A Domain-Grounded Agentic Assistant for Lipid Nanoparticle Formulation Research

Guy D. Rosin
guy@mana.bio
Mana.bio
Tel Aviv, Israel

Pablo Klijnjan
pablo@mana.bio
Mana.bio
Tel Aviv, Israel

Maya Himan
maya@mana.bio
Mana.bio
Tel Aviv, Israel

Daniel Stuczynski
daniel@mana.bio
Mana.bio
Tel Aviv, Israel

Nir Suissa
nir@mana.bio
Mana.bio
Tel Aviv, Israel

Lee Danan Goldfryd
lee@mana.bio
Mana.bio
Tel Aviv, Israel

Abstract

Lipid nanoparticles (LNPs) are multi-component nanoscale carriers used to protect and deliver RNA therapeutics into cells. Designing effective LNP formulations is difficult: evidence is scattered across papers and patents, experimental contexts are inconsistently reported, and useful decisions require reasoning across lipid chemistry, formulation parameters, biological outcomes, and physicochemical properties. General-purpose LLMs provide a convenient interface for exploration, but they do not natively access curated formulation records, chemical similarity tools, or predictive models.

We present Mina, a domain-grounded agentic assistant for LNP formulation research. The system coordinates literature and patent retrieval, natural language access to curated formulation records, chemical similarity search, and predictive modeling, evaluated here through particle size prediction. It maps researcher questions to tool-supported workflows while preserving evidence sources and intermediate outputs for inspection.

We evaluate the assistant in a blinded expert A/B study on 20 realistic LNP questions drawn from real researcher queries. Four domain experts compared anonymized answers from the assistant and a prompted general-purpose LLM baseline without access to the assistant's tools or curated data. Experts preferred the assistant in 73.0% of decisive comparisons, with the largest gain in evidence support and traceability. Correctness ratings were high for both systems, indicating that the main benefit was more traceable and actionable decision support rather than a large correctness difference. These results show that tool-grounded scientific agents can make expert decision support more traceable and actionable in specialized formulation domains.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**; • **Information systems** → *Information retrieval*; • **Applied computing** → *Life and medical sciences.*

Keywords

Agentic AI, scientific assistants, tool use, retrieval-augmented generation, lipid nanoparticles, RNA therapeutics, formulation design

1 Introduction

RNA therapeutics increasingly rely on lipid nanoparticles (LNPs) to protect nucleic-acid payloads, enable cellular uptake, and tune tissue or cell-type targeting [5, 13]. LNPs are typically composed of an ionizable lipid, helper phospholipid, cholesterol or another sterol, and a PEG-lipid; changes in these components and their ratios can strongly affect stability, delivery, biodistribution, and tolerability [4, 5]. Identifying effective formulations is therefore a difficult experimental design problem: the design space spans lipid chemistry, helper lipids, PEG-lipids, molar ratios, payloads, administration routes, manufacturing parameters, assay contexts, and biological readouts.

This complexity creates an information-access problem as well as a wet-lab optimization problem. Relevant evidence is dispersed across papers and patents, experimental conditions are reported inconsistently, and formulation decisions often require combining literature knowledge with structured experimental records, chemical similarity, and quantitative property estimates. Chemical similarity methods such as Morgan fingerprints and Tanimoto similarity support analog reasoning over lipid structures [1, 12, 16], while recent ML-assisted LNP studies show the potential of predictive models for formulation triage [7, 10, 18]. In this paper, we evaluate particle size prediction as a representative callable quantitative tool within the broader agentic workflow.

General-purpose LLMs offer a natural interface for scientific exploration, but they are poorly matched to this setting when used alone. Prior work on retrieval-augmented generation and scientific agents has shown the value of connecting language models to external evidence sources, tools, and validation workflows [8, 15, 17]. However, LNP formulation research poses a specific combination of requirements that is not addressed by document retrieval alone: queries may require normalized formulation records, chemical structure similarity, payload and route context, and quantitative property estimates. General LLMs may summarize well-known concepts, but they do not natively access curated formulation databases, cheminformatics tools, or task-specific predictive models, and they can blur experimentally supported claims with plausible-sounding extrapolations. We integrate these capabilities into a single agentic workflow for LNP formulation research and evaluate whether domain grounding improves expert-rated decision support.

We present Mina, a domain-grounded agentic assistant for LNP formulation research, meaning that answers are conditioned on inspectable outputs from curated formulation records, literature and patent retrieval, chemical structure search, and predictive tools. A deployed version is available at https://mina.mana.bio/. Because the

assistant has been used in researcher-facing workflows, we were able to construct the evaluation set from real researcher questions rather than synthetic prompts. The system coordinates specialized tools over curated scientific data, retrieves supporting evidence, exposes provenance, and uses quantitative models where appropriate. It integrates four complementary sources of domain grounding: structured experimental data curated from papers and patents, a broader indexed literature and patent corpus, chemical structure representations for similarity search [1, 12, 16], and predictive models trained on formulation data [3, 7, 10, 18]. An agent layer routes user questions to the appropriate capability: retrieval-augmented generation for literature synthesis [8], natural language-to-SQL for structured database queries [6, 14, 21], Tanimoto similarity search for analog discovery [1], or particle size prediction for quantitative formulation triage [7, 10, 18].

This paper makes three contributions. First, we present Mina, an expert-facing scientific assistant for LNP formulation research. The agent integrates literature and patent retrieval, structured formulation data access, chemical similarity search, and callable predictive models. We evaluate particle size prediction as a representative predictive tool. Second, we describe a blinded expert A/B evaluation protocol for scientific decision support agents and apply it to 20 realistic LNP formulation queries drawn from real researcher questions, comparing the assistant with a prompted general LLM baseline. Third, we show that experts preferred the assistant in 73.0% of decisive comparisons, with the largest gain in evidence support and traceability and exploratory gains in actionability and specificity. Correctness ratings were high for both systems, suggesting that the main benefit was more traceable and actionable decision support.

2 Related Work

This work builds on three lines of research: LLM agents for scientific workflows, retrieval-augmented generation and natural language access to structured data, and machine-learning-assisted nanoparticle formulation design.

LLM agents for scientific workflows. Tool-using LLM agents extend single-turn prompting by selecting external actions, calling tools, and synthesizing intermediate results. ReAct established a general pattern of interleaving reasoning and action [19]; chemistry-focused systems such as ChemCrow and Coscientist demonstrated the value of domain-specific tools for synthesis, drug discovery, and laboratory workflows [2, 9]. Recent surveys emphasize provenance, validation, and human oversight as central requirements for scientific agents [15, 17]. Our work applies this tool-grounded paradigm to LNP formulation research, where useful answers require joint reasoning over lipid structure, formulation ratios, payload and route context, biological readouts, and physicochemical properties. Methodologically, prior scientific-agent evaluations have often emphasized demonstrations, task completion, tool-use benchmarks, or expert assessment [2, 9, 11, 20]. We complement these approaches with a blinded paired evaluation in which LNP domain experts judge answer quality as decision support.

RAG and structured data access. Retrieval-augmented generation (RAG) improves factual grounding by conditioning generation on retrieved documents and allows knowledge to be updated outside the model parameters [8]. This is useful for scientific literature synthesis, but document retrieval alone is often insufficient when evidence is stored in structured tables, chemical representations, normalized assays, or predictive models. Natural-language-to-SQL systems complement RAG by translating user questions into executable database queries; recent LLM-based approaches use decomposition, schema reasoning, and self-correction to improve execution accuracy on complex database tasks [14, 21]. Newer surveys of LLM-based text-to-SQL systems emphasize persistent challenges in domain generalization, robustness, schema linking, and evaluation on dynamic real-world databases [6]. Our assistant combines literature and patent retrieval with structured access to curated formulation records, allowing qualitative evidence and normalized experimental records to be synthesized within the same workflow.

Cheminformatics and ML-assisted nanoparticle formulation. Chemical similarity search is a standard mechanism for analog reasoning in molecular design. Morgan and extended-connectivity fingerprints provide local-structure molecular representations [12, 16], and Tanimoto similarity is widely used for comparing binary fingerprints [1]. LNP formulation research adds a second challenge: formulation performance depends not only on lipid structure, but also on helper lipids, sterols, PEG-lipids, payloads, ratios, routes, and manufacturing parameters. Reviews of mRNA delivery and LNP technology emphasize the importance of formulation composition and delivery context for RNA therapeutics [5, 13], including tissue-targeting strategies such as SORT nanoparticles [4]. Machine-learning models can help estimate formulation properties or prioritize candidates, and scalable tree-boosting methods such as XGBoost are commonly used for tabular prediction problems [3]. Recent LNP modeling work includes literature-scale ML frameworks for nucleic-acid delivery performance [7], transfection-efficiency prediction from ionizable lipid representations [10], and multimodal deep-learning models for mRNA-LNP properties [18]. These studies motivate computational triage, but also highlight the need for formulation context, data quality, and experimental validation. Together, these lines of work motivate our contribution: an expert-evaluated decision support assistant that integrates retrieval, structured formulation evidence, chemical similarity, and callable predictive tools for LNP formulation research.

3 System Overview

Figure 1 summarizes the system architecture. The assistant translates open-ended formulation questions into tool-supported workflows over domain resources: an indexed literature and patent corpus, a curated formulation database, chemical structure representations, and predictive models. It supports formulation scientists by exposing the evidence, records, similarity matches, predictions, and tool traces used to construct an answer.

Domain resources. The system is grounded in two complementary knowledge sources. First, a curated experimental database extracted from hundreds of scientific papers and patent documents stores lipid structures, formulation conditions, payloads, experimental contexts, and outcomes such as particle size, expression, biodistribution, tropism, and toxicity-related readouts. Where possible, records are normalized for cross-study comparison, including

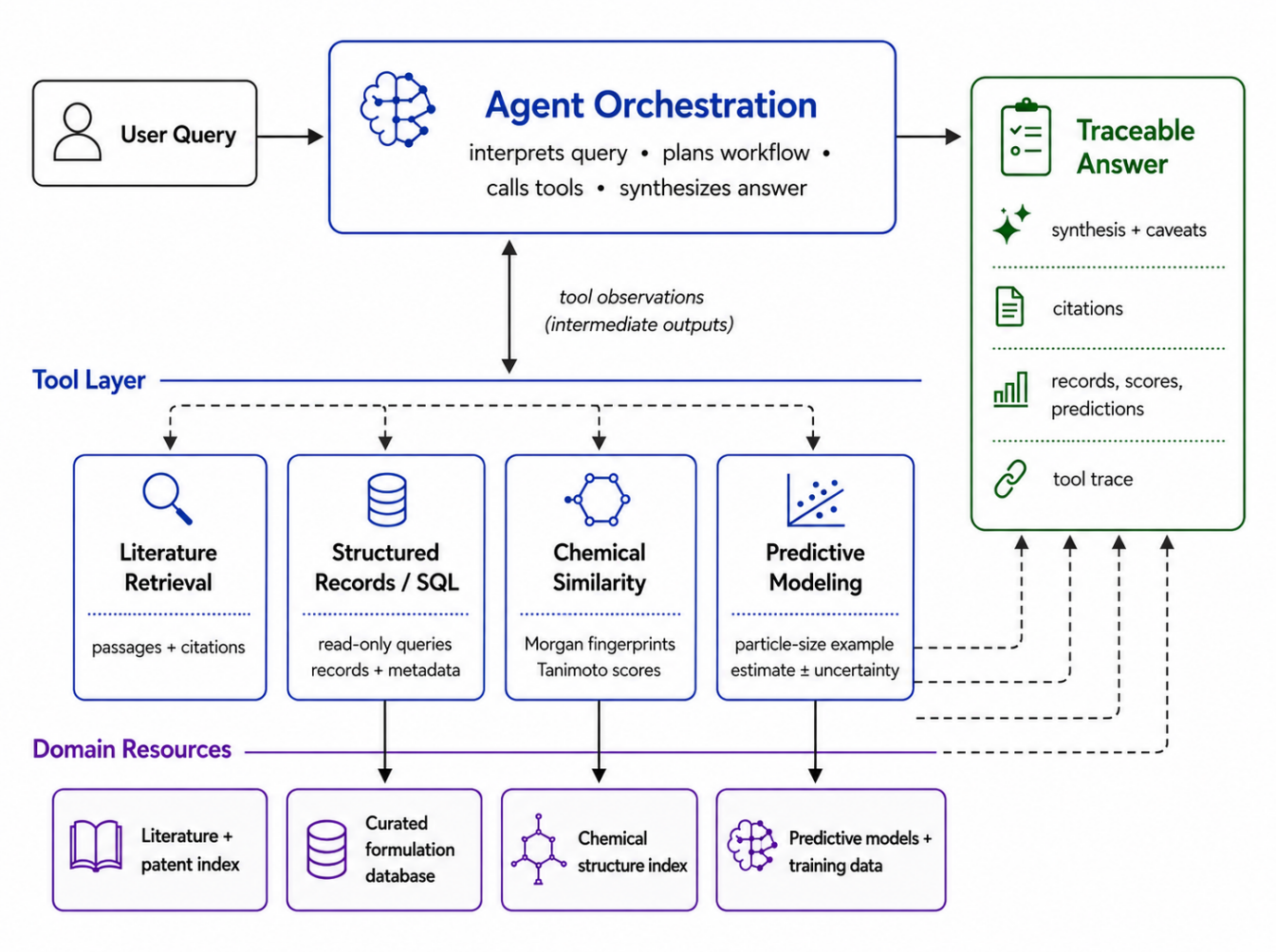

Figure 1: Architecture of the domain-grounded LNP formulation assistant. The agent interprets a user query, plans a workflow, and synthesizes an answer from intermediate tool outputs. Domain tools provide literature passages and citations, read-only structured formulation records, chemical similarity matches, and quantitative predictions, illustrated here with particle size estimates. Retrieved evidence, database records, similarity scores, model outputs, and tool traces are surfaced with the final response to support traceability.

lipid molar ratios, N:P ratios, benchmark formulations, controls, and biodistribution-specificity measures. Second, a broader corpus of LNP-related papers and patents is indexed for literature retrieval. The curated database supports structured comparison, while the broader corpus supports citation-backed synthesis over less structured scientific evidence.

Tool interfaces. The tool layer exposes four main domain capabilities to the agent. Literature tools perform hybrid keyword and vector retrieval over the paper corpus and return passages, article identifiers, and citation metadata for grounded synthesis, following the broader retrieval-augmented generation paradigm [8]. Structured data tools translate natural language requests into SQL queries over the curated database, returning matching formulations and associated experimental records [6, 14, 21]. Cheminformatics tools normalize lipid names or SMILES strings, a standard text notation for molecular structures, then compute Morgan fingerprints and Tanimoto similarities to find related ionizable lipids and supporting evidence [1, 12, 16].

Predictive tools run independently of the LLM and estimate formulation properties; here we evaluate particle size prediction as a representative example. The particle size tool featurizes a normalized specification, including lipid components, component ratios, N:P ratio, payload metadata when available, and chemical descriptors. It returns a point estimate and approximate uncertainty estimate in nanometers. When a query omits required formulation fields, the tool fills missing fields using predefined defaults for common LNP settings. The final response states these assumptions and presents the prediction as conditional on the completed formulation specification.

Agent orchestration. The agent layer is implemented as a tool-calling agent. At each step, the model receives the user query, system instructions, available tool descriptions, and prior tool outputs; it then either invokes another tool or synthesizes a final answer. This design follows the broader tool-using agent paradigm, in which LLMs interleave reasoning, external tool use, and synthesis over intermediate observations [2, 9, 15, 17, 19, 20]. Selected tools are executed externally, their outputs are appended to the model context, and the loop continues until the model emits a final answer or reaches a bounded tool-call budget.

The agent composes workflows based on the query. Literature questions use retrieval and citation-backed synthesis. Formulation-selection questions may use SQL over structured records. Lipid-name or SMILES queries may combine name normalization, chemical similarity, database lookup, and literature retrieval. Formulation-property questions may invoke the particle size model before synthesizing caveats and next steps.

Database access is execution-guided: the agent generates SQL for a read-only relational database tool. Database errors or empty results are returned as observations, allowing the agent to revise the query or report the limitation within the same tool-calling episode.

Traceability and reproducibility. Tool calls, tool inputs, retrieved evidence, intermediate outputs, model outputs, and final responses are logged as a tool trace for inspection and workflow analysis. During synthesis, the agent uses this trace to distinguish claims grounded in retrieved evidence, related-lipid analogs, or predictive-model outputs. Model decoding is configured with temperature 0.0. The non-LLM tools, including SQL execution, chemical similarity search, and particle size prediction, are deterministic for fixed inputs and database/model versions. Fixed prompts, deterministic non-LLM tools, and a bounded tool-call budget reduce workflow variability, although retrieval and LLM generation may still vary across provider or index versions.

4 Evaluation Design

We evaluate the assistant using three complementary analyses. The primary evaluation is a blinded expert A/B study comparing it with a prompted general LLM baseline. This study asks whether domain grounding improves decision support as judged by LNP domain experts. We also inspect workflow traces to understand how tool use helps or fails to translate into useful final answers, and we evaluate particle size prediction as a representative quantitative tool available to the agent.

4.1 Expert A/B Evaluation Set

We use a 20-question evaluation set fixed before answer generation, designed to exercise the assistant's intended workflows while remaining relevant to a prompted general LLM baseline. The set excludes questions that can only be answered by direct lookup in the assistant's curated database. Instead, questions reflect realistic formulation research tasks across four categories: quantitative prediction, structured formulation queries, cheminformatics and similarity, and multi-step correlation/synthesis. The questions were selected from real researcher queries and lightly normalized for clarity. The frozen questions were not used for prompt selection, tool-routing rule development, or model training. The full question list is provided in Appendix A.3.

4.2 Systems Evaluated

We evaluate two systems built on the same base model, Gemini 2.5 Flash. The first is Mina, which has access to the domain tools described in Section 3. The second is a prompted general-purpose LLM baseline, invoked with the same base model but without access to the assistant's curated database, chemical similarity tool, literature index, or particle size model. Each system receives the same user-facing question and the same answer-length constraint; outputs are anonymized before expert review.

Prompted general LLM baseline. The baseline uses Gemini 2.5 Flash with a static expert-assistant prompt and no external tools. The prompt asks the model to act as an LNP formulation assistant, avoid unsupported claims, state uncertainty, and provide formulation-relevant detail. This stronger prompted baseline tests whether domain grounding and tools add value beyond task framing.

Mina. The assistant uses Gemini 2.5 Flash inside the tool-calling workflow described in Section 3, with routing across literature retrieval, structured database access, chemical similarity, and prediction.

4.3 Expert A/B Evaluation

The main evaluation is a blinded A/B study comparing Mina with the prompted general LLM baseline. Both systems answer each question under the same hard length constraint, specified in Appendix A.1. System identities, tool traces, conversational follow-up questions, and other artifacts that could reveal the generating system are removed; citations and evidence statements are preserved because traceability is part of the evaluation. Thus, blinding removes explicit system identity and tool-use artifacts, but not evidence presentation itself, which is one of the evaluated answer qualities. Answer order is randomized independently for each rater and question.

Four LNP domain experts with hands-on experience in nanoparticle formulation or RNA delivery research rated each anonymized answer. Raters did not write, edit, or select the answers they evaluated. They scored each answer on four 1–5 Likert dimensions: scientific correctness, evidence support / traceability, LNP-specificity, and actionability. Raters were instructed not to reward verbosity by itself and to penalize unsupported specificity, fabricated citations, overconfident claims, or recommendations that ignore experimental context. For each question, they also selected which answer they would prefer to provide to an LNP scientist as decision support, with options for either answer, tie, or neither. Optional comments and low-confidence flags were collected for qualitative interpretation only.

For Likert dimensions, each rater-question pair contributes one paired comparison between the assistant and baseline answer; we aggregate scores by reporting system-level means and paired assistant-minus-baseline differences over complete paired score records. For preference labels, we report counts over all usable labels and separately report the decisive preference rate after excluding tie and neither-acceptable judgments. Ties and neither

judgments are therefore retained for descriptive accounting but are not counted as decisive preferences for either system. Optional comments and low-confidence flags are used only for qualitative interpretation and do not change the quantitative aggregation.

We report mean scores, paired differences, preference rates, and inter-rater agreement. We use intraclass correlation for Likert dimensions and Fleiss' kappa for preference labels.

4.4 Workflow Trace Analysis

In addition to expert A/B grading, we inspect assistant traces to understand how tool use contributed to the final answers. Each question is manually labeled with an expected evidence strategy: literature synthesis, formulation-context comparison, chemical analog reasoning, quantitative prediction, or multi-tool workflow. We then review the selected tools and intermediate outputs to identify whether failures arose from routing, retrieval, or final synthesis. This analysis is descriptive and is reported only for the domain-grounded assistant, since the general LLM baseline has no access to the assistant's tool set.

4.5 Particle Size Prediction Tool

We evaluate particle size prediction as a representative quantitative tool because size is routinely measured during LNP development and is used as an early formulation-quality and triage signal. Particle size can affect manufacturability, stability, biodistribution, and cellular uptake, but it does not determine biological efficacy on its own [4, 5]. The particle size prediction model is an XGBoost regression model trained on structured formulation features and chemical descriptors. It is evaluated independently of the LLM agent using nested cross-validation over 6,663 formulation records, comprising 3,549 records curated from papers and patents and 3,114 additional laboratory records. We use 5 outer folds for performance estimation and 3 inner folds for model selection, and report MAE, RMSE, and Pearson correlation. Because this model is called by the assistant but is trained and evaluated independently, these results evaluate a quantitative tool available to the agent, not the agentic workflow itself. Additional model details, including fold construction, missing-value handling, uncertainty estimation, and baseline comparison, are provided in Appendix A.4.

5 Results

5.1 Blinded Expert A/B Evaluation

Table 1 summarizes the blinded expert evaluation. Across 79 evaluable preference judgments from four raters, experts preferred the domain-grounded assistant in 46 cases and the prompted general LLM baseline in 17 cases, with 9 ties and 7 judgments where neither answer was considered acceptable. Among decisive preferences, the domain-grounded assistant was selected in 73.0% of cases (46/63; 95% CI [60.3%, 83.4%]). Of 80 rater-question assignments, one was marked "unable to judge," leaving 79 evaluable preference judgments; 75 rows contained complete paired Likert scores and preference labels. Under the shared 350-word limit, answer lengths were similar overall: the assistant averaged 293 words per response and the prompted baseline averaged 277 words; the assistant was longer in 6 of 20 questions and shorter in 14 of 20.

Table 1: Blinded expert A/B evaluation. Scores are 1–5 Likert means over complete paired rater-question records. Differences are paired assistant-minus-baseline differences; confidence intervals are 95% CIs.

Metric	Assistant	Baseline	Δ	95% CI
Correctness	4.51	4.23	+0.28	[0.01, 0.55]
Evidence / traceability	4.08	2.23	+1.85	[1.49, 2.21]
LNP-specificity	4.55	3.97	+0.57	[0.22, 0.92]
Actionability	3.96	3.16	+0.80	[0.36, 1.24]

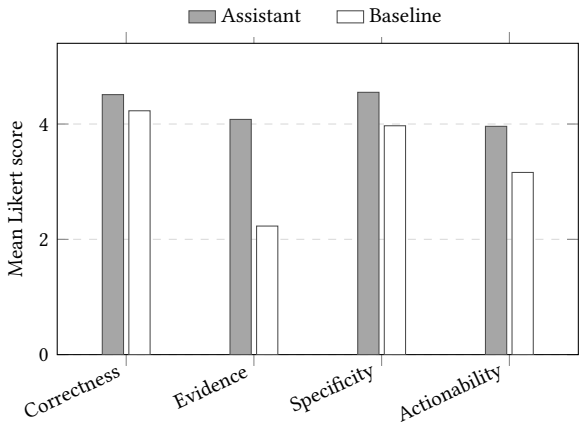

Figure 2: Mean expert Likert scores for the domain-grounded assistant and the prompted general LLM baseline. Paired assistant-minus-baseline differences and 95% confidence intervals are reported in Table 1.

The Likert ratings, summarized in Table 1 and Figure 2, show that the largest and most consistent advantage was evidence support and traceability. The assistant scored substantially higher on evidence support (4.08 vs. 2.23; mean paired difference = 1.85). This result is consistent with the assistant's retrieval-augmented design, but the accompanying preference and actionability gains show that retrieved evidence still needs effective synthesis. The traceability gain was accompanied by improved actionability (3.96 vs. 3.16; $\Delta = 0.80$). LNP-specificity also favored the assistant (4.55 vs. 3.97; $\Delta = 0.57$), although we interpret this dimension cautiously because inter-rater agreement was poor. Correctness ratings were high for both systems and slightly favored the assistant (4.51 vs. 4.23; $\Delta = 0.28$), but the effect was small and should be interpreted cautiously. Thus, the main result is improved traceable decision support, with exploratory gains in actionability and specificity.

We treat dimension-level comparisons as exploratory because Likert scores are ordinal, category-level analyses are small, and no multiple-comparison correction was applied. We therefore emphasize effect sizes, confidence intervals, preference rates, and consistency across dimensions rather than interpreting individual dimensions as confirmatory findings.

5.2 Inter-rater Agreement

Inter-rater agreement was strongest for evidence support (ICC(A,1) = 0.685), moderate for actionability (ICC(A,1) = 0.447), lower for correctness (ICC(A,1) = 0.319), and poor for LNP-specificity (ICC(A,1) = 0.104). Preference-label agreement was fair (Fleiss' $\kappa = 0.321$, computed over 19 questions with complete preference labels from all four raters). Evidence support is easier to judge from citations and explicit claims, whereas LNP-specificity depends more on expert expectations about formulation context. We therefore treat evidence support as the most consistently rated dimension-level improvement, interpret actionability as a moderate exploratory gain, and describe specificity and correctness more cautiously.

5.3 Workflow and Category Analysis

We inspected assistant traces to understand when tool use translated into preferred answers. Figure 3 summarizes preference counts by question category.

The assistant performed best when a question mapped cleanly to a specialized tool. In cheminformatics/similarity questions, experts preferred the assistant in 15 cases versus 3 for the baseline, with 1 tie and 1 neither judgment. These questions benefited from lipid-name normalization, explicit chemical representations, and analog search. Quantitative-prediction questions showed a similar pattern: experts preferred the assistant in 9 cases and never preferred the baseline, with 2 ties, reflecting the value of direct access to the particle size model.

Multi-step synthesis questions showed a broader but less uniform advantage. Experts preferred the assistant in 16 cases versus 6 for the baseline, with 6 ties and 4 neither judgments. These questions benefited mainly from retrieval-grounded evidence support across formulation context, mechanisms, and experimental readouts.

Structured data queries were the main failure mode. Experts preferred the baseline more often in this category (8 baseline preferences versus 6 assistant preferences, with 2 neither judgments), even though the assistant retained an evidence support advantage. Trace inspection showed that the assistant often retrieved relevant records but did not always turn them into concise, ranked, formulation-specific recommendations. This identifies record aggregation, ranking, and recommendation planning as the main bottleneck, rather than database access alone.

5.4 Particle Size Prediction Tool

We assess the particle size prediction model independently of the LLM agent, using the nested cross-validation procedure described in Section 4.5. The evaluation uses 6,663 formulation records and a 5-by-3 nested cross-validation design, and reports MAE, RMSE, and Pearson correlation.

The model achieved MAE = 18.918 nm, RMSE = 26.868 nm, and Pearson $r = 0.750$ under nested cross-validation. This performance supports coarse formulation triage and candidate comparison, but not replacement of experimental size measurement. The assistant therefore presents size predictions with approximate uncertainty estimates and caveats, and treats them as one input alongside literature evidence, structured formulation records, and chemical similarity.

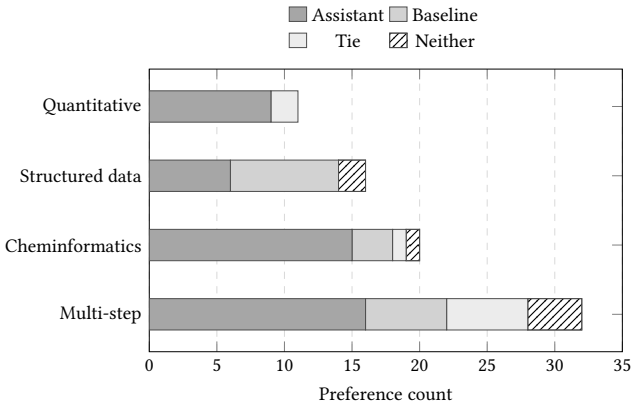

Figure 3: Category-level preference counts. The assistant was strongest on quantitative prediction and cheminformatics/similarity, remained favored on multi-step synthesis, and was weakest on structured data queries.

6 Case Studies

Table 2 illustrates two cases where domain grounding changed the answer. We then discuss a structured data case that illustrates a recurring limitation: retrieving relevant records does not always lead to a concise, decision-oriented recommendation.

Success cases: chemical normalization and design synthesis. In the SMILES-based query, the assistant identified the structure as SM-102, an ionizable lipid used in mRNA-LNP formulations, and connected it to literature, patent records, similar lipids, and formulation outcomes. In the DLin-MC3-DMA comparison query, it synthesized evidence-backed design principles for newer biodegradable ionizable lipids relative to a benchmark ionizable lipid. In both cases, the prompted baseline produced plausible but more generic answers. These examples show that the assistant's advantage came not only from retrieval, but from linking retrieved evidence to chemical normalization, similarity search, and structured formulation context.

Limitation case: structured formulation recommendation. One structured data question asked for an LNP formulation to deliver mRNA to antigen-presenting cells, which are immune cells relevant for vaccine and immunotherapy applications. The assistant retrieved relevant structured evidence, including formulations tested in dendritic-cell assays, and reported concrete parameters such as lipid components, N:P ratio, and antigen-presentation readouts. However, the final answer did not turn those records into a concise, ranked recommendation with a clear starting composition and rationale. By contrast, the baseline gave a simpler starting template and validation plan, but with weaker evidence support. This limitation aligns with the category-level results: structured data queries were the assistant's weakest category despite retaining an evidence support advantage. The example shows that structured retrieval is necessary but not sufficient; future work should improve record aggregation, ranking, and recommendation generation over retrieved formulation evidence.

Table 2: Successful qualitative examples from blinded A/B items. Shortened snippets illustrate two benefits of domain grounding: chemical normalization with similarity search, and evidence-grounded synthesis of formulation-design principles.

Query	Prompted general LLM baseline	Domain-grounded assistant
Literature overview for an ionizable lipid structure.	Described the molecule as a plausible ionizable lipid with expected endosomal-escape behavior, but did not identify the structure as a known named lipid.	Resolved the SMILES to SM-102, retrieved literature and patent records, reported similar lipids with similarity scores, and summarized associated formulation outcomes.
Compare DLin-MC3-DMA with newer biodegradable ionizable lipid designs.	Identified biodegradability as the main shift and described cleavable linkages, but remained mostly generic.	Synthesized evidence-backed design changes including labile linkers, hydroxylated headgroups, branched chains, pKa tuning, reduced hepatotoxicity, and newer synthesis strategies.

7 Discussion

The evaluation indicates that domain grounding adds the most value when the agent combines retrieved evidence with structured records, chemical similarity, and quantitative estimates. The assistant's largest gain was evidence support and traceability. Importantly, this gain was accompanied by higher expert preference and improved actionability, suggesting that retrieved evidence must still be synthesized into useful guidance. Correctness ratings were high for both systems and only slightly favored the assistant (4.51 vs. 4.23; $\Delta = 0.28$ on a 5-point scale), so we interpret this as broadly comparable correctness rather than a definitive correctness improvement. The category analysis points to the main failure mode: useful tool outputs do not help unless the agent turns them into a clear final answer.

Evaluation validity. The evaluation uses expert judgments over open-ended answers, so scores reflect both scientific accuracy and perceived usefulness. To reduce presentation bias, answers were anonymized, answer order was randomized, and both systems used the same length constraint. The question set is compact and usage-derived, so it should not be interpreted as a representative benchmark over all LNP formulation tasks. Because raters evaluated answer text rather than internal tool traces, traceability was assessed through citations, evidence statements, and model outputs shown in the answers.

Limitations. The 20-question set was drawn from real researcher queries, making the evaluation realistic but not exhaustive over the broader LNP formulation task space. The underlying evidence sources are incomplete: curated databases reflect qualification criteria, extraction decisions, and normalization assumptions, while literature retrieval can miss relevant evidence or retrieve studies whose experimental contexts are not directly comparable. Expert evaluation of open-ended scientific answers is inherently subjective, especially when evidence is incomplete or conflicting; this is reflected in lower agreement for LNP-specificity. The assistant was weakest on structured data queries, where baseline answers were preferred more often despite the assistant's stronger evidence support, indicating that access to structured records is not sufficient without better aggregation, ranking, and concise recommendation generation. The evaluation compares the full domain-grounded assistant with a strong prompted LLM baseline and does not isolate the marginal contribution of individual tools or the orchestration layer. Finally, the evaluated particle size prediction tool supports triage but does not establish biological performance, toxicity, manufacturability, or clinical relevance, and retrospective evaluation cannot replace prospective wet-lab validation.

Future work. Three directions follow from this evaluation. First, the field needs broader evaluation sets and ablation studies for LNP-agent assessment, including standardized tasks for evidence retrieval, formulation comparison, analog search, hypothesis generation, and tool-specific contribution analysis. Second, agent outputs should expose more structured provenance, including which records were retrieved, which claims are directly supported, and where evidence is extrapolated or inferred from analogs. Third, prospective studies should test whether agent-supported workflows improve formulation-design efficiency or experimental success rates. These directions would make future systems easier to compare, audit, and validate prospectively.

8 Conclusion

We presented a domain-grounded agentic assistant for LNP formulation research that integrates curated experimental data, literature and patent retrieval, structured database querying, chemical similarity search, and callable predictive tools into a natural language workflow. In a blinded expert study on realistic LNP formulation questions, domain experts preferred the assistant in 73.0% of decisive comparisons. The assistant's strongest advantage was evidence support and traceability, with additional exploratory gains in actionability and specificity. Correctness ratings were high for both systems and slightly favored the assistant, but the effect was small and should not be interpreted as establishing uniformly higher scientific accuracy. These results do not establish autonomous formulation design, prospective experimental success, or clinical utility; the system remains a decision support tool requiring expert interpretation and experimental validation. More broadly, scientific agents for specialized domains should be judged not only by whether they produce plausible answers, but by whether they make evidence, assumptions, and tool-supported reasoning inspectable to domain experts.

References

[1] Dávid Bajusz, Anita Rácz, and Károly Héberger. 2015. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *Journal of Cheminformatics* 7, 20 (2015).

[2] Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. 2023. Autonomous Chemical Research with Large Language Models. *Nature* 624, 7992 (2023), 570–578.

[3] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 785–794.
[4] Qiang Cheng, Tuo Wei, Lukas Farbiak, Lindsay T. Johnson, Sean A. Dilliard, and Daniel J. Siegwart. 2020. Selective Organ Targeting (SORT) Nanoparticles for Tissue-Specific mRNA Delivery and CRISPR–Cas Gene Editing. *Nature Nanotechnology* 15 (2020), 313–320.
[5] Xucheng Hou, Tal Zaks, Robert Langer, and Yizhou Dong. 2021. Lipid Nanoparticles for mRNA Delivery. *Nature Reviews Materials* 6 (2021), 1078–1094.
[6] Yiming Huang, Jiyu Guo, Wenxin Mao, Cuiyun Gao, Peiyi Han, Chuanyi Liu, and Qing Ling. 2025. Exploring the landscape of text-to-sql with large language models: Progresses, challenges and opportunities. *arXiv preprint arXiv:2505.23838* (2025).
[7] Gaurav Kumar and Arezoo M Ardekani. 2025. Machine-learning framework to predict the performance of lipid nanoparticles for nucleic acid delivery. *ACS Applied Bio Materials* 8, 5 (2025), 3717–3727.
[8] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*.
[9] Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. 2024. Augmenting large language models with chemistry tools. *Nature machine intelligence* 6, 5 (2024), 525–535.
[10] Asal Mehradfar, Mohammad Shahab Sepehri, Jose Miguel Hernandez-Lobato, Glen S. Kwon, Mahdi Soltanolkotabi, Salman Avestimehr, and Morteza Rasoulianboroujeni. 2025. LANTERN: A Machine Learning Framework for Lipid Nanoparticle Transfection Efficiency Prediction. *arXiv preprint arXiv:2507.03209* (2025).
[11] Mahmoud Mohammadi, Yipeng Li, Jane Lo, and Wendy Yip. 2025. Evaluation and Benchmarking of LLM Agents: A Survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 6129–6139.
[12] Harry L. Morgan. 1965. The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation* 5, 2 (1965), 107–113.
[13] Norbert Pardi, Michael J. Hogan, Frederick W. Porter, and Drew Weissman. 2018. mRNA Vaccines—A New Era in Vaccinology. *Nature Reviews Drug Discovery* 17 (2018), 261–279.
[14] Mohammadreza Pourreza and Davood Rafiei. 2023. Din-sql: Decomposed in-context learning of text-to-sql with self-correction. *Advances in neural information processing systems* 36 (2023), 36339–36348.
[15] Shuo Ren, Pu Jian, Zhenjiang Ren, Chunlin Leng, Can Xie, and Jiajun Zhang. 2025. Towards Scientific Intelligence: A Survey of LLM-based Scientific Agents. *arXiv preprint arXiv:2503.24047* (2025).
[16] David Rogers and Mathew Hahn. 2010. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* 50, 5 (2010), 742–754.
[17] Jiaqi Wei, Yuejin Yang, Xiang Zhang, Yuhan Chen, Xiang Zhuang, Zhangyang Gao, Dongzhan Zhou, Guangshuai Wang, Zhiqiang Gao, Juntai Cao, et al. 2025. From ai for science to agentic science: A survey on autonomous scientific discovery. *arXiv preprint arXiv:2508.14111* (2025).
[18] Kun Wu, Zixu Wang, Xiulong Yang, Yangyang Chen, Fulvio Mastrogiovanni, Jialu Zhang, and Lizhuang Liu. 2025. TransMA: an explainable multi-modal deep learning model for predicting properties of ionizable lipid nanoparticles in mRNA delivery. *Briefings in Bioinformatics* 26, 3 (2025), bbaf307.
[19] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations (ICLR)*.
[20] Botao Yu, Frazier N Baker, Ziru Chen, Garrett Herb, Boyu Gou, Daniel Adu-Ampratwum, Xia Ning, and Huan Sun. 2025. Tooling or not tooling? the impact of tools on language agents for chemistry problem solving. In *Findings of the Association for Computational Linguistics: NAACL 2025*. 7620–7640.
[21] Xinyu Zhu, Junjie Wang, Lin Zhang, Yuxiang Zhang, Yuhang Gan, Jiaxing Wang, and Yujiu Yang. 2024. Large Language Model Enhanced Text-to-SQL Generation: A Survey. *arXiv preprint arXiv:2410.06011* (2024).

A Expert Evaluation Details

A.1 Answer-generation and rater instructions

Both systems receive the same user-facing question and the same length instruction:

> Answer in no more than 350 words. This is a hard limit.

The same length constraint is applied to both the prompted general LLM baseline and the domain-grounded assistant to reduce presentation bias.

Raters receive the following instruction:

> Please rate each answer as decision support for an LNP formulation scientist. Do not reward verbosity by itself. Prefer answers that are correct, evidence-grounded, formulation-specific, and practically useful. Penalize unsupported specificity, fabricated citations, overconfident claims, or recommendations that ignore experimental context. Use the full 1–5 scale; a score of 3 means acceptable but not strong.

The rubric is:

- **Scientific correctness:** 1 = mostly incorrect, misleading, or makes unsafe overclaims; 3 = plausible but incomplete, oversimplified, or contains minor scientific issues; 5 = scientifically sound, nuanced, appropriately cautious, and unlikely to mislead.
- **Evidence support / traceability:** 1 = key claims are unsupported, unverifiable, or cite questionable/nonexistent evidence; 3 = some key claims are supported, but evidence is incomplete, vague, or not clearly linked to the claim; 5 = key claims are clearly grounded in citations, retrieved evidence, structured data, or model outputs.
- **LNP-specificity:** 1 = generic drug-delivery or biology advice with little LNP formulation context; 3 = some LNP-specific content, but limited formulation detail or experimental context; 5 = concrete LNP-relevant details such as lipids, molar ratios, payloads, routes, assays, readouts, controls, mechanisms, or formulation constraints.
- **Actionability:** 1 = not useful for deciding next steps; 3 = provides some useful direction but limited prioritization, caveats, or experimental guidance; 5 = directly useful for formulation planning, troubleshooting, evidence review, or experimental prioritization.

For each question, raters also answer: "Which answer would you prefer to provide to an LNP scientist as decision support?" with options Answer A, Answer B, Tie / both acceptable, or Neither acceptable. Optional comments and a low-confidence flag are collected for qualitative interpretation.

The full baseline prompt is:

> You are an expert assistant in lipid nanoparticle (LNP) formulation for RNA delivery. Answer the user's question as accurately and specifically as possible. Focus on lipid nanoparticle formulation, RNA delivery, and experimental interpretation. Include citations, DOIs, PMIDs, or paper identifiers only when you can identify them with high confidence. Distinguish established evidence from plausible hypotheses. If evidence is context-dependent, explain the relevant context, such as lipid composition, molar ratios, payload type, route of administration, assay type, species or cell type, controls, and readouts. Do not mention that you lack access to proprietary databases or tools unless directly relevant. Do not include conversational follow-up questions.

A.2 Evaluation accounting and statistical analysis

The final analysis uses four raters and 20 questions, yielding 80 rater-question assignments. Of these, 79 contain usable preference labels and 75 contain complete paired Likert scores together with preference labels. Preference results are reported over usable preference labels. Paired score analyses use the rater-question comparison as the unit of analysis.

Preference labels are summarized both over all usable labels and over decisive labels only, where decisive labels exclude ties and neither-acceptable judgments. The decisive preference rate is reported descriptively with a 95% confidence interval. We avoid treating rater-question preference labels as fully independent because judgments are clustered by both rater and question. Likert dimensions are summarized with assistant and baseline means, paired assistant-minus-baseline differences, and 95% confidence intervals for paired differences. Confidence intervals are Student's t intervals over rater-question-level paired differences. Because Likert scores are ordinal, the evaluation set is compact, and no multiple-comparison correction is applied, dimension-level statistical tests are treated as exploratory.

Inter-rater agreement is summarized with intraclass correlation for Likert dimensions and Fleiss' κ for preference labels. Category-level analyses are descriptive because category sample sizes are small and some rater-question rows are incomplete.

A.3 Frozen evaluation questions

The 20-question evaluation set is grouped by intended workflow category. Questions were selected from real researcher queries, lightly normalized for clarity, and paraphrased where needed to remove identifying or confidential context while preserving the intended evaluation task.

Quantitative prediction.

(1) How large is the mRNA-LNP containing MC3, DSPC, cholesterol, and DMG-PEG2000 at a 50/10/38.5/0.5 molar ratio?
(2) What is the expected size of a formulation using SM-102 and DOPE as the helper lipid at an N:P ratio of 8?
(3) If I change the helper lipid from DSPC to DOPE in a standard DLin-MC3-DMA formulation (50:10:38.5:1.5), how does the predicted size change?

Structured data queries.

(4) I want to make an LNP that will deliver mRNA encoding peptides specifically to antigen-presenting cells. What formulation should I use?
(5) I need an inhalable LNP formulation for siRNA delivery for lung cancer. What formulation evidence and parameters should I consider?
(6) I want to use an LNP for cystic fibrosis. Give me formulation ingredients, parameters, and mixing ratios that are supported by evidence.
(7) Give me examples of LNP functionalization strategies that modify the particle surface with ligands or other molecules.

Cheminformatics and similarity.

(8) Is Lipid 244cis patented?
(9) What is known about SM-102?
(10) Perform a literature overview on `CCCCCCCCCCCOC(=O)CCCCCN(CCCCCCCC(=O)OC(CCCCCCCC)CCCCCCCC)CCO`.
(11) I have a novel lipid with SMILES `C=CCCCCCCCCC(=O)OCCCCCC(NCCN1CCCCC1)C(=O)NCCCCCCCCCCCC`. Is it similar to known ionizable lipids?
(12) Compare DLin-MC3-DMA with newer biodegradable ionizable lipid designs. What design principles appear to have changed?

Multi-step correlation and synthesis.

(13) Provide ionizable lipids that can be used for nasal uptake.
(14) Show examples of LNP size and liver expression level, and summarize the correlation between the two.
(15) Is a zeta potential of -20 mV useful for LNPs?
(16) Can the acidic buffer with RNA contain sucrose before mixing with ethanol?
(17) How can I integrate protein-corona analysis into IVIVC frameworks for LNPs?
(18) What formulation strategies have been reported to reduce LNP-associated hepatotoxicity, and what evidence supports them?
(19) What formulation factors are reported to influence endosomal escape in ionizable lipid nanoparticles?
(20) I want to choose between DOPE and DSPC as helper lipid for an mRNA LNP. What does the evidence suggest, and what caveats should I keep in mind?

A.4 Particle size model details

The particle size target is hydrodynamic diameter measured by DLS, reported in nanometers. Records were drawn from curated literature/patent extractions and additional laboratory measurements. Duplicate measurements for the same formulation were averaged before splitting. For nested cross-validation, data points were grouped by formulation so that measurements from the same formulation did not appear in both training and test folds.

Input features include formulation composition variables, including ionizable lipid, helper lipid, sterol, PEG-lipid, component ratios, N:P ratio, and payload metadata where available, together with chemical descriptors for lipid structures. Rows with missing feature or target values were excluded before model training and evaluation.

The particle size tool returns a point estimate and an approximate uncertainty estimate. Uncertainty is estimated by a quantile model. These intervals are used as approximate decision support signals and are not treated as fully calibrated predictive intervals.

Table 3: Particle size prediction performance under nested cross-validation.

Model	MAE (nm)	RMSE (nm)	Pearson r
Mean baseline	35.40	48.13	–
XGBoost	**18.92**	**26.87**	**0.75**

A.5 Evaluation-set safety and interpretation notes

Questions involving disease settings, delivery routes, or freedom-to-operate are included because they reflect realistic formulation research queries. Raters were instructed to reward evidence-based research guidance while avoiding clinical recommendations, therapeutic claims, or definitive legal/patent conclusions. For freedom-to-operate questions, the expected behavior is to state appropriate caveats and describe the evidence or analysis needed, not to guarantee patent status.