LipoAgent: Coordinating Fine-Tuned LLM Agents for Safer Lipid Design

Anonymous ACL submission

Abstract

Lipid nanoparticles (LNPs) are among the most clinically mature platforms for nucleic acid delivery, yet designing lipids that are both effective and biologically safe remains a major bottleneck. In practical screening, toxicity is a decision-level constraint: if a lipid is toxic, its efficiency prediction is clinically irrelevant. We propose LipoAgent, a safety-aware multi-agent LLM framework for lipid discovery. LipoAgent combines domain-specific fine-tuning with a conditional prediction objective that enforces toxicity as a prerequisite for efficiency prediction, and further improves reliability via multi-agent verification with lightweight human oversight when disagreement persists. Across multiple foundation models, LipoAgent achieves an average 32% relative improvement in mRNA transfection efficiency prediction compared with other reported models for lipid design. **Wet-lab validation** confirms that virtual screening rankings reliably translate to biological transfection outcomes.

1 Introduction

Messenger ribonucleic acid (mRNA) therapeutics have attracted increasing attention due to their ability to regulate disease-related gene expression (Yan et al., 2022). The remarkable success of coronavirus disease 2019 (COVID-19) mRNA vaccines, such as BNT162b2 and mRNA-1273, has highlighted the clinical potential of mRNA-based therapies (Chemaitelly et al., 2022). Lipid nanoparticles (LNPs) play a pivotal role in enabling effective mRNA delivery, owing to their favorable biocompatibility, high encapsulation efficiency, and scalable manufacturing processes (Zong et al., 2023; Lu et al., 2024). As the key functional components of LNPs, ionizable lipids critically determine mRNA transfection efficiency, which is strongly impacted by their molecular structures.

Despite these advances, identifying new lipid candidates that achieve both high delivery efficiency and acceptable safety remains a major bottleneck. Traditionally, optimal lipids are discovered through large-scale combinatorial synthesis followed by extensive in vitro and in vivo screening. Such workflows are extremely time-consuming and resource-intensive, often requiring months of experimental effort and biological validation (Li et al., 2024). As a result, only a small fraction of the vast chemical design space can be practically explored, limiting both discovery speed and diversity.

Recent progress in data-driven modeling has opened new opportunities to accelerate lipid discovery. In particular, large language models (LLMs) have demonstrated strong reasoning and knowledge integration capabilities, enabling promising results in molecular generation and property prediction across the drug discovery pipeline (Zheng et al., 2024; Huang et al., 2020; Yao et al., 2023; Baek et al., 2025; Liu et al., 2025; Bran et al., 2023). These successes suggest that LLMs could serve as powerful assistants for navigating complex chemical spaces and prioritizing candidate molecules. However, directly applying general-purpose LLMs to lipid design presents fundamental challenges.

First, without domain-specific fine-tuning, LLMs often exhibit limited predictive accuracy for biochemical properties such as mRNA transfection efficiency. More critically, existing LLM-based approaches typically lack explicit safety modeling. Toxicity is frequently treated as a post hoc filtering step rather than an integral part of the decision process. This design choice can lead to unsafe high-confidence predictions, where a lipid is recommended as highly efficient despite exhibiting unacceptable toxicity, thereby introducing substantial experimental risk and wasted validation effort (Ramos et al., 2025).

To realize this paradigm, we propose **LipoAgent**, a safety-aware multi-agent LLM framework for lipid discovery. As illustrated in Figure 1, LipoAgent integrates domain-specific fine-tuning, con-

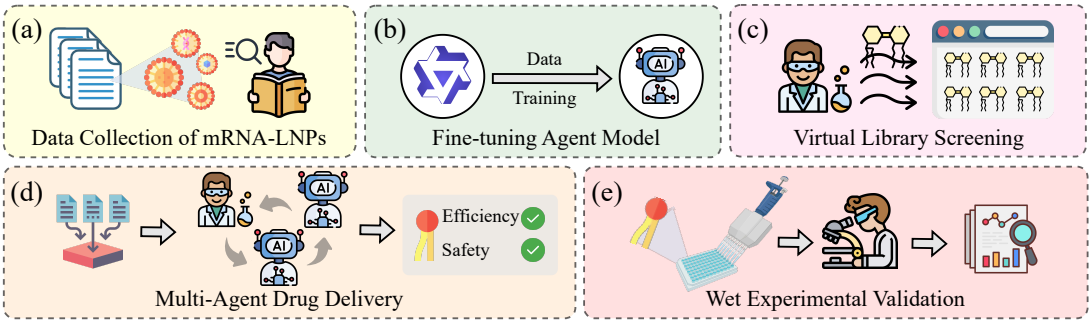

Figure 1: Overview of LipoAgent, with a safety-aware multi-agent LLM framework for lipid discovery.

Method	Multi-Agent	Fine-tuning	Human Feedback	Toxicity Detection
ReAct	✗	✗	✗	✗
ResearchAgent	✓	✗	✗	✗
ChemCrow	✓	✗	✓	✗
DrugAgent	✓	✗	✓	✗
LipoAgent	✓	✓	✓	✓

Table 1: Comparison between LipoAgent and existing LLM-based systems for molecular and drug discovery. Dark-green bold checkmarks indicate supported capabilities; red bold crosses indicate missing components.

ditional prediction, and multi-agent verification into a unified pipeline. The framework consists of two coordinated agents. The *Predictor agent* jointly predicts molecular toxicity and mRNA transfection efficiency while producing structured reasoning traces and confidence estimates. The *Verifier agent* examines low-confidence predictions by checking the consistency between predicted scores and their corresponding explanations. Crucially, we introduce a conditional loss mechanism during both training and inference: when a molecule is predicted to be toxic, the model directly outputs an "unsafe" decision and halts efficiency prediction. This design enforces toxicity as a prerequisite for efficiency modeling and prevents the system from recommending "efficient but toxic" lipid candidates. To support systematic and reproducible evaluation, we further construct and release a new dataset, **TransLipid**. The dataset is manually curated and normalized from multiple published studies and comprises structure–efficiency–toxicity triplets, providing a unified base for training and evaluating safety-aware molecular models. In summary, our contributions are as follows:

Safety-first multi-agent LLM framework: We propose LipoAgent, a multi-agent system for toxicity-aware mRNA transfection efficiency prediction in lipid discovery.

Conditional loss and decision-level safety modeling: We design a conditional loss function that enforces toxicity as a prerequisite for efficiency prediction during both training and inference, reducing unsafe false-positive recommendations.

TransLipid dataset: We construct and release TransLipid, a curated dataset that integrates lipid molecular structures with experimentally reported transfection efficiency data and toxicity annotations when available, enabling systematic evaluation of safety-aware molecular reasoning.

End-to-end experimental validation: Extensive quantitative evaluations show that LipoAgent performs competitively against other reported models for lipid design. Real-cell transfection and cytotoxicity experiments further confirm the accuracy of LipoAgent's predictions and their relevance to practical biological settings.

2 Background and Related Works

2.1 Lipid Design

The structural features of lipid materials play a decisive role in mRNA delivery performance. Subtle variations in molecular architecture can markedly influence particle formation, cellular uptake, endosomal escape, and ultimately protein expression (Zhao et al., 2024). Lipid-library screening is a commonly used strategy for identifying effective materials for mRNA delivery. However, experimental throughput is fundamentally limited by cost, labor, and resource demands, which severely constrain the scale of combinatorial exploration. As a result, promising lipid candidates may remain undiscovered, and the overall optimization process is often slow and inefficient.

These limitations motivate computational approaches that can design lipid candidates in silico. Importantly, such approaches must simultaneously account for delivery efficiency and biological safety, as high transfection efficiency alone is insufficient for clinical applicability.

2.2 LLMs for Drug Discovery and Delivery

In recent years, large language models (LLMs) have been increasingly applied to drug discovery and molecular design owing to their strong reasoning, planning, and knowledge integration capabilities. The early ReAct framework (Yao et al., 2023) introduced a reasoning–acting paradigm that enables models to alternate between thought generation and tool invocation for stepwise problem solving. However, ReAct lacks domain-specific adaptation and does not consider biochemical safety.

Building on this idea, ResearchAgent (Baek et al., 2025) incorporates a multi-agent structure for hypothesis generation, experiment planning, and literature-based reasoning. Despite its effectiveness in scientific knowledge synthesis, it remains limited to textual reasoning and does not perform quantitative prediction of molecular properties. ChemCrow (Bran et al., 2023) further advances LLM-based chemistry applications by integrating domain-specific tools such as molecule generators and property calculators, demonstrating the potential of human–AI collaboration. Nevertheless, it relies heavily on post hoc expert validation and lacks automated toxicity modeling.

More recently, DrugAgent (Liu et al., 2025) proposed a multi-agent system consisting of a planner and an instructor to automate machine learning programming for drug discovery. While it demonstrates agent coordination capabilities, it does not incorporate domain-specific fine-tuning and lacks molecular-level safety assessment.

As summarized in Table 1, existing LLM-based frameworks vary in their use of multi-agent coordination and human feedback. However, most approaches treat toxicity as a post hoc filtering step rather than a decision-level constraint. As a result, efficiency prediction and reasoning are not conditioned on safety, which can lead to high-confidence but unsafe molecular recommendations. In contrast, our proposed framework, LipoAgent, integrates domain-specific fine-tuning, safety-aware prediction, and multi-agent verification within a unified system to jointly optimize lipid transfection efficiency and safety.

2.3 Multi-agent LLM

Multi-agent LLM frameworks aim to improve reasoning reliability and scalability by decomposing complex tasks into cooperative submodules. Systems such as CAMEL (Li et al., 2023), Voyager (Wang et al., 2023), and Reflexion (Shinn et al., 2023) demonstrate how agent interaction, self-reflection, and feedback loops can enhance task performance in scientific reasoning.

Our work adopts a similar multi-agent philosophy but applies it to safety-critical molecular discovery. In LipoAgent, the Predictor agent focuses on joint toxicity and efficiency prediction, while the Verifier agent performs consistency checking between predicted scores and their corresponding reasoning traces. This verification mechanism is particularly important in high-confidence regimes, where single-agent reasoning may still produce incorrect but persuasive explanations. By incorporating safety-aware multi-agent verification, LipoAgent bridges the gap between LLM-based molecular reasoning and experimental reliability.

3 Methodology

We propose LipoAgent, a multi-agent framework for safe, reliable, and interpretable LLM-based lipid discovery for mRNA delivery, comprising two agents and a fine-tuned predictor. (1) a *Predictor Agent* that predicts lipid toxicity and delivery efficiency while generating textual reasoning and uncertainty estimation, and (2) a *Verifier Agent* that inspects low-confidence predictions and validates the logical consistency between the predicted score and its explanation. Through a human-in-the-loop feedback loop, LipoAgent supports iterative correction and continual refinement, as shown in Figure 2.

3.1 Fine-Tuning Strategy for Predictor Agent

The Predictor Agent is trained with a *Conditional Multi-Task Loss* that jointly optimizes toxicity classification and efficiency prediction. We apply LoRA (Hu et al., 2022) to the projection layers of the Predictor Agent while freezing all other parameters, substantially reducing training cost. Given a mini-batch of lipid inputs $\{\mathbf{x}_i\}_{i=1}^{N}$, the model outputs a toxicity logit z_i^{tox} and an efficiency logit vector z_i^{eff} over discrete transfection efficiency levels. The output probabilities are:

$$p_{\text{tox},i} = \sigma(z_i^{\text{tox}}), \qquad p_{\text{eff},i} = \text{Softmax}(z_i^{\text{eff}}).$$

Toxicity Loss. Toxicity prediction is trained using binary cross-entropy:

$$\mathcal{L}_{\text{tox}} = -\frac{1}{N}\sum_{i=1}^{N}\Big[y_i^{\text{tox}}\log\sigma(z_i^{\text{tox}}) + (1-y_i^{\text{tox}})\log(1-\sigma(z_i^{\text{tox}}))\Big] \tag{1}$$

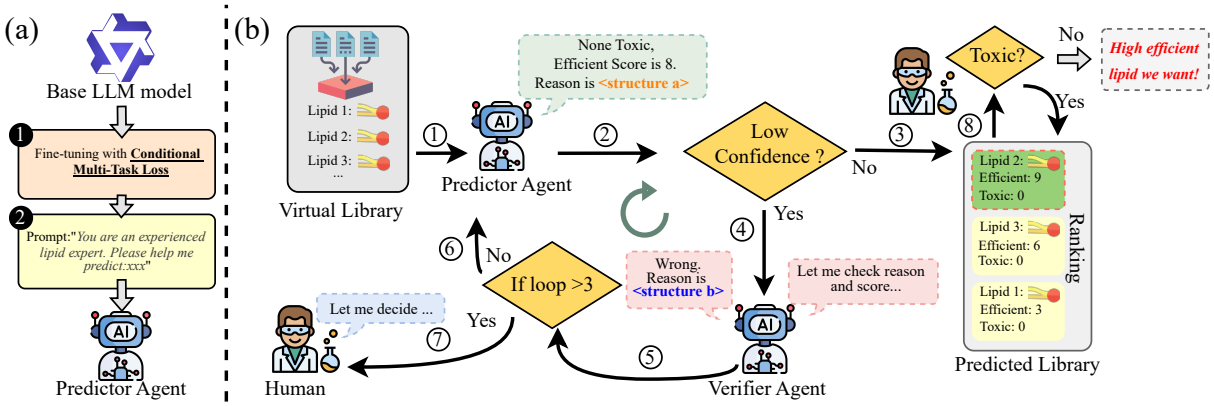

Figure 2: Overview of the LipoAgent framework. (a) Fine-tuning and prompting pipeline for constructing the predictor agent from a base LLM. (b) Multi-agent collaboration in LipoAgent, where agents coordinate with human feedback to iteratively filter and refine candidates toward high-efficiency lipids.

where $y_i^{\text{tox}} \in \{0, 1\}$ is the ground-truth label.

Efficiency Loss (Non-toxic Samples Only). Efficiency supervision is applied only to non-toxic lipids. Let $m_i = 1\{y_i^{\text{tox}} = 0\}$ be a binary mask indicating non-toxic samples. CE(.) is the cross-entropy loss, then the efficiency loss is:

$$\mathcal{L}_{\text{eff}} = \frac{\sum_{i=1}^{N} m_i \, \text{CE}(z_i^{\text{eff}}, y_i^{\text{eff}})}{\sum_{i=1}^{N} m_i + \epsilon}, \tag{2}$$

where $y_i^{\text{eff}} \in \{1, \ldots, 10\}$ is the discrete efficiency score. ϵ is a small constant introduced for stability. The final objective can be expressed as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{tox}} + \alpha \, \mathcal{L}_{\text{eff}}. \tag{3}$$

3.2 Multi-Agent Verification Framework

A key component of LipoAgent is an entropy-based confidence score that determines whether additional verification is needed. Given the efficiency distribution p_{eff}, the entropy is computed as:

$$\mathcal{H}(p_{\text{eff}}) = -\sum_{k=1}^{10} p_{\text{eff}}^{(k)} \log p_{\text{eff}}^{(k)}. \tag{4}$$

We normalize it into a confidence score:

$$\text{Conf}(x) = 1 - \frac{\mathcal{H}(p_{\text{eff}})}{\log 10}, \tag{5}$$

where $\mathcal{H}(p_{\text{eff}}) \in [0, \log 10]$ for a 10-class distribution, and the division by $\log 10$ normalizes the confidence score to the range $[0, 1]$, with higher entropy indicating lower confidence.

LipoAgent forms an iterative verification loop between the Predictor and Verifier Agents. The Predictor outputs $(y_{\text{tox}}, y_{\text{eff}}, r_{\text{pred}}, \text{Conf})$, and predictions with $\text{Conf} > \tau$ are accepted directly. Samples with $\text{Conf} \leq \tau$ are passed to the Verifier for evaluating reasoning–score consistency:

$$y_{\text{ver}} = f_{\text{ver}}(r_{\text{pred}}, y_{\text{eff}}) \in \{0, 1\}. \tag{6}$$

When inconsistency is detected, structured corrective feedback r_{corr} is generated and fed back to the Predictor for the next inference round.

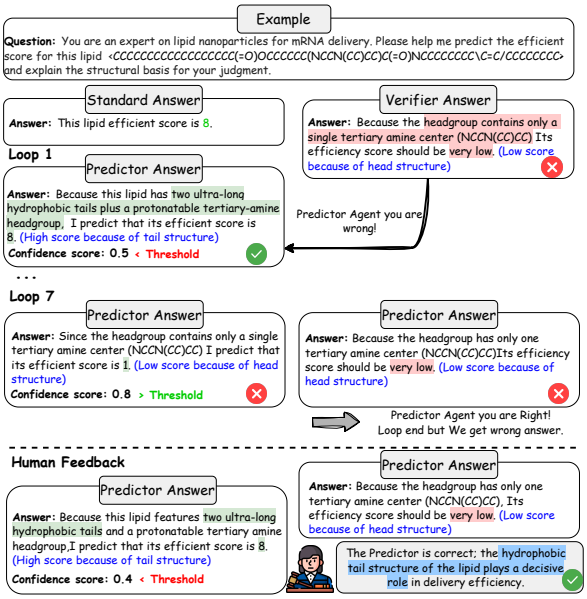

Figure 3: Multi-agent disagreement in iterative molecular efficiency prediction. During multi-step inference, the two agents may generate inconsistent efficiency estimates, leading to erroneous final selections. Incorporating lightweight human feedback resolves these inconsistencies and improves overall prediction accuracy.

3.3 Human-in-the-Loop

To improve safety and reliability, LipoAgent incorporates a human-in-the-loop mechanism. As illustrated in Figure 3, iterative multi-agent inference can result in disagreement between the Predictor and the Verifier. In some cases, the agents enter repetitive loops without reaching consensus; although a decision may eventually emerge after additional iterations (e.g., more than three rounds),

such late-converged predictions are often unreliable or incorrect. To mitigate this failure mode, we introduce lightweight human feedback. When the two agents fail to resolve their disagreement with sufficient confidence, a human expert provides a final judgment. This intervention prevents error accumulation from prolonged inference loops and improves overall system efficiency and accuracy.

4 Evaluation

The training and evaluation data used in this study are drawn from previously published literature (Wang et al., 2024) and organized into a curated dataset termed **TransLipid**. Specifically, we manually curate approximately 1,200 entries from peer-reviewed studies on lipid-based delivery materials, each containing the lipid chemical structure and its experimentally measured mRNA transfection efficiency. To support toxicity modeling and evaluation, we additionally incorporate 400 toxic lipid or lipid-like molecules from the publicly available `toxic_30_datasets` (Lu, 2024). After integration, TransLipid contains a total of 1,600 molecular entries, which are used for subsequent alignment, training, and evaluation.

Since the data originate from multiple independent studies, reported transfection efficiency scores are not directly comparable across sources. To address this issue, we rescale and align all 1,600 samples in TransLipid using a unified evaluation protocol and a consistent scoring data. This alignment is performed solely to ensure cross-study comparability rather than to optimize model performance.

For all experiments, we use a fixed split of the aligned TransLipid dataset, with 800 samples for training and 800 samples for evaluation.

4.1 Experimental Setting

All experiments are conducted on a single node with 4× NVIDIA H800 GPUs. We report averaged results over three independent runs with different random seeds.

Base LLMs and initialization. We consider multiple base LLM checkpoints to instantiate the Predictor Agent, including Qwen3-8B (Yang et al., 2025a), Qwen3-32B, ChemLLM (Zhang et al., 2024), Llama 3.1-8B (Grattafiori et al., 2024), TxGemma-7B (Wang et al., 2025), and TxGemma-27B. All models are initialized from their official pretrained checkpoints and trained under identical experimental settings.

Fine-tuning details. We adopt parameter-efficient fine-tuning with LoRA, applied to the attention projection modules (`q_proj` and `v_proj`), while keeping all remaining parameters frozen. The learning rate is set to 2×10^{-4}. For the conditional multi-task objective in Eq. (3), we set the loss weight $\alpha = 0.1$.

Multi-agent inference and human feedback. During inference, the Predictor and Verifier agents interact in an iterative verification loop. If the loop count exceeds three and the two agents fail to reach a consistent decision on a lipid's mRNA transfection efficiency, lightweight human feedback is triggered to provide a final judgment and terminate the loop. Human intervention is thus used as a fail-safe mechanism rather than a default inference path.

Baselines. We benchmark against diverse architectures on TransLipid, including GNN-based methods (AGILE (Xu et al., 2024), SCENT (Gaiński et al., 2025)), an MLP-based method (LANTERN (Mehradfar et al., 2025)), LLM-based approaches (KnowMol (Yang et al., 2025b), DrugPilot (Li et al., 2025)), and a multi-agent system (DrugAgent (Liu et al., 2025)). Benchmark baselines are evaluated using their released code and pretrained models when available. Since DrugAgent does not provide an official implementation, we reproduce it following the original paper under the same evaluation protocol.

4.2 mRNA Transfection Efficiency and Toxicity

The overall performance of mRNA transfection efficiency and toxicity prediction is summarized in Table 2. Across all base LLM backbones, incorporating the proposed LipoAgent framework leads to substantial improvements in both efficiency and toxicity prediction accuracy. On average, LipoAgent improves prediction accuracy by approximately 32%, with all enhanced models achieving over 70% accuracy on both tasks. These results consistently outperform baselines based on GNNs, MLPs, and prior LLM-based approaches.

Importantly, performance gains arise from complementary components of the framework. Domain-specific fine-tuning primarily improves overall prediction accuracy, while multi-agent verification disproportionately enhances reliability on difficult and extreme cases. As shown in Table 2, LipoAgent achieves over 85% accuracy on extreme efficiency values (scores 1, 2, 9, and 10). Such

Table 2: Comparison of methods and LLM baselines across efficiency/toxicity metrics.

Methods	Architecture	Fine-tuned	Multi-agent	Efficiency accuracy	Extrem accuracy[a]	Middle accuracy[b]	MAE	Toxic accuracy
AGILE (Xu et al., 2024)	GNN	✗	✗	30.80%	87.83%	20.10%	2.28	-
SCENT (Gaiński et al., 2025)	GNN	✗	✗	32.56%	40.33%	28.45%	2.21	-
LANTERN (Mehradfar et al., 2025)	MLP	✗	✗	42.33%	37.22%	51.15%	2.10	58.76%
KnowMol (Yang et al., 2025b)	LLM	✓	✗	62.34%	65.67%	63.26%	1.81	62.12%
DrugPilot (Li et al., 2025)	LLM	✓	✗	53.47%	55.87%	50.86%	1.97	66.05%
DrugAgent (Liu et al., 2025)	LLM	✗	✓	61.21%	63.42%	58.80%	1.98	65.05%
Qwen 3-8B (Yang et al., 2025a)	LLM	✗	✗	60.23%	72.34%	55.27%	1.87	70.40%
Qwen 3-8B w/ Fine-tuned	LLM	✓	✗	70.40%	86.36%	65.30%	1.24	90.30%
Qwen 3-8B w/ LipoAgent	LLM	✓	✓	**76.80%**	**89.60%**	**69.30%**	**1.09**	**100.00%**
Qwen 3-32B (Yang et al., 2025a)	LLM	✗	✗	62.56%	67.23%	72.98%	1.85	70.40%
Qwen 3-32B w/ Fine-tuned	LLM	✓	✗	86.70%	92.31%	75.00%	1.21	93.30%
Qwen 3-32B w/ LipoAgent	LLM	✓	✓	**89.20%**	**92.87%**	**84.32%**	**1.01**	**100.00%**
ChemLLM (Zhang et al., 2024)	LLM	✗	✗	11.10%	13.21%	10.74%	2.96	56.84%
ChemLLM w/ Fine-tuned	LLM	✓	✗	65.57%	69.33%	63.47%	1.87	72.11%
ChemLLM w/ LipoAgent	LLM	✓	✓	**71.41%**	**76.97%**	**69.38%**	**1.33**	**100.00%**
Llama 3.1-8B (Grattafiori et al., 2024)	LLM	✗	✗	35.00%	38.26%	32.33%	1.99	65.33%
Llama 3.1-8B w/ Fine-tuned	LLM	✓	✗	68.83%	70.75%	54.29%	1.57	77.40%
Llama 3.1-8B w/ LipoAgent	LLM	✓	✓	**72.34%**	**74.35%**	**69.38%**	**1.32**	**100.00%**
TxGemma-7B (Wang et al., 2025)	LLM	✗	✗	80.20%	55.45%	83.60%	1.33	85.44%
TxGemma-7B w/ Fine-tuned	LLM	✓	✗	89.50%	84.38%	91.23%	1.13	92.87%
TxGemma-7B w/ LipoAgent	LLM	✓	✓	**94.23%**	**90.83%**	**95.23%**	**0.85**	**100.00%**
TxGemma-27B (Wang et al., 2025)	LLM	✗	✗	82.34%	80.23%	83.12%	1.25	86.70%
TxGemma-27B w/ Fine-tuned	LLM	✓	✗	91.31%	89.47%	92.48%	1.10	94.20%
TxGemma-27B w/ LipoAgent	LLM	✓	✓	**94.23%**	**92.34%**	**95.23%**	**0.81**	**100.00%**

[a] *Extreme accuracy* measures prediction accuracy on lipids whose ground-truth mRNA transfection efficiency scores are at the extremes (i.e., 1, 2, 9, and 10). These cases correspond to either highly ineffective or highly effective lipids and are particularly informative for biological screening, especially for identifying candidates with very high transfection efficiency (scores 9 and 10).
[b] *Middle accuracy* measures prediction accuracy on lipids with intermediate ground-truth efficiency scores ranging from 3 to 8.

extreme cases are particularly critical in practical lipid screening, where highly efficient candidates (scores 9 and 10) are prioritized for costly experimental validation and highly inefficient candidates must be reliably filtered out.

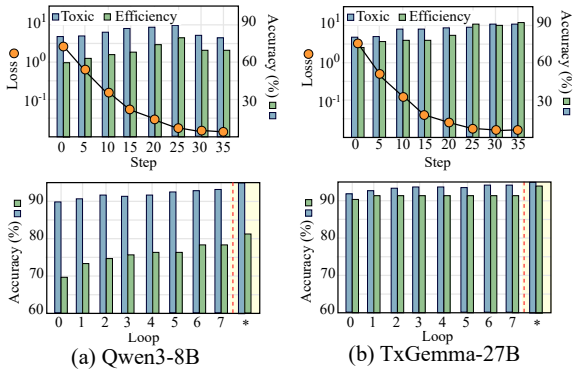

Figure 4: Training and inference dynamics of LipoAgent.Top: training loss, toxicity accuracy, and mRNA transfection efficiency.Bottom: toxicity and efficiency accuracy across verification loops."*" indicates the introduction of human feedback.

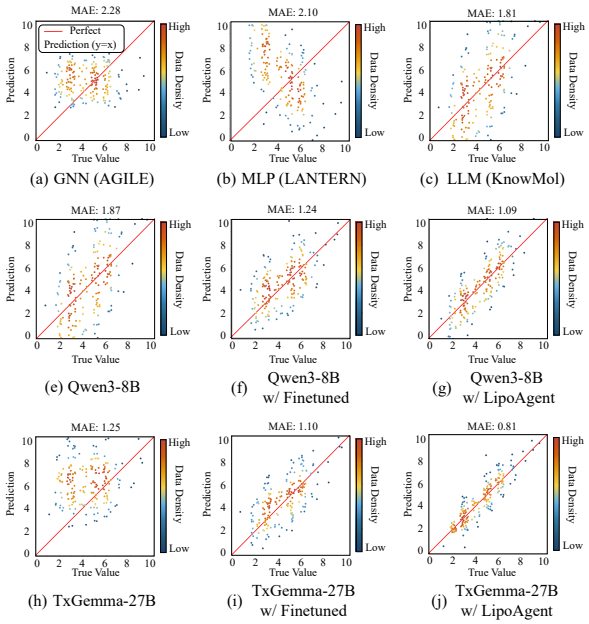

Figure 5: Prediction distributions of multiple baselines and LipoAgent variants on the test set. The red diagonal line ($y = x$) indicates ideal predictions.

Figure 4 further analyzes training and inference dynamics. During fine-tuning, smaller base models occasionally exhibit initial accuracy gains followed by degradation in later stages, indicating overfitting under limited training data. Accordingly, for subsequent experiments and wet-lab validation (Section 4.4), we selected checkpoints with the best validation performance rather than the final training checkpoints.

The figure also highlights the effect of the multi-agent verification loop. Prediction accuracy for both toxicity and efficiency improves steadily with additional verification rounds and converges after several iterations. However, toxicity accuracy does not reach 100% through autonomous multi-agent interaction alone. In safety-critical drug delivery scenarios, even a small number of false-negative toxicity predictions can pose serious risks. By introducing lightweight human feedback only when agent disagreement persists, LipoAgent achieves

Table 3: Ablation study on the timing of human-in-the-loop intervention.

Loop	Efficiency Acc. (%)	Toxic Acc. (%)	MAE
No Human	70.40%	90.30%	1.24
1	72.56%	100.00%	1.12
2	74.20%	100.00%	1.10
3	**76.80%**	**100.00%**	**1.09**
4	75.32%	100.00%	1.07
5	74.01%	100.00%	1.09
6	74.11%	100.00%	1.09

Table 4: Ablation study on the loss weight α.

α	Efficiency Acc. (%)	Toxic Acc. (%)	MAE
0.00	55.23%	90.41%	1.44
0.05	65.44%	90.30%	1.31
0.10	**70.40%**	**90.30%**	**1.24**
0.15	70.50%	88.63%	1.24
0.20	70.40%	86.32%	1.24

perfect toxicity prediction accuracy while further improving efficiency prediction reliability.

Figure 5 visualizes prediction distributions across modeling paradigms. Without domain-specific fine-tuning or verification, most models are biased toward intermediate efficiency scores (typically 4–6), a common failure mode in screening tasks. With fine-tuning and multi-agent verification, predictions increasingly concentrate along the ideal $y = x$ diagonal. Notably, LipoAgent improves accuracy at extreme efficiency values, enabling reliable identification of both highly promising and clearly unsuitable lipid candidates.

4.3 Ablation Study

We conduct ablation studies to analyze two key design choices in LipoAgent: (1) the timing of human-in-the-loop intervention during multi-agent verification, and (2) the weighting factor α in the conditional multi-task loss. All ablation experiments follow the setting in Section 4.1.

Human-in-the-loop intervention timing. As shown in Table 3, the timing of human feedback critically affects accuracy, efficiency, and practical usability. Introducing human feedback too early results in excessive manual review and limits the effectiveness of autonomous multi-agent reasoning. Conversely, introducing it too late leads to prolonged agent disagreement, during which one agent may unduly influence the other, undermining reliable arbitration. Triggering human intervention only after three unsuccessful verification loops provides the best balance between autonomy and safety in this safety-critical setting.

Effect of loss weight α. Table 4 reports ablation results for the loss weighting factor α. When α is too small, optimization is dominated by toxicity prediction, limiting gains in efficiency accuracy. As α increases, efficiency prediction improves accordingly. We find that $\alpha = 0.1$ yields the most favorable trade-off, substantially improving efficiency prediction without degrading toxicity performance. Larger values of α offer no additional benefit and may slightly compromise toxicity accuracy, highlighting the importance of balanced conditional optimization.

4.4 Wet Experimental Validation

To evaluate whether the predictions of LipoAgent translate into real biological performance, we conducted wet-lab experiments on lipid candidates selected via large-scale virtual screening. We first construct a virtual lipid library containing 10,024 lipid molecules that are feasible to synthesize. Using the best-performing configuration, *TxGemma-27B with LipoAgent*, we perform *in silico* screening over the entire library, jointly predicting toxicity and mRNA transfection efficiency for each lipid.

From the screening results, we selected **four** representative lipid candidates (Figure 6A) that are predicted to be non-toxic and span three distinct efficiency levels (high, medium, and low). Importantly, this selection strategy is designed to validate the *relative ordering* of model predictions across the efficiency spectrum, rather than to showcase only top-ranked candidates. The detailed synthetic routes for all four selected lipid molecules are provided in Appendix A. LNPs are formulated using the ethanol injection method (Xu et al., 2025) and evaluated *in vitro* in B16F10 cells (Appendix B).

As shown in Figure 6B, Lipid-1710 exhibits the highest mRNA transfection efficiency among the four tested candidates, consistent with its predicted ranking. The remaining three lipids demonstrate intermediate and lower efficiencies in accordance with their predicted relative ordering, supporting the robustness of the model's ranking capability rather than isolated hit identification. Consistent trends were further confirmed by fluorescence imaging using EGFP mRNA (Figure 6C).

The model-generated explanations provide plausible structure–function hypotheses for the observed trends. Compared with Lipid-1721, Lipid-

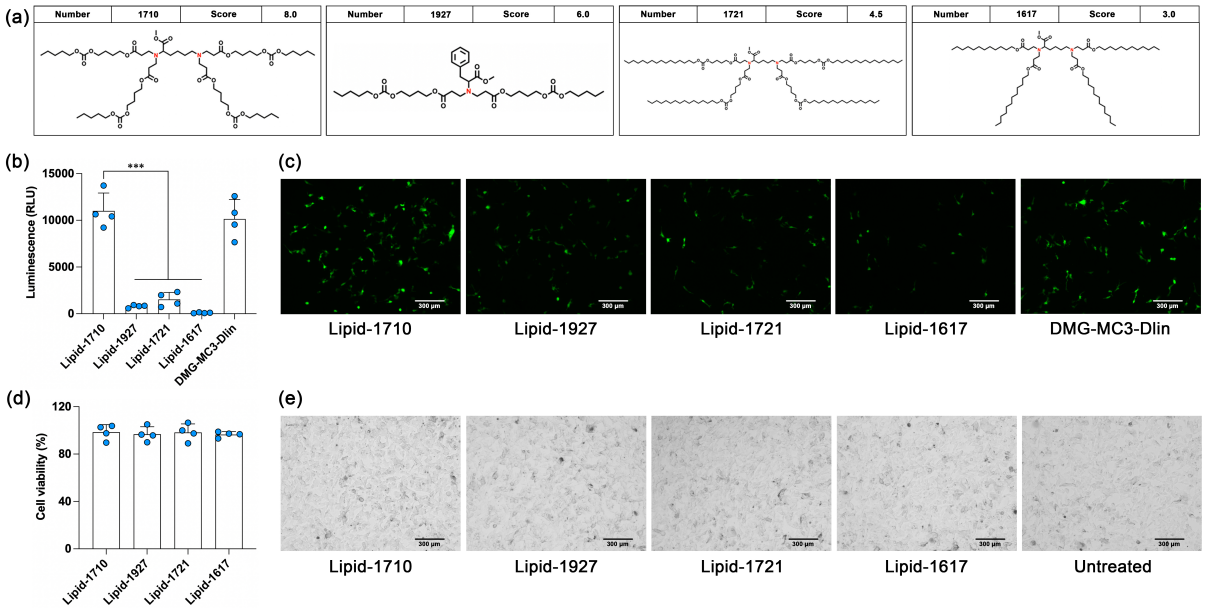

Figure 6: In vitro evaluation of LNPs derived from predicted ionizable lipids. (a) The structures of lipids with different scores. (b) Luciferase (LUC) mRNA transfection efficiency of LNPs derived from the four predicted lipids in B16F10 cells. (c) Fluorescence imaging of B16F10 cells treated with EGFP mRNA-loaded LNPs derived from predicted lipids. (d) Cell viability of LNPs derived from predicted lipids evaluated by MTT assay. (e) Representative imaging of B16F10 cells treated with predicted lipid-derived LNPs. Scale bar: 300 μm.

1710 features an optimized hydrophobic tail length that may facilitate improved interactions with cellular membranes. Relative to Lipid-1617, the presence of a biodegradable carbonate linkage in Lipid-1721 may promote more efficient intracellular mRNA release. In comparison with Lipid-1927, the headgroup architecture and number of hydrophobic tails in Lipid-1710 appear more favorable for effective mRNA delivery. Notably, under the same *in vitro* experimental conditions, Lipid-1710 achieved mRNA transfection efficiency comparable to that of the commercially used lipid DMG-MC3-Dlin, serving as a benchmark.

MTT assays demonstrate negligible cytotoxicity for all four tested lipids (Figure 6D), and no significant differences in cell density or cellular morphology were observed between LNP-treated and untreated cells (Figure 6E). Collectively, these results validate the reliability of LipoAgent in predicting relative mRNA transfection performance and highlight its potential to guide lipid design.

4.5 Human Effort and Time Efficiency

Beyond predictive accuracy, an important objective of LipoAgent is to reduce human effort required for lipid screening. We compare our multi-agent framework with a conventional human-driven workflow, where experts synthesize large lipid libraries and experimentally evaluate candidates without automated filtering or agent-based verification.

In traditional pipelines, lipid synthesis is the dominant bottleneck. In our wet-lab setting, synthesizing **four** lipids required approximately **96** hours of hands-on time, corresponding to about **24** hours per lipid. Extrapolating to a virtual library of **10,024** candidates, exhaustive manual synthesis would require on the order of **240,000** hours, rendering experimental screening infeasible.

In contrast, LipoAgent completes *in silico* screening of the full **10,024**-lipid library within approximately **23** hours and identifies the top **0.1%** candidates (**10** lipids) for downstream validation. Synthesizing these shortlisted lipids requires approximately **10** days, resulting in a total turnaround time of about **264** hours. Overall, LipoAgent reduces the end-to-end time by approximately **99.89%** while maintaining strict safety guarantees, as human supervision is invoked only for a small number of ambiguous cases.

5 Conclusion

We present LipoAgent, a safety-aware multi-agent LLM framework that jointly models toxicity and mRNA transfection efficiency for lipid discovery. Experiments and wet-lab validation show that LipoAgent can greatly improve prediction accuracy and reliability while ensuring safety-critical decision making.

6 Limitations

Despite its effectiveness, LipoAgent has several limitations. First, the training data remain limited in scale and are aggregated from heterogeneous experimental sources, which may introduce residual noise despite manual normalization. Second, mRNA transfection efficiency is modeled using discretized scores, which facilitates ranking but may obscure fine-grained quantitative differences between lipid candidates. Third, toxicity assessment relies on available molecular toxicity datasets and in vitro assays, and does not fully capture complex in vivo safety profiles. Fourth, LipoAgent is designed to predict the mRNA transfection efficiency of given lipid candidates rather than to directly generate novel, high-efficiency lipid molecules in an end-to-end manner. Finally, although wet-lab experiments validate the predictive ordering of selected lipids, experimental evaluation is conducted on a limited number of candidates. Future work will focus on expanding standardized datasets, improving continuous efficiency modeling, incorporating richer toxicity annotations, extending the framework toward generative lipid design, and scaling experimental validation.

References

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2025. Researchagent: Iterative research idea generation over scientific literature with large language models. *Preprint*, arXiv:2404.07738.

Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. 2023. Chemcrow: Augmenting large-language models with chemistry tools. *Preprint*, arXiv:2304.05376.

Hiam Chemaitelly, Houssein H. Ayoub, Sarah AlMukdad, Patrick Coyle, Patrick Tang, Hadi M. Yassine, Hiam A. Al-Khatib, Muna K. Smatti, Mohammad R. Hasan, Zaina Al-Kanaani, Eman Al-Kuwari, Andrew Jeremijenko, Aisha H. Kaleeckal, Anvar N. Latif, Ranya M. Shaik, Hiam F. Abdul-Rahim, Ghada K. Nasrallah, Maryam G. Al-Kuwari, Adeel A. Butt, and 5 others. 2022. Protection from previous natural infection compared with mrna vaccination against sars-cov-2 infection and severe covid-19 in qatar: a retrospective cohort study. *The Lancet Microbe*, 3(12):e944–e955.

Piotr Gaiński, Oussama Boussif, Andrei Rekesh, Dmytro Shevchuk, Ali Parviz, Mike Tyers, Robert A. Batey, and Michał Koziarski. 2025. Scalable and cost-efficient de novo template-based molecular generation. *Preprint*, arXiv:2506.19865.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Kexin Huang, Tianfan Fu, Lucas M Glass, Marinka Zitnik, Cao Xiao, and Jimeng Sun. 2020. Deeppurpose: a deep learning library for drug–target interaction prediction. *Bioinformatics*, 36(22–23):5545–5547.

Bo Li, Ibrahim O. Raji, Alexander G. R. Gordon, Lu Sun, Thomas M. Raimondo, Fisayo A. Oladimeji, A. Y. Jiang, Andrew Varley, Robert S. Langer, and Daniel G. Anderson. 2024. Accelerating ionizable lipid discovery for mrna delivery using machine learning and combinatorial chemistry. *Nature Materials*, 23(7):1002–1008.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for "mind" exploration of large language model society. *Preprint*, arXiv:2303.17760.

Kun Li, Zhennan Wu, Shoupeng Wang, Jia Wu, Shirui Pan, and Wenbin Hu. 2025. Drugpilot: Llm-based parameterized reasoning agent for drug discovery. *Preprint*, arXiv:2505.13940.

Sizhe Liu, Yizhou Lu, Siyu Chen, Xiyang Hu, Jieyu Zhao, Yingzhou Lu, and Yue Zhao. 2025. Drugagent: Automating ai-aided drug discovery programming through llm multi-agent collaboration. *Preprint*, arXiv:2411.15692.

A. Lu, Z. Xu, Z. Zhao, Y. Yan, L. Jiang, J. Geng, H. Jin, X. Wang, X. Liu, Y. Zhu, Y. Shi, L. Liu, H. Dai, and J. C. Wang. 2024. Double braking effects of nanomedicine on mitochondrial permeability transition pore for treating idiopathic pulmonary fibrosis. *Advanced Science*, 11(47):e2405406.

Jiang Lu. 2024. The toxicity data of compounds.

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. *Preprint*, arXiv:2507.03209.

Miguel Ramos, Chen Li, Arjun Patel, Wei Zhang, and Lucia Torres. 2025. A review of large language models and autonomous agents in chemistry. *Chemical Science*, 16(4):1234–1256.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Preprint*, arXiv:2303.11366.

Eric Wang, Samuel Schmidgall, Paul F. Jaeger, Fan Zhang, Rory Pilgrim, Yossi Matias, Joelle Barral, David Fleet, and Shekoofeh Azizi. 2025. Txgemma: Efficient and agentic llms for therapeutics. *Preprint*, arXiv:2504.06196.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Voyager: An open-ended embodied agent with large language models. *Preprint*, arXiv:2305.16291.

Wei Wang, Kai Chen, Tao Jiang, and 1 others. 2024. Artificial intelligence-driven rational design of ionizable lipids for mrna delivery. *Nature Communications*, 15:10804.

S. Xu, Z. Hu, F. Song, Y. Xu, and X. Han. 2025. Lipid nanoparticles: Composition, formulation, and application. *Molecular Therapy – Methods & Clinical Development*, 33(2):101463.

Y. Xu, S. Ma, H. Cui, and 1 others. 2024. Agile platform: a deep learning powered approach to accelerate lnp development for mrna delivery. *Preprint*, Nat Commun 15, 6305 (2024).:2303.11366.

Yu Yan, Xinyu Liu, An Lu, Xinyu Wang, Lixin Jiang, and Jianchun Wang. 2022. Non-viral vectors for rna delivery. *Journal of Controlled Release*, 342:241–279.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Zaifei Yang, Hong Chang, Ruibing Hou, Shiguang Shan, and Xilin Chen. 2025b. Knowmol: Advancing molecular large language models with multi-level chemical knowledge. *Preprint*, arXiv:2510.19484.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. *Preprint*, arXiv:2210.03629.

Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Wanli Ouyang, Dongzhan Zhou, Shufei Zhang, Mao Su, Han-Sen Zhong, and Yuqiang Li. 2024. Chemllm: A chemical large language model. *Preprint*, arXiv:2402.06852.

Y. Zhao, Z. M. Wang, D. Song, M. Chen, and Q. Xu. 2024. Rational design of lipid nanoparticles: overcoming physiological barriers for selective intracellular mrna delivery. *Current Opinion in Chemical Biology*, 81:102499.

Yizhen Zheng, Huan Yee Koh, Maddie Yang, Li Li, Lauren T. May, Geoffrey I. Webb, Shirui Pan, and George Church. 2024. Large language models in drug discovery and development: From disease mechanisms to clinical trials. *Preprint*, arXiv:2409.04481.

Yu Zong, Yifan Lin, Tingting Wei, and Qiang Cheng. 2023. Lipid nanoparticle (lnp) enables mrna delivery for cancer therapy. *Advanced Materials*, 35(51):e2303261.

A Synthetic Routes of Selected Lipids

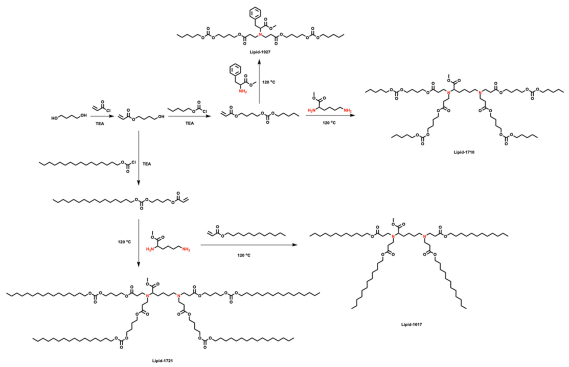

Figure 7: The synthetic routes of four selected lipids.

Lipid synthesis. The chemical structures and synthetic routes of the selected lipids are illustrated in Figure 7. Specifically, Lipid-1710 is synthesized as follows. A mixture of butyleneglycol, acryloyl chloride, and triethylamine is dissolved in tetrahydrofuran (THF) and stirred at room temperature for 30 minutes. After solvent removal under reduced pressure, the crude product is purified by column chromatography to obtain intermediate aT1. Subsequently, aT1 is reacted with pentyl chloroformate in dichloromethane (DCM) at room temperature for 30 minutes, followed by solvent removal and column chromatography to yield intermediate T1. Finally, T1 is reacted with H-Lys-OMe in methanol at 120 °C for 4 hours, and the product is purified by column chromatography to obtain Lipid-1710.

Lipid-1927 is synthesized by reacting intermediate T1 with H-Phe-OMe in methanol at 120 °C for 4 hours, followed by solvent removal and column chromatography purification. For Lipid-1721, intermediate aT1 is first reacted with hexadecyl carbonochloridate in DCM at room temperature for 30 minutes to generate intermediate T2. T2 is then reacted with H-Lys-OMe in methanol at 120 °C for 4 hours, followed by purification to obtain Lipid-1721. Lipid-1617 is synthesized by directly reacting dodecyl acrylate with H-Lys-OMe in methanol at 120 °C for 4 hours, followed by purification.

B Experimental Protocols for LNP Preparation and Evaluation

As shown in Figure 8, the synthesized lipid is used to prepare LNPs loading mRNA, and then the prepared LNPs are evaluated using B16F10 cells. luciferase (LUC) reporter assay and EGFP reporter assay are used to detect the in vitro mRNA transfection. MTT assay is used to detect the cytotoxicity of LNPs.

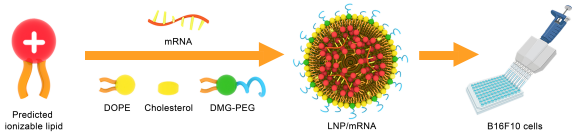

Figure 8: The schematic illustration of LNP preparation and in vitro cellular evaluation procedures.

Preparation of lipid nanoparticles (LNPs). For LNP formulation, mRNA is first dissolved in citrate buffer (pH = 4) to obtain solution A. Ionizable lipid, DOPE, cholesterol (Chol), and DMG-PEG2000 are dissolved in ethanol to obtain solution B. Solution B is then mixed with solution A at a volume ratio of B/A = 1/3. The resulting mixture is dialyzed overnight against RNase-free PBS using a dialysis membrane with a molecular weight cutoff of 10,000–12,000 Da to obtain stable LNP formulations.

Transfection efficiency evaluation. mRNA transfection efficiency is evaluated using B16F10 cells. Cells are seeded into 96-well plates at a density of 20,000 cells per well and incubated with 100 μL of DMEM medium containing LUC mRNA-loaded LNPs for 24 hours. Luciferase expression levels are quantified using the Firefly Luciferase Reporter Gene Assay Kit, with the final LUC mRNA concentration fixed at 100 ng per well. For imaging-based evaluation, B16F10 cells are treated with enhanced green fluorescent protein (EGFP) mRNA-loaded LNPs for 24 hours and subsequently observed under a fluorescence microscope.

MTT assay. B16F10 cells (20000 cells per well) were seeded into 96-well plates and incubated with 100 uL DMEM medium containing mRNA-loaded LNPs for 24 h. Then CyQUANTTM MTT Cell Proliferation Assay Kit was used to detect the cytotoxicity of LNPs.