DISTRIBUTION-INFORMED EFFICIENT CONFORMAL PREDICTION FOR FULL RANKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Ranking models guide decisions in search, recommendation, screening, and information retrieval, yet a deterministic ordering does not express uncertainty about an item's position. Standard conformal prediction is difficult in full-ranking problems because calibration items may be observed only through their relative order, leaving their absolute ranks in the calibration–test pool latent. TCPR handles this uncertainty with high-probability rank envelopes and worst-case scores, a construction that can be conservative. We propose Distribution-informed Conformal Ranking (DCR), which maps the reference distribution induced by uniform rank interleaving into a mixture of potential calibration scores. MDCR approximates the same target through independent Beta–Binomial draws, avoiding dense reference-mixture construction. Our theory quantifies finite-sample coverage error and establishes asymptotic marginal validity under explicit local conditions; it does not infer exact validity from exchangeability alone. Across a controlled Gaussian setting and learned rankers on ESOL and fixed-anchor Yummly similarity, DCR and MDCR substantially reduce conservatism relative to the evaluated TCPR envelope. The experiments also expose predictor-dependent coverage–size trade-offs, including settings where a mean-rank point estimate is competitive, and identify when direct sampling reduces computation.

1 INTRODUCTION

Ranking models are widely used in search engines, recommendation systems, screening pipelines, and information retrieval (Liu et al., 2009; Karatzoglou et al., 2013). Their orderings determine which items receive exposure, which candidates advance, and which results receive further inspection. Yet the ordering alone does not indicate whether a predicted position is reliable. In a screening pipeline, for example, a small ranking error can move a qualified candidate across a decision boundary. A prediction set for that candidate's rank makes this uncertainty visible to the downstream decision maker.

Conformal prediction (CP) calibrates non-conformity scores on held-out examples with observed ground truth (Vovk et al., 2005; Shafer & Vovk, 2008). In transductive full ranking, however, the calibration data may reveal only the relative order of calibration items. Their absolute ranks in the calibration–test pool remain latent, so the scores required for direct calibration are unavailable. TCPR (Fermanian et al., 2025) addresses this problem by assigning each latent rank a high-probability envelope and calibrating the worst-case score inside it. The envelope protects against uncertain positions, but unlikely positions can determine the threshold and produce conservative sets.

We introduce Distribution-informed Conformal Ranking (DCR) to use the structure in these latent ranks. Uniformly interleaving calibration and test items induces a shifted negative-hypergeometric reference distribution for each calibration-relative rank. DCR passes this distribution through the scoring function, aggregates the resulting potential-score distributions, and calibrates a common threshold. It therefore retains probability mass across possible positions instead of reducing rank uncertainty to the worst score in an interval. The reference law describes uniform interleaving; it is not generally the true-rank posterior conditional on all observed features.

Monte Carlo Distribution-informed Conformal Ranking (MDCR) targets the same reference mixture without constructing it densely: an independent Beta–Binomial draw supplies one auxiliary rank for

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.

each calibration item. DCR and MDCR are not drop-in replacements for TCPR's finite-sample envelope guarantee. They exchange worst-case envelope calibration for a distribution-weighted target whose discrepancy from the true calibration-score distribution can be analyzed explicitly. Our theory quantifies this approximation cost, auxiliary sampling error, and realized false coverage proportion under score-array exchangeability and additional local conditions; exact counterexamples show why exchangeability alone is insufficient.

The experiments pair a controlled Gaussian mechanism with three ranking predictors on molecular solubility and fixed-anchor recipe similarity. At target coverage 0.90, DCR reduces ESOL/RankNet–RA size/N from TCPR's 0.5269 to 0.3889 while moving coverage from 0.9687 to 0.9008. Controlled Gaussian experiments separately examine finite-sample conservatism as calibration information changes. The benefit depends on the predictor and score: Mean-rank sometimes gives smaller sets and coverage closer to nominal, while direct auxiliary sampling is most useful when dense mixture construction is costly.

We summarize our contributions as follows:

- We propose DCR for full-pool rank prediction under partial rank observability, using the complete uniform-interleaving reference distribution rather than an envelope worst case, and MDCR as a direct-sampling approximation to the same target.
- We establish finite-sample coverage-error bounds and asymptotic marginal validity under explicit conditions, together with empirical FCP and conditional Monte Carlo error bounds.
- We evaluate the methods in a controlled Gaussian setting and with three ranking predictors on two shared real-data tasks, characterizing conservatism, coverage–size trade-offs, calibration information, and computation. Regression controls and complete sweeps appear in the appendix.

2 Related Work

Learning to rank and rank uncertainty. Learning-to-rank methods estimate orderings from relevance, preference, or comparison signals. Pointwise, pairwise, and listwise approaches encode these signals differently (Liu et al., 2009); representative methods include RankSVM, RankNet, ListNet, and LambdaMART (Joachims, 2002; Burges et al., 2005; Cao et al., 2007; Burges, 2010). Statistical ranking recovery instead asks when noisy comparisons identify an ordering or a top subset (Shah & Wainwright, 2018; Chen et al., 2022), while probabilistic ranking models place distributions over preferences or rankings (Qin et al., 2010). These lines of work primarily model or recover an ordering. We study prediction sets for an item's absolute pooled position when the calibration data reveal only relative ranks.

Conformal prediction for ranking and structured decisions. CP provides a general framework for predictive uncertainty (Vovk et al., 2005; Shafer & Vovk, 2008; Angelopoulos & Bates, 2021), with extensions to regression, adaptive classification, and risk-controlling sets (Lei et al., 2018; Romano et al., 2019; 2020; Bates et al., 2021). Conformal methods have also been developed for retrieval-augmented question answering (Li et al., 2024), recommendation (Angelopoulos et al., 2023), structured matrix prediction (Liang et al., 2024), and label ranking for classification (Huang et al., 2024). Their inferential targets differ from absolute item positions in a calibration–test pool. TCPR (Fermanian et al., 2025) is the closest setting: it also observes calibration-relative ranks and constructs full-pool rank sets using high-probability rank envelopes. DCR instead aggregates a reference distribution over the latent ranks and analyzes its discrepancy from the true calibration scores. Our empirical comparison is with the specified certified numerical TCPR envelope, not every possible envelope construction.

3 Distribution-informed conformal ranking

A ranking model may be fitted on independent training data. We observe the calibration/test pool's features and predictions and the true relative order of calibration items. Even when calibration outcomes can be sorted, unknown test outcomes leave their absolute full-pool ranks latent. We

predict each test item's rank in this fixed pool, studying marginal coverage for a uniformly selected test item and the empirical false coverage proportion (FCP) across the test pool.

Let $n, m \geq 1$, $N = n + m$, and let $\mathcal{I}_c, \mathcal{I}_t$ partition the pool. Rank 1 denotes the lowest outcome. Ties are ordered lexicographically using independent continuous tie breakers; the target is this randomized full rank. Let $R_i \in [N]$ be the true full rank and $r_i \in [n]$ the observed relative rank of a calibration item. A predictor gives full-pool predicted ranks $\widehat{R}_i$ and, when available, ordered predicted values $\widehat{y}_{(j)}$. Potential scores include normalized rank-adaptive (RA) and value-adaptive (VA):

$$s_i^{\mathrm{RA}}(j) = |j - \widehat{R}_i|/N, \qquad s_i^{\mathrm{VA}}(j) = |\widehat{y}_{(j)} - \widehat{y}_i|. \tag{1}$$

Positive common rescaling does not change sets. Given a common threshold T, output $C_i(T) = \{j : s_i(j) \leq T\}$. Write $S_i = s_i(R_i)$, $F_c(t) = n^{-1} \sum_{i \in \mathcal{I}_c} \mathbb{I}\{S_i \leq t\}$, $F_t(t) = m^{-1} \sum_{i \in \mathcal{I}_t} \mathbb{I}\{S_i \leq t\}$, and $\mathrm{FCP}(T) = 1 - F_t(T)$.

Assumption 3.1 (Score-array exchangeability)**.** The entire array $(R_i, s_i(1), \ldots, s_i(N))_{i=1}^N$ is exchangeable across items, and the calibration/test designation is a uniform partition independent of that array. Equivalently, one may use fixed designations with an exchangeable array. The score construction is the same for calibration and test items.

A sufficient implementation uses independent training data and a permutation-equivariant pooled predictor, without fitting scores to the calibration/test designation. Exchangeability of outcomes alone is insufficient for arbitrary data-dependent scores. Our trained predictors use only an independent training split; the same fitted mapping and pooled score construction are used for all calibration and test items.

TCPR handles an unknown absolute calibration rank by replacing it with an envelope and calibrating a worst-case score over that envelope. DCR instead retains the full reference distribution induced by uniform rank interleaving, maps rank uncertainty into a distribution of possible scores, and calibrates against the resulting mixture. This distribution-weighted calibration can be less conservative than envelope calibration when substantial reference mass lies away from the worst case. MDCR targets the same mixture calibration through independent auxiliary draws and can sample the reference ranks directly, without constructing the full mixture. The reference distribution is exact for uniform interleaving of calibration and test ranks, but it is not generally the posterior distribution of a latent rank conditional on all feature information. The analysis therefore asks how well the reference-score mixture approximates the true calibration-score distribution, rather than assuming that the latent-rank posterior is known.

Proposition 3.2 (Reference rank kernel)**.** *For the sorted calibration ranks $J_1 < \cdots < J_n$, $\{J_1, \ldots, J_n\}$ is a uniform n-subset of $[N]$. Thus*

$$K_r(j) = \Pr(J_r = j) = \frac{\binom{j-1}{r-1}\binom{N-j}{n-r}}{\binom{N}{n}}, \quad r \leq j \leq r + m. \tag{2}$$

In particular, $J_r - r$ has a beta-binomial law with parameters $(m, r, n + 1 - r)$.

The kernel also has a shifted negative-hypergeometric characterization. The induced auxiliary CDF and mixture are

$$F_i^{\mathrm{aux}}(t) = \sum_j K_{r_i}(j) \mathbb{I}\{s_i(j) \leq t\}, \qquad G(t) = \frac{1}{n} \sum_{i \in \mathcal{I}_c} F_i^{\mathrm{aux}}(t). \tag{3}$$

They need not equal the conditional distribution of S_i given all observed features and rankings. For example, with $n = m = 1$ and a perfect RA predictor, the true score is always zero, yet $G(0) = 1/2$.

DCR. For $0 < \alpha < 1$, set $k = \lceil (n + 1)(1 - \alpha) \rceil$, $q = k/(n + 1)$ and

$$T_D = Q_G(q) := \inf\{t : G(t) \geq q\}. \tag{4}$$

Use the discrete generalized quantile, without interpolation. When $q = 1$, DCR uses the maximum of its finite auxiliary support; this is not defined to be the full rank set for every test item.

Algorithm 1 Distribution-informed Conformal Ranking (DCR)

Input: Index sets $\mathcal{I}_c, \mathcal{I}_t$; relative calibration ranks r_i;
potential scores $s_i(j)$ from full-pool predictions; sizes n, m;
level $\alpha \in (0, 1)$.
Output: Rank prediction sets for every $i \in \mathcal{I}_t$.

$N \leftarrow n + m$, $k \leftarrow \lceil (n+1)(1-\alpha) \rceil$, $q \leftarrow k/(n+1)$
for each $i \in \mathcal{I}_c$ **do**
  Construct K_{r_i} by (2) and map its support through $s_i(j)$.
  Form the auxiliary score CDF F_i^{aux} in (3).
end for
Aggregate $G \leftarrow n^{-1} \sum_i F_i^{\text{aux}}$.
if $q = 1$ **then**
  $T_D \leftarrow$ maximum of the finite auxiliary score support.
else
  $T_D \leftarrow Q_G(q)$, the discrete generalized quantile in (4).
end if
return $\{C_i(T_D) = \{j \in [N] : s_i(j) \leq T_D\} : i \in \mathcal{I}_t\}$

Algorithm 2 Monte Carlo Distribution-informed Conformal Ranking (MDCR)

Input: Index sets $\mathcal{I}_c, \mathcal{I}_t$; relative calibration ranks r_i;
potential scores $s_i(j)$ from full-pool predictions; sizes n, m;
level $\alpha \in (0, 1)$.
Output: Rank prediction sets for every $i \in \mathcal{I}_t$.

$N \leftarrow n + m$, $k \leftarrow \lceil (n+1)(1-\alpha) \rceil$, $q \leftarrow k/(n+1)$
if $k = n + 1$ **then**
  $T_M \leftarrow +\infty$ {return the full discrete rank support}
else
  for each $i \in \mathcal{I}_c$ **do**
    Independently draw $P_i \sim \text{Beta}(r_i, n + 1 - r_i)$.
    Draw $H_i \mid P_i \sim \text{Binomial}(m, P_i)$; set $J_i \leftarrow r_i + H_i$.
    $Z_i \leftarrow s_i(J_i)$
  end for
  $T_M \leftarrow Z_{(k)}$, the kth order statistic of $\{Z_i : i \in \mathcal{I}_c\}$.
end if
return $\{C_i(T_M) = \{j \in [N] : s_i(j) \leq T_M\} : i \in \mathcal{I}_t\}$

Monte Carlo approximation. Draw one independent $\widetilde{J}_i \sim K_{r_i}$ per calibration item and use the kth order statistic of $s_i(\widetilde{J}_i)$. When $k = n + 1$, MDCR returns $T_M = +\infty$ and full sets. For $k \leq n$, this equals the empirical generalized quantile at q, since $\lceil nq \rceil = k$. MDCR-B is used only to diagnose approximation: draw B times per item and take order statistic $\lceil nBq \rceil$. This index is neither Bk nor a new conformal correction using nB as a calibration sample size. At $q = 1$, the $B > 1$ diagnostic uses the empirical maximum, whereas MDCR keeps its full-set branch. Our coverage comparisons use $B = 1$; the budget analysis studies larger B.

Computation. Each kernel row has $m+1$ entries. Direct evaluation of the mixture uses $O(n(m+1))$ score evaluations, with sorting or score-specific aggregation to obtain a quantile. RA permits integer-distance histogram accumulation. Algorithm 2 avoids forming a full score mixture: the Beta–Binomial draw is distributionally equivalent to sampling K_{r_i} independently for each item (Appendix B.1). MDCR is a stochastic approximation to the DCR target, not a new calibration sample size. Its measured calibration time excludes training and prediction and does not establish the runtime of a different sampler or uncached implementation.

4 STATISTICAL GUARANTEES AND THEIR LIMITS

4.1 FROM REFERENCE SCORES TO COVERAGE

The analysis must account for the difference between the auxiliary score mixture G used for calibration and the true scores that determine coverage. An exact reference rank law does not by itself ensure nominal finite-sample coverage for DCR or MDCR; Appendix C provides explicit exchangeable counterexamples. We therefore relate G to the true calibration-score empirical CDF F_c, quantify the resulting coverage error, and establish asymptotic marginal validity under explicit local conditions.

4.2 TRANSPORT AND A VALID FCP CERTIFICATE

Fix a deterministic integer $d \in \{0, \dots, m\}$ before seeing outcomes. Define

$$B_d(J) = \frac{1}{n} \sum_{r=1}^{n} \sum_{v} K_r(v) \mathbb{I}\{|v - J_r| > d\}, \tag{5}$$

$$\omega_d \geq \max_{i \in \mathcal{I}_c} \max_{|u-v| \leq d} |s_i(u) - s_i(v)|. \tag{6}$$

For RA take $\omega_d = d/N$; for VA take the maximum predicted-value gap over at most d ordered positions. VA does not automatically inherit the normalized RA rate. Let $\kappa = \kappa_{n,m}(d, \delta)$ be any deterministic number with $\Pr(B_d > \kappa) \leq \delta$. Its exact combinatorial quantile or a proved analytic upper bound is allowed; an uncorrected simulated quantile is not.

The displacement radius d is an analysis device, not a DCR tuning parameter. The defect B_d averages reference probability mass farther than d from the realized calibration ranks, with high-probability upper bound κ. The score modulus ω_d transports rank displacement into score error, while e controls the one-sided calibration–test empirical-CDF discrepancy induced by partitioning.

Proposition 4.1 (Pathwise transport)**.** *For every realization, simultaneously for all real t,*

$$G(t - \omega_d) - B_d \leq F_c(t) \leq G(t + \omega_d) + B_d. \tag{7}$$

The pathwise comparison holds simultaneously in t, so it also applies at a data-dependent threshold. Let $e = e_{n,m}(\beta)$ be a deterministic upper quantile of $\sup_t(F_c(t) - F_t(t))$ under uniform score partitioning. An exact lattice-path computation is given in Appendix A.3. A universally valid analytic choice, used in our experiments, is

$$e = \min\left\{1, \sqrt{\frac{\log(4/\beta)}{2n}} + \sqrt{\frac{\log(4/\beta)}{2m}}\right\}. \tag{8}$$

We apply the single-sample DKW bound (Massart, 1990) to both uniform groups and combine the bounds to control partition error.

Theorem 4.2 (FCP certificate)**.** *Under Assumption 3.1, for any random threshold T, with probability at least $1 - \delta - \beta$,*

$$\mathrm{FCP}(T) \leq \min\{1, 1 - G(T - \omega_d) + \kappa + e\}. \tag{9}$$

For DCR also define $\Delta_d = [q - G(T_D - \omega_d)]_+$ and $W_d(T) = F_c(T + \omega_d) - F_c(T)$. On the same event,

$$\mathrm{FCP}(T_D) \leq \min\{1, \alpha + \min(\Delta_d, W_d(T_D)) + \kappa + e\}. \tag{10}$$

The local mass $W_d(T)$ counts true calibration scores in $(T, T + \omega_d]$, excluding the atom at T. Large mass in this window can turn a small score perturbation into appreciable coverage error. Thus rank error passes through score transport and threshold-local mass to coverage error.

4.3 MARGINAL AND ASYMPTOTIC VALIDITY REQUIRE LOCAL CONTROL

Theorem 4.3 (Finite-sample marginal error bound)**.** *Suppose Assumption 3.1 holds and, for deterministic $\gamma, \rho \geq 0$, $\Pr\{W_d(T_D) > \gamma\} \leq \rho$. Put*

$$\ell = \min\{n, \max\{0, \lceil n(q - \kappa - \gamma) \rceil\}\}. \tag{11}$$

For a uniformly chosen test item I, independently chosen within the test index set,

$$\Pr\{R_I \in C_I(T_D)\} \geq \max\{0, \ell/(n+1) - \delta - \rho\}. \tag{12}$$

The bound accounts for approximation and failure terms by comparing to a fixed-index true calibration order statistic and subtracting the probabilities of unsuccessful approximation events.

For an explicit asymptotic choice, let

$$v_{n,m} = \left(\frac{m}{N}\right)^2 \left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\right)^2, \quad d_n = \min\{m, \lceil N\sqrt{2v_{n,m}\log(8n^4)}\rceil\}. \tag{13}$$

Appendix A.2 proves that $\delta_n = \kappa_n = n^{-2}$ is valid and $d_n/N = O(\sqrt{\log n/n})$ for arbitrary $m = m_n \geq 1$.

Corollary 4.4 (RA asymptotics)**.** *As $n \to \infty$, if $W_{d_n}(T_D) \to 0$ in probability, then* $\liminf \Pr\{R_I \in C_I(T_D)\} \geq 1 - \alpha$*, even if m stays bounded. If additionally $m \to \infty$, then* $(\mathrm{FCP}(T_D) - \alpha)_+ \to 0$ *in probability, using $\beta_n = \min(n, m)^{-2}$ in* (8)*.*

Local mass convergence is a substantive assumption. The empirical conclusion controls excess FCP above α on one side. For VA the same deductions require its own modulus and local mass condition. Uniform convergence of the true-score CDF to a continuous limit is one sufficient route. Appendix B proves it for normalized RA under the fixed Gaussian mechanism with $A(X) = X$ and fixed $\sigma > 0$, without a nondegenerate ratio m/n; this is not a result for arbitrary trained predictors.

Practical interpretation. The local mass W_d is an oracle diagnostic, not a deploy-time certificate, and Corollary 4.4 is a sufficient asymptotic result. For finite-sample deployment, Equation (9) is observable from calibration ranks, predictions, and prescribed bounds; it also applies to an MDCR threshold when exact G is evaluated for certification. If its raw right-hand side is at least one, the clipped bound remains valid but provides no informative certificate for that instance. Evaluating exact G for this purpose can remove MDCR's computational advantage.

4.4 Monte Carlo error and envelope comparison

Let $\mathcal{O}$ denote all inputs used to form F_i^{aux} and let $\widehat{G}_B$ be the mean CDF of nB independent, conditionally nonidentically distributed auxiliary scores.

Proposition 4.5 (CDF variance and quantile error)**.** *For fixed t, conditional on $\mathcal{O}$,*

$$\mathbb{E}\widehat{G}_B(t) = G(t), \qquad \mathit{Var}(\widehat{G}_B(t)) = \frac{\sum_i F_i^{\mathrm{aux}}(t)(1 - F_i^{\mathrm{aux}}(t))}{n^2 B}. \tag{14}$$

Conditioning further on the true score array, the MSE relative to $F_c(t)$ equals $(G(t) - F_c(t))^2$ plus the variance in (14)*. For $0 < \varepsilon < q$, the generalized empirical quantile $T_B = Q_{\widehat{G}_B}(q)$ satisfies*

$$\Pr\{G(T_B) < q - \varepsilon \mid \mathcal{O}\} \leq e^{-2nB\varepsilon^2}. \tag{15}$$

Conditional on all observed inputs, the variance is zero when every $F_i^{\mathrm{aux}}(t) \in \{0, 1\}$, and is positive if any summand is positive; fixing t alone does not imply zero variance. Unconditionally, the law of total variance adds $\mathrm{Var}(G(t))$. These are CDF-level statements; a threshold variance comparison would require additional quantile regularity. Taking $\varepsilon = \sqrt{\log(1/\zeta)/(2nB)}$ (the assertion is trivial if $\varepsilon \geq q$) gives $F_c(T_B) \geq q - \varepsilon - \kappa - W_d(T_B)$ outside failure probability $\delta + \zeta$. Thus Theorem 4.3 applies with q replaced by $q - \varepsilon$ and failure $\delta + \zeta + \rho$; the FCP bound adds β and e. MDCR's infinite-threshold branch is fully covering. Under the corresponding local condition, even $B = 1$ suffices asymptotically when $\log(1/\zeta)/(nB) \to 0$. This conclusion does not require threshold consistency.

Proposition 4.6 (Conditional deterministic TCPR comparison)**.** *For realized rank envelopes E_i, suppose $\sum_{j \in E_i} K_{r_i}(j) \geq 1 - \delta_0$ for every calibration item. Let $v_i = \max_{j \in E_i} s_i(j)$, $k_T = \lceil (n+1)(1 - \alpha + \delta_0)\rceil$, and $T_T = v_{(k_T)}$ if $k_T \leq n$, otherwise $T_T = +\infty$. Then $T_D \leq T_T$ and $C_i(T_D) \subseteq C_i(T_T)$.*

This requires realized reference-kernel mass, not merely unconditional coverage of the envelopes. It is a set inclusion comparison, not equal coverage or a transferred validity guarantee. The comparison applies only when the envelope satisfies the stated kernel-mass condition.

Table 1: Main results at target coverage 0.90. Synthetic uses 500 complete trials with fixed $A(X) = X$ at $n = m = 500$ and $\sigma = 1$; ESOL and Yummly average 20 shared splits with learned ranking predictors. Bold jointly marks the non-Oracle method whose unrounded mean coverage is closest to 0.90; exact ties are broken by smaller Size/N. † denotes mean coverage below target. Oracle is shown only as a true-rank reference.

		Oracle (ref.)		TCPR		DCR (ours)		MDCR (ours)	
Score	Predictor	Cov.	Size/N	Cov.	Size/N	Cov.	Size/N	Cov.	Size/N
Panel A: Rank-adaptive (RA)									
Synthetic: $n = m = 500$, $\sigma = 1$									
RA	Fixed $A(X) = X$	0.9003	0.6251	0.9391	0.6938	**0.9001**	**0.6247**	0.9008	0.6260
ESOL									
	RankNet	0.8988†	0.3862	0.9687	0.5269	**0.9008**	**0.3889**	0.9025	0.3902
	LambdaMART	0.9042	0.3860	0.9653	0.5318	**0.9082**	**0.3926**	0.9107	0.3983
	RankSVM	0.9007	0.4950	0.9640	0.6271	**0.9022**	**0.4983**	0.9037	0.5018
Yummly									
	RankNet	0.8971†	0.0729	0.9916	0.2170	0.9560	0.1013	**0.9539**	**0.0998**
	LambdaMART	0.9021	0.2756	0.9645	0.3911	**0.9092**	**0.2833**	0.9107	0.2852
	RankSVM	0.8956†	0.2360	0.9691	0.3580	**0.9032**	**0.2431**	0.9042	0.2443
Panel B: Value-adaptive (VA)									
Synthetic: $n = m = 500$, $\sigma = 1$									
VA	Fixed $A(X) = X$	0.9000	0.6287	0.9458	0.7055	**0.9002**	**0.6289**	0.9008	0.6299
ESOL									
	RankNet	0.9038	0.3996	0.9817	0.5896	**0.9097**	**0.4088**	0.9103	0.4112
	LambdaMART	0.9073	0.4048	0.9695	0.5661	**0.9105**	**0.4097**	0.9105	0.4126
	RankSVM	0.8988†	0.5200	0.9635	0.6560	0.8973†	0.5183	**0.9002**	**0.5256**
Yummly									
	RankNet	0.8994†	0.0751	0.9907	0.2156	**0.9535**	**0.1019**	0.9541	0.1028
	LambdaMART	0.9016	0.2811	0.9656	0.4127	**0.9106**	**0.2908**	0.9110	0.2916
	RankSVM	0.8980†	0.2275	0.9788	0.3740	0.9128	0.2385	**0.9109**	**0.2378**

5 EXPERIMENTS

We study a controlled Gaussian mechanism, ESOL molecular solubility (Delaney, 2004), and fixed-anchor Yummly28K ingredient similarity (Min et al., 2017). The synthetic task uses the fixed predictor $A(X) = X$; the real-data tasks use RankNet, LambdaMART, and linear RankSVM over 20 shared partitions. Ridge and ExtraTrees serve as regression controls in Appendix E.4. The evaluated TCPR baseline implements the published marginal rule with an independently designed and validated numerical quantile envelope, exact union-bound fallback, and joint-failure certification at $\delta = 0.005$ (Appendix D.2).

DCR and MDCR substantially reduce the empirical conservatism of the evaluated TCPR envelope. Table 1 reports both coverage and normalized set size at target coverage 0.90. In the fixed Gaussian setting, DCR–RA attains $0.9001/0.6247$ coverage/size, compared with $0.9391/0.6938$ for TCPR. On ESOL/RankNet–RA, the corresponding values are $0.9008/0.3889$ and $0.9687/0.5269$. The learned-ranker rows also retain unfavorable means: ESOL/RankSVM–VA DCR coverage is 0.8973. These comparisons measure conservatism under distinct calibration targets, rather than efficiency under an identical finite-sample guarantee.

More calibration information reduces conservatism in the controlled synthetic setting. With $Y = X + \sigma\epsilon$, fixed $A(X) = X$, $m = 500$, and $\sigma = 0.1$, increasing n from 100 to 10,000 moves DCR coverage from 99.31% to 90.02% and size/N from 0.1915 to 0.0994 (Figure 1). Small calibration sets leave more reference-rank uncertainty and hence wider sets. The $\sigma = 1$ panels show that the rate depends on score geometry; changing the test-pool size also changes the ranking problem itself. Appendix D.3 retains the test-size and noise scans, heatmaps, and all unfavorable cells.

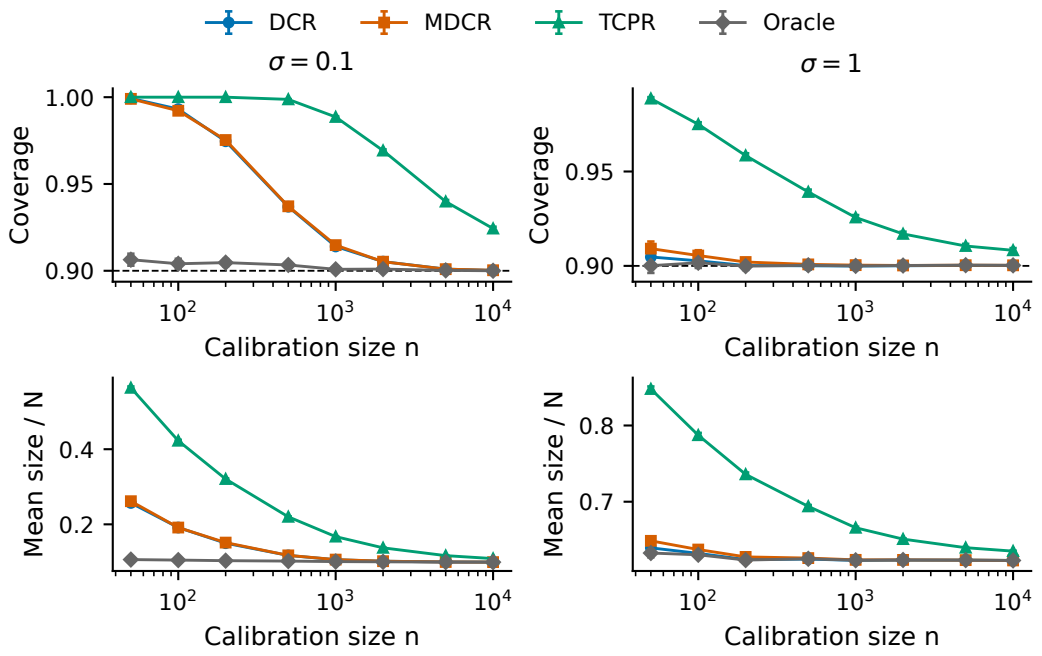

Figure 1: Calibration-size behavior in the fixed Gaussian mechanism, with $m = 500$, RA, $\alpha = 0.1$, and 500 complete trials per point. All eight prespecified calibration sizes are shown for $\sigma = 0.1$ and 1; bars are 95% trial-level intervals.

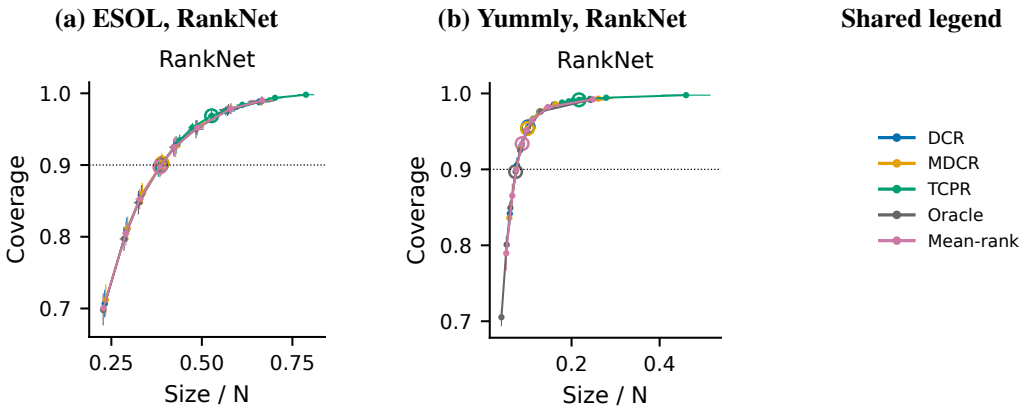

Figure 2: RA coverage–size trade-offs for RankNet on ESOL and Yummly over eight prespecified nominal levels. Curves show DCR, MDCR, TCPR, Oracle, and Mean-rank; hollow circles mark $\alpha = 0.1$. Bars are split-level 95% intervals over 20 shared partitions. Appendix E.5 reports every predictor and both scores.

The full reference distribution is useful but not uniformly better than a point estimate. On Yummly/RankNet–RA, Mean-rank changes coverage/size from DCR's $0.9560/0.1013$ to $0.9339/0.0878$; its paired coverage difference is -0.0221 ± 0.0033. On ESOL/Ridge–VA, however, it moves coverage from 0.8955 to 0.8905, farther below target. Figure 2 places this ablation in the complete eight-level RankNet trade-off. The full reference distribution represents latent-rank dispersion and defines an auxiliary mixture G whose discrepancy from the true calibration-score CDF and Monte Carlo approximation error can be analyzed. It is a principled calibration object, not an empirical claim that a distribution always beats a point estimate; all 20 paired comparisons appear in Appendix E.3.

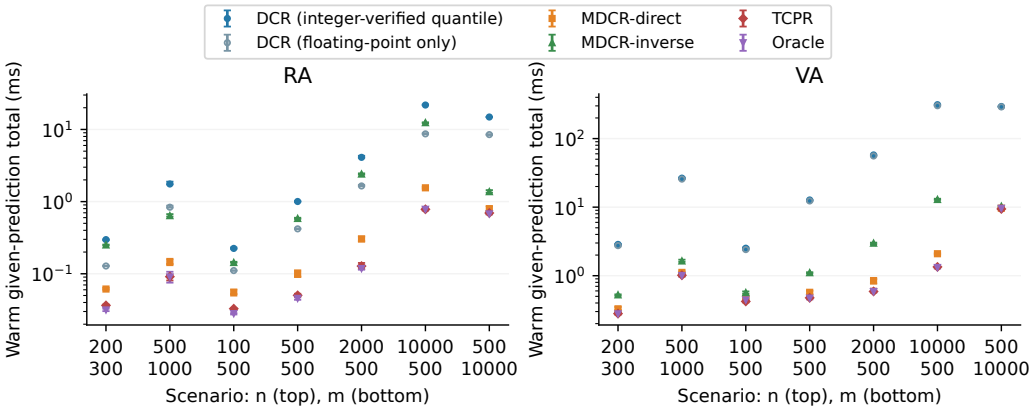

Figure 3: Warm given-prediction costs: medians and IQR over three pools and ten repetitions. RA/VA include sorting, calibration, and identical set output. Filled DCR markers use integer-verified quantiles; hollow markers are the floating-point-only reference. Integer verification certifies the mixture-quantile calculation, not statistical coverage. Categories show (n, m); Appendix E.6 gives every component and cold setup.

Direct auxiliary sampling makes MDCR scalable without constructing the dense reference mixture. At $(n, m) = (10{,}000, 500)$, median warm RA costs are $1.55/12.41/21.80$ ms for direct MDCR, inverse-CDF MDCR, and integer-verified DCR; at $(500, 10{,}000)$ they are $0.78/1.37/14.87$ ms. These implementation-specific costs use a common output representation (Figure 3); Appendix E.6 separates components and cold setup.

6 Conclusion

Latent absolute ranks under partial rank observability are structured rather than merely missing. DCR uses the uniform-interleaving reference structure to calibrate a mixture of potential scores, while MDCR samples the same target without constructing the dense mixture. The analysis quantifies the gap between this reference mixture and the true calibration scores instead of assuming automatic exact validity.

Across the controlled Gaussian and learned-ranker experiments, DCR and MDCR are substantially less conservative than the evaluated TCPR envelope. The gain depends on the predictor and score, and Mean-rank confirms that the full reference distribution is not uniformly preferable in finite samples. Direct sampling reduces calibration cost when dense mixture construction dominates.

Viewed together, the method and evidence separate three questions that a rank-uncertainty procedure must address: how the observed calibration order induces a reference law over latent absolute ranks, how the chosen score maps those ranks into candidate sets, and how finite calibration and Monte Carlo approximation affect the selected threshold. Keeping these roles explicit clarifies why a less conservative set may still under-cover in a particular predictor–score configuration, why Mean-rank can occasionally be competitive, and why direct sampling changes computational cost without strengthening the statistical guarantee.

Limitations. Finite-sample validity depends on the stated assumptions and bounds; exchangeability alone is insufficient. The observable certificate can also be loose: 88.8% of raw bounds in the synthetic certificate study are at least one before clipping (Appendix F.6), limiting their informativeness without equating empirical coverage to the bound. Random-split ESOL and fixed-anchor Yummly do not establish scaffold extrapolation, observed recipe preferences, or broad ranking-task validity. Native multi-query learning-to-rank benchmarks and tasks with observed preference or relevance judgments remain important future work. Runtime ordering is implementation-dependent.

REFERENCES

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Anastasios N Angelopoulos, Karl Krauth, Stephen Bates, Yixin Wang, and Michael I Jordan. Recommendation systems with distribution-free reliability guarantees. In *Conformal and Probabilistic Prediction with Applications*, pp. 175–193, 2023.

Stephen Bates, Anastasios N. Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, pp. 89–96, 2005.

Christopher J. C. Burges. From ranknet to lambdarank to lambdamart: An overview. Technical Report MSR-TR-2010-82, Microsoft Research, 2010.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: From pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning*, pp. 129–136, 2007.

Pinhan Chen, Chao Gao, and Anderson Y Zhang. Optimal full ranking from pairwise comparisons. *The Annals of Statistics*, 50(3):1775–1805, 2022.

John S Delaney. Esol: Estimating aqueous solubility directly from molecular structure. *Journal of Chemical Information and Computer Sciences*, 44(3):1000–1005, 2004.

Jean-Baptiste Fermanian, Pierre Humbert, and Gilles Blanchard. Transductive conformal inference for full ranking. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=3Bxn5XSgrO.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Jianguo Huang, Huajun Xi, Linjun Zhang, Huaxiu Yao, Yue Qiu, and Hongxin Wei. Conformal prediction for deep classifier via label ranking. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20331–20347. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/huang24aa.html.

Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 133–142, 2002.

Alexandros Karatzoglou, Linas Baltrunas, and Yue Shi. Learning to rank for recommender systems. In *Proceedings of the 7th ACM Conference on Recommender Systems*, pp. 493–494, 2013.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018.

Shuo Li, Sangdon Park, Insup Lee, and Osbert Bastani. Traq: Trustworthy retrieval augmented question answering via conformal prediction. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 3799–3821, 2024.

Ziyi Liang, Tianmin Xie, Xin Tong, and Matteo Sesia. Structured conformal inference for matrix completion with applications to group recommender systems. *arXiv preprint arXiv:2404.17561*, 2024.

Tie-Yan Liu et al. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331, 2009.

P. Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The Annals of Probability*, 18(3), July 1990. ISSN 0091-1798. doi: 10.1214/aop/1176990746. URL http://dx.doi.org/10.1214/aop/1176990746.

Weiqing Min, Shuqiang Jiang, Jitao Sang, Huayang Wang, Xinda Liu, and Luis Herranz. Being a supercook: Joint food attributes and multimodal content modeling for recipe retrieval and exploration. *IEEE Transactions on Multimedia*, 19(5):1100–1113, May 2017. ISSN 1941-0077. doi: 10.1109/tmm.2016.2639382. URL http://dx.doi.org/10.1109/TMM.2016.2639382.

Tao Qin, Xiubo Geng, and Tie-Yan Liu. A new probabilistic model for rank aggregation. In *Advances in Neural Information Processing Systems*, volume 23, 2010.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in Neural Information Processing Systems*, 32, 2019.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

Nihar B Shah and Martin J Wainwright. Simple, robust and optimal ranking from pairwise comparisons. *Journal of Machine Learning Research*, 18(199):1–38, 2018.

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

AI USE STATEMENT

Generative AI tools were used to assist with manuscript polishing and editing, to check mathematical derivations and proof steps, and to help write and debug experimental code. All AI-assisted theoretical arguments, code, and manuscript changes were reviewed and verified by the authors. The authors made the final research decisions and take full responsibility for all claims, proofs, experiments, code, and text in this work.

A PROOFS

The following proofs give the dependence on the score-array symmetry assumption, the transport radius, and the local score mass. No local regularity condition is inferred solely from the experiments.

A.1 REFERENCE LAW AND MOMENTS

Proof of Proposition 3.2. There are $\binom{N}{n}$ equally likely calibration subsets. A subset whose rth element is j chooses $r-1$ elements from $[j-1]$ and $n-r$ from $\{j+1,\dots,N\}$. This proves (2), its support, and its normalization. Substituting $j = r + x$ into the beta-binomial mass yields the same expression. This conditioning refers to uniform rank interleaving; observing features that predict the true ranks can change the conditional law of a particular item. □

For completeness, the beta-binomial representation is $U \sim \mathrm{Beta}(r, n+1-r)$ and $J_r = r + Z$ with $Z \mid U \sim \mathrm{Binomial}(m, U)$. Using $\mathbb{E}U = r/(n+1)$ and $\mathrm{Var}(U) = r(n+1-r)/((n+1)^2(n+2))$ gives

$$\mathbb{E}J_r = \frac{r(N+1)}{n+1}, \tag{16}$$

$$\mathrm{Var}(J_r) = \frac{r(n+1-r)m(N+1)}{(n+1)^2(n+2)}. \tag{17}$$

Since $n^{-1}\sum_{r=1}^{n} r(n+1-r) = (n+1)(n+2)/6$, the average variance is $m(N+1)/(6(n+1))$.

A.2 PATHWISE TRANSPORT AND COMPUTABLE DEFECT BOUNDS

Proof of Proposition 4.1. For the calibration item of relative rank r, couple its realized score $s_i(J_r)$ to an auxiliary position v. On $|v - J_r| \le d$, the score difference is at most ω_d. Hence

$$\mathbb{I}\{s_i(v) \le t - \omega_d\} \le \mathbb{I}\{s_i(J_r) \le t\} + \mathbb{I}\{|v - J_r| > d\}.$$

Conversely $\mathbb{I}\{s_i(J_r) \le t\} \le \mathbb{I}\{s_i(v) \le t + \omega_d\} + \mathbb{I}\{|v - J_r| > d\}$. Integrating over $K_r(v)$ and averaging proves both inequalities simultaneously for all t. No probabilistic independence between scores and ranks is used in this deterministic step. □

The defect has a universal distribution under Assumption 3.1. For $\delta \in (0,1)$ one may use its upper quantile, defined by

$$\kappa_{n,m}(d,\delta) = \inf\{b \in [0,1] : \Pr(B_d \le b) \ge 1 - \delta\}.$$

Enumerating uniform subsets gives this quantile exactly for small problems. Its mean is the finite sum

$$\mu_{n,m}(d) = \mathbb{E}B_d = \frac{1}{n}\sum_{r=1}^{n}\sum_{u,v} K_r(u)K_r(v)\mathbb{I}\{|u-v| > d\}. \tag{18}$$

Indeed only the marginal law of J_r enters this expectation; the J_r need not be mutually independent. Markov's inequality permits $\kappa = \min\{1, \mu_{n,m}(d)/\delta\}$. If U_r, V_r are independent copies of J_r, then $\mathbb{E}(U_r - V_r)^2 = 2\mathrm{Var}(J_r)$. Their integer separation on the tail event is at least $d+1$, so

$$\mu_{n,m}(d) \le \min\left\{1, \frac{m(N+1)}{3(n+1)(d+1)^2}\right\}, \qquad B_m = 0. \tag{19}$$

We also prove the exponential bound used for the certificate analysis. Generate two independent groups of n and m iid uniform variables. Their pooled order gives a uniform calibration subset. Write H_n, H_m for their empirical CDFs, and let $U_{(r)}$ be the rth uniform in the calibration group. Almost surely there are no ties, and

$$J_r = r + mH_m(U_{(r)}), \qquad H_n(U_{(r)}) = r/n.$$

Thus

$$\max_r |J_r/N - r/n| \le \frac{m}{N}\{\|H_n - I\|_\infty + \|H_m - I\|_\infty\}. \tag{20}$$

The single-sample DKW inequality states $\Pr(\|H_s - I\|_\infty > x) \le 2e^{-2sx^2}$ (Massart, 1990). Allocate the two deviations in proportion to $1/\sqrt{n}$ and $1/\sqrt{m}$ and use a union bound to get

$$\Pr\{\max_r |J_r/N - r/n| > a\} \le 4\exp(-2a^2/v_{n,m}).$$

For an independent copy J' of the whole rank vector, $|J_r - J'_r|/N > h$ implies that one copy deviates from r/n by more than $h/2$. Applying the preceding inequality to both copies, then averaging over r, proves

$$\mu_{n,m}(d) \le \min\{1, 8\exp(-(d/N)^2/(2v_{n,m}))\}. \tag{21}$$

For $d = d_n$ in (13), if the cap $d_n = m$ is active the defect is zero. Otherwise (21) gives $\mu \le n^{-4}$, so Markov gives $\Pr(B_{d_n} > n^{-2}) \le n^{-2}$. Finally,

$$\sqrt{v_{n,m}} = \frac{m}{N}\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\right) \le \frac{1}{\sqrt{n}} + \frac{1}{2\sqrt{n}},$$

where $\sqrt{m}/(n+m) \le 1/(2\sqrt{n})$. Including the ceiling error $1/N \le 1/n$ proves the asserted normalized radius rate, uniformly over $m \ge 1$.

A.3 Uniform calibration–test partition bounds

Conditional on the unordered true-score multiset, random tie breaking orders score indices without favoring calibration labels. The calibration/test label path is a uniform path with n calibration and m test steps. At any prefix with counts (a, b), the CDF difference is $a/n - b/m$. Thresholds with score ties visit only a subset of these prefixes, so their maximum is no larger.

For a proposed one-sided error e, let $H_e(a, b)$ count paths that have never crossed $a/n - b/m > e$. Set $H_e(0, 0) = 1$, use zero for out-of-grid predecessors, and, for other states, compute

$$H_e(a, b) = \mathbb{I}\{a/n - b/m \le e\}\{H_e(a-1, b) + H_e(a, b-1)\}. \tag{22}$$

The probability of a permitted path is $H_e(n, m)/\binom{N}{n}$. Choosing the smallest lattice value with this probability at least $1 - \beta$ gives a valid exact e. This is an integer counting algorithm, not an uncorrected Monte Carlo estimate.

Alternatively, the same random path is induced by two independent iid uniform groups. Their CDF difference is bounded by the sum of their deviations from the uniform CDF. Two applications of the single-sample DKW bound, each with failure probability $\beta/2$, give (8). This representation proves the bound even though the actual pooled true scores may be dependent. It does not assume iid true scores or apply an iid two-sample theorem to them directly.

Proof of Theorem 4.2. On the joint event $B_d \le \kappa$ and $\sup_t(F_c(t) - F_t(t)) \le e$, whose probability is at least $1 - \delta - \beta$, (7) implies $F_t(T) \ge G(T - \omega_d) - \kappa - e$ for every threshold, including a threshold selected from the observations or auxiliary draws. Clipping FCP by one gives (9). For DCR, $G(T_D) \ge q$. Evaluating the lower transport inequality at $T_D + \omega_d$ gives $G(T_D) \le F_c(T_D + \omega_d) + B_d$. Therefore

$$F_c(T_D) \ge q - B_d - W_d(T_D). \tag{23}$$

The lower transport inequality also gives $F_c(T_D) \ge q - \Delta_d - B_d$. Taking the stronger of these two statements and using $q \ge 1 - \alpha$ yields (10). □

A.4 Marginal rank argument and asymptotics

Proof of Theorem 4.3. Outside an event of probability at most $\delta + \rho$, (23) gives $nF_c(T_D) \ge n(q - \kappa - \gamma)$. Since the left side is integer, at least ℓ true calibration scores are at most T_D. If $\ell \ge 1$, then $T_D \ge S^c_{(\ell)}$, where this is the fixed ℓth true calibration order statistic. For $\ell = 0$ the claimed lower bound is zero. The score of a uniformly chosen test item and the n calibration scores form an exchangeable subarray by Assumption 3.1. Ranking these $n + 1$ scores with independent tie breaking proves $\Pr\{S_I \le S^c_{(\ell)}\} \ge \ell/(n+1)$. Ties can only increase the weak-inequality event. This assertion is unconditional; it is not conditioned on the good approximation event. Subtracting the latter's failure probability by a union bound proves (12). □

Proof of Corollary 4.4. If $W_{d_n}(T_D) \to 0$ in probability, there exist deterministic $\gamma_n \downarrow 0$ such that $\rho_n = \Pr(W_{d_n}(T_D) > \gamma_n) \to 0$. Section A.2 permits $\delta_n = \kappa_n = n^{-2}$. Since $q_n \to 1 - \alpha$, the fixed index satisfies $\ell_n/(n+1) \to 1 - \alpha$. Apply Theorem 4.3; no condition on divergence of m is needed for this marginal argument. If $m \to \infty$ too, choose $\beta_n = \min(n, m)^{-2}$ for all sufficiently large n. Equation (8) then tends to zero, because $\log \min(n, m)/\min(n, m) \to 0$. Apply (10) with $W_{d_n}(T_D) \to 0$. This bounds positive excess FCP in probability; it does not bound conservatism away from the target. □

A.5 Monte Carlo identities, atoms, and coverage transfer

Proof of Proposition 4.5. At fixed t, the nB indicators are conditionally independent Bernoulli variables with probabilities $F_i^{\text{aux}}(t)$ repeated B times. Summing their means and variances proves (14). Conditional on the full true array, $F_c(t)$ is fixed and the auxiliary draws still have these distributions, so the bias–variance decomposition gives the MSE statement. The unconditional identity follows from total variance.

For the quantile bound, let $a = Q_G(q - \varepsilon)$. If $G(T_B) < q - \varepsilon$, monotonicity implies $T_B < a$. The empirical quantile has at least a fraction q of its observations at or below T_B, so the fraction strictly below a is at least q. Its conditional mean is $G(a-) \le q - \varepsilon$. A one-sided Hoeffding bound (Hoeffding, 1963) for independent bounded indicators gives $e^{-2nB\varepsilon^2}$. Strict inequality below a handles atoms correctly. For clarity, the bound itself follows by the Bernoulli centered moment-generating bound $\mathbb{E}e^{\lambda(Z-\mathbb{E}Z)} \le e^{\lambda^2/8}$, multiplying over nB independent terms and minimizing $\exp(-\lambda nB\varepsilon + nB\lambda^2/8)$ at $\lambda = 4\varepsilon$. □

On the quantile event $G(T_B) \ge q - \varepsilon$, the same pathwise proof as (23) yields $F_c(T_B) \ge q - \varepsilon - B_d - W_d(T_B)$. Consequently, if $\Pr(W_d(T_B) > \gamma) \le \rho$, define

$$\ell_B = \min\{n, \max\{0, \lceil n(q - \varepsilon - \kappa - \gamma)\rceil\}\}.$$

Then

$$\Pr\{R_I \in C_I(T_B)\} \ge \max\{0, \ell_B/(n+1) - \delta - \zeta - \rho\}, \tag{24}$$

$$\text{FCP}(T_B) \le \min\{1, \alpha + \varepsilon + \kappa + W_d(T_B) + e\} \tag{25}$$

on an event of probability at least $1 - \delta - \zeta - \beta$ for the second statement. When $q = 1$ and MDCR returns infinity, coverage is one directly. For other MDCR thresholds, use $B = 1$. Choosing $\zeta_n \to 0$ with $\log(1/\zeta_n)/(nB) \to 0$ makes $\varepsilon_n \to 0$. The proof of Corollary 4.4 then applies with the local condition at MDCR's own threshold. The optional $B > 1$ construction is an approximation diagnostic and carries no borrowed sample-size correction.

A.6 Envelope mass and rounding

Proof of Proposition 4.6. More generally, if $K_{r_i}(E_i) \ge 1 - \delta_i$, then for any u,

$$G(u) \ge \frac{1}{n} \sum_{i: v_i \le u} (1 - \delta_i),$$

since all ranks in E_i have scores at most v_i. Suppose the common δ_0 condition holds and $k_T \le n$. Let $h = k_T - k \ge 0$. From the two ceiling definitions, $(n+1)\delta_0 < h + 1$. Therefore

$$G(v_{(k_T)}) \ge \frac{k_T}{n}(1 - \delta_0) > \frac{k}{n+1} + \frac{h(n - k_T)}{n(n+1)} \ge q.$$

The generalized quantile of G is at most $v_{(k_T)}$. If $k_T > n$, the full-set branch makes the comparison immediate. Set inclusion follows because both methods use the same potential scores. Nothing in this argument proves DCR validity from TCPR validity, and envelopes that merely contain true ranks with high unconditional probability need not satisfy the realized kernel-mass premise. □

B SUFFICIENT LOCAL CONDITIONS AND A FIXED GAUSSIAN EXAMPLE

The local condition in Corollary 4.4 is weaker than uniform approximation by a continuous score distribution. The following lemma gives one sufficient route, including data-dependent thresholds; it is not a necessary condition.

Lemma B.1 (Uniform CDF approximation controls local windows). *Let $F_{c,n}$ be the empirical CDF of the true calibration scores. Suppose that F is a continuous CDF, $\|F_{c,n} - F\|_\infty \to 0$ in probability, and $\omega_n \geq 0$ converges to zero in probability. For any possibly data- and auxiliary-randomness-dependent threshold T_n, the window mass*

$$W_n = F_{c,n}(T_n + \omega_n) - F_{c,n}(T_n)$$

converges to zero in probability. For $T_n = +\infty$ the window mass is defined as zero by the CDF limits.

Proof. For every realization and finite T_n,

$$0 \leq W_n \leq 2\|F_{c,n} - F\|_\infty + w_F(\omega_n), \qquad w_F(h) = \sup_x \{F(x+h) - F(x)\}.$$

A continuous CDF is uniformly continuous on $\mathbb{R}$: outside a sufficiently large bounded interval its two tails have arbitrarily small total mass, and on an enlarged compact interval uniform continuity follows from continuity. Thus $w_F(h) \to 0$ as $h \downarrow 0$. Convergence of ω_n in probability implies $w_F(\omega_n) \to 0$ in probability. The bound eliminates T_n uniformly, so no threshold-consistency assumption or independence of the threshold is needed. The infinite-threshold case is immediate. □

Atoms do not by themselves invalidate the weaker local condition. If all true scores are zero and the threshold is nonnegative, the open-at-left window $(T_n, T_n + \omega_n]$ has zero mass.

Proposition B.2 (Fixed Gaussian mechanism, normalized RA). *Let (X_i, E_i) be iid with independent standard-normal coordinates, let $Y_i = X_i + \sigma E_i$ for a fixed $\sigma > 0$, and use the fixed predictor $A(X_i) = X_i$. Calibration comprises the first n items or an independently chosen uniform n-subset of the pool. As $n \to \infty$, for any $m = m_n \geq 1$, the local score-window condition holds for DCR and MDCR with normalized RA and the radius d_n in* (13)*. Consequently each method has asymptotic marginal coverage at least $1 - \alpha$. If also $m \to \infty$, its positive excess empirical FCP above α converges to zero in probability.*

Proof. Write H_N^X, H_N^Y for the full-pool empirical CDFs, with $N = n + m$. With probability one there are no ties, and

$$S_{i,N} = |H_N^Y(Y_i) - H_N^X(X_i)|, \quad Z_i = \left|\Phi\left(Y_i/\sqrt{1+\sigma^2}\right) - \Phi(X_i)\right|.$$

The triangle inequality gives, simultaneously for every item,

$$\max_i |S_{i,N} - Z_i| \leq \Delta_N := \|H_N^Y - F_Y\|_\infty + \|H_N^X - \Phi\|_\infty \xrightarrow{p} 0.$$

Each marginal empirical CDF converges uniformly, for example by the single-sample DKW bound (Massart, 1990). This argument does not require independence of X_i and Y_i.

The pair $(X_i, Y_i/\sqrt{1+\sigma^2})$ is a nondegenerate bivariate normal with correlation $(1+\sigma^2)^{-1/2} < 1$. Applying Φ coordinatewise yields a density on $(0,1)^2$. For every t, the set $\{(u,v) : |v-u| = t\}$ has zero two-dimensional Lebesgue measure. The transformed coordinates have no boundary mass either, so Z_i has no atoms and its CDF F_σ is continuous.

The ideal scores at the calibration indices are iid; an independent uniform subset has the same joint law as the first n indices. Their empirical CDF $H_{c,n}^Z$ therefore converges uniformly to F_σ. The simultaneous score bound implies

$$H_{c,n}^Z(t - \Delta_N) \leq F_{c,n}(t) \leq H_{c,n}^Z(t + \Delta_N),$$

whence

$$\|F_{c,n} - F_\sigma\|_\infty \leq \|H_{c,n}^Z - F_\sigma\|_\infty + w_{F_\sigma}(\Delta_N) \xrightarrow{p} 0.$$

The already established bound $d_n/N \to 0$ holds without a nondegenerate limit for m/n. Normalized RA has $\omega_n = d_n/N$, so Lemma B.1 applies to either selected threshold. Corollary 4.4 gives the DCR conclusion. For MDCR, Proposition 4.5 gives a vanishing auxiliary error at $B = 1$, by choosing, for example, $\epsilon_n = \sqrt{\log n/n}$ and $\zeta_n = n^{-1}$; its transferred local-window bound gives the same limiting marginal conclusion. When $m \to \infty$, the partition error also vanishes using $\beta_n = \min(n, m)^{-2}$, yielding the one-sided empirical FCP conclusion for both methods. □

For $\sigma = 0$, $Y = X$ and all true RA scores are exactly zero. Both methods have nonnegative or infinite thresholds, so coverage is one and every local window has zero mass. This degenerate case does not use the continuous-CDF lemma. The proposition concerns a fixed σ and the fixed predictor $A(X) = X$; it is not uniform over sequences $\sigma_n \to 0$, is not a theorem for arbitrary learned rankers, and does not automatically extend to VA. It asserts neither exact finite-sample coverage nor convergence of FCP to α.

B.1 Independent direct sampling of reference ranks

For calibration relative rank r, draw $P \sim \mathrm{Beta}(r, n+1-r)$ and $H \mid P \sim \mathrm{Binomial}(m, P)$, independently across items and auxiliary repetitions, and put $J = r + H$. Integrating out P, for $h = 0, \ldots, m$,

$$\Pr(H = h) = \binom{m}{h} \frac{\mathrm{B}(r+h, n+1-r+m-h)}{\mathrm{B}(r, n+1-r)} = \frac{\binom{r+h-1}{r-1}\binom{N-r-h}{n-r}}{\binom{N}{n}}.$$

Thus this sampler has exactly the reference law in Proposition 3.2; it changes the implementation, not the statistical definition of MDCR. Its support is $[r, r+m]$, and

$$\mathbb{E}J = \frac{r(N+1)}{n+1}, \qquad \mathrm{Var}(J) = \frac{mr(n+1-r)(N+1)}{(n+1)^2(n+2)}.$$

A separate Beta and Binomial draw for each item is essential. Sharing a Beta variable would change the joint auxiliary law. The direct path stores $O(n)$ random values for $B = 1$ and does not construct an $n \times (m+1)$ rank-CDF array. It is distributionally equivalent to inverse-CDF sampling, not identical under a shared pseudorandom seed. Analytic PMF identities, boundary and moment checks, family-wise bounded empirical CDF diagnostics, and exhaustive small score/set checks are kept separate from this derivation.

C Finite-sample validity boundaries

Items are indexed by true ranks. Fix a predicted rank vector, randomly relabel item identities, and choose a uniform calibration subset. This construction satisfies Assumption 3.1. The vectors are $(1, 2, 3, 6, 5, 4, 7, 8, 9, 10)$ for the first row of Table 2 and $(3, 2, 1)$ for the second. Coverage averages over the random calibration/test partition and a uniformly chosen test item and, for MDCR, all auxiliary draws with their exact rational probabilities. Enumeration of subsets and Cartesian products of auxiliary ranks gives the displayed fractions; these are finite counterexamples, not Monte Carlo estimates.

Table 2: Exact marginal RA coverage under exchangeable rank constructions. Here $N = n + m$, with n calibration items and m test items.

N	n	Nominal coverage $(1-\alpha)$	Exact marginal coverage DCR	MDCR
10	9	$9/10$	$4/5$	$91/100$
3	1	$1/2$	$1/3$	$4/9$

Changing only DCR's quantile denominator does not restore general validity. For $N = 100, n = 99, m = 1, \alpha = 0.1$, start with identity predicted ranks and replace each block starting at $a \in \{45, 49, 53\}$ by $(a+2, a+3, a, a+1)$. Selecting the $90/99$ mixture quantile gives coverage $22/25 < 9/10$ after exact enumeration over the held-out item. This denominator variant is not used in our experiments.

D EXPERIMENTAL PROTOCOL

D.1 DATA AND PREDICTORS

Ridge and ExtraTrees use training-only feature fitting. ESOL contains 1,128 molecules with measured log solubility in mols per litre as the target. We download the processed CSV from the public source used by the DeepChem ESOL loader. Its six predictor columns are Minimum Degree, Molecular Weight, Number of H-Bond Donors, Number of Rings, Number of Rotatable Bonds, and Polar Surface Area. We exclude measured solubility, ESOL-predicted solubility, compound identifiers, and SMILES from the predictor inputs. The downloaded table has no missing entries or duplicated rows, compound IDs, or SMILES. Each random partition uses 628 training, 200 calibration, and 300 test molecules. This evaluates random partitioning of a fixed dataset, not molecular scaffold extrapolation.

The Yummly28K metadata (Min et al., 2017) contain 27,638 ingredient records. We keep the first occurrence of each recipe ID and then of each normalized ingredient text, obtaining 27,244 records. Record 00001 is a fixed anchor, excluded from all training and evaluation, leaving 27,243 usable recipes. We select 5,000 training, 500 calibration, and 1,000 test recipes per split. A TF–IDF representation is fitted only on training ingredient text, using at most 101 alphabetic-token features, English stop-word removal, minimum document frequency two, sublinear term frequency, and ℓ_2 normalization. The target is negative Euclidean distance to the excluded anchor in this training-fitted space. Higher target values denote greater ingredient similarity. The anchor vector and evaluation distances are not predictor inputs. This is a constructed similarity-ranking task based on real ingredients, not measured user preferences; its feature space and hence target distances depend on the training split.

Both datasets use Ridge with penalty 1 and ExtraTrees with 128 trees, minimum leaf size 5, all features available at each split, and one CPU worker. Standardization is fitted on training data only; there is no test-based hyperparameter selection. Regression training uses continuous target labels. Calibration receives only the relative ordering of calibration outcomes; test outcomes and full calibration ranks are restricted to evaluation and Oracle. Independent continuous tie breakers define randomized ranks of tied outcomes and predictions. Each model is trained once per outer split, then cached predictions are reused for both scores, all four calibration methods, and all eight α values. The 20 outer partitions are independently randomized draws from a fixed dataset, not 20 new population samples.

D.2 TCPR NUMERICAL QUANTILE ENVELOPE

Independent joint validation controls numerical envelope construction error. We implement the marginal-coverage TCPR rule in Theorem 3.5 of Fermanian et al. (2025), with the quantile-envelope family of their Section 4.2. The reference rank-vector generator samples uniform n-subsets of $[N]$. For a tail parameter a, the endpoints are $L_r = Q_{K_r}(a)$ and $U_r = Q_{K_r}(1-a)$, restricted to $[r, r+m]$. We use exact marginal kernel CDFs in this family rather than empirical marginal quantiles. On 10,000 independent design vectors, 24 bisection steps choose the largest candidate tail parameter whose observed joint failure is at most $\delta/2$. This is a numerical quantile-envelope adaptation; it is not the theoretical or linear envelope and does not use test outcomes.

We then evaluate the selected envelope on 50,000 new, independent reference vectors. If b vectors violate any endpoint, let u be the one-sided Clopper–Pearson upper bound for joint failure, at confidence $1-\eta$, with $\eta = 10^{-6}$. We accept only if $u \le \delta - \eta$, where $\delta = 0.005$. Otherwise, we use marginal tail level $(\delta-\eta)/(2n)$, which yields joint failure at most $\delta-\eta$ by a union bound. To see the role of validation, conditional on the design sample, the validation failure count is binomial. Except on an event of probability at most η, acceptance implies true joint failure at most $\delta - \eta$; rejection uses the union-bound fallback. Averaging over construction therefore bounds total failure by δ. This statement includes envelope-construction randomness; a successful simulation alone is not an exact certificate for a fixed envelope.

For either score, TCPR uses the maximum score over each envelope and index $k_T = \lceil (n+1)(1-\alpha+\delta) \rceil$, with infinite threshold when $k_T > n$. Integer endpoints, inclusive sets, and the full-set branch are retained. Envelopes depend only on (n, m) and are shared across models, scores, and trials. Cold construction includes design and validation; warm timing excludes it. We verified the public

author code against the paper; the published marginal rule is used here, rather than the notebook's additional FCP adjustment. The comparison concerns this certified numerical envelope and does not rank all TCPR constructions.

D.3 Scales, repeats, and statistical units

The synthetic design separates calibration, pool, and noise effects. Every synthetic setting generates independent standard-normal X and ϵ, sets $Y = X + \sigma\epsilon$, and predicts with $A(X) = X$. The calibration-size sweep uses $n \in \{50, 100, 200, 500, 1000, 2000, 5000, 10000\}$ with $m = 500$ and $\sigma = 0.1, 1$. The test-pool sweep uses the same eight values for m, with $n = 500$ and the same two noise levels. The noise sweep uses $\sigma \in \{0, 0.01, 0.03, 0.1, 0.3, 0.6, 1, 2\}$ at $(n, m) = (500, 1000)$. The heatmap uses $n \in \{50, 100, 200, 500, 1000, 2000\}$ and $\sigma \in \{0, 0.03, 0.1, 0.3, 1, 2\}$ at $m = 500$. Identical settings across these scans share trials. Their union has 60 settings with 500 complete trials each. RA is evaluated throughout, and VA additionally at $n = 100, 500, 2000$, $m = 500$, $\sigma = 0.1, 1$.

The nominal-level sweep uses $\alpha \in \{0.01, 0.025, 0.05, 0.075, 0.1, 0.15, 0.2, 0.3\}$ on both real datasets/models/scores and on synthetic RA at $n = 500, m = 1000, \sigma = 1$. Thresholds are recomputed from the same score distributions, without retraining or changing predictions. Coverage and size are retained at every level, including below-target means and full-set outcomes.

Reported 95% intervals for a mean use $t_{0.975,r-1}$ times the standard error across r complete synthetic trials or real outer splits. The individual test items are not treated as independent uncertainty units. Within-setting method comparisons are paired on these units; paired differences are saved with their own standard errors. Intervals describe Monte Carlo uncertainty from the specified experiment, not uncertainty under arbitrary populations. All grid points, outliers, full-set fractions, thresholds, and timings are retained. Lines connect observed settings only; heatmap cells are actual simulated settings with no interpolation.

E Unified predictors, full results, and calibration comparisons

E.1 Shared tasks and frozen training protocol

Ranking objectives use the same observed training labels. All five predictors share the 20 ESOL and Yummly outer partitions in Table 1. Training, calibration, and test sizes remain $628/200/300$ and $5000/500/1000$, respectively. Actual item IDs, target arrays, training-fitted feature vocabulary, and scaler parameters are matched to the saved partitions; the partitions are not reconstructed merely by reusing a seed. Training features are standardized before fitting each ranker, with no calibration or test labels used for pair selection, binning, stopping, or model selection. Each model is fitted once per split, and one saved raw score vector is shared by both scores, all calibration rules, and all eight nominal levels. Scores increase with the target; their direction is never selected using test performance.

RankNet learns a shared score through logistic preferences. A scalar network has two ReLU hidden layers of widths 64 and 32, followed by a linear output. Its loss on each training pair is

$$\mathrm{BCEWithLogits}(f_\theta(x_i) - f_\theta(x_j), \mathbb{I}\{y_i > y_j\}).$$

We implement analytic backpropagation in NumPy and check the gradients against finite differences. Adam uses learning rate 10^{-3}, weight decay 10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.999)$, and $\epsilon = 10^{-8}$. Each of 40 fixed epochs samples 20,000 ordered pairs with replacement, in batches of 512. Equal-outcome pairs are rejected without adding label noise; independent ordered sampling supplies both preference directions. The saved seeds, per-epoch pair hashes, actual rejection counts, loss sequences, and network weights identify the trained model. The VA input is its raw item score, without a sigmoid. This is the pairwise logistic objective of RankNet (Burges et al., 2005), not regression MSE.

Linear RankSVM fits mirrored training-pair differences. We sample 25,000 unequal-target pairs with replacement, form $d_{ij} = x_i - x_j$ and $s_{ij} = \mathrm{sign}(y_i - y_j)$, and include both (d_{ij}, s_{ij}) and $(-d_{ij}, -s_{ij})$. LinearSVC minimizes

$$\tfrac{1}{2}\|w\|^2 + C \sum_{\text{50,000 instances}} \max\{0, 1 - s_{ij} w^\top d_{ij}\}^2, \qquad C = 1.$$

It uses L2 regularization, squared hinge loss, no intercept, the primal solver, tolerance 10^{-4}, and at most 5,000 iterations. A prespecified convergence-warning fallback permits one retry at 10,000 iterations; all convergence statuses are retained. Standardization is fitted on item features before taking differences, with no second centering of pair features. Mirroring balances the signs but does not create independent observations. This is a sampled-pair linear ranking-SVM variant (Joachims, 2002), not a claim to reproduce the SVMrank binary. Its raw score is $w^\top x$.

LambdaMART uses explicit artificial training candidate lists. LightGBM's LGBMRanker uses the `lambdarank` objective (Burges, 2010), CPU execution, 200 trees, 15 leaves, learning rate 0.05, minimum child size 10, and one thread. An outcome-independent random permutation partitions only the training items into candidate lists of at most 128 items; a terminal singleton, if present, is merged with the preceding list. These are artificial training lists, not native search queries or evaluation pools. Training-target decile boundaries are fitted on the training labels alone, with repeated boundaries merged. Equal targets always receive the same relevance grade. Gains are explicitly linear, $0, 1, \ldots, L-1$, and the LambdaRank truncation level covers the largest actual training group. Group lengths and row permutations are saved and sum exactly to the training count. The executed booster objective, gains, truncation, and CPU setting are checked. Evaluation uses the original continuous targets and raw ranking scores, not discretized relevance or transformed scores.

Value-adaptive sets depend on the score's numerical spacing. RA depends on ordering, whereas a nonlinear monotone transformation can alter VA sets even if it preserves every predicted rank. No predictor receives test-derived score transformations. A common positive affine transformation multiplies all VA scores and their calibrated threshold by the same positive factor and therefore leaves the sets unchanged. The comparisons concern each specified raw scoring function.

E.2 Ranking quality and recipe structure

Spearman correlation and tie-aware Kendall τ_b use raw predictions and raw outcomes on held-out test items. They use average ranks or tie corrections, not the random tie breaking used by CP. The full-pool rank error is instead the mean $|\widehat{R}_i - R_i|/N$ over test items, whose ranks are computed within all calibration and test items. Each predictor/split contributes one quality record, not four records for the calibration methods. Undefined constant-score correlations would be reported as unavailable. Paired model differences and split-level intervals are saved with the results.

Table 3: Mean base-predictor quality over 20 shared outer splits. Correlations use test-only raw values; normalized rank error uses full-pool ranks of the test items.

Task	Predictor	Spearman	Kendall τ_b	Rank error/N
ESOL	RankNet	0.8964	0.7284	0.0965
ESOL	LambdaMART	0.8916	0.7241	0.0965
ESOL	RankSVM	0.8264	0.6370	0.1255
ESOL	Ridge	0.8232	0.6340	0.1263
ESOL	ExtraTrees	0.8878	0.7191	0.0995
Yummly	RankNet	0.9939	0.9475	0.0181
Yummly	LambdaMART	0.9442	0.8026	0.0679
Yummly	RankSVM	0.9246	0.8157	0.0621
Yummly	Ridge	0.9241	0.8036	0.0666
Yummly	ExtraTrees	0.9241	0.7779	0.0767

Ingredient geometry creates a structured similarity target. Before standardization, the excluded anchor has unit norm in all 20 training-fitted vocabularies. Each 1,500-item pool contains 8–24 zero TF–IDF vectors; their count averages 16.4. The duplicate-target fraction $1 - \#\{\text{distinct targets}\}/N$ ranges from 0.0147 to 0.0300. Nonzero vectors and the anchor have unit norm, so

$$-\|x - a\|_2 = -\sqrt{2 - 2x^\top a}$$

holds on that subset (maximum numerical discrepancy 2.1×10^{-8}). Its ordering is therefore equivalent to anchor dot-product ordering. Zero vectors are retained and reported separately; the unit-norm

identity is not applied to them. The anchor, distances, and target are not predictor inputs. This property describes a constructed ingredient-similarity task, not observed preferences or a general ranking benchmark. Per-split norms and target/prediction ties are retained with the results.

E.3 A POINT-MASS CALIBRATION ABLATION

The mean-rank ablation removes reference-rank dispersion. For calibration relative rank r, define

$$\mu_r = \frac{r(N+1)}{n+1}, \qquad \widetilde{r} = \mathrm{clip}(\lfloor \mu_r + 1/2 \rfloor, r, r+m).$$

Mean-rank replaces K_r by a point mass at $\widetilde{r}$, retaining DCR's $q = k/(n+1)$ and discrete mixture quantile. We implement half-up rounding with exact integer arithmetic. Its $q = 1$ threshold is the finite maximum, not the standard-conformal infinite-threshold branch. The ablation receives relative calibration order and predictions only. It is a comparison of calibration constructions, not an additional validity theorem.

Table 4: Mean-rank minus DCR at $\alpha = 0.1$: paired mean differences $\pm$ 95% interval half-width over 20 common splits. Negative size differences alone do not establish improved prediction. Both methods' individual results appear in the full tables.

Task	Predictor	Score	Δ coverage	Δ size/N
ESOL	RankNet	RA	-0.0035 ± 0.0036	-0.0048 ± 0.0037
ESOL	RankNet	VA	-0.0053 ± 0.0031	-0.0089 ± 0.0033
ESOL	LambdaMART	RA	-0.0030 ± 0.0033	-0.0036 ± 0.0031
ESOL	LambdaMART	VA	-0.0018 ± 0.0022	-0.0021 ± 0.0036
ESOL	RankSVM	RA	0.0015 ± 0.0027	0.0010 ± 0.0028
ESOL	RankSVM	VA	-0.0037 ± 0.0029	-0.0053 ± 0.0030
ESOL	Ridge	RA	0.0002 ± 0.0026	0.0012 ± 0.0035
ESOL	Ridge	VA	-0.0050 ± 0.0018	-0.0079 ± 0.0025
ESOL	ExtraTrees	RA	-0.0043 ± 0.0030	-0.0053 ± 0.0033
ESOL	ExtraTrees	VA	-0.0067 ± 0.0037	-0.0090 ± 0.0041
Yummly	RankNet	RA	-0.0221 ± 0.0033	-0.0135 ± 0.0010
Yummly	RankNet	VA	-0.0227 ± 0.0054	-0.0144 ± 0.0012
Yummly	LambdaMART	RA	-0.0029 ± 0.0018	-0.0033 ± 0.0017
Yummly	LambdaMART	VA	-0.0025 ± 0.0015	-0.0029 ± 0.0018
Yummly	RankSVM	RA	-0.0023 ± 0.0012	-0.0023 ± 0.0012
Yummly	RankSVM	VA	-0.0101 ± 0.0031	-0.0070 ± 0.0018
Yummly	Ridge	RA	-0.0036 ± 0.0016	-0.0038 ± 0.0015
Yummly	Ridge	VA	-0.0079 ± 0.0029	-0.0069 ± 0.0022
Yummly	ExtraTrees	RA	-0.0019 ± 0.0022	-0.0017 ± 0.0021
Yummly	ExtraTrees	VA	-0.0037 ± 0.0021	-0.0055 ± 0.0026

E.4 COMPLETE SAME-PROTOCOL RESULTS

All rows in the following tables have zero full-set fraction at $\alpha = 0.1$. Nonzero fractions at other levels remain in the saved trial results. Means use exact recovered integer coverage and cardinality totals, rounded half-up to four decimals; intervals use 20 split-level observations.

Ranking predictors have distinct coverage–size profiles. For ESOL, RankNet and LambdaMART produce smaller mean sets than linear RankSVM in both scores. The calibration methods share each fitted score vector and split. The intervals below quantify variation across the 20 partitions of this fixed molecular dataset; they do not treat the correlated test items within a pool as independent replicates.

Table 5: ESOL, $\alpha = 0.1$: ranking predictors. Mean $\pm$ 95% split-level t interval half-width; 20 shared splits.

		RA		VA	
Predictor	Calibration	Coverage	Size/N	Coverage	Size/N
RankNet	Oracle	$0.8988 \pm .0136$	$0.3862 \pm .0120$	$0.9038 \pm .0092$	$0.3996 \pm .0120$
RankNet	TCPR	$0.9687 \pm .0067$	$0.5269 \pm .0102$	$0.9817 \pm .0033$	$0.5896 \pm .0123$
RankNet	Mean-rank	$0.8973 \pm .0147$	$0.3841 \pm .0128$	$0.9043 \pm .0091$	$0.4000 \pm .0108$
RankNet	DCR	$0.9008 \pm .0129$	$0.3889 \pm .0112$	$0.9097 \pm .0081$	$0.4088 \pm .0100$
RankNet	MDCR	$0.9025 \pm .0137$	$0.3902 \pm .0136$	$0.9103 \pm .0100$	$0.4112 \pm .0116$
LambdaMART	Oracle	$0.9042 \pm .0114$	$0.3860 \pm .0110$	$0.9073 \pm .0144$	$0.4048 \pm .0189$
LambdaMART	TCPR	$0.9653 \pm .0053$	$0.5318 \pm .0101$	$0.9695 \pm .0057$	$0.5661 \pm .0184$
LambdaMART	Mean-rank	$0.9052 \pm .0126$	$0.3890 \pm .0130$	$0.9087 \pm .0144$	$0.4076 \pm .0199$
LambdaMART	DCR	$0.9082 \pm .0122$	$0.3926 \pm .0123$	$0.9105 \pm .0134$	$0.4097 \pm .0176$
LambdaMART	MDCR	$0.9107 \pm .0126$	$0.3983 \pm .0129$	$0.9105 \pm .0134$	$0.4126 \pm .0183$
RankSVM	Oracle	$0.9007 \pm .0126$	$0.4950 \pm .0127$	$0.8988 \pm .0115$	$0.5200 \pm .0131$
RankSVM	TCPR	$0.9640 \pm .0054$	$0.6271 \pm .0126$	$0.9635 \pm .0065$	$0.6560 \pm .0135$
RankSVM	Mean-rank	$0.9037 \pm .0126$	$0.4992 \pm .0137$	$0.8937 \pm .0147$	$0.5130 \pm .0138$
RankSVM	DCR	$0.9022 \pm .0135$	$0.4983 \pm .0131$	$0.8973 \pm .0136$	$0.5183 \pm .0132$
RankSVM	MDCR	$0.9037 \pm .0129$	$0.5018 \pm .0154$	$0.9002 \pm .0156$	$0.5256 \pm .0168$

Regression controls provide a shared-task comparison. Ridge has lower VA coverage than its RA coverage, whereas ExtraTrees stays near the target under both scores. Mean-rank reduces ExtraTrees set size but also lowers its mean coverage. These differences are paired within splits. Comparing the means and their intervals keeps the coverage cost visible instead of ranking calibration methods by size alone.

Table 6: ESOL, $\alpha = 0.1$: regression controls. Mean $\pm$ 95% split-level t interval half-width; 20 shared splits.

		RA		VA	
Predictor	Calibration	Coverage	Size/N	Coverage	Size/N
Ridge	Oracle	$0.9043 \pm .0140$	$0.5049 \pm .0139$	$0.8952 \pm .0126$	$0.5128 \pm .0140$
Ridge	TCPR	$0.9643 \pm .0066$	$0.6307 \pm .0122$	$0.9612 \pm .0064$	$0.6484 \pm .0119$
Ridge	Mean-rank	$0.9048 \pm .0141$	$0.5062 \pm .0146$	$0.8905 \pm .0126$	$0.5062 \pm .0119$
Ridge	DCR	$0.9047 \pm .0133$	$0.5050 \pm .0132$	$0.8955 \pm .0120$	$0.5141 \pm .0123$
Ridge	MDCR	$0.9050 \pm .0133$	$0.5073 \pm .0136$	$0.8962 \pm .0127$	$0.5179 \pm .0143$
ExtraTrees	Oracle	$0.8945 \pm .0141$	$0.3841 \pm .0127$	$0.9012 \pm .0121$	$0.3941 \pm .0128$
ExtraTrees	TCPR	$0.9642 \pm .0048$	$0.5280 \pm .0116$	$0.9738 \pm .0056$	$0.5789 \pm .0123$
ExtraTrees	Mean-rank	$0.8988 \pm .0134$	$0.3878 \pm .0124$	$0.8978 \pm .0110$	$0.3908 \pm .0121$
ExtraTrees	DCR	$0.9032 \pm .0126$	$0.3931 \pm .0114$	$0.9045 \pm .0108$	$0.3997 \pm .0113$
ExtraTrees	MDCR	$0.9053 \pm .0143$	$0.3981 \pm .0151$	$0.9033 \pm .0121$	$0.4011 \pm .0125$

Ranking predictors have distinct coverage–size profiles. On the fixed-anchor recipe task, RankNet has much smaller sets than the other ranking predictors but DCR and MDCR are more conservative. Mean-rank moves its coverage closer to target and reduces size under both RA and VA. This contrast links the ablation to the predictor rather than to a universal ordering of calibration constructions.

Table 7: Yummly, $\alpha = 0.1$: ranking predictors. Mean $\pm$ 95% split-level t interval half-width; 20 shared splits.

		RA		VA	
Predictor	Calibration	Coverage	Size/N	Coverage	Size/N
RankNet	Oracle	$0.8971 \pm .0079$	$0.0729 \pm .0049$	$0.8994 \pm .0096$	$0.0751 \pm .0063$
RankNet	TCPR	$0.9916 \pm .0014$	$0.2170 \pm .0044$	$0.9907 \pm .0016$	$0.2156 \pm .0056$
RankNet	Mean-rank	$0.9339 \pm .0076$	$0.0878 \pm .0050$	$0.9308 \pm .0103$	$0.0875 \pm .0060$
RankNet	DCR	$0.9560 \pm .0060$	$0.1013 \pm .0047$	$0.9535 \pm .0072$	$0.1019 \pm .0052$
RankNet	MDCR	$0.9539 \pm .0060$	$0.0998 \pm .0045$	$0.9541 \pm .0084$	$0.1028 \pm .0049$
LambdaMART	Oracle	$0.9021 \pm .0068$	$0.2756 \pm .0061$	$0.9016 \pm .0060$	$0.2811 \pm .0071$
LambdaMART	TCPR	$0.9645 \pm .0028$	$0.3911 \pm .0058$	$0.9656 \pm .0026$	$0.4127 \pm .0073$
LambdaMART	Mean-rank	$0.9064 \pm .0068$	$0.2800 \pm .0072$	$0.9081 \pm .0059$	$0.2880 \pm .0074$
LambdaMART	DCR	$0.9092 \pm .0068$	$0.2833 \pm .0063$	$0.9106 \pm .0053$	$0.2908 \pm .0069$
LambdaMART	MDCR	$0.9107 \pm .0070$	$0.2852 \pm .0066$	$0.9110 \pm .0055$	$0.2916 \pm .0070$
RankSVM	Oracle	$0.8956 \pm .0097$	$0.2360 \pm .0115$	$0.8980 \pm .0094$	$0.2275 \pm .0117$
RankSVM	TCPR	$0.9691 \pm .0032$	$0.3580 \pm .0107$	$0.9788 \pm .0017$	$0.3740 \pm .0116$
RankSVM	Mean-rank	$0.9009 \pm .0092$	$0.2408 \pm .0107$	$0.9027 \pm .0098$	$0.2314 \pm .0120$
RankSVM	DCR	$0.9032 \pm .0090$	$0.2431 \pm .0105$	$0.9128 \pm .0081$	$0.2385 \pm .0114$
RankSVM	MDCR	$0.9042 \pm .0087$	$0.2443 \pm .0100$	$0.9109 \pm .0092$	$0.2378 \pm .0115$

Regression controls provide a shared-task comparison. Ridge and ExtraTrees use exactly the same recipe targets and outer splits as the ranking predictors. Their MDCR results use inverse-CDF draws from the same reference law. For ExtraTrees, VA increases set size relative to RA, showing that adapting to score spacing need not reduce cardinality. The comparison retains both scores and every calibration method.

Table 8: Yummly, $\alpha = 0.1$: regression controls. Mean $\pm$ 95% split-level t interval half-width; 20 shared splits.

		RA		VA	
Predictor	Calibration	Coverage	Size/N	Coverage	Size/N
Ridge	Oracle	$0.8951 \pm .0091$	$0.2573 \pm .0143$	$0.8965 \pm .0081$	$0.2473 \pm .0135$
Ridge	TCPR	$0.9642 \pm .0040$	$0.3765 \pm .0123$	$0.9775 \pm .0020$	$0.3954 \pm .0134$
Ridge	Mean-rank	$0.8984 \pm .0096$	$0.2607 \pm .0136$	$0.9032 \pm .0087$	$0.2536 \pm .0133$
Ridge	DCR	$0.9019 \pm .0090$	$0.2645 \pm .0128$	$0.9111 \pm .0073$	$0.2605 \pm .0132$
Ridge	MDCR	$0.9017 \pm .0088$	$0.2644 \pm .0134$	$0.9099 \pm .0077$	$0.2600 \pm .0137$
ExtraTrees	Oracle	$0.9082 \pm .0080$	$0.3212 \pm .0071$	$0.8998 \pm .0079$	$0.3414 \pm .0098$
ExtraTrees	TCPR	$0.9685 \pm .0035$	$0.4350 \pm .0074$	$0.9710 \pm .0037$	$0.4874 \pm .0098$
ExtraTrees	Mean-rank	$0.9105 \pm .0081$	$0.3243 \pm .0077$	$0.9016 \pm .0079$	$0.3434 \pm .0095$
ExtraTrees	DCR	$0.9124 \pm .0077$	$0.3260 \pm .0073$	$0.9052 \pm .0078$	$0.3489 \pm .0092$
ExtraTrees	MDCR	$0.9142 \pm .0084$	$0.3289 \pm .0084$	$0.9059 \pm .0078$	$0.3498 \pm .0096$

Full-set behavior depends on the prescribed level. At $\alpha = 0.1$ all fractions are zero. At other levels, 71 dataset/predictor/score/calibrator/level groups have nonzero mean fractions; the complete records are retained with the split-level results. Nonzero groups occur at $\alpha = 0.01$ (61 groups, mean fractions 0.0005–0.6203), 0.025 (9 groups, 0.0025–0.0578), and 0.05 (one group, 0.0018). All other levels have zero full-set fraction.

E.5 Achieved coverage and output size

Target levels expose predictor-dependent conservatism. Figure 4 uses all eight prescribed levels and 20 splits for RankNet on both tasks. The Yummly ranker's high accuracy (mean Spearman correlation 0.9939) accompanies additional conservatism. Intervals describe variation across random partitions, not independent test items. The following panels show all predictors and both scores.

Figures 5–8 use the eight prespecified levels $\alpha \in \{0.01, 0.025, 0.05, 0.075, 0.1, 0.15, 0.2, 0.3\}$. Within each predictor/score, thresholds index a shared nested set family $C_i(t)$, without implying identical mean curves or optimality. Horizontal/vertical bars are split-level 95% intervals for mean size/coverage. Hollow circles mark $\alpha = 0.1$; the largest/smallest size endpoints correspond to $0.01/0.3$. Lines join actual settings, without interpolation or test-based level selection for matched coverage.

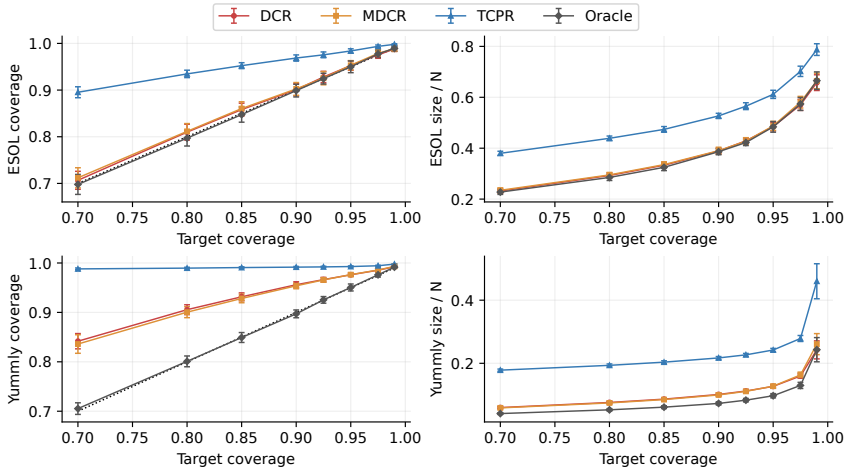

Figure 4: RankNet RA on ESOL (top) and Yummly (bottom): achieved coverage (left) and mean cardinality/N (right) against target coverage. All eight levels use the same 20 splits. Bars are 95% split-level t intervals, not trial-level SD; no experimental point is interpolated.

ESOL RA offers similar DCR and MDCR trade-offs across levels. These panels compare five predictors on the same nested rank-set family. Intervals use 20 shared splits; Oracle uses additional true ranks, so proximity describes empirical behavior rather than optimality.

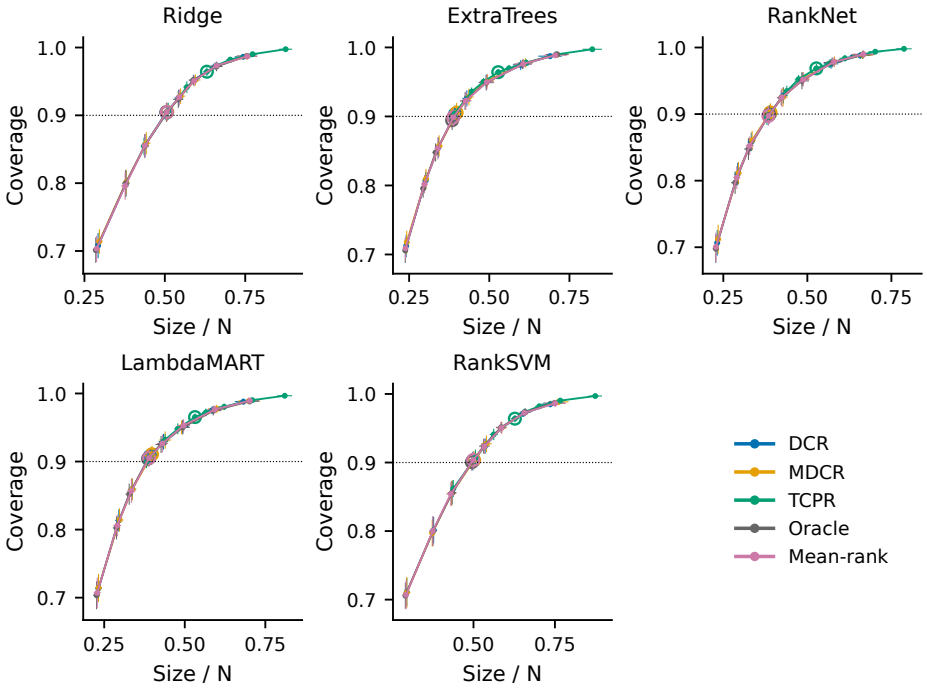

Figure 5: ESOL, RA: all five predictors, five calibration constructions, and eight nominal levels.

VA changes the trade-off through score spacing. ESOL predictors assign different gaps to similarly ordered items. Raw-score constructions are therefore compared separately from RA; Ridge's below-target region remains visible. Both axes use split-level uncertainty.

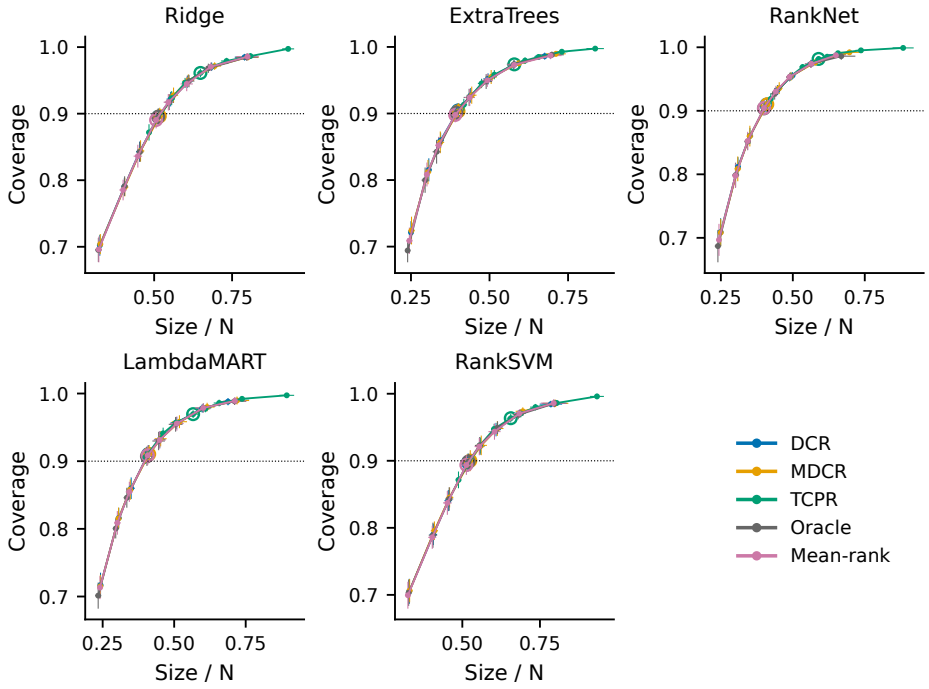

Figure 6: ESOL, VA: raw predictor scores, with the same settings and uncertainty units.

The recipe RankNet is more conservative than the other predictors. On the fixed-anchor task, Mean-rank produces smaller sets and coverage closer to target for RankNet. Other predictors' curves are closer together; all prescribed levels and overlapping curves are retained.

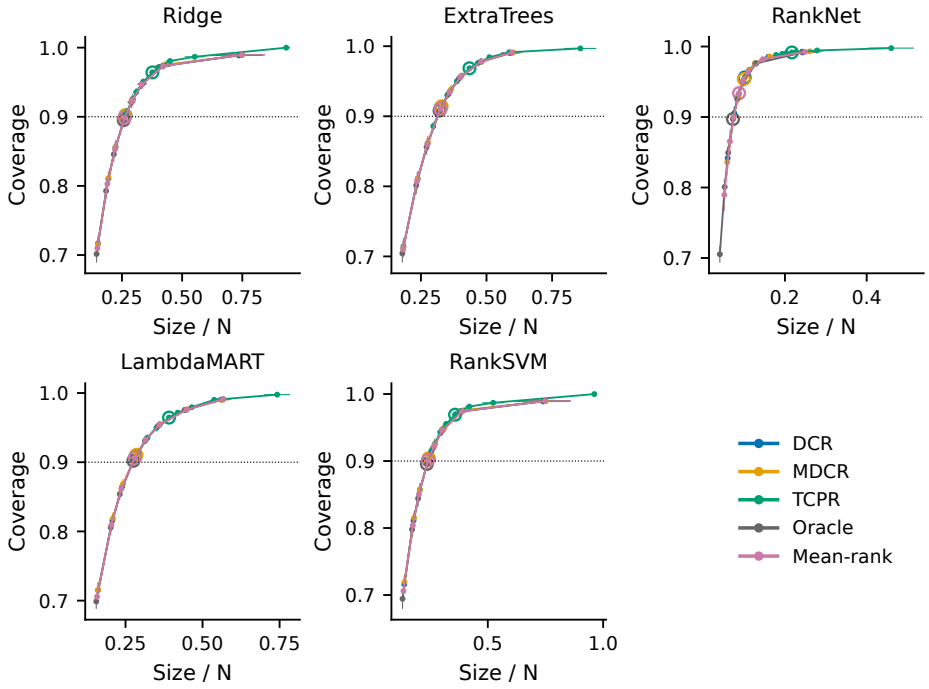

Figure 7: Yummly, RA: the fixed-anchor, training-fitted ingredient-similarity task.

Recipe VA retains the predictor-dependent ablation pattern. Methods share raw predictor scores within each split. RankNet separates from the other predictors in size and conservatism; the full level range also retains below-target means.

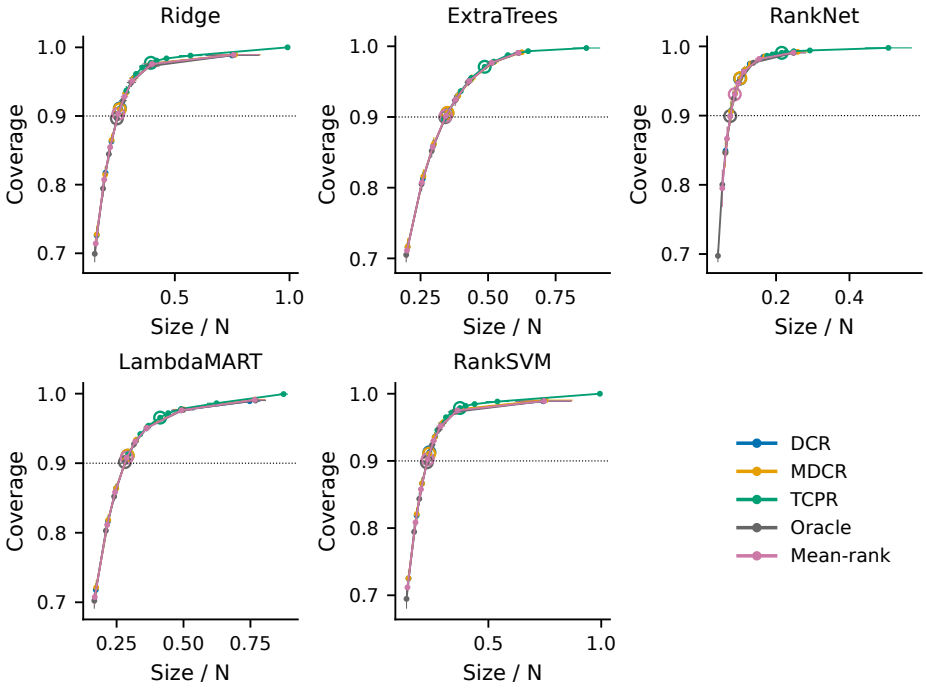

Figure 8: Yummly, VA: all observed nominal-level points, including below-target means.

E.6 Direct sampling and component timing

Component measurements separate setup from repeated calibration.

All paired totals use one CPU thread, seven predetermined (n, m) settings, three saved pools per setting, and ten warm repetitions per pool. Warm totals include prediction sorting, threshold computation, and identical inclusive interval endpoints and cardinalities; training, prediction inference, and outcome-based coverage evaluation are excluded. Warm DCR retains its reference-kernel and integer-count caches after prewarming; cold runs clear both caches. The floating-point-only path is a timing reference rather than a recommended calibration implementation; it is timed on the same pools and operation range for comparison. Quartiles describe repetitions, not independent dataset uncertainty. Separate kernel/envelope setup measurements in Table 10 and fresh-process direct-sampler memory measurements concern unchanged components.

Set construction can dominate direct VA calibration. The component measurements distinguish sorting, auxiliary calibration, and inclusive set construction. At $(500, 10000)$, direct VA set construction costs 8.749 ms, compared with 0.145 ms for calibration. Reducing sampling cost alone therefore does not remove all given-prediction cost. Each total is a measured total; medians of components need not sum to its median.

Table 9: Median warm costs in milliseconds, pooling 30 repetitions from three saved pools. Sort, calibration, and sets describe direct MDCR. DCR includes integer quantile verification.

n	m	Score	Sort	Cal.	Sets	DCR	Inverse	Direct
200	300	RA	0.022	0.035	0.004	0.297	0.251	0.061
200	300	VA	0.026	0.064	0.241	2.860	0.525	0.329
500	1000	RA	0.073	0.067	0.006	1.754	0.639	0.143
500	1000	VA	0.128	0.125	0.859	26.479	1.640	1.117
100	500	RA	0.019	0.031	0.004	0.225	0.143	0.054
100	500	VA	0.020	0.045	0.400	2.518	0.567	0.468
500	500	RA	0.034	0.057	0.005	1.004	0.584	0.098
500	500	VA	0.044	0.118	0.411	12.720	1.106	0.573
2000	500	RA	0.095	0.207	0.005	4.105	2.409	0.303
2000	500	VA	0.161	0.253	0.434	57.727	2.981	0.849
10000	500	RA	0.678	0.861	0.006	21.798	12.411	1.553
10000	500	VA	0.733	0.891	0.450	312.189	12.928	2.092
500	10000	RA	0.650	0.115	0.028	14.871	1.373	0.780
500	10000	VA	0.758	0.145	8.749	293.671	10.300	9.649

Envelope construction and drawing reference ranks are different operations. The envelope column includes independent design and validation simulations; the sampler columns only generate a reference-rank vector. The kernel is separately timed. These measurements explain which costs can be amortized over repeated calibration and cannot be combined into a speed ratio between unlike operations.

Table 10: Separate measurements of unchanged kernel/envelope setup and sampler-only costs. Kernel and TCPR envelope entries are milliseconds; sampler entries are microseconds per draw vector (RA timing records, warm inverse CDF). Envelope construction executes 10,000 design and 50,000 independent validation draws, including its kernel, with the stated exact fallback; it is not a cached-file read.

n	m	Kernel (ms)	Envelope (ms)	Inverse (μs)	Direct (μs)
200	300	0.712	535.4	201.0	24.4
500	1000	4.955	965.8	535.7	40.3
100	500	0.593	374.2	104.6	15.6
500	500	2.481	910.3	517.5	44.4
2000	500	11.042	1260.2	2178.6	153.8
10000	500	61.216	3225.4	14910.1	749.7
500	10000	62.061	1382.0	796.4	47.0

Fresh-process direct runs disable dense-kernel access with raising stubs. Their measured peak process memory is 93.3–99.4 MiB, including Python and imported libraries; this is not the sampler allocation alone. The combined timing process with integer verification peaks at 1066.5 MiB. Direct auxiliary storage is $O(n)$ at $B = 1$; the inverse-CDF table stores $n(m + 1)$ entries. Cold component records and all warm quartiles are retained separately. TCPR cold construction uses an independent design/validation stream; warm TCPR uses the shared frozen envelope.

Cold totals include construction that warm totals amortize. Clearing the required caches changes the relative contribution of kernel and integer-count construction. TCPR additionally constructs and validates its envelope. These totals use the same inclusive interval representation for all methods and exclude model training and prediction inference. Comparing them with warm costs identifies implementation overhead rather than a stronger statistical guarantee.

Table 11: Median cold given-prediction cost (milliseconds) over three saved pools. Required kernel, integer reference counts or actual TCPR envelope construction are included, followed by threshold and set computation.

n	m	Score	Oracle	TCPR	DCR	Inverse	Direct
200	300	RA	0.069	515.612	4.631	0.814	0.108
200	300	VA	0.351	518.762	5.153	1.069	0.378
500	1000	RA	0.174	964.339	96.302	5.663	0.254
500	1000	VA	1.035	945.472	76.954	6.152	1.190
100	500	RA	0.063	361.843	2.018	0.654	0.105
100	500	VA	0.475	360.605	3.802	1.064	0.515
500	500	RA	0.087	915.752	18.818	2.950	0.165
500	500	VA	0.501	915.950	21.950	3.300	0.633
2000	500	RA	0.195	1243.927	59.491	12.234	0.432
2000	500	VA	0.628	1235.175	89.985	12.976	0.867
10000	500	RA	0.849	2987.460	161.500	63.094	1.670
10000	500	VA	1.321	2970.779	421.680	62.449	2.205
500	10000	RA	0.800	976.414	187.054	47.842	0.965
500	10000	VA	9.458	992.652	422.467	56.818	9.908

Integer verification resolves discrete probability boundaries. For $D = \binom{N}{n}$ and $A(r,j) = \binom{j-1}{r-1}\binom{N-j}{n-r}$, let $M(t) = \sum_i \sum_j A(r_i, j)\mathbb{I}\{s_i(j) \le t\}$. We propose a threshold using floating cumulative masses and accept it only after exact integer checks $(n+1)M(T^-) < knD \le (n+1)M(T)$. If either fails, integer comparisons locate the first valid atom. Exact hypergeometric prefix counts are memoized; all support masses remain positive for the finite $q = 1$ branch, regardless of floating underflow. Scores retain the stored binary64 prediction differences and inclusive predicates; exact probability arithmetic does not make model outputs real-arithmetic exact. All 3,200 fitted-predictor DCR thresholds satisfy separately computed integer postconditions, with matching inclusive test sets. A fixed suite of 16,330 small VA cases agrees with separate subset enumeration; this is implementation regression evidence, not a general coverage theorem. Every nontrivial timed DCR call performs integer verification; none of the saved timing pools required relocating its floating candidate. Component logs retain certification time and cold/warm cache scope.

F Score geometry, scales, and computational diagnostics

F.1 Complete-trial distributions and individual rank sets

Trial dispersion adds information beyond mean coverage. With fixed test size and noise, changing calibration size changes the distribution of both FCP and set size. The shared threshold makes test outcomes within a pool dependent, so the complete trial is the unit of comparison. The boxes retain outliers and reveal variation that a confidence interval for the mean would conceal.

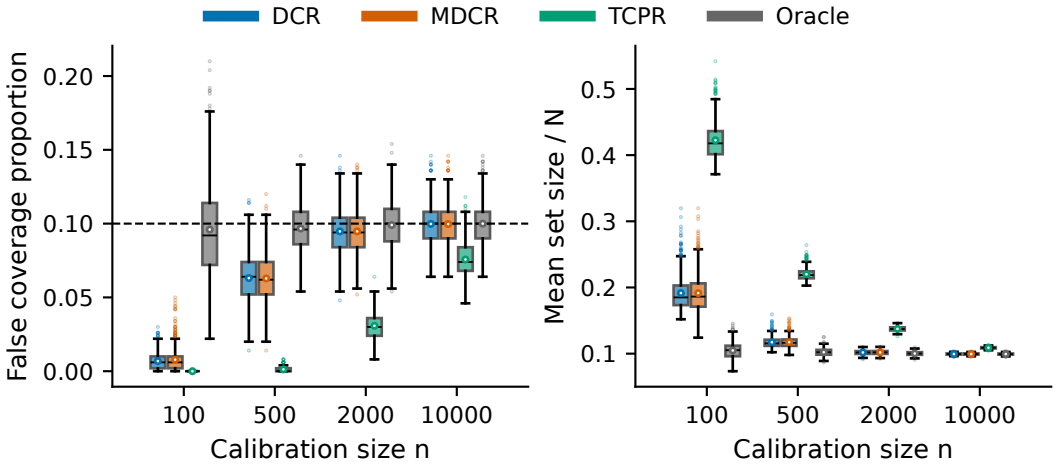

Figure 9: RA trial distributions at $m = 500$, $\sigma = 0.1$, $\alpha = 0.1$; 500 trials per group. Boxes: quartiles; line: median; white dot: mean; whiskers: 1.5 interquartile ranges; points: outliers. Dashed line: target FCP. The full eight-point sweeps appear in Figures 1 and 12.

Full-pool geometry shows how score choice changes rank-set boundaries. The fixed Yummly/ExtraTrees split contains all 1,000 test items and uses the ordering of all $n + m = 1500$ items. RA/VA coverage is $90.7\%/87.8\%$ and size/N is $0.3213/0.3129$. Across 20 splits, DCR–VA coverage is 0.9052 ± 0.0078; the fixed split is an item-level illustration rather than evidence of systematic undercoverage.

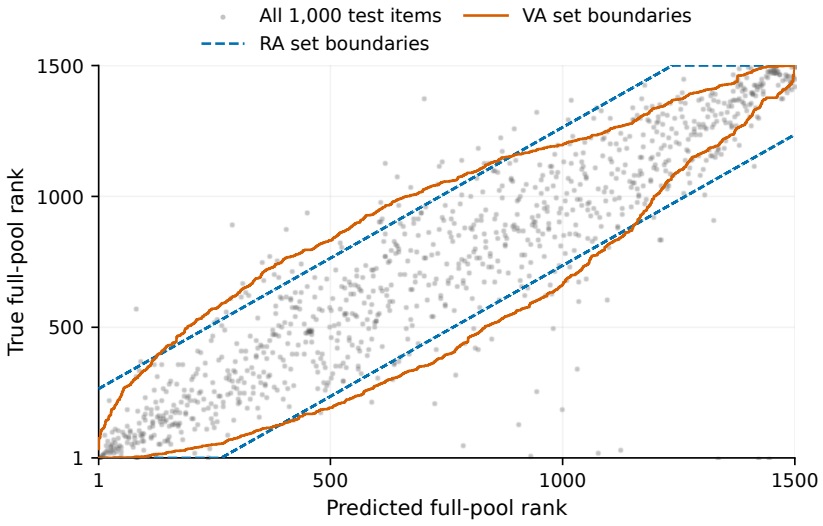

Figure 10: DCR at $\alpha = 0.1$ on Yummly/ExtraTrees split 0. Points include every test item; staircase boundaries are inclusive discrete sets and are not fitted to the cloud. Rank 1 is the lowest outcome.

RA and VA assign different sets to the same items. The 30 items are predetermined members of the same split shown in Figure 10. RA uses rank distances, while VA converts predicted-value gaps into potentially asymmetric rank boundaries. Uncovered points are retained. These item-level sets illustrate score geometry; their coverage is not a separate average over randomized splits.

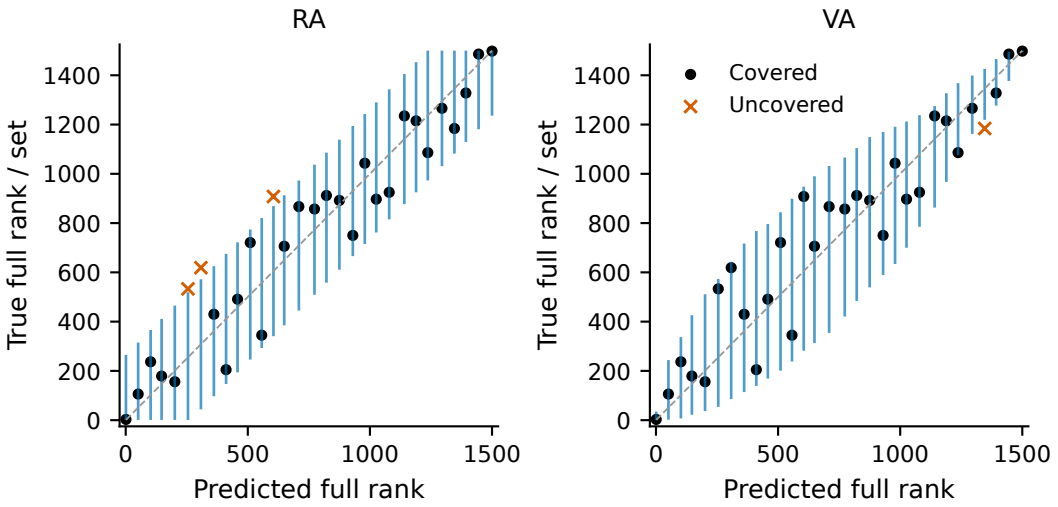

Figure 11: DCR rank sets for 30 predetermined test items, Yummly/ExtraTrees split 0, $\alpha = 0.1$. Vertical segments are prediction sets; points are true full ranks. RA and VA share the same model, pool, and items. The dashed diagonal is predicted rank equals true rank, not a fitted trend.

F.2 Scale, model-quality, and nominal-level results

The complete calibration-size scan is retained in Figure 1. The main-paper figure shows all eight actual sizes, both frozen noise regimes, and trial-level intervals. The remaining panels below preserve the paired test-size, noise, signed-error, and normalized-size analyses.

A larger test pool changes the ranking problem itself. Calibration size stays fixed, but adding test items changes both full-pool positions and the denominator N. The paired coverage and normalized-size panels should therefore be read together. Each point is an independently generated complete-trial setting, with no interpolation between the eight actual pool sizes.

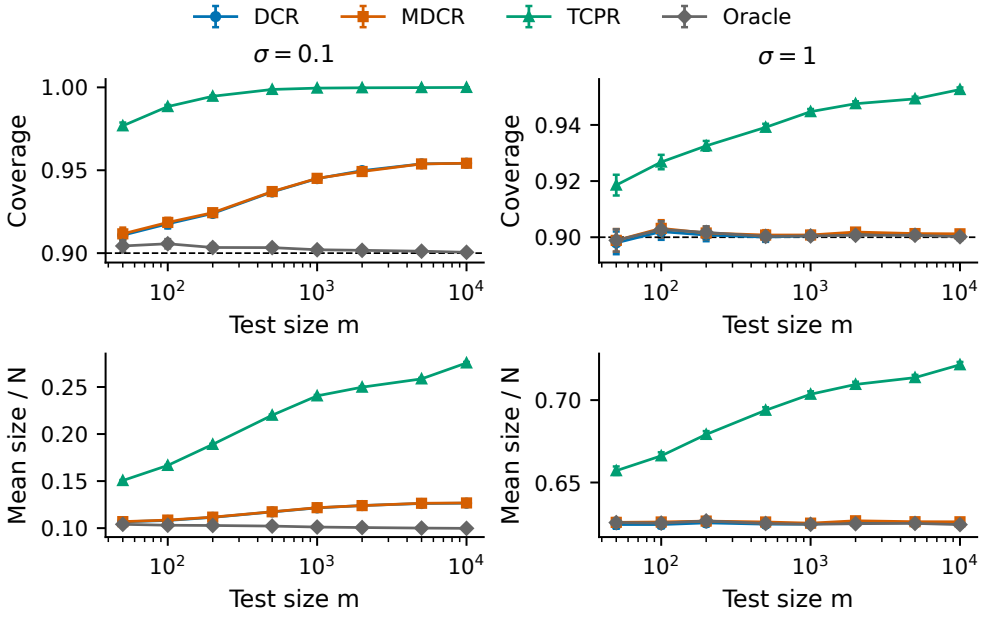

Figure 12: Test-pool sweep with $n = 500$, RA, $\alpha = 0.1$. All eight actual sizes, 500 trials each; bars show 95% trial-level intervals. The ranking target changes with the test pool.

Signed coverage error exposes both conservatism and shortfalls. The grid combines six calibration sizes with six noise levels at fixed test size. A single symmetric color scale permits comparisons between methods; negative cells are not clipped. The grid diagnoses behavior of the fixed Gaussian predictor and does not assert validity for arbitrary fitted ranking functions.

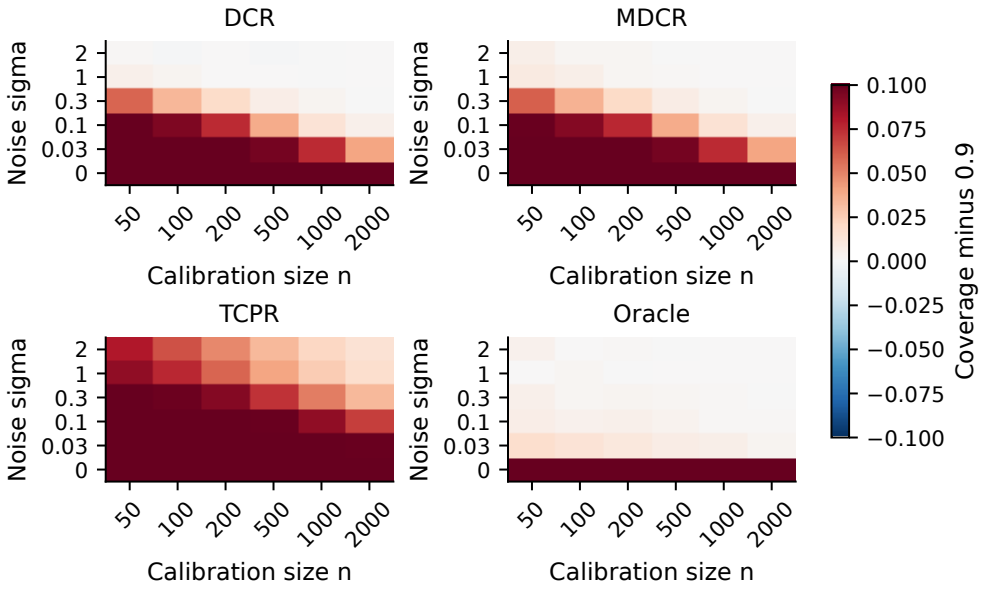

Figure 13: Signed coverage error relative to 0.9, RA, $m = 500$. Each cell is a mean from 500 complete trials; all methods use the same 36 settings. One shared symmetric color scale shows below- and above-target coverage.

Set size complements the signed coverage grid. These panels use exactly the same trials and settings as Figure 13. Smaller cardinality is useful only in conjunction with the achieved coverage, which is why the two grids remain adjacent. A common scale across methods prevents a narrow range of sizes from appearing artificially different through separate color normalization.

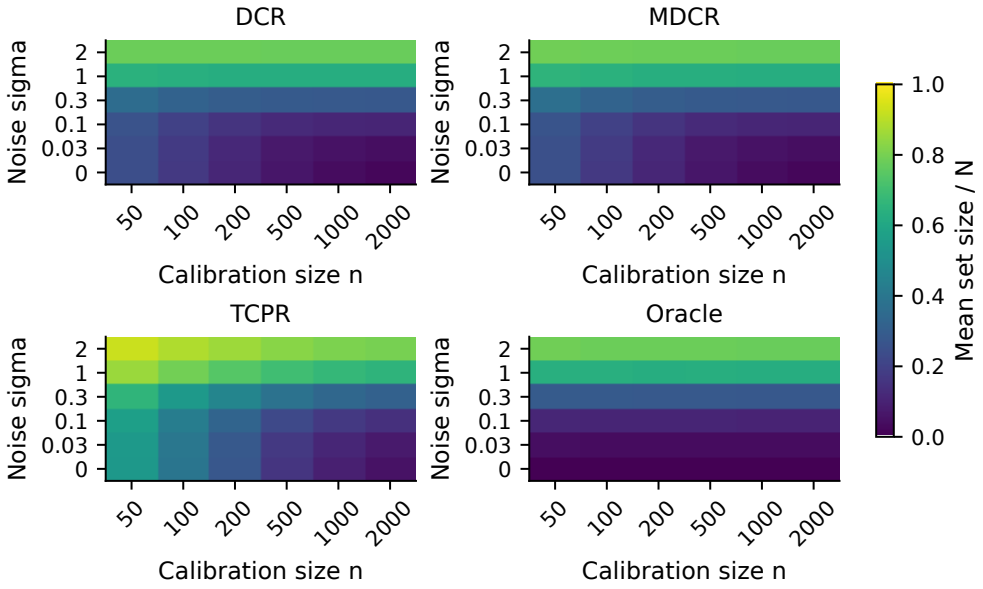

Figure 14: Mean normalized set size on the same 36 RA settings as Figure 13. A shared $[0, 1]$ color scale allows direct comparison. Cells represent observations, not interpolation.

The noise sweep varies predictive quality without fitting a new model. The predictor remains $A(X) = X$ while the outcome noise changes. Zero noise is an explicit setting, not a logarithmic extrapolation. The coverage and size panels preserve all eight actual levels and their trial-level intervals, separating this mechanism from claims about the precision of learned predictors.

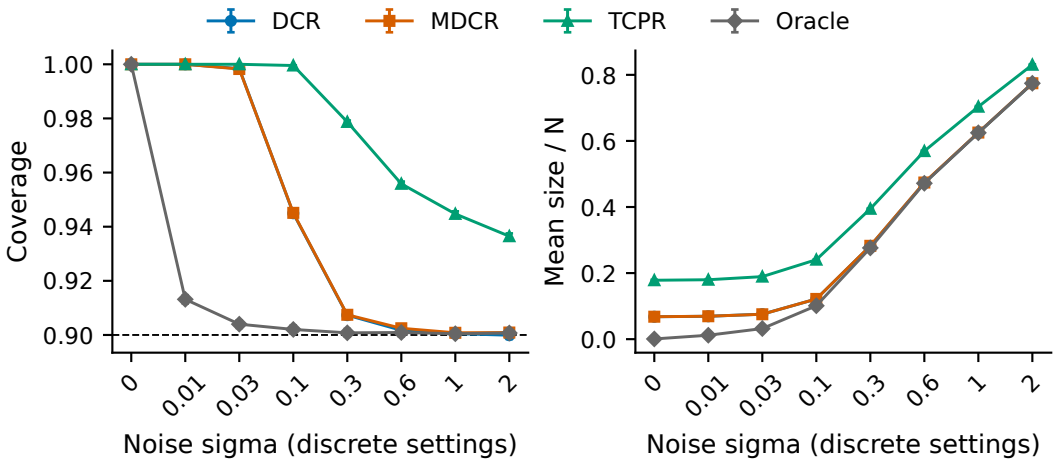

Figure 15: Model-quality sweep, $n = 500, m = 1000$, RA, $\alpha = 0.1$. The eight discrete noise levels include zero as its own category. Means and 95% trial-level intervals use 500 trials per level.

VA and RA need not have the same coverage at a common target. These paired trials keep the predictor and pool fixed while changing the score construction. At low noise, VA is more conservative for small calibration sets, while the gap decreases as n grows. At $\sigma = 1$ and $n = 2000$, the displayed VA means include below-target values. All six settings and all four methods remain in the comparison.

Table 12: Synthetic score comparison at $m = 500$, $\alpha = 0.1$, 500 paired complete trials per row. All entries are means; uncertainty and paired records are saved with the full results.

			Coverage		Size/N	
σ	n	Method	RA	VA	RA	VA
0.1	100	Oracle	0.9040	0.9020	0.1048	0.0955
0.1	100	TCPR	1.0000	1.0000	0.4225	0.4886
0.1	100	DCR	0.9931	0.9994	0.1915	0.2075
0.1	100	MDCR	0.9922	0.9992	0.1915	0.2102
0.1	500	Oracle	0.9033	0.9026	0.1022	0.0940
0.1	500	TCPR	0.9987	1.0000	0.2200	0.2316
0.1	500	DCR	0.9369	0.9580	0.1172	0.1159
0.1	500	MDCR	0.9371	0.9582	0.1174	0.1161
0.1	2000	Oracle	0.9010	0.9005	0.1002	0.0927
0.1	2000	TCPR	0.9692	0.9823	0.1374	0.1334
0.1	2000	DCR	0.9053	0.9103	0.1018	0.0954
0.1	2000	MDCR	0.9052	0.9104	0.1019	0.0954
1	100	Oracle	0.9017	0.9015	0.6304	0.6342
1	100	TCPR	0.9750	0.9840	0.7872	0.8136
1	100	DCR	0.9027	0.9059	0.6322	0.6396
1	100	MDCR	0.9055	0.9083	0.6373	0.6439
1	500	Oracle	0.9003	0.9000	0.6251	0.6287
1	500	TCPR	0.9391	0.9458	0.6938	0.7055
1	500	DCR	0.9001	0.9002	0.6247	0.6289
1	500	MDCR	0.9008	0.9008	0.6260	0.6299
1	2000	Oracle	0.9002	0.8989	0.6236	0.6269
1	2000	TCPR	0.9169	0.9187	0.6509	0.6565
1	2000	DCR	0.9001	0.8989	0.6234	0.6269
1	2000	MDCR	0.9002	0.8991	0.6237	0.6271

ESOL ExtraTrees trades set size for target coverage. The eight levels share predictions and partitions, so their differences arise from calibration thresholds rather than refitting. RA sets depend on predicted ordering, with size normalized by the common full-pool support. These means and intervals describe randomized partitions of the fixed molecular dataset and complement the RankNet comparison in the main experiment.

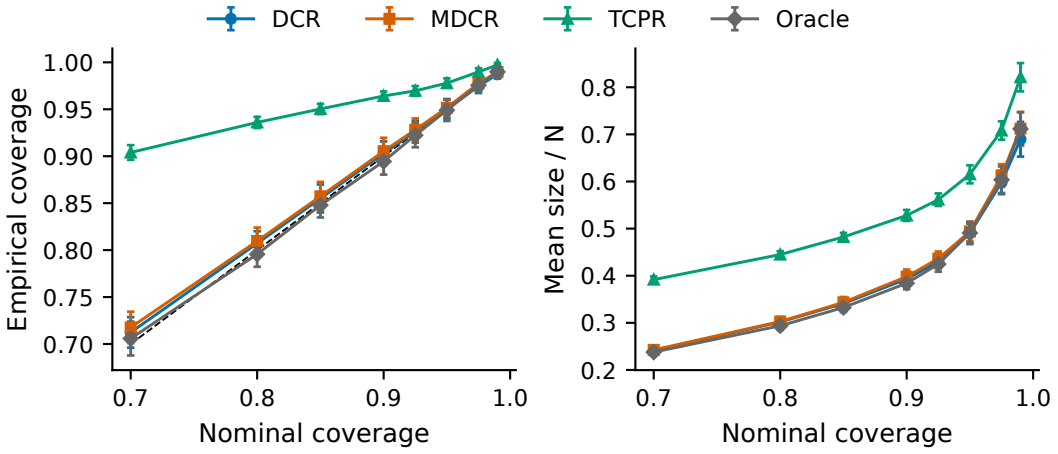

Figure 16: ESOL/ExtraTrees RA nominal-level sweep, 20 splits. The diagonal denotes nominal coverage; bars are split-level 95% intervals. All eight levels share predictions within each split.

Recipe ExtraTrees has a different coverage profile from RankNet. The same fixed-anchor task and split-level aggregation are used here. Raising the target expands sets, but the achieved means must still be compared with the diagonal rather than assumed equal to it. Keeping the complete level sequence makes the coverage–size relationship visible without selecting a favorable operating point.

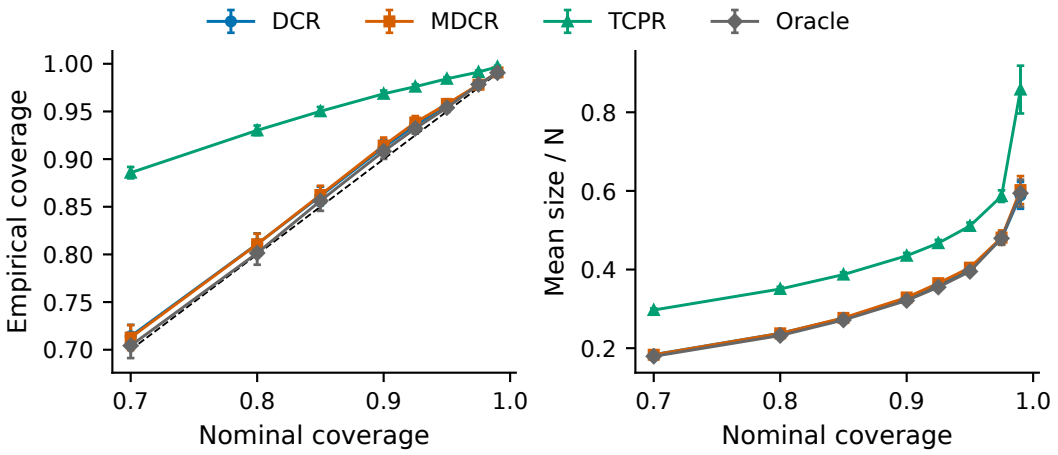

Figure 17: Yummly/ExtraTrees RA nominal-level sweep, 20 splits, with split-level 95% intervals. Larger target coverage trades off against size.

The synthetic nominal-level sweep isolates calibration from model fitting. Here $n = 500$, $m = 1000$, and $\sigma = 1$ remain fixed. Each of the eight levels is evaluated on the same complete trials for all methods. Oracle is a true-rank reference, while TCPR uses its rank envelope. The displayed intervals summarize variation between trials, not conditional Monte Carlo uncertainty.

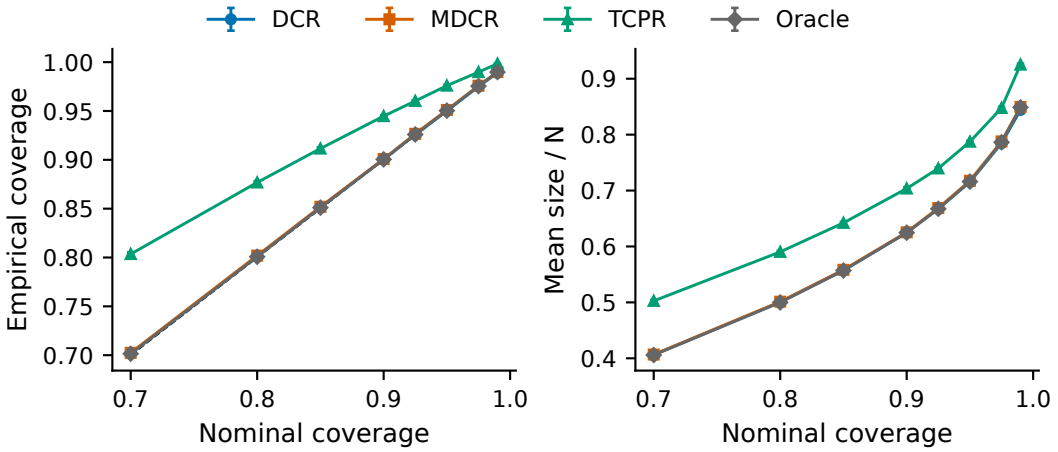

Figure 18: Synthetic RA nominal-level sweep, $n = 500, m = 1000, \sigma = 1$, 500 trials. Bars are complete-trial 95% intervals. The diagonal denotes target coverage.

F.3 Conditional Monte Carlo analysis

Fixed-pool repetitions isolate auxiliary CDF sampling error. We freeze three pools for each $n \in \{100, 500\}$ and $\sigma \in \{0.1, 1\}$, with $m = 500$. Before auxiliary draws, each pool receives 201 evenly spaced, rounded integer-distance grid points between 0 and $N - 1$. For each of seven budgets, 1,000 independent repetitions are processed in batches of 20. We save the exact auxiliary CDFs, empirical CDFs, squared errors, predicted variances, and sampled thresholds. The theoretical grid MSE is the average of (14); each empirical repetition is averaged over the same grid. Its repetition-level standard deviation divided by $\sqrt{1000}$ gives conditional MC SE. The curve in Figure 19 averages six fixed-pool conditional means for each n; its error bar combines independent MC errors, not variation

between randomly drawn populations. The scatter retains all 84 pool/budget cells. Zero-variance CDF endpoints are included in MSE averages without forming undefined pointwise ratios.

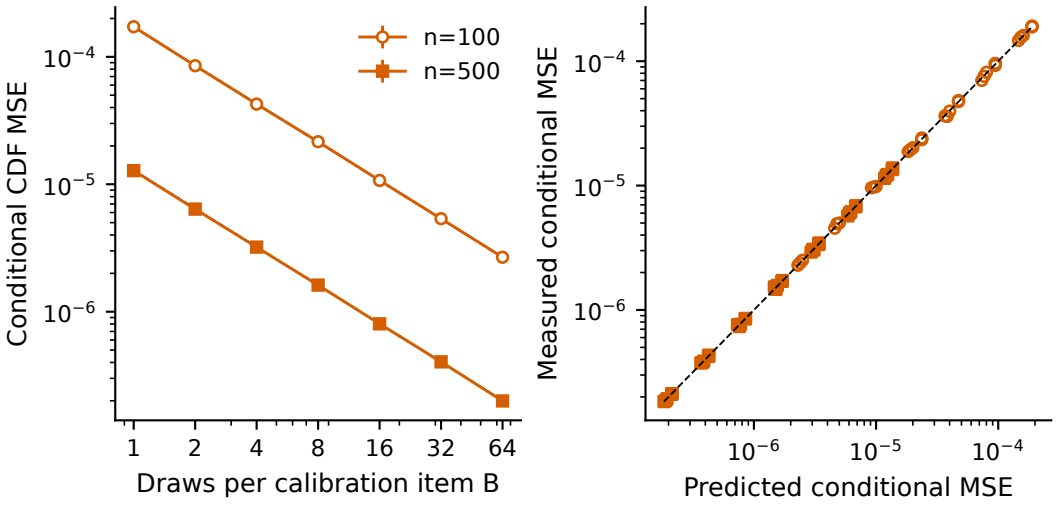

Figure 19: Conditional CDF MSE on 201 frozen integer-distance grid points. Left: empirical means across six fixed pools for each n, with ± 1.96 conditional MC SE; dashed curves: theory. Right: one theory–measurement pair per pool/budget (84 points), with identity line. There is no fitted or smoothed curve.

F.4 Inverse-CDF warm calibration

Computational gains depend on the calibrated operation. Figure 20 measures warm RA calibration with shared kernel and envelope caches. At $n = 500, m = 10{,}000$, median DCR/MDCR costs are $19.45/1.90$ ms; at $n = 10{,}000, m = 500$, they are $30.00/45.96$ ms. Thus MDCR is not uniformly faster in this implementation. Appendix F.5 separately reports cold kernel and envelope setup, training, and prediction costs. We do not extrapolate these operation timings to end-to-end costs or different samplers.

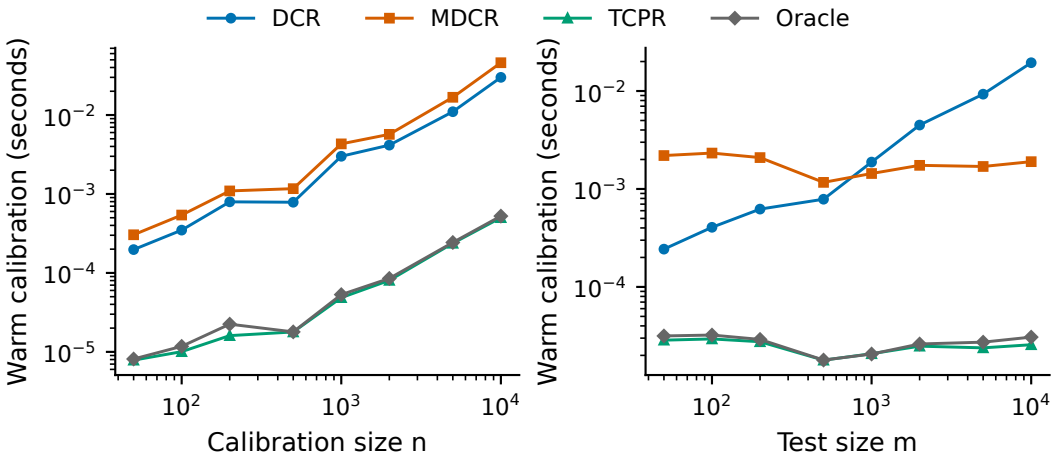

Figure 20: Floating-mixture implementation without systematic integer quantile certification. Median warm RA calibration time over 500 trials per setting, $\sigma = 1$, $\alpha = 0.1$. Left varies n at $m = 500$; right varies m at $n = 500$. Shared kernel and envelope construction, training, prediction, quantile lookup, and set evaluation are excluded. These are cached-operation costs, not end-to-end deployment times.

F.5 CPU COSTS OF THE REGRESSION AND SIMULATION STUDIES

Cached calibration and cold setup have different cost profiles. The fixed-anchor recipe and six-descriptor molecular study, fixed-predictor simulations, and conditional MC analysis used one CPU worker and single-threaded BLAS on an Apple M4 machine with ten available CPU cores, under Python 3.12.14, NumPy 2.3.5, SciPy 1.18.1, and scikit-learn 1.9.1. The three computational stages took 904.1 seconds in total: 282.8 for real-data training/evaluation, 562.6 for synthetic scans, and 58.7 for MC experiments. Peak resident memory across these stages was 1.44 GiB. Times include result/cache writing and incurred envelope construction, and exclude pilot work, plotting, manuscript preparation, and compilation. No formal trial failed or was removed.

Across 80 fitted predictors, recorded model fitting took 166.9 seconds and pool prediction 1.57 seconds; shared feature preprocessing and standardization took 76.7 seconds across 40 dataset/split pairs. These operation totals exclude other stage overhead. The 16 distinct TCPR envelopes took 50.0 seconds to construct and validate in total; 1 used the Bonferroni fallback. They are reused for all compatible pools.

Warm calibration records mixture aggregation/sorting for DCR, inverse-CDF draws and score sorting for MDCR, envelope-score sorting for TCPR, and true-score sorting for Oracle. Shared kernel/envelope setup, training, prediction, repeated quantile lookup, and test-set evaluation are excluded from Figure 20. RA kernels are cached equally for DCR and MDCR; the comparison does not measure an alternative direct sampler. Table 13 measures kernel construction separately by clearing the cache before each repetition. It is therefore inappropriate to infer end-to-end speedups from the warm curves alone. At $(n, m) = (500, 10000)$, median warm RA calibration takes 19.45 ms for DCR and 1.90 ms for MDCR; at $(10000, 500)$ it takes 30.00 and 45.96 ms, respectively. Thus this inverse-CDF implementation is not uniformly faster as calibration size grows.

Table 13: Cold kernel construction: median seconds over three independently cleared-cache calls; no training or envelope simulation included.

n	m	Seconds
200	300	0.0006
500	1000	0.0052
100	500	0.0005
500	500	0.0024
2000	500	0.0101
10000	500	0.0514
500	10000	0.0473

F.6 LIMITS OF ANALYTIC CERTIFICATION

Analytic FCP certificates can be uninformative despite small defects. A separate certificate experiment uses $m = 500$, $n = 50, 100, 200, 500$, $\sigma = 0, 0.1, 1$, and 200 complete trials per setting. We prescribe $\delta = \beta = 0.025$, $\eta_0 = 0.05$, and

$$d = \min\{m, \lceil N\sqrt{2v_{n,m}\log(8/(\delta\eta_0))}\rceil\}, \qquad \kappa = \begin{cases} 0, & d = m, \\ \eta_0, & d < m. \end{cases}$$

Equation (21) and Markov's inequality give $\Pr(B_d > \kappa) \le \delta$. We take $\omega_d = d/N$ and use (8); no threshold or radius is tuned using test outcomes. After calibration we evaluate true B_d and local mass only as oracle diagnostics. Across 4,800 DCR/MDCR evaluations, 88.8% of raw certificates are at least one and there are zero observed clipped-certificate violations. This is an assessment of a conservative bound, not a proof from zero failures. The large transport radius and partition error can erase information even when the defect is tiny.

Defects are evaluated by summing nonnegative kernel tail masses rather than subtracting a CDF from one. Independent exact combinatorial sums for three prescribed (n, m, d) cases agree with floating results to relative error below 1.1×10^{-13}, including values below 10^{-28}. These tiny probabilities do not justify a finely resolved scientific trend. We report certificate informativeness directly. Finite RA/VA quantile and set checks do not replace the proofs in Appendix A.