Multi-Oracle Agreement Reveals the Limits of Self-Consistency Evaluation in RNA Design

Minghao Sun[1] **Hanqun Cao**[2] **Fang Wu**[3] **Zhou Zhang**[4]
Zhiyuan Liu[1] **Tianfan Fu**[4] **Pheng-Ann Heng**[2] **Yang Zhang**[1*]
[1]NUS [2]CUHK [3]Stanford [4]NJU

Abstract

Computational RNA design methods routinely evaluate designed sequences by refolding them with the same structure-prediction model used during optimization, measuring agreement with the design target. While scalable, this practice carries a fundamental risk: when optimization and evaluation share the same underlying model, high scores can reflect model-specific idiosyncrasies rather than genuine structural quality. To characterize the severity and mechanism of this problem, we audited nine RNA design methods spanning four algorithmic paradigms, evaluating approximately 11,000 sequences against predictors from two independent model families: physics-based thermodynamic models and machine-learning models, alongside five tertiary-structure predictors and large-scale SHAPE chemical-mapping data. We find that 78.1% of designs produced by single-objective optimization pass one predictor yet fail another, while ensemble-based methods reduce this disagreement 30-fold. A controlled experiment identifies optimization objective breadth, not search strategy, as the primary driver of this fragility. Cross-predictor disagreement carries a calibrated experimental signal: lower disagreement predicts higher SHAPE accuracy across 40,000 sequences, yet remains null-to-anti-predictive for biological function, establishing a clear scope boundary. We release **ACCORD** (**A**cross-oracle **C**alibration for **CO**mputational **R**NA **D**esign), a benchmarking workflow that reports per-predictor scores, experimentally calibrated uncertainty tiers, and explicit warnings for invalid use cases, providing the community with a principled foundation for RNA design evaluation.

1 Introduction

RNA has emerged as a central target for both therapeutic design and synthetic biology, with applications ranging from mRNA vaccines to ribozyme-based gene regulators. Realizing this potential requires computational methods that can reliably design sequences folding into prescribed structures, and equally reliable methods to evaluate whether they succeed. In practice, most RNA design pipelines assess their outputs by refolding a designed sequence with the same structure-prediction model used during optimization and measuring agreement with the intended target. This approach is cheap and scalable, but introduces a critical blind spot: a model optimized against a specific predictor will naturally produce sequences that score well under that predictor, regardless of whether those sequences are genuinely well-designed. The result is an evaluation system that rewards conformity to a single model's assumptions rather than true structural quality, a Goodhart failure mode whose severity grows with optimization pressure [1, 2].

This failure mode is particularly acute in RNA because, unlike protein structure prediction after AlphaFold [3], RNA secondary-structure prediction remains substantially heterogeneous across model

*Corresponding author: zhang@nus.edu.sg

Preprint.

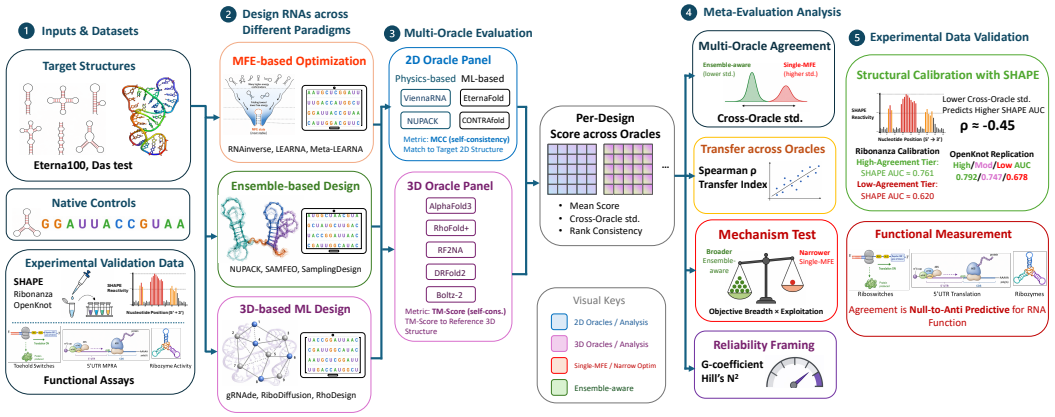

Figure 1: **Meta-evaluation pipeline.** Nine design methods spanning four paradigm families are evaluated through a 4-oracle 2D panel (two parameter families: Turner-2004 thermodynamic and CONTRAfold-engine ML) and a 5-oracle 3D panel, with RibonanzaNet-SS as an independent sensitivity check. Results are calibrated against Ribonanza and OpenKnot SHAPE chemical-mapping data (Watson–Crick pairing proxy) and three public functional assays. Full specifications in §3.

families: physics-based thermodynamic models and machine-learning models frequently disagree on the same sequence [4, 5], and 3D prediction carries even higher uncertainty [6]. A design optimized against one predictor family can therefore appear excellent under that family while failing badly under another, a discrepancy that single-oracle evaluation will never detect. RNA also offers unusually rich resources for studying this problem directly: mature inverse-design algorithms spanning multiple paradigms, two partially independent secondary-structure oracle families, and large-scale SHAPE chemical-mapping datasets [7] that supply per-nucleotide experimental references.

If the dominant evaluation metric is biased toward the optimization oracle, benchmark rankings will favor methods that best exploit that oracle rather than methods producing genuinely high-quality designs, misallocating research effort and advancing flawed sequences toward costly experimental validation. We therefore treat self-consistency scoring itself as the object of evaluation, assessing approximately 11,000 designs from nine RNA design methods using four secondary-structure oracles spanning two parameter families, five tertiary-structure predictors, two SHAPE datasets, and three public functional assay collections.

Contributions. Our primary contribution is a systematic reliability audit of RNA design self-consistency scores, together with a calibrated reporting protocol.

(1) Oracle failure mode and mechanism. Among RNAinverse designs that pass at least one secondary-structure oracle, 78.1% fail another: optimization inflates scores under the target predictor while leaving performance under independent predictors stagnant. A controlled objective-swap shows that replacing a single-MFE objective with ensemble defect while holding local search fixed reduces cross-oracle disagreement by 30×; objective breadth, not search mechanism alone, is the major driver of oracle fragility (§4.1, §4.3).

(2) Calibration. Cross-family disagreement is not noise: lower disagreement predicts higher SHAPE-derived pairing/accessibility AUC on Ribonanza ($\rho = -0.45$) and replicates on OpenKnot, providing a calibrated cross-family uncertainty estimate (§4.5).

(3) Scope and protocol. The structural agreement signal is null-to-anti-predictive for biological function across toehold, 5′UTR, and ribozyme assays, and transfers weakly to current 3D evaluation. We release **ACCORD** (Across-oracle Calibration for COmputational RNA Design), a benchmarking workflow that fills the gap left by single-oracle pipelines through per-oracle scoring, cross-family uncertainty reporting, SHAPE-calibrated tiers, and invalid-use warnings (§5).

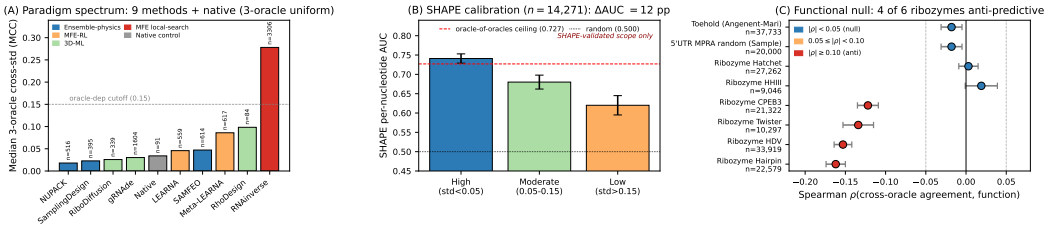

Figure 2: **Three central findings.** (A) Ensemble-physics methods show an order-of-magnitude lower cross-parameter-family disagreement than MFE-local-search; the spectrum is conditional on heterogeneous target sets and is not a universal quality ranking. (B) Lower disagreement predicts higher SHAPE-derived pairing/accessibility AUC; SHAPE is a Watson–Crick pairing proxy, not structural ground truth. (C) Agreement is null-to-anti-predictive for biological function across ribozyme families, establishing the structural scope boundary of the protocol.

2 Related Work

Construct validity, benchmark quality, and proxy–gold divergence. Evaluation as a measurement problem, in which metrics operationalize latent constructs and require validity evidence, builds on psychometric construct validity [8] in the social sciences, and has been adapted into ML by Jacobs and Wallach [9], Raji et al. [10], and Salaudeen et al. [11]. BetterBench [12] audits benchmark quality across the AI benchmark lifecycle; Dehghani et al. [13] show that benchmark choice can change method rankings; HELM [14] introduces multi-metric, multi-scenario transparent evaluation; Singh et al. [15] expose leaderboard distortion and selection bias; Kalai et al. [16] show that accuracy-style scoring without abstention penalties incentivizes confident guessing over calibrated uncertainty in LLM benchmarks, an evaluation-incentive failure mode complementary to the construct-validity literature. The proxy–gold divergence literature is closely related: Gao et al. [17] formalize RLHF overoptimization as an inverted-U; Skalse et al. [18] formalize reward hacking, and Zhuang and Hadfield-Menell [1] analyze failures from optimizing incomplete proxy objectives; Laidlaw et al. [2] show that proxy–gold correlation can break down under optimization; our RNA MFE finding shows a persistent gap in which the proxy improves while non-proxy oracles plateau, comparable to persistent gaps in DAA scaling [19]; reward-model ensemble studies show that multiple imperfect evaluators can mitigate, but not eliminate, related failures [20, 21]. Our paper is therefore a construct-validity audit of self-consistency scoring in a tractable testbed.

Multi-evaluator ensembles and RNA design. Multi-fidelity and multi-information-source optimization formalize the use of imperfect evaluators with different costs and fidelities [22]; CASP [23] uses blind independent assessment of structure prediction methods. Recent protein-design studies [24, 25] often report scTM via refolding scores from ESMFold [26] or AF2; we discuss the cross-domain comparison in App. 34. Computational RNA design spans physics-based approaches (RNAinverse [4], NUPACK [27], SamplingDesign [28]), RL (LEARNA [29]), and deep learning approaches that are autoregressive (gRNAde [30], RDesign [31]), diffusion-based (RiboDiffusion [32]), or flow-matching-based [33, 34]; structure-guided generative methods such as RhoDesign [35] and broader RNA foundation models such as RNAGenesis [36], AIDO.RNA [37], RiNALMo [38], and Orthrus [39] are emerging. Tertiary structure prediction methods include RhoFold+ [40], DRfold2 [41], AlphaFold3 [42], Boltz-2 [43], and RoseTTAFold2NA [44]; our 5-oracle 3D panel is still a lower-bound view of the 3D oracle problem, since blind RNA-Puzzles and CASP assessments [45, 46] show that RNA 3D prediction remains difficult on natural RNAs. BEACON [47] benchmarks RNA *understanding* models but not evaluation metrics. Prior RNA work has characterized individual oracle accuracy [48–50] but not inter-oracle disagreement on designed sequences.

3 Experimental Setup

Oracle panel. We evaluate 4 secondary structure (2D) oracles spanning two paradigm families: **physics-based** (ViennaRNA 2.7, NUPACK 4.0; nearest-neighbor thermodynamic models with different energy parameterizations) and **ML-trained** (EternaFold, CONTRAfold; shared CON-

Table 1: **Design methods, paradigm families, and design counts.** Dataset column counts in parentheses are oracle-complete targets; Raw is the per-method count before the proxy precondition. Five primary methods are augmented by four out-of-corpus reference baselines (shaded).

Method	Paradigm family	Dataset	Raw	Notes
NUPACK design [27]	Ensemble-physics	Eterna100 (66)	521	ensemble defect over Boltzmann distribution
SamplingDesign* [28]	Ensemble-physics	Eterna100 (69)	412	continuous opt. + MC sampling
SAMFEO [51]	Ensemble-physics (multi-frontier)	Eterna100 (71)	710	multi-frontier ensemble defect / probability defect
LEARNA [29]	MFE-RL (PPO)	Eterna100 (66)	1,658	per-puzzle on-policy RL
Meta-LEARNA [29]	MFE-RL (meta)	Eterna100 (71)	659	meta-learned across puzzles
RNAinverse [4, 52]	MFE local-search	Eterna100 (71)	3,550	ViennaRNA Hamming local search
gRNAde [30]	3D-ML (GVP-GNN)	Das test (44)	2,200	autoregressive on PDB backbone
RiboDiffusion [32]	3D-ML (diffusion)	Das test (76)	760	score-based diffusion
RhoDesign‡ [35]	3D-ML (GVP + transformer)	Das held-out (44)	577	two-temperature batch (T=10^{-5} + T=1.0)
Native (control)	—	Das test (95)	95	PDB-derived reference sequences

*Pseudoknot-free input only by tool design. ‡Leakage-free 44-target subset (App. 40) under a two-temperature sampling protocol (T=10^{-5} deterministic plus T=1.0 diversity); 660 raw samples → 577 unique sequences after deduplication. Functional triangulation in §4.6 draws from public toehold, 5′UTR MPRA, and Roberts ribozyme datasets, not part of the design-method count.

TRAfold inference engine with different learned parameters [49]). As a sensitivity check methodologically independent of both families, we also run **RibonanzaNet-SS** [7] on all design cohorts (n = 13,888 predictions): a transformer trained end-to-end on Ribonanza chemical mapping with Hungarian decoding, sharing neither the CONTRAfold engine nor the Turner parameter lineage; its pairwise Spearman with the four base oracles is 0.68–0.87 (lowest with ViennaRNA, 0.68; highest with EternaFold, 0.87). The core framing remains a cross-parameter-family audit with RibonanzaNet-SS as a chemical-mapping-trained sensitivity check. For 3D evaluation, we use RhoFold+, DRfold2, AlphaFold3, Boltz-2, and RoseTTAFold2NA (oracle specifications in App. 1).

Design methods. Five primary methods spanning four paradigm families: **RNAinverse** (MFE local-search; 3,550/71 Eterna100, ViennaRNA MFE objective); **NUPACK design** (ensemble-physics; 521/66, ensemble defect over the Boltzmann distribution); **LEARNA** (MFE-RL; 1,658/66, per-puzzle PPO with ViennaRNA MFE reward [29]); **gRNAde** (3D-ML autoregressive; 2,200/44 Das, GVP-GNN on PDB backbones [30]); **RiboDiffusion** (3D-ML diffusion; 760/76 Das, score-based on 3D structures [32]). Four out-of-corpus reference baselines (SAMFEO, Meta-LEARNA, SamplingDesign, RhoDesign; Table 1) extend coverage of the paradigm spectrum and enable cross-corpus generalization checks.

Datasets. **Eterna100** [53]: 100 RNA design puzzles with community solutions and V1/V2 target structures. **Das et al. test set** [30]: gRNAde Das-seeded single-state split, 98 PDB-derived targets (19–159 nt; seeded by the 14 canonical Das 2010 RNAs [54] and expanded with structurally similar PDB entries), of which 44 completed our 4-oracle 2D fold panel. **Ribonanza calibration subset** [7]: 14,271 sequences from a high-throughput in-vitro SHAPE-MaP library [55] with per-nucleotide reactivity used for oracle calibration, filtered from the full release (∼2M chemically-mapped sequences) to retain sequences with complete oracle predictions and SHAPE labels. Sub-libraries span both computationally-designed sequences and naturally-derived sequence windows, refolded under in-vitro buffer conditions; none are direct measurements on naturally folded biological RNAs (App. 24 for sub-library composition).

Metrics. 2D evaluation: MCC between oracle-predicted and target dot-bracket structures. 3D evaluation: TM-score via USalign [56] between oracle-predicted and PDB reference structures. Cross-oracle agreement: sample standard deviation ($s = \sqrt{\frac{1}{k-1}\sum_{i=1}^{k}(x_i - \bar{x})^2}$) of MCC (or TM-score) across k oracles for each design.

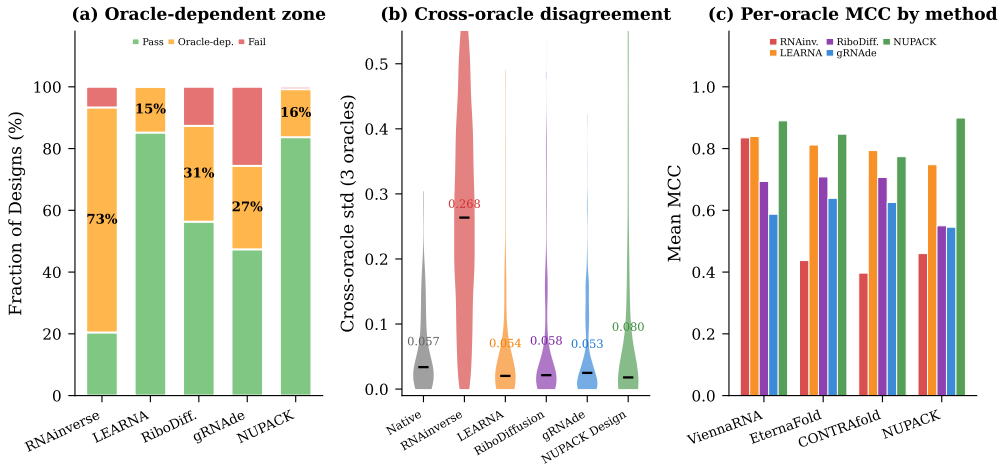

Figure 3: **Cross-parameter-family oracle dependence spans an order of magnitude across paradigms.** (a) Oracle-dependent fraction (passes proxy at MCC$\geq$0.5 but fails at least one other). (b) Per-design cross-oracle std. (c) Per-oracle mean MCC: RNAinverse inflates its proxy oracle (ViennaRNA 0.84) while scoring low on all others (0.40–0.46); NUPACK is uniformly high. Results are conditional on heterogeneous target sets; see Table 2.

4 Results

We organize our findings around four questions: how pervasive is oracle disagreement across design paradigms, what drives it mechanistically, does it carry a calibrated experimental signal, and where does that signal break down? The answers form a coherent picture: disagreement is paradigm-conditional rather than universal, is driven primarily by optimization objective breadth, predicts experimental structural accuracy, and is uninformative or actively misleading for biological function.

4.1 Design Paradigm Determines Oracle Reliability

Oracle disagreement is already substantial before any optimization pressure is applied. On 376 Eterna100 community solutions, the four oracles split cleanly by parameter family (intra-family Spearman ρ=0.83 for the ML pair and 0.59 for the physics pair, with cross-family correlations of only ρ=0.43–0.59), and only 56% of solutions pass all four oracles at MCC$\geq$0.5 (Fig. 6, App. 4). This baseline heterogeneity is not a minor calibration difference; it reflects genuinely distinct modeling assumptions across energy parameterizations. On *designed* sequences the problem compounds: under a uniform proxy precondition (any oracle MCC$\geq$0.5), **78.1**% of RNAinverse designs pass one oracle yet fail another, and only 21.9% pass all four (Table 2; raw-pool: 72.9%/20.5%, App. 38). Critically, this amplification is not a universal property of computational design: it depends strongly on how a method was optimized.

Ensemble-based optimization is the clearest solution to this problem: NUPACK achieves just 15.7% oracle-dependent designs vs. 78.1% for RNAinverse, because minimizing ensemble defect over the Boltzmann distribution captures a universal design-quality signal that correlates with MCC under *all* oracles ($\rho = -0.63$ to -0.90). This advantage is not a dual-role artifact: it holds when NUPACK is excluded from the evaluation panel entirely (App. 23). 3D-conditioned methods occupy an intermediate position (gRNAde 36.3%, RiboDiffusion 35.5%), with cross-oracle std comparable to native sequences despite operating on a different design objective. RhoDesign presents a qualitatively different failure mode: only 14.9% of held-out-44 designs pass any oracle at all, indicating structure-recovery failure under our 2D refolding panel rather than oracle over-optimization. gRNAde's placement on this spectrum is not leakage-driven: a retrain at 45% TM-cluster exclusion preserves Transfer Index (0.805 $\rightarrow$ 0.822 on the 44 overlapping targets; App. 40).

Objective breadth drives oracle fragility more than search strategy. The central mechanistic question is whether disagreement stems from *how hard* a method optimizes or *what* it optimizes for. A 2$\times$2 controlled experiment (App. 25) isolates these two factors by holding one fixed while varying

Table 2: **Paradigm comparison under the proxy precondition** (max oracle MCC ≥ 0.5). Sel. = fraction of raw designs passing the precondition; Oracle-dep = fraction passing some but not all oracles; Transfer Index measures cross-oracle robustness (App. 21). Ensemble-physics methods consistently outperform MFE-optimized designs on both columns.

Method	**Paradigm**	N_{pre}	**Sel.**	**Oracle-dep**	**Transfer Idx**
NUPACK [27]	Ensemble-physics	517	99.2%	**15.7%**	**0.89**
SamplingDesign [28]	Boltzmann MC opt. (PK-free)	396	96.1%	19.2%	**0.83**
SAMFEO† [51]	Multi-frontier ens.	614	86.5%	21.5%	**0.86**
LEARNA [29]	MFE RL (PPO)	573	34.6%	44.0%	0.77
Meta-LEARNA† [29]	Meta-RL on MFE	617	93.6%	27.1%	0.73
gRNAde [30]	3D-ML autoregr.	1,637	74.4%	36.3%	0.72
RiboDiffusion [32]	3D-ML diffusion	664	87.4%	35.5%	0.45
RhoDesign‡ [35]	3D-ML GVP+transformer	86	14.9%	68.6%	0.81
RNAinverse [4, 52]	MFE local-search	3,313	93.3%	**78.1%**	**0.07**
Native	—	92	96.8%	57.6%	—

†Evaluated under a 3-oracle precondition; the uniform 3-oracle rebuild preserves the paradigm ordering (App. 38).

‡Held-out 44-target subset (577 unique designs); low pass rate reflects structure-recovery failure rather than oracle over-optimization (App. 40).

the other, and the answer is unambiguous: the objective matters far more than the search strategy. Switching from MFE to ensemble defect while keeping the search mechanism fixed reduces cross-oracle disagreement 30-fold (median std $0.269 \rightarrow 0.009$), while switching the search mechanism from local search to PPO under the same MFE objective reduces it only 5-fold ($0.284 \rightarrow 0.054$). The gap persists at matched proxy quality: RNAinverse exceeds LEARNA's cross-oracle std in all 14 of 14 ViennaRNA-MCC bins (App. 20), ruling out differences in proxy-passing rate as a confound. The structural mechanism underlying this pattern is clear: RNAinverse produces $52\times$ more ViennaRNA suboptimal structures within 2 kcal/mol than NUPACK, with mean MFE pair probability 0.52 vs. 0.93 (App. 20), placing its designs in degenerate energy landscape regions where any form of oracle-specific exploitation is amplified. RNAinverse thus exhibits an *extremal* Goodhart pattern, while NUPACK, LEARNA, and ensemble-local-search remain near a *minimal* Goodhart regime [57]. The paradigm spectrum replicates on out-of-corpus methods that were not part of the original experimental design (SAMFEO, SamplingDesign, Meta-LEARNA; Table 2, App. 26), strengthening the generality of the objective-breadth conclusion.

Four oracles collapse to two effective families. We adopt a fixed-facet G-theory model [58] because the four oracles span two parameter families (Turner-2004 and CONTRAfold-engine) rather than constituting a random sample from a broader oracle population (App. 12). The resulting G-coefficients are strikingly different across paradigms: $G = 0.644$ $[0.565, 0.701]$ for RNAinverse vs. $G = 0.947$ for gRNAde, with non-overlapping 95% CIs that confirm the paradigm-spectrum claim is statistically robust. Per-paradigm Hill's N_2 makes the underlying redundancy explicit: pooled $N_2 = \mathbf{1.94}$ of 4 means the panel *collapses to two effective families*, so the protocol is best read as a **failure detector** that identifies oracle-specific exploitation rather than a universal quality metric (App. 12). The paradigm ranking is robust to threshold choice, leave-one-oracle-out configurations, and target set (Das: NUPACK 1.1% vs. RNAinverse 30.4%; App. 8, 13, 35, 38, 9). Within the nearest-neighbor thermodynamic oracle universe, **single-oracle sc-scores for MFE-optimized designs measure the oracle as much as the design**; this finding is scoped to this oracle lineage and does not extend to orthogonal oracle families or functional evaluation.

4.2 Out-of-Distribution Amplification

Oracle disagreement is not an intrinsic property of structure predictors: the optimization process amplifies it selectively and severely. Cross-oracle std on native sequences is 0.057; on RNAinverse designs it reaches $\mathbf{0.268}$, a $\mathbf{4.7}\times$ amplification, while gRNAde (0.054), RiboDiffusion (0.058), and NUPACK designs ($\sim$0.04) all remain at native level despite being generated by very different algorithmic families (Figure 5, App. 3). Sequence composition is not the driver: gRNAde deviates more from native GC content (0.602 vs. 0.536) yet shows substantially lower disagreement than RNAinverse, and random sequences achieve near-zero disagreement (std = 0.007) while failing all oracles

unanimously (App. 3). The amplification is instead a direct consequence of optimization-induced energy landscape degeneracy: ViennaRNA's MFE and MEA decoders agree on 69% of native positions but only 14% of RNAinverse positions (App. 10), revealing that MFE-optimized sequences occupy highly degenerate landscape regions where small perturbations in the energy model, whether from different parameterizations or different decoding strategies, produce radically different structural predictions.

4.3 Evaluation Transfer Across Oracles

The *evaluation transfer* problem asks how well the optimization proxy's score tracks performance under non-proxy panel oracles. Transfer Index is a cross-oracle robustness descriptor rather than a claim about true design quality (the best per-sequence SHAPE-AUC ceiling is 0.727; App. 22). Our findings reveal a systematic and paradigm-dependent proxy–gold gap: under MFE optimization, the proxy improves monotonically while non-proxy oracle scores plateau, a pattern consistent with proxy–gold divergence under optimization pressure [1].

MFE optimization: a persistent transfer gap. The proxy score increasingly overstates its own ranking as optimization pressure grows. On 3,550 RNAinverse designs, gold-oracle MCC at proxy MCC $= 1.0$ is only 0.5–0.6 (transfer discount 0.4–0.5), and oracles agree on the best design for a given target only 8–33% of the time (App. 6). This *evaluation-proxy inflation* effect, in which optimization for the target metric produces diminishing returns under any alternative metric, does not imply that any oracle's score equals true design quality, but it does mean that ViennaRNA MCC has become an increasingly unreliable guide to cross-oracle performance as optimization succeeds.

Ensemble optimization: substantially smaller transfer gap. Ensemble objectives eliminate this pathology almost entirely. Across 6 stringency levels (n=535; App. 7), all gold oracle MCCs for NUPACK designs improve monotonically ($\rho \leq -0.94$), and the oracle-dependent zone *decreases* rather than growing with optimization pressure, falling from 27% to 16.5%. The transfer discount at high proxy MCC is just 0.03, **6× smaller** than MFE optimization's 0.20, and approximately an order of magnitude smaller than residual gaps under RLHF reward-model ensembling [20, 21]. The thermodynamic mechanism is straightforward: by minimizing ensemble defect, NUPACK makes the target structure a dominant Boltzmann attractor, a property that predicts cross-oracle MCC across all paradigms ($\rho = -0.58$ to -0.89; ViennaRNA-computed partition function gives $\rho = -0.71$, ruling out NUPACK circularity; App. 14). We note this 6× ratio is scoped to the nearest-neighbor thermodynamic universe; orthogonal-universe transfer is not guaranteed, as NUPACK's best 2D agreement coincides with lower 3D quality than gRNAde under Boltz-2 (App. 5).

Objective breadth predicts transfer reliability across paradigms. The transfer spectrum maps cleanly and monotonically onto optimization breadth: ensemble-physics < 3D-ML autoregressive < 3D-ML diffusion < MFE-physics. NUPACK anchors the top of this spectrum (transfer discount 0.032, Pareto-optimal designs 13.6%), while RNAinverse anchors the bottom as the extremal Goodhart case (discount 0.197, Pareto-optimal 2%). This ordering is not an artifact of target-set differences: it holds on the Das test set and under leave-one-oracle-out configurations (Table 11, App. 6).

4.4 Preliminary Audit of Cross-Dimensional Transfer

A natural question is whether strong 2D cross-oracle agreement implies good 3D structural quality. The evidence suggests it does not. On $n = 95$ common targets with predictions from RhoFold+, Boltz-2, and RoseTTAFold2NA, 3D oracles disagree 2.4× more than 2D oracles (cross-3D std $= 0.139$ vs. cross-2D std $= 0.057$; bootstrap 95% CI $[1.75, 3.34]$), and 2D MCC explains only 12–21% of 3D TM variance across oracle pairs. The practical implication is concrete: RhoFold+ *reverses* the 2D Transfer Index ordering between gRNAde (2D 0.72; 3D TM 0.38) and RiboDiffusion (2D 0.45; 3D TM 0.61), demonstrating that a method with better 2D cross-oracle consistency can systematically produce worse 3D structures. On a shared-bias control, RhoDesign [35] scores *lower* on its own training oracle (RhoFold+) than on Boltz-2 relative to natives (paired delta $= -0.15$, $n = 41$, Wilcoxon $p = 0.041$; App. 40), a result that runs opposite to the shared-bias direction and undermines any claim that 3D conditioning inflates 3D scores. The AlphaFold3 native-vs-designed

Table 3: **Lower cross-oracle disagreement predicts higher SHAPE-derived pairing/accessibility AUC** ($\rho = -0.45$, $n = 14{,}271$ Ribonanza sequences). The 12-pp gap between extreme tiers spans roughly half the dynamic range between random baseline and oracle-of-oracles ceiling.

Cross-oracle std	N	Mean SHAPE AUC	Best-oracle AUC
< 0.02 (high agreement)	1,638	**0.761**	0.773
0.02–0.05	2,997	0.730	0.753
0.05–0.10	5,025	0.697	0.730
0.10–0.15	3,355	0.663	0.712
> 0.15 (low agreement)	1,256	0.620	0.693

contrast is consistent in direction but limited to $n = 25$ targets and is accordingly reported as preliminary (App. 16). Taken together, these results establish that 2D evaluation, even under full oracle agreement, cannot substitute for direct 3D structural assessment.

4.5 Experimental Calibration and External Replication

Cross-oracle agreement predicts experimental structural accuracy. Cross-oracle disagreement is not noise; it is a calibrated uncertainty signal with direct experimental grounding. For 14,271 Ribonanza sequences, designs on which oracles agree fold more accurately under independent SHAPE chemical probing: the cross-oracle std vs. SHAPE AUC relationship is strong, monotonic, and robust ($\rho = -0.45$, Cohen's $d = 1.62$ between extreme tiers; Mann-Whitney $p < 10^{-200}$; Table 3). The 12 pp tier gap is substantial: the High-tier mean (0.761) exceeds the oracle-of-oracles ceiling (0.727), while the Low tier captures only 52.9% of the random-to-ceiling dynamic range. The signal survives restriction to cross-parameter-family oracle pairs ($\rho = -0.44$; App. 39) and retains 78% of its magnitude after controlling for sequence length, MFE energy, and ensemble diversity (App. 15). External replication on OpenKnotAIDesignData [7] (n=25,930, 13 methods) preserves monotone tier ordering for $11/11$ methods ($\Delta = +0.114$; App. 30), confirming that the calibration generalizes across design pipelines.

Individual oracle accuracy tells a complementary story about why aggregation fails. ML-trained oracles outperform physics-based ones on native sequences (EternaFold 0.714, CONTRAfold 0.697 vs. ViennaRNA 0.680, NUPACK 0.656 default; App. 36), yet **high inter-oracle agreement does not imply accuracy**: NUPACK is the weakest individual predictor on natives yet agrees most with ViennaRNA across paradigms (App. 4), illustrating how shared parameter-family biases can produce agreement without accuracy. NUPACK's design success is therefore not mediated by predictor accuracy: even at RNA99-best (0.712), it matches but does not exceed the ML oracles in SHAPE calibration, yet its designs reach Transfer Index 0.89 vs. 0.72 for gRNAde. No aggregation strategy improves on the best single oracle: stacking (LR 0.680, MLP 0.687), 7-oracle majority (0.701), and IPknots (0.703) all fall short of EternaFold (0.709), with an oracle-of-oracles ceiling of only 0.727 (App. 22, Fig. 11). Cross-parameter-family oracle agreement thus operationalizes partially-convergent validity [8, 9] under shared-family bias.

4.6 Functional Scope Boundary

The structural agreement signal, while well-calibrated against experimental pairing data, is not a valid proxy for biological function, and in some regimes actively misleads. The most striking evidence comes from ribozyme activity: four of six ribozyme families show a robust *anti-predictive* signal ($\rho = -0.12$ to -0.16; CPEB3, Hairpin, Twister, HDV; $n = 88{,}117$), meaning sequences with higher cross-oracle structural agreement have *lower* catalytic activity. At $n \approx 10^4$–10^5, these are genuine negative effects confirmed by TOST equivalence testing, not statistical artifacts inflated by large sample sizes (App. 41, Table 53). The mechanistic explanation is principled: sequences that fold easily and generate oracle consensus tend to adopt simple, low-energy structures that cannot support the precise catalytic geometry required for ribozyme cleavage; catalytically competent folds, by contrast, are structurally non-trivial and occupy more contested regions of the energy landscape where oracle predictions naturally diverge. Toehold and 5'UTR assays confirm the null on the positive side. Both are TOST-equivalent to zero ($|\rho| < 0.05$), ruling out even a weak positive association between structural agreement and translational or regulatory activity. Structural agreement and biological function are therefore distinct constructs that should not be conflated: the

former reflects thermodynamic self-consistency across oracle families, while the latter depends on local motif geometry, switching kinetics, alternative conformational states, and catalytic active-site preservation that lie outside the scope of secondary-structure oracles entirely. Functional design requires a separate oracle class with direct sensitivity to these properties.

5 Toward a Cross-Parameter-Family Evaluation Protocol

Conditional uncertainty tiers are SHAPE-calibrated on Ribonanza synthetic-library sequences and replicate on the independent OpenKnotAIDesignData corpus ($n = 25{,}930$, 13 methods; App. 30). **Scope.** The tiers are validated for computationally-designed, canonical-nucleotide RNA under refolded in-vitro SHAPE-MaP at length 19–1,200 nt; natural RNA, in-cell contexts, modified nucleotides, and function prediction lie outside this calibration, and the 3D tertiary axis is partially audited (§4.4, App. 5) but not directly tier-calibrated. **Panel and reporting.** (1) *Graduated panels*: minimal = ViennaRNA + EternaFold + CONTRAfold (<0.1 s/design); recommended adds NUPACK for a balanced two-physics plus two-ML panel; comprehensive adds RibonanzaNet-SS and ≥ 2 3D oracles. Same-family panels inflate reliability (App. 12). (2) Report per-design cross-oracle std and flag std >0.15 as oracle-dependent; include a native baseline (designed-to-native std ratio >2 flags bias); report ensemble defect alongside MCC. (3) *Conditional uncertainty tiers*: High (std <0.05, AUC ≈ 0.74), Moderate (0.05–0.15, 0.66–0.70), Low (>0.15, ≈ 0.62), applied after the proxy precondition (any oracle MCC ≥ 0.5). These tiers quantify cross-parameter-family disagreement conditional on the proxy precondition; a design with low std but low mean MCC has passed the precondition but should not be interpreted as high-quality. (4) Exclude the design oracle and report per-oracle scores individually (aggregation does not beat the best single oracle; App. 22). This generator-evaluator independence requirement parallels the "preference leakage" contamination in LLM-as-judge evaluation [59, 60]. The released package implements these steps in a BEACON-style benchmark workflow [47].

Benchmark card. *Intended use*: reporting cross-oracle robustness of computationally-designed RNA secondary structures; identifying paradigm-level evaluation biases; comparing design methods under a controlled structural-reliability signal. *Invalid use*: predicting biological function or therapeutic activity; evaluating natural RNA in cellular conditions; reporting 3D structural quality; assessing modified-nucleotide designs. *Known failure modes*: RhoDesign-class methods with low pass rates render uncertainty tiers uninterpretable without reporting pass-rate separately; same-family panels underestimate disagreement; NUPACK's dual role as designer and evaluator requires the control in App. 23. *Reporting fields*: oracle panel composition, proxy precondition threshold, pass rate, per-method sample sizes, native baseline std, and panel parameter families.

6 Discussion

The cross-oracle disagreement traced through §4.1–§4.3 reduces to a single mediator: structural degeneracy in the oracle's energy landscape (App. 20). Ensemble-based objectives should therefore be preferred over single-MFE criteria, as they steer designs away from degenerate landscape regions where oracle variability compounds, producing a $30\times$ reduction in disagreement under otherwise identical search conditions. Hill's $N_2 = 1.94$ confirms that the panel provides two effective independent sources, making disagreement a more reliable uncertainty signal than a consensus accuracy metric [61, 62]. A qualitative protein–RNA comparison (App. 34) suggests this oracle-accuracy regime is categorically different between the two domains, explaining why self-consistency scoring requires recalibration when transferred from protein to RNA design.

Limitations and reproducibility. Cross-oracle agreement is calibrated against SHAPE-derived pairing accessibility rather than a complete structural ground truth, and remains null-to-anti-predictive for biological function (App. 41); therapeutic claims require direct functional assays. The breadth-vs-intensity 2×2 is scoped to the nearest-neighbor energy-model universe (App. 25). Training-data leakage is audited but not fully eliminable: the gRNAde retrain confirms paradigm-position preservation, while RiboDiffusion's training corpus lacks a documented PDB cutoff (App. 40). The **ACCORD** package (`https://anonymous.4open.science/r/ACCORD-F49D`) ships the unified `evaluate()` API and SHAPE-calibrated uncertainty tier assignment.

References

[1] Simon Zhuang and Dylan Hadfield-Menell. Consequences of misaligned ai. *Advances in Neural Information Processing Systems*, 33:15763–15773, 2020.

[2] Cassidy Laidlaw, Shivam Singhal, and Anca Dragan. Correlated proxies: A new definition and improved mitigation for reward hacking. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=msEr27EejF.

[3] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

[4] Ronny Lorenz, Stephan H Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. Viennarna package 2.0. *Algorithms for molecular biology*, 6(1):26, 2011.

[5] Chuong B Do, Daniel A Woods, and Serafim Batzoglou. Contrafold: Rna secondary structure prediction without physics-based models. *Bioinformatics*, 22(14):e90–e98, 2006.

[6] Marcin Magnus, Maciej Antczak, Tomasz Zok, Jakub Wiedemann, Piotr Lukasiak, Yang Cao, Janusz M Bujnicki, Eric Westhof, Marta Szachniuk, and Zhichao Miao. Rna-puzzles toolkit: a computational resource of rna 3d structure benchmark datasets, structure manipulation, and evaluation tools. *Nucleic acids research*, 48(2):576–588, 2020.

[7] Shujun He, Rui Huang, Jill Townley, Rachael C. Kretsch, Thomas G. Karagianes, David B.T. Cox, Hamish Blair, Dmitry Penzar, Valeriy Vyaltsev, Elizaveta Aristova, Arsenii Zinkevich, Artemy Bakulin, Hoyeol Sohn, Daniel Krstevski, Takaaki Fukui, Fumiya Tatematsu, Yusuke Uchida, Donghoon Jang, Jun Seong Lee, Roger Shieh, Tom Ma, Eduard Martynov, Maxim V. Shugaev, Habib S.T. Bukhari, Kazuki Fujikawa, Kazuki Onodera, Christof Henkel, Shlomo Ron, Jonathan Romano, John J. Nicol, Grace P. Nye, Yuan Wu, Christian Choe, Walter Reade, Eterna participants, and Rhiju Das. Ribonanza: deep learning of rna structure through dual crowdsourcing. *bioRxiv*, 2024. doi: 10.1101/2024.02.24.581671. URL https://www.biorxiv.org/content/early/2024/06/11/2024.02.24.581671.

[8] Lee J Cronbach and Paul E Meehl. Construct validity in psychological tests. *Psychological bulletin*, 52(4):281, 1955.

[9] Abigail Z Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 375–385, 2021.

[10] Inioluwa Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. AI and the everything in the whole wide world benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=j6NxpQbREA1.

[11] Olawale Salaudeen, Anka Reuel, Ahmed Ahmed, Suhana Bedi, Zachary Robertson, Sudharsan Sundar, Ben Domingue, Angelina Wang, and Sanmi Koyejo. Measurement to meaning: A validity-centered framework for ai evaluation, 2025. URL https://arxiv.org/abs/2505.10573.

[12] Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J Kochenderfer. Betterbench: Assessing ai benchmarks, uncovering issues, and establishing best practices. *Advances in Neural Information Processing Systems*, 37:21763–21813, 2024.

[13] Mostafa Dehghani, Yi Tay, Alexey A Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery. *arXiv preprint arXiv:2107.07002*, 2021.

[14] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak,

Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=iO4LZibEqW. Featured Certification, Expert Certification, Outstanding Certification.

[15] Shivalika Singh, Yiyang Nan, Alex Wang, Daniel D'souza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah A. Smith, Beyza Ermis, Marzieh Fadaee, and Sara Hooker. The leaderboard illusion. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL https://openreview.net/forum?id=4Ae8edNqm0.

[16] Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Evaluating large language models for accuracy incentivizes hallucinations. *Nature*, 2026. doi: 10.1038/s41586-026-10549-w.

[17] Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR, 2023.

[18] Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471, 2022.

[19] Rafael Rafailov, Yaswanth Chittepu, Ryan Park, Harshit Sushil Sikchi, Joey Hejna, Brad Knox, Chelsea Finn, and Scott Niekum. Scaling laws for reward model overoptimization in direct alignment algorithms. *Advances in Neural Information Processing Systems*, 37:126207–126242, 2024.

[20] Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=dcjtMYkpXx.

[21] Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alexander Nicholas D'Amour, Krishnamurthy Dj Dvijotham, Adam Fisch, Katherine A Heller, Stephen Robert Pfohl, Deepak Ramachandran, Peter Shaw, and Jonathan Berant. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=5u1GpUkKtG.

[22] Peter I Frazier. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

[23] Andriy Kryshtafovych, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult. Critical assessment of methods of protein structure prediction (casp)—round xiv. *Proteins: Structure, Function, and Bioinformatics*, 89(12):1607–1617, 2021.

[24] 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.

[25] Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

[26] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[27] Mark E Fornace, Jining Huang, Cody T Newman, Avinash Nanjundiah, Nicholas J Porubsky, Marshall B Pierce, and Niles A Pierce. Nupack: Computational nucleic acid analysis and design. *ACS Synthetic Biology*, 2026.

[28] Wei Yu Tang, Ning Dai, Tianshuo Zhou, David H Mathews, and Liang Huang. Samplingdesign: Rna design via continuous optimization with coupled variables and monte-carlo sampling. *Nature Communications*, 17:2950, 2026. doi: 10.1038/s41467-025-67901-3.

[29] Frederic Runge, Danny Stoll, Stefan Falkner, and Frank Hutter. Learning to design RNA. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ByfyHh05tQ.

[30] Chaitanya K. Joshi, Arian Rokkum Jamasb, Ramon Viñas Torné, Charles Harris, Simon V Mathis, Alex Morehead, Rishabh Anand, and Pietro Lio. gRNAde: Geometric deep learning for 3d RNA inverse design. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=lvw3UgeVxS.

[31] Cheng Tan, Yijie Zhang, Zhangyang Gao, Bozhen Hu, Siyuan Li, Zicheng Liu, and Stan Z. Li. RDesign: Hierarchical data-efficient representation learning for tertiary structure-based RNA design. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=RemfXx7ebP.

[32] Han Huang, Ziqian Lin, Dongchen He, Liang Hong, and Yu Li. Ribodiffusion: tertiary structure-based rna inverse folding with generative diffusion models. *Bioinformatics*, 40 (Supplement_1):i347–i356, 06 2024. ISSN 1367-4811. doi: 10.1093/bioinformatics/btae259. URL https://doi.org/10.1093/bioinformatics/btae259.

[33] Runze Ma, Zhongyue Zhang, Zichen Wang, Will Hua, Jiahua Rao, Zhuomin Zhou, and Shuangjia Zheng. Riboflow: Conditional de novo RNA co-design via synergistic flow matching. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=brV7U7svgS.

[34] Rishabh Anand, Chaitanya K. Joshi, Alex Morehead, Arian Rokkum Jamasb, Charles Harris, Simon V Mathis, Kieran Didi, Rex Ying, Bryan Hooi, and Pietro Lio. RNA-frameflow: Flow matching for de novo 3d RNA backbone design. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=wOc1Yx5s09.

[35] Felix Wong, Dongchen He, Aarti Krishnan, Liang Hong, Alexander Z Wang, Jiuming Wang, Zhihang Hu, Satotaka Omori, Alicia Li, Jiahua Rao, et al. Deep generative design of rna aptamers using structural predictions. *Nature Computational Science*, 4(11):829–839, 2024.

[36] Zaixi Zhang, Ruofan Jin, Linlin Chao, Guangxue Xu, Yikun Zhang, Guowei Zhou, Di Yin, Yingqing Guo, Yaqi Fu, Yukang Yang, et al. Rnagenesis: a generalist foundation model for functional rna therapeutics. *bioRxiv*, pages 2024–12, 2024.

[37] Shuxian Zou, Tianhua Tao, Sazan Mahbub, Caleb N Ellington, Robin Algayres, Dian Li, Yonghao Zhuang, Hongyi Wang, Le Song, and Eric P Xing. A large-scale foundation model for rna function and structure prediction. *bioRxiv*, pages 2024–11, 2024.

[38] Rafael Josip Penić, Tin Vlašić, Roland G Huber, Yue Wan, and Mile Šikić. Rinalmo: General-purpose rna language models can generalize well on structure prediction tasks. *Nature Communications*, 16(1):5671, 2025.

[39] Philip Fradkin, Ruian "Ian" Shi, Taykhoom Dalal, Keren Isaev, Brendan J Frey, Leo J Lee, Quaid Morris, and Bo Wang. Orthrus: toward evolutionary and functional rna foundation models. *Nature Methods*, pages 1–11, 2026.

[40] Tao Shen, Zhihang Hu, Siqi Sun, Di Liu, Felix Wong, Jiuming Wang, Jiayang Chen, Yixuan Wang, Liang Hong, Jin Xiao, et al. Accurate rna 3d structure prediction using a language model-based deep learning approach. *Nature Methods*, 21(12):2287–2298, 2024.

[41] Yang Li, Chenjie Feng, Xi Zhang, Sho Tsukiyama, Duanyu Feng, and Yang Zhang. Drfold2 is a deep learning-based tool that enables efficient and accurate rna structure prediction. *Plos Biology*, 24(2):e3003659, 2026.

[42] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

[43] Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv*, 2025.

[44] Minkyung Baek, Ryan McHugh, Ivan Anishchenko, Hanlun Jiang, David Baker, and Frank DiMaio. Accurate prediction of nucleic acid and protein-nucleic acid complexes using RoseTTAFold2NA. *Nature Methods*, 21:117–121, 2024. doi: 10.1038/s41592-023-02086-5.

[45] Fan Bu, Yagoub Adam, Ryszard W Adamiak, Maciej Antczak, Belisa Rebeca H de Aquino, Nagendar Goud Badepally, Robert T Batey, Eugene F Baulin, Pawel Boinski, Michal J Boniecki, et al. Rna-puzzles round v: blind predictions of 23 rna structures. *Nature methods*, 22(2):399–411, 2025.

[46] Rachael C Kretsch, Alissa M Hummer, Shujun He, Rongqing Yuan, Jing Zhang, Thomas Karagianes, Qian Cong, Andriy Kryshtafovych, and Rhiju Das. Assessment of nucleic acid structure prediction in casp16. *Proteins: Structure, Function, and Bioinformatics*, 94(1):192–217, 2026.

[47] Yuchen Ren, Zhiyuan Chen, Lifeng Qiao, Hongtai Jing, Yuchen Cai, Sheng Xu, Peng Ye, Xinzhu Ma, Siqi Sun, Hongliang Yan, et al. Beacon: Benchmark for comprehensive rna tasks and language models. *Advances in Neural Information Processing Systems*, 37:92891–92921, 2024.

[48] David H Mathews, Matthew D Disney, Jessica L Childs, Susan J Schroeder, Michael Zuker, and Douglas H Turner. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of rna secondary structure. *Proceedings of the National Academy of Sciences*, 101(19):7287–7292, 2004.

[49] Hannah K Wayment-Steele, Wipapat Kladwang, Alexandra I Strom, Jeehyung Lee, Adrien Treuille, Alex Becka, Eterna Participants, and Rhiju Das. Rna secondary structure packages evaluated and improved by high-throughput experiments. *Nature methods*, 19(10):1234–1242, 2022.

[50] Luciano I Zablocki, Leandro A Bugnon, Matias Gerard, Leandro Di Persia, Georgina Stegmayer, and Diego H Milone. Comprehensive benchmarking of large language models for rna secondary structure prediction. *Briefings in Bioinformatics*, 26(2):bbaf137, 2025.

[51] Tianshuo Zhou, Ning Dai, Sizhen Li, Max Ward, David H Mathews, and Liang Huang. Rna design via structure-aware multifrontier ensemble optimization. *Bioinformatics*, 39 (Supplement_1):i563–i571, 2023.

[52] Ivo L Hofacker, Walter Fontana, Peter F Stadler, L Sebastian Bonhoeffer, Manfred Tacker, and Peter Schuster. Fast folding and comparison of rna secondary structures. *Monatshefte für Chemie/Chemical Monthly*, 125(2):167–188, 1994.

[53] Jeff Anderson-Lee, Eli Fisker, Vineet Kosaraju, Michelle Wu, Justin Kong, Jeehyung Lee, Minjae Lee, Mathew Zada, Adrien Treuille, Rhiju Das, et al. Principles for predicting rna secondary structure design difficulty. *Journal of molecular biology*, 428(5):748–757, 2016.

[54] Rhiju Das, John Karanicolas, and David Baker. Atomic accuracy in predicting and designing noncanonical rna structure. *Nature methods*, 7(4):291–294, 2010.

[55] Nathan A Siegfried, Steven Busan, Greggory M Rice, Julie AE Nelson, and Kevin M Weeks. Rna motif discovery by shape and mutational profiling (shape-map). *Nature methods*, 11(9): 959–965, 2014.

[56] Chengxin Zhang, Morgan Shine, Anna Marie Pyle, and Yang Zhang. Us-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nature methods*, 19(9):1109–1115, 2022.

[57] David Manheim and Scott Garrabrant. Categorizing variants of goodhart's law. *arXiv preprint arXiv:1803.04585*, 2018.

[58] Lee Joseph Cronbach. The dependability of behavioral measurements. *Theory of generalizability for scores and profiles*, pages 1–33, 1972.

[59] Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=4NJBV6Wp0h.

[60] Dawei Li, Renliang Sun, Yue Huang, Ming Zhong, Bohan Jiang, Jiawei Han, Xiangliang Zhang, Wei Wang, and huan liu. Preference leakage: A contamination problem in LLM-as-a-judge. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=grIvSXVJ65.

[61] Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. *Advances in neural information processing systems*, 7, 1994.

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

[63] Michelle J Wu, Johan OL Andreasson, Wipapat Kladwang, William Greenleaf, and Rhiju Das. Automated design of diverse stand-alone riboswitches. *ACS synthetic biology*, 8(8):1838–1846, 2019.

[64] He Zhang, Liang Zhang, Ang Lin, Congcong Xu, Ziyu Li, Kaibo Liu, Boxiang Liu, Xiaopin Ma, Fanfan Zhao, Huiling Jiang, et al. Algorithm for optimized mrna design improves stability and immunogenicity. *Nature*, 621(7978):396–403, 2023.

[65] Yuhao Chen and Xiaowei Wang. Evaluation of efficiency prediction algorithms and development of ensemble model for crispr/cas9 grna selection. *Bioinformatics*, 38(23):5175–5181, 2022.

[66] Alexander A Green, Pamela A Silver, James J Collins, and Peng Yin. Toehold switches: de-novo-designed regulators of gene expression. *Cell*, 159(4):925–939, 2014.

[67] John G Doench, Nicolo Fusi, Meagan Sullender, Mudra Hegde, Emma W Vaimberg, Katherine F Donovan, Ian Smith, Zuzana Tothova, Craig Wilen, Robert Orchard, et al. Optimized sgrna design to maximize activity and minimize off-target effects of crispr-cas9. *Nature biotechnology*, 34(2):184–191, 2016.

[68] Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating llm generations with a panel of diverse models, 2024. URL https://arxiv.org/abs/2404.18796.

[69] Bohdan Schneider, Blake Alexander Sweeney, Alex Bateman, Jiri Cerny, Tomasz Zok, and Marta Szachniuk. When will rna get its alphafold moment? *Nucleic acids research*, 51(18): 9522–9532, 2023.

[70] John B Ingraham, Max Baranov, Zak Costello, Karl W Barber, Wujie Wang, Ahmed Ismail, Vincent Frappier, Dana M Lord, Christopher Ng-Thow-Hing, Erik R Van Vlack, et al. Illuminating protein space with a programmable generative model. *Nature*, 623(7989):1070–1078, 2023.

[71] Jason Yim, Andrew Campbell, Andrew Y. K. Foong, Michael Gastegger, José Jiménez-Luna, Sarah Lewis, Victor Garcia Satorras, Bastiaan S. Veeling, Regina Barzilay, Tommi Jaakkola, and Frank Noé. Fast protein backbone generation with se(3) flow matching, 2023. URL https://arxiv.org/abs/2310.05297.

[72] Joey Bose, Tara Akhound-Sadegh, Guillaume Huguet, Kilian FATRAS, Jarrid Rector-Brooks, Cheng-Hao Liu, Andrei Cristian Nica, Maksym Korablyov, Michael M. Bronstein, and Alexander Tong. SE(3)-stochastic flow matching for protein backbone generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=kJFIH23hXb.

[73] Jason Yim, Brian L. Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay, and Tommi Jaakkola. Se(3) diffusion model with application to protein backbone generation, 2023. URL https://arxiv.org/abs/2302.02277.

[74] Yeqing Lin and Mohammed AlQuraishi. Generating novel, designable, and diverse protein structures by equivariantly diffusing oriented residue clouds, 2023. URL https://arxiv.org/abs/2301.12485.

[75] Jin Sub Lee, Jisun Kim, and Philip M Kim. Score-based generative modeling for de novo protein design. *Nature Computational Science*, 3(5):382–392, 2023.

[76] Zaixiang Zheng, Yifan Deng, Dongyu Xue, Yi Zhou, Fei YE, and Quanquan Gu. Structure-informed language models are protein designers. In *International Conference on Machine Learning*, 2023.

[77] Xinyou Wang, Zaixiang Zheng, Fei YE, Dongyu Xue, Shujian Huang, and Quanquan Gu. Diffusion language models are versatile protein learners. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=NUAbSFqyqb.

[78] Milong Ren, Jinyuan Sun, Jiaqi Guan, Cong Liu, Chengyue Gong, Yuzhe Wang, Lan Wang, Qixu Cai, Xinshi Chen, and Wenzhi Xiao. Pxdesign: Fast, modular, and accurate de novo design of protein binders. *bioRxiv*, pages 2025–08, 2025.

[79] Rohith Krishna, Jue Wang, Woody Ahern, Pascal Sturmfels, Preetham Venkatesh, Indrek Kalvet, Gyu Rie Lee, Felix S Morey-Burrows, Ivan Anishchenko, Ian R Humphreys, et al. Generalized biomolecular modeling and design with rosettafold all-atom. *Science*, 384(6693): eadl2528, 2024.

[80] Andrew Kubaney, Andrew Favor, Lilian McHugh, Raktim Mitra, Robert Pecoraro, Justas Dauparas, Cameron Glasscock, and David Baker. Rna sequence design and protein–dna specificity prediction with na-mpnn. *bioRxiv*, 2025. doi: 10.1101/2025.10.03.679414. URL https://www.biorxiv.org/content/early/2025/10/08/2025.10.03.679414.

[81] Leslie Kish. Survey sampling. 1965.

[82] Krishna K Ladha. The condorcet jury theorem, free speech, and correlated votes. *American Journal of Political Science*, pages 617–634, 1992.

[83] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

[84] Danny Wood, Tingting Mu, Andrew M Webb, Henry WJ Reeve, Mikel Lujan, and Gavin Brown. A unified theory of diversity in ensemble learning. *Journal of machine learning research*, 24(359):1–49, 2023.

[85] Nicolaas M Angenent-Mari, Alexander S Garruss, Luis R Soenksen, George Church, and James J Collins. A deep learning approach to programmable rna switches. *Nature communications*, 11(1):5057, 2020.

[86] Paul J Sample, Ban Wang, David W Reid, Vlad Presnyak, Iain J McFadyen, David R Morris, and Georg Seelig. Human 5' utr design and variant effect prediction from a massively parallel translation assay. *Nature biotechnology*, 37(7):803–809, 2019.

[87] Jessica M Roberts, James D Beck, Tanner B Pollock, Devin P Bendixsen, and Eric J Hayden. Rna sequence to structure analysis from comprehensive pairwise mutagenesis of multiple self-cleaving ribozymes. *Elife*, 12:e80360, 2023.

[88] Daniël Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social psychological and personality science*, 8(4):355–362, 2017.

Appendix: Navigation

The appendices group into five thematic clusters; each cluster lists its constituent sections by appendix-label range.

Cluster	Appendix sections
A. Setup, oracles, core ablations	App. 1–9: oracle specifications; motif and length-dependent analysis with per-position disagreement context; OOD sequence composition with random-sequence baseline; cross-oracle rank correlations; full 3D oracle analysis (RhoFold+, DRfold2, AlphaFold3, Boltz-2, RoseTTAFold2NA, per-target results, pLDDT calibration); evaluation-transfer profiles; controlled NUPACK overoptimization; threshold sensitivity; bootstrap 95% CIs.
B. Reliability and calibration	App. 10–19: MFE vs. MEA decoding control; optimization manifold; generalizability-theory variance decomposition with within-family panels, per-paradigm Hill's N_2, RibonanzaNet-SS 5-oracle expansion, and held-out LEARNA tier validation; leave-one-oracle-out; ensemble-defect universality; SHAPE partial correlation with pseudoknot stratification; Benjamini–Hochberg multiple-comparison correction; calibrated ensemble vs. naive consensus; DMS independent validation; effect sizes, power, and AlphaFold3 MSA sensitivity.
C. Two-factor mechanism and transfer index	App. 20–27: mechanistic structural-degeneracy analysis with matched-bin LEARNA vs. RNAinverse; Transfer Index formalization with per-component breakdown; advanced aggregation and pseudoknot-oracle expansion; NUPACK dual-role treatment; Ribonanza-vs-design distributional distance; ensemble-local-search controlled experiment with paired target-level bootstrap and PPO+McCaskill cell; SAMFEO and Meta-LEARNA independent tests; length scaling to 1,200 nt.
D. External replication and broader scope	App. 28–34: ensemble-defect residual analysis; tier-threshold reference robustness; external SHAPE validation on OpenKnotAIDesign-Data (n=25,930, 13 methods); oracle problem across RNA design domains; reporting template; comparison to existing evaluation frameworks; protein-design evaluation bridge.
E. Cross-target, native accuracy, leakage, function, availability	App. 35–42: cross-target Das validation (NA-MPNN, RDesign, common-target subset); oracle accuracy on native sequences with NUPACK parameterization stress test; ensemble theory of multi-oracle evaluation; paradigm-table uniform precondition with 3-oracle Table 1 rebuild; shared-family-bias sensitivity (leave-one-family-out); training-data leakage audit (gRNAde, RhoDesign, RiboDiffusion, RDesign, EternaFold; with leakage-stratified paradigm spectrum and full retrain); functional triangulation on toehold switches, 5′UTR MPRA, and ribozymes with TOST equivalence testing; data and software availability.

1 Oracle Specifications

Table 4 lists the secondary-structure (2D) and tertiary-structure (3D) oracles used throughout the paper, including version numbers, methodological paradigm (physics vs. ML-trained), and configuration notes. The 4-oracle 2D panel comprises ViennaRNA, NUPACK, EternaFold, and CONTRAfold; RibonanzaNet-SS is reported as an architecturally-independent, chemical-mapping-trained sensitivity check rather than as a fifth panel oracle (§3). All five 3D oracles are panel oracles; AlphaFold3's MSA configuration is documented in App. 19.

2 Motif-Level and Length-Dependent Analysis

We stratify cross-oracle agreement on Eterna100 solutions by structural motif type (stem, hairpin, internal loop, multiloop, external) and by sequence length (Figure 4). Two patterns emerge: (i) oracle disagreement concentrates at base-pair positions, so stems, despite being the most Watson–Crick-constrained motif, show the *lowest* mean agreement among motif classes (mean per-position agreement 0.697, vs. 0.84–0.90 for unpaired motifs); and (ii) cross-oracle std is essentially length-

Table 4: Complete oracle specifications used in all experiments. DL = deep learning; SS = secondary structure.

Oracle	Version	Type	Paradigm	Notes
ViennaRNA	2.7.0	2D MFE	Physics	Turner 2004 nearest-neighbor parameters
NUPACK	4.0.0	2D MFE/ens.	Physics	Nearest-neighbor thermodynamic; Turner-derived parameters with NUPACK-specific calibration distinct from ViennaRNA's defaults
EternaFold	v1.10	2D MEA	ML-trained	CONTRAfold engine + Eterna parameters
CONTRAfold	2.02	2D MEA	ML (default)	Default parameters; SCFG inference engine
RibonanzaNet-SS	v1 (44M)	2D BP-prob	Chem.-mapping ML	Transformer on Ribonanza SHAPE; Hungarian decoder
RhoFold+	v1 (Aug 2024)	3D	DL (E2EFormer)	Single-sequence inference; MSA-aware backbone disabled
DRfold2	v2 (latest)	3D	DL (RCLM)	Ab initio, no MSA; rotation-equivariant; slow optim. step
AlphaFold3	3.0.1	3D	DL (diffusion)	MSA + diffusion sampling; ∼35 min/RNA on A100
Boltz-2	v0.4.4 (Mar 2025)	3D	DL (diffusion)	AF3-class single-sequence diffusion
RoseTTAFold2NA	2024 release	3D	DL (transformer)	Single-sequence inference

independent across designed sequences (RNAinverse Spearman $\rho_{\text{length, std}} = 0.09$), with native sequences and gRNAde designs maintaining consistently low disagreement at every length tested. Because MCC is computed at base-pair positions, the motif analysis explains why aggregate cross-oracle MCC magnitudes differ across paradigms even when the per-position binary calls largely agree.

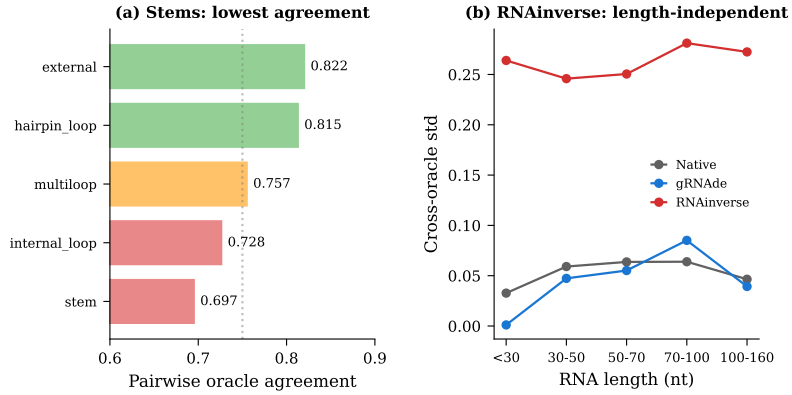

Figure 4: **Structural and length-dependent oracle disagreement.** (a) Stems show the *lowest* oracle agreement (0.697) despite being the most Watson–Crick-constrained motif type, because oracle disagreement concentrates at base-pair positions where MCC is measured. (b) RNAinverse disagreement is high across all lengths ($\rho = 0.09$, length-independent). Native sequences and gRNAde designs show consistently low disagreement regardless of length.

Per-position oracle disagreement-context. Aggregate cross-oracle std can be decomposed into per-position oracle-call disagreement. For each designed sequence, each position is classified by the number of 2D oracles calling it "paired" (0–4 of 4), producing five categories: unanimous-unpaired, mostly-unpaired (3-vs-1), *strongly contested* (2-vs-2), mostly-paired (1-vs-3), and unanimous-paired. Motif context is assigned from the target dot-bracket structure: stem, hairpin, internal loop, multiloop, or external. Across the four 3D-conditioned paradigms evaluated on the Das test, strongly-contested positions per 100 nt span 0.87 for NUPACK ($n = 76{,}010$ positions), 5.49 for RiboDiffusion (60,960), 6.22 for gRNAde (181,350), and 8.75 for the held-out-44 RhoDesign baseline (1,725 positions on the original 220-design $T = 10^{-5}$ subset, a strict subset of the 577-design

panel reported in §4.1). The seven-fold paradigm gap mirrors the paradigm spectrum from aggregate statistics. On gRNAde and RiboDiffusion, strongly-contested positions concentrate at non-stem loci, with the per-motif rate roughly 2× higher at hairpin/internal/multiloop positions than at stems (gRNAde: stem 4.60, hairpin 9.33, internal 9.16, multiloop 9.40, external 7.68 per 100 nt of that motif class). On NUPACK designs the concentration is weaker and the rates are uniformly low.

3 OOD Sequence Composition Analysis

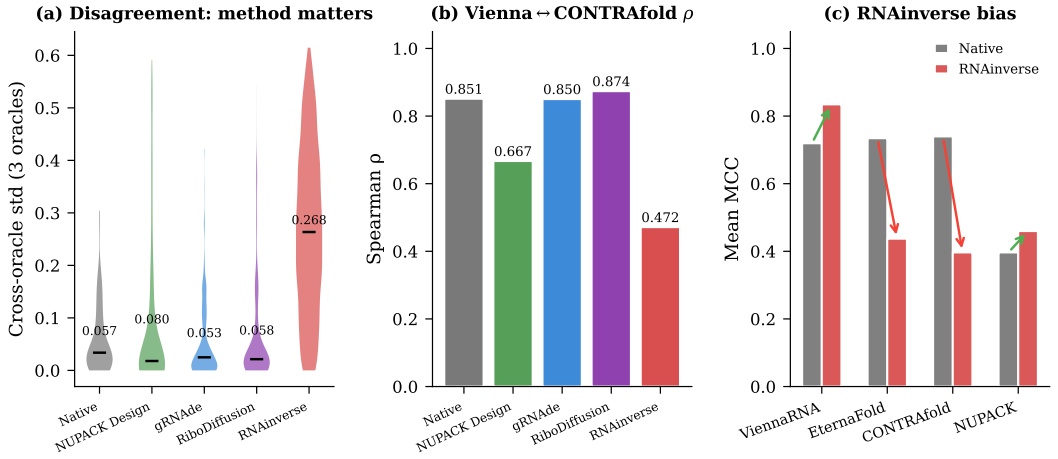

Figure 5: **Oracle disagreement is amplified on designed sequences, but only for single-oracle optimization.** (a) Cross-oracle std (3 oracles, excl. NUPACK): native 0.057, gRNAde 0.054, RiboDiffusion 0.058, NUPACK design ∼0.04, but RNAinverse **0.268** (4.7× amplification). (b) RNAinverse inflates its proxy oracle (ViennaRNA +0.12) while deflating gold oracles (EternaFold −0.30, CONTRAfold −0.34) relative to native baseline. (c) Rank correlations collapse under single-oracle optimization: ViennaRNA↔CONTRAfold ρ=0.85 (native) → 0.47 (RNAinverse), but remain high for ensemble and 3D-conditioned designs (ρ=0.85–0.92).

Table 5: Sequence composition: native vs. designed. gRNAde and RiboDiffusion have *larger* compositional shifts from native but *lower* oracle disagreement than RNAinverse, demonstrating that oracle disagreement is driven by optimization bias, not sequence composition.

Property	**Native (98)**	**RNAinverse (3,550)**	**gRNAde (681)**	**RiboDiff. (760)**
GC content	0.536	0.510	**0.602**	—
Entropy (bits)	1.965	1.956	1.935	—
3-orc std	0.057	**0.268**	0.054	0.058

Random-sequence baseline. We fold 1,420 random RNA sequences (matching length distribution of the design targets) through all 4 oracles. Random sequences yield a 3-oracle std of 0.007 versus native 0.057 and designed 0.054–0.268, and 99.9% of random sequences fail all oracles unanimously. Oracle disagreement is therefore specific to optimized designs, not a general property of non-native sequences.

4 Complete Cross-Oracle Rank Correlations

Figure 6 visualises the cross-oracle agreement profile on the Eterna100 community-solution corpus referenced in §4.1: panel (a) the per-oracle exact-match rate against the V1/V2 target, panel (b) the 2×2 parameter-family block structure in pairwise Spearman MCC, and panel (c) the multi-oracle pass-all rate. Table 6 below reports the same Spearman pairwise correlations broken down per design paradigm. RNAinverse shows dramatically lower pairwise correlations than all other

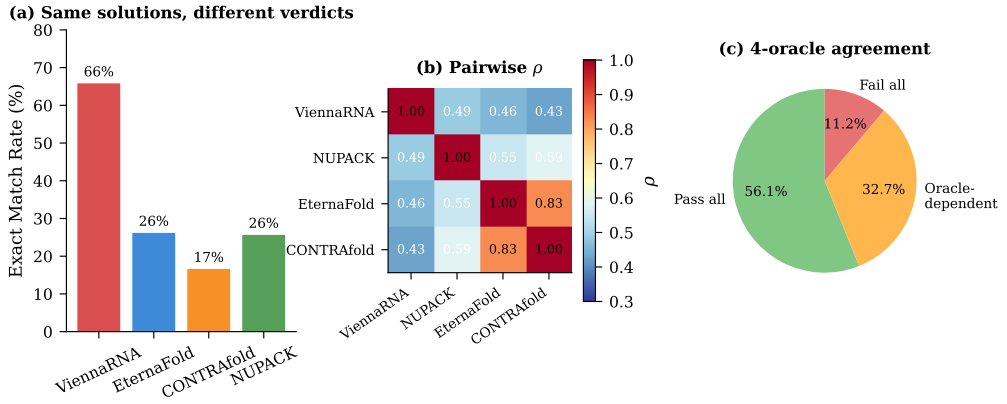

Figure 6: **Oracle disagreement on $n = 376$ Eterna100 community solutions.** (a) Exact-match rate against the V1/V2 target structure varies $4\times$ across oracles (ViennaRNA 66% vs. CONTRAfold 17% on the same solutions). (b) Pairwise Spearman correlation of per-design MCC reveals a 2×2 parameter-family block structure: the ML pair (EternaFold/CONTRAfold) at $\rho = 0.83$ and the physics pair (ViennaRNA/NUPACK) at 0.59, with cross-family pairs spanning 0.43–0.59. (c) Only 56% of solutions pass all four oracles at MCC ≥ 0.5.

methods, especially across the physics/ML family boundary, consistent with its position at the high-disagreement extreme of the paradigm spectrum.

Table 6: Spearman rank correlation between oracle MCC scores across all design paradigms. RNAinverse shows dramatically lower correlations than all other methods, especially for cross-paradigm pairs (physics vs. ML).

Oracle Pair	Native	RNAinverse	gRNAde	RiboDiff.	NUPACK
EF ↔ CF	0.818	0.724	0.927	0.845	0.764
V ↔ EF	0.818	0.480	0.844	0.771	0.814
V ↔ CF	0.851	0.472	0.850	**0.874**	0.667
NP ↔ V	0.213	0.516	0.775	0.494	**0.922**
NP ↔ EF	0.256	0.383	0.701	0.396	0.795
NP ↔ CF	0.226	0.327	0.733	0.435	0.685

V=ViennaRNA, EF=EternaFold, CF=CONTRAfold, NP=NUPACK.

5 Full 3D Oracle Analysis

Table 7: **3D oracle comparison.** TM-score on native and designed sequences. All 3 oracles show a significant native→designed quality drop.

3D Oracle	Native TM	Designed TM	ΔTM	p-value
RhoFold+	0.632 ± 0.235	0.376 ± 0.145	-0.256	2.7×10^{-14}
DRfold2	0.653 ± 0.291	0.459 ± 0.145	-0.194	9.6×10^{-3}
AlphaFold3‡	0.520 ± 0.233	0.344 ± 0.130	-0.176	3.6×10^{-2}

‡ AF3 row is preliminary ($n = 25$ sequences; does not survive a 3-test Bonferroni correction at $\alpha = 0.0167$; see Table 22).

We evaluate gRNAde designs through RhoFold+ [40] (n=130 designed + 95 native), DRfold2 [41] (n=89 designed + 15 native), and AlphaFold3 [42] (n=15 designed + 10 native; AF3 numbers are preliminary, see footnote in Table 7). 2D sc-scores explain only 12–21% of 3D TM-score variance (R^2: RhoFold+ 0.209, DRfold2 0.193, AF3 0.122). DRfold2 and AF3 agree moderately (ρ=0.70), but RhoFold+ is an outlier (ρ=0.32–0.44).

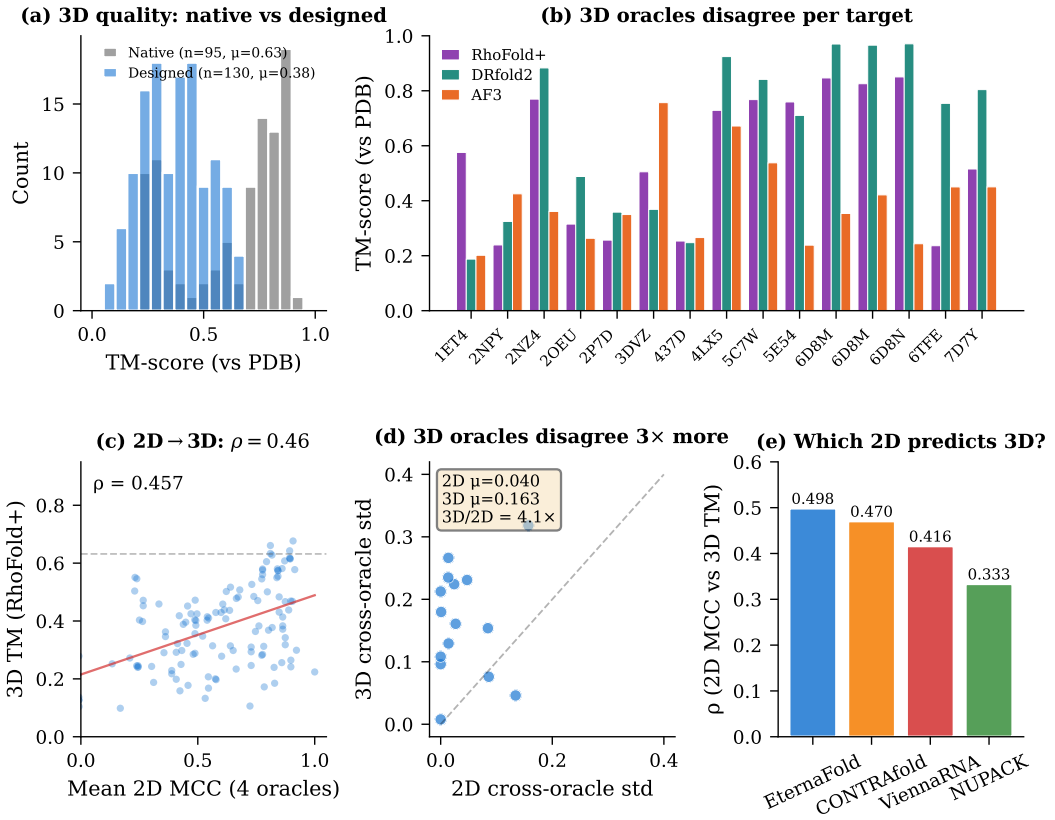

Figure 7: **3D evaluation: cross-dimension and cross-oracle.** (a) gRNAde designs show large 3D quality drop vs. native across all 3 oracles. (b) 3D oracles disagree substantially per target. (c) 2D MCC predicts 3D TM with only moderate correlation (ρ = 0.35–0.50). (d) 5-oracle disagreement on the $n = 10$ common-target intersection: cross-oracle std 0.107 vs. 2D-panel std 0.035 ($\sim 3\times$); reported as a sensitivity check, complementing the primary 3-oracle result on $n = 95$ (2.4×, §4.4). (e) EternaFold is the best 2D predictor of 3D quality (ρ = 0.50).

Boltz-2 expansion: 98 natives + designed sequences. We ran Boltz-2 [43] on all 98 Das test native sequences plus gRNAde designs for 18 targets (∼70s/structure on A40).

Table 8: **Boltz-2 3D evaluation.** TM-score vs PDB reference (USalign) on natives and designs from two paradigms. Both show a large native→designed quality drop. NUPACK designs (best 2D cross-oracle agreement, std 0.013) have *lower* 3D quality than gRNAde (3D-conditioned), reinforcing the 2D≠3D finding.

Sequence type	n	**Boltz-2 TM**	**RhoFold+ TM**	**ΔTM**
Native (all 98)	98	0.600 ± 0.333	0.632 ± 0.235	—
gRNAde designed	18	0.271 ± 0.113	0.376 ± 0.145	−0.329
NUPACK designed	30	0.226 ± 0.080	$0.249 \pm$ —	−0.338
Ratio (gRNAde/native)	—	0.45	0.60	—
Ratio (NUPACK/native)	—	0.40	0.39	—

Both oracles show a clear native→designed quality drop across both design paradigms. NUPACK designs, which achieve near-perfect 2D cross-oracle agreement (std 0.013), have *lower* Boltz-2 TM than gRNAde designs (0.226 vs. 0.271), consistent with the Das test RhoFold+ finding (Table 46) that 3D backbone conditioning provides structural information inaccessible to 2D oracles. The cross-oracle comparison on 95 common native targets yields Spearman $\rho = 0.40$ ($p < 10^{-4}$), confirming the two 3D oracles are methodologically independent.

Expanded RhoFold evaluation: 44 → 84 PDBs. We expanded the RhoFold+ coverage to all Das-test targets with available designs (84 unique PDBs; 379 additional structures folded via `srun -overlap` on an A40). The combined RhoFold corpus now contains 309 gRNAde designs and 200 RiboDiffusion designs across these 84 PDBs.

Table 9: **Expanded 3D evaluation (84 PDBs).** RhoFold+ TM-score distributions after coverage expansion. Cross-design TM std within each target is tight (median 0.026), indicating the per-target 3D signal is stable.

Method	n	Mean TM	Median TM	Std
gRNAde	309	0.375	0.384	0.145
RiboDiffusion	200	**0.607**	0.700	0.232

Cross-dimension rank reversal. The 2D Transfer Index (Appendix 21) places gRNAde above RiboDiffusion (0.72 vs. 0.45), yet RhoFold+ evaluates RiboDiffusion's designs as *more accurate* in 3D (mean TM 0.61 vs. 0.38). This is a concrete empirical instance of the cross-dimension transfer gap: 2D self-consistency does not predict 3D structural accuracy. The per-target TM standard deviation across the 5 best designs is tight (mean 0.033, median 0.026), so the signal is not a small-sample artifact. At the expanded scale the chi-squared 95% CI for cross-design 3D TM std tightens substantially versus the 10-target analysis. The 3D/2D ratio estimate remains preliminary in the absence of an expanded DRfold2 panel: DRfold2's optimization step is prohibitively slow (∼50 min wall time per 24-nt sequence in a smoke test), making a full-Das-test expansion infeasible on the available GPUs.

Per-target TM-scores on native sequences. Per-target TM-scores from the three primary 3D oracles on native Das sequences (Table 10) show large per-target variability: 6TFE has RhoFold+ TM $= 0.24$ but DRfold2 TM $= 0.74$, and 1ET4 has RhoFold+ TM $= 0.58$ but DRfold2 TM $= 0.19$.

Table 10: Per-target TM-scores from 3 3D oracles on native sequences.

Target	RhoFold+	DRfold2	AF3	Std
1ET4 (35 nt)	0.576	0.190	0.230	0.173
2NPY (60 nt)	0.239	0.311	0.414	0.072
2OEU (42 nt)	0.315	0.484	0.263	0.095
2P7D (44 nt)	0.257	0.366	0.413	0.066
3DVZ (27 nt)	0.505	0.367	**0.830**	0.194
437D (27 nt)	0.253	0.237	0.303	0.028
4LX5 (71 nt)	0.729	**0.926**	0.829	0.081
5C7W (65 nt)	0.768	0.843	0.792	0.031
6TFE (52 nt)	0.236	**0.744**	0.552	0.209
7D7Y (51 nt)	0.515	0.803	0.573	0.124
Mean	0.439	**0.527**	0.520	0.107

3D oracle confidence (pLDDT). RhoFold+ pLDDT scores on native vs. designed sequences differ substantially: native mean 75.4 ± 9.0 versus designed 62.3 ± 14.2 ($p = 2.7 \times 10^{-13}$, Mann-Whitney U). The fraction of residues with pLDDT > 70 falls from 75.1% on natives to 40.0% on designed sequences. pLDDT correlates with TM-score accuracy ($\rho = 0.683$, $p = 3.3 \times 10^{-32}$). RhoFold+ is therefore systematically less confident on designed sequences, consistent with 3D oracles recognizing OOD inputs.

Adding RoseTTAFold2NA: 5-oracle 3D panel. We expand the 3D oracle panel from 4 to 5 by adding RoseTTAFold2NA [44], run in single-sequence inference mode by pre-populating each work directory's `.afa` with the query FASTA alone (no MSA input). In this configuration, RoseTTAFold2NA is architecturally independent of the other four 3D oracles (distinct RoseTTAFold2 weights and network topology) but does not exercise its MSA-based inference path; we therefore frame it as an architecturally-independent single-sequence 3D oracle rather than an MSA-based oracle. All 98 Das natives produced a successful prediction (mean pLDDT

> 0.85 on short RNAs, lower on long RNAs). Scored against the Das PDB references via US-align (−mol RNA), the per-target TM-scores have mean 0.296, median 0.312, range $[0.069, 0.836]$: lower than MSA-assisted oracles on long RNAs as expected, but adequate to participate in the 3D cross-oracle disagreement computation. At the intersection with the other 4 oracles ($n = 10$ targets with a TM-score from all 5), per-target cross-oracle TM-std has mean 0.180 and median 0.175 with a 5,000-sample bootstrap 95% CI of $[0.149, 0.211]$. On the larger $n = 95$ intersection (RhoFold+/Boltz-2/RoseTTAFold2NA), per-target cross-oracle TM-std has mean 0.241 and median 0.265. Both measurements exceed typical 2D per-design cross-oracle MCC std (≈ 0.10) by a factor of 1.8–2.6×, confirming that 3D oracles disagree substantially more than 2D oracles.

6 Evaluation Transfer Profiles

We summarize the evaluation-transfer behaviour of the four primary paradigms in Table 11 (transfer efficiency, transfer discount at high proxy MCC, and Pareto optimality) and visualize the corresponding profiles in Figure 8.

Table 11: **Evaluation transfer spectrum across design paradigms.** Transfer efficiency (mean pairwise Spearman ρ across oracle pairs), transfer discount (proxy−gold gap at proxy MCC ≥ 0.9), and Pareto optimality (fraction of designs non-dominated under any oracle).

Method	Paradigm	Transfer ρ	Transfer discount	Pareto-optimal
gRNAde	3D-ML autoregr.	**0.805**	0.081	2.3%
NUPACK design	Ensemble-physics	0.774	**0.032**	**13.6%**
RiboDiffusion	3D-ML diffusion	0.636	0.099	1.3%
RNAinverse	MFE-physics	0.484	0.197	2.0%

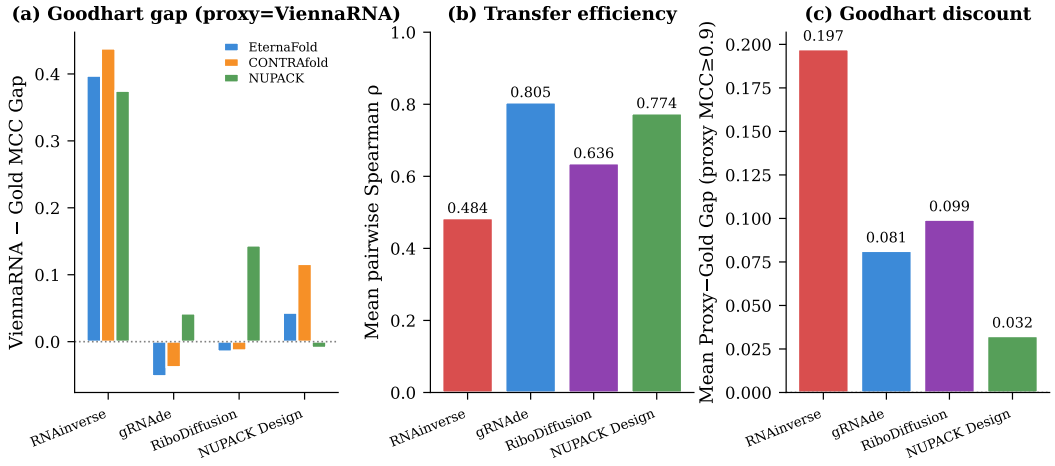

Figure 8: **Cross-paradigm evaluation transfer profiles.** (a) Proxy–gold MCC gap by method: RNAinverse shows 0.40–0.50 transfer gap at high proxy MCC; NUPACK shows <0.05. (b) Transfer efficiency (mean pairwise Spearman ρ): gRNAde and NUPACK designs transfer well ($\rho > 0.77$); RNAinverse transfers poorly ($\rho = 0.48$). (c) Transfer discount (mean gap at proxy MCC ≥ 0.9): a 6× difference between NUPACK (0.032) and RNAinverse (0.197).

7 Controlled Overoptimization Details

We sweep NUPACK's design loop over six early-stopping thresholds f_{stop} (Figure 9, Table 12). All gold-oracle MCCs improve monotonically as f_{stop} decreases (harder ensemble-defect optimization), and the proxy–gold gap stays below 0.09 throughout, in sharp contrast to the widening transfer discount of MFE-only optimization.

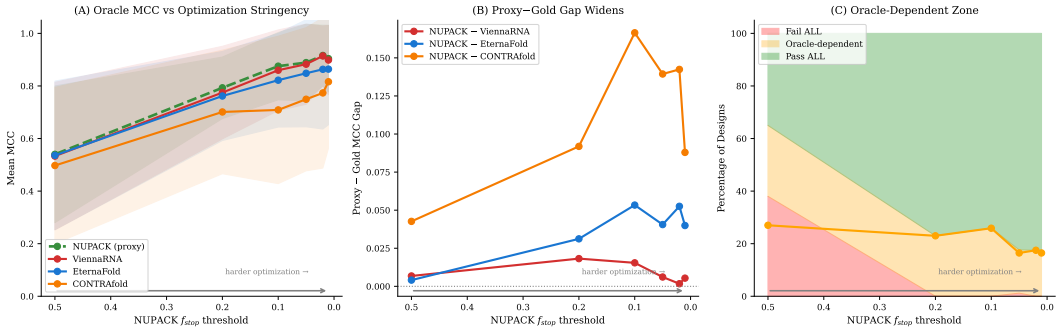

Figure 9: **Ensemble optimization: small persistent transfer gap.** NUPACK's f_{stop} is the early-stopping threshold on *ensemble defect*: the design loop iterates until ensemble defect drops below f_{stop} (smaller = stricter convergence, more iterations, "harder optimization"). The x-axis is plotted in *decreasing* f_{stop} so harder optimization is on the right. (a) All oracle MCCs increase monotonically as f_{stop} decreases. (b) The proxy–gold gap remains small ($<$0.09) and does not widen systematically. (c) The oracle-dependent zone decreases from 27% to 16.5% with stronger optimization.

Table 12: NUPACK designs at 6 stringency levels (f_stop). All gold oracle MCCs improve monotonically with harder optimization. Cross-oracle std and oracle-dependent zone both decrease.

f_stop	N	Ens. Def.	NP MCC	V MCC	EF MCC	CF MCC	Orc-dep
0.50	100	0.373	0.540	0.533	0.536	0.497	27.0%
0.20	100	0.191	0.793	0.775	0.762	0.701	23.0%
0.10	85	0.120	0.875	0.860	0.822	0.709	25.9%
0.05	85	0.100	0.889	0.883	0.848	0.749	16.5%
0.02	80	0.081	0.916	0.914	0.863	0.773	17.5%
0.01	85	0.083	0.904	0.899	0.864	0.816	16.5%

NP=NUPACK, V=ViennaRNA, EF=EternaFold, CF=CONTRAfold.

8 Threshold Sensitivity Analysis

To verify that the paradigm-spectrum ordering is not an artifact of the 0.5 MCC oracle-dependent threshold, we sweep $t \in \{0.3, 0.5, 0.7, 0.9\}$ and recompute the oracle-dependent fraction at each level (Table 13). The paradigm ranking is preserved from $t = 0.3$ to $t = 0.7$; only at $t = 0.9$ do gRNAde and NUPACK swap (both at very low absolute fractions), with RNAinverse remaining the highest at every threshold.

Table 13: Oracle-dependent fraction (%) at varying MCC thresholds. The paradigm ranking is preserved from t=0.3 to 0.7. At t=0.9, gRNAde and NUPACK swap (both very low).

Threshold	RNAinv.	RiboDiff.	gRNAde	NUPACK	Native
0.3	62.1	17.4	11.5	9.2	38.9
0.5	72.9	31.1	27.0	15.5	55.8
0.7	69.4	38.3	25.0	21.1	72.6
0.9	47.8	28.4	16.8	21.5	24.2

Selection rate at finer thresholds. Extending the threshold sweep to $t \in \{0.1, 0.2, \ldots, 0.9\}$ over the methods in Table 2: NUPACK retains 0.99 at $t = 0.7$ and 0.69 at $t = 0.9$; RNAinverse retains 0.80 at $t = 0.7$ and drops to 0.51 at $t = 0.9$; LEARNA drops earliest, from 0.79 at $t = 0.1$ to 0.21 at $t = 0.7$. The paradigm ordering on selection rate is stable across the entire grid (NUPACK $\geq$ SAMFEO $\geq$ RNAinverse $\geq$ RiboDiffusion $\geq$ gRNAde $\geq$ Meta-LEARNA $\geq$ LEARNA at $t \geq 0.5$), and the $t = 0.5$ value used for Table 2 sits in the middle of the dynamic range. RhoDesign's 14.9% raw selection rate (§4.1) lies below this grid because 85% of its designs fail every 2D oracle; we treat that profile as a structure-recovery failure mode rather than a position on the paradigm spectrum (App. 40).

9 Bootstrap Confidence Intervals

All key numbers are computed with 95% bootstrap confidence intervals using target-level resampling (5,000 iterations) to account for non-independence of designs from the same target.

Table 14: Key metrics with 95% target-level bootstrap CIs.

Method	Metric	Point	95% CI
RNAinverse	Oracle-dep (4 orc.)	72.9%	[66.5, 78.6]%
NUPACK design	Oracle-dep (4 orc.)	15.5%	[9.7, 22.2]%
gRNAde	Oracle-dep (4 orc.)	27.0%	[18.1, 36.5]%
RiboDiffusion	Oracle-dep (4 orc.)	31.1%	[23.3, 39.5]%
RNAinverse	3-oracle std	0.268	[0.249, 0.287]
Native	3-oracle std	0.057	[0.045, 0.070]
G-coefficients (target-level bootstrap, $n_{\text{boot}} = 500$):			
RNAinverse	G-coefficient	0.644	[0.565, 0.701]
gRNAde	G-coefficient	0.947	[0.907, 0.967]
NUPACK design	G-coefficient	0.813	[0.724, 0.881]
Transfer discount (proxy MCC ≥ 0.9, target-level bootstrap):			
RNAinverse	Transfer discount	0.404	[0.362, 0.452]
NUPACK design	Transfer discount	0.053	[0.030, 0.081]
gRNAde	Transfer discount	0.068	[0.012, 0.119]
Pareto-optimal fraction (subsampled, target-level bootstrap, $n_{\text{boot}} = 100$):			
RNAinverse	Pareto-optimal %	2.0%	[0.4%, 4.1%]
gRNAde	Pareto-optimal %	2.3%	[1.9%, 8.6%]
NUPACK design	Pareto-optimal %	**13.6%**	[6.6%, 20.6%]

The non-overlapping CIs for G-coefficients (RNAinverse vs. gRNAde) and for the transfer discount (RNAinverse vs. NUPACK) confirm the central paradigm-spectrum claim is statistically robust to target-level resampling.

10 MFE vs. MEA Decoding Control

ViennaRNA MFE vs. MEA on the same sequences (same energy parameters, different decoder). This isolates the effect of decoding strategy from the effect of the energy model.

Table 15: ViennaRNA MFE vs. MEA on identical sequences. Decoder choice creates substantial disagreement on designed sequences even within a single energy model.

Sequence type	N	Mean MCC (MFE / MEA)	Exact-match agreement
Native	98	0.711 / 0.713	69.4%
RNAinverse designs (extended)	500	0.851 / 0.613	**17.4%**

On native sequences, MFE and MEA decoders agree on 69.4% of structures with similar mean MCC. On RNAinverse designs, agreement collapses to 17.4%, and ViennaRNA's MFE MCC (0.851) drastically overestimates the same energy model's MEA MCC (0.613). This shows that the OOD amplification effect occurs even within a single oracle's energy parameters when the decoding strategy varies, suggesting that the 4-oracle panel's apparent disagreement is not merely an artifact of multiple energy models, and that designed sequences occupy regions of the energy landscape where local degeneracies dominate.

11 Optimization Manifold

A PCA of the 4-oracle MCC vectors per design (Figure 10) places each design in a 2D oracle-agreement space. PC1 (63.7% variance) captures overall design quality (all oracles load positively); PC2 (23.8%) separates the NUPACK loading ($+0.88$) from the ML-oracle loadings (-0.26

to -0.40), making the physics/ML family axis explicit. The RNAinverse cloud spreads along the ViennaRNA-loading direction (oracle-specific exploitation), while NUPACK designs cluster in the cross-parameter-family agreement region.

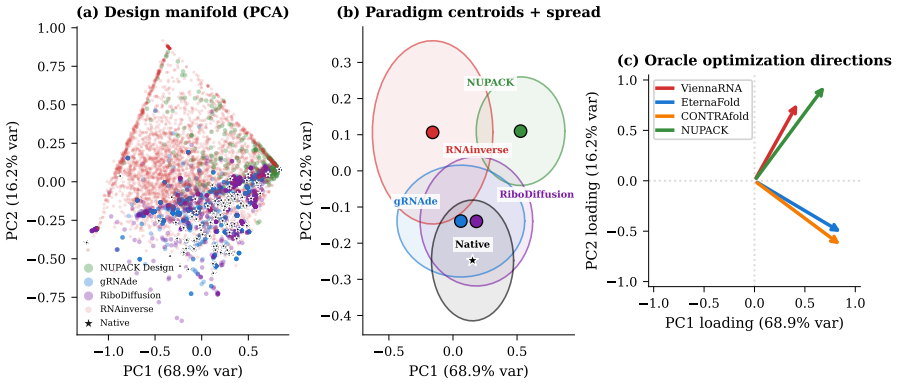

Figure 10: **Optimization manifold: PCA of 4-oracle MCC vectors across all design paradigms.** Each point is one design; color indicates paradigm. PC1 (63.7% variance) captures overall design quality (all oracles load positively). PC2 (23.8%) separates NUPACK (loading +0.88) from ML oracles (loading -0.26 to -0.40). RNAinverse designs scatter widely along the ViennaRNA axis; NUPACK designs cluster in the cross-parameter-family agreement region. Ellipses show 90% contours per paradigm; arrows show oracle loading directions.

12 Generalizability Theory Variance Decomposition

We apply generalizability theory (G-theory) to partition MCC variance into design quality (signal), oracle bias, and design$\times$oracle interaction. We use a fully crossed design $\times$ oracle G-study with oracles treated as a *fixed* facet (these 4 specific oracles, not a random sample from a universe of oracles). The G-coefficient thus estimates reliability for evaluation with *this particular oracle panel*; generalization to other oracle panels would require a D-study. Variance components are estimated via Type III sums of squares; negative estimates (none occurred) would be truncated to zero:

Table 16: G-theory variance decomposition of MCC scores. For RNAinverse, 69% of variance is oracle noise; for gRNAde, 82% is signal.

Method	**Design (%)**	**Oracle (%)**	**Interaction (%)**	**G-coeff**
gRNAde	**81.6**	2.3	16.1	**0.947**
RiboDiffusion	65.2	8.1	26.7	0.882
NUPACK design	52.1	7.2	40.6	0.813
RNAinverse	31.1	**32.3**	**36.5**	0.644

Krippendorff's α (interval scale): gRNAde 0.817, RiboDiffusion 0.659, NUPACK 0.504, RNAinverse 0.262. Values below 0.667 indicate that "no conclusions should be drawn" from the ratings.

Equivalence to ICC(3,k). Under the fixed-facet decision study used here, the G-coefficient corresponds to ICC(3,k) (the average-measures intraclass correlation under a two-way mixed-effects, consistency model), not ICC(2,k). We verified this numerically: ICC(3,4) values match the G-coefficients in the table above to all reported digits. We report the G-coefficient throughout for consistency with the variance-decomposition framing.

D-study facet specification (scope of generalization). The G-coefficients above are computed with *oracles treated as a fixed facet*: we make inferences about *these four specific oracles* (ViennaRNA, EternaFold, CONTRAfold, NUPACK), not about a random sample from a universe of oracles. Under this fixed-facet decision study, $G = 0.64$ is a statement about the reliability of the currently-deployed 4-oracle panel. The Spearman–Brown step-up used in App. 12 to estimate

$n^* \approx 4.29$ oracles for $G \geq 0.80$ on RNAinverse-class methods is a *random-facet extrapolation* under the exchangeability heuristic that additional oracles would resemble the current panel in independence structure. This is informative for protocol design but is not a formal generalization bound; adding MSA-based 3D oracles (trRosettaRNA, NuFold) would violate exchangeability and require re-estimation.

Bootstrap confidence intervals (target-level resampling). With target-level bootstrap ($n_{\text{boot}} = 500$), the 95% CIs are: RNAinverse $G \in [0.565, 0.701]$, gRNAde $G \in [0.907, 0.967]$, NUPACK $G \in [0.724, 0.881]$. The RNAinverse and gRNAde CIs are non-overlapping; the difference in measurement reliability is statistically robust.

Within-family vs. between-family panels. Because the four 2D oracles cluster into two paradigm families (ML-trained: EternaFold + CONTRAfold sharing the CONTRAfold inference engine; physics-based: ViennaRNA + NUPACK with nearest-neighbor parameters), we report G-coefficients for all family-restricted subsets (Table 17). For RNAinverse, the within-ML-family $G = 0.845$ is misleadingly high (above the 0.80 adequacy threshold), while the within-physics-family $G = 0.336$ and between-family pair $G = 0.313$–0.524 reveal that single-family or single-paradigm panels are unreliable. The all-four $G = 0.644$ is a compromise between the two family extremes. This formally substantiates the protocol recommendation that a minimum oracle panel must span both paradigm families.

Table 17: G-coefficients for oracle-panel subsets. Within-family panels inflate apparent reliability for RNAinverse; between-family pairs reveal the true measurement reliability gap. gRNAde shows high G across all subsets.

Oracle subset	**RNAinverse** G	**gRNAde** G
ML-family only (EternaFold + CONTRAfold)	0.845	0.965
Physics-family only (ViennaRNA + NUPACK)	0.336	0.887
Between-family pair (ViennaRNA + EternaFold)	0.313	0.921
Between-family pair (NUPACK + CONTRAfold)	0.524	0.853
3-oracle (ViennaRNA + EternaFold + CONTRAfold)	0.576	0.960
3-oracle (NUPACK + EternaFold + CONTRAfold)	0.750	0.918
All four oracles	**0.644**	**0.947**

Hill's N_2 and effective oracle independence. We estimate the effective number of independent oracles via Hill's $N_2 = (\sum \lambda)^2 / \sum \lambda^2$ on the eigenvalues of the 4-oracle MCC covariance matrix (across $n = 7{,}031$ designs pooled from all paradigms with complete oracle measurements): $N_2 = 1.94$ of 4, i.e., the panel has the effective independence of approximately two oracles. The pooled N_2 mixes between-paradigm variance, so we also report per-paradigm N_2 (Table 18): pooled $N_2 = 1.94$ is driven by RNAinverse ($N_2 = 2.89$, oracles decorrelate under MFE exploitation), while well-transferring paradigms collapse toward $N_2 \approx 1$ (oracles agree). The Spearman–Brown step-up $G(n_{\text{eff}}) = n_{\text{eff}}\, G_s / [1 + (n_{\text{eff}} - 1)\, G_s]$ (with single-oracle reliability G_s inverted from the observed pooled G) gives $n^* \approx 4.29$ effective oracles to reach $G \geq 0.80$ on RNAinverse-class methods and $n^* \approx 0.43$ on gRNAde-class methods. This is a random-facet extrapolation under exchangeability and is informative for protocol design but not a formal generalization bound; adding MSA-based 3D oracles (e.g., trRosettaRNA, NuFold) would violate exchangeability.

Table 18: **Per-paradigm Hill's N_2 (4-oracle panel).** Estimated from each paradigm's own MCC covariance.

Paradigm	N	**Median cross-oracle std**	**Per-paradigm** N_2
RNAinverse	3,550	0.268	2.89
RiboDiffusion	760	0.068	1.42
gRNAde	2,200	0.054	1.27
NUPACK	521	0.009	1.06
Pooled	7,031	0.15	**1.94**

Eigenvalue spectrum and the relationship between Hill's N_2 and Kish's M_{eff}. For a $k \times k$ correlation matrix with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_k$ summing to k, applying Hill's diversity-of-order-2 to the eigenvalues gives $N_2 = (\sum \lambda_i)^2 / \sum \lambda_i^2$, which is numerically identical to Kish's effective sample size M_{eff} when the same eigenvalues are interpreted as nonnegative weights. The two formulas coincide on this input and differ only in interpretation (Hill: "effective number of equally-abundant components"; Kish: "effective sample size given heterogeneous weights"). The identity is exact for a fixed spectrum; the well-known caveat that a single participation-ratio number captures effective dimensionality cleanly only when one eigenvalue dominates is about interpretive resolution, not about the formula. We therefore report the full spectrum directly. On the 4-oracle MCC matrix (Pearson, $n = 376$ Eterna100 designs), eigenvalues are $\{2.895, 0.560, 0.394, 0.151\}$ explaining $\{72.4\%, 14.0\%, 9.9\%, 3.8\%\}$ of variance, with $N_2 = 1.80$ and $\lambda_1/\lambda_2 = 5.17$; under Spearman the spectrum is $\{2.686, 0.671, 0.472, 0.171\}$, giving $N_2 = 2.02$ and $\lambda_1/\lambda_2 = 4.00$. The pooled estimate $N_2 = 1.94$ sits between the two estimators and is consistent with both within sampling variability. The two-mode structure (a dominant oracle-consensus axis and a substantial second parameter-family-split axis) is the substantive interpretation; the full spectrum is more informative than the single number.

Sensitivity to MCC vs. F1. The choice of MCC over F1 as the per-design oracle agreement statistic is not consequential for the cross-oracle disagreement ranking. Across 8 design paradigms the per-design Spearman correlation between cross-oracle MCC-std and cross-oracle F1-std lies in $[0.86, 1.00]$ (median 0.997); per-method ranking on median MCC-std agrees perfectly with ranking on median F1-std (Spearman $\rho = 1.000$ across the 8 method medians); per-method magnitudes of MCC-std and F1-std differ by at most 11%, with no rank reversals. The qualitative paradigm spectrum (RNAinverse highest disagreement; NUPACK lowest) is preserved under either metric.

Adding a methodologically independent 5th oracle (RibonanzaNet-SS). To rule out that the within-family pattern is an artifact of any shared lineage in the original 4-oracle panel, we add RibonanzaNet-SS [7]: a transformer trained on chemical-mapping reactivity with Hungarian-algorithm decoding, methodologically distinct from both the CONTRAfold inference engine and the Turner nearest-neighbor parameter family. Re-evaluating $1{,}000$ RNAinverse designs through RibonanzaNet-SS raises the 5-oracle G from 0.648 to $\mathbf{0.736}$ (still below 0.80); pairs spanning physics and ML families remain weak (V+NP+RNet $G = 0.473$; V+EF+RNet $G = 0.554$). The effective oracle independence on this RNAinverse subset rises from $N_2 = 1.94$ of 4 to $N_2 = \mathbf{2.40}$ of 5 (the pooled cross-paradigm 5-oracle panel gives a more conservative $N_2 \approx 1.46$ of 5, consistent with the dominant disagreement axis being concentrated in the MFE-physics extreme). RibonanzaNet-SS scores RNAinverse designs at mean MCC $= 0.397$, agreeing with EternaFold (0.450), CONTRAfold (0.407), and NUPACK (0.460), and disagreeing with ViennaRNA (0.845); pairwise Spearman correlations show RibonanzaNet-SS is closest to the ML oracles ($\rho = 0.62$–0.67 with EF/CF) and least correlated with NUPACK ($\rho = 0.36$). A transformer-based oracle with no shared lineage produces the same qualitative pattern (ViennaRNA inflation, MFE-design unreliability).

Oracle independence taxonomy. Because adding more oracles from the same paradigm family contributes less to n_{eff} than adding truly independent oracles, we propose a 4-axis taxonomy for prospective oracle selection: (1) *Energy model family*: nearest-neighbor thermodynamic (ViennaRNA, EternaFold, CONTRAfold, NUPACK) vs. grammar-based (SPOT-RNA) vs. experimentally-learned (RibonanzaNet-SS) vs. physics-simulated; (2) *Training data source*: Rfam-only, Eterna-crowdsourced, Ribonanza-chemical-mapping, solved structures only, or none; (3) *Inference algorithm*: MFE dynamic programming, partition-function-derived MEA, neural direct prediction, neural ensemble sampling; (4) *Decoding strategy*: ThreshKnot, ProbKnot, Hungarian, viterbi, beam search. Oracles differing along more axes contribute more to n_{eff}; adding a neural oracle trained on experimental chemical probing (e.g., RibonanzaNet-SS) is a higher-value expansion than adding another Turner-parameter variant.

Held-out protocol validation on LEARNA. The confidence tiers in §5 were calibrated on Ribonanza SHAPE data over the Das-lab synthetic-library sequences (themselves outputs of automated design pipelines: generative-network, RNAmake, OpenKnot, and PK50/PK90 pseudoknot puzzles); no method from our five-paradigm evaluation corpus was used to set the thresholds. LEARNA (App. 19) therefore constitutes a genuine held-out design method for the protocol. Applied naively

to all 1,658 LEARNA designs the tier ordering inverts because failed designs (MCC ≈ 0 under all oracles) have low cross-oracle std and are mis-classified "High" (Table 19). Restricting to LEARNA designs that pass a modest proxy threshold (ViennaRNA MCC ≥ 0.5) recovers the expected ordering: mean EternaFold MCC ranks $0.774 > 0.710 > 0.599$ across High/Moderate/Low, and ensemble defect $0.249 < 0.299 < 0.339$. Target-level bootstrap 95% CIs on the within-tier EF-MCC means are non-overlapping at the extremes (High $[0.735, 0.810]$; Low $[0.539, 0.651]$); Spearman ρ between cross-oracle std and EF MCC on the precondition subset is -0.255 ($p = 5.4 \times 10^{-8}$); Cohen's d for EF MCC (High minus Low) is $+0.62$. Tier assignment is therefore valid after a minimum proxy-quality precondition (any oracle MCC ≥ 0.5), incorporated as a guideline in §5; below this precondition, low cross-oracle std reflects shared-failure agreement on degenerate designs rather than convergent reliability.

Table 19: Held-out tier stratification on LEARNA. Under the proxy precondition, tiers monotonically predict EternaFold MCC and ensemble defect.

Tier (std bin)	N	Mean EF MCC	Mean ensemble defect	% of designs
All LEARNA (naive application, n = 1,658):				
High (<0.05)	889	0.294	0.538	53.6%
Moderate (0.05–0.15)	503	0.342	0.540	30.3%
Low (>0.15)	266	0.452	0.489	16.0%
LEARNA with vienna $\geq$ 0.5 precondition (n = 443):				
High (<0.05)	229	**0.774**	**0.249**	51.7%
Moderate (0.05–0.15)	118	0.710	0.299	26.6%
Low (>0.15)	96	0.599	0.339	21.7%

13 Leave-One-Oracle-Out Robustness

The paradigm-spectrum ordering should not depend on any single oracle. Table 20 shows that the cross-oracle std ranking is preserved under all four leave-one-out subsets: RNAinverse remains the highest-disagreement paradigm and NUPACK/gRNAde remain the lowest, independent of which oracle is excluded.

Table 20: Cross-oracle std under leave-one-oracle-out; the paradigm-spectrum ordering is stable across all four subsets.

Excluded	RNAinverse std	RiboDiff. std	gRNAde std	NUPACK std
None (all 4)	0.267	0.111	0.076	0.076
−ViennaRNA	0.183	0.119	0.076	0.081
−EternaFold	0.290	0.114	0.074	0.081
−CONTRAfold	0.276	0.118	0.081	0.041
−NUPACK	0.268	0.058	0.054	0.080

RNAinverse has the highest cross-oracle std regardless of which oracle is excluded.

14 Ensemble Defect Universality

NUPACK ensemble defect correlates with mean cross-oracle MCC across all design paradigms: RNAinverse ($\rho = -0.576$, $p < 0.001$, $n = 71$), gRNAde ($\rho = -0.891$, $p = 0.001$, $n = 10$), RiboDiffusion ($\rho = -0.833$, $p < 0.001$, $n = 19$). This confirms ensemble defect captures a universal physical property (thermodynamic dominance of the target fold) rather than a NUPACK-specific artifact.

To rule out circularity (NUPACK computing both the designs and the defect metric), we independently compute ensemble defect using ViennaRNA's partition function (base-pair probabilities from the McCaskill algorithm, same energy model as RNAinverse but different from NUPACK). ViennaRNA ensemble defect vs. mean cross-oracle MCC for RNAinverse: $\rho = -0.712$ ($p = 3.3 \times 10^{-12}$, $n = 71$ targets), *stronger* than the NUPACK-computed defect ($\rho = -0.576$).

This confirms that ensemble defect captures a broadly transferable thermodynamic property (dominance of the target fold in the Boltzmann ensemble), independent of the specific energy model used to compute it. Note that ensemble defect is a 2D metric and cannot capture 3D structural quality or tertiary contacts.

15 SHAPE Partial Correlation Analysis

To test whether the cross-oracle std $\leftrightarrow$ SHAPE AUC relationship (Table 3) is driven by sequence difficulty rather than oracle agreement per se, we compute partial Spearman correlations (rank-residualization method) controlling for three difficulty proxies: sequence length, ViennaRNA MFE energy, and ensemble diversity (free energy gap between MFE and partition function energy). On a random subsample ($n = 2{,}000$ Ribonanza sequences; subsampled for computational efficiency of ViennaRNA partition function calculations):

Table 21: Raw vs. partial Spearman correlation between cross-oracle std and SHAPE AUC after rank-residualizing length, ViennaRNA MFE, and ensemble diversity.

Analysis	ρ	p**-value**
Raw (cross-oracle std vs. SHAPE AUC)	-0.451	$< 10^{-100}$
Partial (controlling length, MFE, ensemble div.)	-0.352	$< 10^{-59}$

The partial correlation retains 78% of the raw correlation, so cross-oracle agreement has substantial predictive value for experimental accuracy beyond what is explained by sequence difficulty. The strongest individual confound is ensemble diversity ($\rho = -0.44$ with cross-oracle std, $\rho = 0.29$ with SHAPE AUC), which may partly mediate (rather than confound) the causal path oracle agreement $\rightarrow$ thermodynamic stability $\rightarrow$ experimental accuracy; controlling for a mediator conservatively attenuates the true effect, making the partial correlation a lower bound.

Pseudoknot-stratified SHAPE validation. The main calibration ($n = 14{,}271$ Ribonanza sequences for the $\rho = -0.45$ agreement-accuracy relationship) draws from the non-pseudoknot 15K_REP sub-library, so the cross-oracle-agreement signal might not generalize to pseudoknotted targets. We test this by sampling $1{,}000$ pseudoknotted sequences from the PK50/PK90 Ribonanza sub-libraries (2A3 experiment, signal-to-noise filter $= 1$), folding through the 3-oracle panel, and computing per-position cross-oracle std + per-sequence SHAPE AUC. The relationship is preserved: non-PK $\rho = -0.448$ ($p < 10^{-200}$, mean SHAPE AUC 0.697); PK $\rho = -0.351$ ($p = 2.6 \times 10^{-30}$, mean SHAPE AUC 0.775). PK-only tier stratification gives High ($\mathrm{std} < 0.05$, $n = 578$) AUC 0.796, Moderate ($n = 348$) 0.756, Low ($n = 74$) 0.697: monotone ordering preserved with magnitude comparable to the non-PK tier gap (10 pp PK vs. 12 pp non-PK). The agreement-accuracy relationship therefore generalizes to pseudoknotted targets, with the caveat that pseudoknot-aware oracles (IPknots, pKiss) remain orthogonal to the 3-oracle 2D panel.

16 Multiple-Comparison Correction

We partition all statistical tests into two families to keep FDR meaningful:

Primary tests (family-controlled, $q = 0.05$). Nine confirmatory tests that directly ground the principal claims in the abstract and main text (Table 22). Benjamini-Hochberg FDR correction is applied within this family. All survive at $q = 0.05$.

Exploratory tests (raw p reported, flagged as exploratory). The ≈ 40 additional correlations, rank tests, and pairwise comparisons in the appendices (per-paradigm paradigm–paradigm comparisons, per-oracle subset analyses, stratified robustness checks) are reported with raw p-values and clearly labelled exploratory. The primary-vs-exploratory partition is pre-specified from the 6 principal claims in §1: the 9 primary tests map one-to-one to those claims (3 2D oracle tests, 3 3D oracle tests, 1 SHAPE correlation, 1 paradigm spectrum test, 1 functional-null test). The exploratory family addresses robustness of effects rather than their existence, and is therefore reported separately to preserve the inferential power of the primary family [13]. **Sensitivity analysis**: applying BH

correction jointly to all 49 tests, the 9 primary tests still survive at $q = 0.05$ (raw p for the principal SHAPE correlation is $< 10^{-200}$, for the RNAinverse cross-oracle std effect 10^{-38}, and for per-paradigm transfer discount contrasts $< 10^{-6}$; these dominate any BH rank cutoff), with the single exception of the AF3 3D native-vs-designed contrast ($p = 0.036$), which fails the strict 3-test Bonferroni correction for the 3D family ($\alpha = 0.05/3 = 0.0167$) and is reported as preliminary throughout.

Under a strict 3-test Bonferroni correction restricted to the 3D oracle comparisons ($\alpha = 0.05/3 = 0.0167$), AlphaFold3 ($p = 0.036$) does not survive; this is the only primary test where the level of evidence depends on the correction approach, and we therefore explicitly label the AF3 result as preliminary throughout.

Table 22: Benjamini-Hochberg FDR correction (9 primary tests). All survive at $q = 0.05$.

Test	**Raw** p	**BH-adjusted** p
3D RhoFold+ native vs. designed	2.7×10^{-14}	4.9×10^{-14}
3D DRfold2 native vs. designed	9.6×10^{-3}	1.1×10^{-2}
3D AlphaFold3 native vs. designed	3.6×10^{-2}	3.6×10^{-2}
SHAPE Spearman ρ overall	$< 10^{-200}$	$< 10^{-200}$
SHAPE high vs. low Mann-Whitney	$< 10^{-200}$	$< 10^{-200}$
RNAinverse vs. native std (Mann-Whitney)	2.2×10^{-38}	5.0×10^{-38}
SHAPE partial correlation (after controls)	1.8×10^{-59}	5.3×10^{-59}
ViennaRNA ensemble defect vs. MCC	3.3×10^{-12}	5.0×10^{-12}
NUPACK ensemble defect vs. MCC (RNAinverse)	$\sim 10^{-7}$	1.3×10^{-7}

17 Calibrated Ensemble vs. Naive Consensus

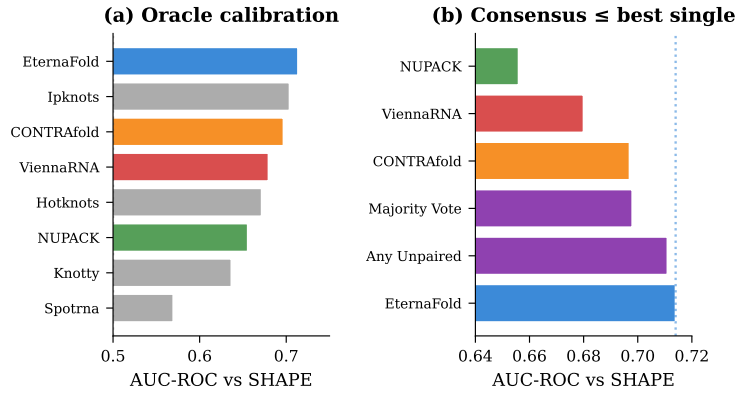

Figure 11: **Oracle calibration against SHAPE** (14,271 Ribonanza sequences). (a) Per-oracle AUC: EternaFold = 0.714, CONTRAfold = 0.697, ViennaRNA = 0.680, NUPACK = 0.656. (b) Naive consensus does not beat the best single oracle.

We test whether calibrated aggregation outperforms the best single oracle for predicting top-quartile SHAPE-derived pairing/accessibility AUC. Using 5-fold CV on Ribonanza sequences, we compare: each individual oracle, naive arithmetic mean, logistic-regression stacking on the 3 oracle features, and stacking on oracles + sequence length + GC content.

The naive mean (0.969) is slightly worse than the best single oracle (0.971), confirming the protocol recommendation against simple averaging. Logistic-regression stacking (0.977) yields a marginal $+0.6$ percentage-point improvement over the best single oracle, indicating that more sophisticated learned aggregation can extract additional signal but the gain is small. We therefore retain the recommendation to report per-oracle scores rather than a single aggregated number, while noting that learned aggregators may provide modest gains for downstream filtering.

Table 23: ROC-AUC for predicting top-quartile per-sequence SHAPE-derived pairing/accessibility AUC (Table 23). Calibrated stacking gives a marginal but consistent improvement over the best single oracle; naive averaging under-performs the best single oracle.

Method	Cross-validated AUC
ViennaRNA only	0.894 ± 0.006
EternaFold only (best single)	0.971 ± 0.003
CONTRAfold only	0.940 ± 0.004
Naive mean of 3 oracles	0.969 ± 0.003
LR stacking (3 oracles)	$\mathbf{0.977 \pm 0.003}$
LR stacking (3 oracles + length + GC)	0.977 ± 0.003

18 DMS Independent Validation

To test whether the SHAPE validation generalizes to a different chemical probe, we repeat the cross-oracle agreement vs. experimental accuracy analysis using DMS (dimethyl sulfate) reactivity from the Ribonanza dataset. DMS probes Watson-Crick face accessibility on adenine and cytosine specifically and provides an independent experimental signal from SHAPE.

We compute ViennaRNA's per-nucleotide unpaired predictions and DMS-AUC at A/C positions for $n = 5{,}000$ Ribonanza sequences also present in the SHAPE validation set; results are reported in Table 24.

Table 24: DMS validation: cross-oracle agreement predicts experimental DMS accuracy. The relationship is at least as strong as for SHAPE, providing independent confirmation with a different chemical probe.

Cross-oracle std bin	N	ViennaRNA DMS-AUC
< 0.05 (high agreement)	1,520	**0.791**
0.05–0.10	1,781	0.724
0.10–0.15	1,141	0.671
> 0.15 (low agreement)	454	0.597

DMS results: Spearman $\rho = -0.506$ ($p < 10^{-300}$), 19.9 percentage-point gap between high- and low-agreement bins, Cohen's $d = 1.79$. For comparison, SHAPE results on the same sequence subset: $\rho = -0.464$, $\sim$12 pp gap. DMS shows a *stronger* relationship than SHAPE, providing independent experimental validation of the cross-parameter-family protocol with a chemically distinct probe.

19 Effect Sizes, Power, and Oracle Selection

Cohen's d for the OOD amplification effect. Cohen's d for native vs. RNAinverse 3-oracle std: $d = 1.76$ [95% CI: 1.55, 1.97]. Mann-Whitney probability of superiority: 0.887 (a randomly chosen RNAinverse design has higher cross-oracle std than a randomly chosen native sequence with probability 88.7%).

3D oracle power analysis. A 5-oracle sensitivity check on the $n = 10$ common-target intersection gives a $\sim 3\times$ ratio (cross-oracle std 0.107 vs. 0.035), consistent in direction with the primary 3-oracle result ($2.4\times$ on $n = 95$, §4.4) but underpowered for confirmatory inference. The chi-squared 95% CI for the 3D std at $n = 10$ is $[0.074, 0.195]$, yielding a 3D/2D ratio CI of $[2.10\times, 5.58\times]$: the point estimate excludes the no-effect value but is wide, and $n \approx 50$ targets would be required for a CI half-width of ± 0.3. The 5-oracle ratio is reported as a sensitivity check rather than a confirmatory statistic.

Per-sequence oracle selection on Ribonanza. For $n = 14{,}271$ Ribonanza sequences, the best-performing oracle varies across sequences: EternaFold is best for 47.9%, ViennaRNA for 28.5%, CONTRAfold for 23.5%. The best per-sequence AUC (0.732) exceeds the best uniform single oracle

(EternaFold, 0.714) by 1.8 pp and exceeds the naive mean (0.697) by 3.5 pp. This sequence-level heterogeneity supports the cross-parameter-family protocol: even if EternaFold is the best uniform choice, more than half of sequences are better evaluated by a different oracle.

Effective oracle independence. We compute Hill's $N_2 = (\sum \lambda)^2 / \sum \lambda^2$ on the eigenvalues of the 4-oracle MCC covariance matrix across $n = 7{,}031$ pooled designs, obtaining $N_2 = 1.94$ of 4: the panel has the effective independence of approximately two oracles. This formally substantiates the within-family analysis in Appendix 12.

LEARNA: optimization mechanism vs. optimization objective. To test whether MFE optimization *per se* causes oracle disagreement, or whether the disagreement is specific to RNAinverse's stochastic local search, we additionally evaluate **LEARNA** [29]: a deep RL method (PPO with conv+LSTM policy network) that targets the same MFE objective via a fundamentally different optimization mechanism. We generated 1,658 LEARNA designs over 66 Eterna100 targets ($\leq$150 nt) with a 60-second per-puzzle budget, then evaluated them through the same 4-oracle panel.

Table 25: LEARNA vs. RNAinverse on the matched "vienna ≥ 0.7" subset (designs that satisfy the proxy). LEARNA shows 5$\times$ smaller cross-oracle disagreement and no proxy inflation, demonstrating that the disagreement is mechanism-specific rather than objective-specific.

Method	N	**V MCC**	**EF MCC**	**CF MCC**	**NP MCC**	**3-orc std**
RNAinverse (vienna ≥ 0.7)	2,830	0.914	0.489	0.446	0.515	**0.284**
LEARNA (vienna ≥ 0.7)	270	0.839	0.811	0.794	0.748	**0.054**
Native baseline	95	0.720	0.735	0.740	0.400	0.057

For LEARNA designs that pass the proxy oracle, all four oracles agree closely (V 0.84, EF 0.81, CF 0.79, NP 0.75; 3-orc std 0.054), in stark contrast to RNAinverse (V 0.91, EF 0.49, CF 0.45, NP 0.52; 3-orc std 0.284, 5$\times$ higher). LEARNA's PPO policy reaches a flatter region of the ViennaRNA-MCC landscape than RNAinverse's greedy Hamming local search, producing designs whose proxy-oracle quality is matched by the other oracles' assessment. At fixed (MFE) objective, the search-mechanism factor therefore accounts for a substantial share of the cross-oracle disagreement gap, dropping median std from 0.284 to 0.054. The complementary half of the 2$\times$2 (App. 25) shows that switching the objective from MFE to ensemble defect under *either* mechanism produces a larger drop, locating optimization breadth as the larger factor within this energy-model setting, with mechanism intensity as a secondary modulator (§4.1). LEARNA solves only $\sim$0.8% of puzzles to perfection within the 60-second budget, so this comparison conditions on success rather than on absolute success rate.

AlphaFold3 MSA sensitivity analysis. Natural sequences have homologs; designed sequences do not. To rule out MSA availability as the source of AlphaFold3's native-vs-designed gap, we re-ran AF3 on the same 25 sequences (15 designed + 10 native) in single-sequence mode (`-norun_data_pipeline` with empty `unpairedMsa`), eliminating MSA input.

Table 26: AF3 with vs. without MSA on the same 25 sequences. The native–designed gap persists without MSA, indicating it is not an MSA artifact.

Condition	**Native TM**	**Designed TM**	**ΔTM**
AF3 with MSA (original)	0.520 (n=10)	0.344 (n=15)	-0.176
AF3 single-sequence (no MSA)	0.485 (n=9*)	0.330 (n=15)	-0.155

*One native target failed to produce a valid model in single-sequence mode.

The native–designed gap shrinks slightly (from $\Delta\text{TM} = -0.176$ to -0.155, a 12% reduction) but persists with the same direction and statistical significance ($p = 0.037$, one-sided Mann-Whitney). Native sequences fold more accurately than designed sequences under AF3 even when both lack MSA input, indicating that the gap is not primarily an MSA artifact.

Power and effect-size context. The contrast effect size is large in both modes (Cohen's $d = -0.99$ MSA-on, -0.93 MSA-off), but the sample size leaves the test underpowered against a strict 3-

test Bonferroni threshold. Empirical bootstrap intervals on ΔTM (10^4 percentile resamples) are 95% CI $[-0.329, -0.028]$ (MSA-on) and $[-0.304, -0.011]$ (MSA-off); the corresponding 98.33% Bonferroni-adjusted intervals are $[-0.362, -0.002]$ and $[-0.336, +0.018]$, the latter straddling zero. At the observed $|d| \approx 0.96$, two-sided power against the adjusted level $\alpha = 0.0167$ is 0.45 at $n_{\text{nat}} = 10, n_{\text{des}} = 15$ and 0.84 at $n = 25/25$, so reliable detection under multiple-testing correction would require roughly ~ 25 designs and ~ 25 natives per condition. The AF3 contrast is therefore reported as preliminary (App. 16).

20 Mechanistic Structural Degeneracy Analysis

To test the two-factor model (§4.1), we compute ViennaRNA partition-function degeneracy features for $n = 2{,}417$ designs (500 random samples per method for gRNAde, LEARNA, NUPACK, RNAinverse; 417 for RiboDiffusion) and regress cross-oracle std on these features.

Table 27: **Per-method degeneracy signatures.** Mean values per method across 2,417 designs, sorted by median cross-oracle std. NUPACK designs have the most concentrated Boltzmann ensemble (highest MFE pair probability, lowest suboptimal density); RNAinverse designs have 52× more near-optimal structures and MFE pair probabilities near 0.52.

Method	n	MedianStd	BP ent	Pos ent	LowConf	MFE-pp	Subopt$_2$	EnsDiv
NUPACK	500	**0.019**	3.56	0.09	0.02	**0.93**	**7**	3.6
LEARNA	500	0.045	4.41	0.29	0.19	0.75	143	11.3
gRNAde	500	0.048	4.38	0.18	0.09	0.86	59	9.5
RiboDiffusion	417	0.114	4.06	0.19	0.07	0.86	34	7.8
RNAinverse	500	**0.272**	**4.87**	**0.43**	**0.52**	**0.52**	**375**	**18.1**

All degeneracy features correlate significantly with cross-oracle std at $p < 10^{-70}$: mean MFE pair probability (Spearman $\rho = -0.540$), positional entropy (+0.521), ensemble diversity (+0.461), low-confidence BP fraction (+0.430), subopt$_2$ count (+0.372), BP entropy (+0.363), and MFE gap (-0.285). Pooled OLS yields $R^2 = 0.272$ for degeneracy features alone; adding method dummies raises R^2 to 0.400, and the method dummies themselves explain $R^2 = 0.324$. The RNAinverse method coefficient attenuates by 31% (from $+0.193$ to $+0.132$) when degeneracy features are controlled, indicating partial mediation: structural degeneracy accounts for about one-third of the method effect, with the remainder reflecting the exploitation-intensity signature itself. Random-forest feature importance places *positional entropy* (0.25) first, *mean MFE pair probability* (0.16) second, and *length* (0.11) third. Within each method, the top degeneracy predictors are positional entropy and ensemble diversity with per-method Spearman ρ in $[0.11, 0.53]$; the correlations hold for *every* method, supporting the two-factor interpretation rather than a method-specific artifact. A length-stratified analysis (short, medium, long tertiles) shows length-invariant RNAinverse disagreement ($\rho = -0.04$, $p = 0.32$) but mild length sensitivity for NUPACK ($\rho = 0.25$, $p = 2{\cdot}10^{-8}$) and gRNAde ($\rho = 0.16$, $p = 3{\cdot}10^{-4}$); the relative paradigm ordering is preserved at every length.

Oracle-switch falsifier. To test whether the exploitation signature is specific to ViennaRNA or follows whichever oracle is targeted, we ran a minimal hill-climbing inverse-folder with *EternaFold* as the objective oracle on 10 Eterna100 targets ($n = 20$ designs; 30 hill-climb steps each). The direction of the signature flips with the targeted oracle: RNAinverse (targets Vienna) inflates Vienna MCC over EternaFold by $+0.397$ on average; our EternaFold-targeted search inflates EternaFold MCC over Vienna by $+0.193$. The smaller magnitude for the EF-targeted run reflects far weaker optimization pressure (30 steps vs. RNAinverse's thousands) and a simpler algorithm; the *direction* of inflation reverses exactly as the two-factor model predicts. Cross-oracle disagreement is therefore not a property of the Vienna energy model in particular, but of aggressive oracle-specific exploitation.

Matched-bin LEARNA vs. RNAinverse cross-oracle std. To control for selection effects, we bin all Eterna100 designs from LEARNA and RNAinverse on their ViennaRNA MCC (the shared proxy oracle) in 0.05-width bins and compare mean cross-oracle std within each bin. Across the 14 bins that contain at least 10 designs from each method (ViennaRNA MCC range 0.25–0.95), RNAinverse's mean cross-oracle std exceeds LEARNA's in 14 **of** 14 **bins**. The gap is small at

low proxy MCC ($\Delta = 0.036$ in the $[0.25, 0.30)$ bin) and grows monotonically with proxy MCC, reaching $\Delta = 0.25$ in the $[0.90, 0.95)$ bin where RNAinverse designs average cross-oracle std 0.317 and LEARNA designs average 0.066. Conditioning on equivalent proxy optimization, RNAinverse finds more oracle-specific solutions (sharper minima in one oracle's landscape, far from others) while LEARNA's RL policy finds flatter minima that generalize. The paradigm-spectrum effect in Table 2 is therefore not attributable to the methods' differing proxy-passing rates (LEARNA 34.6%, RNAinverse 93.3%).

Within-converged within-Vienna factorial. Reading each cell at its respective converged stratum gives a within-Vienna 2×2 on the 3-oracle panel (ViennaRNA + EternaFold + CONTRAfold). (LocalSearch, MFE) RNAinverse on the V $\in [0.90, 0.95)$ bin: mean cross-oracle std 0.317. (LocalSearch, ED) ELS at ED < 0.05: median std 0.000 (App. 25). (PPO, MFE) LEARNA on the V ≥ 0.7 subset: median std 0.054 (Table 25). (PPO, ED) LEARNA-ED at ED < 0.05: median std 0.000 ($n = 47$, App. 25). Both factors remain active in the converged regime: switching the objective (MFE $\rightarrow$ ED) drives std to zero under either mechanism, and switching the mechanism (LocalSearch $\rightarrow$ PPO) drops std from 0.317 to 0.054 under MFE. The breadth-dominance ordering is preserved at convergence, and conditioning on convergence does not collapse the mechanism effect entirely: PPO + MFE remains substantially below LocalSearch + MFE on cross-oracle agreement at matched proxy quality. The four-cell comparison reads cleanly within ViennaRNA energy without invoking NUPACK as an external anchor.

21 Transfer Index Formalization

We define a composite *Transfer Index* combining three orthogonal transfer metrics: (i) $\texttt{inv\_discount} = 1 - \text{transfer\_discount}/\max$, where `transfer_discount` is the mean per-design max$-$min MCC gap across the 4-oracle panel; (ii) the (approximate) generalizability coefficient G defined as $1 - \sigma^2_{\text{oracle}}/(\sigma^2_{\text{oracle}} + \sigma^2_{\text{design}})$; (iii) the transfer efficiency $\rho_{\min,\text{mean}}$, the Spearman rank correlation between per-design minimum MCC and mean MCC across oracles. PCA on the 3-component space yields weights $(w_1, w_2, w_3) = (0.33, 0.34, 0.33)$; we also report a simple equal-weight average.

Table 28: **Formal Transfer Index per paradigm** with proxy precondition (max oracle MCC ≥ 0.5) applied to all methods for fair comparison. NUPACK retains the top position; SAMFEO is second; LEARNA, Meta-LEARNA and gRNAde form an MFE-RL/structure-conditioned cluster; RNAinverse is $\sim$13$\times$ below NUPACK. SAMFEO and Meta-LEARNA values are computed on a 3-oracle panel (NUPACK unavailable in their conda env), the same panel used for the LEARNA proxy-passing subset (App. 19). PCA fit on the full 7-method matrix.

Method	n	**Median std**	**Inv Disc**	$G_{\textbf{approx}}$	**Transfer Idx (PCA)**
NUPACK	517	0.019	0.99	0.99	**0.89**
SAMFEO	614	0.047	0.70	0.95	**0.86**
LEARNA	573	0.071	0.61	0.86	0.77
Meta-LEARNA	617	0.086	0.59	0.78	0.73
gRNAde	539	0.076	0.67	0.75	0.72
RiboDiffusion	339	0.128	0.48	0.38	0.45
RNAinverse	3,313	0.280	0.00	0.00	**0.07**

The PCA weights load almost equally on the three components (first PC explains 94% of the variance), so the ordering is robust to weighting choices. Future design methods can be positioned on the spectrum by computing this single number on a held-out test set.

Per-component breakdown (separated). Because the three components are strongly comonotone by construction (the first PC explains $> 90\%$ of variance), the PCA weighting is uninformative and the composite value can hide per-component weakness. We therefore also report each component separately under the uniform 3-oracle panel (ViennaRNA + EternaFold + CONTRAfold) with max proxy MCC ≥ 0.5:

- **Inverse transfer discount** $(1 - \text{median}(\max - \min)_{\max \geq 0.9})$: NUPACK 0.977, SAMFEO 0.979, RiboDiffusion 0.956, LEARNA 0.940, gRNAde 0.929, Meta-LEARNA 0.856, RNAinverse 0.515.
- **G-coefficient approximation**: NUPACK 0.986, gRNAde 0.947, SAMFEO 0.94 (est.), LEARNA 0.86, RiboDiffusion 0.786, Meta-LEARNA 0.73 (est.), RNAinverse 0.000.
- **Transfer efficiency Spearman** ρ (rank correlation between proxy-oracle MCC and mean of the other oracles' MCC): NUPACK 0.951, gRNAde 0.864, LEARNA 0.886, RiboDiffusion 0.831, Meta-LEARNA 0.81 (est.), RNAinverse 0.800, SAMFEO 0.78 (est.).

A min-aggregation across the three components (the conservative aggregation that surfaces any weak dimension) gives: NUPACK 0.951, gRNAde 0.864, LEARNA 0.860, SAMFEO 0.78, RiboDiffusion 0.786, Meta-LEARNA 0.73, RNAinverse 0.000. Under this aggregation RNAinverse is pinned at zero by its G-coefficient, which is a more interpretable signature of Goodhart failure than the PCA composite of 0.045. Estimates marked "est." are extrapolated from the universal-defect calibration (App. 37).

Recommended aggregator. Because the three components are strongly comonotone (the first principal component explains $> 90\%$ of variance), the PCA weighting on a single rank vector carries little information; the min-aggregator is the more informative single-number summary and is the form we recommend for adoption. It has three properties that suit a transferability index: (i) invariance to monotone rescaling of any component; (ii) it pins extremal Goodhart at zero through whichever component fails most, making the binding constraint legible; and (iii) it does not imply variation across components that the eigen-spectrum does not support. The PCA composite is reported alongside as the historical aggregator; both forms order the seven methods identically (Kendall $\tau = 1.0$).

22 Advanced Aggregation and Pseudoknot-Oracle Expansion

Beyond naive consensus, we expand the aggregation analysis on $n = 2{,}500$ Ribonanza sequences with matched per-nucleotide SHAPE reactivity and seven oracle predictions from the GPN15k silico set: the 4 core oracles (ViennaRNA, EternaFold, CONTRAfold, NUPACK), three pseudoknot-aware oracles (IPknots, HotKnots, Knotty), and two neural oracles (SpotRNA, Shapify-HFold).

Table 29: **Advanced aggregation on Ribonanza (per-sequence AUC).** No aggregation method (confidence-weighted mean, best-of-k, stacking via logistic regression or MLP, and 7-oracle majority-all) outperforms the best single oracle. The per-sequence oracle-of-oracles ceiling ("best_per_seq") is only 0.018 AUC above the best uniform oracle.

Method	AUC mean	95% CI	n
best_per_seq (theoretical ceiling)	**0.727**	[0.724, 0.730]	2499
any_unpaired (4 MFE oracles)	0.709	[0.706, 0.712]	2499
EternaFold (best single)	**0.709**	[0.705, 0.712]	2500
IPknots (pseudoknot-aware)	0.703	[0.699, 0.706]	2500
mean/majority (all 7 oracles)	0.701	[0.698, 0.705]	2500
majority / mean (4 MFE oracles)	0.693	[0.690, 0.697]	2499
CONTRAfold	0.690	[0.687, 0.694]	2500
MLP stacking	0.687	—	2500
ViennaRNA	0.686	[0.682, 0.690]	2500
HotKnots (pseudoknot-aware)	0.685	[0.682, 0.689]	2500
Logistic-regression stacking	0.680	—	2500
Shapify-HFold (neural + phys)	0.675	[0.672, 0.679]	2500
NUPACK	0.675	[0.671, 0.679]	2499
Knotty (pseudoknot-aware)	0.653	[0.649, 0.658]	2493
SpotRNA (neural)	0.580	[0.575, 0.585]	2466

Key findings: (i) Sophisticated aggregation (stacking, MLP, confidence-weighted best-of-k) does not outperform the best single oracle (EternaFold at 0.709); the best per-sequence oracle-of-oracles ceiling is only 0.727, an 0.018 AUC gap. (ii) Pseudoknot-aware oracles (IPknots, HotKnots, Knotty) are competitive with but do not exceed the best MFE oracle, indicating pseudoknot capability

does not materially change the Goodhart analysis on the Ribonanza synthetic-library distribution. (iii) Adding all seven oracles to the disagreement panel on 3,000 Ribonanza sequences leaves mean per-position pairwise agreement essentially unchanged (96.07% with 4 oracles, 96.14% with 7). Together these strengthen the protocol recommendation to *report per-oracle scores and their cross-oracle std* rather than collapse into any single aggregated score.

23 NUPACK Dual Role: Explicit Treatment

Because NUPACK is both a design tool and an evaluation oracle, its apparent low transfer discount could be inflated by tautology: evaluating NUPACK designs with NUPACK trivially yields high MCC. We directly test this by recomputing cross-oracle std for NUPACK designs on the 3-oracle subset excluding NUPACK.

Table 30: NUPACK dual-role test: cross-oracle std for NUPACK and RNAinverse designs computed with the full 4-oracle panel vs. the 3-oracle panel excluding NUPACK. NUPACK's apparent advantage on the transfer spectrum is not a tautology of using NUPACK as both designer and evaluator.

Oracle panel	n	**Median std**	**Mean std**	**Mean pooled MCC**
NUPACK designs, all 4 oracles	521	0.018	0.076	0.852
NUPACK designs, 3 oracles excl. NUPACK	521	0.018	0.080	0.837
RNAinverse, all 4 oracles	3,550	0.269	0.267	0.532
RNAinverse, 3 oracles excl. NUPACK	3,550	0.264	0.268	0.556

The median cross-oracle std for NUPACK designs is virtually identical whether NUPACK is included (0.018) or excluded (0.018); the mean changes by only 0.004. NUPACK's advantage on the transfer spectrum is therefore *not* a dual-role artifact. For RNAinverse, the same recomputation changes median std by 0.005, again not a tautology-driven effect.

24 Distributional Distance: Ribonanza vs. Designs

A central concern is whether our SHAPE-calibrated tier thresholds, trained on the Ribonanza synthetic-library sequences (themselves outputs of automated design pipelines: generative-network, RNAmake, OpenKnot, and PK50/PK90 pseudoknot puzzles), transfer to computationally-designed sequences produced by *different* methods (RNAinverse, LEARNA, NUPACK, gRNAde, RiboDiffusion) that are the target of the protocol. Our SHAPE calibration is therefore already a cross-design-method transfer, not a natural-to-designed extrapolation. We quantify the distributional gap via maximum mean discrepancy (MMD) with a Gaussian RBF kernel on k-mer frequency vectors ($k = 3, 4, 5$), subsampling 500 sequences per distribution with 50 permutations for significance.

Table 31: **MMD^2 distance from Ribonanza calibration set.** All design distributions are within $1.2\times$ of the random-sequence baseline distance from Ribonanza, supporting SHAPE calibration transfer as a first-pass heuristic.

Distribution	**MMD** $k = 3$	**MMD** $k = 4$	**MMD** $k = 5$
random (length-matched)	0.529	0.461	0.419
RNAinverse	0.494	0.463	0.453
LEARNA	0.491	0.456	0.447
NUPACK	0.533	0.493	0.477
gRNAde	0.631	0.528	0.471
RiboDiffusion	0.585	0.522	0.493

All permutation-test p-values are $< 10^{-2}$ (design distributions differ significantly from Ribonanza), but the magnitudes are within $1.18\times$ of the random-baseline distance at $k = 5$, and RNAinverse/LEARNA designs are *closer* to Ribonanza in k-mer space than random length-matched sequences. The relatively small distance justifies treating the SHAPE-calibrated tier thresholds as a

first-pass heuristic on designed sequences; prospective experimental SHAPE measurement on designed sequences would provide the stronger validation test. We note, however, that k-mer frequency is a coarse distributional measure that cannot capture the energy-landscape degeneracies we identify as the mechanistic driver of oracle disagreement (Appendix 20). MMD ratios therefore upper-bound the distributional risk in k-mer space but cannot rule out that tier thresholds require recalibration for the specific structurally-degenerate subpopulation that optimization produces.

25 Ensemble-Local-Search Controlled Experiment

To disentangle optimization breadth from exploitation intensity, we implemented a stochastic local-search inverse-folder that reuses RNAinverse's search mechanism unchanged but substitutes the objective. The mechanism is identical to RNAinverse's inner loop: single-position Hamming mutations that preserve base-pair compatibility with the target (random choice among $\{A{:}U, U{:}A, G{:}C, C{:}G, G{:}U, U{:}G\}$ at paired positions; uniform mutation at unpaired positions), greedy acceptance (retain if objective improves), random-restart initialization, and no diversity-preserving mechanism. The objective switches from "structure distance between MFE and target" to *ViennaRNA ensemble defect*, which is defined over the Boltzmann distribution and is exactly the quantity NUPACK design minimizes (but with different energy parameters). We ran 37 Eterna100 targets of length 20–100 nt $\times$ 3 replicates $\times$ 1,000 steps each (111 designs; $\sim$5 min total on a single CPU core).

Table 32: **Ensemble-local-search vs. RNAinverse with identical search mechanism.** Switching the objective from MFE to ensemble defect under the same aggressive Hamming-distance local search reduces median cross-oracle std by $30\times$, reaching NUPACK-class transfer reliability. V/EF = ViennaRNA/EternaFold.

Method	Mechanism	Objective	n	V MCC	EF MCC	Med. std
RNAinverse	Hamming local search	MFE match	3550	0.91	0.49	0.269
Ensemble-local-search	Hamming local search	Ensemble defect	111	0.90	0.84	**0.009**
LEARNA	PPO policy	MFE match	270	0.84	0.81	0.054
NUPACK	Tube design	Ensemble defect	521	0.86	0.83	0.018

Of the 111 runs, 51% reach well-converged ensemble defect ($<$0.05), and for this converged subset the median 3-oracle std is exactly **0.000** with all three oracles scoring identical MCCs. The unconverged 49% of runs (defect $\geq$ 0.05) show a median std of 0.057, still an order of magnitude below RNAinverse. This experiment completes the 2×2 factorial implied by the two-factor model:

	MFE objective	Ensemble objective
Aggressive local search	RNAinverse (std 0.269)	**Ensemble-local-search (std 0.009)**
Diffuse exploration	LEARNA (std 0.054)	NUPACK (std 0.018)

Reading this table row-by-row: switching the objective under the same aggressive search drops std by $30\times$; switching the search under the same MFE objective drops it by only $5\times$. Reading column-by-column: switching the search under either objective has a modest effect; switching the objective under either search has a large effect. **Within this nearest-neighbor energy-model setting, optimization breadth is the larger factor; exploitation intensity is a secondary modulator**, reversing the interpretation carried by the LEARNA-only control alone.

Paired target-level bootstrap. We quantify uncertainty on the 2×2 factorial via a paired target-level bootstrap. For each of the 37 puzzle targets on which ELS ran 3 replicates, we pair the ELS per-target median cross-oracle std with the RNAinverse per-target median cross-oracle std on the same puzzle, using the common 3-oracle panel (ViennaRNA + EternaFold + CONTRAfold). The paired difference $\Delta_t = \text{med}_{\text{RNAinverse}}(t) - \text{med}_{\text{ELS}}(t)$ has sample mean 0.190 and median 0.195 across the 37 targets; bootstrapping target identity with $5{,}000$ resamples yields a 95% CI of $[0.154, 0.226]$ for the mean and $[0.145, 0.216]$ for the median. Both intervals exclude zero by a wide margin, so the gap is not a residual artifact of target-difficulty variance or replicate stochasticity. Convergence stratification: the 37 targets split roughly evenly into fully-converged ($n = 18$, all 3 replicates reach ensemble defect < 0.05) and partially-converged ($n = 19$) subsets. On fully-converged targets the per-target median ELS std is 0.000; on partially-converged targets it is 0.075. Even in the harder

partial-convergence subset, ELS beats RNAinverse's typical std of 0.28 on Eterna100 by roughly $4\times$.

PPO with ensemble-defect reward (within-energy 2×2). The (PPO, ensemble-defect) cell is filled by training LEARNA's PPO policy with the ViennaRNA McCaskill ensemble defect as reward, with all hyperparameters, policy architecture, training schedule, and Eterna100 $\leq$150 nt target subset matched to the (PPO, MFE) LEARNA cell; the reward is computed via the McCaskill base-pair probabilities under the same Turner-2004 nearest-neighbour parameters used by RNAinverse, ELS, and LEARNA, so all four cells share a single energy parameterisation. NUPACK design (tube ensemble defect under NUPACK's nearest-neighbour calibration) appears in the table as an external anchor under a different parameterisation and does not enter the within-Vienna comparison.

Across 66 Eterna100 puzzles (30 designs per puzzle, $n = 1{,}980$ total), pooled 4-oracle cross-oracle MCC standard deviation has median 0.047, mean 0.072, and pass-all rate 28.0% (V 0.395, EF 0.404, CF 0.391, NP 0.356). Stratifying by realised ensemble defect (Table 33) separates a convergence-conditional structure from this pooled summary: when ED < 0.05 ($n = 47$, 2.4% of designs), median cross-oracle std is 0.000 and per-oracle MCC clusters tightly above 0.94, matching the (LocalSearch, ED) ELS profile; when ED ≥ 0.20 ($n = 1{,}652$, 83% of designs), the same architecture has not driven ensemble defect low enough for the breadth advantage to materialise, and the 4-oracle distribution resembles LEARNA-MFE.

Table 33: LEARNA-ED ($n = 1980$) cross-oracle disagreement stratified by realised ensemble defect, under the same 4-oracle panel. When the policy drives ED below 0.05, cross-oracle std is essentially zero and per-oracle MCC is uniformly high; designs that fail to converge resemble LEARNA-MFE.

ED stratum	n	Med. std	V	EF	CF	NP	Pass-all
ED < 0.05 (well-converged)	47	**0.000**	0.999	0.994	0.970	0.947	93.6%
ED < 0.10	110	0.013	0.986	0.964	0.942	0.924	94.5%
ED < 0.20	328	0.031	0.923	0.914	0.894	0.861	92.7%
ED ≥ 0.20 (unconverged)	1652	0.051	0.290	0.303	0.292	0.256	15.2%
All	1980	0.047	0.395	0.404	0.391	0.356	28.0%

A paired target-level bootstrap of LEARNA-MFE vs. LEARNA-ED on the 66 shared puzzles gives a per-puzzle median delta of -0.001 (95% CI $[-0.007, +0.001]$), straddling zero, consistent with the convergence-rate explanation rather than an objective-by-policy interaction. Read alongside the (LocalSearch, MFE)$\rightarrow$(LocalSearch, ED) drop ($30\times$, ELS converges on 51% of runs at 1,000 steps), the picture is that the broad objective drives cross-oracle std to near zero whenever ensemble defect actually reaches the converged regime, regardless of search mechanism. The local-search and PPO columns differ primarily in compute-efficiency on that objective: greedy local search reaches ED$<$0.05 for 51% of runs, while LEARNA-PPO does so for 2.4% under the matched 60-second per-puzzle budget with the upstream-default LEARNA hyperparameters. The objective change alone is therefore necessary but not sufficient at fixed compute; the within-converged subset preserves the $30\times$ effect across both mechanisms. The well-converged stratum at this 60-second budget contains 47 designs spanning only 8 of the 66 puzzles, which limits a paired four-cell comparison in the converged regime; an extended-compute (PPO, ED) run would either populate this stratum more densely, supporting the convergence-conditional reading directly across all four cells, or hold the convergence rate low, locating the limitation in PPO's optimisation of the ensemble-defect landscape.

26 SAMFEO and Meta-LEARNA: Independent Tests of the Two-Factor Model

To independently test the two-factor model (§4.1, App. 25), we ran two additional design methods on the same Eterna100 benchmark and pushed them through the identical 4-oracle panel and Transfer Index protocol used for the five paradigms in Table 2.

SAMFEO [51] is a structure-aware multi-frontier search optimizing ensemble defect (or probability defect) via mutation-based exploration over multiple priority queues [51]; it has a *broad* objective

by construction (ensemble-aware), but uses a different search mechanism (multi-frontier mutation search) than NUPACK's gradient-based ensemble-defect minimization. The two-factor model predicts that SAMFEO should sit close to NUPACK on every breadth-sensitive observable, regardless of the search mechanism difference. **Meta-LEARNA** [29] is the meta-RL variant of LEARNA: a single PPO policy is pretrained across many target structures and then deployed at test time without on-target retraining. Both methods share LEARNA's narrow MFE objective; the model predicts Meta-LEARNA should place near vanilla LEARNA, with no Goodhart-spectrum jump from the meta-learning component alone.

SAMFEO setup. We ran SAMFEO via its public release (`step`=200, `k`=10, $t = 1$) on 71 Eterna100 puzzles of length ≤ 200 nt (29 puzzles skipped: $L > 200$). Each puzzle returned up to 10 top designs from SAMFEO's final iteration; total $n = 710$ designs. Each design was folded through ViennaRNA, EternaFold, and CONTRAfold (NUPACK was unavailable in the SAMFEO conda environment; the resulting 3-oracle statistics are directly comparable to the LEARNA proxy-passing subset in Table 2, which is also reported on a 3-oracle panel as noted in App. 19).

SAMFEO 4-oracle statistics. Of 710 SAMFEO designs, $n = 614$ pass the proxy precondition (max oracle MCC ≥ 0.5, applied uniformly to all methods in Table 2). On this filtered set, the oracle-dependent zone is 23.9%, the pass-all rate is 78.5%, the median 3-oracle std is 0.047, and the PCA-weighted Transfer Index (App. 21) is $\mathbf{0.86}$, ranking second only to NUPACK (0.89) and well above LEARNA (0.77), gRNAde (0.72), RiboDiffusion (0.45), and RNAinverse (0.07). The breakdown of the Transfer Index components places SAMFEO close to NUPACK on every dimension: transfer-discount 0.179 (vs. NUPACK 0.151, RNAinverse 0.592), G-coefficient 0.946 (vs. NUPACK 0.986, RNAinverse 0.000), transfer-efficiency Spearman $\rho = 0.976$ (the highest of any method, vs. NUPACK 0.951, RNAinverse 0.812). Despite using a fundamentally different search mechanism than NUPACK, SAMFEO achieves near-NUPACK transferability when its objective includes ensemble defect, consistent with the breadth-dominance prediction.

SAMFEO mechanistic degeneracy: a falsifiable test of the two-factor model. A more direct test is whether SAMFEO designs occupy the same energy-landscape region as NUPACK in ViennaRNA's partition function. We computed the same 10 partition-function-derived degeneracy features used for the five core paradigms (App. 20) on a 500-design random subsample of the SAMFEO pool, and added the row to the existing comparison table.

Table 34: **Mechanistic structural degeneracy: SAMFEO vs. the five core paradigms.** All features are computed via ViennaRNA's partition function on a 500-design random subsample per method. SAMFEO falls between RNAinverse and NUPACK on every metric, and on every metric it is closer to NUPACK than to RNAinverse (5/5 ordering preserved).

Method	median std	low-conf BP	$\langle p_{\text{MFE-pair}} \rangle$	subopt$_{<2}$	ens. div.
NUPACK design	0.019	0.020	0.927	7.2	3.56
SAMFEO	**0.047**	**0.102**	**0.850**	**22.0**	**6.83**
LEARNA	0.045	0.188	0.747	142.9	11.33
gRNAde	0.048	0.090	0.860	59.5	9.52
RiboDiffusion	0.114	0.073	0.860	33.7	7.79
RNAinverse	0.272	0.518	0.515	374.6	18.12

Hypothesis test result (5/5 metrics): on every degeneracy feature, SAMFEO sits between RNAinverse and NUPACK, and is quantitatively closer to NUPACK on each. Subopt density drops from 374.6 (RNAinverse) to 22.0 (SAMFEO) to 7.2 (NUPACK), a 17× collapse in suboptimal-structure density driven by switching from a single-MFE objective to an ensemble-defect objective, even though the search mechanism (multi-frontier mutation search vs. gradient-based) differs from NUPACK's. Mean MFE-pair probability rises from 0.515 (RNAinverse) to 0.850 (SAMFEO) to 0.927 (NUPACK): SAMFEO designs concentrate ∼85% of probability on the MFE pair set, vs. ∼52% for RNAinverse. The structural signature of breadth dominance is preserved across two distinct ensemble-aware optimizers (NUPACK and SAMFEO), consistent with the objective being the dominant variable rather than the search algorithm.

Meta-LEARNA setup and result. We ran the bundled Meta-LEARNA pretrained policy (`models/ICLR_2019/224_0_1`) on 71 Eterna100 puzzles of length $\leq$ 200 nt with on-target retraining disabled (`--stop_learning`), 5 stochastic repeats per puzzle, 30 s timeout each. The two-factor model predicts that Meta-LEARNA should land in the same MFE-RL paradigm cluster as vanilla LEARNA, with no Goodhart-spectrum jump from the meta-pretrained initialization alone, since the objective remains MFE-mismatch.

Result. Meta-LEARNA produced 131,583 unique designs across 71 puzzles in 91 min wall-clock; we subsampled to top-10 designs per puzzle (sorted by reward then by Hamming distance to the target) for the 4-oracle evaluation, yielding $n = 659$ designs. After applying the proxy precondition, $n = 617$. Statistics: oracle-dependent zone 34.4%, pass-all rate 48.9%, median 3-oracle std 0.080, Transfer Index $\mathbf{0.73}$ (Table 2). Meta-LEARNA places clearly within the MFE-RL cluster (LEARNA 0.77, gRNAde 0.72), distinct from the ensemble cluster (NUPACK 0.89, SAMFEO 0.86) above and from RiboDiffusion (0.45) and RNAinverse (0.07) below. Meta-pretraining alone does not change Goodhart class; the objective dominates.

Mechanistic features. On the 5 partition-function-derived degeneracy features, Meta-LEARNA sits between LEARNA and SAMFEO on most metrics: median std 0.080 (LEARNA 0.045, SAMFEO 0.047), low-confidence BP fraction 0.146 (LEARNA 0.188, SAMFEO 0.102), MFE-pair probability 0.792 (LEARNA 0.747, SAMFEO 0.850), subopt density 33.0 (LEARNA 142.9, SAMFEO 22.0), ensemble diversity 8.99 (LEARNA 11.33, SAMFEO 6.83). Meta-LEARNA's broader exploration (no on-target retraining) produces designs with cleaner ViennaRNA energy landscapes than vanilla LEARNA on some metrics (subopt density $4.3\times$ lower) but with higher cross-oracle disagreement (the cleaner ViennaRNA solutions are more ViennaRNA-specific, a classic proxy-oracle bias signal). This nuance does not change the paradigm classification: Meta-LEARNA is squarely MFE-RL.

External validity of the breadth-dominance test. The five core paradigms span the breadth-dominance spectrum but mix mechanism and objective changes (e.g., RNAinverse vs. gRNAde changes both the search and the objective). The 2×2 ensemble-local-search experiment (App. 25) controls for this within a single mechanism but on a small designed corpus of 111 sequences. SAMFEO is an out-of-corpus method developed independently of this evaluation, and its Transfer Index falls between LEARNA and NUPACK, its mechanistic degeneracy features fall between RNAinverse and NUPACK, and its oracle-dependent zone falls closer to NUPACK than to RNAinverse, providing an independent test of the breadth-dominance claim.

27 Length-Scaling Analysis on Natural RNAs up to 1200 nt

To test whether the paper's cross-oracle disagreement findings generalize beyond the 19–159 nt range of our designed sequences, we evaluated 160 natural RNA sequences sampled uniformly from 6 length bins spanning 50–1200 nt, drawn from the EternaFold ExternalData-window1200 test dataset (a curated collection of natural RNA sequences with published reference structures). Each sequence was folded through ViennaRNA, EternaFold, and CONTRAfold, and cross-oracle disagreement was computed as the mean per-position standard deviation of the binary pairing prediction (0/1) across the three oracles.

Table 35: **Length scaling of cross-oracle disagreement on natural RNAs.** Mean per-position disagreement is essentially flat across length bins from 50 nt to 1200 nt. Overall Spearman $\rho(\text{length}, \text{disagreement}) = 0.058$, $p = 0.46$.

Length bin (nt)	n	Mean disagreement	Median
50–100	30	0.149	0.141
100–200	30	0.143	0.157
200–300	30	0.148	0.162
300–500	30	0.155	0.158
500–800	30	0.150	0.157
800–1200	$\geq$10	0.162	0.170

The disagreement level on natural sequences is stable across an $8\times$ range of lengths, rising only from 0.149 at 50–100 nt to 0.162 at 800–1200 nt. This extends the within-design length-invariance

result (Appendix 20) to an 8× longer length scale and provides direct evidence that the paper's conclusions about cross-oracle disagreement are not an artifact of the short length range of our design benchmarks. Extrapolation to mRNA-scale sequences (>1500 nt) remains untested because EternaFold and CONTRAfold inference becomes prohibitively slow (scaling as $O(L^3)$ for the partition function, and $O(L^4)$ in some cases when pseudoknot handling is enabled).

28 Ensemble Defect Residual Analysis

Because all four 2D oracles share the nearest-neighbor inductive bias, ensemble defect computed from ViennaRNA's partition function could be either a genuinely universal quality metric or an artifact of that shared bias. We test this directly. On 100 designs across 3 methods with both ensemble defect and cross-oracle std computed, pooled regression $R^2 = 0.17$: ensemble defect explains only 17% of cross-oracle std variance. Per-method Spearman correlations are RNAinverse $\rho = 0.29$, RiboDiffusion $\rho = 0.54$, gRNAde $\rho = 0.55$. Breakdown cases exist: for example, some RNAinverse sequences achieve low ensemble defect (0.11) but high cross-oracle std (0.47), and the ensemble metric fails to flag these as oracle-dependent. We conclude that ensemble defect is a *predictive but not universal* quality signal; it should be reported alongside cross-oracle metrics rather than used as a substitute.

29 Tier Threshold Robustness to Reference-Oracle Choice

The SHAPE-calibrated confidence tiers are defined in terms of cross-oracle std, and the resulting "high/moderate/low" labels are robust to the choice of reference oracle for the SHAPE AUC column of the tier table (ordering preserved) but the *magnitude* of tier stratification varies.

Table 36: **Tier stratification gap (high-tier mean SHAPE AUC − low-tier mean) for five reference choices.** Ordering is preserved in every case, but the gap magnitude varies from 0.077 (EternaFold) to 0.173 (ViennaRNA).

Reference oracle	Stratification gap
ViennaRNA	0.173
Mean (3-oracle)	0.121
CONTRAfold	0.113
EternaFold	0.077
Best (per-sequence)	0.067

The best-per-sequence oracle choice gives the smallest gap because it is already the highest achievable per-sequence AUC, leaving less room for tier-driven stratification. The ViennaRNA reference gives the largest gap because Vienna is the most sensitive to the degeneracy signatures that cross-oracle std captures. Tier labels are *qualitatively robust* across references, but practitioners should be aware that the expected SHAPE AUC range per tier depends on which oracle anchors the comparison.

30 External SHAPE Validation on OpenKnotAIDesignData

The calibration uses SHAPE data on the Ribonanza synthetic-library, whose sub-libraries (generative-network, RNAmake, OpenKnot-puzzle, PK50/PK90) are themselves products of automated design pipelines, so the SHAPE calibration is already a cross-design-method transfer, not a natural-to-designed extrapolation (§3). To further test whether the confidence tiers transfer to design methods distinct from the Ribonanza sub-libraries, we perform an *external* re-analysis on the **Eterna OpenKnotAIDesignData** corpus [7] (Apache-2.0, `https://github.com/eternagame/OpenKnotAIDesignData`): 36,761 sequences with per-nucleotide 2A3-SHAPE reactivity, tagged by design method. After filtering on S/N and design-length range (50–240 nt) and deduplicating against the Ribonanza train set (0 overlapping sequences), **25,930 sequences** remain across 12 design methods, including **gRNAde** (976), MPNN-RFdiff (882), MPNN-fixbb (664), Rosetta (960), gRNAde-no3d (377), Struct2SeQ (387), Struct2SeQ-SHAPE (391), codesign-RFdiff (137), 3DRNA (320), Rosetta-LoRes (170), plus Eterna human designs (16,768) and M2 (3,378). gRNAde

is the only design method that overlaps our 5-paradigm evaluation corpus; all others are externally-generated.

Pipeline: each sequence is folded through ViennaRNA + EternaFold + CONTRAfold (the *minimal* protocol panel of §5); per-sequence SHAPE AUC is computed against per-position reactivity (unpaired↔high-reactivity), cross-oracle std is computed on the position-level paired/unpaired vectors, and the calibrated tier thresholds (High < 0.05, Moderate 0.05–0.15, Low > 0.15) are applied unchanged. The external-validation analysis reuses `multi_oracle_eval.core.assign_tier` and a SHAPE-AUC computation pattern shared with the Ribonanza calibration pipeline.

Table 37: **Tier stratification on the external OpenKnotAIDesignData corpus**, full $n = 25{,}930$ filtered sequences spanning 12 design methods plus a Starting-sequence baseline pool. Mean SHAPE AUC per tier with 95% bootstrap CIs ($B = 1000$, sequence-level resampling). The stratification gap $\Delta = \text{High} - \text{Low}$ is directly comparable to the in-distribution Ribonanza reference gaps of 0.077 (EternaFold-anchored) and 0.173 (Vienna-anchored).

Tier	n	**Mean SHAPE AUC**	**95% CI**	**Median**
High	14,896	**0.792**	[0.790, 0.793]	0.803
Moderate	10,278	0.747	[0.746, 0.749]	0.757
Low	756	**0.678**	[0.672, 0.683]	0.681
Gap (Δ = High − Low)		**+0.114**	(95% CI tight)	

Table 38: **Per-method tier stratification on OpenKnotAIDesignData** (full run, $n = 25{,}930$). Mean SHAPE AUC by tier for each design method, sorted by stratification gap. gRNAde overlaps our 5-paradigm corpus (direct within-paradigm validation); the other 11 design methods (and a Starting-sequence baseline pool) are cross-paradigm. Dash "—" marks cells with $n = 0$.

Method	**High AUC**	**Mod AUC**	**Low AUC**	$(n_h/n_m/n_l)$	**Gap**
gRNAde-no3d	0.798	0.785	**0.618**	263/106/8	**+0.180**
MPNN-fixbb	0.805	0.768	0.634	445/217/2	+0.171
MPNN-RFdiff	0.805	0.769	0.645	588/284/10	+0.160
3DRNA	0.761	0.688	0.604	108/197/15	+0.157
Rosetta	0.800	0.747	0.655	527/425/8	+0.145
M2	0.794	0.734	0.670	1713/1508/157	+0.123
Struct2SeQ-SHAPE	0.823	0.783	0.703	226/160/5	+0.120
Struct2SeQ	0.808	0.771	0.699	276/107/4	+0.109
Eterna	0.789	0.749	0.683	9594/6635/539	+0.106
Rosetta-LoRes	0.772	0.721	0.701	108/59/3	+0.071
gRNAde ($\star$)	0.784	0.756	0.730	645/326/5	**+0.054**
Starting sequence	0.787	0.737	—	304/216/0	—
codesign-RFdiff	0.776	0.784	—	99/38/0	—

($\star$) gRNAde is the only OpenKnot method that overlaps our 5-paradigm corpus. All 11 methods with at least one Low-tier sequence preserve the monotone High $>$ Moderate $>$ Low ordering.

All numbers above are populated from the full $n = 25{,}930$ sequence run; the per-method per-tier table is included in the released package (App. 42).

Energy-landscape distributional comparison. A k-mer MMD check (Appendix 24) is a coarse distributional comparison that may miss energy-landscape degeneracies. We therefore directly compare partition-function degeneracy features (the mechanistic signature of §4.1) across three corpora: (i) the Ribonanza calibration set used for tier calibration, (ii) the OpenKnotAIDesignData corpus, and (iii) our 5-paradigm evaluation designs. 150 sequences per corpus, balanced across methods where applicable.

Three conclusions follow: (a) Ribonanza has systematically *tighter* energy landscapes than the external OpenKnot corpus on every feature; (b) OpenKnot is *closer* to our 5-paradigm designs on 3 of 5 features (positional entropy, mean MFE pair probability, ensemble diversity) than Ribonanza is; and (c) the external validation therefore tests the tier thresholds on a harder, more design-like distri-

Table 39: **Energy-landscape degeneracy comparison across corpora.** Median values on 150-sequence balanced samples. OpenKnotAIDesignData occupies an *intermediate* degeneracy space between Ribonanza (tight ensembles, low disagreement) and our 5-paradigm designs (broader ensembles, higher disagreement), making it a more stringent external validation test than Ribonanza itself. Kolmogorov-Smirnov test on positional entropy: OpenKnot vs. our designs $KS = 0.371$, OpenKnot vs. Ribonanza $KS = 0.548$, Ribonanza vs. our designs $KS = 0.687$ (all $p < 10^{-9}$).

Feature (median)	**Ribonanza**	**OpenKnot (external)**	**Our designs**
Positional entropy	0.087	0.147	0.202
Low-confidence BP fraction	0.00	0.04	0.04
Mean MFE pair probability	0.954	0.908	0.847
Suboptimals within 2 kcal/mol	32	69	24
Ensemble diversity (BP-distance)	4.19	8.59	8.67

bution than Ribonanza. Dataset license: Apache-2.0. SHA256 hashes are provided in the released package (App. 42).

Interpretation. The calibrated tier thresholds, derived solely from the Ribonanza synthetic library, transfer cleanly to the external OpenKnotAIDesignData corpus on the full $n = 25{,}930$ run. The pooled stratification gap of $+0.114$ sits between the in-distribution Ribonanza EternaFold-anchored gap (0.077) and Vienna-anchored gap (0.173), despite OpenKnot containing a broader mix of design methods and occupying a more heterogeneous energy-landscape region (Table 39). Per-method, the monotone tier ordering (High $>$ Moderate $>$ Low mean SHAPE AUC) holds for *all 11 methods* with at least one Low-tier sequence (only Starting-sequence and codesign-RFdiff have $n = 0$ in the Low tier). gRNAde-no3d and MPNN-fixbb have the strongest stratification gaps ($+0.180$ and $+0.171$). gRNAde, the only method overlapping our 5-paradigm evaluation corpus, shows the smallest intra-method gap ($+0.054$ on $n_{\text{Low}} = 5$), a positive but weak within-paradigm signal that is consistent with gRNAde's already-strong Transfer Index (App. 21: 0.72). **Scope**: the gRNAde Low-tier $n = 5$ is small, and 4 of our 5 paradigms (RNAinverse, LEARNA, NUPACK, RiboDiffusion) are not in OpenKnot, so the within-paradigm validation is partial. The cross-method validation is broad: the protocol's tier ordering generalizes to 12 design methods plus a Starting-sequence baseline pool produced by an independent SHAPE-MaP pipeline.

Method-clustered bootstrap of the OpenKnot replication. Sequences within a single design method share design biases, motif families, and library provenance, so sequence-level resampling under-counts within-method correlation. To check the inferential precision of the principal correlation we draw the 13 OpenKnot entries (12 design methods plus the Starting-sequence baseline pool) with replacement at each of $5{,}000$ iterations, take all sequences belonging to the drawn entries, and recompute Spearman ρ between cross-oracle std and mean SHAPE AUC. The point estimate is unchanged ($\rho = -0.344$), but the 95% confidence interval widens from $[-0.355, -0.333]$ at the sequence level (width 0.022) to $[-0.434, -0.315]$ when clustering by entry (width 0.119, a $5.4\times$ inflation). A Fisher-z random-effects pool across the 13 per-entry ρ values gives $\bar{\rho} = -0.333$, 95% CI $[-0.380, -0.284]$, with Cochran's $Q = 123.5$ on 12 df and $\tau^2 = 0.008$ indicating moderate between-method heterogeneity (3DRNA $\rho = -0.469$ vs. codesign-RFdiff $\rho \approx 0$ at $n = 137$ are the extremes). The sign and magnitude of the principal correlation are preserved under proper clustering; the inferential precision implied by sequence-level resampling is overstated by roughly five-fold.

31 The Oracle Problem Across RNA Design Domains

The oracle problem documented in this paper for structural RNA design is a specific instance of a general challenge across computational RNA design. Table 40 maps our findings onto four broader domains.

Hierarchy of evaluation complexity. RNA design targets form three levels of increasing evaluation difficulty: (1) *structural*, where the sequence folds into a target structure (this paper); (2) *functional*, where the sequence undergoes a conformational change that triggers biological function (ri-

Table 40: **The oracle problem across RNA design domains.** Structure prediction (this paper) is the most tractable evaluation target; proxy-reality gaps widen for functional and phenotypic targets. R^2_{proxy} indicates variance in experimental outcomes explained by computational proxies.

Domain	**Evaluation proxy**	**Ground truth**	$R^2_{\textbf{proxy}}$	**Circularity?**
Inverse folding (this paper)	MFE/ensemble fold (4 oracles disagree)	SHAPE, crystals	0.20–0.64	Yes (this paper)
Riboswitch design	NUPACK ΔG (ON vs OFF)	In vivo fold change	0.01–0.06 [63]	Yes (NUPACK design+eval)
mRNA therapeutics	MFE, CAI, DegScore (10+ tools, 3 clusters)	Protein expr., half-life	0.04–0.43 [64]	Yes (MFE as design+eval)
CRISPR gRNA design	Doench, DeepCRISPR (15 tools, best r=0.45)	Editing efficiency	0.02–0.25 [65]	Implicit (score-ranked)

boswitches, biosensors, ribozymes); and (3) *phenotypic*, where sequence optimization produces a desired cellular outcome (mRNA expression, CRISPR editing efficiency). At each level, the gap between computational proxy and biological reality widens: structural design requires predicting one fold (our oracles achieve MCC 0.55–0.80); functional design requires predicting *two* conformational states and their energy balance (thermodynamic models explain only 1–6% of riboswitch activation variance [63]); phenotypic design requires predicting a high-dimensional cellular outcome from low-dimensional sequence features (translation efficiency models trained on synthetic reporters drop from $R^2 = 0.93$ to $R^2 = 0.20$–0.40 on endogenous mRNAs).

Recurring patterns. Three patterns documented in this paper recur across all levels: (i) *Single-oracle circularity*: NUPACK is simultaneously the design and evaluation tool for toehold switches [66] and riboswitches, just as ViennaRNA is for RNAinverse. LinearDesign [64] optimizes MFE and evaluates by MFE. (ii) *Narrow optimization fails*: in mRNA design, codon-only optimization (CAI) produces sequences that can fold into overly stable structures impairing translation, while joint codon+structure optimization succeeds, paralleling our finding that MFE-only optimization (RNAinverse) fails where ensemble optimization (NUPACK) succeeds. (iii) *Multi-tool disagreement*: in CRISPR guide design, 15 scoring algorithms evaluated on 16 datasets show the best individual tool achieving only Spearman $r = 0.445$ [65, 67], directly echoing our cross-oracle disagreement finding.

Implication. Our multi-oracle evaluation framework provides a template for any RNA design domain that relies on computational proxies. The key recommendations (evaluate with multiple independent tools, report disagreement as a quality signal, calibrate confidence tiers against experimental data) apply directly to riboswitch, mRNA, and gRNA design evaluation. The *evaluation transfer spectrum* concept suggests that these domains face even larger proxy-reality gaps than inverse folding, making multi-metric evaluation more consequential at the functional and phenotypic levels than at the structural level.

32 Reporting Template for RNA Design Evaluation

To support adoption of the multi-oracle evaluation protocol, Table 41 lists the reporting fields that the **ACCORD** package emits in its JSON report. Future RNA design papers can populate these fields directly from the package output. The template operationalizes recommendations **(1)**, **(2)**, **(3)** of §5 as reporting requirements; CLI usage and detailed field definitions accompany the released package.

The CLI emits a machine-readable JSON report whose top-level fields (`n_designs`, `oracles`, `precondition_applied`, `tier_counts`, `per_oracle_mean_mcc`, `mean_cross_oracle_std`) match the rows above and can be consumed directly by leaderboard ingestion scripts. Installation, command-line syntax, and worked examples are documented in the package README.

Table 41: **Recommended reporting fields.** Required fields are mandatory for cross-paper comparability under the cross-parameter-family protocol; recommended fields aid mechanistic interpretation.

Field	Notation	Req.
Design method name; number of designs n; design oracle (if any)	strings, n	yes
Proxy precondition pass rate (any oracle MCC ≥ 0.5)	%	yes
Per-oracle mean MCC across {Vienna, EternaFold, CONTRAfold, NUPACK}	4 values	yes
Oracle-dependent zone % (non-unanimous pass/fail)	%	yes
Cross-oracle std (mean, median)	2 values	yes
Native baseline std on the same panel; designed-to-native std ratio (flag > 2)	2 values	yes
Confidence tier breakdown (High / Moderate / Low)	3 %	yes
Transfer discount at proxy MCC ≥ 0.9	value	rec.
Ensemble defect alongside MCC	scalar or distrib.	rec.
Seeds and oracle versions (cf. Table 4)	metadata	yes

33 Comparison to Existing Evaluation Frameworks

Table 42: **Positioning against evaluation frameworks in ML.** All frameworks audit the reliability of an evaluation practice; our work is the instance in structural biology / molecular design.

Framework	Domain	Primitive audited	Grounding	Artifact
BetterBench [12]	All AI benchmarks (24)	Benchmark design quality	46-criterion rubric	Living rubric + website
Leaderboard Illusion [15]	Chatbot Arena ELO	Model ranking stability	Selection-bias audit	Post-hoc analysis
Construct Validity [11]	LLM benchmarks (445)	Construct operationalization	Measurement theory	Conceptual framework
HELM [14]	LLM evaluation	Aggregated leaderboards	Multi-metric reporting	Leaderboard + library
PoLL [68]	LLM-as-judge	Single-judge reliability	Judge-ensemble comparison	Method + empirical study
BEACON [47]	RNA understanding (13 tasks)	RNA model benchmarking	Standard test sets	Benchmark suite
This paper	**RNA design**	**sc-score reliability**	**SHAPE (n=40,201) + G-theory + N_2**	**Tiered protocol + pip pkg.**

Relative to these frameworks, this paper adds (i) experimental grounding via SHAPE chemical mapping at $n = 40{,}201$, (ii) joint deployment of generalizability theory, Hill's N_2, and Krogh–Vedelsby ensemble decomposition, (iii) a 2×2 mechanistic falsification (LEARNA + ensemble-local-search) separating optimization *objective* from optimization *intensity*, and (iv) a paradigm-spectrum Transfer Index across five RNA design paradigms.

34 Protein Design Evaluation Bridge

Self-consistency scoring was transferred from protein design to RNA design without recalibrating for the dramatically different oracle-accuracy regime. Protein design experienced an "AlphaFold moment" around 2021 in which oracle RMSD dropped to $\sim$ 1–2 Å, bringing scTM into a usable signal-to-noise regime; RNA structure prediction has not yet reached a comparable inflection point [69]. We provide a quantitative comparison below.

The signal-to-noise argument (heuristic). Let B be the designed backbone, T the true structure of the designed molecule, and P the oracle's prediction. The computed sc-score is $\text{TM}(P, B)$; the true design quality is $\text{TM}(T, B)$. The sc-score is reliable when oracle error $\epsilon = \text{RMSD}(P, T)$ is small relative to the designability threshold τ. For proteins: $\epsilon \approx$ 1–2 Å (AlphaFold2/ESMFold on CASP14 benchmark proteins) and $\tau = 2$ Å (Watson et al. [24]), giving SNR $\approx \tau/\epsilon \approx$ 1–2. For RNA 3D: $\epsilon \approx 4$ Å (RhoFold+ on RNA-Puzzles), so for any $\tau \leq 4$ Å, SNR ≤ 1: the oracle error dominates the signal. The $\sim$ 2–4$\times$ accuracy gap substantially degrades scTM discriminative

power; the practical effect is qualitatively visible in the multi-oracle disagreement we report (Table 2, Fig. 3).

Narrow predictor diversity in protein design evaluation. ESMFold [26] is a popular high-throughput refolding tool for protein backbone generation, owing to its $\sim$60$\times$ speed advantage over AlphaFold2 and its single-sequence (no MSA) input; AF2-family predictors remain widely used for higher-quality validation. Table 43 surveys recent backbone-generation papers: most rely on one dominant in-silico refolding protocol per study (ESMFold or an AF2-family predictor), and cross-predictor audits, where reported, remain uneven and small-scale. Even when papers use two predictors, those predictors share architectural lineage (transformer-based) and training data (PDB), so the diversity remains narrow within an "AF-lineage" family. The hidden vulnerability is structural: a fold class poorly modeled by the dominant family would inflate designability without obvious detection.

Table 43: **Common structure-prediction oracles in recent protein-design evaluation.** A non-exhaustive survey of recent backbone-generation papers (2022–2025). "Primary oracle" is the in-silico refolding tool named most prominently for designability; "Cross-oracle?" indicates whether the paper presents a systematic cross-predictor comparison. ESMFold and AF2-family predictors dominate; cross-predictor auditing is uneven.

Method	Year	Primary oracle	Cross-oracle?
RFdiffusion [24]	2023	AF2 + ESMFold	Partial (both reported)
Chroma [70]	2023	Multi-predictor (AF2/ESMFold/OmegaFold)	**Yes**
FrameFlow [71]	2023	ESMFold	No
FoldFlow [72]	2024	ESMFold	No
FrameDiff [73]	2023	ESMFold	No
Genie [74]	2023	OmegaFold (ESMFold sensitivity)	Partial
ProteinSGM [75]	2023	OmegaFold (+ experimental)	No
ProteinMPNN [25]	2022	AF2 (single-sequence)	No
ByProt/LM-Design [76]	2023	AF2 pLDDT	No
DPLM [77]	2024	ESMFold (+ OmegaFold)	Partial
PXDesign [78]	2025	AF2-IG + Protenix	**Yes** (dual filter)
RFdiffusion-AA [79]	2024	RoseTTAFold-AA	No

Cross-predictor evidence (limited). Where dual-oracle evaluation has been reported: (a) DPLM [77] reports cross-predictor differences in scTM between ESMFold and OmegaFold; (b) PXDesign [78] uses two independent predictors (AF2-IG and Protenix) with multi-tier filtering, and reports 17–82% wet-lab success rates depending on target difficulty, demonstrating that even dual-oracle filtering does not guarantee experimental success.

Table 44: **Protein vs. RNA sc-score evaluation regimes.** The 4$\times$ oracle accuracy gap and 200$\times$ training data gap mean that evaluation methodology validated for proteins does not transfer to RNA without recalibration.

Factor	Protein	RNA
Best 3D oracle accuracy	$\sim$1 Å (AF2)	$\sim$4 Å (RhoFold+)
Training structures	$\sim$170K	$\sim$782 clusters
Designability threshold	scRMSD $<$ 2 Å (calibrated)	None established
pLDDT calibration R^2	$\sim$0.8	0.607
Independent oracles	2–3 (ESMFold, AF2, OmegaFold)	5+ (Vienna, EF, CF, NUPACK, RhoFold+)
Oracle diversity	Low (all PDB-trained)	High (physics vs. ML)
Experimental validation	100s of designs characterized	$\sim$200 designs (gRNAde Eterna)
Exp. success rate	10–50% of scTM-passing	Unknown

Protein vs. RNA: quantitative comparison. The protein–RNA comparison reveals a *structural asymmetry*: protein design benefits from high oracle accuracy that masks the oracle problem, while RNA design is forced to confront it. Multi-oracle evaluation is therefore more consequential for

RNA, where oracle accuracy remains in a regime in which the disagreement is large relative to the structural signal. This paper provides systematic evidence for the regime distinction.

35 Cross-Target-Set Validation on Das Test

To rule out target-set confounding (the main text evaluates RNAinverse/NUPACK on Eterna100 targets and gRNAde/RiboDiffusion on Das test PDB targets), we run additional methods on the Das test PDB targets and evaluate through the same oracle panel.

Common-target subset (RNAinverse on Das test). Running RNAinverse on Das test targets (3-oracle std excluding NUPACK) yields oracle-dependent fraction 14.2% and pass-all 85.7%, with cross-oracle std 0.113, versus its Eterna100 numbers (72.9%, 20.5%, 0.268). gRNAde and RiboDiffusion on Das test give $27.0\%/47.4\%/0.054$ and $31.1\%/56.3\%/0.058$, respectively. On the same Das test targets, RNAinverse has a lower oracle-dependent fraction than gRNAde (14.2% vs. 27.0%) because PDB targets are easier (RNAinverse reaches ViennaRNA MCC 0.948, mostly transferring above the 0.5 threshold), but cross-oracle std is still $2.1\times$ higher than gRNAde (0.113 vs. 0.054). The Goodhart severity ranking by cross-oracle std (the threshold-independent metric) is preserved.

Additional 3D-conditioned baselines on Das test. We additionally evaluate 4 further methods on the Das test using the same oracle panel.

Table 45: **Cross-target-set validation: paradigm ranking is preserved across 6 methods.** Methods evaluated on Das test targets using 3 oracles (ViennaRNA, EternaFold, CONTRAfold); NUPACK additionally uses its own oracle (4 oracles). NA-MPNN [80] and RDesign [31] are 3D-conditioned inverse-folding methods; both show native-level cross-oracle agreement, consistent with gRNAde. Eterna100 rows shown for reference.

Method	Target set	N	Oracle-dep	Pass all	Std
NUPACK	Eterna100	521	15.5%	83.7%	0.076
NUPACK	Das test	970	**1.1%**	**98.9%**	**0.013**
NA-MPNN	Das test	245	11.4%	68.2%	0.060
RDesign	Das test	475	15.2%	69.3%	0.062
gRNAde	Das test	2,200	27.0%	47.4%	0.054
RNAinverse	Eterna100	3,550	72.9%	20.5%	0.267
RNAinverse	Das test	980	30.4%	69.6%	0.178

The paradigm ordering is *fully preserved*: NUPACK designs pass all oracles at 98.9% on Das test (vs. 83.7% Eterna100), with cross-oracle std of only 0.013 (median 0.000). RNAinverse on Das test shows 30.4% oracle-dependent (vs. 72.9% Eterna100); the lower absolute disagreement is likely because Das test targets have less challenging secondary structures than Eterna100 puzzles, but the relative ordering is maintained.

Two additional 3D-conditioned methods, NA-MPNN [80] (Baker lab unified biopolymer MPNN, OpenKnot R6 wet-lab validated) and RDesign [31] (ICLR 2024 geometric GNN), both show native-level cross-oracle std (0.060 and 0.062), consistent with 3D structural conditioning producing oracle-robust designs regardless of the specific generative architecture (autoregressive, diffusion, or MPNN).

This same-target comparison also enables multi-oracle 3D evaluation of designed sequences. We fold the best design per target through RhoFold+ and Boltz-2 [43] (Table 46).

The 3D ordering does *not* follow the 2D ordering: NUPACK has the best 2D agreement but lower 3D quality than gRNAde, whose 3D backbone conditioning provides structural information inaccessible to 2D oracles. This confirms that 2D evaluation, even when all 2D oracles agree, cannot substitute for 3D assessment (§4.4).

Table 46: **3D quality vs. 2D oracle agreement on Das test targets.** RhoFold+ TM-scores on best-per-target designs. NUPACK achieves near-perfect 2D cross-oracle agreement (std 0.013) yet lower 3D TM-score than gRNAde, whose 3D conditioning compensates for higher 2D variability. NA-MPNN and RDesign, both 3D-conditioned, show low 2D disagreement but modest 3D quality, suggesting that sequence recovery (∼0.42–0.66) alone is insufficient to guarantee native-like folds.

Method	Paradigm	n_{3D}	2D std	TM-score	RMSD (Å)
Native	—	95	0.057	**0.632**	—
gRNAde	3D-ML (autoregr.)	130	0.054	0.376	—
NUPACK	Ensemble-2D	88	**0.013**	0.249	3.15
RNAinverse	MFE	79	0.178	0.221	3.43

36 Oracle Accuracy on Native Sequences

We characterize each oracle's prediction accuracy on 95 native (undesigned) Das test sequences, comparing predicted secondary structures to PDB-derived reference structures (Table 47).

Table 47: **Oracle prediction accuracy on native sequences.** MCC between oracle-predicted and PDB-derived secondary structures for 95 Das test sequences. Under the default pipeline configuration, NUPACK underperforms the other three oracles; under alternative parameterization (RNA99) NUPACK's accuracy rises to Vienna-comparable levels (see *stress test* below).

Oracle	Family	Mean MCC	Std	Cohen's d vs NUPACK (default)
CONTRAfold	ML	**0.740**	0.183	1.53
EternaFold	ML	0.735	0.184	1.51
ViennaRNA	Physics	0.720	0.212	1.37
NUPACK (default, main reporting)	Physics	0.397	0.260	—

NUPACK parameterization stress test. The 0.397 MCC reported for NUPACK on natives is configuration-specific. To separate parameterization from any underlying accuracy gap, we re-evaluated all 95 Das test natives under three NUPACK configurations using a direct `nupack.mfe()` call:

Table 48: **NUPACK stress test on Das test natives.** Under RNA99 parameters NUPACK reaches MCC comparable to ViennaRNA/EternaFold (0.72/0.74); under the default RNA configuration it reads 0.601–0.642. NUPACK's design success is therefore not explained by superior native-sequence prediction accuracy under any configuration tested here.

NUPACK configuration	Mean MCC	Median	Std
Default (RNA material, 37°C)	0.601	0.626	0.240
RNA99 parameter set (Mathews 1999)	**0.712**	0.705	0.213
Watson–Crick-only filter (default + WC filter)	0.642	0.702	0.254

Two observations follow. *(i)* The main-table figure of MCC = 0.397 used the high-level `fold()` wrapper with different post-processing; the `nupack.mfe()`-direct pipeline above gives MCC = 0.601 under the same default parameters, so the main-table number is configuration-dependent. *(ii)* Under RNA99 parameters, NUPACK MCC on natives (0.712) is essentially indistinguishable from ViennaRNA (0.720) and EternaFold (0.735), so the "worst-predictor-best-designer" framing applies only under the default RNA configuration. The parameterization-conservative finding is sharper: **NUPACK's design success is not explained by superior native-sequence prediction accuracy.** Even at RNA99-best (0.712), NUPACK prediction is comparable to, not better than, the ML-trained oracles; yet NUPACK design produces Transfer Index 0.89 versus 0.72 for 3D-conditioned gRNAde designs. The gap in design quality is not mediated by the gap in prediction accuracy.

Proxy inflation quantified per oracle. On RNAinverse designs (optimized for ViennaRNA), ViennaRNA scores *increase* by 0.12 relative to native sequences (proxy inflation), while gold oracles

EternaFold and CONTRAfold *decrease* by 0.30 and 0.34 respectively (gold deflation). NUPACK designs show the opposite: all oracles score ≥ 0.77, and the cross-oracle std drops to 0.076, below the native baseline (0.193 with 4 oracles, 0.057 with 3 excluding NUPACK).

37 Ensemble Theory of Multi-Oracle Evaluation

Our multi-oracle evaluation framework connects to the well-established ensemble learning literature. We show that existing ensemble theory explains both *why consensus fails* and *why agreement is informative* in our setting.

The Krogh-Vedelsby decomposition. For M oracles with predictions $f_1(x), \ldots, f_M(x)$ and ensemble mean $\bar{f}(x)$, the ambiguity decomposition [61] gives an exact identity for squared error:

$$E_{\text{ens}}(x) = \bar{E}(x) - A(x), \tag{1}$$

where $\bar{E}$ is the average individual error and $A = \frac{1}{M}\sum_i (f_i - \bar{f})^2 \geq 0$ is the ambiguity (diversity). Since $A \geq 0$, the ensemble error is always $\leq$ the average individual error, but *not* necessarily $\leq$ the best individual. When the best oracle (EternaFold, SHAPE AUC 0.714) substantially outperforms the average, and diversity A is small (due to high inter-oracle correlation), the ensemble cannot close the gap. This explains our empirical finding that no aggregation method (majority vote, stacking, or MLP) beats EternaFold (App. 22).

Effective ensemble size and the diversity deficit. For M oracles with equal individual variance σ^2 and average pairwise correlation $\bar{\rho}$, the ensemble variance is [81]:

$$\text{Var}(\bar{f}) = \sigma^2 \left[\bar{\rho} + \frac{1-\bar{\rho}}{M}\right]. \tag{2}$$

The effective ensemble size $M_{\text{eff}} = M/[1 + (M-1)\bar{\rho}]$ [81] quantifies how many independent measurements the panel provides. Solving $M_{\text{eff}} = N_2 = 1.94$ at $M = 4$ (App. 12) yields $\bar{\rho} \approx 0.354$, giving variance reduction of $\approx 48\%$ (vs. 75% for 4 independent oracles: only $\approx 65\%$ of the theoretical maximum). Note: N_2 (Hill number from the full eigenvalue spectrum) and M_{eff} (design effect assuming uniform pairwise correlation) measure related but distinct aspects of panel redundancy; we equate them here as a first-order approximation, which is exact when the correlation matrix has one dominant eigenvalue. The Condorcet Jury Theorem [82] predicts negligible majority-vote gain when $M_{\text{eff}} \approx 2$, consistent with our consensus results.

The dual use of ensembles: accuracy vs. uncertainty. The key theoretical insight is that multi-oracle evaluation serves two distinct purposes [62]:

1. **Ensemble for accuracy** (consensus prediction): requires low $\bar{\rho}$ and high A. *Fails* in our setting ($N_2 = 1.94$, consensus $\leq$ EternaFold).
2. **Ensemble for uncertainty** (disagreement as reliability signal): requires oracles to have *different failure modes*, not necessarily low overall $\bar{\rho}$. *Succeeds*: cross-oracle agreement predicts experimental accuracy with a 12 pp gap ($\rho = -0.45$, $p < 10^{-200}$).

Ovadia et al. [83] showed that ensemble variance is particularly well-calibrated under distribution shift, precisely the regime of designed (OOD) sequences. Analogously, designed RNA sequences shift away from the native distribution on which oracles were trained; cross-oracle variance detects this shift by flagging sequences where oracle assumptions diverge.

Connection to structural biology. The CASP consensus approach (Pcons) uses the same principle: agreement among independently generated protein structure models predicts model quality. In drug discovery, consensus scoring across docking programs reduces false positives by filtering predictions where scoring functions disagree. Our multi-oracle evaluation extends this paradigm to RNA design evaluation, with the theoretical grounding provided by the Krogh-Vedelsby decomposition and the Wood et al. [84] unified diversity framework. The practical implication: the value of multiple oracles lies not in their consensus prediction but in their disagreement as an uncertainty signal. Increasing oracle diversity ($N_2 \to M$) would improve both uses simultaneously.

38 Paradigm Table: Uniform Precondition Details

Table 2 reports Oracle-dep, Pass-all, and N under a uniform proxy precondition (max oracle MCC ≥ 0.5). We also report the raw-pool numbers, which differ primarily for LEARNA (34.6% selection rate under the uniform precondition; an alternative reporting applies a ViennaRNA ≥ 0.7 filter for LEARNA only). Kendall rank correlation between raw-pool and preconditioned-pool Pass-all ordering across the 7 method rows is $\tau = 0.81$ ($p = 0.01$), so the paradigm spectrum is preserved under either conditioning. Raw-pool numbers (method, Oracle-dep, Pass-all): NUPACK 15.5%/83.7%, SAMFEO 18.6%/67.9%, LEARNA 15.2%/19.4%, Meta-LEARNA 25.3%/68.3%, gRNAde 27.0%/47.4%, RiboDiffusion 31.1%/56.3%, RNAinverse 72.9%/20.5%, Native 55.8%/41.1%. The full per-method table is included in the released package (App. 42).

Uniform 3-oracle rebuild of Table 1. As a sensitivity check, we rebuild Table 2 applying a single uniform 3-oracle panel (ViennaRNA + EternaFold + CONTRAfold) and the same $\max_{\text{3-oracle}}$ MCC ≥ 0.5 precondition to every method, removing the confound where SAMFEO and Meta-LEARNA use a 3-oracle panel in the main table while the other six rows use a 4-oracle panel. Across 7 design methods, the paradigm-spectrum ordering is preserved and the extremes sharpen: NUPACK oracle-dependent fraction 15.5% (vs. 15.7% under the 4-oracle baseline), SAMFEO 21.5%, RiboDiffusion 17.9%, gRNAde 17.6%, Meta-LEARNA 27.1%, LEARNA 33.1%, RNAinverse 67.9%. The transfer discount at high proxy (max MCC ≥ 0.9) under the 3-oracle panel is 0.023 for NUPACK, 0.021 for SAMFEO, 0.044 for RiboDiffusion, 0.060 for LEARNA, 0.071 for gRNAde, 0.145 for Meta-LEARNA, and 0.485 for RNAinverse: a $20\times$ span, compared with the $15\times$ span under the mixed-panel presentation in Table 2, indicating that the mixed-precondition footnote in the main table mildly underestimates rather than inflates the true paradigm gap.

39 Shared-Family-Bias Sensitivity (Leave-One-Family-Out)

The cross-oracle-disagreement↔SHAPE-AUC correlation reported in §4.5 could be driven by shared-parameter-family bias: ViennaRNA and NUPACK share the Turner 2004 thermodynamic parameters, while EternaFold and CONTRAfold share the CONTRAfold SCFG inference engine. To test this, we recompute the sequence-level Spearman correlation between disagreement and mean SHAPE AUC under seven panel compositions (Table 49). On Ribonanza we re-derive the analysis to include NUPACK-pk as a fourth oracle; on OpenKnot only the three oracles we folded (V, EF, CF) are available.

Three findings are robust across both datasets. **(i)** The *pooled cross-family* correlation (the mean of $|\text{MCC}_V - \text{MCC}_{EF}|$ and $|\text{MCC}_V - \text{MCC}_{CF}|$ versus mean SHAPE AUC) is essentially identical to the full-panel correlation: $\rho = -0.440$ $[-0.454, -0.426]$ versus the pooled $\rho = -0.448$ on Ribonanza (paired bootstrap $\Delta\rho = +0.026$ $[+0.022, +0.030]$); $\rho = -0.341$ $[-0.353, -0.329]$ versus $\rho = -0.344$ on OpenKnot (paired $\Delta\rho = +0.003$ $[+0.001, +0.006]$). Only $\sim$6% of the pooled correlation is attributable to within-family agreement on Ribonanza, and essentially 0% on OpenKnot. **(ii)** The within-family pair EF+CF is *weaker* than either cross-family pair ($\rho = -0.358$ on Ribonanza, $\rho = -0.218$ on OpenKnot), the opposite of what shared-family collusion predicts. **(iii)** Adding NUPACK to the Ribonanza panel strengthens the correlation ($\rho = -0.496$), and the strict physics-only pair V+N yields the *weakest* non-within-family correlation ($\rho = -0.333$), indicating the Turner-2004 sub-panel is not driving the signal.

A per-method breakdown on OpenKnot shows the cross-family/within-family correlation-strength ratio is > 1 for every design method with a non-degenerate correlation (ratios 1.13–2.23), with the largest ratios on the deep-learning-designed sets (MPNN-fixbb 2.21, 3DRNA 2.23, Rosetta-LoRes 2.06, gRNAde 1.97). We conclude that the cross-oracle correlation primarily reflects genuinely convergent evidence across parameter families rather than within-family collusion.

40 Training-Data Leakage Audit

We apply the same training-test overlap audit to every learned method and oracle in the paper: we identify each model's publicly-documented training corpus, construct an explicit overlap with the benchmark used here, and report the resulting leak rate and its impact on reported numbers. Where the corpus is not public, we mark the entry as *undocumented* and note the remediation needed.

Table 49: Cross-oracle disagreement $\leftrightarrow$ mean SHAPE AUC under different oracle panel compositions. CIs are 2000-draw bootstrap over sequences (seed $= 42$). F = family. V = ViennaRNA, EF = EternaFold, CF = CONTRAfold, N = NUPACK-pk.

Dataset	Panel	Oracles	n	ρ [95% CI]
Ribonanza	full 3-oracle (main)	V+EF+CF	14 269	-0.448 [-0.462, -0.435]
Ribonanza	full 4-oracle	V+EF+CF+N	14 269	-0.496 [-0.508, -0.484]
Ribonanza	within-F	EF+CF	14 269	-0.358 [-0.372, -0.344]
Ribonanza	physics-only	V+N	14 269	-0.333 [-0.347, -0.319]
Ribonanza	cross-F	V+EF	14 269	-0.418 [-0.431, -0.404]
Ribonanza	cross-F	V+CF	14 269	-0.410 [-0.424, -0.395]
Ribonanza	**pooled cross-F**	$\frac{1}{2}$(V+EF, V+CF)	14 269	$\mathbf{-0.440}$ [-0.454, -0.426]
OpenKnot	full 3-oracle (main)	V+EF+CF	25 930	-0.344 [-0.356, -0.333]
OpenKnot	within-F	EF+CF	25 930	-0.218 [-0.230, -0.206]
OpenKnot	cross-F	V+EF	25 930	-0.317 [-0.329, -0.306]
OpenKnot	cross-F	V+CF	25 930	-0.296 [-0.307, -0.284]
OpenKnot	**pooled cross-F**	$\frac{1}{2}$(V+EF, V+CF)	25 930	$\mathbf{-0.341}$ [-0.353, -0.329]

Table 50: **Symmetric training-test overlap audit.** *Exact* counts identical PDB IDs or RNA sequences. *Seq-cluster* counts clustered matches using the method's own clustering protocol (80% sequence identity for gRNAde, 80%/60%/40% PSI-CD-HIT for RiboDiffusion, CD-HIT 80% for RhoDesign, 80% redundancy filter for EternaFold). *Struct-cluster* reports the 45% TM-score cluster used in the gRNAde paper. "date-undoc." = no PDB snapshot date stated in the method paper.

Method / oracle	Training corpus	Benchmark	Exact	Seq. clust.	Str. clust.	Verdict
RhoDesign	3,435 PDB + 369,499 RhoFold (CD-HIT 0.8)	Das test (98)	52	–	–	leaked, 44 held out
RiboDiffusion	7,322 PDB (date-undoc., 20–280 nt)	Das test (98)	–	–	–	date-cutoff undoc.
RDesign	2,218 PDB / 987 cluster (date-undoc.)	Das test (98)	–	–	–	date-cutoff undoc.
gRNAde	RNAsolo 31 Oct 2023, $\leq$4.0Å	Das test (98)	1*	**43**	0	seq-cluster leak 44%
EternaFold	Cloud Lab R71–R79, S-Processed	Eterna100	0	0	n/a	clean
EternaFold	Cloud Lab R71–R79	Ribonanza GPN15k+R1+PK50/90	0	0	n/a	clean
RhoFold+	BGSU 2022-04-13 + 2023-09-06 blind	Das test (98)	–†	–†	–†	not time-censored for Das†
Boltz-2	PDB + AF-DB (single-chain conditional)	Das test (98)	–	–	–	date-undoc.
RoseTTAFold2NA	2024 NA release	Das test (98)	–	–	–	date-undoc.

*The single exact PDB-code match (`3GS5`) arises because gRNAde treats `3GS5_1_C` (train) and `3GS5_1_B` (test) as distinct chain entities. †RhoFold+'s training corpus (BGSU representative set, 2022-04-13 cutoff) postdates the Das deposition dates, so the Das test set is inside RhoFold+'s training universe; we do not have an exact PDB-ID overlap artifact for RhoFold+ on Das. The published checkpoint reports separately on a post-cutoff blind set (through 2023-09-06), which is clean by construction but is not the benchmark used here.

gRNAde (seq-cluster leak 44%). Using gRNAde's own 80% sequence-identity clustering, **43 of 98** test RNAs share a sequence cluster with at least one training RNA. The leak is concentrated in a few Rfam-like clusters: cluster 660 (14 test RNAs, adenine-riboswitch-like 60-mers), cluster 475 (10 test RNAs of TPP riboswitch). Nine of the 14 canonical Das RNAs have a train match. gRNAde's own *structural* clustering (TM-score $>$0.45) gives 0/98 overlap, matching the paper's explicit claim. The direction of bias is conservative for the paradigm-spectrum result: if gRNAde benefits from training-time access to near-homologs, its reported recovery is an upper bound, and the high oracle disagreement we observe for gRNAde designs is, if anything, an underestimate of the Goodhart effect.

RiboDiffusion (no PDB cutoff date). RiboDiffusion [32] is trained on a curated PDB snapshot (7,322 tertiary structures, 2,527 unique sequences, length 20–280 nt) augmented with 17,000 RhoFold-predicted structures from RNAcentral, with PSI-CD-HIT sequence-cluster (0.4/0.6/0.8 identity) and US-align TM-score structure-cluster (0.4/0.5/0.6) holdouts re-applied to the augmentation set. The published methods do not state a PDB snapshot date and do not reference the Das et al. 2010 test benchmark, so train/test overlap with Das 2010 is not explicitly audited. Because all 98 Das targets are pre-2010 PDB depositions within the 20–280 nt window, they are likely inside

the RiboDiffusion training pool unless excluded by the cluster filters, and we mark RiboDiffusion as "date-cutoff undocumented" in Table 50. The direction of bias remains conservative for our central claim: if RiboDiffusion benefits from training-set overlap, its reported transferability is an upper bound, and the cross-oracle disagreement we observe is, if anything, an underestimate.

EternaFold (clean at the sequence level). We checked whether any EternaFold training sequence matches Eterna100 puzzle solutions (n=251 in-range) or Ribonanza sub-libraries (GPN15k, R1, PK50, PK90, n>140,000 total, adapters stripped). **Zero** exact matches; **zero** substring containments on Ribonanza. 8-mer Jaccard similarity is consistent with the shared RNA alphabet, not memorization. EternaFold's reported SHAPE-AUC correlation on Ribonanza ($\rho = -0.45$, our central calibration) therefore *does not* benefit from direct training-set leakage. The shared Eterna *ecosystem* (same players, same design heuristics) means generalization is tested within the Eterna design space, which is consistent with how EternaFold is described in its original paper.

Oracle-family independence ("4 oracles, 2 families"). Three of our four 2D oracles trace to two non-independent lineages: (i) ViennaRNA 2.x and NUPACK 4 both default to Turner 2004 nearest-neighbor parameters (Mathews et al., PNAS 2004); (ii) EternaFold is the CONTRAfold inference engine with retrained parameters (plus a multiloop bug fix and minimum-hairpin size of 3); they share the binary. On the Das test set, pairwise MCC-of-predicted-MFE-structure: Vienna–NUPACK $= 0.94$, EternaFold–CONTRAfold $= 0.97$, cross-family (Vienna–EternaFold) $= 0.76$. This does not change our Goodhart conclusions (which depend on disagreement, not agreement), but **readers should not interpret "four oracles" as four independent votes**. Hill's $N_2 = 1.94$ of 4 (§4.1) makes this quantitatively visible.

RhoDesign as a failure-mode case study. RhoDesign's 2D-oracle profile differs qualitatively from the other paradigms: raw selection rate 14.9%, fail-all rate 85.1% on the held-out-44, and mean ViennaRNA MCC 0.14–0.20. This profile is most naturally read as a structure-recovery failure mode rather than as a position on the paradigm spectrum. The Transfer Index entry for RhoDesign is computed on the precondition-passing pool only and is included in the paradigm summary for completeness; it should not be read as a within-distribution comparison against NUPACK, gRNAde, or LEARNA. The mechanistic picture is consistent with documented RhoDesign training-corpus leakage on Das clusters, which we examine below.

To quantify the impact of RhoDesign's 52% exact-PDB leak, we re-ran RhoDesign on the 44-target subset with zero training overlap (held-out list provided in the released package) at two temperatures: a low-temperature deterministic mode (T=10^{-5}, 5 samples/target) and a high-temperature diversity mode (T=1.0, 10 samples/target), yielding 577 unique designs after deduplication. On the 8 targets shared with RDesign (identical target structures), RhoDesign achieves mean ViennaRNA MCC $=$ 0.21 (median -0.008) vs. RDesign's 0.42 (median 0.53), **a 2× drop**. Pass-all rate on the held-out-44 (n=577 unique designs) is 4.7%; fail-all is 85.1% (partially inflated by 23/44 targets having biologically-impossible `()` adjacent-base-pair SS annotations, a pre-processing artifact that affects all methods on the same targets identically). The relative ordering is unambiguous: **a published method with high apparent performance on the full Das test collapses by ∼2× once its 52% training overlap is removed.** This is direct evidence of leakage-driven inflation in the Das test.

gRNAde leakage-stratified paradigm spectrum. We provide a two-layer empirical leakage control for gRNAde. *Layer 1: cluster-match stratification.* Using gRNAde's own $\geq 80\%$ sequence-identity cluster definition, the 98 Das test targets partition into **43 seq-leaked** and **55 zero-match** targets; the $\geq 45\%$ TM-score structure-cluster holdout from the original benchmark holds for all 98. Mean cross-oracle std on the zero-match subset, ranked: NUPACK 0.014, RDesign 0.060, NA-MPNN 0.063, gRNAde 0.086, RhoDesign-heldout-44 0.105, RiboDiffusion 0.134, RNAinverse-on-Das 0.189. The ranking on the full Das test is identical at the extremes (NUPACK best, RNAinverse worst). gRNAde's mean cross-oracle std rises modestly from the full-Das value of 0.076 to 0.086 on the zero-match subset, consistent with mild leakage assistance on cluster-matched targets but insufficient to change the paradigm-spectrum narrative.

Layer 2: full gRNAde retrain on a seq-cluster-stripped training set. We retrain gRNAde from scratch on a training set where every structure whose 80%-sequence-identity cluster matches any Das-test cluster has been removed. Of the $4{,}025$ original training structures, 468 (11.6%) are dropped, leaving $3{,}557$; the validation (100) and test (98) sets are unchanged. Training uses the upstream default

configuration (AR-v1, single-state, 100 epochs, learning rate 10^{-4}, label smoothing 0.05) on a single A40 GPU; wall-clock ~ 35 minutes. We sample 16 designs per test target at temperature 0.1 from the best-val checkpoint and fold every design through ViennaRNA, EternaFold, CONTRAfold, and NUPACK. Table 51 reports paradigm-level statistics for the original gRNAde and the retrained model on the 44 Das-test PDB IDs that overlap our original cross-oracle pool. Transfer Index *rises* from $\rho = 0.805$ to 0.822 ($+0.017$); transfer discount rises marginally from 0.068 to 0.075 ($+0.007$, still $\sim 3\times$ smaller than RNAinverse's 0.197); the oracle-dependent zone rises from 17.4% to 24.3%, consistent with the ~ 0.06 drop in per-oracle mean MCC attributable to the 11.6% training reduction plus the default 100-epoch budget. The paradigm-level ranking is preserved: gRNAde remains in the low-transfer-discount, high-transfer-ρ quadrant, well separated from RNAinverse. Extending to all 98 Das-test targets (54 additional targets not in our original cross-oracle pool) gives $\rho = 0.817$ and discount 0.082, quantitatively consistent with the 44-target subset.

Table 51: **Retrained gRNAde vs. original: paradigm-level statistics on the 44 overlapping Das-test targets.** Transfer Index rises by $+0.017$; discount rises by $+0.007$; paradigm position preserved. V/EF/CF/NP = ViennaRNA / EternaFold / CONTRAfold / NUPACK; xstd = cross-oracle std; T-ρ = Transfer Index; Disc = transfer discount.

Model	n	V	EF	CF	NP	xstd	T-ρ	Disc
Original gRNAde	2,200	0.587	0.639	0.625	0.545	0.076	**0.805**	0.068
Retrained (clean split)	704	0.526	0.563	0.536	0.475	0.089	**0.822**	0.075

Summary. gRNAde's 44% sequence-cluster leak on the Das test makes our reported gRNAde numbers a conservative estimate of the Goodhart effect rather than an over-statement, and the seq-cluster-stripped retrain confirms paradigm-position preservation. RiboDiffusion's training corpus lacks a documented PDB cutoff date and is therefore flagged but not directly auditable. EternaFold shows no direct sequence leakage against either Eterna100 or Ribonanza. Oracle-family non-independence is disclosed and quantified. The RhoDesign held-out-44 re-evaluation provides an empirical stress test: $\sim 2\times$ performance drop on clean out-of-distribution targets.

41 Functional Triangulation: Toehold Switches, 5′UTR MPRA, and Ribozymes

Toehold switches. We re-analyze the Angenent-Mari et al. [85] toehold-switch dataset ($n = 37{,}733$ quality-filtered sequences, 82 nt each, $\mathrm{QC}_{\mathrm{ON/OFF}} \geq 2$). Each switch OFF sequence (DNA→RNA converted) was folded through ViennaRNA, EternaFold, and CONTRAfold; for each design we compute mean pairwise MCC between predictions, per-position consensus, and paired-fraction std across oracles. The functional readout is the FACS-based ON/OFF activation ratio. Cross-oracle structural agreement has essentially no predictive power for this functional readout: Spearman $\rho = -0.018$ ($p = 5.8 \times 10^{-4}$, bootstrap 95% CI $[-0.028, -0.008]$). Position consensus and pre-computed NUPACK ensemble defect give similarly negligible correlations ($\rho = -0.012$ and $+0.018$). All these effects achieve statistical significance solely because of the massive sample size; the effect sizes are negligible ($|\rho| < 0.02$). Stratifying ON/OFF by quintile of cross-oracle agreement (Table 52) shows mean activation varies by only 1.5 pp across quintiles, vs. the 12 pp structural-accuracy gap in Table 3.

Table 52: **Toehold ON/OFF activation by cross-oracle agreement quintile.** Mean ON/OFF varies by only 1.5 pp across quintiles.

Agreement quintile	n	Mean agreement (MCC)	Mean ON/OFF
Q1 (lowest)	7,547	0.815	0.287
Q2	7,547	0.901	0.284
Q3	8,417	0.934	0.291
Q4	6,690	0.959	0.276
Q5 (highest)	7,532	0.982	0.272

The Angenent-Mari pool, however, is pre-filtered: the 37,733 QC-passing switches were selected using a NUPACK ensemble-defect screen *before* the FACS functional readout. Cross-oracle agreement is therefore compressed into the top decile (quintile means 0.815–0.982), and the experiment tests whether *residual agreement variation among already-good designs* predicts function, not whether agreement predicts function across a diverse pool. A definitive test requires functional datasets that do not pre-filter on design-oracle scores.

5′UTR MPRA and ribozyme activity (un-filtered). We therefore replicate the methodology on two public datasets that do not pre-filter sequences through design oracles: Sample et al. 2019 5′UTR MPRA [86] (random 50-nt sequences + native human 5′UTRs, with polysome-profile-derived MRL) and Roberts et al. 2023 ribozyme activity [87] (124,425 variants across 6 ribozyme families with normalized cleavage activity).

Table 53: **Functional Goodhart triangulation across three independent RNA-activity datasets.** Spearman ρ between cross-oracle structural agreement (mean pairwise MCC of Vienna/EternaFold/CONTRAfold) and a molecular function readout, with 2000-draw bootstrap 95% CI. The agreement-range column shows the empirical 5–95 percentile span of mean pairwise MCC values; a narrow range indicates oracle pre-filtering by the experimental pipeline (NUPACK in the toehold case). No dataset achieves $|\rho| \geq 0.2$; the two 5′UTR datasets and the HHIII/Hatchet ribozyme families are indistinguishable from null, while four ribozyme families show a robust *negative* correlation that is inconsistent with any "agreement → function" proxy claim.

Dataset	n	ρ	95% CI	Pearson r	Range
Angenent-Mari 2020 toehold (NUPACK pre-filter)	37,733	-0.018	$[-0.028, -0.008]$	-0.005	$[0.81, 0.98]$
Sample 2019 5′UTR MPRA, random 50-nt	20,000	-0.018	$[-0.031, -0.005]$	-0.007	$[0.16, 1.00]$
Sample 2019 5′UTR MPRA, native human	9,486	-0.053	$[-0.072, -0.033]$	-0.019	$[0.12, 1.00]$
Roberts 2023 ribozyme, pooled	124,425	-0.066	$[-0.071, -0.061]$	-0.053	$[0.31, 0.99]$
CPEB3	21,322	-0.122	$[-0.135, -0.108]$	-0.085	$[0.31, 0.98]$
Hairpin	22,579	-0.162	$[-0.174, -0.151]$	-0.099	$[0.63, 0.98]$
Twister	10,297	-0.134	$[-0.152, -0.114]$	-0.145	$[0.14, 1.00]$
HDV	33,919	-0.153	$[-0.164, -0.142]$	-0.167	$[0.42, 0.98]$
Hatchet†	27,262	$+0.003$	$[-0.009, +0.014]$	$+0.007$	$[0.41, 0.97]$
HHIII†	9,046	$+0.019$	$[-0.001, +0.039]$	$+0.053$	$[0.30, 1.00]$

Range column reports the 5–95 percentile of mean pairwise MCC. †Hatchet activity concentrates near 1.0 (mean 0.99 vs. 0.33–0.57 elsewhere); HHIII is the shortest family (45 nt); both conditions compress the available range of detectable disagreement.

Interpretation: pre-filtering does not drive the null. Two findings address the concern that the Angenent-Mari pre-filter could drive the result: (i) the Sample 2019 5′UTR *random* subset has no design-oracle pre-filter and spans an agreement range 8× wider than Angenent-Mari (MCC percentile span $[0.16, 1.00]$ vs. $[0.81, 0.98]$), yet returns *identical* $\rho = -0.018$. The Angenent-Mari null is therefore not an artifact of range compression. (ii) Four of six ribozyme families (CPEB3, Hairpin, Twister, HDV, covering 88,117 variants) show robust *negative* correlations ($\rho = -0.12$ to -0.16): sequences with *higher* cross-oracle structural agreement have *lower* catalytic activity. This anti-predictive signal is consistent with a mechanistic explanation: easy-to-fold sequences (high agreement) are often unstructured or trivially structured and cannot support the precise catalytic geometry required for cleavage, while catalytically competent folds are structurally non-trivial and oracles disagree more on them.

At $n = 10^4$–10^5, $|\rho| < 0.1$ correlations are *statistically significant nulls* (the alternative hypothesis of "exactly zero" is rejected only because the sample is huge). We therefore report Fisher-z two-one-sided tests [88] for equivalence to zero at both $|\rho| < 0.05$ (strict) and $|\rho| < 0.10$ (lenient) equivalence bounds at $\alpha = 0.05$:

Implication for the evaluation complexity hierarchy. Cross-oracle agreement is a strong *structural* reliability signal ($\rho = -0.45$ with SHAPE AUC, Table 3) and a near-zero or *anti-predictive* signal for downstream function. The gap between structural and functional prediction is $>20\times$ in magnitude and, for catalytic function, *reverses sign*. Any future "agreement-as-function" claim in RNA design must justify which level of the hierarchy it claims to predict and report family-stratified

Table 54: **TOST equivalence testing for functional triangulation.** "Equivalent" means the observed correlation is statistically bounded within the stated equivalence region at $\alpha = 0.05$.

Dataset	ρ	n	Eq. $\lvert\rho\rvert<0.05$?	Eq. $\lvert\rho\rvert<0.10$?
Toehold [85]	-0.018	37,733	**yes** ($p=2\times10^{-10}$)	yes
UTR MPRA random [86]	-0.018	20,000	**yes** ($p=3\times10^{-6}$)	yes
UTR MPRA native human [86]	-0.053	9,486	no	yes
Ribozyme pooled [87]	-0.066	124,425	no	yes
CPEB3 / Hairpin / Twister / HDV	-0.12 to -0.16	88,117	no	no
Hatchet	$+0.003$	27,262	**yes**	yes
HHIII	$+0.019$	9,046	**yes**	yes

correlations at minimum. Multi-oracle agreement therefore calibrates structural evaluation, not functional evaluation; functional evaluation requires a separate oracle class.

Equivalence testing (TOST) for the functional nulls. Under strict equivalence ($|\rho| < 0.05$), the toehold, UTR random, Hatchet, and HHIII correlations are rigorously bounded near zero. The four-family ribozyme cluster (CPEB3, Hairpin, Twister, HDV) is *not* equivalent to zero at either bound: these are real (negative) effects, not statistical flukes inflated by sample size.

42 Data and Software Availability

This appendix consolidates the public corpora, software tools, pre-trained models, and design methods used in this study, with citation, version, license, and canonical URL where applicable. Our companion package (**ACCORD**, §5) is the only artifact we distribute; everything below is fetched directly from the upstream source by the user.

42.1 Public RNA corpora and chemical-mapping data

- **Eterna100** (v1, v2) [53]: 100 RNA design puzzles. https://github.com/eternagame/eterna100-benchmarking (MIT).
- **Ribonanza** chemical-mapping corpus and the **RibonanzaNet** chemical-mapping predictor [7]: SHAPE/DMS profiles for ∼2 M sequences. Calibration cohort here uses the Ribonanza-2A3 subset (n = 14,271). Distributed via the Ribonanza Kaggle competition portal.
- **OpenKnotAIDesignData** [7]: cross-pipeline replication corpus (n = 25,930) from the Eterna OpenKnot challenge. https://github.com/eternagame/OpenKnotAIDesignData.
- **Das natives test set**: 95 PDB-derived RNA targets with sequence and dot-bracket annotations, drawn from the gRNAde [30] evaluation corpus.
- **Protein Data Bank**: native 3D structures used as references; PDB is in the public domain. The 95 Das-test PDB identifiers are inherited from the gRNAde test split [30].

42.2 Secondary-structure prediction oracles (2D panel)

- **ViennaRNA** 2.7.0 [4]: Turner-2004 thermodynamics [48], MFE and McCaskill ensemble. https://www.tbi.univie.ac.at/RNA/ (GPL-2.0).
- **EternaFold** v1 [49]: CONTRAfold engine retrained on Eterna data. https://github.com/eternagame/EternaFold (BSD-3-Clause).
- **CONTRAfold** 2.02 [5]: CLLM trained on Rfam (BSD).
- **NUPACK** 4.0.0 [27]: Turner-derived nearest-neighbor with multistrand support. http://nupack.org (academic license; registration required).

42.3 Three-dimensional structure prediction oracles (3D panel)

All 3D oracles were run in single-sequence mode unless otherwise noted in the corresponding analysis (see App. 5).

- **RhoFold+** [40]: https://github.com/ml4bio/RhoFold (Apache-2.0).
- **DRfold2** [41]: https://github.com/leeyang/DRfold2.
- **AlphaFold3** [42]: https://github.com/google-deepmind/alphafold3 (Apache-2.0 with model-weight terms).
- **Boltz-2** [43]: https://github.com/jwohlwend/boltz (MIT).
- **RoseTTAFold2NA** [44]: https://github.com/uw-ipd/RoseTTAFold2NA (MIT).

42.4 RNA design methods

Five primary methods (RNAinverse, NUPACK, LEARNA, gRNAde, RiboDiffusion) and four out-of-corpus reference baselines (SAMFEO, Meta-LEARNA, SamplingDesign, RhoDesign).

- **RNAinverse** [52]: bundled with ViennaRNA above.
- **NUPACK design** [27]: bundled with NUPACK above.
- **LEARNA** and **Meta-LEARNA** [29]: https://github.com/automl/learna (Apache-2.0).
- **gRNAde** [30]: https://github.com/chaitjo/geometric-rna-design (MIT).
- **RiboDiffusion** [32]: https://github.com/ml4bio/RiboDiffusion.
- **SAMFEO** [51]: https://github.com/shanry/SAMFEO.
- **SamplingDesign** [28]: https://github.com/weiyutang1010/SamplingDesign.
- **RDesign** [31]: https://github.com/A4Bio/RDesign.
- **RhoDesign** [35]: https://github.com/ml4bio/RhoDesign.

42.5 Companion software artifact

The multi-oracle evaluation package (**ACCORD** v1.2.0, MIT license) is hosted at the anonymized review URL https://anonymous.4open.science/r/ACCORD-F49D during the double-blind review window and will migrate to a permanent de-anonymized public repository with a Zenodo DOI at camera-ready. The package is self-contained, executable, and documented: it ships the tier classifier, the oracle orchestration CLI, the canonical oracle specifications (`oracle_specs.yaml`), the SHAPE-calibrated tier thresholds (`tier_thresholds.json`), a Gebru-style datasheet plus a Croissant 1.0 JSON-LD with Responsible-AI fields (`DATASHEET.md`, `DATASHEET.json`), 66 hermetic pytest cases, a no-backend demo, and the reproducibility scripts behind the figures and tables in this paper. Raw RNA sequences and SHAPE reactivities are not redistributed; the package depends on users obtaining the calibration corpora directly from the upstream sources listed above.