Machine Learning-guided Lipid Nanoparticle Design for mRNA Delivery

Daisy Yi Ding [1] **Yuhui Zhang** [2] **Yuan Jia** [3] **Jiuzhi Sun** [3]

Abstract

While RNA technologies hold immense therapeutic potential in a range of applications from vaccination to gene editing, the broad implementation of these technologies is hindered by the challenge of delivering these agents effectively. Lipid nanoparticles have emerged as one of the most widely used delivery agents, but their design optimization relies on laborious and costly experimental methods. We propose to in silico optimize LNP design with machine learning models. On a curated dataset of 572 LNPs from in vivo animal studies, we demonstrate the effectiveness of our model in predicting the transfection efficiency of unseen LNPs, with the multilayer perceptron achieving a classification accuracy of 98% on the test set. Our work represents a pioneering effort in combining ML and LNP design, offering significant potential for improving screening efficiency by computationally prioritizing LNP candidates for experimental validation and accelerating the development of effective mRNA delivery systems.

1. Introduction

RNA-based technologies have the potential to transform life science research and medicine by enabling one's own cells to transiently synthesize therapeutics, mechanistic probes and diagnostics. This science has created new opportunities for therapeutic vaccinations, protein replacement therapies, immunotherapy, gene editing and gene reprogramming (Kaczmarek et al., 2017; Sahin et al., 2014; Chakraborty et al., 2017). However, the single most recognized obstacle to the broad implementation of RNA-based technologies is the delivery of the polar polyanionic RNA across non-polar biological barriers (Dowdy, 2017; Stanton, 2018). Various agents have been developed to encapsulate RNA, among which lipid nanoparticles (LNPs) are one of the most widely investigated.

LNPs are often formulated with four components: (1) an ionizable or cationic lipid for complexing the polyanionic RNA, (2) a helper phospholipid to stabilize the LNP, (3) sterols for facilitating endosomal escape, and (4) lipid-anchored poly(ethylene glycol) (PEG) lipids for increasing circulation time (Eygeris et al., 2021). The functionalities of LNPs are governed by the size, shape, charge, and ratio of each component which can be optimized for specific targeting and disease applications. Seminal studies have demonstrated that the chemical structures of LNP components can influence transfection efficacy and organ selectivity through structure-activity relationships (SAR) analysis. While the previous LNP development is highly fruitful, it often requires the combinatorial synthesis of individual lipids and pooled or individual evaluation of formulated LNPs, the process of which can be laborious and costly.

We hereby aim to expand the scope of traditional experimental methods for designing LNP delivery systems by leveraging the power of computational methods. We realized that LNP design is a suitable application for exploiting ML because it provides a set of modifiable characteristics. As experimental data in this field continues to grow, and ML techniques advance in modeling chemical structures, we see great potential in using ML to optimize the design of delivery agents. Specifically, we formulate the problem as a classification task where we use supervised ML models to predict the transfection efficiency of LNPs, i.e. the ability to successfully deliver mRNA molecules into target cells and facilitate their expression.

We propose to use molecular representation learning techniques and classification models to predict the transfection efficiency of multicomponent LNPs based on their chemical structures. We first curated a dataset of 572 LNPs along with their level of transfection efficiency, characterized by luciferase expression following treatment of the LNP-encapsulated mRNA (Liu et al., 2021). When empirically evaluating different molecular representation learning methods, we found that the representations generated based on rules of chemical domain knowledge are sufficiently predictive, while those generated through large graph neural networks do not necessarily offer superior performance. On

[1]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA [2]Department of Computer Science, Stanford University, Stanford, CA, USA [3]Department of Chemistry, Stanford University, Stanford, CA, USA. Correspondence to: Daisy Yi Ding <dingd@stanford.edu>.

The 2023 ICML Workshop on Computational Biology. Baltimore, Maryland, USA, 2023.

the curated dataset, we showed that ML models effectively predict the transfection efficiency of unseen LNPs, with the multilayer perceptron achieving a classification accuracy of 98% on the test set. Such a model has the potential to significantly improve the screening efficiency by prioritizing LNP candidates for experimental validation.

To summarise, our main contributions in this paper include: (1) we propose to aid the design process of LNPs with ML by formulating it into a classification task where the input is the chemical structure of an LNP and the output is its transfection efficiency, and to our knowledge, our work represents the first attempt in this direction; (2) we curate a dataset of LNPs along with their corresponding levels of transfection efficiency, and will make this dataset public to facilitate collaborative efforts towards advancing the state-of-the-art in the optimization of LNP delivery systems; (3) through proof-of-concept studies, we demonstrate the feasibility of in silico predicting the transfection efficiency of LNP for delivering mRNA, a direction that holds great promise for advancing the development of effective LNP-based mRNA therapeutics.

2. Related Work

Machine learning for molecular representation A variety of ML models have been proposed for molecular property prediction (such as absorption, distribution, metabolism, excretion, and toxicity of small molecules), ranging from support vector machine (SVM) (Zernov et al., 2003; Alvarsson et al., 2016; Hou et al., 2007), random forest (RF) (Zhang & Aires-de Sousa, 2007; Svetnik et al., 2003), to neural networks (Wu et al., 2018; Chen et al., 2018; Rifaioglu et al., 2019; Zhang et al., 2017). Among them, graph neural network (GNN) has recently attracted considerable attention for its capability in learning representations directly from the chemical information encoded by molecular graphs (Wu et al., 2018; Sun et al., 2020; Xiong et al., 2019; Rong et al., 2020). Specifically, molecular graphs are natural representations of chemical structures: while a graph $G = (V, E)$ describes the connectivity relations between a set of nodes V and a set of edges E, a molecule can be considered as a graph consisting of a set of atoms (nodes) and a set of bonds (edges). GNN learns the representation of each atom in the molecule by aggregating the information from its neighboring atoms and connected bonds through message passing across the graph in a recursive process. However, despite the attraction GNNs have gained in the community, it has also been shown that GNNs can be outperformed by traditional descriptor-based methods with SVM and RF in terms of prediction accuracy and computational efficiency, especially in the low-data regimes (Jiang et al., 2021; Mayr et al., 2018).

Machine learning for the design of biomolecule delivery agents There has been previous work on leveraging ML to guide the design of another important type of delivery agents called adeno-associated virus (AAV) capsids (Ogden et al., 2019). The paper characterizes single-codon substitutions, insertions, and deletions of AAV capsids across functions relevant to in vivo delivery of gene therapies. It shows that the ML-guided design of AAV capsids outperforms random mutagenesis. Though its goal is similar to ours as it aims to optimize the delivery agents for gene or mRNA therapies using ML, the data modality is very different: an AAV capsid contains three structural Cap proteins while an LNP often consists of lipids and cholesterol. In addition, LNPs as delivery vehicles are superior to viral vectors such as AAVs in the sense that LNPs have lower immunogenicity and are capable of delivering larger payloads (Swingle et al., 2021).

3. Approach

3.1. Problem Formulation

We formulate the LNP design task as a classification task. The input is a multicomponent LNP consisting of four molecules, along with the corresponding ratios in the LNP. The output is a label $y \in (0, 1)$ indicating if the level of transfection efficiency is satisfying.

3.2. Model Design

Our proposed model consists of two parts: (1) molecular representation learning and (2) downstream classification (Figure 1). Given an LNP with four components and their corresponding ratios, we first use molecular representation learning techniques to generate continuous low-dimensional representations to capture high-level information about the chemical structures and encode the ratio information with one-hot embeddings. We then feed the concatenated representations into a downstream classifier to predict the response, i.e. transfection efficiency. The model performance is affected by both (1) the quality of the generated molecular representations in encoding information of the chemical structures; (2) the capability of the classifier in distinguishing satisfying and unsatisfying designs. In this work, we explore molecular representations generated based on an expert-designed system and on GNNs, respectively, and compare classifiers including SVM, RF, XGBoost (Chen & Guestrin, 2016), and multi-layer perceptron (MLP).

3.2.1. Molecular Representation Learning

We propose to use two approaches to generate molecular representations to encode the structural and semantic information of chemical structures. The first approach utilizes chemical knowledge by designing a set of rules through specified functions to extract different properties of the

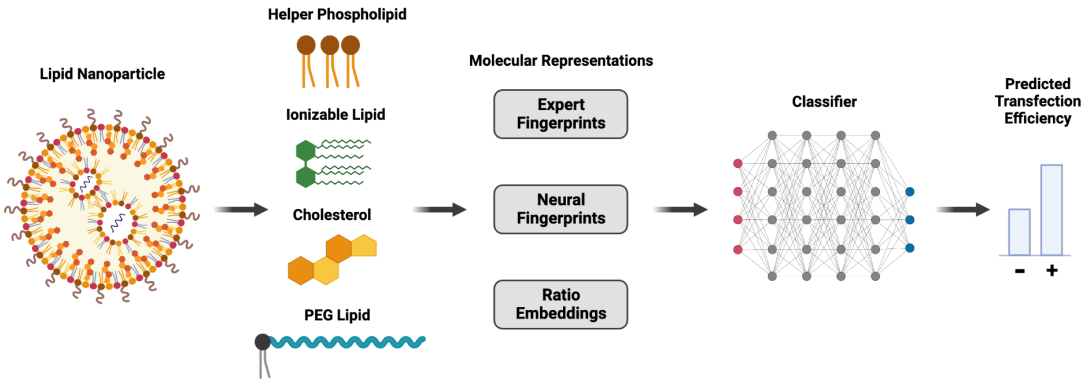

Figure 1. Overview of our approach: (1) Molecular representation learning: we derive "expert fingerprints" and "neural fingerprints" for each of the four molecular components of an LNP, and encode the mixture ratio with one-hot embeddings. (2) Downstream classification: we feed the concatenated molecular representations into classifiers such as SVM or MLP to predict if the transfection efficiency of an LNP design is satisfying. This figure is generated with `BioRender`.

molecules such as atom and bond types. The second approach treats each molecule as a heterogenous graph, where each atom is represented as a node and bond as an edge, and then leverages methods in graph representation learning to generate the representations through graph convolutions. While the first approach using domain expertise is more interpretable, the second approach automates the learning of representations and mitigates the labor involved in handcrafting features. Such representations are usually called molecular fingerprints in the field of chemoinformatics. In this work, we denote the representations generated through the first approach as "expert fingerprints", and through the second approach as "neural fingerprints".

Expert Fingerprint Based on chemical knowledge, one can design a set of specified functions to extract important attributes of molecules that could affect their properties, such as atom types, bond types, chirality, functional groups, etc. These functions are carefully crafted by domain experts and have been made available in chemoinformatics softwares. We use RDKit[1], an open-source cheminformatics software, to generate fingerprints. RDKit provides a number of generation functions, and here we use RDKit2DNormalized, which is one of the most commonly used functions in previous works on molecular property prediction. The generated "expert fingerprints" have 200 dimensions.

Neural Fingerprint Molecules can be viewed as heterogenous graphs consisting of different types of nodes (atoms) and edges (bonds). With graph representation learning, we can learn representations through graph convolutions in an automatic way. Specifically, we use `Grover` to generate neural fingerprints as it has achieved state-of-the-art performance on various molecular prediction tasks (Rong et al., 2020). `Grover` has its architecture based on one of the most expressive GNN variants `GTransformer`. It is pre-trained on a large collection of molecules with self-supervised pre-training, which enables it to learn general representations. We compared `Grover-base` and `Grover-large`, which contain 48M and 100M parameters and output 3400- and 5000-dimension fingerprints, respectively[2].

Component Ratio Embedding The ratio of each molecular component in an LNP can significantly affect its functionality (Kauffman et al., 2015; Hassett et al., 2019). To encode the ratio information, we normalize the ratio of each component to $(0, 1)$ and round it to the nearest number in $\{0.05, 0.10, ..., 1.00\}$. We then use one-hot embeddings to encode the ratio information of each component, each of which has 20 dimensions.

3.2.2. Classification Models

With the generated vector representations of a multicomponent LNP, we concatenate the representations of the different components and their ratios and feed the aggregated representation into a downstream classifier of choice. We experiment with classification models including SVM, RF, XGBoost, and MLP. The classifier learns to differentiate the representations through supervised learning and outputs the likelihood of a molecule being capable of achieving satisfying transfection efficiency.

3.2.3. Evaluation Metric

We use the area under the receiver operating characteristic (AUROC) as our evaluation metric, which measures how well a model ranks examples and distinguishes between classes (i.e. satisfying versus unsatisfying designs of LNPs).

[1]https://www.rdkit.org

[2]We only used `Grover` to generate fingerprints and did not fine-tune its parameters due to the limited amount of training data.

4. Experiments and Discussion

4.1. Data Curation

We curated a dataset of 572 LNPs along with their level of transfection efficiency based on a published paper on membrane-destabilizing ionizable phospholipids for mRNA delivery (Liu et al., 2021), as there were no publicly available data resources that can be readily used to develop ML models for LNP design. We represent the chemical structures of the four components in an LNP with SMILES (Simplified Molecular Input Line Entry System) codes, which provide 2D representations of the atoms and the bonds in molecules. To generate the SMILES codes, the chemical structures were first drawn by domain chemists in a software called `ChemDraw`. We then obtained the readouts on transfection efficiency from the results presented in (Liu et al., 2021), where the readouts are characterized with luciferase expression following treatment of IGROV1 cells with iPhos-delivered mRNA. We used a threshold of 10,000 relative light units to distinguish between LNP designs with satisfying and unsatisfying transfection efficiency, comprising 91 and 481 LNPs, respectively.

The availability of large-scale, high-quality data is often a bottleneck for biomedical applications, including for the optimization of LNP delivery systems. In this study, we address the challenge by curating this new dataset, and by making this dataset publicly available, we aim to contribute to the scientific community and enable more efficient LNP optimization with ML techniques.

4.2. Experiment Setup

We split the dataset into training, validation, and test sets, with 497, 62, 63 data points in each set, respectively (roughly corresponding to a ratio of 8/1/1). For SVM, we experiment with different kernels, including the linear kernel, the radial basis function kernel, and the polynomial kernel, and different regularization strengths. For RF, we experiment with different numbers of trees in the forest and different values for the max depth of each tree. For XGBoost, we experiment with different number of estimators and regularization strengths. For MLP, we use one hidden layer with 128 neurons and train the model with the Adam optimizer. We use the validation set to select the optimal hyperparameters for each model.

4.3. Results

Experimental results of different classifiers trained with different molecular fingerprints (FP) are shown in Table 1. Our methods achieve overall satisfying performances on the LNP transfection efficiency prediction task, with the MLP trained with "expert fingerprints" achieving the highest predictive accuracy of 98% on the test set. Specifically, "expert fingerprints" generated based on domain knowledge are sufficiently predictive by capturing the structural and semantic information of the molecules, while "neural fingerprints" generated with GNN (`Grover`) do not necessarily offer superior performance. This finding is consistent with those presented in Jiang et al. (2021). It may also be caused by the fact that the molecules in LNPs are generally larger than those used to train `Grover`. Since ML often does not perform well on out-of-distribution data, such distribution differences in the size of the molecules could be the reason for GNNs not generating better molecular representations. Overall, our proof-of-concept studies demonstrate the feasibility of using ML to model the chemical structure of multicomponent LNPs in predicting their functionality and show promise in ML-guided LNP design.

Classifier	Expert FP	Neural FP	Val AUC	Test AUC
SVM	✓	✗	0.9092	0.9614
	✗	✓	0.9333	0.9627
	✓	✓	**0.9518**	**0.9654**
	✓	✓(Large)	0.9219	0.9481
Random Forest	✓	✗	0.9326	**0.9521**
	✗	✓	**0.9361**	0.9208
	✓	✓	0.9255	0.9215
	✓	✓(Large)	0.9355	0.9375
XGBoost	✓	✗	**0.9660**	**0.9468**
	✗	✓	0.9489	0.9348
	✓	✓	0.9560	0.9402
	✓	✓(Large)	0.9390	0.9282
MLP	✓	✗	**0.9617**	**0.9815**
	✗	✓	0.9300	0.8892
	✓	✓	0.9383	0.9169
	✓	✓(Large)	0.9533	0.9508

Table 1. Experiment results. Transfection efficiency can be accurately predicted by ML approaches using "expert fingerprints" on our curated dataset.

5. Conclusion

In this work, we proposed to use ML models to predict the transfection efficiency of LNPs for mRNA delivery. We curated a dataset from scratch and empirically evaluated various molecular representation learning techniques and classification models on the task. Our methods achieved overall satisfying performance, demonstrating the feasibility of leveraging ML to model the chemical structures of LNPs in predicting their functionality. Our work shows promise in the direction of ML-guided LNP design, where ML models are used to prioritize LNP candidates for experimental validation and improve screening efficiency. As for future directions, we will leverage the developed ML model to predict and prioritize new LNP designs based on their probability of achieving high transfection efficiency and conduct wet-lab experiments to further validate the effectiveness of our model. We aspire for this prototype to evolve into a powerful tool for assisting researchers in the systematic in silico selection of promising LNP designs.

Reproducibility Statement

We provide all the code and data in https://anonymous.4open.science/r/Lipid_Nanoparticle_Design/. This will help researchers reproduce the results while enabling further exploration of other methods and directions on the curated dataset.

References

Alvarsson, J., Lampa, S., Schaal, W., Andersson, C., Wikberg, J. E., and Spjuth, O. Large-scale ligand-based predictive modelling using support vector machines. *Journal of cheminformatics*, 8 (1):1–9, 2016.

Chakraborty, C., Sharma, A. R., Sharma, G., Doss, C. G. P., and Lee, S.-S. Therapeutic mirna and sirna: moving from bench to clinic as next generation medicine. *Molecular Therapy-Nucleic Acids*, 8:132–143, 2017.

Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., and Blaschke, T. The rise of deep learning in drug discovery. *Drug discovery today*, 23(6):1241–1250, 2018.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Dowdy, S. F. Overcoming cellular barriers for rna therapeutics. *Nature biotechnology*, 35(3):222–229, 2017.

Eygeris, Y., Gupta, M., Kim, J., and Sahay, G. Chemistry of lipid nanoparticles for rna delivery. *Accounts of Chemical Research*, 55(1):2–12, 2021.

Hassett, K. J., Benenato, K. E., Jacquinet, E., Lee, A., Woods, A., Yuzhakov, O., Himansu, S., Deterling, J., Geilich, B. M., Ketova, T., et al. Optimization of lipid nanoparticles for intramuscular administration of mrna vaccines. *Molecular Therapy-Nucleic Acids*, 15:1–11, 2019.

Hou, T., Wang, J., and Li, Y. Adme evaluation in drug discovery: The prediction of human intestinal absorption by a support vector machine. *Journal of chemical information and modeling*, 47(6):2408–2415, 2007.

Jiang, D., Wu, Z., Hsieh, C.-Y., Chen, G., Liao, B., Wang, Z., Shen, C., Cao, D., Wu, J., and Hou, T. Could graph neural networks learn better molecular representation for drug discovery? a comparison study of descriptor-based and graph-based models. *Journal of cheminformatics*, 13(1):1–23, 2021.

Kaczmarek, J. C., Kowalski, P. S., and Anderson, D. G. Advances in the delivery of rna therapeutics: from concept to clinical reality. *Genome medicine*, 9:1–16, 2017.

Kauffman, K. J., Dorkin, J. R., Yang, J. H., Heartlein, M. W., DeRosa, F., Mir, F. F., Fenton, O. S., and Anderson, D. G. Optimization of lipid nanoparticle formulations for mrna delivery in vivo with fractional factorial and definitive screening designs. *Nano letters*, 15(11):7300–7306, 2015.

Liu, S., Cheng, Q., Wei, T., Yu, X., Johnson, L. T., Farbiak, L., and Siegwart, D. J. Membrane-destabilizing ionizable phospholipids for organ-selective mrna delivery and crispr–cas gene editing. *Nature materials*, 20(5):701–710, 2021.

Mayr, A., Klambauer, G., Unterthiner, T., Steijaert, M., Wegner, J. K., Ceulemans, H., Clevert, D.-A., and Hochreiter, S. Large-scale comparison of machine learning methods for drug target prediction on chembl. *Chemical science*, 9(24):5441–5451, 2018.

Ogden, P. J., Kelsic, E. D., Sinai, S., and Church, G. M. Comprehensive aav capsid fitness landscape reveals a viral gene and enables machine-guided design. *Science*, 366(6469):1139–1143, 2019.

Rifaioglu, A. S., Atas, H., Martin, M. J., Cetin-Atalay, R., Atalay, V., and Doğan, T. Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. *Briefings in bioinformatics*, 20(5):1878–1912, 2019.

Rong, Y., Bian, Y., Xu, T., Xie, W., Wei, Y., Huang, W., and Huang, J. Self-supervised graph transformer on large-scale molecular data. *Advances in Neural Information Processing Systems*, 33, 2020.

Sahin, U., Karikó, K., and Türeci, Ö. mrna-based therapeutics—developing a new class of drugs. *Nature reviews Drug discovery*, 13(10):759–780, 2014.

Stanton, M. G. Current status of messenger rna delivery systems. *Nucleic acid therapeutics*, 28(3):158–165, 2018.

Sun, M., Zhao, S., Gilvary, C., Elemento, O., Zhou, J., and Wang, F. Graph convolutional networks for computational drug development and discovery. *Briefings in bioinformatics*, 21(3): 919–935, 2020.

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., and Feuston, B. P. Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of chemical information and computer sciences*, 43(6): 1947–1958, 2003.

Swingle, K. L., Hamilton, A. G., and Mitchell, M. J. Lipid nanoparticle-mediated delivery of mrna therapeutics and vaccines. *Trends in Molecular Medicine*, 27(6):616–617, 2021.

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Xiong, Z., Wang, D., Liu, X., Zhong, F., Wan, X., Li, X., Li, Z., Luo, X., Chen, K., Jiang, H., et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of medicinal chemistry*, 63(16): 8749–8760, 2019.

Zernov, V. V., Balakin, K. V., Ivaschenko, A. A., Savchuk, N. P., and Pletnev, I. V. Drug discovery using support vector machines. the case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *Journal of chemical information and computer sciences*, 43(6):2048–2056, 2003.

Zhang, L., Tan, J., Han, D., and Zhu, H. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug discovery today*, 22(11):1680–1685, 2017.

Zhang, Q.-Y. and Aires-de Sousa, J. Random forest prediction of mutagenicity from empirical physicochemical descriptors. *Journal of chemical information and modeling*, 47(1):1–8, 2007.