AGENTIC ORCHESTRATION OF HETEROGENEOUS PREDICTORS FOR RNA BINDING POCKET DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein surfaces that engage RNA are targeted today by four disjoint families of tools: end to end complex structure prediction, surface pocket detection, binding residue classification, and docking. Each encodes a different inductive bias and is validated under its own protocol, so a practitioner facing a new target has no principled basis for choosing among them. On a structurally disjoint benchmark curated from 1,809 experimental complexes, no single tool is universally best, per target performance spreads more widely *within* a family than *between* families, and the winner is partly predictable from target properties. Per target tool choice, rather than another monolithic model, is therefore the operative design variable. RiboSeer acts on this by separating planning from arithmetic: a language model profiles each pair and selects from a 15 tool library subject to a mandatory core covering all four paradigms, a learned gradient boosted fusion combines per residue scores, cross tool agreement, neighborhood context, and the agent profile into a calibrated binding probability, and an exponential moving average over per category reliability closes the loop across samples. On 107 held out targets separated from training by TM score, RiboSeer reaches Pearson r of 0.593 against 0.439 for the best single tool and 0.515 for a weighted mean over the same tools, with AUROC 0.899. The margin holds across four splits, four contact cutoffs, and strata by RNA length, domain family, and difficulty. Ablations attribute the gain: learned fusion over a paradigm diverse pool supplies most of it, while the language model contributes $+0.039$ Pearson r over a bandit only selector.

1 INTRODUCTION

RNA binding proteins govern nearly every stage of the RNA life cycle, including splicing, editing, subcellular localisation, stability control, and translational regulation, and large scale mapping efforts have catalogued hundreds of such proteins across human and mouse genomes (Moore et al., 2020). Their dysregulation is implicated in cancers, neurodegenerative disease, and viral infection. Which residues on a protein surface contact RNA is the structural prerequisite for interpreting these functions and for prioritising candidates for validation, yet experimental determination remains slow relative to the growth of sequence databases (Berman et al., 2000), making computational identification the practical entry point for most new targets.

Computational support for this task never converged on a single formulation. Four families now coexist, each encoding a different inductive bias. End to end complex structure predictors fold protein and RNA jointly and read the interface off predicted coordinates (Abramson et al., 2024; Passaro et al., 2025; team et al., 2024; Baek et al., 2024), which is powerful when the complex is foldable and uninformative when it is not. Geometric pocket detectors ignore the RNA and score surface cavities by shape and physicochemistry (Krivák & Hoksza, 2018; Le Guilloux et al., 2009; Aggarwal et al., 2022), transferring across targets but unable to distinguish an RNA site from any other. Binding residue classifiers treat the task as per residue labelling from sequence and structure (Roche et al., 2024; Xia et al., 2021; Lam et al., 2019), tolerating novel folds while discarding global geometry. Docking samples placements of an RNA against a protein and scores the poses (Yan et al., 2017; Giulini et al., 2025), capturing extended interfaces but inheriting its input conformer quality. Developed by separate communities on differently curated sets and reported under incompatible protocols, the four offer no common ground to the practitioner choosing among them.

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.

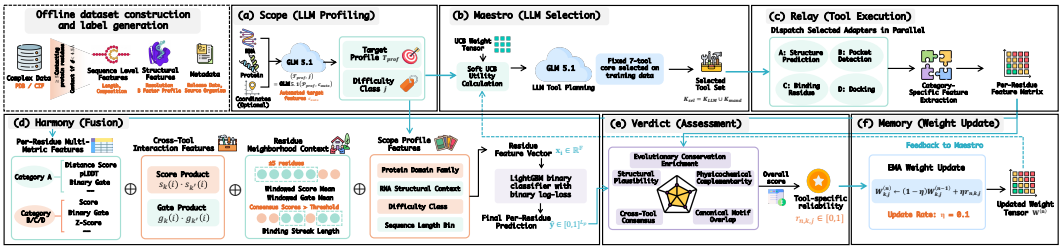


Figure 1: **RiboSeer.** Offline curation builds contact labels at $d_{\text{cut}} = 4.5$ Å. **(a)** SCOPE profiles each target into $(\mathcal{T}_{\text{prof}}, j)$. **(b)** MAESTRO combines a soft UCB utility over the weight tensor with language model planning, then unions the proposal with a fixed seven tool core. **(c)** RELAY dispatches adapters in parallel across the four paradigms and extracts a residue feature matrix. **(d)** HARMONY concatenates four feature groups and fuses them with a LightGBM classifier under log loss. **(e)** VERDICT scores five components. **(f)** MEMORY updates category reliability by EMA and feeds it back to MAESTRO. Shaded blocks are language model driven; all others are deterministic.

The consequence becomes visible only when the families are evaluated together under one protocol, which is what we do here. The best single tool on our benchmark reaches Pearson r of 0.439, but the spread *within* the complex prediction family runs from 0.439 down to -0.087, wider than the spread between families. A parameter free mean over all 15 tools reaches only 0.446, barely above the single best, because averaging admits confidently wrong predictors at full weight alongside the correct one. Two obstacles compound this: tool outputs are not commensurable, since a pLDDT weighted distance, a cavity score, and a docking interface distance occupy different scales with different failure modes, and per target selection is a decision problem with no supervision at inference time, because one cannot know which tool is right until the answer is already known.

We address this by separating planning from arithmetic. RiboSeer is an agentic system in which a language model (Zeng et al., 2026) performs the parts of the problem that are textual and contextual, namely profiling a target by protein domain architecture and RNA structural context and selecting which tools to invoke, while every numerical combination is performed by a learned model that is trained, unit tested, and auditable. This division distinguishes RiboSeer from tool augmented agents that let the model both call and combine (Schick et al., 2023; Qin et al., 2024). A guardrail constrains the agent: whatever subset it proposes is unioned with a mandatory core of seven tools identified by forward greedy search on the training split, guaranteeing the fusion always observes all four paradigms and preventing collapse onto a single family. The fusion consumes per residue scores together with cross tool agreement, sequence neighborhood context, and the agent profile, and emits a calibrated per residue probability clustered into ranked pockets. An exponential moving average over per category tool reliability feeds back into selection across samples, so the system accumulates experience without retraining.

Our contributions are fourfold. First, we construct a structurally disjoint benchmark of 332 cluster representatives drawn from 1,809 experimental complexes, partitioned by TM score (Zhang et al., 2022) so that no test target is homologous to a training target, and evaluate 15 open source predictors spanning all four paradigms under one protocol with one label definition. Second, we quantify the per target variability that motivates orchestration. Third, we present RiboSeer, reaching Pearson r of 0.593, AUROC 0.899, and top k F1 of 0.586 on 107 held out targets, improving on the best single tool by $+0.154$ and on a weighted ensemble over the same tools by $+0.078$. Fourth, we report an ablation that attributes the gain rather than merely demonstrating it: learned fusion supplies most of the improvement, the language model contributes $+0.039$, and four of fifteen tools prove individually dispensable, which we report rather than prune.

2 RELATED WORK

Complex structure prediction. Following the accuracy breakthrough on monomeric proteins (Jumper et al., 2021; Lin et al., 2023), joint folding of protein and nucleic acid has advanced rapidly, with AlphaFold 3 (Abramson et al., 2024), Boltz-2 (Passaro et al., 2025), Chai-1 (team et al., 2024), RoseTTAFold2NA (Baek et al., 2024), and RoseTTAFold All Atom (Krishna et al., 2024)

producing full atomic complexes from sequence, from which an interface can be read directly; RNA only predictors such as RhoFold+ (Shen et al., 2024) supply the partner conformers these systems consume. Their reliability is strongly conditioned on whether a confident complex is attainable, and our results show a spread from 0.439 to -0.087 Pearson r within this single family.

Surface pocket detection. Fpocket (Le Guilloux et al., 2009), P2Rank (Krivák & Hoksza, 2018), DeepPocket (Aggarwal et al., 2022), and equivariant successors such as EquiPocket (Zhang et al., 2023) score cavities from geometry and physicochemistry without reference to any partner molecule. This makes them robust to novel folds and cheap to run, but unable to distinguish an RNA site from a small molecule site. They contribute a surface accessibility prior the other families do not encode.

Binding residue classification. EquiPNAS (Roche et al., 2024), GraphBind (Xia et al., 2021), NucleicNet (Lam et al., 2019), RNABindRPlus (Walia et al., 2014), and BindUP (Paz et al., 2016) label residues directly from sequence and structure features. They remain informative when complex folding fails, which is precisely where Category A degrades, making the two families complementary rather than competing.

Protein RNA docking. HDOCK (Yan et al., 2017) and HADDOCK 3 (Giulini et al., 2025) sample and score placements of an RNA against a protein, optionally guided by templates or restraints. Docking contributes signal on extended interfaces that per residue classifiers tend to fragment, and in our leave one out analysis HDOCK is the second most load bearing member of the library.

Ensembles and tool using agents. Combining heterogeneous predictors spans fixed weight averaging, rank fusion, stacking, and learning to rank, all of which fix a single global weighting policy at training time; a weighted mean over our 15 tools reaches 0.515 against RiboSeer's 0.593, so the residual gain comes from per target conditioning and from features no linear combination of raw scores can express. Language model agents that invoke external tools (Schick et al., 2023; Qin et al., 2024) are closer in spirit, but RiboSeer restricts what the agent may do: it selects and profiles, never performs arithmetic, and cannot override the mandatory core, with exploration governed by a standard bandit utility (Lattimore & Szepesvári, 2019). The agent accounts for $+0.039$ of the $+0.154$ total, which we regard as an honest accounting of where value in an agentic scientific pipeline sits.

3 METHOD

3.1 PROBLEM FORMULATION

Let $\mathbf{r} = (r_1, \ldots, r_{L_r})$ denote an RNA sequence over $\mathcal{A}_r = \{\mathrm{A}, \mathrm{U}, \mathrm{G}, \mathrm{C}\}$ and $\mathbf{p} = (p_1, \ldots, p_{L_p})$ a protein sequence over the 20 letter amino acid alphabet, optionally accompanied by resolved coordinates $\boldsymbol{X}_r \in \mathbb{R}^{L_r \times 3}$ and $\boldsymbol{X}_p \in \mathbb{R}^{L_p \times 3}$. We formulate binding pocket prediction as learning a per residue mapping

$$\mathcal{F} : (\mathbf{r}, \mathbf{p}, \boldsymbol{X}_r, \boldsymbol{X}_p) \longrightarrow \mathbf{y} \in [0,1]^{L_p}, \tag{1}$$

where y_i is the probability that protein residue i contacts the RNA. Ground truth $\mathbf{y}^\star \in \{0,1\}^{L_p}$ is defined by heavy atom distance,

$$y_i^\star = \mathbf{1}\Big[\min_{a \in \mathrm{atoms}(p_i),\, b \in \mathrm{atoms}(\mathbf{r})} \|\boldsymbol{x}_a - \boldsymbol{x}_b\|_2 < d_{\mathrm{cut}}\Big], \qquad d_{\mathrm{cut}} = 4.5\,\text{Å}, \tag{2}$$

yielding a positive rate near 5%. Section 4.5 verifies that all conclusions hold for $d_{\mathrm{cut}} \in [3.5, 6.0]$ Å.

Given a library of $K = 15$ heterogeneous predictors $\{f_1, \ldots, f_K\}$ with tool specific raw outputs $f_k(\mathbf{r}, \mathbf{p}; \theta_k)$, the orchestration problem is to choose a subset $\mathcal{K}_{\mathrm{sel}} \subseteq \{1, \ldots, K\}$ and an aggregation g such that

$$\hat{\mathbf{y}} = g\big(\{P_k\}_{k \in \mathcal{K}_{\mathrm{sel}}},\, \mathbf{c}_{\mathrm{ctx}}\big), \tag{3}$$

where P_k is the standardised output of tool k and $\mathbf{c}_{\mathrm{ctx}}$ encodes target context. RiboSeer instantiates the selection of $\mathcal{K}_{\mathrm{sel}}$ with a language model and the aggregation g with a learned model, keeping the two strictly separate.

Table 1: The MAESTRO tool library. Category A predicts the full complex and yields per residue scores via Eq. equation 6; B detects surface pockets; C performs sequence and structure based binding site prediction; D performs docking. CORE marks membership in the mandatory seven tool set selected by forward greedy search on the training split.

Tool	Cat.	Core	Output signal	Timeout (s)
Boltz-2 (Passaro et al., 2025)	A	✓	Distance score, pLDDT, ipTM	900
Chai-1 (team et al., 2024)	A	✓	Distance score, pLDDT, ipTM	900
RoseTTAFold2NA (Baek et al., 2024)	A		Distance score, pLDDT	600
RoseTTAFold All Atom (Krishna et al., 2024)	A		Distance score, pLDDT	1200
AlphaFold 3 (Abramson et al., 2024)	A		Distance score, pLDDT, ipTM	900
P2Rank (Kriv´ak & Hoksza, 2018)	B		Per residue pocket score	60
Fpocket (Le Guilloux et al., 2009)	B		Per residue pocket score	60
DeepPocket (Aggarwal et al., 2022)	B	✓	Per residue pocket score	300
EquiPNAS (Roche et al., 2024)	C	✓	Binding probability	300
NucleicNet (Lam et al., 2019)	C	✓	Binding probability	600
GraphBind (Xia et al., 2021)	C		Binding probability	600
RNABindRPlus (Walia et al., 2014)	C	✓	Binding probability	300
BindUP (Paz et al., 2016)	C		Binding probability	300
HDOCK (Yan et al., 2017)	D	✓	Interface distance score	600
HADDOCK 3 (Giulini et al., 2025)	D		Interface distance score	600

3.2 ARCHITECTURE OVERVIEW

RiboSeer is a seven stage pipeline (Figure 1). Two stages are language model driven and five are deterministic:

$$\underbrace{\text{GENESIS}}_{\text{curation}} \rightarrow \underbrace{\text{SCOPE}}_{\text{profiling}} \rightarrow \underbrace{\text{MAESTRO}}_{\text{selection}} \rightarrow \underbrace{\text{RELAY}}_{\text{execution}} \rightarrow \underbrace{\text{HARMONY}}_{\text{fusion}} \rightarrow \underbrace{\text{VERDICT}}_{\text{assessment}} \rightarrow \underbrace{\text{MEMORY}}_{\text{update}} \tag{4}$$

Every language model decision is validated against a Pydantic schema before reaching downstream code; failures trigger one retry, then fall back to bandit only selection, which Section 4.4 measures as a -0.039 Pearson r degradation rather than a system failure. Appendix A.13 records why an earlier prototype that placed the fusion arithmetic inside the model was abandoned.

3.3 TOOL LIBRARY

The library spans four paradigms (Table 1). Each tool is wrapped in an adapter $\mathcal{A}_k$ converting native output into a standardised triple

$$P_k = \mathcal{A}_k\big(f_k(\mathbf{r}, \mathbf{p}; \theta_k)\big) = (\mathbf{s}_k, \mathbf{g}_k, \mathbf{c}_k), \qquad \mathbf{s}_k \in [0,1]^{L_p},\ \mathbf{g}_k \in \{0,1\}^{L_p},\ \mathbf{c}_k \in \mathbb{R}^{m_k}, \tag{5}$$

where $\mathbf{s}_k$ is a per residue score, $\mathbf{g}_k$ the tool's discrete prediction, and $\mathbf{c}_k$ global confidence signals such as pLDDT and ipTM. For Categories A and D, the score is derived from the predicted geometry by

$$s_k(i) = \exp\Big(-\tfrac{1}{\sigma} \min_{b \in \text{atoms}(\hat{\mathbf{r}})} \|\hat{\boldsymbol{x}}_{a(i)} - \boldsymbol{x}_b\|_2\Big), \tag{6}$$

with $a(i)$ the representative heavy atom of residue i in the predicted complex. Adapters absorb all format heterogeneity, so adding a tool requires one adapter and no change to the fusion.

3.4 GENESIS: CURATION AND SCOPE: PROFILING

GENESIS enumerates all protein RNA chain pairs per entry, computes the contact set of Eq. equation 2, retains pairs with at least five contact residues, and extracts sequence, structural, and metadata features into $\mathbf{c}_{\text{auto}}$. Retained pairs are doubly clustered by MMseqs2 (Steinegger & Söding, 2017) at 30% protein and 80% RNA identity.

SCOPE then serialises $\mathbf{c}_{\text{auto}}$ into a structured prompt and queries the language model,

$$(\mathcal{T}_{\text{prof}}, j) = \text{LLM}\big(\mathcal{P}_{\text{prof}},\, \mathbf{c}_{\text{auto}}\big), \qquad j \in \{1, \ldots, J\}, \tag{7}$$

where $\mathcal{T}_{\text{prof}}$ is a difficulty analysis and j a categorical label conditioning downstream selection. Profiles span protein domain architecture, assigned by InterPro annotation (Paysan-Lafosse et al.,

2023), and RNA structural context, seeded by ViennaRNA secondary structure (Lorenz et al., 2011) and Rfam family membership (Kalvari et al., 2021).

3.5 MAESTRO: CONSTRAINED TOOL SELECTION

MAESTRO maintains a weight tensor $\boldsymbol{W} \in \mathbb{R}^{K \times J}$ where $W_{k,j}$ records tool k's reliability on category j. It presents a soft upper confidence bound utility (Lattimore & Szepesvári, 2019)

$$u_{k,j} = \hat{\mu}_{k,j} + \beta \sqrt{\frac{\ln n}{N_{k,j}}}, \tag{8}$$

with $\hat{\mu}_{k,j}$ the running quality mean, $N_{k,j}$ the selection count, and n the total round count, and obtains a proposed plan

$$(\mathcal{K}_{\mathrm{LLM}}, \mathcal{T}_{\mathrm{plan}}) = \mathrm{LLM}\big(\mathcal{P}_{\mathrm{sel}},\, \mathcal{T}_{\mathrm{prof}},\, j,\, \boldsymbol{W}\big). \tag{9}$$

A guardrail then fixes the executed set to

$$\mathcal{K}_{\mathrm{sel}} = \mathcal{K}_{\mathrm{LLM}} \cup \mathcal{K}_{\mathrm{mand}}, \tag{10}$$

where $\mathcal{K}_{\mathrm{mand}}$ is the seven tool core of Table 1. The union guarantees $\mathcal{K}_{\mathrm{sel}}$ intersects all four paradigms for every target. Without it, free selection converges to a mean of 5.7 tools concentrated in Categories A and C, and Pearson r falls to 0.569 (Appendix A.5).

3.6 RELAY: CONCURRENT EXECUTION

RELAY dispatches the selected adapters concurrently under per tool timeouts, applying Eq. equation 6 for Categories A and D and reading Categories B and C directly. A tool that times out or fails to converge contributes a zero filled column with a missingness indicator, which the gradient boosted fusion handles natively through missing value splits, so a tool failure degrades rather than aborts the prediction.

3.7 HARMONY: LEARNED FUSION

HARMONY replaces our prototype noisy OR with a learned model over four feature groups.

Group 1, per residue multi metric ($\mathbf{x}^{\mathrm{PR}}$). For Category A tool k we take $\big(s_k(i),\ \ell_k(i),\ g_k(i),\ \mathrm{rank}_k(i)/L_p,\ \bar{\ell}_k,\ \mathrm{ipTM}_k\big)$ with ℓ_k the per residue pLDDT. For Categories B, C, D we take $\big(s_k(i),\ g_k(i),\ \mathrm{rank}_k(i)/L_p,\ z_k(i)\big)$ with

$$z_k(i) = \frac{s_k(i) - \bar{s}_k}{\mathrm{std}(s_k) + \epsilon}. \tag{11}$$

Group 2, cross tool interaction ($\mathbf{x}^{\mathrm{XT}}$). For every unordered pair (k, k'),

$$x^{\mathrm{XT}}_{k,k'}(i) = \big(s_k(i)\, s_{k'}(i),\ g_k(i)\, g_{k'}(i)\big), \qquad v(i) = \sum_{k \in \mathcal{K}_{\mathrm{sel}}} g_k(i), \tag{12}$$

where $v(i)$ is the vote count. Score products amplify agreement and suppress isolated confidence; gate products encode that residues flagged by several tools are disproportionately likely to be true contacts.

Group 3, neighborhood context ($\mathbf{x}^{\mathrm{NB}}$). Interfaces form contiguous segments, so with window $w = 5$,

$$\bar{s}^{(w)}_k(i) = \frac{1}{2w+1} \sum_{|d| \le w} s_k(i+d), \qquad \ell_{\mathrm{streak}}(i) = \max\big\{L : \forall\, i' \in [i-L, i+L],\ \bar{v}(i') > \tau_v\big\}, \tag{13}$$

with $\bar{v}$ the normalised vote count and τ_v a consensus threshold. This demotes isolated spikes and reinforces coherent segments.

Group 4, profile features ($\mathbf{x}^{\mathrm{SC}}$). The SCOPE output $(\mathcal{T}_{\mathrm{prof}}, j)$ is one hot encoded into a 16 dimensional vector over domain family, RNA context, difficulty, and length bin, broadcast to every residue.

Fusion. Concatenating the four groups gives

$$\mathbf{x}_i = \left[\mathbf{x}_i^{\mathrm{PR}};\ \mathbf{x}_i^{\mathrm{XT}};\ \mathbf{x}_i^{\mathrm{NB}};\ \mathbf{x}^{\mathrm{SC}}\right] \in \mathbb{R}^F, \tag{14}$$

and we train a LightGBM classifier (Ke et al., 2017) g_ϕ under binary log loss,

$$\hat{y}_i = \sigma\big(g_\phi(\mathbf{x}_i)\big), \qquad \phi^\star = \arg\min_\phi \; -\sum_n \sum_{i=1}^{L_p^{(n)}} \Big[y_i^{\star,(n)} \log \hat{y}_i^{(n)} + \big(1-y_i^{\star,(n)}\big) \log\big(1-\hat{y}_i^{(n)}\big)\Big]. \tag{15}$$

Section 4.4 reports that Group 1 alone recovers 0.589 of the final 0.593; we retain the remaining groups because Group 4 is the only channel through which target context enters the fusion, which drives the stratified behaviour of Section 4.5.

3.8 VERDICT AND MEMORY

VERDICT produces a label free surrogate reward

$$q = \sum_{m=1}^{5} w_m q_m, \qquad \sum_m w_m = 1, \tag{16}$$

combining structural plausibility, physicochemical complementarity measured by electrostatic correlation (Jurrus et al., 2018) and the shape index of (Lawrence & Colman, 1993), evolutionary conservation enrichment (Ashkenazy et al., 2016), cross tool consensus as mean pairwise Jaccard, and canonical motif overlap (Mistry et al., 2021). The reward is a *selection* signal only and never enters Eq. equation 15, so it cannot leak evaluation information into the prediction. MEMORY then updates reliability by exponential moving average,

$$W_{k,j}^{(n)} \leftarrow (1-\eta)\, W_{k,j}^{(n-1)} + \eta\, r_{n,k,j}, \qquad \eta = 0.1,\ r_{n,k,j} \in [0,1], \tag{17}$$

feeding back into Eq. equation 8 on subsequent samples and closing the loop.

4 EXPERIMENTS

4.1 SETUP

Benchmark. We assemble 1,809 PDB and CIF entries (Berman et al., 2000) and enumerate 45,468 protein RNA chain pairs. After filters on RNA length (20 to 100 nt), protein length ($\leq$ 400 aa), contact count ($\geq$ 5 residues), resolution ($\leq$ 3.5 Å), and quality tier, 2,018 pairs remain, grouping into 332 protein clusters at 30% identity. We retain the highest resolution representative per cluster. To remove structural homology that sequence clustering misses, we compute pairwise TM scores among the 332 representatives with US align (Zhang et al., 2022), threshold at 0.35 to form an affinity graph, and spectrally partition it into 225 training and 107 test targets. The per stage cascade is in Appendix A.1.

Baselines. We evaluate all 15 open source predictors covering the four paradigms, plus three ensembles built on identical tool outputs: a parameter free mean, a weighted mean with per tool weights fit on the training split, and a stacked logistic regression. The ensembles are the informative comparison, since they isolate the value of learned fusion from the value of having many tools. The catalogue is in Appendix A.2.

Metrics. We report residue level quality, pocket geometry following the P2Rank and PDBbind conventions (Krivák & Hoksza, 2018; Wang et al., 2004), and complex quality via DockQ (Basu & Wallner, 2016) for Categories A and D. All are computed per sample then averaged; definitions are in Appendix A.3 and the per sample distribution in Appendix A.9.

Implementation. Tool execution uses 2×NVIDIA A6000 48GB; the LightGBM fusion is CPU only (depth 4, 100 trees, learning rate 0.1). The agent is GLM 5.1 (**?**) at temperature 0.1 for selection and 0.7 for exploration, with chain of thought prompting (Wei et al., 2022) on SCOPE and direct prompting on MAESTRO. Full prompt and schema details are in Appendix A.13.

Table 2: Residue level results on the 107 target structurally disjoint test split. Metrics are averaged over per sample evaluations for which the method produced a valid prediction; n varies because baselines occasionally time out or fail to converge, whereas RiboSeer covers all targets through the fallback of Section 3.6. **Bold** best, underline second best.

Method	Pearson $r \uparrow$	Spearman $\rho \uparrow$	$R^2 \uparrow$	n
Category A: complex structure prediction				
AlphaFold 3	0.246	0.234	0.130	100
Boltz-2	0.439	0.387	0.260	104
Chai-1	0.382	0.339	0.213	105
RoseTTAFold2NA	0.109	0.100	0.048	107
RoseTTAFold All Atom	−0.087	−0.095	0.031	105
Category B: surface pocket detection				
P2Rank	0.180	0.183	0.084	103
Fpocket	0.062	0.085	0.058	105
DeepPocket	0.178	0.155	0.102	105
Category C: binding residue prediction				
EquiPNAS	0.413	0.377	0.220	102
GraphBind	0.301	0.286	0.124	99
NucleicNet	0.321	0.328	0.127	102
RNABindRPlus	0.294	0.300	0.124	106
BindUP	0.250	0.252	0.094	103
Category D: docking				
HDOCK	0.376	0.339	0.232	98
HADDOCK 3	0.219	0.212	0.103	83
Category E: ensembles over the same 15 tools				
Simple mean	0.446	0.400	0.252	107
Weighted mean	<u>0.515</u>	0.440	<u>0.314</u>	107
Stacked logistic reg.	0.502	<u>0.442</u>	0.300	107
RiboSeer (ours)	**0.593**	**0.454**	**0.436**	107

4.2 Main Results

RiboSeer reaches Pearson r of 0.593 against 0.439 for the best single tool, gaining $+0.154$, and improves R^2 from 0.260 to 0.436 (Table 2). Three observations matter more. First, no paradigm dominates, and the within family spread exceeds the between family spread: the best Category A, C, and D tools sit at 0.439, 0.413, and 0.376, while Category A alone ranges from 0.439 to -0.087, so a practitioner selecting by reputation rather than by target would frequently select badly.

Second, tool count is not the mechanism. A parameter free mean over the same 15 tools reaches 0.446, only $+0.007$ above the best tool, because averaging admits wrong predictors at full weight. The weighted mean reaches 0.515 and stacked logistic regression 0.502. The $+0.078$ margin over the weighted mean is therefore attributable to feature construction and target conditioning.

Third, ranking metrics move with correlation metrics: AUROC rises to 0.899 and top k F1 to 0.586 (Appendix A.8), so the gain is not an artefact of score calibration.

4.3 Pocket Level Geometry

Table 3 asks whether the predicted pocket centre coincides spatially with the true interface. RiboSeer attains the best DCC (14.47 Å) and DCA (7.99 Å), improving on the best baseline by 0.56 and 1.34 Å, but is marginally second on residue IoU (0.088 against 0.091) and top 1 success (3.81% against 3.92%). Absolute success rates are low for every method: RNA interfaces are extended and multi segment, so a single 4 Å centre criterion is demanding. We report these because a residue level gain that does not fully transfer to geometry should be visible to the reader.

Table 3: Pocket level geometry on the 107 target test split. DCC and DCA in Å; IoU is residue level intersection over union; success rates in % at the 4 Å criterion. RiboSeer leads on the two distance measures and is second on the two overlap measures.

Method	DCC ↓	DCA ↓	IoU ↑	Top-1 ↑	Top-3 ↑
Boltz-2	17.67	9.46	**0.091**	**3.92**	**12.75**
Chai-1	17.59	10.40	0.086	3.88	5.83
RoseTTAFold2NA	18.94	13.03	0.053	1.90	2.86
P2Rank	17.71	13.61	0.028	2.91	5.83
Fpocket	19.18	13.77	0.027	0.95	2.86
EquiPNAS	15.03	9.33	0.079	0.00	9.80
GraphBind	17.84	10.60	0.060	0.00	2.88
HADDOCK 3	16.61	10.82	0.079	1.20	1.20
RiboSeer	**14.47**	**7.99**	0.088	3.81	9.52

Table 4: Consolidated ablations. **(a)** fusion model class on the identical feature space of Eq. equation 14; **(b)** feature groups, where PR, XT, NB, SC denote Groups 1 to 4; **(c)** agent modules, where – marks replacement by the deterministic fallback. Every row retrains the fusion from scratch and is evaluated on the same 107 targets. Extended versions are in Appendix A.5.

		Configuration	Pearson r ↑	Spearman ρ ↑	R^2 ↑
(a)	Fusion	Mean of raw scores	0.474	0.421	0.270
		Noisy OR (prototype)	0.461	0.421	0.254
		Random forest	0.571	0.448	0.413
		Logistic regression	0.587	**0.458**	0.418
		XGBoost	0.587	0.448	0.431
(b)	Features	PR only	0.589	0.453	0.433
		XT + NB (no PR)	0.549	0.433	0.368
		PR + XT + NB	0.589	0.451	0.432
(c)	Agent	– SCOPE, – MAESTRO	0.554	0.436	0.389
		SCOPE only	0.560	0.438	0.394
		MAESTRO only	0.589	0.451	0.432
RiboSeer (full)			**0.593**	**0.454**	**0.436**

4.4 ABLATIONS

Table 4 consolidates three key ablations; five more appear in Appendix A.5.

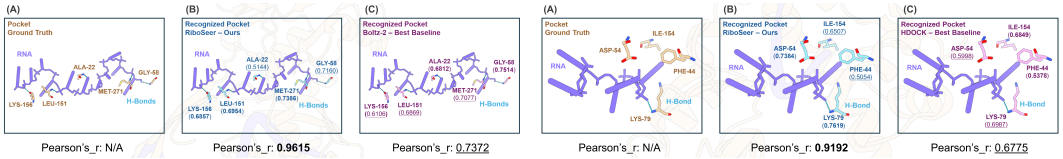
Fusion model. Replacing LightGBM with the noisy OR prototype costs 0.132 Pearson r and with a raw mean 0.119. Among alternatives the spread is small: logistic regression and XGBoost both reach 0.587, random forest 0.571. The gain therefore comes from learning fusion over the engineered space rather than from the learner, which we read as robustness rather than weakness.

Feature groups. Group 1 alone reaches 0.589; the full set reaches 0.593. We do not claim incremental groups drive residue accuracy. Their value lies elsewhere: removing Group 1 while retaining Groups 2 and 3 drops performance to 0.549, confirming those features carry real but redundant signal, and Group 4 is the only channel through which target context reaches fusion.

Agent modules. Bandit only selection gives 0.554. Adding MAESTRO alone gives 0.589; adding SCOPE alone gives 0.560. The agent contributes $+0.039$ in total, most of it through selection rather than profiling. This is a smaller share than the framing of agentic systems often implies, and we state it directly: the primary driver is learned fusion over a paradigm diverse pool, with the agent supplying useful but secondary per target adaptivity.

4.5 STRATIFIED ANALYSIS AND ROBUSTNESS

Where the gain concentrates. By RNA length (Appendix A.6), RiboSeer improves most on medium RNAs of 41 to 70 nt (0.719 against 0.517) and least on long RNAs of 71 to 100 nt (0.362



(a) PDB 3X1L, $r = 0.9615$ (baseline: Boltz-2). (b) PDB 4WKR, $r = 0.9192$ (baseline: HDOCK).

Figure 2: **Case studies on two held out targets.** In each panel, **(A)** shows the ground truth pocket with key RNA bases and the hydrogen bonds to contacting residues (LYS-79 in 4WKR), **(B)** the RiboSeer prediction, and **(C)** the strongest baseline on that target. Higher Pearson r is better (↑).

against 0.295), consistent with the neighborhood features of Eq. equation 13 helping where interfaces are contiguous while every method degrades on the longest targets. By difficulty, the gain is largest on targets SCOPE flags as hard (0.634 against 0.471), indicating the profile identifies cases where orchestration pays.

Where it does not. On the single KH domain target in the test split RiboSeer scores -0.024 while Chai-1 reaches 0.171. With $n = 1$ we draw no conclusion but report it rather than omit the row. The dsRBD, single stranded, internal loop, junction, and G quadruplex strata contain zero test targets, so the benchmark is silent on them.

Tool redundancy. Leave one out (Appendix A.5) shows that removing AlphaFold 3, Fpocket, GraphBind, or BindUP changes Pearson r by at most $+0.002$, so four of fifteen tools are individually dispensable, while removing Boltz-2 costs -0.063 and HDOCK -0.023. We retain the full library because $\mathcal{K}_{\text{mand}}$ is fixed before test evaluation and pruning by test performance would constitute selection on the evaluation set.

Protocol sensitivity. The ordering is stable across four cutoffs from 3.5 to 6.0 Å and four splits, including temporal ($+0.075$) and tight 25% identity ($+0.071$), where absolute values fall for all methods but the margin persists (Appendix A.7). Across five seeds the system varies by ± 0.001 Pearson r. Mean wall clock is 714 s per target, of which 662 s is tool execution and 52 s the two agent calls (Appendix A.10); the fusion itself runs in under 0.1 s.

4.6 CASE STUDIES

Figure 2 visualises two targets where fusion resolves tool disagreement. On 3X1L the fused prediction recovers the hydrogen bonded core of the pocket and reaches Pearson r of 0.9615, against Boltz-2 as the strongest baseline on that target. On 4WKR it recovers the LYS-79 contact and reaches 0.9192, against HDOCK. Both cases sit in the upper decile of the per sample distribution (Appendix A.9) and are chosen to illustrate the mechanism rather than typical performance; Appendix A.12 analyses the bottom decile, where all Category A tools fall below pLDDT 60 and mean pairwise tool agreement drops below Jaccard 0.2.

5 CONCLUSION AND OUTLOOK

We asked whether the right response to four incompatible families of RNA binding site predictors is a fifth model, and found that it is not: under a single protocol on a structurally disjoint benchmark, a learned fusion over a paradigm diverse pool recovers more signal than any member of that pool or any fixed weighting of it, reaching Pearson r of 0.593 and AUROC 0.899 on 107 held out targets with the margin surviving four splits, four cutoffs, and stratification by length, family, and difficulty. Two findings temper this. The language model contributes $+0.039$ of the $+0.154$ total, so orchestration is mostly a fusion problem; and four of fifteen tools can be removed without loss, so future work should select tools by marginal contribution rather than paradigm coverage. Two directions follow. The fusion consumes per residue scalars and discards the geometry Categories A and D produce, so a geometry aware variant should narrow the gap between our residue level and pocket level gains. And the same pattern applies wherever incompatible predictor families coexist under separate protocols, with protein ligand pocket detection an immediate candidate. Better uncertainty estimates could also turn these systems into practical triage tools for experimental validation.

AI USE STATEMENT

Generative AI tools were used solely for manuscript preparation, specifically language polishing and organization of text written by the authors. All AI-assisted text was reviewed and approved by the authors. This is separate from the use of a language model as a component of the proposed system, described in Sections 3.4 and 3.5 and Appendix A.13.

ETHICS STATEMENT

This work uses experimental protein RNA complexes from the PDB and evaluates computational predictors of binding residues; it involves no human subjects, private user data, or personally identifiable information. All results are computational. The benchmark is also silent on several strata: the dsRBD, single stranded, internal loop, junction, and G quadruplex categories have no test representative, and the KH family contains a single target.

REPRODUCIBILITY STATEMENT

We submit the code as part of the supplementary materials. The GENESIS curation cascade, including every filter and the US align spectral partition that separates the 225 training from the 107 test targets, is specified in Section 4.1 and Appendix A.1; the label definition is Eq. equation 2. The tool library with per tool timeouts and mandatory core membership is Table 1, the adapter contract Eq. equation 5, the feature groups Section 3.7, and the fusion objective Eq. equation 15. Hardware, LightGBM settings, backbone, and sampling temperatures are in Section 4.1, with prompts, schemas, and the validation fallback in Appendix A.13. Metric definitions are in Appendix A.3, the baseline catalogue in Appendix A.2, robustness across cutoffs, splits, and five seeds in Appendix A.7, and per sample wall clock and API call counts in Appendix A.10.

REFERENCES

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

Rishal Aggarwal, Akash Gupta, Vineeth Chelur, CV Jawahar, and U Deva Priyakumar. Deeppocket: ligand binding site detection and segmentation using 3d convolutional neural networks. *Journal of Chemical Information and Modeling*, 62(21):5069–5079, 2022.

Haim Ashkenazy, Shiran Abadi, Eric Martz, Ofer Chay, Itay Mayrose, Tal Pupko, and Nir Ben-Tal. Consurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic acids research*, 44(W1):W344–W350, 2016.

Minkyung Baek, Ryan McHugh, Ivan Anishchenko, Hanlun Jiang, David Baker, and Frank DiMaio. Accurate prediction of protein–nucleic acid complexes using rosettafoldna. *Nature methods*, 21 (1):117–121, 2024.

Sankar Basu and Björn Wallner. Dockq: a quality measure for protein-protein docking models. *PloS one*, 11(8):e0161879, 2016.

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1): 235–242, 2000.

Marco Giulini, Victor Reys, João MC Teixeira, Brian Jiménez-García, Rodrigo V. Honorato, Anna Kravchenko, Xiaotong Xu, Raphaëlle Versini, Anna Engel, Stefan Verhoeven, et al. Haddock3: a modular and versatile platform for integrative modeling of biomolecular complexes. *Journal of Chemical Information and Modeling*, 65(13):7315–7324, 2025.

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

Elizabeth Jurrus, Dave Engel, Keith Star, Kyle Monson, Juan Brandi, Lisa E Felberg, David H Brookes, Leighton Wilson, Jiahui Chen, Karina Liles, et al. Improvements to the apbs biomolecular solvation software suite. *Protein science*, 27(1):112–128, 2018.

Ioanna Kalvari, Eric P Nawrocki, Nancy Ontiveros-Palacios, Joanna Argasinska, Kevin Lamkiewicz, Manja Marz, Sam Griffiths-Jones, Claire Toffano-Nioche, Daniel Gautheret, Zasha Weinberg, et al. Rfam 14: expanded coverage of metagenomic, viral and microrna families. *Nucleic acids research*, 49(D1):D192–D200, 2021.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

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

Radoslav Krivák and David Hoksza. P2rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *Journal of cheminformatics*, 10(1):39, 2018.

Jordy Homing Lam, Yu Li, Lizhe Zhu, Ramzan Umarov, Hanlun Jiang, Amélie Héliou, Fu Kit Sheong, Tianyun Liu, Yongkang Long, Yunfei Li, et al. A deep learning framework to predict binding preference of rna constituents on protein surface. *Nature communications*, 10(1):4941, 2019.

Tor Lattimore and Csaba Szepesvári. Bandit algorithms. *Cambridge University*, 355, 2019.

Michael C Lawrence and Peter M Colman. Shape complementarity at protein/protein interfaces, 1993.

Vincent Le Guilloux, Peter Schmidtke, and Pierre Tuffery. Fpocket: an open source platform for ligand pocket detection. *BMC bioinformatics*, 10(1):168, 2009.

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

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

Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik LL Sonnhammer, Silvio CE Tosatto, Lisanna Paladin, Shriya Raj, Lorna J Richardson, et al. Pfam: The protein families database in 2021. *Nucleic acids research*, 49(D1):D412–D419, 2021.

Felix Mölder, Kim Philipp Jablonski, Brice Letcher, Michael B Hall, Peter C van Dyken, Christopher H Tomkins-Tinch, Vanessa Sochat, Jan Forster, Filipe G Vieira, Christian Meesters, et al. Sustainable data analysis with snakemake. *F1000Research*, 10:33, 2025.

Jill E Moore, Michael J Purcaro, Henry E Pratt, Charles B Epstein, Noam Shoresh, Jessika Adrian, Trupti Kawli, Carrie A Davis, Alexander Dobin, et al. Expanded encyclopaedias of dna elements in the human and mouse genomes. *Nature*, 583(7818):699–710, 2020.

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

Typhaine Paysan-Lafosse, Matthias Blum, Sara Chuguransky, Tiago Grego, Beatriz Lázaro Pinto, Gustavo A Salazar, Maxwell L Bileschi, Peer Bork, Alan Bridge, Lucy Colwell, et al. Interpro in 2022. *Nucleic acids research*, 51(D1):D418–D427, 2023.

Inbal Paz, Efrat Kligun, Barak Bengad, and Yael Mandel-Gutfreund. Bindup: a web server for non-homology-based prediction of dna and rna binding proteins. *Nucleic acids research*, 44(W1): W568–W574, 2016.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. In *International Conference on Learning Representations*, volume 2024, pp. 9695–9717, 2024.

Rahmatullah Roche, Bernard Moussad, Md Hossain Shuvo, Sumit Tarafder, and Debswapna Bhattacharya. Equipnas: improved protein–nucleic acid binding site prediction using protein-language-model-informed equivariant deep graph neural networks. *Nucleic Acids Research*, 52 (5):e27–e27, 2024.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551, 2023.

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

Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

Chai Discovery team, Jacques Boitreaud, Jack Dent, Matthew McPartlon, Joshua Meier, Vinicius Reis, Alex Rogozhonikov, and Kevin Wu. Chai-1: Decoding the molecular interactions of life. *BioRxiv*, pp. 2024–10, 2024.

Rasna R Walia, Li C Xue, Katherine Wilkins, Yasser El-Manzalawy, Drena Dobbs, and Vasant Honavar. Rnabindrplus: a predictor that combines machine learning and sequence homology-based methods to improve the reliability of predicted rna-binding residues in proteins. *PloS one*, 9(5):e97725, 2014.

Renxiao Wang, Xueliang Fang, Yipin Lu, and Shaomeng Wang. The pdbbind database: collection of binding affinities for protein- ligand complexes with known three-dimensional structures. *Journal of medicinal chemistry*, 47(12):2977–2980, 2004.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Ying Xia, Chun-Qiu Xia, Xiaoyong Pan, and Hong-Bin Shen. Graphbind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues. *Nucleic acids research*, 49(9):e51–e51, 2021.

Yumeng Yan, Di Zhang, Pei Zhou, Botong Li, and Sheng-You Huang. Hdock: a web server for protein–protein and protein–dna/rna docking based on a hybrid strategy. *Nucleic acids research*, 45(W1):W365–W373, 2017.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

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

Yang Zhang, Zhewei Wei, Ye Yuan, Chongxuan Li, and Wenbing Huang. Equipocket: an e (3)-equivariant geometric graph neural network for ligand binding site prediction. *arXiv preprint arXiv:2302.12177*, 2023.

A Appendix

A.1 Dataset Construction

Table 5 details the GENESIS filtering cascade. Every filter is deterministic and reproducible from the raw PDB or CIF dump, and the entire pipeline is executed as a workflow with pinned tool versions (Mölder et al., 2025).

Table 5: GENESIS data curation cascade. Five filters are applied jointly to the enumerated pairs: RNA length in $[20, 100]$ nt, protein length $\leq$ 400 aa, at least five ground truth contact residues, experimental resolution ≤ 3.5 Å, and quality tier in {strict, standard}. Per filter intermediate counts are not separable because the filters are evaluated as a single predicate.

Stage	# samples
Initial PDB or CIF entries	1,809 (structures)
Enumerated protein RNA chain pairs	45,468
Surviving all five filters	2,018
Protein sequence clusters at 30% identity	332
Train split (US align spectral cut, TM < 0.35)	225
Test split	107

A.2 Baseline Catalogue

Table 6 lists every method evaluated. All 15 tools serve a dual role: each is run standalone as a baseline and is also available to MAESTRO for selection, so the comparison in Table 2 is between the same tools used alone and used in combination. The three Category E ensembles consume identical tool outputs, which isolates the value of the fusion from the value of tool coverage.

Table 6: Full baseline catalogue. Tools marked ✓in IN RIBOSEER are integrated inside the MAESTRO library and also evaluated standalone. Category E ensembles consume the identical tool outputs.

Method	Cat.	In RiboSeer	Paradigm
AlphaFold 3 (Abramson et al., 2024)	A	✓	End to end complex structure prediction
Boltz-2 (Passaro et al., 2025)	A	✓	End to end complex structure prediction
Chai-1 (team et al., 2024)	A	✓	End to end complex structure prediction
RoseTTAFold2NA (Baek et al., 2024)	A	✓	End to end complex structure prediction
RoseTTAFold All Atom (Krishna et al., 2024)	A	✓	End to end complex structure prediction
P2Rank (Krivák & Hoksza, 2018)	B	✓	Protein surface pocket detection
Fpocket (Le Guilloux et al., 2009)	B	✓	Protein surface pocket detection
DeepPocket (Aggarwal et al., 2022)	B	✓	Protein surface pocket detection
EquiPNAS (Roche et al., 2024)	C	✓	Sequence and structure binding residue
GraphBind (Xia et al., 2021)	C	✓	Sequence and structure binding residue
NucleicNet (Lam et al., 2019)	C	✓	Sequence and structure binding residue
RNABindRPlus (Walia et al., 2014)	C	✓	Sequence and structure binding residue
BindUP (Paz et al., 2016)	C	✓	Sequence and structure binding residue
HDOCK (Yan et al., 2017)	D	✓	Template based and free docking
HADDOCK 3 (Giulini et al., 2025)	D	✓	Information driven docking
Simple mean ensemble	E		Parameter free per residue mean of 15 tools
Weighted mean ensemble	E		Per tool weights fit on the training split
Stacked logistic regression	E		Linear stacker on raw tool scores

A.3 Evaluation Metric Definitions

Residue level. Pearson r measures linear correlation between $\hat{\mathbf{y}}$ and $\mathbf{y}^\star$ and is insensitive to absolute scale; Spearman ρ is invariant to monotone transformation and isolates ranking quality; R^2 penalises both correlation weakness and mean or variance mismatch and is the most stringent of

the three. Top k metrics set $k = |B_p|$ per sample, so precision, recall, and F1 coincide. MCC is computed at the per sample optimal F1 threshold and is robust to the $\sim 5\%$ positive rate.

Pocket level. Following (Krivák & Hoksza, 2018) and (Wang et al., 2004), the top scoring contiguous residue cluster is taken as $\hat{P}$. DCC is the distance between the C_α centres of mass of $\hat{P}$ and $P^\star$; DCA is the minimum distance from the predicted centre to any heavy atom of the bound RNA; a pocket is successfully identified when the corresponding distance is at most 4 Å.

Complex level. TM score is computed by US align (Zhang et al., 2022) with normalisation on the experimental protein chain. iRMSD is taken over interface residues after superposition, LRMSD over the RNA chain after superposition on the protein, and Fnat is the recovered fraction of native interfacial contacts. DockQ (Basu & Wallner, 2016) composites these, with ≥ 0.23 acceptable, ≥ 0.49 medium, ≥ 0.80 high.

A.4 Complex Structural Quality

Although RiboSeer's primary output is per residue probabilities rather than coordinates, the orchestrated Category A and D tools each produce full atomic complexes. Table 7 reports complex level metrics, including a consensus row that selects per sample the Category A model whose predicted interface best agrees with the fused probabilities. We state the outcome plainly: HDOCK attains the best DockQ (0.366) and HADDOCK 3 the best iRMSD (4.75 Å), both exceeding the consensus row. Complex selection is not a strength of the current system and we do not claim it as one; the fusion is trained for residue ranking and carries no geometric objective.

Table 7: Complex level structural metrics on the 107 target test split. The last two columns report the percentage of samples reaching at least acceptable and at least medium DockQ quality.

Method	TM ↑	iRMSD ↓	LRMSD ↓	Fnat ↑	DockQ ↑	% Acc.	% Med.+
AlphaFold 3	0.732	16.18	57.92	0.495	0.132	20.79	10.89
Boltz-2	0.733	13.09	52.73	0.525	0.187	30.39	16.67
Chai-1	0.665	18.53	60.09	0.347	0.089	14.56	3.88
RoseTTAFold2NA	0.275	24.54	67.13	0.039	0.012	0.00	0.00
RoseTTAFold All Atom	0.260	21.17	65.93	0.098	0.019	0.00	0.00
HDOCK	**0.830**	11.60	**41.93**	**0.785**	**0.366**	40.38	**38.46**
HADDOCK 3	0.801	**4.75**	65.89	0.012	0.224	**68.67**	1.20
RiboSeer (consensus)	0.692	14.93	53.60	0.540	0.168	27.88	13.46

A.5 Additional Ablations

Fusion method, extended. Table 8 adds the max aggregator and an MLP to the three learned alternatives reported in the main text.

Table 8: HARMONY fusion method ablation on the identical engineered feature space. *Mean* and *Max* operate on raw tool scores with no feature engineering; *Noisy OR* is the prototype aggregator $\hat{y}_i = 1 - \prod_k (1 - c_k s_k(i))$ with uniform c_k.

Fusion method	Pearson r ↑	Spearman ρ ↑	R^2 ↑
Mean of raw scores	0.474	0.421	0.270
Max of raw scores	0.303	0.303	0.115
Noisy OR (prototype)	0.461	0.421	0.254
MLP (3 layers, 256 hidden)	0.457	0.330	0.306
Random forest	0.571	0.448	0.413
Logistic regression	0.587	**0.458**	0.418
XGBoost	0.587	0.448	0.431
LightGBM (ours)	**0.593**	0.454	**0.436**

Feature groups, full grid. Table 9 gives every subset evaluated.

Table 9: Feature group ablation. **PR** Group 1 per residue multi metric; **XT** Group 2 cross tool interaction; **NB** Group 3 neighborhood context; **SC** Group 4 SCOPE profile. Each row retrains LightGBM from scratch on the indicated subset.

PR	XT	NB	SC	Pearson $r \uparrow$	Spearman $\rho \uparrow$	$R^2 \uparrow$
✓				0.589	0.453	0.433
✓	✓			0.587	0.452	0.432
✓		✓		0.587	0.450	0.430
	✓	✓		0.549	0.433	0.368
✓	✓	✓		0.589	0.451	0.432
✓	✓	✓	✓	**0.593**	**0.454**	**0.436**

Leave one out over the tool library. Table 10 removes each tool in turn. Four tools show a non negative Δ, which we report rather than prune, for the reason given in Section 4.5.

Table 10: Leave one out tool ablation. Δ is the change in Pearson r relative to the full system. Positive Δ means the system improves without that tool.

Configuration	Pearson $r \uparrow$	Spearman $\rho \uparrow$	$R^2 \uparrow$	Δ
Full system (15 tools)	**0.593**	**0.454**	**0.436**	–
w/o Boltz-2	0.530	0.416	0.372	−0.063
w/o Chai-1	0.589	0.449	0.433	−0.004
w/o RoseTTAFold2NA	0.586	0.447	0.431	−0.007
w/o RFAA	0.593	0.454	0.436	+0.000
w/o AlphaFold 3	0.596	0.456	0.440	+0.002
w/o P2Rank	0.591	0.452	0.434	−0.002
w/o Fpocket	0.595	0.453	0.439	+0.002
w/o DeepPocket	0.591	0.452	0.435	−0.002
w/o EquiPNAS	0.585	0.447	0.431	−0.008
w/o NucleicNet	0.582	0.445	0.427	−0.011
w/o GraphBind	0.595	0.452	0.440	+0.002
w/o RNABindRPlus	0.586	0.451	0.428	−0.007
w/o BindUP	0.594	0.455	0.437	+0.001
w/o HDOCK	0.570	0.443	0.404	−0.023
w/o HADDOCK 3	0.593	0.452	0.437	−0.001
Cat. A only (3 tools)	0.529	0.421	0.370	−0.064
Best single (Boltz-2)	0.439	0.387	0.260	−0.154

Mandatory core guardrail. Table 11 compares free selection, the proposed union of Eq. equation 10, and always running all 15 tools. Runtime here is measured end to end including sequential fallbacks and is therefore not comparable to the parallel per module breakdown of Appendix A.10.

Table 11: Effect of the mandatory core guardrail in MAESTRO. Free selection is cheapest but loses paradigm coverage; running everything is most expensive and also worse, because uninformative tools dilute the feature space.

Selection policy	Pearson $r \uparrow$	$R^2 \uparrow$	# tools / sample	Runtime (min)
Free LLM selection (3 to 15 tools)	0.569	0.401	5.7	**37.4**
Mandatory core + LLM extras	**0.593**	**0.436**	8.9	55.9
All 15 tools always	0.575	0.422	15.0	74.3

Language model backbone. Table 12 swaps the reasoner while holding prompts and schemas fixed. Absolute values in this table come from an earlier fusion checkpoint retrained per backbone, so they sit below the headline; the comparison across rows is the quantity of interest.

Table 12: Language model backbone ablation. "% Fail" is the schema validation failure rate. All rows share prompts, schemas, tool versions, and the per backbone retraining protocol.

Backbone	Pearson $r \uparrow$	Spearman $\rho \uparrow$	$R^2 \uparrow$	% Fail
No LLM (UCB only)	0.554	0.436	0.389	0.0
DeepSeek V4 Flash	0.554	0.437	0.385	2.3
Qwen 3.7 Max	0.543	0.441	0.376	9.6
Gemini 3.5 Flash	0.547	0.440	0.383	0.0
GLM 5.1 (ours)	**0.565**	**0.446**	**0.395**	0.0

Prompting strategy. Table 13 ablates the chain of thought switch and the sampling temperature, under the same per configuration retraining protocol as Table 12.

Table 13: Prompt engineering ablation. **CoT**: chain of thought scratchpad on SCOPE. **T**: sampling temperature for exploration calls.

CoT	T	Pearson $r \uparrow$	Spearman $\rho \uparrow$	$R^2 \uparrow$	% Fail
–	0.0	0.548	0.433	0.378	0.0
–	0.7	0.549	0.437	0.378	0.1
✓	0.0	0.560	0.434	0.391	0.3
✓	0.7	**0.565**	**0.446**	**0.395**	0.0
✓	1.0	0.561	0.441	0.389	0.2

Memory EMA rate. Table 14 sweeps η in Eq. equation 17. This sweep is run in a streaming protocol where the fusion is frozen and only the weight tensor adapts, which isolates the contribution of online adaptation and yields lower absolute values than the batch trained headline.

Table 14: MEMORY EMA rate ablation under the streaming protocol. The frozen row keeps prior weights unchanged; $\eta = 1.0$ discards all history.

EMA rate η	Pearson $r \uparrow$	Spearman $\rho \uparrow$	$R^2 \uparrow$
0.00 (frozen)	0.527	0.389	0.369
0.05	0.529	0.402	0.375
0.10 (default)	**0.535**	**0.408**	**0.380**
0.20	0.502	0.384	0.361
0.50	0.509	0.388	0.360
1.00 (no memory)	0.490	0.378	0.347

A.6 STRATIFIED RESULTS

Tables 15 to 18 partition the 107 test targets along four axes. Two caveats apply throughout. The strata are unbalanced, with 94 of 107 targets falling into the multi domain or novel fold family, so the canonical RBP families are each represented by a handful of targets at most. And the axes are not independent: longer RNAs correlate with larger and more fragmented interfaces, so the length and difficulty tables describe overlapping populations rather than orthogonal effects.

A.7 ROBUSTNESS

The three tables below vary one protocol choice at a time. Table 19 regenerates labels under four contact cutoffs without retraining; Table 20 rebuilds the partition four ways and retrains per split; Table 21 varies only the seed. Absolute values differ across these because the retraining protocol differs, so the quantity to read is the margin over the strongest baseline, which persists in every case.

Table 15: Performance stratified by RNA length. Short 20 to 40 nt, medium 41 to 70 nt, long 71 to 100 nt.

	Short		Medium		Long	
Method	n	$r \uparrow$	n	$r \uparrow$	n	$r \uparrow$
Boltz-2	27	0.454	48	0.517	29	0.295
EquiPNAS	27	0.434	49	0.483	26	0.260
Chai-1	28	0.417	48	0.461	29	0.216
P2Rank	28	0.235	48	0.205	27	0.077
RoseTTAFold2NA	28	0.089	50	0.182	29	0.002
RiboSeer	28	**0.586**	50	**0.719**	29	**0.362**

Table 16: Performance stratified by protein domain family, assigned by InterPro (Paysan-Lafosse et al., 2023) and cross checked against the SCOPE profile. No dsRBD target survives into the test split, so that family is omitted. The KH column contains a single target, so no conclusion is drawn from it; we report it for completeness.

Family	RRM	KH	ZnF	Multi or novel
n in test split	8	1	4	94
Boltz-2	0.303	0.101	0.227	0.462
EquiPNAS	0.371	0.145	0.330	0.423
Chai-1	0.429	**0.171**	0.326	0.383
RiboSeer	**0.498**	-0.024	**0.554**	**0.603**

A.8 RANKING METRICS

Tables 22 and 23 report threshold free and top k quality, complementing the correlation metrics of Table 2. The ordering is unchanged, which matters because correlation and ranking can dissociate: a method can track the shape of the label vector without placing the right residues at the top. That both move together indicates the fusion improves the ranking itself rather than only its calibration.

A.9 PER SAMPLE DISTRIBUTION

Table 24 confirms the means of Table 2 are not outlier driven: RiboSeer leads at every quantile from the 25th upward. The two case studies of Figure 2 sit above the 90th percentile and are therefore not representative.

A.10 COMPUTATIONAL COST

A.11 FEATURE IMPORTANCE

Table 26 shows that the fusion concentrates on Group 1 signals from the two load bearing tools identified by leave one out, with the cross tool vote count ranking third. This is consistent with Table 9, where Group 1 alone nearly matches the full feature set.

A.12 FAILURE MODE ANALYSIS

Table 27 categorises the bottom decile of test samples. Every failure exhibits both low Category A confidence and high tool disagreement, which suggests the two conditions co-occur rather than acting independently: when no structure predictor is confident, the remaining tools have no anchor to agree on. This is an actionable abstention signal, since both quantities are computable at inference time without labels.

A.13 PROMPT AND SCHEMA DETAILS

An early prototype placed the fusion arithmetic inside the language model and proved untestable: identical inputs produced different aggregates across calls, and no unit test could pin the be-

Table 17: Performance stratified by RNA structural context. The single stranded, internal loop, junction, and G quadruplex contexts have no test representative and are omitted; the benchmark is silent on them.

Context	Stem-loop (n=83)	Unstructured (n=24)
Boltz-2	0.447	0.409
EquiPNAS	0.424	0.380
RiboSeer	**0.599**	**0.550**

Table 18: Performance by SCOPE assigned difficulty. Hard targets are those flagged with low homolog availability or multi domain ambiguity.

	Easy		Medium		Hard	
Method	n	$r \uparrow$	n	$r \uparrow$	n	$r \uparrow$
Boltz-2	6	0.366	16	0.302	82	0.471
EquiPNAS	6	0.382	15	0.267	81	0.443
RiboSeer	7	**0.535**	16	**0.365**	84	**0.634**

haviour down. Moving all arithmetic out of the model makes each numerical stage independently testable, and the complete pipeline is orchestrated as a reproducible workflow with pinned tool versions (Mölder et al., 2025).

All GLM 5.1 calls use the function calling API with Pydantic validated JSON output. SCOPE runs at temperature 0.1 with a chain of thought scratchpad and returns a difficulty analysis string plus a categorical label over domain architecture, RNA context, difficulty tier, and length bin. MAESTRO runs at temperature 0.1 for exploitation and 0.7 for exploration calls, receives the serialised weight tensor and UCB utilities of Eq. equation 8, and returns an ordered tool list with a free text rationale that is logged but never parsed numerically. A validation failure triggers one error correcting retry; a persistent failure falls back to UCB only selection, which Table 12 measures as the "No LLM" row.

Table 19: Sensitivity to the ground truth contact cutoff d_{cut}. Labels are regenerated under each threshold and all methods re-evaluated on the same stored predictions without retraining, which is why the default column differs slightly from the retrained headline of Table 2.

d_{cut}	3.5 Å	4.0 Å	4.5 Å (default)	6.0 Å
Boltz-2	0.365	0.418	0.439	0.497
EquiPNAS	0.376	0.407	0.413	0.422
Chai-1	0.313	0.360	0.381	0.424
P2Rank	0.154	0.176	0.180	0.173
RoseTTAFold2NA	0.094	0.101	0.109	0.124
RiboSeer	**0.505**	**0.569**	**0.588**	**0.598**

Table 20: Robustness across alternative train and test partitions. Each split is rebuilt from the same 332 cluster representatives and the fusion is retrained from scratch per split, so absolute values differ from Table 2; the margin Δ is the quantity of interest.

Split scheme	RiboSeer $r \uparrow$	Boltz-2 $r \uparrow$	Δ
Spectral cut (default)	0.593	0.439	$+0.154$
Random 70 / 30 (cluster level)	0.543	0.400	$+0.143$
Temporal ($\leq$2021 / $\geq$2022)	0.431	0.317	$+0.114$
Tight identity (25%)	0.407	0.296	$+0.111$

Table 21: Variance across five random seeds controlling LightGBM initialisation, the train and validation holdout, and the sampling schedule. Means sit marginally below the single retained configuration of Table 2, which was fixed before test evaluation.

Metric	Mean $\pm$ std (5 seeds)
Pearson r	0.592 ± 0.001
Spearman ρ	0.451 ± 0.001
R^2	0.428 ± 0.001
Top k F1	0.581 ± 0.004

Table 22: Top k metrics on the 107 target test split, with k set per sample to the ground truth pocket size $|B_p|$, so precision, recall, and F1 coincide. MCC is computed at the per sample optimal F1 threshold.

Method	Top k Prec. $\uparrow$	Top k Rec. $\uparrow$	Top k F1 $\uparrow$	MCC $\uparrow$
Boltz-2	0.501	0.501	0.501	0.544
EquiPNAS	0.452	0.452	0.452	0.484
Chai-1	0.438	0.438	0.438	0.475
P2Rank	0.258	0.258	0.258	0.285
RoseTTAFold2NA	0.207	0.207	0.207	0.216
RiboSeer	**0.586**	**0.586**	**0.586**	**0.637**

Table 23: Per sample AUROC and AUPRC on the 107 target test split.

Method	AUROC $\uparrow$	AUPRC $\uparrow$
Boltz-2	0.831	0.539
EquiPNAS	0.838	0.481
Chai-1	0.797	0.472
P2Rank	0.652	0.286
RoseTTAFold2NA	0.590	0.225
RiboSeer	**0.899**	**0.638**

Table 24: Quantiles of the per sample Pearson r distribution on the 107 target test split.

Method	10%	25%	Median	75%	90%
Boltz-2	0.068	0.309	0.482	0.648	0.723
EquiPNAS	**0.113**	0.231	0.455	0.582	0.676
RiboSeer	0.102	**0.389**	**0.699**	**0.820**	**0.901**

Table 25: Per sample wall clock cost, averaged over the 107 test samples on 2×A6000 GPUs with adapters dispatched concurrently. Only SCOPE and MAESTRO issue API calls, one each, averaging 1,517 input and 963 output tokens per call for a per sample total of 3,034 and 1,926; the five deterministic modules issue none.

Module	Wall clock (s)	# API calls
GENESIS (preprocessing)	<0.1	0
SCOPE (profiling)	26.2	1
MAESTRO (selection)	26.2	1
RELAY (tool execution)	661.9	0
HARMONY (fusion)	<0.1	0
VERDICT (assessment)	<0.1	0
MEMORY (weight update)	<0.1	0
Total per sample	714.4	2

Table 26: Top 10 HARMONY features by average LightGBM gain, normalised to the largest. **G1** per residue multi metric, **G2** cross tool interaction, **G4** SCOPE profile.

Rank	Feature	Group	Norm. importance
1	Boltz-2 gate	G1	1.000
2	HDOCK main score	G1	0.700
3	Vote count $v(i)$	G2	0.598
4	Boltz-2 global confidence	G1	0.438
5	Boltz-2 main score	G1	0.305
6	Boltz-2 secondary score	G1	0.186
7	EquiPNAS global confidence	G1	0.171
8	SCOPE RNA length bin	G4	0.107
9	HDOCK main z score	G1	0.105
10	Boltz-2 $\times$ Chai-1 score product	G2	0.102

Table 27: Failure mode breakdown over the bottom 10% of test samples by Pearson r ($n = 11$). Rows do not sum to 100% because a sample may exhibit more than one mode. Two further modes we screened for, RNA secondary structure misprediction and unattributed failure, account for no cases.

Failure mode	# samples	% of failures
All Cat. A tools low confidence (mean pLDDT < 60)	11	100.0
Tool disagreement (mean pairwise Jaccard < 0.2)	11	100.0
Multi domain interface, wrong domain selected	4	36.4