SELECTIVE PREDICTION INTERVALS: JOINTLY LEARNING TO PREDICT, BOUND, AND ABSTAIN

Anonymous authors
Paper under double-blind review

ABSTRACT

Many high-stakes regression decisions ask three things of a model at once: a point estimate to act on, an interval that bounds the risk of acting, and the option to abstain when neither can be trusted. Deep learning provides these in isolation: prediction-interval methods return an interval, and at best a value estimate inside it, for *every* input, while selective-prediction methods learn when to abstain, but only for point predictions and with no statement of how far the truth may lie from the estimate. We introduce SelPI, which trains interval, value and abstention heads jointly under a single loss with an explicit acceptance-rate budget. A natural objection is that interval width already tells us when to abstain. We show that it does not: across 18 tabular benchmarks, width predicts large point errors well (AUROC 0.72) but is *anti*-correlated with interval misses (AUROC 0.38), so abstaining on wide intervals discards precisely the inputs the interval got right. SelPI therefore selects with two gates, one trained on the interval loss and one on the value loss, coupled by a Fréchet penalty that ties their separate budgets to the joint acceptance rate they deliver. With the base predictor held fixed, the learned gate beats width thresholding on a proper interval score on all 18 tabular datasets, closing a median 87% of the gap to an oracle selector, and it also beats selection by deep-ensemble disagreement on every dataset. The advantage grows with the heteroscedasticity of the task, and we report the settings in which it disappears.

1 INTRODUCTION

Consider a pediatric radiologist using a model to estimate skeletal age from a hand radiograph. A bare point estimate is not enough: the clinical question is whether bone age is *consistent* with chronological age, which requires a plausible range. A range alone is not enough either, because the report must state a value. And for some radiographs—unusual poses, occlusions, rare conditions—the honest output is a referral for manual reading. The same triad appears wherever a regression model feeds a decision: these are components of a single decision, and a model that provides only some of them pushes the missing one onto the user.

Two lines of research supply these components separately. *Prediction-interval* (PI) methods learn a lower and an upper bound around the target (Khosravi et al., 2010; Pearce et al., 2018); IPIV (Simhayev et al., 2022) additionally learns *where inside the interval* the target is likely to be, yielding a point estimate coherent with its own bounds. These methods answer every query. *Selective-prediction* methods, exemplified by SelectiveNet (Geifman & El-Yaniv, 2019), learn a gate that abstains on a budgeted fraction of inputs so that the risk on the accepted remainder falls; they return only a point estimate. Neither family delivers the full triad: the accepted set of a selective point predictor carries no statement of risk, and the intervals of a non-selective predictor are only as trustworthy as the model's worst inputs.

The obvious way to close the gap is post hoc: train an interval model and abstain when the interval is wide. Understanding why this baseline falls short is the core of this paper. Width is an *aleatoric* signal, largely blind to *epistemic* failure—sparsely covered regions, distribution shift, intervals that are confidently *wrong*. And when the point estimate lies inside the interval by construction, point error is mechanically bounded by width on every input the interval captures (§3). The failures that matter are therefore the two that width cannot bound: the interval *misses* the target, or it captures the target but places the point estimate at the wrong end of a wide interval.

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.

We propose SelPI (*Sel*ective *P*rediction *I*ntervals), which trains IPIV's interval and value heads jointly with SelectiveNet-style gates under a single selective interval loss with an explicit acceptance budget. Selection is split across two gates—g_{PI}, trained on the selective interval loss and therefore driven by miscoverage, and g_{val}, trained on the selective value loss and therefore driven by value-position error—which must *both* admit an input for it to be answered. Each gate has its own budget, and a Fréchet penalty ties the two budgets to the joint acceptance rate they actually deliver.

Because no existing method produces the same output, the evaluation straddles two fields, and we organize it as three questions (§5, Table 1): does learning to abstain damage the interval model (the PI field's question); is the result a competitive selective regressor (the selective-prediction field's question); and, holding the predictor fixed, does a learned gate select better than the predictor's own uncertainty (the new question). The answer to the third is a clear yes; the answers to the first two are "mostly", and we report the exceptions. Our contributions are:

1. SelPI, the first method to jointly learn prediction intervals, an in-interval value estimate, and a budgeted abstention rule. We identify the structural coupling between an in-interval value estimate and interval width, and derive from it a *dual-gate* selector whose heads specialize to the two failure modes width cannot detect (§4).
2. An evaluation protocol for selective interval prediction: selective risk curves for a proper interval score, a same-predictor comparison that isolates the selector from the base model, and predictor-independent gate diagnostics (§5).
3. A characterization of *when* learned selection helps: the gain over width thresholding grows with residual heteroscedasticity and vanishes on near-deterministic problems, and each gate's rejections are enriched for the failure mode it was trained on (§7).

2 Related Work

Selective prediction. Learning to abstain dates to Chow's error–reject trade-off (Chow, 1970). El-Yaniv & Wiener (2010) introduced the risk–coverage curve, and Cortes et al. (2016) gave the problem a learning-theoretic treatment. Thresholding the softmax response is a strong post-hoc selector for deep classifiers (Geifman & El-Yaniv, 2017). SelectiveNet (Geifman & El-Yaniv, 2019) instead trains the selector jointly with the predictor under an explicit coverage budget, with an auxiliary head that keeps the representation from collapsing onto the accepted subset. Later work recasts abstention as betting (Liu et al., 2019), uses a Gumbel-softmax gate (Salem et al., 2022), adds entropy regularizers (Feng et al., 2022), or calibrates the accepted set (Fisch et al., 2022). The excess-AURC of Geifman et al. (2018) isolates selector quality from base-model quality and underlies our main statistic. Selective *regression* is comparatively unexplored. Apart from a few regression experiments in SelectiveNet and its Gumbel-softmax variant, existing methods select post hoc, by thresholding an uncertainty score of a trained model. That score is the conditional variance, which is the optimal rule for squared loss (Zaoui et al., 2020), a blended variance with a risk guarantee (Jiang et al., 2020), model-agnostic bootstrap uncertainty (Pugnana et al., 2023), a conformal confidence score (Sokol et al., 2026), or the CRPS entropy of a distributional forecast (Zaoui & Dombry, 2025). Cheng et al. (2023) study rejection under an explicit cost and Shah et al. (2022) under fairness constraints. All of these select on a single scalar, and none learns an interval jointly with the selector; our width-threshold and ensemble-disagreement baselines are this family applied to an interval model.

Prediction intervals for deep regression. Distribution-free interval learning began with LUBE (Khosravi et al., 2010), whose coverage–width objective QD (Pearce et al., 2018) made differentiable. QD+ (Salem et al., 2020) added loss terms for a point prediction and a split-normal mixture for aggregation. IPIV (Simhayev et al., 2022) obtains the point prediction from a value head instead, yielding a coherent (point, interval) pair without extra loss terms. Quantile regression (Koenker & Bassett Jr, 1978; Tagasovska & Lopez-Paz, 2019; Chung et al., 2021) and heteroscedastic likelihoods (Nix & Weigend, 1994; Kendall & Gal, 2017) are alternative routes to intervals. Epistemic uncertainty is usually estimated with MC dropout (Gal & Ghahramani, 2016) or deep ensembles (Lakshminarayanan et al., 2017), and Hüllermeier & Waegeman (2021) review the aleatoric/epistemic distinction that organizes our analysis. Conformal prediction (Vovk et al., 2005; Romano et al., 2019; Angelopoulos & Bates, 2023) wraps any of these with a marginal finite-sample guarantee, and recent work calibrates prediction *sets* on a post-hoc selected subset (Xu et al., 2025;

Kumar et al., 2025). All of these either answer every query or take the selector as given. We learn the selector jointly with the intervals (extended discussion in Appendix A).

3 BACKGROUND

Setup and terminology. We observe $(x, y) \in \mathcal{X} \times \mathbb{R}$ drawn i.i.d. from an unknown P. A *prediction interval* at nominal miscoverage α is a pair $L, U : \mathcal{X} \to \mathbb{R}$ intended to satisfy $\Pr[L(x) \le y \le U(x)] \ge 1 - \alpha$; its empirical quality is summarized by the coverage probability $\mathrm{PICP} = \frac{1}{n}\sum_i k_i$, with $k_i = \mathbb{1}[L(x_i) \le y_i \le U(x_i)]$ the *capture* indicator, and the mean width $\mathrm{MPIW} = \frac{1}{n}\sum_i (U(x_i) - L(x_i))$. A *selective predictor* is a pair (f, g) with gate $g : \mathcal{X} \to [0, 1]$ that answers $f(x)$ when $g(x) \ge \tau$ and abstains otherwise. For a loss ℓ, its *selective risk* (El-Yaniv & Wiener, 2010) is the average loss over the accepted subset,

$$R_\ell(f, g; \tau) \;=\; \frac{\sum_i \ell(f(x_i), y_i)\, \mathbb{1}[g(x_i) \ge \tau]}{\sum_i \mathbb{1}[g(x_i) \ge \tau]}. \tag{1}$$

"Coverage" means accepted inputs in the selective literature and captured targets in the PI literature, so we never use the word unqualified: the former is the *acceptance rate* ϕ, the latter PICP. τ_ϕ is the threshold that admits a fraction ϕ of inputs, and $R_\ell(\phi)$ abbreviates $R_\ell(f, g; \tau_\phi)$.

SelectiveNet. SelectiveNet (Geifman & El-Yaniv, 2019) attaches a predictor f, a soft gate g and an auxiliary predictor h to a shared backbone, and minimizes a gate-weighted risk under a soft acceptance constraint,

$$\mathcal{L}_{(f,g)} = \frac{\frac{1}{n}\sum_i \ell(f(x_i), y_i)\, g(x_i)}{\hat{\phi}(g)} + \lambda\, \Psi\big(\phi_{\mathrm{t}} - \hat{\phi}(g)\big), \qquad \hat{\phi}(g) = \tfrac{1}{n}\textstyle\sum_i g(x_i), \tag{2}$$

with ϕ_{t} the target acceptance rate and $\Psi(a) = \max(0, a)^2$. The auxiliary head is trained on *all* inputs with the plain loss $\mathcal{L}_h$ and discarded at test time; the objective $\gamma\, \mathcal{L}_{(f,g)} + (1 - \gamma)\, \mathcal{L}_h$ keeps the backbone from overfitting the accepted subset. One model is trained per ϕ_{t}.

IPIV. IPIV (Simhayev et al., 2022) outputs bounds $U(x), L(x)$ and a mixing coefficient $v(x) \in (0, 1)$, which we call the *value head*,[1] and predicts

$$\hat{y}(x) \;=\; v(x)\, U(x) + \big(1 - v(x)\big)\, L(x). \tag{3}$$

The bounds are trained with the quality-driven objective of QD (Pearce et al., 2018): minimize the width of the intervals that capture their target, $\mathrm{MPIW}_{\mathrm{capt}} = \sum_i k_i (U_i - L_i) / \sum_i k_i$, under a penalty that activates when coverage falls below nominal,

$$\mathcal{L}_{\mathrm{PI}} = \mathrm{MPIW}_{\mathrm{capt}} + \sqrt{n}\, \lambda\, \Psi\big((1 - \alpha) - \widetilde{\mathrm{PICP}}\big), \tag{4}$$

where $\widetilde{\mathrm{PICP}}$ replaces the hard capture indicator with the sigmoid relaxation $\tilde{k}_i = \sigma\big(s(U_i - y_i)\big)\, \sigma\big(s(y_i - L_i)\big)$, and n is the batch size (Appendix B). The value head is trained with a regression loss on equation 3, $\mathcal{L}_v$ (squared error here), on *all* inputs—captured or not—and the total loss is $\beta\, \mathcal{L}_{\mathrm{PI}} + (1 - \beta)\, \mathcal{L}_v$. IPIV motivates this design in the same way SelectiveNet motivates its auxiliary head: a loss that only sees the captured subset overfits to it.

A structural coupling. Equation 3 implies $\hat{y}(x) \in [L(x), U(x)]$, so on any captured input $|\hat{y}_i - y_i| \le U_i - L_i$: *point error is bounded by interval width wherever the interval is correct*. Two consequences shape the rest of the paper. First, abstaining on wide intervals reduces selective point error automatically, so a width threshold is an unusually strong baseline on point-error curves. Second, a learned gate can genuinely separate from width only on the events width does not bound: a *coverage failure* ($y_i \notin [L_i, U_i]$) and a *value failure* (y_i captured, but v_i far from the target's true relative position $v_i^\star = (y_i - L_i)/(U_i - L_i)$). Section 4 builds the selector around exactly these two events.

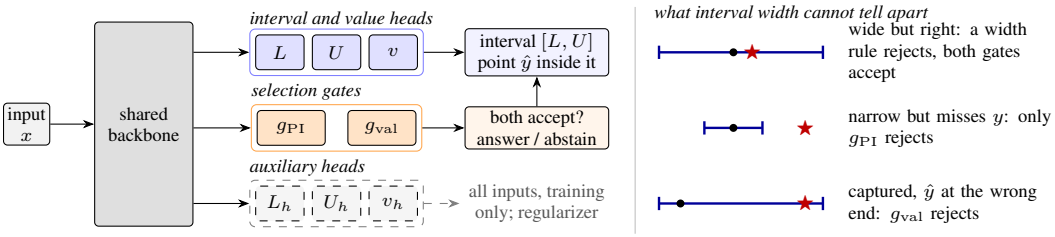


Figure 1: **SelPI in one picture.** *Left:* one backbone feeds the interval and value heads, two selection gates, and auxiliary heads that are trained on all inputs and dropped at inference; a prediction is released only if *both* gates accept. *Right:* the cases a width threshold confuses (bar: predicted interval; dot: point estimate $\hat{y}$; star: true y).

4 SELPI: SELECTIVE PREDICTION INTERVALS

4.1 ARCHITECTURE

SelPI attaches three groups of outputs to a shared backbone (Figure 1): the IPIV triple $\big(L(x), U(x), v(x)\big)$ with $\hat{y}(x)$ as in equation 3; two gates $g_{\text{PI}}(x), g_{\text{val}}(x) \in [0, 1]$; and an auxiliary triple $\big(L_h(x), U_h(x), v_h(x)\big)$ that is trained on all inputs and discarded at inference. The backbone and the (L, U, v) heads are IPIV's for the task, the auxiliary triple duplicates the heads, and each gate is SelectiveNet's selection head (Appendix D). The auxiliary triple plays the role of SelectiveNet's auxiliary head (§3): it is trained on every input, so the shared backbone keeps learning from the inputs the gates reject rather than overfitting the accepted subset. Its loss is given in §4.2 and its contribution is measured in §7. SelPI therefore adds only the gates and the auxiliary triple to an IPIV model, which is what makes the full-acceptance comparison in §6 a fair test of what the additions cost.

4.2 SELECTIVE INTERVAL AND VALUE LOSSES

We instantiate SelectiveNet's gate-weighted risk equation 2 not with a point loss but with IPIV's two objectives, so that both the interval and the value estimate are optimized on the inputs the model will actually answer. For a gate g with $g_i = g(x_i)$, define the gate-weighted capture rate and captured width,

$$\text{PICP}_g = \frac{\sum_i g_i \, \tilde{k}_i}{S_g}, \qquad \text{MPIW}_g = \frac{\sum_i g_i \, k_i \, (U_i - L_i)}{\sum_i g_i \, k_i}, \qquad S_g = \textstyle\sum_i g_i, \qquad \hat{\phi}(g) = S_g / n, \tag{5}$$

where S_g is the effective number of selected samples. The *selective interval loss* carries both constraints—interval coverage and acceptance rate—and the *selective value loss* is the gate-weighted squared error:

$$\mathcal{L}^{\text{sel}}_{\text{PI}}(g) = \text{MPIW}_g + \sqrt{S_g}\, \lambda_{\text{PI}}\, \Psi\big((1-\alpha) - \text{PICP}_g\big) + \lambda_{\text{sel}}\, \Psi\big(\phi_{\text{t}} - \hat{\phi}(g)\big), \tag{6}$$

$$\mathcal{L}^{\text{sel}}_{\text{val}}(g) = \textstyle\sum_i g_i \, (\hat{y}_i - y_i)^2 \;/\; \sum_i g_i. \tag{7}$$

Relative to equation 4, every sample is weighted by the gate, and the coverage penalty is scaled by the *selected* count $\sqrt{S_g}$ rather than the batch size, because that is the amount of data backing the PICP_g estimate (Appendix B). The auxiliary triple is trained with the plain IPIV loss on all samples, $\mathcal{L}_{\text{aux}} = \beta_h \, \mathcal{L}^{(h)}_{\text{PI}} + (1 - \beta_h) \, \mathcal{L}^{(h)}_{v}$; a full triple, rather than one auxiliary regressor, keeps interval-relevant as well as point-relevant features alive for the inputs the gates reject.

4.3 TWO FAILURE MODES, TWO GATES

If a single gate weights both losses, it receives gradient from two sources that reward different behavior: the coverage penalty pushes g down where the interval is likely to miss, the value loss where the point estimate is likely to be far from the target. By the coupling of §3 the second signal is bounded by width on captured inputs and the first is not, so one scalar must encode two different

[1] IPIV calls v its "auxiliary head". We reserve *auxiliary* for SelectiveNet's all-input heads to avoid a clash.

Table 1: **Evaluation design.** SelPI has no like-for-like competitor, so each question is asked on the home ground of the field it comes from, with that field's comparator and metric.

	Question (origin)	Comparator	Metric
Q1	Does learning to abstain damage the interval model? (*PI field*)	IPIV, all inputs answered, IPIV's protocol and datasets	PICP, MPIW, RMSE at $\phi = 1$
Q2	Is it a competitive selective regressor? (*selective-prediction field*)	SelectiveNet-regression, its own grid and settings	selective RMSE / MAE vs. ϕ
Q3	Given an interval model, does a learned gate beat the model's own uncertainty? (*new*)	same predictor, ranked by gate / by width $U-L$ / by an oracle; a deep ensemble ranked by member disagreement	selective interval score and RMSE vs. ϕ; selection efficiency

predictors, with the balance set implicitly by λ_{PI} rather than by the decision-maker. SelPI therefore uses two heads, each weighting the loss whose failures it predicts and each with its own budget $\phi_{\mathrm{t}}^{\mathrm{PI}}, \phi_{\mathrm{t}}^{\mathrm{val}}$. An input is accepted only if *both* gates admit it,

$$\text{accept } x \iff g_{\mathrm{PI}}(x) \geq \tau_{\mathrm{PI}} \ \wedge \ g_{\mathrm{val}}(x) \geq \tau_{\mathrm{val}}, \tag{8}$$

with the two thresholds calibrated independently on validation data, each to its own budget. Figure 1 (right) shows the three cases this separates: a wide interval that is correct, which both gates accept and a width rule rejects; a narrow interval that misses, which only g_{PI} rejects; and a captured target with the value estimate at the wrong end, which g_{val} rejects.

Two gates with two budgets leave the quantity that actually matters, the *joint* acceptance rate, uncontrolled. Since $(1 - g_{\mathrm{PI}})(1 - g_{\mathrm{val}}) \geq 0$ we have $g_{\mathrm{PI}} g_{\mathrm{val}} \geq g_{\mathrm{PI}} + g_{\mathrm{val}} - 1$ pointwise, so the joint rate obeys the Fréchet bound

$$\hat{\phi}_{\mathrm{joint}} \ = \ \tfrac{1}{n} \textstyle\sum_i g_{\mathrm{PI}}(x_i)\, g_{\mathrm{val}}(x_i) \ \geq \ \phi_{\mathrm{t}}^{\mathrm{PI}} + \phi_{\mathrm{t}}^{\mathrm{val}} - 1. \tag{9}$$

We make equation 9 a *trained* constraint: a quadratic penalty fires whenever the gates drift so far apart that their joint mass falls below the floor their budgets imply,

$$\mathcal{L}_{\mathrm{fr}} = \lambda_{\mathrm{fr}}\, \Psi\big(\phi_{\mathrm{t}}^{\mathrm{PI}} + \phi_{\mathrm{t}}^{\mathrm{val}} - 1 - \hat{\phi}_{\mathrm{joint}}\big). \tag{10}$$

Without it the two heads are free to reject disjoint sets, and the intersection collapses well below either budget. With $\mathcal{L}_{g_{\mathrm{PI}}} = \beta_{\mathrm{PI}}\, \mathcal{L}_{\mathrm{PI}}^{\mathrm{sel}}(g_{\mathrm{PI}}) + (1 - \beta_{\mathrm{PI}})\, \mathcal{L}_{\mathrm{fr}}$, the full objective mixes the two selective losses as IPIV mixes its losses, and the result with the auxiliary loss as SelectiveNet does:

$$\mathcal{L}_{2\mathrm{G}} = \gamma \Big[\beta\, \mathcal{L}_{g_{\mathrm{PI}}} + (1 - \beta)\, \mathcal{L}_{\mathrm{val}}^{\mathrm{sel}}(g_{\mathrm{val}})\Big] + (1 - \gamma)\, \mathcal{L}_{\mathrm{aux}}. \tag{11}$$

We fix $\gamma = \beta = \beta_h = 0.5$, the parents' defaults; λ_{fr} and β_{PI}, which have no counterpart in either parent, are tuned per dataset. Tying the gates ($g_{\mathrm{PI}} \equiv g_{\mathrm{val}} \equiv g$, $\lambda_{\mathrm{fr}} = 0$) gives the *single-gate* variant $\mathrm{SelPI}_{1\mathrm{G}}$, which we keep as an ablation; the full model is $\mathrm{SelPI}_{2\mathrm{G}}$. The budget pair (e.g. 0.85 each guarantees a joint rate of at least 0.70) also exposes a *frontier* that trades selective interval quality against selective point accuracy at a fixed joint rate (Appendix I), which is why we do not tie the two budgets to one joint target by formula.

4.4 Training and inference

All heads are trained end to end with Adam; depth, epochs, λ_{PI} and initialization are inherited unchanged from IPIV's published per-dataset configuration, except the learning rate, which we cap at 5×10^{-3} because IPIV's larger values destabilize the multi-term objective. All penalties are active from the first epoch: a warm start that delays the coverage penalty lets the two bounds cross and never recover (Appendix B). One model is trained per target acceptance rate; at inference the thresholds are set on a validation split to accept the budgeted fraction of inputs, and every baseline's threshold is calibrated on the same split. The model returns $\big(\hat{y}(x), [L(x), U(x)]\big)$ on accepted inputs, from a single forward pass, and abstains otherwise.

5 Evaluation Design

Rationale. SelPI produces an output—interval, value and abstention decision—that no existing method produces, so no single comparison is like-for-like. PI methods answer every query and

are scored on the $(\mathrm{PICP}, \mathrm{MPIW})$ pair; selective methods abstain and are scored on risk–coverage curves of a point loss. Neither convention transfers on its own. Under selection the $(\mathrm{PICP}, \mathrm{MPIW})$ pair is no longer a fair summary, because a selector can buy narrow accepted intervals by keeping the misses, or perfect coverage by keeping only easy inputs. A point-loss risk–coverage curve ignores the interval altogether and mechanically favors any width-based selector (§3). We therefore ask three questions, each on the home ground of the field it comes from (Table 1). Q1 and Q2 are the checks each parent field would make first, under that field's protocol. Q3 is the contribution, and it needs two instruments neither field supplies: a *same-predictor* comparison, which ranks the test points of one trained model by the gate, by interval width and by an oracle, so that differences are attributable to the selector and not to the base model; and a *proper* interval score, which charges for a miss rather than only for width and yields one number per (dataset, ϕ) cell.

Datasets. We use 20 regression tasks (Appendix C). *(A) The UCI PI suite*, the ten datasets of QD and IPIV under the protocol of Hernández-Lobato & Adams (2015), and *(B) IPIV's two image tasks*, RSNA bone age (Halabi et al., 2019) and IMDB-WIKI apparent age (Rothe et al., 2018), are the PI field's home ground (Q1). Several UCI sets are too small to evaluate selection (at $\phi = 0.3$ Yacht accepts 9 test points per split), so we add *(C) eight larger OpenML regression tasks* (Fischer et al., 2023; Grinsztajn et al., 2022) with 20k–582k samples, for statistical power. Bone age has a single 200-image test split on which validation-calibrated thresholds transfer poorly, so we treat it as illustrative; IMDB-WIKI (5-fold CV, $\sim$172k images) is the reliable image benchmark.

Baselines. All baselines share splits and seeds, and every threshold is calibrated on validation data.

- *IPIV + width threshold*, the post-hoc rule the central claim is stated against. We evaluate it in two forms: *same-predictor* (ranking SelPI's own intervals by width; used for Q3 and for selection efficiency) and *cross-pipeline* (a separately trained IPIV ranked by its own width; Appendix G).
- *Deep-ensemble disagreement*, the strongest post-hoc rival and the textbook epistemic signal: $M{=}5$ IPIV models per split, combined with the aggregation IPIV's own paper reports (adapted from Lakshminarayanan et al., 2017) and ranked by the variance of the members' point predictions (*point disagreement*) or the summed spread of their bounds (*bound disagreement*), each against the ensemble's own width threshold and oracle. The comparison is cross-pipeline: one SelPI model against a five-model ensemble, at a fifth of the training and inference cost.
- *SelectiveNet-regression*, learned selection on a point regressor; it produces no interval and appears only in point-error comparisons (Q2).
- *Reference points*: IPIV at full acceptance, and an *oracle* that ranks by realized loss and upper-bounds any selector on the same predictor. The non-selective methods of the PI literature (QD, QD+, deep ensembles) are compared at full acceptance in Appendix N.

Two asymmetries should be stated: $\mathrm{SelPI}_{\mathrm{2G}}$'s budgets are tuned by search whereas SelectiveNet runs at its published settings, so we report the *median* search trial, not the best; and SelectiveNet uses its own backbone (Appendix D).

Protocol and metrics. One SelPI model is trained per target acceptance rate on the grid $\phi \in \{0.30, 0.40, \ldots, 0.70, 0.75, \ldots, 0.95\}$, which contains SelectiveNet's regression grid, with $1 - \alpha = 0.95$ throughout; the post-hoc selectors are swept. Tabular results average 5 random 90/10 splits (the provider's split for Year-MSD). Every metric is a selective risk equation 1 as a function of ϕ. The headline is the **selective interval score**, the accepted-set average of the Winkler score (Winkler, 1972),

$$\mathrm{IS}_\alpha(L, U; y) \;=\; (U - L) \;+\; \tfrac{2}{\alpha}\,(L - y)\,\mathbb{1}[y < L] \;+\; \tfrac{2}{\alpha}\,(y - U)\,\mathbb{1}[y > U], \tag{12}$$

a strictly proper scoring rule for interval forecasts (Gneiting & Raftery, 2007). We also report its components (selective PICP, MPIW), selective RMSE/MAE, and the *value-position error* $|v_i - v_i^\star|$ on captured inputs, which g_{val} should select on. Curves are summarized by **AURC**, the normalized area under the selective-risk curve over $\phi \in [0.3, 1]$, and by **selection efficiency** $\mathrm{SE} = (\mathrm{AURC}_{\mathrm{width}} - \mathrm{AURC}_{\mathrm{gate}})/(\mathrm{AURC}_{\mathrm{width}} - \mathrm{AURC}_{\mathrm{oracle}})$, the fraction of the headroom between the same-predictor width threshold and the oracle that the gate captures; SE is the excess-AURC of Geifman et al. (2018) made unit-free so that it can be aggregated across datasets. Gate diagnostics (AUROC for each failure mode, Spearman ρ with width) do not depend on the base predictor. The unit of statistical analysis is the per-dataset AURC, compared with a Wilcoxon signed-rank test (Demšar, 2006) and always accompanied by win/loss counts (Appendices E and H).

Table 2: **Per-dataset results**, 18 tabular datasets, $1-\alpha = 0.95$, SelPI$_{2G}$ median search trial throughout. *Left (Q2):* RMSE of IPIV answering every query; RMSE reduction of SelPI$_{2G}$ at two realized acceptance levels; SelectiveNet-regression vs. SelPI$_{2G}$ at $\phi = 0.7$ ($\Delta > 0$ favors SelPI). *Right (Q3):* learned gate vs. width threshold on the *same* predictor at the realized acceptance nearest 0.7; lower IS$_\alpha$ is better.

	IPIV	SelPI$_{2G}$ RMSE cut vs. IPIV		sel. RMSE at $\phi = 0.7$			SelPI$_{2G}$ sel. PICP at $\phi\approx0.7$		SelPI$_{2G}$ sel. IS$_\alpha$ at $\phi\approx0.7$	
Dataset	ϕ=1	$\phi\approx0.55$	$\phi\approx0.75$	Sel.Net	SelPI$_{2G}$	Δ	gate	width	gate	width
yacht	1.04	63.3%	56.1%	1.62	**0.35**	+78.3%	**0.975**	0.893	**0.10**	0.45
boston	2.64	10.5%	7.5%	2.42	**2.39**	+1.3%	**0.934**	0.869	**1.72**	2.16
energy	1.63	59.4%	50.5%	1.17	**0.82**	+30.2%	**0.945**	0.855	**0.45**	1.41
concrete	6.17	33.5%	22.2%	4.66	**4.59**	+1.5%	**0.850**	0.809	**1.66**	2.03
wine	0.66	16.2%	10.1%	0.56	0.56	0.0%	**0.812**	0.701	**5.39**	9.29
kin8nm	0.08	11.0%	1.5%	**0.07**	0.08	−15.5%	**0.880**	0.786	**1.68**	3.78
power-plant	4.10	7.5%	3.2%	**3.86**	3.92	−1.5%	**0.937**	0.924	**1.14**	1.20
naval	0.0010	15.9%	3.0%	**0.0006**	0.0011	−83.6%	**0.956**	0.812	**0.30**	2.47
protein	4.41	28.3%	14.4%	3.93	**3.56**	+9.4%	**0.947**	0.861	**2.87**	5.19
Year-MSD	9.01	43.1%	28.9%	6.13	**5.96**	+2.8%	**0.932**	0.916	**3.35**	4.38
superconduct	13.05	56.9%	39.5%	8.63	**7.40**	+14.3%	**0.941**	0.887	**1.11**	2.22
sarcos	3.23	51.5%	35.3%	2.40	**1.91**	+20.3%	**0.955**	0.929	**0.46**	0.60
diamonds	563.4	73.3%	56.6%	237.6	**232.5**	+2.2%	**0.973**	0.890	**0.26**	2.28
video_transcoding	2.83	87.5%	80.1%	0.79	**0.50**	+37.4%	**0.982**	0.974	**0.16**	0.23
medical_charges	0.08	30.9%	19.6%	0.07	**0.06**	+10.2%	**0.961**	0.955	**0.60**	0.66
black_friday	0.39	8.1%	4.9%	0.38	**0.37**	+3.5%	**0.950**	0.856	**3.55**	9.25
particulate-matter	0.39	29.7%	19.1%	0.32	**0.30**	+4.1%	**0.943**	0.935	**2.32**	2.51
nyc-taxi-green	0.41	31.7%	21.8%	0.34	**0.31**	+9.6%	**0.958**	0.887	**2.98**	7.20
median [IQR]		31% [16–56]	21% [8–38]			+3.8% [+1.4, +13.3]	**0.946**	0.887	**1.40**	2.25
wins / ties / losses		18/0/0	18/0/0		14 / 2 / 2 for SelPI$_{2G}$			18/0/0		18/0/0

6 RESULTS

Q1: the gates are nearly free; the selective objective is not. At $\phi = 1$ the gates are inactive, so SelPI is an IPIV model trained with the selective objective and carrying extra heads. Across the 18 tabular datasets it deviates from plain IPIV by a median -0.013 in PICP (a small loss of coverage), -6.9% in MPIW (narrower intervals, an improvement) and $+4.4\%$ in RMSE (a loss); per-dataset values are in Table 12. Two effects are mixed here; setting the acceptance target to 1 separates them, because the selective losses then reduce to IPIV's own and only the architecture differs: **the extra heads alone cost a median $+0.6\%$ in RMSE, -0.008 in PICP and -4.2% in MPIW, and improve on IPIV on 7 of 18 datasets.** The rest of the RMSE deviation comes from the selective objective, which optimizes the model for the inputs it will accept; that is the effect the method exists to exploit, and a price only if one trains to select and then declines to. The architecture costs more than 10% in RMSE on four datasets (kin8nm, naval, video_transcoding, black_friday), with no relation to heteroscedasticity ($\rho = -0.22$, $p = 0.39$). The answer to Q1: adding the abstention machinery leaves the interval model essentially intact, with about one point of coverage lost, narrower intervals, and a few percent of RMSE lost mostly to the selective objective rather than to the added heads.

Q2: intervals are usually not paid for in point accuracy. Abstaining pays in absolute terms: at a realized acceptance of ~0.55, SelPI$_{2G}$ cuts RMSE relative to IPIV on 18 of 18 datasets, by a median 31% (Table 2) and by up to 88% where error is concentrated in a minority of inputs (video_transcoding, diamonds); black_friday and naval gain least and are the weak cases of every later analysis. The informative comparison is with SelectiveNet-regression, a method with no interval to maintain. At $\phi = 0.7$ the median SelPI$_{2G}$ trial has the lower selective RMSE on 14 of 18 tabular datasets, ties on two (wine, power-plant) and loses on two (kin8nm, naval), both near-deterministic tasks with little uncertainty for an interval method to characterize. The image tasks are less favorable (Appendix F). On IMDB-WIKI **SelectiveNet wins wherever the two overlap**, by 1.3–1.5 years of MAE (6.29 vs. 7.59 at $\phi \in [0.7, 0.8)$); on bone age the median SelPI$_{2G}$ trial is slightly behind (16.2 vs. 15.9 months at $[0.7, 0.8)$), though the best trial wins every bucket. We attribute the IMDB-WIKI result to headroom: IPIV is already tight there, little error sits in a rejectable minority, and a single-purpose objective extracts what remains more efficiently than our joint one (§7). Part of any point-error advantage is, moreover, the structural coupling of §3, which is why Q3, not Q2, is our evidence that the gates learn something.

Q3: a learned gate beats width thresholding, most clearly on interval quality. Figure 2 shows three datasets; per-dataset values for all 20 are in Figure 9 and Table 15. With the base predictor held fixed, the learned gate attains a lower selective-RMSE AURC than the width threshold on all 20 datasets (Wilcoxon $p = 2 \times 10^{-6}$), with a median selection efficiency of 0.59 (0.65 on the 18 tabular

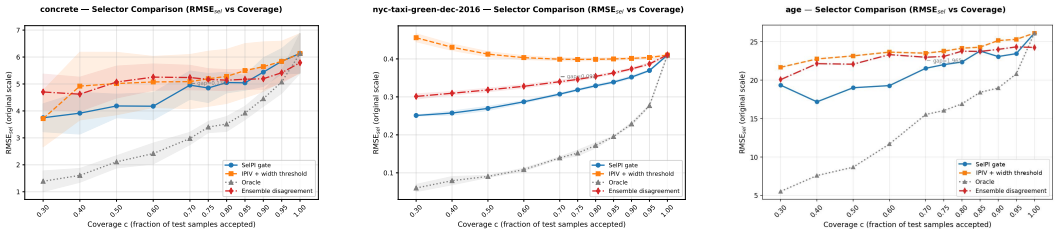


Figure 2: **Q3: selective RMSE vs. acceptance rate on a fixed predictor** (single-gate SelPI$_{1G}$, one model per acceptance rate) for Concrete, NYC Taxi and bone age (illustrative only, §5). Curves: learned gate, width threshold, oracle, and deep-ensemble point disagreement on its own five-model predictor; bands: ± 1 s.d. over runs; lower is better.

Table 3: **What each signal sees, and how well it selects.** AUROC for coverage failure (cov.) and value failure (val.; captured, but error above the captured-set median), and median SE against each selector's *own* predictor's width threshold and oracle (<0: worse than width). The SE on the $g_{\text{PI}}/g_{\text{val}}$ rows is that of the Q3 gate (SelPI$_{1G}$); Pt./Bd. disagr. is the deep-ensemble baseline on its own five-model predictor. Medians over 18 tabular datasets.

	AUROC		selection efficiency	
Signal	cov.↑	val.↑	sel. RMSE	IS$_\alpha$
Width $U - L$	0.378	**0.717**	0 (reference)	
g_{PI}	**0.790**	0.633	**0.59**	**0.87**
g_{val}	0.710	0.647		
Pt. disagr.	0.584	0.606	−0.34	−0.62
Bd. disagr.	0.591	0.581	−0.37	−0.55

datasets, IQR 0.48–0.82). On the interval score the margin is far larger: 20 of 20, median SE **0.87** (IQR 0.74–0.96). The smallest gain is on IMDB-WIKI (SE 0.08: interval score 51.95 against 52.99 for width, oracle 38.20), consistent with its tight, near-homoscedastic predictor (§7). An oracle gap remains everywhere; the gate does not close it, and is closest on the most heteroscedastic datasets.

The gap between the two metrics is informative: a width threshold is a tolerable point-error selector but a poor interval selector. Rejecting wide intervals removes the inputs the model *covered*, so the accepted set retains the misses, and each miss costs $2/\alpha = 40$ times its distance outside the interval. Decomposing IS$_\alpha$ at $\phi \approx 0.7$ confirms the mechanism (right block of Table 2): the gate accepts intervals *wider* than the width threshold's by a median 9.3% and still wins, because it cuts the miss-penalty term by a median 71.0% (miss rate 0.054 against 0.113). Width buys sharpness on the accepted set and pays in misses; the gate makes the opposite trade, and under a proper scoring rule that is the better one.

Q3, continued: the gate also beats ensemble disagreement. Width is not the only post-hoc alternative. The deep-ensemble baseline of §5 ranks a five-model IPIV ensemble by point or bound disagreement across members, the textbook epistemic signal, and Table 3 (right columns) compares it with the gate, each selector measured against its own predictor's width threshold and oracle. Comparing selection efficiency across the two pipelines, the gate wins on every dataset against both signals: 20 of 20 on selective RMSE (Wilcoxon $p = 2 \times 10^{-6}$) and 19 of 19 on the interval score ($p = 4 \times 10^{-6}$). The disagreement selectors mostly fail to beat even their own ensemble's width threshold: on selective RMSE, point disagreement loses to width on 18 of 20 datasets and bound disagreement on 18 of 20; on the interval score, 13 of 19 and 14 of 19 respectively (yacht excluded). The left columns of Table 3 show why: both signals sit well below g_{PI} on coverage failure. Per-dataset values are in Appendix O. Variance across independently trained networks does not substitute for a gate trained jointly with the predictor. (Yacht is excluded from the ensemble's IS$_\alpha$ column, where its SE is undefined; Appendix E.)

Two gates vs. one. At matched acceptance SelPI$_{2G}$ beats SelPI$_{1G}$ in selective RMSE on all 18 tabular datasets while producing *wider* intervals on 16 of them—the opposite of what an "it just picks easy points" explanation predicts—and it reduces the value-position error $|v - v^\star|$ on its accepted set on 18 of 18 (median 10.6%), so g_{val} selects on the quantity it is trained for rather than on width by proxy. The two-budget frontier (Appendix I) shows the mechanism: shifting budget between the

gates at a fixed joint rate trades interval quality against point accuracy, which one scalar gate cannot express. On images the picture is weaker: SelPI$_{2G}$ wins on bone age, but on IMDB-WIKI the single gate is better at three of five comparable acceptance rates.

7 Analysis: What the Gates Learn That Width Does Not

Complementarity and specialization. Treating "did the interval miss?" and "was the value estimate misplaced?" as binary labels, Table 3 asks how well each signal ranks them. Width predicts *value* failure well (0.717), as the coupling of §3 implies, but on *coverage* failure it scores 0.378—below chance, and therefore anti-predictive: wider intervals miss less often, so a reject-if-wide rule systematically discards the inputs whose intervals were correct. g_{PI} reaches 0.790 on that same event. The gates' rank correlation with width stays moderate (per-dataset means from -0.77 to $+0.21$): width is a valid aleatoric signal, but reproducing it would be learning nothing. Splitting rejections by which gate *alone* is responsible confirms the specialization: g_{val}'s unique rejections are enriched for value failure on 17 of 18 datasets, and g_{PI}'s for coverage failure on 13 of the 17 where it rejects anything uniquely (Table 11).

When does learned selection help? We relate each dataset's selection efficiency to three task properties; with 20 datasets this is descriptive, not inferential (Appendix J). Selection efficiency rises with residual heteroscedasticity (Spearman $\rho = 0.54$, $p = 0.013$), as predicted. The deep-ensemble selectors of Table 3 show the opposite pattern (point disagreement $\rho = -0.87$, bound disagreement $\rho = -0.83$, both $p < 10^{-4}$), so the relationship is specific to a gate trained jointly with the predictor. Target kurtosis and full-acceptance miscoverage show no usable relationship (Appendix J). The extremes fit: video_transcoding and diamonds sit near 0.99, while particulate-matter, power-plant and IMDB-WIKI—tight, near-homoscedastic predictors—sit below 0.16, where a width threshold is already close to optimal.

Shift-like subpopulations. Width is aleatoric and shift is epistemic, so shift is where a learned gate should most clearly beat width. Without retraining, we take for each dataset the covariate that best tracks realized error, treat its top quintile as a shifted tail, and ask how much each selector's abstention rate rises there at its $\phi = 0.7$ threshold. The gate's abstention rises by a median $+0.193$ against $+0.046$ for width, and the gate reacts more strongly on 15 of 18 datasets. The deep-ensemble selectors react more than width here too (point disagreement $+0.131$, bound disagreement $+0.165$, on the 15 datasets where the tail is defined for all selectors), but less than the gate: ensembling captures only part of the epistemic signal the gate exploits. The tail is chosen with knowledge of the error, so this is weaker evidence than an injected corruption, but the direction is clear (Appendix K).

Ablations and cost. *Joint training is essential* (Table 13, Appendix L): a gate trained on a frozen IPIV backbone is a median 5.6% worse and loses on 15 of 18 datasets, by up to 54%; the gates shape the representation rather than merely reading it. *The auxiliary triple matters*, though less ($+2.0\%$, worse on 14 of 18). *The $\sqrt{S_g}$ scaling does not pay for itself empirically* ($+0.2\%$, 9 worse and 9 better); we retain it on principle (§4.2). Each gate adds $16h + 65$ parameters for hidden width h (44% on a small tabular model, $\sim 0.2\%$ on the image backbones); training takes a median $1.63\times$ IPIV's wall-clock per model. Inference cost is unchanged: one forward pass (Appendix P).

8 Conclusion

A regression model can learn to predict, bound and abstain jointly. The extra heads cost almost nothing at full acceptance; the resulting selective regressor matches or beats SelectiveNet on most tabular tasks but not on IMDB-WIKI; and on a fixed predictor a learned gate beats width thresholding everywhere on a proper interval score, because width is anti-predictive of the misses that score charges for. It also beats deep-ensemble disagreement on every dataset. Two limitations remain. The accepted set carries no finite-sample coverage guarantee, since abstention breaks the exchangeability that split-conformal calibration relies on; calibrating *conditional on acceptance* would restore one and make SelPI the joint-training counterpart of selective conformal risk control (Xu et al., 2025). And our gates abstain on *hard* inputs, whereas some applications care whether *acting* is expected to lose; retargeting g_{val} to a downstream utility is a natural extension.

AI USE STATEMENT

Large language model (LLM) tools were used solely to assist with the writing of this manuscript, specifically for language polishing, including improving grammar, clarity, phrasing, and overall readability. These tools were not used to generate research ideas, design or conduct experiments, analyze results, or produce scientific content. All AI-assisted edits were reviewed and verified by the authors, who take full responsibility for the content of this paper.

REPRODUCIBILITY STATEMENT

Everything needed to reproduce our results is specified in the appendices. Appendix C lists every dataset with its source, size, target statistics, preprocessing, and split protocol (five random 90/10 splits with fixed seeds for the tabular datasets, the provider's split for Year-MSD, five-fold cross-validation for IMDB-WIKI, and the fixed 200-image test set for bone age). Appendix B states the training objective in full, including the hard and soft capture indicators, the $\sqrt{S_g}$ scaling, and the absence of a warm start. Appendix D gives the architecture and every hyperparameter of SelPI and of each baseline, the values inherited unchanged from IPIV and SelectiveNet, the per-dataset search spaces and tuning budgets, and the rule by which the median search trial is reported. Appendix E defines every metric, the validation-based threshold calibration used for all selectors, and the selection-efficiency normalization; Appendix H describes the statistical tests behind every reported p-value and win/loss count. The deep-ensemble baseline (number of members, aggregation, and ranking signals) is specified in §5 and Appendix O, and compute and hardware in Appendix P. Full per-dataset results for every experiment summarized in the main text are reported in Appendices F–O.

ETHICS STATEMENT

This work uses only publicly available benchmark datasets under their original licenses: the UCI and OpenML regression tasks, the RSNA pediatric bone age challenge data (de-identified radiographs released for research), and IMDB-WIKI (publicly posted celebrity photographs with age labels). No new data involving human subjects were collected, and no attempt was made to identify individuals. Age estimation from faces and radiographs is used here as a benchmark for interval prediction with abstention, not as a deployable system; both datasets are known to be demographically unbalanced, and a selector trained on them may abstain at different rates across groups, which we did not audit. More broadly, the accepted set of SelPI carries no finite-sample coverage guarantee (§8), so an abstention decision should be read as a learned signal of predictive difficulty, not as a certification of safety, and any high-stakes use would require domain-specific validation and human oversight. The authors declare no competing interests.

REFERENCES

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Xin Cheng, Yuzhou Cao, Haobo Wang, Hongxin Wei, Bo An, and Lei Feng. Regression with cost-based rejection. *Advances in Neural Information Processing Systems*, 36:45172–45196, 2023.

Chao Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on information theory*, 16(1):41–46, 1970.

Youngseog Chung, Willie Neiswanger, Ian Char, and Jeff Schneider. Beyond pinball loss: Quantile methods for calibrated uncertainty quantification. *Advances in Neural Information Processing Systems*, 34:10971–10984, 2021.

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *International conference on algorithmic learning theory*, pp. 67–82. Springer, 2016.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010. URL http://jmlr.org/papers/v11/el-yaniv10a.html.

Leo Feng, Mohamed Osama Ahmed, Hossein Hajimirsadeghi, and Amir Abdi. Towards better selective classification. *arXiv preprint arXiv:2206.09034*, 2022.

Adam Fisch, Tommi Jaakkola, and Regina Barzilay. Calibrated selective classification. *arXiv preprint arXiv:2208.12084*, 2022.

Sebastian Felix Fischer, Matthias Feurer, and Bernd Bischl. Openml-ctr23–a curated tabular regression benchmarking suite. In *AutoML Conference 2023 (Workshop)*, 2023.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International conference on machine learning*, pp. 2151–2159. PMLR, 2019.

Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. Bias-reduced uncertainty estimation for deep neural classifiers. *arXiv preprint arXiv:1805.08206*, 2018.

Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35: 507–520, 2022.

Safwan S Halabi, Luciano M Prevedello, Jayashree Kalpathy-Cramer, Artem B Mamonov, Alexander Bilbily, Mark Cicero, Ian Pan, Lucas Araújo Pereira, Rafael Teixeira Sousa, Nitamar Abdala, et al. The rsna pediatric bone age machine learning challenge. *Radiology*, 290(2):498–503, 2019.

José Miguel Hernández-Lobato and Ryan Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International conference on machine learning*, pp. 1861–1869. PMLR, 2015.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

Wenming Jiang, Ying Zhao, and Zehan Wang. Risk-controlled selective prediction for regression deep neural network models. In *2020 international joint conference on neural networks (IJCNN)*, pp. 1–8. IEEE, 2020.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Abbas Khosravi, Saeid Nahavandi, Doug Creighton, and Amir F Atiya. Lower upper bound estimation method for construction of neural network-based prediction intervals. *IEEE transactions on neural networks*, 22(3):337–346, 2010.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pp. 33–50, 1978.

Divake Kumar, Nastaran Darabi, Sina Tayebati, and Amit Ranjan Trivedi. Beyond confidence: Adaptive abstention in dual-threshold conformal prediction for autonomous system perception. *arXiv preprint arXiv:2502.07255*, 2025.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Ziyin Liu, Zhikang Wang, Paul Pu Liang, Russ R Salakhutdinov, Louis-Philippe Morency, and Masahito Ueda. Deep gamblers: Learning to abstain with portfolio theory. *Advances in Neural Information Processing Systems*, 32, 2019.

David A Nix and Andreas S Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 ieee international conference on neural networks (ICNN'94)*, volume 1, pp. 55–60. IEEE, 1994.

Tim Pearce, Alexandra Brintrup, Mohamed Zaki, and Andy Neely. High-quality prediction intervals for deep learning: A distribution-free, ensembled approach. In *International conference on machine learning*, pp. 4075–4084. PMLR, 2018.

Andrea Pugnana, Carlos Mougan, and Dan Saattrup Nielsen. Model agnostic explainable selective regression via uncertainty estimation. *arXiv preprint arXiv:2311.09145*, 2023.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2):144–157, 2018.

Mahmoud Salem, Mohamed Osama Ahmed, Frederick Tung, and Gabriel Oliveira. Gumbel-softmax selective networks. *arXiv preprint arXiv:2211.10564*, 2022.

Tárik S Salem, Helge Langseth, and Heri Ramampiaro. Prediction intervals: Split normal mixture from quality-driven deep ensembles. In *Conference on uncertainty in artificial intelligence*, pp. 1179–1187. PMLR, 2020.

Abhin Shah, Yuheng Bu, Joshua K Lee, Subhro Das, Rameswar Panda, Prasanna Sattigeri, and Gregory W Wornell. Selective regression under fairness criteria. In *International Conference on Machine Learning*, pp. 19598–19615. PMLR, 2022.

Eli Simhayev, Gilad Katz, and Lior Rokach. Integrated prediction intervals and specific value predictions for regression problems using neural networks. *Knowledge-Based Systems*, 247:108685, 2022.

Anna Sokol, Nuno Moniz, and Nitesh V Chawla. Conformalized selective regression. *Discover Data*, 4(1):14, 2026.

Natasa Tagasovska and David Lopez-Paz. Single-model uncertainties for deep learning. *Advances in neural information processing systems*, 32, 2019.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Robert L Winkler. A decision-theoretic approach to interval estimation. *Journal of the American Statistical Association*, 67(337):187–191, 1972.

Yunpeng Xu, Wenge Guo, and Zhi Wei. Selective conformal risk control. *arXiv preprint arXiv:2512.12844*, 2025.

Ahmed Zaoui and Clément Dombry. Distributional regression with reject option. *arXiv preprint arXiv:2503.23782*, 2025.

Ahmed Zaoui, Christophe Denis, and Mohamed Hebiri. Regression with reject option and application to knn. *Advances in Neural Information Processing Systems*, 33:20073–20082, 2020.

A Extended Related Work: Selection with Conformal Guarantees

Selective conformal risk control (Xu et al., 2025) selects post hoc and then calibrates prediction *sets* for classification on the accepted subset; dual-threshold conformal abstention (Kumar et al., 2025) does the same for perception models, and adaptive conformal inference (Gibbs & Candes, 2021) tracks coverage online under drift. Conformalized selective regression (Sokol et al., 2026) goes the other way and uses a conformal score as the confidence function of a post-hoc selective regressor. All of these take either the predictor or the selector as given and add a calibration layer; we learn the selector jointly with the intervals, on regression intervals rather than label sets, and view a conformal wrapper calibrated conditional on acceptance as a natural extension (§8). SelectiveNet's selection machinery is loss-agnostic, but IPIV's heads are regression-specific; a classification analogue would target prediction sets and is left to future work.

B Loss Details

Hard and soft capture in the IPIV loss. In equation 4 the hard indicator k_i has zero gradient almost everywhere, so the coverage term uses the softened capture $\tilde{k}_i = \sigma\big(s\,(U_i - y_i)\big)\,\sigma\big(s\,(y_i - L_i)\big)$ with $s = 160$. The two indicators are *not* interchangeable across the two terms: the soft $\tilde{k}$ enters only the coverage penalty, while $\mathrm{MPIW}_{\mathrm{capt}}$ uses the hard k, so that an interval which misses its target contributes exactly zero to the width objective and is never encouraged to shrink further. The $\sqrt{n}$ factor is the batch size entering as a confidence weight: a larger sample makes the estimate of PICP more trustworthy and a shortfall correspondingly more costly. SelPI preserves this split exactly in equation 5: $\tilde{k}$ in PICP_g, k in MPIW_g.

Why $\sqrt{S_g}$ and not $\sqrt{n}$. When coverage is estimated only on the (softly) accepted subset, the amount of data backing the estimate is $S_g = \sum_i g_i$, not n. Keeping $\sqrt{n}$ over-weights the coverage constraint by a factor $\sqrt{n/S_g}$—$1.8\times$ at $\phi_{\mathrm{t}} = 0.3$, and more early in training, when g sits near 0.5 everywhere. Empirically the choice makes little difference (Table 13); we keep $\sqrt{S_g}$ because it is the consistent generalization of equation 4.

Why there is no warm start. A warm start that holds $\lambda_{\mathrm{PI}} = 0$ for the first $\approx 20\%$ of epochs, letting the gates and the value head settle before interval pressure is applied, seems natural but does not work: with the coverage penalty off nothing stops the two bound heads from crossing, and a non-trivial fraction of runs end the warm-up with $U(x) < L(x)$ and never recover once the penalty is switched on. The coverage penalty is not a constraint on an already-valid interval—it is what makes the interval well ordered in the first place—so we train with all penalties active from the first epoch.

C Dataset Details

Inputs and targets are standardized to zero mean and unit variance; all interval widths and the PICP figures below are therefore in standardized units, while point errors are reported on the original scale. d counts features after one-hot expansion of categoricals for the group-C sets. Naval ships two regression targets and we keep the first, following the convention of the QD and IPIV evaluations. For IMDB-WIKI the two epoch budgets are 90 for SelPI and 29 for IPIV and SelectiveNet-regression; the baselines converge within their shorter budget. Real per-epoch training curves confirm this is measured, not assumed: IPIV and SelectiveNet-regression plateau by epoch $\sim$15–20, while SelPI's harder joint objective — interval, value, selection and auxiliary heads trained together — was still improving at epoch 29, confirmed directly by retraining one fold to 90 epochs, where selective MAE fell from ~ 8.5 to ~ 5.0. The 90-epoch budget is not an inflated allowance chosen to favor SelPI: it matches the original IPIV paper's own IMDB-WIKI training length exactly.

D Hyperparameters and Tuning Budget

Method-level constants, fixed across all 20 datasets. $\gamma = \beta = \beta_h = 0.5$ (the convex-combination weights of equation 11, inherited from SelectiveNet's α and IPIV's β respectively); nominal interval

Table 4: The 20 datasets. n and d are read directly from the data files as loaded by our pipeline. "Test/split" is the held-out count per repetition under the protocol in the last column. Epochs are the per-dataset budgets inherited from IPIV's published configuration for the UCI sets and sized by timing probes for the larger ones. Every tabular dataset uses the same 10-point acceptance grid $\phi \in \{0.30, 0.40, 0.50, 0.60, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95\}$, one model trained per point.

Dataset	n	d	Test/split	Epochs	Reps	Split protocol
(A) UCI PI suite — the standard benchmark of QD and IPIV						
Yacht	308	6	31	2000	5	90/10 random
Boston	506	13	51	300	5	90/10 random
Energy	768	8	77	1200	5	90/10 random
Concrete	1,030	8	103	800	5	90/10 random
Wine	1,599	11	160	1000	5	90/10 random
Kin8nm	8,192	8	819	800	5	90/10 random
Power-plant	9,568	4	957	500	5	90/10 random
Naval	11,934	16	1,193	1000	5	90/10 random
Protein	45,730	9	4,573	600	5	90/10 random
Year-MSD	515,345	90	51,630	300	1	fixed provider split
(C) Larger tabular regression — added for statistical power at low ϕ						
superconduct	21,263	79	2,126	500	5	90/10 random
sarcos	48,933	21	4,894	400	5	90/10 random
diamonds	53,940	26	5,394	400	5	90/10 random
video_transcoding	68,784	24	6,879	350	5	90/10 random
medical_charges	163,065	3	16,307	200	5	90/10 random
black_friday	166,821	23	16,682	200	5	90/10 random
particulate-matter	394,299	26	39,430	120	5	90/10 random
nyc-taxi-green	581,835	9	58,184	100	5	90/10 random
(B) Image regression — IPIV's own two image tasks						
RSNA bone age (VGG-16)	—	—	200	50	1	fixed split
IMDB-WIKI (DenseNet)	∼171,852	—	∼34,370	90 / 29	5	5-fold CV

level $1 - \alpha = 0.95$; softening factor $s = 160$ and acceptance-penalty weight $\lambda_{\text{sel}} = 32$, taken unchanged from IPIV and SelectiveNet; gate hidden width 16 with batch normalization; warm-start fraction 0 (§4.4); weight decay 0, matching IPIV.

Table 5: Per-dataset settings. Everything except the learning rate is inherited unchanged from IPIV's published configuration for that dataset. The learning rate is capped at 5×10^{-3} because IPIV's larger values destabilize the multi-term selective objective (§4.4); where the cap binds, the decay rate is raised to compensate.

Dataset	Hidden	λ_{PI}	σ_{init}	Learning rate	Decay
Boston	50	15.0	0.100	5×10^{-3}	0.990
Concrete	50	15.0	0.200	5×10^{-3}	0.990
Energy	50	17.0	0.350	5×10^{-3}	0.990
Kin8nm	50	12.0	0.436	5×10^{-3}	0.990
Naval	50	4.0	0.100	5×10^{-3}	0.990
Power-plant	50	15.7	0.257	5×10^{-3}	0.990
Wine	50	40.0	0.200	5×10^{-3}	0.990
Yacht	50	6.0	0.200	5×10^{-3}	0.980
Protein	100	40.0	0.200	2×10^{-3}	0.999
Year-MSD and all 8 group-C sets	100	40.0	0.200	2×10^{-3}	0.999

Batch size is 100 for the UCI sets and 256 for Year-MSD, medical_charges, black_friday, particulate-matter and nyc-taxi-green. Optimization is Adam with exponential decay throughout.

Dual-gate search. The two acceptance budgets and the four weights with no counterpart in either parent method are tuned per dataset by Optuna (TPE), minimizing selective RMSE at a nominal joint acceptance rate of 0.7, three seeds per trial. The search space is $\phi_{\text{t}}^{\text{PI}}, \phi_{\text{t}}^{\text{val}} \in [0.5, 0.9]$ uniform; $\lambda_{\text{sel}}^{\text{PI}}, \lambda_{\text{sel}}^{\text{val}} \in [8, 64]$ log-uniform; $\lambda_{\text{fr}} \in [1, 32]$ log-uniform; $\beta_{\text{PI}} \in [0, 1]$ uniform. Because the joint acceptance rate is measured rather than targeted (§4.3), the TPE population clusters in the middle of

the range, so each dataset's main search is supplemented by dedicated low- and high-band batches that constrain the two budgets to $[0.50, 0.58]$ and $[0.82, 0.90]$ (and later $[0.90, 0.97]$) respectively. Trial counts after all batches are 149–287 per tabular dataset (180–287 on the ten UCI sets, 149–165 on the eight larger ones), 177 for bone age and 100 for IMDB-WIKI. Reported trials are filtered to those that are well calibrated ($\mathrm{PICP}_{\mathrm{sel}} \geq 0.85$) and non-degenerate; results are then bucketed by realized joint acceptance rate in 0.1-wide bins, and only buckets with at least 5 trials are reported.

SelectiveNet-regression baseline. Its own canonical configuration, untuned: one hidden layer of 64 ReLU units with batch normalization, selection head of 16 units, learning rate 5×10^{-4}, 800 epochs, 5 runs, over its own published coverage grid. It brings its own backbone rather than sharing IPIV's, which is a real confound on the small nonlinear sets: its full-acceptance RMSE on Yacht is 4.03 against IPIV's 1.04, so a large part of the SelPI margin there is backbone rather than selection.

E Metric Definitions

Write $A_\phi = \{i : x_i \text{ accepted at rate } \phi\}$ for the accepted set, $S = |A_\phi|$, and $C = \{i \in A_\phi : k_i = 1\}$ for its captured subset. Every metric below is the instantiation of equation 1 with the stated per-sample loss.

- **Selective PICP**: $\mathrm{PICP}(\phi) = \frac{1}{S}\sum_{i \in A_\phi} k_i$, reported as the deviation $|\mathrm{PICP}(\phi) - (1-\alpha)|$, since over- and under-coverage are both failures.
- **Selective MPIW**: $\frac{1}{S}\sum_{i \in A_\phi}(U_i - L_i)$. Targets are standardized, so this is already in units of the target standard deviation.
- **Selective RMSE / MAE**: $\left(\frac{1}{S}\sum_{i \in A_\phi}(\hat{y}_i - y_i)^2\right)^{1/2}$ and $\frac{1}{S}\sum_{i \in A_\phi}|\hat{y}_i - y_i|$, on the original target scale. MAE is used for the two image tasks, following their convention; RMSE elsewhere.
- **Selective interval score**: $\frac{1}{S}\sum_{i \in A_\phi} \mathrm{IS}_\alpha(L_i, U_i; y_i)$ with IS_α as in equation 12.
- **Value-position error**: $\frac{1}{|C|}\sum_{i \in C}|v_i - v_i^\star|$, $v_i^\star = (y_i - L_i)/(U_i - L_i)$, defined only on captured inputs since $v^\star \notin [0,1]$ otherwise; we report $|C|/S$ alongside it.
- **AURC**: the trapezoidal area of any of the above against ϕ over the grid $[0.3, 1]$, divided by 0.7 so it reads as an average rather than an area. **E-AURC** subtracts the oracle computed on the same base predictor.
- **Selection efficiency** $\mathrm{SE} = (\mathrm{AURC}_{\text{width}} - \mathrm{AURC}_{\text{method}})/(\mathrm{AURC}_{\text{width}} - \mathrm{AURC}_{\text{oracle}})$: computed against the *same-predictor* width threshold and oracle, never a separately trained IPIV's. $\mathrm{SE} = 1$ is the oracle, $\mathrm{SE} = 0$ the width threshold, and $\mathrm{SE} < 0$ a selector that is worse than a width threshold on its own predictor; SE is undefined when the width and oracle AURC coincide (yacht, for the ensemble's interval score). For the deep-ensemble selectors the "own predictor" is the five-model ensemble of §5, and width is its aggregated interval width, computed the same way on all three groups: IPIV's published aggregate (mean bounds widened by $z\,\sigma/\sqrt{M}$ across members). The disagreement signals' failure-mode AUROC (Table 3, and the corresponding columns of Table 15) is reported on the 18 tabular datasets only, matching the scope of the diagnostic it extends; Appendix O's image rows report selection efficiency only, marked "–" in the AUROC columns.

Two events are used as binary labels for the gate diagnostics. A *coverage failure* is $k_i = 0$: the interval missed. A *value failure* is $k_i = 1$ together with $|\hat{y}_i - y_i|$ above the median absolute error among captured inputs: the interval was right but the value head was misplaced within it. The median is taken per dataset, so a value failure is by construction a within-dataset relative event, and roughly half of all captured inputs qualify. The pipeline standardizes targets before training, so widths are already in s.d. units, while point errors are converted back to the original scale — width is therefore comparable across datasets and error is not, which is why selection efficiency, not raw RMSE, is the cross-dataset statistic.

F Image Regression Results

Table 6: **Full image-regression grids**, every trained point rather than the three buckets of Table 7. *Top*: bone age, single gate, one run over 200 test images — note the severe drift between target and realized acceptance rate, which is why Table 7 is keyed by the realized rate. *Middle*: IMDB-WIKI, single gate then the symmetric dual-gate grid, 5-fold CV at 90 epochs. *Bottom*: SelectiveNet-regression on IMDB-WIKI over its own published coverage grid.

Target ϕ	Realized ϕ	$\mathrm{PICP}_{\mathrm{sel}}$	$\mathrm{MPIW}_{\mathrm{sel}}$	$\mathrm{MAE}_{\mathrm{sel}}$
bone age, single gate — IPIV at $\phi = 1$: 0.915, 2.189, 20.71				
0.30	0.000	—	—	—
0.40	0.070	0.929	1.589	10.901
0.50	0.075	0.933	1.729	16.505
0.60	0.190	1.000	2.063	14.099
0.70	0.225	0.889	1.889	15.371
0.75	0.480	0.917	1.775	16.381
0.80	0.505	0.950	1.946	14.526
0.85	0.625	0.904	1.869	16.198
0.90	0.680	0.963	2.075	14.548
0.95	0.865	0.879	1.933	18.027
IMDB-WIKI, single gate, 90 epochs — IPIV at $\phi = 1$: 0.933, 2.835, 7.683				
0.30	0.302	0.921	2.480	7.261
0.40	0.401	0.943	2.854	7.691
0.50	0.503	0.934	2.847	7.834
0.60	0.602	0.935	2.857	7.826
0.70	0.702	0.940	2.942	7.796
0.75	0.750	0.942	3.019	7.975
0.80	0.799	0.937	3.113	8.455
0.85	0.851	0.938	3.137	8.416
0.90	0.900	0.948	3.326	8.433
0.95	0.959	0.926	3.259	9.113
IMDB-WIKI, dual gate, symmetric budgets; second column is the realized joint *rate*				
0.5	0.490	0.942	2.884	7.641
0.6	0.591	0.942	3.142	8.252
0.7	0.694	0.945	3.217	8.446
0.8	0.795	0.939	3.159	8.376
0.9	0.898	0.932	3.201	8.841
IMDB-WIKI, SelectiveNet-regression — cols. are realized ϕ, sel. MAE, full MAE, aux. MAE				
0.70	0.700	6.294	7.257	7.256
0.75	0.751	6.486	7.314	7.303
0.80	0.799	6.506	7.202	7.204
0.85	0.851	6.755	7.328	7.338
0.90	0.901	6.847	7.242	7.242
0.95	0.951	6.999	7.214	7.213

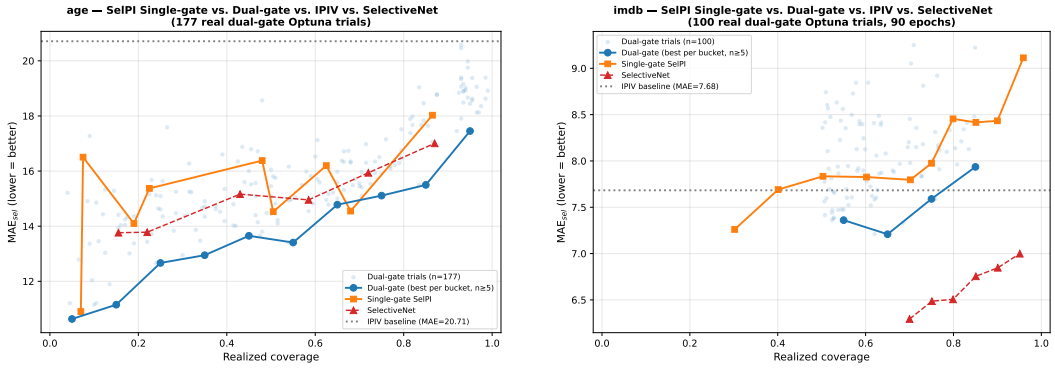


Figure 3: **All four comparisons on the two image tasks**, at $1 - \alpha = 0.95$: single-gate SelPI (orange), dual-gate SelPI (blue: every Optuna trial faintly, best-per-bucket as the line), plain IPIV at full acceptance (gray dotted), and SelectiveNet-regression (red). *Left, bone age*: both SelPI variants beat IPIV everywhere and the dual gate beats SelectiveNet at every comparable acceptance rate. *Right, IMDB-WIKI*: the dual gate beats IPIV over the 0.5–0.8 range but **SelectiveNet wins clearly and consistently**, by 0.7–1.6 years of MAE. IMDB-WIKI is the largest dataset in the study and this is the paper's clearest negative result (§6).

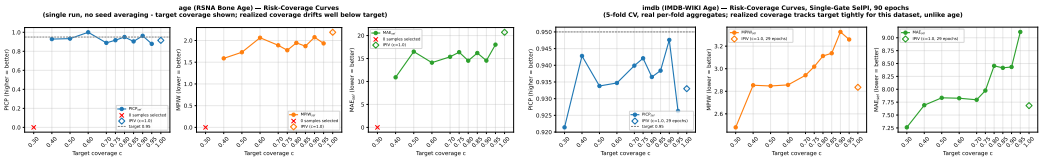


Figure 4: **Risk–coverage curves for the two image tasks**: selective PICP, MPIW and MAE against target acceptance rate, with plain IPIV at full acceptance as the hollow diamond. Note the contrast in calibration transfer: IMDB-WIKI's realized acceptance rate tracks its target almost exactly ($0.70 \to 0.700$, $0.95 \to 0.959$), whereas bone age drifts severely (target $0.90 \to$ realized 0.68; target $0.30 \to$ realized 0.00). The cause is sample size — thresholds calibrated on validation transfer cleanly to IMDB-WIKI's $\sim$34k-sample folds and poorly to bone age's 200-image test set. Bone age's "target coverage" axis should therefore be read as a training knob, not as a description of what fraction of test inputs was accepted.

Table 7: **Image regression**, selective MAE at $1 - \alpha = 0.95$ (bone age in months, IMDB-WIKI in years), grouped by *realized* acceptance rate. Plain IPIV at full acceptance is 20.71 (bone age) and 7.68 (IMDB-WIKI). For the dual gate we give the best and the median trial in each bucket, because its budgets are tuned by search while SelectiveNet is run once at its published settings — the median is the fair comparison. *Bone age*: every SelPI variant beats IPIV by a wide margin, and the dual gate's best trial beats SelectiveNet in all three buckets. *IMDB-WIKI*: the dual gate beats IPIV in the two lower buckets, but **SelectiveNet wins decisively wherever the two overlap**, by 1.3–1.5 years. Rows are keyed by realized rather than target acceptance rate, which matters most for bone age, whose drift is severe (target 0.90 gives realized 0.68) because a threshold calibrated on validation transfers poorly to a 200-image test set; IMDB-WIKI has no such problem and is the more trustworthy of the two.

	RSNA bone age			IMDB-WIKI		
Realized acceptance rate	$[0.5, 0.6)$	$[0.7, 0.8)$	$[0.8, 0.9)$	$[0.5, 0.6)$	$[0.7, 0.8)$	$[0.8, 0.9)$
IPIV ($\phi = 1$, no selection)		20.71			7.68	
SelectiveNet-regression	14.96	15.94	17.01	—	**6.29**	**6.76**
$\text{SelPI}_{1\text{G}}$ single gate	14.53	—	18.03	7.83	7.80	8.42
$\text{SelPI}_{2\text{G}}$ dual gate (median trial)	15.07	16.23	17.83	7.92	8.16	8.21
$\text{SelPI}_{2\text{G}}$ dual gate (best trial)	**13.41**	**15.11**	**15.50**	**7.36**	7.59	7.94

G Full Per-Dataset Results

Table 8: **Dual gate vs. plain IPIV by acceptance bucket**, median search trial, reduction in selective RMSE; buckets are the realized joint acceptance rate. The gain decays as acceptance rises on every dataset, as it must: as $\phi \to 1$ the accepted set converges on the full test set and there is nothing left to select.

Dataset	$[0.5, 0.6)$	$[0.6, 0.7)$	$[0.7, 0.8)$	$[0.8, 0.9)$
yacht	63.3%	61.6%	56.1%	36.5%
boston	10.5%	8.5%	7.5%	4.2%
energy	59.4%	51.0%	50.5%	42.3%
concrete	33.5%	25.5%	22.2%	16.6%
wine	16.2%	14.8%	10.1%	8.1%
kin8nm	11.0%	7.1%	1.5%	-6.0%
power-plant	7.5%	4.7%	3.2%	1.7%
naval	15.9%	14.7%	3.0%	-6.4%
protein	28.3%	22.1%	14.4%	11.0%
Year-MSD	43.1%	35.3%	28.9%	21.2%
superconduct	56.9%	47.3%	39.5%	29.5%
sarcos	51.5%	44.7%	35.3%	29.4%
diamonds	73.3%	65.0%	56.6%	44.9%
video_transcoding	87.5%	84.8%	80.1%	73.1%
medical_charges	30.9%	25.7%	19.6%	16.1%
black_friday	8.1%	5.4%	4.9%	4.0%
particulate-matter	29.7%	24.7%	19.1%	15.1%
nyc-taxi-green	31.7%	28.0%	21.8%	18.5%

Table 9: **What abstention buys.** Single-gate SelPI$_{1G}$, one model trained per acceptance rate (not a searched trial population). Selective RMSE at four acceptance rates, against plain IPIV answering every query; 18 tabular datasets, $1 - \alpha = 0.95$. The last column is the reduction at $\phi = 0.5$. Naval's full-acceptance error is 0.001, so its percentage is a small-denominator artifact rather than a meaningful loss.

	IPIV	SelPI$_{1G}$ selective RMSE at $\phi =$				Cut at
Dataset	$(\phi = 1)$	0.3	0.5	0.7	0.9	$\phi = 0.5$
yacht	1.04	0.45	0.44	0.55	1.07	57.7%
boston	2.64	2.23	2.18	2.63	2.55	17.5%
energy	1.63	0.61	0.91	0.84	1.10	44.4%
concrete	6.17	3.81	4.31	5.03	5.60	30.1%
wine	0.66	0.47	0.52	0.54	0.60	21.2%
kin8nm	0.08	0.07	0.07	0.08	0.09	13.6%
power-plant	4.10	3.53	3.73	3.78	3.99	9.1%
naval	0.001	0.002	0.002	0.001	0.002	-27.5%
protein	4.41	2.99	3.21	3.66	4.09	27.3%
Year-MSD	9.01	4.26	5.10	6.14	7.77	43.4%
superconduct	13.05	3.37	5.50	7.74	10.76	57.9%
sarcos	3.23	1.10	1.42	1.90	2.59	55.9%
diamonds	563.40	103.29	137.61	221.58	350.09	75.6%
video_transcoding	2.83	0.27	0.33	0.49	1.02	88.3%
medical_charges	0.08	0.05	0.06	0.06	0.07	30.2%
black_friday	0.39	0.34	0.35	0.37	0.38	8.4%
particulate-matter	0.39	0.22	0.27	0.31	0.35	31.2%
nyc-taxi-green	0.41	0.25	0.27	0.31	0.35	34.3%

Table 10: **Dual gate vs. SelectiveNet-regression**, selective RMSE at 70% acceptance. "Best" is the strongest configuration the search found, "median" the typical one; the median column is the fair comparison, because SelectiveNet is run once at its published settings while the dual gate is tuned (§5). Positive percentages favor SelPI. Naval and kin8nm are the two clear losses, and both are near-deterministic tasks with little uncertainty for an interval method to characterize.

Dataset	SelectiveNet	SelPI$_{2G}$ best	SelPI$_{2G}$ med.	Best	Median
yacht	1.62	0.27	0.35	83.1%	78.3%
boston	2.42	2.20	2.39	9.2%	1.3%
energy	1.17	0.63	0.82	46.2%	30.2%
concrete	4.66	4.24	4.59	9.1%	1.5%
wine	0.56	0.53	0.56	5.6%	0.0%
kin8nm	0.07	0.08	0.08	-9.0%	-15.5%
power-plant	3.86	3.81	3.92	1.3%	-1.5%
naval	0.0006	0.0009	0.0011	-51.3%	-83.6%
protein	3.93	3.37	3.56	14.3%	9.4%
Year-MSD	6.13	5.71	5.96	6.8%	2.8%
superconduct	8.63	6.81	7.40	21.1%	14.3%
sarcos	2.40	1.76	1.91	26.4%	20.3%
diamonds	237.56	205.32	232.45	13.6%	2.2%
video_transcoding	0.79	0.43	0.50	45.3%	37.4%
medical_charges	0.07	0.06	0.06	12.2%	10.2%
black_friday	0.38	0.35	0.37	6.2%	3.5%
particulate-matter	0.32	0.30	0.30	6.3%	4.1%
nyc-taxi-green	0.34	0.28	0.31	17.0%	9.6%

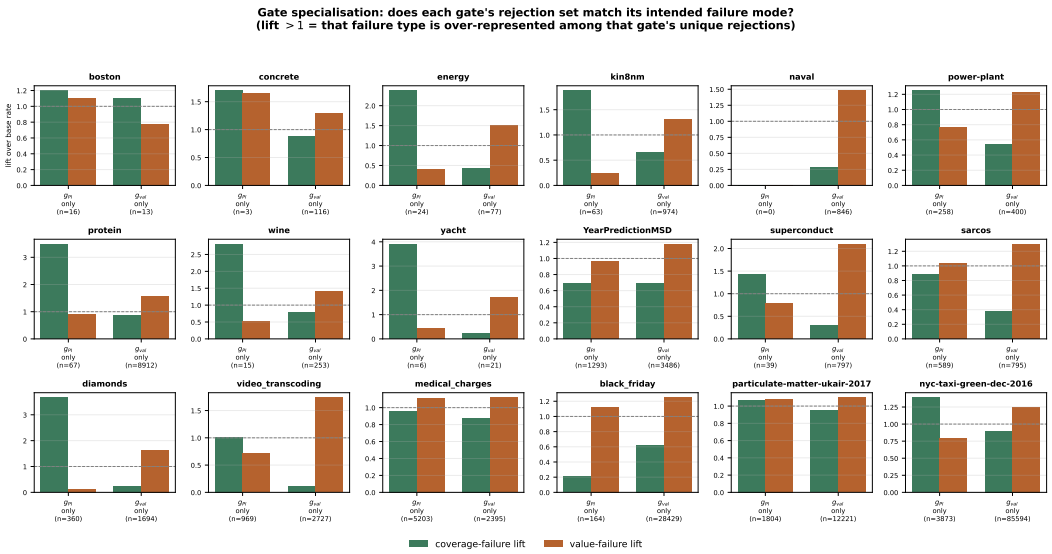


Figure 5: **Gate specialization, all 18 tabular datasets.** For each dataset, the two bars show how over-represented each failure type is among the inputs that *only* g_{PI} rejects and the inputs that *only* g_{val} rejects, relative to that dataset's own base rate (dashed line at 1). The signature of successful specialization is a tall green bar on the left pair and a tall orange bar on the right pair. g_{val} shows it on 17 of 18; g_{PI} on a clear majority, with the exceptions discussed in §7.

Table 11: **Each gate's unique rejections are enriched for the failure mode it was trained on.** For inputs rejected by *only* one gate, how over-represented each failure type is relative to that dataset's base rate; 1.0 means no enrichment. g_{val}'s unique rejections are enriched for value failure on 17 of 18 datasets; g_{PI}'s are enriched for coverage failure on 13 of the 17 where it has any unique rejections at all (on naval it has none, its rejection set being a subset of g_{val}'s). The value gate is therefore the more consistently load-bearing of the two, which the aggregate results do not reveal.

	rejected by g_{PI} only			rejected by g_{val} only		
Dataset	n	cov. lift	val. lift	n	val. lift	cov. lift
yacht	6	3.91	0.42	21	1.69	0.22
boston	16	1.20	1.10	13	0.78	1.11
energy	24	2.38	0.42	77	1.51	0.43
concrete	3	1.70	1.66	116	1.29	0.88
wine	15	2.81	0.51	253	1.40	0.78
kin8nm	63	1.88	0.24	974	1.32	0.66
power-plant	258	1.25	0.77	400	1.23	0.55
naval	0	—	—	846	1.48	0.29
protein	67	3.49	0.89	8912	1.56	0.85
Year-MSD	1293	0.69	0.96	3486	1.18	0.69
superconduct	39	1.43	0.78	797	2.10	0.30
sarcos	589	0.89	1.03	795	1.30	0.38
diamonds	360	3.67	0.12	1694	1.63	0.25
video_transcoding	969	1.01	0.71	2727	1.74	0.12
medical_charges	5203	0.96	1.11	2395	1.12	0.88
black_friday	164	0.21	1.12	28429	1.25	0.62
particulate-matter	1804	1.07	1.08	12221	1.10	0.95
nyc-taxi-green	3873	1.39	0.79	85594	1.25	0.90

H Statistical Tests

Two comparisons are kept separate because only the second isolates the selector. The learned gate against a width threshold *on the same predictor*: 20 wins, 0 losses, $p = 2 \times 10^{-6}$, median selection efficiency 0.587. The whole pipeline against a separately trained IPIV + width: 16 wins, 3 losses, $p = 5.2 \times 10^{-4}$. On the interval score the same-predictor comparison gives 20 wins, 0 losses, median 0.872. The gate against deep-ensemble disagreement (a cross-pipeline comparison of selection efficiency): 20 wins, 0 losses ($p = 2 \times 10^{-6}$) on selective RMSE and 19 wins, 0 losses ($p = 4 \times 10^{-6}$) on the interval score, for both disagreement signals. Coverage within width quintiles of the accepted set (Angelopoulos & Bates, 2023) is nearly uniform on the six largest datasets (gap ≤ 0.017) but not on the small UCI sets (up to 0.375 on yacht), where accepted-set calibration is partly carried by one easy sub-population. Friedman and Nemenyi tests are not applicable, since they rank a field of methods and each of our comparisons has two.

I Two-Budget Frontier

J "When Does It Help": Regression Details

The figure supports only one of the three relationships §7 predicts. **Heteroscedasticity: confirmed**, $\rho = 0.54$, $p = 0.013$, $n = 20$, where the heteroscedasticity statistic is Spearman $\rho(U - L, |y - \hat{y}|)$ on the plain IPIV baseline. **Excess kurtosis: no relationship**, $\rho = -0.03$, $p = 0.91$. **Full-acceptance miscoverage: weakly negative and not significant**, $\rho = -0.32$, $p = 0.20$ — if anything the gate does *worse* where there is more miscoverage, the opposite of the intuitive direction.

K Distribution-Shift Protocols and Results

For each dataset and run we regenerate the exact test split the model saw, identify the covariate that best tracks realized error, and treat its top quintile as a shifted tail; no retraining is needed. Holding each selector's threshold fixed at its $\phi = 0.7$ value, we measure how much its abstention rate rises

Table 12: **Per-dataset results, merged.** All 18 tabular datasets, $1-\alpha = 0.95$. $\rho(g, U-L)$: Spearman correlation between the gate and interval width, mean and range over the acceptance grid. *vs. IPIV+width*: improvement in selective RMSE over a *separately trained* IPIV with a width threshold — a cross-pipeline comparison, which is why it goes sharply negative where plain IPIV is simply the better predictor; this is distinct from the same-predictor SE columns, which stay confound-free in $[0, 1]$ by construction (§5). *SE*: selection efficiency of the learned gate over a width threshold on the same base predictor, on selective RMSE and on the interval score (1 = matches the oracle, 0 = matches the width threshold). $\phi = 1$ *vs. IPIV*: deviation from plain IPIV at full acceptance — the cost of training with the selective objective. *Ablations*: change in selective RMSE at $\phi = 0.5$ when one component is removed or replaced; positive is worse.

	$\rho(g, U-L)$		vs. IPIV + width		Selection efficiency		$\phi = 1$ vs. IPIV			Ablations ($\Delta\text{RMSE}_{\text{sel}}$)		
Dataset	mean	range	$\phi = 0.5$	$\phi = 0.7$	RMSE	IS_α	ΔPICP	ΔMPIW	ΔRMSE	no aux	frozen	$\sqrt{n}$
yacht	-0.55	[-0.76, -0.37]	-14.9%	26.7%	0.833	0.868	-0.065	-13.7	+36.1	+41.1	+19.7	+20.4
boston	-0.18	[-0.33, -0.04]	6.3%	-1.2%	0.456	0.766	+0.024	+20.2	+4.2	-4.3	-7.7	-3.1
energy	-0.57	[-0.71, -0.36]	-55.8%	-12.9%	0.732	0.951	+0.005	+45.6	+21.4	-36.8	+6.8	+8.1
concrete	-0.13	[-0.37, 0.25]	16.0%	10.4%	0.572	0.841	+0.004	+5.6	+4.7	+1.3	+6.3	+1.2
wine	0.21	[0.06, 0.48]	5.3%	4.3%	0.423	0.795	-0.011	-9.9	-3.6	-1.1	-0.1	-2.6
kin8nm	-0.14	[-0.50, 0.35]	-13.3%	-12.6%	0.710	0.884	-0.014	+38.3	+51.7	+1.8	+11.1	-0.8
power-plant	-0.23	[-0.47, 0.24]	3.0%	0.6%	0.160	0.526	-0.009	-2.2	+0.1	+2.8	+2.9	+0.6
naval	-0.37	[-0.70, 0.06]	-255.0%	-110.7%	0.891	0.972	-0.063	+91.0	+103.1	+26.7	-20.0	-6.7
protein	-0.15	[-0.39, 0.40]	12.0%	8.0%	0.602	0.866	-0.025	-6.7	+2.0	+4.1	+4.4	+1.4
Year-MSD	-0.56	[-0.73, -0.01]	10.0%	12.3%	0.567	0.799	-0.026	-11.9	-0.7	+1.9	+3.9	-0.2
superconduct	-0.32	[-0.67, 0.54]	9.3%	14.9%	0.868	0.981	-0.031	-16.5	-0.1	+41.2	+13.2	-2.7
sarcos	-0.63	[-0.74, -0.38]	23.1%	18.2%	0.543	0.876	-0.011	-7.1	+0.7	+2.4	+33.3	-0.0
diamonds	-0.37	[-0.85, 0.44]	-1.5%	4.1%	0.949	0.997	-0.012	-8.5	+11.0	+33.3	+8.2	-12.2
video_transcoding	-0.27	[-0.70, 0.58]	53.3%	40.1%	0.987	0.998	-0.019	-14.2	+35.9	+31.2	+54.2	+39.5
medical_charges	-0.71	[-0.88, -0.38]	3.7%	1.8%	0.321	0.661	-0.006	-3.3	+0.4	+0.4	+4.9	+0.4
black_friday	0.19	[-0.32, 0.53]	8.2%	4.1%	0.694	0.959	-0.054	-18.5	+28.9	+0.6	+3.7	-0.4
particulate-matter	-0.77	[-0.85, -0.63]	5.7%	3.6%	0.147	0.337	-0.004	-3.6	-0.6	-0.1	+3.7	+0.7
nyc-taxi-green	-0.05	[-0.26, 0.11]	34.0%	22.0%	0.785	0.959	-0.028	-10.5	+10.5	+2.1	+12.9	+2.0

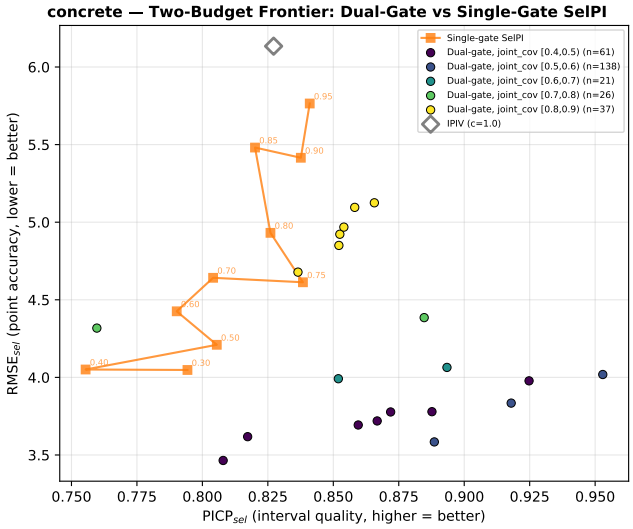


Figure 6: **The two-budget frontier**, Concrete. Each dot is one dual-gate configuration, colored by the joint acceptance rate it actually realized; the orange line is the single-gate model traced over its acceptance grid. Down and to the right is better (lower selective error, higher selective interval coverage). The dual-gate cloud sits below and to the right of the single-gate curve at matched acceptance — the trade-off a single scalar gate cannot express. Per-dataset versions for all 18 tabular datasets are in Figure 13.

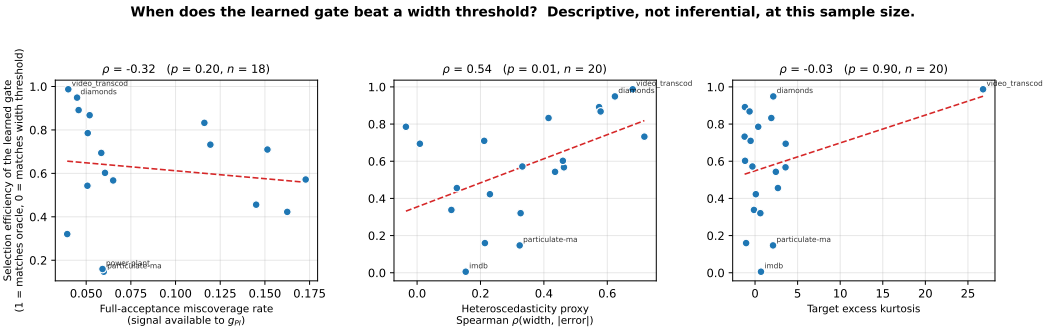


Figure 7: **When joint selection helps.** Per-dataset selection efficiency of the learned gate against (left) the full-acceptance miscoverage rate, (middle) a heteroscedasticity statistic, (right) target excess kurtosis. Each point is one of 20 datasets (18 for the left panel, where the two image tasks lack a full-acceptance miscoverage figure); Spearman ρ and its p-value are in each panel title, with a least-squares line for orientation only. Only the heteroscedasticity relationship is distinguishable from noise at this sample size.

on the tail. The learned gate raises abstention by a median $+0.193$ against $+0.046$ for the width threshold, and reacts more strongly on 15 of 18 datasets; on diamonds, black_friday and nyc-taxi-green the gate abstains on 73–96% of the tail while the width threshold's rate moves by less than 0.3, sometimes in the *wrong* direction. On two datasets (naval, particulate-matter) both selectors abstain *less* on the tail, i.e. the chosen covariate isolates an easier region there. Because the tail is constructed from a held-out covariate rather than injected, the protocol does not let us claim that the advantage grows with shift severity, and it does not cover the image tasks.

L Full Ablation Results

The three ablations of Table 13 were run on all 18 tabular datasets; per-dataset values are in Table 12 and Figures 14–16.

Table 13: **Ablations**: change in selective RMSE at $\phi = 0.5$ vs. the full recipe (positive: ablation is worse); 18 tabular datasets. W/B: datasets worse/better.

Variant	Med.	W/B	Range (%)
no auxiliary heads	$+2.0\%$	14/4	$-37, +41$
frozen IPIV, post-hoc gate	$+5.6\%$	15/3	$-20, +54$
$\sqrt{n}$ for $\sqrt{S_g}$	$+0.2\%$	9/9	$-12, +39$

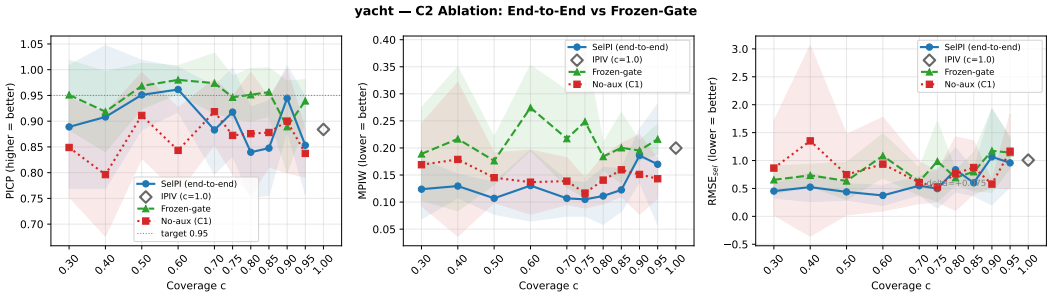


Figure 8: **Ablations on Yacht**, the dataset where the effects are largest: end-to-end SelPI against a gate trained on a frozen IPIV model, and against removing the auxiliary head. Selective PICP, MPIW and RMSE against target acceptance rate; shaded bands are ± 1 s.d. over 5 runs. End-to-end training dominates on both interval width and selective error across the whole grid. Equivalent figures exist for all 18 tabular datasets.

M Per-Dataset Figures

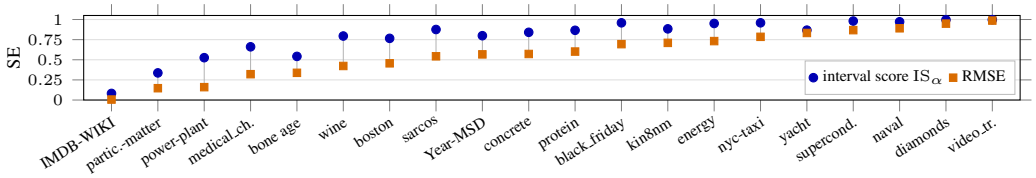


Figure 9: **Selection efficiency per dataset** (single-gate SelPI$_{1G}$): the share of the gap between the same-predictor width threshold (SE $= 0$) and the oracle (SE $= 1$) that the learned gate closes, on selective RMSE and on the selective interval score. The gate beats the width threshold on every dataset on both metrics, and by more on the interval score on all 20. The two image tasks are illustrative (§5); SE is scale-free, so their real-unit widths do not affect comparability.

Each grid shows the same plot for all 18 tabular datasets, in the order used throughout the appendix: the nine UCI sets, then the nine larger ones.

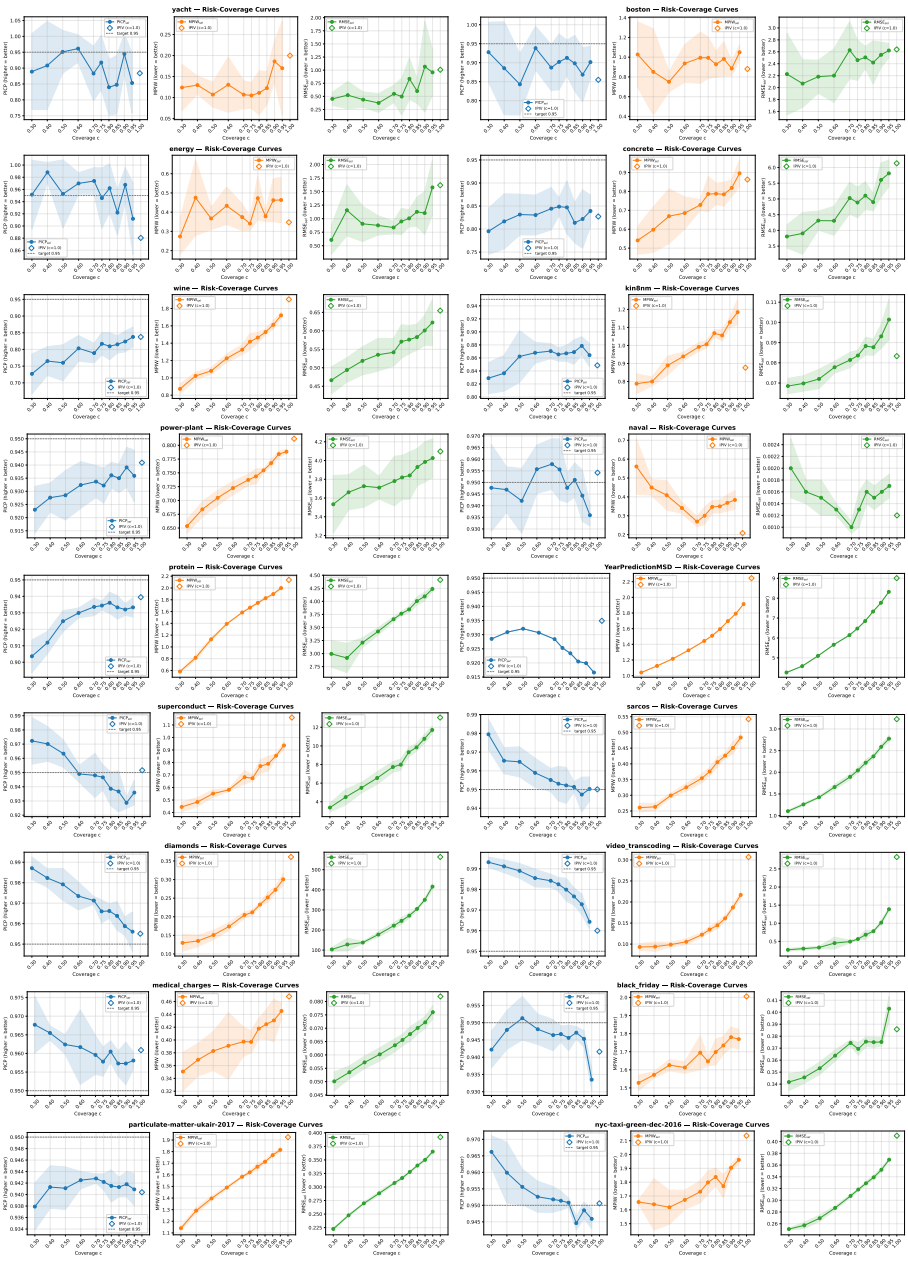


Figure 10: **Risk–coverage curves**: selective PICP, MPIW and RMSE against target acceptance rate, with plain IPIV at full acceptance as the hollow diamond. Bands are ± 1 s.d. over runs.

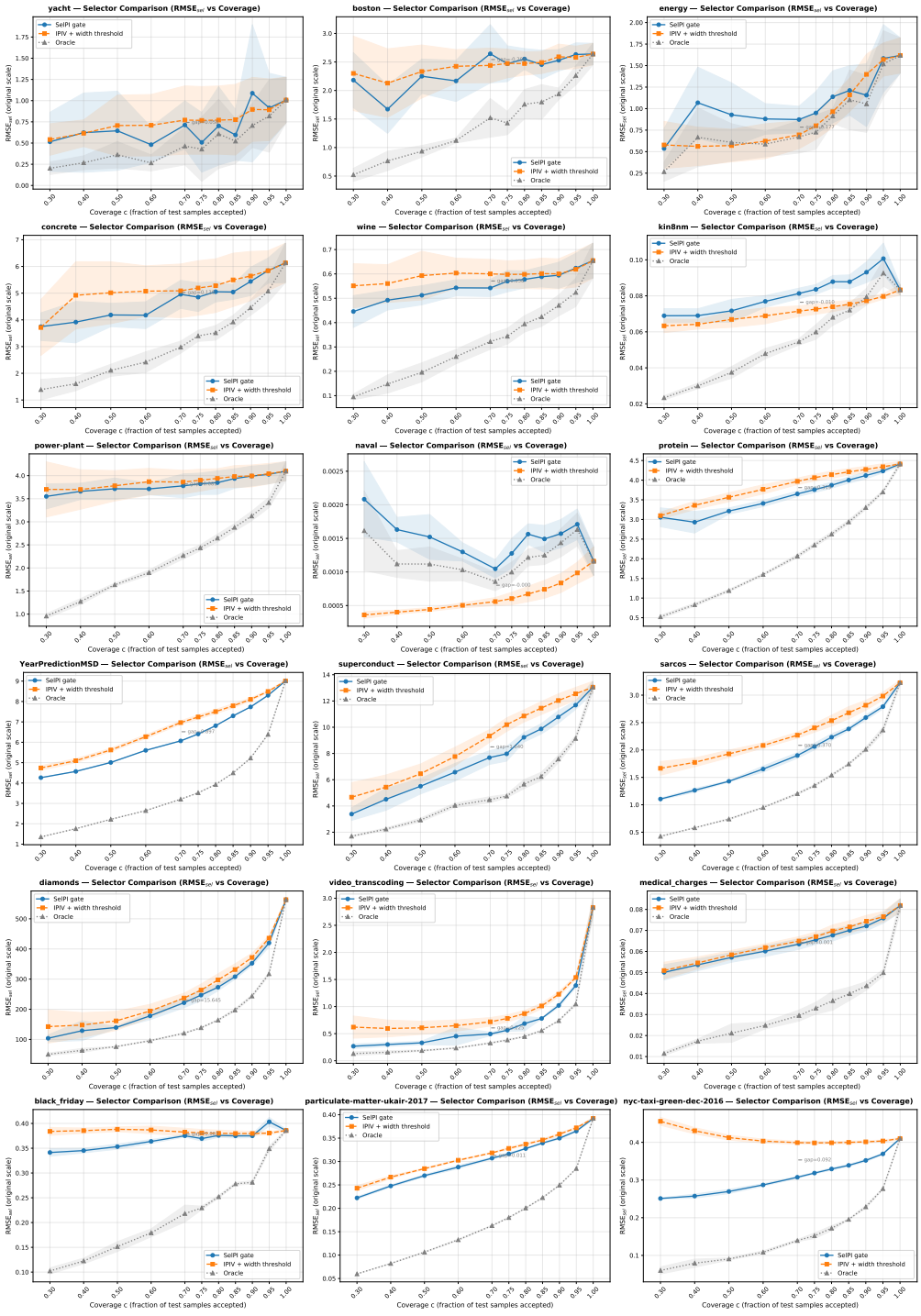


Figure 11: **Selector comparison**: the learned gate against IPIV + width threshold and against an oracle ranking by realized error. The persistent oracle gap is the headroom that remains on every dataset.

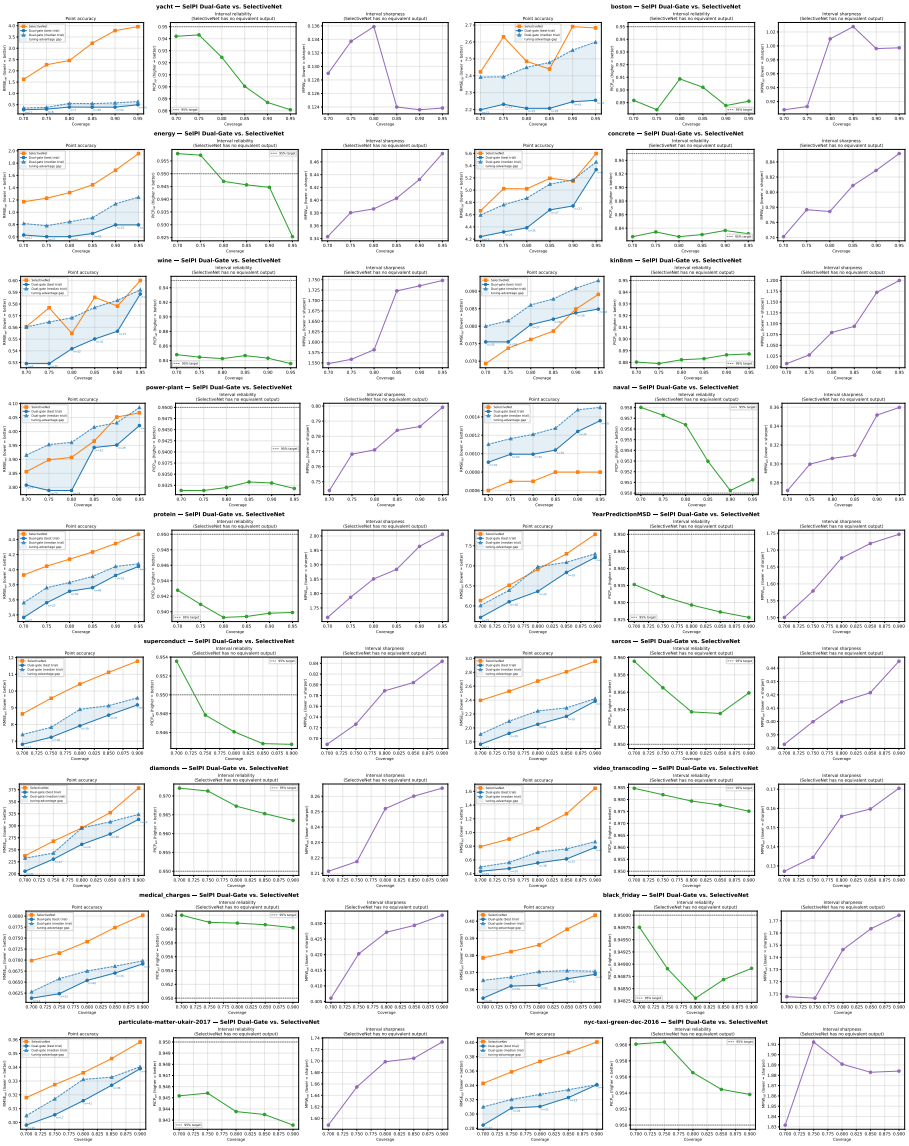


Figure 12: **Dual gate vs. SelectiveNet-regression**: point accuracy, then the interval reliability and sharpness SelectiveNet cannot produce at all. Shading between best and median trial is the tuning advantage of §5.

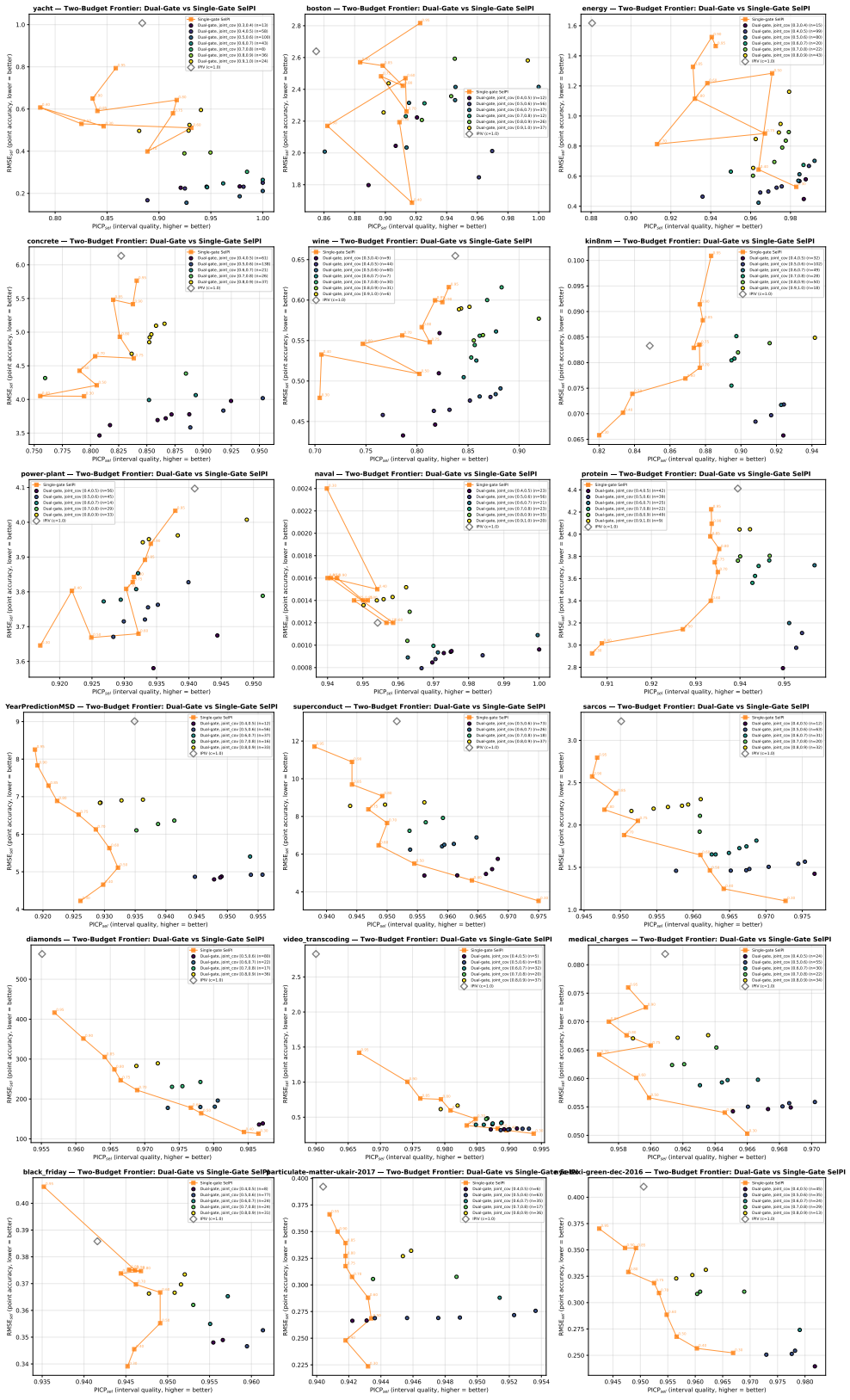


Figure 13: **Two-budget frontier**: every dual-gate configuration, colored by realized joint acceptance rate, against the single-gate curve. Down and to the right is better.

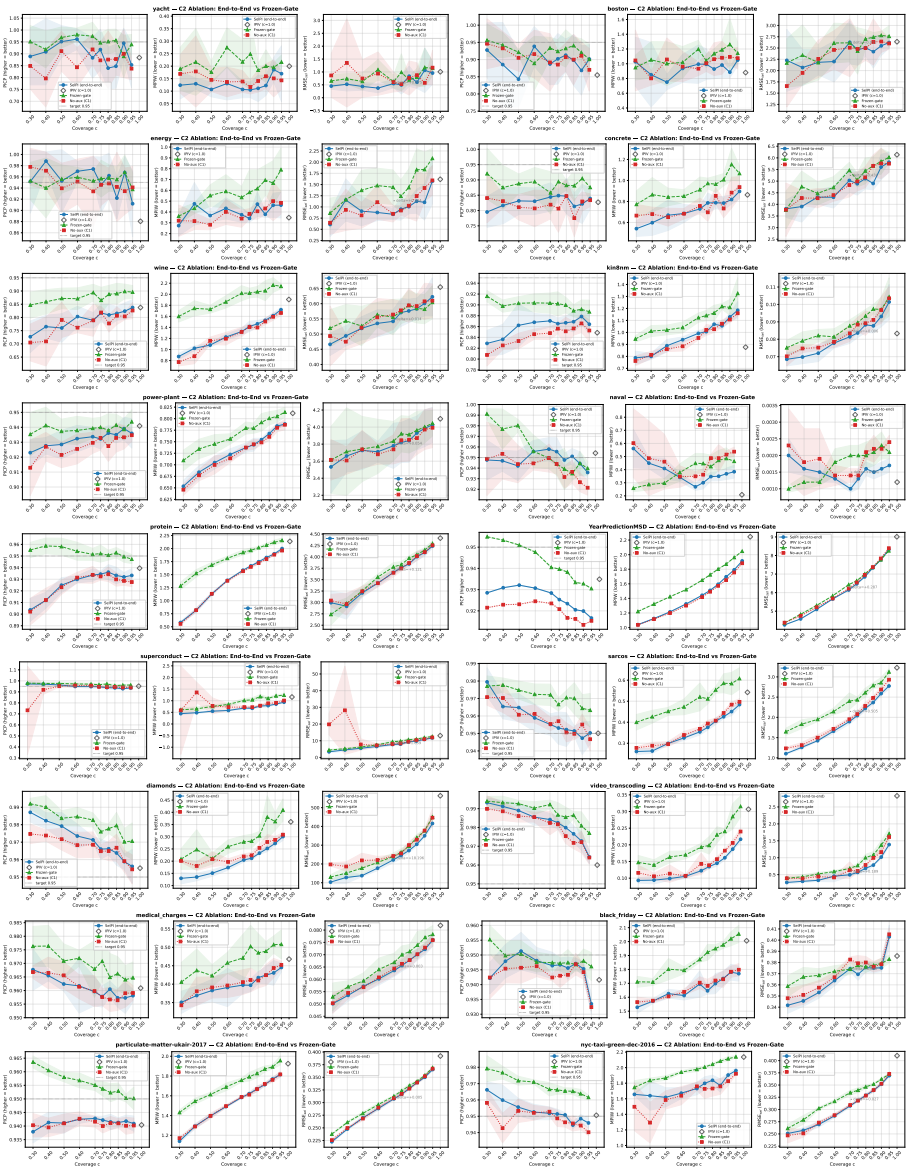


Figure 14: **Ablation: end-to-end vs. frozen-gate training**, with the no-auxiliary-head variant overlaid. Frozen-gate is worse on 15 of 18 datasets (Table 13). This overlay also contains the no-auxiliary-head comparison shown on its own in Figure 15 below.

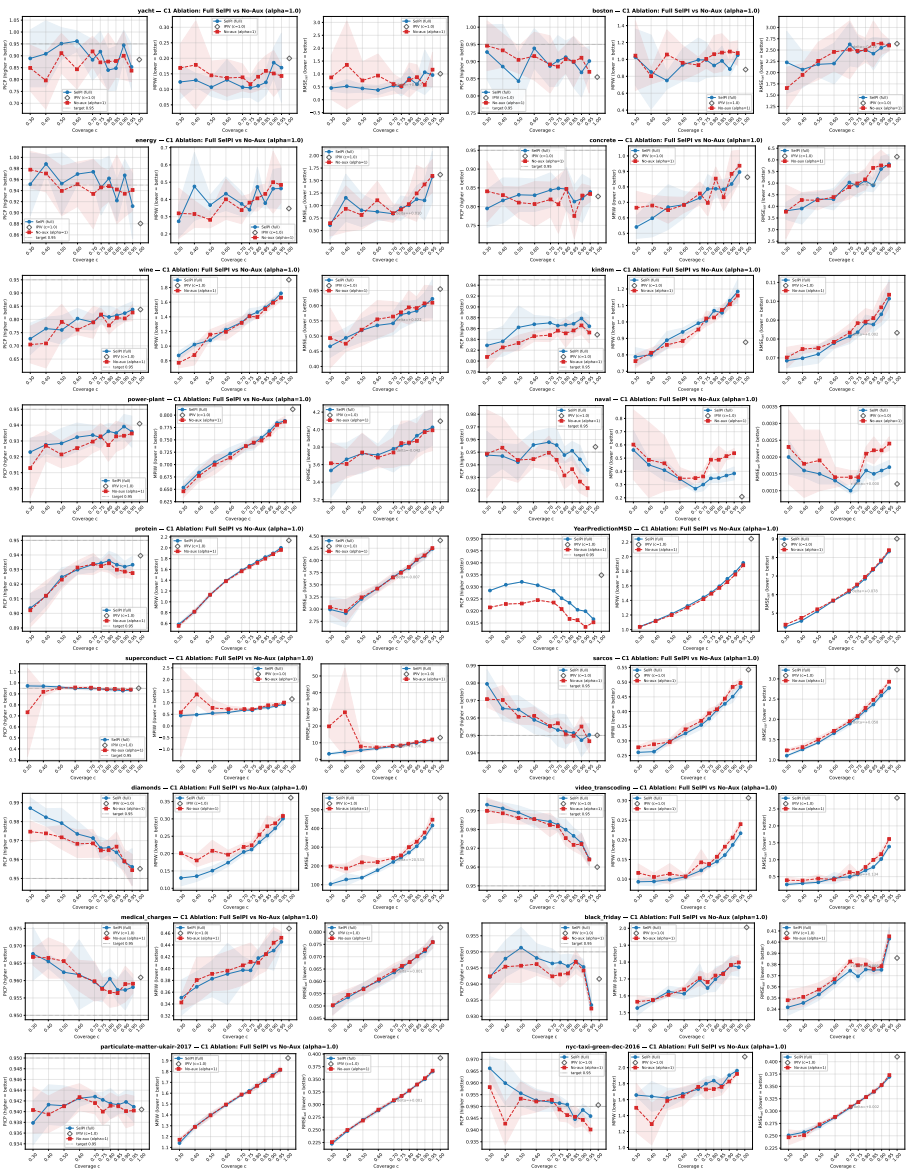


Figure 15: **Ablation: with vs. without the auxiliary head**, isolated into its own panel (also visible as part of the overlay in Figure 14).

N Full-Acceptance Baselines from the Literature

Our evaluation compares selectors, so it needs baselines that abstain. Neither parent paper does: IPIV and its comparators answer every query. Their published numbers are nonetheless useful in two ways, and we reproduce them here rather than leave the impression that no wider comparison exists. First, they place our reproduction of IPIV — the base predictor every SelPI number is built on — against the numbers its authors report, on the same datasets and protocol. Second, they show where the interval-quality field stood at full acceptance before a reject option is added.

Our IPIV reproduction tracks the published numbers closely on the tabular sets and is somewhat worse on the two image tasks, where we train a single model rather than an ensemble. Our own *Ens.* column closes most of that gap: on IMDB-WIKI it improves *repro*'s MAE from 7.68 to 7.42, and on bone age from 20.71 to 19.46, tracking the published DE/IPIV figures more closely once the base predictor is itself an ensemble rather than a single model. It still trails the published DE figure on IMDB-WIKI (6.66), which is expected: DE is a different architecture (a Gaussian-NLL

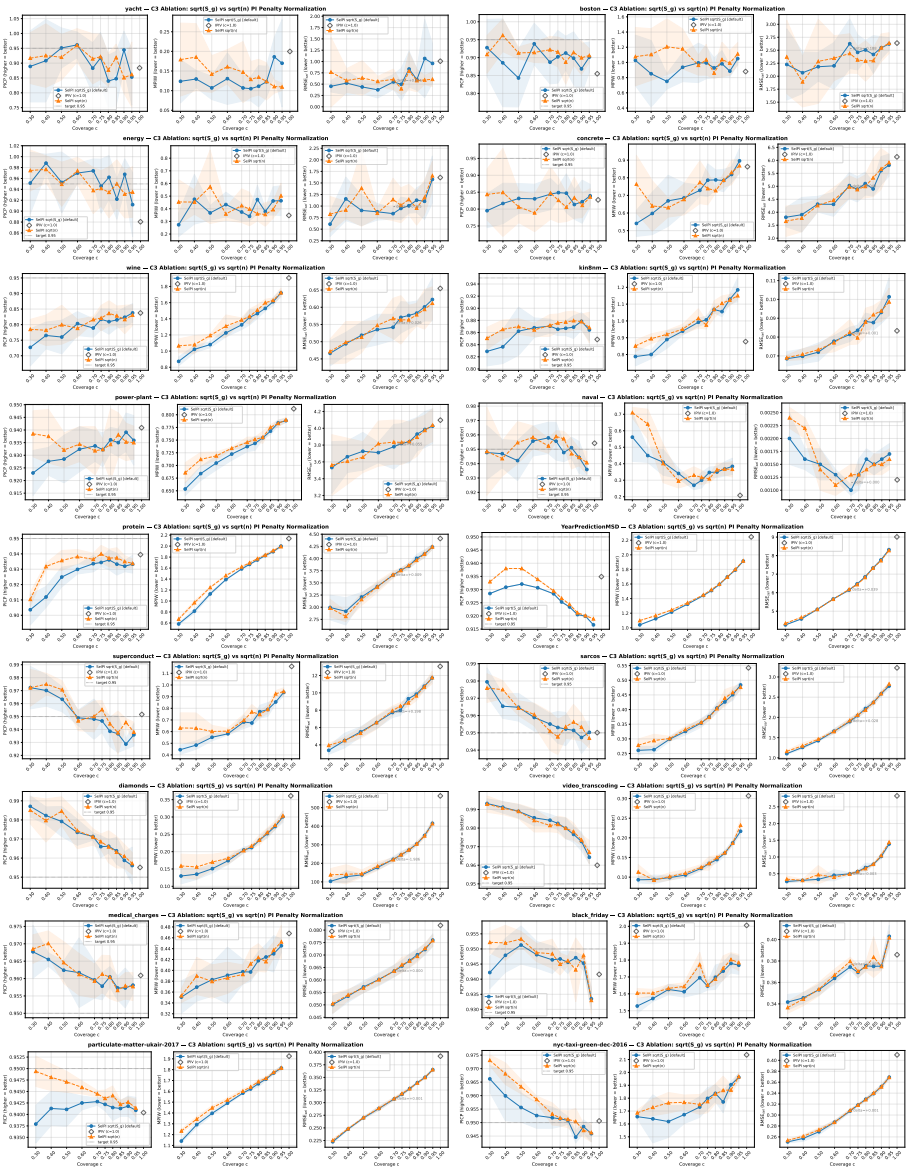


Figure 16: **Ablation: $\sqrt{S_g}$ vs. $\sqrt{n}$ penalty normalization.** The curves track each other closely on most datasets, which is the visual form of the null result in Table 13.

ensemble) evaluated over 20 splits, ours is an IPIV ensemble over 5. On the tabular sets the *Ens.* MPIW column tracks the published IPIV column almost exactly, as it should, since the published IPIV figures are themselves five-model ensembles. This *Ens.* predictor is also the one behind the deep-ensemble selector baseline of §6–§7, which we compare directly against SelPI's learned gates.

O Deep-Ensemble Selector: Per-Dataset Results

Table 15 gives the per-dataset values behind the selection-efficiency columns of Table 3: the selection efficiency of the learned gate and of the two disagreement selectors on both metrics, each against its own predictor's width threshold and oracle, and the two failure-mode AUROCs of the disagreement signals.

Table 14: Full-acceptance ($\phi = 1$) results as published in the IPIV paper, with our own reproduction of IPIV alongside (the *repro* columns — this is plain IPIV retrained in our environment, not SelPI). QD, QD+ and DE are that paper's own baselines. *Ens.* is our own M=5-model IPIV ensemble (§5), the base predictor behind the deep-ensemble selector of §6. Published figures use 20 splits and 5-model ensembles; ours use 5 splits (and, for *Ens.*, 5 members per split), so small differences are expected and the comparison is one of fidelity, not of ranking. RMSE/MAE are on the original target scale; MPIW is in standardized units.

	RMSE						MPIW					
Dataset	DE	QD	QD+	IPIV	*repro*	*Ens.*	DE	QD	QD+	IPIV	*repro*	*Ens.*
Boston	2.87	3.39	6.12	3.13	*2.64*	*2.70*	0.87	1.15	4.98	1.09	*0.88*	*1.09*
Concrete	5.21	5.88	11.87	5.43	*6.17*	*5.79*	1.01	1.08	3.43	1.02	*0.86*	*1.08*
Energy	1.68	2.28	7.05	1.65	*1.63*	*1.70*	0.49	0.45	1.52	0.42	*0.35*	*0.42*
Kin8nm	0.08	0.08	0.19	0.07	*0.08*	*0.08*	1.14	1.18	2.64	1.10	*0.88*	*1.10*
Naval	0.00	0.00	0.01	0.00	*0.00*	*0.00*	0.31	0.27	2.85	0.24	*0.21*	*0.27*
Power	3.99	4.14	12.00	4.08	*4.10*	*4.09*	0.91	0.86	1.51	0.86	*0.81*	*0.86*
Protein	4.36	4.99	5.10	4.35	*4.41*	*4.36*	2.68	2.27	3.30	2.26	*2.14*	*2.25*
Wine	0.62	0.67	0.70	0.63	*0.66*	*0.62*	2.50	2.24	4.89	2.22	*1.91*	*2.20*
Yacht	1.38	1.10	8.19	0.98	*1.04*	*0.95*	0.33	0.18	1.25	0.17	*0.20*	*0.28*
Year-MSD	8.95	9.30	10.11	8.93	*9.01*	*8.93*	2.91	2.45	4.25	2.42	*2.25*	*2.47*
IMDB-WIKI (MAE)	6.66	10.23	10.08	7.03	*7.68*	*7.42*	2.61	3.47	5.01	2.87	*2.84*	*3.28*
Bone age (MAE)	18.69	20.24	25.18	18.13	*20.71*	*19.46*	2.17	1.99	5.24	2.09	*2.19*	*3.38*

Table 15: Per-dataset selection efficiency (SE) on selective RMSE and on the interval score for the learned gate (SelPI$_{1G}$) and for the two deep-ensemble disagreement selectors, and AUROC of each disagreement signal for coverage failure (cov.) and value failure (val.). Each SE is computed against its own predictor's width threshold and oracle; "–" marks an undefined SE (width and oracle AURC coincide).

	SE, selective RMSE			SE, interval score			AUROC, point disagr.		AUROC, bound disagr.	
Dataset	gate	point	bound	gate	point	bound	cov.	val.	cov.	val.
yacht	0.833	−0.961	−1.062	0.868	–	–	0.761	0.735	0.840	0.747
boston	0.456	0.108	−0.001	0.766	0.225	0.170	0.603	0.575	0.509	0.572
energy	0.732	−1.456	−1.471	0.951	−1.060	−0.552	0.681	0.795	0.674	0.837
concrete	0.572	−0.331	−0.269	0.841	−0.403	−0.235	0.605	0.519	0.598	0.527
wine	0.423	−0.088	−0.158	0.795	0.050	0.047	0.548	0.577	0.583	0.494
kin8nm	0.71	−0.326	−0.308	0.884	−1.015	−0.887	0.486	0.533	0.494	0.537
power-plant	0.16	−0.114	−0.174	0.526	0.147	0.040	0.486	0.465	0.472	0.444
naval	0.891	−0.773	−0.803	0.972	−30.2	−12.7	0.913	0.741	0.958	0.726
protein	0.602	−0.131	−0.365	0.866	0.230	0.187	0.565	0.620	0.599	0.425
Year-MSD	0.567	−0.358	−0.371	0.799	−0.254	−0.334	0.546	0.638	0.523	0.626
superconduct	0.868	−0.624	−0.579	0.981	−0.858	−0.659	0.693	0.686	0.692	0.666
sarcos	0.543	−0.758	−0.741	0.876	−2.332	−2.077	0.555	0.592	0.554	0.591
diamonds	0.949	−0.849	−0.693	0.997	−1.635	−1.408	0.746	0.739	0.742	0.744
video_transcoding	0.987	−1.020	−0.618	0.998	−3.189	−1.632	0.865	0.787	0.876	0.790
medical_charges	0.321	−0.453	−0.423	0.661	−0.988	−0.946	0.508	0.523	0.505	0.544
black_friday	0.694	−0.001	0.001	0.959	0.006	−0.102	0.498	0.524	0.486	0.551
particulate-matter	0.147	−0.457	−0.453	0.337	−0.874	−0.854	0.523	0.516	0.518	0.509
nyc-taxi-green	0.785	0.264	0.286	0.959	0.571	0.676	0.669	0.633	0.751	0.635
bone age	0.338	−0.072	−0.271	0.542	−0.621	−1.030	–	–	–	–
IMDB-WIKI	0.006	−0.167	−0.197	0.082	−0.396	−0.470	–	–	–	–
median (tabular)	0.648	−0.34	−0.37	0.872	−0.62	−0.55	0.584	0.606	0.591	0.581

P COST

SelPI adds exactly two things to a IPIV model: the gate heads and the auxiliary triple. For a tabular backbone with hidden width h, each gate costs $16h + 65$ parameters ($16h + 16$ for the hidden layer, 32 for batch normalization, 17 for the output unit) and the auxiliary triple costs $3h + 3$. On Concrete ($h = 50$, $d = 8$) that takes a 603-parameter IPIV model to 1,621 for the single gate and 2,486 for the dual gate; on superconduct ($h = 100$, $d = 79$) it takes 8,303 to 11,936, a 44% increase. The relative cost falls as the backbone grows and is negligible for the image tasks: on the bone-age

VGG-16 stack each gate reads a 1024-unit representation for $16{,}449$ parameters against a backbone of roughly 15M, so the two gates and the auxiliary triple together add about 0.2%.

Training cost is the more meaningful number, and we measure it directly rather than estimate it. Across the nine UCI datasets, training one SelPI model takes a median $\mathbf{1.63}\times$ the wall-clock of training one IPIV model on the same data and hardware (range $1.00\times$ on Yacht to $2.03\times$ on Protein, the largest). The whole-pipeline cost is higher than this ratio suggests, because SelPI trains one model per acceptance rate: a 10-point curve costs roughly $16\times$ a single IPIV fit.

Inference is where the architecture pays off. Both gates are scalar heads on the shared trunk, so an accept/reject decision, an interval and a point estimate all come from *one* forward pass — the same cost as plain IPIV, and unlike an ensemble or MC-dropout selector, which need K or T passes for their uncertainty signal. The deep-ensemble selector of §5 is the concrete case: it costs five IPIV fits per split to train and five forward passes per query at inference, against one SelPI fit per acceptance rate and one pass per query; the comparison in Table 3 (right columns) is made at that fivefold inference cost.

Q Additional Metrics Discussed in the Literature

The coverage-width criterion of the LUBE/QD line combines coverage and width into one number, $\mathrm{CWC} = \mathrm{NMPIW}\left(1 + \gamma(\mathrm{PICP})\, e^{-\eta(\mathrm{PICP}-\mu)}\right)$, where NMPIW is the mean width divided by the target range, μ is the nominal level, and γ is 1 only when coverage falls below nominal. We computed it for completeness but do not headline it, for two reasons. It is not a proper scoring rule, so it cannot be used to rank forecasters without the ranking depending on η; and its exponential penalty makes it extremely sensitive near the nominal level, which shows plainly in our numbers — on wine, a drop in selective coverage from 0.838 to 0.791 moves CWC from 86 to 605, a sevenfold change from a five-point coverage difference. The interval score of equation 12 measures the same trade-off with a linear miss penalty and a decision-theoretic justification, which is why it is our headline.