Research Article

Evaluation of Vector Representations of Lipid Nanoparticles in Cheminformatic Predictions of Transfection Efficiency

Shruti Patel

BASIS Independent Silicon Valley, 1290 Parkmoor Ave, San Jose, CA 95126, United States

ABSTRACT

Lipid nanoparticles (LNPs) are a revolutionary drug delivery system for RNA-based therapeutics, as they are difficult to degrade and can efficiently transport their contents to distant target cells. Optimizing LNP formulations is essential for improving therapies, yet the best method to computationally represent these formulations for predictive modeling remains unexplored. This study analyzes different strategies for constructing the formulation vector of LNPs to evaluate their impact on predicting transfection efficiency. Specifically, three approaches were examined: fully describing all lipid components using molecular descriptors, fully describing only the cationic lipid while incorporating molar ratios for other components, and fully describing both the ionizable and helper lipids while using molar ratios for rest. Machine learning models were trained using each formulation representation, revealing minimal differences in prediction accuracy. The results suggest that the structures of the cationic and helper lipids are most critical, and including molecular descriptors for the PEGylated (PEG) lipid and cholesterol may introduce excess noise in the neural network without improving its performance. This can streamline LNP formulation research, which traditionally takes years of testing to design specific LNPs. By identifying effective strategies to represent LNP formulations, this work contributes to optimizing RNA-based drug delivery systems.

Keywords: Lipid nanoparticles; drug delivery; machine learning; neural networks; RNA therapeutics; transfection efficiency; cheminformatics; molecular descriptors

INTRODUCTION

Lipid nanoparticles (LNPs) have transformed drug delivery due to their ability to transfer nucleic acids, target specific tissue or cell types, and produce quicker than viral vectors (1). Specifically, their clinical relevance was notably demonstrated during the COVID-19 pandemic of delivering mRNA vaccines to muscle cells, signaling the production of the SARS-CoV-2 spike protein (2). This triggers an immune response to help the body recognize and eliminate the virus (2). Despite these advances, optimizing LNP formulations for efficient intracellular delivery remains challenging due to the chemical diversity of lipid components and the complexity of their interactions (3).

Conventional experimental approaches to optimize LNP formulation are time-consuming and resource-intensive, presenting significant barriers during urgent public health crises. Hence, there is a critical need for

Corresponding author: Shruti Patel, E-mail: shrutipatel2026@gmail.com.
Copyright: © 2025 Shruti Patel. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
Received May 27, 2025**; Accepted** July 8, 2025
https://doi.org/10.70251/HYJR2348.342328

computational models that accurately represent LNPs. Machine learning, in combination with cheminformatics, could predict delivery efficiency based on molecular representations of lipid components.

LNPs are composed of four primary lipid components: ionizable (cationic) lipids, helper phospholipids, cholesterol, and PEGylated (PEG) lipids (4). Each component plays a distinct role in nanoparticle stability, cellular uptake, endosomal escape, and additional functions. The cationic lipids carry a positive charge at a low pH which allows them to bind to the negatively charged mRNA being transported (5). The phospholipids act as the structural backbone of the LNP, while the cholesterol is the stiffening agent of the membrane, as it maintains its fluidity and prevents early degradation. The PEGylated lipid is the protective coating of the LNP that prevents clearance from the body and enhances stability and circulation time of the formulation in the bloodstream (6). The chemical properties of these lipids, such as molecular weight affect the overall delivery performance of LNPs (7). Molecular descriptors derived from SMILES (Simplified Molecular Input Line Entry System) strings capture key physicochemical and structural properties and have been applied to various molecular property prediction tasks. However, their application to modeling LNP delivery performance remains underexplored (8).

This study investigates different strategies for representing LNP formulations in machine learning models to evaluate their impact on the accuracy of transfection efficiency predictions. Three approaches were compared: 1) using molecular descriptors for all lipid components, 2) fully describing only the cationic lipid while adding molar ratios for the remaining components, and 3) fully describing both the ionizable and helper lipids while using molar ratios for the rest. We hypothesized that fully describing the cationic and helper lipids would yield the highest prediction accuracy, given their greater structural variation across formulations, while detailed inclusion of PEGylated lipids and cholesterol – which are relatively invariant – could introduce noise and diminish model performance. By identifying efficient representations of LNP formulations, this study aims to assist in the development of future LNP drug delivery systems.

METHODS AND MATERIALS

Data preparation

The independent variable in this experiment is the strategy used to construct the formulation vector representation, while the dependent variable is the prediction accuracy of the unnormalized delivery of the LNPs. In this study, the unnormalized delivery refers to the measure of how effectively the LNPs facilitate the intracellular delivery of the mRNA cargo. It is a metric derived from the fluorescence intensity emitted by a fluorescent protein expressed within the target cells after the mRNA is delivered, so it is directly proportional to the amount of LNPs that successfully enter the cells and release their mRNA (9). A greater amount of delivery is indicated by a higher fluorescence intensity. The unnormalized delivery value is a raw measure of this transfection efficiency before any adjustments are applied. A limitation of this target is that the metric does not control for cell types. However, at this time, this dataset is the largest and most diverse dataset of LNP delivery efficiency. The control variable is the strategy that fully describes all four structural components of the LNPs, as it represents the raw and complete data as a baseline for comparison to the other models.The data preparation process begins with the importation of essential Python libraries and the uploading of the dataset. The libraries used include, Mordred for molecular descriptor calculations, RDKit for chemical informatics, Keras for building and training machine learning models, and NumPy, Pandas, and Scikit-learn for data manipulation and preprocessing.

The dataset consists of columns containing information about the SMILES string of the cationic lipid, the helper lipid type, the molar ratio of each LNP component, and the unnormalized delivery value (9). SMILES strings for the three types of helper lipids (DOPE, DSPC, DOTAP) and PEG2000, which was used in most of the LNP formulations, were obtained from the PubChem website (10). Molecular descriptor calculations were conducted using the RDKit library, beginning with the preparation of a descriptor calculator for a given SMILES string. The molecular descriptors for the helper lipids, PEG lipid, and cholesterol were then computed and added to the dataset. Additionally, the molecular descriptors of the cationic lipids were calculated, and the normalized molar concentrations were determined as ratios based on the given values of each component.

As shown in Figure 1 below, the histogram of the unnormalized delivery values were heavily skewed right with outliers reaching as far as 37,800, creating an output that was almost unusable to train the model by. This is likely because the dynamic range of unnormalized delivery is too wide, making it difficult for the model to train and learn a distribution that covers multiple orders of magnitude. Thus, the data was shifted to become positive

and re-expressed using the logarithm function. Since the lowest negative unnormalized delivery value was around -3.221, 3.3 was added to all the data points. When this transformed data was plotted, the right skew was reduced and the distribution became more normal, resulting in an output that the model could be easily trained on under one order of magnitude, as expressed in Figure 1B below.

Formulation representation strategy

As displayed in Figure 2, there are three different strategies used in this experiment to analyze which method is the most accurate. In Strategy 1, which is the control variable, all four components are fully described from the dataset using the RDKit library, including the cationic lipid, helper lipid, PEG lipid, and cholesterol. Each component is multiplied by their normalized molar ratios, and then all the components are concatenated into the final formulation vector. In Strategy 2, only the cationic lipids are fully described and multiplied by their molar ratios. The helper lipid, PEG lipid, and cholesterol molar ratios are directly added to the concatenation along with the described cationic lipids. In Strategy 3, the

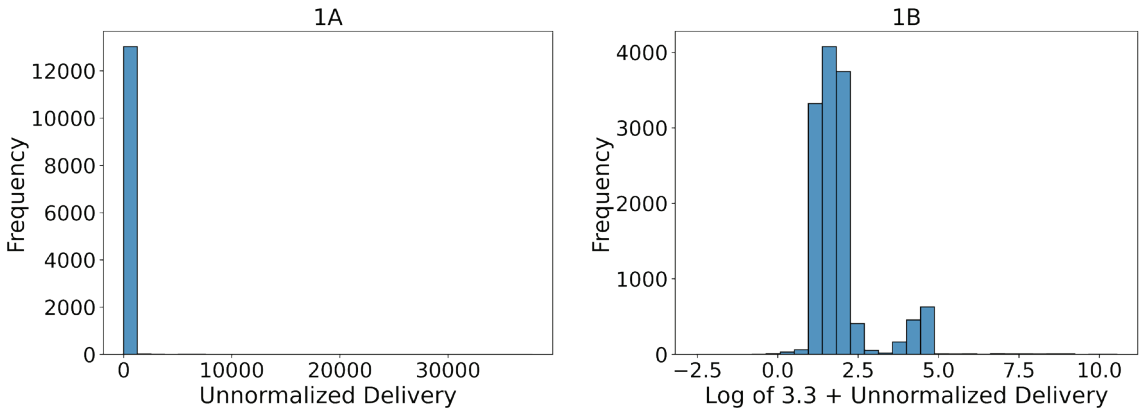

Figure 1. Distribution of target value. (1A) Unnormalized delivery distribution (1B) log-transformed delivery distribution. In 1B, 3.3 was added to the data points prior to the logarithmic transformation to ensure all data points were positive. After applying the logarithmic function, there is a reduced right skew compared to 1A, so the model could easily train on the re-expressed data.

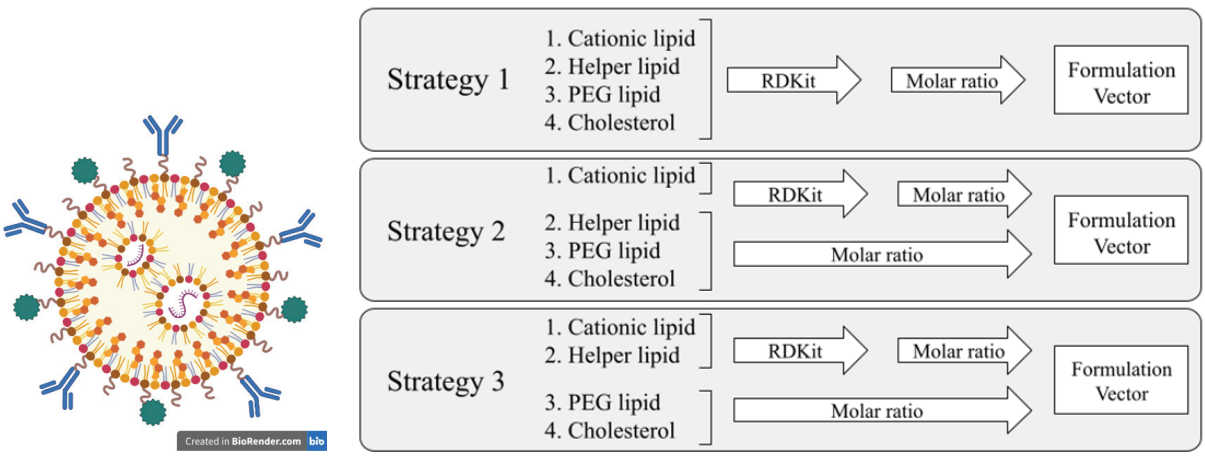

Figure 2. LNP structure and the 3 ML representation strategies. (left) Schematic depicting an LNP loaded with mRNA (right) 3 strategies for encoding LNP formulations into ML input vectors. The four main components of an LNP's structure are cationic lipids, helper lipids, PEG lipids, and cholesterol. In Strategy 1, all components were fully described using RDKit to construct the concatenated formulation vector. In Strategy 2, only the cationic lipid was described, and in Strategy 3, only the cationic and helper lipids were described, while the remaining components were simply weighed by their molar ratios before being added to the formulation vector and inputted into the neural network.

cationic lipids and helper lipids are fully described with the PEG lipid and cholesterol molar ratios directly added to the final formulation vector. Strategies 2 and 3 were crafted this way to describe only the cationic and helper lipids due to their primary roles in delivery performance, such as preventing degradation; the cholesterol and PEG lipids do not vary between each LNP in the dataset and are not as significant in delivery. Each strategy is now prepared to be used in the models, which will be created in the following step.

Data splitting and model training

The dataset was randomly split into training and testing sets to evaluate the models' performance. Subsequently, neural networks were constructed using Keras framework for each of the three strategies. Each model structure included an input layer, a hidden layer, a dropout layer to mitigate overfitting, a second hidden layer, a second dropout layer, and an output layer. The target variable Y was re-expressed by applying a logarithmic transformation to the shifted unnormalized delivery values, thereby accounting for variance and enhancing the model's predictive accuracy. Each model was compiled using an appropriate optimizer and loss function, and training was performed with predefined hyperparameters. The metrics evaluated were mean squared error (MSE), mean absolute error (MAE), and the coefficient of determination (R^2).

RESULTS

Strategy 1, which fully described all lipid components (including ionizable lipids, helper lipids, cholesterol, and PEG lipids) using RDKit, performed the least well in terms of both MAE and R^2 (Figure 3, Table 1). This suggests that including molecular descriptors for all lipid components resulted in a model that, although comprehensive, may

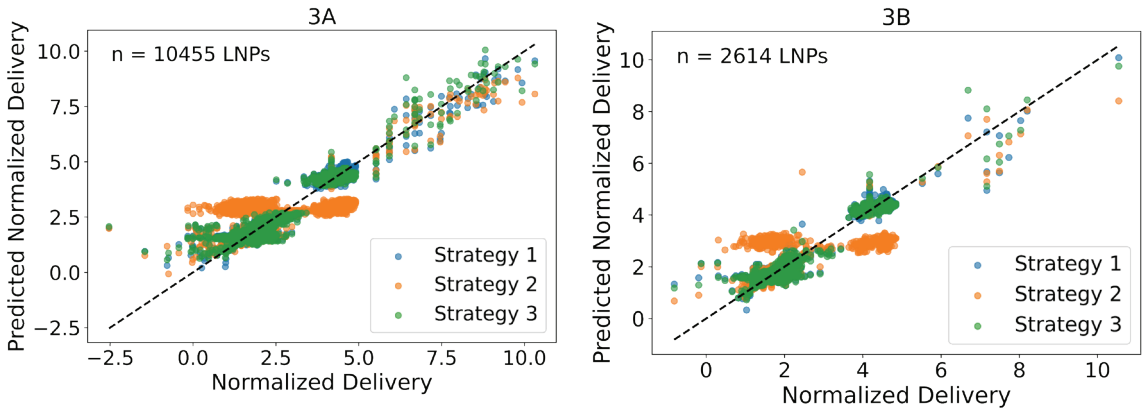

Figure 3. Parity plots comparing predicted and actual delivery values for all three strategies. (3A) Results on the training set (n = 10,455 LNPs). (3B) Results on the test set (n = 2,614 LNPs). Each point represents a single observation; the dashed line indicates perfect parity. Different colors represent different strategies: Strategy 1 (blue), Strategy 2 (orange), Strategy 3 (green). Strategy 2 shows the broadest dispersion around the parity line, resulting in the highest error of all the strategies. Strategies 1 and 3 mostly overlap, but Strategy 3 is clustered more closely to the parity line. Thus, Strategy 3's parity plot demonstrates the most accurate predictions and highest agreement between actual and predicted delivery values.

Table 1. Results of each metric (mean absolute error (MAE), mean squared error (MSE), and coefficient of determination (R^2)) for the three LNP formulation strategies

Metric	Strategy 1	Strategy 2	Strategy 3
MAE	0.16	0.38	0.15
MSE	0.05	0.42	0.05
R^2	93.98%	54.21%	94.42%

One-way ANOVA testing revealed a statistically significant difference in R^2 values across strategies with cross validation ($F = 63.34$, $p < 0.00001$, $\alpha = 0.01$), but no significant difference for MAE values ($F = 0.04021$, $p = 0.960721$, $\alpha = 0.01$). The results indicate that formulation representation significantly impacts model performance.

have introduced extraneous noise into the prediction, thereby compromising the predictive value. The additional data from the cholesterol and PEG lipids did not contribute significantly to the accuracy of the predictions. However, future work should quantitatively assess multicollinearity, low variance, or redundancy among the molecular descriptors of cholesterol and PEG lipids. Such analysis would confirm whether these components genuinely introduce noise or simply lack predictive value due to low variability across formulations. Additionally, from Table 1, the R^2 value of 93.98% indicates that while the model explained a reasonable proportion of the variability in unnormalized delivery, it still left substantial room for improvement.

Strategy 2, which described only ionizable lipids with RDKit and used only molar ratios for other components, performed slightly worse than Strategy 3 with a MAE of 0.38 and an R^2 of 54.21% (Figure 3). This notably low R^2 value may seem surprising, but after looking at the parity plot in Figure 3, Strategy 2 had the largest outliers relative to the other two strategies. Although ionizable lipids were expected to be a key factor, the exclusion of descriptors for the helper lipids significantly reduced the model's accuracy relative to Strategy 3, which included these descriptors. This finding shows that the helper lipids play an essential role, similar to the cationic lipids, and should be considered when representing LNP formulations.

Strategy 3, which described only cationic and helper lipids but not cholesterol and PEG, produced the highest R^2 value of 94.42% and a relatively low MAE of 0.15 (Figure 3, Table 1). This suggests that the cationic and helper lipids play the most significant role in the delivery efficiency of LNPs, and therefore, accurately modeling these components is sufficient to achieve optimal prediction performance. By focusing only on the components with the most variance, the formulation vector for this strategy avoided the noise introduced by the helper lipids and cholesterol, therefore achieving higher accuracy.

Statistical analysis using one-way ANOVA confirmed that the differences in R^2 values across strategies were statistically significant ($F = 63.34$, $p < 0.00001$, $\alpha = 0.01$), which validates the claim that formulation representation influences model performance. However, the differences in MAE were not statistically significant ($F = 0.04021$, $p = 0.960721$, $\alpha = 0.01$) and were well above the significance level. This suggests that although Strategy 3 had the lowest MAE, the observed differences could have occurred due to random variation. MAE is also known to be sensitive to outliers, which may have skewed the results and masked potential differences.

Strategy 3 likely outperformed the others because it focused on the components with the most variation (cationic and helper lipids) that have the most direct influence on the LNPs' transfection efficiencies. This approach streamlined the model by excluding components, such as cholesterol and PEG lipids, which did not seem to significantly improve prediction accuracy. It seems that cationic and helper lipids are the primary contributors to the delivery mechanism of the LNPs. By focusing on these two components, the model was able to more effectively predict transfection efficiency. These findings suggest that researchers working on lipid nanoparticle formulations should prioritize the cationic and helper lipids when developing prediction models for RNA delivery systems. This focus may help reduce the time required for experimental optimization of LNP formulations, minimizing the need for extensive data on other lipid components, such as PEG lipids and cholesterol. Despite there being only three types of helper lipids in the dataset, these results display that they play a significant role in prediction accuracy. Additionally, these models show that scaling features by the mole ratios provided by the dataset are a good way to encode this type of information relating to transfection efficiency.

CONCLUSION

This study evaluated different formulation vector representation strategies for predicting the unnormalized delivery of lipid nanoparticles. The results indicate that Strategy 3, which fully describes only cationic and helper lipids using RDKit but not cholesterol and PEGylated lipid, achieved the highest prediction accuracy, with the lowest mean absolute error (MAE = 0.15) and the highest R^2 value (94.42 %). These findings provide compelling evidence that including the molecular descriptors of the cationic and helper lipids are crucial for accurately predicting transfection efficiency. However, including components like PEG lipids and cholesterol does not detract the model's performance, as much as solely describing the cationic lipids. The model was able to discern that the PEG lipids and cholesterol are less influential than the cationic and helper lipids, as shown in Strategy 1.

Nevertheless, an important limitation of this study is the restricted diversity of lipid types in the dataset as only three types of helper lipids and a single PEG lipid were included. This narrow compositional scope may constrain the generalizability of the conclusions, particularly those related to the relative importance of helper lipids. Further work using a more chemically diverse dataset is needed to

validate these findings.

Beyond this project, more research should be conducted to fully understand the mechanistic properties that drive differences in transfection efficiency. It remains unclear whether the key factor is primarily particle fluidity, which is structural rigidity induced by different lipid compositions, or the presence of higher-order features that cannot be captured solely through molecular descriptors. Further investigation of these properties may reveal formulation strategies that result in a higher correlation with delivery efficiency when modeled. Exploring the broader properties of LNPs could contribute to the development of a more comprehensive model for predicting transfection efficiency. Additionally, further research could be done to understand the interactions between lipid composition, nanoparticle stability, and intracellular trafficking drive the next models aiming to optimize LNP formulations.

DECLARATION OF CONFLICT OF INTERESTS

The author declares that there are no conflicts of interest regarding the publication of this article.

REFERENCES

1. Tenchov R, Bird R, Curtze AE, Zhou Q. Lipid nanoparticles─From liposomes to mRNA vaccine delivery, a landscape of research diversity and advancement. *ACS Nano*. 2021; 15 (11): 16982–17015. https://doi.org/10.1021/acsnano.1c04996
2. Zhang L, More KR, Ojha A, Jackson CB, *et al*. Effect of mRNA-LNP components of two globally-marketed COVID-19 vaccines on efficacy and stability. *NPJ Vaccines*. 2023; 8 (1): 1–14. https://doi.org/10.1038/s41541-023-00751-6
3. Wang J, Ding Y, Chong K, Cui M, *et al*. Recent advances in lipid nanoparticles and their safety concerns for mRNA delivery. *Vaccines*. 2024; 12 (10): 1148. https://doi.org/10.3390/vaccines12101148
4. Hald Albertsen C, Kulkarni JA, Witzigmann D, Lind M, Petersson K, Simonsen JB. The role of lipid components in lipid nanoparticles for vaccines and gene therapy. *Adv Drug Deliv Rev*. 2022; 188: 114416. https://doi.org/10.1016/j.addr.2022.114416
5. Guéguen C, Ben Chimol T, Briand M, Renaud K, *et al*. Evaluating how cationic lipid affects mRNA-LNP physical properties and biodistribution. *Eur J Pharm Biopharm*. 2024; 195: 114077. https://doi.org/10.1016/j.ejpb.2023.08.002
6. Tenchov R, Sasso JM, Zhou QA. PEGylated lipid nanoparticle formulations: Immunological safety and efficiency perspective. *Bioconjug Chem*. 2023; 34 (6): 941–960. https://doi.org/10.1021/acs.bioconjchem.3c00174
7. A Complete Guide to Understanding Lipid Nanoparticles (LNP). Inside Therapeutics. Available from: https://insidetx.com/review/complete-guide-to-understanding-lipid-nanoparticles/ (accessed on 2025-01-26).
8. Quirós M, Gražulis S, Girdzijauskaitė S, Merkys A, Vaitkus A. Using SMILES strings for the description of chemical connectivity in the Crystallography Open Database. *J Cheminform*. 2018; 10 (1): 23. https://doi.org/10.1186/s13321-018-0279-6
9. Witten J, Raji I, Manan RS, Beyer E, *et al*. Artificial intelligence-guided design of lipid nanoparticles for pulmonary gene therapy. *Nat Biotechnol*. 2024. https://doi.org/10.1038/s41587-024-02490-y
10. PubChem Compound Summary for CID 406952, Peg2000 DSPE. National Center for Biotechnology Information. Available from: https://pubchem.ncbi.nlm.nih.gov/compound/Peg2000-dspe (accessed on 2025-02-23).
11. Binici B, Rattray Z, Zinger A, Perrie Y. Exploring the impact of commonly used ionizable and PEGylated lipids on mRNA-LNPs: A combined in vitro and preclinical perspective. *J Control Release*. 2025; 377: 162–173. https://doi.org/10.1016/j.jconrel.2024.11.010
12. Ramadan E, Ahmed A, Naguib YW. Advances in mRNA LNP-based cancer vaccines: Mechanisms, formulation aspects, challenges, and future directions. *J Pers Med*. 2024; 14 (11): 1092. https://doi.org/10.3390/jpm14111092
13. Hou X, Zaks T, Langer R, Dong Y. Lipid nanoparticles for mRNA delivery. *Nat Rev Mater*. 2021; 6 (12): 1078–1094. https://doi.org/10.1038/s41578-021-00358-0
14. Jung HN, Lee SY, Lee S, Youn H, Im HJ. Lipid nanoparticles for delivery of RNA therapeutics: Current status and the role of in vivo imaging. *Theranostics*. 2022; 12 (17): 7509–7531. https://doi.org/10.7150/thno.77259