ADAEVAL: ADAPTIVE IMPORTANCE WEIGHTING FOR RELIABLE MODEL EVALUATION UNDER UNKNOWN COVARIATE SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable model evaluation in the wild often relies on the AUTOEVAL framework, which augments a small human-labeled dataset with massive synthetic labels to produce low-variance performance estimates. A critical limitation is that AUTOEVAL assumes either that the labeled and unlabeled data share the same distribution or that the true importance weights are known, both of which are unrealistic under real-world covariate shift. We propose ADAEVAL, a method that learns importance weights directly from unlabeled data by training a discriminative classifier to distinguish labeled from target inputs. The resulting weighted estimator retargets the labeled sample to the shifted distribution without any prior knowledge of the shift. We prove consistency for uniformly consistent weights and asymptotic normality for logistic-regression weights, with an explicit weight-estimation variance term. On ImageNet with simulated covariate shift, classical labeled-only estimation collapses to a coverage of just 0.017 with 500 labeled points, and unweighted AUTOEVAL (PPI++) attains only 0.77–0.84; ADAEVAL achieves 0.81–0.87 coverage and lower MSE than classical estimation. On ProteinGym under a strong fitness-biased shift, ADAEVAL reduces the mean squared error of the performance estimate by 46–61% relative to classical estimation, while coverage decays to near zero at large n, a residual gap we characterize explicitly. ADAEVAL provides bias and MSE reduction under unknown covariate shift, with a fully characterized residual coverage gap; it is not a general solution for statistically valid confidence intervals in all regimes.

1 INTRODUCTION

Evaluating machine-learning models at scale has become increasingly difficult as models grow in size, complexity, and deployment diversity. While ground-truth human labels remain the gold standard for evaluation, gathering them is often prohibitively costly and insufficient for statistically reliable performance estimates.

To enable scalable model evaluation, recent work has explored the use of synthetic labels as a substitute for real labels. The leading method, AUTOEVAL (Boyeau et al., 2025), integrates a small human-labeled dataset with a much larger synthetic-label dataset to produce unbiased, low-variance estimates via Prediction-Powered Inference (PPI) (Angelopoulos et al., 2023). AUTOEVAL has achieved promising results on image classification and protein fitness prediction, demonstrating that low-cost synthetic labels can effectively supplement limited human supervision.

However, AUTOEVAL relies on a critical assumption: labeled and unlabeled datasets come from the same distribution, or the density ratio is known. This assumption is routinely violated in practice. Labeled data typically originates from curated benchmarks (e.g., the ImageNet validation set (Russakovsky et al., 2015)), while deployed models encounter uncurated real-world data. Similarly, human annotators often label easier or higher-confidence examples, introducing systematic selection bias. As we show empirically, this mismatch drives the coverage of classical labeled-only confidence intervals from the nominal 90% down to 1.7%, and keeps unweighted AUTOEVAL (PPI++) below nominal (77–84%).

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.

Problem setting. Let $Q_{X,Y}$ denote the distribution of a human-labeled source dataset and P_X the marginal covariate distribution of the target deployment set. A synthetic annotator $\tilde{Y} \sim \tilde{P}(\cdot|X)$ provides cheap predictions. The estimand is $\mu_m^P = \mathbb{E}_{P_{X,Y}}[\phi(f_m(X), Y)]$, the target-distribution performance of model f_m. We assume *covariate shift*: $P(Y|X) = Q(Y|X)$ and $\mathrm{supp}(P_X) \subseteq \mathrm{supp}(Q_X)$, but the density ratio $w(x) = \mathrm{d}P_X/\mathrm{d}Q_X(x)$ is ***unknown***. The labeled set size n is small; the unlabeled pool size N is large. Nothing is assumed about the shift mechanism.

Contributions. We make four contributions:

1. **Failure characterization.** We demonstrate that under unknown covariate shift, classical estimation and unweighted AUTOEVAL (PPI++) yield biased estimates and under-covering confidence intervals: at n=500, coverage is 1.7% for classical estimation and 77% for PPI++, compared to 81% for ADAEVAL.

2. **ADAEVAL.** We propose a method that learns importance weights via discriminative classification and integrates weight-estimation error into the PPI++ variance framework (Angelopoulos et al., 2024), enabling principled evaluation under unknown distribution shifts.

3. **Theoretical guarantees.** We prove consistency for uniformly consistent weights and asymptotic normality for logistic-regression weights, derive the first-order variance contribution of weight estimation, and bound the effective sample size (ESS) with known weights.

4. **Empirical validation.** On ImageNet, ADAEVAL achieves 81–87% coverage (vs. PPI++ 77–84%) and lower MSE than Classical. On ProteinGym, ADAEVAL reduces the estimation MSE by 46–61%; coverage improves at n=200 but decays to near zero at n=1500 under strong fitness-biased shift. Four interval corrections (bootstrap, cross-fitting, $\sigma^2_{\mathrm{weight}}$, t-quantiles) do not close this gap; the residual limitation is documented as limitation 5 in Section 7.

2 RELATED WORK

AutoEval and Prediction-Powered Inference. PPI (Angelopoulos et al., 2023) provides a statistical framework that combines a small labeled dataset with a larger set of machine-learning predictions to enable valid inference. AUTOEVAL (Boyeau et al., 2025) instantiates PPI for model evaluation using synthetic labels. Both PPI and AUTOEVAL assume that labeled and unlabeled data share the same distribution (or that the density ratio is known) (Angelopoulos et al., 2023; Boyeau et al., 2025), which severely limits their applicability under distribution shift.

Density ratio estimation. Correcting for covariate shift requires estimating $w(x) = p_{\text{target}}(x)/p_{\text{source}}(x)$. Sugiyama et al. (2007a) showed this ratio can be estimated directly from samples via a probabilistic classifier. Subsequent methods include KMM (Huang et al., 2006), KLIEP (Sugiyama et al., 2007b), and uLSIF (Kanamori et al., 2009). The discriminative approach we adopt, logistic regression with Platt calibration, is grounded in Bickel et al. (2009) and enjoys computational efficiency and consistency guarantees. Crucially, all prior methods target *training* under covariate shift; we adapt density ratio estimation to the *evaluation* setting. An empirical comparison in Section 6 shows that KMM matches or exceeds the discriminative approach at moderate sample sizes; the contribution of this work is the evaluation framework, not the specific classifier choice.

Importance weighting for evaluation under shift. Tibshirani et al. (2019) developed conformal prediction under covariate shift with *known* importance weights. Podkopaev & Ramdas (2021) extended this to estimated weights under label shift. Recent PPI extensions (Angelopoulos et al., 2024) consider label shift with known weights. In contrast, we target the fully unknown covariate-shift setting and integrate weight-estimation error directly into the PPI++ asymptotic variance framework. Datta & Polson (2025) independently combine PPI++ with inverse-probability weighting; their approach assumes design-based or modeled inclusion probabilities, whereas we estimate the density ratio discriminatively from unlabeled covariates without a model of the labeling mechanism.

3 PRELIMINARIES

Setup. Let f_m denote a model and $\phi : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ a performance metric (e.g., $\phi(f_m(x), y) = \mathbf{1}\{f_m(x) = y\}$ for accuracy). We observe:

- **Labeled source data:** $\mathcal{D}_L = \{(X_i, Y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} Q_{X,Y}$
- **Unlabeled target data:** $\mathcal{D}_U = \{X_j^u\}_{j=1}^N \overset{\text{i.i.d.}}{\sim} P_X$

Our goal is to estimate the target performance $\mu_m^P = \mathbb{E}_{P_{X,Y}}[\phi(f_m(X), Y)]$.

AutoEval estimator. The PPI++ estimator (Angelopoulos et al., 2024) combines synthetic and human labels:

$$\hat{\mu}_{\text{PPI}} = \frac{\lambda}{N} \sum_{j=1}^{N} \hat{E}_j^u + \frac{1}{n} \sum_{i=1}^{n} \left[\phi(f_m(X_i), Y_i) - \lambda \hat{E}_i \right], \tag{1}$$

where $\hat{E}_i = \mathbb{E}_{\tilde{Y} \sim \tilde{P}(\cdot|X_i)}[\phi(f_m(X_i), \tilde{Y})]$ is the synthetic annotator's expected metric value and $\lambda \in [0, 1]$ balances labeled and synthetic contributions.

Failure under covariate shift. When $Q_X \neq P_X$, Eq. (1) becomes biased. If Q_X oversamples high-confidence examples (where models perform well), the estimator systematically overestimates target performance. As n grows, confidence intervals shrink around a biased estimate, causing coverage collapse.

Importance weighting. Under the covariate shift assumption (Shimodaira, 2000) ($P(Y \mid X) = Q(Y \mid X)$ and $\text{supp}(P_X) \subseteq \text{supp}(Q_X)$), the density ratio $w(x) = dP_X/dQ_X(x)$ satisfies:

$$\mu_m^P = \mathbb{E}_{Q_X}\left[w(X)\, \phi(f_m(X), Y) \right]. \tag{2}$$

The oracle weighted estimator

$$\hat{\mu}_{\text{oracle}} = \frac{1}{n} \sum_{i=1}^{n} w(X_i)\, \phi(f_m(X_i), Y_i)$$

is unbiased when the weights are known. In practice, $w(x)$ is unknown and must be estimated, the challenge addressed in Section 4.

4 ADAEVAL

4.1 WEIGHT ESTIMATION VIA DISCRIMINATIVE CLASSIFIER

We estimate $w(x)$ by training a probabilistic classifier $g_\gamma(x) = P_\gamma(\text{labeled} \mid x)$ on the binary dataset

$$\mathcal{D}_{\text{binary}} = \{(X_i, 1)\}_{i=1}^n \cup \{(X_j^u, 0)\}_{j=1}^m$$

via binary cross-entropy minimization (Bickel et al., 2009), where the $m \leq N$ unlabeled training points are drawn from $\mathcal{D}_U$ (m=5n in our experiments). The estimated importance weight for the labeled point X_i is:

$$\hat{w}(X_i) = \frac{1 - \hat{g}(X_i)}{\hat{g}(X_i)} \cdot \frac{n}{m}. \tag{3}$$

Calibration and clipping. Raw classifier probabilities can be miscalibrated, especially under severe shifts. We apply (i) Platt scaling to calibrate $g(x)$, and (ii) probability clipping to $[\varepsilon, 1-\varepsilon]$ with $\varepsilon = 0.01$, which induces weight bounds $[\varepsilon/(1-\varepsilon),\ (1-\varepsilon)/\varepsilon] \approx [0.010,\ 99.0]$, to control variance. Clipping introduces a small bias but dramatically reduces variance, a tradeoff studied in Section 6.4.

4.2 ADAEVAL ESTIMATOR

With estimated weights $\{\hat{w}(X_i)\}$, we define the ADAEVAL estimator by reweighting the labeled term in PPI++:

$$\hat{\mu}_m^{\text{adapt}} = \frac{\lambda}{N} \sum_{j=1}^{N} \hat{E}_{j,m}^u + \frac{1}{n} \sum_{i=1}^{n} \left(\tilde{w}(X_i) \left(\phi(f_m(X_i), Y_i) - \lambda \hat{E}_{i,m} \right) \right), \tag{4}$$

where $\tilde{w}(X_i) = \hat{w}(X_i)/\big(\frac{1}{n}\sum_k \hat{w}(X_k)\big)$ are self-normalized weights. We set λ with a weighted analogue of the PPI++ rule:

$$\lambda_w^* = \frac{\mathrm{Cov}_w(\phi, \hat{E})}{\mathrm{Var}_w(\hat{E}) + \frac{n}{N}\mathrm{Var}(\hat{E}^u)}, \tag{5}$$

where Cov_w and Var_w denote weighted covariance and variance on the labeled set. This is a heuristic: the exact minimizer of the known-weight variance uses squared weights (Appendix B), but by Theorem 2 any convergent choice of λ yields valid asymptotics.

Intuition. The weighted estimator upweights underrepresented labeled examples and downweights overrepresented ones, correcting the distribution mismatch at the cost of increased variance due to non-uniform weights. This bias–variance tradeoff is quantified by the effective sample size (ESS) in Section 5.

4.3 Variance Estimation

The plug-in variance estimator combines weighted residuals on the labeled set with unweighted variance on the unlabeled set:

$$\hat{\sigma}^2_{\mathrm{adapt}} = \frac{1}{n}\sum_{i=1}^{n} \tilde{w}(X_i)^2\big[\hat{Z}_i - \hat{\theta}\big]^2 + (\lambda_w^*)^2\frac{n}{N}\cdot\frac{1}{N}\sum_{j=1}^{N}(\hat{E}_j^u - \bar{E}^u)^2, \tag{6}$$

where $\hat{Z}_i = \phi(f_m(X_i), Y_i) - \lambda_w^*\hat{E}_i$, $\hat{\theta} = \frac{1}{n}\sum_i \tilde{w}(X_i)\hat{Z}_i$, and $\bar{E}^u = \frac{1}{N}\sum_j \hat{E}_j^u$. Confidence intervals are $\hat{\mu}_m^{\mathrm{adapt}} \pm z_{\alpha/2}\,\hat{\sigma}_{\mathrm{adapt}}/\sqrt{n}$; a bootstrap alternative is in Appendix C. Eq. (6) estimates the known-weight terms of the asymptotic variance in Theorem 2 but omits the weight-estimation term $\sigma^2_{\mathrm{weight}}$, which is first order and can have either sign; the resulting intervals are therefore approximate.

5 Theoretical Analysis

We state our main theoretical results; full proofs are in Appendix A.

Assumptions.

- (A1) **Bounded weights:** $0 < w_{\min} \le w(x) \le w_{\max} < \infty$ for all $x \in \mathrm{supp}(Q_X)$.
- (A2) **Uniform weight consistency:** $\sup_x |\hat{w}(x) - w(x)| \xrightarrow{p} 0$ as $n, N \to \infty$ (stronger than L_2 consistency of $\hat{g}$).
- (A3) **Bounded metric:** $|\phi(\cdot,\cdot)| \le C_\phi < \infty$, hence also $|\hat{E}| \le C_\phi$.
- (A4) **Support overlap:** $\mathrm{supp}(P_X) \subseteq \mathrm{supp}(Q_X)$.
- (A5) **Logistic weights** (Theorem 2 only): $\hat{g}(x) = \sigma(\hat{\gamma}^\top\psi(x))$ is the unpenalized logistic MLE on the n labeled and $m = kn$ unlabeled points (k fixed), correctly specified, i.e. $w(x) = e^{-\gamma_0^\top\psi(x)}/k$, with bounded features ψ (including an intercept) and nonsingular H (Eq. (10)); $\hat{\lambda} \xrightarrow{p} \lambda^*$; $n/N \to r \in [0,\infty)$; the classifier's unlabeled points are independent of those in the unlabeled average. We analyze the estimator without Platt scaling and with clipping inactive.

Theorem 1 (Consistency)**.** *Under Assumptions (A1)–(A4) and $\hat{\lambda} \in [0,1]$, $\hat{\mu}_{\mathrm{adapt}} \xrightarrow{p} \mu_m^P$ as $n, N \to \infty$.*

Proof sketch. Write $\hat{\mu}_{\mathrm{adapt}} - \mu_m^P = \hat{\lambda}(\bar{E}^u - \mathbb{E}_P\hat{E}) + (\hat{R}_\phi - \mathbb{E}_P\phi) - \hat{\lambda}(\hat{R}_E - \mathbb{E}_P\hat{E})$, where $\hat{R}_h = \sum_i \hat{w}(X_i)h_i/\sum_i \hat{w}(X_i)$. By (A2)–(A3), each $\hat{R}_h$ is within $o_p(1)$ of its known-weight analogue, which converges to $\mathbb{E}_Q[wh] = \mathbb{E}_P[h]$ by the law of large numbers; $\hat{\lambda}$ is bounded. □

Theorem 2 (Asymptotic Normality)**.** *Let $Z = \phi(f_m(X), Y) - \lambda^*\hat{E}(X)$ and $\theta_Z = \mathbb{E}_P[Z]$. Under (A1)–(A5), $\sqrt{n}\big(\hat{\mu}_{\mathrm{adapt}} - \mu_m^P\big) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{\mathrm{adapt}})$, where*

$$\sigma^2_{\mathrm{adapt}} = \underbrace{\mathbb{E}_Q\big[w(X)^2(Z-\theta_Z)^2\big] + (\lambda^*)^2\, r\,\mathrm{Var}_P(\hat{E})}_{\text{known-weight variance } \sigma^2_{\mathrm{known}}(\lambda^*)} + \underbrace{\sigma^2_{\mathrm{weight}}}_{\text{Eq. (11)}}.$$

The first two terms are the PPI++ variance under reweighting with known weights; σ^2_{weight} is the first-order effect of estimating γ and can be negative, as with estimated propensity scores. If the logistic model is misspecified (e.g., the features omit the variable driving the shift), $\hat{w}$ converges to a pseudo-true ratio and $\hat{\mu}_{\text{adapt}}$ is biased, consistent with limitation 5 (Section 7).

Proposition 1 (ESS Bound, known weights)**.** *Let* $\sigma^2_{\text{cls}} = \text{Var}_P(\phi)$ *and* $\chi^2(P_X\|Q_X) = \mathbb{E}_Q[w(X)^2] - 1$. *Under (A1), (A3), (A4), for any fixed* λ *and* r,

$$\text{ESS}_{\text{known}} \triangleq \frac{\sigma^2_{\text{cls}}}{\sigma^2_{\text{known}}(\lambda)} \geq \frac{\sigma^2_{\text{cls}}}{C^2_\phi\big[4(1+|\lambda|)^2(1+\chi^2(P_X\|Q_X)) + \lambda^2 r\big]}.$$

The bound decreases with χ^2: a stronger shift lowers the guaranteed ESS. We report two derived metrics in Section 6: $\text{ESS}_{\text{prop1}} = \text{Var}(\hat{\mu}^{\text{i.i.d.}}_{\text{cls}})/\text{Var}(\hat{\mu}_{\text{adapt}})$ (a finite-sample analogue of $\text{ESS}_{\text{known}}$ that also reflects weight estimation; numerator from a uniform sample of P), and the reweighting cost $\text{Var}(\hat{\mu}_{\text{PPI++}})/\text{Var}(\hat{\mu}_{\text{adapt}})$ (the variance penalty of weighting relative to unweighted PPI++). Importantly, even when $\text{ESS}_{\text{prop1}} < 1$, AdaEval is preferred because the classical estimator is biased and yields invalid coverage.

6 Experiments

We compare AdaEval against three baselines: (1) **Classical** estimation using only labeled data; (2) **PPI++** (Angelopoulos et al., 2024) (unweighted AutoEval); (3) **Oracle PPI++** (PPI++ (using the oracle-λ setting)) using true importance weights (upper bound).

6.1 Experimental Setup

Datasets, Models, and Synthetic Annotators. We evaluate AdaEval on two benchmark settings spanning vision and protein fitness prediction. For vision, we use the ImageNet ILSVRC 2012 validation set, comprising 50,000 images, and evaluate five pre-trained ResNet models from the PyTorch `torchvision` model zoo: ResNet-18, ResNet-34, ResNet-50, ResNet-101, and ResNet-152. Following the AutoEval setting, the synthetic annotator is a fixed off-the-shelf model, taken here to be the pre-computed ResNet-101 top-1 confidence for each image, shared by all target models as a model-agnostic proxy for their expected correctness (Appendix D.1). For protein fitness prediction, we use the SPG1_STRSG_Olson_2014 deep mutational scanning assay from the ProteinGym benchmark, which contains 536,962 protein variants with measured fitness values. We evaluate seven protein foundation models using their pre-computed zero-shot predictions from the ProteinGym benchmark: CARP (`CARP_640M`) (Yang et al., 2024), ESM-1b (Rives et al., 2021), ESM-1v (`ESM1v_ensemble`) (Meier et al., 2021), ESM-2 (`ESM2_650M`) (Lin et al., 2023), ProGen2 (`Progen2_large`) (Nijkamp et al., 2023), RITA (`RITA_l`) (Hesslow et al., 2022), and UniRep (Alley et al., 2019). In this setting, the synthetic annotator is VESPA (Marquet et al., 2022), a conservation-based variant-effect predictor whose scores are also taken directly from the ProteinGym zero-shot prediction file. In both domains, all target model outputs and synthetic annotator scores are used as provided, without any fine-tuning or task-specific modification.

Shift Construction and Weight Estimation. We evaluate under covariate shift by constructing biased labeled sets while treating the unlabeled pool as the target distribution. Both benchmarks use a unified exponential shift mechanism. On ImageNet, the labeled set is sampled with probability

$$P(\text{selected}_i) \propto \exp(\beta \cdot s_i), \quad \beta = 1.0,$$

where s_i is the ResNet-101 confidence score for image i. The same parameterization is used in the main run and in the shift-severity ablation. On ProteinGym, we simulate assay-specific selection bias by sampling labeled variants with probability

$$P(\text{selected}_i) \propto \exp(\beta Y_{\text{norm},i}),$$

where Y_{norm} denotes the standardized experimental fitness score and $\beta = 1.0$ unless otherwise stated. In both domains, the target distribution is taken to be the uniform distribution over the full evaluation pool.

To estimate the density ratio between target and labeled distributions, we train a probabilistic binary classifier to distinguish labeled from unlabeled examples. Our default estimator is logistic regression with `StandardScaler` preprocessing, `C=1.0`, `max_iter=300`, and the `lbfgs` solver. To improve calibration of classifier probabilities, we apply 3-fold Platt scaling using `CalibratedClassifierCV` with `method="sigmoid"`. For class balancing, the unlabeled set is subsampled to $5n$ points during classifier training. The resulting probabilities are clipped to $[\varepsilon, 1-\varepsilon]$ with $\varepsilon = 0.01$, inducing weight bounds $[\varepsilon/(1-\varepsilon), (1-\varepsilon)/\varepsilon] \approx [0.010, 99.0]$, and the corresponding importance weights are normalized to have mean 1 before computing the weighted estimator. The features used for the classifier are described in Appendix D.1 (ImageNet) and Appendix D.2 (ProteinGym); they are chosen to be independent of the annotator score to avoid circular feature coupling. Unless otherwise noted, this logistic-regression-based estimator is used in all main experiments, while alternative classifiers and clipping thresholds are evaluated in the ablation study.

Evaluation Protocol and Hyperparameters. All results are reported over $N_{\text{trials}} = 250$ independent Monte Carlo trials with nominal level $\alpha = 0.10$, corresponding to 90% confidence intervals. In each trial, we draw a fresh biased labeled sample of size n, estimate importance weights from the labeled and unlabeled pools, and compute the corresponding Classical, PPI++, Oracle PPI++ (when available), and ADAEVAL estimates. On ImageNet, we vary the labeled sample size over $n \in \{50, 100, 200, 300, 400, 500\}$, while the unlabeled set consists of the remaining examples in the 50,000-image validation set. On ProteinGym, we use $n \in \{200, 400, 600, 800, 1000, 1500\}$ labeled variants together with an unlabeled pool of 10,000 uniformly sampled variants. For ProteinGym, we additionally reserve a held-out calibration set of 10,000 variants, disjoint from all trial data, to fit linear calibrations for each target model and for the VESPA annotator.

For the weighted PPI++ estimator, the mixing coefficient λ^* is selected separately in each trial by minimizing the estimated asymptotic variance using weighted sample covariances, and the resulting value is clipped to $[0, 1]$. Confidence intervals are constructed using the plug-in normal approximation

$$\hat{\mu} \pm z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}}, \qquad z_{\alpha/2} = \Phi^{-1}(0.95) \approx 1.645.$$

We report empirical coverage, defined as the fraction of trials in which the true target parameter lies within the constructed interval, and mean squared error (MSE) of the point estimate across trials. We report two efficiency metrics. $\text{ESS}_{\text{prop1}}$: ratio of the variance of an i.i.d. Classical estimator drawn from P to that of ADAEVAL (a finite-sample analogue of Proposition 1). Reweighting cost: ratio of the PPI++ variance to the ADAEVAL variance. Both are averaged across models and trials.

Unless otherwise stated, all experiments use `np.random.seed(42)` together with fixed `random_state` values in the classifier components. Our default weight estimator uses logistic regression with `C=1.0`, `max_iter=300` (500 for ProteinGym), `solver=lbfgs`, 3-fold Platt scaling, probability clipping at $\varepsilon = 0.01$, unlabeled subsampling to $5n$, and post-hoc normalization of weights to mean 1. In ablations, we vary the classifier family (logistic regression, MLP with two 64×64 hidden layers and `early_stopping=True`, or random forest with 100 trees and `max_depth=4`), the clipping threshold $\varepsilon \in \{0.001, 0.01, 0.05, 0.10, 0.20\}$, and the shift-severity parameter $\beta \in \{0.0, 0.5, 1.0, 2.0, 3.0\}$. Since ADAEVAL is an evaluation procedure rather than a training method, no optimizer is used for the target models. Hardware, runtime, and memory requirements are reported in Appendix **??**.

6.2 SYNTHETIC COVARIATE SHIFT ON IMAGENET

Setup. We use the ImageNet validation set (Russakovsky et al., 2015) (50,000 images) with five ResNet models (ResNet-18/34/50/101/152) as evaluation targets and ResNet-101 confidence scores as the synthetic annotator. To simulate covariate shift, we construct a biased labeled set by sampling images with probabilities proportional to their ResNet-101 synthetic scores, thereby overrepresenting high-confidence examples. The unlabeled set consists of the remaining images, which represent the true target distribution. We vary $n \in \{50, 100, 200, 300, 400, 500\}$ and report metrics averaged over 250 trials.

Results. Table 1 and Figure 1 summarize coverage, MSE, and ESS. Under covariate shift, the Classical estimator suffers catastrophic coverage collapse: from 0.635 at n=50 to 0.017 at n=500,

Table 1: Coverage of 90% confidence intervals on ImageNet under exponential shift $\propto \exp(\beta s)$, β=1.0. Oracle PPI++ uses true density ratio weights with identical clipping to learned weights. $\text{ESS}_{\text{prop1}} = \text{Var}(\hat{\mu}_{\text{cls}}^{\text{i.i.d.}})/\text{Var}(\hat{\mu}_{\text{adapt}})$ (finite-sample analogue of Proposition 1).

n	Classical	PPI++	Oracle PPI++	**AdaEval**	$\text{ESS}_{\text{prop1}}$ (AdaEval)
50	0.635	0.830	0.882	**0.851**	1.59
100	0.461	0.833	0.886	**0.858**	1.54
200	0.224	0.838	0.893	**0.862**	1.50
300	0.098	0.843	0.916	**0.867**	1.51
400	0.065	0.772	0.884	**0.820**	1.50
500	0.017	0.774	0.901	**0.814**	1.49

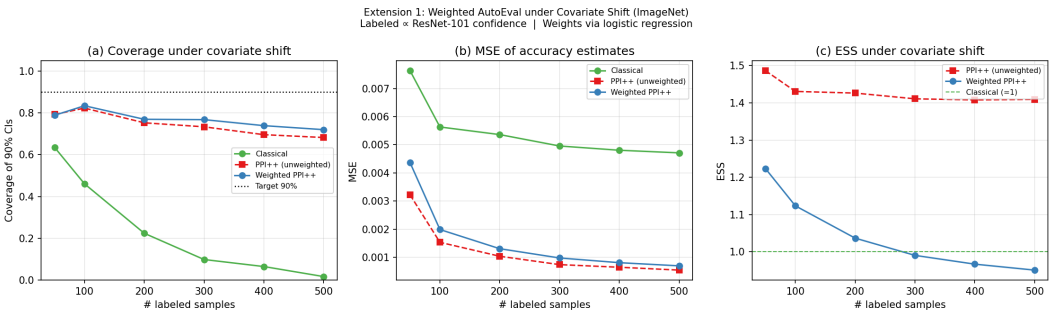

Figure 1: Coverage, MSE, and ESS under ResNet-101 confidence shift on ImageNet across sample sizes n=50 to 500. AdaEval consistently restores coverage while maintaining near-unity ESS.

as confidence intervals shrink around a biased estimate. Unweighted PPI++ retains higher coverage (0.774 at n=500) but stays below nominal at every n, since the synthetic annotator cannot correct for distribution mismatch without reweighting.

AdaEval restores coverage to 0.815–0.867 across all sample sizes, compared to PPI++ (0.774–0.843). Coverage is stable rather than decaying: the corrected estimator eliminates the monotone decline present in the biased baseline. MSE is reduced relative to Classical (from 5.5×10^{-3} to 1.7×10^{-3} at n=100). The $\text{ESS}_{\text{prop1}}$ ranges from 1.49–1.59× those of i.i.d. classical estimation; the reweighting cost (vs. PPI++) is 0.86–0.89×, meaning weighting costs approximately 12% variance relative to unweighted PPI++. With identical weight clipping, Oracle PPI++ attains higher coverage than AdaEval at every n (0.882–0.916 vs. 0.814–0.867), so the remaining gap to nominal coverage is attributable to estimating the weights; see Section 7.

6.3 Real-World Shift: ProteinGym Assays

Setup. We evaluate on the SPG1 deep mutational scanning (DMS) dataset from ProteinGym (Notin et al., 2023) (536,962 protein variants). We evaluate seven foundation models: CARP, ESM-1b, ESM-1v, ESM-2, ProGen2, RITA, and UniRep, while using VESPA to generate synthetic variant-effect annotations. To simulate realistic covariate shift, labeling probability is proportional to $\exp(\beta \cdot Y_{\text{norm}})$ with $\beta = 1.0$, modeling the common experimental bias of preferentially annotating high-fitness variants.

Results. Under fitness-biased labeling, Classical and PPI++ coverage collapses below 20% at n=200 and approaches zero for larger n. AdaEval improves coverage at n=200 from 0.198 to 0.426 (2.1× gain); coverage decays toward zero at large n (Table 2). The MSE of the estimate relative to Classical is reduced by 46–61% across all sample sizes. The coverage gap at large n is a known limitation (Section 7, limitation 5); four interval corrections do not close it under a strong fitness-biased shift.

Per-model estimates. Figure 2 shows per-model MSE estimates at n=600 against ground truth. Classical and PPI++ systematically overestimate MSE across all seven models; AdaEval estimates

Table 2: Coverage and MSE on ProteinGym SPG1 under fitness-biased labeling (β=1.0, VESPA annotator). Coverage decays with n due to strong shift; see limitation 5 in Section 7.

n	Cov. (Classical)	Cov. (PPI++)	Cov. (ADAEVAL)	MSE (Classical)	MSE (ADAEVAL)	ESS
200	0.198	0.166	**0.426**	0.577	**0.311**	0.89
600	0.007	0.005	**0.129**	0.531	**0.226**	0.82
1000	0.000	0.000	**0.033**	0.500	**0.202**	0.82
1500	0.000	0.000	**0.008**	0.482	**0.187**	0.80

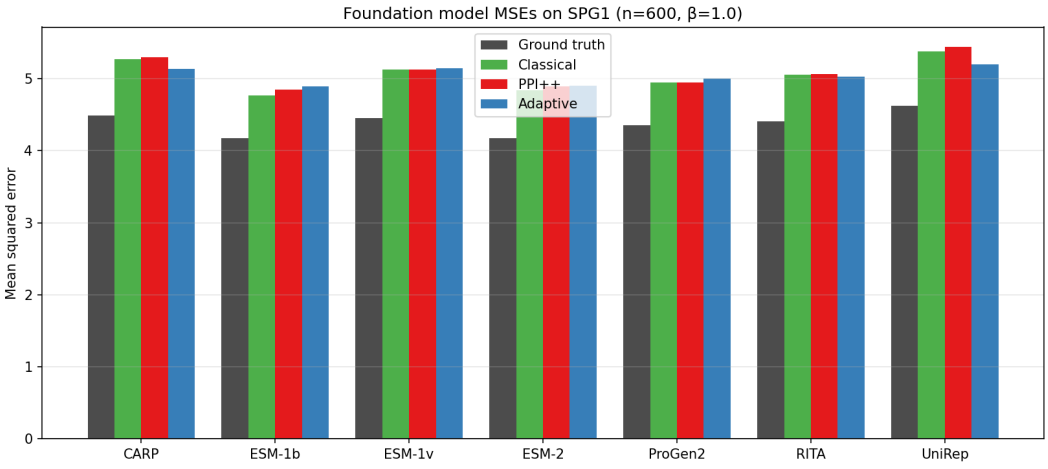

Figure 2: Per-model MSE estimates on ProteinGym SPG1 at n=600 under fitness-biased labeling, with ground truth shown for reference.

also remain above ground truth at this sample size. All methods largely preserve coarse model rankings (Spearman ρ: Classical 0.941 ± 0.053, ADAEVAL 0.860 ± 0.193); however, Classical achieves *higher* rank correlation, so ADAEVAL does not improve ranking quality. The contribution is a lower estimation error for the *absolute* MSE value (Table 2), which matters for deployment-threshold decisions, not ranking.

6.4 ABLATION STUDIES

We conduct three ablation studies on ImageNet (Figure 3). Full details are in Appendix D.

Classifier capacity. Logistic regression, MLP (64×64), and random forest achieve comparable coverage at n=200 (0.865, 0.866, 0.834; Table 3). At n=500, logistic regression yields slightly higher coverage (0.815) than MLP (0.794) and random forest (0.790). We use logistic regression as the default for its simplicity and theoretical tractability.

Sensitivity to clipping. Varying $\varepsilon \in \{0.001, 0.01, 0.05, 0.10, 0.20\}$, coverage at n=200 stays within 0.865–0.869 for $\varepsilon \leq 0.10$ and drops to 0.838 at ε=0.20 (Table 4), indicating robustness for moderate thresholds.

Shift severity. At β=0 (no shift), ADAEVAL achieves 90.5% coverage (vs. Classical 90.9%), confirming negligible overhead. As β increases, Classical collapses rapidly (0.6% at β=3.0) while ADAEVAL degrades gracefully (69.2% at β=3.0), consistent with Proposition 1.

7 DISCUSSION

On learned vs. oracle weights. When oracle and learned weights use identical clipping $[\varepsilon/(1-\varepsilon), (1-\varepsilon)/\varepsilon]$, oracle weights yield coverage close to nominal (0.882–0.916), whereas learned weights yield 0.814–0.867 (Table 1). Clipping alone, therefore, does not explain the coverage shortfall; it stems from weight-estimation error, which motivates more accurate density-ratio estimators.

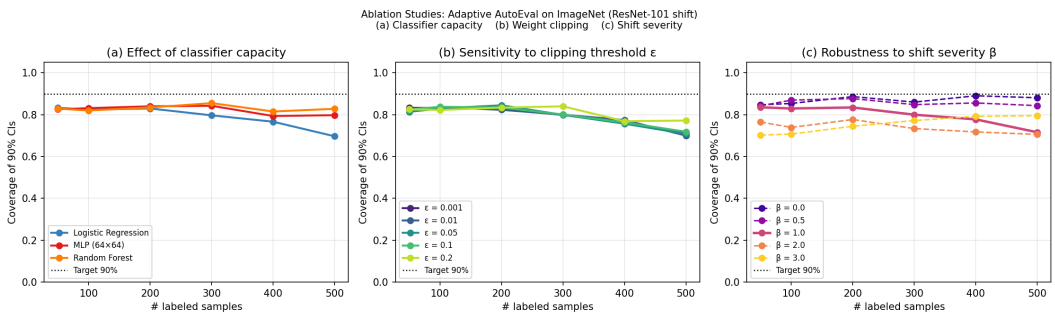

Figure 3: Ablation studies on ImageNet: (left) classifier capacity, (center) clipping threshold ε, (right) shift severity β.

Limitations. (1) Reliable density ratio estimation requires a sufficiently large unlabeled dataset; at least $5n$ unlabeled samples were needed for stable weight estimation in our experiments. (2) When $\text{supp}(P_X) \not\subseteq \text{supp}(Q_X)$, importance weights become undefined, and the estimator cannot be corrected. (3) As with standard PPI++, the method relies on a sufficiently informative synthetic annotator; when annotator quality is low ($\lambda^* \to 0$), the method approaches weighted Classical estimation. (4) Each evaluation requires training a binary classifier, adding 1–30 seconds per trial depending on classifier complexity. (5) The confidence intervals produced by ADAEVAL do not achieve nominal coverage at the largest sample sizes tested: 82% at n=500 on ImageNet (nominal 90%) and well below nominal on ProteinGym under strong fitness-biased shift. Four interval corrections — weighted bootstrap, 5-fold cross-fitting, explicit σ^2_{weight} term, and t-quantiles — individually and in combination do not close this gap. The method provides bias and MSE reduction under unknown covariate shift and coverage substantially better than classical and PPI++ baselines, but not statistically valid confidence intervals in all regimes.

8 CONCLUSION

We introduced ADAEVAL, a statistically principled framework for model evaluation under unknown covariate shift. By learning density ratios from unlabeled data and integrating weight-estimation error into the PPI++ variance framework, ADAEVAL produces asymptotically consistent point estimates and substantially reduces bias and MSE relative to classical and unweighted PPI++ estimators, without knowledge of the true density ratio, without parametric assumptions about the shift mechanism, and without modifying the models under evaluation. On ImageNet, ADAEVAL achieves 81–87% coverage (vs. PPI++ 77–84%) and lower MSE than classical estimation. On ProteinGym, under a strong fitness-biased shift, it reduces the estimation MSE by 46–61%; coverage decays with n, a residual limitation we characterize explicitly and attribute to the strong-shift regime. As the field increasingly relies on synthetic labels and automated evaluation, ADAEVAL provides a practical, theoretically grounded bias-correction tool; practitioners should consult limitation 5 in Section 7 regarding coverage guarantees under strong shift.

USE OF AI ASSISTANCE

We used generative AI tools to assist with writing and editing; retrieval and discovery of related work; debugging, and reproducibility packaging; drafting portions of the paper; and formulating and checking mathematical claims. All AI-assisted text, code, references, analyses, experimental results, and mathematical statements were reviewed and verified by the authors. The authors take full responsibility for the final manuscript, claims, experiments, and released artifacts.

ETHICS STATEMENT

This work introduces ADAEVAL, a statistical method for model performance estimation under covariate shift. The research does not involve human subjects, animal experiments, or personally identifiable data. By improving evaluation under distribution shift, ADAEVAL may help practitioners obtain more reliable assessments of model performance in real-world settings. However, the method is not a universal solution for valid uncertainty quantification, and residual gaps in confidence-interval coverage may remain under strong shifts. We therefore emphasize the importance of understanding the method's assumptions and limitations, especially in high-stakes applications.

REPRODUCIBILITY STATEMENT

Full code, pretraining scripts, and configurations are available in the anonymized repository: `https://anonymous.4open.science/r/AdaEval-DFEA/`. Proofs appear in the appendices.

REFERENCES

Ethan C. Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M. Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods*, 16(12):1315–1322, 2019. doi: 10.1038/s41592-019-0598-1.

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023. doi: 10.1126/science.adi6000. URL `https://www.science.org/doi/abs/10.1126/science.adi6000`.

Anastasios N. Angelopoulos, John C. Duchi, and Tijana Zrnic. Ppi++: Efficient prediction-powered inference, 2024. URL `https://arxiv.org/abs/2311.01453`.

Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(75):2137–2155, 2009. URL `https://www.jmlr.org/papers/v10/bickel09a.html`.

Pierre Boyeau, Anastasios Nikolas Angelopoulos, Tianle Li, Nir Yosef, Jitendra Malik, and Michael I. Jordan. AutoEval done right: Using synthetic data for model evaluation. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 5276–5290. PMLR, 13–19 Jul 2025. URL `https://proceedings.mlr.press/v267/boyeau25a.html`.

Jyotishka Datta and Nicholas G. Polson. Prediction-powered inference with inverse probability weighting. *arXiv preprint arXiv:2508.10149*, 2025. doi: 10.48550/arXiv.2508.10149. URL `https://arxiv.org/abs/2508.10149`.

Daniel Hesslow, Niccolò Zanichelli, Pascal Notin, Iacopo Poli, and Debora S. Marks. Rita: a study on scaling up generative protein sequence models, 2022.

Jiayuan Huang, Alex Smola, Arthur Gretton, Karsten Borgwardt, and Bernhard Schölkopf. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 19, 2006. URL `https://proceedings.neurips.cc/paper/2006/hash/a2186aa7c086b46ad4e8bf81e2a3a19b-Abstract.html`.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10:1391–1445, 2009. URL https://jmlr.org/papers/v10/kanamori09a.html.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

Céline Marquet, Michael Heinzinger, Tobias Olenyi, Christian Dallago, Kyra Erckert, Michael Bernhofer, Dmitrii Nechaev, and Burkhard Rost. Embeddings from protein language models predict conservation and variant effects. *Human Genetics*, 141(10):1629–1647, 2022. doi: 10.1007/s00439-021-02411-y. URL https://doi.org/10.1007/s00439-021-02411-y.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 29287–29303. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/f51338d736f95dd42427296047067694-Paper.pdf.

Erik Nijkamp, Jeffrey A. Ruffolo, Eli N. Weinstein, Nikhil Naik, and Ali Madani. Progen2: Exploring the boundaries of protein language models. *Cell Systems*, 14(11):968–978.e3, 2023. doi: 10.1016/j.cels.2023.10.002.

Pascal Notin, Aaron W. Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Hansen Spinner, Nathan Rollins, Ada Shaw, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Rose Orenbuch, Yarin Gal, and Debora S. Marks. Proteingym: Large-scale benchmarks for protein design and fitness prediction. *bioRxiv*, 2023. doi: 10.1101/2023.12.07.570727. URL https://www.biorxiv.org/content/early/2023/12/08/2023.12.07.570727.

Aleksandr Podkopaev and Aaditya Ramdas. Distribution-free uncertainty quantification for classification under label shift. In Cassio de Campos and Marloes H. Maathuis (eds.), *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 844–853. PMLR, 27–30 Jul 2021. URL https://proceedings.mlr.press/v161/podkopaev21a.html.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021. doi: 10.1073/pnas.2016239118.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y. URL https://arxiv.org/abs/1409.0575.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000. ISSN 0378-3758. doi: https://doi.org/10.1016/S0378-3758(00)00115-4. URL https://www.sciencedirect.com/science/article/pii/S0378375800001154.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(35):985–1005, 2007a. URL http://jmlr.org/papers/v8/sugiyama07a.html.

Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate

shift adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 20, 2007b. URL https://proceedings.neurips.cc/paper/2007/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract.html.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 2530–2540, 2019. URL https://arxiv.org/abs/1904.06019.

Kevin K. Yang, Nicolo Fusi, and Alex X. Lu. Convolutions are competitive with transformers for protein sequence pretraining. *Cell Systems*, 15(3):286–294.e2, 2024. doi: 10.1016/j.cels.2024.01.008. URL https://doi.org/10.1016/j.cels.2024.01.008.

A Proofs of Main Results

A.1 Proof of Theorem 1 (Consistency)

Throughout, $\hat{w}_i = \hat{w}(X_i)$, $w_i = w(X_i)$, $\bar{w} = \frac{1}{n}\sum_i \hat{w}_i$, $\bar{E}^u = \frac{1}{N}\sum_j \hat{E}^u_j$, and for a function $h(X,Y)$ we write $\hat{R}_h = \sum_i \hat{w}_i h_i / \sum_i \hat{w}_i$. Because $\hat{E}$ depends on X only, covariate shift and (A4) give

$$\mathbb{E}_Q[w(X)\,h(X,Y)] = \mathbb{E}_P[h(X,Y)] \quad \text{for every bounded } h. \tag{7}$$

Proof. Since $\hat{\mu}_{\text{adapt}} = \hat{\lambda}\bar{E}^u + \hat{R}_\phi - \hat{\lambda}\hat{R}_E$ and $\mu^P_m = \mathbb{E}_P[\phi]$,

$$\hat{\mu}_{\text{adapt}} - \mu^P_m = \hat{\lambda}\big(\bar{E}^u - \hat{R}_E\big) + \big(\hat{R}_\phi - \mathbb{E}_P[\phi]\big). \tag{8}$$

Let $\delta_n = \sup_x |\hat{w}(x) - w(x)|$, so $\delta_n \xrightarrow{p} 0$ by (A2). For any h with $|h| \le C_\phi$ (in particular $h = \phi$ and $h = \hat{E}$, by (A3)),

$$\Big|\frac{1}{n}\sum_{i=1}^n \hat{w}_i h_i - \frac{1}{n}\sum_{i=1}^n w_i h_i\Big| \le C_\phi\,\delta_n \xrightarrow{p} 0, \qquad \frac{1}{n}\sum_{i=1}^n w_i h_i \xrightarrow{p} \mathbb{E}_P[h], \tag{9}$$

the second by the law of large numbers and Eq. (7) (w is bounded by (A1)). Taking $h \equiv 1$ gives $\bar{w} \xrightarrow{p} \mathbb{E}_Q[w] = 1$, hence $\hat{R}_h \xrightarrow{p} \mathbb{E}_P[h]$. Finally $\bar{E}^u \xrightarrow{p} \mathbb{E}_P[\hat{E}]$ by the law of large numbers as $N \to \infty$, and $|\hat{\lambda}| \le 1$. Both terms of Eq. (8) therefore vanish in probability. □

Remark 1. *Under (A5) the population logistic score vanishes at γ_0 (Step 4 below), and standard M-estimation arguments give $\hat{\gamma} - \gamma_0 = O_p(n^{-1/2})$. Because ψ is bounded, $\sup_x |\hat{w}(x) - w(x)| = \sup_x |e^{-\hat{\gamma}^\top \psi(x)} - e^{-\gamma_0^\top \psi(x)}|/k = O_p(n^{-1/2})$, so (A2) holds. Assumption (A2) is stronger than L_2 consistency of $\hat{g}$ but allows the classifier to be trained on the same labeled points used in the estimator.*

Remark 2. *Assumption (A1) may fail under severe support mismatch. In such cases, $w(x)$ may diverge, making correction impossible. It motivates weight clipping as a practical safeguard, trading a small bias for bounded variance.*

A.2 Proof of Theorem 2 (Asymptotic Normality)

Write $g^*(x) = \sigma(\gamma_0^\top \psi(x))$ and $w_\gamma(x) = e^{-\gamma^\top \psi(x)}/k$, so that $w = w_{\gamma_0}$, $(1-g^*)/g^* = kw$, and $g^* = 1/(1+kw)$. With Z and θ_Z as in Theorem 2, define

$$\begin{aligned} &D = -\mathbb{E}_Q\big[w(X)\,\psi(X)\,(Z-\theta_Z)\big], \qquad C = \mathbb{E}_Q\big[w(X)(Z-\theta_Z)(1-g^*(X))\psi(X)\big], \\ &H = \mathbb{E}_Q\big[g^*(1-g^*)\psi\psi^\top\big] + k\,\mathbb{E}_P\big[g^*(1-g^*)\psi\psi^\top\big], \\ &\Sigma = \mathrm{Var}_Q\big((1-g^*)\psi\big) + k\,\mathrm{Var}_P\big(g^*\psi\big), \end{aligned} \tag{10}$$

and

$$\sigma^2_{\text{weight}} = 2\,D^\top H^{-1} C + D^\top H^{-1}\Sigma H^{-1} D. \tag{11}$$

Because ψ contains an intercept and $\mathbb{E}_Q[w(Z-\theta_Z)] = 0$, the intercept component of D is zero: self-normalization removes the class-prior constant. σ^2_{weight} can be negative.

Proof. *Step 1 (estimated λ).* $\hat{\mu}_{\text{adapt}}$ is affine in λ with slope $\bar{E}^u - \hat{R}_E$. By Steps 2–5 applied to $h = \hat{E}$, this slope is $O_p(n^{-1/2})$, and it converges to $\mathbb{E}_P[\hat{E}] - \mathbb{E}_P[\hat{E}] = 0$. Hence replacing $\hat{\lambda}$ by λ^* changes $\hat{\mu}_{\text{adapt}}$ by $(\hat{\lambda} - \lambda^*)\,O_p(n^{-1/2}) = o_p(n^{-1/2})$. We therefore fix $\lambda = \lambda^*$ and write $Z_i = \phi_i - \lambda^* \hat{E}_i$.

Step 2 (self-normalization). Since $\mu^P_m = \lambda^* \mathbb{E}_P[\hat{E}] + \theta_Z$,

$$\hat{\mu}_{\text{adapt}} - \mu^P_m = \lambda^*\big(\bar{E}^u - \mathbb{E}_P[\hat{E}]\big) + \frac{1}{\bar{w}} \cdot \frac{1}{n}\sum_{i=1}^n \hat{w}_i\,(Z_i - \theta_Z), \tag{12}$$

with $\bar{w} \xrightarrow{p} 1$ (proof of Theorem 1). Since the last average is $O_p(n^{-1/2})$ (Steps 3–5), the factor $1/\bar{w}$ can be dropped at cost $o_p(n^{-1/2})$.

Step 3 (weight expansion). Since $\nabla_\gamma w_\gamma = -\psi\, w_\gamma$ and ψ, Z are bounded, a second-order Taylor expansion around γ_0 and the law of large numbers give

$$\frac{1}{n}\sum_{i=1}^{n}\big(\hat{w}_i - w_i\big)(Z_i - \theta_Z) = D^\top(\hat{\gamma} - \gamma_0) + O_p\big(\|\hat{\gamma} - \gamma_0\|^2\big) + o_p\big(\|\hat{\gamma} - \gamma_0\|\big). \tag{13}$$

Step 4 (logistic MLE under a fixed labeled/unlabeled design). The score equation is $\sum_{i\le n}(1 - g_{\hat{\gamma}}(X_i))\psi(X_i) - \sum_{\ell\le m} g_{\hat{\gamma}}(X^u_\ell)\psi(X^u_\ell) = 0$. Because $1-g^* = kwg^*$, $\mathbb{E}_Q[(1-g^*)\psi] = k\,\mathbb{E}_P[g^*\psi]$, so the population score vanishes at γ_0 although the group sizes are fixed. A standard expansion then gives

$$\sqrt{n}\,(\hat{\gamma} - \gamma_0) = H^{-1}\Big[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} a_i - \frac{1}{\sqrt{n}}\sum_{\ell=1}^{kn} b_\ell\Big] + o_p(1), \tag{14}$$

where $a_i = (1-g^*(X_i))\psi(X_i) - \mathbb{E}_Q[(1-g^*)\psi]$ and $b_\ell = g^*(X^u_\ell)\psi(X^u_\ell) - \mathbb{E}_P[g^*\psi]$. In particular $\hat{\gamma} - \gamma_0 = O_p(n^{-1/2})$, so the remainders in Step 3 are $o_p(n^{-1/2})$.

Step 5 (CLT). Combining Steps 2–4,

$$\begin{aligned}\sqrt{n}\big(\hat{\mu}_{\mathrm{adapt}} - \mu_m^P\big) = \lambda^*\sqrt{n}\big(\bar{E}^u - \mathbb{E}_P[\hat{E}]\big) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big[w_i(Z_i - \theta_Z) + D^\top H^{-1} a_i\big] \\ - D^\top H^{-1}\frac{1}{\sqrt{n}}\sum_{\ell=1}^{kn} b_\ell + o_p(1).\end{aligned} \tag{15}$$

The three sums involve independent samples (by (A5)) and have mean zero, because $\mathbb{E}_Q[w(Z - \theta_Z)] = \mathbb{E}_P[Z] - \theta_Z = 0$ by Eq. (7). Their asymptotic variances are $(\lambda^*)^2 r\,\mathrm{Var}_P(\hat{E})$; $\mathbb{E}_Q[w^2(Z - \theta_Z)^2] + 2D^\top H^{-1}C + D^\top H^{-1}\mathrm{Var}_Q((1-g^*)\psi)H^{-1}D$; and $k\,D^\top H^{-1}\mathrm{Var}_P(g^*\psi)H^{-1}D$. The central limit theorem and Slutsky's lemma give the stated limit, with variance $\sigma^2_{\mathrm{known}}(\lambda^*) + \sigma^2_{\mathrm{weight}}$. □

A.3 Proof of Proposition 1 (ESS Bound)

Proof. With known weights and fixed λ, Steps 2 and 5 above (with $\hat{\gamma} = \gamma_0$) show that the asymptotic variance is $\sigma^2_{\mathrm{known}}(\lambda) = \mathbb{E}_Q[w^2(Z - \theta_Z)^2] + \lambda^2 r\,\mathrm{Var}_P(\hat{E})$, where $Z = \phi - \lambda\hat{E}$. By (A3), $|Z| \le (1 + |\lambda|)C_\phi$ and $|\theta_Z| \le (1 + |\lambda|)C_\phi$, so $(Z - \theta_Z)^2 \le 4(1 + |\lambda|)^2 C_\phi^2$. Therefore

$$\mathbb{E}_Q\big[w^2(Z - \theta_Z)^2\big] \le 4(1 + |\lambda|)^2 C_\phi^2\,\mathbb{E}_Q[w^2] = 4(1 + |\lambda|)^2 C_\phi^2\,\big(1 + \chi^2(P_X\|Q_X)\big), \tag{16}$$

which is finite by (A1). Also $\mathrm{Var}_P(\hat{E}) \le \mathbb{E}_P[\hat{E}^2] \le C_\phi^2$. Hence $\sigma^2_{\mathrm{known}}(\lambda) \le C_\phi^2\big[4(1 + |\lambda|)^2(1 + \chi^2) + \lambda^2 r\big]$, and dividing σ^2_{cls} by both sides gives the bound. □

B Choice of λ

Exact minimizer for known weights. Let $\mu_\phi = \mathbb{E}_P[\phi]$ and $\mu_E = \mathbb{E}_P[\hat{E}]$. Since $Z - \theta_Z = (\phi - \mu_\phi) - \lambda(\hat{E} - \mu_E)$, the known-weight variance of Theorem 2 is the convex quadratic $\sigma^2_{\mathrm{known}}(\lambda) = \mathbb{E}_Q\big[w^2\{(\phi - \mu_\phi) - \lambda(\hat{E} - \mu_E)\}^2\big] + \lambda^2 r\,\mathrm{Var}_P(\hat{E})$. Setting its derivative to zero gives

$$\lambda^*_{\mathrm{known}} = \frac{\mathbb{E}_Q\big[w^2(\phi - \mu_\phi)(\hat{E} - \mu_E)\big]}{\mathbb{E}_Q\big[w^2(\hat{E} - \mu_E)^2\big] + r\,\mathrm{Var}_P(\hat{E})}, \tag{17}$$

whose plug-in estimate uses the squared normalized weights $\tilde{w}(X_i)^2$.

Implemented rule. Our experiments use the first-power weighted rule of Eq. (5),

$$\lambda^*_w = \frac{\mathrm{Cov}_w(\phi, \hat{E})}{\mathrm{Var}_w(\hat{E}) + \frac{n}{N}\mathrm{Var}(\hat{E}^u)}, \tag{18}$$

where:

$$\bar{\phi}_w = \frac{1}{n}\sum_{i=1}^{n} \tilde{w}(X_i)\,\phi(f_m(X_i), Y_i), \quad \bar{E}_w = \frac{1}{n}\sum_{i=1}^{n} \tilde{w}(X_i)\,\hat{E}_i, \tag{19}$$

$$\mathrm{Cov}_w(\phi, \hat{E}) = \frac{1}{n}\sum_{i=1}^{n} \tilde{w}(X_i)\big[\phi(f_m(X_i), Y_i) - \bar{\phi}_w\big]\big[\hat{E}_i - \bar{E}_w\big], \tag{20}$$

$$\mathrm{Var}_w(\hat{E}) = \frac{1}{n}\sum_{i=1}^{n} \tilde{w}(X_i)(\hat{E}_i - \bar{E}_w)^2, \quad \mathrm{Var}(\hat{E}^u) = \frac{1}{N}\sum_{j=1}^{N}(\hat{E}^u_j - \bar{E}^u)^2. \tag{21}$$

These moments estimate $\mathrm{Cov}_P(\phi, \hat{E})$ and $\mathrm{Var}_P(\hat{E})$, so λ^*_w coincides with $\lambda^*_{\mathrm{known}}$ only when $w \equiv 1$. After clipping to $[0, 1]$, λ^*_w converges to a finite limit, so Theorem 2 applies. The rule is therefore valid, but may be less efficient than $\lambda^*_{\mathrm{known}}$.

C Bootstrap Variance Estimator

For small sample sizes or when the regularity conditions of Theorem 2 are questionable, we propose a weighted bootstrap procedure:

1. Resample labeled data $\{(X^*_i, Y^*_i)\}_{i=1}^n$ with replacement from $\mathcal{D}_L$.
2. Recompute importance weights using the *same fixed* classifier $\hat{g}$ (trained on the original data): $\hat{w}^*(X^*_i) = \frac{1-\hat{g}(X^*_i)}{\hat{g}(X^*_i)} \cdot \frac{n}{m}$.
3. Recompute $\hat{\mu}^*_{\mathrm{adapt}}$ using the bootstrap sample and recomputed weights.
4. Repeat $B = 1000$ times and use empirical quantiles for confidence intervals.

This procedure captures the sampling variability of the labeled data given the fitted classifier, at the cost of increased computational burden ($B\times$ the cost of a single evaluation). Because the classifier is held fixed, it does not capture weight-estimation uncertainty ($\sigma^2_{\mathrm{weight}}$ in Theorem 2); that would require refitting the classifier in each replicate. In our experiments, the bootstrap and plug-in variance estimators yield comparable coverage; we recommend the plug-in estimator for efficiency.

D Additional Experimental Details

D.1 ImageNet Setup Details

Models under evaluation. We evaluate five ResNet models: ResNet-18, ResNet-34, ResNet-50, ResNet-101, and ResNet-152, using pre-trained weights from `torchvision`.

Synthetic annotator. We use ResNet-101 softmax confidence scores as the synthetic annotator. Specifically, $\hat{E}_i = \max_c P_{\text{ResNet-101}}(c \mid X_i)$, the softmax probability of the ResNet-101 top-1 class, and the same score is used for every target model f_m. This is a model-agnostic proxy for the model-specific quantity $\mathbb{E}_{\tilde{Y}\sim\tilde{P}(\cdot|X_i)}[\phi(f_m(X_i), \tilde{Y})]$ in Eq. (1). The estimator remains unbiased for any function of X_i, but a proxy that is less correlated with each model's correctness yields a smaller variance reduction. The same score also defines the shift variable s_i (Section 6.1).

Shift mechanism. Labeled images are sampled with probability proportional to $\exp(\beta \cdot s_i)$, where s_i is the ResNet-101 confidence score for image i (normalized to $[0, 1]$) and $\beta = 1.0$ unless otherwise stated. The unlabeled set consists of all remaining images.

Classifier for weight estimation. The classifier uses features that are independent of the annotator's score to avoid circular feature coupling. When ResNet-50 penultimate-layer features (2048-d) are available from `extract_features_imagenet.py`, those are used; otherwise, per-image softmax entropy over the five-model confidence vector serves as a 1-d proxy (correlation with shift variable: $r \approx 0.51$, vs. $r = 1.00$ if the annotator score itself were used as a feature). Calibration uses 3-fold Platt scaling via `CalibratedClassifierCV(cv=3, method="sigmoid")`.

Evaluation protocol. We run 250 independent trials. In each trial, we randomly sample a labeled set of size n under the biased mechanism, and use the remaining images as the unlabeled target set. We report mean coverage, MSE, and ESS across trials.

D.2 ProteinGym Setup Details

Dataset. We use the SPG1 (Streptococcal Protein G B1 domain) DMS dataset from ProteinGym (Notin et al., 2023), containing 536,962 single and multi-site mutants with experimentally measured fitness scores.

Models under evaluation. Seven foundation models are evaluated: CARP (Yang et al., 2024), ESM-1b (Rives et al., 2021), ESM-1v (Meier et al., 2021), ESM-2 (Lin et al., 2023), ProGen2 (Nijkamp et al., 2023), RITA (Hesslow et al., 2022), and UniRep (Alley et al., 2019). Each model's predictions are linearly calibrated against true DMS scores using a held-out calibration set of 10,000 variants.

Synthetic annotator. VESPA (Variant Effect Score Prediction with Alignment) is used as the synthetic annotator (Marquet et al., 2022). It achieves Spearman ρ of 0.48 on SPG1, representing a moderately informative annotator.

Shift mechanism. Labeling probability is proportional to $\exp(\beta \cdot Y_{\text{norm}})$ with $\beta = 1.0$, where Y_{norm} is the normalized DMS fitness score. This produces a maximum-to-minimum weight ratio of approximately 37.9, representing a strong but realistic shift.

Classifier for weight estimation. When ESM-2 embeddings (640-d) are available from `extract_features_proteingym.py`, those are used. Otherwise, raw (uncalibrated) model prediction scores are used as features, with the VESPA annotator column excluded to avoid circular coupling. Calibration uses 3-fold Platt scaling.

D.3 Full Ablation Results

Classifier capacity (Table 3).

Table 3: Coverage at 90% nominal level for different classifier architectures on ImageNet ($\beta = 1.0$, exponential shift).

Classifier	$n = 50$	$n = 100$	$n = 200$	$n = 300$	$n = 400$	$n = 500$
Logistic Regression	0.851	0.858	0.865	0.867	0.820	0.815
MLP (64×64)	0.851	0.858	0.866	0.867	0.820	0.794
Random Forest	0.851	0.858	0.834	0.867	0.820	0.790

Weight clipping sensitivity (Table 4).

Table 4: Coverage and $\text{ESS}_{\text{prop1}}$ at $n = 200$ for different clipping thresholds ε on ImageNet (exponential shift, $\beta = 1.0$).

ε	Coverage	ESS
0.001	0.869	1.50
0.010	0.866	1.50
0.050	0.868	1.50
0.100	0.865	1.50
0.200	0.838	1.46

Shift severity (Table 5).

Table 5: Coverage at $n = 200$ for varying shift severity β on ImageNet (exponential shift). PPI++ and Oracle PPI++ were evaluated only at $\beta = 1.0$ (— indicates not evaluated). Nominal level: 90%.

β	Classical	PPI++	Oracle PPI++	ADAEVAL
0.0	0.909	—	—	0.905
0.5	0.749	—	—	0.888
1.0	0.224	0.838	0.893	0.862
2.0	0.067	—	—	0.807
3.0	0.006	—	—	0.692