GLADE: ADAPTIVE STRUCTURAL TRANSFER FOR POWERFUL OUT-OF-DISTRIBUTION TESTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Auxiliary labeled data from historical records, related domains, or synthetic sources can be leveraged to improve the power of out-of-distribution testing. However, differences between auxiliary and current anomaly distributions can make directly transferred detection scores ineffective or even harmful. The challenge is to turn auxiliary category information into detection power while controlling false discoveries and guarding against negative transfer. We propose GLADE, an adaptive structural-transfer procedure that uses auxiliary category labels to guide local score learning from current unlabeled data and reference inliers. GLADE further combines these local scores with a global anchor and selects their contribution according to estimated discovery utility. A symmetric learning-and-selection construction ensures finite-sample false discovery rate (FDR) control under exchangeability conditions. Our theory characterizes the potential power gains from structural transfer, the costs of learning and selection, and conditions for protection against negative transfer. Experiments on synthetic and real datasets demonstrate power improvements over matched global baselines, with empirical FDR below the nominal level. Together, these results establish a principled approach to exploiting auxiliary structure for out-of-distribution detection with rigorous control of false discoveries.

1 INTRODUCTION

Out-of-distribution (OOD) testing seeks to identify observations that do not follow a reference distribution. In many applications, several distinct outlier types coexist within the same test data. For example, in network security, malicious traffic may arise from different attack categories; in industrial equipment monitoring, abnormal sensor readings may reflect different types of equipment faults. These anomaly types may involve different feature subsets, depart from the reference distribution in different directions, and occur at widely varying frequencies.

Capturing heterogeneous anomaly patterns with a single global scoring function may require a complex model that is difficult to estimate reliably with limited data. Simpler models, however, may lack the flexibility to capture distinct feature subsets and directions of departure. We study how auxiliary information from labeled outlier examples can help simple models capture this heterogeneity and improve OOD testing. Such examples may be available from historical investigations or related domains, or generated by simulating specific anomaly mechanisms or using generative models. The category labels can guide local learning from current data, allowing simple models with different scoring directions to be combined into a more flexible detection rule.

A natural way to use these labeled examples is to learn category-specific scoring functions from the auxiliary data. This approach raises two challenges. First, it is unclear which auxiliary categories are relevant to each new observation and how to combine the corresponding local scores. Second, auxiliary anomaly distributions may differ from those in the current data because of temporal shifts, environmental differences, or discrepancies between synthetic and observed anomalies. Such source–target mismatch can make directly transferred scores less effective at distinguishing current outliers from reference inliers, potentially inducing *negative transfer* and reducing OOD testing power.

In addition to transfer challenges, large-scale OOD testing requires error control across many simultaneous tests. *False discovery rate* (FDR; Benjamini & Hochberg, 1995) control ensures that the expected proportion of false discoveries among reported outliers does not exceed a prespecified

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.

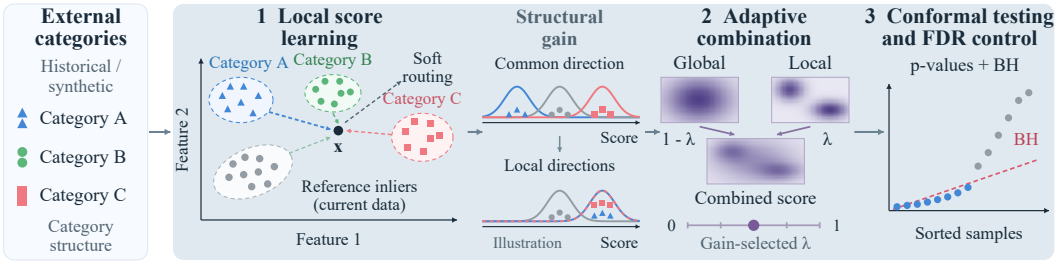

Figure 1: Overview of GLADE. Auxiliary category structure guides local score learning from current data. Adaptive shrinkage combines local and global scores, followed by conformal testing with FDR control.

level, helping reduce the burden of unnecessary follow-up investigations. Conformal inference provides a model-free approach to constructing p-values; under suitable exchangeability and symmetry conditions, applying the Benjamini–Hochberg (BH) procedure to these conformal p-values yields finite-sample FDR control. However, adaptively learning and selecting detection scores with guidance from auxiliary labeled outliers introduces substantial challenges for conformal calibration. Our goal is to develop a procedure that exploits informative category structure to improve power, guards against negative transfer, and preserves FDR control after adaptive score learning and selection.

Contributions. We introduce GLADE (Global–Local Adaptive shrinkage for Discovery Enhancement), a structural-transfer framework for OOD testing. Our contributions are threefold:

1. **Adaptive structural transfer.** GLADE provides a novel framework for improving existing outlier detectors through adaptive score refinement. Taking an existing detector's score as a global anchor, GLADE uses external category information to guide local score learning from current unlabeled data and reference inliers. Adaptive shrinkage combines these complementary scores, retaining local refinements only when their estimated discovery gains meet a prespecified threshold.
2. **Validity and transfer guarantees.** Our symmetric learning-and-selection design ensures model-free, finite-sample FDR control under exchangeability, even when auxiliary and current anomaly distributions differ. Our theory characterizes the potential power gains from structural transfer, identifies conditions under which these gains persist despite routing, estimation, and selection errors, and bounds negative-transfer losses when candidate evaluation is sufficiently accurate.
3. **Empirical validation.** Simulations and experiments on four real datasets demonstrate power improvements over matched global anchors, with empirical FDR below the nominal level. GLADE can adapt to the quality of auxiliary information, retaining beneficial local refinements while limiting negative transfer from unreliable guidance.

Related work. (1) **Conformal outlier testing.** Conformal outlier testing constructs finite-sample calibrated evidence from learned anomaly scores (Bates et al., 2023), with extensions that exploit unlabeled target observations through semi-supervised or target-adaptive learning (Mary & Roquain, 2022; Marandon et al., 2024), incorporate labeled outliers (Liang et al., 2024), and support adaptive score selection or collective detection (Liang et al., 2023; Bai & Jin, 2024; Zhao & Sun, 2025; Sun et al., 2026; Magnani et al., 2026). By contrast, GLADE introduces a structural transfer strategy that exploits auxiliary category structure without directly transferring scores learned from auxiliary data.

(2) **Structured multiple testing.** Structured multiple-testing methods exploit heterogeneity through grouping and hypothesis weighting (Genovese et al., 2006; Benjamini & Heller, 2007; Sun & Cai, 2009; Sun et al., 2015), while procedures based on independent hypothesis weighting, adaptive p-value thresholding, and the structure-adaptive Benjamini–Hochberg algorithm learn data-adaptive weights or thresholds (Ignatiadis & Huber, 2021; Lei & Fithian, 2018; Li & Barber, 2019). Related methods based on covariate-assisted ranking and screening, grouping–adjusting–pooling, and locally adaptive weighting further demonstrate the value of covariate and local structural information (Cai et al., 2019; Xia et al., 2020; Cai et al., 2022). The distinction here is that GLADE incorporates observation-dependent structure directly into score fitting and aggregation. The resulting dependence

must be accounted for in the validity argument (Blanchard & Roquain, 2008), motivating calibration of the complete adaptive score while preserving symmetry throughout learning and selection.

(3) **Transfer learning and domain adaptation.** Transfer learning and domain adaptation reuse source information through importance weighting or representation alignment (Pan & Yang, 2009; Zhuang et al., 2021; Sugiyama et al., 2007; Ganin et al., 2016; Farahani et al., 2021), with open-set extensions accommodating target-private classes and label-space mismatch (Jang et al., 2022; Garg et al., 2022). However, source–target mismatch can induce negative transfer (Wang et al., 2019; Zhang et al., 2022), while predictive confidence may also become unreliable under distribution shift (Ovadia et al., 2019; Park et al., 2020; Gong et al., 2021); outlier exposure faces a related issue when auxiliary anomalies do not represent current anomalies (Hendrycks et al., 2019). GLADE addresses these concerns by transferring auxiliary category information, relearning local scores from current data, and shrinking toward a global anchor when estimated local discovery gains are insufficient.

2 Structural Transfer via Global–Local Learning

2.1 Problem Setup

Let $X \in \mathcal{X} \subseteq \mathbb{R}^d$ denote a feature vector, where d is the feature dimension, and let P_0 denote the reference inlier distribution. We observe n_{tr} inliers for training, n_{cal} inliers for calibration, and m unlabeled test observations. Set $n = n_{\mathrm{tr}} + n_{\mathrm{cal}}$ and define the disjoint index sets $\mathcal{I}_{\mathrm{tr}} = \{1, \dots, n_{\mathrm{tr}}\}$, $\mathcal{I}_{\mathrm{cal}} = \{n_{\mathrm{tr}} + 1, \dots, n\}$, and $\mathcal{I}_{\mathrm{test}} = \{n + 1, \dots, n + m\}$. Write the training, calibration, and test data as $\mathcal{D}_{\mathrm{tr}} = (X_i)_{i \in \mathcal{I}_{\mathrm{tr}}}$, $\mathcal{D}_{\mathrm{cal}} = (X_i)_{i \in \mathcal{I}_{\mathrm{cal}}}$, and $\mathcal{D}_{\mathrm{test}} = (X_i)_{i \in \mathcal{I}_{\mathrm{test}}}$, with $X_i \sim P_0$ for $i \in \mathcal{I}_{\mathrm{tr}} \cup \mathcal{I}_{\mathrm{cal}}$. For each test observation X_{n+j}, let the fixed but unknown state $\theta_j \in \{0, 1\}$ indicate whether it is an inlier ($\theta_j = 0$) or an outlier ($\theta_j = 1$). Test outliers may belong to K categories, with distributions $P_{1,1}, \dots, P_{1,K}$ different from P_0. Conditional on $\theta_j = 1$, we write $X_{n+j} \sim P_{1,z_j}$, where $z_j \in \{1, \dots, K\}$ is the fixed but unobserved category label.

The hypotheses of interest are $H_{0,j} : \theta_j = 0$ versus $H_{1,j} : \theta_j = 1$, for $j = 1, \dots, m$. Let $\mathcal{H}_0 = \{j : \theta_j = 0\}$ and $\mathcal{H}_1 = \{j : \theta_j = 1\}$, with $m_0 = |\mathcal{H}_0|$ and $m_1 = |\mathcal{H}_1|$. A testing procedure returns a discovery set $\widehat{\mathcal{R}} \subseteq \{1, \dots, m\}$, where $j \in \widehat{\mathcal{R}}$ means that $H_{0,j}$ is rejected and X_{n+j} is declared an outlier. Define the false-discovery proportion and true-discovery count as

$$\mathrm{FDP} = \frac{|\widehat{\mathcal{R}} \cap \mathcal{H}_0|}{1 \vee |\widehat{\mathcal{R}}|}, \qquad \mathrm{TDC} = |\widehat{\mathcal{R}} \cap \mathcal{H}_1|, \tag{1}$$

where $a \vee b = \max\{a, b\}$. Our goal is to improve the average power $\mathrm{AP} = \mathbb{E}[\mathrm{TDC}/(1 \vee m_1)]$ subject to $\mathrm{FDR} = \mathbb{E}[\mathrm{FDP}] \leq \alpha$ for a prescribed $\alpha \in (0, 1)$.

We additionally observe an independent sample of labeled auxiliary outliers,

$$\mathcal{D}_{\mathrm{aux}} = ((X_\ell^A, C_\ell))_{\ell \in \mathcal{I}_{\mathrm{aux}}}, \qquad \mathcal{I}_{\mathrm{aux}} = \{1, \dots, n_{\mathrm{aux}}\}$$

where $C_\ell \in \{1, \dots, K_0\}$ identifies one of K_0 auxiliary outlier categories. For each category c, let Q_c denote its source feature distribution, so that $X_\ell^A \sim Q_{C_\ell}$. We assume that $\mathcal{D}_{\mathrm{aux}}$ is independent of the reference and test data. Anomaly patterns can change over time and across environments through covariate and concept shift (Moreno-Torres et al., 2012; Vovk, 2020). Accordingly, we allow both *the set of anomaly categories and their category-specific feature distributions to change*: K need not equal K_0, and we assume neither a one-to-one correspondence between auxiliary and current categories nor equality between Q_c and $P_{1,k}$. Thus, some auxiliary categories may be absent from the current data, previously unrepresented categories may occur, and even corresponding categories may differ distributionally.

Conformal OOD testing constructs calibrated p-values to achieve finite-sample FDR control under suitable exchangeability conditions (Bates et al., 2023; Marandon et al., 2024). Existing formulations, however, do not use auxiliary category labels for structural routing. Appendix A.3 provides a more detailed comparison with Marandon et al. (2024) and Liang et al. (2024), focusing on how unlabeled target data and labeled outliers are incorporated. Our objective is to exploit auxiliary category structure to improve the detection of heterogeneous current anomalies while preserving FDR control and guarding against negative transfer.

2.2 THE GLADE PROCEDURE

We describe GLADE in three steps. Algorithm 1 and Figure 4 in Appendix A.1 provide the complete implementation and a schematic overview, respectively.

Step 1. Structural routing and local target learning.

Fit a multiclass probabilistic classifier to $\mathcal{D}_{\mathrm{aux}}$ and use fitted category probabilities as routing weights:

$$A(x) = \big(A_1(x), \ldots, A_{K_0}(x)\big) \in \Delta^{K_0-1}, \qquad \Delta^{K_0-1} = \Big\{a \in [0,1]^{K_0} : \textstyle\sum_{c=1}^{K_0} a_c = 1\Big\}.$$

Here $A_c(x)$ is the soft routing weight for observation x in the local comparison associated with auxiliary category c, allowing participation in multiple comparisons. These weights encode similarity to auxiliary categories rather than current outlier probabilities.

Remark 1. *When an independently pretrained model is available, its category predictions can directly supply the routing weights, without direct access to the auxiliary labeled sample $\mathcal{D}_{\mathrm{aux}}$. For example, a vision–language model such as CLIP can compare observations with prespecified textual descriptions of anomaly categories, with softmax-normalized similarity scores defining $A(x)$ (Radford et al., 2021). Jeong et al. (2023) explores a related use of textual descriptions for anomaly detection.*

Let $G : \mathcal{X} \to \mathbb{R}$ be a global score function, with larger values indicating stronger evidence against P_0. Define $\mathcal{I}_{\mathrm{mix}} = \mathcal{I}_{\mathrm{cal}} \cup \mathcal{I}_{\mathrm{test}}$ and the pooled sample $\mathcal{D}_{\mathrm{mix}} = (X_i)_{i\in\mathcal{I}_{\mathrm{mix}}}$, of size $N = n_{\mathrm{cal}} + m$. Here $\mathcal{D}_{\mathrm{tr}}$ contains reference training inliers, whereas $\mathcal{D}_{\mathrm{mix}}$ pools calibration inliers with unlabeled test observations. For each auxiliary category $c = 1, \ldots, K_0$, fit a weighted binary classifier $\widehat{g}_{G,c}$ that contrasts $\mathcal{D}_{\mathrm{tr}}$ with $\mathcal{D}_{\mathrm{mix}}$, weighting observations in both samples by $A_c(x)$. This common weighting focuses each comparison on the region associated with category c. For $x = (x_1, \ldots, x_d)^\top$, the classifier uses the augmented feature vector $Z_G(x) = \big(G(x), x_1, \ldots, x_d\big)^\top$. Including $G(x)$ reuses information captured by the global score, while the original features allow adaptation to local anomaly patterns. Aggregate the fitted scores using the routing weights:

$$S_{\mathrm{local},G}(x) = \textstyle\sum_{c=1}^{K_0} A_c(x)\widehat{g}_{G,c}\{Z_G(x)\}. \tag{2}$$

With informative routing, the weighted aggregate captures heterogeneous departures through simple local models with distinct scoring directions, reducing the complexity required of a global model.

Step 2. Global–local shrinkage and utility-oriented selection.

The global anchor provides a stable common scoring rule, while the local score can capture heterogeneous departures. Local estimates may nevertheless be unreliable when routing is uninformative or weighted comparisons have small effective sample sizes. We therefore apply the classical shrinkage principle (James & Stein, 1961; Efron & Morris, 1975; Hansen, 2007; Johnstone, 2019), shrinking the local score toward the global anchor to balance modeling flexibility with estimation stability. For a finite grid $\Lambda \subset [0, 1]$ containing zero, form

$$S_{G,\lambda}(x) = (1-\lambda)G(x) + \lambda S_{\mathrm{local},G}(x), \qquad \lambda \in \Lambda. \tag{3}$$

The tuning parameter λ controls the magnitude of the local correction. Smaller values reduce both the correction and its sensitivity to routing and estimation errors, whereas larger values retain more of the structure captured by local learning. Including $\lambda = 0$ preserves the global anchor as a candidate.

We select λ using estimated discovery utility, an instance of *conformalized automated model selection* (Marandon et al., 2024; Liang et al., 2024; Bai & Jin, 2024; Magnani et al., 2026). Repeated cross-fitting produces held-out scores for reference training inliers and pooled observations, using the same splits across candidates and withholding calibration/test roles within the pool. Comparing these scores yields pseudo-p-values, to which we apply BH at the target FDR level α. Let $R^{(b)}_{G,\lambda}$ denote the resulting pseudo-discovery count in repeat $b = 1, \ldots, B$, and let $\overline{R}_{G,\lambda} = B^{-1}\sum_{b=1}^{B} R^{(b)}_{G,\lambda}$. We evaluate candidates using

$$J_{G,\lambda} = \overline{R}_{G,\lambda} - \kappa_{\mathrm{se}}\widehat{\mathrm{se}}\big(R^{(1)}_{G,\lambda}, \ldots, R^{(B)}_{G,\lambda}\big), \tag{4}$$

where $\widehat{\mathrm{se}}$ is the empirical standard deviation of the repeat counts divided by $\sqrt{B}$, and $\kappa_{\mathrm{se}} \geq 0$ controls the variability penalty, set to zero when $B = 1$. This objective rewards larger pseudo-discovery

counts while discounting sensitivity to data splits. Let $\lambda_G^\dagger \in \arg\max_{\lambda\in\Lambda} J_{G,\lambda}$, with ties resolved in favor of smaller λ. For a prespecified minimum gain $\Delta_{\min} \geq 0$, select

$$\widehat{\lambda}_G = \begin{cases} \lambda_G^\dagger, & J_{G,\lambda_G^\dagger} - J_{G,0} \geq \Delta_{\min}, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

The exact anchor candidate and minimum-gain rule provide the basis for protection against negative transfer under the conditions of Theorem 4. Appendix A.2 provides the cross-fitting details and interpretation of the utility proxy.

Step 3. Conformal calibration and FDR control.

Classical conformal calibration relies on exchangeability of the scores entering the rank comparison, which symmetric score construction preserves under exchangeable sampling (Vovk et al., 2005; Shafer & Vovk, 2008). When scores are learned using the calibration–test pool, this symmetry must also account for data-adaptive fitting and selection (Marandon et al., 2024; Angelopoulos et al., 2025). GLADE therefore enforces *pooled permutation equivariance* (Definition B.1) throughout fold construction, fitting, selection, and refitting. Appendix B.1 gives the construction requirements and calibration argument. With the fitted router A and selected weight $\widehat{\lambda}_G$ fixed, refit the global anchor according to its training specification and each local classifier using the full $\mathcal{D}_{\text{tr}}$ and unordered $\mathcal{D}_{\text{mix}}$. Combine the refitted scores using the shrinkage rule in Step 2 to obtain $S_G^{\text{GLADE}} = S_{G,\widehat{\lambda}_G}^{\text{full}}$, and evaluate this same score on all calibration and test observations. For each test observation, define

$$p_j = \frac{1+\sum_{i\in\mathcal{I}_{\text{cal}}} \mathbf{1}\{S_G^{\text{GLADE}}(X_i) \geq S_G^{\text{GLADE}}(X_{n+j})\}}{n_{\text{cal}}+1}, \qquad j = 1,\ldots,m. \tag{6}$$

Applying BH at level α yields the discovery set $\widehat{\mathcal{R}} = \text{BH}_\alpha(p_1,\ldots,p_m)$.

3 Theory: validity, structural gain, and adaptive protection

This section establishes finite-sample FDR control for GLADE and characterizes the potential power gains from structural transfer. We then identify conditions under which GLADE protects against negative transfer.

3.1 Finite-sample validity

We first establish finite-sample FDR control for GLADE under a conditional formulation of the inlier exchangeability assumption commonly used in conformal OOD testing (Bates et al., 2023).

Assumption 1 (Conditional exchangeability of inliers)**.** *Let $\mathcal{H}_0 = \{j : \theta_j = 0\}$ denote the set of null test indices. Conditional on $\mathcal{D}_{\text{tr}}$, $\mathcal{D}_{\text{aux}}$, the test states, and the non-null test observations, the calibration observations $(X_i)_{i\in\mathcal{I}_{\text{cal}}}$ and null test observations $(X_{n+j})_{j\in\mathcal{H}_0}$ are jointly exchangeable.*

Theorem 1 (Finite-sample FDR control for GLADE)**.** *Let $\widehat{\mathcal{R}}_G^{\text{GLADE}} \subseteq \{1,\ldots,m\}$ denote the rejection set obtained by applying BH at level α to the conformal p-values in equation 6, computed by Algorithm 1 with global anchor G. Under Assumption 1,*

$$\text{FDR} = \mathbb{E}\left[\frac{|\widehat{\mathcal{R}}_G^{\text{GLADE}} \cap \mathcal{H}_0|}{|\widehat{\mathcal{R}}_G^{\text{GLADE}}| \vee 1}\right] \leq \alpha,$$

where the expectation is over the data and algorithmic randomization.

Theorem 1 separates inferential validity from the quality of structural transfer: FDR control does not require informative auxiliary structure, accurate local scores, or reliable utility estimates. Under the stated exchangeability assumption, validity rests on pooled permutation equivariance throughout fold construction, fitting, selection, and refitting, together with symmetric tie handling. Proposition B.1 verifies that GLADE satisfies the required equivariance.

3.2 Potential gains from structural transfer

Capturing heterogeneous anomaly patterns with a single global score may require substantial model complexity, making reliable estimation difficult with limited current data. Auxiliary category structure

offers a way to increase scoring flexibility by combining simple local models. We examine this mechanism in a heterogeneous Gaussian model with a shared covariance. Within each group, the log-likelihood ratio is linear in the features, making linear scores natural optimal benchmarks. Different groups may nevertheless require different scoring directions. We therefore compare group-specific linear scores with the best common linear direction, and then examine how differences between auxiliary and target directions affect direct score transfer.

We quantify structural gains through an asymptotic power analysis at a fixed nominal FDR level $\alpha \in (0, 1)$ (Cai et al., 2022; Liang et al., 2024). Specifically, we compare the large-m limits of the proportion of non-null hypotheses rejected by BH. Suppose $m_0/m \to \rho_0 \in (0, 1)$, and let $\rho_1 = 1 - \rho_0$. Let P_1 denote the aggregate alternative distribution, so that $P_\rho = \rho_0 P_0 + \rho_1 P_1$ describes the limiting test composition.

For a fixed score S with a continuous null distribution, define the null-tail p-value $p_S(x) = P_0\{S(X') \geq S(x)\}$, where $X' \sim P_0$, and its alternative distribution function $F_{1,S}(q) = P_1\{p_S(X) \leq q\}$. The limiting rejection threshold and asymptotic power are

$$\tau_\alpha(S) = \sup\big\{q \in [0, 1] : q \leq \alpha\{\rho_0 q + \rho_1 F_{1,S}(q)\}\big\}, \qquad \mathrm{Pow}_\alpha(S) = F_{1,S}\{\tau_\alpha(S)\}.$$

These quantities describe the large-m limits of the BH cutoff and the fraction of non-null hypotheses rejected, respectively, under uniform convergence of the empirical p-value distributions and a stable crossing of the limiting BH curve (Genovese & Wasserman, 2002; Storey et al., 2004). They are defined using exact null calibration; extending this interpretation to empirically calibrated p-values additionally requires vanishing calibration error. Appendix B.2.1 gives the precise conditions.

We first compare a common scoring direction with group-specific directions in an idealized setting where group membership $Z \in \{1, \ldots, K\}$ is observed, and auxiliary categories coincide with the target groups, so that $K_0 = K$. Write $X = (Z, W)$, where $W \in \mathbb{R}^p$, and let $H = 0$ and $H = 1$ denote inlier and outlier status, respectively. Assume that group proportions are identical under the null and alternative: $P(Z = k \mid H = h) = \pi_k > 0$ for $h \in \{0, 1\}$, with $\sum_{k=1}^K \pi_k = 1$. Within each group, suppose

$$W \mid (H = 0, Z = k) \sim N(0, \Sigma), \qquad W \mid (H = 1, Z = k) \sim N(\mu_k, \Sigma), \tag{7}$$

where $\mu_k \in \mathbb{R}^p$ and Σ is positive definite. Define $\|u\|_{\Sigma^{-1}} = (u^\top \Sigma^{-1} u)^{1/2}$.

Theorem 2 (Strict power gain from group-specific directions)**.** *Under the preceding model, suppose $\|\mu_k\|_{\Sigma^{-1}} = a > 0$ for every k, and define $T_L(Z, W) = \mu_Z^\top \Sigma^{-1} W / a$. Let $\mathcal{S}_{\mathrm{lin}}$ consist of all constant scores and all strictly increasing transformations of $b + v^\top W$, where $b \in \mathbb{R}$ and $v \in \mathbb{R}^p$. If the group means $\mu_1, \ldots, \mu_K$ are not all identical, then* $\mathrm{Pow}_\alpha(T_L) > \sup_{S \in \mathcal{S}_{\mathrm{lin}}} \mathrm{Pow}_\alpha(S)$.

Theorem 2 illustrates how group structure improves detection: the oracle score T_L combines features along the optimal direction for each group, even though group membership provides no direct anomaly evidence on its own. Although the dominance result is relative to the best common-direction score and does not establish superiority over unrestricted global scores, it demonstrates how simple local models can collectively produce a flexible and effective scoring rule. This provides a statistical rationale for GLADE: auxiliary structure can expand the expressive capacity of simple models to capture heterogeneous anomalies, offering a practical strategy when current data are too limited to estimate a complex global score reliably. Theorem C.2 in Appendix C.3 shows that the power advantage of group-specific scoring can persist despite moderate deviations from target-optimal directions. Appendix C.2.2 also gives conditions under which merging target groups preserves a power advantage, showing that useful structural guidance need not distinguish every target category.

Structural gains versus learning costs. We quantify how learning and selection affect structural gains. Fix an anchor G in Algorithm 1. Let $\widehat{S}_\lambda$ denote the full-data candidate fit, $\widehat{G} = \widehat{S}_0$ the fitted anchor, $\widehat{\lambda}$ the selected weight, and $d_{\min} = \Delta_{\min}/N$. Let S°_λ and G° be the population scores for the specified routing and learner classes. Suppose some $\lambda^\star \in \Lambda$ provides a structural gain $\Delta^\circ_{\mathrm{struct}} = \mathrm{Pow}_\alpha(S^\circ_{\lambda^\star}) - \mathrm{Pow}_\alpha(G^\circ) > 0$. Define the limiting discovery proportion $\mathcal{Q}_\alpha(S) = \rho_0 \tau_\alpha(S) + \rho_1 \mathrm{Pow}_\alpha(S)$, estimated by $J_{G,\lambda}/N$ for each fitted candidate. On an event $\mathcal{E}_{\mathrm{learn}}$, assume the router, anchor, and local scores satisfy the approximation bounds in Theorem B.1. Suppose $S^\circ_{\lambda^\star}$ and G° satisfy Assumption B.1, with errors small enough for Lemma B.2 to apply. All fitted candidates are assumed to have continuous null distributions and nontrivial exact BH crossings under

the same P_ρ. On an event $\mathcal{E}_{\mathrm{sel}}$, assume $\max_{\lambda\in\Lambda}|J_{G,\lambda}/N - \mathcal{Q}_\alpha(\widehat{S}_\lambda)| \le \eta_{\mathrm{sel}}$. Let $\varepsilon_{\mathrm{learn}}$ denote the resulting combined power-approximation error for the candidate and anchor. Exact crossings imply $\mathrm{Pow}_\alpha(\widehat{S}_\lambda) = c_\rho \mathcal{Q}_\alpha(\widehat{S}_\lambda)$, where $c_\rho = (1-\alpha\rho_0)/\rho_1$. Set $\varepsilon_{\mathrm{select}} = c_\rho(2\eta_{\mathrm{sel}} + d_{\min})$.

Theorem 3. *Under the preceding conditions, on $\mathcal{E}_{\mathrm{learn}} \cap \mathcal{E}_{\mathrm{sel}}$,*

$$\mathrm{Pow}_\alpha(\widehat{S}_{\widehat{\lambda}}) - \mathrm{Pow}_\alpha(\widehat{G}) \ge \Delta^\circ_{\mathrm{struct}} - \varepsilon_{\mathrm{learn}} - \varepsilon_{\mathrm{select}}. \tag{8}$$

Structural transfer therefore improves asymptotic power when its population gain exceeds learning and selection costs. Appendix B.4.2 details the sources of evaluation error.

3.3 Adaptive protection against negative transfer

Misleading auxiliary information can reduce detection power: Theorem C.3 in Appendix C.3 shows that sufficiently misaligned group-specific scores can underperform the best common-direction score. We therefore show that adaptive selection can avoid negative transfer when evaluation error is sufficiently small relative to the minimum-gain threshold. For the fixed anchor, define $\mathcal{U}_\lambda = \mathcal{Q}_\alpha(\widehat{S}_\lambda)$, $\widehat{\mathcal{U}}_\lambda = J_{G,\lambda}/N$, and $g_\lambda = \mathcal{U}_\lambda - \mathcal{U}_0$. Recall that $d_{\min} = \Delta_{\min}/N$, and assume that Λ contains zero and at least one nonzero candidate.

Theorem 4 (Fallback activation and negative-transfer protection)**.** *Let $\widehat{\lambda}$ be selected by equation 5, with ties resolved in favor of smaller λ. Suppose that, for some $\eta_{\mathrm{sel}} \ge 0$ and $\delta_{\mathrm{sel}} \in [0,1]$,*

$$\mathbb{P}\left\{ \max_{\lambda\in\Lambda} |\widehat{\mathcal{U}}_\lambda - \mathcal{U}_\lambda| \le \eta_{\mathrm{sel}} \right\} \ge 1 - \delta_{\mathrm{sel}}. \tag{9}$$

Define $g_{\max} = \max_{\lambda\in\Lambda\setminus\{0\}} g_\lambda$ and $[u]_+ = \max\{u, 0\}$. On the event in equation 9, GLADE selects the anchor ($\widehat{\lambda} = 0$) whenever $g_{\max} + 2\eta_{\mathrm{sel}} < d_{\min}$. Moreover, on the same event,

$$[\mathcal{U}_0 - \mathcal{U}_{\widehat{\lambda}}]_+ \le [2\eta_{\mathrm{sel}} - d_{\min}]_+. \tag{10}$$

The uniform accuracy condition bounds the error in each estimated gain by $2\eta_{\mathrm{sel}}$. The first conclusion ensures that GLADE retains the global anchor when no local candidate can reach the minimum-gain threshold, even allowing for evaluation error. The second bounds utility loss relative to the anchor, including when a local correction is selected. On the stated event, $d_{\min} \ge 2\eta_{\mathrm{sel}}$ guarantees that the selected score has utility at least as high as the anchor, although larger thresholds may forgo modest gains. These guarantees permit inaccurate local models, provided their utilities are evaluated reliably. Appendix B.4 provides the proof and implications for power and threshold choice.

4 Simulation studies

We consider heterogeneous Gaussian anomalies in $d = 100$ dimensions. Let e_1 denote the first standard basis vector, and let $P_0 = N(0, I_d)$ denote the inlier distribution. The target outlier distribution is $P_1 = 0.8N(se_1, I_d) + 0.2N(-se_1, I_d)$, where $s > 0$ controls signal strength. Each test sample contains $m = 1{,}000$ observations drawn from $0.8P_0 + 0.2P_1$, giving an expected outlier proportion of 20%. Independently, we generate $2{,}000$ training inliers and $1{,}000$ calibration inliers from P_0. We consider two settings that preserve the auxiliary category structure while varying its distributional mismatch or sample size.

We compare five baseline detectors with their corresponding GLADE variants: OCC, trained on reference inliers alone; BIC, trained to distinguish reference inliers from auxiliary outliers; MCC, which additionally uses auxiliary category labels; ICP (Liang et al., 2024), which incorporates labeled outliers through integrative conformal calibration; and AdaDetect (Marandon et al., 2024), which contrasts reference inliers with unlabeled target observations. All methods apply BH at $\alpha = 0.10$. Each configuration is evaluated over 200 matched repetitions. We estimate average power and FDR by averaging $\mathrm{TDC}/(m_1 \vee 1)$ and FDP, respectively. Appendix D provides the data-splitting and learner specifications.

Robustness to auxiliary signal mismatch. We generate $n_{\mathrm{aux}} = 1{,}000$ labeled auxiliary outliers from $N(-se_1, I_d)$ and $N(ase_1, I_d)$, with mixture weights 0.2 and 0.8, respectively. The three columns

of Figure 2 correspond to $s \in \{2, 3, 4\}$; within each column, we vary $a \in \{1, 1.2, 1.4, 1.6, 1.8, 2\}$. At $a = 1$, the auxiliary and target outlier distributions coincide. Increasing a strengthens only the prevalent positive auxiliary category, introducing signal mismatch while leaving the target distribution unchanged. MCC performs best when auxiliary and target signals match, but its power declines sharply as the positive auxiliary component exceeds its target counterpart. G-MCC is more stable, maintaining high power at moderate and strong target signals. BIC exhibits similar sensitivity to mismatch, whereas G-BIC and G-ICP maintain or improve power as the mismatch increases. Under weak target signals, all methods have low power, suggesting that structural guidance alone offers limited gains in this regime. These results illustrate how local learning from current data can preserve power when auxiliary category structure remains informative despite differences in signal strength.

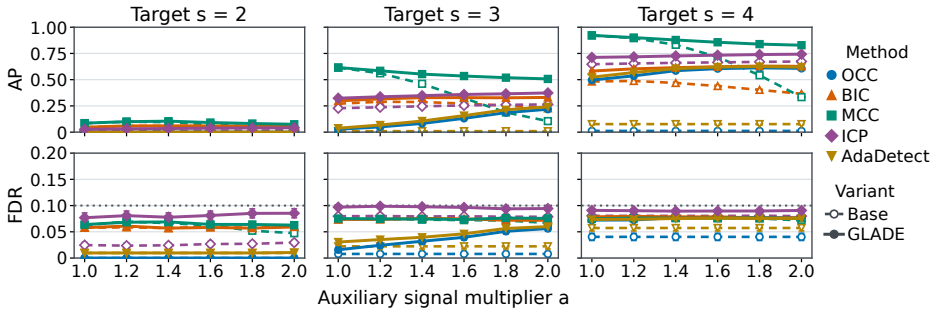

Figure 2: Robustness to auxiliary signal mismatch. Top and bottom panels report empirical AP and FDR over 200 repetitions, respectively. The dotted line marks the nominal FDR level $\alpha = 0.10$.

Benefit of additional labeled auxiliary outliers. We retain the target data-generating model and draw auxiliary outliers from a balanced mixture of $N(1.5se_1, I_d)$ and $N(-0.5se_1, I_d)$. We vary the auxiliary sample size $n_{\text{aux}} \in \{100, 500, 1{,}000\}$ and target signal strength $s \in \{1.5, 2, 2.5, 3, 3.5, 4\}$. In Figure 3, larger auxiliary samples improve GLADE's power when target signals are sufficiently strong, consistent with more reliable routing. The gains are particularly pronounced for G-OCC, G-BIC, G-ICP, and G-AdaDetect. G-MCC achieves the highest power even with a small auxiliary sample and improves further as more labeled outliers become available. With target samples matched across columns, OCC and AdaDetect remain essentially unchanged, as neither standalone method uses auxiliary labels. Under weak target signals, increasing the auxiliary sample size yields limited gains. Across both experiments, empirical FDR remains below the nominal level.

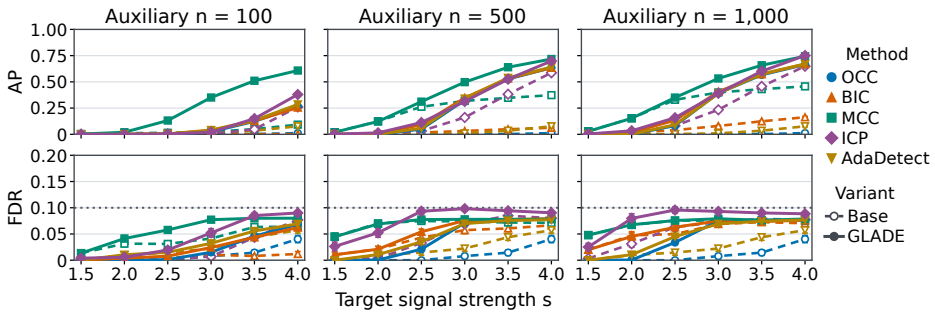

Figure 3: Effect of auxiliary sample size. Top and bottom panels report empirical AP and FDR over 200 repetitions, respectively. The dotted line marks the nominal FDR level $\alpha = 0.10$.

5 Real-data experiments

We evaluate whether auxiliary category information can improve existing detectors through target-adaptive score refinement under source–target shift. Table 1 summarizes four complementary benchmarks: UNSW-NB15 and CSE-CIC-IDS2018 study cross-domain and cross-day network

intrusion detection, Covertype considers geographic transfer across wilderness areas, and Gas Sensor Array Drift examines temporal sensor drift. These datasets cover diverse distribution shifts while retaining meaningful auxiliary category structure. All methods are evaluated over 100 matched repetitions using BH at $\alpha = 0.10$. Detailed constructions and additional comparisons are provided in Appendix E.

Table 1: Primary real-data results over 100 repetitions. Entries report mean TDC and FDP with Monte Carlo standard errors in parentheses; the largest mean TDC for each dataset is shown in **bold**.

	UNSW-NB15		Covertype		CSE-CIC-IDS2018		Gas Sensor Drift	
Method	TDC	FDP	TDC	FDP	TDC	FDP	TDC	FDP
OCC	236.31 (2.40)	0.081 (0.003)	47.30 (5.20)	0.053 (0.006)	63.54 (0.75)	0.084 (0.005)	7.45 (1.17)	0.053 (0.009)
G-OCC	1191.56 (5.97)	0.079 (0.001)	**686.56** (2.29)	0.077 (0.001)	282.62 (1.58)	0.079 (0.002)	81.84 (2.32)	0.074 (0.005)
BIC	1117.47 (6.15)	0.079 (0.001)	63.04 (2.87)	0.072 (0.005)	274.02 (1.33)	0.074 (0.002)	106.58 (0.59)	0.072 (0.004)
G-BIC	1232.59 (3.43)	0.080 (0.001)	654.25 (3.22)	0.077 (0.002)	284.02 (0.78)	0.077 (0.002)	**110.03** (0.58)	0.075 (0.004)
MCC	1142.83 (4.33)	0.080 (0.001)	321.83 (3.43)	0.076 (0.002)	285.16 (1.17)	0.077 (0.002)	101.66 (0.70)	0.070 (0.004)
G-MCC	**1249.79** (3.97)	0.080 (0.001)	651.85 (2.94)	0.077 (0.002)	**287.86** (0.99)	0.079 (0.002)	108.25 (0.83)	0.075 (0.004)
ICP	350.62 (4.51)	0.077 (0.002)	221.82 (3.42)	0.076 (0.003)	0.00 (0.00)	0.000 (0.000)	42.24 (0.91)	0.067 (0.005)
G-ICP	1211.79 (6.45)	0.080 (0.001)	666.89 (2.96)	0.078 (0.001)	283.27 (1.54)	0.079 (0.002)	83.46 (1.90)	0.081 (0.005)
AdaDetect	797.18 (4.94)	0.081 (0.001)	383.37 (10.73)	0.076 (0.002)	277.89 (1.74)	0.075 (0.002)	88.21 (2.12)	0.072 (0.004)
G-AdaDetect	1217.97 (5.54)	0.080 (0.001)	649.65 (3.08)	0.077 (0.002)	282.88 (1.19)	0.078 (0.002)	91.74 (1.90)	0.072 (0.004)

Useful structure despite weak direct transfer. Covertype provides a clear example: the five global anchors differ substantially in performance, whereas their GLADE variants are uniformly much stronger. This suggests that auxiliary category labels can retain useful structural information even when the corresponding source observations are imperfect templates for the target distribution. By using these labels to organize local learning from current data rather than directly transferring scores learned from auxiliary data, GLADE can recover substantial detection power under source–target mismatch.

Gains depend on the strength of the global anchor. Across all four datasets, every GLADE variant improves the mean TDC of its matched global anchor while maintaining mean FDP below the nominal level. The magnitude of the gain depends on how much useful structure is left unexploited by the anchor: improvements are larger when the global detector is weak or poorly aligned with the target, whereas gains are more modest when the anchor is already strong. Importantly, there is no uniformly best global anchor across datasets: MCC is strongest on UNSW-NB15 and CSE-CIC-IDS2018, AdaDetect on Covertype, and BIC on Gas Sensor Array Drift. This variation motivates an anchor-agnostic refinement framework rather than relying on a single global detector to transfer well across applications. Correspondingly, the strongest GLADE variant also varies across datasets. Overall, the results are consistent with GLADE exploiting informative auxiliary structure when it is useful, while adaptive shrinkage and utility-oriented selection limit unhelpful local refinements and help avoid negative transfer.

6 Discussion

GLADE provides a structural-transfer framework for refining existing outlier scores. Auxiliary category information guides local comparisons, while current data determine the corresponding detection scores. This separation allows simple local models to capture heterogeneous departures even when directly transferred scores are poorly aligned with current anomalies. Global–local shrinkage and the minimum-gain rule adapt the contribution of local learning to estimated discovery utility. The symmetric learning and selection procedure preserves finite-sample FDR control under the stated exchangeability assumption, while protection against negative transfer depends on reliable candidate evaluation. These principles motivate further work on structural transfer guided by downstream objectives. One direction is to integrate complementary category structures from multiple auxiliary sources. Another is to extend GLADE to sequential detection, updating structural guidance as anomaly patterns evolve. More broadly, separating transferable structure from task-specific scoring may inform methods for conformal prediction and selective decision-making.

AI USE STATEMENT

We used generative AI tools to improve the manuscript's language, clarity, and readability. The authors reviewed and revised the AI-assisted text to ensure that it accurately reflected the intended scientific content. We take full responsibility for the final manuscript, including its claims and conclusions.

ETHICS STATEMENT

This work focuses on statistical methodology and theory for out-of-distribution testing, supported by experiments on synthetic data and existing benchmark datasets. We do not identify any specific ethical concerns arising from the theoretical developments or benchmark experiments reported in this work.

REPRODUCIBILITY STATEMENT

Section 2.2 specifies the score construction and final conformal test, while Appendix A provides the complete algorithm and cross-fitting details. Sections 4 and 5, together with Appendices D and E, describe the experimental designs and data protocols. The complete code for reproducing all experiments is included in the supplementary material.

REFERENCES

Anastasios N. Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction, 2025. URL https://arxiv.org/abs/2411.11824.

Tian Bai and Ying Jin. Optimized conformal selection: Powerful selective inference after conformity score optimization. *arXiv preprint arXiv:2411.17983*, 2024. doi: 10.48550/arXiv.2411.17983. URL https://arxiv.org/abs/2411.17983.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023.

Yoav Benjamini and Ruth Heller. False discovery rates for spatial signals. *Journal of the American Statistical Association*, 102(480):1272–1281, 2007. doi: 10.1198/016214507000000941.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

Jock A. Blackard and Denis J. Dean. Covertype. UCI Machine Learning Repository, 1998. URL https://archive.ics.uci.edu/dataset/31/covertype.

Gilles Blanchard and Etienne Roquain. Two simple sufficient conditions for FDR control. *Electronic Journal of Statistics*, 2:963–992, 2008. doi: 10.1214/08-EJS180.

T. Tony Cai, Wenguang Sun, and Weinan Wang. Covariate-assisted ranking and screening for large-scale two-sample inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(2):187–234, 03 2019. ISSN 1369-7412. doi: 10.1111/rssb.12304. URL https://doi.org/10.1111/rssb.12304.

T. Tony Cai, Wenguang Sun, and Yin Xia. Laws: A locally adaptive weighting and screening approach to spatial multiple testing. *Journal of the American Statistical Association*, 117(539): 1370–1383, 2022. doi: 10.1080/01621459.2020.1859379.

Bradley Efron and Carl Morris. Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350):311–319, 1975. doi: 10.1080/01621459.1975.10479864.

Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R. Arabnia. A brief review of domain adaptation. In *Advances in Data Science and Information Engineering*, pp. 877–894, Cham, 2021. Springer International Publishing.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Saurabh Garg, Sivaraman Balakrishnan, and Zachary Lipton. Domain adaptation under open set label shift. In *Advances in Neural Information Processing Systems*, volume 35, pp. 22531–22546, 2022.

Christopher R. Genovese and Larry Wasserman. Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):499–517, 2002. doi: 10.1111/1467-9868.00347.

Christopher R. Genovese, Kathryn Roeder, and Larry Wasserman. False discovery control with p-value weighting. *Biometrika*, 93(3):509–524, 2006. doi: 10.1093/biomet/93.3.509.

Yunye Gong, Xiao Lin, Yi Yao, Thomas G. Dietterich, Ajay Divakaran, and Melinda Gervasio. Confidence calibration for domain generalization under covariate shift. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8958–8967, 2021.

Bruce E. Hansen. Least squares model averaging. *Econometrica*, 75(4):1175–1189, 2007. doi: 10.1111/j.1468-0262.2007.00785.x.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyxCxhRcY7.

Nikolaos Ignatiadis and Wolfgang Huber. Covariate powered cross-weighted multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(4):720–751, 2021. doi: 10.1111/rssb.12411.

Willard James and Charles Stein. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 361–379. University of California Press, 1961. URL https://projecteuclid.org/euclid.bsmsp/1200512173.

JoonHo Jang, Byeonghu Na, Dong Hyeok Shin, Mingi Ji, Kyungwoo Song, and Il-Chul Moon. Unknown-aware domain adversarial learning for open-set domain adaptation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16755–16767, 2022.

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19606–19616. IEEE, 2023.

Iain M. Johnstone. Gaussian estimation: Sequence and wavelet models. Book manuscript, draft version of September 16, 2019, 2019. URL https://imjohnstone.su.domains/GE_09_16_19.pdf.

Lihua Lei and William Fithian. Adapt: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(4):649–679, 2018.

Ang Li and Rina Foygel Barber. Multiple testing with the structure-adaptive benjamini—hochberg algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(1):45–74, 2019.

Ziyi Liang, Yanfei Zhou, and Matteo Sesia. Conformal inference is (almost) free for neural networks trained with early stopping. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 20810–20851. PMLR, 2023. URL https://proceedings.mlr.press/v202/liang23i.html.

Ziyi Liang, Matteo Sesia, and Wenguang Sun. Integrative conformal p-values for out-of-distribution testing with labelled outliers. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(3):671–693, 2024.

Chiara G. Magnani, Matteo Sesia, and Aldo Solari. Collective outlier detection and enumeration with conformalized closed testing. *Journal of the American Statistical Association*, 2026. doi: 10.1080/01621459.2026.2699463. Advance online publication.

Ariane Marandon, Lihua Lei, David Mary, and Etienne Roquain. Adaptive novelty detection with false discovery rate guarantee. *The Annals of Statistics*, 52(1):157–183, 2024.

David Mary and Etienne Roquain. Semi-supervised multiple testing. *Electronic Journal of Statistics*, 16(2):4926–4981, 2022. doi: 10.1214/22-EJS2050.

Jose G. Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530, 2012. doi: 10.1016/j.patcog.2011.06.019.

Nour Moustafa and Jill Slay. UNSW-NB15: A comprehensive data set for network intrusion detection systems. In *2015 Military Communications and Information Systems Conference (MilCIS)*, pp. 1–6, 2015. doi: 10.1109/MilCIS.2015.7348942.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

Sangdon Park, Osbert Bastani, James Weimer, and Insup Lee. Calibrated prediction with covariate shift via unsupervised domain adaptation. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3219–3229. PMLR, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(12):371–421, 2008. URL `https://www.jmlr.org/papers/v9/shafer08a.html`.

Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorbani. Toward generating a new intrusion detection dataset and intrusion traffic characterization. In *Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP)*, pp. 108–116, 2018. doi: 10.5220/0006639801080116.

John D. Storey, Jonathan E. Taylor, and David Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):187–205, 2004. doi: 10.1111/j.1467-9868.2004.00439.x.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, 2007.

Rongyi Sun, Wenguang Sun, and Zinan Zhao. Structure-adaptive conformal inference for large-scale out-of-distribution testing, 2026. URL `https://arxiv.org/abs/2605.26429`.

Wenguang Sun and T. Tony Cai. Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):393–424, 2009. doi: 10.1111/j.1467-9868.2008.00694.x.

Wenguang Sun, Brian J. Reich, T. Tony Cai, Michele Guindani, and Armin Schwartzman. False discovery control in large-scale spatial multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(1):59–83, 2015. doi: 10.1111/rssb.12064.

Alexander Vergara. Gas sensor array drift dataset. UCI Machine Learning Repository, 2012. URL https://archive.ics.uci.edu/dataset/224/gas+sensor+array+drift+dataset.

Vladimir Vovk. Testing for concept shift online. arXiv preprint arXiv:2012.14246, 2020. URL https://arxiv.org/abs/2012.14246.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, NY, 2005. doi: 10.1007/b106715.

Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11293–11302, 2019.

Yin Xia, T. Tony Cai, and Wenguang Sun. Gap: A general framework for information pooling in two-sample sparse inference. *Journal of the American Statistical Association*, 115(531):1236–1250, 2020. doi: 10.1080/01621459.2019.1611585.

Wen Zhang, Lingfei Deng, Lei Zhang, and Dongrui Wu. A survey on negative transfer. *IEEE/CAA Journal of Automatica Sinica*, 10(2):305–329, 2022.

Zinan Zhao and Wenguang Sun. False discovery rate control for structured multiple testing: Asymmetric rules and conformal q-values. *Journal of the American Statistical Association*, 120(550): 805–817, 2025.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2021. doi: 10.1109/JPROC.2020.3004555.

A METHODOLOGICAL DETAILS AND SUPPLEMENTARY COMPARISONS

This appendix gives the complete GLADE algorithm, its cross-fitted evaluation, the global-anchor variants, and an example illustrating the importance of accounting for adaptation before calibration.

A.1 ALGORITHM AND WORKFLOW

Algorithm 1 implements the three-step procedure in Section 2.2 for a fixed anchor specification. The main text defines the GLADE score, utility objective, and selection rule; this appendix specifies the cross-fitting construction used to evaluate the candidate scores and their discovery utility. Figure 4 gives a schematic overview.

The implementation follows the construction rules in Appendix B.1.1: pooled folds do not use calibration/test roles, fitting is invariant to training-row order, and evaluations are equivariant to query-row order. Proposition B.1 verifies the resulting symmetry of learning, selection, and refitting.

Algorithm 1 GLADE for a fixed global anchor

Input: $\mathcal{D}_{\mathrm{tr}}, \mathcal{D}_{\mathrm{cal}}, \mathcal{D}_{\mathrm{test}}, \mathcal{D}_{\mathrm{aux}}$; global anchor G; grid Λ with $0 \in \Lambda$; FDR level α; folds V; repeats B; κ_{se} and Δ_{min}.

Output: Selected $\widehat{\lambda}_G$, final score S_G^{GLADE}, conformal p-values, and discoveries $\widehat{\mathcal{R}}$.

1. **Learn structural routing.** Fit the router A on $\mathcal{D}_{\mathrm{aux}}$ and form $\mathcal{D}_{\mathrm{mix}} = \mathcal{D}_{\mathrm{cal}} \cup \mathcal{D}_{\mathrm{test}}$.
2. **Cross-fit candidate scores.** For each repeat, partition $\mathcal{D}_{\mathrm{tr}}$ and the unordered $\mathcal{D}_{\mathrm{mix}}$ into V folds, using the same folds for all $\lambda \in \Lambda$. For each fold, fit the global anchor and local target–reference classifiers on the remaining folds, construct $S_{G,\lambda}$ according to equation 2–equation 3 for all $\lambda \in \Lambda$, and evaluate the candidate scores on the held-out observations.
3. **Evaluate and select.** From the concatenated held-out scores, compute the pseudo-discovery counts $R_{G,\lambda}^{(1)}, \ldots, R_{G,\lambda}^{(B)}$, evaluate $J_{G,\lambda}$ by equation 4, and select $\widehat{\lambda}_G$ by equation 5.
4. **Refit.** Using the full $\mathcal{D}_{\mathrm{tr}}$ and unordered $\mathcal{D}_{\mathrm{mix}}$, refit the global anchor and local classifiers and form $S_G^{\mathrm{GLADE}} = S_{G,\widehat{\lambda}_G}^{\mathrm{full}}$.
5. **Calibrate and test.** Compute the final conformal p-values by equation 6 and return $\widehat{\mathcal{R}} = \mathrm{BH}_\alpha(p_1, \ldots, p_m)$.

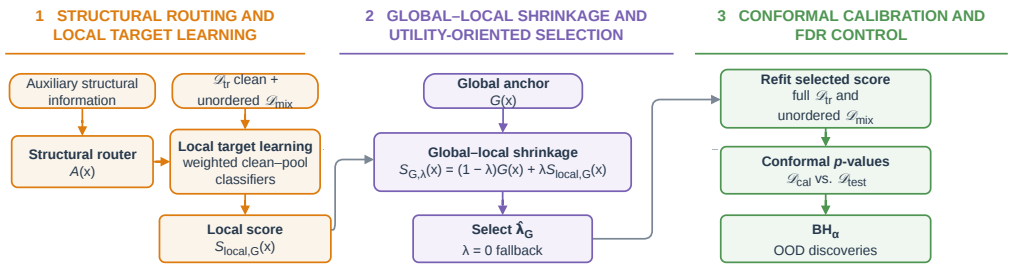

Figure 4: The three steps of GLADE for a chosen global anchor.

A.2 UTILITY-ORIENTED SELECTION: RATIONALE AND IMPLEMENTATION

This section details the utility-oriented selection component of Step 2 in Section 2.2. The main text defines the candidate objective $J_{G,\lambda}$ in equation 4 and the final selection rule in equation 5. Here we specify how the cross-fitted pseudo-discovery counts $R_{G,\lambda}^{(b)}$ are constructed and interpret the resulting utility proxy. We treat the final conformal calibration in Step 3 separately in Appendix B.1.2.

Selection objective and prior work. The shrinkage parameter determines how much local flexibility is used in the final test. Selecting it therefore requires both evaluating downstream discovery utility and preserving the symmetry needed for valid inference. Conventional classification loss need not rank candidate scores by their eventual detection power, while directly choosing the most favorable

final testing result can introduce selection bias. Related conformalized model-selection ideas have been developed for adaptive or selected OOD scores (Marandon et al., 2024; Liang et al., 2024), early stopping and model selection (Liang et al., 2023), collective outlier detection (Magnani et al., 2026), and conformity-score optimization for FDR-controlled selection (Bai & Jin, 2024). GLADE uses this perspective to evaluate the strength of the local correction according to current discovery evidence while retaining the global anchor as an exact candidate.

Preserving symmetry during selection. The calibration/test roles within $\mathcal{D}_{\mathrm{mix}}$ are withheld from fold construction, fitting, evaluation, and refitting. Conditional on $\mathcal{D}_{\mathrm{tr}}$, $\mathcal{D}_{\mathrm{aux}}$, and the algorithmic randomization, permuting the pooled observations must carry their fold assignments. The fitted models must be invariant to training-row order, and their evaluations must follow any permutation of the query observations.

Consequently, a permutation of the pooled observations merely permutes the held-out score and pseudo-p-value vectors. The associated BH rejection counts, their variability penalty, and the fixed tie-breaking and minimum-gain rules remain unchanged. Using common folds ensures that candidate scores are compared on the same observations, but common folds or cross-fitting alone do not guarantee the required symmetry. Definition B.1 states the full requirement, and Proposition B.1 verifies it for Algorithm 1.

Cross-fitted utility evaluation. For repeat $b = 1, \ldots, B$, partition $\mathcal{I}_{\mathrm{tr}}$ and $\mathcal{I}_{\mathrm{mix}}$ into V disjoint folds. The same partitions are used for every $\lambda \in \Lambda$, and the pooled folds are constructed without using calibration/test roles. The router, learned from the independent auxiliary sample $\mathcal{D}_{\mathrm{aux}}$, is held fixed.

Suppressing the repeat index temporarily, let $v(i)$ denote the fold containing index i, and write $\mathcal{I}_h^{(-v)} = \mathcal{I}_h \setminus \mathcal{I}_h^{(v)}$ for $h \in \{\mathrm{tr}, \mathrm{mix}\}$. Refit the anchor and its local learners using $\mathcal{D}_h^{(-v)} = (X_i)_{i \in \mathcal{I}_h^{(-v)}}$, with $\mathcal{D}_{\mathrm{aux}}$ available to anchors that use it. For each held-out observation, record $\widetilde{S}_{G,\lambda,i} = S_{G,\lambda}^{(-v(i))}(X_i)$, so that every observation is evaluated by a fit excluding its own fold. Collect the resulting scores in index order as $\widetilde{\mathbf{S}}_{G,\lambda}^{\mathrm{tr}} = (\widetilde{S}_{G,\lambda,i})_{i \in \mathcal{I}_{\mathrm{tr}}}$ and $\widetilde{\mathbf{S}}_{G,\lambda}^{\mathrm{mix}} = (\widetilde{S}_{G,\lambda,i})_{i \in \mathcal{I}_{\mathrm{mix}}}$.

For each pooled index $i \in \mathcal{I}_{\mathrm{mix}}$, define the pseudo-$p$-value

$$\widetilde{p}_{G,\lambda,i} = \frac{1 + \sum_{a \in \mathcal{I}_{\mathrm{tr}}} \mathbf{1}\{\widetilde{S}_{G,\lambda,a} \geq \widetilde{S}_{G,\lambda,i}\}}{n_{\mathrm{tr}} + 1}. \tag{11}$$

For repeat b, the corresponding pseudo-discovery count is $R_{G,\lambda}^{(b)} = |\,\mathrm{BH}_\alpha((\widetilde{p}_{G,\lambda,i}^{(b)})_{i \in \mathcal{I}_{\mathrm{mix}}})|$. These quantities are used only to compare candidate scores; they are not the final inferential p-values or discoveries.

Stability adjustment and minimum-gain selection. The repeated pseudo-discovery counts are aggregated using the stability-adjusted objective in equation 4. Its variability penalty discounts candidate improvements that are highly sensitive to the cross-fitting splits. Because repeated splits reuse the same observations, this penalty should be interpreted as a measure of split sensitivity rather than as a confidence bound based on independent replications.

We then select the final mixing parameter according to equation 5. A nonzero local contribution is retained only when its estimated improvement over the global anchor clears the prespecified margin Δ_{min}. The two safeguards play complementary roles: the variability penalty discounts unstable candidate performance, whereas the minimum-gain rule provides an explicit fallback to the anchor. Theorem 4 formalizes their implications for negative-transfer protection.

Interpreting discovery utility and its accuracy. At a fixed FDR target, rejection counts provide an observable selection criterion even though the current anomaly labels are unavailable. They are not true-discovery counts and are not, in general, unbiased estimates of average power. For the final discovery set, write $R = |\widehat{\mathcal{R}}|$. The identity $\mathrm{TDC} = R(1 - \mathrm{FDP})$ shows that controlling $\mathbb{E}[\mathrm{FDP}]$ does not by itself make maximizing $\mathbb{E}[R]$ equivalent to maximizing average power. The theoretical connection instead arises through the large-m analysis: under the exact-crossing conditions for

equation 23, asymptotic power is proportional to the limiting rejection proportion for a fixed test composition.

The normalized empirical criterion $J_{G,\lambda}/N$ therefore serves as a surrogate for the final-task discovery utility $\mathcal{Q}_\alpha(\widehat{S}_\lambda)$. Its discrepancy includes finite-sample evaluation error, differences between fold-trained and full-data-refitted scores, the explicit variability penalty, and differences between the selection pool and the final test population. In particular, $\mathcal{D}_{\mathrm{mix}}$ contains the calibration inliers whereas the final BH procedure acts only on $\mathcal{D}_{\mathrm{test}}$.

Appendix B.4.2 separates these error sources, and Proposition B.2 gives sufficient conditions for consistent candidate evaluation. The pseudo-p-values and pseudo-discoveries are used only for selection. Final FDR control relies instead on Assumption 1, pooled permutation equivariance, and the final calibration in Step 3; it does not require accurate utility estimation.

Implementation conventions. The fitted routing function A is denoted by $\widehat{A}$ in the routing-error analysis. For fold notation, write $\mathcal{D}_h^{(b,v)} = (X_i)_{i\in\mathcal{I}_h^{(b,v)}}$ and $\mathcal{D}_h^{(b,-v)} = (X_i)_{i\in\mathcal{I}_h^{(b,-v)}}$ for $h \in \{\mathrm{tr}, \mathrm{mix}\}$, and suppress the repeat index b when only one repeat is being considered.

The pooled fold rule must satisfy the symmetry requirements above. In particular, partitioning designated calibration and test subsets separately would use their withheld roles and would not satisfy the required construction.

We use $V = 3$ folds in the numerical studies. The corresponding experimental protocols specify B, Λ, κ_{se}, and $\Delta_{\min}$. For completeness, BH_α applied to M values $(u_1, \ldots, u_M)$ orders them as $u_{(1)} \le \cdots \le u_{(M)}$ and sets $r^\star = \max\{r \in \{1, \ldots, M\} : u_{(r)} \le \alpha r/M\}$, with $r^\star = 0$ if the set is empty. It then rejects $\{i : u_i \le \alpha r^\star/M\}$. The selection task uses $M = N$, whereas the final OOD testing task uses $M = m$.

A.3 Comparison with Marandon et al. (2024) and Liang et al. (2024)

AdaDetect (Marandon et al., 2024) and integrative conformal p-values (ICP,Liang et al., 2024) are closely related to GLADE, but they assign different roles to the available data. AdaDetect learns anomaly evidence from the current unlabeled population, whereas ICP integrates evidence from labeled inliers and outliers. GLADE connects these two perspectives by using auxiliary category labels to guide local comparisons while learning the corresponding detection scores from current data.

AdaDetect: pooled target-adaptive learning. AdaDetect (Marandon et al., 2024) learns a score by distinguishing reference training inliers from a pooled sample containing calibration inliers and unlabeled test observations. This positive–unlabeled (PU) construction adapts the score to the current alternatives without requiring labeled examples of them. Its basic score-learning procedure does not use auxiliary category labels to guide local comparisons. Indeed, the authors identify combining labeled historical outliers with target-adaptive learning under nonstationarity as a direction for further development. This is not a fundamental restriction of their general framework: auxiliary-informed scores can also be accommodated when the required permutation symmetry is preserved. GLADE provides a specific construction of this kind, using auxiliary category labels to guide local target–reference comparisons. The distinction is therefore not that a pooled classifier cannot represent heterogeneous alternatives in principle. Rather, auxiliary-guided localization provides a useful inductive bias when a single fitted learner struggles to recover several different or unequally represented anomaly mechanisms from finite data.

ICP: integration through outlier-informed weighting. ICP (Liang et al., 2024) uses labeled outliers to complement inlier-based evidence. In its basic construction, preliminary quantities $\widehat{u}_0(x)$ and $\widehat{u}_1(x)$ summarize conformity to labeled inliers and outliers, respectively. The ratio $\widehat{r}(x) = \widehat{u}_0(x)/\widehat{u}_1(x)$ is then recalibrated against inliers, accounting for the dependence between the primary evidence and the outlier-informed weight. ICP also supports automatic selection among one-class and binary classifiers. Thus, its contribution goes beyond simply training a binary classifier on pooled labeled data. Importantly, the marginal validity of the resulting p-value does not require exchangeability between auxiliary and target outliers. The relevance of source–target mismatch is

instead that conformity to auxiliary outliers may become less informative or even misleading for ranking current anomalies. ICP's defining mechanism uses this conformity information for weighting and model selection; it does not construct an auxiliary-routed collection of target–reference comparisons. Its transductive calibration and cross-validation variants should therefore be distinguished from learning such local contrasts from the current unlabeled batch. For multiple testing, the original paper additionally develops conditional FDR calibration, rather than obtaining a general BH guarantee from marginal validity alone.

GLADE: structural transfer with a global fallback. GLADE uses auxiliary category labels to determine *where* to learn specialized comparisons, while the current data determine *what* departure is present. Specifically, the routing function A is learned from $\mathcal{D}_{\mathrm{aux}}$, and the kth local learner contrasts $\mathcal{D}_{\mathrm{tr}}$ with $\mathcal{D}_{\mathrm{mix}} = \mathcal{D}_{\mathrm{cal}} \cup \mathcal{D}_{\mathrm{test}}$, applying the same routing weight to observations on both sides. The fitted local scores are combined as $S_{\mathrm{local},G}(x) = \sum_{k=1}^{K_0} A_k(x)\widehat{g}_{G,k}\{Z_G(x)\}$. Consequently, auxiliary outlier observations need not remain unchanged prototypes of current anomalies. For example, an auxiliary category can remain useful for local fitting even when its signal strength or within-region distribution changes, as long as the learned router still exposes informative structure.

Because auxiliary structure can also be unhelpful, GLADE combines the local component with a global anchor through $S_{G,\lambda}(x) = (1-\lambda)G(x) + \lambda S_{\mathrm{local},G}(x)$, with $\lambda = 0$ included as an explicit fallback. Both AdaDetect and ICP already allow data-adaptive model selection; the additional purpose of selection here is to determine how much auxiliary-guided localization the current data support. Repeated cross-fitted evaluation, the split-instability penalty, and the minimum-gain rule discourage local corrections that do not provide a sufficiently stable improvement. The resulting negative-transfer protection depends on the chosen anchor and on the accuracy of the selection criterion, as our theoretical analysis makes explicit. It is not an unconditional guarantee of improvement, nor does it remove source–target mismatch already present in a source-trained anchor.

B Main theoretical results and proofs

This appendix collects the supporting conditions and proofs for Theorems 1–4, following their order in Section 3. Complementary population and Gaussian results are given with their proofs in Appendix C.

B.1 Finite-sample FDR control

We verify pooled permutation equivariance, establish the conformal–BH argument for Theorem 1, and conclude with an example illustrating why adaptation must precede calibration.

B.1.1 Symmetry of the algorithmic construction

Write the pooled observations in index order as $\mathbf{X}_{\mathrm{mix}} = (X_i)_{i\in\mathcal{I}_{\mathrm{mix}}}$, with $N = n_{\mathrm{cal}} + m$. Calibration/test roles are withheld during score construction and used only for the final p-values. The following definition formalizes the symmetry of this construction.

Definition B.1 (Pooled permutation equivariance)**.** *Let U collect any internal randomization used by the learning and cross-fitting algorithms, independently of the data. Conditional on $(\mathcal{D}_{\mathrm{tr}}, \mathcal{D}_{\mathrm{aux}}, U)$, let*

$$\mathfrak{A}(\mathbf{x}) = \big\{\widehat{\vartheta}(\mathbf{x}), \mathbf{S}(\mathbf{x})\big\}, \qquad \mathbf{S}(\mathbf{x}) = (S_1(\mathbf{x}), \ldots, S_N(\mathbf{x})),$$

denote the complete selection-and-refitting map, where $\widehat{\vartheta}$ is the selected λ for the fixed anchor. The procedure is pooled permutation equivariant *if, for every permutation π of $\{1,\ldots,N\}$, $\widehat{\vartheta}(\pi\mathbf{x}) = \widehat{\vartheta}(\mathbf{x})$, $\mathbf{S}(\pi\mathbf{x}) = \pi\mathbf{S}(\mathbf{x})$. Here $\pi\mathbf{x}$ and $\pi\mathbf{S}$ denote the correspondingly permuted input and score vectors. For a transductive learner, this definition concerns the returned score vector and does not require the existence of a pooled-data independent scalar score function.*

The definition concerns score construction before any final randomized tie-breaking. Algorithm 1 enforces it through the following component rules, with all score-construction randomness fixed:

(D1) fit the structural router without using calibration/test identities, and evaluate it rowwise;

(D2) for every repeat, use a permutation-equivariant pooled fold rule, with the same assignment for all λ candidates;

(D3) use anchor and local learners that are invariant to training-row order and whose evaluations are equivariant to query-row order;

(D4) compute candidate objectives from the held-out BH counts and cross-split variability; apply the minimum-gain fallback and the fixed smaller-λ tie rule to these objectives; and

(D5) after selection, refit using $\mathcal{D}_{\mathrm{mix}}$ as an unordered sample and evaluate the fitted score rowwise on the pool.

These are implementation requirements, not distributional assumptions. In particular, hiding calibration/test roles alone is insufficient if fold assignment or fitting still depends on the arbitrary input-row order. The rules permit any learning algorithm satisfying these requirements; the result does not presume them for unrestricted software implementations.

Proposition B.1 (Symmetry of the GLADE construction)**.** *Algorithm 1, implemented using rules (D1)–(D5), is pooled permutation equivariant in the sense of Definition B.1.*

Proof of Proposition B.1. Let π be a permutation of the pooled observations. By rule (D2), permuting $\mathcal{D}_{\mathrm{mix}}$ merely permutes the fold labels attached to its observations. In each fold, the training sample is therefore unchanged as an unordered multiset, while the held-out query rows are permuted. Rules (D1) and (D3) imply that the corresponding global and local out-of-fold score vectors are permuted in the same way. The out-of-fold reference-inlier score vector is unchanged, because π acts only on $\mathcal{D}_{\mathrm{mix}}$.

Pseudo-conformalization compares each permuted pooled score with the same reference-inlier reference multiset. Hence the pseudo-p-value vector is permuted, while its BH rejection count is unchanged. This holds for every repeat and every λ candidate. By rule (D4), all cross-fitted objectives are therefore unchanged, so the selected λ and the minimum-gain decision are unchanged.

Finally, rule (D5) implies that the selected full-data fit is identical after permuting its pooled training rows, and that its vector of pooled evaluations is permuted by π. This is precisely the pooled permutation equivariance required by Definition B.1. □

B.1.2 Conformal calibration and proof of FDR control

Step 3 calibrates the selected, full-data-refitted score. The cross-fitted pseudo-p-values used in Step 2 serve only to choose the candidate and are not used for the final testing decisions. To justify the final conformal comparisons, the symmetry requirement must cover the complete construction, including selection and refitting. Definition B.1 and Proposition B.1 ensure that a permutation of the pool leaves the selected parameter unchanged and merely permutes the final scores. Under Assumption 1, this transfers inlier exchangeability to the null scores. Symmetric tie handling and Lemma B.1 then justify the conformal p-values and BH decisions in Step 3, yielding Theorem 1.

Symmetric treatment of ties. When ties occur, draw i.i.d. marks $E_i \sim \mathrm{Unif}(0,1)$ for pooled observations, independently of the data and score-construction randomness, and compare pairs (S_i, E_i) lexicographically. Thus a mark is used only to break a score tie. Marks travel with their observations under permutation; they are not conditioned on in the exchangeability argument. Equivalently, replace the pairs by their distinct pooled ranks. This preserves all strict score comparisons and conditional exchangeability of the null score coordinates while removing ties. The final comparisons in equation 6 are interpreted in this way when ties are present.

From exchangeable scores to conformal–BH. The following result allows scores to depend on all calibration and test observations. Its key requirement is conditional exchangeability of the null scores, which the GLADE construction will supply.

Lemma B.1 (Adaptive conformal–BH)**.** *Let $(T_1, \ldots, T_{n_{\mathrm{cal}}}, T_{n_{\mathrm{cal}}+1}, \ldots, T_{n_{\mathrm{cal}}+m})$ be a vector of real-valued scores. Let $\mathcal{H}_0$ and $\mathcal{H}_1$ index the null and non-null test coordinates. Suppose that, conditionally on the non-null test scores, $\big(T_1, \ldots, T_{n_{\mathrm{cal}}}, T_{n_{\mathrm{cal}}+j} : j \in \mathcal{H}_0\big)$ is exchangeable and that*

the complete score vector has no ties almost surely. For $j = 1, \dots, m$, define

$$p_j = \frac{1 + \sum_{i=1}^{n_{\mathrm{cal}}} \mathbf{1}\{T_i > T_{n_{\mathrm{cal}}+j}\}}{n_{\mathrm{cal}} + 1}. \tag{12}$$

Then the null p-values are super-uniform and the family $(p_1, \dots, p_m)$ has positive regression dependence on the subset $\mathcal{H}_0$ (PRDS). Consequently, BH at level α satisfies

$$\mathrm{FDR} \le \alpha \frac{|\mathcal{H}_0|}{m}. \tag{13}$$

The same conclusion holds with the weak inequality in equation 12 after applying an independent, permutation-equivariant tie-breaking rule.

Proof of Lemma B.1. Under the stated conditional exchangeability and no-ties assumptions, the empirical conformal p-values in equation 12 are super-uniform and PRDS on the null coordinates; this is the adaptive-score conformal result of Marandon et al. (2024), extending the fixed-score result of Bates et al. (2023). The standard BH theorem under PRDS then yields equation 13 conditionally on the non-null scores. Taking expectations over the conditioning variables proves the unconditional claim. Independent symmetric tie-breaking reduces the case with ties to the no-ties case without changing exchangeability. □

Proof of the main validity theorem.

Proof of Theorem 1. Let $\mathcal{F}$ be the sigma-field generated by $\mathcal{D}_{\mathrm{tr}}$, $\mathcal{D}_{\mathrm{aux}}$, the score-construction randomness U, the test-state vector, and all non-null test observations. Final tie-breaking marks are not included in $\mathcal{F}$. Conditional on $\mathcal{F}$, the calibration observations and true-null test observations are exchangeable by Assumption 1 and independence of U. Apply the complete GLADE map to the full pooled sample. Proposition B.1 establishes its permutation equivariance. Consequently, any permutation of the calibration and true-null test observations, leaving the non-null coordinates fixed, induces the same permutation of their final scores and leaves the non-null score coordinates unchanged. Therefore

$$\big(S_G^{\mathrm{GLADE}}(X_{n_{\mathrm{tr}}+1}), \dots, S_G^{\mathrm{GLADE}}(X_n), S_G^{\mathrm{GLADE}}(X_{n+j}) : j \in \mathcal{H}_0\big)$$

is exchangeable conditionally on $\mathcal{F}$ and on the non-null final scores. The independent tie-breaking marks preserve this symmetry. The resulting distinct rank-score vector is likewise exchangeable in its null coordinates conditional on $\mathcal{F}$ and its non-null coordinates. Lemma B.1, applied to the resulting rank comparisons in the final conformal p-values in equation 6, gives

$$\mathbb{E}\Big[\mathrm{FDP}\big(\widehat{\mathcal{R}}_G^{\mathrm{GLADE}}\big) \mid \mathcal{F}\Big] \le \alpha \frac{m_0}{m}.$$

Taking expectations over $\mathcal{F}$ yields the stated FDR bound. □

B.1.3 Why observation-dependent routing can distort null evidence

The preceding validity argument calibrates the complete adaptive score. The following example shows why applying observation-dependent routing after calibration can distort null evidence instead.

Example B.1 (Gaussian null distortion)**.** *Suppose that, under an all-null model,*

$$(T, B) \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right), \qquad 0 < \rho < 1,$$

where T is a primary statistic and B is a scalar auxiliary routing score. The one-sided p-value $p = 1 - \Phi(T)$ is marginally uniform under the null. However, after selecting the auxiliary-positive region $\{B > c\}$, for $0 < u < 1/2$ and $c \ge 0$, $\mathbb{P}(p \le u \mid B > c) = \frac{\mathbb{P}\{T \ge z_u, B > c\}}{\mathbb{P}(B > c)} > u$, $z_u = \Phi^{-1}(1-u)$. Indeed, positive correlation implies $\mathbb{P}(B > c \mid T \ge z_u) > \mathbb{P}(B > c)$, whereas $\mathbb{P}(T \ge z_u) = u$. Thus the marginally valid p-value becomes anti-conservative inside the auxiliary-selected region.

Example B.1 is diagnostic rather than a model assumption for GLADE. It shows that an observation-dependent router cannot in general be appended to already calibrated target p-values as though it were independent side information. Structural adaptation must instead be included in constructing the detection score, and calibration must be performed after that adaptation.

B.2 Structural gain from group-specific directions

We first specify the limiting BH threshold and asymptotic power, then prove Theorem 2. The extension to source–target directional mismatch is developed in Appendix C.3.

B.2.1 Limiting rejection threshold and asymptotic power

The quantities $\tau_\alpha(S)$ and $\mathrm{Pow}_\alpha(S)$ in Section 3.2 are defined using the exact null distribution for a fixed score. Their interpretation as large-m limits follows the threshold analysis of Genovese & Wasserman (2002) and the empirical-process formulation of Storey et al. (2004). The test states may be fixed, with $m_0/m \to \rho_0 \in (0,1)$; $P_\rho = \rho_0 P_0 + \rho_1 P_1$ records limiting proportions rather than a prior on those states.

Let $\widehat{F}_{0,m}$ and $\widehat{F}_{1,m}$ be the empirical distribution functions of $p_S(X_i)$ among null and non-null tests. A sufficient setting is $\sup_{q\in[0,1]} |\widehat{F}_{0,m}(q) - q| \longrightarrow 0$, $\sup_{q\in[0,1]} |\widehat{F}_{1,m}(q) - F_{1,S}(q)| \longrightarrow 0$, in probability, with $F_{1,S}$ continuous. These conditions hold, for example, for independent samples within the null and alternative classes. Let $F_S(q) = \rho_0 q + \rho_1 F_{1,S}(q)$ and $\tau = \tau_\alpha(S) \in (0,1)$. To ensure stability, suppose that every left neighborhood of τ contains a q with $\alpha F_S(q) - q > 0$, whereas $\alpha F_S(q) - q < 0$ for every $q > \tau$. Uniform convergence then implies that the empirical BH cutoff $\widehat{\tau}_m = \alpha|\widehat{\mathcal{R}}_m|/m$ converges to τ. Indeed, a strictly feasible point just below τ bounds the cutoff from below, and continuity gives uniform separation from feasibility on each compact interval bounded away to its right. Consequently, $\frac{|\widehat{\mathcal{R}}_m \cap \mathcal{H}_1|}{m-m_0} \longrightarrow \mathrm{Pow}_\alpha(S)$, $\frac{|\widehat{\mathcal{R}}_m|}{m} \longrightarrow \mathcal{Q}_\alpha(S)$. The bounded first ratio also converges in expectation, so $\mathrm{Pow}_\alpha(S)$ is the large-m limit of average power. At an exact crossing, $\tau = \alpha\mathcal{Q}_\alpha(S)$, which yields equation 23. A constant score has $p_S(x) = 1$ for every x and yields no rejections for $\alpha < 1$; its asymptotic power is therefore zero.

For a random fitted score, we evaluate the same functionals conditional on the fitted rule, using P_ρ and the exact null tail. Under the preceding convergence and crossing conditions, they characterize performance on an independent test sequence with that composition. Interpreting them as limits for a reused test sample requires additional control of learning and calibration error. In particular, $m \to \infty$ alone does not replace a finite calibration sample by the exact null law. One sufficient route is uniform approximation of the empirical conformal p-values by $p_S(X_i)$, together with the convergence and crossing conditions above. The Gaussian comparisons in Theorems 2, C.2, and C.3 use exact null calibration; the adaptive conformal procedure has the separate finite-sample FDR guarantee in Theorem 1.

B.2.2 Gaussian projections and proof

For $q \in (0,1)$, let $z_q = \overline{\Phi}^{-1}(q)$. Under the model in equation 7, define $p_L(Z,W) = \overline{\Phi}\{T_L(Z,W)\}$. Its alternative distribution is

$$F_{1,L}(q) = \overline{\Phi}(z_q - a). \tag{14}$$

Let $\tau_L = \sup\{q : q \le \alpha[\rho_0 q + \rho_1 F_{1,L}(q)]\}$, $\mathbb{S}_\Sigma = \{v : v^\top \Sigma v = 1\}$. The uniform restricted-class gap at τ_L is

$$\underline{\Delta}_{\mathrm{lin},K} = F_{1,L}(\tau_L) - \max_{v\in\mathbb{S}_\Sigma} \sum_{k=1}^{K} \pi_k \overline{\Phi}(z_{\tau_L} - v^\top \mu_k).$$

Proof of Theorem 2. Under $H = 0$, $T_L \mid Z = k \sim N(0,1)$, while under $H = 1$, $T_L \mid Z = k \sim N(a,1)$. Thus p_L is uniform under the null and equation 14 holds. Because the group proportions are the same under the null and alternative, the joint likelihood ratio is

$$\begin{aligned}\frac{\mathrm{d}P_1}{\mathrm{d}P_0}(Z,W) &= \exp\left\{\mu_Z^\top \Sigma^{-1} W - \frac{1}{2}\|\mu_Z\|^2_{\Sigma^{-1}}\right\} \\ &= \exp\{aT_L(Z,W) - a^2/2\}.\end{aligned}$$

It is strictly increasing in T_L. Therefore, at every nontrivial null size q, thresholding T_L gives the unique Neyman–Pearson rejection region, up to a null boundary.

Every nonconstant common-slope score has a rejection region of the form $B_{v,c} = \{(z,w) : v^\top w \ge c\}$, with the same normal vector in every group. The likelihood-ratio region at null size q is

$B_L(q) = \{(z, w) : \mu_z^\top \Sigma^{-1} w \geq a z_q\}$. If these two regions agreed P_0-almost surely for some $q \in (0, 1)$, their halfspaces would agree in every group of positive probability. Hence all μ_k would be positive multiples of one common normal vector. Equal Mahalanobis norms would then force all μ_k to be identical, contrary to the hypothesis. The strict Neyman–Pearson lemma gives

$$F_{1,L}(q) > F_{1,S}(q), \qquad q \in (0, 1), \quad S \in \mathcal{S}_{\text{lin}} \text{ nonconstant.} \tag{15}$$

Mills' ratio implies

$$\frac{F_{1,L}(q)}{q} = \frac{\overline{\Phi}(z_q - a)}{\overline{\Phi}(z_q)} \longrightarrow \infty \quad \text{as } q \downarrow 0,$$

so $\tau_L > 0$; also $\tau_L < 1$ because $\alpha < 1$. Every BH threshold feasible for a common-slope score is feasible for T_L, and thus $\tau_S \leq \tau_L$. By equation 15,

$$\text{Pow}_\alpha(S) = F_{1,S}(\tau_S) \leq F_{1,S}(\tau_L) < F_{1,L}(\tau_L) = \text{Pow}_\alpha(T_L).$$

A constant score has zero nontrivial power.

For uniformity over the common-slope class, normalize every nonzero direction to $v \in \mathbb{S}_\Sigma$. Its alternative p-value CDF is $F_{1,v}(q) = \sum_{k=1}^K \pi_k \overline{\Phi}(z_q - v^\top \mu_k)$. The ellipsoid $\mathbb{S}_\Sigma$ is compact and $v \mapsto F_{1,v}(\tau_L)$ is continuous. Since the strict inequality holds at every v, the attained maximum remains strictly below $F_{1,L}(\tau_L)$. Hence $\underline{\Delta}_{\text{lin},K} > 0$, establishing the strict power gap over the entire common-direction linear class. □

B.3 Structural gains under learning and selection

This subsection collects all approximation conditions and auxiliary arguments needed for Theorem 3. We first establish stability of asymptotic power, then derive the routing and score error bound, and finally account for selection. We keep an anchor index ℓ to allow the result to apply to any fixed anchor specification: the main-text shorthand is $\widehat{S}_\lambda = \widehat{S}^{\text{full}}_{\ell,\lambda}$, $\widehat{G} = \widehat{G}^{\text{full}}_\ell$, $S^\circ_\lambda = S^\circ_{\ell,\lambda}$, and $G^\circ = G^\circ_\ell$.

B.3.1 Threshold stability and its proof

For a score S, define the score-scale upper tails $Q_{h,S}(t) = P_h\{S(X) \geq t\}$, $Q_{\rho,S}(t) = \rho_0 Q_{0,S}(t) + \rho_1 Q_{1,S}(t)$, and $\Psi_S(t) = \alpha Q_{\rho,S}(t) - Q_{0,S}(t)$. For a nonconstant score with a continuous null law, let $\mathcal{C}_\alpha(S) = \{t : \Psi_S(t) \geq 0,\ Q_{\rho,S}(t) > 0\}$, $t_\alpha(S) = \inf \mathcal{C}_\alpha(S)$, and set $\text{Pow}_\alpha(S) = Q_{1,S}\{t_\alpha(S)\}$. If the crossing set is empty, set $t_\alpha(S) = +\infty$ and $\text{Pow}_\alpha(S) = 0$. These definitions agree with the null-tail p-value formulation in Section 3.2 after the monotone change from score thresholds to null-tail p-value thresholds.

Assumption B.1 (Regular BH crossing)**.** *For a reference score S, there exist $r_S > 0$, $\kappa_S > 0$, and finite constants $M_{0,S}, M_{1,S}$ such that, with $t_S = t_\alpha(S)$ and $M_{\rho,S} = \rho_0 M_{0,S} + \rho_1 M_{1,S}$,*

(i) $\Psi_S(t_S) = 0$*;*

(ii) $S(X)$ *has a density bounded by $M_{h,S}$ under P_h, $h \in \{0, 1\}$;*

(iii)

$$\begin{aligned} &\Psi_S(t) \leq -\kappa_S(t_S - t), && t \in [t_S - r_S, t_S], \\ &\Psi_S(t) \geq \kappa_S(t - t_S), && t \in [t_S, t_S + r_S], \\ \sup_{t < t_S - r_S} &\Psi_S(t) \leq -\kappa_S r_S. \end{aligned}$$

Lemma B.2 (Stability of asymptotic power)**.** *Suppose S satisfies Assumption B.1, both $S(X)$ and $\widetilde{S}(X)$ have continuous null distributions, and $\|\widetilde{S} - S\|_\infty \leq e$. Set $d_S(e) = (M_{0,S} + \alpha M_{\rho,S})e$. If $d_S(e) < \kappa_S r_S$, then*

$$|t_\alpha(\widetilde{S}) - t_\alpha(S)| \leq \frac{d_S(e)}{\kappa_S}, \tag{16}$$

$$|\,\text{Pow}_\alpha(\widetilde{S}) - \text{Pow}_\alpha(S)| \leq C_S e, \qquad C_S = M_{1,S}\left(1 + \frac{M_{0,S} + \alpha M_{\rho,S}}{\kappa_S}\right). \tag{17}$$

Proof of Lemma B.2. The uniform score bound implies, for every t and $h \in \{0, 1\}$,

$$Q_{h,S}(t + e) \le Q_{h,\widetilde{S}}(t) \le Q_{h,S}(t - e).$$

The density bound therefore gives $\sup_t |Q_{h,\widetilde{S}}(t) - Q_{h,S}(t)| \le M_{h,S}e$, and hence

$$\sup_t |\Psi_{\widetilde{S}}(t) - \Psi_S(t)| \le d_S(e). \tag{18}$$

Write $d = d_S(e)$. For every $t < t_S - d/\kappa_S$, Assumption B.1 and equation 18 give $\Psi_{\widetilde{S}}(t) < 0$. On the other hand, for every $\eta > 0$ with $d + \eta < \kappa_S r_S$, evaluation at $t = t_S + (d + \eta)/\kappa_S$ gives $\Psi_{\widetilde{S}}(t) \ge \eta > 0$. Taking the first crossing and then letting $\eta \downarrow 0$ proves equation 16.

Finally,

$$\begin{aligned}
&\left|Q_{1,\widetilde{S}}\{t_\alpha(\widetilde{S})\} - Q_{1,S}(t_S)\right| \\
&\quad \le \left|Q_{1,\widetilde{S}}\{t_\alpha(\widetilde{S})\} - Q_{1,S}\{t_\alpha(\widetilde{S})\}\right| \\
&\qquad + \left|Q_{1,S}\{t_\alpha(\widetilde{S})\} - Q_{1,S}(t_S)\right| \\
&\quad \le M_{1,S}e + M_{1,S}|t_\alpha(\widetilde{S}) - t_S|.
\end{aligned}$$

Substitution of equation 16 proves equation 17. □

B.3.2 Routing and score approximation

The limiting threshold and power functionals are defined in Section 3.2. This result makes the learning term in Theorem 3 explicit. It evaluates the asymptotic power functional at a fitted score, conditional on that fit; it does not bound the expected true-discovery proportion in the finite conformal–BH sample used to fit and select that score.

To distinguish the learned router used by GLADE from its population target, write $\widehat{A}$ for the fitted map and A° for a comparison routing map with the same K_0 coordinates. This comparator need not recover all K current outlier categories; the structural gap measures its usefulness for the current task. Fix an anchor index ℓ. Let G°_ℓ be the population target of the global learner and let $g^\circ_{\ell,k}$ be the corresponding oracle local learner on the anchor-augmented feature $Z^\circ_\ell(x) = \{G^\circ_\ell(x), x^\top\}^\top$. Define $L^\circ_\ell(x) = \sum_{k=1}^{K_0} A^\circ_k(x) g^\circ_{\ell,k}\{Z^\circ_\ell(x)\}$, $S^\circ_{\ell,\lambda}(x) = (1 - \lambda)G^\circ_\ell(x) + \lambda L^\circ_\ell(x)$. Let $\widehat{G}^{\text{full}}_\ell$ and $\widehat{g}^{\text{full}}_{\ell,k}$ denote the corresponding full-data fits that would be obtained if (ℓ, λ) were fixed, and define $\widehat{L}^{\text{full}}_\ell(x) = \sum_{k=1}^{K_0} \widehat{A}_k(x)\widehat{g}^{\text{full}}_{\ell,k}\{Z_{\widehat{G}_\ell}(x)\}$, $\widehat{S}^{\text{full}}_{\ell,\lambda}(x) = (1 - \lambda)\widehat{G}^{\text{full}}_\ell(x) + \lambda\widehat{L}^{\text{full}}_\ell(x)$, where $Z_{\widehat{G}_\ell}(x) = \{\widehat{G}^{\text{full}}_\ell(x), x^\top\}^\top$.

Theorem B.1 (Preservation of structural gain under score estimation)**.** *Suppose that, on an event $\mathcal{E}_\ell$ of probability at least $1 - \delta_\ell$,*

$$\begin{aligned}
\sup_x \|\widehat{A}(x) - A^\circ(x)\|_1 &\le e_A, \\
\|\widehat{G}^{\text{full}}_\ell - G^\circ_\ell\|_\infty &\le e_{G,\ell}, \\
\max_{k \le K_0} \sup_x \left|\widehat{g}^{\text{full}}_{\ell,k}\{Z_{\widehat{G}_\ell}(x)\} - g^\circ_{\ell,k}\{Z^\circ_\ell(x)\}\right| &\le r_{g,\ell}, \\
\max_{k \le K_0} \sup_x \left|g^\circ_{\ell,k}\{Z^\circ_\ell(x)\}\right| &\le B_{g,\ell}.
\end{aligned}$$

Then, for every $\lambda \in [0, 1]$,

$$\left\|\widehat{S}^{\text{full}}_{\ell,\lambda} - S^\circ_{\ell,\lambda}\right\|_\infty \le e_{\ell,\lambda} := (1 - \lambda)e_{G,\ell} + \lambda\{r_{g,\ell} + B_{g,\ell}e_A\}. \tag{19}$$

Assume that $S^\circ_{\ell,\lambda}$ and G°_ℓ satisfy the regular BH-crossing condition in Assumption B.1, and that the errors are small enough for Lemma B.2 to apply. Assume also that the fitted combined and global scores have continuous null laws. Define the oracle structural gap $\Delta^\circ_{\ell,\lambda} = \text{Pow}_\alpha(S^\circ_{\ell,\lambda}) - \text{Pow}_\alpha(G^\circ_\ell)$. Then, on $\mathcal{E}_\ell$,

$$\text{Pow}_\alpha(\widehat{S}^{\text{full}}_{\ell,\lambda}) - \text{Pow}_\alpha(\widehat{G}^{\text{full}}_\ell) \ge \Delta^\circ_{\ell,\lambda} - C_{\ell,\lambda}e_{\ell,\lambda} - C_{\ell,0}e_{G,\ell}, \tag{20}$$

where the corresponding crossings of the limiting BH curve determine the constants. Thus a candidate retains a strict power advantage whenever its oracle structural gap exceeds the router, local-learning, and anchor errors.

More generally, if $\widehat{G}_\ell^{\mathrm{full}}$ belongs almost surely to a restricted global class $\mathcal{S}_\ell$, then

$$\mathrm{Pow}_\alpha(\widehat{S}_{\ell,\lambda}^{\mathrm{full}}) - \mathrm{Pow}_\alpha(\widehat{G}_\ell^{\mathrm{full}}) \geq \Delta_{\ell,\lambda}^{\mathrm{class}} - C_{\ell,\lambda} e_{\ell,\lambda}, \tag{21}$$

where $\Delta_{\ell,\lambda}^{\mathrm{class}} = \mathrm{Pow}_\alpha(S_{\ell,\lambda}^\circ) - \sup_{S\in\mathcal{S}_\ell} \mathrm{Pow}_\alpha(S)$.

Proof of Theorem B.1. Insert and subtract $\sum_k \widehat{A}_k(x) g_{\ell,k}^\circ\{Z_\ell^\circ(x)\}$. On $\mathcal{E}_\ell$,

$$\begin{aligned}
&|\widehat{L}_\ell^{\mathrm{full}}(x) - L_\ell^\circ(x)| \\
&\quad\leq \sum_{k=1}^{K_0} \widehat{A}_k(x) \left|\widehat{g}_{\ell,k}^{\mathrm{full}}\{Z_{\widehat{G}_\ell}(x)\} - g_{\ell,k}^\circ\{Z_\ell^\circ(x)\}\right| \\
&\qquad + \sum_{k=1}^{K_0} |\widehat{A}_k(x) - A_k^\circ(x)| \left|g_{\ell,k}^\circ\{Z_\ell^\circ(x)\}\right| \\
&\quad\leq r_{g,\ell} + B_{g,\ell} e_A.
\end{aligned}$$

Consequently,

$$\begin{aligned}
&\left|\widehat{S}_{\ell,\lambda}^{\mathrm{full}}(x) - S_{\ell,\lambda}^\circ(x)\right| \\
&\quad\leq (1-\lambda)|\widehat{G}_\ell^{\mathrm{full}}(x) - G_\ell^\circ(x)| + \lambda|\widehat{L}_\ell^{\mathrm{full}}(x) - L_\ell^\circ(x)| \\
&\quad\leq e_{\ell,\lambda}.
\end{aligned}$$

Taking the supremum proves equation 19. Applying Lemma B.2 to the combined and global scores yields

$$\begin{aligned}
\mathrm{Pow}_\alpha(\widehat{S}_{\ell,\lambda}^{\mathrm{full}}) &\geq \mathrm{Pow}_\alpha(S_{\ell,\lambda}^\circ) - C_{\ell,\lambda} e_{\ell,\lambda}, \\
\mathrm{Pow}_\alpha(\widehat{G}_\ell^{\mathrm{full}}) &\leq \mathrm{Pow}_\alpha(G_\ell^\circ) + C_{\ell,0} e_{G,\ell}.
\end{aligned}$$

Subtracting proves equation 20. For the restricted-class version, $\mathrm{Pow}_\alpha(\widehat{G}_\ell^{\mathrm{full}}) \leq \sup_{S\in\mathcal{S}_\ell} \mathrm{Pow}_\alpha(S)$; combining this with the combined-score lower bound proves equation 21. □

The quantity $r_{g,\ell}$ is intentionally defined at the level of the fitted local score. It may include the effect of estimating the router used as training weights, the estimation error of the anchor entering $Z_{\widehat{G}_\ell}$, and the ordinary finite-sample error of the local learner. For logistic probability scores, one may take $B_{g,\ell} \leq 1$.

B.3.3 Selection and proof of the main bound

For a fixed candidate $\lambda^\star$, let $\mathcal{E}_{\mathrm{learn}}$ be the event on which the approximation bounds above hold. The regular crossing conditions are imposed on $S_{\ell,\lambda^\star}^\circ$ and G_ℓ°, and their errors must lie in the range specified by Lemma B.2. The approximation conditions of Theorem B.1 and the regular-crossing condition of Assumption B.1 give

$$\varepsilon_{\mathrm{learn}} = C_{\ell,\lambda^\star}\left[(1-\lambda^\star)e_{G,\ell} + \lambda^\star\{r_{g,\ell} + B_{g,\ell}e_A\}\right] + C_{\ell,0}e_{G,\ell}. \tag{22}$$

The event can be the event $\mathcal{E}_\ell$ in Theorem B.1, or the intersection of separate router, anchor, and local-fitting accuracy events.

For every fitted candidate, define $\widehat{\mathcal{U}}_\lambda = J_{G_\ell,\lambda}/N$ and $\mathcal{U}_\lambda = \mathcal{Q}_\alpha(\widehat{S}_\lambda)$, where $\mathcal{Q}_\alpha(S) = \rho_0\tau_\alpha(S) + \rho_1 \mathrm{Pow}_\alpha(S)$. These functionals evaluate the fitted function under the fixed final-task law P_ρ, conditionally on that function. The uniform-selection event is $\mathcal{E}_{\mathrm{sel}} = \{\max_{\lambda\in\Lambda} |\widehat{\mathcal{U}}_\lambda - \mathcal{U}_\lambda| \leq \eta_{\mathrm{sel}}\}$. The empirical criterion evaluates an out-of-fold pseudo-task; its discrepancy from the full-data score's final-task utility includes evaluation, variability penalization, refitting, and test-composition effects. equation 9 gives this event a high-probability interpretation, while Appendix B.4.2 separates these

sources of error. The proof below is deterministic on the event and does not require an unproved accuracy rate.

For candidates with continuous null laws and nontrivial exact BH crossings under the same P_ρ, $\tau_\alpha(S) = \alpha\mathcal{Q}_\alpha(S)$ and

$$\mathrm{Pow}_\alpha(S) = c_\rho \mathcal{Q}_\alpha(S), \qquad c_\rho = \frac{1-\alpha\rho_0}{\rho_1}. \tag{23}$$

The common mixture and exact crossings ensure that the same conversion factor applies to every candidate.

Proof of Theorem 3. Let $\lambda^\dagger$ maximize $\widehat{\mathcal{U}}_\lambda$ using the prescribed tie rule. The minimum-gain rule either retains $\lambda^\dagger$ or selects zero when its estimated advantage is less than $d_{\min}$. Therefore $\widehat{\mathcal{U}}_{\widehat{\lambda}} \geq \max_\lambda \widehat{\mathcal{U}}_\lambda - d_{\min}$. On $\mathcal{E}_{\mathrm{sel}}$,

$$\begin{aligned}\mathcal{U}_{\widehat{\lambda}} &\geq \widehat{\mathcal{U}}_{\widehat{\lambda}} - \eta_{\mathrm{sel}} \\ &\geq \widehat{\mathcal{U}}_{\lambda^\star} - d_{\min} - \eta_{\mathrm{sel}} \\ &\geq \mathcal{U}_{\lambda^\star} - d_{\min} - 2\eta_{\mathrm{sel}}.\end{aligned}$$

By equation 23, $\mathrm{Pow}_\alpha(\widehat{S}_{\widehat{\lambda}}) \geq \mathrm{Pow}_\alpha(\widehat{S}_{\lambda^\star}) - \varepsilon_{\mathrm{select}}$. On $\mathcal{E}_{\mathrm{learn}}$, Theorem B.1 and equation 22 give

$$\mathrm{Pow}_\alpha(\widehat{S}_{\lambda^\star}) - \mathrm{Pow}_\alpha(\widehat{G}) \geq \Delta^\circ_{\mathrm{struct}} - \varepsilon_{\mathrm{learn}}.$$

Combining these inequalities proves equation 8. □

B.3.4 Interpretation and scope of the conditions

The routing comparison A° need not recover every target category. Its target usefulness enters through $\Delta^\circ_{\mathrm{struct}}$; its estimation error enters through e_A and the local-fitting error $r_{g,\ell}$. Routing therefore affects both the observations emphasized while fitting and the weights used to aggregate the local scores. For fixed oracle scores, shrinkage attenuates local approximation errors in equation 19, but also changes the oracle structural gain. This is the tradeoff quantified by the theorem, not a guarantee that finer routing or a larger candidate collection always helps.

The uniform bounds are convenient sufficient conditions, not generic rates for arbitrary learners. Bounded probability scores satisfy the oracle magnitude bound, but accurate estimation of them must still be established for the learner and data regime. The regular-crossing requirement prevents score perturbations from selecting an earlier, nearly feasible threshold; flatter crossings produce larger power-error constants. Uniform utility accuracy is needed because the candidate is selected from the same set whose performance is evaluated. None of these accuracy conditions follows from permutation symmetry.

The Gaussian comparison in Theorem 2 does not imply a positive structural gap against an arbitrarily flexible anchor. Its unbounded linear scores also do not automatically satisfy the bounded-score and uniform-approximation conditions above; transferring that comparison to an estimated Gaussian learner requires additional error control. Theorem 3 bounds the asymptotic power functional conditional on the fitted scores and stated accuracy events. It does not guarantee an increase in realized true discoveries in each test sample. Under Assumption 1, Algorithm 1 retains the finite-sample FDR guarantee in Theorem 1 independently of these power-accuracy events.

B.4 Adaptive protection and selection accuracy

We prove the fallback and protection guarantees in Theorem 4, then derive regret bounds and separate the sources of selection error. A consistency result and a task-mismatch counterexample clarify when reliable final-task evaluation is available.

B.4.1 Fallback, protection, and regret

Use the fixed-anchor notation of Section 3.3 and assume the finite grid Λ contains zero and at least one positive value. Write $\widehat{g}_\lambda = \widehat{\mathcal{U}}_\lambda - \widehat{\mathcal{U}}_0$ for the estimated gain. Let $\lambda^\dagger$ maximize $\widehat{\mathcal{U}}_\lambda$, with ties favoring

smaller λ. For $d_{\min} > 0$, the implemented rule selects zero exactly when $\max_{\lambda>0} \widehat{g}_\lambda < d_{\min}$. For $d_{\min} = 0$, the tie rule selects zero exactly when that maximum is nonpositive. Neither decision uses the unknown true gains.

Proof of Theorem 4. On the uniform-accuracy event, every gain estimate satisfies

$$|\widehat{g}_\lambda - g_\lambda| \le |\widehat{\mathcal{U}}_\lambda - \mathcal{U}_\lambda| + |\widehat{\mathcal{U}}_0 - \mathcal{U}_0| \le 2\eta_{\text{sel}}.$$

If $g_{\max} + 2\eta_{\text{sel}} < d_{\min}$, all positive candidates have estimated gain below the required margin, proving the stated fallback condition. If a nonzero candidate is retained, then $\widehat{g}_{\widehat{\lambda}} \ge d_{\min}$ and hence $g_{\widehat{\lambda}} \ge d_{\min} - 2\eta_{\text{sel}}$. When zero is selected, the loss is zero. Taking positive parts in the remaining case proves equation 10. □

Fallback activation versus utility protection. The stated activation condition is sufficient; fallback can also occur outside this condition, depending on the realized evaluation errors. A margin $d_{\min} > 2\eta_{\text{sel}}$ forces fallback whenever all local gains are nonpositive. The weaker condition $d_{\min} \ge 2\eta_{\text{sel}}$ ensures that the selected utility is at least the anchor utility, even if a correction is adopted. If the margin is smaller, an overestimated correction can be harmful, with loss bounded by $2\eta_{\text{sel}} - d_{\min}$. All these conclusions hold jointly on the uniform-accuracy event, whose probability is at least $1 - \delta_{\text{sel}}$.

Protection requires reliable evaluation of the fitted candidates; it does not require informative routing or accurate local score estimates. A poor local fit can therefore be discarded when its utility is assessed correctly. This separates the accuracy needed to learn a useful correction from the accuracy needed to protect against an unhelpful one.

The same argument also clarifies the strength of the accuracy assumption. It is sufficient to control gain errors directly: if $\zeta_{\text{sel}} = \max_{\lambda>0} |\widehat{g}_\lambda - g_\lambda|$, then $g_{\max} + \zeta_{\text{sel}} < d_{\min}$ forces fallback and $[\mathcal{U}_0 - \mathcal{U}_{\widehat{\lambda}}]_+ \le [\zeta_{\text{sel}} - d_{\min}]_+$. Absolute utility accuracy gives $\zeta_{\text{sel}} \le 2\eta_{\text{sel}}$, but is stronger than necessary: an additive bias common to every candidate cancels from their gains. What threatens protection is an error in their relative performance.

On the same uniform-accuracy event, the complementary oracle-regret bound is

$$\max_{\lambda\in\Lambda} \mathcal{U}_\lambda - \mathcal{U}_{\widehat{\lambda}} \le 2\eta_{\text{sel}} + d_{\min}. \tag{24}$$

If every fitted candidate has a continuous null law and a nontrivial exact crossing of the limiting BH curve under the same P_ρ, then equation 23, with $c_\rho = (1 - \alpha\rho_0)/\rho_1$, gives

$$\max_{\lambda\in\Lambda} \text{Pow}_\alpha(\widehat{S}_\lambda) - \text{Pow}_\alpha(\widehat{S}_{\widehat{\lambda}}) \le c_\rho(2\eta_{\text{sel}} + d_{\min}),$$
$$[\text{Pow}_\alpha(\widehat{G}) - \text{Pow}_\alpha(\widehat{S}_{\widehat{\lambda}})]_+ \le c_\rho[2\eta_{\text{sel}} - d_{\min}]_+.$$

Proof of the regret and power bounds. Whether or not the minimum-gain rule replaces $\lambda^\dagger$,

$$\max_{\lambda\in\Lambda} \widehat{\mathcal{U}}_\lambda - \widehat{\mathcal{U}}_{\widehat{\lambda}} \le d_{\min}.$$

Adding the two endpoint evaluation errors proves equation 24. Multiplying this bound and equation 10 by c_ρ proves the power bounds. In particular, if a fitted candidate improves anchor power by Δ^{full}, the selected improvement is at least $\Delta^{\text{full}} - c_\rho(2\eta_{\text{sel}} + d_{\min})$. □

Practical interpretation of the margin. The protection and regret bounds quantify both sides of the margin choice. Increasing $d_{\min}$ reduces possible loss against the anchor but can discard genuine improvements that are too small to clear it. GLADE applies this rule using the estimated gains $\widehat{g}_\lambda$; the theorem does not supply a certified value of η_{sel}. In particular, the implemented choice $\Delta_{\min} = 1$, or $d_{\min} = 1/N$, does not ensure $d_{\min} \ge 2\eta_{\text{sel}}$. The variability penalty measures only part of the discrepancy between the selection criterion and final-task utility. Refitting and differences in test composition can also affect candidate comparisons, as detailed below. Calibrating the margin to a rigorous error bound would require additional control of these sources of uncertainty. The bounds concern utility functionals and, under the stated crossing conditions, asymptotic power; they do not ensure that realized true discoveries never decrease in a particular test sample.

B.4.2 SOURCES OF SELECTION ERROR AND ASYMPTOTIC REGRET

The cross-fitted pseudo-p-values are used only as a model-selection device. They combine scores from different fold-trained learners and are not asserted to be final inferential p-values. This section separates the sources of error between the empirical selector and the utility of the final full-data fit.

For this fixed-anchor analysis, set $\Theta = \Lambda$ and write θ for a candidate λ. Define the unpenalized empirical cross-fitted rejection proportion and the normalized penalty by $\widehat{\mathcal{Q}}_\theta^{\mathrm{cf}} = \frac{1}{B}\sum_{b=1}^{B} \frac{R_\theta^{(b)}}{N}$, $\widehat{\Gamma}_\theta = \kappa_{\mathrm{se}} \frac{\widehat{\mathrm{se}}\{R_\theta^{(1)},\ldots,R_\theta^{(B)}\}}{N}$. Thus $\widehat{\mathcal{U}}_\theta = \widehat{\mathcal{Q}}_\theta^{\mathrm{cf}} - \widehat{\Gamma}_\theta$. Let $\mathcal{U}_\theta^{\mathrm{cf}}$ denote the population utility of the fold-trained pseudo-task obtained by replacing the empirical held-out reference-inlier and pooled score distributions by their population counterparts. Let $\mathcal{U}_\theta^{\mathrm{mix}}$ denote the limiting rejection proportion of the full-data-refitted score under the composition represented by the selection pool $\mathcal{D}_{\mathrm{mix}}$, and let $\mathcal{U}_\theta$ be the final test-task utility of the full-data-refitted score under P_ρ.

Define

$$\eta_{\mathrm{emp}} = \sup_{\theta\in\Theta} \left| \widehat{\mathcal{Q}}_\theta^{\mathrm{cf}} - \mathcal{U}_\theta^{\mathrm{cf}} \right|, \qquad \eta_{\mathrm{pen}} = \sup_{\theta\in\Theta} \widehat{\Gamma}_\theta,$$
$$\eta_{\mathrm{refit}} = \sup_{\theta\in\Theta} \left| \mathcal{U}_\theta^{\mathrm{cf}} - \mathcal{U}_\theta^{\mathrm{mix}} \right|, \qquad \eta_{\mathrm{task}} = \sup_{\theta\in\Theta} \left| \mathcal{U}_\theta^{\mathrm{mix}} - \mathcal{U}_\theta \right|.$$

The term η_{emp} includes finite reference-inlier and pooled- sample error as well as the comparability error induced by concatenating scores from different fold-trained learners. The term η_{refit} measures the difference between training on $V-1$ folds and refitting on all available observations. The term η_{task} is needed because $\mathcal{D}_{\mathrm{mix}}$ contains the additional null observations in $\mathcal{D}_{\mathrm{cal}}$, whereas the final BH procedure is applied only to $\mathcal{D}_{\mathrm{test}}$. Finally, η_{pen} records the deliberate cross-split stability penalty.

Lemma B.3 (Decomposition of the cross-fitted selection error)**.** *For every realization,*

$$\sup_{\theta\in\Theta} |\widehat{\mathcal{U}}_\theta - \mathcal{U}_\theta| \le \eta_{\mathrm{emp}} + \eta_{\mathrm{pen}} + \eta_{\mathrm{refit}} + \eta_{\mathrm{task}}. \tag{25}$$

Consequently, equation 9 holds with any η_{sel} that upper-bounds the right-hand side with probability at least $1-\delta_{\mathrm{sel}}$.

Proof of Lemma B.3. For every θ, add and subtract $\widehat{\mathcal{Q}}_\theta^{\mathrm{cf}}$, $\mathcal{U}_\theta^{\mathrm{cf}}$, and $\mathcal{U}_\theta^{\mathrm{mix}}$. Since $\widehat{\Gamma}_\theta \ge 0$, the triangle inequality gives

$$\begin{aligned} |\widehat{\mathcal{U}}_\theta - \mathcal{U}_\theta| \le{}& |\widehat{\mathcal{Q}}_\theta^{\mathrm{cf}} - \mathcal{U}_\theta^{\mathrm{cf}}| + \widehat{\Gamma}_\theta \\ &+ |\mathcal{U}_\theta^{\mathrm{cf}} - \mathcal{U}_\theta^{\mathrm{mix}}| + |\mathcal{U}_\theta^{\mathrm{mix}} - \mathcal{U}_\theta|. \end{aligned}$$

Taking the supremum over Θ proves equation 25. □

Remark B.1 (What the decomposition does and does not establish)**.** *Lemma B.3 identifies the quantities that must be controlled; it does not establish that they are small. Out-of-fold scores come from different fitted functions and share training data across folds. Repeated splits also reuse the same observations, so the split penalty is not a confidence bound based on independent replications. The learning and BH-stability results in Appendix B.3 explain how score perturbations can affect population utility, but do not themselves bound the empirical cross-fitted objective or the task-composition gap. Proposition B.2 supplies one sufficient regime for evaluation consistency. Example B.2 shows why a persistent composition gap cannot in general be ignored.*

Corollary B.1 (Vanishing oracle regret)**.** *Suppose $\eta_{\mathrm{emp}} + \eta_{\mathrm{pen}} + \eta_{\mathrm{refit}} + \eta_{\mathrm{task}} = o_{\mathbb{P}}(1)$ and $d_{\min} = o(1)$. Then the fixed-anchor oracle regret in equation 24 is $o_{\mathbb{P}}(1)$. In particular, a fixed minimum gain of one rejection gives $d_{\min} = 1/N \to 0$.*

Proof of Corollary B.1. The regret inequalities are deterministic with the actual uniform discrepancy substituted for η_{sel}. Lemma B.3 bounds that discrepancy by $o_{\mathbb{P}}(1)$. Together with $d_{\min} = o(1)$, this proves the claim. □

B.4.3 A SUFFICIENT REGIME FOR RELIABLE EVALUATION

The preceding decomposition and regret bound are conditional. The next result gives explicit sufficient conditions under which the implemented criterion consistently evaluates the final fitted candidates. It allows misspecified score functions; what matters is that fold-trained and full-data scores approach the same limits and are evaluated on compatible tasks.

Proposition B.2 (Consistency of candidate evaluation)**.** *Consider a fixed finite grid Λ, fixed numbers of folds V and repeats B, and bounded κ_{se}. Suppose that:*

(i) Reference observations are independent draws from P_0. Conditional on fixed test states, test observations are independent of one another and of the reference data, with null law P_0, alternative law P_1, and non-null fraction converging to $\rho_1 \in (0,1)$. The training, calibration, and test sample sizes tend to infinity, and $n_{\mathrm{cal}}/m \to 0$.

(ii) For each λ, all fold-trained scores $\widehat{S}_\lambda^{(b,-v)}$ and the full-data score $\widehat{S}_\lambda$ converge uniformly in probability to the same deterministic function S_λ°:

$$\max_{\lambda,b,v} \|\widehat{S}_\lambda^{(b,-v)} - S_\lambda^\circ\|_\infty + \max_{\lambda} \|\widehat{S}_\lambda - S_\lambda^\circ\|_\infty = o_{\mathbb{P}}(1).$$

(iii) Each S_λ° has a continuous null distribution, $p_{S_\lambda^\circ}(X)$ has a continuous distribution under P_1, and its BH crossing under P_ρ is nontrivial and stable in the sense specified in Appendix B.2.1.

Then

$$\max_{\lambda\in\Lambda} |\widehat{\mathcal{U}}_\lambda - \mathcal{U}_\lambda| = o_{\mathbb{P}}(1). \tag{26}$$

In particular, equation 9 holds for some sequences $\eta_{\mathrm{sel}} \to 0$ and $\delta_{\mathrm{sel}} \to 0$. If $d_{\min} \to 0$, the selected utility has vanishing regret relative to the best fitted candidate. If, in addition, for some $\gamma > 0$, $\max_{\lambda>0}\{\mathcal{Q}_\alpha(S_\lambda^\circ) - \mathcal{Q}_\alpha(S_0^\circ)\} \le -\gamma$, then $\mathbb{P}(\widehat{\lambda} = 0) \to 1$.

Proof. Fix a candidate and a repeat. The common uniform score approximation replaces each concatenated held-out score by S_λ° at the same observation with an $o_{\mathbb{P}}(1)$ error, simultaneously across folds. Let $H_{0,\lambda}^\circ$ be the continuous null score CDF. The empirical CDF of S_λ° on the reference training data converges uniformly to $H_{0,\lambda}^\circ$. Bracketing the held-out scores by $S_\lambda^\circ \pm \epsilon$ and using uniform continuity of $H_{0,\lambda}^\circ$ shows that the pseudo-p-values in equation 11 differ uniformly from $p_{S_\lambda^\circ}(X_i)$ by $o_{\mathbb{P}}(1)$. The add-one correction vanishes as the training sample size grows. This comparison is pathwise on the score-approximation event and does not assume independence of the fold-trained scores.

The fraction of alternatives in the pooled data converges to ρ_1 because $n_{\mathrm{cal}}/m \to 0$. Independence within the fixed null and alternative classes gives uniform convergence of their empirical CDFs of the exact-null p-values. Their continuous limits, together with the pseudo-p-value approximation, imply uniform convergence of the pooled pseudo-p-value CDF to $F_\lambda^\circ(q) = \rho_0 q + \rho_1 F_{1,S_\lambda^\circ}(q)$. The stable-crossing argument in Appendix B.2.1 yields

$$R_\lambda^{(b)}/N \longrightarrow \mathcal{Q}_\alpha(S_\lambda^\circ) \quad \text{in probability.}$$

This holds simultaneously over the fixed finite sets of candidates and repeats. Consequently the average normalized count has the same limit, and its cross-split penalty tends to zero.

The uniform approximation for the full-data fit also gives $\sup_x |p_{\widehat{S}_\lambda}(x) - p_{S_\lambda^\circ}(x)| = o_{\mathbb{P}}(1)$ by the same null-CDF bracketing argument. Continuity of the limiting p-value CDFs and stability of the BH crossing therefore give $\mathcal{U}_\lambda \to \mathcal{Q}_\alpha(S_\lambda^\circ)$ in probability. Comparing the two limits proves equation 26. The deterministic regret bound proves the second conclusion when $d_{\min} \to 0$. Under the strict negative gap, all estimated local gains are negative with probability tending to one, so the global candidate maximizes the criterion and is selected. □

Condition (ii) makes explicit the needed stability of both the numerical scores across folds and their full-data refits. On a bounded feature space, uniform convergence can, for example, follow from convergence of the coefficients of a fixed-dimensional, Lipschitz parametric score family. Such

coefficient convergence must still be justified for the chosen learners. The score and BH-stability tools in Appendix B.3 give quantitative perturbation bounds under their stronger regularity conditions; neither those bounds nor cross-fitting establishes learner consistency automatically. Condition (i) is one sufficient way to remove task mismatch, not a requirement for finite-sample FDR control. If the calibration/test size ratio stays positive, an alternative bound on task mismatch or direct control of gain errors is needed.

Proposition B.2 proves an asymptotic evaluation guarantee in this regime, rather than a computable finite-sample value of η_{sel}. Even vanishing evaluation error need not be smaller than the implemented $1/N$ margin. A strictly negative limiting gap does imply eventual fallback; near-zero gains instead require attention to the relative sizes of gain error and the margin.

B.4.4 Why task agreement matters

Example B.2 (Persistent mismatch between selection and testing)**.** *Take $\alpha = 1/10$ and two fixed candidate scores $S_0(X) = 1 - X_1$ and $S_1(X) = 1 - X_2$, corresponding to the endpoints $\Lambda = \{0, 1\}$ of the shrinkage family. Under P_0, let X_1 and X_2 be independent $\mathrm{Unif}(0, 1)$ variables. Under P_1, let them be independent with $X_1 \sim \mathrm{Unif}(0, 3/200)$, $X_2 \sim \frac{1}{2}\,\mathrm{Unif}(0, 1/1000) + \frac{1}{2}\,\mathrm{Unif}(0, 1)$. Thus the p-values computed from the exact null distribution are X_1 and X_2, respectively, with alternative CDFs $F_{1,0}(q) = \min\{200q/3, 1\}$ and $F_{1,1}(q) = \frac{1}{2}\min\{1000q, 1\} + q/2$. Let the final alternative fraction be $\rho_1 = 1/5$ and let $n_{\mathrm{cal}}/m \to 1$. The selection pool then has limiting alternative fraction $1/10$. Direct substitution into the limiting BH inequality gives*

	Selection pool		*Final test population*	
Candidate	*Limiting cutoff*	*Utility*	*Limiting cutoff*	*Utility*
Global S_0	0	0	$1/46$	$5/23$
Local S_1	$1/181$	$10/181$	$1/91$	$10/91$

For verification, at an alternative fraction r, the global score has a positive crossing $\alpha r/\{1-\alpha(1-r)\}$ only when it reaches $3/200$; otherwise its cutoff is zero. The local crossing, which exceeds $1/1000$ at both displayed fractions, is $(\alpha r/2)/\{1 - \alpha(1 - r/2)\}$. Utility equals the cutoff divided by α.

For independent observations and growing reference samples, the cross-fitted criterion converges to the selection-pool utilities because the scores are fixed. The zero crossing for S_0 is stable from above: its BH inequality is strictly infeasible at every positive threshold. Any $d_{\min} \to 0$, including $1/N$, therefore selects S_1 with probability tending to one. Yet its final-task power is $46/91$, compared with power one for S_0. Harmful transfer is retained despite the absence of score-learning error. The two tasks favor different rankings.

This example is a limitation of the selection surrogate, not a counterexample to Theorem 4, whose accuracy condition is not small here. Both candidates have exact uniform null p-values; the power loss does not require failure of FDR control. Together, the error decomposition, the consistency proposition, and this example specify what is needed to connect the fallback rule to reliable protection. They do not establish uniform evaluation accuracy for every black-box implementation or for the sample proportions used in every reported experiment.

C Supplementary structural-power results and proofs

This appendix complements the main theoretical analysis with results on population ranking, partial group information, and source–target directional mismatch. Each result is presented together with its proof; the conditions and proofs for the main theorems remain in Appendix B.

C.1 Population oracle and groupwise ranking equivalence

We first distinguish the role of structural routing from the unrestricted population oracle. The result below concerns a population marginal false discovery rate (mFDR) constraint, rather than finite-sample optimality of conformal–BH.

Let $P_1 = \sum_{k=1}^{K} \omega_k P_{1,k}$ be the aggregate target alternative law, where $\omega_k \geq 0$ and $\sum_k \omega_k = 1$, and let $P_\rho = \rho_0 P_0 + \rho_1 P_1$, where $\rho_0 + \rho_1 = 1$ and $\rho_0, \rho_1 > 0$. For a measurable rejection region B, define $\mathrm{mFDR}(B) = \frac{\rho_0 P_0(B)}{\rho_0 P_0(B) + \rho_1 P_1(B)}$, $\mathrm{TPR}(B) = P_1(B)$. Let f_0 and f_1 denote densities with respect to a common dominating measure and write $L^\star = f_1/f_0$. For a simplex-valued routing map $A = (A_1, \ldots, A_{K_0})$, define $\mu_{hk} = \mathbb{E}_{P_h}\{A_k(X)\}$, $h \in \{0, 1\}$, and, for $0 < \gamma < 1$, $P_\gamma = (1-\gamma)P_0 + \gamma P_1$, $\mu_{\gamma k} = (1-\gamma)\mu_{0k} + \gamma \mu_{1k}$. Whenever the normalizing constants are positive, let $f^A_{0,k}(x) = \frac{A_k(x) f_0(x)}{\mu_{0k}}$, $f^A_{\gamma,k}(x) = \frac{A_k(x) f_\gamma(x)}{\mu_{\gamma k}}$. Here $f_\gamma = (1-\gamma)f_0 + \gamma f_1$.

Proposition C.1 (Population oracle and groupwise ranking equivalence)**.** *Assume $0 < \alpha < \rho_0$, f_0 and f_1 are positive almost everywhere, and there exists $t_\alpha > 0$ such that $B^\star_\alpha = \{x : L^\star(x) \geq t_\alpha\}$ satisfies $\mathrm{mFDR}(B^\star_\alpha) = \alpha$. Then $\mathrm{TPR}(B^\star_\alpha) \geq \mathrm{TPR}(B)$ for every measurable B satisfying $\mathrm{mFDR}(B) \leq \alpha$. If $P_0\{L^\star(X) = t_\alpha\} = 0$, the inequality is strict whenever $P_0(B \triangle B^\star_\alpha) > 0$.*

Moreover, for every $k = 1, \ldots, K_0$ and every x with $A_k(x) > 0$,

$$\frac{\mu_{\gamma k}}{\mu_{0k}} \frac{f^A_{\gamma,k}(x)}{f^A_{0,k}(x)} = \frac{f_\gamma(x)}{f_0(x)} = (1-\gamma) + \gamma L^\star(x). \tag{27}$$

Hence localization does not create new pointwise likelihood-ratio information for an unrestricted population learner. Its benefit must arise through a restricted representation, improved finite-sample estimation, or both.

Proposition C.1 explains why GLADE is a structural-transfer method rather than a new population oracle. The same routing weight appears on the reference-inlier and pooled sides of each local comparison and therefore cancels from the exact within-group likelihood ratio. In finite samples, however, the router can expose group-specific directions that a single restricted learner would otherwise have to estimate simultaneously.

C.1.1 Proof

Proof of Proposition C.1. Set

$$c_\alpha = \frac{(1-\alpha)\rho_0}{\alpha \rho_1}.$$

Because $P_1(B) = \int_B L^\star \, \mathrm{d}P_0$, the constraint $\mathrm{mFDR}(B) \leq \alpha$ is equivalent to

$$\int_B \{L^\star(x) - c_\alpha\} \, \mathrm{d}P_0(x) \geq 0. \tag{28}$$

The equality $\mathrm{mFDR}(B^\star_\alpha) = \alpha$ implies

$$c_\alpha = \frac{P_1(B^\star_\alpha)}{P_0(B^\star_\alpha)} = \mathbb{E}_{P_0}\{L^\star(X) \mid X \in B^\star_\alpha\} \geq t_\alpha.$$

First suppose $t_\alpha < c_\alpha$, and define $\eta_\alpha = \frac{t_\alpha}{c_\alpha - t_\alpha}$, $h_\alpha(x) = L^\star(x) + \eta_\alpha\{L^\star(x) - c_\alpha\}$. Then $h_\alpha(x) = \frac{c_\alpha}{c_\alpha - t_\alpha}\{L^\star(x) - t_\alpha\}$, so $B^\star_\alpha$ is the nonnegative region of h_α, up to its boundary. For every feasible B,

$$P_1(B) \leq \int_B h_\alpha \, \mathrm{d}P_0 \leq \int_{B^\star_\alpha} h_\alpha \, \mathrm{d}P_0 = P_1(B^\star_\alpha).$$

The first inequality follows from equation 28; the last equality uses equality in that constraint for $B^\star_\alpha$.

If $t_\alpha = c_\alpha$, equality of the conditional mean with its lower bound implies $L^\star = c_\alpha$ almost surely on $B^\star_\alpha$, while $L^\star < c_\alpha$ on its complement. Any feasible set must therefore be contained in $B^\star_\alpha$, up to a P_0-null set. The same optimality conclusion follows. When $P_0\{L^\star(X) = t_\alpha\} = 0$, the positive and negative regions of h_α are strict, so the power inequality is strict whenever the two regions differ on a set of positive P_0-measure.

For the groupwise identity, the common routing weight cancels whenever $A_k(x) > 0$:

$$\frac{f^A_{\gamma,k}(x)}{f^A_{0,k}(x)} = \frac{A_k(x) f_\gamma(x)/\mu_{\gamma k}}{A_k(x) f_0(x)/\mu_{0k}} = \frac{\mu_{0k}}{\mu_{\gamma k}} \frac{f_\gamma(x)}{f_0(x)}.$$

Multiplying by $\mu_{\gamma k}/\mu_{0k}$ and using $f_\gamma = (1-\gamma)f_0 + \gamma f_1$ proves equation 27. The right-hand side is a strictly increasing affine function of $L^\star$. Thus each groupwise likelihood ratio has the same within-group ranking as the unrestricted oracle. Finally, since $A(X)$ is a measurable function of X, adjoining it cannot enlarge the sigma-field generated by X. □

C.2 Gaussian group coarsening

C.2.1 Geometry of the structural advantage

Theorem 2, proved in Appendix B.2, assumes full observation of group membership. We now examine how much of its structural advantage can remain when the auxiliary representation merges two target groups.

The between-group signal variation can be summarized by $V_K = \sum_{k=1}^K \pi_k \|\mu_k - \bar{\mu}\|^2_{\Sigma^{-1}} = a^2 - \|\bar{\mu}\|^2_{\Sigma^{-1}}$, $\bar{\mu} = \sum_{k=1}^K \pi_k \mu_k$. The weighted variance identity proves the equality, and strict convexity of the squared norm shows that $V_K > 0$ exactly when the component means are not all identical.

C.2.2 Power under group coarsening

We next quantify a simple form of routing mismatch in which the auxiliary representation merges two target groups. For distinct $r, s \in \{1, \dots, K\}$, define $\mathcal{P}_{K-1} = \big\{\{k\} : k \notin \{r,s\}\big\} \cup \big\{\{r,s\}\big\}$. Let $\bar{\mu} = \sum_{k=1}^K \pi_k \mu_k$. For each block $H \in \mathcal{P}_{K-1}$, let

$$\pi_H = \sum_{k \in H} \pi_k, \qquad \mu_H = \frac{1}{\pi_H} \sum_{k \in H} \pi_k \mu_k, \qquad a_H = \|\mu_H\|_{\Sigma^{-1}}. \tag{29}$$

The coarsened mean-difference learner uses direction μ_H within block H. This result concerns a population linear target rather than the exact finite-sample logistic implementation.

For an alternative null-tail p-value CDF F, write $\tau(F) = \sup\{q : q \le \alpha[\rho_0 q + \rho_1 F(q)]\}$, $\mathrm{Pow}_{\mathrm{BH}}(F) = F\{\tau(F)\}$.

Theorem C.1 (Power after merging two Gaussian groups)**.** *Assume equation 7, but do not require equal component norms. Suppose every μ_k and every centroid in equation 29 is nonzero. Let $a_k = \|\mu_k\|_{\Sigma^{-1}}$ and $z_q = \overline{\Phi}^{-1}(q)$. Then:*

(i) Geometric loss. *For a partition $\mathcal{P}$, define $V(\mathcal{P}) = \sum_{H \in \mathcal{P}} \pi_H \|\mu_H - \bar{\mu}\|^2_{\Sigma^{-1}}$. If $\mathcal{P}_K = \{\{1\}, \dots, \{K\}\}$, then*

$$V(\mathcal{P}_K) - V(\mathcal{P}_{K-1}) = \frac{\pi_r \pi_s}{\pi_r + \pi_s} \|\mu_r - \mu_s\|^2_{\Sigma^{-1}}. \tag{30}$$

(ii) Coarsened alternative distribution. *Define $T_{K-1}(Z,W) = \frac{\mu_{H(Z)}^\top \Sigma^{-1} W}{a_{H(Z)}}$, $p_{K-1}(Z,W) = \overline{\Phi}\{T_{K-1}(Z,W)\}$, where $H(Z)$ is the block containing Z. Then p_{K-1} is uniform under P_0. For $k \in H$, let $\delta_{kH} = \frac{\mu_H^\top \Sigma^{-1} \mu_k}{a_H}$. The alternative CDF is*

$$F_{K-1}(q) = \sum_{H \in \mathcal{P}_{K-1}} \sum_{k \in H} \pi_k \overline{\Phi}(z_q - \delta_{kH}). \tag{31}$$

Moreover, $\sum_{k \in H} \frac{\pi_k}{\pi_H} \delta_{kH} = a_H > 0$, so $F_{K-1}(q)/q \to \infty$ as $q \downarrow 0$, and the coarsened score has positive asymptotic power.

(iii) Retaining dominance over common-slope scores. *Let $F_K(q) = \sum_{k=1}^K \pi_k \overline{\Phi}(z_q - a_k)$, $\tau_K = \tau(F_K)$, $\mathrm{Pow}_K = F_K(\tau_K)$, and, for $v^\top \Sigma v = 1$, let $F_v(q) = \sum_{k=1}^K \pi_k \overline{\Phi}(z_q - v^\top \mu_k)$, $\mathrm{Pow}^\star_{\mathrm{lin}} = \sup_{v: v^\top \Sigma v = 1} \mathrm{Pow}_{\mathrm{BH}}(F_v)$. If the μ_k's do not all lie on one common positive ray, then*

$$\Delta_{\mathrm{lin},K} := \mathrm{Pow}_K - \mathrm{Pow}^\star_{\mathrm{lin}} > 0. \tag{32}$$

Furthermore,

$$0 \le F_K(q) - F_{K-1}(q) \le \varepsilon_{rs}, \qquad \varepsilon_{rs} = \frac{L_{rs}}{\sqrt{2\pi}}, \tag{33}$$

where $L_{rs} = \pi_r a_r + \pi_s a_s - (\pi_r + \pi_s) a_{\{r,s\}} \geq 0$. *Suppose there exist* $\kappa > 0$, $M_K < \infty$, *and* $r_0 \in (0, \tau_K)$ *such that, for* $0 \leq u \leq r_0$,

$$\Psi_K(\tau_K - u) \geq \kappa u, \qquad \Psi_K(q) = \alpha\{\rho_0 q + \rho_1 F_K(q)\} - q,$$
$$|F_K(q) - F_K(q')| \leq M_K |q - q'|, \qquad q, q' \in [\tau_K - r_0, \tau_K].$$

If $\alpha\rho_1\varepsilon_{rs}/\kappa \leq r_0$, *then*

$$\mathrm{Pow}_{\mathrm{BH}}(F_{K-1}) \geq \mathrm{Pow}_K - \left(1 + \frac{\alpha\rho_1 M_K}{\kappa}\right)\varepsilon_{rs}. \tag{34}$$

Hence the coarsened score remains strictly more powerful than every common-slope linear score whenever $\left(1 + \frac{\alpha\rho_1 M_K}{\kappa}\right)\varepsilon_{rs} < \Delta_{\mathrm{lin},K}$. *If* μ_r *and* μ_s *lie on the same positive ray, then* $\varepsilon_{rs} = 0$, *so the merge loses no asymptotic power.*

C.2.3 Proof

Proof of Theorem C.1. Only the terms indexed by r and s change after merging. The weighted variance identity in the Mahalanobis inner product gives

$$\begin{aligned}&\pi_r\|\mu_r\|^2_{\Sigma^{-1}} + \pi_s\|\mu_s\|^2_{\Sigma^{-1}} - (\pi_r + \pi_s)\|\mu_{\{r,s\}}\|^2_{\Sigma^{-1}} \\ =&\frac{\pi_r\pi_s}{\pi_r + \pi_s}\|\mu_r - \mu_s\|^2_{\Sigma^{-1}}.\end{aligned}$$

The global centroid is unchanged, proving equation 30.

Conditional on $H = 0$ and $Z = k \in H$, $T_{K-1} \sim N(0,1)$. Conditional on $H = 1$ and $Z = k \in H$, $T_{K-1} \sim N(\delta_{kH}, 1)$. This proves null uniformity and equation 31. Also,

$$\sum_{k\in H}\frac{\pi_k}{\pi_H}\delta_{kH} = \frac{\mu_H^\top\Sigma^{-1}}{a_H}\sum_{k\in H}\frac{\pi_k}{\pi_H}\mu_k = a_H.$$

Thus every block contains at least one component with positive δ_{kH}. Mills' ratio then gives $F_{K-1}(q)/q \to \infty$.

For any v with $v^\top\Sigma v = 1$, Mahalanobis Cauchy–Schwarz gives $v^\top\mu_k \leq a_k$, and hence $F_v(q) \leq F_K(q)$. If the component means do not all lie on one positive ray, at least one inequality is strict for every fixed v. Compactness of $\{v : v^\top\Sigma v = 1\}$, continuity, and the same BH-feasibility argument used in the proof of Theorem 2 yield equation 32.

Only the merged components contribute to $F_K - F_{K-1}$. For $k \in \{r, s\}$, Cauchy–Schwarz gives $\delta_{k,\{r,s\}} \leq a_k$. Since $d \mapsto \overline{\Phi}(z - d)$ is increasing with derivative bounded by $1/\sqrt{2\pi}$,

$$0 \leq F_K(q) - F_{K-1}(q) \leq \frac{1}{\sqrt{2\pi}}\sum_{k\in\{r,s\}}\pi_k\{a_k - \delta_{k,\{r,s\}}\}.$$

Moreover, $\pi_r\delta_{r,\{r,s\}} + \pi_s\delta_{s,\{r,s\}} = (\pi_r + \pi_s)a_{\{r,s\}}$, which proves equation 33.

Set $u_{rs} = \alpha\rho_1\varepsilon_{rs}/\kappa$. If $u_{rs} \leq r_0$, then

$$\begin{aligned}&\alpha\{\rho_0(\tau_K - u_{rs}) + \rho_1 F_{K-1}(\tau_K - u_{rs})\} - (\tau_K - u_{rs}) \\ &\quad\geq \Psi_K(\tau_K - u_{rs}) - \alpha\rho_1\varepsilon_{rs} \geq 0.\end{aligned}$$

Hence $\tau(F_{K-1}) \geq \tau_K - u_{rs}$, and

$$\begin{aligned}\mathrm{Pow}_{\mathrm{BH}}(F_{K-1}) &\geq F_{K-1}(\tau_K - u_{rs}) \\ &\geq F_K(\tau_K - u_{rs}) - \varepsilon_{rs} \\ &\geq F_K(\tau_K) - M_K u_{rs} - \varepsilon_{rs},\end{aligned}$$

which is equation 34. Combining this inequality with equation 32 proves the dominance condition. If the merged means lie on the same positive ray, their standardized directions coincide, so $\delta_{r,\{r,s\}} = a_r$ and $\delta_{s,\{r,s\}} = a_s$; hence $\varepsilon_{rs} = 0$. □

C.3 POWER UNDER DIRECTIONAL MISMATCH

Unlike group coarsening, directional mismatch preserves the group labels but changes the scoring directions. The following results identify gain and loss regimes for transferred scores relative to the best common-direction score.

Retain the target model and assumptions of Theorem 2. The source has the same null law and group proportions, but its group-k alternatives follow $N(\nu_k, \Sigma)$, with $\|\nu_k\|_{\Sigma^{-1}} = b > 0$ for every k. Let $\varphi_k \in [0, \pi]$ be the angle between $\Sigma^{-1/2}\mu_k$ and $\Sigma^{-1/2}\nu_k$, and define $T_H(Z, W) = \nu_Z^\top \Sigma^{-1} W / b$. Write $\overline{\Phi} = 1 - \Phi$ and $z_q = \overline{\Phi}^{-1}(q)$.

Theorem C.2 (Power gain under directional mismatch)**.** *Under the preceding model, set* $c_{\alpha,\rho} = (1 - \alpha\rho_0)/(\alpha\rho_1)$ *and define*

$$F^\star_{\mathrm{lin}}(q) = \max_{v : v^\top \Sigma v = 1} \sum_{k=1}^{K} \pi_k \overline{\Phi}(z_q - v^\top \mu_k), \qquad \tau_\star = \sup\{q \in [0,1] : F^\star_{\mathrm{lin}}(q) \geq c_{\alpha,\rho} q\}, \qquad r_\star = c_{\alpha,\rho}\tau_\star, \tag{35}$$

using the continuous endpoint extensions of $F^\star_{\mathrm{lin}}$*. Then* $0 < \tau_\star < 1/c_{\alpha,\rho}$*, and the critical angle*

$$\varphi_\star = \arccos\left(\frac{z_{\tau_\star} - z_{r_\star}}{a}\right) \in (0, \pi/2) \tag{36}$$

is well defined. If $\varphi_k \in [0, \varphi_\star]$ *for every* k*, with at least one* $\varphi_k < \varphi_\star$*, then*

$$\mathrm{Pow}_\alpha(T_H) > \sup_{S \in \mathcal{S}_{\mathrm{lin}}} \mathrm{Pow}_\alpha(S). \tag{37}$$

The structural advantage can therefore survive nonzero directional mismatch, although sufficiently large mismatch reverses the comparison.

Theorem C.3 (Power loss under directional mismatch)**.** *Under the source and target models above, let* $\varphi_\star$ *be defined by equation 36. If* $\varphi_k \in [\varphi_\star, \pi]$ *for every* k*, with at least one* $\varphi_k > \varphi_\star$*, then*

$$\mathrm{Pow}_\alpha(T_H) < \sup_{S \in \mathcal{S}_{\mathrm{lin}}} \mathrm{Pow}_\alpha(S). \tag{38}$$

If every $\varphi_k \geq \pi/2$*, then* $\mathrm{Pow}_\alpha(T_H) = 0$*. If every* $\varphi_k = \varphi_\star$*, equality holds in place of equation 38.*

These angle ranges are sufficient conditions when angles differ across groups; for a common angle, $\varphi_\star$ is the exact boundary. The loss is relative to the best common-direction score, not to every global anchor. This possibility motivates assessing local corrections against the chosen anchor, as in Section 3.3.

C.3.1 PROOFS

Proof of Theorem C.2. Under the target null, $T_H \sim N(0,1)$; conditional on $H = 1$ and $Z = k$, $T_H \sim N(a\cos\varphi_k, 1)$. Its alternative p-value CDF is therefore

$$F_{1,H}(q) = \sum_{k=1}^{K} \pi_k \overline{\Phi}(z_q - a\cos\varphi_k). \tag{39}$$

For each Gaussian score considered here, the BH feasibility condition is $F_{1,S}(q) \geq c_{\alpha,\rho} q$. Continuity at the largest feasible cutoff yields

$$\mathrm{Pow}_\alpha(S) = c_{\alpha,\rho}\tau_\alpha(S), \tag{40}$$

with both sides zero if the cutoff is zero.

Every nonconstant member of $\mathcal{S}_{\mathrm{lin}}$ has the same p-values as a normalized projection $v^\top W$ with $v^\top \Sigma v = 1$, whose alternative CDF is $F_{1,v}(q) = \sum_k \pi_k \overline{\Phi}(z_q - v^\top \mu_k)$. The compactness of the normalized directions and the boundedness of the shifts $v^\top \mu_k$ imply that the envelope in equation 35 is attained and continuous, including at the endpoints. Choosing $v = \Sigma^{-1}\mu_1/a$ gives $F_{1,v}(q)/q \to \infty$ as $q \downarrow 0$, so $\tau_\star > 0$. Since $c_{\alpha,\rho} > 1$ and $F^\star_{\mathrm{lin}}(q) < 1$ for $q < 1$, continuity gives $\tau_\star < 1/c_{\alpha,\rho}$ and

$F^{\star}_{\mathrm{lin}}(\tau_\star) = r_\star$. No common direction has a feasible cutoff above $\tau_\star$, while one attaining the envelope at $\tau_\star$ has exactly that cutoff. Equation 40 thus gives

$$\sup_{S \in \mathcal{S}_{\mathrm{lin}}} \mathrm{Pow}_\alpha(S) = r_\star, \tag{41}$$

since constant scores have zero power.

By equation 15 and compactness, $F^{\star}_{\mathrm{lin}}(q) < F_{1,L}(q) = \overline{\Phi}(z_q - a)$ for every $q \in (0,1)$. Hence $\tau_\star < r_\star < \overline{\Phi}(z_{\tau_\star} - a)$, which implies $0 < z_{\tau_\star} - z_{r_\star} < a$ and proves the stated range of $\varphi_\star$. Under the angle condition, $a\cos\varphi_k \geq a\cos\varphi_\star$ for every k, with at least one strict inequality. Consequently, equation 39 gives $F_{1,H}(\tau_\star) > \overline{\Phi}(z_{\tau_\star} - a\cos\varphi_\star) = r_\star$. The BH inequality for T_H is strictly feasible at $\tau_\star$, so continuity implies $\tau_\alpha(T_H) > \tau_\star$. Combining equation 40 and equation 41 proves equation 37. □

Proof of Theorem C.3. Set $d_\star = a\cos\varphi_\star > 0$ and $F_\star(q) = \overline{\Phi}(z_q - d_\star)$. The angle condition and equation 39 imply $F_{1,H}(q) < F_\star(q)$ for every $q \in (0,1)$. Moreover, $F_\star(\tau_\star) = r_\star = c_{\alpha,\rho}\tau_\star$ by equation 36.

Since $F'_\star(q) = \exp(d_\star z_q - d_\star^2/2)$ is strictly decreasing, $F_\star$ is strictly concave. Together with $F_\star(0) = 0$, this implies that $F_\star(q)/q$ is strictly decreasing. Thus $F_\star(q) < c_{\alpha,\rho}q$ for every $q > \tau_\star$. Consequently, the BH inequality for T_H is infeasible for every $q \geq \tau_\star$. Its feasible set is closed, so $\tau_\alpha(T_H) < \tau_\star$. Equations equation 40 and equation 41 prove equation 38.

If every $\varphi_k \geq \pi/2$, all alternative shifts are nonpositive, so $F_{1,H}(q) \leq q < c_{\alpha,\rho}q$ for every $q > 0$ and the power is zero. If every $\varphi_k = \varphi_\star$, then $F_{1,H} = F_\star$, its largest feasible cutoff is $\tau_\star$, and its power equals $r_\star$. □

D Simulation details

The two experiments in Section 4 use a common two-sided Gaussian target and alter either the auxiliary signal strength or the number of labeled auxiliary outliers. They therefore isolate two practical requirements for structural transfer: robustness to source–target signal mismatch and sufficient data to learn the auxiliary category structure.

D.1 Data generation and splitting

Let e_1 be the first coordinate vector in $\mathbb{R}^{100}$. All inlier observations in $\mathcal{D}_{\mathrm{tr}}$, $\mathcal{D}_{\mathrm{cal}}$, and the null portion of $\mathcal{D}_{\mathrm{test}}$ are drawn independently from $N(0, I_d)$. Each of the $m = 1{,}000$ test observations is non-null independently with probability 0.2. Conditional on being non-null, its type $C \in \{+,-\}$ and feature vector satisfy

$$\Pr(C = +) = 0.8, \qquad X^{\mathrm{test}} \mid C = + \sim N(se_1, I_d), \qquad X^{\mathrm{test}} \mid C = - \sim N(-se_1, I_d). \tag{42}$$

The reference samples have sizes $|\mathcal{D}_{\mathrm{tr}}| = 2{,}000$ and $|\mathcal{D}_{\mathrm{cal}}| = 1{,}000$. The unlabeled mixture available to adaptive learners is

$$\mathcal{D}_{\mathrm{mix}} = \mathcal{D}_{\mathrm{cal}} \cup \mathcal{D}_{\mathrm{test}},$$

with $|\mathcal{D}_{\mathrm{mix}}| = 2{,}000$.

For the signal-mismatch experiment, the auxiliary sample has size $n_{\mathrm{aux}} = 1{,}000$, uses the same type probabilities as equation 42, and follows

$$X^A \mid C = + \sim N(ase_1, I_d), \qquad X^A \mid C = - \sim N(-se_1, I_d). \tag{43}$$

We use $s \in \{2,3,4\}$ and $a \in \{1, 1.2, 1.4, 1.6, 1.8, 2\}$. Only the positive auxiliary component changes with a; the target distribution is fixed within each column.

For the auxiliary-size experiment, the test distribution remains equation 42, whereas the auxiliary types are balanced and obey

$$X^A \mid C = + \sim N(1.5se_1, I_d), \qquad X^A \mid C = - \sim N(-0.5se_1, I_d). \tag{44}$$

We vary $n_{\mathrm{aux}} \in \{100, 500, 1{,}000\}$ and $s \in \{1.5, 2, 2.5, 3, 3.5, 4\}$.

D.2 METHODS AND EVALUATION

We compare OCC, BIC, MCC, ICP, and AdaDetect with G-OCC, G-BIC, G-MCC, G-ICP, and G-AdaDetect. The router is multinomial logistic regression. Target–reference and local comparisons use logistic regression with inverse regularization strength $C = 0.1$; OCC is an RBF one-class SVM with $\nu = 0.05$ and `gamma=scale`.

The candidate grid is $\Lambda = \{0, 0.25, 0.5, 0.75, 1\}$. In each of the two repeats, $\mathcal{D}_{\mathrm{tr}}$ and the unordered $\mathcal{D}_{\mathrm{mix}}$ are separately partitioned into three folds, shared by all candidates. Fits excluding each fold produce held-out reference and pooled scores. After concatenating these scores, BH is applied once per repeat to the complete pseudo-p-value vector. Selection uses equation 4 with $\kappa_{\mathrm{se}} = 1$ and equation 5 with $\Delta_{\min} = 1$. After full-data refitting, final conformal p-values compare test observations with $\mathcal{D}_{\mathrm{cal}}$, and BH is applied at $\alpha = 0.10$.

For Monte Carlo repetition $r = 1, \ldots, M$, record TDC_r and FDP_r using the definitions in Section 2.1. Count summaries use $\overline{\mathrm{TDC}} = M^{-1} \sum_{r=1}^{M} \mathrm{TDC}_r$, and false-discovery summaries use $\overline{\mathrm{FDP}} = M^{-1} \sum_{r=1}^{M} \mathrm{FDP}_r$, the empirical estimate of FDR. When comparing detection rates, report $\widehat{\mathrm{AP}} = M^{-1} \sum_{r=1}^{M} \mathrm{TDC}_r / (1 \vee m_{1,r})$, where $m_{1,r}$ is the realized number of test outliers. This is the average normalized TDC, not a count. In particular, random outlier counts do not permit recovering mean TDC by multiplying $\widehat{\mathrm{AP}}$ by the expected number of outliers.

E REAL-DATA DETAILS AND ADDITIONAL RESULTS

This appendix gives the common protocol and the construction of each benchmark, followed by repetition-level TDC distributions and selected shift or sensitivity comparisons. Table 1 in Section 5 contains the primary mean TDC and mean FDP results. Each additional figure addresses a distinct question rather than repeating that table.

E.1 COMMON REAL-DATA PROTOCOL

Target evaluation labels are hidden from every method until performance is assessed. Each repetition uses disjoint target-domain reference training, reference calibration, and test samples. No separate tuning sample is reserved. The selector uses common repeated three-fold partitions of $\mathcal{D}_{\mathrm{tr}}$ and the unordered calibration–test pool. We concatenate held-out scores before computing the pseudo-BH count, and final testing follows the full-data refit in Section 2.2.

All comparisons use 100 matched repetitions and the five anchor–G pairs introduced in Section 4. The candidate grid is $\Lambda = \{0, 0.25, 0.5, 0.75, 1\}$. The router is multinomial logistic regression; local target–reference classifiers use logistic regression with $C = 0.1$ and the anchor-augmented feature Z_G from Step 1. OCC uses an RBF one-class SVM with $\nu = 0.05$ and `gamma=scale`.

We use the TDC and FDP summaries defined in equation 1. The main table reports their Monte Carlo means and standard errors, while boxplots show TDC variation across repetitions. Shift comparisons additionally report $\widehat{\mathrm{AP}}$, the mean of $\mathrm{TDC}/(1 \vee m_1)$, because changing anomaly prevalence changes the number of available true discoveries. Mean FDP estimates FDR; an individual FDP need not lie below α. The same admissibility requirements apply to each anchor and its GLADE variant.

For UNSW-NB15 and Covertype, we use $|\mathcal{D}_{\mathrm{tr}}| = 12{,}000$ and $|\mathcal{D}_{\mathrm{cal}}| = 6{,}000$, with a test sample of $10{,}000$ observations, including $2{,}000$ outliers. The auxiliary sample contains $2{,}500$ observations per auxiliary category. For CSE-CIC-IDS2018, the corresponding sample sizes are $6{,}000$, $3{,}000$, and $3{,}000$, with 600 outliers in the test set and $1{,}250$ auxiliary observations per auxiliary category. In the primary Gas Sensor Array Drift setting, we use $|\mathcal{D}_{\mathrm{tr}}| = 500$ and $|\mathcal{D}_{\mathrm{cal}}| = 300$, with a test sample of 600 observations, including 120 outliers. The auxiliary sample contains 350 observations per auxiliary gas category; this smaller sample size reflects the limited number of observations available in the later temporal batches.

E.2 UNSW-NB15

We use the two official UNSW-NB15 subsets as source and target domains (Moustafa & Slay, 2015). The larger subset supplies labeled auxiliary attacks, and the smaller subset supplies target-domain reference, calibration, and test observations. We retain normal flows and five attack types: Generic,

Exploits, Fuzzers, DoS, and Reconnaissance. Numeric variables are imputed using source-domain medians; `proto`, `service`, and `state` are one-hot encoded using the source vocabulary, yielding 194 features.

A hash-disjoint partition rule prevents identical raw feature vectors from appearing in different partitions of a repetition. The primary mixture balances the attack types at alternative prevalence $\rho_1 = 0.20$.

All G variants improve their anchors in the primary comparison, with G-MCC leading. BIC and MCC already benefit from labeled auxiliary attacks. At the same time, the larger improvements for OCC and ICP are consistent with local target comparisons adding evidence not effectively used by those anchors. Figure 5 shows the repetition-level variation behind the mean TDC comparison.

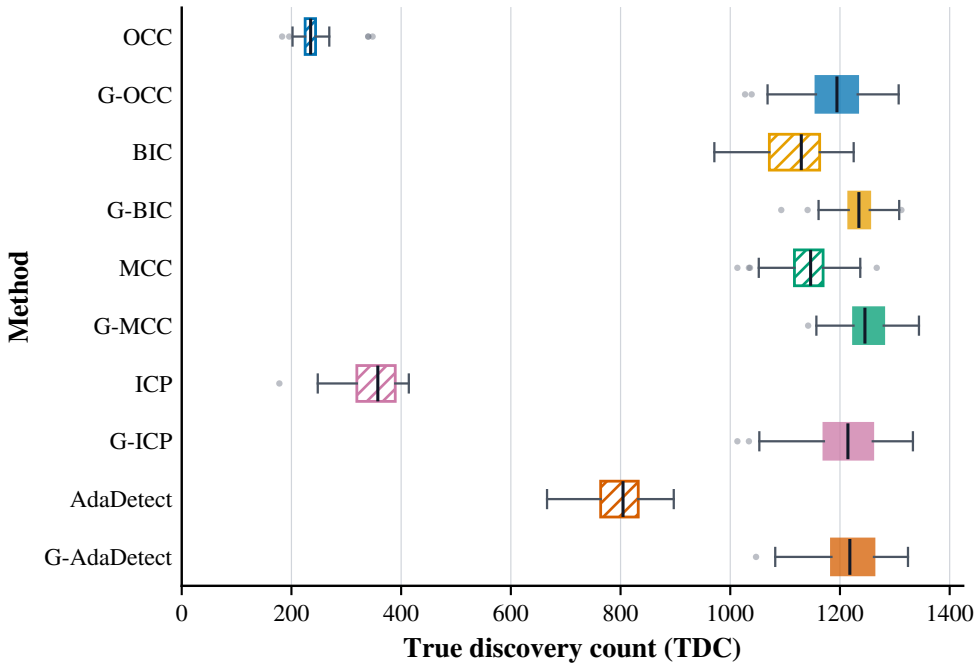

Figure 5: UNSW-NB15 primary comparison: empirical TDC distributions across 100 matched repetitions.

The target-shift comparisons in Figure 6 support useful structural transfer beyond the primary mixture, although the leading G variant varies across settings. Prevalence changes both the information available for target learning and the multiple-testing problem; normalized TDC distinguishes detection rates from a simple increase in the number of outliers.

E.3 COVERTYPE

Covertype evaluates transfer between Rawah Wilderness Area (source) and Comanche Peak Wilderness Area (target) (Blackard & Dean, 1998). Lodgepole Pine is the reference class; Spruce/Fir, Aspen, and Krummholz are the alternatives shared across the two regions. The primary representation contains ten continuous terrain variables, excluding wilderness-area and soil-type indicators that may encode geographic location. Stable row identifiers keep partitions disjoint. The primary alternative mixture is balanced with $\rho_1 = 0.20$.

Every G variant substantially improves its anchor, and G-OCC has the largest primary mean TDC. In particular, weak BIC performance contrasts with the stronger G variants, consistent with auxiliary observations being imperfect templates for target anomalies. At the same time, their cover-type labels remain useful for local specialization. G-AdaDetect also improves the pooled target-adaptive anchor, indicating that auxiliary category structure can add value beyond target adaptation alone. Figure 7 gives the corresponding TDC distributions.

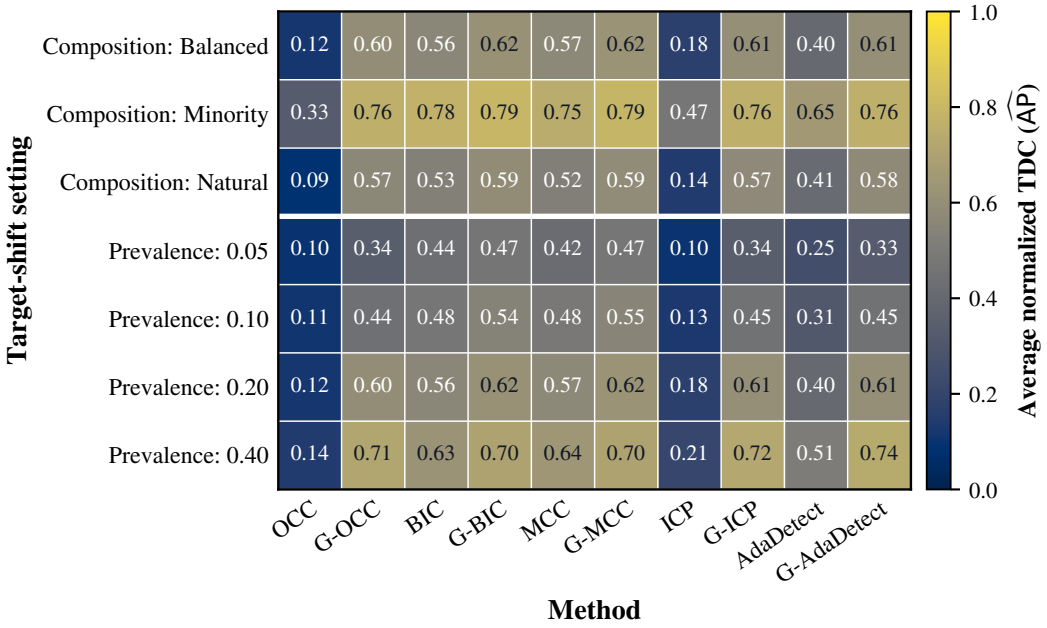

Figure 6: UNSW-NB15 target shifts across attack-composition and alternative-prevalence settings. (a) Average normalized TDC ($\widehat{\mathrm{AP}}$).

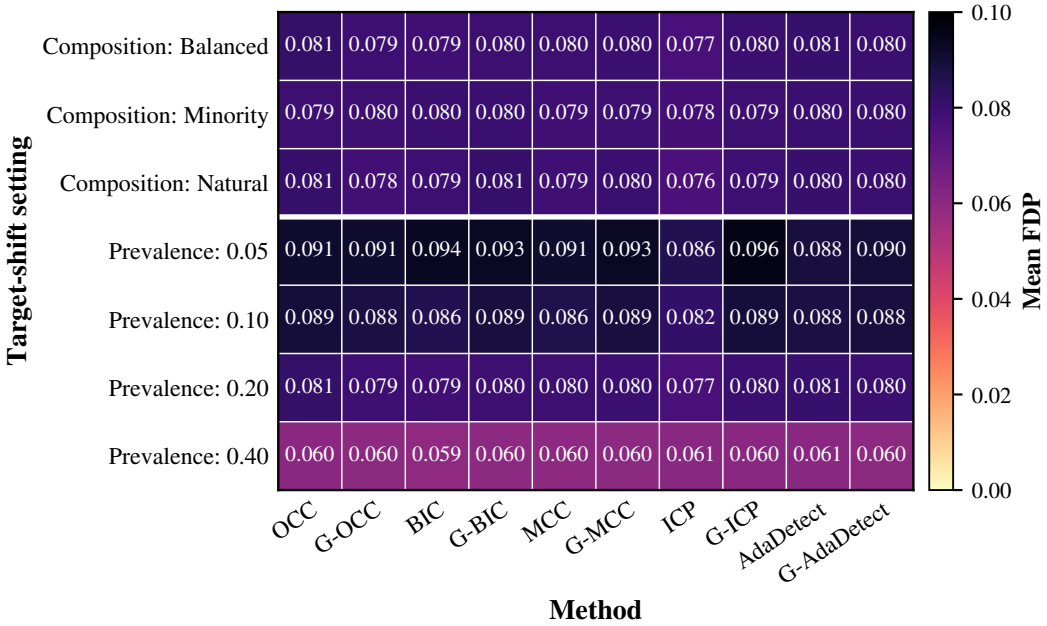

Figure 6: UNSW-NB15 target shifts (continued). (b) Mean FDP; the nominal level is $\alpha = 0.10$.

Changes in target composition and prevalence alter the relative difficulty of the ecological testing problem, but the G variants remain competitive across the reported settings (Figure 8). Reversing the geographic transfer preserves their advantage. Adding soil indicators substantially benefits several anchors but has smaller, mixed effects on the G variants, so this comparison assesses sensitivity to the available representation rather than geographic generalization (Figure 9).

E.4 CSE-CIC-IDS2018

CSE-CIC-IDS2018 contains benign traffic and attacks collected on different days (Sharafaldin et al., 2018). The source combines FTP/SSH brute force from February 14, GoldenEye/Slowloris DoS

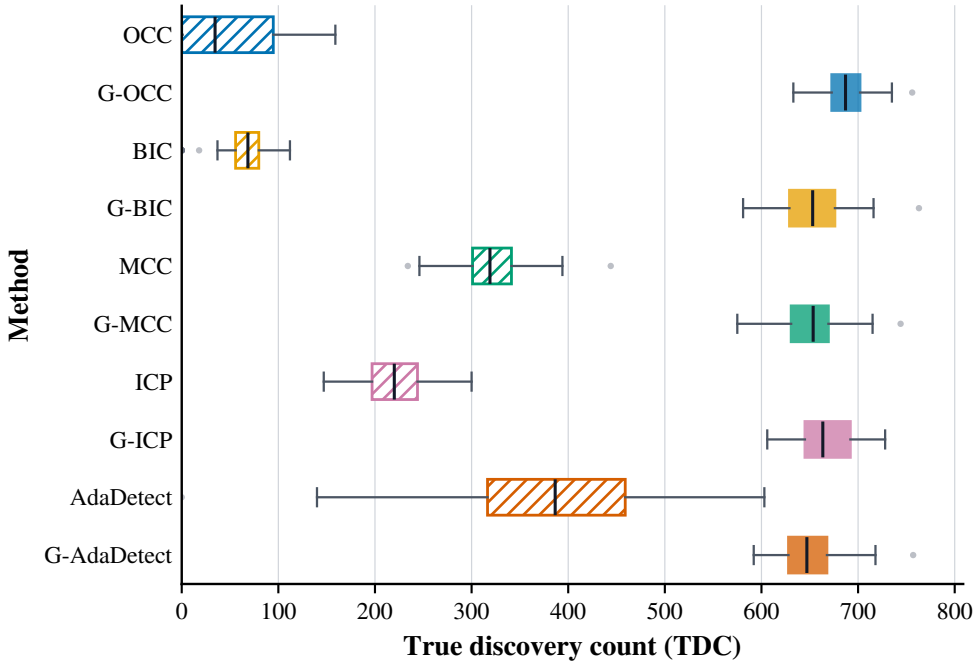

Figure 7: Covertype primary comparison: empirical TDC distributions across 100 matched repetitions.

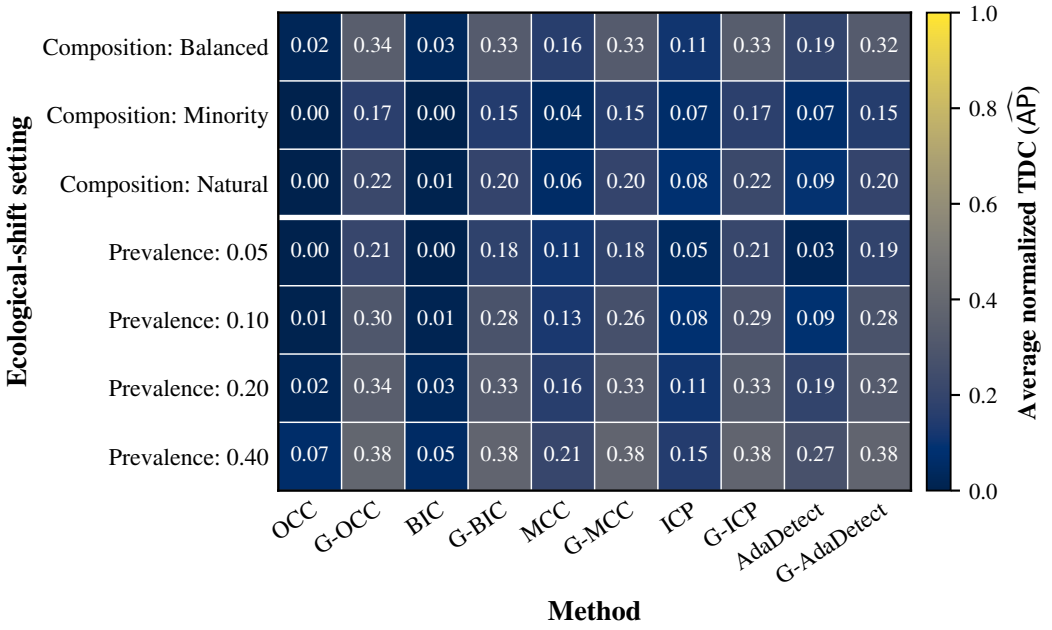

Figure 8: Covertype ecological shifts under changes in alternative composition and prevalence. (a) Average normalized TDC ($\widehat{\mathrm{AP}}$).

from February 15, and infiltration from February 28. The target combines web brute force/XSS/SQL injection from February 22, SlowHTTPTest/Hulk DoS from February 16, and infiltration from March 1. We group these implementations into BruteForce, DoS, and Infiltration. Botnet is excluded because it occurs on only one day.

We remove timestamps and labels, retain 78 numeric CICFlowMeter features, and replace nonfinite values using source-domain medians. A deterministic capped reservoir contains 220,362 flows,

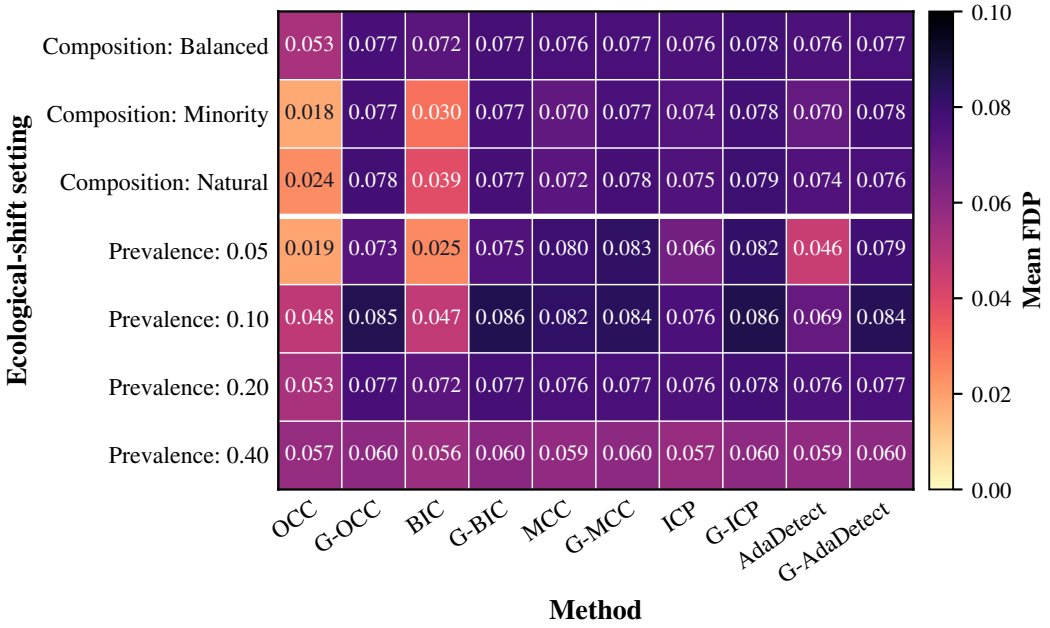

Figure 8: Covertype ecological shifts (continued). (b) Mean FDP; the nominal level is $\alpha = 0.10$.

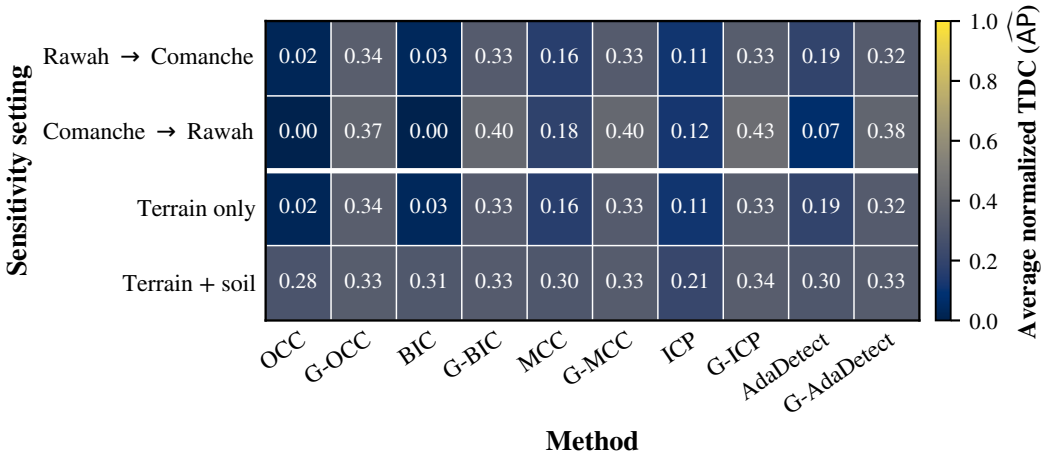

Figure 9: Covertype geographic transfer and representation sensitivity for the two transfer directions and the terrain-only and terrain-plus-soil representations. (a) Average normalized TDC ($\widehat{\mathrm{AP}}$).

with each split retaining up to 80,000 benign flows and 12,000 flows per attack family. The target BruteForce day supplies only 362 observations and is not oversampled. Feature hashes are disjoint across reference, calibration, test, and auxiliary partitions. The primary mixture is balanced with $\rho_1 = 0.20$; additional settings change family composition or set $\rho_1 \in \{0.10, 0.30\}$. All reported settings use 100 matched repetitions; preliminary runs are excluded.

G-MCC has the largest primary mean TDC, with smaller gains over the already competitive MCC anchor than over OCC or ICP. The improvement of G-ICP shows that local target learning can strengthen a baseline that makes no discoveries at the chosen level, without implying that its original ranking contains no information. The TDC distributions in Figure 10 complement these mean comparisons.

The target-shift results in Figure 11 show uniformly nonnegative gains for the GLADE variants over their matched baselines across all reported settings, with strict improvements in nearly all comparisons. The leading method varies with the target composition and prevalence, while mean FDP remains below the nominal level throughout.

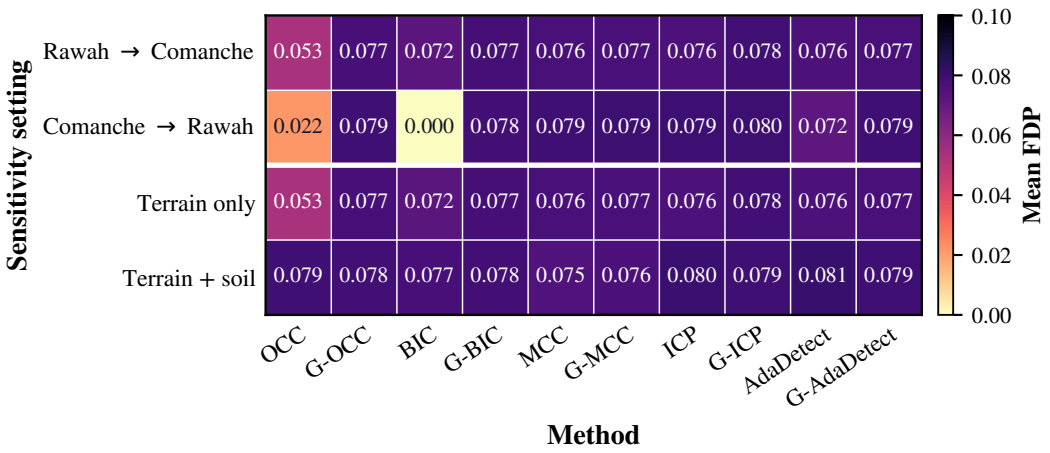

Figure 9: Covertype geographic-transfer and representation sensitivity (continued). (b) Mean FDP; the nominal level is $\alpha = 0.10$.

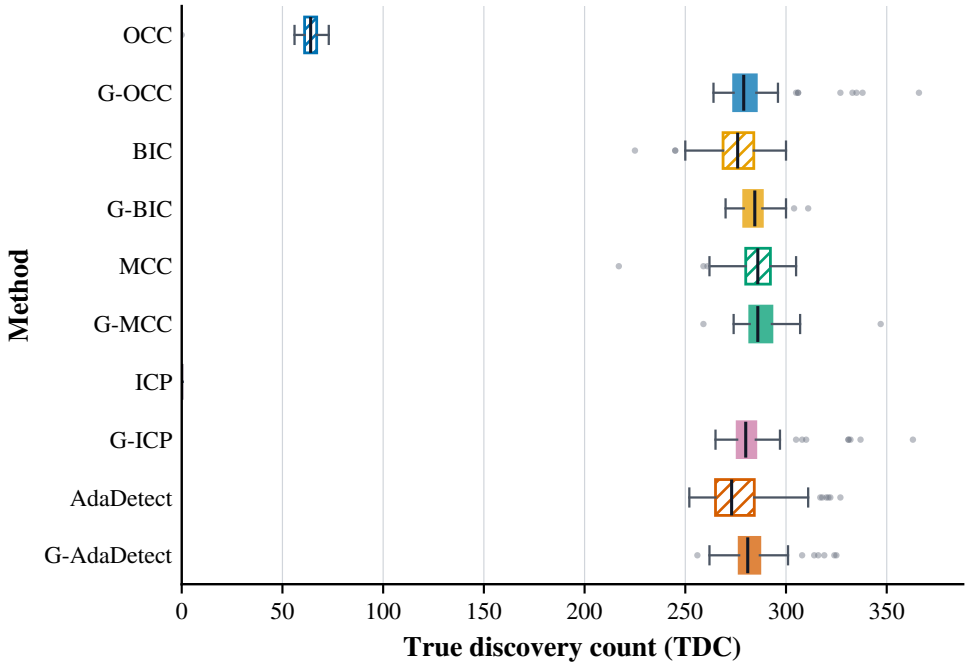

Figure 10: CSE-CIC-IDS2018 primary comparison: empirical TDC distributions across 100 matched repetitions.

E.5 Gas Sensor Array Drift

The Gas Sensor Array Drift dataset contains 13,910 measurements of six pure gases collected by 16 chemical sensors over 36 months (Vergara, 2012). Each observation has 128 steady-state and transient features, and the archive comprises ten temporal batches. Batches 1–6 supply historical observations and batches 7–10 form the target domain. Ammonia is the reference class, while Ethanol, Ethylene, Acetaldehyde, Acetone, and Toluene form the alternatives. The primary mixture balances these five gases; additional settings vary composition and prevalence.

BIC is already a strong anchor, and G-BIC improves it further while attaining the largest primary mean TDC. The other G variants also improve their anchors, with larger gains where the original detector is weaker. This pattern is consistent with gas-category labels from earlier batches retaining useful structure despite sensor drift: local learning can adapt the scoring directions to current data while retaining a strong global anchor. Figure 12 displays repetition-level variability.

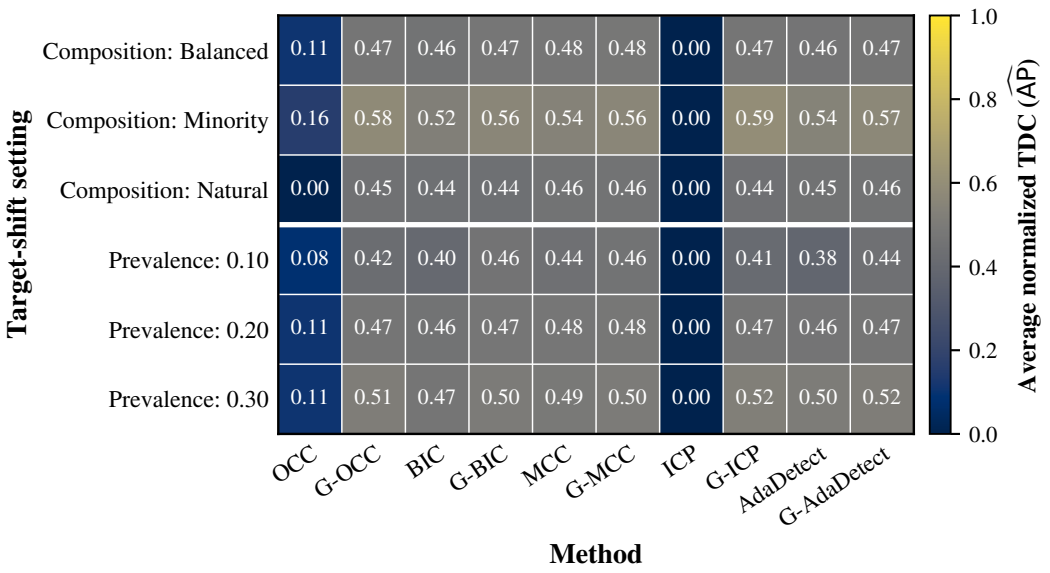

Figure 11: CSE-CIC-IDS2018 target shifts across attack-family composition and alternative-prevalence settings. (a) Average normalized TDC ($\widehat{\mathrm{AP}}$).

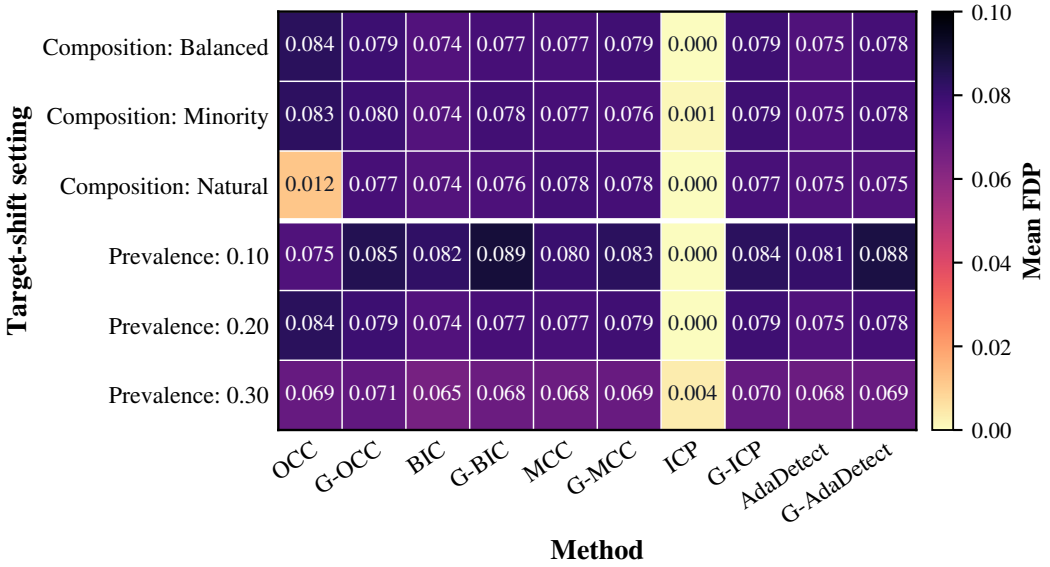

Figure 11: CSE-CIC-IDS2018 target shifts (continued). (b) Mean FDP; the nominal level is $\alpha = 0.10$.

G-BIC remains the strongest method across the reported composition and prevalence settings, and the matched G-versus-anchor mean differences are generally positive. Mean FDP stays below the nominal level (Figure 13).

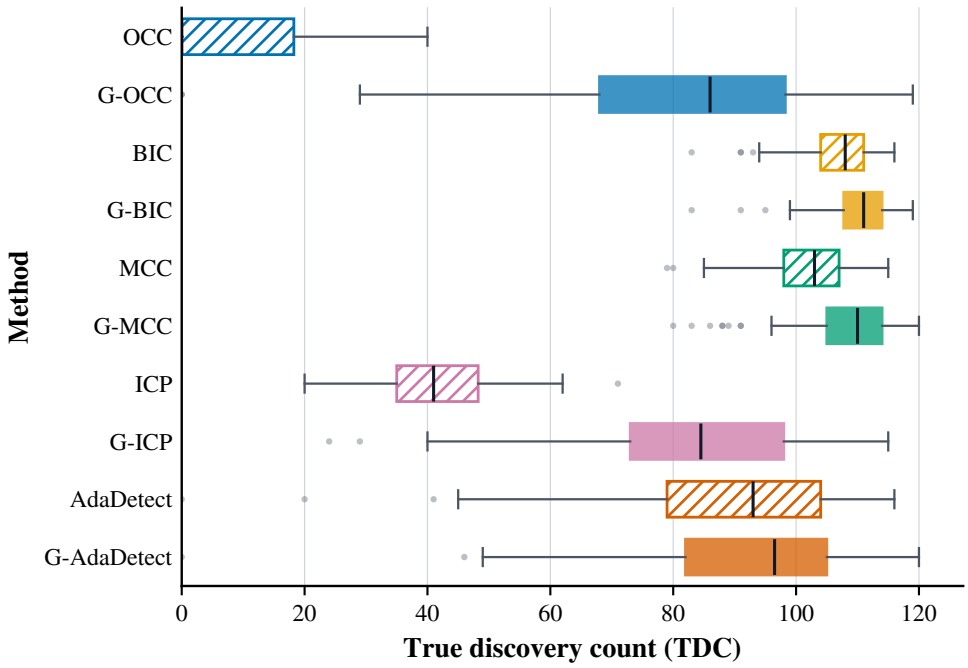

Figure 12: Gas Sensor Array Drift primary comparison: empirical TDC distributions across 100 matched repetitions.

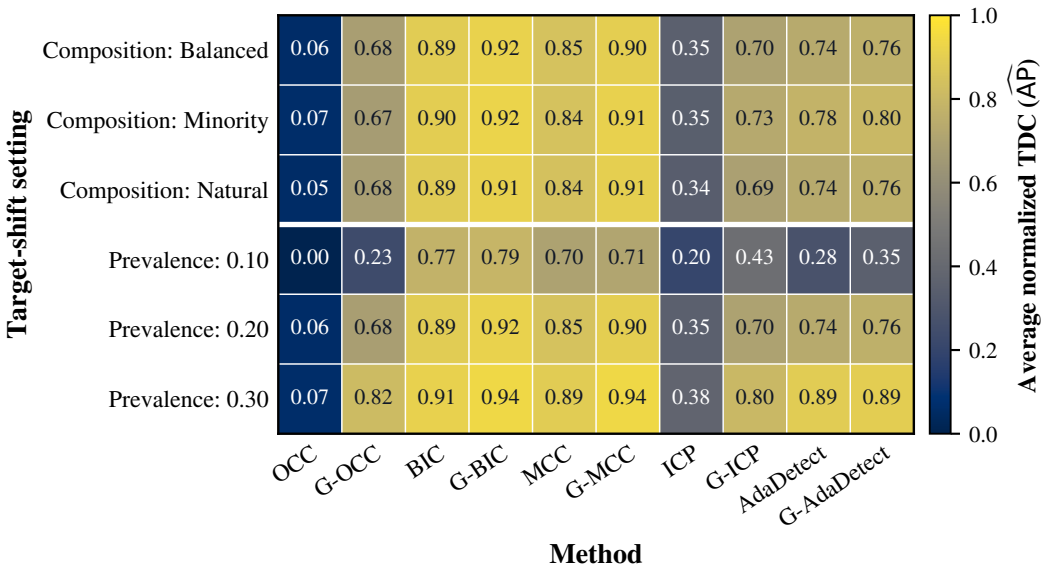

Figure 13: Gas Sensor Array Drift target shifts under changes in gas composition and anomaly prevalence. (a) Average normalized TDC ($\widehat{\mathrm{AP}}$).

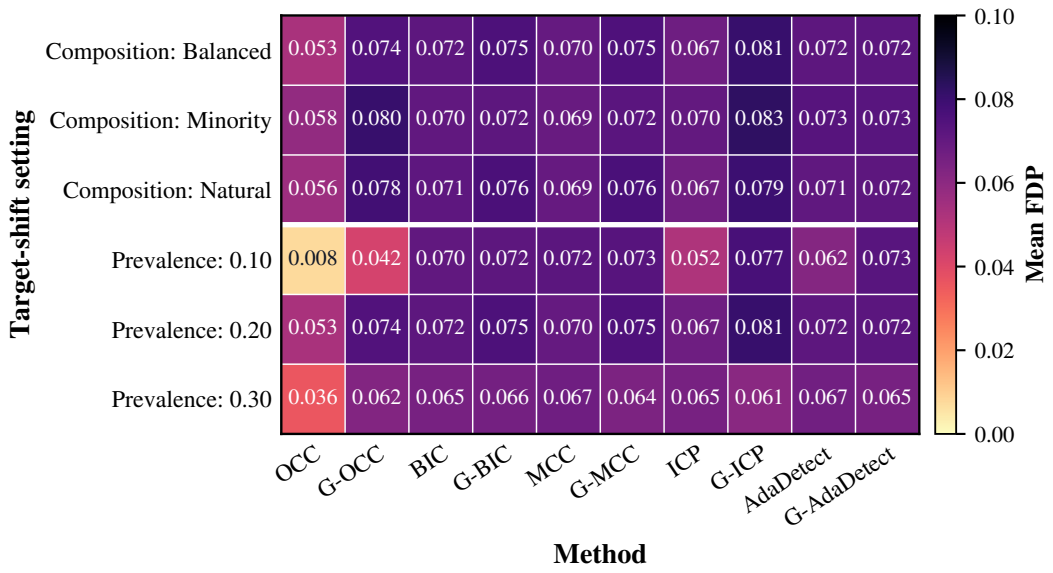

Figure 13: Gas Sensor Array Drift target shifts (continued). (b) Mean FDP; the nominal level is $\alpha = 0.10$.

Overall findings. Across the four benchmarks and the reported shift settings, GLADE typically improves the detection power of its matched global anchor and is never worse in mean performance. When the auxiliary structure is informative, the learned local refinements extract additional target-relevant information and can yield substantial power gains. When such information provides little additional benefit, the utility-oriented selection tends to shrink the correction toward the global anchor, so that performance falls back to that of the original detector rather than suffering negative transfer. This pattern is observed across substantially different global anchors and across domain, temporal, geographic, and sensor shifts, indicating that GLADE can refine both direct source-transfer methods and target-adaptive detectors without requiring a single anchor to be uniformly optimal. Meanwhile, mean FDP remains below the nominal level throughout the reported experiments. Overall, these results support GLADE's two main roles: exploiting useful auxiliary structure to improve power when transferable information is available, and adaptively limiting its influence when it is not.

E.6 Scope and Limitations

The reported mean FDP values describe repeated-sampling performance under these constructions; they do not establish exchangeability under unrestricted future shifts. Network flows may remain dependent within connections, hosts, or time windows despite hash-disjoint splitting. CSE-CIC-IDS2018 also uses coarse attack-family groupings and has a limited target BruteForce inventory. Covertype lacks pixel coordinates for spatially blocked validation. Gas measurements share sensors, concentration regimes, and temporal batches, and treating Ammonia as the reference defines an application-specific testing problem. The results therefore support flow-, pixel-, and measurement-level improvements under the stated designs, not unrestricted generalization or a verification of the theoretical learning- and selection-accuracy conditions.