AGENTIC SELECTION WITHOUT FEEDBACK: AN LLM AGENT DESIGNS CANCER PANELS BY KNOWING WHAT NOT TO SEQUENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Designing the target panel for a cell-free DNA cancer assay is a budgeted selection problem with no execution feedback. The panel fixes which genomic regions are sequenced, it dominates development cost, and its only test is a clinical study that arrives after the design is frozen. Current practice mines labeled case-control cohorts, so panels inherit the population they were trained on and can degrade in populations with a shifted cfDNA background, such as patients under with chronic liver disease, where labeled positives are scarce. We ask whether selection from prior information alone can compete with supervised design. We present an LLM agent that constructs a multi-cancer methylation panel for a chronic liver disease population from three inputs, the cancer literature, a reference cell-type methylation atlas, and unlabeled plasma from cancer-free donors in the target population. No tumor sample and no detector response is available at any design stage. On a titration benchmark with 700 background controls and a single fixed detector, the frozen panel reaches 0.7323 AUC against 0.7050 for the strongest commercial panel at matched sequencing cost, and exceeds all comparators under paired tests. Ablations suggests the gain is a function of the panel as a whole and not any specific markers. The agent's contribution is the systematic exclusion of weak regions, which requires prior knowledge but no outcome data. The agent were highly reproducible with any two of five independent runs share 86% to 92% of their selected CpGs. The agent provide unique region when substituting the target population shifts selection far beyond run-to-run variation (Jaccard index 0.73 vs. 0.92) supporting on demand per population panel design. These results suggest that, for budgeted decisions made without feedback, an agent's value lies in reliably rejecting poor candidates rather than in identifying the best ones.

1 INTRODUCTION

Circulating cell-free DNA (cfDNA) contains fragments shed by cells throughout the body, where disease-specific DNA fragments introduce a detectable signal against a background. Consequently, cancer detection relies on isolating tumor-derived DNA from this mixture. However, in early-stage disease, tumor fraction often drops well below 1% of total circulating cfDNA, making detection hardest precisely when intervention is most beneficial (Bettegowda et al., 2014). In practice, to recover this rare signal cfDNA are sampled at a restricted set of genomic regions, using a target enrichment panel, at high sequencing depth often under a strict genomic footprint budget to make the liquid biopsy commercially viable. Abedini & Damangir (2026) showed that in cancer tumor cfDNA lie on a manifold with concentrated deviation from background, and that the detection performance is bound by the design of such panel.

Historically designing this targeted panel demanded the largest effort and resource in cancer liquid biopsies. Panels are selected by comparing labeled positives against controls (Xu et al., 2017; Liu et al., 2020), so a panel inherits the population it was trained on. However, target populations present distinct cfDNA backgrounds; for instance, patients under liver cancer surveillance exhibit elevated liver-derived background DNA compared to healthy donors (Moss et al., 2018; Lehmann-Werman et al., 2018; Foda et al., 2023). Adapting to this shift using supervised techniques demands labeled positives within specific population, which are not only scarce in early stages of the disease (Lin

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.

et al., 2022) but prohibitively expensive. As a result, panels are designed on general populations and reused across different risk groups. Importantly as many problems in medicine, the feedback loop consist of a clinical study which means the selection under strict budget should happen without immediate execution feedback.

This represents a broader class of budgeted selection problems under distribution shift, where a resource budget must be committed without execution feedback using only indirect information such as domain knowledge, literature, potentially a signal reference atlas, and unlabeled target data. Sensor placement (Krause et al., 2008), assay design and the choice of which experiments to fund all have this shape. The lack of immediate feedback has made this class of problem challenging for human and agents alike. The agent literature is split on whether LLM agents use experimental feedback at all. Gupta et al. (2025) find no sensitivity to it, since permuting outcome labels leaves performance unchanged. Wainrib et al. (2026) find the opposite once the model is strong enough, reporting a measurable gain in discoveries from real feedback that disappears under permuted labels. Work that asks agents to judge a design before any feedback is pessimistic: ArchEval (Wang et al., 2026) reports median self-assessment errors of 81 to 99% on submitted designs.

We ask the complementary question. When the loop cannot exist, what is prior-driven selection worth, and what is it made of? Can an agent reliably explore possibilities and commit to a solution?

We design and evaluate an LLM agent for panel design. We ask the agent to find a multi-cancer panel to be used for a population with chronic liver disease. The agent-designed panel achieves an AUC of 0.7323 compared to 0.7050 for a hand tuned commercial panel of the same sequencing cost. Analysis reveals that this performance gain is distributed across the entire panel, operating reliably across independent runs and adapting dynamically to the target population background. We perform a battery of ablation studies to isolate the reason the agent outperform hand tuned model and demonstrating that the agent's primary mechanism is the systematic exclusion of loosing regions.

The contributions of this paper are as follows.

1. We formalize a budgeted selection problem with no positive samples, under a background shift and propose an LLM agent that integrates domain text, a cell-type atlas, and unlabeled target plasma.
2. On a controlled titration benchmark, the frozen agent panel beats a hand tuned commercial panel, a published academic panel, each information source alone, a deterministic baseline, and random content.
3. We demonstrate that the agent's gain is a distributed property of the entire panel, driven by the systematic filtering of losing hypothesis rather than the isolated identification of high-signal loci.
4. We show that independent agent executions are highly stable and dynamically adapt panel content to specified population backgrounds, providing a practical tool for rapid target-tailored panel design.

2 Related work

Existing agentic systems for experimental design, such as BioDiscoveryAgent (Roohani et al., 2025) and LLAMBO (Liu et al., 2024), select sequential interventions by closing the loop over previous experimental outcomes. However, the extent to which language models truly leverage this empirical feedback remains contested. Gupta et al. (2025) find that permuting outcome labels leaves performance unchanged, suggesting models rely predominantly on priors and proposing that classical methods handle state updates. Conversely, Wainrib et al. (2026) report substantial performance gains from real feedback given sufficient model capacity, while benchmark frameworks like PerturbTrace (Yu et al., 2026) and AssayBench-Loop (Edwards et al., 2026) attempt to disentangle raw discovery capacity from active feedback utilization. Critically, these paradigms all presuppose an active closed loop; we instead address the zero-feedback setting, quantifying the foundational performance ceiling achievable purely through prior-driven synthesis.

In the domain of automated feature selection, approaches such as LLM-Select (Jeong et al., 2025) choose features directly from candidate text labels, matching standard baselines like LASSO. In cancer liquid biopsy setting, however, features cannot simply be filtered from a pre-existing tabular matrix, they must be actively constructed from literature claims, subject to a strict genomic footprint

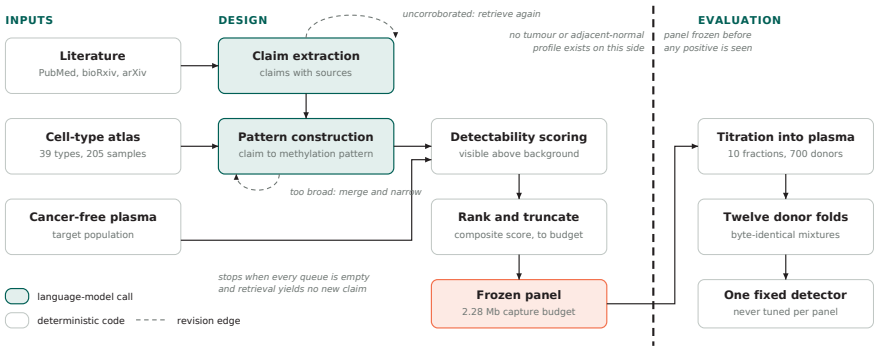


Figure 1: The design loop and the evaluation that follows it. Shaded stages are language-model calls; unshaded stages are code. A vertical rule marks the boundary, and no tumour or adjacent-normal profile exists to its left.

budget, and evaluated against the benchmark of available hand tuned commercial assays. This aligns broadly with open-loop execution in agentic control (Sun et al., 2023), where plans are constructed without environmental feedback. While classical subset selection framework, such as sensor placement (Krause et al., 2008) and LLM-guided optimization (Romera-Paredes et al., 2024; Yang et al., 2024) assume that the target objective can be continually evaluated during search, our task renders the objective unobservable during selection. The agent must therefore construct candidates purely from domain text and screen them against unlabeled background data.

Conventional cfDNA assay design typically mines differentially methylated regions from labeled tissue samples and optimizes candidates using case-control cohorts (Xu et al., 2017; Liu et al., 2020; Kandimalla et al., 2021). We bypass this supervised mining step entirely, relying instead on reference cell-type atlases established in the literature (Sun et al., 2015; Moss et al., 2018; Loyfer et al., 2023). While agents have been applied downstream of target identification, such as PrimeGen (Wang et al., 2025) for multiplex primer optimization, or to nominate general cfRNA gene panels tested in labeled cohorts (Gaudio et al., 2026), our approach differs fundamentally. We select continuous genomic intervals under a hard footprint constraint, condition dynamically on specified clinical populations, and freeze the assay before observing a single labeled positive sample.

Finally, our benchmark design follows standard clinical protocol by evaluating assays using in-silico titration against controlled tumor fractions (Li et al., 2018; Jamshidi et al., 2022; Bredno et al., 2021). Evaluating performance as a function of tumor fraction reflects clinical limit-of-detection standards in multi-cancer early detection, where circulating tumor content serves as a stronger predictor of classifier accuracy than traditional clinical stage or tumor origin (Jamshidi et al., 2022).

3 The panel-design agent

Problem Formulation. Each candidate genomic region is one feature. Given a candidate set $\mathcal{C}$ and a budget b of total captured sequence, out of a 3 Gigabase pairs genome, the agent returns a ranked subset $S \subseteq \mathcal{C}$ whose total length is at most b. The objective, how well tumour signal can be detected in the deployment population, is never observed during agents work.

Inputs. The agent uses three inputs, each addressing a distinct property of a candidate region. First, structured literature knowledge captures published findings linking genes, pathways, cell types, or tissue locations related to cancer; all source literature citing the evaluation tissue cohorts is strictly filtered out prior to execution to prevent target leakage. Second, a reference cell-type atlas across 39 sorted cell types from 205 healthy tissue samples (Loyfer et al., 2023) determines whether a region reflects tissue-specific identity or stochastic variation. Third, unlabeled population plasma from cancer-free donors within the target population (e.g., patients with chronic liver disease) establishes background, allowing the agent to design against population-specific baseline shift and retain only features relevant for that background.

Execution Loop. The agent maintains three queues holding unprocessed, corroboration-pending, and overly broad claims. Three tasks move claims between them. During claim extraction and

retrieval, the agent extracts biological assertions with source citations from literature searches for uncorroborated claims. Next, pattern construction integrates claims, genomic annotations, pathway databases, and the reference atlas to convert biological assertions into specific molecular signatures using the atlas data. Broad claims that fail to map to localized regions are returned to the queue for aggregation and narrowing. Finally, detectability validation validates candidate loci against population plasma, and scoring them based on their detectability in the background. Claims iterate through this pipeline until they yield a validated pattern, are rejected, or exhaust the retrieval budget. Execution stops when the literature stops yielding new claims. Validated regions are then ranked by their composite scores, and truncated at the budget b.

Zero-Feedback Boundary. A defining feature of this architecture is the complete absence of execution feedback. cfDNA from patient with chronic disease who have cancer (i.e. positives) were not available to the agent. The agent was also blind to the evaluation strategy (Figure 1; details in Appendix B).

Panel Budget. Panel footprint sets the cost of an assay, since probe synthesis and read depth scale with captured bases. TWISTALLIANCE spans 2.36 Mb and as such we set a budget of 2.3 Mb for the agents.

Computational Cost and Initialization. A complete design run requires approximately 41 hours on a single `c8i.12xlarge` instance ($93 compute cost) using Claude Opus 4.6 model supplemented by literature retrieval (PubMed, bioRxiv, arXiv), a Python execution environment, and standard genomic annotation toolkits. Operators set four fixed global parameters prior to execution: the target malignancy scope (broad multi-cancer), the target population background, available reference databases, and the budget b (matched to the sequencing footprint of commercial comparators).

Experimental Suite. We evaluate fourteen total panel configurations, consisting of seven agent-generated panels and seven comparative baselines (Table 1). The agent suite comprises five repeat executions of a reference configuration, a sixth execution substituting targeting alcoholic liver disease cohort, and a seventh execution ablating the reference cell-type atlas.

4 EVALUATION

Synthetic Ground Truth. Public whole-genome bisulfite tumor and adjacent-normal methylomes are titrated into cancer-free plasma from 700 chronic liver disease donors from six cohorts at ten geometrically spaced fractions (0.005% to 10%). To ensure that a detector cannot succeed by simply recognizing normal tissue markers, adjacent-tissue mixtures serve as negative controls. No cfDNA in the evaluation cohort originates from a patient used during the execution of the agent. Genomic intervals are padded by 30 bp, merged if within 500 bp, and tiled into 500 bp windows; within each window, fragments are assigned to five equal-width bins based on their observed CpG methylation fraction.

Detector. To isolate panel feature quality from downstream modeling choices, we employ a single detector fixed beforehand applied uniformly across all panels, folds, and titration fractions. The detector establishes a negative baseline with a pseudocount of 1, floors per-cell variance at the multinomial shot-noise limit, estimates cell-level mixture responses via least squares, and combines cells exceeding three standard errors into an inverse-variance score. Fitting and evaluation utilize donor-disjoint splits. Tissue and plasma donors are partitioned prior to mixture generation, weights are fitted strictly on training mixtures, and validation instances are scored using frozen weights.

Evaluation Endpoints. We evaluate panels across two complementary metrics. The primary endpoint is AUC, discriminating tumor mixtures from fraction-matched adjacent-normal mixtures, averaged across informative fractions ($\geq 0.063\%$). The secondary endpoint is the standardized partial AUC above 90% specificity on pooled negatives. Because partial AUC isolates the top decile of negative scores, it becomes informative primarily above 0.8% tumor fraction.

Sequencing Depth and Cross-Validation. Mixtures are simulated at $300\times$ whole-genome coverage; at this depth sampling noise is negligible, so differences between panels reflect their content. Statistical inference uses a 12-fold, cohort-stratified donor cross-validation strategy. Panels are evaluated on byte-identical mixtures within each fold to enable paired sign-flip significance testing (min-

Table 1: The comparator set. Each entry answers a different question about whether the design step is doing work. Internal identifiers are mapped in Appendix A.

Panel	Construction	Question it settles
AGENT×7	The agent-designed panel, no cancer-positive specimen at any design stage	The paper's claim
TWISTALLIANCE	A commercial pan-cancer methylation capture panel covering 31 cancer types	Can positive-free design beat a tuned product built with positives?
EPIPANGIDX	A commercial/academic panel for gastrointestinal cancers	Can it beat a published tumor tissue driven design?
ASSOCIATION-ONLY	Disease-gene association ranking from a public target-association platform, with promoters and gene bodies added in rank order until the budget is filled	Is knowing what is relevant enough?
DETECTABILITY-ONLY	Intended-population plasma only: after coverage and uniformity filtering, regions with low background variance whose baseline methylation sits opposite to the expected cancer direction	Is knowing what is observable enough?
Random ×3	Random draws from the Twist Human Methylome Panel, a 123 Mb commercial capture design covering 3.98M CpG sites, taken at a matched capture size	Does selection beat spending the budget arbitrarily?

imum attainable $p \leq 0.0005$), with step-down family-wise error correction. Donor-level confidence intervals are derived via percentile bootstrapping over tissue donors.

Comparators. We benchmark the agent against seven comparator panels (Table 1), including independent non-agent baselines (ASSOCIATION-ONLY, DETECTABILITY-ONLY) and cost-matched random region draws. Plasma donors seen during design are excluded from evaluation for every arm, leaving 700 background controls.

5 RESULTS

Four primary questions organize the results: Can a panel designed without positive cancer specimens compete with panels engineered from labeled cohorts? Where inside the panel does its performance advantage reside, and what selection mechanism is at work? Is the design process stable across runs and responsive to the target population? And finally, does the agent add genuine reasoning value beyond what its individual information sources or a deterministic merge can achieve? Every comparison is paired across donor-balanced folds on byte-identical mixtures and evaluated using exact sign-flip tests.

5.1 THE AGENT PANEL OUTPERFORMED HUMAN ENGINEERED PANEL

Table 2 reports the prespecified endpoints across all fourteen panels (the seven agent panels of Section 3 and the seven comparators of Table 1). The reference AGENT panel reaches a primary AUC of 0.7323, compared to 0.7050 for the TWISTALLIANCE panel, which is the strongest human engineered panel. Every agent variant outperforms every comparator across both primary AUC and secondary partial AUC endpoints, including the ablation arm built without the cell-type atlas. Testing the gent panel against each of the seven comparators on both endpoints gives fourteen paired comparisons, and all fourteen resolve under the exact sign-flip test, with a worst p-value of 0.0161. What makes this result compelling is what the baselines possessed during design. Both TWISTALLIANCE and EPIPANGIDX panels were engineered using labeled case-control cohorts and had full access to public cell-type atlases as well as cfDNA data. They are nevertheless surpassed by a panel whose design observed neither a positive specimen nor a downstream detector response.

Cost-matched random content (0.6598 AUC) provides a baseline performance anchor. Selection can therefore move the endpoint across 0.16 above chance, and the agent covers $+0.0726$ of that range. Under a donor-level bootstrap on the one-score-per-tissue AUC, the AGENT panel had a level of detection (LoD) of 0.34% tumour fraction.

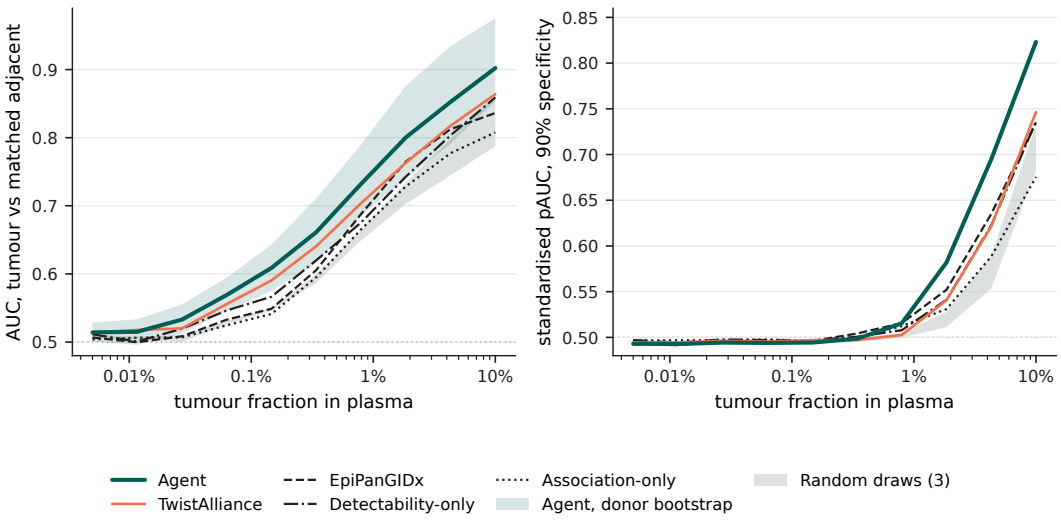


Figure 2: Endpoints across titration fractions on 700 background controls. The reference panel includes a donor-level bootstrap interval; the shaded envelope spans three cost-matched random draws (full set in Appendix C).

Table 2: Frozen endpoints for all fourteen panels on 700 background controls. Footprint (Mb) reflects manufacturing and sequencing cost. Δ is the gap to the best non-agent panel on that endpoint, TWISTALLIANCE on primary AUC and EPIPANGIDX on secondary pAUC. Differences within the agent family are assessed against the detectable-effect floors of Section 5.3.

				primary AUC			secondary pAUC		
Panel	Mb	CpGs	CpG/kb	AUC	Δ	p	pAUC	Δ	p
AGENT-ALD	2.72	213,052	78.2	**0.7406**	+0.0356	—	**0.5920**	+0.0304	—
agent run 3	2.37	226,830	95.7	0.7326	+0.0276	—	0.5852	+0.0236	—
AGENT	2.28	233,081	102.4	0.7323	+0.0273	—	0.5860	+0.0244	—
agent run 5	2.46	223,147	90.8	0.7307	+0.0257	—	0.5839	+0.0223	—
agent run 4	2.42	225,301	93.2	0.7299	+0.0249	—	0.5861	+0.0245	—
agent run 2	2.42	224,278	92.7	0.7290	+0.0240	—	0.5838	+0.0222	—
AGENT-NOATLAS	3.03	204,704	67.6	0.7248	+0.0198	—	0.5853	+0.0237	—
TWISTALLIANCE	2.36	192,308	81.3	0.7050	0	0.0020	0.5572	−0.0044	$\leq$0.0005
DETECTABILITY-ONLY	2.00	264,190	132.1	0.6877	−0.0173	0.0024	0.5574	−0.0042	0.0010
EPIPANGIDX	0.85	65,045	76.2	0.6841	−0.0209	$\leq$0.0005	0.5616	0	0.0161
random 3	1.95	56,408	29.0	0.6761	−0.0289	0.0029	0.5470	−0.0146	0.0015
ASSOCIATION-ONLY	1.39	105,952	76.1	0.6628	−0.0422	0.0010	0.5429	−0.0187	$\leq$0.0005
random 2	1.94	57,532	29.7	0.6524	−0.0526	$\leq$0.0005	0.5475	−0.0141	0.0010
random 1	1.93	56,211	29.1	0.6508	−0.0542	$\leq$0.0005	0.5332	−0.0284	$\leq$0.0005

5.2 THE ADVANTAGE IS SPREAD ACROSS THE PANEL, AND THE MECHANISM IS EXCLUSION

A panel can achieve superior performance either by isolating a small set of exceptional markers, or by systematically excluding weak ones. Evaluating sub-panel slices shows the agent operates via the latter mechanism. Scored standalone across the same twelve folds (Table 3, Figure 3), the 4,019 regions private to the reach 0.7342 AUC, while the 3,755 shared regions reach 0.7294 AUC (versus 0.7323 for the full panel). Ten random half-panels average 0.7336 AUC (0.7137–0.7447), with private content showing no margin over random halves (+0.0006, $p = 0.93$). Crucially, every sub-slice, including the lowest-scoring random half, outperforms the full commercial product (0.7050 AUC). This means that the agent did not rely on a few privileged markers. It constructed a uniformly high-quality feature pool.

Ablations reveal how this selection works at the region level. Removing the cell-type atlas admits 5,896 replacement regions that score 0.6268 standalone (versus 0.7244 for random content from the same panel), whereas the 1,813 atlas-contributed regions achieve 0.7461 AUC (+0.0154 over random draws, $p = 0.040$). Substituting the population background admits 3,746 novel regions scoring 0.6188 standalone (versus 0.7435). Both inputs change the quality of the regions that enter the panel.

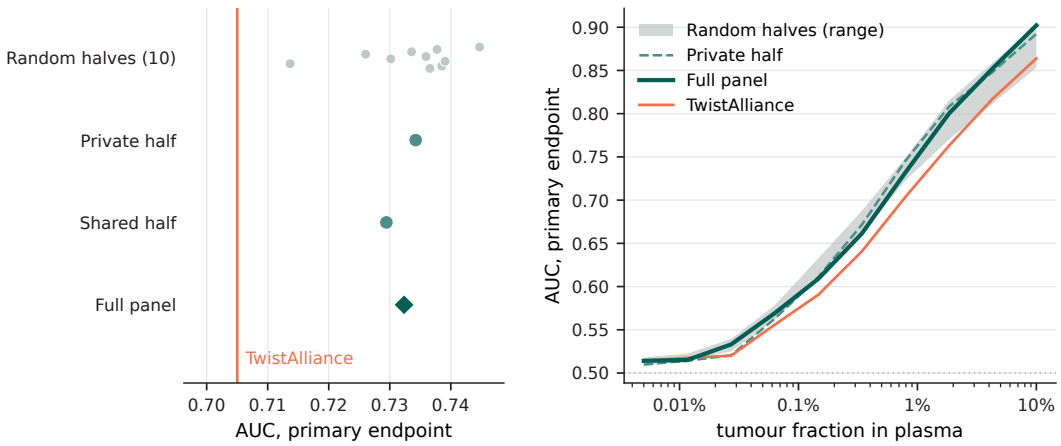


Figure 3: Sub-panel performance. Left: Standalone primary AUC for ten random halves, private vs. shared sub-slices, and full panel against TWISTALLIANCE (0.7050). Right: Per-fraction AUC tracking the random half envelope.

Table 3: Standalone performance of panel sub-slices and differential regions between ablation arms across twelve evaluation folds. Upper block: Reference panel slices. Lower block: Differentially selected regions scored against same-size random draws.

Region set	Regions	Standalone	Same-size random	Diff	p
Private to the agent	4,019 / 7,774	0.7342	0.7336 (0.7137–0.7447)	+0.0006	0.93
Shared with a comparator	3,755 / 7,774	0.7294	—	—	—
Full reference panel	7,774	0.7323	—	—	—
Regions					
the atlas contributed	1,813 / 7,774	0.7461	0.7307	+0.0154	0.040
the atlas-free run added	5,896 / 11,667	0.6268	0.7244	−0.0977	≤0.0005
the population input added	3,746 / 10,324	0.6188	0.7435	−0.1247	≤0.0005

At the panel level, removing the atlas costs 0.0061 AUC ($p = 0.057$). The candidate pool holds more usable signal than a 2 Mb budget can carry, so weaker admitted regions cost little at the endpoint. The cost concentrates where detection is hardest. 0.0201 AUC at 0.063% tumour fraction, reversing only above 1.85%. The agent's work is raising the quality floor of the pool so that weak regions never enter.

5.3 THE PROCESS IS STABLE, AND IT RESPONDS TO THE POPULATION

To be practical, an automated panel generator must be reproducible across independent runs on the same target and sensitive to changes in population background. Across five independent executions of the reference configuration, primary AUC spans a tight 0.7290 to 0.7326. The maximum pairwise gap between any two runs (0.0036) falls strictly below our powered detectable effect floors (0.0043–0.0086), confirming true performance stability (Figure 4). Content agreement mirrors score stability: pairwise CpG distance (1 − Jaccard Index) ranges between 0.078 and 0.143, with 72% of unique CpGs (97,398 out of 136,208) shared across all five runs.

In contrast to intra-family stability, substituting plasma from an alcoholic liver disease (ALD) cohort shifts selection distance to 0.272–0.284 from reference runs, a clean 0.129 gap above the family ceiling. Removing the atlas shifts it further, to 0.400–0.403.

The AGENT-ALD that targeted alcoholic liver disease population (a subpopulation of chronic liver disease) reaches 0.7406 AUC, +0.0097 above the family mean. Its added regions are weak on their own (Section 5.2), so the retained core contribute to the performance. The apparent margin sits entirely where detection is easy. Per fraction the AGENT-ALD is behind the reference at every fraction up to 0.341% (−0.0047 at 0.063%, −0.0008 at 0.147%, +0.0005 at 0.341%) and pulls ahead only from 0.794% upward (+0.0139, rising to +0.0178 at 10%). A gain confined to high

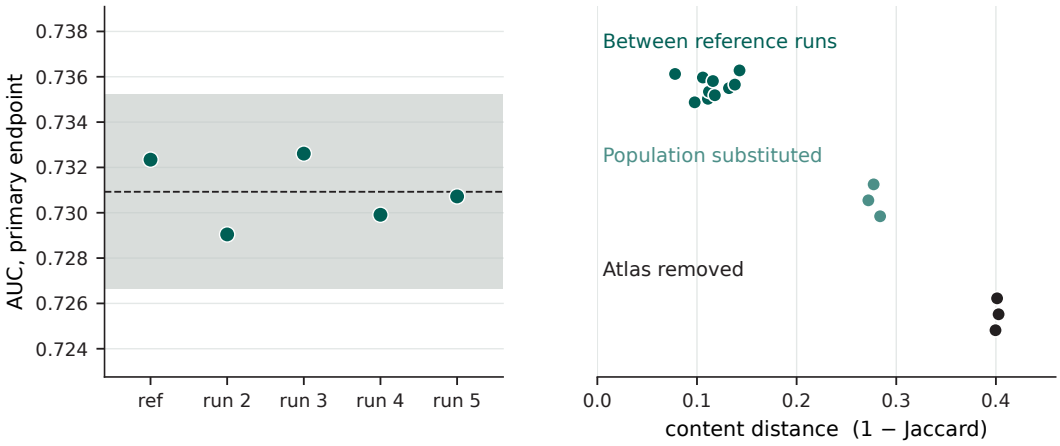


Figure 4: Stability and population response. Left: Primary AUC for five reference runs against detectable effect bounds. Right: Pairwise CpG content distance (1 − Jaccard) across arms.

Table 4: Single source performance and agent margin over deterministic combinations. Upper block: Lift over random mean (0.6598 AUC). Lower block: Paired agent margin (p-values from sign-flip tests).

	primary AUC		secondary pAUC	
	Δ	p	Δ	p
Gain over cost-matched random content				
ASSOCIATION-ONLY	+0.0030	0.888	+0.0003	0.963
DETECTABILITY-ONLY	+0.0279	0.084	+0.0148	0.036
AGENT	+0.0726	≤0.0005	+0.0434	≤0.0005
The agent's margin over				
ASSOCIATION-ONLY	+0.0695	0.0010	+0.0431	≤0.0005
DETECTABILITY-ONLY	+0.0447	0.0024	+0.0286	0.0010
DETERMINISTIC-MERGE	+0.0276	0.0205	+0.0245	0.0020

tumor fraction is not an exclusion effect. Selection responds to the population even across two related chronic liver disease population.

5.4 THE AGENT IS BETTER THAN THE SUM OF ITS PARTS

Could a simpler heuristic pipeline achieve the same result without an agentic loop? Table 4 answers this directly. Literature association alone (ASSOCIATION-ONLY) matches cost-matched random content (+0.0030 AUC, $p = 0.89$), while observability screening alone (DETECTABILITY-ONLY) recovers only one-third of the agent's gain over random draws (+0.0279 AUC, $p = 0.084$). A pre-registered deterministic merge (DETERMINISTIC-MERGE) combines both single-source panels by a fixed rule, ranks the result by CpG count, and truncates to agent's footprint at 2.28 Mb. It therefore competes at identical cost while carrying more material: 261,240 CpGs against the agent's 233,081, at 114.8 per kilobase against 102.4. It reaches 0.7047 AUC, tying the commercial baseline (TWISTALLIANCE) and capturing 62% of the agent's margin over random content. The agent nevertheless leads DETERMINISTIC-MERGE by +0.0276 AUC ($p = 0.021$) on primary AUC and +0.0245 ($p = 0.0020$) on secondary pAUC. Having both inputs is worth most of the gain, and the agent adds the remaining 0.0276. The agent shares only 17% of its footprint with its two sources, so that margin covers both how it combines them and the candidates it generates itself.

6 DISCUSSION AND CONCLUSION

We set out to ask what an agent is worth when it cannot check its own work. The answer, in this instance, is that it is worth more than the products built by people who could. A panel committed from literature, a cell-type atlas and unlabeled plasma, with no tumour specimen and no detector

response anywhere in the loop, reaches 0.7323 AUC against 0.7050 for a hand-tuned commercial panel at the same sequencing cost, and leads every one of seven comparators on both endpoints. The panels it competes against were built by groups whose job this is, from labeled case-control cohorts, with access to the same public atlases.

The more useful finding is where that advantage lives. It is not in any particular region. Any half of the panel performs as well as the whole, the regions no comparator covers score like any other slice, and the worst of ten random halves still beats the commercial panel. The advantage is a property of the pool rather than of its members. The ablations say why. Take away an input and the regions that replace it score 0.62 to 0.63 on their own, against 0.72 to 0.74 for random content from the same panel. What the agent did was keep such regions out. This is the one thing a selector without an objective can do well. It cannot rank winners, because ranking winners requires seeing outcomes, but it can recognise a loser from prior knowledge alone and refuse it. Exclusion is the mechanism available to an open-loop agent, and it is enough here because the candidate pool holds more usable signal than a 2 Mb budget can carry.

That reading fits the wider picture in the agent literature. Gupta et al. (2025) found that scrambling experimental feedback leaves an agent's performance untouched and concluded that the value sits in the prior. ArchEval (Wang et al., 2026) found that agents cannot reliably judge how a design will perform before it is measured. Both results are consistent with what we see. The agent never predicts which region will be most informative, and it does not need to. It needs only to be right about which regions are not worth the budget, and that judgment comes from knowledge rather than from measurement.

The practical consequence is that panel design becomes cheap enough to repeat. A run costs about two days and a hundred dollars, five independent runs agree to within 0.0036 AUC while sharing 72% of their content, and substituting the background plasma moves selection four times further than the run-to-run spread. Design can therefore be conditioned on the population a test will actually serve, rather than done once on a general cohort and reused. For a field where the scarce input is labeled positives from the target population, removing that requirement changes what is feasible, not just what is cheap.

An open question is what the language model itself contributes. DETERMINISTIC-MERGE is the closest non-agent selector over the same inputs, and the agent beats it by 0.0276 at $p = 0.021$, but a purpose-built deterministic selector would place that margin more precisely. The agent also shares only 17% of its footprint with its two sources, so its margin covers both how it combines them and the candidates it generates itself. Present experiments cannot distinguish these two effects.

The evaluation has its own boundaries. Mixtures are simulated at $300\times$ depth, which removes sampling noise by construction, but these are nonetheless sampled from a limited pool of molecules. While the presented performance is not the clinical performance, we believe that the experiment is robust to ordering. Tissue identity is common-mode across panels scored on byte-identical mixtures, and restricting the average to the four lowest informative fractions reranks all fourteen panels at Spearman 0.9956 against the frozen ranking (Appendix F).

Nothing in the method is specific to methylation or to cancer. The ingredients are a text prior over what might matter, a reference that decomposes the signal into components and define a weakly specified manifold over the data, and unlabeled data from the setting where the decision will be deployed. Similar arrangement can be seen in other scientific research and engineering where a number of sources defining a prior and a manifold of possibility is available at runtime yet no direct feedback is available. Allocation of experimental budget is another clear example that share the property that a closed-loop system is too slow or costly to be useful.

The result suggests that for this class of problem the right question is not whether an agent can find the best answer, which it cannot verify, but whether it can reliably avoid the bad ones, which it can.

AI USE STATEMENT

In this work, we used generative AI tools for cleaning and reformatting datasets. We have not used generative AI tools for proposing or refining hypotheses, developing theoretical models or conceptual frameworks, or supporting qualitative and thematic data analysis, and formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of

proofs, and assisting with translation are not applicable to this work. Additionally, we used generative AI tools for creating and editing software code, creating and modifying figures, identifying and summarising existing literature, formatting references and editing parts of this paper. We have reviewed all AI-assisted work. AI-assisted code was re-derived by the authors, and every number, figure and table reported here was recomputed from the released per-sample score files. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Population specific cfDNA cohorts were part of XXXX, YYYY, ZZZZ clinical studies (redacted for anonymity during peer review), and collected under IRB approval from the respective host institution ethics committee. No additional risk was incurred by participants beyond standard care as a result of this study.

REPRODUCIBILITY STATEMENT

Whole-genome bisulfite sequencing data contain genetic information that could re-identify participants and are available from investigators under a non-commercial material transfer agreement for reasonable IRB-approved research. Requests for data should be addressed to XXXX (redacted for anonymity during peer review).

REFERENCES

Fereshteh Abedini and Soheil Damangir. Why cancer cfDNA models fail on chronic disease: A geometric information-theoretic bound and its architectural implications. In *The Fortieth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=7aV3mkNAEJ.

Chetan Bettegowda, Mark Sausen, Rebecca J. Leary, et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. *Science Translational Medicine*, 6(224):224ra24, 2014. doi: 10.1126/scitranslmed.3007094.

Joerg Bredno, Jafi Lipson, Oliver Venn, Alexander M. Aravanis, and Arash Jamshidi. Clinical correlates of circulating cell-free DNA tumor fraction. *PLoS ONE*, 16(8):e0256436, 2021. doi: 10.1371/journal.pone.0256436.

Carl Edwards, Edward De Brouwer, Xiner Li, Namkyeong Lee, Ehsan Hajiramezanali, Anne Biton, Sara Mostafavi, and Gabriele Scalia. Biology-in-the-loop: Amortized adaptive hit discovery in crispr screens, 2026. URL https://arxiv.org/abs/2609.11877.

Zachariah H. Foda, Akshaya V. Annapragada, Kavya Boyapati, et al. Detecting liver cancer using cell-free DNA fragmentomes. *Cancer Discovery*, 13(3):616–631, 2023. doi: 10.1158/2159-8290.CD-22-0659.

Hunter A. Gaudio, Andrew Bliss, Conor J. Loy, Daniel Eweis-LaBolle, Anne E. Gardella, and Iwijn De Vlaminck. Benchmarking large language models for cell-free RNA diagnostic biomarker discovery. *Nature Communications*, 2026. doi: 10.1038/s41467-026-74077-x.

Rushil Gupta, Jason Hartford, and Bang Liu. LLMs for Bayesian optimization in scientific domains: Are we there yet? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025. arXiv:2509.21403.

Arash Jamshidi et al. Evaluation of cell-free DNA approaches for multi-cancer early detection. *Cancer Cell*, 2022. URL https://www.cell.com/cancer-cell/fulltext/S1535-6108(22)00513-X.

Daniel P. Jeong, Zachary C. Lipton, and Pradeep Ravikumar. LLM-Select: Feature selection with large language models. *Transactions on Machine Learning Research*, 2025. arXiv:2407.02694.

Raju Kandimalla, Jianfeng Xu, Alexander Link, et al. EpiPanGI Dx: A cell-free DNA methylation fingerprint for the early detection of gastrointestinal cancers. *Clinical Cancer Research*, 27(22): 6135–6144, 2021. doi: 10.1158/1078-0432.CCR-21-1982.

Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9:235–284, 2008. URL https://jmlr.org/papers/v9/krause08a.html.

Roni Lehmann-Werman et al. Monitoring liver damage using hepatocyte-specific methylation markers in cell-free circulating DNA. *JCI Insight*, 2018. URL https://insight.jci.org/articles/view/120687.

Wenyuan Li, Qingjiao Li, Shuli Kang, et al. CancerDetector: ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free DNA methylation sequencing data. *Nucleic Acids Research*, 46(15):e89, 2018. doi: 10.1093/nar/gky423.

Nan Lin, Yongping Lin, Jianfeng Xu, Dan Liu, Diange Li, Hongyu Meng, Maxime A. Gallant, Naoto Kubota, Dhruvajyoti Roy, Jason S. Li, Emmanuel C. Gorospe, Morris Sherman, Robert G. Gish, Ghassan K. Abou-Alfa, Mindie H. Nguyen, David J. Taggart, Richard A. Van Etten, Yujin Hoshida, and Wei Li. A multi-analyte cell-free DNA–based blood test for early detection of hepatocellular carcinoma. *Hepatology Communications*, 6(7):1753–1763, 2022. doi: 10.1002/hep4.1918.

M. C. Liu, G. R. Oxnard, E. A. Klein, C. Swanton, M. V. Seiden, et al. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Annals of Oncology*, 31(6):745–759, 2020. doi: 10.1016/j.annonc.2020.02.011.

Tennison Liu, Nicolás Astorga, Nabeel Seedat, and Mihaela van der Schaar. Large language models to enhance Bayesian optimization. In *International Conference on Learning Representations*, 2024. arXiv:2402.03921.

Netanel Loyfer, Judith Magenheim, Ayelet Peretz, et al. A DNA methylation atlas of normal human cell types. *Nature*, 613(7943):355–364, 2023. doi: 10.1038/s41586-022-05580-6.

Joshua Moss, Judith Magenheim, Daniel Neiman, et al. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nature Communications*, 9:5068, 2018. doi: 10.1038/s41467-018-07466-6.

Bernardino Romera-Paredes et al. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024. doi: 10.1038/s41586-023-06924-6.

Yusuf Roohani, Andrew H. Lee, Qian Huang, Jian Vora, Zachary Steinhart, Kexin Huang, Alexander Marson, Percy Liang, and Jure Leskovec. BioDiscoveryAgent: An AI agent for designing genetic perturbation experiments. In *International Conference on Learning Representations*, 2025. arXiv:2405.17631.

Haotian Sun, Yuchen Zhuang, Lingkai Kong, Bo Dai, and Chao Zhang. AdaPlanner: Adaptive planning from feedback with language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. arXiv:2305.16653.

Kun Sun, Peiyong Jiang, K. C. Allen Chan, et al. Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proceedings of the National Academy of Sciences*, 112(40):E5503–E5512, 2015. doi: 10.1073/pnas.1508736112.

Gilles Wainrib, Barbara Bodinier, Haitem Dakhli, Josep Monserrat, Almudena Espin Perez, Sabrina Carpentier, Roberta Codato, and John Klein. Can AI scientist agents learn from lab-in-the-loop feedback? evidence from iterative perturbation discovery. In *International Conference on Machine Learning*, 2026. arXiv:2603.26177.

Chenyu Wang et al. ArchEval: Measuring AI agents as computer architects. *arXiv preprint arXiv:2607.03601*, 2026.

Y. Wang, Y. Hou, L. Yang, et al. Accelerating primer design for amplicon sequencing using large language model-powered agents. *Nature Biomedical Engineering*, 2025. doi: 10.1038/s41551-025-01455-z.

Rui-hua Xu, Wei Wei, Michal Krawczyk, et al. Circulating tumour DNA methylation markers for diagnosis and prognosis of hepatocellular carcinoma. *Nature Materials*, 16(11):1155–1161, 2017. doi: 10.1038/nmat4997.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *International Conference on Learning Representations*, 2024. arXiv:2309.03409.

Changjie Yu, Siyu Liu, Guanren Qiao, Mingrui Luo, Yujia Xiang, and Zhiping Xu. PerturbTrace: Evaluating feedback use by AI co-scientist agents in perturbation discovery. *bioRxiv*, 2026. doi: 10.64898/2026.08.18.745260.

A PANEL PROVENANCE AND IDENTIFIERS

Semantic names are used throughout the main text; this table maps each to its internal identifier and source. Region and interval counts are raw BED values, and CpG counts are against the full 29,464,876-CpG GRCh38.d1.vd1 reference. Pooling drops regions containing no CpG, which is why post-transformation region counts run below the raw counts quoted here.

Table 5: All fourteen panels with provenance, checksum prefix and size.

Name	Source	BED lines	Intervals	Mb	CpGs	CpG/kb
AGENT	agent, run 1 (reference)	7,819	5,883	2.28	233,081	102.4
AGENT, run 2	agent, repetition	8,730	6,895	2.42	224,278	92.7
AGENT, run 3	agent, repetition	8,458	6,597	2.37	226,830	95.7
AGENT, run 4	agent, repetition	8,720	6,866	2.42	225,301	93.2
AGENT, run 5	agent, repetition	8,918	7,072	2.46	223,147	90.8
AGENT-ALD	agent, substituted plasma input	10,396	8,628	2.72	213,052	78.2
AGENT-NOATLAS	agent, atlas removed	11,751	9,944	3.03	204,704	67.6
TWISTALLIANCE	Twist Alliance pan-cancer, vendor files	8,129	6,324	2.36	192,308	81.3
DETECTABILITY-ONLY	intended-population plasma, low background variance	6,331	4,483	2.00	264,190	132.1
EPIPANGIDX	EpiPanGI Dx, published targets	2,367	1,357	0.85	65,045	76.2
ASSOCIATION-ONLY	Open Targets association ranking	4,116	2,419	1.39	105,952	76.1
Random 1	random draw, public methylome panel	7,587	6,657	1.93	56,211	29.1
Random 2	random draw, public methylome panel	7,649	6,718	1.94	57,532	29.7
Random 3	random draw, public methylome panel	7,764	6,895	1.95	56,408	29.0

B DATA AVAILABLE AT EACH DESIGN STAGE

Table 6: What each stage could see. No tumour or adjacent-normal molecular profile is available at any design stage, which is the boundary Figure 1 draws.

Stage	Cancer literature	Cell atlas	Cancer-free plasma	Tumour/normal WGBS
Claim extraction	yes	no	no	no
Pattern construction	yes	yes	no	no
Detectability validation	yes	yes	yes	no
Panel ranking and freezing	yes	yes	yes	no
Detector fitting	no	no	yes	yes
Evaluation	no	no	yes	yes

C PER-FRACTION RESULTS

Table 7: Primary endpoint by tumour fraction, all fourteen panels. The final row is the frozen headline, the mean over the seven fractions at or above 6.3×10^{-4}.

Fraction	Agent	Run2	Run3	Run4	Run5	ALD	NoAtl	Comm	Det	GI	Assoc	Rnd1	Rnd2	Rnd3
0.005%	.5142	.5136	.5117	.5176	.5152	.5139	.5136	.5140	.5114	.5066	.5038	.5050	.5106	.5048
0.0116%	.5150	.5159	.5193	.5174	.5136	.5166	.5189	.5177	.5008	.4998	.5064	.5081	.5073	.5042
0.0271%	.5331	.5238	.5318	.5303	.5283	.5253	.5229	.5200	.5196	.5085	.5070	.5003	.5045	.5100
0.063%	.5692	.5653	.5710	.5643	.5678	.5645	.5491	.5558	.5464	.5329	.5246	.5317	.5303	.5257
0.147%	.6089	.6044	.6060	.6036	.6088	.6081	.5980	.5905	.5666	.5489	.5413	.5460	.5501	.5495
0.341%	.6614	.6545	.6579	.6552	.6602	.6619	.6483	.6407	.6201	.6057	.5952	.5869	.5916	.5996
0.794%	.7321	.7289	.7338	.7329	.7292	.7460	.7144	.7038	.6754	.6878	.6653	.6490	.6496	.6641
1.85%	.8003	.7929	.7992	.7970	.7937	.8124	.7988	.7631	.7428	.7648	.7284	.7028	.7069	.7340
4.30%	.8525	.8521	.8555	.8530	.8541	.8711	.8569	.8174	.8033	.8125	.7772	.7446	.7509	.7978
10%	.9022	.9053	.9048	.9034	.9011	.9200	.9083	.8639	.8590	.8362	.8077	.7945	.7877	.8621
grid mean	**.7323**	**.7290**	**.7326**	**.7299**	**.7307**	**.7406**	**.7248**	**.7050**	**.6877**	**.6841**	**.6628**	**.6508**	**.6524**	**.6761**

D REPRODUCIBILITY ACROSS RUNS

See figure 5.

Table 8: Secondary endpoint, standardised pAUC above 90% specificity, by tumour fraction. The three lowest informative fractions sit at or below chance.

Fraction	Agent	Run2	Run3	Run4	Run5	ALD	NoAtl	Comm	Det	GI	Assoc	Rnd1	Rnd2	Rnd3
0.005%	.4930	.4930	.4926	.4935	.4925	.4930	.4942	.4945	.4971	.4923	.4965	.4961	.4982	.4981
0.0116%	.4927	.4937	.4937	.4929	.4923	.4930	.4919	.4944	.4943	.4916	.4968	.4951	.4952	.4949
0.0271%	.4943	.4950	.4937	.4932	.4930	.4939	.4938	.4962	.4978	.4954	.4968	.4930	.4945	.4983
0.063%	.4939	.4946	.4936	.4939	.4947	.4933	.4941	.4954	.4969	.4936	.4978	.4954	.4968	.4984
0.147%	.4943	.4959	.4947	.4950	.4943	.4948	.4938	.4970	.4969	.4961	.4961	.4952	.4971	.4969
0.341%	.4985	.4997	.4973	.4970	.4970	.4996	.4974	.4971	.5007	.5042	.4991	.4950	.4984	.4998
0.794%	.5154	.5156	.5127	.5120	.5132	.5157	.5113	.5027	.5078	.5156	.5123	.4998	.5050	.5043
1.85%	.5819	.5772	.5827	.5788	.5767	.5856	.5792	.5406	.5411	.5520	.5309	.5112	.5277	.5167
4.30%	.6947	.6833	.6950	.6982	.6919	.7083	.6960	.6213	.6225	.6350	.5884	.5536	.5895	.5803
10%	.8231	.8206	.8201	.8280	.8195	.8471	.8251	.7464	.7358	.7350	.6755	.6822	.7183	.7323
grid mean	**.5860**	**.5838**	**.5852**	**.5861**	**.5839**	**.5920**	**.5853**	**.5572**	**.5574**	**.5616**	**.5429**	**.5332**	**.5475**	**.5470**

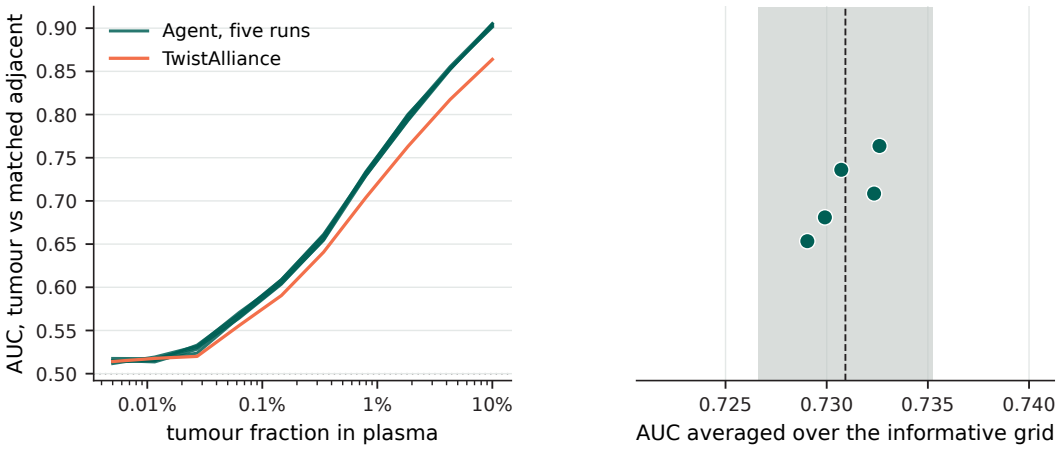


Figure 5: Five runs of one configuration. (a) The band is the spread across runs at each fraction, against the best comparator. (b) Each run's frozen endpoint against the shaded bands, which are the smallest and largest per-comparison effects the design can detect at 80% power. Every run sits inside the band, so this is a null the study had the power to break.

E SELECTION AGAINST DENSITY

Two bounds hold the selection claim from opposite sides. Cost-matched random draws lose by $+0.0726$ at $p \leq 0.0005$, a margin that covers selection and CpG density together, since those panels carry a quarter of the agent's CpGs at 29 per kilobase against 102. From the other side, DETECTABILITY-ONLY carries more CpGs at higher density and still loses by $+0.0447$ at $p = 0.0024$. Among designed panels the endpoint does not rise monotonically with density. Together the two bounds show that the margin over random content is not density alone and that part of it is selection. Density-matched random content would give the split. Figure 6 plots the endpoint against density for all fourteen panels.

F ROBUSTNESS OF THE RANKING TO THE FRACTION GRID

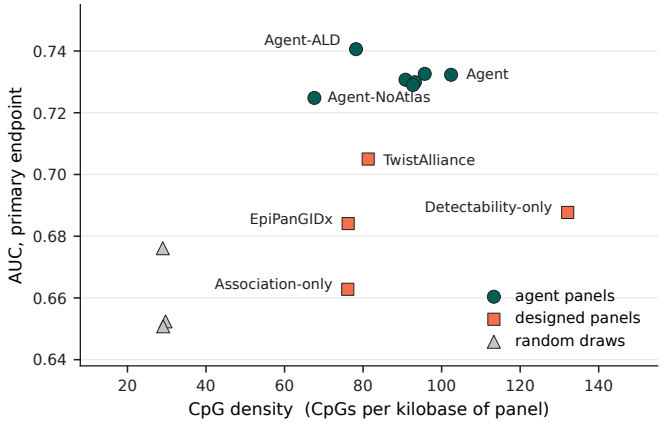


Figure 6: Frozen endpoint against CpG density for all fourteen panels. Density does not order the panels: the densest panel in the study ranks ninth. The shaded region marks the density range in which no random panel exists, which is why the margin over random content cannot yet be partitioned between selection and density.

Table 9: The frozen seven-point endpoint against the four-point low-band mean, which uses only fractions at or below 0.79%. Spearman between the two rankings is 0.9956; the only swap is two panels 0.0003 apart.

Panel	7-point (frozen)	rank	4-point low band	rank
AGENT-ALD	0.7406	1	0.6451	1
Agent, run 3	0.7326	2	0.6422	3
AGENT	0.7323	3	0.6429	2
Agent, run 5	0.7307	4	0.6415	4
Agent, run 4	0.7299	5	0.6390	5
Agent, run 2	0.7290	6	0.6382	6
AGENT-NOATLAS	0.7248	7	0.6275	7
TWISTALLIANCE	0.7050	8	0.6227	8
DETECTABILITY-ONLY	0.6877	9	0.6021	9
EPIPANGIDX	0.6841	10	0.5938	10
Random 3	0.6761	11	0.5847	11
ASSOCIATION-ONLY	0.6628	12	0.5816	12
Random 2	0.6524	13	0.5804	13
Random 1	0.6508	14	0.5784	14