

D3 Chat-Log Analysis: Margarita (b.next) × IGOR V2.1

Source: `margarita_d3_PAP_r1-r2.md` (workspace `d3_PAP_r1-r2`, staging). Timestamps are UTC, 2026-10-20 02:49 → 2026-10-21 07:16. Line refs `L####` point into the original log.

0. Shape of the session

Duration	~28.5 h wall-clock; active blocks on 10-20 (02:49–12:50, 19:33–20:58) and 10-21 (04:53–07:16)
Messages	103 from Margarita, 147 from IGOR; 429 events including tool calls
Agents	CopilotAgent (145 calls), ProblemFramingAgent (26), intent_elicitation_agent (3)
Proposals	40 cards: 30 applied, 8 rejected, 2 no response. Most rejections were superseded drafts (volume constraint ×4, retrain ×2, Pmix_dispende exclusion)
Spreadsheet audits	8 "Review the file" uploads: 1 dataset CSV, 5 × <code>261020_PAP-LC_expt-setup.xlsx</code> , 2 × <code>..._redo.xlsx</code>
Most-used tools	<code>search_knowledge</code> 19, <code>get_file_as_plain_text</code> 18, <code>create_visualization</code> 11, <code>propose_edit_experiment</code> 8, <code>get_predictions</code> 6, <code>search_openalex</code> 6, <code>get_ges_justification_context</code> 2, <code>get_control_diagnostics</code> 2
Data	229 rows (228 active) over 8 LabCraft runs (r1007–r1018); 16 real inputs, 3 categorical inputs, 3 outputs

The arc in one line. She set up a workspace in about 1.5 h. Most of the session was then spent designing, auditing and re-designing a 28-condition plate. The plate came back with **zero signal**. IGOR then root-caused the failure with her and helped her build a conservative 19-condition redo plate. The redo's bench results are not in this log. Per the team, the redo partly failed (probably an air bubble), and 22 of its 26 experiments were salvaged through outlier triage (V12).

Where the conversation went (time spent, not value). Most of the chat is plate design, spreadsheet audits and troubleshooting. That tells us what she needed *conversation* for. It does not tell us where the value came from, for three reasons. - **The log under-records the engine.** - GES were generated in the UI, not through chat tools. - A suggestion batch or a model retrain shows up as one line, while auditing a spreadsheet takes pages. - A chat log naturally over-represents whatever needs back-and-forth. - **The agent layer's best moments rest on engine outputs.** The insights that mattered most were the agents interpreting numbers the engine computed: - the evidence packets (V3); - the replicate noise floor and outlier tests (V4, V12); - the model's uncertainty map that exposed the design gap (V5); - the confounded sensitivity (V7); - the `get_predictions` scores behind every chat-designed candidate (Exp 1–7).

The engine and the agents are coupled, so splitting the credit between them is not meaningful from this log. - **Value is ultimately decided at the bench.** We can't say whether the engine-generated GES (GES-1, GES-2, GES-S, GES-10), the agent-designed candidates, the literature rows or her own controls performed best until the redo and r3 results are compared by source (as in V11's "Endpoint RFU by source").

What the log *does* support: IGOR got used across the whole workflow (setup, curation, explaining the optimizer, design, QA and troubleshooting), not only as a suggestion generator. The value points below are judged on their outcomes, not on how much of the chat they took up.

1. Where she got the most value

Each point is broken down as **Problem / goal** → **Method / how** → **What happened** → **What we learned**.

Sources for each point: - **V1–V7** come from the chat log, with L#### refs. - **V8** happened outside the log (between Margarita and Scott) and is written from Scott's account. - **V9–V12** come from the FWM team's account and supporting assets.

Rounds (as the team uses them)

Round	What it was
r0	Control experiments fed to IGOR: everything before 10/18. SOP recipes run in replicate to confirm the setup worked and the results repeated.
r1	The 10/18 run (r1018). The first test of IGOR's agentic suggestions, alongside concentration sweeps chosen by MK.
r2	The 10/20 run. The second test of IGOR's agentic plus Bayesian (GES) suggestions. Two plates: the first 261020 plate gave zero signal; the redo kept 22 of 26 experiments after outlier triage (V12).
r3	The 10/21 run. The third test of IGOR's agentic plus Bayesian suggestions.

This chat log (workspace d3_PAP_r1-r2) trains on r0 + r1. It covers designing the r2 plate, the zero-signal result, and setting up the redo.

V1. Plate QA that caught real errors before the bench

- **Problem / goal:** A 28-condition, 84-well LabCraft plate is planned in a multi-tab spreadsheet (conditions , reagents , LC reagents , LC wells , platemap). Many things have to agree for it to print correctly:
- the stock concentrations and the volumes they imply per well;
- the 2 µL well budget;
- droplet sizes of 9–10 nL (and 50–80 nL stream drops);
- the channel count and the well count.

One wrong cell costs a plate, or a whole round. - **Method / how:** She uploaded the setup sheet for review 8 times ("Review the file"). Each time, IGOR read every tab (get_file_as_plain_text) and checked them against: - one another; - the reaction volume; - the stock concentrations she gave it; - the droplet resolution; - the platform constraints in her corpus. - **What happened:** IGOR caught a series of real errors:

When	What IGOR caught	Ref
10-20 09:16	MAX() references for K-phos and pyruvate swapped between tabs	L9810
10-20 09:16	GES-1 overflowing the well at 2.028 µL	L9810
10-20 09:16	Two near-duplicate rows below droplet resolution	L9810

When	What IGOR caught	Ref
10-20 09:16	A mismatch between the well count and the master-mix budget	L9810
10-20 12:50	The DNA stock linked to the wrong prep (13.98 vs 26.98 nM), giving -21 nL of water	L12832
10-20 20:34	The instrument's "level collapse" warnings mapped onto 4-D droplet vectors; only 2 real collisions out of 28 conditions	L13023
10-21 06:15	Two redo conditions that print identical droplet vectors (9,2,3,2)	L13978
10-21 07:15	Three conditions with negative nL left after a stock change, each fixed with a single-droplet nudge	L14392

Along the way, IGOR also proposed a single 1000 mM K-phos stock. That makes 5 mM printable as one drop without adding a 10th cartridge. - **What we learned:** - Reviewing the bench plan was the most frequent use in the chat, alongside the optimizer, not instead of it. This is where D2's lesson ("hardware realities must be encoded") now shows up day to day. - It also exposed the weak spot. IGOR did this arithmetic in LLM prose and got some of it wrong (F3). The next step is to move plate, volume and droplet math into a deterministic tool.

V2. Separating the factors behind the PAP yield recovery (see Fig. 4a)

- **Problem / goal:** Earlier LabCraft PAP runs gave low or no signal. Margarita made two changes at once, dispensing the proteins (PMix, AckA, KatE, Pox5) separately from the master mix and removing RNAsin, and yields recovered. She didn't know which change mattered: "haven't isolated yet which of the two changes was what helped" (L10876).
- **Method / how:** IGOR compared her existing runs (r1007, r1008, r1015, r1016, r1018) with `get_control_diagnostics` and her corpus (failure-modes register, storage-buffer doc, Slack thread on assembly order, a `platemap` note "pmix bubbly in reservoir"), plus OpenAlex literature. We then re-analyzed the same runs with all three changing factors shown side by side (Fig. 4a).
- **What happened:**
- **IGOR got one part right:** removing RNAsin did not explain the recovery (r1007 vs. r1008, median 0.62 vs. 0.56 μM).
- **IGOR missed the main driver.** It credited protein separation (`Pmix_dispense`, effect 0.363, $p \approx 9\text{e-}23$) and saved that as a team convention in workspace memory (L11102–L11866). But the plasmid also switched from pOpen to pET21a between the premixed and separated runs, and IGOR didn't account for it.
- **The within-plate comparison resolves it.** On r1016, with the same separated protein dispense, pOpen wells gave a median of 0.64 μM and pET21a wells 2.99 μM (about 4.7 $\times$). Separating the proteins with pOpen did no better than premixing (0.64 vs. 0.56–0.62 μM). In base Cytosol without PAP, neither plasmid nor dispense mode mattered (r1015, 0.93–1.09 μM), so the plasmid effect appears specific to PAP reactions.
- **What remains open:** the full flatlines Margarita described ("happened before when the proteins were premixed", L13694) are not in this dataset, so protein separation may still matter for that failure mode. The pET21a wells on r1016 also had more template DNA (1.9 vs. 1.0 nM), although r1018 reached a 2.35 μM median with pET21a at 1.0 nM.
- **What we learned:**

- The dataset did hold the answer, but only a comparison that controls for co-varying factors reveals it. Plasmid was already a modeled categorical input, so IGOR had what it needed to catch the confound and didn't.
- **Product:** before attributing an effect, IGOR should check whether other factors changed alongside it and say so. Conventions saved to memory should carry their evidence strength.
- **Action:** update the workspace-memory convention so future rounds don't treat protein separation as the established cause. The pET21a-vs-pOpen effect in PAP reactions is now the stronger hypothesis to test directly.

V3. IGOR explaining its own optimizer, which changed how she configured it

- **Problem / goal:** The suggested experiments (GES) didn't make sense to her. "No matter what I set the completion time goals to, it keeps selecting GESs with high completion times yet steady states below my lower goal" (L6564). Later she asked why GES-3 and GES-4 were predicted so low, when IGOR's chat-designed experiments were predicted high (L10777).
- **Method / how:** IGOR pulled each suggestion's evidence packet (`get_ges_justification_context`): predictions, uncertainty, hit probability and bounds. It then explained the result in terms of how the engine actually works, and pinned that explanation to the suggestion row.
- **What happened:** IGOR identified four causes: 1. Her yield floor (5.0 μM) was above the best yield ever measured (3.99 μM). Hit probability was therefore maximized by chasing uncertainty far outside the data (CP = 100 mM vs 42 mM max observed). 2. The long completion time was a side effect of that extrapolation, not a target. 3. Her "exactly 2 μL " constraint had become impossible once DNA was locked at 1 nM, so the solver fell back to 1.781 μL . 4. For GES-3 and GES-4: with 19 of 23 inputs held constant, the diversity (similarity) filter blocked the rest of the high-yield ridge. That pushed GES-4 into a dead 2 mM-Mg corner (hit probability 1.7e-20).

She reset her goals, bounds and constraint, and IGOR proposed tighter Mg and pyruvate floors (L6999, L10835).

- **What we learned:** - When the optimizer seems to misbehave, it is usually following the user's own settings to the letter. An agent that reads the evidence packet turns "the AI is wrong" into a configuration fix. - This is the clearest demonstration of the agents and the engine working together. - It also gives us product work: guard against goals above the observed maximum, report infeasible constraints, and drop suggestions with near-zero hit probability (F8).

V4. Data curation that matched what she meant scientifically

- **Problem / goal:** Several wells had zero yield, and the curve fitter gave them meaningless completion times (1.9–16 h). Margarita: "the steady state and t_{final} values are real but the fits are poor... I want it captured that these are bad conditions though. Should we change completion time to n/a or what" (L5396).
- **Method / how:** IGOR used three sources:
 - the outlier test (`get_outliers`);
 - the replicate noise diagnostics (`get_control_diagnostics`);
 - a yield-vs-completion-time plot.
- **What happened:**
 - IGOR showed that every completion time above 6.8 h belonged to a near-zero-yield well. Unexplained replicate variance was 51.8% for completion time vs about 4% for yield.
 - It excluded one dead well that disagreed with its 5 replicates.

- On 7 zero-yield wells it cleared only the completion-time value and kept the zero yields.
- It then set the completion-time goal as a band ([3.5, 6.5] h, later [3, 5] h) rather than "maximize" (L4268, L5380–L6166).
- **What we learned:**
- The model still learns that low Mg kills expression, without learning from fitting artifacts.
- This is the engine's design principle ("unmeasured outputs are masked, never imputed") reached through conversation about what the scientist meant.

V5. Finding the hole in her own DOE

- **Problem / goal:** "What areas of Mg/CP/Pyr/K-phos are unexplored and should I collect more data" (L9478).
- **Method / how:** IGOR crossed three sources: her full dataset, the model's uncertainty (`get_predictions`), and the literature.
- **What happened:**
- IGOR found a "diagonal hole" in the r1018 design: whenever phosphate was mid-range or CP was high, Mg happened to be set low (~9.5 mM). Whenever Mg was high, phosphate was always high.
- So the model had never seen 15–22 mM Mg with 8–20 mM phosphate. That is the quadrant both Yadav and her own best low-phosphate runs point to.
- IGOR ranked 5 gaps, said which to skip (the extreme corners), and designed a condition specifically to break that link (L9594–L10010).
- **What we learned:** The AI's view of the design surfaced a confound in the human-designed experiment that a factor-by-factor review would miss. This sets up point 8.

V6. Literature → bench translation

- **Problem / goal:** "What conditions from Yadav's paper haven't been tested for us yet" (L9989). She also wanted to know which ones were worth plate space.
- **Method / how:** IGOR extracted Yadav's DOE table from the supplementary info (Tables S3 and S5). It cross-referenced every condition against her 229 experiments and the planned plate. It then formatted the result in her sheet's exact columns, as TSV and then CSV on request.
- **What happened:**
- **The CP difference:** IGOR pointed out that Yadav ran PAP with **CP = 0**, while her plate added about 19 mM CP on top. That explains why Yadav's Axial 14 gave 193 RFU but her version gave only 0.6 μ M: the CP binds up the magnesium.
- **Triage:** it sorted the literature conditions into redundant, poor and worth running.
- **Convergence:** it noted that the engine's top suggestions had independently gravitated toward Yadav's best corner (high pyruvate, low phosphate). *Verify this before featuring it.*
- **Result:** paste-ready rows went onto her plate (L10050–L10318).
- **What we learned:** Published results don't transfer to a new system as-is; they need translating into it. IGOR did that translation in minutes, and explained the mechanism behind the difference.

V7. Honest handling of what the model does not know

- **Problem / goal:** "Any insights on ribosome concentration?" and "how much of an effect will reducing ribos have?" (L8733, L9174).

- **Method / how:** IGOR used three sources: local sensitivities (`get_sensitivities`), predictions across a ribosome sweep (`get_predictions`), and the PURE literature.
- **What happened:**
 - IGOR reported the model's steep ribosome slope. It also said it should not be trusted.
 - Ribosome concentration had only varied by about 9%, through dilution of the master mix in wells that also carried heavy salt loads. So the slope is confounded with the salt penalty.
 - It recommended a 3-point titration to separate the two effects (L9449).
- **What we learned:**
 - This is the "computed, cited, or labelled as hypothesis" contract at its best: the agent argued against the model's own number.
 - Behaviour like this is what earns trust. It contrasts with the places where IGOR stated mechanisms as fact (F6).

V8. Breaking out of one-factor-at-a-time thinking (*outside the log; from Scott's account*)

- **Problem / goal:**
 - Humans find it hard to think in high dimensions. Yet working in high dimensions is where AI, and Bayesian optimization in particular, helps most.
 - Bench science is taught one factor at a time (OFAT): start from a baseline recipe and vary one thing.
 - That mental model has to change before a scientist can get the full benefit of a multivariate model. This gap between a new information technology and the habits of the people using it recurs throughout history: a new tool pays off only once people learn to think in its terms.
- **Method / how:** We used IGOR's 2-D prediction plots. Each one shows predicted yield over two inputs, with every other input held fixed at a chosen *reference formulation*. We compared two reference points:
 - Yadav's centerpoint in Cytosol-PAP space (pyruvate 27.5, phosphate 27.5, Mg 12.5 mM; the red dot in the Mg-vs-phosphate panel). At that point IGOR reproduced the published landscape.
 - A different formulation, where the landscape looked very different from the paper.
- **What happened:**
 - Margarita wanted to check that her experiments reproduced the published work. She generated a 2-D prediction plot, and it did not match the Yadav figure.
 - She told Scott and was worried the data couldn't be trusted.
 - Scott asked what all the other inputs were held at in the Yadav figure. Margarita realized her plot had fixed the non-plotted inputs at values very different from Yadav's.
 - She re-plotted at Yadav's reference formulation, and the result matched the paper.
 - Seeing the same model give two different-looking landscapes depending on where the other inputs sit made the coupling concrete. The Mg optimum moves with phosphate, pyruvate and CP. There is no single "Mg curve."
 - From then on she started reasoning about formulations as points in a multivariate space rather than as one-factor variations from a baseline.
- **What we learned:**
 - Giving the user the right insight at the right moment produced what may be the most important shift of the engagement. She began to see her biological system as a truly high-dimensional, coupled system, rather than a baseline landscape you explore away from one factor at a time.
 - That shift is what lets a user get the full benefit of AI, and of Bayesian techniques in particular. The same thinking shows up in V5 (the design confound) and V6 (CP changes Yadav's Mg optimum).

- **Product implication:**
 - 2-D slices should always show their reference formulation prominently.
 - Slices should offer "compare at another reference point" side by side.
 - Plots should snap to literature or control formulations.
 - IGOR should proactively explain a mismatch like this one, which today took a human (Scott) to resolve.
- **For the D3 devnote:** a figure pair would make the point immediately: the same model, sliced at Yadav's point vs. her formulation.

V9. Optimizing over categorical choices (DNA prep, ribosome lot)

- **Problem / goal:** The data contains categorical features with several possible values, such as DNA ID (plasmid prep) and Ribosome ID (ribosome lot). She needed GES that could choose among those options, not only tune concentrations.
- **Method / how:** In the UI, the user can now pick, for each categorical feature column, which of its values the optimizer may search over.
- **What happened:**
 - Margarita wanted to generate GES across different DNA IDs and Ribosome IDs.
 - The chat log shows she first kept `Plasmid`, `DNA ID` and `Ribosome ID` as categorical inputs during schema review (L1161).
 - IGOR's model then estimated a lot effect: about 0.6 μM between ribosome lots at the same recipe (L9165).
- **What we learned** (*draft; please confirm or rewrite*):
 - Lot and prep identity is a real experimental variable in cell-free systems, not noise to average away. This matches the failure-modes register's advice: "record every reagent lot / artifact ID as a categorical feature."
 - Letting the scientist choose which lots are in scope for the optimizer turns an inventory question into an optimization question.
 - Example: which ribosome lot should the next plate use?
 - It also gives the model a way to separate lot effects from recipe effects.
- **Caveat from the log:** IGOR mixed up which lot was the 53 μM stock (F5). Narration of categorical features needs to be tied to the actual lot metadata before we lean on it.

V10. Round 1: designing a plate when the data has almost no variance

- **Problem / goal:** Round 0 was SOP recipes run in replicate, to confirm that the setup worked and the results repeated. That makes a good noise floor but gives the model almost no variation to learn from. How do you choose the next formulations when the formulations have barely varied?
- **Method / how:** IGOR combined several information sources to propose formulations, each with a mechanistic rationale:
 - the literature it retrieved;
 - the connectors (Slack, Notion, Drive);
 - the uploaded corpus;
 - what the model could already say.

It then ranked which ones to run and produced a table for the user. - **What happened:** - Margarita needed suggestions for r1, but the starting dataset was mostly replicates of baseline Cytosol and PAP. - Through

question-and-answer with IGOR, she arrived at the formulations that went onto the r1 plate (10/18), alongside her own concentration sweeps. - Supporting asset: [assets/CDK Steady State Optimization – 10-Experiment STE Engineering Summary.pdf](#). - **What we learned:** - The human-agent loop was what made progress possible when the starting data carried so little information. The scientist asks; the agents draw on the Bayesian engine and the information repository; IGOR returns insights; the scientist applies judgment or asks a follow-up. - A model alone has nothing to extrapolate from in this cold-start case, and the literature alone doesn't fit the user's system. Together they gave her a defensible first plate. - The same pattern recurs in the r2 design in this log (V5, V6).

V11. Benchmark: IGOR vs. Claude on the same task

- **Problem / goal:** We want to benchmark IGOR against alternatives, and understand the strengths and weaknesses of each.
- **Method / how:** We compared formulations generated by IGOR with formulations generated by Claude (a frontier general-purpose LLM) for the same r1 objective. We then compared the results at the bench.
- **What happened (round 1):** 1. **Claude reviewing both sets:** Margarita asked both Claude and IGOR to generate formulations. When she then gave Claude IGOR's formulations, Claude judged them better than its own and recommended running IGOR's. (*Supporting doc: [to add]*) 2. **Asked which technique would reach Margarita's goals faster,** Claude recommended Bayesian optimization: "*To be completely honest with you: No, I would likely not beat a well-tuned Gaussian Process (GP) Bayesian Optimizer in the long run for this specific problem.*" 3. **At the bench,** the results so far suggest IGOR's formulations were more likely to hit the targets, and they came with mechanistic reasons. This needs a closer look. (*Figure: section 17, "Endpoint RFU by source", in [assets/20261018_PAP-LC_Labcraft_QC_mScarLet_g75.html](#) .*)
- **What we learned (draft):**
- A general LLM can reason about the chemistry. Yet even by its own account, it lacks the model of *this* system that a Bayesian engine trained on the user's data provides. That is the core of the "Agentic Machine Learning" thesis, stated by a competitor.
- **How to present it:**
 - Claude's preference and Claude's quote are color. They are not evidence: an LLM deferring to a list it is shown is a weak signal.
 - The evidence is item 3, the bench comparison by source.
 - For any external use (especially Anthropic), report n, replicates, the noise floor, and how the conditions were chosen for each arm. Also note that this was one round with a small number of conditions.
- **Accuracy note:** Claude's quote says "GP." IGOR's engine is a deep ensemble, not a GP. Quote it verbatim, but don't let it describe our method.

V12. Round 2: salvaging a partly failed run with interactive outlier triage

- **Problem / goal:** Some r2 wells worked and some didn't. Ideally the good wells would be kept rather than throwing out the whole run.
- **Method / how:** Margarita and IGOR walked through FWM's outlier detection together. IGOR flagged outliers and suggested a likely cause for each. It also judged whether each one was more likely an experimental failure or a genuinely unusual experiment.
- **What happened:**

- Margarita ran the r2 PAP experiments; some wells worked and some did not. She believes an air bubble in the LabCraft dispense was the cause.
- She would normally discard the run and re-run the plate. Instead, working through the outliers with IGOR, she identified **4 bad experiments out of 26**, removed them, and kept the other **22** for model training in r3.
- **What we learned:**
- Interactive outlier triage turned a lost plate into 85% usable data. It saved a re-run's worth of reagents, instrument time and calendar time, and kept the campaign on schedule for r3.
- Two things made this work: the statistics (a Rosner test against the replicate noise floor) and the agent's explanation of likely causes, which together let the scientist decide well by well with confidence.
- The same mechanism appears in the log on r1 data: one dead well excluded, 7 artifact cells cleared, and the rest kept (V4).
- **Which plate this was (confirmed by the team):** r2 involved two plates, and they should be reported separately.
 1. **The first 261020 plate** came back with **zero signal in all 84 reaction wells** (L13251). That is a separate, complete failure: nothing was salvageable, and the root cause is still open (see F1).
 2. **The redo of 261020** was cut back to the 19-condition design built in this log (L13911–L14392). Some of its wells worked and some did not. That is the run where 22 of 26 experiments were kept.
- **Write-up framing:** this makes r2 a stronger story than either event alone. One plate failed outright. The redo was designed with IGOR, partly failed (probably an air bubble), and was still salvaged by working through the outliers with IGOR.

2. Where it fell short (product feedback, and claims we should not make)

#	Issue	Evidence	Why it matters
F1	The redesigned plate failed completely (zero signal). Several things changed at once: concentrated tRNA (70 µg/µL) and ribosome (53 µM) stocks; a master mix shrunk to 0.8 µL; protein dispense moved late; water moved late. Some of these IGOR had endorsed. It never warned against changing several variables in one run, and its final audit said "100% good to run this plate".	L13229, L13251	A must-report for D3, in the D2 spirit. The product lesson: the audit should flag "first use of new lots + protocol changes" as risk.
F2	Flip-flop under challenge. IGOR recommended proteins <i>late</i> , citing r1016's record at 3.99 µM (L11886, L13931). When she asked whether late dispense caused the failure, it reversed to "stick with early, like r1018" (L13955).		Agreeing with whoever spoke last undermines trust. It needs a stable recommendation with stated uncertainty.
F3	LLM arithmetic in prose, visibly self-correcting ("Wait—", "HOLD ON!", "Aha!"), and wrong several times. Examples: it proposed GES-10 at 5 mM K-phos from a 2000 mM stock (not printable), and an Exp-4-BL that overflows. She corrected both (L14162). It also mixed CP stock assumptions in the 1000 mM K-phos math, so its "+2.8 nL" became –8.2 nL on her sheet.	L9963, L13419, L14095– L14392	This contradicts "code owns every number" for plate math. Build a deterministic volume/droplet/feasibility tool and route all plate arithmetic through it.
F4	Wrong description of its own engine: it called the surrogate a "Gaussian Process" several times.	L13045, L13110, L14242	It is a deep ensemble. Fix this in the prompt or the tool descriptions.

#	Issue	Evidence	Why it matters
F5	Lot-identity mix-up: it called AR-907 the "53 μM " stock and credited it with a +0.6 μM batch advantage (L9165, L9384). Later it said AR-907 was 18 μM and AR-1220 was 53 μM . This plausibly encouraged the switch to concentrated stocks.		Categorical-feature narration must be tied to the actual lot metadata.
F6	Mechanistic stories presented as fact, not labelled as hypotheses. Examples: Pox5/KatE "smoking gun" bubbles; RNAsin oxidation; "each μM of ribosomes coordinates 2–3 mM Mg^{2+} "; NTP+TPP stripping Mg in the master mix. Some citations are a stretch.	L11124, L11492, L9454, L13770	Contradicts the grounding contract's "labelled as a hypothesis" clause. Needs a hypothesis tag.
F7	Misattributed the yield recovery: IGOR credited protein separation, but the within-plate comparison points to the plasmid switch (pOpen $\rightarrow$ pET21a), which it didn't account for (V2, Fig. 4a).	L11102	It saved a convention to memory that the data don't support. Attribution should check co-varying factors.
F8	The optimizer emits useless or infeasible suggestions silently: a GES with hit probability $1.7\text{e-}20$; an infeasible equality constraint solved as 1.781 μL ; GES-1 at CP = 100 mM, outside her LabCraft limits.	L6999, L10835	Drop near-zero-hit GES. Report infeasibility instead of falling back silently. Import instrument limits as bounds.
F9	UX friction		
	• A card was not rendered ("you didn't show me a card").	L2696	
	• Bounds edited in the dashboard looked unsaved because the read cache was stale ("can you ask me for each one, I'm not able to save").	L3050–L3273	
	• One card per row: "continue" clicked 7 $\times$ to clear 7 cells.		
	• Equality constraints need a [1.999, 2.0] workaround.		
	• It could not export a markdown file to PDF.		
F10	Literature search noise: the first <code>search_knowledge</code> hit returned a mitochondrial-carrier review.	L131	Retrieval relevance.

What the expert supplied (a good human-in-the-loop story, and evidence the expert remains essential): - She corrected IGOR's RNAsin reasoning: "if there's rnasin, wouldn't I see some signal in some of the wells?" - She supplied the hard physical constraints: water must print near-last because of evaporation, and the master mix needs an Mg floor for ribosome integrity. - She caught the droplet math errors and the stock-feasibility errors. - She relayed the instrument's own level-collapse warnings.

3. D2 lessons: did V2.1 close them?

D2 lesson	What happened in D3	Verdict
Equal goal weighting under-prioritized yield	Goals set as explicit floors (3.5, then 3.8 μM). The completion-time band was designed from the data, and IGOR explained the floor-above-max failure mode.	Improved , but still needs expert iteration. Users can still set an unreachable floor; worth adding a guard.

D2 lesson	What happened in D3	Verdict
Hardware and mixing constraints not encoded	Volume constraint built from the setup sheet; droplet quantization; dispense-order convention saved to memory; failure-modes register used.	Much improved as a conversation. Not yet enforced in code (F3, F8).
Need to elicit expert feedback more naturally	Column-by-column schema review; "ask me again for each one"; memory entries for decisions.	Improved.
Mixing-order and precipitation issues	Surfaced from Slack and FM-01 and used in the advice. The redo still failed for an undetermined reason.	Partially. Root cause still open.

4. What to highlight, by audience

D3 devnote (evidence first) - Treat the workflow as the result. Show the plate-design-and-audit loop: errors caught per audit, and the droplet/volume feasibility work. - V2 and Fig. 4a: what drove the PAP yield recovery. RNasin removal and protein separation did not; the plasmid switch did. Report IGOR's partial misattribution openly. - V5 (the DOE "diagonal hole") and V6 (Yadav cross-reference and CP-chelation explanation). - Report the zero-signal plate and the redo honestly. State which changes were co-introduced and which were IGOR's recommendations. - Bench outcomes for the redo plate are pending. No yield claims until then.

Customer case study (outcomes) - "Caught N plate-breaking errors across 8 audits": swapped formula, well overflow, wrong DNA stock link, droplet-identical conditions, negative water. - From messy multi-plate history to a curated, modeled dataset in about 1.5 h, with decisions saved to memory. - Literature conditions turned into paste-ready plate rows in minutes.

Investor / Anthropic (capability) - V3: the agent explains the engine's behaviour from the evidence packet (unreachable goal floor; similarity filter; infeasible constraint). - The engine's top suggestions converged on the literature optimum on their own (verify first). - V7: IGOR refuses to trust a confounded slope. - Avoid "no hallucinations." This log contains LLM arithmetic errors and unlabelled mechanisms. Instead say: "engine numbers are computed; we are moving plate arithmetic into code next" (F3). That turns a weakness into a roadmap item.

5. Open questions to close before writing

1. What were the redo plate's results for its 22 kept experiments, especially GES-10, the Exp candidates, the Yadav rows and the controls? Which 4 were removed, and why? Also, what is the unit for "26"? The final redo layout in the log is 19 conditions in about 56 reaction wells (L14392), so 26 doesn't match either the conditions or the wells.
2. Was the flatline root cause found: tRNA lot, ribosome lot, dispense order, or the Mg floor?
3. Can we get `margarita_d3_PAP_r1-r2.events.json` ? Tool results are cut at 3,000 characters in this export. The full output is needed to verify the `Pmix_dispense` effect size, the GES evidence packets and the Yadav table extraction.

4. Is Margarita willing to give a quote? She gives no explicit praise in the log. The trust signal is behavioural: 8 audit uploads, coming back after the failure, and adopting the saved conventions.