and Telco→Iranian, and lower than the matched target-only MLP on WDBC→Mammographic Masses (Table 2). We therefore treat Census as a motivating low-label positive case and WDBC as an explicit negative-transfer boundary. Appendix E.3 reports the descriptive equal-pair summary for completeness.
Within the motivating Census-Income→Adult case, across the prespecified 0.5–20% regime, ADLG-Frozen gains 2.36 points (95% CI: 2.02–2.72) over the trained-from-scratch model and 1.88 points (1.63–2.17) over the tuned target-only MLP. Figure 3(a) shows the corresponding balanced-accuracy learning curve, and Figure 2 expands the prespecified 0.5%, 2%, and 10% budgets to AUROC, PR-AUC, balanced accuracy, and minority recall. For ADLG-Frozen, the aggregate improvement is not only a recall shift: recall and specificity rise by 3.41 and 1.32 points versus the trained-from-scratch model. Complete learning curves, secondary metrics, uncertainty, and the target-native linear/tree context appear in Appendix H.1 and Section H.2. The main comparison isolates the question ADLG is designed to answer: whether knowledge in an available source hierarchy can be transferred through a compact learned interface under scarce target labels. At 0.5%, 2%, and 10% labels, ADLG-Frozen improves over the architecture-matched model trained from scratch by 1.52, 2.56, and 2.89 balanced-accuracy points, respectively; Appendix Table 49 gives the complete budget-specific results.

7.2 Matched Target-Native and Pretrained Baselines

The matched target-native benchmark uses identical target splits and labeled subsets and shows clear task-level heterogeneity with no uniform ordering. Relative to the matched MLP, ADLG is favored on two pairs, inconclusive on two, and lower on WDBC. Comparisons with pretrained and tree-based target-native methods likewise vary across pairs and budgets. On Census, logistic regression and the tuned tree are 0.98 and 1.39 points higher, respectively; ADLG's use case is source-knowledge transfer, not replacement of routine target-native baselines. The complete comparison—including HyperFast, linear/tree models, pair-level effects, uncertainty, and access assumptions—is reported without omission in Appendices E.2, E.3, and E.4. External prediction files have stable test IDs, but the available ADLG cache does not; we therefore do not form unmatched paired calibration or PR-AUC contrasts.

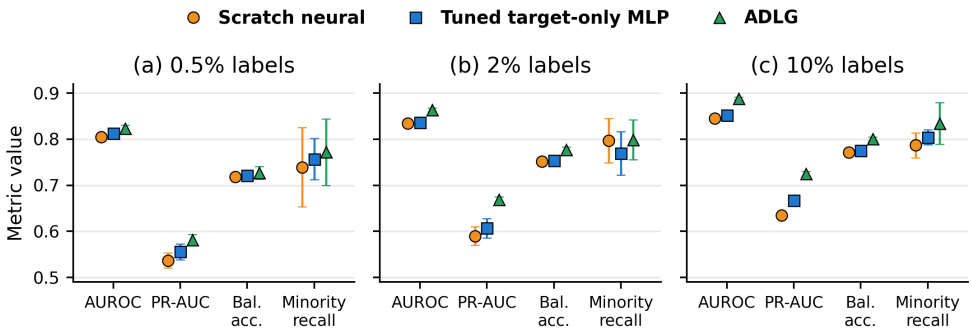

Figure 2: Census-Income→Adult at the prespecified 0.5%, 2%, and 10% target-label budgets. ADLG-Frozen, an architecture-matched model trained from scratch, and a tuned target-only MLP are compared on AUROC, class-1 PR-AUC, balanced accuracy, and minority recall. Points average repeats within each outer split; error bars show ±1 SEM over five splits. Complete results appear in Appendix H.1.

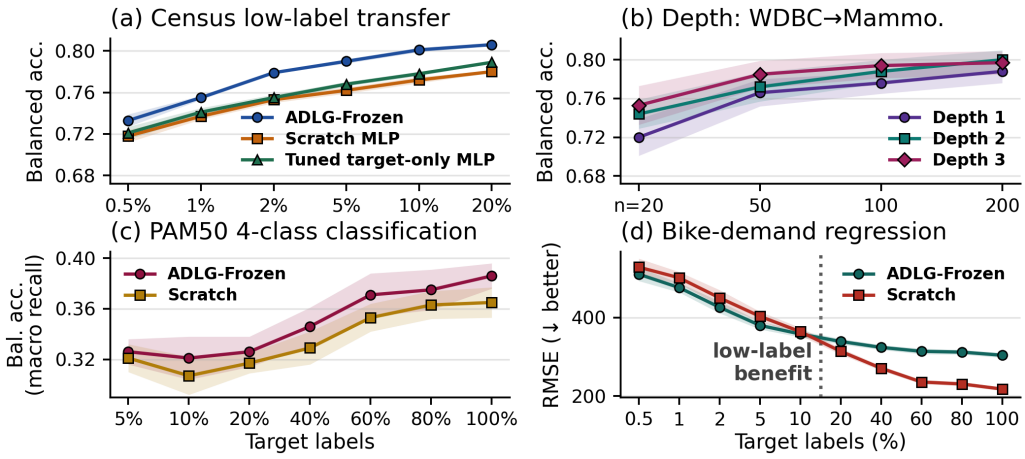

Figure 3: ADLG-Frozen results: (a) Census-Income→Adult BA; (b) WDBC→Mammographic fixed-depth BA; (c) PAM50 multiclass BA; (d) bike-demand RMSE (lower is better). Bands show ±1 SEM over split means (five splits in (a,c,d), ten in (b)). Full results: Appendices H.1 (a), I.2 (b), and A.2–A.3 (c,d).

WDBC→Mammographic is an explicit negative-transfer boundary relative to the matched target-only MLP, which exceeds ADLG by 2.68 BA points. One plausible interpretation is that the five-feature target can be modeled efficiently in its native representation, whereas grafting constrains prediction through a frozen hierarchy learned from the higher-dimensional WDBC source. Consistent with this view, correctly pretrained tails do not uniformly outperform random-frozen tails in the source-condition audit. These controls support a transfer-bias interpretation but do not identify a unique cause. A corrected low-budget audit still places ADLG above the model trained from scratch; the boundary is specifically relative to the higher-performing matched target-only MLP.

7.3 External Heterogeneous-Transfer Comparisons

Matched Telco and Taiwan studies compare ADLG with linear HFA and an XTab-style source-pretraining configuration. In the 5–20% regime, ADLG improves BA by 6.60 and 1.84 points over HFA and by 0.83 and 3.50 points over XTab-style, respectively; all but the Telco XTab interval exclude zero. Executed TransTab and CATTLE comparisons provide additional implementation-scoped context. Protocols, fidelity limits, paired intervals, and four-metric curves appear in Appendices E.6–E.8. Supplementary Figure 16 reports additional minority-recall operating-point diagnostics for Telco→Iranian, Taiwan→German Credit, and eICU→MIMIC.

Table 2: Five-pair ADLG-Frozen versus architecture-matched target-only MLP. Entries are mean ADLG−MLP balanced-accuracy points over prespecified budgets and matched outer splits; 95% CIs cluster-bootstrap outer splits. The equal-pair summary is in Appendix E.3.

Transfer pair	ADLG − MLP	95% CI	Interpretation
Census-Income→Adult	$+0.65$	$[+0.34, +0.97]$	ADLG favored
HCV→ILPD	$+1.52$	$[+0.07, +3.06]$	ADLG favored
Student→OULAD	-0.14	$[-0.46, +0.15]$	inconclusive
Telco→Iranian	$+0.16$	$[-0.80, +1.25]$	inconclusive
WDBC→Mammographic	-2.68	$[-4.20, -1.15]$	target MLP favored

7.4 Multimodal and Inter-modal Transfer

ADLG also operates across multimodal and inter-modal settings, with gains that depend on direction and label budget. On HAM10000→Derm7pt, ADLG improves BA over the target multimodal model by 2.99 points across the prespecified 5–20% regime (95% CI: 1.55–4.17); the ten paired runs share one official test split and characterize repeated optimization on that fixed partition. On METABRIC→TCGA-BRCA, the multimodal BA gain is supported at 10% labels (+3.3 points [1.1,5.5]), while other BA budgets are inconclusive. Table 3 records the corresponding endpoint evidence with its statistical unit and scope. The multimodal results are not pooled with the five-pair benchmark: HAM10000→Derm7pt uses repeated runs on one fixed official test partition, whereas the METABRIC studies use paired target-split summaries. Clinical→omics is favorable at 10–40% labels but reverses at full supervision (-3.14 BA points, 95% CI $[-5.41, -0.04]$). In the reverse direction, omics→clinical improves BA by 6.33, 6.06, and 6.71 points at 40%, 80%, and 100% labels, respectively (Appendices H.11–H.12). These directional boundaries show that the value of a frozen source hierarchy depends on modality direction and target supervision.

7.5 Multiclass and Regression Extensions

Figures 3(c)–(d) extend the frozen-tail graft to multiclass classification and regression. Both prespecified low-label summaries favor the frozen-tail graft: PAM50 balanced accuracy improves by 0.95 points (95% CI: 0.44–1.42), and bike-demand RMSE decreases by 11.50 (95% CI: 7.08–15.92). All five PAM50 split aggregates are positive; ADLG lowers RMSE at 0.5–10%, while the trained-from-scratch model is better from 20% onward. Complete secondary metrics and all-budget boundaries are in Sections A.2–A.3. The regression study is an empirical endpoint extension; the theoretical bound is invoked only under its stated bounded-loss and Lipschitz assumptions. For both endpoint extensions, the applicable CKA rule selected depth 3 in every evaluated cell. These studies therefore support the frozen-tail graft across new endpoint types, but not effective adaptive depth selection on these two pairs (Appendices A.2–A.3).

7.6 Attachment Depth and Source-Tail Controls

Attachment depth materially affects prediction: Figure 3(b) shows the finalized fixed-depth WDBC→Mammographic Masses audit. In a separate constant-width six-depth audit, the best mean depth varies across pairs, so bridge parameter count alone does not explain the observed depth differences; this analysis is descriptive (Appendix G.15). WDBC is also the pair on which the matched target-only MLP outperforms ADLG. The panel therefore shows that attachment location can matter materially even when transfer is not beneficial relative to direct target learning.

Across 150 locked Census cells, CKA and validation give 0.79309 and 0.79313 BA (difference -0.004 points, CI $[-0.125, +0.185]$), but CKA uses 1.041 versus 2.628 s median time and two versus four neural fits. Cheaper one-fit fixed and random policies are about 0.05 points lower with intervals including zero. This is an accuracy–cost menu, not CKA superiority (Appendix I.2).

At Census depth 1, correct pretraining exceeds shuffled-label pretraining by 0.98 BA points, but this advantage contracts at deeper attachments. On WDBC, pretrained tails are 0.83 points above shuffled-label tails yet 1.41 points below random-frozen tails (Appendix G.3). Thus, source-label information is detectable, especially at shallow interfaces, but architectural inductive bias also contributes materially.

The affine+ReLU bridge is capacity-controlled, not claimed universally optimal: tested richer bridges and tail tuning show no consistent advantage. Shallow width and retained depth remain partly coupled outside the constant-width audit; full bridge, ordering, probe, and grid controls are in Appendices G.8–G.12.

Table 3: Summary of the main empirical evidence. Positive values indicate an ADLG advantage. BA denotes balanced-accuracy points; RMSE reduction denotes trained-from-scratch RMSE minus ADLG RMSE. Brackets denote paired 95% CIs.

Study / question	Comparator	Main result	Scope
Census-Income→Adult	Trained-from-scratch model	$+2.36$ **BA**	0.5–20% labels
Census-Income→Adult	Tuned target-only MLP	$+1.88$ **BA**	0.5–20% labels
Telco / Taiwan	HFA	$+6.60/+1.84$ **BA**	5–20% labels
HAM10000→Derm7pt	Target multimodal model	$+2.99$ **BA** $[1.55, 4.17]$	5–20%; 10 paired runs on one official test split
METABRIC→TCGA-BRCA	Target multimodal model	$+3.30$ **BA** $[1.10, 5.50]$	10% labels only
Clinical→omics	Target-only omics MLP	$+1.19/+2.49/+2.86$ **BA**	10%/20%/40% labels
Omics→clinical	Target-only clinical MLP	$+6.33/+6.06/+6.71$ **BA**	40%/80%/100% labels
PAM50 multiclass	Trained-from-scratch model	$+0.95$ **BA** $[0.44, 1.42]$	5–20%; five outer splits
Bike-demand regression	Trained-from-scratch model	**11.50 RMSE reduction** $[7.08, 15.92]$	0.5–20% aggregate; gains at 0.5–10%; five outer splits

Selector efficiency: CKA and inner validation achieve nearly identical BA, while CKA reduces median selection time from 2.628 s to 1.041 s and uses two neural fits in the Census audit (Appendix I.2).

8 Discussion and Limitations

ADLG-Frozen studies when knowledge in a pretrained source hierarchy can be transferred through a compact target-to-hidden bridge under incompatible schemas and scarce target labels. The benefit is pair-, modality-, and label-budget-dependent. Census and HCV favor ADLG in the matched MLP comparison, Student and Telco are inconclusive, and WDBC favors the target-only MLP. WDBC illustrates the other side of the tradeoff: when direct target learning is already effective, transfer bias from a frozen source tail can outweigh its estimation benefit. Source-condition controls support this interpretation without identifying a unique cause. The theory exposes how retained-tail bias, bridge mismatch, and estimation interact with depth; it does not select an optimal interface, and the controlled affine–ReLU bridge does not rule out richer adapters. Optional CKA screening is supervised, can inherit probe and low-sample bias, and is not a risk predictor; the perturbation and debiased-CKA/RV2 audits test specific sensitivities. Finally, the benchmark concerns frozen feed-forward MLP tails: endpoint and modality breadth does not establish transfer to Transformer, tree, or foundation-model internals. These boundaries motivate richer interfaces, broader backbones, and matched cross-table comparisons.

Practical scope. ADLG is useful when a task-aligned source MLP already exists, target schemas are incompatible, and target labels are scarce. It should be tested against well-tuned target-native baselines on identical labeled subsets; fixed depth or inner validation remain valid when CKA is unstable. Cross-institutional medical use also requires local validation, calibration, subgroup analysis, governance, and deployment-specific oversight.

9 Conclusion

ADLG turns internal layers of an existing source MLP into transfer interfaces for targets with incompatible schemas, without paired samples or a supplied feature correspondence. The evidence shows a clear low-label benefit on Census, mixed pair-specific behavior across five heterogeneous tabular pairs, direction- and budget-dependent multimodal/inter-modal transfer, and empirical extensions to multiclass and regression endpoints. Theory and controlled audits support the central lesson that attachment depth changes the balance between what must be learned from scarce target labels and what can be retained from the source. Interface selection remains modular; the main contribution is the graftable target-to-hidden interface and an auditable protocol for testing source-knowledge transfer.

AI USE STATEMENT

Generative AI tools were used as assistive tools for literature search, implementation and debugging, presentation, manuscript editing, and feedback on mathematical exposition and experimental design. The authors independently verified every assumption, derivation, citation, executed analysis, and reported value; they made all final methodological and scientific decisions. AI tools did not create or alter empirical observations. The authors take full responsibility for the work.

ETHICS STATEMENT

The clinical studies use previously released research datasets and do not constitute clinical validation or deployment guidance. Any deployment would require institution-specific validation, calibration, fairness assessment, and applicable oversight.

REPRODUCIBILITY STATEMENT

The supplement records dataset construction and preprocessing provenance, executed architectures and schedules, paired statistical procedures, fidelity limits, and artifact checks. Unavailable historical provenance is marked explicitly rather than reconstructed. The accompanying supplementary ZIP contains anonymized code, configurations, frozen outputs, and analysis scripts for recomputing the reported summaries. Workflows that require third-party or restricted data identify those inputs explicitly.

REFERENCES

Snigdhatanu Acharya and Mrinal Das. Map wisely for efficient transfer learning across heterogeneous data sources. *Neurocomputing*, 660:131762, 2026. doi: 10.1016/j.neucom.2025.131762.

Kazi F. Akhter, Ibna Kowsar, and Manar D. Samad. Generalized context in cross attention for transfer learning of disjoint tabular data. *Neural Networks*, 205:109568, 2026. doi: 10.1016/j.neunet.2026.109568.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

David Bonet, Daniel Mas Montserrat, Xavier Giró-i Nieto, and Alexander G. Ioannidis. HyperFast: Instant classification for tabular data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2024. doi: 10.1609/aaai.v38i10.28988.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2017. doi: 10.1109/TPAMI.2016.2615921.

MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. Reliability of CKA as a similarity measure in deep learning. In *International Conference on Learning Representations*, 2023.

Lixin Duan, Dong Xu, and Ivor W. Tsang. Learning with augmented features for heterogeneous domain adaptation. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

Kyle D. Feuz and Diane J. Cook. Transfer learning across feature-rich heterogeneous feature spaces via feature-space remapping (FSR). *ACM Transactions on Intelligent Systems and Technology*, 6 (1):3:1–3:27, 2015. doi: 10.1145/2629528.

Josh Gardner, Juan C. Perdomo, and Ludwig Schmidt. Large scale transfer learning for tabular data via language modeling. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025. doi: 10.1038/s41586-024-08328-6.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 2019.

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

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019.

Ibna Kowsar, Kazi F. Akhter, and Manar D. Samad. LLM attention transplant for transfer learning of tabular data across disparate domains. *arXiv preprint arXiv:2511.06161*, 2025.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

Jingjing Li, Ke Lu, Zi Huang, Lei Zhu, and Heng Tao Shen. Heterogeneous domain adaptation through progressive alignment. *IEEE Transactions on Neural Networks and Learning Systems*, 30 (5):1381–1391, 2019. doi: 10.1109/TNNLS.2018.2868854.

Facundo Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 11(4):417–487, 2011. doi: 10.1007/s10208-011-9093-5.

Ari S. Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems*, volume 31, pp. 5732–5741, 2018.

Junki Mori, Ryo Furukawa, Isamu Teranishi, and Jun Sakuma. Heterogeneous domain adaptation with positive and unlabeled data. *arXiv preprint arXiv:2304.07955*, 2023. doi: 10.48550/arXiv.2304.07955.

Zizheng Pan, Jianfei Cai, and Bohan Zhuang. Stitchable neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6076–6085, 2017.

Junhong Shen, Liam Li, Lucio M. Dery, Corey Staten, Mikhail Khodak, Graham Neubig, and Ameet Talwalkar. Cross-modal fine-tuning: Align then refine. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31030–31056. PMLR, 2023.

Ruxue Shi, Yili Wang, Mengnan Du, Hangting Ye, Yi Chang, and Xin Wang. TAROT: Task-adaptive refinement of LLM-prior graphs for few-shot tabular learning. *arXiv preprint arXiv:2606.11640*, 2026.

Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. Fused gromov–wasserstein distance for structured objects. *Algorithms*, 13(9):212, 2020. doi: 10.3390/a13090212.

Chang Wang and Sridhar Mahadevan. Heterogeneous domain adaptation using manifold alignment. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, pp. 1541–1546, 2011. doi: 10.5591/978-1-57735-516-8/IJCAI11-259.

Zifeng Wang and Jimeng Sun. TransTab: Learning transferable tabular transformers across tables. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Chao Ye, Guoshan Lu, Haobo Wang, Liyao Li, Sai Wu, Gang Chen, and Junbo Zhao. Towards cross-table masked pretraining for web data mining. *arXiv preprint arXiv:2307.04308*, 2024. doi: 10.48550/arXiv.2307.04308.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, volume 27, pp. 3320–3328, 2014.

Bingzhao Zhu, Xingjian Shi, Nick Erickson, Mu Li, George Karypis, and Mahsa Shoaran. XTab: Cross-table pretraining for tabular transformers. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 43181–43204. PMLR, 2023a.

Max Zhu, Katarzyna Kobalczyk, Andrija Petrovic, Mladen Nikolic, Mihaela van der Schaar, Boris Delibasic, and Pietro Lio. Tabular few-shot generalization across heterogeneous feature spaces. *arXiv preprint arXiv:2311.10051*, 2023b. doi: 10.48550/arXiv.2311.10051.

APPENDIX

Overview. These appendices provide proofs, protocols, complete results, controls, and reproducibility checks. We use *n/r* only when the released artifacts do not establish a historical quantity; such values are neither estimated nor reconstructed.

Reader map. The supplement proceeds from endpoint extensions and theory to data provenance, the canonical protocol, and baseline fidelity. It then separates developmental variants from estimator ablations, reports complete pair–budget results, audits selector accuracy and cost, specifies the statistical methodology, and ends with reproducibility and artifact coverage. Existing labels are retained to keep main-text cross-references stable. In every table, – means not run, and boldface denotes only the larger matched mean, not statistical significance.

Canonical protocol. ADLG-Frozen is a controlled method for testing whether knowledge in a pretrained source hierarchy can be transferred through a compact target-to-hidden bridge when schemas are incompatible and target labels are scarce. Its benefit is pair-, modality-, and label-budget-dependent. It is the only canonical method in this paper and uses a neutral Xavier-initialized affine–ReLU bridge, a frozen source tail, target-side fitting, and no test-based selection. The canonical candidate set is $\mathcal{K} = \{1, 2, 3\}$. CKA is a supervised pre-fit screening heuristic because its target representations come from a split-local probe trained on the permitted labeled target subset; fixed-depth and inner-validation policies use the same estimator. Initializer variants and tail adaptation are noncanonical controls or exploratory extensions.

Table 4: Claim-to-evidence ledger for canonical ADLG-Frozen.

Claim	Supporting artifact	Scope	Limitation
Frozen source-hierarchy transfer can be tested across unmatched schemas	Five-pair ADLG-Frozen run tables and saved split manifests	Five heterogeneous binary pairs	Pair-specific effects are heterogeneous; WDBC favors the matched target-only MLP
Attachment depth can materially affect accuracy	Fixed-depth WDBC audit and constant-width six-depth benchmark	Executed backbones and budgets	Does not identify one generally optimal depth
Census is a favorable low-label case	Locked 150-cell Census rerun; paired outer-split contrasts	Census→Adult, 0.5–20% labels	One pair; linear and tree baselines remain strong
CKA can reduce selection cost	End-to-end Census selector timing and accuracy audit	Three candidate depths on Census	Not consistently more accurate than validation or fixed policies
Shallow interfaces can retain a source-label signature	Correct, shuffled-label, and random-frozen source controls	Fixed-depth Census/WDBC audits	Architectural inductive bias explains part of the effect
Endpoint scope extends beyond binary classification	PAM50 multiclass and bike-demand regression artifacts	One pair per endpoint extension	Does not establish broad endpoint generalization

Main-paper crosswalk. Section 7.1 is supported by Appendix H and Tables 2 and 50–52; Section 7.3 by Appendix E; Section 7.4 by Appendices H.9–H.12; Section 7.5 by Appendix A; Section 7.6 by Appendices G and I; and the reproducibility statement by Appendix K.

For eICU and MIMIC, "Raw n" refers to the supplied study-cohort extracts before outcome filtering, not to the complete underlying clinical databases.

A ADDITIONAL ENDPOINT STUDIES

This section gives the complete evidence behind Section 7.5. It appears first because these experiments extend the paper's empirical scope. Both use the same affine+ReLU bridge, three candidate interfaces $\{1, 2, 3\}$, 128 quantiles, and a frozen source tail as the binary experiments. Only the endpoint-specific output, loss, ordering rule, and metric change. Each comparison averages three optimization repeats within each of five target outer splits; paired inference treats those five split summaries as the units. Percentile-bootstrap intervals resample outer splits, and paired Student-t intervals use $df = 4$.

Table 5: Four-class PAM50 transfer. ΔBA is ADLG minus the trained-from-scratch model in points; intervals resample five outer-split summaries.

Labels	ADLG BA	Trained-from-scratch BA	ΔBA [95% CI]	ΔAUROC
5%	.327	.321	+0.56 [−0.04,1.07]	+1.33
10%	.321	.307	+1.37 [−0.33,3.16]	+1.73
20%	.326	.317	+0.91 [0.33,1.69]	+1.45
40%	.346	.329	+1.74 [−0.03,4.75]	+0.98
60%	.371	.353	+1.78 [0.19,2.87]	+1.11
80%	.375	.363	+1.22 [−0.98,3.13]	+1.54
100%	.386	.365	+2.10 [0.66,3.23]	+2.34

A.1 Datasets, Protocol, and Information Isolation

Four-class PAM50 transfer. METABRIC clinical data are the source ($n_S = 1{,}608$, 16 semantic and 40 encoded features) and TCGA-BRCA clinical data are the target ($n_T = 945$, 13 semantic and 95–101 split-dependent encoded features). The four shared labels are LumA, LumB, HER2-enriched, and Basal; source counts are 700/475/224/209 and target counts are 499/197/78/171, respectively. Patient identifiers and the PAM50 target are excluded from predictors. Source preprocessing is fit on the source training partition and target preprocessing separately within each outer target-training pool; no target-test patient enters preprocessing, selection, or fitting, and no source–target feature match is supplied. The seven nested target-label budgets are $\{5, 10, 20, 40, 60, 80, 100\}\%$, with 5–20% prespecified as low-label.

The source, target probe, trained-from-scratch model, and graft use four logits and class-weighted cross-entropy. For each candidate depth and class c, source and target activations are independently ordered by $p(y = c \mid x)$, aligned on the quantile grid, and assigned a CKA score. The selector uniformly averages the four scores, preventing the majority class from receiving extra weight.

Bike-demand regression. Washington/DC Bike Sharing is the source ($n_S = 17{,}379$, 12 semantic and 58 encoded features) and Seoul Bike Sharing is the target ($n_T = 8{,}760$, 14 semantic and 36 encoded features); hourly rental count is the shared response. The domains are encoded independently without a supplied feature map. Target transformations are fit only on each outer training pool, target splits are group-disjoint, and all labeled subsets lie inside that pool. The ten nested budgets are $\{0.5, 1, 2, 5, 10, 20, 40, 60, 80, 100\}\%$, with 0.5–20% prespecified as low-label. Models use a scalar output with squared-error training, and quantile alignment orders examples by predicted demand. RMSE is primary; MAE and R^2 are secondary. Automated checks confirm frozen tails, finite predictions, nested subsets, and absence of target-test access. Ordinary squared-error training is outside the bounded-loss theorem as written; this study is therefore an empirical endpoint extension rather than an application of that bound without additional assumptions.

A.2 Four-Class PAM50 Results

Table 5 reports every budget. Across the prespecified 5–20% regime, ADLG improves BA by 0.95 points (0.324 versus 0.315), with a bootstrap 95% CI of [0.44,1.42] points and paired-t CI of [0.20,1.69]. All five outer-split aggregate differences are positive. The corresponding macro-F1 and macro-AUROC differences are +0.97 points [0.34,1.82] and +1.51 points [0.87,2.14]. Budget-specific intervals are wider; they are not treated as seven independent replications or datasets.

PAM50 balanced accuracy is shown in main Figure 3(c); the complete budget table, secondary metrics, selector audit, and per-class diagnostics are reported here.

The class-conditional selector is compared on the same already-trained fixed-depth bank, so this audit changes selection only. It has higher selected BA, more exact oracle hits, and lower regret than either scalar compression (Table 6). Class-specific preferred depths disagree in 74.3% of cells, while the macro choice matches the majority class-specific preference in every cell. Nevertheless, class-conditional CKA selects depth 3 in all 105 cells. Thus this experiment supports the grafting mechanism beyond binary outcomes and the class-conditional rule relative to two alternatives; it does not demonstrate adaptive depth variation on this pair.

Per-class recall is heterogeneous rather than uniformly shifted. At 5%, ADLG-minus-trained-from-scratch changes are +0.27/+0.17/−2.92/+4.71 points for LumA/LumB/HER2/Basal; at 20% they are +0.27/+4.10/−2.50/+1.76; at 100% they are +0.53/−2.56/+4.17/+6.27. Figure 4 shows the full class and selector diagnostics.

Table 6: PAM50 selector audit on a shared fixed-depth result bank. Regret is the retrospective-oracle BA minus selected BA, in points.

Ordering	Mean BA	Oracle hit	Within 1 pt	Regret
Class-conditional	.350	52.4%	61.0%	1.39
Top confidence	.339	25.7%	42.9%	2.54
Probability margin	.343	35.2%	46.7%	2.09

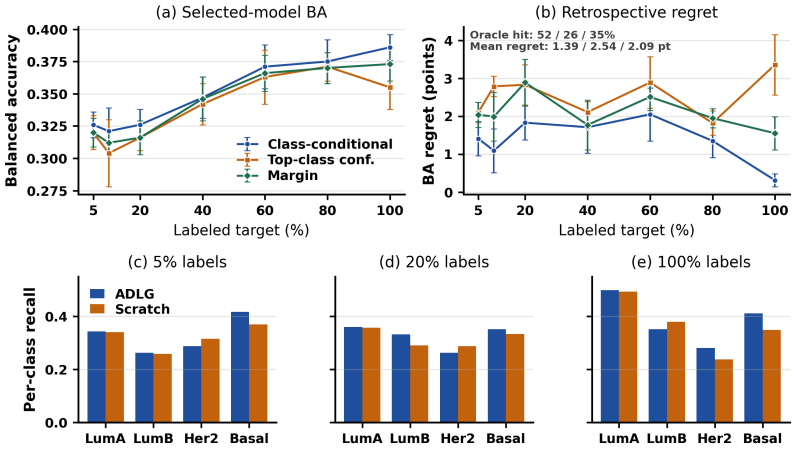

Figure 4: PAM50 diagnostics. Per-class and selector diagnostics for the four-class endpoint extension across label budgets. This is an endpoint extension study rather than an additional replication of the five-pair suite.

A.3 Bike-Demand Regression Results

Positive values in the final column of Table 7 favor ADLG. Across the prespecified 0.5–20% regime, trained-from-scratch minus ADLG RMSE is 11.50 [7.08,15.92] (paired-t: [4.22,18.78]), a 2.70% reduction, and ADLG minus trained-from-scratch R^2 is +0.0289 [0.0183,0.0408]. All five outer-split aggregate RMSE and R^2 differences favor ADLG. The MAE difference is -1.93 [-5.42,1.00] and therefore inconclusive. Budget-level results locate the boundary precisely: RMSE favors ADLG at 0.5–10%, then favors the trained-from-scratch model from 20% onward. This crossover is consistent with, but does not causally identify, the retained-tail mismatch term δ_k in Theorem 1.

Table 7: Bike-demand regression. ΔRMSE is trained-from-scratch minus ADLG, so positive values favor ADLG; intervals resample five outer-split summaries.

Labels	ADLG	Trained from scratch	ΔRMSE [95% CI]	Labels	ADLG	Trained from scratch	ΔRMSE [95% CI]
0.5%	510.19	527.84	+17.65 [3.73,31.57]	20%	338.86	313.48	−25.39 [−33.32,−16.53]
1%	475.60	501.06	+25.47 [9.52,43.96]	40%	323.90	271.09	−52.81 [−61.99,−44.79]
2%	426.09	449.28	+23.19 [12.18,34.16]	60%	314.42	235.48	−78.94 [−96.98,−60.90]
5%	380.06	402.54	+22.48 [15.74,29.22]	80%	311.73	231.26	−80.47 [−90.83,−68.85]
10%	358.16	363.76	+5.60 [0.07,11.36]	100%	304.31	218.29	−86.02 [−97.98,−74.06]

Bike-demand RMSE is shown in main Figure 3(d); the complete all-budget results, secondary metrics, fixed-depth results, and diagnostics are reported here.

The regression selector chooses depth 3 in every cell, while the retrospective oracle varies. Its exact oracle-hit rate is 14.7%, 28.7% of choices fall within 2% of oracle RMSE, and mean relative regret is 9.66%. Hence the regression result supports low-label transfer by the selected graft, but not a claim that CKA adapts depth effectively on this pair. The all-budget mean RMSE spread across fixed depths is 20.66, confirming that interface choice remains consequential even though this ordering rule is not sufficiently discriminative. Figure 5 reports fixed-depth and stress diagnostics.

Under synthetic residual corruption, trained-from-scratch-minus-ADLG RMSE at 5% is 22.88 [13.15,32.60] for clean targets, 43.67 [33.20,52.90] for moderate Gaussian noise, 44.24 [28.58,64.04] for moderate Student-t noise, and 87.61 [73.88,101.33] for severe Student-t noise. At 20%, the

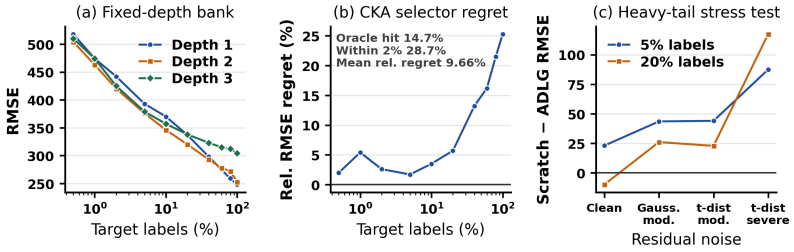

Figure 5: Regression diagnostics. Selected and fixed-depth behavior together with prespecified synthetic residual-noise stress tests. These diagnostics are exploratory robustness analyses and are not additional natural-domain benchmarks.

corresponding differences are -10.04 [-23.91,7.54], 26.14 [13.55,38.74], 22.67 [11.33,37.15], and 117.74 [98.75,132.61]. These controlled perturbations show robustness under the executed noise models; they do not establish distributional robustness outside the executed noise models. All configurations, split manifests, per-cell predictions, fixed-depth fits, and validation checks accompany the result archive.

B Theory and Proofs

This appendix gives proofs and scope conditions for Section 5. The fixed-depth result concerns the frozen-tail estimator. The CKA results characterize interface compatibility and recovery of the population CKA-score maximizer; they do not equate CKA with target risk. The bounds below assume a bounded Lipschitz loss. The Bike-demand experiment uses ordinary squared-error training and is therefore treated as an empirical endpoint extension rather than an application of these bounds without additional boundedness assumptions.

For candidate attachment depth k, recall the grafted predictor

$$f_T^{(k)} = F_k \circ G^{(k)}, \tag{9}$$

where F_k is the retained source tail and $G^{(k)} \in \mathcal{H}_G^{(k)}$ is a target-to-hidden bridge.

For any predictor f, define

$$R_T(f) = \mathbb{E}_{(x,y)\sim T}\left[\ell(f(x), y)\right], \qquad R_T^\star = \inf_f R_T(f). \tag{10}$$

For each candidate depth k, let $\mathcal{G}_k$ be the class of measurable maps from target inputs into the representation expected by F_k. Assume the following infimum is attained and choose

$$g_k^\dagger \in \arg\min_{g\in\mathcal{G}_k} R_T(F_k \circ g). \tag{11}$$

Define

$$\delta_k = R_T(F_k \circ g_k^\dagger) - R_T^\star \tag{12}$$

and

$$\epsilon_k^\star = \inf_{G\in\mathcal{H}_G^{(k)}} \mathbb{E}_{x\sim T}\left[\rho_k\left(G(x), g_k^\dagger(x)\right)\right]. \tag{13}$$

Here ρ_k is a fixed metric on the representation space expected by the retained tail at depth k.

We assume throughout that target inputs are bounded, the loss is bounded and L_{loss}-Lipschitz in its prediction argument, and F_k is $\mathrm{Lip}(F_k)$-Lipschitz with respect to ρ_k. The bridge parameters lie in a bounded set, and

$$\ell \circ F_k \circ \mathcal{H}_G^{(k)}$$

admits a finite uniform-convergence bound.

B.1 Proof of the Target Excess-Risk Decomposition

Theorem 2 (Target excess-risk decomposition). *Let $\widehat{G}_k$ be an approximate empirical-risk minimizer satisfying*

$$\widehat{R}_T(F_k \circ \widehat{G}_k) \le \inf_{G \in \mathcal{H}_G^{(k)}} \widehat{R}_T(F_k \circ G) + \epsilon_{\mathrm{opt},k}. \tag{14}$$

Suppose that with probability at least $1-\delta$,

$$\sup_{G \in \mathcal{H}_G^{(k)}} \left| R_T(F_k \circ G) - \widehat{R}_T(F_k \circ G) \right| \le U_k(m_T, \delta). \tag{15}$$

Define

$$\Gamma_k(m_T, \delta) = 2U_k(m_T, \delta). \tag{16}$$

Then, on the same event,

$$R_T(F_k \circ \widehat{G}_k) - R_T^\star \le \delta_k + L_{\mathrm{loss}} \operatorname{Lip}(F_k)\epsilon_k^\star + \Gamma_k(m_T, \delta) + \epsilon_{\mathrm{opt},k}. \tag{17}$$

Proof. Fix $\varepsilon > 0$. By the definition of $\epsilon_k^\star$, there exists $G_{k,\varepsilon} \in \mathcal{H}_G^{(k)}$ such that

$$\mathbb{E}_{x \sim T}\left[\rho_k\left(G_{k,\varepsilon}(x), g_k^\dagger(x)\right)\right] \le \epsilon_k^\star + \varepsilon. \tag{18}$$

On the event in Equation 15,

$$\begin{aligned} R_T(F_k \circ \widehat{G}_k) &\le \widehat{R}_T(F_k \circ \widehat{G}_k) + U_k(m_T, \delta) && (19)\\ &\le \widehat{R}_T(F_k \circ G_{k,\varepsilon}) + \epsilon_{\mathrm{opt},k} + U_k(m_T, \delta) && (20)\\ &\le R_T(F_k \circ G_{k,\varepsilon}) + 2U_k(m_T, \delta) + \epsilon_{\mathrm{opt},k}. && (21) \end{aligned}$$

For any (x, y), Lipschitz continuity gives

$$\begin{aligned} &\left|\ell\left(F_k(G_{k,\varepsilon}(x)), y\right) - \ell\left(F_k(g_k^\dagger(x)), y\right)\right| && (22)\\ &\quad\le L_{\mathrm{loss}} \operatorname{Lip}(F_k)\rho_k\left(G_{k,\varepsilon}(x), g_k^\dagger(x)\right). && (23) \end{aligned}$$

Taking expectations and using Equation 18,

$$R_T(F_k \circ G_{k,\varepsilon}) - R_T(F_k \circ g_k^\dagger) \le L_{\mathrm{loss}} \operatorname{Lip}(F_k)(\epsilon_k^\star + \varepsilon). \tag{24}$$

By Equation 12,

$$R_T(F_k \circ g_k^\dagger) - R_T^\star = \delta_k.$$

Combining these inequalities and letting $\varepsilon \downarrow 0$ yields Equation 17.

□

Interpretation. The bound contains

$$\underbrace{\delta_k}_{\text{retained-tail insufficiency}} + \underbrace{L_{\mathrm{loss}} \operatorname{Lip}(F_k)\epsilon_k^\star}_{\text{bridge-induced transfer bias}} + \underbrace{\Gamma_k(m_T, \delta)}_{\text{estimation}} + \underbrace{\epsilon_{\mathrm{opt},k}}_{\text{optimization}}. \tag{25}$$

The decomposition makes the role of depth explicit. A poorly matched retained tail increases δ_k; a restricted bridge increases $\epsilon_k^\star$; and a larger bridge class can increase estimation error. Positive transfer occurs only when these costs are smaller than the benefit of estimating fewer target-specific parameters.

B.2 Affine-Bridge Complexity and Depth Dependence

The fixed-depth theorem leaves Γ_k abstract because it applies to a generic bridge class. In the executed standard Census core the trainable affine preactivation is followed by ReLU,

$$G_k(x) = \mathrm{ReLU}(W_k x + b_k), \qquad P_k = d_k(d_T + 1).$$

The ReLU is parameter free and 1-Lipschitz, so the trainable affine parameter count remains P_k; the complexity specialization below therefore counts the affine weights and biases while absorbing the fixed ReLU into the composed Lipschitz map. To make the depth dependence explicit, suppose target inputs satisfy $\|x\|_2 \le B_x$, the affine parameter vector belongs to a bounded P_k-dimensional set, the retained tail is fixed, and the composed loss is uniformly bounded and Lipschitz in the bridge parameters. Standard finite-dimensional covering-number arguments then yield a uniform-convergence term of the form

$$U_k(m_T, \delta) \le C\sqrt{\frac{P_k \log(c\, m_T) + \log(1/\delta)}{m_T}}, \tag{26}$$

for constants C, c depending on the stated norm and Lipschitz bounds. Consequently,

$$\Gamma_k(m_T, \delta) = \widetilde{\mathcal{O}}\left(\sqrt{\frac{P_k + \log(1/\delta)}{m_T}}\right). \tag{27}$$

Why this matters for depth. For fixed d_T, moving to a candidate with smaller hidden width d_k reduces the number of target-specific affine parameters, while δ_k, $\epsilon_k^\star$, and $\mathrm{Lip}(F_k)$ may vary in the opposite direction. Thus the analysis admits a genuinely depth-dependent bias–estimation tradeoff. For Census–Adult the affine bridge counts are $6976, 3488, 1744$ for widths $64, 32, 16$, respectively. Section G.2 extends this to a six-depth ladder spanning $P_k \in \{872, \dots, 27904\}$ (widths 256 down to 8), where the wider dynamic range in P_k makes the estimation term's dependence on depth more visible empirically. The separate matched bridge audit in Section G.8 provides complementary empirical context: richer bridge classes do not yield monotone performance gains, particularly under scarce target supervision. This is consistent with, but does not constitute a direct test of, the bound's capacity–estimation interpretation.

Scope. Equation 26 is a finite-dimensional specialization under explicit boundedness assumptions; it is not claimed as the sharpest possible complexity bound.

B.3 Simultaneous Generalization for Data-Dependent Depth Selection

Theorem 2 is stated for a fixed depth. Because ADLG chooses its depth from a finite candidate set, the fixed-depth events must hold simultaneously.

Corollary 1 (Simultaneous bound for adaptive depth selection)**.** *Let $\widehat{k}$ be any measurable rule taking values in the finite set $\mathcal{K}$. Apply Theorem 2 to each candidate with confidence level $\delta/|\mathcal{K}|$. Then, with probability at least $1 - \delta$,*

$$R_T(F_{\widehat{k}} \circ \widehat{G}_{\widehat{k}}) - R_T^\star \le \delta_{\widehat{k}} + L_{\mathrm{loss}}\,\mathrm{Lip}(F_{\widehat{k}})\epsilon_{\widehat{k}}^\star + \Gamma_{\widehat{k}}\left(m_T, \frac{\delta}{|\mathcal{K}|}\right) + \epsilon_{\mathrm{opt},\widehat{k}}. \tag{28}$$

Proof. For each k, let $\mathcal{E}_k$ be the event on which the fixed-depth bound holds at failure probability $\delta/|\mathcal{K}|$. By the union bound,

$$\Pr\left(\bigcap_{k\in\mathcal{K}} \mathcal{E}_k\right) \ge 1 - \sum_{k\in\mathcal{K}} \frac{\delta}{|\mathcal{K}|} = 1 - \delta.$$

On this simultaneous event the bound holds for every candidate and hence for the random, data-dependent index $\widehat{k}$. □

Scope. This corollary resolves the fact that the deployed depth is selected from data. It does not show that CKA chooses the candidate with smallest target risk or smallest upper bound. That selection quality is assessed empirically through retrospective oracle regret, fixed-depth comparisons, and validation-selected controls.

B.4 Estimation Savings Versus Transfer Bias

Suppose a target-from-scratch estimator satisfies

$$R_T(\widehat{h}_{m_T}) - R_T^\star \le \delta_{\mathrm{scratch}} + \Gamma_{\mathrm{scratch}}(m_T, \delta) + \epsilon'_{\mathrm{opt}}. \tag{29}$$

Define

$$\mathrm{EstimationSaving}(m_T) = \Gamma_{\mathrm{scratch}}(m_T, \delta) - \Gamma_k(m_T, \delta), \tag{30}$$

$$\mathrm{TransferBias}(k) = \delta_k + L_{\mathrm{loss}}\,\mathrm{Lip}(F_k)\epsilon_k^\star - \delta_{\mathrm{scratch}}, \tag{31}$$

$$\Delta_{\mathrm{opt}} = \epsilon_{\mathrm{opt},k} - \epsilon'_{\mathrm{opt}}. \tag{32}$$

Corollary 2 (Bound-comparison condition)**.** *If*

$$\mathrm{EstimationSaving}(m_T) > \mathrm{TransferBias}(k) + \Delta_{\mathrm{opt}}, \tag{33}$$

then the grafted upper bound is smaller than the corresponding trained-from-scratch upper bound.

Proof. Subtracting the trained-from-scratch bound from the grafted bound gives

$$\mathrm{TransferBias}(k) - \mathrm{EstimationSaving}(m_T) + \Delta_{\mathrm{opt}}, \tag{34}$$

which is negative under Equation 33.

□

Important limitation. The corollary compares upper bounds rather than realized target risks.

B.5 Simplified Crossover Regimes

For qualitative interpretation, suppose

$$\Gamma_k(m, \delta) \lesssim \frac{C_G}{\sqrt{m}}, \qquad \Gamma_{\mathrm{scratch}}(m, \delta) \lesssim \frac{C_S}{\sqrt{m}}, \tag{35}$$

and define

$$A_G = \delta_k + L_{\mathrm{loss}}\,\mathrm{Lip}(F_k)\epsilon_k^\star, \qquad A_S = \delta_{\mathrm{scratch}}. \tag{36}$$

Ignoring optimization-slack differences for exposition,

$$A_G + \frac{C_G}{\sqrt{m}} < A_S + \frac{C_S}{\sqrt{m}}. \tag{37}$$

Case I: $A_G \le A_S$ and $C_G < C_S$. The grafted upper bound is favorable throughout this simplified comparison.

Case II: $A_G > A_S$ and $C_G < C_S$. The grafted model has larger transfer/approximation bias but lower estimation complexity and can therefore be favored under limited supervision. Equality occurs at

$$m^\star = \left(\frac{C_S - C_G}{A_G - A_S}\right)^2. \tag{38}$$

Case III: $A_G \le A_S$ and $C_G \ge C_S$. The grafted predictor may instead become favorable only at sufficiently large m.

Case IV: $A_G > A_S$ and $C_G \ge C_S$. The grafted upper bound is not favorable under this simplified model.

These cases are explanatory rather than predictive. The theory does not establish that the bridge class necessarily has smaller statistical complexity than the trained-from-scratch class and does not predict the empirical regime of a new source–target pair from its description alone. The exploratory six-depth ladder in Section G.2 exhibits a pattern qualitatively consistent with Case II across candidate depths of a single pair (larger bridges are favored only once m_T is large enough), which we report as suggestive empirical alignment with this qualitative picture rather than as a test of it.

B.6 Proof of the Spectral CKA–Compatibility Result

Let $X_0 = \widetilde{H}_T^{(k)} \in \mathbb{R}^{m\times p}$ and $Y_0 = \widetilde{H}_S^{(k)} \in \mathbb{R}^{m\times q}$ be the centered target and source representation matrices on the common quantile grid. Standardize them as $X = X_0/\|X_0\|_F$ and $Y = Y_0/\|Y_0\|_F$, so $\|X\|_F = \|Y\|_F = 1$. This leaves linear CKA unchanged. Define

$$S = X^\top X, \qquad C = X^\top Y, \qquad \chi = \|C\|_F^2.$$

We require only that $X \neq 0$; S may be singular and the representation widths may differ.

B.6.1 Exact rank-robust identity

The quantile-aligned linear compatibility error is

$$\epsilon_{\text{QA-lin}}(X,Y) = \min_{B\in\mathbb{R}^{p\times q}} \|XB - Y\|_F^2.$$

Its minimizers include $B^\star = X^\dagger Y$. Since $P_X = XX^\dagger = XS^\dagger X^\top$ is the orthogonal projector onto $\mathrm{col}(X)$,

$$\begin{aligned}\epsilon_{\text{QA-lin}}(X,Y) &= \|(I-P_X)Y\|_F^2 &(39)\\ &= 1 - \mathrm{tr}\big(Y^\top X S^\dagger X^\top Y\big) &(40)\\ &= 1 - \mathrm{tr}\big(C^\top S^\dagger C\big)\,. &(41)\end{aligned}$$

The first equality follows by decomposing $XB - Y$ into its orthogonal components in $\mathrm{col}(X)$ and its complement. The second uses $\|Y\|_F = 1$ and idempotence of P_X.

Proposition 1 (Spectral CKA–compatibility bounds)**.** *Let $\lambda_{\max}(S)$ and $\lambda^+_{\min}(S)$ be the largest and smallest positive eigenvalues of S. Then*

$$\max\left\{0, 1 - \frac{\chi}{\lambda^+_{\min}(S)}\right\} \le \epsilon_{\text{QA-lin}}(X,Y) \le 1 - \frac{\chi}{\lambda_{\max}(S)}, \tag{42}$$

where

$$\chi = \mathrm{CKA}(X,Y)\|X^\top X\|_F\|Y^\top Y\|_F. \tag{43}$$

Proof. Every column of $C = X^\top Y$ lies in $\mathrm{range}(X^\top) = \mathrm{range}(S)$. On this range, the eigenvalues of $S^\dagger$ lie between $1/\lambda_{\max}(S)$ and $1/\lambda^+_{\min}(S)$. Applying these Rayleigh-quotient bounds columnwise gives

$$\frac{\chi}{\lambda_{\max}(S)} \le \mathrm{tr}(C^\top S^\dagger C) \le \frac{\chi}{\lambda^+_{\min}(S)}.$$

Substitution into Equation 41 proves the two inequalities; the lower bound is clipped at zero because a squared residual is nonnegative. Equation 43 is the definition of linear CKA. □

Corollary 3 (A sufficient regime for CKA ordering)**.** *Suppose X has rank r, unit Frobenius norm, and equal nonzero singular values. Then the positive eigenvalues of S all equal $1/r$, and*

$$\epsilon_{\text{QA-lin}}(X,Y) = 1 - r\chi. \tag{44}$$

Across candidate depths with common r, $\|X^\top X\|_F$, and $\|Y^\top Y\|_F$, minimizing this reconstruction error is equivalent to maximizing CKA.

Proof. Unit Frobenius norm gives $\mathrm{tr}(S) = 1$. Equal positive eigenvalues and rank r therefore imply $S^\dagger = rP_{\mathrm{range}(S)}$. Because the columns of C lie in that range, $\mathrm{tr}(C^\top S^\dagger C) = r\|C\|_F^2 = r\chi$. The ordering statement follows from Equation 43. □

Interpretation and scope. The exact identity separates alignment strength χ from the conditioning of the target representation. CKA orders the linear reconstruction diagnostic in the corollary's isotropic, fixed-rank regime; outside it, different spectra and Gram factors can reverse the ordering. Moreover, this diagnostic concerns the quantile-indexed matrices used by the selector. It is neither the population training error of the nonlinear bridge nor a target-risk bound.

Proposition 2 (Limit of representation-only selection)**.** *Without assumptions linking target labels to the candidate representations, no selector that observes only unlabeled representation distributions can be guaranteed to choose the target-risk-optimal candidate.*

Proof. Let x be uniform on $\{-1, +1\}$, and consider two fixed candidates $f_1(x) = \mathbf{1}\{x = +1\}$ and $f_2(x) = \mathbf{1}\{x = -1\}$. Define target problem T_1 by $y = f_1(x)$ and T_2 by $y = f_2(x)$. The input distribution, candidate representations, and every statistic computed only from their unlabeled distributions are identical under T_1 and T_2, so a representation-only selector must make the same choice on both. Under zero-one loss, however, f_1 is optimal on T_1 and f_2 is optimal on T_2. □

The proposition does not apply to the full ADLG selector as an impossibility theorem because its target probe is trained with labels. It instead states the assumption boundary for representation similarity alone: task information or additional structural assumptions are necessary for universal target-risk ordering.

B.7 Proof of Conditional Depth-Selection Stability

Let C_k denote the population representation-alignment score and $\widehat{C}_k$ its empirical estimate. Define

$$k^\star_{\mathrm{CKA}} = \arg\max_{k \in \mathcal{K}} C_k \tag{45}$$

and

$$\gamma = C_{k^\star_{\mathrm{CKA}}} - \max_{j \neq k^\star_{\mathrm{CKA}}} C_j. \tag{46}$$

Lemma 1 (Conditional CKA selector stability)**.** *If*

$$\max_{k \in \mathcal{K}} \left| \widehat{C}_k - C_k \right| \leq \eta. \tag{47}$$

and

$$\gamma > 2\eta, \tag{48}$$

then

$$\arg\max_{k \in \mathcal{K}} \widehat{C}_k = k^\star_{\mathrm{CKA}}. \tag{49}$$

Proof. For the population-best depth,

$$\widehat{C}_{k^\star_{\mathrm{CKA}}} \geq C_{k^\star_{\mathrm{CKA}}} - \eta. \tag{50}$$

For every competing $j \neq k^\star_{\mathrm{CKA}}$,

$$\widehat{C}_j \leq C_j + \eta \leq C_{k^\star_{\mathrm{CKA}}} - \gamma + \eta. \tag{51}$$

Therefore

$$\widehat{C}_{k^\star_{\mathrm{CKA}}} - \widehat{C}_j \geq \gamma - 2\eta > 0. \tag{52}$$

Hence the empirical maximizer equals the population CKA maximizer.

□

Scope. This auxiliary lemma concerns recovery of the population alignment-score maximizer. It does not state that this depth minimizes target risk and is not used as a downstream-performance guarantee in the main paper.
The error η may depend on finite labeled target data used to fit the probe, finite source and target samples, probe optimization, quantile interpolation, and representation estimation. No universal concentration rate is asserted. This is consistent with the low exact-match rate observed for the six-depth ladder in Section G.2: with 6 closely-performing candidates and only 3 repeats, the empirical margin $\widehat{\gamma}$ is frequently small relative to plausible estimation error, so exact recovery of $k^\star_{\mathrm{CKA}}$ is not expected even when the lemma's qualitative picture holds.

B.8 Quantile Alignment as a Common Rank Coordinate

ADLG compares unpaired source and target domains by indexing each domain according to a task-relevant within-domain rank. The construction does not assume that equal quantiles correspond to identical examples or identical latent states.

Assumption 1 (Monotone domain-specific task orderings)**.** *Let Z_S and Z_T denote scalar latent task variables in the source and target domains, with possibly different distributions F_{Z_S} and F_{Z_T}. Suppose*

$$r_S = \phi_S(Z_S), \qquad r_T = \phi_T(Z_T), \tag{53}$$

where ϕ_S and ϕ_T are strictly increasing.
Suppose further that representations at depth k admit domain-specific curves

$$h_S^{(k)} = \psi_S^{(k)}(Z_S), \qquad h_T^{(k)} = \psi_T^{(k)}(Z_T). \tag{54}$$

Proposition 3 (Population quantile indexing)**.** *Under Assumption 1, equal quantiles of r_S and r_T index equal* within-domain latent ranks*:*

$$F_{Z_S}(Z_S) = q, \qquad F_{Z_T}(Z_T) = q. \tag{55}$$

Thus quantile alignment establishes a common rank coordinate $q \in (0, 1)$, but does not imply $Z_S = Z_T$ or equality of the corresponding representations.

Proof. Strict monotonicity gives

$$F_{r_S}(r_S) = F_{Z_S}(Z_S), \qquad F_{r_T}(r_T) = F_{Z_T}(Z_T). \tag{56}$$

Selecting the same score quantile q in each domain yields

$$Z_S = F_{Z_S}^{-1}(q), \qquad Z_T = F_{Z_T}^{-1}(q). \tag{57}$$

These states share the same within-domain rank q, but need not have the same numerical value or distribution. The corresponding representations are

$$\psi_S^{(k)}\left(F_{Z_S}^{-1}(q)\right) \qquad \text{and} \qquad \psi_T^{(k)}\left(F_{Z_T}^{-1}(q)\right). \tag{58}$$

□

Scope. The proposition formalizes the indexing operation rather than semantic correspondence. In practice, r_T is estimated from limited labeled target data. Probe error, rank noise, ties, finite-sample interpolation, and optimization randomness can perturb the idealized ordering. These effects are evaluated empirically through ordering controls, probe-sensitivity experiments, and cross-fitting.

B.9 Structural GW Interpretation

Let

$$C_S \in \mathbb{R}^{d_k \times d_k}, \qquad C_T \in \mathbb{R}^{d_T \times d_T} \tag{59}$$

denote source-hidden-unit and target-feature dissimilarity matrices.
The unregularized relational term of the GW objective is

$$\mathcal{L}_{\mathrm{GW}}(T) = \sum_{u,v,a,b} \left(C_S[u,v] - C_T[a,b]\right)^2 T_{ua} T_{vb}. \tag{60}$$

The coupling therefore compares within-domain relational geometry rather than direct feature values. A low GW objective indicates relative compatibility under the chosen dissimilarity matrices. It does not imply semantic identity between target features and source hidden units, nor does it imply lower target risk.

B.10 Residual Reconstruction Non-Degradation

Suppose

$$W = W_0^{\mathrm{GW}} + \Delta \tag{61}$$

and let $\mathcal{L}_{\mathrm{rec}}(W)$ be the source-side reconstruction objective used to train the residual.
If the residual class contains $\Delta = 0$, then

$$\inf_{\Delta} \mathcal{L}_{\mathrm{rec}}\left(W_0^{\mathrm{GW}} + \Delta\right) \le \mathcal{L}_{\mathrm{rec}}\left(W_0^{\mathrm{GW}}\right). \tag{62}$$

This property concerns only source reconstruction and does not imply

$$R_T\left(W_0^{\mathrm{GW}} + \Delta\right) \le R_T\left(W_0^{\mathrm{GW}}\right). \tag{63}$$

B.11 Frozen-Head Reparameterization at the Deepest Attachment

Proposition 4 (Deepest frozen-head equivalence)**.** *Let $a, \tilde{a} \in \mathbb{R}^d$ have nonzero entries and the same numbers of positive and negative signs. For fixed c, define*

$$\mathcal{F}_{a,c} = \{x \mapsto \sigma(a^\top \mathrm{ReLU}(Wx + b) + c) : W \in \mathbb{R}^{d\times p},\ b \in \mathbb{R}^d\}.$$

Then $\mathcal{F}_{a,c} = \mathcal{F}_{\tilde{a},c}$.

Proof of Proposition 4.

Proof. Permute hidden units to align the signs of corresponding head coefficients. For each aligned unit, set

$$\tilde{w}_i = \frac{|a_i|}{|\tilde{a}_i|} w_i, \qquad \tilde{b}_i = \frac{|a_i|}{|\tilde{a}_i|} b_i.$$

Positive homogeneity of ReLU and the aligned signs give

$$\tilde{a}_i \,\mathrm{ReLU}(\tilde{w}_i^\top x + \tilde{b}_i) = a_i \,\mathrm{ReLU}(w_i^\top x + b_i).$$

Summing over units preserves the logit and hence the predictor. The positive rescalings and permutation are invertible, proving both inclusions. □

Scope and training implications. The result concerns unrestricted expressivity at the deepest attachment only. It does not assert equivalence for different fixed output biases, constrained bridges, or arbitrary heads containing zero entries. Zero coefficients yield inactive units and must be treated separately; the same argument applies to active units when their positive and negative counts match. If the output bias is trainable over the same range in both classes, the equivalence also holds by taking the union over that common range.
Equal function classes do not imply equal regularized solutions or optimization trajectories. For an explicit bridge-weight L_2 penalty, writing $u_i = |a_i| w_i$ gives

$$\lambda \sum_i \|w_i\|_2^2 = \lambda \sum_i \frac{\|u_i\|_2^2}{|a_i|^2}.$$

If hidden biases are also penalized, their squared terms transform analogously. This illustrates how head magnitudes can induce different penalties in the same functional coordinates. It is not an exact characterization of AdamW or gradient clipping and does not demonstrate a pretraining benefit. The proposition compares frozen-head classes; it does not characterize the trajectories of the subsequent jointly fine-tuned models.

C Dataset and Preprocessing Provenance

For every source–target pair, Appendix C reports raw and final sample counts, semantic feature counts, encoded dimensions, class distributions, missing-data treatment, duplicate removal, categorical processing, label construction, and leakage-prevention rules.
Entries below are populated only when the executed notebooks, preprocessing audits, raw-data manifests, or frozen encoded matrices establish them directly. A historical preprocessing detail not supported by these materials is described as not available in the archived preprocessing record rather than inferred from external knowledge.

Table 8: Verified dataset statistics from the supplied executed notebooks, preprocessing audits, frozen encoded matrices, and, for WDBC and Mammographic Masses, direct comparison with the supplied raw files. "Raw" refers to the pre-encoding source artifact. MIMIC has 140 rows from 100 unique subjects.

Dataset	Raw n	Final n	Semantic d	Encoded d	Positive n	Positive rate
WDBC	569	569	30	30	212	0.3726
Mammographic Masses	961	961	5	5	445	0.4631
Telco Churn	7043	7043	19	45	1869	0.2654
Iranian Churn	3150	3150	13	13	495	0.1571
Taiwan Credit	30000	29965	23	29	6630	0.2213
German Credit	1000	1000	20	61	300	0.3000
HCV	615	608	12	12	75	0.1234
ILPD	583	583	10	10	416	0.7136
eICU	2520	2492	34	68	212	0.0851
MIMIC	140	140	34	68	20	0.1429
Census-Income	299285	292550	40	511	18568	0.0635
Adult	48842	48790	14	108	11681	0.2394
Student Dropout	4424	3630	36	238	1421	0.3915
OULAD	32587	32587	46	85	10150	0.3115

Table 9: Verified replacements in the historical Mammographic input matrix. The values establish the transformation, not its original fitting scope.

Predictor	Missing entries	Replacement
BI-RADS assessment	2	4
Age	5	57
Shape	31	4
Margin	48	1
Density	76	3

Evidence scope. Within each outer split, preprocessing and scaling parameters are fitted on the permitted training partition and applied without refitting to validation/test partitions whenever the executed pipeline starts from raw or auditable intermediate inputs. Some experiments instead receive already encoded CSVs. For those inputs, result files establish the downstream split-local operations but cannot by themselves establish how the upstream encoded matrix was imputed or encoded; we rely only on supplied preprocessing notebooks, manifests, and hashes, and describe missing provenance as not available in the archived preprocessing record. Repeated outer splits are paired analysis units, not independent datasets.

For eICU and MIMIC, "Raw n" refers to the supplied study-cohort extracts before outcome filtering, not to the complete underlying clinical databases.

C.1 WDBC → Mammographic Masses

Task and verified label alignment. Binary malignancy classification. Direct comparison of all supplied raw and processed records verifies WDBC diagnosis encoding as benign (B) $= 0$ and malignant (M) $= 1$ for all 569 observations (357/212). The Mammographic target retains the raw severity encoding, benign $= 0$ and malignant $= 1$, for all 961 observations (516/445), as documented in the supplied dataset description. Thus class-1 scores have the same outcome orientation in both domains. This establishes label alignment, without asserting identical measurement processes or patient populations across domains.

Verified raw-to-processed transformations. All 30 WDBC predictors exactly match the supplied raw records. The with-ID export retains the original unique identifiers; the model loader removes that column, and the no-ID export is identical after its removal. The Mammographic matrix retains all raw rows in order, all labels, and every observed predictor value. Its only value changes are replacement of 162 missing entries across 131 rows. Age is filled with 57; BI-RADS assessment, shape, margin, and density are filled with 4, 4, 1, and 3, respectively. These values coincide with the full-table age median and the respective predictor modes.

Table 10: Leakage-safer WDBC→Mammographic Masses rerun: ADLG-Frozen minus the architecture-matched trained-from-scratch neural model. This three-arm raw-data rerun evaluates preprocessing isolation and does not replace the canonical comparison against the tuned target-only MLP, for which WDBC remains a negative-transfer boundary. Brackets show paired 95% bootstrap intervals over ten outer splits.

Metric	5% labels	10% labels	20% labels
AUROC	0.0355 [0.0155, 0.0567]	0.0459 [0.0263, 0.0659]	0.0367 [0.0223, 0.0521]
PR-AUC	0.0273 [−0.0022, 0.0577]	0.0337 [0.0053, 0.0623]	0.0278 [0.0112, 0.0485]
Balanced accuracy	0.0612 [0.0407, 0.0800]	0.0543 [0.0434, 0.0672]	0.0262 [0.0100, 0.0429]
Minority recall	0.0225 [−0.0068, 0.0539]	0.0124 [−0.0169, 0.0416]	0.0427 [0.0191, 0.0742]

Exact reconstruction and its scope. The newly reconstructed notebook `WDBC_Mammo_Exact_Preprocessing.ipynb` generates the WDBC with-ID, WDBC no-ID, and Mammographic processed CSVs directly from the supplied raw files. SHA-256 comparisons establish byte-for-byte equality with all three supplied historical CSVs, including column order and serialization. The notebook and its manifest are reconstruction artifacts, not the recovered original preprocessing implementation. They neither alter nor constitute a rerun of the reported model experiments.

Leakage-safe raw-data rerun. We retain the historical reconstruction above only as a provenance record; it is not used as the target input for the corrected experiment. The corrected three-arm WDBC→Mammographic Masses rerun loads the raw target file, creates each stratified 80/20 outer split first, and fits missing-value replacements using only covariates from its 768-row outer target-training pool. Age receives the training-pool median; BI-RADS assessment, shape, margin, and density receive training-pool modes. These values are frozen before application to both the training pool and its 193 held-out rows. Target standardization is then fitted on the completed training pool only, applied unchanged to the test rows, and clipped to $[-8, 8]$. The whole outer training pool is available unlabeled for this preprocessing; no test covariates, test labels, or labels outside the current supervision budget are used.

The per-split audit records modes $(4, 4, 1, 3)$ for BI-RADS, shape, margin, and density on all ten repeats, while the training-pool age median is 56, 56.5, or 57 (historical full-table value: 57). Thus the corrected results need not be numerically identical to the historical run, but they test the same hypothesis without relying on its unknown imputation fitting scope. Table 10 reports the paired low-label gains from this corrected protocol.

This audit compares ADLG with the trained-from-scratch model to test split-first preprocessing; it does not alter the main Table 2 comparison against the stronger tuned target-only MLP.

Repeated profiles and BI-RADS assessment. No raw rows were removed: the raw Mammographic table contains 272 duplicate rows beyond first occurrences, and the processed table contains 322. Identical profiles do not establish patient identity, and the supplied target file does not support a patient-independent splitting claim. The five-predictor model retains BI-RADS assessment, which the supplied description distinguishes as a radiologist assessment/comparator rather than one of the four predictive attributes. Raw assessment values are retained without range correction. Accordingly, this audit supports split-first preprocessing under the supplied row structure, but it does not make a patient-independent splitting claim or an assessment-free sensitivity claim.

C.2 Telco Churn → Iranian Churn

Task. Binary customer-churn prediction.

Preprocessing. The executed notebook receives already-numeric preprocessed matrices. Telco contains 7,043 raw and final rows, 19 semantic predictors, and 45 encoded predictors, with 1,869 churn positives (26.54%). Iranian Churn contains 3,150 raw and final rows, 13 semantic and 13 encoded predictors, with 495 churn positives (15.71%). The harmonized mapping is 0 = non-churn and 1 = churn. Remaining missing values are rejected. Source standardization is fit on permissible source data; target standardization is fit only on the current Iranian outer-training pool,

applied unchanged to held-out rows, and clipped to $[-8, 8]$. The raw categorical-encoding and duplicate-removal provenance are not available in the archived preprocessing record.

Leakage prevention. No held-out target-test statistic is used during preprocessing, selection, or training.

C.3 TAIWAN CREDIT → GERMAN CREDIT

Task. Binary credit-risk prediction.

Preprocessing. The raw Taiwan Credit source contains 30,000 observations and 23 semantic predictors; after preprocessing, 29,965 observations remain and 29 encoded predictors are used. German Credit contains 1,000 observations and 20 semantic predictors, with 61 encoded predictors in the final matrix. The harmonized mapping is $0 =$ non-default/good credit and $1 =$ default/bad credit. Taiwan has 6,630 positives (22.13%); German Credit has 300 positives (30.00%). The model-input matrices are numerical and remaining missing values are rejected. Source standardization is fit on permissible source data; target standardization is fit only on the current German outer-training pool, applied unchanged to held-out rows, and clipped to $[-8, 8]$. Raw categorical-processing and duplicate-removal provenance are not available in the archived preprocessing record.

C.4 HCV → ILPD

Task. Binary liver-related classification. Exact 0/1 label semantics and their harmonization provenance are not recoverable from the supplied frozen preprocessing artifacts.

Preprocessing. The raw HCV artifact contains 615 rows, of which 608 remain in the final modeling table. HCV has 12 semantic and 12 encoded predictors, with 75 class-1 observations (12.34%). ILPD contains 583 rows, 10 semantic and 10 encoded predictors, with 416 class-1 observations (71.36%). The matrices reaching ADLG are numerical and remaining missing values are rejected. Source standardization is fit on permissible source data; target standardization is fit only on the current ILPD outer-training pool, applied unchanged to held-out rows, and clipped to $[-8, 8]$. The exact semantic interpretation or cross-dataset harmonization of the 0/1 labels is not established by the supplied preprocessing artifacts. We therefore describe the task as binary classification without assigning a stronger cross-dataset label interpretation.

C.5 EICU → MIMIC

Task. Hospital-mortality prediction using structured clinical variables.

Input setting and claim scope. This is a clinically harmonized, matched-variable transfer setting. The 34 clinical variables and mortality outcome, together with intended unit and availability-window compatibility, are upstream design choices made before ADLG is applied. The supplied records do not fully document every unit match, which remains a benchmark limitation. ADLG receives no feature-to-feature map during bridge learning, but this benchmark does not test whether clinical schema harmonization can be omitted and is excluded from claims about transfer without supplied semantic matching.

Preprocessing. The supplied eICU study-cohort extract contains 2,520 rows; after outcome filtering, 2,492 labeled rows remain. The source contains 212 mortality-positive observations (8.51%). The supplied MIMIC study-cohort extract contains 140 rows from 100 unique subjects, with 20 mortality-positive rows (14.29%). The executed ICU notebook uses 34 prespecified clinically matched variables for both domains. Numerical variables are median-imputed using permissible training data and accompanied by explicit missingness indicators; categorical variables, where present, are frequency encoded using training-only frequencies. The 34 semantic variables therefore produce 68 encoded model inputs. Target standardization is fit only on the current MIMIC training pool. The binary outcome is hospital mortality, with class 1 denoting mortality; MIMIC contains 20 mortality-positive rows (14.29%). The supplied workbook does not document units for every matched clinical variable, so semantic/unit compatibility of the 34 declared pairs is an explicit benchmark assumption rather than an inferred fact.

Patient-level leakage prevention. MIMIC is split by unique subject in 10 repeated outcome-stratified 80/20 splits. Each repeat contains 80 training subjects and 20 held-out subjects with zero subject overlap. Because a subject can contribute multiple rows, row-level train/test counts can vary across repeats. All target-label budgets and target preprocessing are constructed strictly inside the training-subject pool.

Table 11: Exact number of labeled Adult training examples at each target-label budget.

Target	0.5%	1%	2%	5%	10%	20%	40%	60%	80%	100%
Adult	195	390	781	1951	3903	7806	15613	23419	31226	39032

Table 12: Exact class composition of each Adult labeled subset.

Class	0.5%	1%	2%	5%	10%	20%	40%	60%	80%	100%
Minority	47	93	187	467	934	1869	3738	5607	7476	9345
Majority	148	297	594	1484	2969	5937	11875	17812	23750	29687

C.6 CENSUS-INCOME → ADULT

Task. Binary income classification.

Final encoded data. The Census-Income source contains 299,285 raw observations and 40 semantic predictors. After preprocessing, 292,550 observations remain and the categorical expansion produces 511 encoded predictors. The final source class counts are 273,982/18,568 (negative/positive; 6.35% positive). Adult contains 48,842 raw observations and 14 semantic predictors. After preprocessing, 48,790 observations remain and the final representation contains 108 encoded predictors, with class counts 37,109/11,681 (negative/positive; 23.94% positive). The final binary label is income class.

Preprocessing and leakage prevention. The final matrices are preprocessed but unscaled. Target scaling is fit inside the Adult outer-training pool and then applied unchanged to validation/test rows. Detailed categorical encoding, missing-value handling, and duplicate/removal provenance are reported by the frozen preprocessing audit and are not inferred beyond what that artifact establishes.

C.7 STUDENT DROPOUT → OULAD

Task. Heterogeneous student-outcome prediction.

Preprocessing. The raw Student Dropout table contains 4,424 observations and 36 semantic predictors. After preprocessing, 3,630 observations remain and the model input contains 238 encoded predictors. There are 1,421 dropout-positive observations (39.15%).
OULAD contains 32,587 observations, 46 semantic predictors, and 85 encoded model inputs, with 10,150 dropout-positive observations (31.15%). The harmonized outcome mapping is $0 =$ non-dropout/graduate and $1 =$ dropout. Student Dropout contains 1,421 positives (39.15%) and OULAD contains 10,150 positives (31.15%). The matrices reaching ADLG are numerical and remaining missing values are rejected. Source standardization is fit on permissible source data; target standardization is fit only on the current OULAD outer-training pool, applied unchanged to held-out rows, and clipped to $[-8, 8]$. The raw OULAD aggregation, excluded-class construction, original categorical encoding, and duplicate-removal provenance are not available in the archived preprocessing record.

C.8 ABSOLUTE TARGET-LABEL BUDGETS

For every target dataset and target-label fraction, Appendix C.8 reports both the number of labeled examples and the minority-class count. For Adult, the realized counts are taken directly from the saved final split manifest rather than reconstructed by rounding.
For the executed 10-repeat standard three-arm protocol, the target outer-training pool is stratified and the 5–100% subsets are nested class-wise. Exact total labeled counts recoverable from the notebooks are:
Exact class-wise counts for WDBC→Mammographic and HCV→ILPD, and complete repeat-specific ICU row counts, are not inferred when the corresponding frozen split/result manifests are unavailable. The uploaded engineering Census notebook also uses 5–100% budgets with counts 1951, 3903, 7806, 15613, 23419, 31226, and 39032, but those values are not used to replace the finalized 5-outer-split Census protocol above.

D CANONICAL IMPLEMENTATION PROTOCOL

D.1 OUTER SPLITS, LABEL BUDGETS, AND INFORMATION ISOLATION

For every pair, target data are first divided into an outer training pool and a held-out test set.

Table 13: Exact labeled-target counts in the executed 10-repeat three-arm notebooks for targets whose counts are explicitly recoverable from the supplied artifacts.

Target	5%	10%	20%	40%	60%	80%	100%
Iranian Churn	126	252	504	1008	1512	2016	2520
German Credit	40	80	160	320	480	640	800
OULAD	1303	2607	5214	10428	15641	20855	26069
Derm7pt development	31	61	123	246	370	493	616

All target-label budgets are constructed strictly inside the outer training pool. Where possible, budget subsets are nested within an outer split and repeat.
Within the executed training code, the target test set is excluded from the following fitting and selection operations. For the corrected raw-data WDBC→Mammographic rerun, this also includes imputation: its median/mode statistics are fitted only on the outer target-training pool (Appendix C.1). This statement does not certify upstream preprocessing of other already-encoded dataset inputs.

- fitting standardization statistics within the executed training code;
- target-probe training;
- task-score construction;
- quantile ordering;
- CKA computation;
- depth selection;
- bridge initialization;
- GW construction;
- early stopping;
- hyperparameter selection;
- threshold selection;
- model selection.

All paired methods inside a

$$(\text{outer split}, \text{budget}, \text{repeat})$$

cell use identical labeled target indices and identical held-out test examples.
Optimization/subset repeats are nested within an outer split and are not treated as independent dataset replications.

Executed non-ICU standard three-arm protocol. For WDBC→Mammographic, Telco→Iranian, Taiwan→German, HCV→ILPD, and Student→OULAD, the notebooks use 10 repeated stratified 80/20 target splits with random state $23051991 + \text{repeat}$. Nested class-wise labeled subsets use fractions $\{5, 10, 20, 40, 60, 80, 100\}\%$ and RNG $23051991 + 1000 + \text{repeat}$. The held-out target-test rows never enter the labeled pool. The uploaded engineering Census standard three-arm notebook uses the same 10-repeat/7-budget construction, but it is not the source of the finalized Census $5\times3\times10$ headline statistics.

Executed eICU → MIMIC protocol. The target split is made at the subject level in 10 repeated outcome-stratified 80/20 splits, with 80 training and 20 held-out subjects per repeat and zero subject overlap. Label budgets are formed only within the training-subject pool; row counts vary because subjects may contribute multiple observations. These safeguards address subject-level leakage but do not alter the benchmark's upstream clinically matched-variable construction.

Table 14: Pair-specific architecture and transfer settings. Source and probe networks use hidden widths $64, 32, 16$ followed by a scalar output. Candidate depths are $\{1, 2, 3\}$ and bridges are affine+ReLU. Phase 1 freezes the source tail; Phase 2 jointly fine-tunes it (in the engineering notebook for Census). Input widths and recorded source checkpoint information available in the supplied run records are listed below. The descriptions distinguish saved checkpoint bundles from the WDBC audit checkpoints and the Census release/configuration audit.

Pair	Source input width	Probe input width	Source checkpoint evidence
WDBC→Mammo	30	5	Three audit checkpoints; seed IDs not recorded
Telco→Iranian	45	13	Saved checkpoint bundle
Taiwan→German	29	61	Saved checkpoint bundle
HCV→ILPD	12	10	Saved checkpoint bundle
eICU→MIMIC	68	68	Saved checkpoint bundle
Census→Adult	511	108	Released final configuration / notebook audit
Student→OULAD	238	85	Saved checkpoint bundle
HAM10000 → Derm7pt	529	524–526	Saved validated checkpoint bundle

Final Census-Income → Adult split settings. Paired Adult outer splits: 5.
Paired repeats per outer split: 3.
Outer train/test fraction: stratified $80/20$.
Target-label budgets: $\{0.5, 1, 2, 5, 10, 20, 40, 60, 80, 100\}\%$.
Total core paired cells: $5 \times 10 \times 3 = 150$.
The exact outer-split and labeled-subset seeds are released with the final configuration and split manifests.

D.2 Binary Benchmark Source Network

The executed source networks use hidden widths $64 \to 32 \to 16$, ReLU after each hidden affine layer, and a final $16 \to 1$ logit head. Source training uses class-weighted `BCEWithLogitsLoss`, AdamW with learning rate 4×10^{-3}, weight decay 10^{-4}, and gradient clipping at 5.0. WDBC, Telco, Taiwan, and Student use 180 source epochs; HCV and eICU use 500. The uploaded engineering Census notebook uses 8 source epochs; this training schedule is an implementation-audit fact and does not replace the finalized Census statistical protocol. The executed notebooks use fixed epoch schedules rather than early stopping. Exact source random seeds, source split and checkpoint provenance, and the finalized Census source schedule remain tied to the released configuration where not explicit in the supplied artifacts. Section G.2 reports an exploratory extension of this architecture to six hidden layers for Census–Adult only.

D.3 Binary Target-Side Probe

The target probe is trained using only the labeled target subset available at the current budget and is refit for every repeat–budget cell. Each probe mirrors the source hidden widths $64, 32, 16$ while using the target input dimension shown in Table 14; its class-1 probability is the target ordering score r_T. With a scalar output head, the exact trainable parameter count is $64d_T + 2689$: 3009, 3521, 6593, 3329, 7041, 9601, and 8129 for Mammographic Masses, Iranian Churn, German Credit, ILPD, MIMIC, Adult, and OULAD, respectively. The HAM→ Derm7pt probe has $d_T = 524$–526 after labeled-subset-fitted metadata encoding, hence $36{,}225$–$36{,}353$ parameters. The shared widths define the diagonal candidate-depth comparison; CKA then tests empirical compatibility at those corresponding levels rather than assuming that separately trained representations are identical.
Probe training uses class-weighted `BCEWithLogitsLoss`, AdamW with learning rate 5×10^{-3}, weight decay 10^{-4}, and gradient clipping at 5.0. WDBC, Telco, Taiwan, Student, and the uploaded engineering Census notebook use 45 probe epochs; HCV and eICU use 80; HAM→Derm7pt uses 45. These schedules and architectures are fixed before examining test outcomes. No dropout or validation-based early stopping is used; the stated weight decay, class weighting, and gradient clipping are the regularization safeguards, while the fixed schedule avoids consuming scarce target labels for an additional stopping split. The probe is not the deployed predictor and its probability values

are not evaluated as calibrated estimates: it supplies hidden activations and a rank ordering for interface screening. Because the sigmoid is strictly monotone, this ordering is exactly the same as sorting the probe logits and is unchanged by any global logit shift. The Census sensitivity audit in Section G.11 separately varies optimization seed, architecture, and regularization; across its representative budgets, the largest within-budget range in selected-model balanced accuracy is 0.0043. The low-label protocols should not be conflated: the finalized Census 0.5% cell contains 195 labeled Adult examples, the smallest HAM→ Derm7pt cell contains 31 labeled development examples at 5%, and the 20-example cells belong to the separate fixed-budget five-pair benchmark. In every case the applicable subset size and probe input width are recorded with the executed cell.

D.4 Operational Definition of the Quantile-CKA Selector

For every candidate depth k, let $h_S^{(k)}(x)$ denote the activation of the pretrained source network at that depth and $h_T^{(k)}(x)$ the depth-compatible activation of the target probe.

Source representations. On the permissible source setup set,

$$H_S^{(k)} = \left[h_S^{(k)}(x_1^S); \ldots; h_S^{(k)}(x_{n'_S}^S)\right]. \tag{64}$$

Target representations. On the permissible target outer-training pool,

$$H_T^{(k)} = \left[h_T^{(k)}(x_1^T); \ldots; h_T^{(k)}(x_{N'_T}^T)\right]. \tag{65}$$

The held-out target test set is excluded.

Ordering scores. The source ordering score r_S is the source model's task score. The target ordering score r_T is the corresponding score from the target probe.
Binary studies use positive-class probabilities; sigmoid and logit sorting are equivalent. HAM10000 → Derm7pt is specifically melanoma versus non-melanoma, not seven-class prediction. The PAM50 extension instead computes one ordering and CKA value per class probability and averages uniformly across the four classes. The regression extension orders by the predicted continuous outcome. In every case, the frozen output head requires a shared endpoint and coding between source and target. For all binary selector runs, the candidate depth set is $\{1, 2, 3\}$, the common quantile grid uses $Q = 128$, interpolation is linear, and the target probe is refit within the target-training partition. These shared settings apply to WDBC→Mammographic, Telco→Iranian, Taiwan→German, HCV→ILPD, eICU→MIMIC, Census→Adult, Student→OULAD, and HAM10000→Derm7pt. The multiclass and regression rules are fully specified in Appendix A.1.

Quantile interpolation. Source and target examples are independently sorted by r_S and r_T.
For a fixed grid

$$q_1, \ldots, q_Q \in [0, 1],$$

every representation coordinate is interpolated along its domain-specific empirical rank axis.
The common quantile grid is a rank-index coordinate and does not create pseudo-paired source and target observations.
This gives

$$\widetilde{H}_S^{(k)} \quad \text{and} \quad \widetilde{H}_T^{(k)},$$

which have a common number of rows Q but may have different widths.

CKA and selection. After centering,

$$\widehat{C}^{(k)} = \mathrm{CKA}\left(\widetilde{H}_S^{(k)}, \widetilde{H}_T^{(k)}\right), \tag{66}$$

and

$$k^\star = \arg\max_{k \in \mathcal{K}} \widehat{C}^{(k)}. \tag{67}$$

Centering is performed across the Q quantile locations. The executed selector does not whiten columns or invert an empirical covariance matrix. Consequently, rank-deficient representations are permitted as long as neither centered Gram matrix is identically zero; such a degenerate candidate is otherwise undefined and must be rejected rather than assigned an inferred score.
No held-out target-test metric participates in this selection.

Table 15: Affine weight-plus-bias parameter counts for the executed target bridge before the parameter-free ReLU.

Pair	Target d_T	Depth 1 (64)	Depth 2 (32)	Depth 3 (16)
WDBC→Mammo	5	384	192	96
Telco→Iranian	13	896	448	224
Taiwan→German	61	3968	1984	992
HCV→ILPD	10	704	352	176
eICU→MIMIC	68	4416	2208	1104
Census→Adult	108	6976	3488	1744
Student→OULAD	85	5504	2752	1376

Table 16: Endpoint-specific executed configurations. The grafting estimator and frozen-tail rule are unchanged across endpoints.

Endpoint	Output	Training loss	Selector ordering	Primary metric
Binary classification	1 logit	Class-weighted BCE	Positive-class probability	BA
Multiclass PAM50	4 logits	Class-weighted CE	Class probabilities, macro averaged	BA (macro recall)
Regression	Scalar linear output	MSE	Predicted continuous outcome	RMSE

D.5 Bridge Architecture

The executed bridge at candidate depth k is $G_k(x) = \mathrm{ReLU}(W_k x + b_k)$, mapping the target input to width $d_k \in \{64, 32, 16\}$. The ReLU adds no trainable parameters; the affine weight-plus-bias count is $d_k(d_T + 1)$. The completed capacity audit in Section G.8 replaces this bridge with low-rank and multi-layer alternatives while keeping the surrounding experimental cell matched; separate controls vary activation and weight decay.

D.6 Canonical Bridge Training and Endpoint Configurations

ADLG-Frozen trains only the bridge; all retained source hidden layers and the output head remain fixed. In the binary benchmark, AdamW, weight decay 10^{-4}, class-weighted `BCEWithLogitsLoss`, and gradient clipping at 5.0 are used. Target networks trained from scratch and target probes use learning rate 5×10^{-3}. The ordinary pair notebooks use batch size $\min(32, n)$; the uploaded engineering Census notebook uses $\min(8192, n)$. Training uses fixed epoch schedules rather than early stopping.

All candidate depths use the same neutral bridge initialization so that depth is not confounded with initialization. The theory and canonical experiments apply to this frozen-tail estimator. Supplement J documents a separate exploratory adaptation extension.

D.7 Endpoint Metrics and Classification Thresholds

Balanced accuracy is primary for binary and multiclass classification. The multiclass extension additionally reports macro-F1 and one-vs-rest macro-AUROC and average precision. RMSE is primary for regression, with MAE and R^2 secondary. Where binary score artifacts permit, we also report AUROC, class-1 PR-AUC, and minority-class recall.

In the canonical ADLG-Frozen analyses, balanced accuracy and minority-class recall are computed at the prespecified probability threshold 0.5. AUROC is threshold free, and PR-AUC is computed with `average_precision_score`, i.e. class-1 average precision. Minority recall is explicitly evaluated for the minority class, which need not be class 1 on every target. Precision/specificity and matched-operating- point analyses are separate audits where corresponding artifacts are available; any validation-selected operating point is chosen without held-out target-test labels.

D.8 Software, Hardware, and Reproducibility Metadata

Software, hardware, seeds, dataset hashes, and configuration details are reported from the experiment-specific artifacts where available. Historical metadata not established by those artifacts are reported as unavailable rather than inferred. The completed TabPFN v2, LATTLE, CARTE, and TABULA-8B runs were executed on Kaggle GPU T4×2 sessions with the environment pinned to 2026-07-01;

Table 17: Executed target-side training schedules. Epoch counts are fixed; the engineering Census schedule is used only for implementation auditing and does not replace the finalized Census 5×3×10 statistical protocol.

Pair group	Trained from scratch	Probe	Bridge epochs	Bridge LR
WDBC, Telco, Taiwan, Student	120	45	120	8×10^{-3}
HCV, eICU	290	80	220	8×10^{-3}
Engineering Census notebook	120	45	120	8×10^{-3}
HAM10000 → Derm7pt	120	45	120	8×10^{-3}

method-specific package versions and execution records are stated with the corresponding baseline in Appendix E. The HyperFast archive records two visible Tesla T4 devices, although the completed 200-cell benchmark ran serially on one device after the parallel launch check failed. We therefore report the hardware environment and measured per-cell times without attributing a two-GPU speedup.

CARTE execution. The official `carte-ai` package completed the full five-pair benchmark: 200/200 matched pair–split–budget cells, with no failures, in a Kaggle T4×2 environment. The fixed target-only configuration and measured runtime are reported with the CARTE results in Appendix E.4; no source rows or target-test outcomes enter fitting or selection.

E Baseline Implementations and Fidelity

Baseline comparisons control target splits and target-label budgets while allowing each method to operate according to its intended formulation.

An external method is reported under its published name only when its implementation status is explicit. For Census, we report an evaluated TransTab-Transfer configuration and a CATTLE-Supervised (reduced-schedule) configuration. Additional evaluated TransTab-Transfer and CATTLE-Supervised comparisons are shown for Telco Churn → Iranian Churn and Taiwan Credit → German Credit in the complete pair-specific results below. These configurations are labeled exactly as run and are not presented as exact reproductions of every published training detail. Linear HFA and XTab-style comparisons are reported on Telco and Taiwan under their exact executed configurations and paired directly to the saved ADLG cells. Progressive Alignment is positioned in related work. Under the paper's fidelity rule, numerical comparisons require a verified implementation and a completed protocol-compatible run; the available artifacts do not establish either condition for that method.

FLAT, PU-HUDA/PADA, TAROT, and Stitchable Neural Networks are therefore used for conceptual positioning rather than numerical comparison. Their evaluation contracts differ materially from ADLG's matched source–target protocol: FLAT meta-trains over a collection of tasks and generates a GAT for an unseen dataset (Zhu et al., 2023b); PU-HUDA/PADA assumes a positive-only labeled source and an unlabeled target and uses adversarial alignment (Mori et al., 2023); TAROT obtains an LLM-derived semantic graph from task descriptions and feature names (Shi et al., 2026); and vision stitching joins pretrained networks with compatible activation geometry on a common task (Bansal et al., 2021; Pan et al., 2023). None is omitted from the executed tables after an unfavorable result: no protocol-compatible run was performed, so reporting a numerical entry would conflate distinct supervision and pretraining assumptions.

E.1 Target-Only Controls and Contextual References

The standard three-arm protocol contains a target model trained from scratch, source-only- initialized ADLG, and ADLG with the optional GW–OT residual. The trained-from-scratch arm is the common mechanistic reference across the seven-pair notebooks. The finalized Census stress test additionally includes a separately tuned target- only MLP, a stronger Census-specific practical reference rather than a member of the standard three-arm protocol.

For the finalized Census benchmark, Appendix E.1 also reports regularized logistic regression and a boosted-tree model as well-tuned target-native tabular references. These models test target-only prediction rather than transfer of source knowledge, and provide a practical reference for the neural results.

Census neural comparison. ADLG-Frozen exceeds both neural controls in mean balanced accuracy at every budget. Across the prespecified 0.5–20% low-label range, the paired gains are 2.36

Table 18: Target-native comparison on the fixed five-pair protocol. ΔBA is the balanced-accuracy difference, in percentage points, from the inner-validation-selected boosted-tree reference. The displayed interval is a paired bootstrap summary over 50 pair–outer-split units after averaging budgets; it is descriptive across the five transfer pairs. W/T/L uses a 0.1-point tie threshold.

Method	Mean BA rank	ΔBA	95% interval	W/T/L
ADLG	**2.60**	**+1.67**	$[+0.01, +3.32]$	30/0/20
Target-only MLP (parameter-matched)	2.64	+1.77	$[+0.22, +3.35]$	32/1/17
CatBoost	2.98	+0.05	$[-0.50, +0.61]$	21/9/20
XGBoost	3.09	-1.65	$[-2.48, -0.86]$	17/4/29
Histogram GBM	3.70	-4.90	$[-5.86, -3.98]$	2/0/48

Table 19: ADLG minus the inner-validation-selected GBDT balanced accuracy, in percentage points. Positive values favor ADLG. Each budget-specific entry averages the ten matched outer splits.

Pair	20 labels	50 labels	100 labels	200 labels	Mean
Census-Income → Adult	+4.46	+3.44	+2.03	+1.94	+2.97
HCV → ILPD	-4.06	+3.27	+5.72	+8.07	+3.25
Student Dropout → OULAD	-5.03	-6.34	-5.72	-5.82	-5.73
Telco Churn → Iranian Churn	+13.89	+10.86	+8.08	+7.41	+10.06
WDBC → Mammographic Masses	-0.97	-3.35	-2.71	-1.84	-2.22

points over the trained-from-scratch model (95% outer-split bootstrap CI: 2.02–2.72) and 1.88 points over the tuned target-only MLP (1.63–2.17). All 150 locked split–repeat–budget cells and all 750 method evaluations completed without failed-run substitution.

Secondary non-neural context. The non-neural tabular references are reported only in Appendix E.1. Across the prespecified low-label range, ADLG-Frozen is 0.98 points below regularized logistic regression and 1.39 points below the boosted-tree reference. These results are retained as contextual evidence and are not used to define the primary neural representation-transfer claim.

E.2 Five-Pair Target-Native Boosted-Tree Benchmark

We additionally evaluate CatBoost, XGBoost, and histogram gradient boosting (GBM) on the fixed five-pair protocol. These target-native baselines use the exact outer target splits and labeled-target subsets used by ADLG. They use only the numerical target representation: no source data, unlabeled target data, ADLG outputs, or outer-test outcomes are used for fitting or selection. Within every pair, outer split, and label budget, each tree family selects between two prespecified configurations using up to three-fold stratified cross-validation on the current labeled target subset. The final comparison reference is the tree family with the highest inner-validation balanced accuracy in that matched cell.

Aggregate comparison. Table 18 averages the four absolute label budgets within each pair–outer-split unit. ADLG has the best mean balanced-accuracy rank among the five evaluated methods and averages 1.67 points above the inner-validation-selected boosted-tree reference. This cross-pair aggregate has a paired interval of $[+0.01, +3.32]$ points, which narrowly excludes zero under the executed bootstrap. It remains descriptive because repeated splits within a pair overlap and the benchmark contains only five transfer pairs. The pair-level results in Table 19 show the relevant heterogeneity.

The advantage is strongest and consistent across budgets on Telco→Iranian. ADLG is also favorable overall on Census→Adult and HCV→ILPD, whereas target-native trees are stronger on Student Dropout→OULAD and WDBC→Mammographic Masses. Thus, the benchmark supports practical competitiveness rather than uniform dominance.

Minority-class sensitivity. Figure 6 reports balanced accuracy (solid) and minority recall (dashed) separately. Selection uses inner-validation balanced accuracy only. Across matched cells, ADLG improves mean minority recall by 3.19 points relative to the parameter-matched target-only MLP and by 18.45, 23.86, and 31.87 points relative to CatBoost, XGBoost, and GBM, respectively. These are secondary held-out outcomes, not separately optimized selection objectives.

Configurations and selection. CatBoost uses depth $\{4, 6\}$, learning rate 0.05, 300 iterations, and ℓ_2 leaf regularization $\{3, 5\}$. XGBoost uses maximum depth $\{3, 4\}$, learning rate 0.05, 200 estimators,

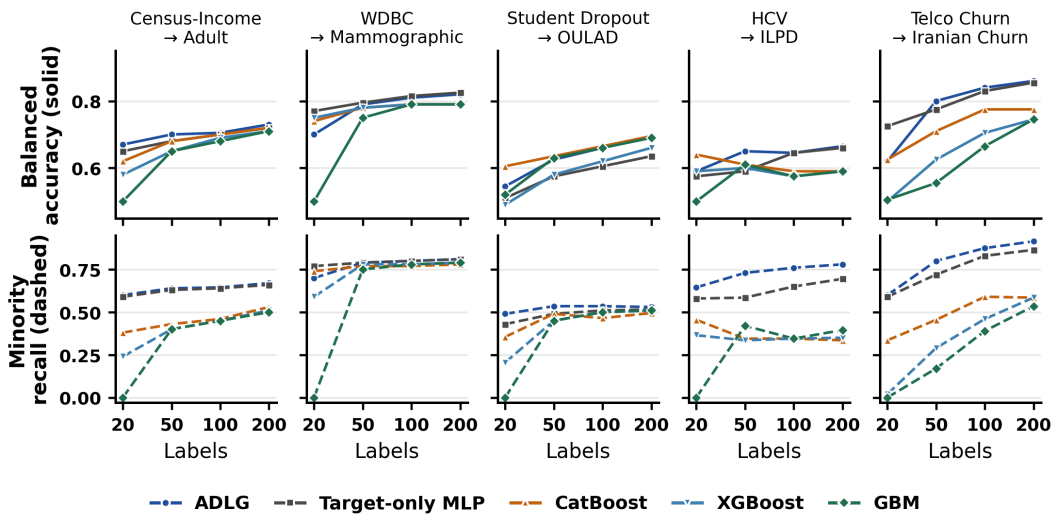

Figure 6: Target-native baselines on the fixed five-pair protocol. Solid curves show balanced accuracy and dashed curves show minority recall. All methods use the same outer splits and labeled subsets. Tree configurations are selected using inner-validation balanced accuracy within the current labeled subset.

and minimum child weight $\{1, 3\}$. Histogram GBM uses 200 iterations and learning rate 0.05, with maximum leaf nodes $\{15, 31\}$ and ℓ_2 regularization $\{0, 0.1\}$. This is a controlled lightweight search rather than exhaustive optimization of each tree family. When the current labeled subset contains fewer than two minority examples, stratified validation is unavailable; the first prespecified configuration is fitted and the selection status is recorded.

Uncertainty and scope. For a fixed pair and budget, uncertainty is computed by paired bootstrap over the ten matched outer splits. The aggregate interval in Table 18 instead resamples pair–split summaries after averaging budgets. Because repeated splits within a pair overlap and the benchmark contains only five heterogeneous transfer pairs, that aggregate is reported descriptively. Although the displayed ADLG interval narrowly excludes zero under the executed resampling scheme, we do not treat it as population-level evidence of uniform improvement on unseen tasks. CatBoost receives the same numerical benchmark representation as the other methods; categorical-feature-specific processing is not enabled.

E.3 Direct ADLG–Target MLP Comparison

Table 20 reports the five pair-specific ADLG–MLP contrasts and, for completeness, an equal-pair descriptive aggregate. Pair-specific intervals resample the ten matched outer splits as clusters, retaining all four budgets within a sampled split. The final row weights the five pair means equally and uses a hierarchical bootstrap over transfer pairs and outer splits. Because the benchmark contains only five heterogeneous transfer pairs, this aggregate is a descriptive summary rather than evidence of equivalence or a common cross-task treatment effect. Its interval indicates no clear aggregate difference. W/T/L counts use a 0.1-point tie threshold over the 40 matched pair–budget cells per row.

E.4 TabPFN, HyperFast, CARTE, and LATTLE Five-Pair Comparisons

All four methods use the same five target datasets, ten saved outer splits, and four absolute target-label budgets (20, 50, 100, and 200) as the five-pair ADLG benchmark. All 200 matched cells completed for each method. Comparison files join to ADLG by pair, outer split, and budget. Target-test outcomes and ADLG predictions are excluded from fitting and selection.

TabPFN v2 provenance and protocol. We use TabPFN v2 (Hollmann et al., 2025) (`tabpfn==2.0.9`) through the official `TabPFNClassifier` interface, with eight estimators and the prespecified threshold of 0.5. Each cell is fitted only on its saved labeled target subset; source data and unlabeled target examples are not used. The captured runtime records all 200 completed cells. The reported archive is the direct output of a Kaggle GPU T4×2 session with the environment pinned

Table 20: Paired ADLG minus parameter-matched target-only MLP balanced accuracy, in percentage points. Intervals are 95% cluster-bootstrap intervals over outer splits within each pair; the final row also resamples pairs. W/T/L uses a 0.1-point tie threshold.

Transfer pair	Δ BA [95% CI]	W/T/L
Census-Income→Adult	$+0.65$ $[+0.34, +0.97]$	29/2/9
HCV→ILPD	$+1.52$ $[+0.07, +3.06]$	23/2/15
Student Dropout→OULAD	-0.14 $[-0.46, +0.15]$	17/6/17
Telco Churn→Iranian Churn	$+0.16$ $[-0.80, +1.25]$	21/1/18
WDBC→Mammographic Masses	-2.68 $[-4.20, -1.15]$	11/0/29
Equal-pair aggregate	-0.10 $[-1.62, +1.19]$	101/11/88*

* W/T/L in the aggregate row counts the 200 matched pair–budget cells; the aggregate interval instead uses the pair/split hierarchical bootstrap described above.

to 2026-07-01; the model itself runs on one visible GPU and the configuration is fixed across cells. The earlier template phrase "not Kaggle-tested" is not the provenance of these completed outputs and has been corrected in the execution manifest. Standardization is fit only on the current labeled target subset. No feature is truncated or dropped; a cell would fail rather than silently substitute an unsupported representation.

LATTLE reimplementation. LATTLE (Kowsar et al., 2025) is a two-stage source-to-target transplant method. The executed reimplementation follows the paper's design. Its source stage trains DistilGPT2 with a gated FT-Transformer using at most 10,000 source rows and five epochs; a seeded 10% source validation split selects the source checkpoint. For each target cell, the transferred key/value projections initialize the first target-transformer layer and remain frozen, while the target stage uses only the matched labeled subset. Fixed, predeclared hyperparameters replace the paper's 100-trial Optuna search so that the smallest target budgets remain fully available for training. The run uses fixed target epochs and a reduced source schedule, with one worker assigned to each of two visible Tesla T4 GPUs.
Source rows are serialized as feature-name/value text and tokenized individually with left truncation at 1,024 tokens. The saved audit records the fraction of source rows whose encoded length reaches this maximum: 1.0 for Census-Income (511 source features) and Student Dropout (238), and 0.0 for WDBC (30), HCV (12), and Telco (45). Thus, the Census and Student results measure the transplant under a fixed source context budget. The target stage does not place labeled examples in an LLM prompt, so an effective-shot field does not apply to this method.

HyperFast provenance and protocol. We use the official HyperFast classifier (Bonet et al., 2024) (`hyperfast==1.0.6`) as a target-only pretrained baseline. Every cell uses the saved labeled target subset, split-local standardization, eight ensemble members, stratified sampling, and the fixed 0.5 threshold. The configuration was declared once for the benchmark; no target hyperparameter search or test-set selection was performed. Source data and unlabeled target examples are not used. The released result files contain all 200 completed cells and their predictions, with no failed cells. The run used a single Tesla T4 serially; the second visible T4 did not contribute to the completed benchmark. Completion fields, cell records, predictions, and the worker log document the completed Kaggle execution.

CARTE provenance and protocol. We use CARTE (Kim et al., 2024) through the official `carte-ai` package (version 0.0.26) and its released pretrained `Table2GraphTransformer` checkpoint. In this benchmark CARTE is a target-only pretrained baseline, not a cross-table transfer run: each cell uses only the saved labeled target subset, with split-local preprocessing, and does not use the source table, unlabeled target rows, or target-test outcomes. We fix `num_model=10`, the package seed, and all other settings before examining the results. The run completed all 200 cells without failure. This comparison therefore tests whether target-only tabular pretraining can recover the benefit of ADLG's source-to-target interface transfer under the same label budget and outer split. Across the 20 pair–budget cells, paired 95% intervals favor ADLG over TabPFN in 11 comparisons, favor TabPFN in five, and include zero in four. For the LATTLE reimplementation, nine favor ADLG, three favor LATTLE, and eight include zero. The pair-specific intervals first average the four budget differences within each outer split and then resample the ten paired outer splits. The five-pair averages are descriptive summaries of these evaluated tasks.

Table 21: Balanced accuracy and minority recall averaged over ten matched outer splits for each pair and absolute target-label budget. Recall uses the prespecified threshold of 0.5. Values are means.

Pair	Labels	Balanced accuracy			Minority recall		
		ADLG	TabPFN	LATTLE	ADLG	TabPFN	LATTLE
Census-Income→Adult	20	.667	.556	.660	.628	.149	.481
	50	.699	.658	.688	.661	.398	.533
	100	.709	.682	.701	.649	.442	.543
	200	.730	.736	.719	.669	.547	.586
HCV→ILPD	20	.586	.583	.541	.644	.309	.362
	50	.652	.529	.576	.728	.132	.418
	100	.651	.556	.574	.760	.191	.409
	200	.669	.543	.608	.778	.171	.438
Student Dropout→OULAD	20	.528	.542	.547	.491	.111	.330
	50	.574	.542	.589	.534	.146	.401
	100	.608	.638	.625	.539	.322	.466
	200	.631	.701	.653	.532	.487	.501
Telco Churn→Iranian Churn	20	.726	.596	.650	.600	.215	.441
	50	.804	.693	.757	.775	.417	.591
	100	.843	.740	.818	.876	.518	.704
	200	.852	.809	.838	.915	.655	.735
WDBC→Mammographic Masses	20	.721	.778	.725	.703	.706	.703
	50	.766	.821	.762	.756	.785	.743
	100	.776	.821	.783	.769	.808	.773
	200	.788	.829	.772	.787	.810	.753

Table 22: Matched ADLG–HyperFast balanced accuracy. Budget columns report ADLG/HyperFast means over ten outer splits. The final column is the pair-average ADLG-minus-HyperFast difference in percentage points with a 95% cluster-bootstrap interval over outer splits; all four budgets are retained within each resampled split (20,000 resamples; seed 60718).

Pair	20 labels	50 labels	100 labels	200 labels	ΔBA [95% CI]
Census-Income→Adult	.667/.650	.699/.716	.709/.753	.730/.775	-2.19 $[-2.81, -1.57]$
HCV→ILPD	.586/.583	.652/.628	.651/.642	.669/.644	$+1.55$ $[+0.19, +2.96]$
Student Dropout→OULAD	.528/.555	.574/.593	.608/.627	.631/.632	-1.64 $[-2.74, -0.75]$
Telco Churn→Iranian Churn	.726/.707	.804/.781	.843/.799	.852/.808	$+3.24$ $[+1.06, +5.38]$
WDBC→Mammographic Masses	.721/.787	.766/.799	.776/.796	.788/.782	-2.85 $[-3.75, -1.90]$
Equal-pair mean	ADLG .699; HyperFast .703				-0.38 (descriptive)

At the pair–budget level, paired intervals favor HyperFast in nine of the 20 cells, favor ADLG in four, and include zero in seven. The pair-average results show where the methods differ: ADLG is higher on HCV and Telco, whereas HyperFast is higher on Census, Student, and WDBC. The small equal-pair mean difference therefore summarizes heterogeneous task-level outcomes and is not treated as evidence of a uniform ordering.

ADLG has the higher mean in 14 of the 20 pair–budget cells. The cell-wise paired bootstrap intervals favor ADLG in 11 cells, CARTE in five, and include zero in four. The largest pair-average differences favor ADLG on Telco, Census, and HCV. CARTE is higher on Student at 50–200 labels and on WDBC at 50–200 labels, showing that its target-only pretraining is effective on selected targets even though it does not match ADLG's equal-pair average.

E.5 TABULA-8B Three-Pair Target-Only Comparison

TABULA-8B (Gardner et al., 2024) is evaluated as a frozen, target-only few-shot classifier on three complete matched pairs. The targets are HCV→ILPD, Telco Churn→Iranian Churn, and WDBC→Mammographic Masses. The study contains 120 completed cells (three pairs, ten outer splits, and four budgets). A single capped Census smoke-test cell is excluded from quantitative summaries, and Student Dropout→OULAD was not run. Accordingly, we use this experiment as a focused three-pair check and do not form a five-pair TABULA-8B rank.

The executed configuration uses the frozen `mlfoundations/tabula-8b` model, raw finalized target values serialized as text, an 8,192-token context, seeded class-interleaved shot order, and probabilities from the next-token logits for labels 0 and 1. Examples come only from the saved labeled target subset. Source data, unlabeled target data, ADLG outputs, and target-test outcomes are excluded from fitting and selection. The context policy retains as many complete labeled examples as fit and records the resulting k_{used} for every cell.

Table 23: Matched ADLG–CARTE balanced accuracy. Each budget entry is ADLG/CARTE averaged over ten matched outer splits. The final column is the pair-average ADLG-minus-CARTE difference in percentage points with a 95% cluster-bootstrap interval over outer splits; all four budgets remain grouped within a resampled split (20,000 resamples; seed 60718).

Pair	20 labels	50 labels	100 labels	200 labels	ΔBA [95% CI]
Census-Income→Adult	.667/.529	.699/.594	.709/.659	.730/.706	$+7.93$ $[+6.35, +9.52]$
HCV→ILPD	.586/.563	.652/.551	.651/.585	.669/.564	$+7.39$ $[+5.44, +9.36]$
Student Dropout→OULAD	.528/.510	.574/.594	.608/.644	.631/.666	-1.79 $[-3.12, -0.46]$
Telco Churn→Iranian Churn	.726/.625	.804/.578	.843/.685	.852/.761	$+14.39$ $[+12.66, +16.11]$
WDBC→Mammographic Masses	.721/.688	.766/.802	.776/.811	.788/.817	-1.68 $[-3.85, +0.47]$
Equal-pair mean	ADLG .699; CARTE .647				$+5.25$ (descriptive)

Table 24: LATTLE source rows reaching the 1,024-token maximum in the executed source-stage audit. Feature counts exclude the source label.

Source dataset	Source features	Fraction at maximum
Census-Income	511	1.00
Student Dropout	238	1.00
WDBC	30	0.00
HCV	12	0.00
Telco Churn	45	0.00

At the fixed operating threshold, ADLG has higher pair-averaged minority recall on HCV and Telco, while TABULA-8B has higher recall on WDBC. Averaging the three pair means equally gives an ADLG-minus-TABULA-8B difference of 15.6 points. The corresponding matched three-pair BA means are 0.736 for ADLG and 0.714 for TABULA-8B, a $+2.20$-point ADLG difference. This focused result complements the complete five-pair TabPFN and LATTLE studies by testing a large frozen few-shot model under the same saved target partitions.

E.6 CATTLE-Supervised (2026): executed reimplementation

We compare with CATTLE-Supervised (Akhter et al., 2026), the supervised configuration of the method introduced by Akhter et al. The result is an *executed reimplementation*, not an official authors' run. It uses the saved target splits and label budgets of the present study and a fixed, predeclared reduced training schedule. We therefore label the reported configuration as *CATTLE-Supervised (2026, reduced-schedule reimplementation)* and do not represent it as a reproduction of the full published tuning schedule.

On Census-Income → Adult, ADLG has higher balanced accuracy and minority recall across the displayed supervision range. CATTLE-Supervised closes much of the gap on AUROC and PR-AUC as target supervision increases, so the result supports a class-balanced and minority-sensitivity advantage rather than uniform metric dominance.

E.7 TransTab

The Census comparison uses the evaluated *TransTab-Transfer* configuration under the same Adult target-label budgets. The method is reported under this implementation label to avoid implying identity with every possible TransTab pretraining/fine-tuning schedule.

Across Census budgets, ADLG maintains higher balanced accuracy and minority recall. TransTab-Transfer becomes competitive or stronger on AUROC and PR-AUC over parts of the supervision range. This metric split is retained in full rather than collapsed into a single win/loss summary.

E.8 Matched HFA and XTab-Style Comparisons

We compare ADLG with source-subsampled linear HFA and a single-source, XTab-motivated pretraining setup (Zhu et al., 2023a) on Telco→Iranian and Taiwan→German. Each comparison contains 10 matched repeats and seven nested target-label budgets; methods share the target test partition and labeled-target indices cell by cell. Target-test observations are excluded from preprocessing, tuning, and training. Lines in Figures 7 and 8 are repeat means, and shaded regions are 95% bootstrap confidence intervals over repeats.

On Telco→Iranian, ADLG exceeds linear HFA in balanced accuracy, PR-AUC, and minority recall at all seven budgets and in AUROC at six; relative to XTab-style pretraining, it improves AUROC and

Table 25: TABULA-8B three-pair comparison. Values average ten matched outer splits. k is the effective number of in-context labeled examples; an asterisk marks a context-limited budget. Recall uses threshold 0.5.

Pair	Labels	k	Balanced accuracy		Minority recall	
			ADLG	TABULA-8B	ADLG	TABULA-8B
HCV→ILPD	20	20	.586	.622	.644	.465
	50	50	.652	.548	.728	.179
	100	66^*	.651	.671	.760	.709
	200	66^*	.669	.669	.778	.815
Telco Churn→Iranian Churn	20	20	.726	.593	.600	.227
	50	50	.804	.620	.775	.277
	100	69^*	.843	.688	.876	.427
	200	69^*	.852	.845	.915	.877
WDBC→Mammographic Masses	20	20	.721	.820	.703	.800
	50	50	.766	.834	.756	.804
	100	100	.776	.833	.769	.802
	200	165^*	.788	.825	.787	.837

Table 26: Low-label paired effects in percentage points. For each repeat, the effect is averaged over the prespecified 5%, 10%, and 20% target-label budgets; brackets are 95% bootstrap confidence intervals over the 10 repeat- level averages. Intervals containing zero are treated as inconclusive.

Pair	Metric	ADLG − linear HFA	ADLG − XTab-style
Telco→Iranian	Balanced accuracy	$+6.60\ [+4.50, +8.74]$	$+0.83\ [-0.15, +1.90]$
	Minority recall	$+15.15\ [+8.96, +21.21]$	$-1.62\ [-3.57, +0.24]$
	AUROC	$+1.75\ [+1.37, +2.20]$	$+1.98\ [+1.39, +2.54]$
	PR-AUC	$+4.57\ [+3.51, +5.74]$	$+6.72\ [+4.90, +8.85]$
Taiwan→German	Balanced accuracy	$+1.84\ [+0.67, +3.23]$	$+3.50\ [+1.16, +6.12]$
	Minority recall	$-0.61\ [-3.28, +2.28]$	$+9.22\ [+2.89, +16.00]$
	AUROC	$+4.00\ [+2.68, +5.26]$	$+5.74\ [+3.20, +8.76]$
	PR-AUC	$+4.46\ [+2.85, +6.13]$	$+5.48\ [+2.59, +8.74]$

PR-AUC at all seven budgets and balanced accuracy at five. Taiwan→German is more heterogeneous: the aggregated low-label effects in Table 26 favor ADLG on balanced accuracy, AUROC, and PR-AUC against both configurations, while several higher-budget comparisons reverse. We therefore claim broad Telco gains and a Taiwan low-label advantage, not uniform superiority.

HFA is the executed linear formulation with per-domain weighting and 1,000 source rows (of 7,043 Telco or 29,965 Taiwan rows). XTab-style uses a two-layer, width-64 Transformer pretrained on the full source dataset (20 Telco epochs or 10 Taiwan epochs); it is not official multi-table XTab. These labels prevent the matched comparisons from being interpreted as unrestricted reproductions of either broader method family.

E.9 Baseline Fidelity and Comparison Scope

The external comparison is reported under the configurations actually executed, and Table 27 records the scope of every configuration. Progressive Alignment is positioned in related work under the same fidelity rule: a numerical comparison requires a verified implementation and a completed protocol-compatible run. This separation keeps the reported evidence directly traceable to the saved experimental artifacts.

F Developmental and Noncanonical Variants

Internal exploratory variants considered during method development are not used as published baselines or as evidence for the primary claims unless a frozen, reproducible implementation and result artifact is available. We therefore omit unfinished Raw-GW, FGW-adapter, GW-token-transfer, and CCGR templates from the quantitative appendices rather than presenting planned settings as completed experiments.

F.1 Bridge-Initialization Priors

The following mechanisms are optional bridge-initialization priors. Neither is assumed to establish semantic feature correspondence or guarantee improved target risk.

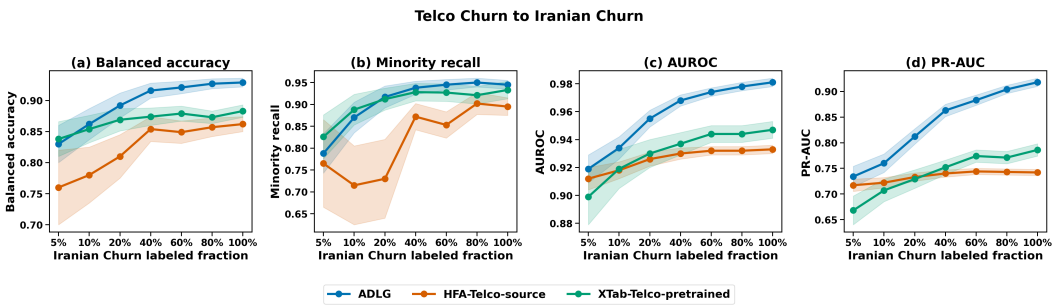

Figure 7: Telco Churn→Iranian Churn matched comparison across balanced accuracy, minority recall, AUROC, and PR-AUC. ADLG shows broad gains over linear HFA and consistently improves ranking metrics over XTab-style pretraining; minority-recall differences against XTab-style pretraining are mostly inconclusive. Curves are means over 10 matched repeats and shaded regions are 95% bootstrap confidence intervals over repeats.

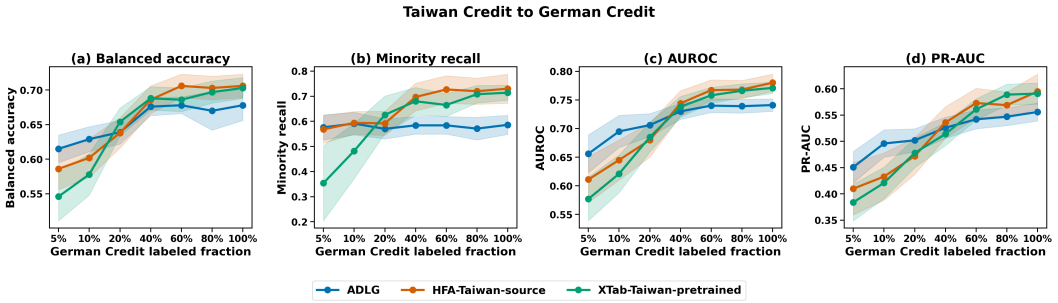

Figure 8: Taiwan Credit→German Credit matched comparison across balanced accuracy, minority recall, AUROC, and PR-AUC. ADLG is strongest or competitive in the low-label regime, while HFA and XTab-style pretraining match or exceed it on several higher-budget outcomes. Curves are means over 10 matched repeats and shaded regions are 95% bootstrap confidence intervals over repeats.

F.2 Structural GW Initialization

For standardized target features X_T, define

$$C_T[a,b] = 1 - \left|\mathrm{corr}\left(X_{T,:,a}, X_{T,:,b}\right)\right|. \tag{68}$$

For source hidden activations $A_S^{(k)}$, define

$$C_S[u,v] = 1 - \left|\mathrm{corr}\left(A^{(k)}_{S,:,u}, A^{(k)}_{S,:,v}\right)\right|. \tag{69}$$

The entropically regularized GW coupling is

$$T^\star = \arg\min_{T\in\Pi(p,q)} \sum_{u,v,a,b} \left(C_S[u,v] - C_T[a,b]\right)^2 T_{ua} T_{vb} - \varepsilon H(T). \tag{70}$$

The resulting coupling is transformed into bridge orientation and scaled according to the frozen implementation, producing

$$W_0^{\mathrm{GW}}.$$

The executed notebooks use uniform marginals, square loss, and entropic Gromov–Wasserstein with $\varepsilon = 0.05$ through the POT implementation with the PGD solver. The initial coupling is $G_0 = pq^\top$; the outer maximum is 500 iterations with tolerance 10^{-9}, and the internal solver is allowed up to 5,000 iterations with stopping threshold 10^{-10}. The implementation stabilizes correlation-derived structures, balances the resulting coupling to the requested marginals, transposes it into target-to-hidden orientation, row normalizes it, and scales the bridge initializer by $\sqrt{d_k}$. Zero-variance handling is performed in the correlation construction so that non-finite entries do not propagate. These operations define an initialization heuristic only and are not interpreted as semantic feature matching.

Table 27: Fidelity and scope of the executed baseline comparisons.

Method	Quantitative status	Target protocol	Fidelity/reporting scope
ADLG	Executed	Reference protocol	Paper method; full implementation described here
Target-only MLP / trained-from-scratch	Executed	Matched	Target-only controls using the same target splits/budgets
TabPFN v2 target-only	Executed	Five matched pairs	`tabpfn==2.0.9`; official `TabPFNClassifier`; fixed eight-estimator configuration; labeled target rows only
HyperFast target-only	Executed	Five matched pairs	`hyperfast==1.0.6`; official classifier; fixed eight-member ensemble; labeled target rows only; 200/200 cells completed serially on one T4
LATTLE transplant	Executed reimplementation	Five matched pairs	Fixed hyperparameters and reduced source schedule; source context audit in Appendix E.4
TABULA-8B target-only	Executed subset	Three matched pairs	Frozen few-shot configuration; effective shots recorded; focused scope in Appendix E.5
CARTE target-only	Executed	Five matched pairs	`carte-ai==0.0.26`; official pretrained checkpoint; fixed ten-model configuration; labeled target rows only; 200/200 cells completed
TransTab-Transfer	Executed	Matched where reported	Implementation-specific configuration; not claimed as an exact reproduction of every published schedule
CATTLE-Supervised (2026)	Executed reimplementation	Matched where reported	Published supervised method; fixed, predeclared reduced schedule; not an official run or a full tuning-schedule reproduction
Linear HFA	Executed on Telco and Taiwan	Matched	Source-subsampled configuration using 1,000 source rows; per-domain weighting
XTab-style source pretraining	Executed on Telco and Taiwan	Matched	Single-source two-layer Transformer; not official multi-table XTab

Table 28: Scope of developmental variants. Canonical claims use ADLG-Frozen only; these variants test whether richer initialization or adaptation changes that controlled reference.

Variant	Purpose	Canonical?	Used in main claims?
Structural GW initializer	Geometry-based bridge initialization	No	No
Source-only initializer	Source reconstruction prior	No	No
Residual combination	Combine structural and source priors	No	No
Two-phase tail adaptation	Test controlled unfreezing	No	No

F.3 Source-Only Self-Supervised Prior

Let Ω_j denote a synthetic source corruption.
The source reconstruction objective at depth k has the form

$$\mathcal{L}_j = \mathbb{E}_{x\sim\mathcal{D}_S}\left[\left\|G_j(\Omega_j(x)) - H_S^{(k)}(x)\right\|_2^2\right]. \tag{71}$$

A source-trained initializer or hypernetwork is conditioned at transfer time on a permutation-invariant descriptor of the permissible target training pool and the selected depth.

Table 29: Exploratory ADLG-2P audits. Census reports the mean gain over the tuned target-only MLP across 0.5–20% labels; WDBC reports ADLG-2P minus ADLG-Frozen at matched selected depth.

Audit	Paired effect	95% CI
Census	$+2.12$ BA pt	$[+1.92, +2.36]$
WDBC	-0.24 BA pt	$[-0.44, -0.09]$

The permutation-invariant descriptor uses six per-feature statistics: mean, standard deviation, skewness, kurtosis, and the 10th and 90th percentiles. Feature statistics are processed by a $6 \to 24 \to 24$ ReLU token network; pooled information feeds a $48 \to 24 \to 8$ network, yielding a rank-8 latent initializer with depth-specific output mappings for widths 64, 32, and 16. Synthetic source corruptions include feature-subset corruptions with Gaussian noise of standard deviation 0.05 and random-projection corruptions using a $\tanh$ transformation plus noise with scale sampled approximately from 0.02 to 0.15. WDBC, Telco, Taiwan, and Student use six corruptions (four meta-train, two holdout) and six meta epochs; HCV and eICU use 42 corruptions (36 meta-train, six holdout) and 25 meta epochs. The meta optimizer is Adam with learning rate 2×10^{-3} and weight decay 10^{-5}. At transfer time the hypernetwork consumes only the permissible target-training-pool descriptor and selected depth; no target-test information is used.

F.4 Residual Source-Trained Correction

When structural and source-trained priors are combined,

$$W_0 = W_0^{\mathrm{GW}} + \Delta_\phi. \tag{72}$$

The residual Δ_ϕ is trained through the source-side reconstruction/meta-training mechanism.
Because the residual class contains $\Delta_\phi = 0$, the optimum cannot increase the source reconstruction objective relative to W_0^{GW} alone.
This does not imply lower target risk.

F.5 Exploratory Two-Phase Fine-Tuning

ADLG-2P is a historical, noncanonical variant that jointly adapts the bridge, retained hidden layers, and output head after frozen Phase 1 training. It was not evaluated across the full benchmark and is excluded from the primary evidence chain. Table 29 reports the two available audits without pooling them with ADLG-Frozen.
The WDBC audit shows no Phase-2 benefit: lower-rate joint adaptation reduced balanced accuracy by 0.24 points relative to the matched frozen estimator. The favorable historical Census result is pair-specific and does not justify promoting tail adaptation to the canonical method. We therefore retain ADLG-Frozen as the sole method used for the main claims.

G Estimator Ablations and Source Controls

The final ablation package separates attachment-depth effects from source knowledge, ordering information, unlabeled-target access, initialization, bridge capacity, and retained-tail adaptation.

G.1 Adaptive-Depth Ablation

The Census depth ablation compares every fixed candidate depth with random, CKA-based, and validation-based selection using the same neutral bridge initialization. The retrospective test oracle is included only as a diagnostic reference and is never used to select a model.
For metric M, define

$$\mathrm{Regret}_M = M(k_{\mathrm{oracle}}) - M(k_{\mathrm{CKA}}). \tag{73}$$

The tables below report CKA–oracle agreement, selector regret, and paired comparisons with the deployable depth policies. Oracle results remain retrospective and are not treated as achievable deployment performance.

G.2 Wider-Range Depth Ladder (Exploratory)

Section G.2 introduces an exploratory extension of the Census depth ablation above to a source network and target probe with six hidden layers, widths $256, 128, 64, 32, 16, 8$ (a $32\times$ range, versus

Table 30: Census balanced accuracy by neutral-initialization attachment policy. †The test oracle is retrospective and diagnostic only.

Labels	CKA	Shallow	Middle	Deep	Random	Validation	Oracle†
0.5%	0.733	0.737	0.730	0.725	0.730	0.730	0.740
1%	0.750	0.756	0.751	0.749	0.756	0.756	0.760
2%	0.775	0.770	0.770	0.778	0.771	0.775	0.780
5%	0.790	0.781	0.786	0.791	0.786	0.789	0.793
10%	0.801	0.794	0.800	0.800	0.799	0.799	0.802
20%	0.807	0.803	0.807	0.808	0.806	0.808	0.809
40%	0.811	0.807	0.814	0.815	0.813	0.815	0.816
60%	0.816	0.810	0.816	0.818	0.815	0.817	0.819
80%	0.818	0.814	0.819	0.820	0.818	0.820	0.820
100%	0.819	0.813	0.818	0.820	0.816	0.819	0.821

Table 31: Census CKA selector quality. "Within-1-point" is the prespecified diagnostic criterion $M(k_{\text{oracle}}) - M(k_{\text{CKA}}) \leq 0.01$ and is never used for selection.

Labels	Oracle hit	Within 1 pt	Mean regret (pt)	Median regret (pt)	CKA margin
0.5%	0.467	0.733	0.70	0.59	0.018
1%	0.267	0.733	1.01	0.47	0.022
2%	0.600	0.800	0.48	0.00	0.014
5%	0.667	0.733	0.40	0.00	0.026
10%	0.667	1.000	0.17	0.00	0.035
20%	0.333	1.000	0.24	0.16	0.041
40%	0.267	0.867	0.46	0.33	0.040
60%	0.600	0.933	0.26	0.00	0.028
80%	0.467	1.000	0.21	0.06	0.016
100%	0.400	1.000	0.21	0.17	0.060

the finalized ladder's $4\times$ range). It retains the neutral (Xavier) bridge initializer at every candidate depth and the same quantile-aligned CKA construction (Appendix D.4) applied over 6 candidate depths instead of 3, but uses a distinct source/probe architecture, candidate bank, repeat count, budget grid, and noncanonical training schedule documented in Supplement J. It therefore supports only descriptive conclusions about depth effects, CKA's imperfect oracle agreement, and constant-width behavior; it is not used for canonical selector accuracy, cost, or method comparisons.

Protocol differences from the finalized statistical protocol. This extension uses 3 outer-split repeats (rather than 5) and 8 target-label budgets, $\{0.5, 1, 2, 5, 10, 20, 40, 100\}\%$ (rather than 10), to keep the pilot tractable. We report it as an exploratory result and do not claim outer-split-level confidence intervals for it; the statistical procedures of Appendix J should be applied before treating any of the following numbers as final.

Selector diagnostics. Table 35 reports the same selector-quality diagnostics as Table 31 for the six-depth ladder. Exact CKA–oracle agreement (4.2% overall) is low, and at several budgets no repeat selects the exact oracle depth; with only 3 repeats per budget (24 repeat–budget cells overall), this exact-match statistic is estimated from very few trials and should not be read as a precise agreement rate – a 6-way uniform-random selector would itself be expected to match the oracle in roughly 16.7% of cells, so an observed rate of 4.2% over 24 trials is consistent with considerable sampling noise rather than necessarily indicating below-chance selection. The within-one-point rate (87.5% overall) and mean regret (0.59 points overall) are comparable in magnitude to the finalized three-depth ladder (88.0%/0.41 points), which we read as the more informative comparison given how closely several candidate depths perform at most budgets (Table 32).

CKA-selected versus each fixed depth. Table 36 reports the paired mean difference in balanced accuracy between the CKA-selected depth and every fixed depth, averaged across all repeat–budget cells – the comparison most directly relevant to whether adaptive selection outperforms simply committing to one depth. In this pilot, CKA-selected performance is within ± 0.5 points of every fixed depth on average, i.e. statistically indistinguishable from several reasonable fixed policies at the current repeat count.

Table 32: Six-depth Census ladder, balanced accuracy (exploratory, 3 repeats; not part of the finalized statistical protocol). †Retrospective, diagnostic only, never used for model selection.

Labels	D1 (256)	D2 (128)	D3 (64)	D4 (32)	D5 (16)	D6 (8)	CKA	Oracle†	Trained from scratch
0.5%	0.732	0.755	0.762	0.767	0.763	0.762	0.763	0.767	0.732
1%	0.752	0.772	0.778	0.779	0.778	0.774	0.769	0.781	0.719
2%	0.773	0.781	0.783	0.787	0.787	0.784	0.784	0.788	0.746
5%	0.794	0.795	0.792	0.791	0.791	0.794	0.792	0.797	0.771
10%	0.804	0.802	0.799	0.798	0.796	0.802	0.801	0.804	0.797
20%	0.812	0.808	0.803	0.801	0.800	0.805	0.805	0.812	0.812
40%	0.818	0.814	0.812	0.812	0.809	0.813	0.812	0.818	0.807
100%	0.819	0.821	0.816	0.817	0.818	0.819	0.818	0.821	0.808

Table 33: Modal retrospective-oracle attachment depth by budget across 3 repeats (six-depth ladder). Depth 1 = width 256 (widest), depth 6 = width 8 (narrowest). The oracle depth moves toward the widest bridge (depth 1) in the 10–40% range and toward depth 2 at 100%.

Labels	0.5%	1%	2%	5%	10%	20%	40%	100%
Modal oracle depth	4	3	4	2	1	1	1	2

Interpretation of the widened ladder. This extension establishes that attachment depth has a materially larger effect on balanced accuracy once the candidate range is widened (up to $\sim$3.5 points at the smallest budget between the best and worst fixed depth, versus $\lesssim$1 point on the finalized three-depth ladder), directly supporting the premise that depth selection is a real problem worth solving. Because the widened ladder uses development-scale repeats and does not cross source condition, ordering, and unlabeled-target access, we use it only as supporting evidence that the candidate range can matter. The paper's primary mechanism claims rely on the finalized constant-width and source-condition controls reported below.

G.3 Source-Condition Controls

The ADLG-Frozen Census audit separates source-label information from the frozen architecture under matched attachment depth, bridge architecture and seed, target subset, and training schedule. It crosses three source conditions, three independently seeded checkpoints, five outer splits, three target-side optimization repeats, ten budgets, and depths 1–3. All 4,050 expected fits complete without failed-run substitution. Inference first averages checkpoints and optimization repeats within each outer split, then treats the five outer splits as paired units. Test outcomes are used only for final evaluation; CKA and test-based depth selection are disabled in this fixed-depth audit.

The five-pair constant-width audit in Section G.15 adds an interface-specific view of this question. It holds bridge size fixed across six depths and shows that the pretrained-minus-random contrast is strongest at some shallow interfaces but can become negligible at deep ones. We therefore treat source knowledge as an interface- and pair-dependent contribution rather than a uniform explanation for every grafting gain.

At depth 1, correct pretraining exceeds shuffled-label pretraining by 0.98 BA points; the corresponding paired-t interval is $[+0.79, +1.17]$. Its advantage over shuffled pretraining is 0.85 points larger at depth 1 than at depth 3, with both the bootstrap interval in Table 37 and the paired-t interval $[+0.37, +1.33]$ excluding zero. Meaningful source labels therefore leave their clearest transferable signature at the shallow interface. Correct pretraining does not uniformly exceed random-frozen tails, which remain strong across depths. The audit consequently supports a composite mechanism: task-conditioned information is detectable in shallow layers, while architecture and regularization can account for material predictive differences.

Independent WDBC fixed-depth audit. The WDBC→Mammographic Masses mechanism audit repeats this question more strictly across all three fixed attachment depths and multiple independently seeded source checkpoints. Task-pretrained tails are not uniformly superior to random-frozen tails; the largest differences favor the random tail at shallow attachment and contract at the deepest interface. The shuffled-label condition provides a second architecture/optimization control. This result does not imply that source pretraining is useless. Pooled over the matched audit cells, pretrained-minus-random-frozen is -1.41 balanced-accuracy points (95% CI: -2.07 to -0.79), whereas pretrained-minus-shuffled-label is $+0.83$ points (95% CI: $+0.43$ to $+1.24$; Table 38). Thus the control identifies task-label information relative to shuffling, but does not establish that task-pretrained tails outperform

Table 34: Affine bridge parameter counts for the six-depth ladder ($d_T = 108$), before the parameter-free ReLU. The $32\times$ range in P_k (versus $4\times$ for the finalized three-depth ladder) is the mechanism through which the estimation term of Equation 27 can dominate at small budgets.

Depth	1	2	3	4	5	6
Width d_k	256	128	64	32	16	8
$P_k = d_k(d_T + 1)$	27904	13952	6976	3488	1744	872

Table 35: Six-depth Census ladder CKA selector quality (exploratory, 3 repeats per budget). Cf. Table 31 for the finalized three-depth ladder. Overall (pooled across budgets): oracle hit $= 0.042$, within-1-point $= 0.875$, mean regret $= 0.59$ points, median regret $= 0.55$ points.

Labels	Oracle hit	Within 1 pt	Mean regret (pt)
0.5%	0.333	1.000	0.36
1%	0.000	0.333	1.16
2%	0.000	1.000	0.43
5%	0.000	1.000	0.54
10%	0.000	1.000	0.39
20%	0.000	0.667	0.77
40%	0.000	1.000	0.69
100%	0.000	1.000	0.37

random-frozen tails on this pair. The complete paired contrasts are retained in Table 38; the redundant source-condition plot is omitted.

The 3.89-point statistic is defined within this fully crossed source-condition mechanism audit. Because the finalized three-depth result bank plotted in main Figure 3(b) uses a different aggregation, we report the 3.89-point spread only for the mechanism audit and do not transfer it to the main-paper panel.

WDBC is also the transfer pair on which the matched target-only MLP exceeds ADLG in the five-pair benchmark. The source-condition audit shows that task-pretrained frozen tails are not uniformly better than random-frozen tails, while fixed-depth performance still varies materially. Thus attachment depth can be consequential even when source-knowledge transfer does not outperform direct target learning. Correct-pretrained performance above shuffled-label controls identifies a source-label signature, whereas performance below random-frozen controls shows that the favorable mechanism cannot be attributed solely to task-specific source knowledge.

G.4 Scope of the Source-Condition Audit

The source-condition audit separates task-specific source information from effects associated with the frozen architecture by comparing correctly pretrained, shuffled-label-pretrained, and randomly initialized source networks under the same grafting protocol. In the fully crossed ADLG-Frozen Census audit, correct pretraining has its clearest advantage over shuffled-label pretraining at depth 1, and the advantage contracts significantly by depth 3. The separate five-pair constant-width study also finds that source-condition effects contract at deep interfaces, although their signs vary by pair and protocol. Together these studies localize source-label effects without claiming that correct pretraining uniformly dominates a random frozen architecture.

These controls establish an interface-dependent source effect. They do not fully separate the parameterization of the deepest frozen head from associated optimization and regularization effects. Such a decomposition would require jointly varying head initialization, head trainability, and source-training condition under the same outer splits. This finer decomposition concerns the origin of the deepest-interface effect; the present experiments address the paper's primary question of whether attachment location materially changes heterogeneous transfer.

G.5 Quantile-Ordering Controls

The ordering audit compares the true task-relevant ordering with randomized target ordering and a shuffled-label target probe. The reported CKA values test whether independent sorting and interpolation alone can produce high similarity; random target ordering remains near zero, while the shuffled-label probe is intermediate (Table 39). This diagnostic separates the score from an arbitrary row-order artifact; it does not by itself prove that CKA orders downstream risk.

Table 36: Paired mean difference in balanced accuracy (percentage points) between the CKA-selected depth and each fixed depth, averaged across all repeat–budget cells of the six-depth ladder. All differences are within ± 0.5 points.

vs. Depth	1	2	3	4	5	6
CKA − fixed (pt)	+0.50	−0.06	−0.04	−0.12	+0.01	−0.13

Table 37: ADLG-Frozen Census source-condition effects in the prespecified 0.5–20% regime. Entries are correct-pretrained minus the matched control in BA points; brackets are 95% bootstrap intervals over five outer-split aggregates. The final row compares the source-condition effect at depth 1 with the corresponding effect at depth 3.

Depth	Correct − random	Correct − shuffled
1	-0.43 $[-0.55, -0.34]$	$+0.98$ $[+0.87, +1.11]$
2	-0.45 $[-0.72, -0.23]$	$+0.25$ $[+0.06, +0.39]$
3	-0.14 $[-0.26, -0.00]$	$+0.13$ $[-0.06, +0.32]$
Depth 1 − depth 3	-0.29 $[-0.43, -0.13]$	$+0.85$ $[+0.57, +1.17]$

G.6 Unlabeled-Target Fairness

This fairness audit compares two ADLG selectors on Census: one uses the full permitted target-training pool, and the other uses only the currently labeled target examples. The held-out test set is excluded from both. The comparison tests whether access to unlabeled target covariates changes ADLG's selection or performance; by itself, it does not isolate source-transfer benefit from the effects of unlabeled data across different model families.

Across the ten label budgets, full-pool minus labeled-only balanced accuracy averages $+0.028$ points, the maximum absolute change is 0.27 points, and the selected depth agrees in 92.7% of cells. Because the source network, bridge training, outer splits, labeled subsets, and test observations are held fixed, this comparison isolates the selector's access to unlabeled target covariates. The near-identical performance shows that such access is not a material explanation for the Census low-label result.

G.7 Initialization Ablation

The Census initialization ablation compares neutral, structural GW, source-only self-supervised, and combined-residual initializers with the full ADLG configuration (Table 42). The comparison fixes the attachment depth so that initialization and depth effects are not confounded.

G.8 Bridge Capacity, Activation, and Regularization Audit

Purpose and protocol. This completed auxiliary Census-Income→Adult audit tests whether the core affine+ReLU interface is an arbitrary architectural restriction. Within each fixed attachment depth $k \in \{1, 2, 3\}$, the audit compares alternative bridge families under matched target splits, labeled subsets, initialization/training budget, and evaluation. It uses 10 repeated target splits and seven label budgets $\{5, 10, 20, 40, 60, 80, 100\}\%$. The retained source tail is frozen throughout this audit. The same source checkpoint is used throughout, target-test rows are excluded from all fitting and selection, and the automated validation confirms the expected matched design. Decision thresholds in this auxiliary audit are selected using its inner validation partition and are therefore not fixed at 0.5; balanced accuracy follows that executed rule. No tail adaptation is used (the recorded fine-tuning time is zero). After deduplicating shared reference conditions across the capacity, activation, and regularization sweeps, the audit contains 1,470 unique trained cells. This is the supplied engineering 10-repeat/7-budget protocol and is *not* substituted for the finalized $5 \times 3 \times 10$ Census headline experiment.

Capacity families. The capacity sweep compares the executed affine+ReLU bridge with a low-rank ReLU bridge, a small MLP, and a larger MLP. The trainable bridge-parameter ranges across depths are, respectively, $1{,}744-6{,}976$, $1{,}016-1{,}448$, $4{,}016-5{,}600$, and $23{,}248-26{,}368$. The following summary accompanies the depth-wise results and paired effects.

Capacity result. The principal conclusion is not that affine+ReLU is optimal across settings, but that increasing bridge expressivity does not monotonically improve target performance under this matched audit. At the 5%, 20%, and 100% budgets, the affine+ReLU summary balanced accuracies

Table 38: Complete paired WDBC→Mammographic Masses mechanism-audit summary. Its estimates use the fully crossed source-condition cells rather than the aggregation used for main Figure 3(b). Estimates are averaged over the prespecified audit cells; "BA points" means percentage points of balanced accuracy. The rank-correlation contrast is unitless. These results scope the paper's mechanism claims to what this audit supports.

Audit contrast	Estimate	95% bootstrap CI	Interpretation
Pretrained − random frozen (BA points)	-1.41	$[-2.07, -0.79]$	Random-frozen tails perform better on average.
Pretrained − shuffled-label (BA points)	$+0.83$	$[+0.43, +1.24]$	Pretraining retains some task-label information relative to the shuffled control.
Within-cell fixed-depth spread in this mechanism audit (BA points)	$+3.89$	$[+3.35, +4.51]$	Attachment depth materially affects performance within this audit.
CKA − simple fixed/random policies (BA points)	$+0.19$	$[-0.48, +0.81]$	Inconclusive.
CKA − validation selection (BA points)	$+0.33$	$[-0.58, +1.23]$	Inconclusive.
$\rho(\text{bridge}, \text{BA}) - \rho(\text{CKA}, \text{BA})$	-0.05	$[-0.163, +0.057]$	Neither association is demonstrably stronger.

Table 39: Census ordering control by candidate depth (mean ± SD over the saved audit runs). Random target ordering remains near zero; shuffled-label probe ordering is intermediate.

Depth	True ordering	Shuffled-label probe	Random target order
1	0.742 ± 0.061	0.249 ± 0.061	0.042 ± 0.014
2	0.716 ± 0.117	0.402 ± 0.131	0.016 ± 0.017
3	0.767 ± 0.084	0.546 ± 0.115	0.014 ± 0.017

are 0.796 ± 0.002, 0.813 ± 0.002, and 0.822 ± 0.001, respectively; the corresponding low-rank values are 0.781 ± 0.003, 0.808 ± 0.002, and 0.821 ± 0.001. The small and larger MLPs are lower throughout these representative budgets, with the largest separation under limited labels. Across the 210 matched repeat–budget–depth cells used for the parameter-count diagnostic, no cell shows non-decreasing balanced accuracy as bridge parameter count increases; the mean Spearman association between trainable bridge parameter count and balanced accuracy is $\rho = -0.6781$. We use this as evidence for a capacity–estimation tradeoff, not as a claim that parameter count alone determines transfer quality.

Activation control. To separate bridge capacity from the choice of nonlinearity, the affine bridge is also evaluated with ReLU and GELU under the same fixed-depth protocol.
The activation comparison is therefore interpreted conservatively: nonlinearity can matter locally, but the observed activation differences do not reverse the broader capacity result that additional MLP expressivity is not uniformly beneficial.

Regularization control. A separate sweep weakens and strengthens bridge weight decay around the default while leaving the remaining fixed-depth protocol unchanged.

Interpretation and scope. Together, the capacity, activation, and regularization sweeps support the use of affine+ReLU as a controlled low-capacity default for isolating attachment-depth selection. They do not establish global optimality over all possible bridge architectures, nonlinearities, or regularizers. Jointly selecting both attachment depth and bridge architecture remains a broader problem outside the core ADLG estimator considered here. The separate parameter-versus-performance plot is omitted because the same diagnostic is already summarized numerically by $\rho = -0.6781$ and visually by the paired-effect heatmap in the main paper.

G.9 Retained-Tail Fine-Tuning

This control asks whether favorable grafting primarily reflects a fixed source interface or target-driven adaptation of the retained hierarchy.
The exploratory WDBC tail-adaptation comparison is retained in Supplement J; it is not part of the canonical estimator or its headline evidence.

G.10 CKA Estimator Robustness

The deployed selector uses conventional centered linear CKA. On the separate Census engineering audit, we recompute each repeat–budget–depth score using debiased CKA and RV2 on the same

Table 40: Census unlabeled-target fairness control. The full-pool selector has negligible average advantage over labeled-only selection.

Labels	Full-pool BA	Labeled-only BA	Gain (pt)	Depth agreement
0.5%	0.733	0.733	+0.00	1.000
1%	0.750	0.749	+0.09	0.933
2%	0.775	0.772	+0.27	0.667
5%	0.790	0.790	+0.00	1.000
10%	0.801	0.801	+0.00	1.000
20%	0.807	0.807	+0.01	0.933
40%	0.811	0.813	-0.15	0.800
60%	0.816	0.816	+0.00	1.000
80%	0.818	0.818	+0.06	0.933
100%	0.819	0.819	+0.00	1.000

Table 41: Critical ADLG controls and ablations. The neutral-initialization adaptive graft used for causal depth comparisons is distinct from the plotted three-arm adaptive graft, which is initialized by the source-only hypernetwork in the executed notebooks.

Variant	Adaptive depth	Pretrained tail	GW	SS
Fixed-depth graft	No	Yes	No	No
Random-tail control	Yes	No	No	No
Adaptive graft (neutral-init depth ablation)	Yes	Yes	No	No
Executed three-arm adaptive graft (source-only init.)	Yes	Yes	No	Yes
Adaptive + GW	Yes	Yes	Yes	No
Adaptive + SS	Yes	Yes	No	Yes
Full ADLG	Yes	Yes	Yes	Yes

quantile-aligned representations. Table 45 reports all 70 cells aggregated by depth. All three estimators have their largest aggregate mean at depth 3. Debiased CKA and RV2 make identical cell-level depth choices; they agree with conventional linear CKA in 82.9% of cells.

We also map each estimator's selected depth to the corresponding held-out result in the common fixed-depth bank. Conventional linear CKA exactly selects the retrospective best depth in 57.1% of cells, is within 0.01 balanced accuracy in 91.4%, and has mean regret 0.0029. The identical debiased-CKA and RV2 choices achieve 70.0%, 97.1%, and 0.0015, respectively. These are diagnostic re-selections from a fixed-depth result bank rather than newly trained ADLG arms; primary results retain conventional linear CKA. Under randomized source or target ordering, mean debiased-CKA/RV2 similarity falls to approximately 0.002–0.003, compared with approximately 0.764 under true ordering, providing a scale-corrected check that the aligned signal is not reproduced by arbitrary row order.

G.11 Probe Robustness and Cross-Fitting

We perturb the target probe using an alternate optimization seed, a simpler architecture, a modestly stronger architecture, and stronger regularization at the representative 5%, 20%, and 100% budgets. Across these settings, the largest within-budget range of selected-model balanced accuracy is 0.0043, although agreement with the default depth decreases as the label budget grows. This separates sensitivity of the discrete depth identity from sensitivity of the resulting predictor.

Three-fold cross-fitting uses only permissible outer-training labels. It agrees with the current probe's depth in 44.3% of the 70 cells, but mean selected-model balanced accuracy remains close (0.8070 versus 0.8079), with mean retrospective regret 0.0038 versus 0.0029. The audit therefore supports performance stability more strongly than exact depth-selection stability.

G.12 Quantile-Grid Sensitivity

We recompute conventional linear CKA using 64, 128, and 256 equally spaced quantile locations. Each alternative grid agrees with the deployed 128-point selector in 82.9% of repeat–budget cells. Mean selected-model balanced accuracy is 0.8067, 0.8079, and 0.8093 for 64, 128, and 256 points, respectively; corresponding mean retrospective regret is 0.0041, 0.0029, and 0.0015. Interpolated grid points are deterministic summaries of the original representations and are not treated as independent statistical samples.

Table 42: Census balanced-accuracy initialization ablation. Effects are modest and the structural-GW-only initializer is not uniformly best.

Initializer	1%	5%	20%
Full ADLG	0.754	0.797	0.810
GW + residual SS, frozen tail	0.754	0.796	0.810
Neutral Xavier	0.750	0.790	0.807
Source-only self-supervised	0.750	0.792	0.809
Structural GW	0.750	0.785	0.804

Table 43: Auxiliary fixed-depth Census-Income→Adult bridge-capacity audit. Balanced-accuracy entries are the audit's reported mean ± 95% interval summaries aggregated across the three fixed depths; the parameter range reflects the depth-dependent trainable bridge count. This controlled engineering audit is not the finalized Census headline experiment.

Bridge family	Bridge params.	5% BA	20% BA	100% BA
Affine+ReLU	1,744–6,976	**0.796 ± 0.002**	**0.813 ± 0.002**	**0.822 ± 0.001**
Low-rank+ReLU	1,016–1,448	0.781 ± 0.003	0.808 ± 0.002	0.821 ± 0.001
Small MLP	4,016–5,600	0.754 ± 0.004	0.792 ± 0.002	0.813 ± 0.002
Larger MLP	23,248–26,368	0.747 ± 0.006	0.775 ± 0.004	0.793 ± 0.002

G.13 Scope of the GW Initializer

The GW–OT residual is evaluated as an optional bridge initializer, not as a semantic feature-matching procedure. The initialization ablation (Table 42) shows that the GW-only condition is not uniformly best. The available experiments do not provide a full sensitivity analysis over regularization and structural controls, so we make no claim that the GW coupling identifies feature correspondences.

G.14 Scope of the Source-Only Initializer

The source-only self-supervised initializer is compared with neutral and structural initializers in Table 42. These results measure downstream transfer performance for the tested configurations; they do not establish that reconstruction quality predicts target risk or that the corruption procedure is optimal.

G.15 Five-Pair Constant-Width Benchmark

This complementary audit evaluates five heterogeneous tabular pairs with six width-64 source hidden layers and an affine+ReLU bridge of equal trainable size across attachment depths within each pair. Ten stratified outer splits, three paired optimization repeats, four absolute target-label budgets (20, 50, 100, and 200), and three source conditions yield all 10,800 expected fixed-depth records, without duplicate experimental keys. The 1,206 recorded integrity checks pass. This architecture and absolute-label protocol differ from the finalized Census percentage-budget study.

Interpretation. This audit answers a narrower question than the finalized Census study: does attachment location matter after holding the bridge parameterization constant? The fixed-depth results answer yes. They do not establish that CKA is uniformly better than validation, random, fixed-depth, or target-only alternatives. They also show that the contribution of task-specific source pretraining can depend on the selected interface.

Attachment-depth patterns. The best mean fixed depths are 5 for Census and WDBC, 6 for Student and HCV, and 4 for Telco. Figure 9 shows these descriptive patterns with bridge size held constant within each pair.

Depth-policy comparisons. ADLG's graft estimator does not depend on CKA. Across these five pairs, inner validation improves equal-pair mean BA over the prespecified CKA screen by 0.13 points. CKA instead selects before graft fitting and is the resource-aware policy used in the reported ADLG results. It has favorable average differences over validation and random selection on HCV and Student, but unfavorable differences on the other three pairs. None of the 180 supplied budget-specific selector contrasts survives the reported Holm correction; this does not establish equivalence. The standalone classification performance of the neural target-only probe exceeds CKA-selected grafting on four pairs and nearly ties it on HCV. Census CKA outcomes exactly match fixed depth 1 at all four budgets. Thus the audit establishes that interface location is empirically consequential and that selection presents an accuracy–cost tradeoff, not that CKA minimizes target risk. The ADLG-Frozen

Table 44: Census retained-tail tuning ablation, balanced accuracy. Frozen, partial, and extensive tuning are close at the representative budgets.

Tail policy	1%	5%	20%
Extensive fine-tuning	0.754	0.789	0.806
Frozen tail	0.755	0.790	0.806
Partial fine-tuning	0.750	0.790	0.807

Table 45: Census CKA-estimator robustness in the separate engineering audit: mean $\pm$ standard deviation across 70 repeat–budget cells. The estimators are compared on identical quantile-aligned representations.

Estimator	Depth 1	Depth 2	Depth 3
Linear CKA	0.742 ± 0.061	0.716 ± 0.117	0.767 ± 0.084
Debiased CKA	0.748 ± 0.065	0.745 ± 0.114	0.801 ± 0.069
RV2	0.748 ± 0.065	0.744 ± 0.114	0.800 ± 0.069

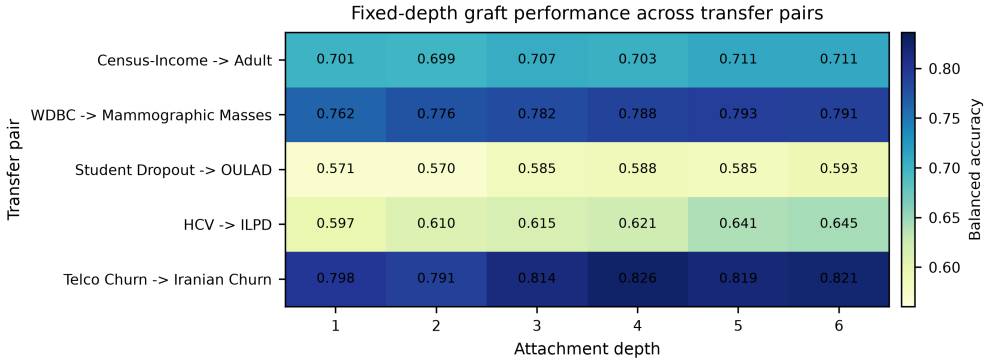

Figure 9: Exploratory six-depth fixed-graft diagnostic across transfer pairs. This diagnostic is separate from the canonical candidate set $\mathcal{K} = \{1, 2, 3\}$ and is not used for model selection. Cells average balanced accuracy over four absolute label budgets, ten outer splits, and three optimization repeats. Bridge size is constant across depths within each pair; retained-tail length varies.

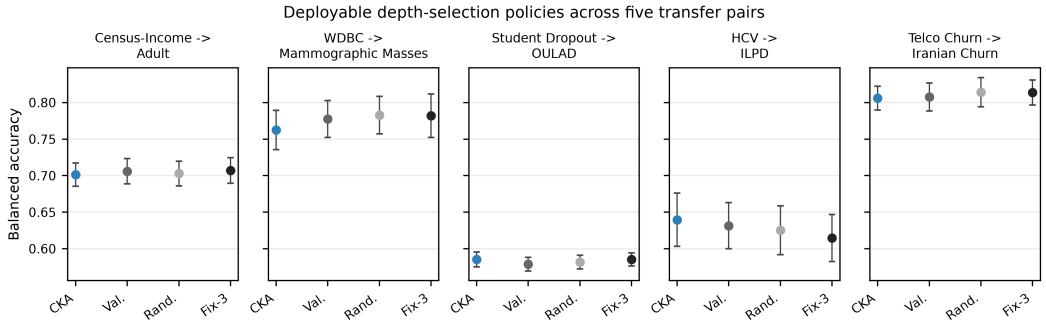

Figure 10: Exploratory replay of deployable depth-selection policies across five transfer pairs. Fixed depth 3 is prespecified in this replay; the figure does not replace the canonical Census selector-cost audit. Points average optimization repeats and the four label budgets; error bars summarize outer-split uncertainty. The corresponding paired budget-specific tests are reported separately.

Census audit includes probe construction and final fitting: CKA uses 39.4% of validation's median total time at effectively the same mean BA (Section I.2).

Source-condition controls. Source effects also vary across depth. At depth 1, correct pretraining exceeds random-frozen performance on all five pairs, but Census correct pretraining is 3.33 BA

Table 46: Census target-probe sensitivity in the separate engineering audit. †Mean exact selected-depth agreement of the four alternative probe settings with the default. BA and regret are obtained from the common fixed-depth result bank.

Budget	BA range across probes	Regret range	Mean agreement†
5%	0.7908–0.7912	0.0003–0.0006	0.925
20%	0.8040–0.8063	0.0022–0.0046	0.500
100%	0.8114–0.8156	0.0045–0.0088	0.475

Table 47: Census quantile-grid sensitivity: mean conventional linear CKA across 70 cells in the separate engineering audit. Downstream selected-model performance is stable even where the exact depth identity changes.

Grid	Depth 1	Depth 2	Depth 3
64	0.746	0.697	0.744
128	0.742	0.716	0.767
256	0.729	0.729	0.786

points below shuffled-label pretraining. WDBC depth-1 correct pretraining exceeds random and shuffled controls by 26.23 and 10.28 points, respectively. At depth 6, pretrained-minus-random differences range from -0.58 to $+0.29$ points across pairs. The pattern indicates that task-specific source effects vary with the interface: shallow grafts can retain useful source information, whereas deep grafts can increasingly behave as an architectural prior. These descriptive contrasts identify where task-specific source knowledge is most visible. The supplied source-knowledge effect file compares correct pretraining against the per-split test maximum of random and shuffled controls; that quantity is a conservative diagnostic, not a deployable selection policy.

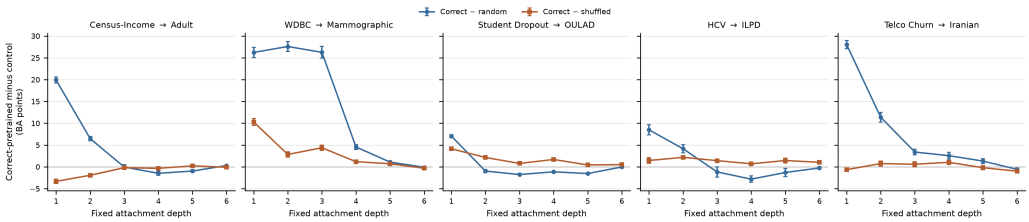

Figure 11: Depth-dependent source-condition contrasts. Each point is correct-pretrained minus the named control in balanced-accuracy points, after averaging optimization repeats and the four absolute label budgets within each outer split; error bars are one standard error over ten outer splits. Positive values favor correct pretraining. These are fixed-depth comparisons and use no test outcome for depth selection.

Statistical and provenance scope. The benchmark averages optimization repeats within outer splits and uses Student-t confidence intervals with sign-flip tests, not bootstrap intervals. Budget-specific tests and descriptive four-budget averages are distinct; no aggregate significance is inferred from the table. Repeated outer splits overlap and do not constitute independent datasets. Inputs are previously preprocessed tables: split-local scaling and the recorded isolation checks do not establish training-only provenance for upstream imputation or encoding. Cross-fitted probe hidden representations and source-checkpoint variability remain relevant scope checks; the ADLG-Frozen Census audit supplies complete selector costs, and the archive includes the available run-level metrics and budget-specific effects.

Table 48: Five-pair audit: descriptive balanced-accuracy averages over four absolute label budgets, ten outer splits, and three optimization repeats. Differences are CKA minus the named comparator in percentage points. Fixed depth 3 is prespecified. These averages do not imply significance.

Pair	CKA BA (%)	Δ validation	Δ random	Δ fixed 3	Δ probe
Census→Adult	70.14	−0.45	−0.15	−0.57	−1.56
HCV→ILPD	63.97	+0.81	+1.43	+2.49	+0.03
Student→OULAD	58.54	+0.64	+0.37	+0.00	−0.63
Telco→Iranian	80.61	−0.14	−0.80	−0.75	−1.85
WDBC→Mammographic	76.26	−1.49	−2.01	−1.95	−2.77

H Complete Pair-and-Budget Results

This section contains the primary neural comparison for ADLG, secondary non-neural target-only context, verified heterogeneous-transfer baselines when available, and principal controls. The main paper presents selected balanced-accuracy panels for readability; this section retains the complete budget curves, additional metrics, paired summaries, and configuration scope.
No learning curve or dataset–budget condition is omitted solely because it is unfavorable to ADLG.

H.1 Primary Neural Comparison

H.2 Secondary Non-Neural Tabular Context

For Census, the external-transfer comparison now includes TransTab-Transfer and CATTLE-Supervised (reduced-schedule). ADLG is stronger on balanced accuracy and minority recall across the displayed budgets, whereas ranking metrics are more mixed. HFA and Progressive Alignment are not reported on Census. Matched linear-HFA and XTab-style results for Telco and Taiwan are reported in Appendix E.8; no proxy values are inserted for unexecuted pair–method combinations.

H.3 WDBC → Mammographic Masses

The executed WDBC→Mammographic notebook uses 569 source rows with 30 modeled predictors and 961 target rows with 5 predictors, 10 repeated stratified 80/20 target splits, and nested 5–100% label budgets. It saves the four principal metrics for the trained-from-scratch target model and the historical developmental graft variants. These variants are noncanonical; their adaptation component is scoped in Supplement J. The corrected experiment starts from the raw Mammographic file and uses outer-training-pool-only imputation and standardization as specified in Appendix C.1. The reviewer-targeted mechanism audit is analyzed separately in Sections G.3, G.9, and I.
The complete WDBC outcomes are reported in the corresponding result and mechanism tables; the redundant learning-curve panel is omitted.

H.4 Telco Churn → Iranian Churn

The executed Telco→Iranian notebook uses 7,043 source rows/45 predictors and 3,150 target rows/13 predictors, 10 repeated stratified 80/20 target splits, and nested 5–100% budgets with labeled counts 126, 252, 504, 1008, 1512, 2016, and 2520. It saves the four principal metrics for the three-arm comparison; the base adaptive-depth arm is source-only-hypernetwork initialized. Exact full tables and paired uncertainty are reported only where established by the saved result artifacts. Precision/specificity and matched-operating-point results are not treated as completed unless separate artifacts are supplied.

External heterogeneous-transfer comparison. Figure 12 compares ADLG with the evaluated TransTab-Transfer and CATTLE-Supervised configurations across the seven Iranian-Churn label budgets. ADLG has the highest displayed minority recall, PR-AUC, balanced accuracy, and AUROC at every shown budget. The separation from TransTab-Transfer is especially visible in minority recall and PR-AUC as target supervision decreases, while CATTLE-Supervised has lower displayed scores on this pair. Because the figure does not display paired uncertainty for the external baselines, the comparison is interpreted descriptively rather than as a claim of statistical dominance.

H.5 Taiwan Credit → German Credit

The executed Taiwan→German notebook uses 29,965 source rows/29 predictors and 1,000 target rows/61 predictors, 10 repeated stratified 80/20 target splits, and nested 5–100% budgets with

Table 49: ADLG-Frozen Census-Income → Adult neural comparison on balanced accuracy. Values are mean ± standard deviation of the five outer-split means after paired-repeat aggregation.

Adult labels	ADLG	Tuned target-only MLP	Trained-from-scratch neural
0.5%	0.733 ± 0.013	0.721 ± 0.021	0.718 ± 0.015
1%	0.755 ± 0.007	0.741 ± 0.012	0.736 ± 0.005
2%	0.779 ± 0.003	0.755 ± 0.008	0.753 ± 0.007
5%	0.790 ± 0.006	0.767 ± 0.005	0.762 ± 0.008
10%	0.801 ± 0.005	0.778 ± 0.007	0.772 ± 0.007
20%	0.806 ± 0.003	0.789 ± 0.003	0.780 ± 0.002
40%	0.812 ± 0.002	0.795 ± 0.002	0.789 ± 0.002
60%	0.818 ± 0.002	0.800 ± 0.002	0.795 ± 0.005
80%	0.818 ± 0.003	0.809 ± 0.003	0.806 ± 0.004
100%	0.819 ± 0.002	0.803 ± 0.004	0.799 ± 0.003

Table 50: ADLG-Frozen Census means over the prespecified 0.5–20% regime. Secondary-metric differences are descriptive; the paired balanced-accuracy contrasts are reported in Table 58.

Metric	Trained-from-scratch neural	Tuned target-only MLP	ADLG-Frozen
Balanced accuracy	.754	.758	.777
Minority recall	.778	.776	.812
Specificity	.729	.741	.743
AUROC	.833	.841	.864
PR-AUC	.602	.624	.669

labeled counts 40, 80, 160, 320, 480, 640, and 800. It saves the four principal metrics for the three-arm comparison; the base adaptive-depth arm is source-only-hypernetwork initialized. Exact full tables and paired uncertainty are reported only where established by the saved result artifacts. Precision/specificity and matched-operating-point results are not treated as completed unless separate artifacts are supplied.

External heterogeneous-transfer comparison. Figure 13 compares ADLG with the evaluated TransTab-Transfer and CATTLE-Supervised configurations across the seven German-Credit label budgets. ADLG is clearly above both external baselines on PR-AUC, balanced accuracy, and AUROC throughout the displayed range. Minority recall is more variable with supervision: ADLG is generally higher at intermediate budgets, while the smallest-budget point does not show the same clear ordering. We therefore describe the external comparison as broadly favorable rather than uniformly dominant on every metric and budget. As with the Telco comparison, the plotted external-baseline curves are interpreted descriptively because this figure does not display paired uncertainty.

H.6 HCV → ILPD

The executed HCV→ILPD notebook uses 608 source rows/12 predictors and 583 target rows/10 predictors, 10 repeated stratified 80/20 target splits, and nested 5–100% label budgets. It saves AUROC, class-1 PR-AUC, balanced accuracy, and minority recall for the three-arm comparison. Exact outcomes are retained in the complete five-pair tables; class semantics or preprocessing facts not recoverable from the frozen artifacts are not inferred.

H.7 eICU → MIMIC

The executed eICU→MIMIC notebook uses 2,492 labeled eICU rows, 140 MIMIC rows from 100 unique subjects, 34 prespecified clinically matched variables expanded to 68 encoded inputs, and 10 repeated subject-grouped outcome-stratified 80/20 splits with zero subject overlap. It saves the four principal metrics for the three-arm comparison; the base adaptive-depth arm is source-only-hypernetwork initialized. We interpret these results only as testing transfer through the source representation after valid clinical variables have been defined, not as evidence that ADLG replaces clinical schema harmonization. Exact repeat/budget numerical tables and paired uncertainty are not inferred when the corresponding frozen result tables are unavailable; precision/specificity and matched-operating-point analyses are not treated as completed without separate artifacts.

H.8 Census-Income → Adult

Data and protocol. The encoded source contains $292{,}550$ observations and 511 predictors; Adult contains $48{,}790$ observations and 108 predictors. Adult positive prevalence is 23.94%. Five indepen-

Table 51: Secondary non-neural target-only context for Census-Income → Adult. These methods are retained for transparency but are not treated as the main neural representation-transfer comparison.

Adult labels	ADLG	Boosted tree	Logistic regression
0.5%	0.733 ± 0.013	0.738 ± 0.012	0.744 ± 0.017
1%	0.755 ± 0.007	0.760 ± 0.009	0.769 ± 0.008
2%	0.779 ± 0.003	0.790 ± 0.006	0.785 ± 0.008
5%	0.790 ± 0.006	0.809 ± 0.002	0.800 ± 0.005
10%	0.801 ± 0.005	0.819 ± 0.002	0.810 ± 0.002
20%	0.806 ± 0.003	0.831 ± 0.002	0.814 ± 0.001
40%	0.812 ± 0.002	0.835 ± 0.001	0.818 ± 0.003
60%	0.818 ± 0.002	0.838 ± 0.003	0.818 ± 0.003
80%	0.818 ± 0.003	0.839 ± 0.002	0.819 ± 0.002
100%	0.819 ± 0.002	0.840 ± 0.001	0.820 ± 0.001

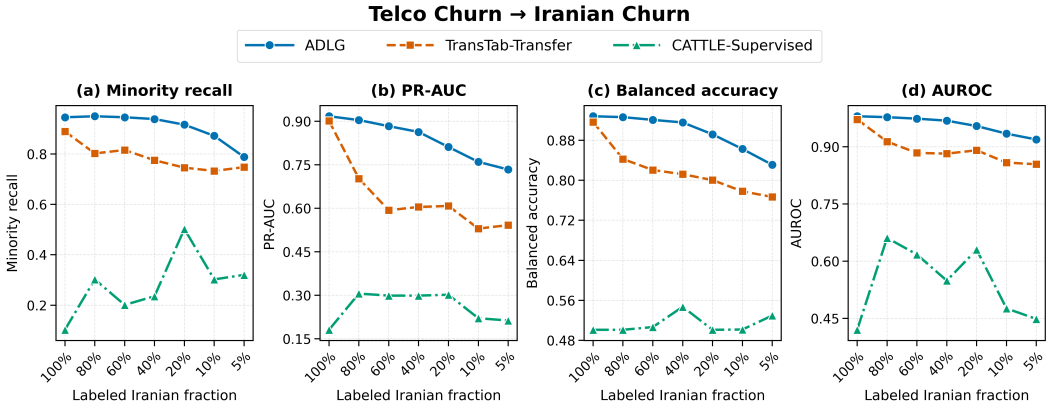

Figure 12: Telco Churn → Iranian Churn external-transfer comparison across minority recall, PR-AUC, balanced accuracy, and AUROC. ADLG is above the evaluated TransTab-Transfer and CATTLE-Supervised curves throughout the displayed label-budget range. The curves are descriptive; paired uncertainty is not displayed in this figure.

dent stratified 80/20 outer splits, three paired repeats, and ten label budgets from 0.5% to 100% yield 150 core paired cells.

External heterogeneous-transfer comparison. The Census benchmark now additionally includes TransTab-Transfer and CATTLE-Supervised (reduced-schedule). ADLG is consistently stronger on balanced accuracy and minority recall across the shown budgets. The ranking metrics are deliberately retained because they are more mixed: TransTab-Transfer is competitive or stronger on AUROC/PR-AUC over parts of the range, while CATTLE approaches ADLG as the label budget grows. This supports a positive but metric-specific interpretation centered on class-balanced prediction and minority sensitivity.

Primary neural target-only comparison. On balanced accuracy, ADLG-Frozen has the higher mean than both neural target-only references at all ten budgets. Across the prespecified 0.5–20% range, its paired gains are 2.36 points over the trained-from-scratch model (95% bootstrap CI: 2.02–2.72) and 1.88 points over the tuned target-only MLP (1.63–2.17). The low-label aggregate also favors ADLG descriptively on recall, specificity, AUROC, and PR-AUC (Table 50).

Secondary non-neural context. For completeness, Appendix H.2 retains two non-neural tabular references. Across the prespecified low-label regime, ADLG-Frozen is 0.98 points below regularized logistic regression and 1.39 points below the boosted-tree reference. These target-native comparisons delimit the neural-transfer claim.

Attachment depth and CKA selector. The source exposes three candidate widths $64, 32, 16$. Across budgets, CKA selection exactly matches the hindsight test oracle in 47.3% of cells, but lies within the prespecified one balanced-accuracy point tolerance in 88.0%. Mean oracle regret is 0.41 points, so the selector is performance-near oracle more often than it is depth-identical to oracle. An exploratory extension to a six-depth ladder with a wider bridge-width range (Section G.2) shows a

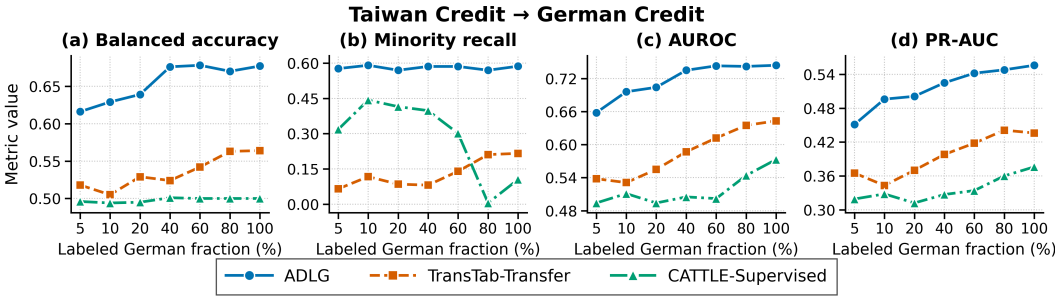

Figure 13: Taiwan Credit → German Credit external-transfer comparison across minority recall, PR-AUC, balanced accuracy, and AUROC. ADLG shows a clear advantage on PR-AUC, balanced accuracy, and AUROC across the displayed budgets, while minority recall varies more strongly with the label budget. The curves are descriptive; paired uncertainty is not displayed in this figure.

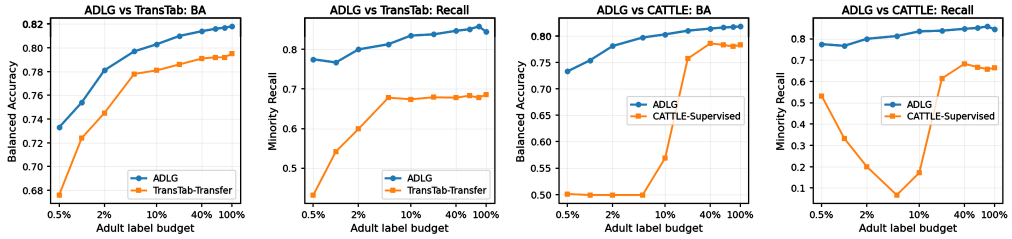

Figure 14: Census-Income→Adult external comparisons on balanced accuracy and minority recall. Panels (a)–(b) compare ADLG with TransTab-Transfer; panels (c)–(d) compare ADLG with CATTLE-Supervised.

larger fixed-depth performance spread, directly supporting the premise that depth selection matters, while leaving open how reliably CKA selects among six closely performing candidates at production repeat counts.

Mechanism. Averaged over budgets, within-run Spearman correlations are $\rho = 0.39$ for CKA versus negative quantile-aligned compatibility error, $\rho = 0.23$ for CKA versus downstream balanced accuracy, and $\rho = 0.41$ for negative compatibility error versus balanced accuracy. The evidence therefore supports a partial CKA–compatibility–performance chain rather than a direct monotone CKA–performance law.

Alternative explanations. The pretrained source tail differs from the random frozen source tail by only -0.08 points on average across budgets, and from the shuffled-label source tail by +0.25 points. This does not establish a strong task-specific source-knowledge effect on Census. In contrast, true task-relevant ordering yields higher CKA than random source/target ordering. Restricting CKA selection to only labeled Adult rows changes balanced accuracy by just +0.028 points on average (maximum absolute change 0.27 points) with 92.7% depth agreement, consistent with the matched fairness control in Table 40. Unlabeled-target access is therefore not a material explanation for the Census result.

H.9 HAM10000 → Derm7pt Multimodal Transfer

HAM10000 → Derm7pt combines a shared frozen image interface with unmatched metadata schemas. Despite its name, Derm7pt is not treated as a seven-class target. The prespecified endpoint is binary melanoma detection: HAM10000 diagnosis `mel` and all Derm7pt melanoma variants map to class 1, while the remaining diagnoses map to class 0. Accordingly, both the source network and target probe have one logit, and quantile ordering uses its scalar sigmoid score. No probability-simplex ordering is performed.

The shared image component consists of frozen 512-dimensional embeddings from ImageNet-pretrained ResNet18. Source metadata (age, sex, localization) and target metadata (sex, location, elevation) are fitted independently without a cross-table feature map. The resulting source input has 529 coordinates. The target input has 524–526 coordinates across labeled subsets because its

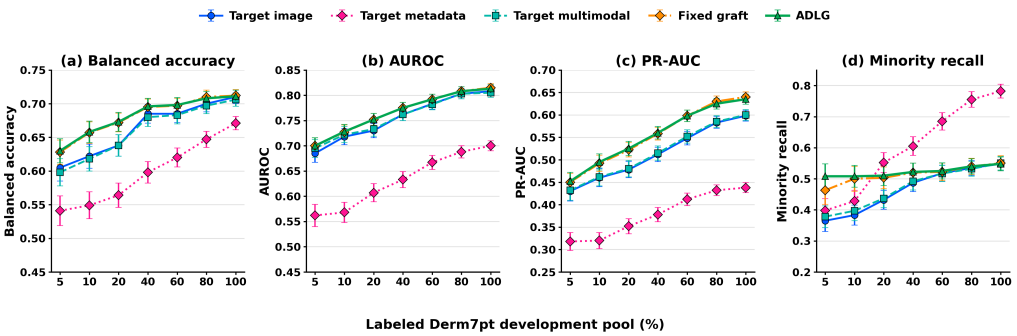

Figure 15: HAM10000 → Derm7pt multimodal transfer across labeled fractions of the Derm7pt development pool. Curves show means and 95% bootstrap confidence intervals over ten paired repeated evaluations on the fixed official test split. Image-only, metadata-only, and multimodal target controls are shown with the fixed-depth graft and ADLG.

categorical encoder is fitted anew within each subset. The target probe is therefore $d_T \to 64 \to 32 \to 16 \to 1$, with 36,225–36,353 trainable parameters. It is trained for the prespecified 45 epochs with class-weighted binary cross-entropy, AdamW learning rate 5×10^{-3}, weight decay 10^{-4}, and gradient clipping at 5.0. The 5% minimum budget contains 31 labeled development cases; all seven budgets are nested and the probe is refit for every one of the ten repeats.

For this experiment, both probe fitting and CKA use only the current labeled target subset. At each candidate depth, examples are stably sorted by their scalar risk and each hidden coordinate is linearly interpolated onto the fixed 128-location quantile grid. The grid is a common coordinate system, not 128 independent observations; at the minimum budget it is an interpolation of the 31 observed labeled ranks. Since sigmoid is monotone, probability calibration and global logit shifts cannot alter this ordering. Candidate depths $\{1, 2, 3\}$ correspond to widths $\{64, 32, 16\}$, with ties resolved in favor of the shallower depth.

The official Derm7pt development pool contains 616 cases and the official test split contains 395. The official test split is excluded from preprocessing, target-probe fitting, CKA computation, depth selection, bridge training, fine-tuning, and threshold selection. The ten runs share that test set and are treated as paired repeated evaluations rather than independent outer test splits. Accordingly, their intervals quantify optimization and initialization stochasticity on this fixed partition, not population-level test-set sampling uncertainty. The released validation audit passes all checks, including nested target subsets, identical method subset IDs, frozen image embeddings, test exclusion from preprocessing and CKA, three candidate depths, complete repeat/depth records, finite predictions, and reproducible aggregates. It contains 70 completed selector cells, 210 depth-specific CKA scores, 490 repeat–method–budget metric rows, and 193,550 test predictions. Figure 15 reports all four metrics. Across the prespecified 5–20% regime, ADLG exceeds the target multimodal model by 2.99 balanced-accuracy points, with paired 95% interval [1.55,4.17]. Because all ten paired runs share one official test split, this interval describes repeated optimization on that fixed partition rather than independent new-population uncertainty. The relative benefit outside this summary remains metric- and budget-dependent.

H.10 METABRIC → TCGA-BRCA Multimodal Transfer

This cross-cohort experiment transfers a METABRIC source trained on baseline clinical variables plus mRNA microarray measurements to a TCGA-BRCA target represented by independently processed baseline clinical variables plus RNA-seq. The shared, prespecified endpoint is binary 60-month overall survival: class 1 is death by 60 months and class 0 is known survival to at least 60 months; patients censored before the horizon are excluded. The eligible source cohort contains 1,917 patients (1,490 class 0, 427 class 1), and the eligible target cohort contains 348 patients (249 class 0, 99 class 1). The cohorts have no overlapping patient identifiers.

The protocol uses five stratified patient-level 80/20 target splits and nested target-label fractions $\{5, 10, 20, 40, 60, 80, 100\}\%$. Each target training pool contains 278 patients and each held-out test partition 70. We retain the 500 most variable genes separately for the source cohort and within each target training pool; gene sets are not forced to match, and target clinical encoding, gene selection, imputation, scaling, probe fitting, CKA selection, bridge training, and threshold choice exclude

Table 52: METABRIC→TCGA-BRCA BA: target-only MLP and paired ADLG–MLP differences (points; five target splits). Only 10% excludes zero.

Target labels	Target MLP	ADLG–MLP
5%	.529	$+0.7$ $[-1.2, 2.6]$
10%	.570	$+3.3$ $[+1.1, +5.5]$
20%	.581	$+0.7$ $[-3.1, 3.7]$
40%	.576	$+2.3$ $[-0.9, 5.7]$
60%	.591	$+2.2$ $[-2.3, 6.6]$
80%	.646	$+0.6$ $[-2.0, 3.3]$
100%	.672	-1.2 $[-4.1, 1.7]$

target-test patients. The primary comparison is multimodal ADLG versus a target-only multimodal MLP, with a fixed-depth graft and clinical-only/omics-only ADLG ablations recorded in the executed manifest.

The 10% result is the only paired balanced-accuracy contrast supported under the executed percentile-bootstrap summary. We therefore use this study to document a verified cross-cohort multimodal regime in which ADLG is favorable, not to claim a uniform multimodal advantage. The full five-method, four-metric curves and integrity checks are contained in the accompanying supplementary ZIP.

H.11 METABRIC Clinical → Omics

This experiment transfers a clinical source representation to an omics target. Each of five repeats partitions eligible patients into disjoint source and target groups and reserves a held-out target test split. Target feature selection, imputation, scaling, CKA computation, and model fitting use only the target outer-training partition. The relative ordering changes across budgets and metrics, so we report the result as direction-specific evidence rather than uniform improvement. Complete curves and paired results are reported here and in the accompanying supplementary ZIP; the main paper summarizes the direction- and budget-specific findings. The supported BA gains are +1.19, +2.49, and +2.86 points at 10%, 20%, and 40% labels, respectively. At 100%, the contrast reverses to -3.14 points (95% CI: $[-5.41, -0.04]$). The 5% BA contrast is not treated as supported.

H.12 METABRIC Omics → Clinical

The reverse direction uses omics as the source and clinical variables as the target. It follows the same patient-disjoint, held-out-target protocol. The strongest and most persistent advantage is in minority recall from 20% labels onward, with supporting balanced-accuracy and ranking-metric gains at several budgets. Balanced accuracy improves by +6.33, +6.06, and +6.71 points at 40%, 80%, and 100% labels, respectively, locating the supported BA evidence in the moderate/high-supervision regime rather than primarily at low labels. Because the effects remain budget- and metric-dependent, we do not aggregate the two directions into a single performance claim.

H.13 Student Dropout → OULAD

The executed Student Dropout→OULAD notebook uses 3,630 source rows/238 predictors and 32,587 target rows/85 predictors, 10 repeated stratified 80/20 target splits, and nested 5–100% budgets with labeled counts 1303, 2607, 5214, 10428, 15641, 20855, and 26069. It saves the four principal metrics for the same three-arm comparison. Exact outcomes are retained in the complete five-pair tables; the redundant learning-curve panel is omitted.

H.14 Operating-Point Interpretation of Minority Recall

Minority recall is reported at the prespecified classification threshold of 0.5, using the same saved predictions and matched test observations as the other metrics. It provides a class-sensitive description at that operating point, not a threshold-invariant measure of class discrimination. Balanced accuracy averages minority recall and majority specificity, and therefore penalizes a pure operating-point shift that improves one by sacrificing the other. AUROC and class-1 PR-AUC provide complementary ranking measures where the corresponding score vectors are available.

For the TabPFN v2, HyperFast, CARTE, LATTLE, and TABULA-8B artifacts, the matched records allow the fixed-threshold comparison to be decomposed into balanced accuracy, minority recall, and majority specificity. Recall and specificity are computed at 0.5, with specificity $\mathrm{TN}/(\mathrm{TN}+\mathrm{FP})$. The four comparisons show that the observed recall differences are accompanied by lower majority specificity; Table 53 displays both quantities rather than presenting recall in isolation. They are

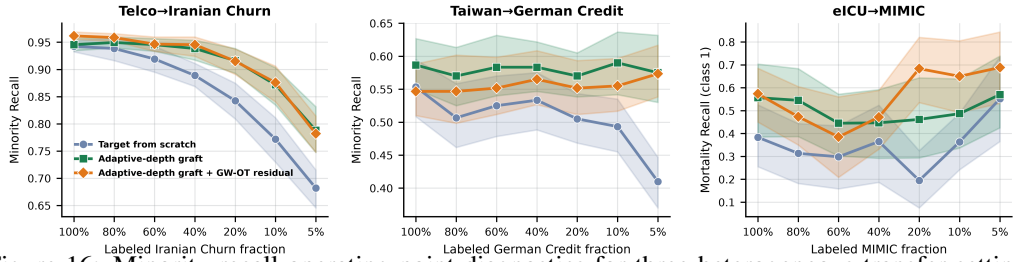

Figure 16: Minority-recall operating-point diagnostics for three heterogeneous transfer settings: Telco→Iranian Churn, Taiwan→German Credit, and eICU→MIMIC. Values use the saved matched predictions at the prespecified 0.5 threshold. The figure is supporting evidence for scarce-label minority sensitivity, not a replacement for the balanced-accuracy and ranking-metric analyses in the main evidence chain.

Table 53: Operating-point decomposition for the executed pretrained baselines. Each value first averages budgets within a pair and then weights pairs equally. All deltas are ADLG minus the named baseline in percentage points. Recall and specificity use threshold 0.5. The class-1 PR-AUC column is the external baseline's mean, not an ADLG difference. All comparisons are descriptive.

Baseline	Δ BA	Δ recall	Δ specificity	Baseline PR-AUC
TabPFN v2 (five pairs)	+3.14	+27.37	−21.10	.713
LATTLE (five pairs)	+1.97	+14.41	−10.47	.650
HyperFast (five pairs)	−0.38	+4.04	−4.80	.661
CARTE (five pairs)	+5.25	+17.77	−7.28	.677
TABULA-8B (three pairs)	+2.20	+15.60	−11.20	.764

therefore reported as an operating-point description, not as a stand-alone discrimination claim. The run artifacts record AUROC and class-1 PR-AUC for each external method. The newly supplied ADLG `predictions.csv` contains probability arrays only for Census source-condition cells (one outer split, 20/50-label budgets) and does not contain the stable test identifiers used by the five-pair TabPFN, CARTE, and LATTLE files. It therefore cannot be safely joined to those external predictions, and no new paired calibration, Brier, or PR-AUC contrast is inferred. By contrast, the saved score-based curves for the Telco and Taiwan external-transfer studies report PR-AUC, AUROC, balanced accuracy, and recall together in Figures 12 and 13.

Because classifiers may differ in calibration, the recall differences are not interpreted as guarantees at matched precision or specificity. Accordingly, minority recall is supporting evidence rather than the basis of the principal comparative claim. The main conclusions rest on balanced accuracy and matched-split analyses; PR-AUC is used as the threshold-independent companion only in comparisons for which paired score-based evidence is available.

The reported quantities obey the identity $\Delta\mathrm{BA} = (\Delta\mathrm{recall} + \Delta\mathrm{specificity})/2$ under the same pair-weighted aggregation. This identity provides a direct consistency check on every row of Table 53.

I Selector Accuracy, Regret, Stability, and Total Cost

This section gives the complete controls behind Section 7.6. The main paper reports the depth spread and Census selector regret because they directly test the method's design. The panels and tables below retain neutral and adverse selector outcomes. Estimator, source-condition, ordering, and initializer controls are reported once in Appendix G.

I.1 What Does CKA Measure? Compatibility Versus Target Performance

For every candidate depth k, let $C_k = \widehat{C}^{(k)}$ and let E_k denote the normalized quantile-aligned linear compatibility error defined by

$$E_k = \frac{\min_A \|\widetilde{H}_T^{(k)} A - \widetilde{H}_S^{(k)}\|_F^2}{\|\widetilde{H}_S^{(k)}\|_F^2}. \tag{74}$$

Table 54: Census within-run Spearman mechanism statistics, summarized by budget. The across-budget means are 0.390, 0.233, and 0.410, respectively.

Labels	ρ(CKA,−bridge)	ρ(CKA,BA)	ρ(−bridge,BA)
0.5%	0.567	-0.133	-0.367
1%	0.600	-0.200	-0.200
2%	0.700	0.333	0.467
5%	0.567	0.567	0.467
10%	0.533	0.600	0.500
20%	0.267	0.133	0.600
40%	-0.033	0.033	0.533
60%	0.200	0.400	0.833
80%	0.300	0.267	0.700
100%	0.200	0.333	0.567

Table 55: ADLG-Frozen Census selector audit over 150 cells per policy. Times are end-to-end medians, not kernel-only timings.

Policy	Mean BA	Median time (s)	Neural fits
CKA	0.793091	1.041	2
Inner validation	0.793127	2.628	4
Random depth	0.792606	0.639	1
Fixed depth 1	0.791216	0.758	1
Fixed depth 2	0.792626	0.680	1
Fixed depth 3	0.792582	0.616	1

Within each pair/split/budget/repeat cell we compute Spearman correlations $\rho(C_k, -E_k)$, $\rho(C_k, M_k)$, and $\rho(-E_k, M_k)$ across candidate depths. This avoids treating heterogeneous budget/depth observations as independent and keeps compatibility distinct from downstream risk.

Independent WDBC interface audit. The WDBC mechanism audit adds a diagnostic closer to the deployed interface by fitting a training-only regularized map from raw target inputs to the candidate source interface and evaluating held-out quantile-grid error. The within-cell rank associations are reported in Table 38. CKA is positively associated with bridgeability, but the paired difference $\rho(\mathrm{bridge}, \mathrm{BA}) - \rho(\mathrm{CKA}, \mathrm{BA}) = -0.05$ has a 95% bootstrap CI of $[-0.163, 0.057]$ (Table 38). Thus neither diagnostic is demonstrably more associated with downstream balanced accuracy on this pair. Across Census and WDBC, CKA is interpreted as a representation/interface compatibility statistic whose downstream usefulness is pair dependent, not as a target-risk estimator. This is the precise scope of Proposition 1.

I.2 Selector Quality, Regret, and Computational Cost

ADLG-Frozen end-to-end Census audit. Six prespecified deployable policies are evaluated in every combination of five outer splits, three repeats, and ten label budgets (150 cells per policy; 900 total), with no failures and no test-set oracle. Total time includes split-local preprocessing, target-probe training where required, alignment, selection, and the final graft, using one warm-up and three timed runs. Table 55 shows nearly identical observed mean BA for CKA and validation, while CKA uses 39.4% of validation's median time and half as many neural fits. The paired BA difference is -0.0036 points (95% bootstrap CI $[-0.125, +0.185]$); this interval supports uncertainty around a small observed difference, not a formal equivalence claim. The mean time saving is 2.436 s (95% CI $[2.392, 2.475]$). CKA exceeds fixed depth 3 and random depth by 0.051 and 0.048 BA points, with paired intervals $[-0.067, +0.146]$ and $[-0.102, +0.206]$, respectively. Those one-fit policies remain the cheaper nonadaptive choices. At 0.5–20% labels, CKA is 0.055 points higher than validation descriptively and uses about 60% less median time.

The oracle depth $k_{\mathrm{oracle}} = \arg\max_k M_k$ is retrospective and never participates in deployment. For Census, the complete fixed-depth and selector results are given in Tables 30 and 31. Exact CKA–oracle agreement is 47.3%, the within-one-balanced-accuracy-point rate is 88.0%, and mean oracle regret is 0.41 points.

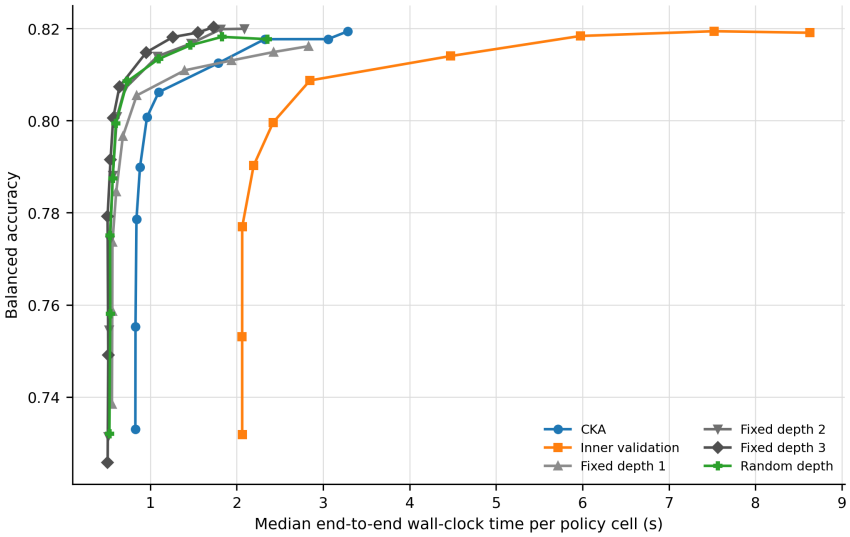

Figure 17: Accuracy–cost choices in the ADLG-Frozen Census audit. CKA attains nearly the same observed mean BA as validation at lower complete selection time; fixed and random policies are cheaper nonadaptive references.

Table 56: WDBC→Mammographic Masses depth-policy results. The oracle uses held-out test outcomes retrospectively and is never deployable. CKA's mean is comparable to the tested deployable selectors, not established as superior.

Policy	Mean balanced accuracy
Fixed deep	0.8116
Fixed middle	0.8030
CKA	0.8005
Random depth	0.7990
Validation	0.7971
Bridgeability	0.7962
Fixed shallow	0.7809
Oracle (test-only diagnostic)	0.8163

Independent WDBC selector audit. The WDBC audit evaluates CKA, validation-based selection, bridgeability-based selection, random depth, every fixed depth, and a test-only oracle on one common fixed-depth result bank. Table 56 shows that the fixed deep policy has the highest mean balanced accuracy, while CKA is statistically indistinguishable from the tested simple fixed/random policies and from validation selection (Table 38). We therefore do not interpret CKA as an oracle optimizer. Conditional on the common target probe already being computed, its scoring stage costs 0.00485 seconds, compared with 1.28697 seconds for validation selection and 1.03428 seconds for bridgeability selection in this implementation. These legacy component timings exclude probe training and are not used for the end-to-end claim; the ADLG-Frozen Census audit above supplies that comparison.

I.3 CKA Margin and Selection Stability

We measure the CKA selection margin as

$$\widehat{\gamma} = \widehat{C}_{\text{best}} - \widehat{C}_{\text{second-best}}. \tag{75}$$

The theoretical stability result concerns perturbations of the estimated score maximizer; it does not imply that this empirical margin predicts downstream regret. We therefore interpret the margin only as a score-separation diagnostic, not as a guarantee of selector quality.

Related estimator controls. Source-knowledge, ordering, and initializer controls appear with the other estimator ablations in Appendix G; the constant-width five-pair audit is Appendix G.15. We do not repeat those controls here because this section concerns selector behavior.

Table 57: Census computational summary for the executed original three-arm Phase 1+2 protocol. Counts after joint unfreezing are 1,761, 4,033, or 9,601 for depths 3, 2, or 1; these are not frozen-tail counts. Tree-model counts do not use the same neural-parameter convention.

Method	Mean time / cell	Target-trainable	Total params.
Executed Census ADLG (Phase 1+2)	1.47	1761–9601	1761–9601
Boosted tree	3.98	0	0
Regularized logistic regression	4.73	109	109
Tuned target-only MLP	5.53	9601	9601
Trained-from-scratch neural	2.07	9601	9601

Cost accounting. The ADLG-Frozen audit records complete selector wall time, including target-probe training, on a 2-physical/4-logical-core CPU host with 31.35 GB RAM; visible T4 devices were unused. Across all 150 cells, CKA uses 1.041 s median total time versus 2.628 s for validation, with effectively identical BA (Table 55). This establishes an implementation-specific end-to-end advantage over validation, not over cheaper nonadaptive one-fit policies. Peak memory and FLOPs were not instrumented.

For the standard Census graft, bridge-only counts are 6,976, 3,488, and 1,744 at depths 1–3. After Phase-2 unfreezing, the corresponding trainable graft counts are 9,601, 4,033, and 1,761. Thus the previously reported 1,744/1,761 values refer to the deepest bridge alone versus that bridge plus the 17-parameter output head, respectively.

J Statistical Methodology

All primary method comparisons exploit the paired experimental design. For the final Census analysis, five repeated paired outer splits are the split-level analysis units after averaging the three optimization repeats within each split. Because their training/test samples overlap, they are not described as independent datasets. Thus the 150 core cells are not treated as 150 independent replications.

For methods a, b, outer split s, label budget f, and nested repeat r, define

$$\Delta^{a,b}_{s,f,r} = M^{(a)}_{s,f,r} - M^{(b)}_{s,f,r}. \tag{76}$$

J.1 Outer-Split-Aware Paired Differences

Optimization/subset repeats are nested inside outer splits and are not treated as independent dataset replications.

First aggregate within outer split:

$$\overline{\Delta}^{a,b}_{s,f} = \frac{1}{R_s} \sum_{r=1}^{R_s} \Delta^{a,b}_{s,f,r}. \tag{77}$$

Then define the budget-level summary

$$\overline{\Delta}^{a,b}_{f} = \frac{1}{S} \sum_{s=1}^{S} \overline{\Delta}^{a,b}_{s,f}. \tag{78}$$

J.2 Confidence Intervals

Confidence intervals resample the split-level paired summaries while preserving the paired method structure.

For the finalized Census analysis, repeats are averaged within each outer split, and confidence intervals resample the five paired outer-split summaries while preserving paired methods. Since a nonparametric percentile bootstrap with $S = 5$ has limited resampling support, we treat it as a finite-split uncertainty summary and also compute the paired Student's t-interval with four degrees of freedom as a complementary sensitivity check. We do not base an inferential statement on the bootstrap alone if these summaries disagree. The exploratory six-depth ladder has only three repeats per budget and is not included in this finalized outer-split bootstrap analysis.

Table 58: ADLG-Frozen Census-Income→Adult low-label balanced-accuracy effects in percentage points (paired 95% outer-split percentile-bootstrap CIs). Three optimization repeats are averaged within each of five paired outer splits before inference; paired Student's t-intervals with four degrees of freedom are used as a complementary small-S sensitivity check. The final row averages the prespecified low-label budgets within split before resampling.

Labels	ADLG–trained-from-scratch	ADLG–tuned MLP
0.5%	$+1.52$ $[+0.37, +2.62]$	$+1.20$ $[-0.61, +2.61]$
1%	$+1.88$ $[+1.19, +2.54]$	$+1.45$ $[+0.43, +2.48]$
2%	$+2.56$ $[+2.14, +3.03]$	$+2.32$ $[+1.56, +3.05]$
5%	$+2.76$ $[+2.41, +3.11]$	$+2.25$ $[+2.17, +2.37]$
10%	$+2.89$ $[+2.73, +3.08]$	$+2.30$ $[+2.03, +2.57]$
20%	$+2.58$ $[+2.43, +2.72]$	$+1.75$ $[+1.37, +2.16]$
0.5–20% avg.	$+2.36$ $[+2.02, +2.72]$	$+1.88$ $[+1.63, +2.17]$

J.3 Selector Statistics

The selector comparisons report oracle agreement and regret alongside paired comparisons with deployable policies where those outputs are available. The test oracle is used only as a retrospective diagnostic; it is not a candidate selection rule.

J.4 Win–Tie–Loss and Average Rank

Win–tie–loss summaries and average ranks may be reported across dataset–budget conditions as descriptive complements.
The tie definition must be prespecified before inspecting the final comparison.
These summaries do not replace metric-specific paired comparisons or confidence intervals.

J.5 Multiple Metrics and Reporting Policy

Balanced accuracy is the prespecified primary metric.
AUROC, PR-AUC, and minority-class recall are prespecified secondary metrics.
The reported metric is not selected post hoc according to which method performs best.
Negative and neutral outcomes remain in the complete results and controls in Appendix I.

K Reproducibility, Artifact Coverage, Hashes, and Limitations

K.1 Automated Validity Audit

The final implementation verifies at minimum:

1. target-test rows never appear in training;
2. target-test labels are inaccessible before final evaluation;
3. paired methods use identical target labeled indices;
4. paired methods use identical outer splits;
5. every candidate depth is evaluated in the depth ablation;
6. CKA-selected depth is selected using CKA only;
7. validation-selected depth uses validation information only;
8. retrospective oracle depth is marked diagnostic and never deployed;
9. threshold-dependent metrics use the prespecified 0.5 threshold in the executed three-arm notebooks, or an explicitly validation-selected operating point in dedicated audits, never target-test labels;
10. final metrics are calculated through a common evaluator;
11. every reported external baseline is verified or explicitly marked unverified/not run;
12. every final figure is reproducible from saved source-data tables.

Final Census audit. All twelve automated integrity checks pass. The audit verifies 750/750 saved prediction groups, including test isolation, paired split/index identity, complete candidate-depth evaluation, CKA-only depth selection, validation-only validation selection, diagnostic-only oracle use, threshold/test-label isolation, executed-or-explicitly-not-run arms, and agreement between the common evaluator and saved predictions. The exploratory six-depth ladder (Section G.2) satisfies the same test-isolation and diagnostic-only-oracle rules but has not yet been passed through this full audit at the finalized repeat count.

Independent WDBC mechanism audit. The completed mechanism audit passes its implemented training/selection checks; these checks do not certify upstream imputation provenance. It contains 420 tail-adaptation runs, 1,890 source-condition×fixed-depth runs, 630 interface-bridgeability cells, 630 validation-search cells, 1,680 depth-policy cells, 70 logistic controls, 70 target-from-scratch controls, and 10 imputation-audit records; all expected counts are observed. Historical WDBC target-test identities are reconstructed exactly for all ten repeats. Target-test isolation within the executed training code, training-only deployable selectors, matched source-condition initialization, frozen-tail enforcement, Phase-2 drift measurement, and diagnostic-only oracle use all pass the automated checks. Separately, the supplied historical predictions reproduce all 840 metric entries in the 210-row three-arm result table to numerical precision (maximum absolute difference approximately 1.1×10^{-16}); their labels match the raw target severity field at the stored row indices. This verifies consistency of the saved predictions and summaries, not a new training run or an audit of the original imputation fitting scope for the historical artifact. The corrected raw-data WDBC three-arm rerun is reported separately in Appendix C.1.

ADLG-Frozen Census source-condition audit. All 4,050 expected fits are present: three source conditions, three source checkpoints, five outer splits, three optimization repeats, ten budgets, and three fixed depths. The audit verifies identical architecture and target-side training, paired bridge seeds, frozen tails, absence of tail adaptation or GW initialization, within-partition label shuffling with preserved class counts, split-local target scaling, validation-only threshold fitting, and target-test use only after fitting. Checkpoint, partition, feature, label, and index hashes are retained in the run-level artifact.

K.2 Protocol and Runtime Record

The released artifacts record the finalized configurations and validation checks available for each experiment. Where historical provenance is not available, this supplement labels it as unavailable and does not use it to support a claim. Documented implementation corrections and changes to the modeling procedure are distinguished in the relevant experiment-specific sections.

K.3 Result and Figure Provenance

The accompanying supplementary ZIP maps reported figures and tables to the available source result files, analysis code, configurations, method names, and metric definitions. Its notebooks recompute the listed analyses from the supplied frozen outputs without retraining.

K.4 Negative-Result Reporting Policy

The reported results retain unfavorable as well as favorable outcomes. For example, Appendix I gives the mixed CKA-versus-selector results, the target-only probe comparisons, and the depth-dependent pretrained-versus-random tail contrasts. No seed, pair, budget, or metric is excluded solely because it is unfavorable to ADLG. The exploratory six-depth ladder's low exact CKA–oracle agreement is also reported in full (Section G.2).

K.5 Released Artifacts

The accompanying supplementary ZIP contains run-level metrics, available split and labeled-subset manifests, configurations, analysis utilities, and notebooks. Example- or patient-level clinical predictions, restricted raw datasets, and access-controlled model artifacts are not redistributed. The package supports recomputation of the reported summaries from frozen outputs; training workflows that require original datasets or licensed models identify those external inputs explicitly.

- run-level metrics and frozen aggregate outputs;
- available split and labeled-subset manifests;
- model and training configurations;

- analysis utilities and executable notebooks;
- available environment, seed, and fidelity metadata.

K.6 Software, Hardware, Seeds, Hashes, and Release Checklist

Detailed hardware and package records appear in Appendix D.8, with integrity checks and release contents in Appendices K.1–K.5.

K.6.1 Reproducibility Status

The main finalized Census experiment, ADLG-Frozen 900-cell selector-cost audit, and independent WDBC mechanism audit include automated training-isolation/ selection checks and frozen result artifacts. For WDBC/Mammographic, raw-record audit and exact reconstruction resolve label semantics and output reproducibility, while the leakage-safe rerun removes the historical imputation-fitting concern by fitting imputation within each outer target-training pool (Appendix C.1). The appendices report only dataset, preprocessing, seed, environment, or baseline- fidelity details established by those artifacts; unavailable provenance is reported as not available in the archived record rather than inferred. Artifact coverage is experiment-dependent and is stated in Appendix K.5; unavailable or restricted material is not implied by these checks.