# DESIGN YIELD IS A TAIL FUNCTIONAL: WHY DISTRIBUTIONAL METRICS CANNOT RANK GENERATIVE DESIGNERS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Generative models for molecules, binders, and materials are commonly evaluated with distributional metrics and then deployed by drawing $N$ candidates and retaining the best. We study the mismatch between these two procedures. Best-of-$N$ yield depends on a generator only through the upper tail of its pushforward score distribution. For scores bounded in $[0, B]$, yield is Lipschitz in total variation with sharp constant $BN$, so the guarantee weakens linearly with the deployment budget. For scores unbounded above, yield has no modulus of continuity: metrics built from finitely many bounded test functions, including validity, uniqueness, novelty, FID-like scores, MMD, and precision/recall, can remain arbitrarily close while yield differs arbitrarily. The two metrics that respond to this construction, $W_1$ and forward KL to the data, penalise the same tail mass that improves design yield. We give a family with rank correlation $-1$ between distributional quality and yield, and construct generators with identical forward divergence and identical moments to any prescribed order whose sharp yield gap is determined by a linear program. We also evaluate a peaks-over-threshold estimator of $\mathbb{E}[M_N]$ and measure its calibration. Across 27 trained samplers from three architectures and five objectives, with yield held out from every predictor, the standard distribution-learning score is negatively correlated with design yield on three objectives. Two tail statistics are positively correlated with yield on all five, with bootstrap intervals excluding zero. On $\log P$, the sampler closest to the data trails the highest-yield sampler by about a hundred bootstrap standard errors. These quantities can be computed from samples that a generative benchmark already draws.

## 1 INTRODUCTION

Generative models in molecular and biomolecular design are often deployed as a filtering pipeline. A model proposes many candidates; a docking score, learned affinity model, interface-quality predictor, or synthesisability heuristic scores them; and only the best candidates advance. The relevant deployment quantity is therefore the best score obtained within a fixed budget.

Evaluation usually focuses on different quantities. Likelihood, divergence to the data distribution, validity, uniqueness, and novelty are all aggregate properties, and the benchmarks that popularised several of these metrics (Brown et al., 2019; Polykovskiy et al., 2020) report them directly. We study how much such distributional summaries can say about best-of-$N$ design yield.

**Setup.** A generator is a distribution $q$ on designs; $s$ is the score (higher is better); $F = s\#q$ is the pushforward. The pipeline draws $x_1, \dots, x_N \sim q$ and delivers $M_N = \max_{i \le N} s(x_i)$, with

$$\mathbb{P}(M_N \le z) = F(z)^N, \qquad \mathbb{E}[M_N] = \int z \, d\big(F(z)^N\big) \,. \tag{1}$$

Thus, yield depends on $q$ only through $F$, and at large $N$ it is governed by the upper tail of $F$ near the $(1 - 1/N)$ quantile.

**Contributions.**



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

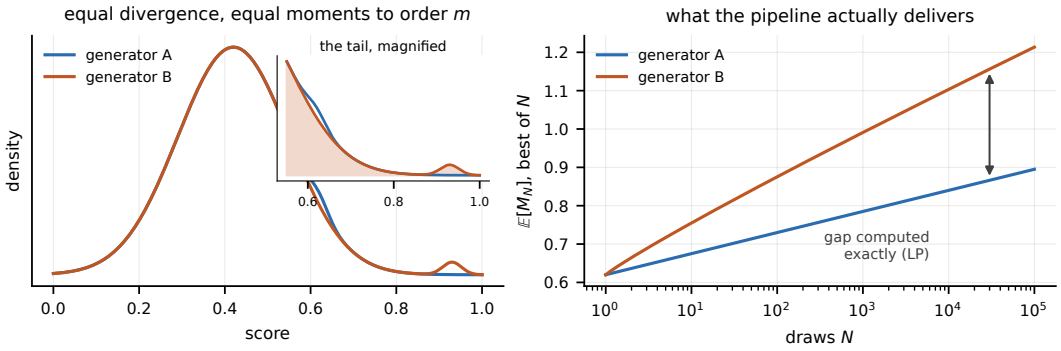


Figure 1: **Best-of-$N$ yield is determined by the score tail. Left:** two generators with identical forward divergence to the reference and identical score moments up to order $m$; the inset magnifies the only region that differs. **Right:** what a generate-and-filter pipeline delivers, $\mathbb{E}[M_N]$, separates as $N$ grows. Theorem 3 makes the construction exact and the sharp size of the gap is the value of a linear program.

1. **Continuity of design yield.** Theorem 1 and Corollary 2 show that, for scores bounded in $[0, B]$, yield is Lipschitz in total variation with sharp constant $BN$. For scores unbounded above, it has no modulus of continuity, and metrics built from finitely many bounded test functions—including validity, uniqueness, novelty, FID, MMD, precision/recall, and the GuacaMol distribution score—can be arbitrarily close while yield differs arbitrarily.
2. **Separation under matched moments.** Theorem 3 constructs generators with equal forward divergence and equal moments to any prescribed order but different best-of-$N$ yield. The sharp gap is given by a linear program. Proposition 4 gives a family with rank correlation exactly $-1$ between distributional quality and yield.
3. **Tail-based prediction.** Section 3 evaluates a peaks-over-threshold estimator of $\mathbb{E}[M_N]$ for budgets beyond the sampled pool and reports its empirical coverage.
4. **Surrogate tilting.** Proposition 5 shows that tilting toward a surrogate can improve the surrogate score while making true yield non-monotone, with an interior optimum that surrogate statistics alone cannot identify.
5. **Evaluation on trained generators.** Section 5.2 studies 48 samplers from three architectures on five objectives, with yield held out from every predictor. The standard distribution-learning score is negatively correlated with design yield on three objectives and positively correlated on two, whereas two tail statistics are positively correlated with yield on all five. On $\log P$, the sampler closest to the data is lower in yield by 2.20, about a hundred bootstrap standard errors.

## 2 The Separation

### 2.1 Yield is a discontinuous functional of the generator

Write $d_{\mathrm{TV}}(q, q') = \frac{1}{2}\|q - q'\|_1$, and let $d_{\mathrm{K}}(F, F') = \sup_z |F(z) - F'(z)|$ denote the Kolmogorov distance between the pushforward distributions. Since $d_{\mathrm{K}} \le d_{\mathrm{TV}}$, closeness of the generators in total variation implies closeness of their score distributions. The next result characterises when such closeness controls best-of-$N$ yield.

**Theorem 1** (continuity dichotomy)**.** *Fix a budget $N \ge 1$.*

*(i)* Bounded score. *If $s$ takes values in $[0, B]$ then for all $q, q'$*

$$\left|\mathbb{E}[M_N(q)] - \mathbb{E}[M_N(q')]\right| \;\le\; B\,N\,d_{\mathrm{K}}(F, F') \;\le\; B\,N\,d_{\mathrm{TV}}(q, q'), \tag{2}$$

*and the factor $N$ cannot be improved: for every $\delta \in (0, 1)$ there are $q, q'$ at $d_{\mathrm{TV}} = \delta$ whose gap is exactly $B\big(1 - (1-\delta)^N\big)$, which is $B\,N\delta\,(1 + o(1))$ as $N\delta \to 0$.*

*(ii)* Unbounded score. *If $s$ is unbounded above on the design space then for every $\delta > 0$ and every $Y > 0$ there are $q, q'$ with $d_{\mathrm{TV}}(q, q') \le \delta$ and $\mathbb{E}[M_N(q)] - \mathbb{E}[M_N(q')] \ge Y$. Hence $\mathbb{E}[M_N]$ admits no modulus of continuity in total variation, at any budget.*

**Corollary 2** (every bounded-test-class metric is blind)**.** *Let $D(q) = \Phi\big(\mathbb{E}_q[\varphi_1], \dots, \mathbb{E}_q[\varphi_k]\big)$ with $\|\varphi_j\|_\infty \le 1$ and $\Phi$ locally Lipschitz with constant $L$. If $s$ is unbounded above then for every $\varepsilon > 0$, every $N$ and every $Y > 0$ there are generators with $|D(q) - D(q')| \le \varepsilon$ and yield gap at least $Y$: apply Theorem 1(ii) with $\delta = \varepsilon/(2Lk)$.*

The corollary applies directly to several metrics used in generative-model evaluation. Validity, uniqueness, and novelty are averages of indicators. Precision, recall, density, and coverage (Sajjadi et al., 2018; Naeem et al., 2020) average support-membership indicators. FID (Heusel et al., 2017) is determined by the mean and covariance of a feature map, while MMD with a bounded kernel is a norm of a mean embedding. The GuacaMol distribution-learning score (Brown et al., 2019) aggregates divergences between descriptor histograms, which are themselves vectors of bounded-function averages. These quantities are continuous in total variation. Consequently, when the objective is unbounded above, they cannot certify best-of-$N$ yield at a fixed budget. Examples include $\log P$, penalised $\log P$, and unnormalised learned rewards.

**Metrics that respond to the tail can rank it in the wrong direction.** Two familiar quantities fall outside Corollary 2: the 1-Wasserstein distance and forward KL between the generator's score distribution and the data distribution. Neither has a bounded test class on an unbounded score, so both respond to the construction in Theorem 1(ii). In the instance we verify, two distributions at total variation $10^{-3}$ have a best-of-100 gap of $10^3$; every bounded test function changes by at most $2 \times 10^{-3}$, whereas $W_1$ changes by 10.5. The change is nevertheless in the direction that penalises the high-scoring tail, because that tail lies away from the data distribution. Proposition 4 gives an exact rank reversal, and §5.1 measures correlations of $\rho = -0.700$ and $-0.745$ on $\log P$. Thus, a distributional metric may either be insensitive to the relevant tail or penalise movement into it.

**Scope of the theorem.** The witness in Theorem 1(ii) is the mixture $(1-\delta)q' + \delta\,\delta_{x_H}$, which can be realised by a sampler that emits a fixed design with probability $\delta$. The construction therefore does not require an unrestricted class of probability measures. The theorem concerns what a metric can *certify*: generators that are nearly indistinguishable under the metric can have arbitrarily different yields. Section 5.2 separately measures the size of this effect among trained models.

**The bounded-score case.** Part (i) gives the corresponding positive result. When the score is bounded, closeness in distribution does control closeness in yield, with Lipschitz constant $BN$. The constant grows linearly with the deployment budget. For example, with $B = 1$ and $N = 10^3$, a Kolmogorov distance of $10^{-3}$ still permits a yield difference spanning the full range of the objective. Boundedness therefore restores continuity, but the resulting guarantee weakens in proportion to the campaign size.

### 2.2 Matching moments, and what it buys

Theorem 1(ii) moves a small amount of probability mass to a large score, so a statistic sensitive to large values, such as a high-order moment, may distinguish the constructed distributions. The next result quantifies how moment constraints affect the separation. Let $p$ be a reference distribution on a finite score support and let $\delta \in (0, 1)$.

**Theorem 3** (equal divergence, equal moments, unequal yield)**.** *Fix $m \ge 1$ and $\delta \in (0, 1)$. There exist generators $q_a, q_b$ with*

$$\mathrm{D}(p\|q_a) = \mathrm{D}(p\|q_b) = -\log(1-\delta), \qquad \mathbb{E}_{q_a}[s^k] = \mathbb{E}_{q_b}[s^k] \quad (k = 1, \dots, m), \tag{3}$$

*and $\mathbb{E}[M_N(q_a)] - \mathbb{E}[M_N(q_b)] \to \max \mathrm{supp}(q_a) - T_m^\star$ as $N \to \infty$, where $T_m^\star$ is the smallest support maximum of any measure of mass $\delta$ matching the same $m$ moments. $T_m^\star$ is the value of a linear program and is computed by bisection on that program.*

*Construction.* Let $q_a = q_b = (1-\delta)p$ on $\mathrm{supp}(p)$ and place the remaining mass $\delta$ off the support of $p$. Then $\mathrm{D}(p\|q) = \sum_x p(x) \log \frac{p(x)}{(1-\delta)p(x)} = -\log(1-\delta)$ for both, *independently* of where the

off-support mass goes: forward divergence is blind to the generator on sets the data do not occupy. Choose $q_a$'s off-support atoms freely and solve for $q_b$'s by the moment-matched LP with the support maximum minimised. □

The construction isolates the relevant blind spot: $\mathrm{D}(p\|q)$ does not depend on the part of $q$ outside the support of $p$, although best-of-$N$ sampling can be dominated by that mass. On a 40-atom reference with $\delta = 10^{-3}$, matching only the mean permits a yield gap of 13.2; matching the first two moments reduces the gap to 6.35, and matching the first three reduces it to 0.75. The corresponding sharp reachable maxima are $T_1^\star = 6.80$, $T_2^\star = 13.65$, and $T_3^\star = 19.25$, compared with $\max \mathrm{supp}(q_a) = 20$. Higher-order moment matching therefore constrains the tail, but progressively.

**Proposition 4** (perfect anti-alignment)**.** *There is a one-parameter family $\{q_t\}$ and a budget $N$ such that $t \mapsto \mathrm{D}(p\|q_t)$ and $t \mapsto \mathbb{E}[M_N(q_t)]$ are both strictly monotone* in the same direction*, so the rank correlation between model quality by the divergence (i.e. $-\mathrm{D}$) and design yield is exactly $-1$. In our instance the generator with $\mathrm{D} = 0.0010$ delivers $\mathbb{E}[M_{50}] = 2.15$ and the one with $\mathrm{D} = 0.2231$ delivers 8.00.*

## 3 What governs yield, and how well can it be predicted?

Extreme-value theory provides a direct model for the quantity of interest. If $F$ lies in the domain of attraction of a generalised extreme-value law with index $\xi$, then $(M_N - u_N)/a_N$ converges and $\mathbb{E}[M_N]$ is governed by $(u_N, a_N, \xi)$ (de Haan & Ferreira, 2006). We estimate the upper tail with a peaks-over-threshold model (Pickands, 1975), fitting a generalised Pareto distribution to exceedances. The same limit law underlies Pareto-smoothed importance sampling (Vehtari et al., 2024). For heavy-tailed cases, we also report the Hill estimator (Hill, 1975) as a check on $\hat{\xi}$. Related sample-size results for importance sampling (Chatterjee & Diaconis, 2018) reflect the same dependence on a tail, although there the relevant tail belongs to the importance weights rather than the design score. Following Balkema & de Haan (1974); Pickands (1975); Smith (1987); Coles (2001); Beirlant et al. (2004), we fit exceedances above the 95th percentile of a pool of 20,000 samples and predict $\mathbb{E}[M_N]$ at $N = 10^5$ and $10^6$. A moderate sample can therefore support projection to a larger campaign when the extrapolation is short or the tail is light.

**Calibration.** We evaluate coverage and bias over 30 independent pools per cell:

| | coverage of nominal 90% | | median rel. bias | | mean $\hat{\xi}$ | |
|---|---|---|---|---|---|---|
| tail | 0.70 dec. | 1.70 dec. | 0.70 dec. | 1.70 dec. | 0.70 dec. | 1.70 dec. |
| bounded (Beta 2,5) | 0.80 | 0.77 | 0.013 | 0.017 | $-0.236$ | $-0.236$ |
| light (normal) | 0.60 | 0.50 | 0.056 | 0.092 | $-0.129$ | $-0.129$ |
| light (exponential) | 0.80 | 0.83 | 0.071 | 0.096 | $+0.008$ | $+0.008$ |
| heavy (Pareto $\alpha$=3) | 0.80 | 0.80 | 0.222 | 0.286 | $+0.341$ | $+0.286$ |

For the bounded case, median relative bias is 1.3% at 0.70 decades and 1.7% at 1.70 decades. Examples of bounded design scores include interface predicted TM-score, normalised rewards, QED, and docking scores within a fixed pocket. Across all four tail families, nominal 90% bootstrap intervals achieve only 50–83% empirical coverage. We therefore report realised coverage rather than treating the nominal level as calibrated.

## 4 Surrogate tilting: the reward-hacking mechanism

**Proposition 5.** *Let $q_\lambda \propto p\, e^{\lambda \hat{s}}$ be the exponential tilt of $p$ towards a surrogate $\hat{s}$. Then $\lambda \mapsto \mathbb{E}_{q_\lambda}[\hat{s}]$ is strictly increasing. If $\hat{s}$ over-ranks a decoy relative to a rare true winner, then $\lambda \mapsto \mathbb{E}[M_N(q_\lambda)]$ under the* true *score is non-monotone, with an interior maximum and $\mathbb{E}[M_N] \to s(\text{decoy})$ as $\lambda \to \infty$. Tilting towards the true score $s$ itself can never decrease $\mathbb{E}[M_N]$ for any $N$.*

In the three-atom example, the surrogate mean rises from 0.95 to 4.00, while the true $\mathbb{E}[M_{100}]$ peaks at $\lambda = 0.25$ and then falls to 1.01, an 83% loss in yield. By contrast, tilting toward the true score produces a monotone increase in yield. The failure therefore arises from surrogate misranking in the rare, high-value region, and the yield-optimal tilt cannot be identified from surrogate statistics alone.

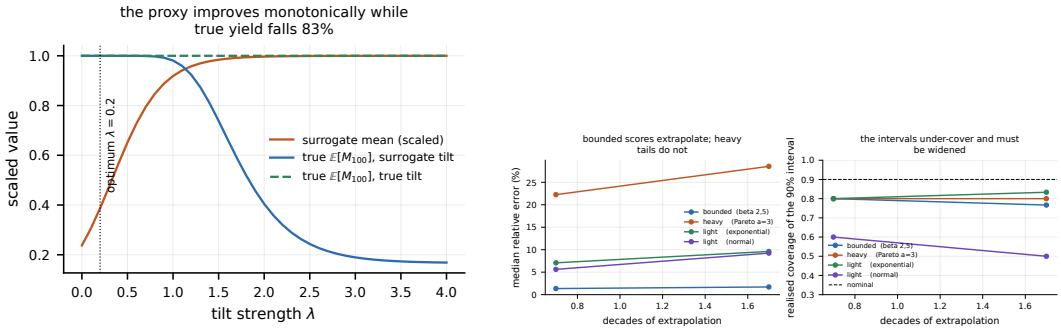


Figure 2: **Left:** reward-weighted fine-tuning against a surrogate. The surrogate's mean rises monotonically in the tilt while true best-of-100 peaks at $\lambda = 0.1$ and collapses; tilting the true score (dashed) never decreases yield. The optimum is interior and cannot be located from the surrogate's own statistics. **Right:** the calibration study. Point extrapolation is dependable for bounded scores and degrades with the extreme-value index; nominal 90% bootstrap intervals realise 50–83% coverage and must be widened.

## 5 Experiments on real score distributions

### 5.1 An exact construction on real molecular scores

Theorem 3 gives an exact construction; we next measure the effect on real molecular scores. We use the 49,662 distinct product molecules in USPTO-50k, with half designated as a training set so that novelty and distributional distances retain their standard interpretation, leaving a pool of 24,853 molecules. On this pool we construct eleven samplers representing several behaviours observed in trained models: tilting toward a surrogate property, mode collapse onto a molecular-weight window, tilting toward the true score, and exact matching of the training score histogram. Each sampler is an explicit probability vector, which makes best-of-$N$ yield and all reported metrics exactly computable. Tilt strengths are expressed in units of each score's standard deviation, allowing the same sampler family to be evaluated on three objectives: QED, the negated synthetic-accessibility score of Ertl & Schuffenhauer (2009), and $\log P$, which is unbounded above in this pool and reaches 21.7. Each predictor receives the same 1,000 scored draws and predicts $\mathbb{E}[M_{20000}]$. We keep all budgets below the finite pool size so that the experiment measures the sampler rather than exhaustion of the library (§B.1).

| predictor of $\mathbb{E}[M_{20000}]$, Spearman | QED | −SA | $\log P$ |
|---|---|---|---|
| −1-Wasserstein to train *(no scores)* | −0.027 | −0.064 | **−0.700** |
| −KL to train *(no scores)* | −0.045 | −0.118 | **−0.745** |
| sample mean | +0.955 | +0.800 | +0.882 |
| sample maximum | +0.991 | +0.973 | +0.909 |
| sample 99th percentile | +0.991 | +0.973 | +0.945 |
| generalised-Pareto tail fit | +0.582 | +0.939 | +0.964 |
| generators with an undefined tail fit | 0/11 | 1/11 | 4/11 |

The separation is also visible on the real score distributions. For $\log P$, the two distributional metrics are negatively correlated with design yield, at $-0.700$ and $-0.745$. High-$\log P$ molecules lie away from the training distribution, so closeness to the data and design yield induce opposing rankings in this example. Among predictors that use scored samples, the high sample quantile gives the strongest rank correlation, whereas the tail fit gives the smallest error in predicted magnitude (0.0481 versus 0.0835 on QED). The tail fit is undefined for samplers concentrated at the pool maximum: 4 of 11 samplers on $\log P$ and 1 of 11 on SA. We exclude those cases from the corresponding row rather than substituting a value, and report the counts in the table.

**Bounded and unbounded objectives.** Section 3 uses $\hat{\xi}$ to distinguish regimes for tail extrapolation. QED (Bickerton et al., 2012), interface predicted TM-score, and normalised rewards lie in $[0, 1]$;

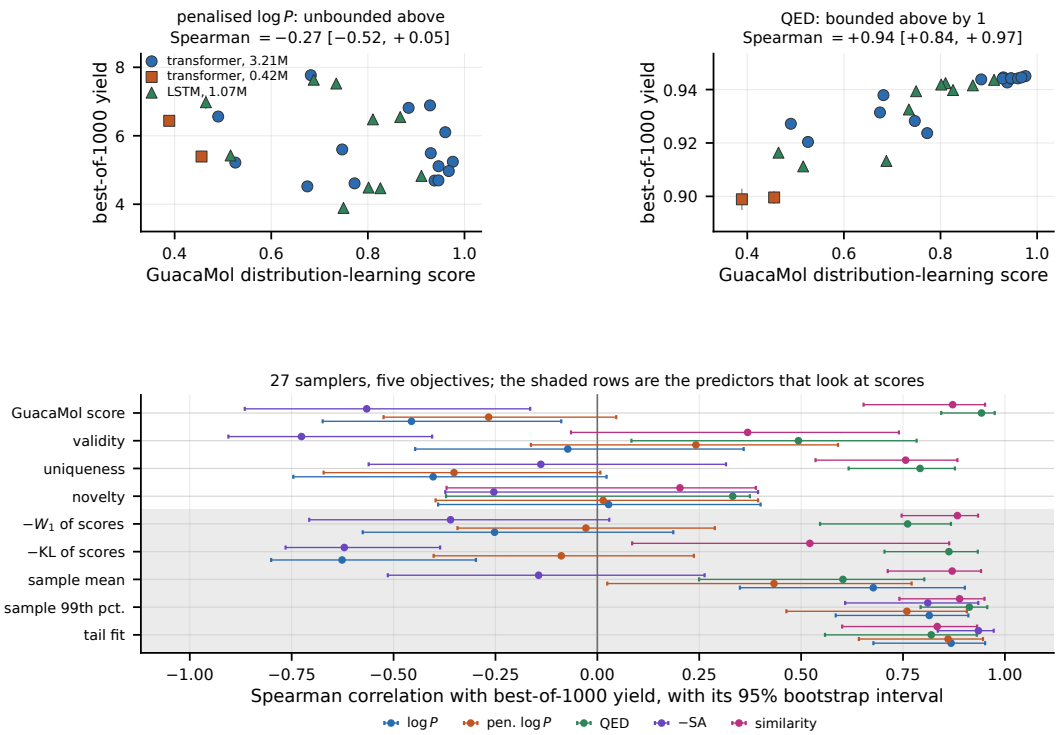


Figure 3: **A population of trained samplers. Top:** what a distribution-learning benchmark reports against what a campaign obtains, on an objective that is unbounded above and on one that is bounded. Marker shape is the architecture; error bars are three bootstrap standard errors of the yield, measured on draws held out from every predictor. **Bottom:** the rank correlation of each predictor with best-of-1000 yield, with 95% bootstrap intervals, for all five objectives. Shaded rows are the predictors that look at scores. The distribution score changes sign with the objective; the two tail statistics do not.

docking scores in a fixed pocket are also bounded by the finite geometry of the pocket. In our calibration study, a pool of $10^4$ samples predicts campaigns of $10^5$–$10^6$ to within a few percent for such bounded cases. Unnormalised learned rewards and ratio-like objectives need not be bounded, so their best-of-$N$ projections should be accompanied by uncertainty intervals.

**A tail-aware diagnostic.** Alongside distributional metrics, a generative-design benchmark can report (i) the estimated extreme-value index $\hat{\xi}$ of the score under each model, with a standard error, and (ii) the extrapolated $\mathbb{E}[M_N]$ at the deployment budget, with a bootstrap interval adjusted using the empirical calibration in Section 3. Both quantities use the same samples already generated for evaluation.

### 5.2 A POPULATION OF TRAINED GENERATORS

The constructed samplers above make the separation exact but do not establish whether it appears among fitted models. We therefore assemble a population of trained generators from three architectures: the 3.21M-parameter character-level transformer of §B.3, a 0.42M-parameter transformer, and a 1.07M-parameter LSTM of the kind commonly used as a SMILES baseline. We save multiple checkpoints from each architecture and sample each at four settings: temperatures 0.8, 1.0, and 1.2 at full support, and temperature 1.0 with nucleus $p = 0.9$. This yields **48 samplers**. All use the same USPTO product split, tokeniser, and reference set as §B.3, and each produces 40,000 strings.

**Predictors are evaluated on held-out yield.** We split each sampler's draw stream at its 1,000th valid molecule. All predictor quantities—validity, uniqueness, novelty, the GuacaMol distribution-learning score over nine descriptors, distance between the score histogram and the reference, sample

Table 1: **Rank correlations across trained samplers.** Rank correlation between each predictor and best-of-1000 design yield across 27 trained samplers. Bold marks a correlation whose 95% bootstrap interval over samplers excludes zero; the intervals themselves are drawn in Figure 3 and tabulated in §B.2. Rows marked † are the metrics a distribution-learning benchmark reports, and need no scores at all. Their sign is set by the objective rather than by the generator: negative on $\log P$, penalised $\log P$ and $-$SA, positive on QED and similarity. Boundedness does not predict which, since $-$SA is bounded. The two tail statistics are positive on all five.

| | *unbounded above* | | *bounded above* | | |
|---|---|---|---|---|---|
| Spearman with $\mathbb{E}[M_{1000}]$ | $\log P$ | pen. $\log P$ | QED | $-$SA | sim. |
| GuacaMol distribution score † | $\mathbf{-0.46}$ | $-0.27$ | $\mathbf{+0.94}$ | $\mathbf{-0.57}$ | $\mathbf{+0.87}$ |
| validity † | $-0.07$ | $+0.24$ | $\mathbf{+0.49}$ | $\mathbf{-0.73}$ | $+0.37$ |
| uniqueness † | $-0.40$ | $-0.35$ | $\mathbf{+0.79}$ | $-0.14$ | $\mathbf{+0.76}$ |
| novelty † | $+0.03$ | $+0.01$ | $+0.33$ | $-0.25$ | $+0.20$ |
| $-W_1$ of the score histogram | $-0.25$ | $-0.03$ | $\mathbf{+0.76}$ | $-0.36$ | $\mathbf{+0.88}$ |
| $-$KL of the score histogram | $\mathbf{-0.63}$ | $-0.09$ | $\mathbf{+0.86}$ | $\mathbf{-0.62}$ | $\mathbf{+0.52}$ |
| sample mean of scores | $\mathbf{+0.68}$ | $\mathbf{+0.43}$ | $\mathbf{+0.60}$ | $-0.14$ | $\mathbf{+0.87}$ |
| sample 99th percentile | $\mathbf{+0.81}$ | $\mathbf{+0.76}$ | $\mathbf{+0.91}$ | $\mathbf{+0.81}$ | $\mathbf{+0.89}$ |
| tail fit (§3) | $\mathbf{+0.87}$ | $\mathbf{+0.86}$ | $\mathbf{+0.82}$ | $\mathbf{+0.94}$ | $\mathbf{+0.83}$ |
| spread of the yield across samplers | 3.21 | 3.88 | 0.0461 | 0.479 | 0.153 |
| mean bootstrap s.e. of the yield | 0.0208 | 0.0281 | 0.000309 | 0.00185 | 0.00114 |

mean, sample 99th percentile, and tail fit—are computed before this cut. Thus every sampler contributes exactly 1,000 scored molecules to the predictors, independent of its validity rate. This controls predictor sample size across samplers. Validity remains a predictor in its own right, and the campaign-level analysis below separately charges for invalid draws. Yield $\mathbb{E}[M_N]$ is measured after the cut. We include a sampler if it produces at least $2 \times 1{,}000$ valid molecules in 40,000 draws; below this threshold there are too few valid samples to estimate both predictors and held-out yield. 21 of the 48 samplers are excluded: ten of the twelve settings of the smallest transformer, which never exceeds 5% validity at any checkpoint, and the earliest checkpoints of the other two model families. The most valid excluded sampler has validity $0.041$, while the least valid included sampler has validity $0.050$. The remaining 27 samplers span validity 0.05–0.99, uniqueness 0.07–1.00, and distribution score 0.39–0.98.

**The distribution score changes sign across objectives.** The GuacaMol distribution-learning score correlates with design yield at $+0.94$ $[+0.84, +0.97]$ on QED and $+0.87$ $[+0.65, +0.95]$ on similarity to a fixed target, but at $-0.46$ $[-0.67, -0.09]$ on $\log P$, $-0.27$ $[-0.52, +0.05]$ on penalised $\log P$, and $-0.57$ $[-0.87, -0.16]$ on negated SA score. Three of the five correlations are negative, and two of those three have bootstrap intervals excluding zero. The metric is unchanged across columns; only the objective changes. Validity, uniqueness, and novelty show the same qualitative variation and are individually weaker. By contrast, both tail statistics are positively correlated with yield on all five objectives, with intervals excluding zero.

**The sign depends on the location of high-scoring regions.** Boundedness alone does not determine whether a distributional score tracks yield. The negated SA score is bounded, yet its distribution score is negatively correlated with yield. This is consistent with Theorem 1(i): the factor $BN$ can already be uninformative at $N = 1000$. Empirically, the relevant distinction is whether the objective's high-scoring region lies near the data distribution or away from it. This property can be assessed from the reference set's own best-of-$N$ behaviour before fitting a model. Measured here, the reference set's own best-of-1000 minus the population median, in units of the spread of yield across the population, is $\log P$ $-0.06$, pen. $\log P$ $-0.05$, QED $+0.09$, $-$SA $-0.41$, sim. $+0.29$, and the sign of that difference agrees with the sign of the distribution score's correlation on 5 of the 5 objectives. When the data already nearly maximise an objective, moving toward the data also tends to improve that objective. When high-scoring designs lie away from the data, the two directions can oppose. This distinction is a property of the objective rather than the generator, and a metric computed only between generator and data does not encode it.

**A paired comparison.** On $\log P$, the sampler closest to the data among all 27 models is the best-trained checkpoint of the largest model at temperature $1.0$. It has distribution score 0.98 and yields $\mathbb{E}[M_{1000}] = 7.43$. The highest-yield sampler reaches 9.63 at distribution score 0.69, a yield gap of 2.20 compared with a mean bootstrap standard error of 0.021. On penalised $\log P$, the same closest-to-data sampler yields 5.24, whereas the highest-yield sampler reaches 7.77, a gap of 2.53. Thus the sampler selected by closeness to the data is not the sampler with the highest yield on these two objectives.

**Tail statistics at different budgets.** Both score-based tail predictors are positively correlated with yield across the objectives and budgets we examine. At $N = 100$, which lies within the sample available to each predictor, the sample 99th percentile gives the stronger rank statistic ($+0.93$ to $+0.96$ across the five objectives). At $N = 1000$, the tail fit gives the stronger result ($+0.82$ to $+0.94$ versus $+0.76$ to $+0.91$), as it extrapolates beyond the observed quantile. These results support using an empirical high quantile when the deployment budget lies within the sampled range and a calibrated tail fit when extrapolation is required, together with the coverage assessment in §3.

**Accounting for invalid draws.** The main analysis equalises the number of valid molecules used by each predictor. We also repeat the analysis at a fixed budget of $1000$ generated strings, so invalid outputs consume the same campaign budget as valid ones. Charging for invalid draws does help the distributional metrics, since validity then contributes to yield: the distribution score moves to $\log P$ $-0.26$, pen. $\log P$ $+0.19$, QED $+0.94$, $-$SA $-0.52$, sim. $+0.92$, changing sign on pen. $\log P$. What the accounting does not move is the ranking done by the tail statistics, which stay positive with intervals excluding zero on all five objectives ($+0.76$ to $+0.93$ and $+0.64$ to $+0.86$, as they are under the per-molecule accounting). Which accounting is used therefore changes how badly a distributional metric misleads, not whether a tail statistic is the thing to report.

## 6 Related work

**Distributional metrics for generative models.** Common evaluation tools include FID-style two-sample statistics (Heusel et al., 2017), precision/recall and density/coverage (Sajjadi et al., 2018; Naeem et al., 2020), and molecular benchmarks that aggregate related quantities (Brown et al., 2019; Polykovskiy et al., 2020). These metrics are functions of finitely many bounded-test-function averages, placing them within the scope of Corollary 2.

**Goal-directed benchmarks.** Molecular optimisation benchmarks already report extreme outcomes at fixed budgets. The goal-directed suite of Brown et al. (2019) scores the top-1, top-10, and top-100 molecules, while Gao et al. (2022) report the area under the top-$k$ curve within a fixed oracle budget. These are tail statistics, but a top-$k$ value at one budget does not by itself determine performance at a much larger deployment budget. Section 3 addresses that extrapolation problem and measures its calibration. Corollary 2 separately shows why distribution-learning metrics cannot substitute for goal-directed evaluation when the relevant objective is unbounded. Renz et al. (2019) document cases in which benchmark scores can improve without corresponding improvements in the generated designs; Theorem 1 and Proposition 5 provide two mechanisms that produce such a mismatch.

**Best-of-$N$ as an object of study.** In alignment, Beirami et al. (2024) study the KL divergence between a best-of-$N$ policy and its base policy, showing that the commonly quoted $\log N - (N-1)/N$ expression is an upper bound rather than an equality. For repeated sampling in language models, Brown et al. (2024) find that coverage—the fraction of problems solved by at least one of $k$ samples—is often log-linear in $k$ over four orders of magnitude. Schaeffer et al. (2025) explain related power laws from the distribution of per-problem success probabilities. Our analysis uses the upper tail of the score distribution in the corresponding role and asks which model-selection metrics preserve that information.

**Design as search over a generator.** Generate-and-filter pipelines appear throughout molecular and biomolecular design, including GFlowNets (Bengio et al., 2021; Jain et al., 2022), Bayesian optimisation in learned latent spaces (Stanton et al., 2022), and structure or sequence generators (Watson et al., 2023; Hayes et al., 2025). In these settings, the deployment outcome is governed by the

best candidates found within the sampling budget, making $\mathbb{E}[M_N]$ a natural summary of campaign yield.

**Tail-aware generative optimisation.** Concurrent work studies CVaR-constrained fine-tuning of flow models and related tail-shaping objectives (De Santi et al., 2025; Wang et al., 2026; Kopf et al., 2026). These methods modify the generator to improve tail behaviour. Our focus is complementary: we study when standard distributional metrics control tail-dependent yield and how that yield can be estimated from samples.

**Reward overoptimisation.** Gao et al. (2023) measure how optimisation against a learned reward model can diverge from the true objective as the KL budget grows. Related reward-gaming effects are formalised by Skalse et al. (2022) and arise in preference-based fine-tuning (Stiennon et al., 2020). Proposition 5 gives a design-yield analogue in which surrogate misranking is sufficient to create an interior optimum in true yield.

**Related sources of tail error.** Concurrent work estimates tail risk when samples come from a generative model (Zhao et al., 2026), while work on test-time scaling with noisy verifiers studies how verifier error limits best-of-$N$ selection (Dorner et al., 2025; Setlur et al., 2025). We hold the verifier fixed and exact, isolating the contribution of the generator's score distribution.

## 7 Conclusion

Generative design is often deployed as a search over samples, so its outcome is governed by the upper tail of the generator's score distribution. We show that standard distributional metrics do not generally control this tail-dependent quantity. For bounded objectives, yield is continuous but the sharp Lipschitz constant grows linearly with the campaign budget. For unbounded objectives, no modulus of continuity exists, so generators that are arbitrarily close under bounded-test-function metrics can have arbitrarily different best-of-$N$ yield.

The empirical results show the same distinction among trained samplers. Whether a distributional score correlates positively or negatively with yield depends on where the objective's high-scoring region lies relative to the data distribution, whereas the tail statistics remain positively correlated across all five objectives we study. A peaks-over-threshold fit provides a direct way to estimate yield beyond the observed sample, with calibration assessed empirically. Reporting the tail index together with projected best-of-$N$ yield therefore aligns evaluation more closely with the quantity a generate-and-filter campaign ultimately uses.

## AI use statement

GPT was used only for writing and language polishing.

## Reproducibility statement

**Metric pre-registration.** The primary metric is $\mathbb{E}[M_N]$, and all comparison metrics are computed on the same samplers. We report ranking performance and magnitude accuracy separately for every predictor, including the tail-based predictors. Coverage is reported empirically rather than only at its nominal level, and all budgets remain below the finite pool size. In §5.2, yield is measured on draws held out from every predictor under the inclusion rule stated with the result. The theory and calibration study are implemented in `code/s1_exact.py` (65 checks, 0 failures), which cross-checks the exact best-of-$N$ formula against an independent closed form to $10^{-13}$.

## References

August A Balkema and Laurens de Haan. Residual life time at great age. *The Annals of Probability*, 2(5):792–804, 1974. doi: 10.1214/aop/1176996548.

Ahmad Beirami, Alekh Agarwal, Jonathan Berant, Alexander D'Amour, Jacob Eisenstein, Chirag Nagpal, and Ananda Theertha Suresh. Theoretical guarantees on the best-of-$n$ alignment policy. *arXiv preprint arXiv:2401.01879*, 2024.

Jan Beirlant, Yuri Goegebeur, Johan Segers, and Jozef Teugels. Statistics of extremes. *Wiley Series in Probability and Statistics*, 2004. doi: 10.1002/0470012382.

Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow network based generative models for non-iterative diverse candidate generation. *Advances in Neural Information Processing Systems*, 34, 2021.

G. Richard Bickerton, Gaia V. Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L. Hopkins. Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4:90–98, 2012. doi: 10.1038/nchem.1243.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

Nathan Brown, Marco Fiscato, Marwin H. S. Segler, and Alain C. Vaucher. Guacamol: Benchmarking models for de novo molecular design. *Journal of Chemical Information and Modeling*, 59(3):1096–1108, 2019. doi: 10.1021/acs.jcim.8b00839.

Sourav Chatterjee and Persi Diaconis. The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2):1099–1135, 2018. doi: 10.1214/17-AAP1326.

Stuart Coles. An introduction to statistical modeling of extreme values. *Springer Series in Statistics*, 2001. doi: 10.1007/978-1-4471-3675-0.

Laurens de Haan and Ana Ferreira. *Extreme Value Theory*. Springer, 2006. doi: 10.1007/0-387-34471-3.

Riccardo De Santi et al. Flow density control: Generative optimization beyond entropy-regularized fine-tuning. In *Advances in Neural Information Processing Systems*, 2025.

Florian E. Dorner, Yatong Chen, André F. Cruz, and Fanny Yang. Roc-n-reroll: How verifier imperfection affects test-time scaling. 2025.

Peter Ertl and Ansgar Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics*, 1(1):8, 2009. doi: 10.1186/1758-2946-1-8.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. 2023.

Wenhao Gao, Tianfan Fu, Jimeng Sun, and Connor W Coley. Sample efficiency matters: a benchmark for practical molecular optimization. *Advances in Neural Information Processing Systems (Datasets and Benchmarks)*, 35, 2022.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025. doi: 10.1126/science.ads0018.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.

Bruce M. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174, 1975. doi: 10.1214/aos/1176343247.

Moksh Jain, Emmanuel Bengio, Alex-Hernandez Garcia, Jarrid Rector-Brooks, Bonaventure F P Dossou, Chanakya Ekbote, Jie Fu, Tianyu Zhang, Micheal Kilgour, Dinghuai Zhang, et al. Biological sequence design with gflownets. *International Conference on Machine Learning*, 2022.

Sarina Kopf, Cristina Nevado, and Philippe Schwaller. Sample efficient generative optimization for molecular design. 2026.

Andrey A. Markov. On certain applications of algebraic continued fractions. 1884. Thesis, St Petersburg; the moment problem used in Theorem 1 is the modern grid form, see Karlin and Studden (1966).

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. *International Conference on Machine Learning*, 2020.

James Pickands. Statistical inference using extreme order statistics. *The Annals of Statistics*, 3(1): 119–131, 1975. doi: 10.1214/aos/1176343003.

Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, et al. Molecular sets (moses): A benchmarking platform for molecular generation models. *Frontiers in Pharmacology*, 11:565644, 2020. doi: 10.3389/fphar.2020.565644.

Philipp Renz, Dries Van Rompaey, Jörg Kurt Wegner, Sepp Hochreiter, and Günter Klambauer. On failure modes in molecule generation and optimization. *Drug Discovery Today: Technologies*, 32–33:55–63, 2019. doi: 10.1016/j.ddtec.2020.09.003.

Mehdi S M Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *Advances in Neural Information Processing Systems*, 31, 2018.

Rylan Schaeffer, Joshua Kazdan, John Hughes, Jordan Juravsky, Sara Price, Aengus Lynch, Erik Jones, Robert Kirk, Azalia Mirhoseini, and Sanmi Koyejo. How do large language monkeys get their power (laws)? *International Conference on Machine Learning*, 2025.

Amrith Setlur, Nived Rajaraman, Sergey Levine, and Aviral Kumar. Scaling test-time compute without verification or rl is suboptimal. 2025.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35, 2022.

Richard L Smith. Estimating tails of probability distributions. *The Annals of Statistics*, 15(3): 1174–1207, 1987. doi: 10.1214/aos/1176350499.

Samuel Stanton, Wesley Maddox, Nate Gruver, Phillip Maffettone, Emily Delaney, Peyton Greenside, and Andrew Gordon Wilson. Accelerating bayesian optimization for biological sequence design with denoising autoencoders. *International Conference on Machine Learning*, 2022.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33, 2020.

Aki Vehtari, Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry. Pareto smoothed importance sampling. *Journal of Machine Learning Research*, 25(72):1–58, 2024.

Zifan Wang, Riccardo De Santi, Xiaoyu Mo, Michael M. Zavlanos, Andreas Krause, and Karl H. Johansson. Efficient tail-aware generative optimization via flow model fine-tuning. 2026.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023. doi: 10.1038/s41586-023-06415-8.

Shuning Zhao, Patrick Wong, Leran Zhang, and Xiaolin Hu. Diachronic sample integration: Robust tail-risk estimation with generative models. 2026.

# A PROOFS

## A.1 PROOF OF THEOREM 1 AND COROLLARY 2

*(i)* For a score in $[0, B]$, $\mathbb{E}[M_N] = \int_0^B \big(1 - F(z)^N\big)dz$, so

$$\big|\mathbb{E}[M_N(q)] - \mathbb{E}[M_N(q')]\big| \le \int_0^B \big|F(z)^N - F'(z)^N\big|dz$$
$$\le \int_0^B N \max(F, F')^{N-1}\big|F - F'\big|dz \;\le\; B\,N\,d_{\mathrm{K}}(F, F'),$$

using $|a^N - b^N| \le N \max(a, b)^{N-1}|a - b|$ and $\max(F, F') \le 1$, and $d_{\mathrm{K}} \le d_{\mathrm{TV}}$ holds because $|F(z) - F'(z)| = |q(s \le z) - q'(s \le z)|$ is the discrepancy of a single event. For sharpness take $q' = \delta_{x_0}$ with $s(x_0) = 0$ and $q = (1-\delta)\delta_{x_0} + \delta\,\delta_{x_B}$ with $s(x_B) = B$. Then $d_{\mathrm{TV}} = \delta$, $\mathbb{E}[M_N(q')] = 0$ and $\mathbb{E}[M_N(q)] = B\big(1 - (1-\delta)^N\big)$, whose ratio to $BN\delta$ tends to 1 as $N\delta \to 0$.

*(ii)* Fix any $q'$ with $\mathbb{E}[M_N(q')]$ finite, any $\delta > 0$, and $x_H$ with $s(x_H) = H$ above the essential supremum of $s$ under $q'$; such $x_H$ exists for every $H$ because $s$ is unbounded above. Put $q = (1-\delta)q' + \delta\,\delta_{x_H}$, so $d_{\mathrm{TV}}(q, q') = \delta\,d_{\mathrm{TV}}(\delta_{x_H}, q') \le \delta$. Conditioning $N$ independent draws from $q$ on whether $x_H$ appears, and noting that $q$ conditioned on $\{x \ne x_H\}$ is exactly $q'$,

$$\mathbb{E}[M_N(q)] = \big(1 - (1-\delta)^N\big)H + (1-\delta)^N\,\mathbb{E}[M_N(q')], \tag{4}$$

which exceeds $\mathbb{E}[M_N(q')] + Y$ once $H \ge \big(Y + |\mathbb{E}[M_N(q')]|\big)/\big(1 - (1-\delta)^N\big)$. Since $\delta$ was arbitrary, no function of $d_{\mathrm{TV}}$ can bound the gap.

*Corollary.* $|D(q) - D(q')| \le L\sum_{j \le k}\big|\mathbb{E}_q\varphi_j - \mathbb{E}_{q'}\varphi_j\big| \le 2Lk\,d_{\mathrm{TV}}(q, q')$, because $|\mathbb{E}_q\varphi - \mathbb{E}_{q'}\varphi| \le \|\varphi\|_\infty\|q - q'\|_1 = 2\|\varphi\|_\infty d_{\mathrm{TV}}$. Applying *(ii)* with $\delta = \varepsilon/(2Lk)$ gives a pair inside the $\varepsilon$-sublevel set of $D$ with yield gap at least $Y$. □

**Verification.** We check Equation (2) against the exact best-of-$N$ formula on 200 random pairs at six budgets; the largest ratio to the bound is 0.66. The sharp family reproduces $B(1 - (1-\delta)^N)$ to $10^{-12}$, and its ratio to $BN\delta$ approaches 1 to within $10^{-6}$. Equation (4) is checked to $10^{-9}$. In the example from §2.1, $d_{\mathrm{TV}} = 10^{-3}$ and the best-of-100 gap is $1.003 \times 10^3$; the largest change among five bounded test functions is $9.4 \times 10^{-4}$, while $W_1$ of the score distribution changes by 10.5.

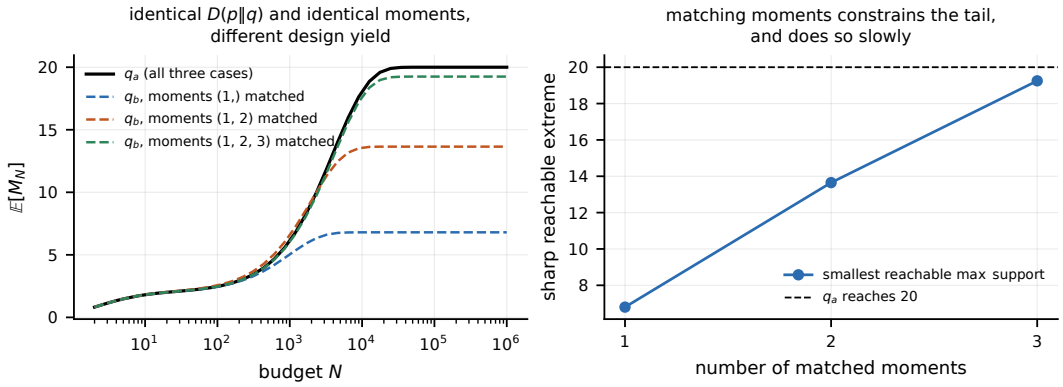


Figure 4: **Left:** best-of-$N$ for the two generators of Theorem 3. $q_a$ is the same curve in all three cases; $q_b$ is the moment-matched competitor with the smallest reachable extreme. Both have forward divergence exactly $-\log(1-\delta)$ to the reference and matched moments to the stated order, and the yield gap survives. **Right:** the sharp reachable extreme as a function of how many moments are matched, computed by the grid Markov moment linear program. Matching additional moments raises the smallest reachable extreme.

## A.2 THE MOMENT-CONSTRAINED EXTREME, AND WHY THEOREM 3 NEEDS AN LP

Fix mass $\delta$, orders $m_1 < \cdots < m_r$, and targets $\tau_k$. The set of non-negative measures $w$ on a finite grid $\{h_1, \ldots, h_G\}$ satisfying $\sum_i w_i = \delta$ and $\sum_i w_i h_i^{m_k} = \tau_k$ is a polytope. For any threshold $T$,

requiring the measure to be supported on $\{h_i \le T\}$ gives a linear feasibility problem. Feasibility is monotone in $T$, so bisection over $T$ with a linear program at each step returns the exact $T^\star$ used in our implementation. The corresponding continuum problem is the classical Markov moment problem (Markov, 1884). Matching $r$ moments implies $T^\star \ge (\tau_{m_r}/\tau_{m_{r-1}})^{1/(m_r - m_{r-1})}$ through the support bound $h^{m_r} \le T^{m_r - m_{r-1}} h^{m_{r-1}}$. Additional moments therefore raise the minimum reachable support maximum through ratios of moments. In our instance, $T_1^\star = 6.80$, $T_2^\star = 13.65$, and $T_3^\star = 19.25$, compared with $\max \mathrm{supp}(q_a) = 20$.

### A.3 Proof of Proposition 4

Take $p$ on a finite score support, an extra atom at $h$ above that support, and $q_t = (1-t)p + t\delta_h$. Then $\mathrm{D}(p\|q_t) = -\log(1-t)$, strictly increasing in $t$, and $\mathbb{E}[M_N(q_t)]$ is strictly increasing in $t$ because $\mathbb{P}(M_N \le z) = F_t(z)^N$ with $F_t$ strictly decreasing in $t$ at every $z < h$. So model quality by the divergence, $-\mathrm{D}(p\|q_t)$, is strictly *decreasing* in $t$ while yield is strictly increasing, and the two rank orderings are exact reverses. □

### A.4 Proof of Proposition 5

$\frac{d}{d\lambda}\mathbb{E}_{q_\lambda}[\hat{s}] = \mathrm{Var}_{q_\lambda}(\hat{s}) > 0$ whenever $\hat{s}$ is non-constant on the support, giving strict monotonicity of the surrogate mean. As $\lambda \to \infty$, $q_\lambda$ concentrates on $\arg\max \hat{s}$, so $M_N \to s(\arg\max \hat{s})$ almost surely; if the surrogate's maximiser is the decoy then this limit is $s(\text{decoy})$, which is below $\mathbb{E}[M_N(q_0)]$ for $N$ large enough that $q_0$ finds the true winner. Since $\mathbb{E}[M_N(q_\lambda)]$ is continuous in $\lambda$ and exceeds its limit somewhere, it is non-monotone and attains an interior maximum. For the last claim, tilting in $s$ itself gives $q_\lambda(x)/q_\lambda(y) = \big(q_0(x)/q_0(y)\big)e^{\lambda(s(x)-s(y))}$, which is non-decreasing in $\lambda$ whenever $s(x) > s(y)$; hence $F_\lambda$ decreases pointwise in $\lambda$ and $\mathbb{E}[M_N]$ is non-decreasing for every $N$. □

## B Experimental detail

### B.1 The constructed-sampler study in full

**Budgets below the library size.** At $N = 10^5$ with a pool of 24,853 molecules, samplers with full support effectively exhaust the finite library and return the pool maximum. The resulting yield has little dynamic range and reflects library saturation rather than sampler quality. We therefore keep evaluation budgets well below the pool size and report correlations as a function of budget.

**Distributional metrics on QED.** Across budgets from 10 to $10^4$, the Spearman correlation between design yield and the reported distributional metrics is small on QED. At $N = 10$, the correlations are $+0.018$ for the 1-Wasserstein distance to the training score distribution, $-0.027$ for KL, and $-0.055$ for the distinct fraction. Over the same range, the yield spread across samplers decreases from 0.88 at $N = 10$ to 0.36 at $N = 10^4$. The mean of the true score reaches a Spearman correlation of $+0.991$.

**Ranking and magnitude use different statistics.** Each predictor below receives the same 1,000 scored draws and predicts $\mathbb{E}[M_{20000}]$ on QED, where the yield spread is 0.2813:

| predictor | Spearman | mean abs. error |
|---|---|---|
| $-$1-Wasserstein to train *(needs no scores)* | $-0.027$ | — |
| $-$KL to train *(needs no scores)* | $-0.045$ | — |
| sample mean | $+0.955$ | 0.3452 |
| sample maximum | $+\mathbf{0.991}$ | 0.0524 |
| sample 99th percentile | $+\mathbf{0.991}$ | 0.0835 |
| generalised-Pareto tail fit | $+0.582$ | $\mathbf{0.0481}$ |

Predictors that use the scored samples are substantially more informative than the two distribution-to-reference metrics in this experiment. Among score-based predictors, the sample maximum and 99th percentile give the strongest rank correlation, both at $+0.991$, compared with $+0.582$ for the generalised-Pareto fit. The tail fit instead gives the smallest error in predicted magnitude, 0.0481 compared with 0.0835 for the 99th percentile, because it extrapolates beyond the observed quantile.

### B.2 THE POPULATION STUDY: INTERVALS AND THE SMALLER BUDGET

Table 1 reports point estimates. The following tables give the corresponding 95% bootstrap intervals at the main-text budget and at a budget ten times smaller.

Table 2: As Table 1, with 95% bootstrap intervals over samplers.

| | *unbounded above* | | *bounded above* | | |
|---|---|---|---|---|---|
| Spearman with $\mathbb{E}[M_{1000}]$ | $\log P$ | pen. $\log P$ | QED | $-$SA | sim. |
| GuacaMol distribution score † | $\mathbf{-0.46}$ [−0.67, −0.09] | −0.27 [−0.52, +0.05] | $\mathbf{+0.94}$ [+0.84, +0.97] | $\mathbf{-0.57}$ [−0.87, −0.16] | $\mathbf{+0.87}$ [+0.65, +0.95] |
| validity † | −0.07 [−0.45, +0.36] | +0.24 [−0.16, +0.59] | $\mathbf{+0.49}$ [+0.08, +0.78] | $\mathbf{-0.73}$ [−0.91, −0.41] | +0.37 [−0.06, +0.74] |
| uniqueness † | −0.40 [−0.75, +0.02] | −0.35 [−0.67, +0.01] | $\mathbf{+0.79}$ [+0.62, +0.88] | −0.14 [−0.56, +0.32] | $\mathbf{+0.76}$ [+0.54, +0.88] |
| novelty † | +0.03 [−0.39, +0.40] | +0.01 [−0.40, +0.39] | +0.33 [−0.37, +0.37] | −0.25 [−0.37, +0.39] | +0.20 [−0.37, +0.39] |
| $-W_1$ of the score histogram | −0.25 [−0.58, +0.19] | −0.03 [−0.34, +0.29] | $\mathbf{+0.76}$ [+0.55, +0.87] | −0.36 [−0.71, +0.03] | $\mathbf{+0.88}$ [+0.75, +0.93] |
| −KL of the score histogram | $\mathbf{-0.63}$ [−0.80, −0.30] | −0.09 [−0.40, +0.24] | $\mathbf{+0.86}$ [+0.70, +0.93] | $\mathbf{-0.62}$ [−0.76, −0.39] | $\mathbf{+0.52}$ [+0.09, +0.86] |
| sample mean of scores | $\mathbf{+0.68}$ [+0.35, +0.90] | $\mathbf{+0.43}$ [+0.02, +0.77] | $\mathbf{+0.60}$ [+0.25, +0.80] | −0.14 [−0.51, +0.26] | $\mathbf{+0.87}$ [+0.71, +0.94] |
| sample 99th percentile | $\mathbf{+0.81}$ [+0.58, +0.91] | $\mathbf{+0.76}$ [+0.46, +0.91] | $\mathbf{+0.91}$ [+0.79, +0.96] | $\mathbf{+0.81}$ [+0.61, +0.93] | $\mathbf{+0.89}$ [+0.74, +0.95] |
| tail fit (§3) | $\mathbf{+0.87}$ [+0.68, +0.95] | $\mathbf{+0.86}$ [+0.64, +0.95] | $\mathbf{+0.82}$ [+0.56, +0.93] | $\mathbf{+0.94}$ [+0.84, +0.97] | $\mathbf{+0.83}$ [+0.60, +0.93] |
| spread of the yield across samplers | 3.21 | 3.88 | 0.0461 | 0.479 | 0.153 |
| mean bootstrap s.e. of the yield | 0.0208 | 0.0281 | 0.000309 | 0.00185 | 0.00114 |

Table 3: The same at $N = 100$, a budget well inside the sample each predictor was given. The sample quantile is the stronger rank statistic here; the tail fit overtakes it at the larger budget.

| | *unbounded above* | | *bounded above* | | |
|---|---|---|---|---|---|
| Spearman with $\mathbb{E}[M_{100}]$ | $\log P$ | pen. $\log P$ | QED | $-$SA | sim. |
| GuacaMol distribution score † | $\mathbf{-0.59}$ [−0.80, −0.23] | −0.20 [−0.52, +0.12] | $\mathbf{+0.94}$ [+0.83, +0.97] | $\mathbf{-0.76}$ [−0.94, −0.47] | $\mathbf{+0.90}$ [+0.74, +0.96] |
| validity † | +0.03 [−0.40, +0.45] | $\mathbf{+0.48}$ [+0.12, +0.74] | $\mathbf{+0.55}$ [+0.16, +0.81] | $\mathbf{-0.67}$ [−0.89, −0.29] | $\mathbf{+0.48}$ [+0.07, +0.79] |
| uniqueness † | $\mathbf{-0.61}$ [−0.82, −0.31] | $\mathbf{-0.48}$ [−0.73, −0.13] | $\mathbf{+0.76}$ [+0.57, +0.85] | −0.33 [−0.71, +0.11] | $\mathbf{+0.74}$ [+0.52, +0.87] |
| novelty † | +0.03 [−0.40, +0.41] | +0.09 [−0.39, +0.39] | +0.27 [−0.37, +0.38] | −0.25 [−0.37, +0.39] | +0.31 [−0.38, +0.39] |
| $-W_1$ of the score histogram | −0.33 [−0.68, +0.14] | +0.02 [−0.32, +0.35] | $\mathbf{+0.78}$ [+0.57, +0.88] | $\mathbf{-0.58}$ [−0.82, −0.24] | $\mathbf{+0.90}$ [+0.78, +0.94] |
| −KL of the score histogram | $\mathbf{-0.65}$ [−0.83, −0.35] | −0.00 [−0.34, +0.34] | $\mathbf{+0.85}$ [+0.68, +0.92] | $\mathbf{-0.65}$ [−0.78, −0.41] | $\mathbf{+0.52}$ [+0.08, +0.87] |
| sample mean of scores | $\mathbf{+0.85}$ [+0.65, +0.95] | $\mathbf{+0.61}$ [+0.19, +0.89] | $\mathbf{+0.62}$ [+0.28, +0.84] | +0.08 [−0.33, +0.45] | $\mathbf{+0.91}$ [+0.79, +0.95] |
| sample 99th percentile | $\mathbf{+0.94}$ [+0.82, +0.98] | $\mathbf{+0.96}$ [+0.86, +0.98] | $\mathbf{+0.93}$ [+0.83, +0.97] | $\mathbf{+0.94}$ [+0.84, +0.98] | $\mathbf{+0.95}$ [+0.87, +0.97] |
| tail fit (§3) | $\mathbf{+0.77}$ [+0.59, +0.86] | $\mathbf{+0.72}$ [+0.39, +0.92] | $\mathbf{+0.80}$ [+0.52, +0.92] | $\mathbf{+0.90}$ [+0.76, +0.96] | $\mathbf{+0.81}$ [+0.57, +0.92] |
| spread of the yield across samplers | 3.22 | 4.7 | 0.0857 | 0.479 | 0.117 |
| mean bootstrap s.e. of the yield | 0.0341 | 0.041 | 0.000861 | 0.00456 | 0.00133 |

### B.3 THE SINGLE-ARCHITECTURE PILOT

We first ran a smaller study with a single architecture before the population experiment in §5.2. A character-level decoder is trained on USPTO product SMILES (3.21M parameters, 14,797 molecules, 90 epochs). Checkpoints at epochs 22, 45, and 90 are each evaluated with two sampling settings: $T = 0.7$, and nucleus sampling with $p = 0.9$ at $T = 1.0$. The resulting six samplers span validity from 0.41 to 0.80 and uniqueness from 0.04 to 0.95. Each setting is sampled once and then scored under both objectives, so the sampling cost does not depend on the objective.

Table 4: Six trained samplers, ranked against best-of-1000 design yield. Under $\log P$, which is unbounded above in this pool, the distributional metrics are uninformative or anti-correlated; under QED, which is bounded near 0.95, they happen to agree.

| Spearman with $\mathbb{E}[M_{1000}]$ | $\log P$ | QED |
|---|---|---|
| $-\mathrm{W}_1$ to the reference | +0.143 | +0.829 |
| −KL to the reference | $\mathbf{-0.714}$ | +0.771 |
| validity | +0.429 | −0.543 |
| uniqueness | +0.486 | +0.943 |
| novelty | −0.486 | −0.486 |
| sample mean of scores | +0.029 | −0.429 |
| sample 99th percentile | $\mathbf{+0.829}$ | +0.943 |
| peaks-over-threshold prediction | +0.486 | +0.943 |

On $\log P$, yield ranges from 6.42 to 8.75, with a mean bootstrap standard error of 0.010. The checkpoint with the smallest divergence to the reference (KL = 0.0411) yields 7.63, whereas a

sampler with divergence $0.7197$ yields $\mathbf{8.75}$. Across the six samplers, the tail statistic has Spearman correlation $+0.829$ with yield, while the sample mean has correlation $+0.029$.

Under QED, the same six samplers show a different pattern. QED is bounded near $0.95$, best-of-$1000$ approaches saturation, and the yields span only $0.032$. With only six samplers, the corresponding rank correlations have wide uncertainty intervals; this is one reason for the larger population study in §5.2.

At $p = 0.9$, nucleus sampling also produces a distinct finite-sample effect. Uniqueness falls to $0.04$–$0.11$, the same high-$\log P$ molecule is drawn hundreds of times, and best-of-$1000$ becomes deterministic, with bootstrap standard error $0.0000$. Here the small uncertainty reflects diversity collapse rather than a broader or better-resolved tail.

# C Discussion

## C.1 Why forward divergence is the relevant blind spot

Theorem 3 uses the fact that $\mathrm{D}(p\|q)$ does not depend on the mass that $q$ places outside the support of $p$. The reverse divergence $\mathrm{D}(q\|p)$ behaves differently: it becomes $+\infty$ whenever $q$ assigns positive mass outside the support of $p$. This does not provide a useful design metric in the setting considered here. Likelihood-based training and several common evaluations are forward-type comparisons to the data, while a design pipeline explicitly seeks novel candidates outside the observed data support. Section 3 therefore evaluates the relevant tail directly rather than replacing one divergence with another.