REPRESENTABLE BUT UNLEARNED: ENCODING RANK AND THE INTERACTION-PREDICTION FLOOR

Anonymous authors
Paper under double-blind review

ABSTRACT

Input encodings can restrict which measured contrasts a predictor can jointly reproduce, even when no single contrast is forced to vanish. We compute the attainable contrast space from an encoder's equivalence classes and a fixed contrast design, without labels, loss, or a fitted model; projecting the recorded contrasts onto that space gives an empirical error floor for any unrestricted decoder on those classes. On a 140-rectangle siRNA interaction panel, a graph neural network's training-only feature mask merges 165 endpoint states into 90 classes and cuts the rank of the 140 interaction contrasts to 72. The resulting floor is 0.009980, which is 14.6% of the fitted model's interaction squared error; the fitted model reaches 0.068335, slightly worse than a control predicting no interaction at all. A minimum of three restored chemistry columns recovers full rank. Refitting without the mask removes the floor entirely, yet interaction MSE improves by only 0.000017 under the reported protocol, and the restored columns remain absent from every training input. On a released RNA-splicing predictor, whose encoding is injective on the measured states, the same computation returns the full design rank of 1,986 and a floor of exactly zero. These results separate what an encoding permits from what a fitted model achieves; they do not identify what limits the remaining error. The rank check needs no fits and bounds what any amount of training under a fixed encoding can recover.

1 INTRODUCTION

Chemically modified small interfering RNAs (siRNAs) are designed by changing sugars, backbones and bases at named positions, so the quantity of interest is a *contrast* between modified conditions on a fixed background. Pooled error or correlation on heterogeneous activity tables evaluates endpoint prediction, not contrasts between matched chemical conditions; each claim needs its own evaluation. Our scope is the datasets and predictors examined here. Neither published modified-siRNA route passed its readiness gate (Section 7: running natively end to end on these evaluation states), so the only new fits on measured biological datasets are two interventions: removing the training-only mask of the audited graph neural network (GNN), and restoring training support in one published splicing predictor. The historical training campaigns and the present analysis are kept separate.

Contributions. We develop an exact finite-panel diagnostic for a fixed encoding and contrast design. First, we characterize the joint contrast vectors attainable by unrestricted deterministic decoders, exposing restrictions that individual-contrast cancellation checks can miss. The rank calculation is label-independent; projecting recorded contrasts yields an empirical squared-error floor attained by an unrestricted decoder. Second, we localise the siRNA panel's rank loss to the training-only feature mask and identify minimum-size column restorations that recover the full design rank. Third, we evaluate the distinction between representability and fitted accuracy: mask-restoration refits produce little improvement under the reported protocol, while a released splicing predictor with an injective encoding preserves the full design contrast space and has zero encoding floor.

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

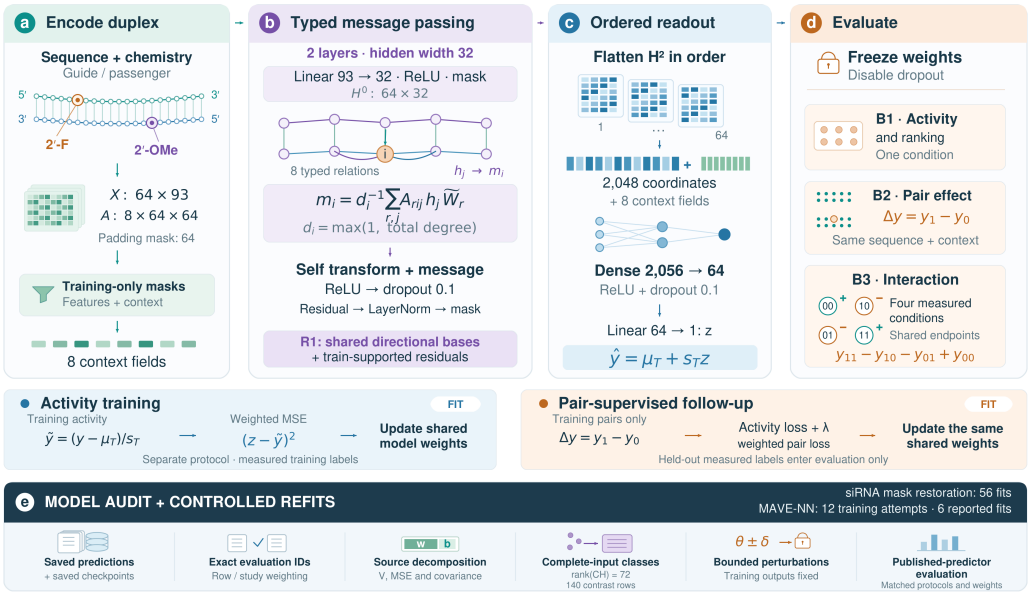

Figure 1: Architecture and evaluation (Appendix A2). Panels (a)–(d) show the siRNA representation, typed message passing, ordered readout, and evaluation with frozen weights. The two supervision bands distinguish activity training from pair-supervised follow-up; held-out labels enter evaluation only. Band (e) summarizes saved-model auditing, published-predictor evaluation and controlled refits: 56 siRNA mask-restoration fits and 12 MAVE-NN attempts, six of which supply reported results.

1.1 Benchmarks and cohorts

Three cohorts supply measurements. The *grouped ENsiRNA benchmark* is our eligible subset of the activity table released with ENsiRNA (Tan et al., 2025): 2,927 of its 3,527 released rows meet the admission rules of Appendix A1, in 150 sequence-linked and ten study-linked components never split between training and evaluation. Of the admitted rows, 1,924 come from Bramsen et al. (2009, PubMed 19282453) and 594 from the patent family US20120088815A1/EP2415869A1. The *APP cohort* holds 1,839 measurement rows against the amyloid-beta precursor protein gene (APP) from two Alnylam patent families, curated in CMsiRNAdb (He et al., 2026) and measured in 17 overlapping assay pools. *Davis S7* is the 118-row Supplementary Table 7 of Davis et al. (2025). Activity models are trained and selected on the grouped benchmark only, so APP and S7 are held out; saved models come from campaigns v3 (released labels) and v4 (Bramsen excluded or relabelled from its primary assay).

Three benchmarks use them. *B1* is activity and ranking: held-out knockdown prediction on all three cohorts, and top-five selection within each APP assay pool. *B2* is a measured chemistry bundle: 156 APP pairs over 26 duplex backgrounds in twelve sequence components, where guide positions 6, 8 and 9 change from 2-Fluoro to 2-O-Methyl while position 7 changes to glycol nucleic acid (GNA), so no single change is isolated. *B3* is a complete antisense-by-sense sub-rectangle we extract from the 48-by-45 (2,160-duplex) screen of Bramsen et al. (2009); the 15-by-11 selection is ours, not a design that screen reports. Its 15 antisense and 11 sense chemistries on one background give 165 measured states, 140 four-condition rectangles, and 14 antisense and 10 sense marginal effects against a shared reference.

2 An exact representational restriction

Fix a finite panel with distinct raw endpoint states $s_1, \ldots, s_m$, signed contrast coefficients $C \in \mathbb{R}^{q \times m}$, and an encoding indicator $H \in \{0,1\}^{m \times k}$ sending each state to one of k complete-input classes. An unrestricted deterministic decoder on those classes is an arbitrary $g \in \mathbb{R}^k$ and its

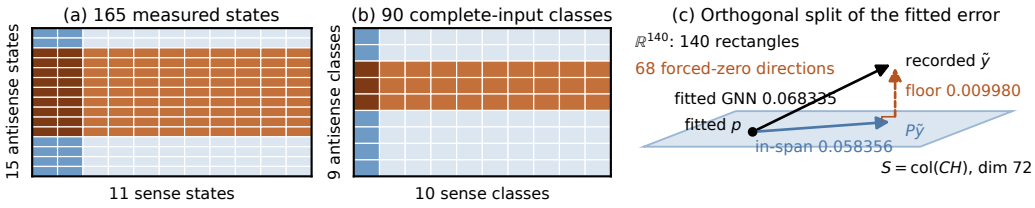

Figure 2: How the encoding collapses the panel, and what the collapse costs. (a) The panel is a 15 by 11 antisense-by-sense product, 165 states. (b) The mask zeroes 44 node-feature columns, merging three antisense triples and one sense pair (rust rows, blue columns) into 9 by 10 classes. A minimum of three restored chemistry columns recovers full rank. (c) The 140 contrasts then span only $S = \mathrm{col}(CH)$, of dimension 72, leaving 68 forced-zero directions; the fitted error splits orthogonally into an in-span part and a residual off S whose mean square is the floor.

contrast vector is CHg. Representability here means attainability by an unrestricted decoder on these classes, not by a fitted family. The next two statements hold for any finite panel and encoding, with no labels, loss or fitted model.

Proposition 2.1. *The contrast vectors attainable by unrestricted deterministic decoders on the encoding classes are exactly* $\mathrm{col}(CH)$. *A fixed linear functional* $u^\top$ *of the* q *contrasts vanishes for every such decoder if and only if* $u^\top CH = 0$, *so the forced-zero functionals form a space of dimension* $q - \mathrm{rank}(CH)$. *If in addition* $C\mathbf{1}_m = 0$ *then* $\mathrm{rank}(CH) \le k - 1$.

Corollary 2.2. *If the classes factor as* a *antisense by* b *sense classes and the panel is the full antisense-by-sense product with every rectangle anchored at one shared reference state, the interaction rank is exactly* $(a-1)(b-1)$.

With 0 the reference state's class, each rectangle's contrast is one interaction coordinate $t_{ab} = g_{ab} - g_{a0} - g_{0b} + g_{00}$ or zero; decoders set the $(a-1)(b-1)$ nonreference coordinates independently and the full product hits each, so the rank is exactly their number. Both $\mathrm{rank}(CH)$ and the forced-zero dimension follow by exact rational elimination on integer matrices from the encoder and design alone, so the procedure applies to any categorical panel; proofs are in Appendices A10.6 and A10.11.1. For the recorded contrast vector $\widetilde{I} \in \mathbb{R}^q$ the encoding floor is

$$F = \min_{g \in \mathbb{R}^k} \tfrac{1}{q}\|\widetilde{I} - CHg\|^2 = \tfrac{1}{q}\|(\mathrm{Id} - P)\widetilde{I}\|^2, \qquad P \text{ the orthogonal projector onto } \mathrm{col}(CH).$$

The rank needs no labels, whereas the floor is empirical, since it uses the recorded contrasts; "achievable" means attained by an unrestricted class decoder, not necessarily by the fitted family.

Everything below is specific to the audited panel (Fig. 2). With $q = 140$, $m = 165$ and $k = 90$, Proposition 2.1 bounds $\mathrm{rank}(CH) \le 89$ and exact elimination gives 72: the encodings impose 68 independent linear constraints on the joint *predicted* contrast vector, which the recorded vector need not satisfy. No row of CH is zero, so the complete-input test finds no forced cancellation in any of the 140 rectangles. The partition, shared by all 40 audited instances and verified on all 27,225 state pairs, factors into 9 antisense and 10 sense classes, so Corollary 2.2 gives $(9-1)(10-1) = 72$ and duplicate-row equalities generate all 68 constraints. Distinct sugar chemistries merge within each matched position pattern; all four families share this preprocessing, so the finding does not extend to unrelated published pipelines. On a released splicing predictor the same computation returns a floor of exactly zero (Section 9).

Every fitted contrast vector lies in $S = \mathrm{col}(CH)$ to within 10^{-16}, so the fitted error splits orthogonally into floor and in-span parts, per family and method in Table 1. For the GNN the floor is 38.1% of fitted squared error on antisense marginals, 0.4% on sense marginals and 14.6% on interactions, each against its own family's fitted error. The rank loss localises to the mask alone: parsed chemistry, the pre-mask tensors and the mask-restored encoding all keep 165 classes and rank 140 (appendix Fig. 6). A target contrast vector is representable exactly when it lies in $\mathrm{col}(CH)$, and every contrast attainable from the raw design is preserved exactly when $\mathrm{rank}(CH) = \mathrm{rank}(C)$; representability does not imply learnability, which Section 8 examines separately. These numbers are specific to one panel and one background; Appendix A10.2 gives the complete explainer counterexample.

Family	Method	q	Floor	In-span	MSE_q	Ratio
Antisense	R1 GNN	14	0.018972	0.030885	0.049856	2.627909
Antisense	R1 no-message	14	0.018972	0.031574	0.050546	2.664273
Antisense	Chemistry tree	14	0.018972	0.023931	0.042903	2.261392
Antisense	Token CNN	14	0.018972	0.030339	0.049311	2.599177
Sense	R1 GNN	10	1.478×10^{-5}	0.003320	0.003335	225.665139
Sense	R1 no-message	10	1.478×10^{-5}	0.003221	0.003236	218.972986
Sense	Chemistry tree	10	1.478×10^{-5}	0.008408	0.008423	569.941309
Sense	Token CNN	10	1.478×10^{-5}	0.004025	0.004040	273.371096
Interaction	R1 GNN	140	0.009980	0.058356	0.068335	6.847534
Interaction	R1 no-message	140	0.009980	0.058301	0.068281	6.842116
Interaction	Chemistry tree	140	0.009980	0.061959	0.071939	7.208643
Interaction	Token CNN	140	0.009980	0.058429	0.068409	6.854911

Table 1: Encoding-floor decomposition of the recorded B3 error, equal weights: Floor is $q^{-1}\|(\mathrm{Id} - P)\widetilde{I}\|^2$ with P the projector onto $\mathrm{col}(CH)$ for that family, In-span is $q^{-1}\|P\widetilde{I} - p\|^2$, and MSE_q their sum, exact before rounding. The floor is shared across methods in a family because every fitted contrast vector lies in $\mathrm{col}(CH)$; denominators are the 14 and 10 marginal contrasts and the 140 rectangles. All entries are rounded from stored values; floor, in-span and ratio are each computed before rounding. The tree is extremely randomized trees, the token CNN a convolutional baseline.

3 Evaluation framework and identifiability

Sections 3 and 4 fix the conventions and provenance rules used later; results begin in Section 5. We distinguish three claims. **(C1)** An explanation is *faithful* if it describes the fixed predictor it is computed from; **(C2)** a dependence is *training-supported* if the training computation constrains it; **(C3)** a prediction *agrees with a measured effect* if it matches a comparable recorded intervention. A faithful input derivative answers C1 alone; a good pooled score answers none.

Four definitions make these operational. A *complete-input encoding class* groups raw states whose audit signatures, defined once in Appendix A8.4, are byte-identical. An *evaluation identity* hashes identifiers, labels, weights and response convention, so two numbers compare only when these match. A *weighting* is one unit per row, or inverse study-component size normalised to average one; the two define different estimands. A *training-absent block* provably cannot change any training prediction, for every parameter value, not merely at one gradient snapshot.

The duplex occupies 64 ordered slots with 93 features; two width-32 layers over eight directed relation types feed a 2,056–64–1 ordered readout with degree-normalised typed messages (Appendix A2). R0 (original GNN) uses independent W_r; R1 (training-gated GNN) shares directional bases and gates residuals by positive training occurrence. Four relation types are absent from every training graph in each layer, 8,192 independently parameterised scalars: support changes which routes receive data gradients, not the parameter count. Appendices A2 and A3 give the forward map and training protocol.

4 Evidence, provenance and exact evaluation identities

In the grouped benchmark the Bramsen rows are 65.73% of the total in one sequence component and the patent family's 594 rows span 132 components. Source identifiers are categories, not independent studies, and released per-row assay provenance is incomplete, so no assay-level partition is claimed. Denominators matter. One 20-row source has label variance 0.000711 and MSE 0.06739, so its R^2 is about -93.78 even though its absolute error is not large (`variance_sources`; Appendices A4 and A10.7).

Two uses of the same 1,003 scored rows must not be confused: rescoring the v3 models, trained and selected on data including Bramsen, gives tree $R^2 = +0.002992$ and GNN $R^2 = -0.1142$, while Table 2 reports models trained *and selected* with that source excluded, giving -0.000692 and -0.2115 on the same rows. Equal row counts establish neither equal labels nor equal model identities, so each comparison carries its identity hash.

Primary-source adjudication is preserved and unresolved: parsing the printed chemistry links 1,604 of 1,924 admitted Bramsen rows to a unique primary measurement, quarantines 320, and leaves 299 differing from the released labels by more than 0.05, maximum 1.0852 — unresolved discrepancies, not certified label errors. The strand orientation of Davis Supplementary Table S1 stays quarantined, APP spans 92 sequence components, and rows, pairs and states are different units that do not count replication. The 792 focused v4 fits are historical (Appendix A10.16).

5 Activity results and training-fitted controls

Grouped benchmark: 1,924 of 2,927 rows (65.73%) from one source in one sequence component							
Sensitivity	Method	Grouped MSE	R^2	APP MSE	R^2	S7 MSE	R^2
Source-excluded	R1 GNN	0.0641	-0.212	0.1631	-0.005	0.1191	-0.324
	R1 no-message	0.0625	-0.182	0.1637	-0.009	0.1221	-0.358
	Chemistry tree	0.0529	-0.001	0.1655	-0.020	0.0974	-0.082
	Token CNN	0.0628	-0.187	0.1626	-0.002	0.1103	-0.226
	Training row mean	0.0571	-0.079	0.1810	-0.115	0.1587	-0.764
	Oracle test mean	0.0529	0.000	0.1623	0.000	0.0900	0.000
Primary-assay	R1 GNN	0.0747	0.005	0.1628	-0.003	0.1052	-0.169
	R1 no-message	0.0746	0.007	0.1624	-0.001	0.1073	-0.193
	Chemistry tree	0.0692	0.079	0.1899	-0.171	0.0934	-0.039
	Token CNN	0.0753	-0.002	0.1641	-0.011	0.1019	-0.132
	Training row mean	0.0759	-0.011	0.1742	-0.074	0.1450	-0.612
	Oracle test mean	0.0751	0.000	0.1623	0.000	0.0900	0.000

Table 2: Row-weighted fixed-ensemble prediction on matched rows within each sensitivity; cohorts and models are those of Section 1.1 and Table 1. Grouped coverage is 1,003 and 2,607 rows; APP and S7 retain 1,839 and 118. The training row mean is a deployable constant from each training partition; the oracle test mean reads the evaluation labels and is a diagnostic, not a risk floor.

On the source-excluded grouped target the retrospective test-mean MSE is 0.05289. The tree beats both training-mean predictors under row weighting (-0.004148, -0.001961) but not under equal study-component weighting (-0.001319, $+0.000535$), and all four paired conditional 95% intervals include zero over nine resampled components (Appendix A10.12): no general advantage. The GNN has higher observed error than both constants, and on the primary-assay cohort its R^2 changes from $+0.004856$ by rows to -0.3288 by equal-component weighting.

Negative per-source R^2 with positive pooled R^2 does not force a between-source explanation: only 3.98% of released label variance lies between source categories, and the between-source covariance is negative (Appendix A10.7). This corrects score interpretation; it is not a contribution.

6 Calibration, protocol and ranking sensitivity

Results from the v3 released-label campaign are in Table 13. On the v3 APP evaluation squared bias is 99.53% of the tree-minus-GNN gap, which describes it without explaining it, and GNN correlation -0.018 with prediction SD 0.0072 against label SD 0.4028 means a relative MSE advantage does not establish useful discrimination. Comparisons are protocol-specific: the GNN-minus-no-message difference on the same 118 S7 rows changes sign across protocols, while the APP tree comparison favours the GNN throughout; interval endpoints must not be subtracted to infer the uncertainty of a change. Ranking is descriptive for the same reason: populations, support rules, optimisation and selection changed jointly, so a reversal identifies no single cause (Appendix A6).

7 Measured chemistry contrasts

For B2 (Section 1.1), both endpoints of every pair and their sequence groups are excluded from each outer training and selection population. Pair-GNN equal-component MSE minus the training-fold

mean is -0.0125 $[-0.0707, 0.0464]$ and its within-assay correlation 0.266 $[-0.045, 0.493]$, both unresolved; all pair predictions carry the same operational sign, giving 111/137 correct signs, matching the majority rule under the ± 0.02 band. Sequence-specific measured chemistry prediction is not established.

Panel/model	Inputs	Forced	MSE_{165}	$(\Delta\mu)^2$	AS	SS	NA	Pred. NA
Measured panel	165	–	0.079894	–	0.041341	0.020415	0.018138	–
R1 GNN	90	0	0.077350	0.000041	0.042930	0.016196	0.018183	0.000002
R1 no-message	90	0	0.077353	0.000005	0.042811	0.016371	0.018166	0.000002
Chemistry tree	90	0	0.078546	0.000012	0.043859	0.016079	0.018596	0.000146
Token CNN	90	0	0.077276	0.000211	0.042819	0.016072	0.018174	0.000005

Table 3: Primary B3 on the complete 15 by 11 panel, equal weights, reusing source-held-out endpoints. Exact complete inputs among the 165 states force no cancellation in any of the 140 rectangles, for all ten preprocessing instances. MSE_{165} is endpoint MSE over 165 states, not the contrast MSE_q of Table 1; AS, SS and NA are antisense, sense and non-additive components, and row one, being centered energy, has no grand-mean term. The four components sum to MSE_{165} exactly; Pred. NA is each model's own predicted nonadditive energy. These component sums hold to 10^{-15}.

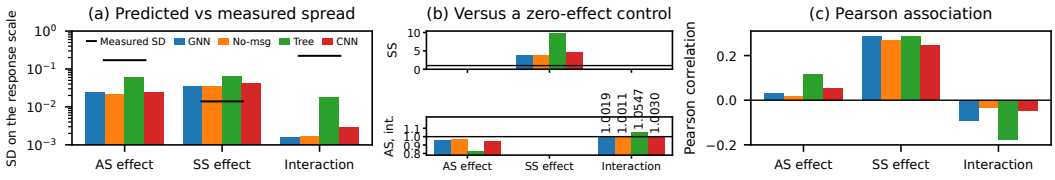

Figure 3: Source-held-out predictions on the primary panel, from the ten saved seeds. (a) Predicted against measured spread per family. (b) Each family's error over its zero-effect control, unity being the control, in two sub-panels because sense reaches 9.85; interaction ratios are printed. (c) Pearson correlation with the measured effect.

Endpoint prediction on the same panel is weak: GNN $R^2 = 0.0318$, Pearson 0.1848, predicted SD 0.04026 against recorded 0.28266, so this is no interaction-specific failure of an otherwise strong predictor. Nor is every effect attenuated: sense-marginal predicted SD is 0.03447 against recorded 0.01395, while antisense and interaction predictions have too little spread (Fig. 3).

The 140 contrasts $I_{AB} = \mu_{11} - \mu_{10} - \mu_{01} + \mu_{00}$ share states and one background, so no independent-rectangle interval is computed. All 140 GNN ensemble contrasts are nonzero with seed magnitudes above 10^{-4}, their SD 0.0015893 against a measured 0.2224421, and MSE 0.068335 sits only 0.19% above the 0.068205 zero-interaction control, which matters less than the near-constant predictions and weak effect recovery. Splitting the panel into a grand mean, antisense and sense main effects and a nonadditive residual, measured energy divides 51.7%, 25.6% and 22.7%, and each model's error splits exactly (Table 3). For the GNN ensemble it is 0.04293 antisense and 0.01620 sense against 0.01818 nonadditive: the deficit is in the main effects, not specifically interactions, and every nonadditive error exceeds the zero-residual control. This is descriptive algebra on the recorded scale, a different object from the encoding-class projection of Section 2 (Appendix A10.23).

Fixed-prediction B3 sensitivities qualify that comparison: reference shifts of one reported SD either way preserve the ordering, none of 14 antisense omissions reverses it, and one of 10 sense omissions (JC5) does, with GNN-minus-zero MSE -0.000009. These are dependent influence diagnostics, not confidence limits or new folds (Appendix A8.5).

Reproducing the published predictors. The published modified-siRNA predictors ENsiRNA-mod (Tan et al., 2025) and MEG-mod (Chen et al., 2026) were re-examined natively. MEG-mod's forward assembles modification tokens from variables its release never defines; ENsiRNA-mod needs licence-gated per-duplex Rosetta geometry that is absent. All 156 B2 pairs, 165 B3 endpoints and 140 contrasts therefore stay unassessed, with unknown supervised overlap (Appendix A10.18).

8 TRAINING-SUPPORT DIAGNOSTICS

A relation absent from every training graph contributes zero whatever its transform, so a layer-by-layer induction proves its parameter block cannot change any training prediction. The induction (Appendix A10.3) concerns the training computation, not one zero gradient, and licenses an outcome-independent perturbation test that fits nothing. It does not imply independence of validation selection, regularisation or pretraining, and AdamW's decoupled weight decay can still move a zero-gradient parameter: these are not never-updated weights. The 8,192 scalars belong to the original architecture; the corrected models' 8,192 false-gated residual scalars are a different object.

The perturbation protocol used three saved seeds per arm, fixed Rademacher directions and scales 0 and ± 0.25 of each block's root-mean-square magnitude. All 2,518 training predictions per checkpoint stayed bitwise equal. In the original model (R0) the largest of the 1,839 APP prediction changes is 0.003300 and 1,601 to 1,721 global ranks move, yet top-five selection within the seventeen assay pools changes in only 15/102 cases (minimum overlap 0.800), against 0/102 for R1; overlapping pools are not independent trials. The largest raw input-gradient change, 0.005870 on sixteen queries, is a local sensitivity, not a feasible chemical intervention. R1's gated residuals give the expected null, zero-scale controls pass and B3 changes are exactly zero, so this mechanism does not explain the B3 attenuation.

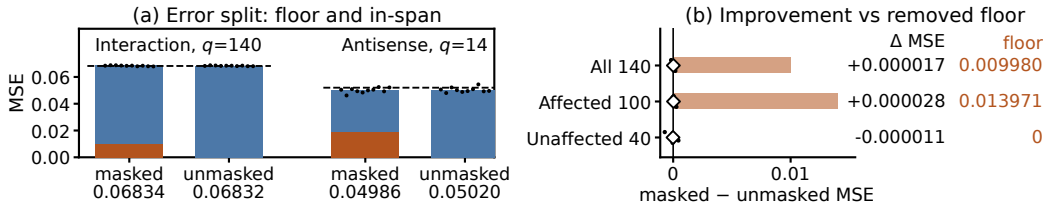

Figure 4: Mask-restoration follow-up on the R1 GNN, ten paired seeds; interaction $q = 140$, antisense $q = 14$. (a) Ensemble floor (rust) and in-span error (blue), each seed's total a dot, control dashed. The restored encoding has zero oracle floor. (b) Improvement per seed (dots) and ensemble (diamond) against the masked floor per subset (bar); the 100 affected rectangles hold all of the floor.

Mask-restoration follow-up. Refitting the R1 GNN without the training-only node-support mask restores 165 classes and rank 140, reducing the unrestricted oracle floor from 0.009980 to zero. Architecture, optimizer, selection rule, splits and seeds were unchanged; the re-run grid chose the same configuration, and the masked refit reproduced saved predictions to 6×10^{-17}. MSE moved from 0.068335 to 0.068318, an improvement of 0.000017 or 0.17% of the removed floor. The original protocol hypothesized an improvement no larger than the removed floor. This was not a valid general upper bound, because refitting can also change error within the original contrast space. We retain the removed floor only as a descriptive reference scale. The observed paired-seed changes ranged from -0.000176 to $+0.000262$ (Fig. 4), inside the protocol's near-zero category of $|\Delta| < 0.004990$. The mask zeroes 44 of the 93 node-feature columns, all zero on every training node, so their embeddings receive no training-data gradient; only those separating otherwise-distinct chemistries collapse states, so the rank loss localises to a few named columns (Appendix Table 16). This does not by itself imply unchanged weights under AdamW; validation selection and weight decay remain distinct from the training-data term. Removing the encoding restriction did not materially improve accuracy under this refit protocol. The restored columns remained zero on every training node, so restoration did not add direct training support for those feature coordinates. These observations do not identify what limited the remaining accuracy. These 56 fits (21,513 optimizer updates) are the only siRNA fits in this work (Appendix A10.11.3). Neither published modified-siRNA pipeline passed its readiness gate (Section 7), so generalisation needs a predictor that runs end to end; MAVE-NN provides one.

9 GENERALIZATION TEST: A COMPLETE PUBLISHED PREDICTOR

MAVE-NN (Tareen et al., 2022), a framework for multiplex assays of variant effect, ships a model of the BRCA2 exon-17 subset of the massively parallel splicing assay (MPSA) of Wong et al. (2018),

which measures 5-prime splice-site selection, here as log10 percent spliced in (PSI): not modified siRNA, and nothing here closes that gap.

The two encoders fall on opposite sides of the diagnostic: the siRNA encoder removes 68 of the B3 design's 140 contrast directions, leaving rank 72 and a floor of 0.009980, 14.60% of its fitted interaction error, whereas the splicing encoder removes none, leaving $\mathrm{rank}(CH) = \mathrm{rank}(C) = 1{,}986$ of 2,083 and a floor of exactly zero. Under the complete-input rule of Section 2 its one-hot input keeps all 3,696 endpoint states distinct. The forced-zero functionals differ in kind: the siRNA panel's 68 are encoding dependencies from merging 165 states into 90 classes, which the recorded contrasts do not satisfy, so the shortfall is unreachable error; the splicing panel's 97 are design dependencies the measured contrasts satisfy too, costing nothing. Only the first kind is paid for, and all of the equal-weight interaction MSE 0.3334 is in-span; the same 2,083 rectangles under the nested primary weights of Table 4 give 0.3617, a different weighting of the same contrasts rather than a different result (library 2: rank 1,856 of 1,947, 91 design dependencies, floor zero; Fig. 5b). This null second instance shows that the diagnostic distinguishes the two declared encodings and does not automatically assign a positive floor to a new panel. A frozen-protocol extension finds the same split (Appendix A10.22): a wider panel from the same Bramsen screen loses 730 of its 1,845 contrast directions, while a protein-binding panel under the same one-hot encoding loses none; neither adds an independent encoder.

The released pairwise global-epistasis (GE) checkpoint records $N = 24{,}405$, matching the released training-plus-validation count, but the metadata do not list all training-sequence identities, so the count alone does not certify held-out provenance. Our native evaluation gave $I_{\mathrm{var}} = 0.3059$ and $I_{\mathrm{pred}} = 0.3424$ against tutorial values 0.3066 and 0.3580: not bitwise reproduction (Appendix A10.21).

Quantity	N	Model MSE	Matched control	Difference	Pearson
Endpoints, released test split	6078	0.084224	0.259250	-0.175026	0.822247
Endpoints, rectangle population	3696	0.089748	0.246508	-0.156759	0.797282
Single substitutions	12486	0.155262	0.302181	-0.146919	0.697187
Interactions (primary)	2083	0.361690	0.456701	-0.095011	0.462959
Interactions, library 2	1947	0.327946	0.455823	-0.127877	0.531564

Table 4: Released MAVE-NN pairwise GE predictor on the released MPSA test split, in $\log_{10}$ PSI; the checkpoint-provenance limitation is stated in Appendix A10.21. Endpoint controls are the constant fitted to the released training-plus-validation population; contrast controls are matched zero-effect predictions on identical identities and weights. Primary rectangle weights give equal mass to each position pair, then each background within a pair, then each rectangle within a background; for library 2 they are restricted to surviving identities and renormalized. Negative differences favour the model; differences are computed before rounding.

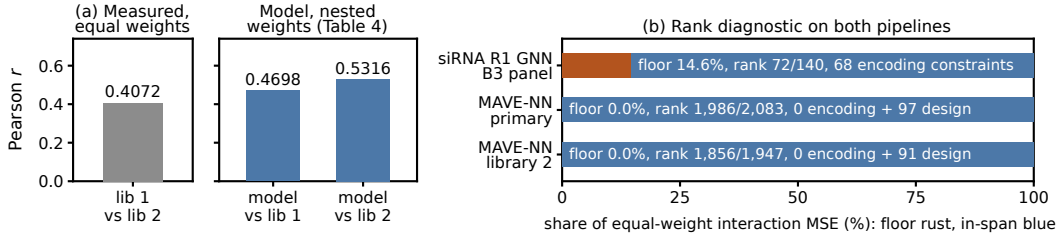

Figure 5: MAVE-NN on RNA splicing, not modified-siRNA efficacy. (a) Two separate comparisons on the $1{,}947$ shared rectangles: measured library agreement under equal rectangle weights, model agreement under the nested weights of Table 4. Neither is a noise ceiling or a bound on model error. (b) Encoding floor (rust) and in-span error (blue) for both pipelines, as reported in the text.

Every quantity in Table 4 beats its matched control and the direction holds unweighted (-0.07940). Measured-effect performance is not implied by endpoint performance: interaction Pearson is 0.4630 against 0.7973 on the linked endpoint population, and predicted interaction SD 0.3678 against measured 0.6756. The replicate comparison shows limited agreement between measured

interactions but does not apportion the model–measurement discrepancy between prediction error, measurement error and shared bias.

Fixed-size support intervention. At training size $14{,}238$ and fixed architecture, the restored-support arm improved affected held-out interaction MSE by 0.04415 relative to the restricted arm; the matched zero-effect MSE was 0.3756. Unaffected-interaction MSE increased by 0.01441, and all three paired seeds showed improvement on the affected subset (Appendix Fig. 8). These concern one replacement design for one feature with unmatched backgrounds; they do not isolate feature support from other consequences of replacing training observations. Twelve attempts ran, six supplying the retained results and $62{,}424$ optimizer updates (Appendix A10.24). The withheld coefficients receive no data gradient but remain under the L2 penalty, so prediction independence and regularized-objective independence stay distinct; no training-independent block is claimed.

10 Related work

Faithfulness and agreement with a scientific effect are distinct targets (Chen et al., 2020). The closest prior art is Kuskova et al. (2026), who prove population identifiability of normalized minimal gated neural additive autoregression decompositions under distributional support conditions. Their continuous-support population theorem does not apply to this finite discrete panel and supplies no finite-panel encoding floor; we compute an exact rational rank and an achievable empirical floor over unrestricted class decoders. Lengerich et al. (2020) canonicalise a fitted model's terms against a specified feature distribution; ours is reachability under a fixed encoder, without a population-distribution assumption or a fitted predictor. The statistical ancestor is factorial estimability (Box & Hunter, 1961; Wu & Chen, 1992); here the raw endpoint design permits all 140 contrasts, and the encoding introduces the additional restriction when it merges 165 states into 90 classes. Bilodeau et al. (2024) show complete, linear attributions can fail to beat chance at spurious-feature and recourse tasks; full rank of the finite contrast map neither establishes nor refutes their hypotheses, and we claim no formal separation. Azzolin et al. (2025; 2026) study architectural limitations and degenerate explanations in regular and self-explainable GNNs. Our result characterises which contrast vectors a predictor can attain on the measured panel under a fixed complete-input encoding. Gradient correction on the one-hot simplex removes normal components, not every direction absent from support (Majdandzic et al., 2023); Tareen et al. (2022) model genotype–phenotype and measurement maps separately. Chemistry-aware predictors already include ENsiRNA-mod and MEG-mod (Tan et al., 2025; Chen et al., 2026); sequence determinants and message passing are established (Reynolds et al., 2004; Khvorova et al., 2003; Gilmer et al., 2017; Schlichtkrull et al., 2017), and our interaction-index and derivative conventions follow Grabisch & Roubens (1999) and Simonyan et al. (2013). Underspecification can give similar in-domain performance with different deployment behaviour (D'Amour et al., 2022); ours proves training-prediction independence for specified blocks, not equal validation performance.

11 Discussion and limitations

This study separates input distinguishability, training support and measured-effect accuracy. It explains the finite-panel rank restriction and measures training-independent prediction changes. The input-collision restriction leaves 85.40% of GNN squared error inside the permitted contrast space. Removing the mask changes MSE only slightly, and the reference-independent decomposition places the larger error components in the main effects. These observations do not identify the cause of the remaining error; what they do establish is that rank and support are distinct checks on representation and observed training support, and that on this panel opening the first left the second untouched.

Independent validation of modified-siRNA efficacy remains incomplete: neither the splicing result nor the support intervention transfers to chemical modification. B3 has one background, B2 jointly changes four positions, APP and S7 are retrospective with related sources, and shared-control covariance, the 299 primary/released discrepancies, Davis Table S1 orientation and dose units stay unresolved. Seeds are not biological replication; no wet-lab, causal or clinical validation is claimed, and no broadly useful reliability score or independent generalisation is established.

AI USE STATEMENT

AI tools were used throughout this work.

The authors have checked the whole paper: its mathematical statements and their proofs, every reported number against the recorded evidence store, and its text.

The authors take responsibility for the scientific claims, the citations, the data permissions and the submitted manuscript.

ETHICS STATEMENT

This study reanalyzes published measurements and trains computational predictors. It includes no newly collected biological specimens, wet-lab procedures, participant recruitment or clinical validation. Predicted activity and chemistry contrasts are not recommendations for clinical use. Missing replication and assay information, provenance limits and negative outcomes remain explicit.

REPRODUCIBILITY STATEMENT

Appendices A1–A7 specify inputs, grouping, models, objectives, selection, metrics, uncertainty and resources. The supplementary archive contains the analysis code, the frozen protocols, the execution records, the evidence store from which every reported number is drawn, and the encoding matrices and rank records behind the representational results. Data that cannot be redistributed is replaced by its extraction script and a checksum manifest.

REFERENCES

Charles R. Allerson, Namir Sioufi, Russell Jarres, Thazha P. Prakash, Nishant Naik, Andres Berdeja, Lisa Wanders, Richard H. Griffey, Eric E. Swayze, and Balkrishen Bhat. Fully 2'-modified oligonucleotide duplexes with improved in vitro potency and stability compared to unmodified small interfering RNA. *Journal of Medicinal Chemistry*, 48(4):901–904, 2005. doi: 10.1021/jm049167j. URL https://doi.org/10.1021/jm049167j.

Steve Azzolin, Antonio Longa, Stefano Teso, and Andrea Passerini. Reconsidering faithfulness in regular, self-explainable and domain invariant GNNs. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2406.15156.

Steve Azzolin, Stefano Teso, Bruno Lepri, Andrea Passerini, and Sagar Malhotra. GNN explanations that do not explain and how to find them. *arXiv preprint arXiv:2601.20815*, 2026. URL https://arxiv.org/abs/2601.20815.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. URL https://arxiv.org/abs/1607.06450. arXiv:1607.06450; author manuscript.

Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks, 2018. URL https://arxiv.org/abs/1806.01261. arXiv:1806.01261.

Blair Bilodeau, Natasha Jaques, Pang Wei Koh, and Been Kim. Impossibility theorems for feature attribution. *PNAS*, 121(2):e2304406120, 2024. doi: 10.1073/pnas.2304406120. URL https://doi.org/10.1073/pnas.2304406120.

Amanda Birmingham, Emily M. Anderson, Angela Reynolds, Diane Ilsley-Tyree, Devin Leake, Yuriy Fedorov, Scott Baskerville, Elena Maksimova, Kathryn Robinson, Jon Karpilow, William S. Marshall, and Anastasia Khvorova. 3' UTR seed matches, but not overall identity, are associated with RNAi off-targets. *Nature Methods*, 3(3):199–204, 2006. doi: 10.1038/nmeth854.

G. E. P. Box and J. S. Hunter. The 2^{k-p} fractional factorial designs part I. *Technometrics*, 3(3):311–351, 1961. doi: 10.1080/00401706.1961.10489951. URL https://doi.org/10.1080/00401706.1961.10489951.

Jesper B. Bramsen, Maria B. Laursen, Anne F. Nielsen, Thomas B. Hansen, Claus Bus, Niels Langkjær, B. Ravindra Babu, Torben Højland, Mikhail Abramov, Arthur Van Aerschot, Dalibor Odadzic, Romualdas Smicius, Jens Haas, Cordula Andree, Jharna Barman, Malgorzata Wenska, Puneet Srivastava, Chuanzheng Zhou, Dmytro Honcharenko, Simone Hess, Elke Müller, Georgii V. Bobkov, Sergey N. Mikhailov, Eugenio Fava, Thomas F. Meyer, Jyoti Chattopadhyaya, Marino Zerial, Joachim W. Engels, Piet Herdewijn, Jesper Wengel, and Jørgen Kjems. A large-scale chemical modification screen identifies design rules to generate siRNAs with high activity, high stability and low toxicity. *Nucleic Acids Research*, 37(9):2867–2881, 2009. ISSN 0305-1048. doi: 10.1093/nar/gkp106.

Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

Hugh Chen, Joseph D. Janizek, Scott Lundberg, and Su-In Lee. True to the model or true to the data? *arXiv preprint arXiv:2006.16234*, 2020. URL https://arxiv.org/abs/2006.16234.

Lu Chen, Siyu Lou, Keyan Zhang, Jin Huang, and Quanshi Zhang. HarsanyiNet: Computing accurate shapley values in a single forward propagation. In *International Conference on Machine Learning, PMLR*, pp. 4804–4825, 2023. URL https://proceedings.mlr.press/v202/chen23s.html.

Yuanting Chen, Mengyu Tong, Long Chen, Weihua Li, Keyun Zhu, Jiaying Li, Yun Tang, and Guixia Liu. MEG-mod: A multiview enhanced graph neural network for knockdown efficiency prediction of chemically modified siRNA. *Journal of Medicinal Chemistry*, 69:13434–13451, 2026. doi: 10.1021/acs.jmedchem.6c00411. URL https://doi.org/10.1021/acs.jmedchem.6c00411.

Alexander D'Amour, Katherine Heller, Dan Moldovan, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022. URL https://jmlr.org/papers/v23/20-1335.html.

Sarah M. Davis, Samuel Hildebrand, Hannah J. MacMillan, Kathryn R. Monopoli, Julianna Buchwald, Jacquelyn Sousa, David Cooper, Socheata Ly, Dimas Echeverria, Nicholas McHugh, Chantal Ferguson, Andrew Coles, Vignesh N. Hariharan, Daniel O'Reilly, Qi Tang, Raymond Furgal, Ken Yamada, Julia F. Alterman, James W. Gilbert, Emily Knox, Yamilett Pineda, Caitlyn N. Weston, Christina E. Baer, Athma A. Pai, and Anastasia Khvorova. Systematic analysis of siRNA and mRNA features impacting fully chemically modified siRNA efficacy. *Nucleic Acids Research*, 53 (12):gkaf479, 2025. ISSN 1362-4962. doi: 10.1093/nar/gkaf479.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7: 1–26, 1979. doi: 10.1214/aos/1176344552. URL https://doi.org/10.1214/aos/1176344552.

Sayda M. Elbashir, Jens Harborth, Winfried Lendeckel, Abdullah Yalcin, Klaus Weber, and Thomas Tuschl. Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature*, 411(6836):494–498, 2001. doi: 10.1038/35078107. URL https://doi.org/10.1038/35078107.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63:3–42, 2006. doi: 10.1007/s10994-006-6226-1. URL https://doi.org/10.1007/s10994-006-6226-1.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1263–1272, 2017. URL https://proceedings.mlr.press/v70/gilmer17a.html.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Michel Grabisch and Marc Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory*, 28:547–565, 1999. doi: 10.1007/s001820050125. URL https://doi.org/10.1007/s001820050125.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs, 2017. URL https://arxiv.org/abs/1706.02216. arXiv:1706.02216; author manuscript.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, et al. Array programming with NumPy. *Nature*, 585:357–362, 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

Sicheng He, Cheng Chen, Xianrun Pan, Gaogao Xue, Yu Yang, Juan Feng, Hasan Zulfiqar, Yang Zhang, and Kejun Deng. CMsiRNAdb: a database of chemically modified siRNA silencing efficiency for nucleic acid drug design. *BMC Bioinformatics*, 27:33, 2026. doi: 10.1186/s12859-025-06359-y. URL https://doi.org/10.1186/s12859-025-06359-y.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970. doi: 10.1080/00401706.1970.10488634. URL https://doi.org/10.1080/00401706.1970.10488634.

Dieter Huesken, Joerg Lange, Craig Mickanin, Jan Weiler, Fred Asselbergs, Justin Warner, Brian Meloon, Sharon Engel, Avi Rosenberg, Dalia Cohen, Mark Labow, Mischa Reinhardt, François Natt, and Jonathan Hall. Design of a genome-wide siRNA library using an artificial neural network. *Nature Biotechnology*, 23(8):995–1001, 2005a. doi: 10.1038/nbt1118. URL https://www.nature.com/articles/nbt1118.

Dieter Huesken, Joerg Lange, Craig Mickanin, Jan Weiler, Fred Asselbergs, Justin Warner, Brian Meloon, Sharon Engel, Avi Rosenberg, Dalia Cohen, Mark Labow, Mischa Reinhardt, François Natt, and Jonathan Hall. Corrigendum: Design of a genome-wide siRNA library using an artificial neural network. *Nature Biotechnology*, 23(10):1315, 2005b. doi: 10.1038/nbt1005-1315a. URL https://www.nature.com/articles/nbt1005-1315a.

Dieter Huesken, Joerg Lange, Craig Mickanin, Jan Weiler, Fred Asselbergs, Justin Warner, Brian Meloon, Sharon Engel, Avi Rosenberg, Dalia Cohen, Mark Labow, Mischa Reinhardt, François Natt, and Jonathan Hall. Corrigendum: Design of a genome-wide siRNA library using an artificial neural network. *Nature Biotechnology*, 24(8):1033, 2006. doi: 10.1038/nbt0806-1033e. URL https://www.nature.com/articles/nbt0806-1033e.

John D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9:90–95, 2007. doi: 10.1109/MCSE.2007.55. URL https://doi.org/10.1109/MCSE.2007.55.

Aimee L. Jackson, Steven R. Bartz, Janell Schelter, Sumire V. Kobayashi, Julja Burchard, Mao Mao, Bin Li, Guy Cavet, and Peter S. Linsley. Expression profiling reveals off-target gene regulation by RNAi. *Nature Biotechnology*, 21(6):635–637, 2003. doi: 10.1038/nbt831. URL https://doi.org/10.1038/nbt831.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4:100804, 2023. doi: 10.1016/j.patter.2023.100804. URL https://par.nsf.gov/servlets/purl/10513990.

Anastasia Khvorova, Angela Reynolds, and Sumedha D. Jayasena. Functional siRNAs and miRNAs exhibit strand bias. *Cell*, 115(2):209–216, 2003. doi: 10.1016/s0092-8674(03)00801-8. URL https://doi.org/10.1016/s0092-8674(03)00801-8.

Yoon Kim. Convolutional neural networks for sentence classification, 2014. URL https://arxiv.org/abs/1408.5882. arXiv:1408.5882; author manuscript.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. URL https://arxiv.org/abs/1412.6980. arXiv:1412.6980; author manuscript.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks, 2016. URL https://arxiv.org/abs/1609.02907. arXiv:1609.02907; author manuscript.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, et al. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning, PMLR*, pp. 5637–5664, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Valentina Kuskova, Dmitry Zaytsev, and Michael Coppedge. When are neural interaction discoveries real? identifiability, recoverability, and a pre-fit diagnostic. *arXiv preprint arXiv:2606.08390*, 2026. URL https://arxiv.org/abs/2606.08390.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles, 2016. URL https://arxiv.org/abs/1612.01474. arXiv:1612.01474; author manuscript.

Benjamin Lengerich, Sarah Tan, Chun-Hao Chang, Giles Hooker, and Rich Caruana. Purifying interaction effects with the functional ANOVA: An efficient algorithm for recovering identifiable additive models. In *International Conference on Artificial Intelligence and Statistics*, pp. 2402–2412, 2020. URL https://proceedings.mlr.press/v108/lengerich20a.html.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2017. URL https://arxiv.org/abs/1711.05101. arXiv:1711.05101; author manuscript.

Antonio Majdandzic and Peter K. Koo. Statistical correction of input gradients for black box models trained with categorical input features. In *ICML 2021 Workshop on Computational Biology*, 2021. URL https://icml-compbio.github.io/2021/papers/WCBICML2021_paper_56.pdf.

Antonio Majdandzic, Chandana Rajesh, and Peter K. Koo. Correcting gradient-based interpretations of deep neural networks for genomics. *Genome Biology*, 24:109, 2023. doi: 10.1186/s13059-023-02956-3. URL https://doi.org/10.1186/s13059-023-02956-3.

Jayaprakash K. Nair, Jennifer L. S. Willoughby, Amy Chan, Klaus Charisse, Md. Rowshon Alam, Qianfan Wang, Menno Hoekstra, Pachamuthu Kandasamy, Alexander V. Kel'in, Stuart Milstein, Nate Taneja, Jonathan O'Shea, Sarfraz Shaikh, Ligang Zhang, Ronald J. van der Sluis, Michael E. Jung, Akin Akinc, Renta Hutabarat, Satya Kuchimanchi, Kevin Fitzgerald, Tracy Zimmermann, Theo J. C. van Berkel, Martin A. Maier, Kallanthottathil G. Rajeev, and Muthiah Manoharan. Multivalent N-acetylgalactosamine-conjugated siRNA localizes in hepatocytes and elicits robust RNAi-mediated gene silencing. *Journal of the American Chemical Society*, 136(49):16958–16961, 2014. ISSN 0002-7863. doi: 10.1021/ja505986a.

C. Anders Olson, Nicholas C. Wu, and Ren Sun. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Current Biology*, 24(22):2643–2651, 2014. doi: 10.1016/j.cub.2014.09.072. URL https://doi.org/10.1016/j.cub.2014.09.072.

Adam Paszke, Sam Gross, Francisco Massa, et al. PyTorch: An imperative style, high-performance deep learning library, 2019. URL https://arxiv.org/abs/1912.01703. arXiv:1912.01703; author manuscript.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. URL https://jmlr.org/papers/v12/pedregosa11a.html.

Angela Reynolds, Devin Leake, Queta Boese, Stephen Scaringe, William S. Marshall, and Anastasia Khvorova. Rational siRNA design for RNA interference. *Nature Biotechnology*, 22(3):326–330, 2004. doi: 10.1038/nbt936. URL https://doi.org/10.1038/nbt936.

Rachapun Rotrattanadumrong and Carlo De Donno. OligoGym: Curated datasets and benchmarks for oligonucleotide drug discovery. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/4927dd556ff6e49671e3a77ac1f12487-Abstract-Datasets_and_Benchmarks_Track.html.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks, 2017. URL https://arxiv.org/abs/1703.06103. arXiv:1703.06103; author manuscript.

Dianne S. Schwarz, György Hutvágner, Tingting Du, Zuoshang Xu, Neil Aronin, and Phillip D. Zamore. Asymmetry in the assembly of the RNAi enzyme complex. *Cell*, 115(2):199–208, 2003. doi: 10.1016/s0092-8674(03)00759-1. URL https://doi.org/10.1016/s0092-8674(03)00759-1.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. URL https://arxiv.org/abs/1312.6034.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014. URL https://www.jmlr.org/papers/v15/srivastava14a.html.

Jiyuan Tan, Jose Blanchet, and Vasilis Syrgkanis. Consistency of neural causal partial identification. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2405.15673. Updated arXiv version 3, June 2025.

Wenchong Tan, Mingshu Dai, Shimin Ye, Xin Tang, Dawei Jiang, Dong Chen, and Hongli Du. ENsiRNA: A multimodality method for siRNA-mRNA and modified siRNA efficacy prediction based on geometric graph neural network. *Journal of Molecular Biology*, 437:169131, 2025. doi: 10.1016/j.jmb.2025.169131. URL https://pubmed.ncbi.nlm.nih.gov/40194620/.

Ammar Tareen, Mahdi Kooshkbaghi, Anna Posfai, William T. Ireland, David M. McCandlish, and Justin B. Kinney. MAVE-NN: learning genotype-phenotype maps from multiplex assays of variant effect. *Genome Biology*, 23:98, 2022. doi: 10.1186/s13059-022-02661-7. URL https://doi.org/10.1186/s13059-022-02661-7.

Kumiko Ui-Tei, Yuki Naito, Fumitaka Takahashi, Takeshi Haraguchi, Hiroko Ohki-Hamazaki, Aya Juni, Ryu Ueda, and Kaoru Saigo. Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference. *Nucleic Acids Research*, 32(3):936–948, 2004. doi: 10.1093/nar/gkh247. URL https://doi.org/10.1093/nar/gkh247.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2. URL https://doi.org/10.1038/s41592-019-0686-2.

Mandy S. Wong, Justin B. Kinney, and Adrian R. Krainer. Quantitative activity profile and context dependence of all human 5' splice sites. *Molecular Cell*, 71(6):1012–1026.e3, 2018. doi: 10.1016/j.molcel.2018.07.033. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC6179149/.

C. F. J. Wu and Youyi Chen. A graph-aided method for planning two-level experiments when certain interactions are important. *Technometrics*, 34(2):162–175, 1992. doi: 10.1080/00401706.1992.10484905. URL https://doi.org/10.1080/00401706.1992.10484905.

Kevin Xia, Yushu Pan, and Elias Bareinboim. Neural causal models for counterfactual identification and estimation. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.00035.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ryGs6iA5Km.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3391–3401, 2017. URL https://papers.nips.cc/paper/6931-deep-sets.pdf.

A1 DATA, CHEMISTRY, EXPERIMENT, AND GROUPING

Appendix roadmap and evidence map. Data, chemistry and partitions: Section A1; full forward map, losses and derivatives: Sections A2–A3; metrics and uncertainty: Section A4; primary-panel evidence: Section A5.1; adjudication: Section A8. Table 2 also uses the complete decomposition proof (A10.7). Table 3 and Figures 3, 6 resolve to the endpoint/marginal tables, exact rank proof (A10.6), chemical factorization (A10.11.1) and sensitivities (A8.5). The perturbation diagnostics quoted in Section 8 resolve to the induction in Section A10.3, the numerical utility protocol in Section A10.20 and rows of `perturbation_summary`. Figure 1 resolves to the architecture; the focused grouped sensitivity plot in the canonical store carries the matched activity and exact metric decomposition. Table 22 and Figure 7 resolve to the complete utility protocol (A10.20); the independent attempt is in Section A10.18. The canonical report supplies each object's store/CSV mapping; every required mathematical argument is written here.

A1.1 NOTATION AND THE OBSERVED EXPERIMENT

Definition A1.1 (Observation and endpoint)**.** *An observation i consists of source identifier u_i, guide and passenger sequences $s_i = (s_i^g, s_i^p)$ in reported $5'$–$3'$ order, chemistry state m_i, assay metadata a_i, and a reported activity y_i. Chemistry includes positional modification names, available linkage types, terminal groups, and reported stereochemistry. Metadata include target, accession, cell/species, concentration, time, delivery, table, and response normalization where recorded. Unavailable fields remain missing. A parsed observation is not assumed to be an independent biological replicate.*

Symbol	Meaning
y_i	Inhibition fraction, by cohort: the grouped benchmark uses the released inhibition fraction; APP uses $(100 -$ primary mean percent mRNA remaining$)/100$; the primary B3 panel uses $1 -$ relative eGFP (Appendix A5.1); Davis S7 is kept on its original response scale. No clipping.
$\mu(s, m, a)$	Population mean of that assay endpoint; not directly known from a reported sample mean.
$\hat{f}_k(x_i)$	Committed prediction of model k from the permitted molecule/context features x_i.
$g(i), q(i)$	Study/source-linked component and sequence-only component, respectively.
w_i	Within-partition observation weight $n/(Gn_{g(i)})$, where G is the number of study components and n_g is its size.
X, A, b, C	Node features, relation adjacency tensor, node mask, and training-masked numeric context; here C is the context tensor, not the contrast matrix C of Section 2.
L, M, F, R, d	Maximum strand length 32; total node slots 64; 93 node features; 8 relation types; hidden width 32.
a, S_a, S_a^*	Ranking pool, five model-selected candidates, and five highest measured candidates.
$d_j, \hat{d}_j$	Measured and predicted differences for matched chemistry pair j.
B	Number of sequence-component bootstrap draws, fixed at 10,000 in the current revision.

Table 5: Notation and conventions.

This is a retrospective, heterogeneous observation collection. No shared latent Gaussian experiment, independent five-stratum sampling model, private-anchor law, pathwise differentiability, or unknown-profile search class is assumed. Ordinary finite moments are enough to define empirical losses because all admitted numeric values and saved predictions are finite. We do not infer population MSE bounds from those finite empirical sums.

A1.2 Row eligibility and identity reconciliation

All raw rows are preserved before exclusions. ENsiRNA source identifiers must be seven-to-nine-digit numeric publication identifiers or the explicit US20120088815A1/EP2415869A1 family string. Other release identifiers remain unresolved. Guide and passenger bases must consist of uppercase A, C, G, U, T. Source, both complete parsed chemistry states, reported concentration, and outcome define an exact-release duplicate; repeated observations with distinct outcomes remain measurements but are grouped together for splitting. Two exact duplicate records are removed.

Each concentration must be positive and each outcome finite. Each strand must admit the graph alignment specified below. Chemistry annotations with inconsistent modification/position arity, indices outside the strand, unknown base changes, or unresolved geometry-changing entries are excluded. The strings "Mutation", "Inverted abasic", and "Assymetric siRNA (3-End Overhang)" are not silently translated into canonical bases or guessed duplex structures.

APP rows are first excluded if the cell field is "Transgenic mice" or administration time is not exactly 24h. Empty chemistry is not unmodified RNA. Three description/base disagreements are rejected, as are unsupported base strings, nonpositive dose, nonfinite outcome, or failed duplex alignment. Exclusions are assigned the first failing reason in this order, so counts are mutually exclusive within the parser. A row with several deficiencies is not counted several times.

The first reconciliation admits 5,033 rows: 2,927 ENsiRNA and 2,106 APP. The primary-table step subsequently excludes 267 APP rows for unreconciled exact chemical strings, leaving 4,766. Its separate status supersedes the intentionally preserved earlier-stage "pending" B2 field. No scientific fit used the earlier population.

A1.3 Complete chemical parser

For ENsiRNA, `anti raw seq`/`sense raw seq` give bases, `anti mod`/`sense mod` list modification names separated by an asterisk, and matching position fields list comma-separated one-based indices within each modification group. A position token is converted to an integer and checked against strand length. The literal zero/zero pair denotes "unmodified as annotated"; an empty field does not. Multiple entries at a position become a sorted set, and an otherwise untagged position is recorded as unmodified within the release's annotation scope. This is not a claim of complete physical chemical characterization.

Only explicit aliases are merged: 2-O-Methyl ribose and 2-Methoxy become 2-O-Methyl; 2-Deoxy-2-Fluoro becomes 2-Fluoro; 2-Deoxyribonucleotide and 2-Deoxythymidine become 2-Deoxy; Un-Locked nucleic acid becomes Unlocked nucleic acid; 5-phosphate ribose becomes 5-Phosphate. Other names remain distinct, including source spelling. Indexed phosphorothioate and boranophosphate tags remain node features with an unresolved direction flag, and ENsiRNA backbone edges are encoded as unresolved under their annotation-only PO default, so an indexed tag is never promoted to a resolved directed PS bond. Terminal annotations are retained for 5-Phosphate and Dodecyl derivative; unreported stereochemistry stays unreported.

APP indexed descriptions are separated by `||`, with each nucleotide item `index*description`. Base words adenosine, guanosine, cytidine, uridine and thymidine must agree with the raw base string. Sugar chemistry is selected in the order hexadecyl, Glycol Nucleic Acid, Deoxy, Methyl, Fluoro; a description matching none is rejected and every position must be described exactly once. A phosphorothioate descriptor tags its node and the following within-strand bond, becoming terminal PS at the final position, while unmarked bonds are PO under this convention. A vinyl descriptor adds a vinylphosphonate tag and terminal entry. The raw token `(Tgn)` identifies the reported S-GNA state; other stereochemistry remains unreported.

An unnumbered GalNAc description is localized to the final position only if the raw chemistry ends in `L96`; indexed GalNAc descriptions retain their stated index, and L96, vinylphosphonate, phosphate and other terminal names are kept in parsed records. Graph terminal coordinates are capped at the last observed nucleotide as an implementation convention, with original coordinates and names retained in provenance. No chemical string is lowercased, since case can encode sugar modifications, and T and U remain distinct model bases whose equivalence is used only in sequence grouping and canonical pairing recognition.

A1.4 EXACT PRIMARY-TABLE VERIFICATION

The APP rows come from two Alnylam patent families, US20220002724A1 (priority 19 December 2018) and US20240117349A1 (priority 29 January 2021); the CMsiRNAdb APP release also carries an Ionis family, US20220380773A1, which is not admitted here. For US20220002724A1, extracted tables 4 and 7 provide means and SDs at 10 and .1nM for two cell types each. Tables 17 and 18 provide agent, mean, SD, and dose for Be(2)C and Neuro2A, respectively. Table 29 is also parsed but contributes no final admitted pool. US20240117349A1 Table 6 provides Be(2)C means and SDs at 10, 1, and .1nM. Lookup keys are publication, agent, exact cell label, and numeric dose. Tables are parsed in the foregoing order; later occurrences replace earlier identical keys. Every admitted row also agrees with its released inhibition within .015 percentage points. There are zero primary-label corrections in the final records.

Chemical strings are reconstructed from wrapped primary rows, joining continuation cells until a blank row or a changed row width. Spaces are removed for string comparison; case and chemical tokens are preserved. An exact agent match must yield the exact guide and passenger chemical strings in the database export. The accepted numeric response is $(100 - \text{primary mean remaining})/100$, with the primary SD retained and divided by 100 when used on the activity scale. Negative inhibition is retained. Tables 17/18 identify the AD-1955-relative mRNA endpoint in accompanying source text. No assumption that all normalized assays have identical biological scales is made.

The final 17 assay pools are: Table 4, two cells times two doses, 209 rows each; Table 7, two cells times two doses, 91 each; Tables 17 and 18, one cell times three doses each, 90 each; US20240117349A1 Table 6, one cell times three doses, 33 each. These 17 pools hold $209 \cdot 4 + 91 \cdot 4 + 90 \cdot 6 + 33 \cdot 3 = 1{,}839$ measurement rows over a far smaller number of distinct duplexes, so the cohort size is a row count, not a count of siRNAs. All use the corresponding source RNAiMAX screening protocol and 24h incubation. Primary tables contain endpoint SDs but do not resolve replicate n, raw wells, or covariance from shared normalizers for these admitted records. We do not infer these quantities from the SD.

A1.5 PROTECTED COMPONENTS AND CURRENT GROUPED PARTITIONS

Two transitive closures define the grouping units. The study-linked union-find joins all observations sharing a known source family and all observations connected by a shared strand 13-mer. The sequence-only union-find uses the latter criterion alone. Both strands contribute contiguous 13-mers after replacing T by U; a match can connect opposite strands or connect records indirectly. This is a conservative sequence-identity safeguard, not a proof of biological independence. Group identities are deterministic hashes of the retained memberships, and source, record and outcome-bearing identifiers never become model inputs.

The existing five outer ENsiRNA test groups are retained exactly. Every protected study-linked component occurs in one outer test with no sequence or source-family intersection with that test's training or validation partition; historical internal-test observations are retired into this grouped development population, and the APP patent family and all Davis S7 rows remain outside ENsiRNA model selection.

Within each outer training population, sort whole study-linked components by decreasing number of observations, breaking ties by their identifier. Assign the next component to the currently smallest of three inner folds, ties broken by index, and for each inner validation fold train on the other components; no protected component is split to balance counts. The deployment inner design applies the same construction to all ENsiRNA components. The final validation fold is the one with fewest validation observations, with an index tie-break, so the choice uses only membership counts; final-seed fits use its complementary training components and monitor this grouped holdout. The deployment fit therefore updates parameters on a declared subset of ENsiRNA while the remaining rows serve validation, and there is no all-data refit.

Pair outer tests retain the previous assignment of the twelve sequence components to four tests. Three inner folds are greedily balanced by pair count with whole sequence components protected, endpoint training uses only the measured endpoints in the appropriate component set, and the final validation fold follows the same smallest-count rule. Both endpoint identities and all their sequence-

linked observations are excluded from the held-out fold's training and selection; membership tables give every role, and row counts and group counts are distinct.

A2 Exact representation, forward map and support claims

A2.1 Graph construction and all coordinates

For each duplex, guide position i occupies slot i and passenger position j slot $32 + j$, using zero-based indices. Strands longer than 32 are rejected, never truncated. $X \in \mathbb{R}^{64\times 93}$ and binary active-node mask $b \in \{0,1\}^{64}$ represent these ordered slots. Coordinates 0–4 are A,C,G,U,T, coordinates 5–6 identify strand, and 7–38 give within-strand position. Coordinates 39–83 are the 45 positional modification names of the fixed schema, enumerated in the saved preprocessing support and the preserved v10 appendix rather than reprinted here. Coordinates 84–88 indicate unreported stereochemistry, S-GNA, unresolved indexed-linkage direction, strand start and strand end; 89–92 indicate L96 GalNAc, vinylphosphonate, 5-phosphate and other terminal groups. Schema exposure included historical label-free test molecules; this is disclosed rather than described as a training-only vocabulary. Effective masks in the current campaign are fitted separately on each training partition.

Adjacency $A \in \{0,1\}^{8\times 64\times 64}$ uses $A_{r,v,u} = 1$ for source u and receiver v. For adjacent positions, indices 0,1,2 encode forward PO, PS and unresolved linkages, and 3,4,5 their reverse directions. The annotation-only ENsiRNA PO assumption is unresolved, including indexed PS tags whose bond direction was not established. Relations 6/7 are bidirectional canonical/mismatched pairing. Counts sum directed tensor entries, so each paired strand position contributes two edges and a backbone bond contributes one forward and one reverse edge. Padded nodes contribute no edges.

For guide/passenger lengths l_g, l_p, enumerate offsets $o = -l_p + 1, \ldots, l_g - 1$. Match guide i to passenger $j = l_p - 1 - (i - o)$ whenever both indices are valid. Require overlap at least 14, then maximize lexicographically the number of Watson–Crick matches, negative mismatch count, overlap, negative absolute offset and negative offset. Complementarity uses A–U and C–G, treating T as U only for matching. Admit a selected pairing fraction of at least .8. The inferred alignment, not an experimental tertiary structure, supplies these edges. Overhangs have no pairing edge, and there is no target-mRNA or inter-observation graph.

A2.2 Training-fitted masks and context

Let T contain only the current training indices. A node column is enabled if it is nonzero at any active training node; otherwise set that column to zero on every split. Padding is excluded from this support computation. This mask is identical for original/corrected graph and no-message arms. It may discard meaningful chemistry that has no training support, including shifted-cohort terminal flags. No held-out outcome or held-out support count selects the mask.

The eight raw context coordinates are verified $\log(1 + c)/\log(101)$ in nM (zero when the unit is unverified), an unverified-unit flag, time divided by 24 (zero when missing), a missing-time flag, a recorded-cell flag, inferred pairing fraction, guide length/32 and passenger length/32. Every coordinate constant across the training partition is set to zero on all splits. No test-fitted centering, variance denominator or clipping is used. The recorded-cell flag is not a cell-identity embedding, and an unverified dose is not converted into nM. This is the conservative historical C1 rule, held fixed across the primary four-arm intervention.

A2.3 Original and corrected relation transforms

Every graph or no-message model owns two tensors $W^\ell \in \mathbb{R}^{8\times 32\times 32}$. R0 uses these matrices directly. R1 uses

$$\widetilde{W}_0 = W_2 + q_0 W_0, \qquad \widetilde{W}_1 = W_2 + q_1 W_1, \qquad \widetilde{W}_2 = W_2, \tag{A2.1}$$

$$\widetilde{W}_3 = W_5 + q_3 W_3, \qquad \widetilde{W}_4 = W_5 + q_4 W_4, \qquad \widetilde{W}_5 = W_5, \tag{A2.2}$$

$$\widetilde{W}_6 = q_6 W_6, \qquad \widetilde{W}_7 = q_7 W_7, \tag{A2.3}$$

where $q_r = \mathbf{1}\{\sum_{i\in T,u,v} A_{irvu} > 0\}$. This is the complete fallback rule. A new explicit PO or PS edge routes through the common base for its direction; its residual is used only when that type

appeared in training. Every training backbone edge uses the corresponding base through the sum above, even if no edge is labeled unresolved. Its realized gradient contribution also depends on the loss and activation state. A positive count is eligibility only: a relation represented by two edges is still rare. Pairing relations have no invented backbone fallback; they are gated separately.

A2.4 Forward pass and exact parameter count

All hidden states are row vectors. With D denoting dropout at .1 and LN_ℓ affine LayerNorm,

$$h_i^0 = b_i \,\mathrm{ReLU}(X_i E + e), \tag{A2.4}$$

$$d_i = \max\left(1, \sum_{r,j} A_{rij}\right), \tag{A2.5}$$

$$m_i^\ell = d_i^{-1} \sum_{r,j} A_{rij} h_j^\ell \widetilde{W}_r^\ell, \tag{A2.6}$$

$$h_i^{\ell+1} = b_i \,\mathrm{LN}_\ell\left[h_i^\ell + D\{\mathrm{ReLU}(h_i^\ell U_\ell + u_\ell + m_i^\ell)\}\right], \quad \ell = 0, 1. \tag{A2.7}$$

R0 replaces $\widetilde{W}$ by W. Concatenate the 64 states in slot order and the eight context coordinates to obtain $v \in \mathbb{R}^{2056}$. The scalar normalized output is

$$z = D\{\mathrm{ReLU}(vW_o + b_o)\}\beta + b, \qquad \hat{y} = \mu_T + s_T z. \tag{A2.8}$$

For a 32-coordinate state u, the normalization is

$$\mathrm{LN}(u)_k = \gamma_k \frac{u_k - \bar{u}}{\sqrt{32^{-1}\sum_{j=1}^{32}(u_j - \bar{u})^2 + 10^{-5}}} + \beta_k, \qquad \bar{u} = 32^{-1}\sum_{j=1}^{32} u_j.$$

The variance divisor is 32, not 31. In training, $D(u)_k = M_k u_k/.9$ with Bernoulli(.9) masks drawn by the recorded PyTorch random stream; in evaluation $D(u) = u$. ReLU is coordinatewise $\max(0, u_k)$. These definitions fix the actual normalization and dropout conventions.

There is no sigmoid, output clipping, global pooling or permutation-invariance claim. LayerNorm uses PyTorch's 10^{-5} epsilon over the 32 state coordinates and learned scale/bias. For a linear layer with fan-in f, weights and biases initialize uniformly on $[-f^{-1/2}, f^{-1/2}]$. Width-three convolution uses fan-in $32 \cdot 3 = 96$. LayerNorm scales initialize to one and biases to zero. Each full $(8, 32, 32)$ relation tensor receives Xavier uniform initialization with fan-in $32 \cdot 32 = 1024$ and fan-out $8 \cdot 32 = 256$, hence bound $\sqrt{6/1280}$; this is not eight independently initialized $(32, 32)$ matrices. Dropout uses inverted scaling in training and is disabled for prediction and the post-fit data-gradient diagnostic.

The embedding has $93 \cdot 32 + 32 = 3,008$ scalars, self transforms $2(32^2 + 32) = 2,112$, normalization $2(2 \cdot 32) = 128$, relation tensors $2 \cdot 8 \cdot 32^2 = 16,384$, and readout $(2056 \cdot 64 + 64) + (64 + 1) = 131,713$. Their sum is 153,345. R1 reuses the owned tensors and adds no parameters. Structurally unsupported residuals remain in the optimizer; total parameter count is therefore not a count of effectively data-trained scalars. Counts, data gradients and parameter changes are all reported separately. The final audit counts exactly nonzero scalar gradients of both the full training activity loss and the actual training objective at every selected final neural checkpoint, with dropout disabled. The latter includes the selected pair term for pair-supervised fits. It also compares every scalar to the same seeded initialization. These realized diagnostic counts and the structural upper bounds are distinct: weight decay can change a zero-gradient scalar, and a scalar with zero gradient at selection could have received data gradients earlier. No full-trajectory ever-active count is claimed.

Proposition A2.1 (Unsupported route and the no-message intervention)**.** *In R0, a relation absent from every training graph has identically zero training-data gradient through its relation matrix at every layer. A parameter can nevertheless change under decoupled weight decay. In the no-message arm, each receiver's hidden state depends only on its own initial features, its relation-degree vector and the common fitted parameters; it has no dependence on other nodes' hidden states through message exchange.*

Proof. For a matrix entry $(W_r^\ell)_{ak}$ the message derivative is $\partial m_{ik}^\ell / \partial (W_r^\ell)_{ak} = d_i^{-1} \sum_j A_{rij} h_{ja}^\ell$. When $A_r = 0$ throughout training, every summand is zero. All downstream derivatives, including

weighted residuals, multiply this zero derivative. The conclusion holds for activity loss and for a sum of endpoint difference losses, because both endpoints are training graphs. At ReLU boundaries the implementation uses its fixed subgradient; the zero multiplicative factor remains zero. AdamW additionally multiplies an updated parameter by a weight-decay factor, so a data gradient of zero is not a statement of zero numerical parameter change.

The no-message aggregation is $\sum_j A_{rij} h_i^\ell = (\sum_j A_{rij}) h_i^\ell$. Its derivative with respect to h_j^ℓ is zero for $j \neq i$. At layer zero the receiver is a function of X_i alone. Substitution into the displayed update proves by induction for both layers that no other node state enters that receiver. The ordered global readout still uses all receiver outputs. Thus this is a precise neighbor-exchange intervention, not a claim that the final predictor ignores the rest of the molecule.

In R1, each directional backbone message contains W_2 or W_5, so its derivative with respect to the base matrix sums over all backbone types in that direction. An unsupported type's residual is multiplied by $q_r = 0$. The fallback therefore supplies a train-supported computational route whenever training contains any backbone edge in that direction. Nonzero support does not guarantee a nonzero realized gradient under every loss/activation state, nor improved test risk. □

The historical four absent transforms per layer account for $4 \cdot 2 \cdot 32^2 = 8,192$ scalars. They are a subset of the 16,384 relation scalars. The count applies to that training partition and encoding; it is not a universal count for every pair-trained or corrected fit.

A2.5 NON-GRAPH BASELINES AND ARITHMETIC

The token CNN embeds 93 to 32 coordinates, applies two 32-to-32 width-three Conv1d layers with padding one and ReLU, with .1 dropout between them. It adds their output to the embedded states, masks padding and uses the same ordered readout. Its 140,929 trainable scalars include $3,008 + 2(32 \cdot 32 \cdot 3 + 32) + 131,713$. The convolution operates on all 64 slots in fixed order, including the padded gap, and does not read adjacency.

The chemistry tree receives flattened masked node features, each node's eight relation degrees and context. Columns constant in the training partition are removed. Extra-Trees uses 300 estimators, squared-error splitting, no bootstrap, feature fraction .3, no maximum depth and leaf size 2 or 8; minimum split size is two, minimum weighted leaf fraction and impurity decrease are zero, leaf count is unrestricted, pruning alpha is zero, and warm start, out-of-bag scoring and monotonic constraints are disabled. Guide ridge receives 160 guide base/position indicators and eight context fields; pairwise ridge adds all 276 position pairs among the first 24 guide slots, each with 25 ordered base categories, and both omit passenger chemistry. Nonconstant columns are standardized with unweighted training means/SDs, replacing SD below 10^{-6} by one. Ridge uses weighted residual sum of squares, $\alpha \in \{.1, 1, 10\}$, LSQR tolerance 10^{-4}, no maximum-iteration limit, a copied design matrix, no positivity constraint and an unpenalized intercept. The iterative solver may terminate at its installed automatic iteration limit; no exact-arithmetic minimizer certificate is asserted. These are explicit current settings, not an assertion that the historical ridge settings were identical.

For exact matched pairs with equal guide/context inputs, any deterministic chemistry-invariant predictor returns equal endpoints and hence a zero difference, by equality of inputs and deterministic evaluation rather than a biological no-effect argument. Floating-point differences near machine precision are not chemistry evidence.

Neural arithmetic is float32; exported metrics are float64. Dense graph storage costs $O(RM^2 + MF)$ per observation, one layer $O(RM^2 d + RMd^2)$ operations and the ordered readout $O(Md \cdot 64)$. These follow from the stated matrix products and are not certified bit-complexity or wall-time bounds. No numerical integration, quantile estimator, fitting certificate or unrelated theoretical backend is used.

A3 COMPLETE SELECTION, FITTING AND CHECKPOINT PROTOCOL

A3.1 EXPERIMENT MATRIX AND INPUT ROLES

The historical v3 experiment matrix enumerates every population, method, configuration, development fold and seed before fitting. The five historical ENsiRNA outer tests are unchanged. The sixth activity population supplies APP/S7 follow-up prediction; its development groups are ENsiRNA only. Every outer training population has three inner grouped validation folds as defined in Appendix A1. We report the final training/validation partition separately from the total development population. Protected groups are never subdivided to achieve a desired row count.

The four controlled arms are original GNN, training-gated GNN, original no-message and corrected no-message. The token CNN, chemistry tree and guide/pairwise ridge are matched on admissible observation roles. All four graph arms use the same input masks, context, losses, initialization convention, parameter count, sampling and optimization budget. Their independently selected configuration is part of the fitted procedure. Thus the loss interaction compares four declared learning procedures, not a post hoc perturbation of one fixed parameter vector.

Development seeds are 701 and 907. The six neural configurations cross learning rates $10^{-4}, 3 \cdot 10^{-4}, 10^{-3}$ with weight decays $10^{-6}, 10^{-4}$. For each configuration, average the equal-protected-component validation MSE over three folds and both development seeds; select the minimum, breaking ties by the lexicographic configuration name. The tree crosses leaves 2 or 8 with the same development folds/seeds. Ridge selects $\alpha = .1, 1, 10$ over the three folds and is deterministic. Every selected configuration and competing score is preserved in `v3/tables/selection_scores.csv`.

The final seeds, including all unfavorable finite outcomes, are

$$1103,\ 2207,\ 3301,\ 4409,\ 5519,\ 6637,\ 7753,\ 8867,\ 9973,\ 11027. \tag{A3.1}$$

They are matched across the controlled arms. Deterministic methods are fitted once per selected partition. These are ten independently initialized training runs on fixed data, not ten independent data samples. The original three-seed campaign is not silently reused: its selection, schedules and input procedures differ from this frozen revision.

A3.2 TARGETS, WEIGHTS AND ACTIVITY OBJECTIVE

For training partition T with n observations, G protected groups, group membership $g(i)$ and counts n_g, define

$$w_i = \frac{n}{G n_{g(i)}}, \qquad \mu_T = \frac{1}{n}\sum_{i\in T} w_i y_i, \qquad s_T = \max\left\{.05, \sqrt{\frac{1}{n}\sum_{i\in T} w_i (y_i - \mu_T)^2}\right\}. \tag{A3.2}$$

Each group has total weight n/G, so $n^{-1}\sum_i w_i = 1$. Study-linked groups define activity weights, while pair-endpoint training uses sequence components. Targets for the normalized neural output are $\widetilde{y}_i = (y_i - \mu_T)/s_T$. Both location and scale are computed from training labels alone and saved with the checkpoint. Prediction restores $\hat{y}_i = \mu_T + s_T z_i$; no test mean, test variance or clipping enters this map. Activity tree and ridge estimators fit original-scale y_i directly; their recorded training means/SDs serve metadata and constant controls, not a hidden target normalization. Direct pair ridge/tree likewise fit original-scale d_p.

The full activity objective is $L_{\text{act}} = n^{-1}\sum_{i\in T} w_i (z_i - \widetilde{y}_i)^2$. Ordinary activity fitting draws a fresh uniform permutation of training rows each epoch and uses batches of at most 128. Each batch minimizes the mean of its weighted squared errors; the last smaller batch uses its own batch denominator. This specifies the actual stochastic objective, including the last-batch convention. No balanced-group sampler or outcome-based resampling is used.

A3.3 OPTIMIZATION, SELECTION AND RESUME

AdamW uses the selected learning rate and weight decay, moment coefficients $(.9, .999)$, denominator epsilon 10^{-8}, no AMSGrad, no maximization, and the installed automatic foreach/fused selection (both arguments left at None). Capturable and differentiable modes are false and there is

no learning-rate scheduler. Weight decay applies to every optimizer parameter including biases and normalization scales, global gradient norm is clipped to five before each update, and loss and gradient finiteness are checked. Default linear initialization, whole-tensor Xavier relation initialization, NumPy/Python/PyTorch seeds and a dedicated CPU permutation generator seeded by final/development seed plus 10,000 make the random sources explicit, with deterministic PyTorch algorithms enabled. GPU0 is used when available with a .30 allocator fraction and two CPU threads; no second GPU or paid resource is acquired.

Every epoch evaluates validation with dropout disabled. The selection score is the validation equal-group MSE on the original activity scale for ordinary activity models. The best checkpoint updates only on improvement greater than 10^{-8}, and training ends at 500 epochs or after at least 50 epochs and 50 epochs since the best qualifying improvement. A best checkpoint from epoch one is retained if it remains best, and later execution is not misrepresented as later selected training. Final-seed models select checkpoints on the predeclared final grouped holdout and are not refitted on it.

The best checkpoint stores weights, optimizer state, epoch/update count, target mean/SD, preprocessing and the full fit signature. The last checkpoint adds epoch history, best score/epoch, permutation-generator state and CPU/CUDA random states for continuation, and atomic temporary-file replacement avoids advertising a partial checkpoint as complete. A fit is reusable only when protocol, code, preservation-input manifest, train/validation/test indices, method, configuration, seed, loss coefficient and selection criterion match and each output hash verifies; a stage likewise verifies its code, dependency-stage hashes and output hashes. The resumable entry point continues missing stages, and a hash check is not a new fit.

A nonfinite training loss, gradient or validation score is recorded as a failed fit with its history and state; no poor finite seed is removed or restarted on a held-out score. The declared failure output is a flagged training-fold mean prediction with failed-fit rate reported separately, a technical preflight failure is distinguished from a training failure with its command and cost retained, and no configuration or budget expansion is selected from APP/S7 predictions.

A3.4 Controls and implementation checks

Zero difference is an input-invariant prediction control. The training-fold mean difference uses equal sequence-component weights on training pairs. Majority direction chooses the largest training weight among negative, operational tie and positive calls, breaking an exact mass tie toward the smaller sign. It is a direction-only rule; its numeric encoding $-1, 0, 1$ is not treated as a regression effect magnitude in the current results. The fixed band is $|d| \leq .02$ for operational ties. Neither the band nor majority direction uses the held-out fold.

Preflight checks verify protected group exclusion, input-hash preservation, training-only feature support, the exact unseen-relation route, masked padding, nonzero neighbor dependence for graph aggregation and zero off-receiver derivative for the no-message aggregation. Pair labels are checked against their endpoint subtraction. A tied constant is checked to yield exact random-tie ranking. Recorded optimizer updates, finite gradients, selected parameter changes and data-only relation gradients supply execution evidence; a global weight change is not used as proof that every transform learned.

A4 Metric definitions, uncertainty and all finite-sample identities

A4.1 Activity estimands and constants

Let $w_i \geq 0$ sum to one on the declared evaluated observations. Row weighting uses $w_i = 1/n$; equal-component weighting assigns each component mass $1/G$ and divides that mass equally among

its records. For fixed predictions $\hat{y}_i$ define

$$\mathrm{MSE}_w = \sum_i w_i(\hat{y}_i - y_i)^2, \qquad \mathrm{MAE}_w = \sum_i w_i|\hat{y}_i - y_i|, \tag{A4.1}$$

$$\bar{y}_w = \sum_i w_i y_i, \qquad \bar{\hat{y}}_w = \sum_i w_i \hat{y}_i, \tag{A4.2}$$

$$V_y = \sum_i w_i(y_i - \bar{y}_w)^2, \qquad V_{\hat{y}} = \sum_i w_i(\hat{y}_i - \bar{\hat{y}}_w)^2, \tag{A4.3}$$

$$R^2_w = 1 - \mathrm{MSE}_w / V_y, \qquad \mathrm{bias}_w = \bar{\hat{y}}_w - \bar{y}_w. \tag{A4.4}$$

Prediction/label SDs are $\sqrt{V_{\hat{y}}}$ and $\sqrt{V_y}$ with population denominators, not uncertainty in a biological mean. Weighted correlation is the weighted centered cross-product over $\sqrt{V_y V_{\hat{y}}}$; a zero or negligible denominator is reported undefined and no finite correlation is imputed for a constant. Squared error and R^2 use the same observations and weighting.

The training-row constant is the arithmetic mean of training activity; the equal-training-group constant uses the training weights from Appendix A3. Out-of-fold constants are recomputed in each corresponding training partition. The oracle test mean $\bar{y}_w$ is a separately marked diagnostic. It uses evaluation labels and is not deployed, included in model selection, or described as a statistical lower bound.

Proposition A4.1 (Constant and ensemble decompositions)**.** *For every finite weighted dataset,* $\sum_i w_i(y_i - c)^2 = V_y + (c - \bar{y}_w)^2$. *If* $\bar{f}_i = K^{-1}\sum_{k=1}^K f_{ik}$ *averages* K *fixed fitted predictions, then*

$$\frac{1}{K}\sum_k \sum_i w_i(f_{ik} - y_i)^2 = \sum_i w_i(\bar{f}_i - y_i)^2 + \sum_i w_i \frac{1}{K}\sum_k (f_{ik} - \bar{f}_i)^2. \tag{A4.5}$$

Proof. Write $y_i - c = (y_i - \bar{y}_w) + (\bar{y}_w - c)$, expand the square and use $\sum_i w_i(y_i - \bar{y}_w) = 0$. For the second identity write $f_{ik} - y_i = (f_{ik} - \bar{f}_i) + (\bar{f}_i - y_i)$. The cross-term vanishes after averaging over k because $\sum_k (f_{ik} - \bar{f}_i) = 0$ for each i. Summing weighted residual squares gives the expression. The nonnegative final term explains why mean seed risk can exceed ensemble risk. It does not certify uncertainty coverage or improved future risk. □

The first identity makes the empirical oracle mean the best constant on that weighted test sample, not a lower bound on nonconstant prediction risk. Poor R^2 may coexist with a meaningful relative MSE difference but limits practical usefulness, motivating the spread, bias and ranking diagnostics, and no test-fitted centering correction is presented as a model prediction. Calibration plots bin each model/cohort by prediction quantiles into up to five bins, dropping duplicate cut points, then average prediction and measured activity within each nonempty bin. This uses evaluation labels for a descriptive display, not for a fitted correction.

A4.2 Direct matched loss contrasts and conditional bootstrap

For predictors a, b evaluated on identical record IDs, retain each loss difference $D_i = (\hat{y}_i^a - y_i)^2 - (\hat{y}_i^b - y_i)^2$ with sequence, study, source and pool identifiers in the activity loss export (and endpoint/background/fold identifiers for B2). Negative mean difference favors a. For row-weighted comparisons, resample whole sequence components with replacement, with the number of draws equal to the number of observed components. The model-versus-constant intervals of Table 17 are the exception: they resample whole study-linked components under both weightings (10 groups, 9 with the source excluded), so their width is not comparable to sequence-component intervals. If component g is sampled m_g times, recompute

$$\widehat{\Delta}^* = \frac{\sum_g m_g \sum_{i \in g} D_i}{\sum_g m_g n_g}. \tag{A4.6}$$

Thus large sampled components retain their row mass; replacing this ratio by an average of component means would target a different statistic. For equal-component comparisons use $G^{-1}\sum_g m_g \bar{D}_g$.

The grouped equal-study comparison resamples study-linked components, whereas the shifted-cohort equal-component comparison uses sequence components. Tables identify the unit explicitly.

There are 10,000 bootstrap draws with analysis seed 90413. Endpoints are empirical 2.5th/97.5th percentiles of the resampled statistic. Fitted predictions and all selection decisions are held fixed. This is conditional uncertainty within the observed cohort under a component-resampling approximation; it does not include training randomness, unobserved laboratory shifts, source-family variation in a one-family cohort or unresolved shared-control measurement covariance. No nominal unconditional coverage theorem is asserted. In particular, the small number of study-linked components limits the grouped comparison.

The support-by-message interaction is computed from the joint row-level contrast

$$D_i^{\text{int}} = (f_{R1,G,i} - y_i)^2 - (f_{R1,N,i} - y_i)^2 - (f_{R0,G,i} - y_i)^2 + (f_{R0,N,i} - y_i)^2. \tag{A4.7}$$

It receives one joint component bootstrap; subtracting independently computed interval endpoints is not used. This contrast concerns the four selected procedures on fixed observed outcomes. It is not a causal claim about chemical activity or an isolated single-weight intervention.

A4.3 Training-seed variability

For every stochastic method, all ten seed risks and matching seed-level loss differences are reported, along with their mean, sample SD, minimum and maximum. Each seed's prediction file is tied to its selected and executed update count and failure status. Seed-average risk and ensemble risk occupy separate fields. No three-seed t test replaces a fixed-ensemble comparison, and no seed-level interval is claimed to represent new-study uncertainty. With deterministic methods there is one fitted object, not ten invented independent replicates. Failed fits and their declared fallbacks remain explicit.

A4.4 Exact candidate-ranking ties and dependence

An eligible APP pool is an exact assay/table/cell/dose/time group with at least five observations. Larger activity means stronger inhibition. Assign label rank r_i using average ranks for measured ties, and define percentile $q_i = (r_i - 1)/(n - 1)$ for $n > 1$. Set $k = \min(5, n)$. Let c be the kth largest prediction, $H = \{i : \hat{y}_i > c\}$ and $T = \{i : \hat{y}_i = c\}$. The selection mass is one on H, $(k - |H|)/|T|$ on T and zero below. Report $k^{-1} \sum_i s_i q_i$ and $k^{-1} \sum_i s_i y_i$. The equality test is exact equality of saved float predictions, not an outcome-selected tolerance.

Proposition A4.2 (Tie-aware constant reference)**.** *For a constant prediction on a pool of size $n > 1$, expected top-k mean measured percentile under uniform tie breaking is exactly $1/2$.*

Proof. Every observation is tied and has inclusion probability k/n. Average ranking preserves the sum of untied ranks, $\sum_i r_i = n(n + 1)/2$, even when measured labels tie. Therefore $n^{-1} \sum_i q_i = [(n+1)/2-1]/(n-1) = 1/2$, and $k^{-1} \sum_i (k/n) q_i = 1/2$. For a partial boundary tie, exchangeability gives each tied candidate inclusion probability $(k - |H|)/|T|$, proving the fractional rule above. □

All pools and their sizes, shared sequence components and direct tree–GNN / GNN–no-message percentile differences are reported. We do not perform an independent-pool significance test: the pools belong to one related patent family, and backgrounds can recur across pools. A ranking advantage is a descriptive candidate-selection property in those pools, not a biological validation or population guarantee.

A4.5 Measured pair metrics and uncertainty limits

For pair p, compute $d_p = y_{b_p} - y_{a_p}$ and $\hat{d}_p = \hat{y}_{b_p} - \hat{y}_{a_p}$, or the direct pair method's prediction. Pair MSE, MAE, correlation, prediction/observed SD and mean bias use the same formulas as activity, reporting row-, equal-sequence- and equal-exact-background weightings distinctly. A sign call is zero for absolute difference at most .02 and otherwise its sign. We report all negative/tie/positive calls, accuracy over all pairs, and exact correct counts among measured non-ties. The majority-direction control has sign metrics only; arbitrary numeric sign codes are not regression predictions.

Direct pair comparisons bootstrap complete sequence components with all their pairs and shared endpoints together. This retains known sequence/endpoint dependence but cannot represent unknown shared-control covariance between different components. Endpoint SDs alone do not determine a mean-difference SE: $\mathrm{Var}(\bar{Y}_b - \bar{Y}_a) = \mathrm{Var}(\bar{Y}_b) + \mathrm{Var}(\bar{Y}_a) - 2\,\mathrm{Cov}(\bar{Y}_b, \bar{Y}_a)$ requires replicate counts and covariance information that these records do not supply. We leave it missing. The operational sign band is not calibrated to those unknown variances.

A4.6 Within-assay B2 variation diagnostic

To distinguish sequence/background variation from a common assay shift, we additionally center measured and predicted pair differences within each exact primary assay. This is an explicitly oracle diagnostic, not an altered deployed prediction. For observation weights w_p, assay $a(p)$ and $W_a = \sum_{p:a(p)=a} w_p$, define $d_p^c = d_p - W_{a(p)}^{-1} \sum_{q:a(q)=a(p)} w_q d_q$ and the analogous $\hat{d}_p^c$. Their weighted correlation and SDs use the formulas above. The main diagnostic uses equal row weights. Each of 10,000 sequence-component resamples recomputes both weighted assay means and the centered correlation with that draw's component multiplicities. A missing variance gives an undefined correlation, not zero. The number of finite correlations is retained. This assessment conditions on the fitted ensembles and observed cohort; it does not supply missing replicate covariance. No centered score enters fitting, stopping or method selection. A low raw pair error without within-assay discrimination cannot by itself establish sequence-specific chemistry prediction.

A5 Primary-source comparisons and unresolved reproduction inputs

Interaction attribution and identification. HarsanyiNet computes attributions of its specially structured predictor under its receptive-field and masking requirements (Chen et al., 2023); this network has neither those units nor their attribution theorem. Neural causal identification uses graph-consistent structural models and matched distributions (Xia et al., 2023); molecular adjacency supplies neither a causal diagram nor those distributional constraints. Continuous neural partial-identification results require bounded Lipschitz mechanisms, latent-support regularity and controlled approximation and matching schedules (Tan et al., 2024). Those assumptions are not verified here, and local neural extrema would not provide certified outer bounds. Functional-ANOVA purification fixes a weighting-dependent decomposition of a predictor (Lengerich et al., 2020); it does not supply an unobserved joint assay mean. The G-NAVAR preprint's population result assumes fixed direct-sum edge spaces, positive product support and regularity, with additional HOFD and faithfulness conditions in its shared-modulator formulation (Kuskova et al., 2026). A covariance-rank diagnostic is not sufficient identification. None of these results is invoked as a guarantee for this siRNA learner. Four forward passes, encoding-equivalence checks and optimization over a compatible finite response class do not by themselves establish a new ML contribution.

A5.1 New primary measured four-condition panel

The bounded B3 search recovered the original Bramsen supplement (Bramsen et al., 2009), from primary article PMC2685080 and `gkp106_nar-02440-c-2008-File010.xls`. Primary Table 1 identifies antisense and sense chemical states, and Supplemental Table 1 supplies the corresponding normalized eGFP means and SDs. The assay uses HeLa-eGFP cells, 10 nM duplex, INTERFERin and a 72-hour readout normalized against a mismatch control, with triplicate assays repeated twice. Recorded viability values are not model inputs, training targets or admission filters, and an eGFP reduction is not shown to be viability-independent or a direct mRNA measurement. The release lacks the raw shared-control covariance needed for a defensible SE on every contrast, so we retain the printed SDs without inventing those covariances.

The source sequence reference is AS W053 and SS W207. Their base strings are, respectively, `ACUUGUGGCCGUUUACGUCGC` and `GACGUAAACGGCCACAAGUUC`. A is a specified whole-antisense chemistry replacement with the same bases and length; B is a specified whole-sense replacement, also with the same bases and length. Every admitted rectangle contains measured 00, 10, 01 and 11 conditions, all in this assay. On the activity scale $y = 1 -$ relative eGFP, let μ_{ab} denote

the population assay mean in condition ab. Distinguish

$$I_{AB} = \mu_{11} - \mu_{10} - \mu_{01} + \mu_{00}, \tag{A5.1}$$

$$I_{AB}^{\text{obs}} = y_{11} - y_{10} - y_{01} + y_{00}, \tag{A5.2}$$

$$\widehat{I}_{AB} = f(x_{11}) - f(x_{10}) - f(x_{01}) + f(x_{00}). \tag{A5.3}$$

The tables evaluate the predictor against I_{AB}^{obs}, the contrast of reported measured means, not an exactly known population interaction; reading it as an estimate of I_{AB} requires comparable condition-specific means on the same response scale, and unresolved replication and shared-control covariance limit its uncertainty. It describes nonadditivity on the reported assay scale and is not an isolated GNA effect, a mechanism proof, new wet-lab validation or evidence of private-anchor calibration. Additive changes on another response scale need not remain additive after a nonlinear transformation.

Admission first parses the base characters and explicit chemical subscripts in primary Table 1. The literal aliases and normalized modification names appear in Table 6. Unknown aliases, including the ambiguous AP annotation, are excluded, and only parseable states exactly matching each reference's base string and length are considered. A state must differ chemically from its reference, and duplicate primary (AS,SS) identities and nonfinite means are excluded. Every such AS/SS pair is enumerated and a rectangle retained only when all four unique measured endpoints exist, with neither the observed interaction magnitude nor any prediction selecting it. Graphs use the frozen positional vocabulary and the original deterministic pairing rule, and bonds are encoded as unresolved under the same annotation-only convention as ENsiRNA, so no directional linkage is invented from a node label.

An initial protocol required the unmodified reference to be present in the historical chemistry-only release and yielded no eligible rectangle; that failed restriction and its evidence remain preserved. An explicit source-driven amendment then allowed the directly measured primary reference, before any prediction on the panel. It yields 165 measured conditions and 140 complete rectangles. Each label comes from the primary workbook. All rectangles share one sequence background and a common reference, so we report descriptive errors and all seeds without an independent-rectangle confidence interval.

The complete Bramsen source belongs to the original outer-0 test component, so only final outer-0 checkpoints trained and selected without that source are used. We verify the absence of every primary-panel strand 13-mer from their training and validation records, verify saved input transformations and reuse fixed checkpoint weights; no new fit, pair supervision, calibration or model selection uses this panel. Its discovery is retrospective, and source separation does not make it a prospectively collected external replication. The original APP inventory still has no four-state block.

Primary subscript	Frozen name
OMe	2-O-Methyl
F	2-Fluoro
DNA	2-Deoxy
AEM	2-Aminoethoxymethyl
APM	2-Aminopropoxymethyl
EA	2-Aminoethyl
CE	2-Cyanoethyl
GE	2-Guanidinoethyl
HM	4-C-Hydroxymethyl Deoxyribonucleic acid
LNA	Locked nucleic acid
ALN	Alfa-L-Locked nucleic acid
ADA	2-N-Adamant-1-yl-Methylcarbonyl-2-Amino-Locked nucleic acid
PYR	2-N-Pyren-1-yl Methyl-2-Amino-Locked nucleic acid
OX	Oxetane-Locked nucleic acid
CENA	2,4-Carbocyclic-Ethylene-bridged nucleic acid-Locked nucleic acid
CLNA	2,4-Carbocyclic-Locked nucleic acid-Locked nucleic acid
UNA	Unlocked nucleic acid
ANA	Altritol nucleic acid

Primary subscript	Frozen name
AENA	2-Deoxy-2-N,4-C-Ethylene-Locked nucleic acid
HNA	Hexitol nucleic acid

Table 6: Literal source aliases used in primary B3 admission. AP was excluded from the original B3 admission; its primary sugar alias and duplicate release mappings are adjudicated in Appendix A8.

A5.2 Original primary versus released comparison

The identity reconciliation maps exact primary base/chemical states to the historical ENsiRNA release without using prediction quality. At the first rounding tolerance of 0.000500001 activity units, some mapped means disagree; allowing 0.005000001 explains additional small differences. A substantial set differs by more than 0.05. Table 7 gives all counts and the reconciliation scope. These discrepancies are conditional on the documented chemical identity mapping and require further primary curation. They must not all be labeled biological errors: annotation mapping and release rounding can contribute. None of the frozen labels or historical checkpoints is overwritten. We therefore describe ENsiRNA activity tables as released-label benchmark results, with separately measured primary B3 evaluation. APP primary-table agreement and S7 provenance are different evidence checks and are not negated or certified by this one-source audit.

Reconciliation diagnostic	Count/value
identity_matched_pairs	1604
within_half_tenth_percentage_point	1209
within_half_percentage_point	1258
greater_than_five_percentage_points	299
max_absolute_activity_difference	1.0852170324990895

Table 7: Primary/release audit conditional on the stated identity mapping. Rounding tolerances are activity fractions; no frozen label is changed.

A5.3 Prior-work hypotheses and current use

No published asymptotic theorem supplies a risk guarantee for the trained model. The foundational demonstration of siRNA-mediated silencing in mammalian cells (Elbashir et al., 2001), together with sequence-design, strand-bias and off-target work (Reynolds et al., 2004; Ui-Tei et al., 2004; Khvorova et al., 2003; Schwarz et al., 2003; Jackson et al., 2003; Birmingham et al., 2006), supplies biological context and separates efficacy from specificity; chemical potency and delivery results concern their stated chemistries, assays and exposures, which we do not assume for every record (Allerson et al., 2005; Bramsen et al., 2009; Nair et al., 2014). The early neural siRNA design paper and its corrections are provenance precedents, not additional training cohorts (Huesken et al., 2005a;b; 2006), and CMsiRNAdb, OligoGym and the Davis release motivate explicit dataset identities rather than certifying every raw measurement (He et al., 2026; Rotrattanadumrong & De Donno, 2025; Davis et al., 2025).

MPNNs and relational graph convolutions motivate the typed aggregation notation, while GCN and GraphSAGE make different normalisation or sampling choices (Gilmer et al., 2017; Schlichtkrull et al., 2017; Kipf & Welling, 2016; Hamilton et al., 2017; Battaglia et al., 2018); the ordered position-aware readout is not invariant set pooling, and expressivity results do not imply finite-data generalisation here (Zaheer et al., 2017; Xu et al., 2019). The token CNN is a conventional local-sequence control (Kim, 2014), and the ridge penalty, randomized trees and their ensemble precedent use the exact objectives and library settings stated above (Hoerl & Kennard, 1970; Geurts et al., 2006; Breiman, 2001). Dropout, layer normalisation and decoupled weight decay have the axes, probability and updates stated in this appendix (Srivastava et al., 2014; Ba et al., 2016; Kingma & Ba, 2014; Loshchilov & Hutter, 2017); deep ensembles motivate retaining member variation rather than

treating ten predictors as ten biological cohorts (Lakshminarayanan et al., 2016). Proper-scoring distinctions concern their defined predictive targets (Gneiting & Raftery, 2007) and the finite-sample identities used here are proved directly; bootstrap resampling is applied to the declared observed groups without a population-coverage theorem (Efron, 1979); leakage and shift benchmarks explain why group protection and deployment scope are recorded separately (Kapoor & Narayanan, 2023; Koh et al., 2021); and software citations identify the actual numerical and plotting implementations, whose installed versions are recorded independently (Paszke et al., 2019; Pedregosa et al., 2011; Harris et al., 2020; Virtanen et al., 2020; Hunter, 2007).

A6 Preserved results and negative evidence

Table 8 reports every activity procedure of the frozen released-label (v3) campaign with row weighting. It contains no new fits; cohorts, weighting and target definitions are those of the campaign. The four audited models on APP and S7 repeat Table 13; the original R0 arms, both ridge baselines, the three constants and the whole grouped cohort are recorded only here. On APP and S7 both ridge baselines have higher MSE than every constant. Measured B2 pair results for every procedure are in Table 21, APP ranking across the seventeen assay pools in Table 20, and focused source-excluded and primary-assay results in Appendix A9.

A7 Historical reproduction, arithmetic and resources

A7.1 Actual resource accounting and failures

Quantity and scope	Actual value
Core fit objects	2312
Executed optimizer updates	1407967
Selected-checkpoint updates	406367
Summed core-fit CPU seconds	12523.769
GPU training-region elapsed seconds	11689.879
Completed stage child CPU seconds	13202.199
Maximum reported stage child RSS, KiB	3306840
Exact published siRNA predictors fitted	0
Published splicing-architecture fits in this work	12 attempts, 6 reported
siRNA mask-restoration fits in this work	56 (21,513 updates)
Administrative inspection/authoring	Additional unmetered nonzero cost

Table 9: Resource accounting with distinct scopes. The core-fit and process-cost entries summarize the historical campaign; the intervention rows separately report new work. These entries do not form one cumulative resource total. Fit-region, stage and GPU-region time measurements overlap and are not added. The MAVE-NN attempt count includes both invocations; the retained-fit update count and incomplete cumulative accounting are described in Appendix A10.24.

Selected-checkpoint updates describe the saved predictor. Executed updates describe optimization work, including the trajectory after that checkpoint. Development fits are part of executed work. GPU fitting-region elapsed time is a synchronized host measurement around training; it overlaps stage wall time and is not GPU kernel time, energy or a measure of independent work. CPU process time, child CPU time and aggregate command wall time likewise have different scopes. They must not be summed as if disjoint.

Direct filesystem inspection, authoring, citation browsing, preflight document builds, rendering, copying and hashing have additional nonzero administrative costs; not every interactive call has a process-level meter. We report that unmetered remainder explicitly. No historical allowance wrapper or ledger is used. Failed dependency/acquisition commands retain their measured costs. An unused

Cohort	Model	MSE	R^2	r	Pred. SD	Bias
APP	R1 GNN	0.1625	-0.001	-0.018	0.0072	-0.0081
	R1 no-message	0.1630	-0.005	0.015	0.0085	-0.0286
	R0 GNN	0.1625	-0.001	-0.026	0.0073	-0.0049
	R0 no-message	0.1629	-0.004	-0.001	0.0085	-0.0241
	Chemistry tree	0.1896	-0.168	-0.021	0.0103	-0.1644
	Token CNN	0.1644	-0.013	0.080	0.0036	-0.0483
	Guide ridge	0.2949	-0.818	-0.014	0.1420	0.3330
	Pairwise ridge	0.2064	-0.272	0.029	0.0619	0.2044
	Training row mean	0.1760	-0.085	–	0.0000	0.1172
	Training group mean	0.1768	-0.090	–	0.0000	0.1207
	Oracle test mean	0.1623	0.000	–	0.0000	0.0000
S7	R1 GNN	0.1093	-0.215	0.112	0.0059	0.1403
	R1 no-message	0.1047	-0.164	0.117	0.0066	0.1231
	R0 GNN	0.1102	-0.225	0.116	0.0058	0.1437
	R0 no-message	0.1057	-0.175	0.110	0.0067	0.1270
	Chemistry tree	0.0919	-0.021	0.017	0.0089	-0.0441
	Token CNN	0.1017	-0.130	0.108	0.0034	0.1092
	Guide ridge	0.7815	-7.686	0.123	0.1150	0.8287
	Pairwise ridge	0.9698	-9.780	0.031	0.0619	0.9366
	Training row mean	0.1488	-0.654	–	0.0000	0.2426
	Training group mean	0.1505	-0.673	–	0.0000	0.2461
	Oracle test mean	0.0900	0.000	–	0.0000	0.0000
Grouped	R1 GNN	0.0712	0.011	0.163	0.0574	0.0307
	R1 no-message	0.0712	0.010	0.158	0.0575	0.0293
	R0 GNN	0.0712	0.011	0.163	0.0574	0.0307
	R0 no-message	0.0712	0.010	0.158	0.0575	0.0293
	Chemistry tree	0.0656	0.087	0.312	0.0756	0.0250
	Token CNN	0.0717	0.003	0.143	0.0680	0.0205
	Guide ridge	0.0938	-0.304	0.056	0.1172	-0.1078
	Pairwise ridge	0.0737	-0.024	0.058	0.0535	-0.0234
	Training row mean	0.0724	-0.006	-0.073	0.0088	0.0031
	Training group mean	0.0769	-0.069	-0.095	0.0250	0.0553
	Oracle test mean	0.0719	0.000	–	0.0000	0.0000

Table 8: All original v3 ensemble activity procedures with row weighting on the full cohorts: APP 1,839 rows, S7 118 and the grouped benchmark 2,927, with label SD 0.4028, 0.3000 and 0.2682. R1 is the training-gated and R0 the original GNN. Training means are constants fitted on each training partition; the grouped cohort pools five outer test folds, so they vary there and have a defined correlation. The oracle test mean reads the evaluation labels and is a diagnostic, not a risk floor. A dash marks a correlation undefined for a constant prediction; bias is mean prediction minus mean label. Alternate weightings and further metrics: `v3/tables/activity_metrics.csv`.

published-model fit reserve is not an executed experiment. The frozen plan set a 48-hour campaign wall ceiling; the supervising session monitored elapsed time, but the runner does not implement an automatic wall-clock termination timer. Its per-fit epoch and stopping limits are implemented in code. The actual completed runtime, rather than the planning estimate, is reported above.

A8 Source integrity and representation robustness

This section specifies the follow-up independently of the historical campaign. The original released-label observations, predictions, selected checkpoints and negative results remain unchanged. All new evaluations are retrospective sensitivities. In particular, an observation previously used in model development does not become an independent external test because the present protocol assigns it a new role.

A8.1 SOURCE RECORDS, CHEMICAL IDENTITIES AND ADJUDICATION

Let $D_R = \{(r_i, x_i, y_i, s_i, g_i)\}$ denote the admitted released benchmark, with persistent record identifier r_i, annotated duplex/context x_i, released inhibition fraction y_i, source label s_i, and protected study/sequence component g_i. The Bramsen source is $s_B = 19282453$. There are 1,972 raw workbook records with this source label and 1,924 admitted records. The latter form one sequence component and contribute $1924/2927 = 0.657328\ldots$ of the grouped benchmark. The denominator 2,927 excludes APP and Davis S7; it is not a count of independent biological experiments.

The primary identity is an ordered pair of the antisense and sense strand identifiers in Table 1 of Bramsen et al. (2009). For each strand, the comparison key retains the printed base sequence, its length, and the complete list of reported position-specific modification names. Antisense and sense are not interchanged. Printed oligonucleotide order is retained; missing per-row terminal direction labels do not authorize reversal of sequence or chemical positions. The key does not infer unreported linkage stereochemistry, terminal groups or assay metadata. A unique match of the reported fields is therefore narrower than independently verified complete chemical identity.

The primary XML is parsed in two ways. The historical implementation traverses elements while associating each chemical subscript with its preceding nucleotide. The follow-up replaces chemical tags by lexical tokens, removes markup and whitespace, and tokenizes the resulting base/chemical stream. These algorithms agree for all 91 historically parsed strands. The released raw workbook annotations are separately expanded from modification/position lists and checked against the frozen observation objects. Neither implementation reads model predictions to select identities. Outcome agreement is not used to choose among candidate mappings, strand orders or chemical aliases.

The primary Table 1 footnote defines AP as a $2'$-aminopropyl substitution. This supports the sugar alias, not the distinct base alias in the frozen vocabulary. Expanding that literal alias exposes 76 admitted released rows with multiple rows mapping to the same primary pair; these are quarantined rather than resolved by their activity values. The final admitted-source partition is 1,604 uniquely linked reported identities, 201 unmatched identities, 76 rows with multiple released mappings, and 43 rows lacking a unique primary measurement. These categories sum to 1,924. The 320 quarantined rows are not replaced by predictions or synthetic measurements.

Supplemental Table 1 is read by named columns, the eGFP value taken from the source cell in column D with the spreadsheet row retained; direct cell access rechecks the row, both identifiers and the numeric value. Both readers use the XLS backend, so their agreement is a cell-address check, not two independent binary-format implementations. The 91 available Table 1 marginal eGFP means agree with the corresponding supplement values to the precision printed in Table 1. This independently supports the chosen endpoint column. Viability columns and unrelated spreadsheet summaries are not substituted for eGFP activity.

The primary screen used HeLa cells stably expressing eGFP, 10 nM siRNA and INTERFERin, with confocal readout 72 h after transfection. eGFP was normalized to a mismatch-siRNA control; viability used a different control. The article describes triplicate assays repeated twice. These facts do not provide all per-run observations or shared-control covariances. The released records do not individually provide a complete assay provenance key. Matching source, reported dose and molecular annotations therefore does not establish the origin of every discrepancy.

The comparison convention is fixed by the documented endpoint definitions:

$$y_i^R = \mathrm{PCT}_i/100, \qquad y_j^P = 1 - \mathrm{eGFP}_j. \tag{A8.1}$$

The primary eGFP quantities are fractions, not percentages. The transformation preserves negative inhibition values when relative expression exceeds one, and preserves the reported SD scale. There is no clipping, fitted offset, outcome-selected row shift or fitted permutation. Every one of the 299 historical residuals exceeding 0.05 is checked at its mapped source cell; 32 additional records are selected in deterministic SHA256 order without reference to their model errors. These 299 discrepancies are conditional on the 1,604 matched identities. They are not an estimated error rate for all 1,924 source records or the complete benchmark, and do not identify whether a disagreement originated in an annotation, transcription, assay convention or measurement.

The reconciliation preserves workbook/sheet/row, persistent ID, raw sequences and annotations, primary identifiers and source cell, both numeric fields, the fixed transformation, match cardinality

and unresolved fields; the manifest pins the exact primary XML and workbook bytes. This permits the primary-assay sensitivity below without certifying that the historical release has been repaired.

A8.2 Separately versioned primary-assay sensitivity

Define J by unique reported identity, unique primary measurement, and the documented nominal dose, before considering any model error. The versioned sensitivity target is

$$D_P = \{(r_i, x_i, y_i^P, s_i, g_i) : s_i = s_B,\ r_i \in J\} \cup \{(r_i, x_i, y_i^R, s_i, g_i) : s_i \neq s_B\}. \tag{A8.2}$$

It contains 1,604 primary-linked Bramsen rows and 1,003 other-source rows, for 2,607 observations. Its manifest retains each original released value, primary value, source-cell provenance and original graph index. The other-source labels are unchanged and are not declared newly adjudicated. The same frozen graph vocabulary and original graph tensors are indexed; chemical encodings are not selected by efficacy. Protected group identifiers are retained because the operation removes rows without inventing new biological independence.

Released-label and primary-assay outcomes are separate targets. Restricting a released-trained predictor to J and rescoring it against y^P answers a different question from refitting all affected training and selection stages on D_P. The matched-subset table reports both operations separately. Methods within each comparison use identical eligible observations. The missing 320 source rows are disclosed alongside every primary-assay coverage statement.

A8.3 Exact calibration decomposition

For fixed nonnegative weights w_i with $\sum_i w_i = 1$, write $\bar{y} = \sum_i w_i y_i$, $\bar{p} = \sum_i w_i p_i$, $b = \bar{p} - \bar{y}$, $v_y = \sum_i w_i (y_i - \bar{y})^2$, $v_p = \sum_i w_i (p_i - \bar{p})^2$ and $c = \sum_i w_i (y_i - \bar{y})(p_i - \bar{p})$. Since both centered sums vanish,

$$\sum_i w_i (p_i - y_i)^2 = \sum_i w_i \big[(p_i - \bar{p}) - (y_i - \bar{y}) + b\big]^2 \tag{A8.3}$$

$$= v_p + v_y - 2c + b^2. \tag{A8.4}$$

The cross term with b is $2b \sum_i w_i [(p_i - \bar{p}) - (y_i - \bar{y})] = 0$. This proves the identity without a sampling or model assumption. Subtracting two models' identities cancels v_y. A large difference in squared mean bias therefore describes the observed error gap but does not establish why either model acquired that bias. No held-out intercept correction is applied.

With uniform weights, the reported coefficient of determination is $R^2 = 1 - \sum_i (p_i - y_i)^2 / \sum_i (y_i - \bar{y})^2$. Equivalently it is $1 - \mathrm{MSE}\,/v_y$, with both quantities normalized by n. Replacing v_y by the $n-1$ sample-variance estimate would change the statistic. A pooled source-subset R^2 need not equal any average of within-source R^2 values because both the center and denominator change. When the denominator vanishes, R^2 is undefined rather than assigned a favorable value.

A8.4 Frozen input identities and deterministic cancellation

Let $E_T(x)$ be the actual preprocessing fitted on training partition T. It applies the saved node-feature mask, saved varying-context mask, imputation, ordered padding and the model's graph/context convention. The audit hashes named arrays, their shapes and canonical little-endian float32 bytes with signed zeros normalized, without approximate prediction-based matching. The audit signature of a state is the byte content of the tensors each model family receives: X, A, mask, C for the GNN and for the no-message model, whose forward reads only receiver relation counts from A, so retaining the full A is conservative; X, mask, C for the token CNN, which reads no adjacency; and the selected feature columns for the tree. Every reported siRNA H was built from this signature, and the splicing predictor's H from its one-hot input (Appendix A10.21). On B3 the no-message convention does not matter: for each of the ten seeds its full-A partition equals the partition from X, mask, C alone, and receiver counts are a function of A, so they induce the same 90 classes; all four families give that partition. All ten saved preprocessors are checked, not inferred from one favorable seed.

For corrected graph routing, an additional algebraic representation groups equal directional backbone bases and gates unsupported residuals, while retaining the original total-degree denominator.

This real-arithmetic quotient is reported separately from literal input identity, since floating-point reassociation is not an exact bitwise equivalence claim. Loss of a named chemistry feature does not imply the intervention disappears entirely, because base identity, modification position and the unmodified indicator can still distinguish its encoded state; conversely, distinct encoded states do not establish adequate expressivity or learned chemical effects.

Proposition A8.1 (Finite input cancellation)**.** *Fix preprocessing and a deterministic prediction function f. If the multisets $\{E_T(x_{11}), E_T(x_{00})\}$ and $\{E_T(x_{10}), E_T(x_{01})\}$ are equal, then the predicted four-term contrast is zero.*

Proof. For every distinct encoded value u, multiset equality gives equal multiplicities on the two sides. Multiplying each multiplicity by $f(u)$ and summing over the finitely many encoded values gives $f(E_T(x_{11})) + f(E_T(x_{00})) = f(E_T(x_{10})) + f(E_T(x_{01}))$. Rearranging proves the claim. The statement requires no fitted-parameter, smoothness or biological assumption. It applies to deterministic inference, with dropout disabled. Purely algebraic routing equivalence is qualified by floating-point rounding. The converse is false: a constant f gives zero contrast even when all four inputs differ. □

Proposition A8.2 (Observed collision error floor)**.** *Partition the observed inputs into exact encoding classes C. For squared error, every deterministic function of the encoded input has weighted empirical risk at least $\sum_C \sum_{i\in C} w_i(y_i - \bar{y}_C)^2$, where $\bar{y}_C = \sum_{i\in C} w_i y_i / \sum_{i\in C} w_i$ for positive-mass classes.*

Proof. The function assigns one prediction p_C to each class. Expanding $p_C - y_i = (p_C - \bar{y}_C) + (\bar{y}_C - y_i)$ gives

$$\sum_{i\in C} w_i(p_C - y_i)^2 = \Big(\sum_{i\in C} w_i\Big)(p_C - \bar{y}_C)^2 + \sum_{i\in C} w_i(y_i - \bar{y}_C)^2,$$

because $\sum_{i\in C} w_i(\bar{y}_C - y_i) = 0$. The first term is nonnegative. Summing over classes proves the lower bound. An unrestricted function assigning $p_C = \bar{y}_C$ attains it on these observed inputs; the particular neural architecture need not do so. Zero-mass classes contribute nothing. This is a finite-data identity, not a population minimax or generalization guarantee. □

A8.5 Shared-reference and leave-margin B3 sensitivities

The 140 distinct contrasts cross 14 antisense and 10 sense replacements. They use 165 measured states on one sequence background and share reference/marginal conditions. Writing all endpoint means as a vector Y and the signed contrast matrix as C, the contrast covariance would be $C\,\mathrm{Cov}(Y)C^\top$. Endpoint SDs and the aggregate replication description do not supply this covariance. No independent-rectangle bootstrap or inferred mean-difference standard error is used.

For fixed model contrasts p_j and measured contrasts I_j, a common reference perturbation δ sends every I_j to $I_j + \delta$. With uniform weights, direct expansion gives

$$M(\delta) = M(0) - 2\delta\,\overline{(p - I)} + \delta^2, \tag{A8.5}$$

$$Z(\delta) = Z(0) + 2\delta\,\bar{I} + \delta^2, \tag{A8.6}$$

$$M(\delta) - Z(\delta) = M(0) - Z(0) - 2\delta\,\bar{p}, \tag{A8.7}$$

where M is model MSE and Z is zero-contrast MSE. The last line follows by subtraction, since $\overline{(p - I)} + \bar{I} = \bar{p}$. The audit reports shifts of minus one, zero and plus one reported reference SD solely as descriptive sensitivity values; they are not confidence limits for the reference mean.

Leaving one antisense replacement removes its ten contrasts; leaving one sense replacement removes its fourteen contrasts. Predictions stay fixed and remaining errors are recomputed against the same zero control. These 24 omissions expose margin dependence but are neither new fitted folds nor independent biological replications. Observed assay nonadditivity, a population biological interaction, representability of an input change and the trained model's contrast are different objects.

The reported reference SD is 0.021415097 on the fractional response scale. Shifts retain 165 endpoints/140 contrasts; an antisense omission retains 154/130 and a sense omission 150/126. Predictions remain fixed. Previously reported MSEs and differences are reproduced to 2×10^{-15};

added spreads use the same identities. Table 10 retains the GNN reversal after omitting JC5 and no-message reversals after omitting JC10 or JC5. Tree and CNN remain worse than zero throughout; no method reverses under reference shifts.

For every reduced panel, rebuild C_J over retained endpoints, restrict the input-class indicator to H_J, and recompute an orthonormal basis for $\mathrm{col}(C_J H_J)$ using SVD with relative cutoff 10^{-10}; rational elimination verifies the rank. The full-panel projector is never sliced and reused. The residual $(I - P_J)t_J$ has energy share 13.39–15.89% under reference shifts, 8.00–15.83% under antisense omissions and 13.69–16.65% under sense omissions. These are finite-panel influence values, not sampling uncertainty or causal error allocation. Full reduced maps, bases and outcomes remain in the store.

Sensitivity	Method	m/q	Model MSE	Zero MSE	$10^3\Delta$	Flips
Omit AS	Tree	154/130	0.059628–0.077251	0.056062–0.073257	2.456–4.031	0
Omit AS	GNN	154/130	0.056195–0.073399	0.056062–0.073257	0.015–0.143	0
Omit AS	No-msg	154/130	0.056160–0.073339	0.056062–0.073257	-0.045–0.098	1
Omit AS	CNN	154/130	0.056290–0.073475	0.056062–0.073257	0.007–0.229	0
Omit SS	Tree	150/126	0.055818–0.079524	0.053743–0.075358	1.176–4.188	0
Omit SS	GNN	150/126	0.053882–0.075507	0.053743–0.075358	-0.009–0.151	1
Omit SS	No-msg	150/126	0.053848–0.075443	0.053743–0.075358	-0.032–0.104	1
Omit SS	CNN	150/126	0.053966–0.075589	0.053743–0.075358	0.000–0.231	0
Ref. shift	Tree	165/140	0.066229–0.078566	0.062803–0.074524	3.426–4.042	0
Ref. shift	GNN	165/140	0.062923–0.074664	0.062803–0.074524	0.120–0.140	0
Ref. shift	No-msg	165/140	0.062871–0.074608	0.062803–0.074524	0.069–0.084	0
Ref. shift	CNN	165/140	0.062985–0.074749	0.062803–0.074524	0.183–0.225	0

Table 10: Ranges over three reference shifts, fourteen antisense omissions or ten sense omissions. m/q: retained endpoints/contrasts. Paired differences Δ are scaled by 10^3 and computed before taking ranges, not by subtracting range endpoints. Flips count reversals from the unperturbed ordering. These are dependent descriptive diagnostics, not confidence limits.

Sensitivity	Method	Predicted SD	Recorded SD	Largest influence	Gap change
Omit AS	Tree	0.008806–0.018376	0.204264–0.229500	JC10	-0.001278
Omit AS	GNN	0.000252–0.001648	0.204264–0.229500	JC10	-0.000115
Omit AS	No-msg	0.000299–0.001714	0.204264–0.229500	JC10	-0.000122
Omit AS	CNN	0.001113–0.002984	0.204264–0.229500	JC10	-0.000197
Omit SS	Tree	0.007219–0.018653	0.203193–0.230994	JC5	-0.002558
Omit SS	GNN	0.000327–0.001674	0.203193–0.230994	JC5	-0.000139
Omit SS	No-msg	0.000348–0.001740	0.203193–0.230994	JC5	-0.000108
Omit SS	CNN	0.000642–0.003028	0.203193–0.230994	JC5	-0.000204
Ref. shift	Tree	0.017767	0.222442	1	-0.000308
Ref. shift	GNN	0.001589	0.222442	1	-0.000010
Ref. shift	No-msg	0.001652	0.222442	-1	0.000008
Ref. shift	CNN	0.002879	0.222442	1	-0.000021

Table 11: Spread ranges and the omitted margin (or reference multiplier) with largest absolute change of paired error difference. A positive reference multiplier adds one reported SD to the shared activity reference.

A8.6 Ranking reconciliation and uncertainty scope

Historical and current predictions are joined by exact persistent observation and assay-pool identifiers. All 1,839 APP records and all 17 pool memberships agree. Larger measured inhibition is the preferred direction. Every campaign is rescored with the same outcome average-rank percentile and fractional selection mass among exact prediction ties at the top-five boundary. A saved ensemble row is an aggregate, not an extra seed. Ensemble predictions are reconstructed from actual member rows and the historical reported scores are retained alongside the common-definition scores.

The original, frozen-diagnostic, factorial and all-ENsiRNA deployment protocols remain separate. Training populations, support/context handling, selection effort, optimization budget and ensemble sizes changed jointly. The ranking reversal limits protocol-stability claims; it does not make a well-defined descriptive ranking score meaningless or identify one cause of the reversal. Pools share one related patent family and overlap in sequence components. Their number does not justify an independent-pool significance test.

Component resampling conditions on fixed fitted predictions and the observed cohorts. The independent review calculation uses RNG seed 914221; campaign comparisons retain the previously defined seed and 10,000 draws. Differences between valid resampling RNG realizations are distinguished from errors in the point estimate. Ten-seed variability is reported separately. An interval containing zero denotes an unresolved comparison, not proof of equivalence; no post-hoc equivalence margin is introduced.

A9 Complete focused numerical summaries and evidence index

All numbers here derive from completed v4 fits or explicitly reused fixed predictions; no unfavourable seed or failed attempt is removed. The focused v4 cohort results appear in main Table 2, with matched training-constant comparisons in Appendix A10.12. Historical v3 results are retained separately in Appendix A10.8; the metric identities are given in Appendix A4. The complete per-cohort, per-weighting and per-seed rows are columns of `activity_metrics` and `variance_decomposition` in the canonical store rather than reprinted pages.

A9.1 Evidence index and complete row-level records

The separately indexed scientific archive uses paths relative to its root. Every listed table retains its full precision, column schema and SHA256 in the archive manifest. Figures are generated from the same named records.

Archive path	Preserved evidence
`v4/adjudication/row_reconciliation.csv`	All 1,972 raw source rows (1,924 admitted): cells, chemistry, cardinality, original/primary values, categories
`v4/dataset/primary_reanchored_observations.jsonl`	2,607 versioned observations, original values and graph indices
`v4/review_checks/source_scores.csv`	All-source, Bramsen, other-source, every individual source and constant; three weightings
`v4/evaluation/per_source_metrics.csv`	Every source after focused refitting; exact matched coverage
`v4/evaluation/ensemble_predictions.csv`	Exact member means and cohort identifiers
`v4/fits/`	Best/last checkpoints, optimizer states, histories, selected transforms and memberships
`v4/visibility/B3_rectangle_visibility.csv`	All method/rectangle identities and forced-cancellation flags
`v3/tables/all_fit_summary.csv`	All original 2,312 fits, preserved without new fitting
`v3/b3_primary/`	Original primary panel, rectangles, fixed predictions and source-exclusion checks

Table 12: Indexed detailed evidence; no repetitive per-fit dump is printed in the manuscript.

A10 STRUCTURAL ATTRIBUTION AUDIT: COMPLETE SPECIFICATION

A10.1 EXACT CONTRAST CERTIFICATE AND ITS LIMITS

Let $s_1, \dots, s_m$ be the distinct raw endpoint states used by a finite collection of contrasts. An encoding instance includes its fitted training-only preprocessing. Write $e(s_i)$ for the complete inputs consumed by its deterministic decoder, including all ordered slots and ancillary branches. Partition these states into k classes of exactly equal encodings and define $H \in \{0,1\}^{m\times k}$ by $H_{ij} = 1$ precisely when state i belongs to class j. Let $C \in \mathbb{R}^{q\times m}$ contain the signed contrast coefficients. Repeated endpoints contribute their coefficients with multiplicity. An unrestricted deterministic decoder on these classes is an arbitrary vector $g \in \mathbb{R}^k$, and its contrast vector is CHg.

Proposition A10.1. *$CH = 0$ if and only if $CHg = 0$ for every unrestricted deterministic decoder g on the observed encoding classes.*

Proof. If $CH = 0$, multiplication by any g gives $CHg = 0$. Conversely, assume $CHg = 0$ for every $g \in \mathbb{R}^k$. For each $j = 1, \dots, k$, choose the coordinate vector $g = e_j$. Then CHe_j is the jth column of CH and is zero. Since every column is zero, $CH = 0$. Equivalently, if any entry $(CH)_{ij}$ is nonzero, the decoder taking value one on class j and zero on the other classes witnesses a nonzero ith contrast. This proves both implications without continuity, probability or learning assumptions. □

For coefficients $(1, -1, -1, 1)$, the condition is exactly equality of the positive and negative encoding multisets. It concerns this contrast, not every explanation functional. Restricted decoders may have additional limitations when $CH \neq 0$: the basis-vector decoder used in the converse need not belong to a fixed neural family. The identity is elementary and is not an algorithmic novelty claim. For an ensemble, a sufficient certificate is that each constituent's own complete-input encoding passes the test. A hidden-state coincidence found after fitting is a checkpoint-specific observation, not a pre-fit certificate.

The implementation compares named tensor arrays with their shapes, dtype and exact C-order bytes. It never rounds, reorders slots, substitutes graph isomorphism, or turns an invalid record into a zero input. SHA256 fingerprints locate identities; actual byte payloads decide equality. The tensors compared are those of the single audit signature defined in Appendix A8.4. Original node and edge conventions and their unresolved chemical assumptions remain as specified in the data appendix. The additional real-arithmetic routing quotient remains distinct from literal equality.

For a set F of verified forced-zero contrasts and nonnegative weights w_q with positive total, every decoder covered by this certificate incurs at least $\sum_{q\in F} w_q \tilde{I}_q^2 / \sum_q w_q$ recorded-label squared error. Each flagged prediction is zero, so its squared residual equals $\tilde{I}_q^2$; all other residual squares are nonnegative, proving the inequality. This is a finite recorded-label bound with the full denominator. It neither removes measurement noise nor supplies a biological population-risk lower bound. Here F is empty, so its contribution is zero.

A10.2 WHY FOUR COINCIDENT CORNERS DO NOT SETTLE OTHER EXPLAINERS

The reference scientific model target is $f(e(s_{11})) - f(e(s_{10})) - f(e(s_{01})) + f(e(s_{00}))$. Its players are the declared antisense and sense replacements, its baseline is the measured 00 condition, and masking means selecting the corresponding full, re-encoded measured state. Other annotations and assay context are fixed. No continuous interpolation is needed.

For the analytic counterexample let $e(a, b, z) = ab(1 - z)$ and $g_\theta(t) = \theta t$. At $z = 1$, all four a, b encodings equal zero, and the four-term contrast equals zero for every θ. Define a three-player coalition game with baseline $(0, 0, 0)$, target $(1, 1, 1)$ and masking that sets precisely the players in a coalition to one. Re-encoding gives $v(S) = \theta\, 1_{\{a,b\}\subseteq S}(1-1_{z\in S})$. For a pair, the Shapley interaction index weights background $S \subseteq \{z\}$ by $|S|!(3 - |S| - 2)!/(3 - 1)!$ (Grabisch & Roubens, 1999). Both weights are $1/2$. The second difference at $S = \varnothing$ is θ, while that at $S = \{z\}$ is zero. Their weighted sum is $\theta/2$. Thus equality at four fixed-background corners does not fix this interaction index. All eight binary inputs are admissible for this explicitly defined analytic game; they are not synthetic biological outcomes. The implemented checks use $\theta = -2, 0, 3$ and retain the null case.

For continuous-path explanation, equality at endpoints alone cannot determine intermediate function derivatives or values. A term vanishing at the selected corners can be nonzero between them or in other backgrounds. No integrated-gradient or Hessian biological interpretation is inferred from the four-corner certificate. Our numerical gradient diagnostic instead evaluates $\nabla_X \hat{y}$ at the supplied encoded query, holding A, mask and context fixed (Simonyan et al., 2013). It is a signed local Euclidean derivative, with no baseline, masking or integration path. It is not an attribution to a feasible molecular edit, and its directions need not preserve one-hotness or joint chemical validity.

A10.3 Training-independent blocks: proof before perturbation

Fix a finite training collection, its preprocessing, and all dropout masks. For R0 suppose $A_{rij} = 0$ for every training record and all nodes i, j for relation r. Consider changing only independently parameterized matrices W_r^ℓ for these absent relations. Embedding states are unchanged because they do not depend on these matrices. If the states entering layer ℓ are unchanged, every absent-relation term $A_{rij} h_j^\ell W_r^\ell$ is zero for every value of W_r^ℓ. Terms on the other relations are unchanged. The total degree depends only on A, and the self transform, bias, residual, fixed dropout mask, LayerNorm and padding mask therefore return unchanged next states. Induction proves identical states through every layer and identical readout predictions on every training example for every value of the selected matrices. This holds for every dropout realization, batch and parameter value of the other blocks, not merely at one zero-gradient checkpoint.

Consequently any data loss that is a function only of these training predictions and fixed labels is independent of the selected blocks, including sums of endpoint losses and pair differences when both endpoints belong to the same covered training collection. Its parameter derivative is identically zero wherever defined. Our diagnostic checkpoints use activity training and do not introduce pair endpoints. The proof does not infer global prediction invariance: a held-out graph with that relation can use the selected block. Nor does it assert invariance of a selection rule that examines additional validation inputs; training-data independence and every possible source of training-procedure information are different statements.

For R1 the effective forward/backward backbone transforms use the respective shared bases W_2, W_5 plus supported residuals. Only residuals with a fixed false gate are independently absent; shared bases cannot be included merely because their own relation count is zero. False-gated residuals cancel from every forward, including held-out graphs, so the same perturbation has a stronger, architecture-imposed invariance. The verified original partition has four absent relations $0, 1, 3, 4$, two layers and 32×32 matrices: $4 \cdot 2 \cdot 32^2 = 8192$ scalars. The corrected models happen to have 8192 false-gated residual scalars in the inspected partitions; these are not 8192 independently unconstrained effective backbone routes.

A zero data derivative does not mean an unchanged parameter. With zero moments and zero data gradient, AdamW still applies the decoupled factor $1 - \eta_t \lambda$ to a parameter represented in the optimizer; regularization can also add an objective derivative. Weight changes do not prove label learning. Conversely an input gradient is not a parameter-loss gradient. The present architecture has no pretrained message matrices or unrecorded shared parameter aliases in these blocks. Structural independence concerns the data term; optimizer/regularization choices remain disclosed.

A10.4 Bounded diagnostic protocol and arithmetic

The protocol used saved v3 deployment checkpoints for R0 and R1 GNN, seeds 1103, 2207 and 3301; all 2518 training inputs, all 1839 APP inputs and all 165 primary-panel inputs; first sixteen lexicographic APP identifiers for raw input derivatives. For each eligible layer/relation block draw fixed Rademacher signs with RNG seed 20260914 and multiply by that block's original root mean square parameter magnitude. Apply scales $0, -0.25, +0.25$ to this fixed direction. Thus the perturbation Frobenius norm in each block is at most one quarter of its original Frobenius norm. Neither labels, diagnostic outcomes nor rank changes select a direction, scale or query. No optimization occurs.

Predictions use deterministic CPU float32 inference in batches of 128, two Torch threads, disabled dropout, original scale restoration, and no clipping or recalibration. Every training prediction must be bitwise equal to its unperturbed value. All endpoint predictions are reused across rectangles.

APP ranks use average ranks among all 1839 rows; these are global diagnostic ranks, not the seventeen assay-pool efficacy scores. Raw-gradient changes cover sixteen queries per checkpoint, not all APP records. B3 diagnostic deployment checkpoints include the source in historical training; they are separate from the source-held-out checkpoints used for the B3 measured performance table. Their zero changes are not a new external-evaluation result. Models are perturbed only in memory, restored exactly, and original checkpoint SHA256 hashes are checked. No modified checkpoint is saved. No training optimizer is constructed or advanced.

A10.5 Experimental support versus the one-hot simplex

For a four-letter one-hot input x, $\sum_c x_c = 1$. The normal vector is proportional to $(1, 1, 1, 1)$, so the gradient projection is $G_c - \frac{1}{4}\sum_d G_d$ (Majdandzic et al., 2023). A fixed-G experimental position occupies a single vertex within this simplex; a Y position permits only C and U. Their missing directions include tangent directions of the full simplex, so subtracting its normal component does not resolve experimental-support absence. For example $f(x) = \theta x_A$ at a fixed-G position is zero on all observed inputs, but its projected gradient is $\theta(3/4, -1/4, -1/4, -1/4)$ in A,C,G,U order and remains unconstrained by those observations.

The actual MAVE-NN input release has 30483 RNA sequences of length nine, with G fixed at position four and C/U at position five. There are three absent nucleotide columns at position 4 and two at position 5, totaling five. Using a DNA T channel for this RNA-U release would incorrectly count thirteen absent columns; the audit corrects that adapter error. An active constant G channel is not absent: for preactivation $u = w_G x_G + b$, $\partial L/\partial w_G = (\partial L/\partial u) x_G$ can be nonzero when $x_G = 1$. Its effect can be confounded with the intercept under this design. In an additive position-specific map five absent columns correspond to five θ_{lc} coefficients; with a dense width-h first layer they can affect $5h$ weights; a shared convolution requires a separate analysis across all receptive-field positions.

MAVE-NN explicitly models genotype-to-phenotype and measurement maps (Tareen et al., 2022). Its BRCA2 exon-17 splice-site assay measures exon inclusion, not siRNA efficacy. The pinned release and layer definitions establish this support calculation. No supplied BRCA2 workshop checkpoint, implementation identifier or claimed attribution percentages could be identified. The categorical input-gradient workshop paper of Majdandzic & Koo (2021) uses synthetic motif data, so it is not silently substituted as that missing experiment. No workshop percentages or additional fitted-model result are invented.

A10.6 Exact rank of the encoding-class contrast map

We continue the notation above: $C \in \mathbb{R}^{q\times m}$ holds the signed contrast coefficients over the m distinct raw endpoint states of a finite recorded panel, and $H \in \{0,1\}^{m\times k}$ assigns each state to its complete-input equivalence class for one encoding instance. An unrestricted deterministic decoder on those classes is a vector $g \in \mathbb{R}^k$ and realises the contrast vector CHg.

Proposition A10.2. *The set of contrast vectors attainable by unrestricted deterministic decoders on these classes is exactly the column space of CH. A fixed linear functional $u^\top$ of the q contrasts is identically zero for every such decoder if and only if $u^\top CH = 0$, so the space of forced-zero functionals has dimension $q - \operatorname{rank}(CH)$. If in addition $C\mathbf{1}_m = 0$, then $\operatorname{rank}(CH) \le k - 1$.*

Proof. Attainability is the definition of a column space. If $u^\top CH = 0$ then $u^\top(CHg) = 0$ for every g; conversely if $u^\top CHg = 0$ for every g, taking $g = e_j$ for each j shows every entry of $u^\top CH$ vanishes. The forced-zero functionals are therefore the left null space of CH, of dimension $q - \operatorname{rank}(CH)$ by rank–nullity. Finally $H\mathbf{1}_k = \mathbf{1}_m$, so $CH\mathbf{1}_k = C\mathbf{1}_m = 0$: the columns of CH satisfy one linear relation and at most $k - 1$ of them are independent. □

Proposition A10.2 bounds a dimension. It does not say that any particular row cancels: row i is forced to zero exactly when the ith row of CH is the zero vector, which is a strictly stronger requirement than a rank deficit. Both statements were computed, not assumed.

For the primary panel, $q = 140$, $m = 165$ and every audited encoding instance has $k = 90$. The forty instances — four methods by ten seeds — induce the same partition of the 165 states, verified

by exact label comparison, so the concatenated constituent blocks of an ensemble are identical and the combined column space equals the single constituent space; every ensemble weight is $1/10$ and none is zero. No rank bound is transferred between differing partitions. With integer C and H and rational forward elimination, $\operatorname{rank}(C) = 140$ and $\operatorname{rank}(CH) = 72$, against the dimension bound 89. Hence the audited encodings impose $140-72 = 68$ independent homogeneous linear constraints on the joint *predicted* contrast vector: every u in the 68-dimensional left null space of CH satisfies $u^\top CHg = 0$ for every decoder g. The recorded contrast vector need not satisfy those constraints, and they are not 68 vanishing individual interactions nor 68 independent biological observations. Explicit exact left-null witnesses were verified to annihilate CH; the machine-readable witnesses are rows of the canonical store rather than printed coefficient dumps. No row of CH is zero, which is consistent with the complete-input test finding no forced cancellation in any of the 140 rectangles.

Let P be the orthogonal projector onto the column space of CH and let $\widetilde{I} \in \mathbb{R}^q$ be the recorded contrast vector. For equal weights $1/q$ and any decoder g,

$$\tfrac{1}{q}\big\|CHg - \widetilde{I}\big\|^2 \;\geq\; \tfrac{1}{q}\big\|(\mathrm{Id} - P)\widetilde{I}\big\|^2,$$

because CHg ranges over the column space and orthogonal projection minimises the distance to it. Moreover, for any fitted contrast vector $p \in S = \operatorname{col}(CH)$ the decomposition $\widetilde{I} - p = (\mathrm{Id} - P)\widetilde{I} + (P\widetilde{I} - p)$ has orthogonal summands, so

$$\tfrac{1}{q}\big\|\widetilde{I} - p\big\|^2 = \tfrac{1}{q}\big\|(\mathrm{Id} - P)\widetilde{I}\big\|^2 + \tfrac{1}{q}\big\|P\widetilde{I} - p\big\|^2.$$

For a fitted masked contrast vector $p \in S = \operatorname{col}(CH)$, write $F = q^{-1}\|(\mathrm{Id} - P)\tilde{I}\|^2$. For a restored predictor p' that need not lie in S, the improvement is

$$\frac{\|\tilde{I} - p\|^2}{q} - \frac{\|\tilde{I} - p'\|^2}{q} = F + \frac{\|P\tilde{I} - p\|^2}{q} - \frac{\|\tilde{I} - p'\|^2}{q}.$$

Consequently, F does not upper-bound improvement after refitting. Only the unrestricted oracle minimum decreases by exactly F when the restored contrast space contains the recorded target vector. Numerically $\frac{1}{q}\|\widetilde{I}\|^2 = 0.06820465$ and $\frac{1}{q}\|(\mathrm{Id} - P)\widetilde{I}\|^2 = 0.00997952$, so the encoding constraints exclude 14.63% of the recorded contrast energy. For the GNN ensemble the split reads $0.06833507 \approx 0.00997952 + 0.05835556$: the same residual is 14.6038% of the fitted squared error, while 85.3962% of it lies *within* the permitted contrast subspace. These two percentages have different denominators — recorded contrast energy and fitted squared error — and must not be conflated. Most fitted-model squared error therefore lies inside the permitted subspace, and the mechanism producing the nearly constant interaction predictions remains unresolved. The residual is a finite-panel representational restriction for these recorded measurements and weights: not a noise-corrected biological error, not a population risk bound, and not a causal allocation of attenuation to representation, architecture, optimisation or data support. The unrestricted decoder used here need not be realizable by the fitted GNN, tree or CNN family. As an implementation check, each fitted ensemble contrast vector lies in S to within 10^{-16}, as Proposition A10.2 requires. This applies the general contrast-space characterization to one recorded panel on one sequence background. The resulting floor is specific to the recorded measurements, encoding and weights. It provides no population-risk guarantee, and no shared-control interval is constructed without a covariance model.

A10.7 Source decomposition identities and what they do not imply

Let $w_i > 0$ be normalised evaluation weights, y_i recorded targets and p_i fixed predictions. For a partition into groups s put $a_s = \sum_{i\in s} w_i$, $\mu = \sum_i w_i y_i$, $\nu = \sum_i w_i p_i$, and let $\mu_s, \nu_s, \operatorname{Var}_s(y), \operatorname{Var}_s(p), \operatorname{Cov}_s(y,p)$ be the group means, variances and covariance under the renormalised weights w_i/a_s.

Proposition A10.3. *With* $V = \sum_i w_i(y_i - \mu)^2$ *and* $\mathrm{MSE} = \sum_i w_i(p_i - y_i)^2$,

$$V = \underbrace{\sum_s a_s \operatorname{Var}_s(y)}_{V_{\text{within}}} + \underbrace{\sum_s a_s(\mu_s - \mu)^2}_{V_{\text{between}}},$$

$$\mathrm{MSE} = \sum_s a_s\big[\operatorname{Var}_s(y) + \operatorname{Var}_s(p) - 2\operatorname{Cov}_s(y,p) + (\nu_s - \mu_s)^2\big],$$

and $\operatorname{Cov}(y,p) = \sum_s a_s \operatorname{Cov}_s(y,p) + \sum_s a_s(\mu_s - \mu)(\nu_s - \nu)$.

Proof. Write $y_i - \mu = (y_i - \mu_{s(i)}) + (\mu_{s(i)} - \mu)$ and expand. The cross term is $2\sum_s a_s(\mu_s - \mu)\sum_{i\in s}(w_i/a_s)(y_i - \mu_s) = 0$ because each inner sum vanishes. The same splitting applied to p gives the covariance identity. For the second display, within group s the weighted mean of $(p_i - y_i)^2$ equals $\mathrm{Var}_s(p-y) + (\nu_s - \mu_s)^2$ and $\mathrm{Var}_s(p-y) = \mathrm{Var}_s(y) + \mathrm{Var}_s(p) - 2\,\mathrm{Cov}_s(y,p)$; summing with weights a_s gives the claim. □

Pooled $R^2 = 1 - \mathrm{MSE}/V$ uses the pooled denominator V. It is not an average of the per-group $R_s^2 = 1 - \mathrm{MSE}_s / \mathrm{Var}_s(y)$, which use different denominators, and R_s^2 is undefined when $\mathrm{Var}_s(y) = 0$. Three distinct mechanisms make most per-group values negative while the pooled value is positive: a small $\mathrm{Var}_s(y)$ in the denominator, a group-specific calibration offset $(\nu_s - \mu_s)^2$, and unequal group sizes, since a count of groups is not a row weight.

These identities do not license the inference that between-group variance dominates or that a model has learned group levels. On the released grouped benchmark under equal row weights, V_{between} is 3.98% of V across the seventeen source categories and 3.33% across the ten study-linked components, so the between-group part is the small part. Moreover, for this GNN evaluation the covariance splits as $+0.00306249$ within and -0.00055903 between, so the positive pooled score is carried by the within-source component while the between-source component is negative. That is an observed evaluation covariance for this arm only: the chemistry tree shows a positive between-source covariance on the same rows, and an unfavourable covariance does not establish the absence of source-dependent learning. Source identifiers, study-linked components and assay identifiers are three different partitions of the same rows and are not renamed as one another; the recorded panel does not resolve assay identity, so no assay-level partition is asserted.

Two further qualifications are kept explicit. A group-mean oracle, and any score computed after centring predictions on evaluation-group means, read the evaluation labels themselves; they are retrospective diagnostics, not deployable corrected models and not model-risk lower bounds. And a negative within-group R^2 does not prove zero discrimination or the absence of chemistry information: calibration, a small label variance and measurement noise each produce it on their own. All scenarios here reuse the same recorded rows and the same saved model families, so the reported comparisons are dependent re-analyses rather than independent replications.

A10.8 Retained calibration results

The retained calibration table supplements the main results. Cohort and adjudication counts appear in the data and robustness sections; Appendix A10.16 gives deduplicated historical fit counts.

Cohort	Original v3 model	MSE	R^2	r	Pred. SD	Bias
APP	R1 GNN	0.1625	-0.001	-0.018	0.0072	-0.0081
	R1 no-message	0.1630	-0.005	0.015	0.0085	-0.0286
	Chemistry tree	0.1896	-0.168	-0.021	0.0103	-0.1644
	Token CNN	0.1644	-0.013	0.080	0.0036	-0.0483
S7	R1 GNN	0.1093	-0.215	0.112	0.0059	0.1403
	R1 no-message	0.1047	-0.164	0.117	0.0066	0.1231
	Chemistry tree	0.0919	-0.021	0.017	0.0089	-0.0441
	Token CNN	0.1017	-0.130	0.108	0.0034	0.1092

Table 13: Unchanged v3 released-trained activity results on original response scales. Label SD is 0.4028 on APP and 0.3000 on S7. Prediction spread and mean calibration are distinct from discrimination; no test-fitted adjustment is used.

A10.9 Current constants and model comparisons

For each evaluation identity, weights are either one per row or inverse study-component size normalised to average one. Means, variances and losses divide by total weight. We report pooled $R^2 = 1 - \mathrm{MSE}/V$; it is undefined when $V = 0$, and it is never averaged across folds or groups. Training constants are fold-specific training-mean predictors; their pooled prediction SD need not be zero. The test-mean constant is retrospective, not deployable and not a universal risk floor. APP

and S7 labels and weights are identical across scenarios and each contains a single study component, so their row and equal-component weightings coincide and are reported once. The full identity hash includes identifiers, labels, weights and the declared response convention, and every per-evaluation row remains in the canonical store table `activity_metrics`. Grouped coverage, targets and retrospective oracles for all four cohort/weighting combinations (2,607 and 1,003 rows over 150 and 149 sequence components; oracle MSE 0.062573, 0.075091, 0.059018 and 0.052895) are rows of `activity_metrics`, with APP and S7 in main Table 2; each has one study component, so their two weightings coincide. The oracle is the retrospective evaluation mean, not a deployable control; essential controls appear in Appendix A10.12.

Indexed source decompositions. The indexed decomposition rows reuse saved predictions; nothing was refitted. Each identity of Proposition A10.3 was verified to machine precision, and each variance identity was verified a second time by centring rows on their own group means. The summaries concern multi-source grouped cohorts, where the decomposition is informative; APP and S7 each fall in one source family and one study component, so $V_b = 0$ there by construction. All 304 decomposition records for every method, weighting and partition, the retrospective group-mean oracle (which equals V_w exactly), the group-centred loss and the within/between covariance split quoted in the main text are rows of `variance_decomposition`; neither retrospective quantity is a deployable corrected model.

Per-source denominators. Per-source $R_s^2 = 1 - \mathrm{MSE}_s / V_s$ uses the group label variance as its denominator and is undefined when that variance is zero. These are category-level summaries, not independent replications. All 1,144 per-source records for every campaign, cohort, weighting and method, each with its own retained denominator V_s and row count, are rows of `variance_sources`; counts of positive and negative ratios do not weight observations.

A10.10 Endpoint and marginal-effect reconstruction

The source-held-out export is the authoritative input; the source-included deployment checkpoints of the perturbation experiment are not used. The 560 rectangle-corner appearances reduce to 165 measured endpoint states (one shared reference, fourteen antisense, ten sense, 140 combinations) whose repeated labels agree exactly. Ensembles are the mean of the ten actual final seeds, excluding the saved ensemble row and the deterministic seed-0 rows; they reproduce the stored ensemble to 1.3×10^{-16} and the interaction errors of Table 1 exactly.

Method	MSE	R^2	r	Pred. SD	Label SD	Retro. mean	Train const
GNN	0.077350	0.031844	0.184788	0.040265	0.282656	0.079894	0.080173
No-msg	0.077353	0.031810	0.185135	0.038465	0.282656	0.079894	0.080173
Tree	0.078546	0.016873	0.167266	0.076857	0.282656	0.079894	0.080173
CNN	0.077276	0.032767	0.190711	0.045136	0.282656	0.079894	0.080173

Table 14: Endpoint prediction over the 165 deduplicated states on the original response scale. The retrospective column is the MSE of the evaluation-mean predictor; the final column is the MSE of a genuine training-fitted constant recovered from the ten fit records, never from these evaluation labels.

Method	Family	n	MSE	Zero MSE	r	Pred. SD	Label SD
GNN	AS effect	14	0.049856	0.051986	0.028779	0.024352	0.170616
GNN	SS effect	10	0.003335	0.000855	0.286065	0.034469	0.013953
GNN	Interaction	140	0.068335	0.068205	-0.090263	0.001589	0.222442
No-msg	AS effect	14	0.050546	0.051986	0.019753	0.021060	0.170616
No-msg	SS effect	10	0.003236	0.000855	0.269464	0.034474	0.013953
No-msg	Interaction	140	0.068281	0.068205	-0.034429	0.001652	0.222442
Tree	AS effect	14	0.042903	0.051986	0.116779	0.061459	0.170616
Tree	SS effect	10	0.008423	0.000855	0.286572	0.064857	0.013953
Tree	Interaction	140	0.071939	0.068205	-0.177024	0.017767	0.222442
CNN	AS effect	14	0.049311	0.051986	0.055427	0.023715	0.170616
CNN	SS effect	10	0.004040	0.000855	0.244159	0.041626	0.013953
CNN	Interaction	140	0.068409	0.068205	-0.046847	0.002879	0.222442

Table 15: The three effect families with directions and response scale fixed: antisense $f(A,0) - f(0,0)$, sense $f(0,B) - f(0,0)$ and the original reference-based interaction. Seed dispersion is separate and no independent-marginal interval is claimed.

A10.11 Encoding and rank ablation

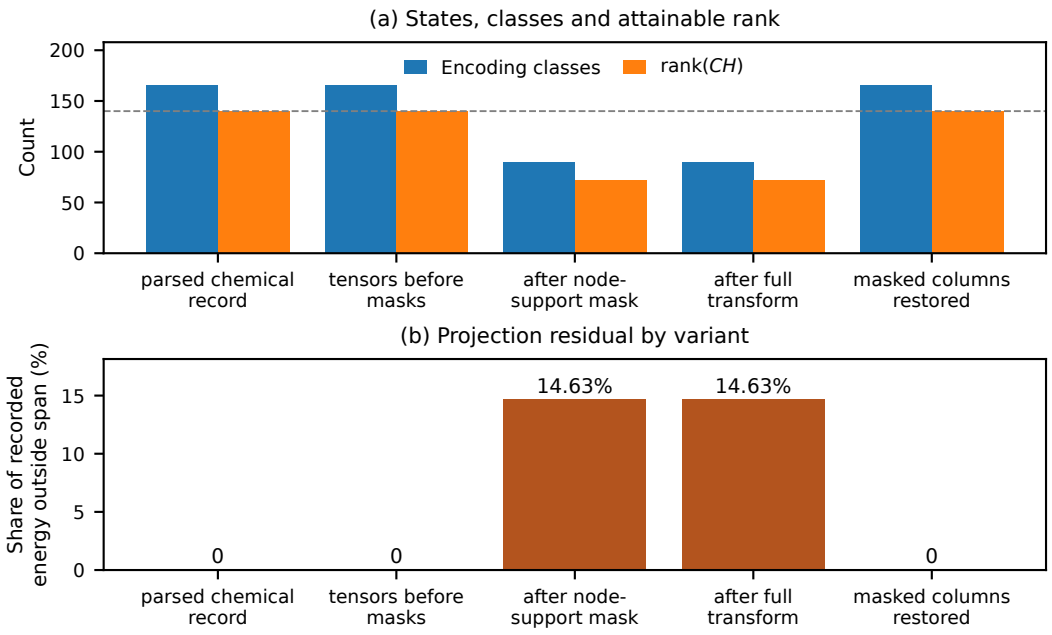

Figure 6: Label-independent encoding ablation on the fixed panel, with the raw states, endpoint identities, response convention and C held fixed. (a) Encoding classes and the exact attainable rank per variant; the dashed line is $\mathrm{rank}(C) = 140$. (b) Share of recorded contrast energy outside the attainable span, linear scale. Full-row-rank variants have exact zero residual in real arithmetic; their computed residuals are roundoff. Parsed chemistry, the pre-mask tensors and the mask-restored diagnostic all keep 165 classes and full rank; applying the training-only node-support mask alone collapses them to 90 classes and rank 72. Higher rank means more representable contrasts, not better prediction; exact class counts, ranks, null dimensions and residual shares per variant are rows of `encoding_ablation`.

Strand	Indistinguishable variants	Positions	Size
AS	JC-A1, JC-F1, JC-S1	3, 18	3
AS	JC-A2, JC-F2, JC-S2	4, 18	3
AS	JC-A3, JC-F3, JC-S3	3, 4, 18	3
SS	DO003, DO004	17	2

Table 16: All non-singleton strand classes, determined by exact complete inputs across every partner. Other strand states are singletons. Full positional names, raw records and equality witnesses remain in the store.

A10.11.1 Chemical factorization and complete rank argument

The raw panel is the full product of fifteen antisense and eleven sense states, including reference W053/W207. Two antisense states are equivalent only when their complete post-mask inputs coincide for every sense partner; sense equivalence is defined symmetrically. This compares X, A, mask, C with shapes, dtypes and exact bytes. All 27,225 ordered state pairs satisfy the stronger biconditional: complete inputs coincide if and only if both strand classes coincide. There are nine antisense and ten sense classes; both reference classes are singletons.

Within each antisense position pattern in Table 16, JC-A denotes the recorded 2-deoxy-2-N,4-C-ethylene-locked chemistry, JC-F the recorded 2,4-carbocyclic-ethylene-bridged locked chemistry, and JC-S the recorded 2,4-carbocyclic locked chemistry. Their raw names differ but the fitted mask removes their distinguishing columns. Sense DO003 (2-aminoethyl) and DO004 (2-guanidinoethyl), both at position 17, also merge. Modification positions and modified versus unmodified indicators can remain distinguishable; an encoded equality is not chemical equivalence. Restoring masked columns to the encoding recovers rank, but not additively. A minimum of three restored chemistry columns recovers full rank; exhaustive search over all smaller sets establishes that lower bound, and exactly six distinct three-column sets attain it. One of these six minimal triples, the lexicographically first, is the two carbocyclic locked-nucleic-acid columns, indices 39 and 40, and the 2-aminoethyl column, index 42. In isolation those three restore 27, 27 and 8 of the lost ranks, 62 in total, against 68 jointly, and no single column restores more than half, so these marginal rank gains do not provide an additive attribution of the joint rank recovery. Rows are in `column_localization`, `column_localization_minimal` and `column_localization_summary`.

Index product classes by (a, b), reference $(0, 0)$, and arbitrary decoder values g_{ab}. The unique interaction coordinates are $t_{ab} = g_{ab} - g_{a0} - g_{0b} + g_{00}$ for nonreference a, b. These $(9-1)(10-1) = 72$ coordinates are independent: for any vector v, set $g_{ab} = v_{ab}$ on nonreference pairs and every reference-row/reference-column value to zero, yielding $t = v$. Every original rectangle maps to one coordinate, and every coordinate occurs because the panel is a full product. Thus $CH = RK$, with K this surjective coordinate map and R its repetition matrix. The columns of R have nonempty disjoint supports, hence are independent, proving rank 72.

For each identical-row group choose representative q_0. Each $e_q - e_{q_0}$ for a nonrepresentative row annihilates CH. These vectors are independent because each nonrepresentative coordinate appears in only its own vector. There are $140-72 = 68$ of them, equal to the left-null dimension, so they span it: all restrictions are duplicate-contrast equalities, with no additional general linear dependencies or individually zero row. Rational elimination independently checks the rank and every witness. The forty audited instances share this partition, not forty unrelated representation designs. Restoring masked columns here is an encoding analysis; the refit that feeds restored features to the predictor is the intervention described below in Appendix A10.11.3.

A10.11.2 Constructed factorial sweep: what the floor binds

The orthogonal split $\|y - p\|^2 = \|(\mathrm{Id} - P)y\|^2 + \|Py - p\|^2$ holds for any $p \in S$, and a predictor that consumes only the merged classes has $p = CHg \in S$ by construction, so its out-of-subspace error equals the analytic floor as algebra. To test whether that equality is a property of the input restriction or merely of the measurement, the sweep runs two arms per cell. A protocol whose SHA256 (`744c6053...`) the fitting stage re-asserts, and which the append-only execution ledger records before the fits, fixed eight surjective level-merging encoders on two constructed antisense-by-sense designs (6×5 and 8×6), spanning rank deficit 0 to 29 (at most $19/20$ of q), ground truth

$y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$ with γ drawn once per setting and zeroed on the reference row and column, marginal-only training cells so every interaction is extrapolation, two architectures (a 16-unit MLP and a two-node message-passing model) and three retained seeds each.

All 48 merged-class fits (96,000 optimizer updates, 0 failures) passed a gate stricter than the stated 10^{-10} floor tolerance: it also requires zero within-class prediction deviation and a split residual below 10^{-10}. The largest absolute floor gap is 1.1×10^{-16}, the largest split residual 5.6×10^{-16}, within-class prediction deviation is exactly zero in all 48, and an independent exact-rational floor from the left null space of CH matches the projector floor to 5.6×10^{-17} in all eight settings. The zero-deficit control has floor exactly zero in the rational arithmetic, with realized perpendicular residual $q^{-1}\|(\mathrm{Id} - P)(y - p)\|^2$ below 2×10^{-31}; this is distinct from the stored `zero_control` column, which is the no-interaction baseline. Because γ is redrawn per setting, floor magnitudes are not comparable across deficits.

A second pre-registered arm (SHA256 `7dce9922...`, frozen before any of its fits) changes only the input: the same 48 fits receive a one-hot over the $a + b$ raw levels instead of the $a_* + b_*$ merged classes, so their contrast vectors are not confined to S. The registered prediction was that this arm's out-of-subspace error would fall below the floor wherever the deficit is nonzero. *It did not.* The raw-index arm does not enforce equality within the original merged classes: its within-class prediction deviation ranges from 0.55 to 3.76, compared with exactly zero in every merged-class fit. Its out-of-subspace error differs from the original floor, but among the 42 nonzero-deficit fits only 8 fall below that floor and 34 lie above it, with median ratio 1.33, maximum 15.97 and minimum 0.81. No setting has every seed below the floor, and four have every seed above it. The raw-index arm is not confined to the original contrast subspace. Paired total prediction MSE on the same settings, models and seeds is higher for the raw-index arm in 28 of the 42 nonzero-deficit fits and lower in 14, with median ratio 1.179 over the range 0.577 to 3.903. Under the reported marginal-only training protocol, removing the original class-equality restriction did not systematically reduce prediction error. The training data contain no observed targets for the nonreference joint cells. Predictions on those cells still depend on the fitted architecture, shared parameters and optimization procedure. The six zero-deficit pairs are identical, as expected where the two inputs coincide. These are synthetic-only fits on constructed data. A smaller out-of-subspace error does not by itself establish a smaller total prediction error, and these experiments provide no evidence about siRNA efficacy. Per-setting rows are in `floor_synthetic`, `floor_synthetic_summary`, `floor_synthetic_rawindex` and `floor_synthetic_rawindex_outcome`.

A10.11.3 MASK-RESTORATION PROTOCOL

The recorded protocol uses the removed oracle floor $F = 0.009980$ as a comparison scale. It labels the ensemble improvement Δ on all 140 rectangles as near floor ($F/2 \le \Delta \le F$), near zero ($|\Delta| < F/2$), above the floor ($\Delta > F$), or worse ($\Delta \le -F/2$). These are descriptive outcome categories, not bounds on the improvement attainable after refitting. The unmasked arm re-ran the 36-fit development grid and chose the masked configuration, and ten masked refits reproduced the historical signatures, epochs and predictions. Training inputs are identical across arms, while 173 of 198 validation rows and 157 of 165 B3 states carry restored columns. Comparing each seed's selected parameters between the unmasked and masked refits, 5 of the 10 seeds are bit-identical: exactly those whose validation selection chose the same epoch in both arms. The 100 affected rectangles improved by $+0.000028$ and the 40 unaffected by -0.000011, against a zero-interaction control of 0.068205 in both arms; antisense MSE went from 0.049856 to 0.050204 with its 0.018972 floor removed and sense from 0.003335 to 0.003073. Per-seed rows are in `mask_scores` and `mask_paired`.

A10.12 Model versus training constants under both weightings

Scen.	Weights	Model	Constant	Difference	Lower	Upper	Grp
primary	rows	GNN	row const	-0.001217	-0.004579	0.020462	10
primary	rows	GNN	grp const	-0.006226	-0.010077	0.021086	10
primary	rows	Tree	row const	-0.006784	-0.013080	0.003109	10
primary	rows	Tree	grp const	-0.011792	-0.016886	0.000805	10
primary	equal study	GNN	row const	0.015532	-0.007321	0.038432	10
primary	equal study	GNN	grp const	0.015232	-0.007620	0.038511	10
primary	equal study	Tree	row const	-0.007124	-0.025018	0.008232	10
primary	equal study	Tree	grp const	-0.007423	-0.024537	0.007632	10
excluded	rows	GNN	row const	0.007004	-0.005002	0.040376	9
excluded	rows	GNN	grp const	0.009191	-0.001189	0.038475	9
excluded	rows	Tree	row const	-0.004148	-0.013327	0.009386	9
excluded	rows	Tree	grp const	-0.001961	-0.008954	0.006106	9
excluded	equal study	GNN	row const	0.019636	-0.001169	0.042041	9
excluded	equal study	GNN	grp const	0.021490	0.003128	0.042072	9
excluded	equal study	Tree	row const	-0.001319	-0.019699	0.014955	9
excluded	equal study	Tree	grp const	0.000535	-0.012901	0.012998	9

Table 17: Fixed-ensemble error minus each training-fitted constant, with paired conditional 95% percentile intervals from resampling whole study-linked components under both weightings at the existing fixed analysis seed over 10,000 draws, predictions and selection held fixed. Constants are training-fitted and never an evaluation mean. The tree has lower observed MSE than both constants under row weighting, but not against the group-mean predictor under equal-component weighting on the excluded cohort; all four source-excluded tree-versus-constant intervals include zero. Ten and nine components respectively: conditional, few-component intervals whose endpoints are not subtracted.

Leave-one-study-component-out rescoring omits one evaluation component and renormalises the stated weights, leaving the fitted models and their training-partition constants unchanged. Across 152 omissions the descriptive ordering changes in 8 cases and no retrospective denominator was near zero. These are influence diagnostics, not new independent studies.

A10.13 Practical size of the bounded perturbation

Method	Seed	Scale	Changed	Mean ovl.	Min ovl.	Med. disp.	Max disp.
GNN	1103	-0.250000	0	1.000000	1.000000	0.000000	0.000000
GNN	1103	0.250000	0	1.000000	1.000000	0.000000	0.000000
GNN	2207	-0.250000	0	1.000000	1.000000	0.000000	0.000000
GNN	2207	0.250000	0	1.000000	1.000000	0.000000	0.000000
GNN	3301	-0.250000	0	1.000000	1.000000	0.000000	0.000000
GNN	3301	0.250000	0	1.000000	1.000000	0.000000	0.000000
original_gnn	1103	-0.250000	4	0.952941	0.800000	0.009615	0.093750
original_gnn	1103	0.250000	3	0.964706	0.800000	0.009615	0.089888
original_gnn	2207	-0.250000	0	1.000000	1.000000	0.004808	0.077778
original_gnn	2207	0.250000	4	0.952941	0.800000	0.004808	0.093750
original_gnn	3301	-0.250000	4	0.952941	0.800000	0.009615	0.067416
original_gnn	3301	0.250000	0	1.000000	1.000000	0.009615	0.062500

Table 18: Within each of the seventeen exact APP assay pools, top-five selection mass under the existing fractional boundary-tie rule, and average-rank displacement normalised by $n - 1$. The pools overlap and come from one patent family, so these are not independent trials.

Arm	Seed	Scale	Max coordinate	Max L_2	Max relative L_2
R1	1103	-0.250000	0.000000	0.000000	0.000000
R1	1103	0.250000	0.000000	0.000000	0.000000
R1	2207	-0.250000	0.000000	0.000000	0.000000
R1	2207	0.250000	0.000000	0.000000	0.000000
R1	3301	-0.250000	0.000000	0.000000	0.000000
R1	3301	0.250000	0.000000	0.000000	0.000000
R0	1103	-0.250000	0.005795	0.071476	0.240187
R0	1103	0.250000	0.005870	0.088127	0.336519
R0	2207	-0.250000	0.004085	0.055905	0.212644
R0	2207	0.250000	0.005719	0.075019	0.294461
R0	3301	-0.250000	0.003599	0.054309	0.288555
R0	3301	0.250000	0.002808	0.019713	0.108521

Table 19: Raw input-gradient changes on the same sixteen fixed APP queries. Each statistic maximizes across those queries; the coordinate statistic also maximizes across coordinates. Relative change divides each query gradient-change norm by its baseline norm, with threshold 10^{-8}; no near-zero references occur. Other graph inputs are fixed. These are local encoded sensitivities, not feasible chemical interventions, and R1 is the expected structural null control.

A10.14 Official published-route preflight

The official repository documents a Docker route and an `ENsiRNA-mod/easy_run.py` entrypoint. The manifest for the documented tag resolves to digest `sha256:1a9c8b80a2d5b5943770fb5e736264cb5234997181df98d7704bde673f09167f` with 35 layers and 20.60 GiB of compressed layers, against 11.12 GiB free on the working filesystem, and no container runtime is installed; no pull was started because the known requirement does not fit. The documented Linux route additionally requires a licensed Rosetta installation and pre-folded PDB geometry, and the released inference code shells out to Rosetta binaries and consumes atom identities and RNA sequence embeddings. The training partition of the released checkpoint is not declared, so overlap with Bramsen is unknown and no held-out predictive claim is made. The earlier pinned partial-branch chemistry audit is unchanged.

Protocol	GNN	No-message	Tree	CNN
v1 original	0.521	0.497	0.565	0.453
v2 factorial	0.538	0.498	0.574	0.530
v2 deployment	0.397	0.367	0.631	0.510
v3 group selected	0.534	0.577	0.445	0.583
Source-excluded	0.400	0.374	0.696	0.421
Primary-assay	0.565	0.588	0.514	0.563

Table 20: Common tie-aware APP mean top-five percentile across the identical seventeen assay pools; larger is preferred. The pools overlap and come from one patent family, so they are not seventeen independent experiments.

Method	Row MSE	Comp. MSE	r	Pred. SD	Sign
Activity GNN	0.2351	0.3560	0.083	0.036	111/137
Pair GNN	0.2221	0.3211	0.248	0.038	111/137
Pair no-message	0.2220	0.3196	0.228	0.043	111/137
Pair tree	0.2438	0.3101	0.120	0.167	109/137
Pair ridge	0.2193	0.2888	0.226	0.148	105/137
Zero	0.2808	0.4342	–	0.000	0/137
Training mean	0.2638	0.3336	-0.076	0.064	111/137

Table 21: Unchanged held-out measured B2 errors. Comp. gives each of twelve sequence components equal mass. Sign counts use 137 non-ties under the operational 0.02 band; the majority control also reaches 111/137.

Exact algebra records. The exact matrix rows, all left-null witnesses and full-precision projection components remain in `chemical_exact_matrices`, `chemical_left_null_witnesses` and `contrast_projection`. Section A10.6 gives the complete proof; the main text gives the numerical error decomposition.

Complete perturbation records. Per-method, per-scale perturbation outcomes – maximum training change, APP change, rank range, gradient change and B3 contrast change – are rows of `perturbation_summary` in the canonical store.

A10.15 Published preprocessing: exact scope

ENsiRNA-mod commit `028824341635903f3c661f5d1cc737de106493d5` uses radius-two 512-bit Morgan chemistry fingerprints. Its partial branch admits 140/140 B3 rectangles with no cancellation; complete inference also requires atom IDs, geometry and RNA-FM. MEG-mod commit `c335a4c69d56ef73754677da8bfe627c8112b350` supplies the UniMol dictionary: 117/140 rectangles pass its modification branch without cancellation, 23 have unsupported annotations. Undeclared strand-token construction/dimensions block its complete forward pass. Unknown inputs are never zero-filled.

Eight conformance pairs per method include an unchanged control, sugar/phosphate edits and unsupported names. ENsiRNA has three partial collisions, three distinct pairs, one control and one unsupported pair; MEG has two, two, one and three. ENsiRNA fingerprints coincide for locked/altritol nucleic acid, 2-fluoro/fluoroarabino and uracil/pseudouridine; MEG embeddings coincide for the first and third. Other branches can distinguish them: these are neither chemical-equivalence claims nor whole-predictor certificates. ModMapper exposes a PHP form but no local parser/ontology release, blocking all 140 rectangles and eight conformance inputs; no job is submitted. Raw names, downloaded bytes, hashes and witnesses remain in the store.

A10.16 Historical fit accounting

Preserved campaigns contain 37, 210, 2,312 and 792 unique siRNA fit records (3,351 total and 1,934,535 optimizer updates); the separate v1 twenty-update profile is not an efficacy fit. This work adds 56 siRNA mask-restoration fits (21,513 optimizer updates; Appendix A10.11.3), twelve MAVE-NN support-intervention attempts, six with retained per-seed results (Appendix A10.24), and 96 synthetic fits on constructed factorial designs, 48 in each of the merged-class and raw-index arms, each arm with 96,000 optimizer updates and no failures (no siRNA checkpoint read or modified); the raw-index arm ran outside the pipeline's execution ledger, and its counts come from that run's own record. The reported 62,424 MAVE-NN updates cover the retained invocation, not a complete total across both invocations. Historical fits, new attempts, retained outcomes and failed workflow invocations are distinct accounting units. Source, training/selection, response and common-row changes remain identified in `protocol_comparisons` and `protocol_matched_changes`; prior ranking corrections and source discrepancies are retained above.

A10.17 Canonical execution and evidence mapping

Set `PYTHONDONTWRITEBYTECODE=1` and run `python analysis/structural_attribution_audit/audit.py all`, then replace `all` with `verify-resume` to check unchanged reuse. The same directory holds the manifest, SQLite results and single report with exact historical dependency paths. No scientific input is copied or modified. Styles and builds remain outside the four-entry source; the source ZIP independently compiles. Recorded diagnostic and build costs exclude additional nonzero administrative work.

A10.18 Independent-predictor fidelity and coverage

Two candidates were allowed: ENsiRNA-mod at commit `028824341635903f3c661f5d1cc737de106493d5` and MEG-mod at `c335a4c69d56ef73754677da8bfe627c8112b350`, both still at those commits. Each recorded failure was reclassified against a compatible environment supplying Torch 2.11 with CUDA, PyTorch Geometric, ViennaRNA 2.6.4 and `rna-fm` with cached weights. The Python

dependencies listed above were available in the compatible environment. The remaining obstacles included missing native structure-generation software and unresolved release semantics.

MEG-mod's released `BAN_graph.py` contains no import statements at all, which is why its documented entry point stops at line 10 with `NameError: torch`: a released-source defect, not a missing package. Restoring the omitted block is additive, verified by deleting it again and recovering the released bytes exactly. The repaired module then executes, and its forward's only remaining unbound globals are `s_tokens`, `a_tokens`, `max_L_mod_sense` and `max_L_mod_anti`, which build the padded modification-token tensors consumed by `bcn_mod`, one of the two halves entering the output block. The released checkpoint holds sixteen `bcn_mod` tensors, so that branch is required under the strict load its own prediction script performs, and their shapes fix `embed_dim`=1536, an admissible binding. No shape states what a modification token contains, in what order, or how many there may be; that semantics appears in no released file, so the route stopped rather than guessed. The release also omits the `requirements.txt` its README directs users to install.

The earlier record that ENsiRNA-mod has no local checkpoint is corrected: five released checkpoints ship inside the repository at 9,523,522 bytes each, and `checkpoint_1.ckpt` deserialises unrepaired into `model.mask_model.RNAmaskModel` with 2,365,808 parameters exposing the documented `test` interface. Exactly one input is missing. `X`, the per-residue coordinates, is that interface's first argument and is consumed twice, once to build the nine-nearest-neighbour graph and once as coordinates for the equivariant network, so without real structures there is no graph and substituted coordinates would be invented geometry. The release obtains them only from Rosetta `rna_denovo`, absent here and distributed behind a licence-and-registration page; the linked prefolded archive covers the authors' duplexes, not these cohorts. The container route was not retried, since no Docker or Podman is installed and the recorded 22.12 GB compressed image still exceeds the free space.

Zero of 156 B2 pairs (312 endpoints, 12 sequence components, 26 distinct base duplexes), zero of 165 B3 endpoints and zero of 140 B3 contrasts therefore received a complete external prediction, and 477 endpoint records would have needed native inputs. Support was fixed from native input requirements, never from observed error, and nothing was zero-filled: these are unassessed cases, not model errors. Supervised efficacy-training and selection overlap stays undocumented for both checkpoints, so neither could have supported a held-out claim even had it run. The frozen cap was two candidates, 1,200 seconds of setup, two CPU threads, no GPU and at most six fitting attempts only after fidelity and a runtime estimate; with no executable forward and no geometry the fitting budget stayed zero and all six attempts are unspent. No optimizer was constructed, no container pulled and no repository file added. Pins, diffs, provenance, checkpoint identities, executed commands, denominators and a frozen but never executed external evaluation rule are in the `recovery` tables of the canonical store. Independent predictive validation remains incomplete.

A10.19 Does the audit identify unreliable predictions?

Two fixed scores test usefulness: training-support absence among chemical entries relevant to an input or edit (A), and maximum ensemble-effect change over the two existing bounded probes (B); lower scores retain supposedly more reliable cases. Three fixed seeds define each predictor, its disagreement score and its probes, and R0 deployment diagnostics are never mixed with source-held-out R1 errors. Novelty is nearest-training positional-chemistry Jaccard distance, coverage is 20, 40, 60, 80 and 100% with fractional ties and equal sequence-component base weights, and a constant score fractionally retains every case, supplying no ranking information (Appendix A10.20).

Retention score	Model MSE	Zero MSE	Excess MSE
Support absence	0.434741	0.434166	+0.000575
Probe sensitivity	0.356843	0.356330	+0.000513
Ensemble SD	0.363586	0.362827	+0.000759
Chemical novelty	0.434741	0.434166	+0.000575
Random	0.451729	0.451135	+0.000594

Table 22: R0 B2 at 60% observation mass: 93.6 fractional pairs out of 156, twelve sequence components. Each zero control uses identical retained membership and weights. Random averages 200 frozen selections. These three-seed activity-only predictions are distinct from historical pair-supervised estimates.

The primary comparison, B minus ensemble disagreement in excess-over-zero MSE at 60% coverage, is -0.000246 with paired conditional 95% interval [-0.001059, +0.000561]. It does not establish a utility advantage. Lower raw pair error under B mostly follows lower zero-effect error on the retained pairs, which does not demonstrate learned sequence-specific chemistry. A and novelty are constant across B2, B is constant for R1, and on source-held-out B3 it remains zero despite poor interaction recovery. Secondary APP and within-assay comparisons establish no general diagnostic superiority.

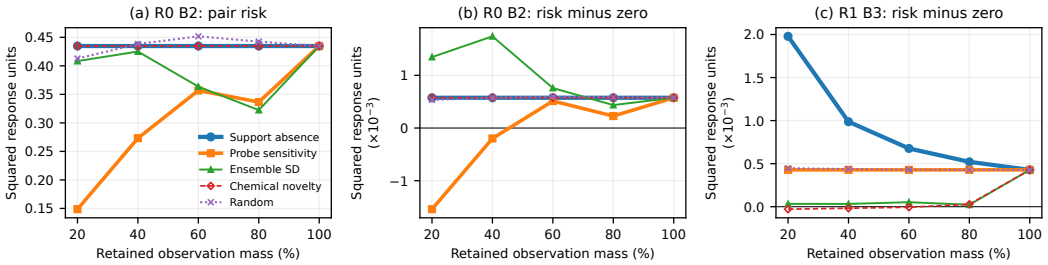

Figure 7: Retrospective utility at all frozen coverage levels. Raw and excess-over-zero risk differ because selection changes the measured-effect distribution. Panels (a,b) use the identical R0 deployment ensemble on B2; (c) uses the separate source-held-out R1 ensemble on B3. B3 has one background; these curves are descriptive, and the frozen primary comparison is at 60%. Constant scores coincide: support absence with chemical novelty on B2 (a,b), probe sensitivity with random retention for R1 (c).

A10.20 Diagnostic utility: complete protocol and estimands

All cohorts were previously inspected: this is a frozen retrospective analysis, not preregistration. R0 and R1 deployment ensembles use seeds 1103, 2207 and 3301 with identical members for predictions, disagreement and probes; APP's 1,839 endpoints and B2's 156 bundles were excluded from their supervised training and selection, and one patent family with existing sequence components does not certify external generalization. Deployment B3 is source-exposed and descriptive, while a separate R1 outer0 ensemble excludes Bramsen and scores one previously inspected B3 background. Checkpoints, memberships, masks, response conventions and evaluation identities are recorded jointly; no deployment score is paired with an outer0 error.

Let p_{ij} denote seed j's endpoint prediction, $J = 3$, $q_{cj} = \sum_i a_{ci} p_{ij}$ and $\bar{q}_c = J^{-1} \sum_j q_{cj}$. Endpoints use one coefficient one, B2 changed minus reference, and B3 its marginal and four-corner coefficients; the target is the same signed combination of measured activities, response values are unchanged and no calibration reads evaluation labels.

Support absence (A). Use the actual schema's half-open chemical columns $[39, 84)$ and original adjacency. For an endpoint the relevant set is its active chemistry occurrences and directed edges. For a contrast it is the union of active entries that differ across corners. If r_{ce} indicates this set and u_e is one when the corresponding node-feature or relation type never occurs in the exact training

partition, define $A_c = \sum_e r_{ce} u_e / \sum_e r_{ce}$, or zero for an empty set. Entries count once; this is an occurrence score, not a learned effect size, and all evaluated names are in the fixed schema. A preliminary offset $[36, 81)$ was corrected against the manifest and encoders; superseded scores and protocol remain preserved, and the correction was not chosen for better outcomes.

Probe sensitivity (B). The existing RNG 20260914 generates fixed Rademacher directions at each block's original RMS. With $s \in \{-0.25, +0.25\}$ define $B_c = \max_s |J^{-1} \sum_j q_{cj}^{(s)} - \bar{q}_c|$. This is a maximum over two tested ensemble probes, not a certified worst case. R0 perturbs only absent independent transforms and R1 only false-gated residuals, never shared bases (Appendix A10.3). Existing deployment zero/probe checks were reused, and three outer0 checkpoints received full-training checks at zero and both scales with bitwise-invariant predictions, unchanged B3 outputs, exactly restored parameters and retained checkpoint hashes. No perturbed checkpoint is saved.

Comparators. Disagreement is $[J^{-1} \sum_j (q_{cj} - \bar{q}_c)^2]^{1/2}$ using actual per-seed contrasts, preserving endpoint covariance. Novelty is nearest-training Jaccard distance between sets of (strand-slot, pre-mask chemistry-column) pairs with empty/empty distance zero, a contrast taking its largest endpoint distance, and the same training partition defines novelty and support. Higher always means less reliable; no reversal, combination or threshold is optimized. Random retention averages 200 fixed uniform-score vectors per model/cohort block at RNG 20260916.

Retention and risk. For n objects and $f \in \{.2, .4, .6, .8, 1\}$ let t be the smallest score threshold reaching fn objects. Set $r_i = 1$ below t, zero above it, and $(fn - n_{<t})/n_{=t}$ at equality. Thus $\sum_i r_i = fn$ without ID or outcome tie-breaking and a constant score gives $r_i = f$ everywhere. Base weights w_i are inverse sequence-component counts (equal object weights for B3); renormalize $v_i = w_i r_i / \sum_k w_k r_k$ and compute $R = \sum_i v_i (\bar{q}_i - y_i)^2$, $R_0 = \sum_i v_i (b_i - y_i)^2$ and $E = R - R_0$. Endpoint b_i averages the actual checkpoint training means; contrast $b_i = 0$. Report fractional count, positive-membership count, retained base-weight mass and effective sample size $(\sum_i w_i r_i)^2 / \sum_i (w_i r_i)^2$. A smaller R may just retain smaller measured effects, hence the matched E comparison.

The primary estimand is $E_B(.6) - E_{\mathrm{SD}}(.6)$ for R0 B2, negative favoring B; all other coverages, scores and comparators are secondary. Within-assay checks require at least ten objects and the same fractional rule, and overlapping pools are descriptive, not independent replication. No candidate-selection improvement is inferred from global rank movement.

Uncertainty. Draw 2,000 multinomial resamples of observed sequence components at RNG 20260916, applying identical multiplicities to the retained numerator and denominator for each score. Fits and original selection memberships stay fixed; recompute paired risk and excess-risk differences and take the 2.5th/97.5th percentiles. No-retained-mass resamples are undefined and counted, not zero-filled (up to four in a secondary comparison); all primary resamples are defined, and random averages its defined frozen selections within each resample. Twelve B2 components from one family limit interpretation; no B3 interval or rectangle bootstrap is used.

Secondary checks on endpoints and B3. The conclusion also depends on its measured target. On source-held-out B3 novelty selects contrasts with much smaller measured squared effects, its lower raw MSE tracking its lower zero-effect risk (rows of `utility_curves`), and constant B retains every interaction at fractional mass, so it cannot identify poor recovery. These are 140 dependent contrasts on one background, not 140 independent validations.

On APP endpoints the probe-minus-disagreement excess-risk difference is +0.006818, conditional 95% interval [-0.008608, +0.023869]. Within the 17 overlapping APP assay pools, the probe score has lower excess risk in 7 and higher in 10; on the 6 B2 assay contexts it is lower in 5 and higher in 1. These descriptive counts preserve context without creating independent biological replications; they do not overturn the primary null, and no candidate-selection improvement follows.

Results and limits. The primary interval includes zero. B2 support absence and novelty are constant and R1 probe sensitivity is zero even on the poorly predicted source-held-out B3 contrasts; the historical ten-seed B3 result is preserved separately from this three-seed ensemble. All

predictions, controls, seed contrasts, risks, denominators, paired intervals and within-assay outcomes are in `utility_observations`, `utility_curves`, `utility_comparisons` and `utility_within_assay`. The primary null does not prove every diagnostic useless; it rejects a claimed advantage unsupported by this test.

Coverage	Pair mass	Probe MSE	Its zero MSE	SD MSE	ΔE	95% interval
20%	31.2	0.148697	0.150237	0.408344	-0.002887	[-0.004963,-0.000467]
40%	62.4	0.273094	0.273292	0.425092	-0.001938	[-0.003812,-0.000147]
60%	93.6	0.356843	0.356330	0.363586	-0.000246	[-0.001059,+0.000561]
80%	124.8	0.336510	0.336282	0.322389	-0.000208	[-0.000668,+0.000103]
100%	156.0	0.434741	0.434166	0.434741	+0.000000	[+0.000000,+0.000000]

Table 23: R0 B2 across all frozen coverage levels. Delta E is probe minus disagreement in excess-over-identical-subset-zero risk; 60 percent is primary. Equal-component weights are renormalized after fractional retention. Different methods retain different zero risks.

A10.21 Complete published predictor: MAVE-NN measured effects

Native reference comparison and provenance. The intended split holds out sequences within one assay and gene context; it is not a new-gene-context test or independent biological validation. The native evaluation gave $I_{\text{var}} = 0.305865$ and $I_{\text{pred}} = 0.342436$, versus tutorial values 0.306635 and 0.357971, respectively. The differences are -0.000770 and -0.015535. The tutorial reports 0.357971 ± 0.013754 for I_{pred},[1] so the present estimate is not inside that displayed interval. This observation alone does not establish a statistically significant discrepancy. No combined-uncertainty criterion or prediction-level numerical tolerance is established here; this is a reference comparison, not bitwise reproduction. Identity enumeration preceded scoring: 2083 rectangles whose endpoints all belong to the released test split over 28 position pairs and 1853 backgrounds, and 12486 single substitutions, with position 4 fixed G and position 5 restricted to C and U, so no globally unmeasured substitution is scored.

Implementation, data and checkpoint are pinned: mavenn 1.1.3, TensorFlow 2.21.0, Keras 3.10.0 (newer Keras dereferences an argument specification that mavenn's error-handling decorator hides) and Python 3.13.12, outside the repository. The released weights are `mpsa_ge_pairwise.weights.h5` (SHA256 `90e955f8...`) with metadata `.pickle` (`d41a069a...`); the data are `mpsa_data.csv.gz` (`df125142...`, 30,483 rows) and `mpsa_replicate_data.csv.gz` (`c18a8247...`, 30,697 rows). The first hash equals the independently pinned GitHub copy used for the input-support calculation. Settings are the released ones: pairwise G-P map, GE regression, SkewedT noise, heteroskedasticity order two, 50 hidden nodes, θ and η regularisation 0.1. The checkpoint metadata record $N = 24{,}405$ and aggregate training statistics, but not the full sequence memberships. The author training notebook removes released test rows before setting data and saves the corresponding model name. This supports the intended split; count equality alone does not establish the exact source–data–checkpoint provenance or exclude test use from every model-selection step. The released response values use a consensus-sequence assay normalizer before the random split. This shared assay-scale normalization is distinct from model-fitted preprocessing. Its common additive offset on the log scale cancels from within-library contrasts whose coefficients sum to zero, but remains relevant to endpoint-scale interpretation.

Predictions use the complete native observable map `x_to_yhat`; the latent ϕ is never substituted for a measurement. An additive latent map can still produce nonadditive observed contrasts through the nonlinear measurement map, so observed nonadditivity is not evidence of a coding error. No calibration transform is fitted on test or replicate labels.

A rectangle is eligible only when all four sequences that differ at exactly two declared positions are released test rows; a single substitution needs both of its sequences there. Orientation is lexicographic in both position indices and both state pairs, so identities are label-independent, and

[1] https://mavenn.readthedocs.io/en/v1.1.3/tutorials/3_splicing_mpra_multiple_gpmaps.html

duplicates are removed. Weights are three nested equal levels: each of the 28 eligible position pairs carries equal total weight, each of its backgrounds equal weight within the pair, and each eligible rectangle equal weight within its background; the weights sum to one. Single substitutions use the analogous rule over positions, backgrounds and substitution pairs. Enumeration precedes scoring, and 2083 of 2083 enumerated rectangles are supported, with zero unsupported and zero unresolved rectangle identities, because eligibility is decided by membership of the released test split alone.

Encoding rank and floor for this pipeline. The diagnostic of Appendix A10.6 was applied to the splicing encoder on its own terms. The complete consumed input is `x_ohe` as produced by MAVE-NN's `x_to_stats`, float32, 36 coordinates per sequence, and classes are formed by byte equality of that tensor, the same rule used for the siRNA encoder. All 3,696 endpoint states of the primary rectangle population remain distinct, so H restricts to the identity on them and $\mathrm{rank}(CH) = \mathrm{rank}(C) = 1{,}986$ against $q = 2{,}083$ contrasts. The 97 forced-zero directions are therefore dependencies of the rectangle enumeration rather than of the encoding. Because $\mathrm{col}(CH) \subseteq \mathrm{col}(C)$ and the two ranks are equal, the column spaces coincide; the recorded contrast vector is C applied to the recorded endpoint means and so lies in $\mathrm{col}(CH)$, making the floor exactly zero. The computed value is 6.2×10^{-30}, at float rounding. Library 2 gives 3,521 states, $\mathrm{rank}(CH) = \mathrm{rank}(C) = 1{,}856$ against $q = 1{,}947$, 91 design dependencies and floor zero on the same argument. Ranks were obtained by exact sparse rational elimination and cross-checked against GF(p) elimination at two primes and against the SVD rank. This is the second instance of the diagnostic and the one in which it does not fire; per-pipeline rows, certificates and encoding descriptions are in `rank_floor_generalization`.

The additive comparator checkpoint was not available in the installed example-model directory used for this evaluation. Historical additive assets exist elsewhere in the author repository, but a compatible evaluation of those assets was not established here. The additive model was not refitted; the authors' tutorial values ($I_{\mathrm{var}} = 0.216834$, $I_{\mathrm{pred}} = 0.219266$) are external context, not measurements from our evaluation.

The replicate library is matched by exact sequence only, never by its own set column: 5884 of 6078 test sequences recur, 1947 of 2083 rectangles and 11937 of 12486 single substitutions survive, and the losses are reported rather than back-filled. It is measured assay replication of the same library design, not independent biological replication of each sequence or rectangle. No replicate-agreement threshold was chosen after seeing model errors and replicate agreement is not treated as an exact noise ceiling. Full identities, weights and per-contrast values are in `mavenn_fidelity`, `mavenn_coverage`, `mavenn_endpoints`, `mavenn_contrasts`, `mavenn_replicate` and `mavenn_comparator`.

A10.22 Encoding breadth audit: further measured panels

A separate audit applied the rank and floor calculation of Appendix A10.6 beyond the primary panel. Its candidate routes, eligibility rules and size caps were frozen before any new rank, floor or prediction error was computed (protocol SHA-256 `8a3c05d2...`), and every selected case is reported, including zeros. It constructed no model or optimizer and ran no checkpoint inference. Classes use the byte-equality rule of Appendix A8.4: for the GNN, all four forward arguments X, A, mask, C after the historical training-support preprocessing, with no audit-chosen threshold; for MAVE-NN, the released one-hot input. The restriction added by an encoding is counted as $d_{\mathrm{enc}} = \mathrm{rank}(C) - \mathrm{rank}(CH)$, because $q - \mathrm{rank}(CH)$ also counts dependencies already present in the contrast design. Ranks are exact, by rational elimination or, where C has full row rank, a signed identity minor. Floors use equal weights $1/q$ on each panel's recorded response scale. Every value in Table 24 was recomputed independently from the saved C, H and measured contrasts.

Bramsen screen, expanded. The expanded panel uses the primary spreadsheet of Bramsen et al. (2009), the same assay context (HeLa eGFP, 10 nM, 72 h), response (1−relative eGFP, unclipped) and W053/W207 anchor as B3. Of 2,208 parsed rows, 2,158 are unique finite measured pairs once empty rows and duplicated identities are removed. Excluding 181 unresolved or missing strand identities and 45 pairs that fail the original pairing-admission rule leaves 1,932 endpoints, and 1,845 anchored rectangles have all four corners measured (221 candidates lack one). All 165 original states are retained, and their raw tensors, labels and 90-class partition reproduce. The 1,767 added states

Panel	m	k	q	$\operatorname{rank} C$	$\operatorname{rank} CH$	d_{enc}	Floor	Energy	Share
B3 (reproduced)	165	90	140	140	72	68	0.009980	0.068205	14.63%
Bramsen, expanded	1932	1183	1845	1845	1115	730	0.007565	0.055180	13.71%
new contrasts only	1792	1111	1705	1705	1043	662	0.007367	0.054111	13.61%
MAVE-NN splicing	3696	3696	2083	1986	1986	0	0	0.412839	0%
GB1 binding	17689	17689	15856	15856	15856	0	0	0.549741	0%

Table 24: Encoding rank and empirical floor on further measured panels, equal weights $1/q$. d_{enc} counts restrictions added by the encoding; $q - \operatorname{rank} C$ counts design dependencies (97 for splicing, none elsewhere). Energy is $q^{-1}\|\widetilde{I}\|^2$ and Share is the floor as a percentage of it, a different denominator from the fitted-error shares of Table 1. Zero floors are exact: H is injective, so $\operatorname{rank}(CH) = \operatorname{rank}(C)$ and the recorded contrasts lie in $\operatorname{col}(CH)$. The new-contrasts row is a follow-up check run after the frozen audit. Response scales differ between panels, so floors are compared within a panel, not across panels.

change strand sequence or length, so only the original 165 satisfy the chemistry-only restriction of B3: the expanded contrasts measure anchored strand replacement, a broader estimand. No row of CH is zero, so the 730 restrictions are joint, as in B3. After the frozen audit, a follow-up check removed the 140 original contrasts, each matched exactly once by its signed endpoint identities; the remaining 1,705 contrasts alone lose 662 directions with floor 0.007367, so the positive floor is not carried by the original panel. All these contrasts share the assay, controls and reference state of B3. They are within-study coverage, not an independent encoding or biological replication, and because the populations differ, the lower expanded floor is not an encoding improvement.

GB1 under the same encoder. The released protein-G binding assay of Olson et al. (2014), distributed with MAVE-NN 1.1.3 on its log2 enrichment scale, is encoded by the same one-hot rule with a 20-letter alphabet over 55 residues. A frozen size cap selects the first 50 of 1,485 position-pair blocks in lexicographic order, residue 2 against residues 3–52, so coverage is deliberately limited rather than random. Within each block, rectangles are anchored at the first measured amino-acid pair and need four measured double mutants; no wild-type or single-mutant corner is imputed. Of 17,907 rows in the selected blocks, 17,689 form 15,856 complete rectangles (218 candidates lack a corner). H is injective, so the floor is exactly zero. The selected rows mix the released training, validation and test splits and no saved predictions exist for them, so this is a descriptive encoding diagnostic, not held-out prediction. It adds an assay under an existing encoder, not an independent encoder. The splicing panel of Appendix A10.21 reproduces with floor zero under both equal and nested weights.

Blocked routes and scope. ENsiRNA-mod and MEG-mod were rechecked at unchanged commits and remain blocked (Appendix A10.18): ENsiRNA-mod builds its graph from per-duplex Rosetta coordinates that are not available for these states, and the forward pass of MEG-mod uses per-example modification tokens that its release never defines. No geometry, token meaning or partial branch was substituted, so no independent complete modified-siRNA encoding passed the eligibility gate. The audit therefore supports a reproducible diagnostic and a same-study extension, not prevalence across independent encoders. It fits nothing, so it cannot say when restoring representability improves accuracy; a controlled design crossing encoding restriction with training coverage, at matched training sizes and with a correctly specified interaction model, would address that question.

A10.23 Reference-independent B3 decomposition

The verified 165-state panel is a complete 15 by 11 matrix with antisense rows and sense columns, including the shared reference state. For measured or predicted Y set $\mu = \overline{Y}$, $a_i = \overline{Y_{i\cdot}} - \mu$, $b_j = \overline{Y_{\cdot j}} - \mu$ and $R_{ij} = Y_{ij} - \mu - a_i - b_j$, with equal weights over all 165 cells. Reconstruction is exact: $\sum_i a_i = \sum_j b_j = 0$, every row and column sum of R vanishes, the three centred components are mutually orthogonal and total centred energy splits exactly, the largest violation over all nine checks, a residual column sum, being 2.7e-15 against a 10^{-12} tolerance. The split is linear, so prediction error decomposes exactly as $\text{MSE} = (\Delta\mu)^2 + \overline{(\Delta a_i)^2} + \overline{(\Delta b_j)^2} + \overline{(\Delta R_{ij})^2}$, and the realised identity gap is at most 10^{-17}.

Only saved full-precision endpoint predictions are used; no B3 fit or inference was performed. The historical ten-seed ensemble and its seeds are analysed separately from the three-seed utility ensemble, and seed spread is optimisation variability, not biological replication: the GNN's nonadditive error component has mean 0.018185 and standard deviation 0.000067 across ten seeds. Two columns of the historical export, the `original` arms, reproduce the `corrected` arms bitwise over all 165 states and ten seeds although their checkpoints differ; this unresolved anomaly is retained in `b3decomp_export_anomalies` and carries no reported result, since every published B3 number uses the four audited families. Absolute amplitudes are kept beside fractions, because a large fraction of a tiny spread is not accurate recovery: predicted nonadditive energies are 1.6e-06 to 1.5e-04 against a measured 0.018138. Reference-based contrasts, their weighting and the 140-rectangle results are unchanged, and no projection fitted to evaluation labels is reported as a predictor.

A10.24 Training-support intervention: design and accounting

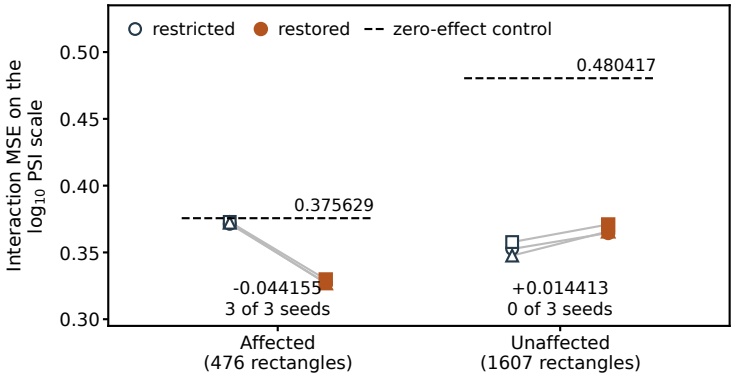

Figure 8: Fixed-size training-support comparison, three paired seeds individually. Restoration lowers interaction MSE on the affected rectangles and not the unaffected ones; the zero-effect control is identical across arms within a subset and differs between them, so it is drawn per subset. This is one controlled training-data comparison on one feature (position 9, base G, smallest of 28 four-state training counts), only 2000 training rows are replaced and the shared validation set carries none of it. It does not show that missing support always causes error, nor that support alone makes an explanation valid.

The feature is selected from training input counts: among positions admitting all four nucleotides, choose the smallest trainval count, breaking ties by position index and then nucleotide. That selects position 9, G, with 5286 trainval occurrences, which varies in the library and has measured restoration examples, so no globally unmeasured substitution is used. Both arms train on $14{,}238$ observations and share a core of $12{,}238$. The restricted arm adds $2{,}000$ feature-free observations to that core; the restored arm instead adds the first $2{,}000$ feature-carrying trainval sequences in lexicographic order. The removed feature-free observations are the last $2{,}000$ restricted-arm training rows in that order. Selection uses identities only, not response labels. Backgrounds are not individually matched, so other aspects of the training distribution can differ along with feature support. A shared validation set of 4881 feature-free rows keeps early stopping blind to the withheld feature, test identities are identical and excluded from every selection step, and representation and vocabulary are unchanged, so observations are removed, not the feature channel. Affected rectangles are those with the feature in at least one corner, fixed by identity rather than observed effect or error size, and the unaffected set is a specificity check. Three paired seeds per arm supplied six retained fits, with $62{,}424$ reported optimizer updates, no recorded training failure in that retained invocation and no GPU use. An earlier invocation completed six fits but failed while serializing an empirical-gauge coefficient undefined for a zero-support feature, before its per-seed results were persisted. The recovery therefore consumed twelve attempts against a six-attempt authorization. The same feature, design, identities, weights and seeds were reused; this was a repeated execution, not a second specification. The $62{,}424$-update count covers the retained invocation; a complete two-invocation update total is not established here. The attempt ledger and retained per-seed results are recorded separately in `support_attempt_ledger` and `support_seed_results`; unpersisted first-invocation

outcomes are not represented as available per-seed evidence. The zero-effect control is identical across arms on identical identities, so an improvement in training loss alone would not count.