Agentic Systems for Sample-Efficient Drug Formulation Design

Michael Craig [1] **Pauric Bannigan** [1] **Gary Tom** [1] **Yingshan Ma** [1] **Rémi Piché-Taillefer** [1] **Christine Allen** [1] **Riley Hickman** [1]

Abstract

Self-emulsifying drug delivery systems (SEDDS) can enhance the oral bioavailability of poorly soluble drugs, but their development remains constrained by how formulation space is explored in practice. The problem is not only that the space is large, but that existing workflows typically explore only a small, familiar, and conservative subset of it, restricting innovation and limiting opportunities to discover differentiated formulations. Traditional approaches based on excipient screening, pseudo-ternary phase diagrams, and design-of-experiments (DoE) are labor-intensive and better suited to local refinement than to broad, discovery-oriented exploration. Bayesian optimization (BO) has improved experimental efficiency, but still depends on multiple sequential rounds of wet-lab testing. In this work, we present a closed-loop agentic system, grounded in structured experimental data, that autonomously designs SEDDS formulations and executes them in an integrated miniaturized laboratory where preparation, physico-chemical characterization and dispersion assay run end-to-end, evaluating 128 formulations in just over a week of platform time. The system matches BO performance with 4× fewer experiments. An ablation removing the experimental data shows that this evidence accounts for a 45% improvement in mean drug solubilization, establishing that structured experimental evidence, rather than LLM parametric knowledge alone, drives the advantage. A second agentic batch improves formulation quality by 33% over the first batch's results, with no model retraining. Taken together, these results are a first step toward practical and scalable autonomous formulation design, enabling broader exploration of formulation space with substantially reduced experimental burden.

[1]Intrepid Labs, MaRS Centre, 661 University Ave, Toronto, ON M5G 0B7, Canada. Correspondence to: Michael Craig <michael@intrepidlabs.ai>.

Accepted by ICML 2026 AI for Science Workshop.

1. Introduction

A growing proportion of drug candidates emerging from modern discovery pipelines exhibit poor aqueous solubility, with estimates suggesting that over 70% of compounds in development are poorly water-soluble and classified in BCS Class II or IV (Lipinski et al., 2001; Butler & Dressman, 2010; Stegemann et al., 2023). This trend has been linked to the increasing molecular weight, lipophilicity, and structural complexity of contemporary small-molecule candidates (Shultz, 2019). For many of these molecules, especially BCS Class II compounds, oral absorption is limited by dissolution and solubilization in gastrointestinal fluids; for BCS Class IV compounds, low permeability can present an additional barrier (Amidon et al., 1995).

Self-emulsifying drug delivery systems (SEDDS) are isotropic lipid-based mixtures of oils, surfactants, and sometimes cosolvents that spontaneously emulsify upon dilution with gastrointestinal fluids. They have emerged as an important enabling formulation strategy for this growing class of compounds (Pouton, 1997; 2000; 2006). Several marketed oral products rely on SEDDS or related lipid formulations, including cyclosporine, ritonavir, and tipranavir (Savla et al., 2017), and the approach continues to attract interest for highly lipophilic compounds that are poorly served by conventional crystalline formulations (Cherniakov et al., 2015). As discovery programs continue to yield molecules with increasingly challenging developability profiles, efficient SEDDS development methods will become increasingly important for translating promising hits into viable oral drug products.

However, designing SEDDS requires navigating a high-dimensional, mixed-type combinatorial space where excipient identities, their relative proportions, and drug loading interact nonlinearly to determine self-emulsification efficiency, droplet size and polydispersity, and sustained drug solubilization (Pouton, 2006; Siepmann et al., 2019). In principle, this space is broad; in practice, systematic exploration of it is labor intensive. Conventional development workflows typically begin with screening drug solubility across candidate oils, surfactants, and cosolvents, followed by construction of pseudo-ternary phase diagrams to identify self-emulsifying regions and iterative bench evaluation

of dispersion behavior, precipitation resistance, and formulation stability under stress conditions (Pouton, 1997; 2000). Optimization within selected regions then commonly relies on classical design-of-experiments (DoE) approaches such as factorial, Box-Behnken, or mixture designs, combined with expert judgment (Gao et al., 2021). As a result, scientists often either accept slow, resource-intensive development or restrict exploration to previously used excipients and familiar formulation regions, limiting broader exploration and the discovery of novel formulations.

Bayesian optimization (BO) and machine learning methods have begun to address this. Surrogate-model-based approaches using Gaussian processes, gradient-boosted trees, or neural networks retrained after each wet-lab batch (Greenhill et al., 2020; Shields et al., 2021; Tom et al., 2024; Hickman et al., 2025) can replace broad screening with directed exploration, substantially reducing the number of experiments needed relative to classical DoE. Our BO baseline exemplifies this paradigm: a probabilistic model surrogate with iterative retraining converges on high-quality formulations in 5 **batches** (80 formulations), a significant improvement over conventional workflows that might require several hundred formulations across weeks of manual iteration, to achieve the same performance. Additional ML approaches have shown promise for predicting formulation outcomes from composition features (Bannigan et al., 2023; Hsueh et al., 2023; Bennett-Lenane et al., 2021a), and in silico screening methods can pre-filter candidates based on thermodynamic models (Brinkmann et al., 2020; Bennett-Lenane et al., 2021b). Yet even BO-based approaches require multiple sequential batches for surrogate retraining, and each new compound demands a fresh optimization campaign with no transfer of learned knowledge across APIs (Active Pharmaceutical Ingredients).

Large language models (LLMs) offer a qualitatively different approach. Trained on large scientific corpora, LLMs may capture broad knowledge relevant to SEDDS formulation, including excipient roles and properties, principles governing solubilization and self-emulsification, HLB-based design logic, and practical heuristics linking composition to drug loading, precipitation behavior, and formulation stability (Brown et al., 2020; Jablonka et al., 2023). Recent work has demonstrated that LLMs can serve as autonomous experimental planners in chemistry (Boiko et al., 2023) and can be augmented with domain-specific tools for chemical reasoning (Bran et al., 2024). In the SEDDS domain specifically, Craig et al. (2025) showed that frontier LLMs can act as virtual instruments for predicting formulation outcomes when provided with structured "formulation cards" encoding composition and physicochemical descriptors. However, LLM parametric knowledge is generic: it reflects published literature rather than the specific behaviors of particular instruments, excipient lots, and assay protocols in a given laboratory. We hypothesized that augmenting an LLM with structured in-house experimental data (excipient pair performance statistics, similar-API retrieval, and level/response trends computed from prior campaigns on the same instruments) could bridge this gap, enabling sample-efficient formulation design strategies that match or exceed iterative BO baselines.

We evaluate this hypothesis on Clofazimine, a BCS Class II phenazine dye derivative (MW 473, LogP $\approx$ 7.7, aqueous solubility $\approx$ 0.23 mg/L) whose extreme lipophilicity and high precipitation risk upon aqueous dilution make it a challenging SEDDS target (Shultz, 2019). The system operates as a closed-loop design-build-test workflow: the LLM designs each batch, formulations are prepared and assayed in the wet lab, and results feed back into the next design cycle, all without manual formulation selection by a human scientist. Our evaluation is itself enabled by a closed-loop autonomous formulation laboratory: liquid handling, dispersion, dissolution sampling, and analytical quantification are integrated such that each 16-formulation batch executes within a single working day. This throughput, roughly an order of magnitude beyond typical manual SEDDS development, is what makes head-to-head comparison of an LLM-guided and a BO-guided design loop tractable at all. Prior closed-loop work in autonomous chemistry has focused primarily on synthesis planning and execution (Boiko et al., 2023); extending autonomy to lipid-based formulation design-build-test cycles over a pharmaceutically relevant dissolution assay is itself a contribution of this work. Our contributions are:

1. A two-agent (proposer-reviewer) architecture that generates validated SEDDS formulations for direct wet-lab execution, matching the performance of a 4-batch BO baseline (64 formulations) in a single batch of 16 experiments.

2. An ablation study demonstrating that in-house experimental data accounts for a 45% improvement in mean drug solubilization, producing interpretable strategy shifts in surfactant loading, cosolvent usage, and oil selection.

3. Evidence of in-context batch-to-batch learning: a second agentic batch improves formulation quality by 33% through prompt-based integration of prior results, with no model weight updates.

4. Demonstration of a closed-loop autonomous SEDDS laboratory in which 128 formulations across four experimental conditions were designed, prepared, and assayed within a single campaign in 8 days.

2. Methods

2.1. Autonomous Formulation Laboratory

All four experimental conditions were conducted in a closed-loop autonomous laboratory integrating automated liquid handling systems with auxiliary instrumentation to enable reproducible formulation preparation, dispersion assays, time-resolved sampling, and analytical quantification. Experimental metadata, composition vectors, and endpoint measurements are automatically captured in a structured database, which serves both as the evaluation record and—specifically for Agentic Batch 2—as the in-context evidence layer for subsequent design cycles. This infrastructure enables the complete execution of a 16-formulation batch—including design, preparation, a three-hour dispersion assay, and post-sampling analysis—within approximately 24 hours, with high reproducibility and minimal human intervention. This represents a substantial reduction in experimental time, manual effort, and potential for human error relative to the multi-week cycle times typical of manual SEDDS development. The throughput afforded by this system is essential for the comparisons presented here: evaluating 128 formulations across four design strategies under consistent instruments, operators, and protocols would not be feasible using conventional manual workflows.

2.2. Agentic Architecture

We implemented a two-agent pipeline in which each wet-lab batch of 16 formulations is generated through sequential LLM calls:

Agent 1 (Proposer). The proposer receives the full assembled context window and generates 16 candidate formulations as a structured JSON array. Each formulation includes all 8 parameter values (Section 2.4) plus a free-text `reasoning` field explaining the scientific rationale. The proposer system prompt captures key principles of SEDDS formulation science, including self-emulsification thermodynamics, HLB-based surfactant selection, precipitation inhibitor mechanisms, and level-sum constraints, and instructs the model to maximize the probability of achieving the Target Product Profile (TPP).

For Batch 1 (exploration), the proposer balances conservative formulations with ambitious hypotheses while ensuring diversity across oil types, surfactant HLB ranges, and precipitation inhibitor options. For Batch 2 (exploitation), a dedicated system prompt shifts the objective: the proposer must analyze the Batch 1 dataset, identify top performers, and allocate $\sim$70–80% of slots to systematic perturbations around winners, 10–20% to cross-combinations, and $\sim$10% to justified exploration.

Agent 2 (Reviewer). The reviewer receives the same context plus the proposer's output and acts as a quality gate, evaluating each formulation for TPP achievability, chemical soundness, redundancy elimination, and precipitation risk. The reviewer may replace unsound or redundant formulations.

Both agents use a large language model (LLM) accessed via a unified inference gateway. The architecture is model-agnostic by design: any sufficiently capable LLM that supports structured output and large context windows can serve as the backbone (Craig et al., 2025).

2.3. In-House Experimental Data

The context window comprises three independently computed layers:

Layer 1: Excipient Performance Summary. Aggregate statistics derived from an internal SEDDS dataset comprising hundreds of thousands of instrument-level measurements across prior formulation campaigns, with all Clofazimine data held out to prevent leakage. This layer includes oil $\times$ surfactant pair performance summaries, top historical formulations, and level–response trends.

Layer 2: Similar-API Retrieval. A small set of the APIs most physicochemically similar to Clofazimine, identified by distance in a low-dimensional property space of standard drug-like descriptors, together with their top formulations.

Layer 3: Chemistry Context. Cheminformatics descriptors for Clofazimine and for all excipients in the design space, enabling first-principles reasoning through formulation generation and prediction.

For the ablation condition ("Untrained"), Layers 1 and 2 are disabled, meaning the model receives the drug profile, parameter space, and TPP, but no historical experimental data.

2.4. Design Space

The design space comprises 8 mixed-type parameters spanning drug loading, oil and surfactant identity and level, precipitation inhibitor identity and level, and cosolvent level, with a hard sum-to-one constraint on the five continuous level parameters. The resulting combinatorial space contains on the order of 10^8 discrete compositions before the sum-to-one constraint is applied.

2.5. Target Product Profile

Three dissolution objectives were directly measured (i.e., API concentration at 15, 60, and 180 minutes in the FeSSIF stage of the dispersion assay) and their corresponding dispersion fractions (calculated from the measured concentrations) were obtained, with the 180-min endpoint weighted 1.5$\times$ to reward sustained drug solubilization in the distal

gastrointestinal tract. We note that this weighting rewards maintained drug solubilization at later timepoints in the two-stage dispersion assay rather than transiently high drug concentration per se; the goal is to ensure that dissolved drug remains available for absorption throughout the full transit window captured by the assay.

2.6. Experimental Conditions

Four conditions were evaluated: (1) **Agentic B1** (trained, 16 formulations, 1 batch); (2) **Agentic B2** (trained + Batch 1 prior results, 16 formulations); (3) **Ablation** (untrained, no in-house data, 16 formulations); (4) **BO baseline** (Bayesian optimization with a gradient-boosted surrogate model and oracle, retrained after each batch, 80 formulations across 5 batches). The BO system represents an alternative to conventional SEDDS development: rather than relying on manual excipient screening, phase diagrams, and classical DoE, it uses a surrogate model to direct each successive batch toward regions of the design space predicted to yield the best dissolution performance. Both the agentic and the BO baseline approaches thus represent advances over traditional formulation workflows; the question is whether the agentic system can achieve comparable results with fewer experimental iterations. All four conditions used the same Clofazimine FeSSIF dispersion assay with identical instruments, operators, and protocols.

2.7. Evaluation Metrics

Concentration $AUC_{15–180}$ is the trapezoidal integral of drug concentration in FeSSIF from 15 to 180 minutes (units: mg·min/mL), providing a single measure of total drug exposure. **Formulation quality score** is the fraction of four quality gates passed per formulation: $AUC_{15–180} \geq 10$ mg·min/mL, $C_{180} \geq 0.05$ mg/mL, $F_{180} \geq 15\%$, and precipitation ratio (C_{180}/C_{15}) ≥ 0.85.

2.8. Interpretable Reasoning Traces

A distinctive feature of the agentic architecture is that both the proposer and reviewer generate free-text scientific reasoning alongside each design decision. Unlike surrogate-model-based optimization, where the mapping from model state to experimental proposal is opaque, the LLM produces a natural-language trace that can be audited by domain experts before wet-lab execution. Each batch generation yields three categories of reasoning output: (i) a *strategy summary* in which the proposer articulates its overall design logic, including API-specific risk assessment, excipient selection rationale, and batch-level resource allocation; (ii) *per-formulation reasoning* explaining the scientific hypothesis behind each of the 16 compositions; and (iii) a *reviewer critique* that evaluates chemical soundness, redundancy, and TPP achievability, with replacement formulations where warranted.

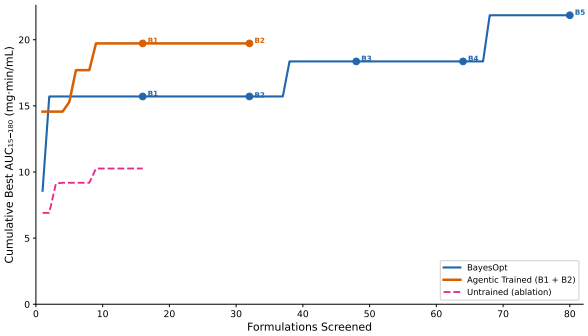

Figure 1. Cumulative best $AUC_{15–180}$ vs. formulations screened across all four conditions. Agentic B1 (amber) discovers a formulation achieving AUC = 19.7 within 16 experiments (1 batch), matching the BO baseline (blue) through its first four batches. BO surpasses the agentic system only at batch 5 (21.8 mg·min/mL after 80 experiments, 5 batches). The untrained ablation (dashed red) is lower at 10.3. Agentic B2 maintains the peak through 32 total formulations. Batch markers show cumulative best at each boundary.

3. Results

3.1. Sample-Efficient Formulation Discovery

Figure 1 shows the cumulative best concentration $AUC_{15–180}$ as a function of formulations screened. The agentic system with in-house data discovered a formulation achieving AUC = 19.7 mg·min/mL within its first batch of 16 formulations (1 batch), matching or exceeding the BO baseline through its first four batches (cumulative best 18.3 after 64 formulations). BO ultimately surpassed the agentic system only at its fifth batch (AUC = 21.8). The agentic system therefore reached BO-B4-equivalent performance using 4× fewer experiments and four fewer batches of wet-lab work. Notably, the BO baseline is itself a substantial advance over conventional SEDDS development: its surrogate-model-guided search converges on high-quality formulations in 5 batches, whereas traditional workflows based on excipient screening, phase diagram construction, and classical DoE optimization typically require considerably more formulations and weeks of manual iteration (Pouton, 2000; Gao et al., 2021). The agentic system's achievement is therefore not merely matching a weak baseline but reaching, in a single batch, the performance level that a state-of-the-art automated optimization pipeline requires four additional batches to achieve.

The BO baseline exhibited the expected learning curve: its first two batches achieved a cumulative best of 15.7, with a substantial jump at B3 (18.3) as the retrained surrogate identified high-performing regions, eventually reaching 21.8 after the fifth batch. The agentic system's first batch matched BO through B4 and was surpassed only at the fifth

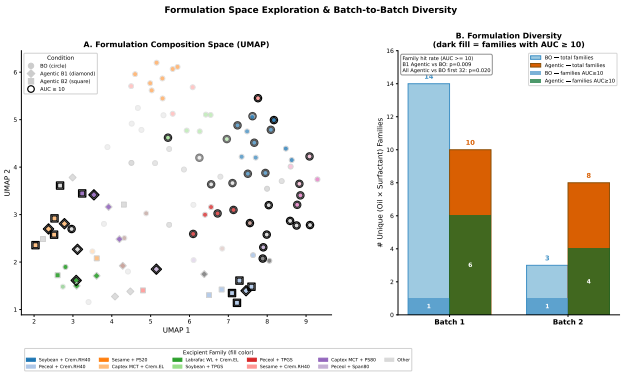

Figure 2. **Formulation space exploration and batch-to-batch diversity. (A)** UMAP projection of all evaluated formulations in composition space. Circles: BO baseline (80 formulations across 5 batches); diamonds: Agentic Batch 1; squares: Agentic Batch 2. Fill color indicates excipient family (oil × surfactant pair); grey denotes less-frequent "Other" families. Bold black outlines mark high-performing formulations ($AUC_{15\text{–}180} \geq 10$ mg·min/mL). **(B)** Number of unique (oil × surfactant) excipient families evaluated per batch. Light bars: total families; dark-filled bars: families containing at least one formulation with AUC $\geq$ 10.

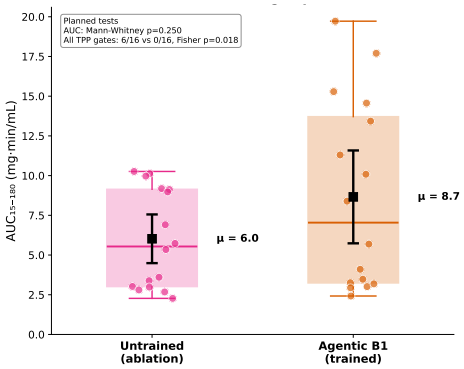

Figure 3. Impact of in-house SEDDS data on first-batch AUC. Same agent, same compound, $n = 16$ each. The trained condition achieves 45% higher mean drug exposure. Black bars: mean ± 95% bootstrap CI.

BO batch.

The untrained ablation reached a cumulative best AUC of only 10.3 mg·min/mL, compared with 19.7 for the trained condition. This establishes that the LLM's parametric knowledge of formulation science, while sufficient to produce structurally valid SEDDS compositions, is insufficient to identify high-performing formulations for a specific API on specific instruments. Agentic B2 did not exceed the peak AUC of 19.7 through 32 total formulations; however, the mean AUC increased from 8.7 to 10 (Figure S1). Per-endpoint comparisons can also be observed in Figure S2.

3.2. Formulation Space Exploration and Batch-to-Batch Diversity

Beyond raw performance, an important question is whether the agentic system and the BO baseline explore comparable regions of the formulation space. Figure 2A shows a UMAP projection of every evaluated formulation, computed over the full 8-parameter composition vector. The agentic batches (diamonds, squares) and the BO batches (circles) occupy largely distinct regions of the projection: BO concentrates its 80 formulations in a band dominated by Soybean + Cremophor RH40 and Sesame + PS20 families, while the agentic batches populate regions dominated by Labrafac WL + Cremophor EL, Peceol + Span 80, and Captex MCT + Cremophor EL families. High-performing formulations (bold outlines, AUC $\geq$ 10) appear in both regions, indicating that the two approaches are discovering distinct high-performance formulations rather than converging on a single shared optimum.

Figure 2B quantifies this at the excipient-family level. In Batch 1, BO evaluates 14 unique (oil × surfactant) families of which only 1 contains a formulation with AUC $\geq$ 10; the agentic system evaluates 10 families and produces 6 productive ones. In Batch 2, BO evaluates 3 distinct families (1 productive), whereas the agentic system still evaluates 8 families and produces 4 productive ones. On a per-formulation basis, the fraction of experiments landing in a productive excipient family is higher for the agentic system in both batches (6/16 and 4/16) than for the corresponding BO batches (1/16 each), though we note that absolute counts are small and these ratios should be interpreted as descriptive rather than statistically rigorous. This indicates that the agentic system is not simply a more sample-efficient version of BO operating over the same chemistries, but is exploring a structurally distinct region of formulation space and producing a higher density of chemically diverse, high-performing formulations per experiment.

3.3. In-House Data Improves Performance and Shifts Formulation Strategy

Figure 3 shows the AUC improvement from in-house data: the trained condition achieves mean AUC = 8.7 vs. 6.0 for untrained (+45%). Beyond raw performance, the in-house data induced qualitative changes in formulation strategy. Surfactant loading increased from a mean of 46% (untrained) to 57% (trained), reflecting instrument-specific evidence that higher surfactant-to-oil ratios produce smaller droplets and more robust self-emulsification. Cosolvent loading decreased sharply from 5.0% to 1.8%, as the historical data contained examples of cosolvent-induced precipitation upon aqueous dilution, a failure mode that is particularly relevant for highly lipophilic compounds like Clofazimine. Oil selection narrowed from 8 types to 4, converging on the medium-chain lipids that were top performers in the historical dataset. These shifts represent a move from generic,

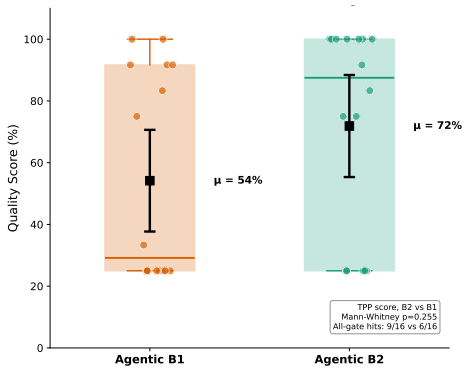

Figure 4. SEDDS quality score (percentage of four quality gates passed per formulation) from Batch 1 to Batch 2. Quality improves from 54% to 72% as the system shifts from exploration to exploitation mode via in-context integration of Batch 1 results. Black bars: mean ± 95% bootstrap CI.

textbook-driven formulation design to a focused strategy calibrated to the specific instruments and assay conditions in use.

3.4. Batch-to-Batch Learning Improves Formulation Quality

Figure 4 shows the quality score from Batch 1 (54%) to Batch 2 (72%). The Batch 1 distribution had many points at 25% (1 of 4 gates passed), reflecting the exploratory first batch. By Batch 2, the majority of formulations passed all four quality gates, with a median of 100%.

This improvement results from the Batch-2 prompt architecture. When Batch 1 results are injected, the proposer receives the full master sheet scored against the TPP. The Batch-2 system prompt instructs: (i) identify top performers and shared compositional features; (ii) allocate 70–80% of slots to systematic perturbations around winners; (iii) reserve the remainder for cross-combinations or justified exploration. The reviewer enforces this by protecting exploitation moves.

The shift in quality demonstrates learning within a single API campaign, not through model weight updates but through in-context learning over structured experimental evidence.

3.5. Qualitative Analysis of LLM Reasoning

Beyond quantitative metrics, the proposer's reasoning traces illustrate how the system translates structured data into formulation strategy. We highlight two excerpts that demonstrate the shift from data-grounded design (Batch 1) to evidence-driven refinement (Batch 2).

Batch 1: Instrument-calibrated design. When designing the first batch, the proposer grounded its excipient selection directly in the historical performance data from Layer 1, rather than relying on generic formulation heuristics:

Proposer — Batch 1 Strategy

Model Explanation

"I anchor on oil/surfactant pairs with proven small droplet sizes on your instruments: captex_355_mct + cremophor_el (median ∼127 nm) and captex_355_mct + polysorbate_80 (∼119 nm) [. . .] I use high surfactant fractions (0.55–0.70) in most formulas to push droplet size down (historical trend: 60–80% surfactant → ∼116 nm median)."

The model cites instrument-specific median droplet sizes and level–response trends computed from the Layer 1 historical dataset, translating quantitative evidence into compositional decisions. This is the instrument calibration that the untrained ablation lacks: without Layer 1, the ablation proposer defaulted to textbook HLB reasoning and produced the lower mean AUC underlying the 45% improvement.

Batch 2: In-context learning from experimental results. When the Batch 1 dispersion data was injected into the Batch 2 context window, the proposer identified a counter-intuitive top performer and generated a mechanistic hypothesis to guide the next round of experiments:

Proposer — Batch 2 Strategy

Model Explanation

"Clear winner chemistry: Peceol + Span 80 + VA64 (Sample 6) gave very strong late fraction (0.935 @180m) despite low initial fraction—suggesting a digestion/mixed-micelle enabled 'ramp-up' mechanism worth exploiting (surprising vs typical 'high-HLB surfactant only' assumptions)."

This excerpt illustrates three aspects of the system's understanding: (i) identification of the top performer from scored Batch 1 results; (ii) recognition that the outcome contradicts conventional SEDDS design logic, since a low-HLB surfactant (Span 80, HLB 4.3) outperformed high-HLB alternatives at the critical 180-minute endpoint; and (iii) generation of a mechanistic hypothesis, that lipase-mediated digestion of Peceol produces mixed micelles that progressively solubilize the drug, to explain the observation. We note that the specific mechanism invoked here reflects a generic SEDDS literature assumption: lipase-mediated digestion applies to lipolysis-coupled assays, whereas the dispersion assay used in this work is lipase-free. The compositional move was nonetheless supported by the empirical Batch 1 results, and this mismatch points to a concrete prompt augmentation improvement, namely supplying explicit assay context alongside the excipient and chemistry layers. The system allocated 69% of Batch 2 slots to systematic perturbations around this and two other winners, executing the explore-to-exploit transition through prompt architecture alone. This shift from data-grounded design in Batch 1 to

hypothesis-driven refinement in Batch 2 is the mechanism underlying the quality improvement reported in Figure 4.

4. Discussion

Our results demonstrate that an agentic LLM architecture, augmented with structured in-house experimental data, can match the performance of an established Bayesian optimization pipeline while using 4× fewer experiments. Several aspects of these findings merit discussion.

The role of in-house data. The 45% AUC improvement from in-house data, combined with interpretable strategy shifts, establishes that the value proposition is not the LLM alone but the combination of LLM reasoning with structured experimental evidence. The untrained model produces chemically valid formulations, consistent with demonstrations that LLMs encode substantial pharmaceutical knowledge (Jablonka et al., 2023), but lack the instrument-specific calibration needed to identify high-performing compositions for a particular API on particular instruments. This finding parallels observations in ML-guided formulation design more broadly, where domain-specific training data consistently outweighs model architecture choices (Bannigan et al., 2023; Gao et al., 2021). The strategy shifts we observe (higher surfactant loading, reduced cosolvent, focused oil selection) are not arbitrary but align with known SEDDS design principles (Pouton, 2006; Cherniakov et al., 2015): the in-house data effectively encodes instrument-specific instantiations of general formulation heuristics that a traditional formulation scientist would acquire through months or years of bench experience.

Diversity as a complement to sample efficiency. The performance comparison with BO understates a second advantage of the agentic approach. As shown in Figure 2, the agentic and BO systems are not competing within the same chemistry; they explore largely separate regions of the formulation space. BO concentrates its budget inside a narrow, high-expected-value region identified by its surrogate, behavior that is optimal under a pure exploitation objective but that tends to produce many near-duplicates of a single chemical family. The agentic system, guided by pair-level historical performance and similar-API retrieval rather than a pointwise surrogate, distributes its budget across a wider set of excipient families and recovers high-performing formulations in several of them. For pharmaceutical development, where patent differentiation, supplier risk, excipient availability, and regulatory precedent all matter alongside raw *in vitro* performance, this diversity is a first-class outcome rather than a by-product. A campaign that identifies one top performer together with several chemically distinct near-top performers can be operationally more valuable than one that identifies multiple variants of a single composition, even when single-point peak performance is comparable.

Advances over both traditional and BO-based workflows. It is worth emphasizing that the BO baseline is itself a significant advance over conventional SEDDS development. Traditional workflows involving excipient solubility screening, pseudo-ternary phase diagram construction, classical DoE optimization, and iterative bench testing (Pouton, 1997; 2000) may require hundreds of formulations across weeks of manual effort for a single compound. The BO baseline compresses this to 80 formulations over 5 batches by replacing broad screening with surrogate-model-directed search. The agentic system achieves a further compression: matching four BO batches (64 formulations) of performance in 16 formulations and 1 batch, and reaching a peak (19.7) that BO only exceeds at its fifth and final batch. This represents a progression from traditional (weeks, hundreds of formulations) to BO-guided (5 days, 80 formulations) to LLM-guided (1 day, 16 formulations to reach BO-B4-equivalent performance), where each step reduces experimental burden by roughly an order of magnitude while maintaining or improving formulation quality.

In-context vs. parametric learning. The quality improvement from B1 to B2 occurs entirely through prompt engineering, without gradient updates or fine-tuning. This in-context learning mechanism (Brown et al., 2020) is particularly valuable in pharmaceutical settings where data is scarce and campaign-specific. The approach contrasts with conventional surrogate-model-based optimization (Greenhill et al., 2020; Shields et al., 2021), which requires retraining a statistical model after each batch. Our system achieves a similar explore-to-exploit transition through prompt architecture alone, with the Batch-2 system prompt and injected prior results serving the functional role of a retrained surrogate.

Comparison to prior ML approaches for SEDDS. Previous computational approaches to SEDDS design have focused on predictive modeling, training neural networks or thermodynamic models to predict outcomes from formulation composition (Bennett-Lenane et al., 2021a;b; Brinkmann et al., 2020), or on building curated datasets for downstream modeling (Zaslavsky & Allen, 2023). These approaches require substantial labeled data per API and typically serve as screening filters rather than autonomous design agents. The agentic architecture presented here operates at a different level of abstraction: it uses the LLM as a hypothesis generator that reasons over structured evidence to propose complete experiments, rather than as a property predictor that scores individual candidates. This enables the system to incorporate qualitative reasoning (e.g., precipitation risk assessment, HLB compatibility) alongside quantitative evidence, and to produce novel formulations rather than interpolating within a trained feature space.

Relation to LLM-guided experimental design. Our work

extends the emerging paradigm of LLM-guided autonomous experimentation (Boiko et al., 2023) to pharmaceutical formulation. While Boiko et al. (2023) demonstrated autonomous chemical synthesis planning, our system operates in a regime where the LLM replaces rather than augments the optimization loop entirely. The proposer-reviewer architecture provides a form of self-critique that reduces the rate of chemically unsound proposals, analogous to the tool-augmented reasoning in Bran et al. (2024) but operating through natural-language scientific review rather than computational verification. Building on our prior finding that frontier LLMs can predict SEDDS outcomes with accuracy that scales with in-context example count (Craig et al., 2025), the present work demonstrates that this predictive capability translates into effective experimental design when embedded in an agentic pipeline with structured context.

Model Explanations. The model explanations provide direct evidence of this mechanism. In Batch 1, the proposer cited instrument-specific droplet statistics to justify excipient selection; in Batch 2, it identified a counterintuitive result, proposed a mechanistic explanation, and reallocated the experimental budget accordingly (Section 3.5). This capacity for structured scientific reasoning, articulating hypotheses, interpreting surprises, and translating evidence into design changes, distinguishes the agentic approach from surrogate-model optimization, where the mapping from model state to experimental proposal is opaque. The chain-of-thought traces (Wei et al., 2022) also serve as auditable scientific documentation, providing a transparency mechanism that is absent from black-box approaches and increasingly valued in pharmaceutical development (Caldas Ramos et al., 2025).

Limitations and future directions. This study evaluates a single API (Clofazimine) in a single dispersion assay. Generalizability to structurally diverse compounds, alternative SEDDS platforms, and different assay conditions remain to be demonstrated. The sample sizes and limited number of batches limit statistical power for pairwise comparisons.

Additionally, the agentic system requires access to structured historical data from prior campaigns on the same instruments; laboratories without such data would need to begin with the untrained condition and accumulate evidence over successive compounds. A natural next step is fine-tuning a foundation model on the accumulated in-house SEDDS corpus and integrating it into the agentic pipeline, which could internalize instrument-specific patterns as learned weights rather than relying entirely on in-context presentation of evidence. Because the architecture is backbone-agnostic, such a domain-adapted model could be deployed as a drop-in replacement for the current backbone, potentially improving both calibration and the quality of generated batches while enabling fully on-premise deployment for sensitive pharmaceutical data.

Finally, while the reasoning traces presented in Section 3.5 provide valuable transparency into the system's design logic, access to such traces depends on the deployment configuration. Commercial models may expose chain-of-thought outputs to varying degrees depending on the provider and pricing tier, and some configurations return only final outputs without intermediate reasoning. Self-hosted or open-weight models offer full access to reasoning traces by default, which is one additional motivation for the fine-tuning direction described above. The interpretability advantages documented here should therefore be understood as achievable under configurations that expose model reasoning, rather than as an inherent guarantee of all possible deployment modes.

Broader implications. The agentic architecture is not specific to SEDDS: any formulation design problem with a well-defined parameter space, measurable objectives, and available historical data could benefit from the same proposer-reviewer/context pipeline. Recent work on ML-guided design of long-acting injectables (Bannigan et al., 2023), LNP formulations (Xu et al., 2024), and ocular drug delivery systems (Hsueh et al., 2023) suggests that data-driven formulation optimization is broadly applicable across delivery modalities. On the spectrum from passive tool to autonomous scientist, our system occupies an intermediate position: it autonomously generates and quality-gates experimental designs, but relies on human execution in the wet lab and human judgment for campaign-level decisions (e.g., whether to run a third batch).

5. Conclusion

We presented an agentic LLM architecture for SEDDS formulation design that matches the performance of a 4-batch BO baseline (64 formulations) in a single batch of 16 experiments, and is surpassed by BO only at its fifth batch. The system's effectiveness depends critically on structured in-house experimental data (+45% AUC improvement), produces interpretable formulation strategy shifts, and demonstrates in-context batch-to-batch learning without model retraining. These results suggest that LLM-guided formulation design, grounded in laboratory-specific experimental evidence, offers a practical path toward more sample-efficient pharmaceutical development. These gains were obtained in a closed-loop autonomous laboratory; the combination of agentic design and physical automation, rather than either component in isolation, is what makes single-day formulation discovery a tractable avenue for future development.

References

Amidon, G. L., Lennernas, H., Shah, V. P., and Crison, J. R. A theoretical basis for a biopharma-

ceutic drug classification: the correlation of in vitro drug product dissolution and in vivo bioavailability. *Pharmaceutical Research*, 12:413–420, 1995. doi:10.1023/A:1016212804288.

Bannigan, P. et al. Machine learning models to accelerate the design of polymeric long-acting injectables. *Nature Communications*, 14:35, 2023. doi:10.1038/s41467-022-35343-w.

Bennett-Lenane, H. et al. Artificial neural networks to predict the apparent degree of supersaturation in supersaturated lipid-based formulations: A pilot study. *Pharmaceutics*, 13:1398, 2021a. doi:10.3390/pharmaceutics13091398.

Bennett-Lenane, H. et al. Applying computational predictions of biorelevant solubility ratio upon self-emulsifying lipid-based formulations dispersion to predict dose number. *Journal of Pharmaceutical Sciences*, 110:164–175, 2021b. doi:10.1016/j.xphs.2020.10.055.

Boiko, D. A., MacKnight, R., Kline, B., and Gomes, G. Autonomous chemical research with large language models. *Nature*, 624:570–578, 2023. doi:10.1038/s41586-023-06792-0.

Bran, A. M., Cox, S., Schilter, O., Baldassari, C., White, A. D., and Schwaller, P. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6:525–535, 2024. doi:10.1038/s42256-024-00832-8.

Brinkmann, J., Exner, L., Luebbert, C., and Sadowski, G. In-silico screening of lipid-based drug delivery systems. *Pharmaceutical Research*, 37:249, 2020. doi:10.1007/s11095-020-02955-0.

Brown, T. B. et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020. doi:10.48550/arXiv.2005.14165.

Butler, J. M. and Dressman, J. B. The developability classification system: application of biopharmaceutics concepts to formulation development. *Journal of Pharmaceutical Sciences*, 99:4940–4954, 2010. doi:10.1002/jps.22217.

Caldas Ramos, M., Collison, C. J., and White, A. D. A review of large language models and autonomous agents in chemistry. *Chemical Science*, 16:2514–2572, 2025. doi:10.1039/D4SC03921A.

Cherniakov, I., Domb, A. J., and Hoffman, A. Self-nano-emulsifying drug delivery systems: an update of the biopharmaceutical aspects. *Expert Opinion on Drug Delivery*, 12:1121–1133, 2015. doi:10.1517/17425247.2015.999038.

Craig, M., Tom, G., Bannigan, P., Allen, C., and Hickman, R. LLMs as virtual instruments for drug formulation. In *NeurIPS 2025 Workshop on AI Virtual Cells and Instruments*, 2025. URL https://openreview.net/forum?id=U9K6UJQrGN.

Gao, H. et al. Integrated in silico formulation design of self-emulsifying drug delivery systems. *Acta Pharmaceutica Sinica B*, 11:3585–3594, 2021. doi:10.1016/j.apsb.2021.04.017.

Greenhill, S., Rana, S., Gupta, S., Vellanki, P., and Venkatesh, S. Bayesian optimization for adaptive experimental design: A review. *IEEE Access*, 8:13937–13948, 2020. doi:10.1109/ACCESS.2020.2966228.

Hickman, R. J., Sim, M., Pablo-García, S., Tom, G., Woolhouse, I., Hao, H., Bao, Z., Bannigan, P., Allen, C., Aldeghi, M., et al. Atlas: a brain for self-driving laboratories. *Digital Discovery*, 4(4):1006–1029, 2025. doi:10.1039/D4DD00115J.

Hsueh, H. T. et al. Machine learning-driven multifunctional peptide engineering for sustained ocular drug delivery. *Nature Communications*, 14:2509, 2023. doi:10.1038/s41467-023-38056-w.

Jablonka, K. M. et al. 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon. *Digital Discovery*, 2:1233–1250, 2023. doi:10.1039/D3DD00113J.

Lipinski, C. A., Lombardo, F., Dominy, B. W., and Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 46:3–26, 2001. doi:10.1016/S0169-409X(00)00129-0.

Pouton, C. W. Formulation of self-emulsifying drug delivery systems. *Advanced Drug Delivery Reviews*, 25:47–58, 1997. doi:10.1016/S0169-409X(96)00490-5.

Pouton, C. W. Lipid formulations for oral administration of drugs: non-emulsifying, self-emulsifying and self-microemulsifying drug delivery systems. *European Journal of Pharmaceutical Sciences*, 11:S93–S98, 2000. doi:10.1016/S0928-0987(00)00167-6.

Pouton, C. W. Formulation of poorly water-soluble drugs for oral administration: physicochemical and physiological issues and the lipid formulation classification system. *European Journal of Pharmaceutical Sciences*, 29(3-4):278–287, 2006. doi:10.1016/j.ejps.2006.04.016.

Savla, R., Browne, J., Plassat, V., Wasan, K. M., and Wasan, E. K. Review and analysis of FDA approved drugs using lipid-based formulations. *Drug Development and Industrial Pharmacy*, 43:1743–1758, 2017. doi:10.1080/03639045.2017.1342654.

Shields, B. J., Stevens, J., Li, J., Parasram, M., Damber, F., Janey, J. M., Adams, R. P., and Doyle, A. G. Bayesian reaction optimization as a tool for chemical synthesis. *Nature*, 590:89–96, 2021. doi:10.1038/s41586-021-03213-y.

Shultz, M. D. Two decades under the influence of the Rule of Five and the changing properties of approved oral drugs. *Journal of Medicinal Chemistry*, 62:1701–1714, 2019. doi:10.1021/acs.jmedchem.8b00686.

Siepmann, J. et al. Lipids and polymers in pharmaceutical technology: Lifelong companions. *International Journal of Pharmaceutics*, 558:128–142, 2019. doi:10.1016/j.ijpharm.2018.12.080.

Stegemann, S. et al. Trends in oral small-molecule drug discovery and product development based on product launches before and after the Rule of Five. *Drug Discovery Today*, 28:103344, 2023. doi:10.1016/j.drudis.2022.103344.

Tom, G., Schmid, S. P., Baird, S. G., Cao, Y., Darvish, K., Hao, H., Lo, S., Pablo-García, S., Rajaonson, E. M., Skreta, M., et al. Self-driving laboratories for chemistry and materials science. *Chemical Reviews*, 124(16):9633–9732, 2024. doi:10.1021/acs.chemrev.4c00055.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. doi:10.48550/arXiv.2201.11903.

Xu, Y., Ma, S., Cui, H., Chen, J., Xu, S., Gong, F., Golubovic, A., Zhou, M., Wang, K. C., Varley, A., Lu, R. X. Z., Wang, B., and Li, B. Agile platform: a deep learning powered approach to accelerate lnp development for mrna delivery. *Nature Communications*, 15:6305, 2024. doi:10.1038/s41467-024-50619-z.

Zaslavsky, J. and Allen, C. A dataset of formulation compositions for self-emulsifying drug delivery systems. *Scientific Data*, 10:914, 2023. doi:10.1038/s41597-023-02812-w.

A. Bayesian Optimization Baseline

The BO baseline is a batched, surrogate-guided optimizer over the same parameter design space described in Section 2.4. A gradient-boosted regression surrogate is fit to all previously observed (composition, $\mathrm{AUC}_{15\text{–}180}$) pairs and retrained from scratch after each batch. Each batch of 16 formulations is selected by scoring candidate compositions with an acquisition function that balances predicted performance with predictive uncertainty, estimated from an ensemble of surrogate models trained on bootstrap resamples. Candidates violating the sum-to-one constraint are rejected at sampling time. The first batch is seeded from a space-filling design over the parameter space; subsequent batches are drawn from the acquisition-ranked pool with a diversity filter to avoid near-duplicate proposals within a batch.

B. Statistical Analysis and Figure Generation

All statistical analyses were performed at the formulation level. Each candidate formulation was measured in triplicate, and replicate measurements were averaged before hypothesis testing to avoid treating technical replicates as independent experimental units. For each formulation, FeSSIF exposure was summarized as the area under the concentration–time curve from 15 to 180 minutes using the linear trapezoidal rule:

$$\mathrm{AUC}_{15\text{–}180} = \sum_{i=1}^{n-1} \frac{t_{i+1} - t_i}{2} \left(C_i + C_{i+1}\right), \qquad (1)$$

evaluated at $(t_1, t_2, t_3) = (15, 60, 180)$ minutes. With concentrations in mg/mL, this yields:

$$\mathrm{AUC}_{15\text{–}180} = \tfrac{45}{2}\left(C_{15} + C_{60}\right) + \tfrac{120}{2}\left(C_{60} + C_{180}\right), \quad (2)$$

in units of mg·min/mL.

In addition to mean AUC, we evaluated whether each formulation met a target product profile (TPP). A formulation was considered to satisfy the full TPP if it met all four gates:

- $\mathrm{AUC}_{15\text{–}180} \geq 10$ mg·min/mL
- $C_{180} \geq 0.05$ mg/mL
- Fraction solubilized at 180 minutes ≥ 0.15
- Precipitation ratio $C_{180}/C_{15} \geq 0.85$

We also computed a continuous TPP score as the fraction of these four gates satisfied by each formulation.

Because formulation discovery is not solely an average-performance problem, we also evaluated diversity-aware endpoints. We defined an excipient family as a unique oil

$\times$ surfactant pair. For each family, we retained the best-performing formulation by mean AUC. Diverse hit yield was defined as the number of unique families whose best formulation exceeded $\mathrm{AUC} \geq 10$ mg·min/mL.

Planned comparisons were chosen to reflect the experimental claims: trained agentic versus untrained ablation, trained agentic Batch 1 versus BO Batch 1 at equal first-batch budget, and Agentic Batch 2 versus Agentic Batch 1 to assess batch-to-batch improvement. For continuous endpoints, including mean AUC and TPP score, we used two-sided Mann–Whitney U tests. For binary endpoints, including full-TPP success and family-level hit rates, we used two-sided Fisher exact tests. Bootstrap 95% confidence intervals on group means were computed by resampling formulations with replacement ($B = 10{,}000$ resamples).

C. Per-Batch AUC Distributions Across All Conditions

Figure S1 reports the full per-batch $\mathrm{AUC}_{15\text{–}180}$ distribution for every condition, complementing the cumulative-best trajectory in Figure 1 and the first-batch comparison in Figure 3. Each point is one formulation ($n = 16$ per batch); box plots show median, interquartile range, and 1.5$\times$ IQR whiskers; black squares mark the mean.

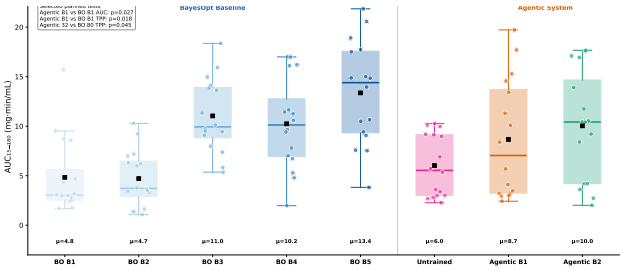

Figure S1. **Per-batch $\mathrm{AUC}_{15\text{–}180}$ distributions across all conditions.** BO baseline (blue, left of divider) across five sequential batches, and the agentic system with its untrained ablation and two agentic batches (right of divider). Each point is a single formulation ($n = 16$ per batch, triplicate replicates per formulation); horizontal lines denote medians, box edges the interquartile range, whiskers extend to 1.5$\times$ IQR, and black squares mark batch means (annotated below each box). The BO baseline shows the expected surrogate-model learning curve: batch means rise from $\mu = 4.8$ (B1) and $\mu = 4.7$ (B2) to $\mu = 11.0$ (B3) once the retrained surrogate identifies high-performing regions, settling at $\mu = 13.4$ by B5 with a top score of 21.8. The untrained ablation ($\mu = 6.0$) and Agentic B1 ($\mu = 8.7$) sit between BO B1–B2 and BO B3–B4 in mean performance; Agentic B2 reaches $\mu = 10.0$, comparable to BO B3–B4 ($\mu = 11.0$ and 10.2) in mean but not matching BO B5. The cumulative-best advantage of the agentic system reported in Figure 1 therefore reflects the upper tail of the Agentic B1 distribution rather than a mean-level parity across all formulations in the batch.

D. Raw Replicate Data

The following figures show mean $\pm$ standard deviation across experimental replicates for all measured endpoints. Each formulation was assayed in triplicate. Low standard deviations confirm that observed performance differences between conditions reflect formulation design quality rather than measurement noise.

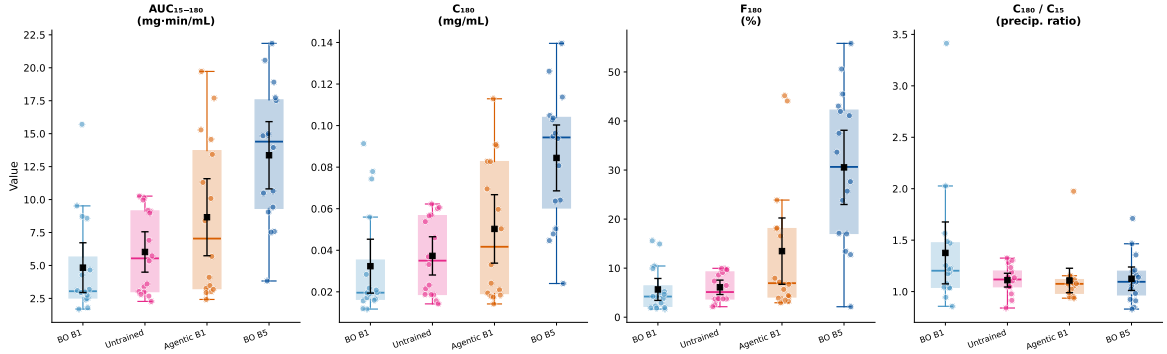

Figure S2. Per-endpoint comparison showing first-batch performance (BO B1, Untrained, Agentic B1; each $n = 16$) alongside the BO baseline's final batch (B5) after four rounds of surrogate retraining. Agentic B1 matches BO B5 levels on late-timepoint and total exposure metrics from its first batch.

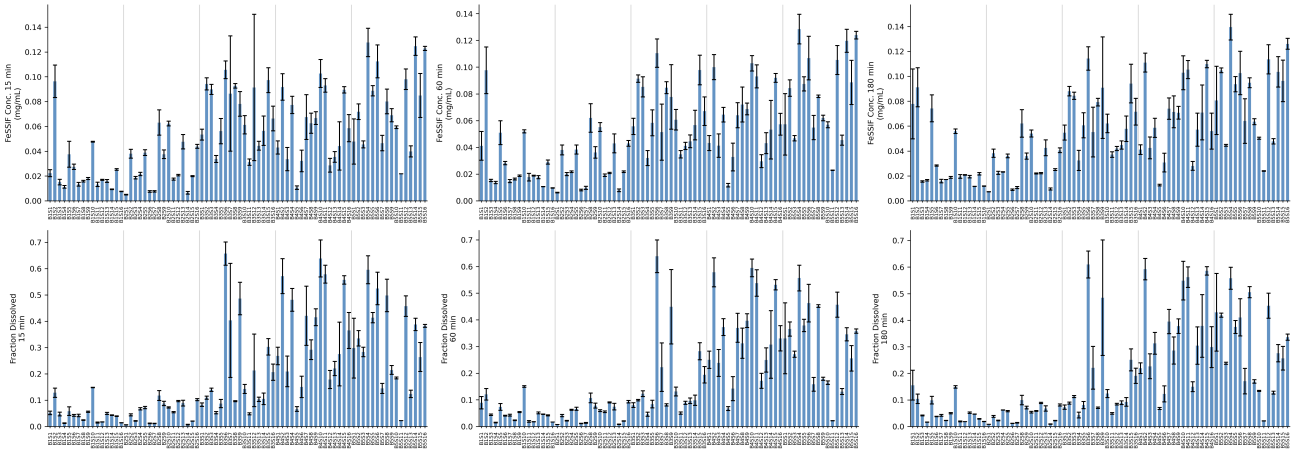

Figure S3. **Bayesian Optimization Baseline: raw replicate data.** Mean ± SD across replicates for all dissolution endpoints across all five batches. Error bars represent ±1 SD ($n = 3$ replicates per formulation).

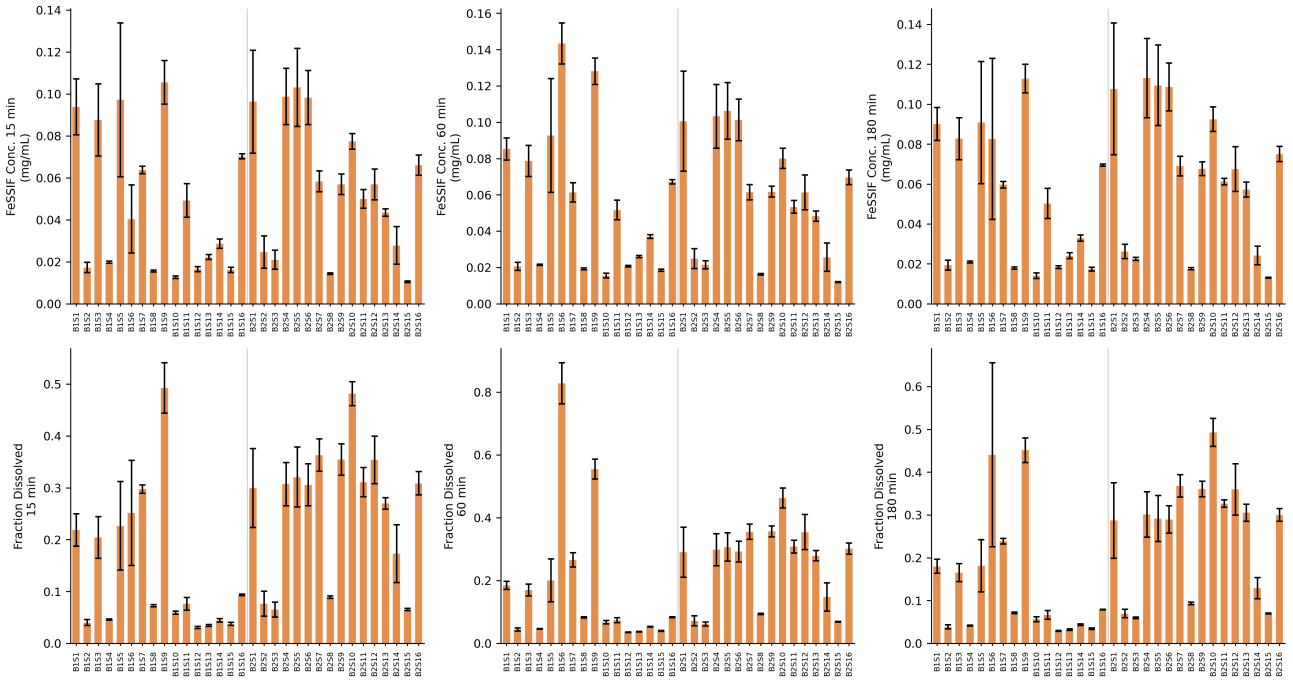

Figure S4. **Agentic Trained: raw replicate data.** Mean ± SD across replicates for all dissolution endpoints. Error bars represent ±1 SD ($n = 3$ replicates per formulation).

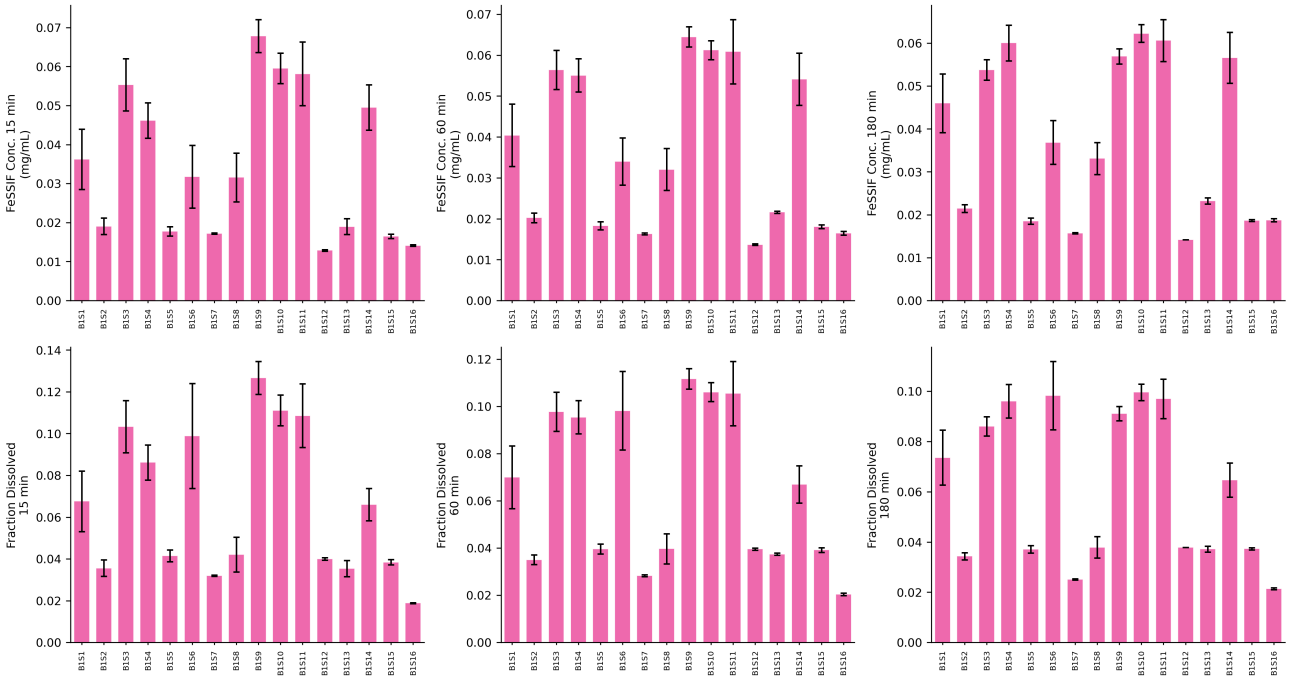

Figure S5. **Agentic Untrained (Ablation): raw replicate data.** Mean ± SD across replicates for all dissolution endpoints. This condition received no in-house experimental data. Error bars represent ±1 SD ($n = 3$ replicates per formulation).