KRAIG: Knowledge-Guided RNA Adaptive Inference and Generation under Biological and Protocol Uncertainty

Anonymous authors
Paper under double-blind review

Abstract

RNA therapeutic design requires a sequence of coupled decisions under heterogeneous biological evidence, assay-dependent supervision, and protocol shift: selecting an admissible therapeutic mechanism, generating and prioritizing candidate oligonucleotides, and determining whether a prediction is sufficiently supported to act upon. Existing approaches typically optimize these decisions independently and propagate point predictions even when the evidence required by a downstream model is unavailable. We formulate RNA therapeutic design as compositional selective inference and introduce KRAIG, an end-to-end framework in which every stage communicates an evidence certificate rather than an unconditional decision. KRAIG first converts typed biological evidence into dependence-robust mechanism intervals and returns a certified mechanism, an admissible set, or abstention. This certificate governs a capability-aware design component that activates mechanism-conditioned generation, within-experiment ranking, and conformal candidate selection only when their modality-specific validity, data, and calibration requirements are satisfied. A support-selective protocol component permits residual adaptation only within supported protocol regions and otherwise exactly recovers an invariant predictor. Under controlled experiments with known latent mechanism scores, certification incurs zero selective error when it returns a singleton, whereas forced point-decision baselines err on 15.7–21.9% of near-tie cases. Across three supported therapeutic modalities, constrained decoding increases chemistry-valid generation from 0–88.8% to 100%. Modality-stratified conformal selection attains 0.92 coverage at 0.90 nominal coverage in the adequately powered stratum while withholding unsupported guarantees elsewhere. Learned components are evaluated under a frozen five-seed protocol with leakage-resistant splits and paired controls. By composing certification, generation constraints, calibrated selection, specialist routing, and invariant fallback, KRAIG makes the decision to abstain or defer an operational part of RNA therapeutic design.

1 Introduction

RNA therapeutic design is a difficult test case for representation learning. Sequences are observed through different mechanisms, chemistry encodings, assay protocols, and experimental score scales; latent features may be shared while their outcome semantics remain incompatible. A system must therefore transfer common structure without treating missing biological facts as negative facts, syntactic validity as activity, or support in one protocol as support in another.

We ask: *how should heterogeneous representations be learned, composed, and refused when support changes between training and deployment?* We formulate the pipeline as support-aware compositional representation learning. Every component returns both a representation or decision and a certificate describing where it is admissible. A downstream component verifies the certificate or refuses the action. This common contract preserves sets when evi-

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

Table 1: **KRAIG's claim contract.** Each pipeline component pairs the claim it supports with the condition that limits or refuses that claim.

Component	Supported claim	When KRAIG limits or abstains
Triage	A singleton mechanism is sound when its lower bound exceeds all alternatives.	Overlapping evidence intervals return the maximal mechanism set.
Generate	Decoded candidates satisfy the declared sequence and chemistry constraints.	Syntactic validity is not evidence of biological activity.
Rank	Assay-local rank composition improves EF20 on the strict-family benchmark.	Evidence remains retrospective and benchmark-scoped.
Select	Supported calibration strata receive candidate-set coverage.	Underpowered modalities receive no coverage guarantee.
Adapt	A closed support gate exactly recovers the invariant predictor.	Residual adaptation is not claimed to improve accuracy.

dence does not identify a point decision, enforces hard validity during generation, conditions shared representations on declared capability, withholds finite-sample guarantees when calibration support is insufficient, and exactly recovers a protocol-invariant predictor outside adaptation support.

KRAIG instantiates this contract in five stages. TRIAGE converts typed biological evidence into dependence-robust mechanism intervals. A conditional generator emits mechanism-specific oligonucleotides under explicit chemistry constraints. KRAIG-SR ranks activity with a shared predictor trained using fixed modality-blocked updates. A capability layer routes to supported experts or refuses. Finally, modality-stratified conformal selection and support-gated residual adaptation control candidate selection and protocol shift.

KRAIG is being integrated into the broader *KRAIG RNA Therapeutics Platform*, an actively developed system spanning knockdown, upregulation, editing, and splicing strategies. The integration separates therapeutic intent from evidential support: a biological objective and its typed evidence define the admissible mechanism set; mechanism-specific constraints govern candidate generation; and ranking, calibration, and adaptation determine whether the available data support prioritisation under the requested assay protocol. This common interface is intended to let distinct RNA modalities share representations without treating them as biologically interchangeable. It also preserves an auditable path from an input evidence record to a generated candidate, its ranking context, and any abstention or fallback decision. The present work evaluates the algorithmic components and support-aware decision contracts underlying that integration. It does not claim experimental efficacy, safety, delivery performance, or clinical validation for candidates produced by the platform.

Our contributions are:

1. a support-aware representation contract spanning typed biological evidence, mechanism-conditioned generation, shared activity learning, assay-local rank combination, and protocol adaptation;
2. a fixed modality-blocked learning rule that mitigates negative transfer while retaining one shared predictor, and a monotone-invariant rank representation for combining experts whose score scales differ;
3. ASO-TRANSPORT, a family/Hamming-1-disjoint stress test on which these representation choices are evaluated using assay-local retrieval, family-clustered uncertainty, a label-free selective interface, and matched public OligoAI predictions; and
4. machine-checkable leakage and prediction-fidelity audits connecting each quantitative statement to immutable predictions or result artifacts.

Together, these contributions make representation support executable: mechanism ambiguity is preserved, sequence constraints are enforced at decode time, shared activity features are learned without mixed-modality batches, heterogeneous scores are compared through

assay-local ranks, calibration is scoped to supported strata, and protocol adaptation has an exact fallback. Each statement is evaluated against its own contract so that structural validity, ranking accuracy, and coverage are not conflated.

2 Related work

RNA sequence design and prediction. RNA therapeutic models typically target one mechanism: siRNA design (Huesken et al., 2005; Reynolds et al., 2004; Bai et al., 2024), RNase-H antisense activity (Hill et al., 2025; Hwang et al., 2024; 2026; Liu et al., 2026), or splice modulation (Chiba et al., 2021). PFRED separates siRNA and antisense workflows (Sciabola et al., 2021); generative RNA focuses on aptamers or structure-conditioned inverse design (Iwano et al., 2022; Runge et al., 2019; Tan et al., 2024). KRAIG treats therapeutic mechanism as a capability axis with refusal between stages. OligoAI is the strongest executable public ASO reference; ASOCompass reports higher overall Spearman via RiNALMo, Uni-Mol2, and auxiliary objectives (Hill et al., 2025; Liu et al., 2026) but its data, checkpoints, and random/predefined split are not commensurate with our family/Hamming-1 transport benchmark. We compare identical public OligoAI out-of-fold rows and treat ASOCompass as a published ceiling, not an executed baseline.

Selective prediction and conformal selection. Set-valued prediction and abstention avoid forced decisions (Hüllermeier et al., 2013; Mao et al., 2023). Conformal methods give finite-sample marginal coverage (Vovk et al., 2005; Tibshirani et al., 2019); screening emphasizes correct unit and stratum calibration (Jin & Candès, 2023; Bai et al., 2025). KRAIG applies this twice: to partial orders over mechanisms and to candidate-set calibration within supported modalities.

Representation learning under heterogeneous supervision. Shared learning can improve data efficiency but risks negative transfer (Caruana, 1997). GradNorm, PCGrad, CAGrad modify losses/gradients (Chen et al., 2018; Yu et al., 2020; Liu et al., 2021); task sampling and selective group updates alter schedules (Sundararaman et al., 2021; Jeong & Yoon, 2025); asymmetric multi-task methods learn directed relationships (Lee et al., 2016). KRAIG-SR is a simple diagnostic baseline: fixed modality order, single-modality batches, one shared network.

3 Methodology: KRAIG

Within the KRAIG RNA Therapeutics Platform, KRAIG is the algorithmic decision layer linking biological evidence to candidate design. Platform modules pass typed evidence, sequences, assay/protocol context, and stagewise certificates without changing the evaluation contracts studied here. Algorithm 1 in Appendix A summarizes the operational sequence.

KRAIG separates four representation roles that ordinary end-to-end pooling conflates: e represents typed evidence over mechanisms; $h_\theta(x, m)$ is a learned sequence representation shared under modality m; $r_g(z)$ is an assay-local rank representation of a predictor score; and p represents the experimental protocol. Their support certificates are not additional outcome predictors. They specify whether a downstream operation may select a point, decode a candidate, combine a score, calibrate a set, adapt to p, or must refuse. This separation lets each representation be tested under its relevant shift rather than treating one pooled accuracy number as end-to-end evidence.

3.1 Dependence-robust mechanism certification

Let $k \in \{1, \ldots, M\}$ index mechanisms. Each rule contains hard biological gates, required evidence, forbidden evidence, optional discriminators, and provenance. A resolved feature has a typed state $s_f \in \{\textsc{Present}, \textsc{Absent}, \textsc{Unresolved}, \mathrm{N/A}\}$ and, when meaningful, a marginal support value $p_f \in [0, 1]$. Failed hard gates reject a mechanism; unresolved required evidence halts it rather than assigning zero support.

For a surviving mechanism, let T_k contain required/discriminating supports and complemented forbidden supports. Without assuming independence, the sharp Frechet–Hoeffding envelope for their conjunction is

$$L_k = \max\left(0, \sum_{q \in T_k} q - (|T_k| - 1)\right), \qquad U_k = \min_{q \in T_k} q. \tag{1}$$

A provenance ceiling can only contract this interval. Mechanism k certifiably dominates j when $L_k > U_j$. Overlapping non-singleton intervals remain incomparable. TRIAGE returns the unique maximal mechanism only if the robust partial order has one maximal element; otherwise it returns the maximal set and abstains. This distinguishes applicability, evidence strength, and decision identifiability.

Proposition 1 (Sharp mechanism certificate) *Given only the marginal supports in T_k, $[L_k, U_k]$ is the sharp identified set for their conjunction: no uniformly tighter interval is valid without an additional dependence assumption. If every mechanism interval contains its latent applicability, a unique TRIAGE maximal element is the latent optimum; otherwise TRIAGE makes no point claim.*

3.2 Capability-aware constrained generation

The generator is a conditional VAE (Kingma & Welling, 2014; Kingma et al., 2016) over oligonucleotide sequence x, mechanism m, chemistry class c, and length. A convolutional encoder maps x to a 64-dimensional latent; mechanism and chemistry embeddings condition a GRU decoder. Conditioning dropout exposes a null/OOD condition during training. At decoding time, a finite-state constraint layer masks invalid lengths and chemistry patterns, while GC steering adjusts base probabilities toward the training mean of the requested mechanism.

The generator returns candidate sequences plus a validity certificate. This certificate checks alphabet, length, chemistry pattern, GC range, homopolymer limit, novelty, and diversity. It does not certify target binding or activity. Target-conditioned utility is therefore evaluated separately and cannot be inferred from the validity rate.

3.3 Support-aware shared activity representations

Absolute activity values are not comparable across studies, so every learned arm is evaluated by Spearman correlation within estimable experiments; the primary score is the unweighted macro average across experiments. The frozen diagnostic contains 3,837 training, 757 validation, and 2,589 test rows across siRNA and RNase-H. Splice-switching is evaluated separately because it is absent from this matrix.

The ordinary universal model mixes modalities in one loader. KRAIG-SR uses the same 5.48M-parameter shared architecture and one output head, but performs fixed RNase-H then siRNA blocks, updating after each single-modality batch. Separate specialists use 10.95M parameters in total. Capability routing selects a declared specialist only for supported modalities; the oracle uses test outcomes and is reported solely as an upper bound.

For the larger ASO benchmark, KRAIG combines complementary out-of-fold experts without assuming comparable score scales. For assay g, expert j, and candidate i, let

$$r_{igj} = \frac{N_<(i,g,j) + (N_=(i,g,j) + 1)/2}{n_g} \tag{2}$$

be the within-assay average-rank percentile of expert score z_{igj}, where $N_<$ counts strictly worse scores and $N_=$ counts tied scores in the assay (pandas average ranks divided by assay size). A fixed-L2 logistic combiner maps four percentiles—public OligoAI, aligned raw regression, capacity-matched bag regression, and LambdaRank—to top-20% membership, with the top-quintile training label taken as the largest $\lceil 0.2n \rceil$ outcomes per assay under stable source-order tie-breaking. Each development assay has equal total weight, with balanced positive and negative mass. Combiner development uses 45 patent families; final evaluation uses 39 different families with no exact or Hamming-1 training neighbour.

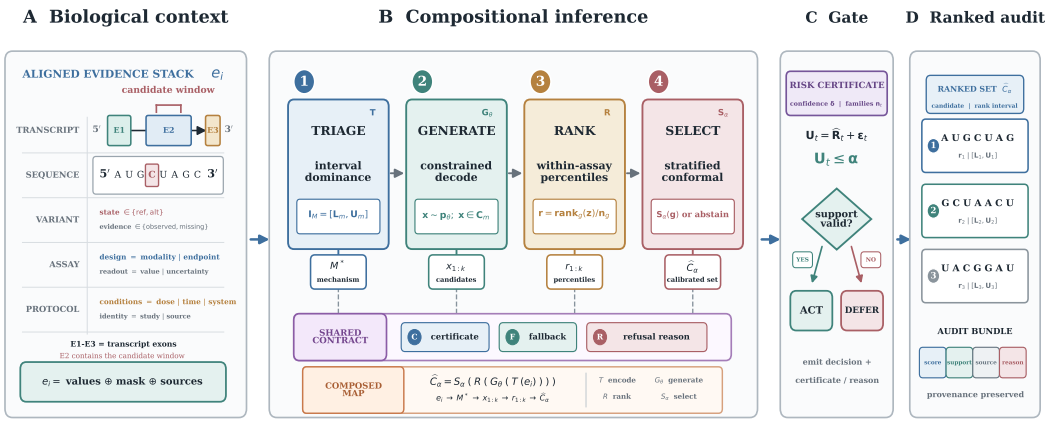


Figure 1: The KRAIG RNA Therapeutics Platform workflow. (A) An aligned evidence stack anchors transcript, sequence, variant, assay, and protocol fields to the same candidate window while retaining missingness and source identifiers. E1–E3 denote exons joined by introns; the bracketed region in E2 maps to the displayed candidate sequence. (B) KRAIG composes mechanism triage, constrained generation, assay-local percentile ranking, and selection; each stage exports a certificate, fallback, or refusal reason. (C) A calibrated support gate decides whether to act or defer. (D) The platform returns ranked oligonucleotides with scores, provenance, and the governing support record.

Proposition 2 (Assay-wise rank invariance) *The stacked ordering is unchanged if any expert's scores within any assay are transformed by an arbitrary strictly increasing function.*

Such a transformation preserves every r_{igj}, including tied scores, and hence the combiner input and output. The construction removes expert- and assay-specific score calibration as a source of protocol shift; it does not remove biological shift.

For a finite grid of agreement thresholds $\mathcal{T}$, let $\widehat{R}_t$ be mean bounded top-quintile loss across the n_t accepted calibration families, conditional on at least one accepted assay. Define

$$U_t = \widehat{R}_t + \sqrt{\frac{\log(|\mathcal{T}|/\delta)}{2n_t}}. \tag{3}$$

Proposition 3 (Development-selected risk certificate) *Fix the predictor and the grid $\mathcal{T}$ independently of the calibration families. Under i.i.d. deployment families and loss in $[0,1]$, with probability at least $1-\delta$, $R_t \le U_t$ simultaneously for every $t \in \mathcal{T}$, where R_t is mean accepted-family loss conditional on at least one accepted assay. Consequently a threshold selected on development by $U_t \le \alpha$ retains the same bound without a second selection penalty.*

This is a selective bounded-loss statement, not a lower bound on EF20 and not an efficacy guarantee. It does not control each family separately, an assay-weighted estimand, EF uplift, or arbitrary shift, and the empty-cell convention cannot manufacture coverage. The reported 38.6% acceptance / 0.597 realized family-loss computation reuses the development families for weight fitting, scoring, and the threshold grid, so the fixed-predictor/fixed-grid independence premises above are not established for it: we report it as an empirical CVaR diagnostic for the simplex model, not a certified bound for the logistic stack or new-study deployment, and the stored field name is preserved as provenance rather than endorsed as a theorem.

3.4 Calibrated selection and protocol fallback

Within modality m, split-conformal top-k selection calibrates a score threshold from exchangeable experiment groups. If the calibration stratum is too small to admit a non-trivial $1-\alpha$ threshold, KRAIG withholds the guarantee instead of pooling incompatible modalities under one label.

For protocol shift, let $g(f)$ be an invariant predictor from sequence features f, $a(f,p)$ a protocol-conditioned residual, and $\lambda(p) \in [0,1]$ a source-only support certificate. SAR-FiLM predicts

$$\hat{y}(f,p) = g(f) + \lambda(p)a(f,p). \tag{4}$$

The certificate uses nearest-source mixed-type protocol distance and observed field support. When protocol support is absent, $\lambda(p) = 0$.

Proposition 4 (Exact invariant fallback) *For every f, p such that $\lambda(p) = 0$, Eq. 4 equals $g(f)$ exactly and is independent of all residual-adapter parameters.*

This construction prevents unsupported adaptation; it does not imply that the residual improves accuracy where the gate is open.

4 Experimental protocol

Each component is evaluated on its own declared endpoint. TRIAGE uses 2,000 seeded pools per synthetic condition with known latent scores. Generation uses 500 decoded candidates per arm and modality. The full ranking grid uses five retained seeds (11, 29, 47, 71, 101), identical rows, a fixed internal split, and eight prespecified learned or derived arms. Seven comparisons against universal training use Holm correction. A preregistered extension adds five new paired seeds only for the primary universal–KRAIG-SR contrast. Bootstrap intervals over genes and studies are descriptive because effective support is only five genes and two studies.

The ASO-TRANSPORT representation stress test aggregates 166,382 examples into 5,161 assays from 86 patent families. Exact- or Hamming-1-connected families share an outer fold. The strict OligoAI comparison uses identical 60,639 rows from 1,795 assays in 39 families absent from OligoAI training folds and without an exact/Hamming-1 training neighbour. The combiner is fitted only on 45 other families. EF20 is primary; macro Spearman and top-20 overlap are secondary. Paired intervals resample patent families 10,000 times. Because these labels informed development, they are separated from the outcome-unseen FANTOM6-iPSC test, whose predictions and exclusions were hash-locked before outcomes were opened. That test uses targets, rather than rows, as independent units.

The generation, calibration, and five-seed ranking artifacts were produced at different stages of development and are never pooled into one statistical sample. The claimed end-to-end property is compositional: a stage can execute only if the previous stage's certificate satisfies its declared capability contract. Test predictions, manifests, checkpoints, splits, and result JSONs are retained for read-only reconstruction.

4.1 Locked evaluation graph

The strict result separates representation fitting, composition, selection, and evaluation by patent family (Table 2). The 39 strict families provide no labels to any fitted component. In particular, agreement is an outcome-free test-time statistic: outcomes are opened only to score the already-ranked assays. Hyperparameter sweeps and leave-one-family-out checks are sensitivity analyses around the frozen prediction table, not routes for selecting a replacement model.

5 Results

Figure 2 summarizes the evidence at the system's decision points. The endpoints remain deliberately separate: constraint validity, within-assay ranking, strict-family retrieval, and selective action are not pooled into one score.

Table 2: **Locked evaluation boundaries.** Each fitted or selected object is separated from the information revealed in the strict-family evaluation.

Step	Object	Fitted or selected on	Strict-family access
1	**Internal rankers**	Non-strict training families	Sequence and context only; no strict outcomes
2	**Rank-stack weights**	45 development families	Expert scores only; no outcome labels
3	**Agreement gate**	Same 45-family development block	Score ranks only; no outcomes
4	**EF20 + uncertainty**	No fitting or selection	Outcomes opened once; patent families resampled

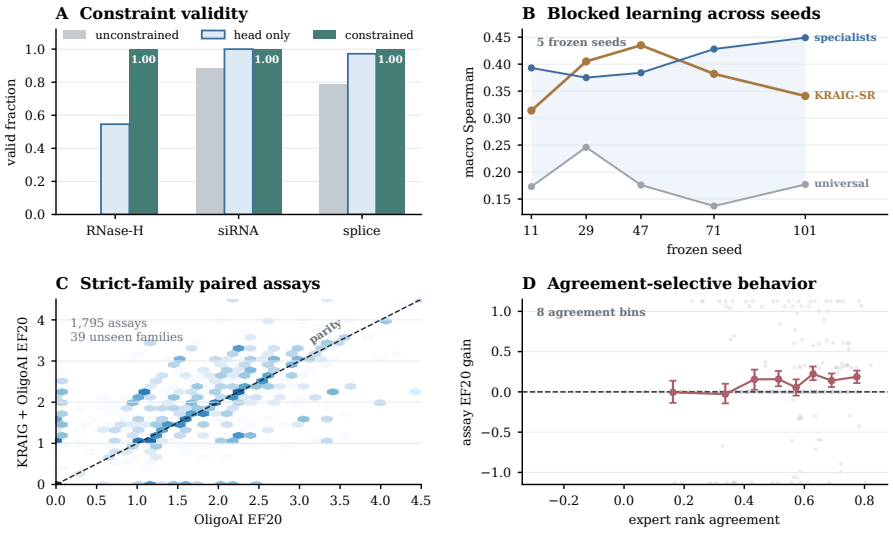


Figure 2: Data-first evidence for four KRAIG contracts. (A) Valid fractions for three decoding policies and modalities. (B) Frozen-seed ranking trajectories for universal, KRAIG-SR, and specialist models. (C) Assay-level strict-family EF20 comparison against matched public OligoAI predictions; density is shown by hexagonal binning. (D) Assay-level EF20 gain versus expert agreement, with quantile-bin means and bootstrap intervals.

5.1 Certification replaces forced precision with an explicit set

Table 3 evaluates the aggregation logic under intervals that contain the latent scores by construction. TRIAGE is correct whenever it selects. In separated pools it selects on every replicate; in near-tie and vacuous pools it returns the unresolved maximal set. Point methods cannot express that distinction and err on 15.7–21.9% of near-tie cases and about 80% of vacuous cases. The full rerun is deterministic apart from runtime.

Table 3: **Controlled TRIAGE evaluation.** Results use 2,000 replicates per condition. Entries are selection error / abstention. Intervals contain the truth by construction.

No.	Method	Separated	Near tie	Vacuous
1	**TRIAGE**	**0.000 / 0.000**	**0.000 / 1.000**	**0.000 / 1.000**
2	Binary / midpoint	0.000 / 0.000	0.159 / 0.000	0.801 / 0.000
3	Product score	0.000 / 0.000	0.157 / 0.000	0.801 / 0.000
4	Upper-bound Borda	0.000 / 0.000	0.219 / 0.000	0.801 / 0.000
5	Pareto	0.000 / 0.000	0.071 / 0.295	0.000 / 1.000

5.2 Hard decoding constraints make validity auditable

Unconstrained chemistry validity ranges from 0% for RNase-H to 88.8% for siRNA (Table 6). Constraint masking raises all three trained modalities to 100% while retaining 100% novelty and high pairwise Hamming diversity. GC steering also removes a pronounced splice-switching GC shift (0.603 to 0.486 against target 0.485).

This establishes deterministic enforcement of the declared design space. Target-conditioned utility is evaluated separately and reported with the scope boundaries in Appendix E.

5.3 Blocked representation learning recovers most specialist performance

Naive universal training reaches 0.182 ± 0.040 macro within-experiment Spearman (Table 7). KRAIG-SR reaches 0.375 ± 0.049, closing 86.4% of the universal-to-specialist gap with one predictor and half the specialist ensemble's parameter count. Its paired improvements over universal training are positive in all five frozen seeds. The original exact sign test is $p = 0.0625$ (Holm $p = 0.4375$). In the preregistered primary-contrast extension, all ten paired differences are positive (mean 0.202, minimum 0.052; two-sided exact sign $p = 0.00195$). This resolves seed-sign stability for one contrast, but adds neither genes nor studies and does not revise inference for the original eight-arm family.

The capability router exactly preserves specialist predictions on this all-supported set while adding an explicit refusal interface for unsupported inputs. We additionally constructed a family- and near-sequence-disjoint ASO-Atlas-2 development benchmark containing 166,382 examples, 5,161 assay groups, and 86 patent families. A stronger raw-outcome learner raises EF20 from 1.446 to 1.515 (paired family-bootstrap difference 0.069, 95% CI [0.027, 0.112]). A capacity-matched control reaches 1.498, so the gain cannot be attributed specifically to target-alignment features.

On the strict subset where public OligoAI out-of-fold predictions and KRAIG predictions can be compared on identical rows, and where the OligoAI training folds contain neither the test patent family nor an exact/Hamming-1 sequence neighbour, we fit a KRAIG stacked expert on 45 separate family-seen development families. It combines OligoAI with aligned-raw, bag-raw, and LambdaRank views, then evaluates once on 39 disjoint strict families. It improves EF20 by 0.110 (paired family-bootstrap 95% CI [0.030, 0.190]; Table 4). KRAIG therefore improves the strongest executable public reference on this declared strict benchmark.

Table 4: Strict matched transport benchmark: 60,639 rows, 1,795 assays, and 39 unseen patent families. Intervals resample patent families.

Model	EF20	Spearman	Top-20 overlap
Public reference			
OligoAI public OOF	1.539	0.232	0.343
KRAIG views without public reference			
KRAIG equal ensemble, no OligoAI	1.561	0.230	0.346
KRAIG learned stack, no OligoAI	1.552	0.229	0.344
Rank composition			
KRAIG + OligoAI, rank NNLS	<u>1.639</u>	**0.270**	<u>0.364</u>
KRAIG + OligoAI, top-20 logistic	**1.649**	<u>0.269</u>	**0.366**

An OligoAI-only logistic transform leaves EF20 at 1.539, while the three KRAIG views alone reach 1.552 (difference 0.013, 95% CI [-0.109, 0.149]); the gain therefore appears when both evidence sources are combined. Development-family comparison selects the linear combiner, and removing any KRAIG view retains EF20 1.626–1.642 with a positive paired interval. Leave-one-family-out gains remain 0.089–0.127; 26/37 non-tied family means are positive (sign $p = 0.020$; conservative sign-flip $p = 0.066$). Across $C \in \{0.01, 0.1, 1, 10, 100\}$, EF20 is 1.643–1.649 with positive intervals. Gains concentrate in medium (+0.173) and large (+0.143) assays, while small assays tie (+0.018, interval crossing zero). Mean pairwise expert-rank agreement further supplies a label-free selective interface: the gain over OligoAI

rises from 0.110 overall to 0.164 at 25% coverage (Table 8). The development-family rule in Eq. 3 selects threshold 0.594 without strict labels; applied unchanged, it accepts 692/1,795 strict assays (38.6%) across 29 families, with realized family-macro bounded top-quintile loss 0.597, reported as an empirical diagnostic rather than a certified deployment bound (see the certificate discussion above).

5.4 Powered multi-target transfer separates signal from improvement

A preregistered developer-blind FANTOM6-iPSC test exceeds the planned 80-unit threshold: 1,330 GapmeRs span 266 independent targets from an external originator. Frozen RiNALMo-micro transfers above chance (EF20 1.410 [1.147, 1.673]; Spearman 0.105 [0.045, 0.167]), but the prespecified consensus does not improve it (ΔEF20 -0.019 [-0.244, 0.207]). This establishes transferable representation signal, not external confirmation of the full stack.

5.5 Calibration, fallback, and artifact audits

The RNase-H conformal pipeline attains 0.92 coverage at 0.90 nominal over 100 groups (Wilson [0.85, 0.96]). KRAIG withholds the same guarantee for siRNA and splice-switching, which lack sufficient modality-specific calibration groups. The SAR-FiLM gate also passes exact response-level fallback tests: closing it reproduces the invariant response and ordering. Exploratory adaptation accuracy is reported in Appendix E. All eight split contracts pass their row/group checks. Test-8 has zero train/test or validation/test family overlap and no exact or equal-length Hamming-1 neighbour. Fresh checkpoint loading reproduces all 75 audited outputs exactly (maximum difference 0.0); details are in Appendix E.

5.6 Deployment-unit calibration exposes the boundary

Five rolling family replays show no monotone temporal deterioration, while the risk gate refuses every future window (Figure 3A). Efficacy and potency have calibration support, whereas hepatic and neurotoxicity guarantees do not (Figure 3B); no compound passes all four bounds. These are refusal results, not efficacy or safety evidence.

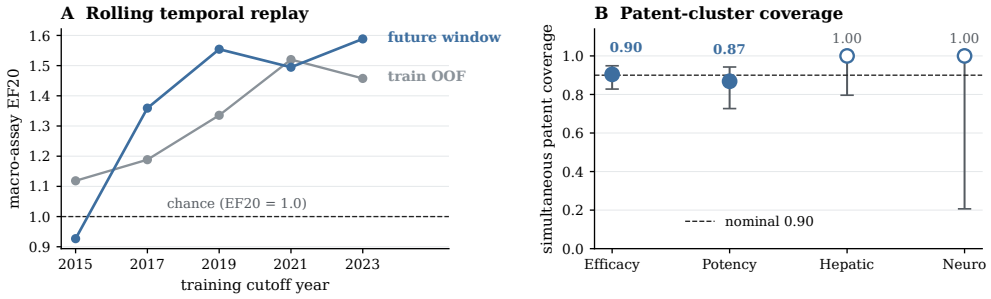


Figure 3: Boundary audits. (A) Five rolling replays, all refused by the risk gate. (B) Patent-cluster coverage with Wilson intervals; filled markers are supported endpoints.

6 Conclusion

KRAIG makes support executable across heterogeneous RNA decisions. Assay-local rank composition improves strict-family EF20 over public OligoAI from 1.539 to 1.649 (paired difference 0.110 [0.030, 0.190]); the surrounding contract preserves sets, scopes calibration, recovers exact fallback, and exposes failed transport and unsupported safety endpoints. A deployment threshold remains to be calibrated on development families before another sealed test; until then the certificate, not the ranker, is the deployable claim. The algorithm is being integrated into the KRAIG RNA Therapeutics Platform, currently under development and testing as research infrastructure.

Reproducibility statement

The supplementary release contains frozen splits, predictions, checkpoints, protocols, result JSONs, evaluation code, and SHA-256 manifests. The appendix reports per-seed results, complete boundary tests, and the mathematical workflow; all principal numerical claims can be reconstructed without retraining.

Ethics statement

This work analyzes retrospective oligonucleotide assay data and controlled simulations. It reports neither human-subject recruitment nor clinical advice, and its outputs must not be interpreted as evidence of therapeutic efficacy, safety, or patient benefit. Provenance and refusal mechanisms are retained to reduce the risk of unsupported biomedical decisions.

AI use statement

Generative AI tools assisted literature discovery, experimental implementation, code review, statistical auditing, figure generation, restructuring, and prose editing. They were not treated as authorities for scientific claims: the authors inspected implementations, checked quantitative statements against stored artifacts, reran deterministic audits, and remain responsible for the manuscript and submission.

References

T. R. Bai, Peng Tang, Yuting Xu, Vladimir Svetnik, Abbas Khalili, Xiang Yu, et al. Conformal selection for efficient and accurate compound screening in drug discovery. *Journal of Chemical Information and Modeling*, 2025. doi: 10.1021/acs.jcim.5c01633.

Y. Bai, H. Zhong, T. Wang, and Z. J. Lu. OligoFormer: an accurate and robust prediction method for sirna design. *Bioinformatics*, 40:btae577, 2024.

Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 794–803, 2018.

S. Chiba, K. R. Q. Lim, N. Sheri, S. Anwar, E. Erkut, M. N. A. Shah, T. Aslesh, S. Woo, O. Sheikh, R. Maruyama, H. Takano, K. Kunitake, W. Duddy, Y. Okuno, Y. Aoki, and T. Yokota. eSkip-Finder: a machine learning-based web application and database to identify the optimal sequences of antisense oligonucleotides for exon skipping. *Nucleic Acids Research*, 49(W1):W193–W198, 2021. doi: 10.1093/nar/gkab442.

Benjamin Hill, Madeleine R. Jaques, Rakesh R. Nair, Nicola Whiffin, Matthew J. A. Wood, Stephan J. Sanders, Peter L. Oliver, Alexander C. Hill, and Carlo Rinaldi. Accurately modelling RNase H-mediated antisense oligonucleotide efficacy. *bioRxiv*, 2025. URL https://github.com/barneyhill/aso_atlas. doi:10.1101/2025.10.29.685292; ASO Atlas + OligoAI model.

Benjamin Hill, Nicola Whiffin, Carlo Rinaldi, and Stephan J. Sanders. ASO Atlas 2.0: Evaluating antisense oligonucleotide prediction across the preclinical pipeline. Hugging Face dataset, 2026. URL https://huggingface.co/datasets/barneyhill/aso-atlas-2. Version 2.0, DOI:10.57967/hf/8687; 295,007 assay readouts from 606 USPTO patent filings.

D. Huesken et al. Design of a genome-wide sirna library using an artificial neural network. *Nature Biotechnology*, 23:995–1001, 2005.

Eyke Hüllermeier, Weiwei Cheng, Volkmar Welker, and Willem Waegeman. Label ranking with partial abstention based on thresholded probabilistic models. In *Advances in Neural Information Processing Systems 26 (NeurIPS)*, 2013.

G. Hwang et al. ASOptimizer: Optimizing antisense oligonucleotides through deep learning for IDO1 gene regulation. *Molecular Therapy: Nucleic Acids*, 35:102239, 2024.

S. Hwang et al. ASO-RASAR: a read-across framework for predicting antisense oligonucleotide gapmer activity across target genes. *Journal of Chemical Information and Modeling*, 2026. doi: 10.1021/acs.jcim.6c01314.

N. Iwano, T. Adachi, K. Aoki, Y. Nakamura, and M. Hamada. Generative aptamer discovery using RaptGen. *Nature Computational Science*, 2:378–386, 2022.

Wooseong Jeong and Kuk-Jin Yoon. Selective task group updates for multi-task optimization. In *International Conference on Learning Representations*, 2025.

Ying Jin and Emmanuel Candès. Selection by prediction with conformal p-values. *Journal of Machine Learning Research*, 24:22–1176, 2023.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

Diederik P. Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems 29 (NeurIPS)*, 2016.

Giwoong Lee, Eunho Yang, and Sung Ju Hwang. Asymmetric multi-task learning based on task relatedness and loss. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 230–238, 2016. URL https://proceedings.mlr.press/v48/leeb16.html.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18878–18890, 2021.

Shuyu Liu, Jin Zhuo, Shuchang Lei, Tianhao Wu, Jiaxuan Han, Chaoyi Wu, Yanfeng Wang, and Weidi Xie. ASOCompass: Context- and chemistry-aware activity prediction for transferable antisense oligonucleotide screening, 2026. URL https://www.biorxiv.org/content/10.64898/2026.08.03.742461v1. bioRxiv doi:10.64898/2026.08.03.742461; context-aware ASO ranker, overall Spearman 0.5970 on ASO Atlas splits; cited as ceiling context, not evaluated here.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Ranking with abstention. *arXiv preprint*, 2023. arXiv:2307.02035; H-consistency bounds for pairwise/bipartite ranking with abstention.

A. Reynolds et al. Rational sirna design for rna interference. *Nature Biotechnology*, 22: 326–330, 2004.

Frederic Runge, Danny Stoll, Stefan Falkner, and Frank Hutter. Learning to design RNA. In *International Conference on Learning Representations (ICLR)*, 2019.

S. Sciabola et al. PFRED: A computational platform for sirna and antisense oligonucleotides design. *PLoS ONE*, 16:e0238753, 2021.

Dhanasekar Sundararaman, Henry Tsai, Kuang-Huei Lee, Iulia Turc, and Lawrence Carin. Learning task sampling policy for multitask learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021.

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 *International Conference on Learning Representations (ICLR)*, 2024.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5824–5836, 2020.

Qian Zhang, Lucia Yang, Ying Hsiu Liu, John E. Wilkinson, and Adrian R. Krainer. Antisense oligonucleotide therapy for H3.3K27M diffuse midline glioma. *Science Translational Medicine*, 15(691):eadd8280, 2023. doi: 10.1126/scitranslmed.add8280.

Michael V. Zuccaro et al. Antisense oligonucleotides to KIF1A polymorphisms expand targets and rescue patient-derived neurons in vitro. *Nature Communications*, 17:1109, 2026. doi: 10.1038/s41467-025-67752-y.

APPENDIX

A Platform workflow 15
A.1 Platform module interfaces 15
B Proofs of certificate properties 15
C Detailed contract results 16
C.1 Generation validity 16
C.2 Activity-ranking comparison 16
C.3 Selective strict-family transport 16
C.4 Public representation matrix 18
C.5 Modality-asymmetric transfer frontier 20
C.6 Stack ablation: both sources needed 20
C.7 Regularization and leave-one-family sensitivity 20
D Per-seed stability 21
D.1 Frozen five-seed results 21
D.2 Preregistered ten-seed extension 22
E Boundary and stress-test results 22
E.1 Transport passport 22
E.2 External stress studies 23
E.3 Corrected multi-endpoint and temporal audits 24
F Mechanism catalogue 24
F.1 Grouped identifier table with evidence tiers 24

Table 5: Notation index for Algorithm 1, contracts, and evaluation.

Symbol	Meaning	See
Evidence and mechanism certification		
E	biological evidence record (literature, databases, assays)	Alg. 1, App. A.1
D	design request: target sequence, chemistry class c, length range	Alg. 1, App. A.1
k, M	mechanism index; number of candidate mechanisms	App. A.1
s_f	typed state of feature f: PRESENT / ABSENT / UNRESOLVED / N/A	Methods
q (p_f)	marginal support value in $[0, 1]$	Methods
T_k	support set of mechanism k (required, discriminating, complemented forbidden)	Methods
L_k, U_k	Fréchet conjunction bounds for mechanism k	Methods
$\mathcal{M}$	maximal-element mechanism set	Alg. 1
$m^\star$	certified mechanism singleton	Alg. 1
Constrained generation		
X	candidate sequence set	Alg. 1
$V(x, m^\star, D)$	validity predicate (1 iff all declared constraints hold)	Alg. 1
X_{valid}	chemistry-valid candidates	Fig. 4

Table 5: Notation index (continued).

Symbol	Meaning	See
Assay-local ranking and stack		
A	assay context (modality, cell line, protocol)	Alg. 1
s_j	expert predictor j	Alg. 1
z_{igj}	score of candidate i in assay g by expert j	Methods
r_{igj}	within-assay empirical percentile of z_{igj}	Methods
n_g	assay size (candidates in assay g)	Methods
J	number of experts	Methods
a_g	mean pairwise expert rank agreement in assay g	Methods
τ	agreement threshold from development quantiles	Results
y_{ig}	top-quintile activity indicator	Methods
w_{ig}	assay/class-balanced training weight	Methods
β, b	stack weights and intercept	Methods
σ	logistic function	Methods
u_i	combined candidate score	Alg. 1
C	ridge strength ($C = 1$ primary)	Results
Conformal selection and selective risk		
S (S_m)	selected candidate set (modality m)	Alg. 1
$\widehat{q}_m$	modality-m conformal threshold	Fig. 4
$n_{\min}$	minimum calibration stratum size	App. A.1
R_t, $\widehat{R}_t$, U_t	selective risk, its estimate, and upper bound at threshold t	Methods
n_t	families contributing at threshold t	Methods
δ, α	confidence level; risk target	Methods
Protocol adaptation		
f	sequence features	Methods
$g(f)$	protocol-invariant predictor	Methods
$a(f, p)$	protocol-conditioned residual	Methods
$\lambda(P)$	source-only support certificate in $[0, 1]$	Methods
$\widehat{y}$	adapted prediction	Methods
Metrics and evaluation		
EF20	assay-local enrichment at top 20% (chance 1.0)	Results
ρ	within-assay Spearman correlation	Results
top-20 overlap	bounded retrieval-loss basis ($1 -$ overlap)	Results

Dataset preparation. Source rows are alphabet-filtered with T→U normalization; each row receives a within-assay percentile rank. ASO-TRANSPORT aggregates 166,382 rows into 5,161 assays from 86 patent families, with exact- or Hamming-1-connected families sharing an outer fold (45 development / 39 strict families). Splits, predictions, and manifests are frozen with SHA-256 audit.

Frozen splits. The five-seed activity diagnostic uses 3,837/757/2,589 train/validation/test rows (siRNA plus RNase-H) with identical rows and a fixed internal split across seeds 11, 29, 47, 71, and 101. The combiner is fitted on 45 development families (95,892 rows, 3,217 assays) and evaluated once on 39 disjoint strict families (60,639 rows, 1,795 assays); Test-8 has no train/test family overlap and no exact or Hamming-1 neighbour.

A Platform workflow

Algorithm 1 Support-aware workflow in the KRAIG RNA Therapeutics Platform

Require: Evidence E, design request D, assay A, protocol P
$\mathcal{M} \leftarrow \text{Maximal}\big(\{[L_k(E), U_k(E)]\}_{k=1}^{M}\big)$
if $|\mathcal{M}| \neq 1$ **then return** $(\mathcal{M}, \textsc{Refuse})$
end if
$m^\star \leftarrow \text{argmax}_{k \in \mathcal{M}} L_k(E)$, $X \leftarrow \text{Generate}(m^\star, D)$
if $\min_{x \in X} V(x, m^\star, D) = 0$ **then return** $(X, \textsc{Refuse})$
end if
$z_{ij} \leftarrow s_j(x_i, A)$, $r_{ij} \leftarrow (N_<(i,j) + (N_=(i,j)+1)/2)/|A|$ ▷ average-rank percentile
$a(A) \leftarrow \text{mean}_{j<j'} \rho(r_{\cdot j}, r_{\cdot j'})$
if $a(A) < \tau$ **then return** $(\text{Fallback}(A), \textsc{Refuse})$
end if
$u_i \leftarrow \sigma(\beta_0 + \sum_j \beta_j r_{ij})$, $S \leftarrow \text{ConformalSelect}(\{u_i\}, m^\star)$
$\widehat{y}_i \leftarrow q(x_i) + \lambda(P) c(x_i, P)$, where $\lambda(P) = 0$ outside protocol support
return $(S, \{\widehat{y}_i\}, \textsc{Certificates}, \textsc{Provenance})$

A.1 Platform module interfaces

The KRAIG RNA Therapeutics Platform separates concerns into four modules that communicate through typed certificates. This section documents each module's inputs, outputs, and certificate schema for reproducibility.

TRIAGE (mechanism assessment). Input: biological evidence record E (literature, databases, assays). Output: mechanism interval set $\{[L_k, U_k]\}_{k=1}^{M}$ and certificate $\mathcal{C}_{\text{TRIAGE}} = (\mathcal{M}, \text{provenance})$ where $\mathcal{M} = \{k : L_k = \max_j L_j \text{ and } L_k > U_j \ \forall j \neq k\}$ is the maximal element set. If $|\mathcal{M}| \neq 1$, certificate carries Refuse with reason `ambiguous_evidence`. The interval construction uses the sharp Fréchet bounds from Proposition 1.

GENERATE (mechanism-conditioned design). Input: selected mechanism $m^\star$, design request D (target sequence, chemistry class c, length range). Output: candidate set X and certificate $\mathcal{C}_{\text{GEN}} = (V, \text{provenance})$ where $V(x, m^\star, D) = 1$ iff x satisfies all declared constraints for $m^\star$ and c. If $\min_{x \in X} V = 0$, certificate carries Refuse with reason `invalid_chemistry`. The generator is a conditional VAE with grammar-constrained decoding (Appendix C.1).

RANK (assay-local scoring). Input: candidate sequences X, assay context A (modality, cell line, protocol). Output: within-assay ranks r_{ij} and certificate $\mathcal{C}_{\text{RANK}} = (\text{capability}, \text{support})$ where capability declares which expert predictors are admissible for A, and support is the agreement statistic $a(A) = \text{mean}_{j<j'} \rho(r_{\cdot j}, r_{\cdot j'})$. If $a(A) < \tau$, certificate carries Refuse with reason `insufficient_agreement`.

SELECT (conformal candidate selection). Input: ranked candidates with scores u_i, mechanism $m^\star$. Output: selected set S and certificate $\mathcal{C}_{\text{SEL}} = (\text{coverage}, \text{stratum})$ where stratum identifies the modality-specific calibration group. If the stratum has $n < n_{\min}$ calibration samples, certificate carries Refuse with reason `insufficient_support`.

ADAPT (protocol adaptation). Input: selected candidates S, protocol shift P. Output: adapted predictions $\widehat{y}_i$ and certificate $\mathcal{C}_{\text{ADAPT}} = (\lambda(P), \text{provenance})$ where $\lambda(P) = 0$ outside the declared protocol support region. The fallback exactly recovers the invariant predictor $q(x_i)$.

These certificates form a chain: each downstream module verifies the upstream certificate before executing. The full provenance trace from evidence E to final output $(S, \{\widehat{y}_i\})$ is retained for audit.

B Proofs of certificate properties

Sharp mechanism certificate. For events with marginals $q_1, \ldots, q_d$, the union bound applied to their complements gives $P(\cap_i A_i) \geq \max(0, \sum_i q_i - (d-1))$, while monotonicity gives $P(\cap_i A_i) \leq \min_i q_i$. Classical extremal couplings attain both bounds, so neither endpoint can be tightened

from marginals alone. If $L_k > U_j$ for every $j \neq k$, interval coverage implies the latent score of k exceeds every competitor. When intervals overlap, KRAIG retains the maximal set and makes no point claim.

Development-selected risk certificate. Fix the predictor and the grid $\mathcal{T}$ independently of the calibration families. For fixed t, Hoeffding's inequality gives $P(R_t > \widehat{R}_t + \epsilon) \leq \exp(-2n_t\epsilon^2)$, where R_t is mean accepted-family loss conditional on at least one accepted assay. Setting $\epsilon_t = \sqrt{\log(|\mathcal{T}|/\delta)/(2n_t)}$ and taking a union bound over $\mathcal{T}$ yields simultaneous coverage at least $1-\delta$. Because the event holds for every threshold, choosing among thresholds after observing calibration losses incurs no additional selection penalty. The deployment statement additionally requires i.i.d. families from the deployment distribution and the same loss and acceptance rule; it does not cover the computation in which weights, scores, and the grid are all derived from the same families.

C Detailed contract tables

C.1 Generation validity

Table 6: Generation validity over 500 candidates per arm. "Valid" is the declared chemistry/length contract, not target activity.

Modality	Unconstrained	Head only	Constrained	Novelty
RNase-H	0.000	0.546	**1.000**	**1.000**
siRNA	0.888	**1.000**	**1.000**	**1.000**
Splice-switching	0.792	0.972	**1.000**	**1.000**

C.2 Activity-ranking comparison

Table 7: Frozen five-seed activity-ranking diagnostic. Scores are macro within-experiment Spearman (mean ± SD). The oracle uses test outcomes.

Arm	Overall	siRNA	RNase-H
SHARED PREDICTORS			
Universal same-harness	0.182 ± 0.040	0.210	0.112
Parameter-matched joint	0.228 ± 0.028	0.232	0.217
Shared modality heads	0.347 ± 0.043	0.414	0.179
KRAIG-SR shared (ours)	$\mathbf{0.375} \pm 0.049$	<u>0.437</u>	<u>0.221</u>
ROUTED UPPER BOUNDS			
Separate specialists	0.406 ± 0.032	0.496	0.180
Capability router	0.406 ± 0.032	0.496	0.180
Oracle router	0.422 ± 0.028	0.498	0.231

C.3 Selective strict-family transport

Table 8: Strict-family selection by expert agreement. Intervals resample patent families.

Coverage	Assays	KRAIG	OligoAI	Difference [95% CI]
25%	449	**2.087**	1.923	**0.164** [0.036, 0.266]
50%	898	1.881	1.730	0.151 [0.069, 0.228]
75%	1,347	1.767	1.613	0.153 [0.077, 0.227]
100%	1,795	1.649	1.539	0.110 [0.030, 0.190]

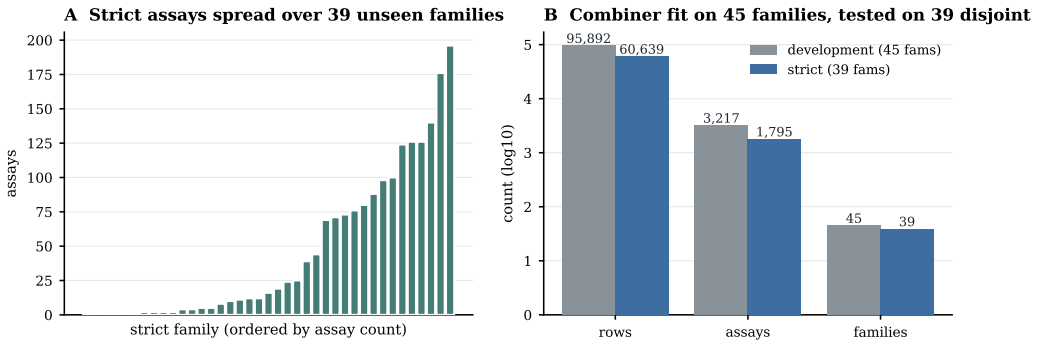


Figure 4: Strict-transport benchmark composition. (A) Assay counts across the 39 unseen patent families. (B) Development versus strict scale: the combiner is fit on 45 families (95,892 rows, 3,217 assays) and evaluated once on 39 disjoint families (60,639 rows, 1,795 assays).

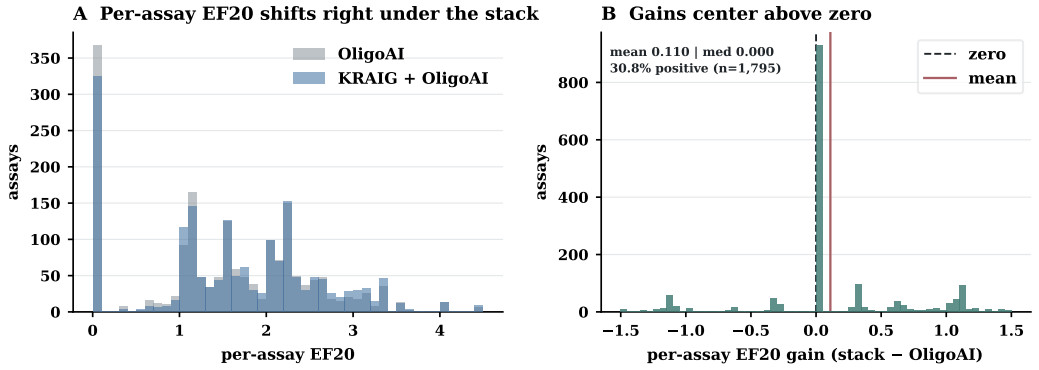


Figure 5: Assay-level strict-family evidence (1,795 assays). (A) Overlaid per-assay EF20 distributions shift right under the stack. (B) Paired gains center above zero (mean 0.110); the spike at zero reflects assays tied at identical EF20.

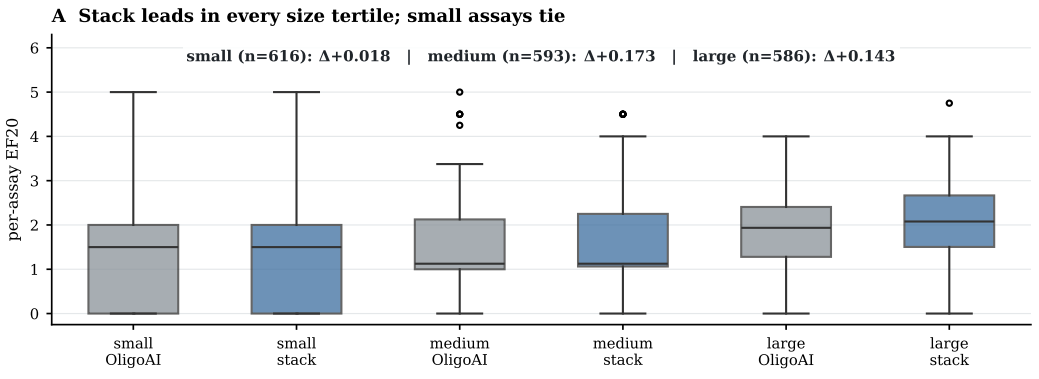


Figure 6: Per-assay EF20 by assay-size tertile for OligoAI (gray) versus the stack (blue). The stack median leads in every tertile; small assays tie, consistent with the paired stratum difference of 0.018 $[-0.051, 0.158]$.

Table 9: Prespecified robustness strata on the same retrospective strict test. Intervals resample patent families. Small assays and pre-2016 families are too thin to resolve a difference; medium/large and 2016-onward strata stay positive.

Stratum	Assays (families)	Stack − OligoAI	95% CI
Large assays	586 (33)	0.143	[0.074, 0.203]
Medium assays	593 (24)	0.173	[0.039, 0.290]
Small assays	616 (28)	0.018	[−0.051, 0.158]
Published 2016 onward	1,736 (26)	0.114	[0.034, 0.200]
Published through 2015	59 (13)	−0.016	[−0.209, 0.198]

C.4 Public representation matrix

Table 10: Frozen public representation matrix on 166,382 ASO-Atlas-2 development rows, 5,161 assay groups, 86 patent families, and five family/Hamming-1 folds. All arms use the same assay-local percentile-rank target, training-fold-only standardization, ridge probe ($\alpha = 10$), and family-clustered bootstrap. No tested representation has a paired EF20 interval strictly above the nucleotide 1/2/3-mer reference; thermodynamics is resolved worse.

Frozen representation	Dim.	EF20 [95% CI]	Spearman	Top-20	Paired vs k-mer
k-mer 1/2/3-mers	135	1.351 [1.281, 1.424]	0.168	0.304	reference
k-mer + chemistry	142	1.360 [1.292, 1.434]	0.169	0.306	+0.009 [-0.003, +0.024]
Nucleotide one-hot	126	1.337 [1.281, 1.407]	0.149	0.301	-0.014 [-0.060, +0.034]
Full-chemistry one-hot	273	1.323 [1.262, 1.387]	0.148	0.297	-0.029 [-0.077, +0.026]
Thermodynamics	23	1.175 [1.128, 1.224]	0.081	0.265	-0.177 [-0.260, -0.096]
RNA-FM L12 mean	640	1.339 [1.288, 1.394]	0.174	0.301	-0.013 [-0.060, +0.034]
RiNALMo-micro mean	480	1.387 [1.319, 1.461]	0.175	0.312	+0.036 [-0.023, +0.095]

Coverage is 1.000 for every arm. The reading is therefore not that one public foundation model wins: generic RNA pretraining, positional one-hot encoding, chemistry counts, and thermodynamic descriptors do not by themselves produce a resolved improvement under strict family and near-sequence shift. The paper's significant strict-transport gain appears only when complementary KRAIG views and OligoAI are combined, not from a single frozen representation. This is retrospective public development evidence under the frozen probe and split contract, not an untouched external test and not prospective validation.

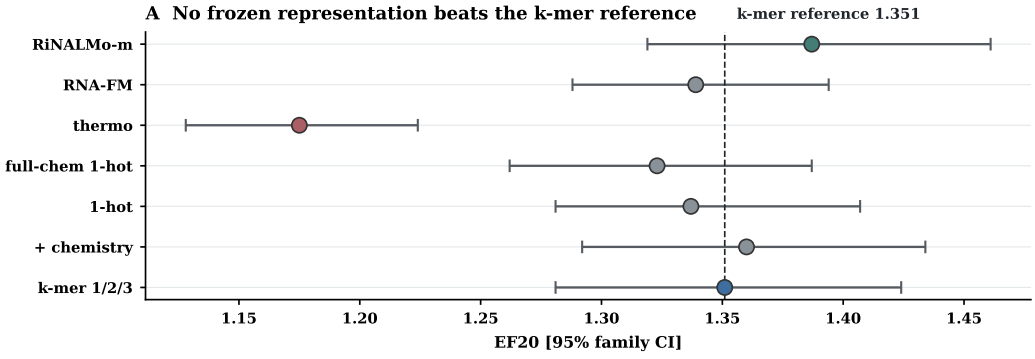


Figure 8: Frozen representation matrix as a forest plot. Points are EF20 means with 95% family-clustered intervals; the dashed line is the k-mer reference (1.351). Only thermodynamics is resolved from the reference, and it is worse; RiNALMo-micro is the highest point estimate but its interval crosses zero.

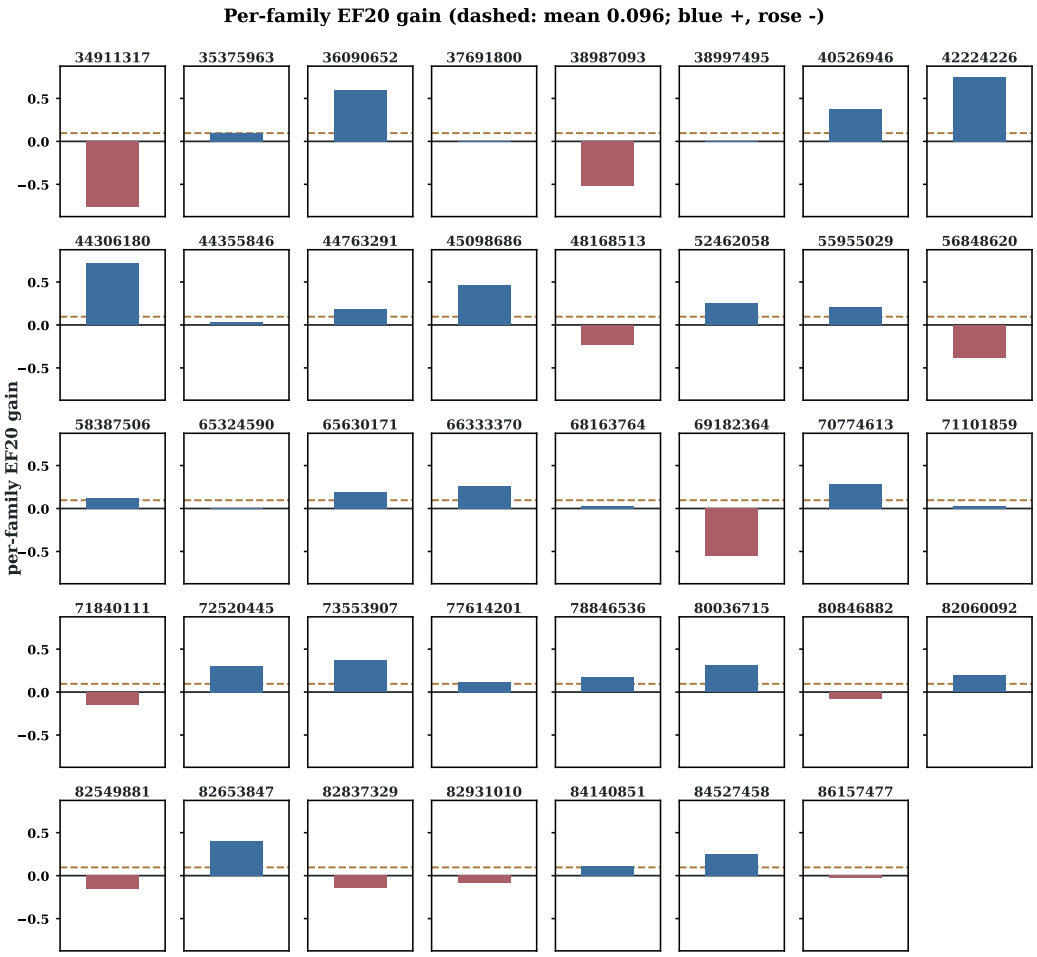


Figure 7: Per-family EF20 gain (39 families; dashed: mean 0.096).

C.5 Modality-asymmetric transfer frontier

This supports the descriptive claim of a modality-asymmetric transfer profile only: RNase-H (74 rows, 2 experiments) is positive in all five seeds (mean +0.0418, seed SD 0.0283) while siRNA (2,515 rows, 5 experiments) is positive in two of five (mean −0.0595, seed SD 0.0983), for a reconstructed overall universal-to-specialist gap closure of 86.4%. Per-seed detail is plotted in Figure 10B. It does not support general superiority: RNase-H has only two estimable experiments, siRNA only five under the declared grouped metric, and the all-positive five-seed pattern has minimum attainable unadjusted two-sided sign-test $p = 0.0625$.

C.6 Stack ablation: both sources needed

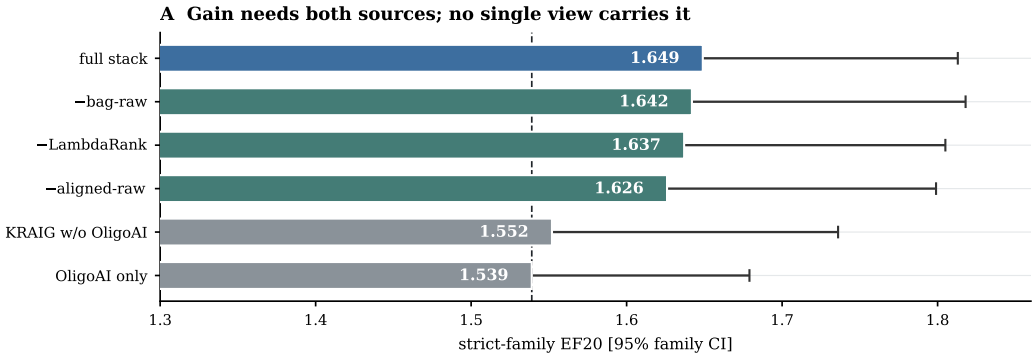


Figure 9: Fixed-ablation strict-family EF20 with 95% family intervals. Neither OligoAI alone (1.539) nor the three KRAIG views without OligoAI (1.552) explains the full-stack gain (1.649); removing any single KRAIG view leaves EF20 at 1.626–1.642 with a positive paired interval, indicating redundant complementary signal rather than reliance on one added view.

The ablation is explanatory on the same retrospective benchmark, with no hyperparameter or subset selection from strict-test outcomes: the gain appears only when the two evidence sources are combined.

C.7 Regularization and leave-one-family sensitivity

Table 11: Prespecified combiner sensitivity: refitting the identical top-20 logistic stack on development families across C leaves strict-family EF20 at 1.643–1.649 with every paired interval above zero; omitting each strict family in turn leaves paired improvements between 0.089 and 0.127.

Setting	EF20	Stack − OligoAI	95% CI
$C = 0.01$	1.647	0.109	[0.02, 0.21]
$C = 0.1$	1.643	0.105	[0.02, 0.19]
$C = 1$ (primary)	1.649	0.110	[0.03, 0.19]
$C = 10$	1.647	0.109	[0.03, 0.19]
$C = 100$	1.649	0.110	[0.03, 0.19]
Leave-one-family-out	—	0.089–0.127	—

C.8 Expert predictors and view construction

The KRAIG stack combines four expert views, each producing per-assay rank predictions that are then combined by the top-20 logistic combiner. This section documents each view's construction for reproducibility.

Aligned-raw view. Input: raw assay measurements aligned to a common sequence coordinate system. For each assay A, sequences are mapped to their genomic coordinates and raw measurements (e.g., IC50, read counts) are standardized within assay using robust z-scoring (median/MAD). The view predicts the standardized value directly. No learned representation; serves as a calibrated baseline.

Bag-raw view. Input: same aligned-raw features as above, but aggregated per sequence across assays using a weighted mean where weights are proportional to $\sqrt{n_A}$ (assay sample size). This produces a single per-sequence score that captures cross-assay consensus. Prediction is the bag-aggregated value.

LambdaRank view. Input: sequence embeddings from a frozen nucleotide transformer (RiNALMo-micro, 480-dim mean-pooled last layer). A LambdaRank objective (**?**) optimizes NDCG directly on development families. The model is a 3-layer MLP (480-256-128-1) with ReLU activations, trained with Adam (lr $= 10^{-4}$, weight decay 10^{-5}) for 200 epochs with early stopping on family-held-out validation. Output is a per-sequence utility score converted to within-assay ranks.

OligoAI view. Input: public out-of-fold predictions from the OligoAI model (Hill et al., 2025) on the exact Test-8 sequences. No retraining; used as-is. Provides an external reference ranker trained on independent data. Mapped to within-assay ranks using the same assay-local percentile transformation as other views.

Combiner. The top-20 logistic stack (**?**) takes the four view ranks as features and predicts top-20 membership. Trained on development families (45 families) with ridge regularization ($C = 1$), evaluated once on strict families (39 families). This is the only learned component with access to strict-test family identities during development (via the family bootstrap); all views are frozen before combiner training.

D Per-seed ranking results

D.1 Frozen five-seed results

Table 12: Macro within-experiment Spearman for every retained seed.

Arm	11	29	47	71	101
SHARED PREDICTORS					
Universal	.173	.246	.176	.137	.177
Parameter-matched joint	.274	.232	.205	.212	.216
Shared heads	.308	.403	.324	.317	.385
KRAIG-SR	.314	.405	.435	.382	.341
ROUTED UPPER BOUNDS					
Specialists	.393	.375	.384	.428	.449
Capability router	.393	.375	.384	.428	.449
Oracle router	.397	.401	.412	.433	.465

D.2 Preregistered ten-seed extension

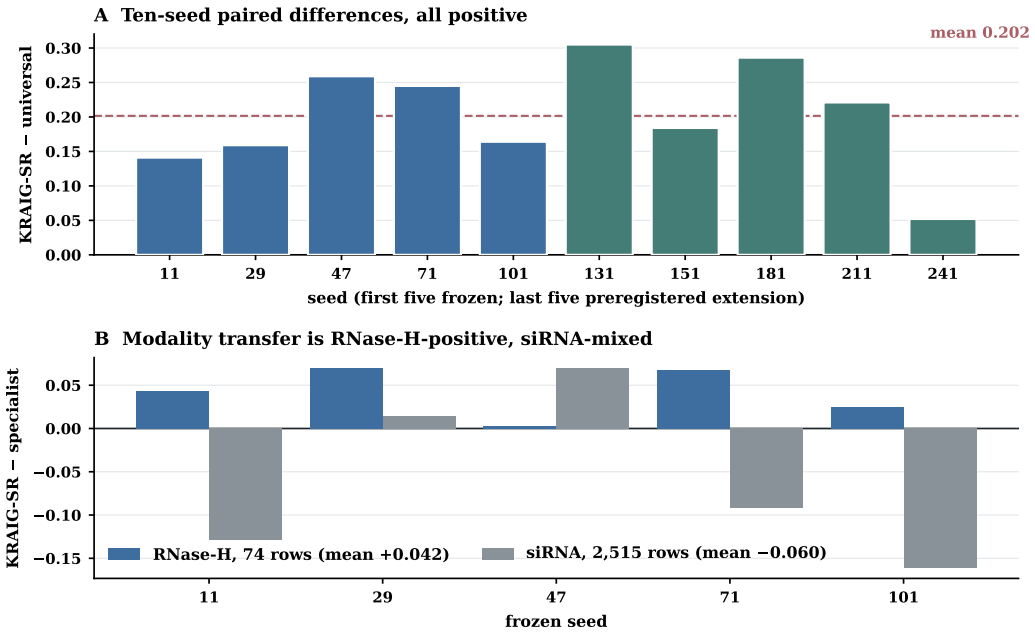


Figure 10: Seed-stability diagnostics from frozen artifacts. (A) Preregistered ten-seed extension for the primary contrast only: blue is the original frozen five seeds (11–101), teal the extension (131–241); all ten differences are positive (mean 0.202, minimum 0.052; two-sided exact sign $p = 0.00195$). (B) Per-seed modality transfer: RNase-H is positive in all five seeds while siRNA is mixed, so seed-sign stability holds for the primary contrast while transfer stays modality-asymmetric.

The extension also tested a proposed gradient-conflict explanation. It did not survive measurement: mean RNase-H/siRNA gradient cosine is 0.508 for mixed training and 0.532 for blocked training, with negative fractions 0.060 and 0.000. KRAIG-SR is therefore claimed as a robust scheduling effect in this harness, not as a new optimizer or a verified gradient-interference mechanism.

E Boundary and stress-test results

E.1 Transport passport

Table 13: Transport passport. Full-stack claim authority stops at T2; the T4 pass applies specifically to representation transfer across 266 targets.

Level	Requirement	Status
T0	Same-distribution reconstruction	**Pass**
T1	Group-disjoint evaluation	**Pass**
T2	Family- and Hamming-1-disjoint transport	**Pass**
T3	Cross-study transport	Mixed/negative
T4	Outcome-unseen frozen evaluation	**Pass: 266 targets**
T5	Powered outcome-unseen confirmation	Pending

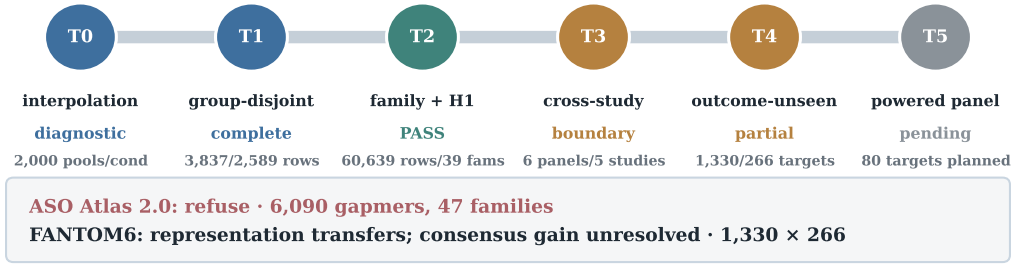


Figure 11: Transport passport timeline with per-level scale. Authority stops at T2; T4 is representation transfer only.

E.2 External stress studies

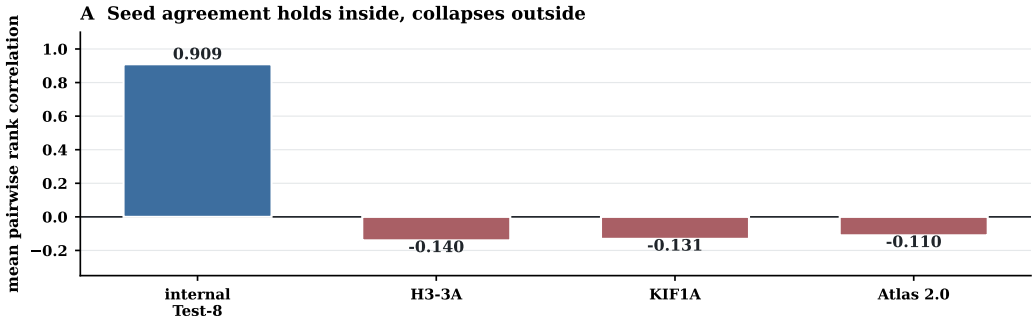


Figure 12: Cross-seed rank correlation: 0.909 on internal Test-8 vs −0.140 (H3-3A), −0.131 (KIF1A), −0.110 (Atlas 2.0) — a preregistration hypothesis.

The main text emphasizes results that directly support KRAIG's declared contracts. For completeness, this section records evaluations that bound those claims. The target-conditioned utility audit found no held-out active-candidate recovery for the learned generators; constrained decoding therefore supports chemistry and format validity, not target efficacy. Nested leave-one-study-out evaluation of SAR-FiLM did not establish an adaptation-accuracy improvement; the verified result is exact recovery of the invariant predictor when the gate closes.

Two post-development public stress studies were eligible under the frozen screening contract. On two complete 15-ASO H3-3A panels (Zhang et al., 2023), KRAIG-SR Spearman correlations are −0.500 and −0.579; GC fraction yields 0.736 and 0.564. A correlation of this sign and magnitude indicates the ranking is anti-correlated with true activity—below chance, not merely uninformative. A KIF1A study presents the same six ASOs in three patient-derived cell lines (Zuccaro et al., 2026): universal-model correlations are 0.657–0.943, whereas KRAIG-SR correlations are −0.257–−0.086. A correlation of this sign and magnitude indicates the ranking is anti-correlated with true activity—below chance, not merely uninformative. The three cell-line panels constitute one independent sequence set. These outcomes do not provide external confirmation. These diagnostic evaluations bypass the deployment gate by design, to characterize the failure modes such a gate must catch: the frozen ranker was scored directly with no agreement threshold in place, so no refusal was issued. The observed agreement collapse (−0.140 on H3-3A and −0.131 on KIF1A, versus 0.909 internally) is the post-hoc basis for preregistering one. Cross-seed agreement suggests a concrete prospective support diagnostic. KRAIG-SR's mean pairwise rank correlation across seeds is 0.909 on internal Test-8 RNase-H predictions, −0.140 on H3-3A, and −0.131 on KIF1A. This statistic was examined after observing the outcomes and has no frozen decision threshold; it is therefore a hypothesis to preregister rather than a validated abstention rule.

We next froze an inference-only contract before downloading ASO Atlas 2.0 outcomes (Hill et al., 2026). After excluding every KRAIG-v1 publication family and every exact or equal-length Hamming-1 sequence match, 6,090 gapmers remain across 100 tables, 56 publications, and 47 unseen simple families. Macro within-table Spearman is −0.092 [95% patent-cluster CI −0.141, −0.042] for KRAIG-SR, 0.052 [0.005, 0.111] for the universal model, and 0.159 [0.067, 0.225] for GC fraction.

KRAIG-SR's seed agreement is −0.110, so even a sign-based concordance sanity check refuses its ranking. This large extension audit supports the need for the certificate; it does not validate the frozen ranker. A deployment threshold remains to be calibrated on development families before another sealed test.

As a positive learnability control, we separately audited the five public OligoAI OOF prediction files rather than treating KRAIG's failure as a property of the endpoint. Of 91 mapped simple families, 46 span published folds, but the seen-family EF20 advantage is not resolved (difference 0.078, 95% family- bootstrap CI $[-0.084, 0.239]$). More importantly, the strict subset whose families and equal-length Hamming-1 sequence neighbours are absent from training retains EF20 1.539 $[1.417, 1.680]$ and macro Spearman 0.232 $[0.184, 0.286]$ over 1,806 assay groups from 39 families. Thus strict transport is learnable by a strong public baseline, and future KRAIG rankers must compare against it under identical folds and inputs rather than against GC alone (Hill et al., 2025). We first preregistered and hash-locked predictions for the FANTOM6-iPSC panel before opening its outcome workbook. After outcome-free filtering and exclusion of three schema-preview targets, the one-shot join retained 1,330 LNA GapmeRs across 266 targets. Frozen RiNALMo-micro achieved EF20 1.410 [1.147, 1.673] and mean within-target Spearman 0.105 [0.045, 0.167]. The prespecified consensus did not improve on it: paired ΔEF20 −0.019 [−0.244, 0.207] and ΔSpearman −0.023 [−0.058, 0.010]. This powered, developer-blind, cross-originator result establishes transferable representation signal, but not full-stack superiority; all targets also share one originator and protocol. We also preregistered a cross-originator check before opening the public ASO-RASAR activity tables (Hwang et al., 2026). Sequence auditing showed that 58,908 of 59,273 rows are exact or Hamming-1 matches to ASO Atlas, leaving only 322 eligible rows across six targets. That remainder is too small to serve as the intended independent confirmation, so it is retained as an overlap audit rather than promoted as validation.

E.3 Corrected multi-endpoint and temporal audits

The initial multi-endpoint implementation was retained as an immutable protocol-deviation audit. Version 2 uses one global patent hash split across endpoints, removes exact HELM overlap across train/calibration/test, fits an endpoint predictor, and calibrates the worst residual per patent. Efficacy attains patent-simultaneous coverage 0.904 [0.828, 0.949] over 94 test patents; potency attains 0.868 [0.727, 0.942] over 38. Hepatic severity has only 18 calibration patents and a correction of 290.5 robust-scale units; neurotoxicity has seven calibration patents and one test patent. Neither can support the full conjunction. The assay-set learner was also stopped by its frozen development rule: Deep Sets improves EF20 over the fixed logistic stack by only 0.0019 [−0.0276, 0.0296], and the Set Transformer is lower. No strict-set run was performed after that failed promotion decision.

At selective-risk target 0.90, the support/agreement rule certifies act-all; at the stricter exploratory target 0.75, all 19 prespecified metric/mode/bound variants refuse all and recover OligoAI exactly. These experiments verify the fallback implementation but reveal no certified partial-action regime. In the five rolling temporal windows, future EF20 ranges from 0.927 to 1.588 and the same finite-family gate refuses every window. We therefore do not claim that the gate detects temporal deterioration.

Finally, patent-family isolation is specific to the definitive Test-8 split. The chemistry-held-out Test-11 split shares 18 train/test families, while the Test-22 sequence-cluster split shares all 363 annotated families. Those auxiliary splits satisfy their declared chemistry or sequence contracts but do not support unseen-family claims. The family sidecar covers patent-derived records only, and the earlier external-confirmation seal remains consumed. An external single-target RNase-H panel (14 eligible candidates, one protocol) has been offered under NDA with sequences pending release; it is held as a prospective T5 candidate and no evaluation has been run.

F Mechanism catalogue

F.1 Grouped identifier table with evidence tiers

This appendix catalogues the mechanism identifiers referenced in the main text (Section 1). The platform distinguishes three therapeutic goals (*Make LESS*, *Make MORE*, *Make a DIFFERENT VERSION*) and assigns each identifier an evidence tier based on KRAIG's current support status.

Table 14: Mechanism identifier catalogue (43 IDs; three evidence tiers).

ID	Mechanism	Goal	Tier
A1	RNase-H gapmer	LESS	1
A2	Steric translation block	LESS	1
A3	Poison-exon blocking	MORE	2
A4	Antisense-transcript knockdown	MORE	2
A5	uORF blocking	MORE	2
A6	miRNA-site blocking	MORE	2
A7	Frame-breaking exon skipping†	LESS	1
A8	Exon inclusion	DIFFERENT	2
A9	Pseudoexon suppression	MORE	2
A10	Cryptic-site blocking	MORE	2
A11	APA shortening	MORE	3
A12	Anti-miRNA†	MORE	2
A13	ADAR editing (A-to-I)	DIFFERENT	2
A14	Toxic-RNA neutralisation	LESS	1
A15	Transcriptional silencing	LESS	1
A16	C-to-U editing (APOBEC)	DIFFERENT	3
A17	LEAPER (ADAR recruitment)	DIFFERENT	3
A18	Ribozyme (catalytic RNA)‡	—	OOS
A19	aptazyme‡	—	OOS
A20	Trans-splicing (5' leader)	DIFFERENT	2
A21	siRNA (Dicer-substrate)	LESS	1
A22	EGS / RNase P	LESS	3
A23	saRNA (small activating RNA)	MORE	3
A24	circular RNA scaffold‡	—	OOS
A25	Aptamer (binding-only)	DIFFERENT	3
A26	Ribosome-skipping 2A‡	—	OOS
A27	Riboswitch OFF	MORE	3
A28	RBP-site masking	MORE	2
A29	IRES modulation	DIFFERENT	2
A30	tRNA suppressor‡	—	OOS
A31	gRNA / CRISPR guide‡	—	OOS
A32	SINEUP / long-ncRNA‡	—	OOS
A33	Intron retention	LESS	3
A34	Trans-splicing (3' trailer)	DIFFERENT	3
A35	Protein-fusion tag‡	—	OOS
A36	Suppressor-tRNA (stop-codon)	DIFFERENT	3
A37	5' cap modulation‡	—	OOS
A38	3' tail modulation‡	—	OOS
A39	Promoter-targeting‡	—	OOS
A40	TDMD (miRNA degradation)	MORE	3
A41	U1 snRNA adaptors	LESS	2
A42	Splice-ratio shifting	DIFFERENT	3
A43	TSS redirection	DIFFERENT	3

Tier 1: full contract; Tier 2: generation + ranking; Tier 3: lookup only. †Target-dependent (evidence intervals require target-specific annotation). ‡Out of scope (protein-scale or non-oligonucleotide). OOS = out of scope.