ONE AGENT, MANY STAINS: AGENTIC ENSEMBLE DESIGN FOR MULTI-STAIN COMPUTATIONAL PATHOLOGY

Anonymous authors
Paper under double-blind review

ABSTRACT

Pathological diagnosis routinely relies on multiple stains of the same tissue, including H&E and special or immunohistochemical stains, and multi-stain slide sets are increasingly available in digitized archives. Yet computational pathology models typically read a single stain or fuse stains using a single fixed architecture applied to every task, even though sections of different stains are unregistered and frequently missing, and labeled cohorts are small. We present ARISE (Agentic Rule Evolution for Integrating Stain-specific Experts), a simple and effective alternative: one well-trained multiple-instance model per stain is kept fixed, and the rule that integrates their predictions is designed separately for each task. ARISE builds a pool of candidate rules that abstain when a stain is missing and scores each candidate by its margin over the strongest conventional method. An LLM agent evolves the pool, the objective, and the search, while a numerical inner loop fits the weights. The output is a frozen program, a weighted list of pool members, that runs without an LLM or a GPU. We evaluate ARISE on four cohorts covering three organs (kidney, head and neck, lymphoma), panels of four to eight stains, both classification and time-to-event prediction, and one external site. The comparison methods, run under one protocol, include single-stain models, conventional ensembles, and published multi-stain fusion architectures. ARISE outperforms the strongest conventional method on every task in every cohort. On kidney biopsy lesion scoring, the margin holds in the internal cohort and at the external site (mean ΔAUC $+0.014$ and $+0.037$, both $p < 0.001$). Applied without modification, the same procedure improves head-and-neck survival and recurrence prediction (ΔAUC $+0.009$ and $+0.032$) and lymphoma survival prediction (ΔC-index $+0.035$). These results suggest that multi-stain integration can be treated as a task-specific design problem rather than a fixed architectural choice. ARISE provides a common procedure for constructing such integration programs across organs, stain panels, and prediction tasks.

1 INTRODUCTION

Histopathological assessment often relies on multiple stains that reveal complementary aspects of the same tissue, from morphology and tissue structure to protein expression (Trimarchi et al., 2017; Hans et al., 2004). Integrating these views is important for both diagnostic assessment and outcome prediction. As multi-stain datasets become available, developing and evaluating methods that combine information across stains becomes increasingly important (Vrabac et al., 2021; Dörrich et al., 2025). Computational pathology foundation models provide representations for individual slides and perform well on diverse downstream tasks, including biomarker and survival prediction (Xu et al., 2024; Wang et al., 2024). Still, these representations alone do not specify how to integrate complementary evidence from different stains for a particular prediction task.

Integrating information across stains requires accounting for several practical constraints. Tissue distortion and sectioning artifacts can prevent precise spatial correspondence between differently stained sections, even after registration (Weitz et al., 2024; Lotz et al., 2023). The available stains can also vary across patients, resulting in incomplete stain panels (Gallagher-Syed et al., 2025; Jaume et al., 2024). In addition, some multi-stain cohorts contain relatively few patients with out-

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.

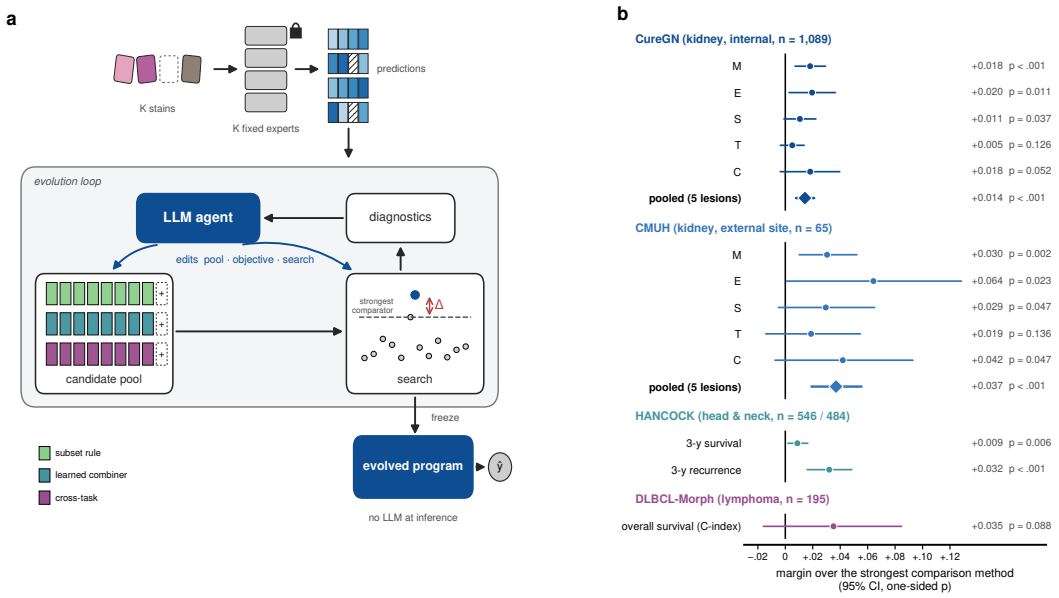

Figure 1: (a) Overview of ARISE. One fixed expert per stain scores each patient, and a patient with a missing stain (dashed, hatched) is scored from the predictions that exist. The predictions enter an evolution loop: a numerical search selects an integration program from a typed candidate pool by its margin over the strongest comparison method, and an LLM agent, reading the search diagnostics, revises the pool, the objective, and the search itself. The final program is frozen and runs without an LLM at inference. (b) Margin of the evolved program over the strongest comparison method in every task–cohort cell (AUROC; C-index for DLBCL-Morph), with the pooled margin over the five lesions for each kidney cohort (diamonds); bars are two-sided 95% confidence intervals from a patient-level paired bootstrap ($B = 40{,}000$), p is one-sided for the margin being positive.

come labels, limiting the data available to develop and evaluate integration methods (Vrabac et al., 2021; Cheng et al., 2025). Several multi-stain methods use dedicated modules to fuse stain-specific features, requiring researchers to choose a fusion architecture and train it for the prediction task (Koch et al., 2026; Gallagher-Syed et al., 2025).

A practical alternative is to keep the stain-specific models fixed and integrate their predictions (Kittler et al., 1998; Wolpert, 1992). Prediction-level integration requires no spatial correspondence between sections, and a patient with a missing stain can still be scored from the predictions that exist. It does, however, leave the integration rule itself as a design choice. The space of such rules is large, covering which models to combine and how to transform and weight their predictions (Kittler et al., 1998; Wolpert, 1992; Caruana et al., 2004). Committing to a single rule in advance is therefore unreliable. Recent systems such as AlphaEvolve and OpenEvolve show that an LLM can generate and revise programs using evaluation feedback (Novikov et al., 2025; Sharma, 2025). Building on this idea, we introduce ARISE (Agentic Rule Evolution for Integrating Stain-specific Experts), a framework in which an LLM agent evolves a task-specific integration program over fixed stain-specific predictions, searching for the integration rule rather than fixing it in advance.

We evaluate ARISE on four cohorts spanning three organs, covering classification and survival prediction and including an external-site evaluation. We compare it with single-stain models, conventional ensembles, and published multi-stain fusion methods. ARISE outperforms the strongest conventional method on every task in every cohort, and the margin holds at the external site. Our contributions are:

- An agentic framework for designing task-specific multi-stain integration programs over fixed stain-specific predictions.
- An evaluation across four cohorts that compares prediction-level integration with single-stain models and multi-stain fusion methods.

- An analysis of the final integration designs, including a comparison between agent-designed search and conventional ensemble selection on the same candidate pool.

2 Related Work

Multi-stain integration. Stains can be combined at the input, feature, or prediction level (Baltrušaitis et al., 2019; Schouten et al., 2025). Combining corresponding pixels across stains requires spatial alignment, which is imperfect for serial sections (Borovec et al., 2020; Weitz et al., 2024; Lotz et al., 2023). Feature-level methods learn a joint representation with a dedicated module: UNICORN applies a transformer to stain-specific tokens and handles missing stains natively (Koch et al., 2026), PathMoE gates stain-specific experts (PathMoE), and further feature-level models have been reported for colorectal cancer, rheumatology and immunohistochemistry panels (Foersch et al., 2023; Nakhli et al., 2023; Gallagher-Syed et al., 2023; 2025); MADELEINE instead uses cross-stain alignment during pretraining to learn slide representations for downstream H&E tasks (Jaume et al., 2024). Prediction-level integration combines the outputs of separately trained models through rules such as averaging or stacking (Kittler et al., 1998; Wolpert, 1992); in kidney pathology, models trained independently on four stains have been integrated for treatment-response prediction (Cheng et al., 2025). Language models have also been placed inside the fusion model, for example, to arbitrate among expert predictions at inference (Debsarkar & Prasath, 2026). ARISE instead uses an LLM agent offline to search for a prediction-level integration program for each task, and the resulting program runs at inference without an LLM.

Ensemble selection and stacking. Combining the outputs of several predictors has a long history, from fixed combination rules (Kittler et al., 1998) and stacked generalization (Wolpert, 1992) to selecting an ensemble from a library of models by greedy forward selection (Caruana et al., 2004; 2006). Auto-sklearn uses this selection procedure (Feurer et al., 2015), while AutoGluon combines multi-layer stacking with ensemble selection (Erickson et al., 2020). Recent work learns the combination end-to-end, with differentiable model selection (Kotary et al., 2023) or a neural ensembler that assigns per-instance weights (Pineda Arango et al., 2025). These methods select or weight a set of predictors they are given. ARISE differs in two ways: the agent adds task-specific candidates to the pool, and it runs its own search over that pool. In §5 we test each difference separately by running greedy selection and the agent's search on both the original pool and the agent-extended pool.

LLM-driven program evolution. FunSearch showed that a frozen LLM paired with an executable evaluator can evolve programs toward high scores (Romera-Paredes et al., 2024). AlphaEvolve and OpenEvolve extend this approach to broader program optimization (Novikov et al., 2025; Sharma, 2025), while ShinkaEvolve focuses on sample-efficient evolutionary search (Lange et al., 2026). AIDE frames machine-learning engineering as a tree search over solution code (Jiang et al., 2025). ContextEvolve compresses optimization history into context for code generation (Su et al., 2026), while Procedural Graphs revises explicit procedural graphs that guide an agent's actions (Lu et al., 2026). ARISE applies evaluation-guided program evolution to a clinical prediction problem, where the program is an integration rule over fixed, stain-specific predictors, and the evaluator compares each candidate against a panel of conventional methods.

3 Method

ARISE starts with one fixed model per stain, the *experts*. Each expert is a gated-attention multiple-instance model (ABMIL; Ilse et al., 2018) trained on frozen Virchow2 tile features (Zimmermann et al., 2024) with the same recipe for every stain, and it outputs one prediction per patient: a probability for a binary task, a probability vector for a multiclass task, or a risk score for a time-to-event task. A patient without a stain receives no prediction from that expert. The experts are trained once and never revised. ARISE designs the *integration program*, the rule that maps the K expert predictions to one score per patient, from the experts' validation predictions and labels alone. The method rests on three design choices. The program space is a pool of candidate rules that an LLM agent can extend (§3.1). A program's fitness is its margin over the strongest conventional method (§3.2). The search is bilevel: the agent evolves the design while a numerical inner loop fits the program (§3.3).

3.1 CANDIDATE POOL

ARISE designs a weighted program composed of *integration rules*: each rule maps the expert predictions of a patient to one score, and the program is a weighted mean of the rule outputs in the task's own space (within-cohort ranks for a binary task, log-probabilities for a multiclass task, risk ranks for a time-to-event task). A missing stain contributes no value to a rule: a rule over several stains uses only those available for that patient, and a rule for which all required stains are missing, including a single-stain rule when that stain is absent, outputs the neutral score 0.5. No prediction is imputed for the missing stain itself.

The *base pool* is the set of rules and their *members* from which a conventional ensemble method would select, and it requires no labels. It includes every subset of experts under each of a fixed list of pooling rules (mean, median, rank mean, maximum, minimum, trimmed mean), the experts themselves, and power means, copula combinations, and fixed gating rules over all experts. It enumerates once per cohort from K alone and supports up to about 1,800 members for $K = 8$.

The agent extends this pool with families of rules that the enumeration cannot express when the diagnostics of the current best program suggest one; the pool is instantiated from a specification that names the families, so adding a family is one edit of that specification. Three families arose in our experiments. *Learned combiners* are linear, kernel, and tree models fitted, on validation rows only, to features derived from the predictions (the predictions themselves, their ranks, pairwise differences, and agreement and uncertainty statistics); a combiner that scores one validation block must be fitted on the others, so this family is possible only on the kidney cohorts, whose folds provide several validation blocks. *Cross-task members* let a task borrow members from a related task's pool so that a lesion score can draw on predictions for other lesions from the same biopsy. *Anchored order-statistic rules*, added to the head and neck tasks, pool a subset that always contains two anchor stains with an order statistic, a power mean, or a threshold count, after filling an absent stain at the neutral rank, which changes an order statistic where it leaves a mean unchanged. With these additions, the pool holds about 800 members per task in the kidney cohorts and about 2,400 in the head and neck cohort. §5.1 measures what the additions contribute.

3.2 STRONGEST-COMPARATOR OBJECTIVE

The fitness of a candidate program is its margin over the strongest conventional method, measured on validation predictions using the same paired-bootstrap statistic as in §4 for the evaluation data. A fixed *comparison panel* holds the single experts, the standard pooling rules over all experts, and the published multi-stain fusion architectures trained on the same features and splits. On each validation split, the reference is the panel member that is strongest on that split, chosen after the fact. A program is scored by the paired-bootstrap statistic $t = \Delta/\mathrm{SE}(\Delta)$ of its margin Δ over the reference in the task's native metric (AUROC, macro one-versus-rest AUROC, or Harrell's C), with patients resampled within the split. Scoring a statistic rather than a raw metric discounts a margin on a small split. When a cohort has several validation splits, the objective is a soft minimum of the per-split statistics plus a small term on their sum,

$$J(\text{program}) = \mathrm{softmin}_s \; t_s + \lambda \sum_s t_s, \tag{1}$$

so that a program is rewarded only if it improves on every split. The bootstrap draw used during the search is separate from the one used for the final verdict.

3.3 BILEVEL EVOLUTION

ARISE separates design from fitting. The agent, an LLM (Claude Opus 5) run offline once per task, edits a *program specification* with four fields: the member families of the pool, the feature set of the learned combiners, the shape of the objective, and the settings of the numerical search. It cannot change the experts, the comparison panel, the metric, or the verdict procedure. One outer iteration has four steps. *Propose*: the agent writes or edits the specification. *Evaluate*: the pool is instantiated, and the inner search, bagged greedy forward selection followed by local moves and simulated annealing on the weights, from several random starts, selects members and weights and returns the best program with its objective value. *Observe*: the agent reads the program's diagnostics: its per-split margins, member composition, agreement between members, and behavior on patients with missing

Table 1: Evaluation cohorts. Evaluation: five stratified random patient-level splits per cohort, each about 60/20/20 train/validation/evaluation, with results averaged over the splits; *CMUH is an external cohort with the same stains and labels as CureGN, scored by the five CureGN models. Patients: evaluated patients; Missing: share of them lacking at least one stain. Every method is scored on every patient.

Cohort	Stains	Task (metric)	Evaluation	Patients	Missing
CureGN (kidney) (Mariani et al., 2019)	H&E, PAS, TRI, SIL	MEST-C (AUROC)	5 splits	1,089	12%
CMUH* (kidney)	same	MEST-C (AUROC)	external	65	12%
HANCOCK (head and neck) (Dörrich et al., 2025)	H&E + 7 IHC	3-year survival (AUROC)	5 splits	546	10%
		3-year recurrence (AUROC)		484	10%
DLBCL-Morph (lymphoma) (Vrabac et al., 2021)	H&E + 5 IHC	overall survival (C-index)	5 splits	195	45%

stains. *Refine*: the agent edits the specification; it discards revisions that do not improve the objective and restores the previous specification. The loop stops when the agent judges that further revisions would not improve the objective; the specification went through six revisions on the kidney tasks, six on the head and neck tasks, and eleven on the lymphoma task. The best program is then frozen: a member list with its weights and the fitted combiners, shared across all cohort splits, that runs at inference on stored predictions without an LLM. Add a new stain by training one more expert and re-running the loop.

4 Experiments

4.1 Cohorts and tasks

Table 1 lists the four cohorts, chosen to test the same procedure across organs, stain panels, and task types, with one external site. Each kidney biopsy carries the five Oxford MEST-C lesion scores (M, E and S binary; T and C three-level), and each lesion is a task of its own. CMUH comes from a second institution and a different scanner and takes no part in training, validation, or the search. The two public cohorts take their case lists, labels, and exclusions from the source papers and official repositories.

4.2 Comparison methods and protocol

All methods use the same Virchow2 tile features and the same patient-level splits. The single-stain experts are the ABMIL models of §3, trained once per cohort with one schedule and no early stopping.

Comparison methods. Table 2 compares the evolved program with four groups of methods. (i) The single-stain experts, each used on its own. (ii) Eight fixed pooling rules (Kittler et al., 1998) applied to the expert predictions over the stains a patient has: mean, mean logit, median, rank mean, maximum, minimum, harmonic mean, and trimmed mean (on the time-to-event task, applied to within-split risk ranks). (iii) Two published end-to-end multi-stain architectures, UNICORN (Koch et al., 2026) and PathMoE (PathMoE), retrained on each cohort and using their own handling of missing stains. (iv) Learned ensembles of the experts, in the tradition of stacking (Wolpert, 1992): validation-weighted averaging, convex weights, Bayesian model averaging, ridge, elastic-net and gradient-boosted stacking, each fitted on the validation predictions of the experts.

Protocol. In each task–cohort cell, the reference is the strongest method of groups (i)–(iii), the *comparison panel*. We estimate the uncertainty of the program's margin relative to the reference using a patient-level paired bootstrap ($40{,}000$ resamples), and report the margin, its two-sided 95% confidence interval, and the one-sided p-value for the margin being positive. Program selection uses training and validation data only; the evaluation data take no part in the search and are scored once, on the frozen program.

4.3 MAIN RESULTS

Table 2 lists every comparison method and ARISE in the 13 task–cohort cells, and Fig. 1b plots ARISE's margin over the strongest panel method in each cell, with its 95% confidence interval and one-sided p-value. ARISE is ahead of the strongest panel method in all 13 cells and ahead of every learned ensemble in all 13. On kidney lesion scoring the mean margin over the five lesions is $+0.014$ AUROC on CureGN (95% CI $+0.008$ to $+0.021$) and $+0.037$ on CMUH ($+0.019$ to $+0.056$), both $p < 0.001$, so the margin holds at a second institution and scanner without adaptation; per lesion, M, E and S reach $p < 0.05$ on both cohorts, C on the external cohort only, and T on neither. On HANCOCK, the program shows the strongest fixed-effect effect: $+0.009$ on survival ($p = 0.006$) and $+0.032$ on recurrence ($p < 0.001$), and all 24 methods are significant at $p < 0.05$. On DLBCL-Morph, the smallest cohort with 67 events, the margin over the median rule is $+0.035$ ($p = 0.088$). Against the two fusion models alone, ARISE is ahead in all 26 cells. The strongest panel method varies between cells (the mean, the mean logit, the rank mean, the median, the single H&E expert, or UNICORN), and for three of the five kidney lesions it differs between the internal and the external cohort, which is why a rule chosen once for all tasks is not enough.

5 ANALYSIS

Table 2 shows ARISE's margin over every comparison method. This section asks where that margin comes from, and whether it depends on the patients with missing stains. The ablations use the two kidney cohorts: the same frozen programs are scored on the internal splits of CureGN and at the external site CMUH, so a gain attributed to a component cannot be an artifact of one split. ARISE differs from conventional ensemble selection in three respects. It selects from the candidate pool of §3.1 rather than from the stain predictions directly; the agent extends that pool; and the agent designs the search. §5.1 measures the three steps in turn, replacing each with its conventional counterpart while the objective stays fixed. We freeze every program before evaluation, run it without an LLM, and score it once on the held-out splits. §5.2 then repeats the main comparison separately on patients with a complete and with an incomplete stain panel.

5.1 CONTRIBUTIONS OF POOL EXPANSION AND SEARCH

Table 3 reports the evaluation AUROC for the program each procedure selects, with the strongest comparison method in Table 2 as the floor.

The first step is the pool without the agent's cross-task members: the base pool of §3.1 together with the learned combiners, about 166 members per lesion. Greedy forward selection (Caruana et al., 2004) on this pool is above the strongest comparison method in every cell but CMUH C, where the two are level: by 0.003 to 0.017 AUROC on CureGN and by 0.010 to 0.037 on CMUH. Given subset rules and learned combiners to mix, a conventional procedure therefore already accounts for much of the margin.

The second step is the agent's pool extension. During the loop, the agent added cross-task members borrowed from the pools of the other four lesions, bringing the pool to about 792 members. With greedy selection on both pools, the additions change the CureGN result by at most 0.008 (M) but raise four of the five CMUH lesions: M from 0.750 to 0.767, E from 0.721 to 0.740, S from 0.788 to 0.800 and C from 0.759 to 0.784, with T unchanged. The gain from the extended pool appears at the external site.

The third step is the search. On the extended pool, we compare greedy forward selection, bagged greedy selection (Caruana et al., 2006), random search with the same number of objective evaluations, and the agent-designed search of §3.3, reporting the median over five search seeds for the stochastic procedures. Bagged greedy and the agent-designed search stay within 0.011 of greedy selection in every cell; random search is below greedy in nine of the ten cells, by up to 0.020 (CMUH C). The committed program of Table 2, a single run of the agent-designed search, is within 0.006 of the best procedure in every cell and exceeds greedy selection on the same pool by at most 0.017 (CMUH C).

Across the three steps, the margin over conventional ensembles comes mainly from the candidate pool: from the base pool and the learned combiners on both cohorts, and from the cross-task mem-

Table 2: Every comparison method and ARISE in each task–cohort cell (AUROC; C-index for DLBCL-Morph; mean over the five splits). In each column, the strongest method of the comparison panel, groups (i)–(iii), is underlined and ARISE is in bold; Fig. 1b gives ARISE's margin over the underlined method with its confidence interval. Single-stain experts on HANCOCK are scored on the patients that the stain covers.

	CureGN (internal)					CMUH (external)					HANCOCK		DLBCL
Method	M	E	S	T	C	M	E	S	T	C	SURV	REC	OS
Single-stain experts (i)													
HE	.690	.799	.794	.893	.749	.646	.566	.778	.886	.680	.703	.734	.666
PAS	.767	.758	.776	.885	.743	.690	.585	.690	.706	.619	–	–	–
TRI	.735	.724	.743	.861	.714	.680	.658	.655	.847	.657	–	–	–
SIL	.730	.743	.763	.866	.719	.624	.601	.632	.788	.604	–	–	–
CD3	–	–	–	–	–	–	–	–	–	–	.698	.667	–
CD8	–	–	–	–	–	–	–	–	–	–	.706	.746	–
CD56	–	–	–	–	–	–	–	–	–	–	.709	.732	–
CD68	–	–	–	–	–	–	–	–	–	–	.706	.713	–
CD163	–	–	–	–	–	–	–	–	–	–	.694	.700	–
PDL1	–	–	–	–	–	–	–	–	–	–	.683	.637	–
MHC1	–	–	–	–	–	–	–	–	–	–	.681	.698	–
CD10	–	–	–	–	–	–	–	–	–	–	–	–	.639
BCL6	–	–	–	–	–	–	–	–	–	–	–	–	.666
MUM1	–	–	–	–	–	–	–	–	–	–	–	–	.628
BCL2	–	–	–	–	–	–	–	–	–	–	–	–	.539
MYC	–	–	–	–	–	–	–	–	–	–	–	–	.652
Fixed pooling rules (ii)													
mean	.787	.816	.806	.916	.807	.722	.675	.733	.852	.696	.754	.743	.691
mean logit	.783	.813	.804	.915	.801	.715	.628	.724	.828	.695	.754	.741	.689
median	.782	.811	.802	.914	.806	.735	.649	.734	.842	.690	.748	.743	.697
rank mean	.784	.814	.806	.890	.789	.740	.644	.732	.850	.718	.750	.746	.691
max	.773	.805	.798	.907	.774	.689	.669	.725	.847	.668	.748	.733	.639
min	.768	.789	.796	.897	.757	.673	.572	.712	.780	.648	.744	.731	.644
harmonic mean	.771	.794	.800	.905	.775	.686	.574	.711	.785	.659	.753	.742	.679
trimmed mean	.782	.811	.802	.914	.806	.735	.649	.734	.842	.690	.753	.744	.677
End-to-end fusion (iii)													
UNICORN	.748	.801	.800	.943	.842	.734	.684	.743	.898	.759	.712	.717	.683
PathMoE	.736	.784	.787	.928	.795	.652	.621	.772	.818	.664	.628	.696	.607
Learned ensembles of the experts (iv)													
validation-weighted mean	.764	.819	.805	.916	.804	.721	.674	.723	.844	.667	.753	.747	.673
convex stacking	.764	.822	.806	.917	.805	.721	.668	.729	.843	.670	.714	.728	.617
Bayesian model averaging	.704	.800	.786	.896	.763	.652	.566	.703	.844	.659	.716	.724	.698
ridge stacking	.766	.814	.800	.933	.829	.716	.644	.714	.892	.677	.674	.731	.551
ridge stacking, cross-task inputs	.784	.816	.800	.932	.850	.691	.711	.721	.895	.717	–	–	–
elastic-net stacking	.769	.812	.805	.935	.841	.710	.614	.735	.879	.675	.635	.740	.534
gradient-boosted stacking	.738	.788	.782	.923	.813	.715	.682	.672	.850	.577	.623	.574	.573
ARISE	**.805**	**.836**	**.816**	**.948**	**.860**	**.771**	**.748**	**.807**	**.917**	**.801**	**.763**	**.778**	**.732**

bers at the external site. Given that pool, a conventional selection procedure recovers most of the result, and the agent-designed search adds little.

5.2 Performance with complete and incomplete stain panels

A fixed pooling rule applies to whichever stains a patient has, whereas an ARISE program can put its weight on members that do not need the missing stain (§3.1). The margin of Table 2 could therefore be a property of the patients with missing stains alone. To test this, we split the evaluated patients of every cell into those with every stain and those lacking at least one, and score ARISE and the strongest fixed pooling rule on each group (Fig. 2). On patients with every stain, ARISE is ahead in 12 of the 13 cells: by 0.010 to 0.047 on CureGN, 0.027 to 0.069 on CMUH, and 0.010 and 0.030 on the two HANCOCK tasks. The exception is DLBCL-Morph, where the median rule leads by

Table 3: Where the margin comes from, on the kidney cohorts (evaluation AUROC). First row: the strongest comparison method of Table 2. Then the program each search procedure selects on the pool without the cross-task members (the base pool of §3.1 plus the learned combiners) and on the pool with them; without the cross-task members only greedy selection and the agent-designed search. Stochastic procedures are the median over five search seeds; the committed program is the single run of Table 2.

	CureGN (internal)					CMUH (external)				
Pool / search	M	E	S	T	C	M	E	S	T	C
strongest comparison method (Table 2)	.787	.816	.806	.943	.842	.740	.684	.778	.898	.759
Pool without the cross-task members (base pool + learned combiners)										
greedy forward selection	.795	.833	.815	.946	.852	.750	.721	.788	.909	.759
agent-designed search	.798	.833	.816	.947	.851	.757	.728	.793	.915	.777
Pool with the agent's cross-task members										
greedy forward selection	.803	.839	.816	.946	.852	.767	.740	.800	.909	.784
bagged greedy selection	.804	**.841**	**.818**	.947	.855	.766	.742	.796	.912	.779
random search, equal evaluations	.799	.836	.814	.944	.851	.760	.729	.787	.909	.764
agent-designed search	**.805**	**.841**	.816	**.948**	.858	.770	**.748**	.803	**.918**	.795
committed program (ARISE)	**.805**	.836	.816	**.948**	**.860**	**.771**	**.748**	**.807**	.917	**.801**

0.006 in the 112 patients with all six stains. For patients lacking a stain, the gap is wider than in the complete group in seven of the ten cells that can be scored, and widest for HANCOCK recurrence (0.824 against 0.693) and DLBCL-Morph (0.858 against 0.768). The advantage therefore does not depend on missing stains, except for DLBCL-Morph, where the whole margin comes from the 83 patients with an incomplete panel.

6 Conclusion

We introduced ARISE, which treats multi-stain integration as a task-specific program rather than a one-time architecture choice. The stain-specific experts stay fixed; an LLM agent evolves the candidate pool, the objective, and the search; and a numerical inner loop fits the resulting program. Across four cohorts, three organs and two task types, these programs outperform single-stain models, conventional ensembles and published fusion architectures, and the margin holds at an external site. The strictest fixed rule varies across tasks and sites, so it must be designed for each setting rather than chosen in advance. Ablations attribute most of the advantage to the candidate pool, including the members the agent added, and only a small part to the search. The agent's contribution is to design what is searched; the fitting itself can be conventional. These results support treating multi-stain integration as a task-specific design problem, with ARISE as one procedure for building such programs across organs, stain panels, and prediction tasks. Future work could test whether a scripted proposer, under the same revision budget, reaches comparable designs and evaluate the procedure on larger external cohorts.

AI use statement

Generative AI is part of the method: the outer loop of ARISE is driven by an LLM agent (Claude Opus 5), which writes and edits the specification of the candidate pool, the objective, and the search settings, and reads the diagnostic reports, as described in §3.3; every program it produced was scored by the fixed numerical pipeline, and the frozen programs contain no LLM. We also used LLM coding assistants to write and refactor experiment code and figure scripts, and LLM writing assistants to edit the prose of this paper. All LLM-written code was reviewed and tested by the authors; all numbers in the paper were recomputed from stored predictions by scripts the authors checked; and the authors take responsibility for the final content, claims, and artifacts.

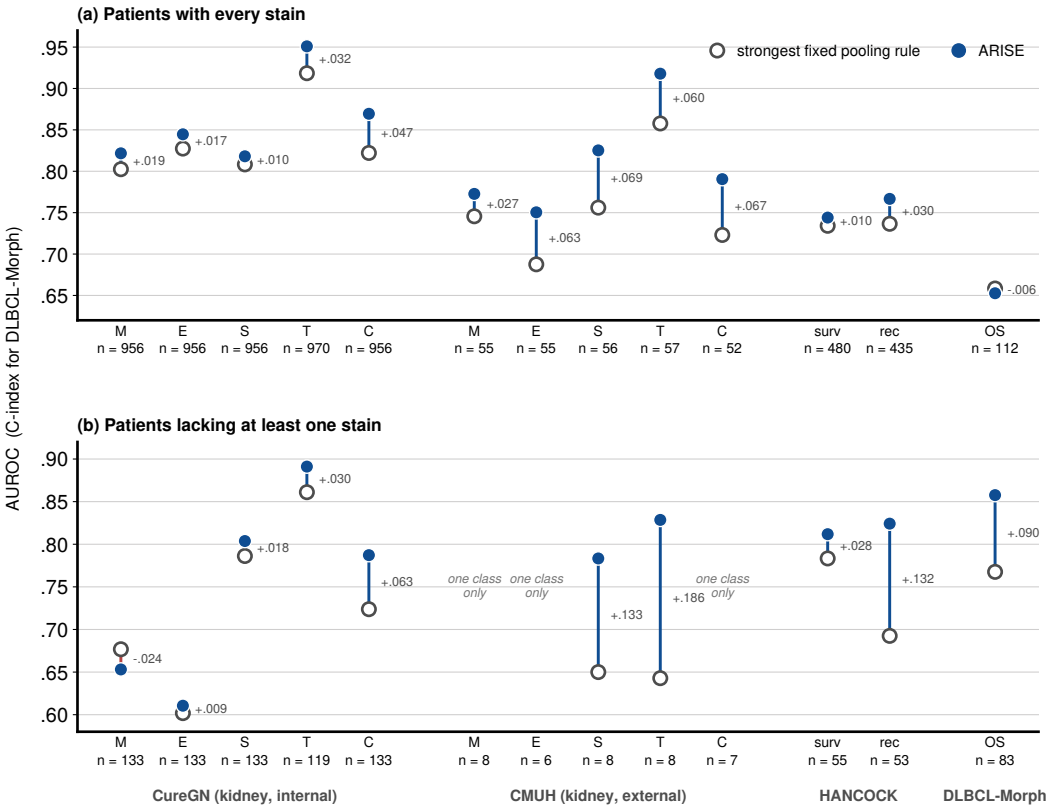

Figure 2: ARISE (filled) and the strongest fixed pooling rule (open) on (a) patients with every stain and (b) patients lacking at least one stain, in every task–cohort cell (AUROC; C-index for DLBCL-Morph; mean over the five splits; n is the group size). CMUH M, E and C have one class among their incomplete panels and cannot be scored there.

Ethics Statement

All data are de-identified human biopsy slides and clinical labels. CureGN and CMUH data were used under their respective data-use agreements and institutional review board approvals and are not redistributed; HANCOCK and DLBCL-Morph are public cohorts used under their release terms. The programs are research prototypes and are not intended for clinical use.

Reproducibility Statement

The stain-specific experts, comparison methods, and evaluation protocol are specified in §3 and §4; the ARISE member families, objective, and search procedure are described in §3.1–§3.3; and the results for all comparison methods are reported in Table 2. HANCOCK and DLBCL-Morph are public cohorts, and their case lists, labels, and exclusions follow the source publications and official repositories. Code for the experts, candidate-pool construction, search, and evaluation, together with the frozen programs and the prediction outputs needed to reproduce the reported results where permitted by the relevant data-use agreements, will be released upon publication.

References

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE TPAMI*, 41(2):423–443, 2019. doi: 10.1109/TPAMI.2018.2798607.

Jiří Borovec, Jan Kybic, Ignacio Arganda-Carreras, et al. ANHIR: Automatic non-rigid histological image registration challenge. *IEEE Transactions on Medical Imaging*, 39(10):3042–3052, 2020. doi: 10.1109/TMI.2020.2986331.

Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. Ensemble selection from libraries of models. In *ICML*, 2004. doi: 10.1145/1015330.1015432.

Rich Caruana, Art Munson, and Alexandru Niculescu-Mizil. Getting the most out of ensemble selection. In *ICDM*, 2006. doi: 10.1109/ICDM.2006.76.

Cheng Cheng, Bin Li, Jie Li, Yiqin Wang, Han Xiao, Xingji Lian, Lizhi Chen, Junxian Wang, Haiyan Wang, Shuguang Qin, Li Yu, Tingbo Wu, Sui Peng, Weiping Tan, Qing Ye, Wei Chen, and Xiaoyun Jiang. Multi-stain deep learning prediction model of treatment response in lupus nephritis based on renal histopathology. *Kidney International*, 107(4):714–727, 2025. doi: 10.1016/j.kint.2024.12.007.

Shyam Sundar Debsarkar and V. B. Surya Prasath. A multi-expert deep learning framework with LLM-guided arbitration for multimodal histopathology prediction. *Computerized Medical Imaging and Graphics*, 128:102704, 2026. doi: 10.1016/j.compmedimag.2026.102704.

Marion Dörrich, Matthias Balk, Tatjana Heusinger, et al. A multimodal dataset for precision oncology in head and neck cancer. *Nature Communications*, 16:7163, 2025. doi: 10.1038/s41467-025-62386-6. Data: The Cancer Imaging Archive, `https://www.cancerimagingarchive.net/collection/hancock/`.

Nick Erickson, Jonas Mueller, Alexander Shirkov, et al. AutoGluon-Tabular: Robust and accurate AutoML for structured data. *arXiv preprint arXiv:2003.06505*, 2020.

Matthias Feurer, Aaron Klein, Katharina Eggensperger, et al. Efficient and robust automated machine learning. In *NeurIPS*, 2015. URL `https://proceedings.neurips.cc/paper/2015/file/11d0e6287202fced83f79975ec59a3a6-Paper.pdf`.

Sebastian Foersch, Christina Glasner, Ann-Christin Woerl, et al. Multistain deep learning for prediction of prognosis and therapy response in colorectal cancer. *Nature Medicine*, 29:430–439, 2023. doi: 10.1038/s41591-022-02134-1.

Amaya Gallagher-Syed, Luca Rossi, Felice Rivellese, Costantino Pitzalis, Myles Lewis, Michael Barnes, and Gregory Slabaugh. Multi-stain self-attention graph multiple instance learning pipeline for histopathology whole slide images. In *British Machine Vision Conference (BMVC)*, 2023. URL `https://proceedings.bmvc2023.org/789/`.

Amaya Gallagher-Syed et al. BioX-CPath: Biologically-driven explainable diagnostics for multistain IHC computational pathology. In *CVPR*, pp. 10372–10383, 2025. URL `https://arxiv.org/abs/2503.20880`.

Christine P. Hans, Dennis D. Weisenburger, Timothy C. Greiner, et al. Confirmation of the molecular classification of diffuse large B-cell lymphoma by immunohistochemistry using a tissue microarray. *Blood*, 103(1):275–282, 2004. doi: 10.1182/blood-2003-05-1545.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *ICML*, 2018.

Guillaume Jaume, Anurag Vaidya, Andrew Zhang, et al. Multistain pretraining for slide representation learning in pathology. In *ECCV*, 2024. URL `https://arxiv.org/abs/2408.02859`.

Zhengyao Jiang, Dominik Schmidt, Dhruv Srikanth, et al. AIDE: AI-driven exploration in the space of code. *arXiv preprint arXiv:2502.13138*, 2025.

Josef Kittler, Mohamad Hatef, Robert P. W. Duin, and Jiri Matas. On combining classifiers. *IEEE TPAMI*, 20(3):226–239, 1998. doi: 10.1109/34.667881.

Valentin Koch et al. UNICORN: a deep learning model for integrating multi-stain data in histopathology. *npj Digital Medicine*, 9:480, 2026. doi: 10.1038/s41746-026-02829-6. arXiv:2409.17775.

James Kotary, Vincenzo Di Vito, and Ferdinando Fioretto. Differentiable model selection for ensemble learning. In *IJCAI*, 2023. URL https://www.ijcai.org/proceedings/2023/217.

Robert Tjarko Lange, Yuki Imajuku, and Edoardo Cetin. ShinkaEvolve: Towards open-ended and sample-efficient program evolution. In *ICLR*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/7886b9bafe76c52fd568db10ff9772df-Abstract-Conference.html. arXiv:2509.19349.

Johannes Lotz, Nick Weiss, Jeroen van der Laak, and Stefan Heldmann. Comparison of consecutive and restained sections for image registration in histopathology. *Journal of Medical Imaging*, 10 (6):067501, 2023. doi: 10.1117/1.JMI.10.6.067501.

Yuxing Lu, Yicheng Chen, Shanchan Wu, and Sercan Ö. Arık. Procedural graphs: Self-evolving execution structures for LLM agents. *arXiv preprint arXiv:2609.09153*, 2026.

Laura H. Mariani, Andrew S. Bomback, Pietro A. Canetta, et al. CureGN study rationale, design, and methods: establishing a large prospective observational study of glomerular disease. *American Journal of Kidney Diseases*, 73(2):218–229, 2019. doi: 10.1053/j.ajkd.2018.07.020.

Ramin Nakhli, Puria Azadi Moghadam, Haoyang Mi, et al. Sparse multi-modal graph transformer with shared-context processing for representation learning of giga-pixel images. In *CVPR*, 2023. URL https://arxiv.org/abs/2303.00865.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, et al. AlphaEvolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

PathMoE. PathMoE: AI-enabled prediction of therapy response and prognosis from multistain biopsies in breast cancer. Code repository, https://github.com/yyyhd/PathMoE, n.d. Accessed September 25, 2026.

Sebastian Pineda Arango, Maciej Janowski, Lennart Purucker, Arber Zela, Frank Hutter, and Josif Grabocka. Regularized neural ensemblers. In *Proceedings of the Fourth International Conference on Automated Machine Learning*, volume 293 of *Proceedings of Machine Learning Research*, pp. 8/1–33. PMLR, 2025. URL https://proceedings.mlr.press/v293/arango25a.html.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, et al. Mathematical discoveries from program search with large language models. *Nature*, 625:468–475, 2024. doi: 10.1038/s41586-023-06924-6.

Daan Schouten, Giulia Nicoletti, Bas Dille, Catherine Chia, Pierpaolo Vendittelli, Megan Schuurmans, et al. Navigating the landscape of multimodal AI in medicine: A scoping review on technical challenges and clinical applications. *Medical Image Analysis*, 105:103621, 2025. doi: 10.1016/j.media.2025.103621.

Asankhaya Sharma. OpenEvolve: an open-source evolutionary coding agent. https://github.com/algorithmicsuperintelligence/openevolve, 2025.

Hongyuan Su, Yu Zheng, and Yong Li. ContextEvolve: Multi-agent context compression for systems code optimization. *arXiv preprint arXiv:2602.02597*, 2026.

Hernán Trimarchi, Jonathan Barratt, Daniel C. Cattran, et al. Oxford classification of IgA nephropathy 2016: an update from the IgA nephropathy classification working group. *Kidney International*, 91(5):1014–1021, 2017.

Damir Vrabac, Akshay Smit, Rebecca Rojansky, et al. DLBCL-Morph: Morphological features computed using deep learning for an annotated digital DLBCL image set. *Scientific Data*, 8:135, 2021. doi: 10.1038/s41597-021-00915-w.

Xiyue Wang, Junhan Zhao, Eliana Marostica, et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, 634:970–978, 2024. doi: 10.1038/s41586-024-07894-z.

Philippe Weitz, Masi Valkonen, Leslie Solorzano, et al. The ACROBAT 2022 challenge: Automatic registration of breast cancer tissue. *Medical Image Analysis*, 97:103257, 2024. doi: 10.1016/j.media.2024.103257.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992. doi: 10.1016/S0893-6080(05)80023-1.

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630:181–188, 2024. doi: 10.1038/s41586-024-07441-w.

Eric Zimmermann, Eugene Vorontsov, Julian Viret, et al. Virchow2: Scaling self-supervised mixed-magnification models in pathology. *arXiv preprint arXiv:2408.00738*, 2024.