Using Machine Learning to Probe the Transfection Efficiency of DNA-Lipid Nanoparticle Formulations

Brendan Lee, Yining Zhu, Ruochen Shen, Hai-Quan Mao

Johns Hopkins University | Department of Materials Science and Engineering | Baltimore, MD

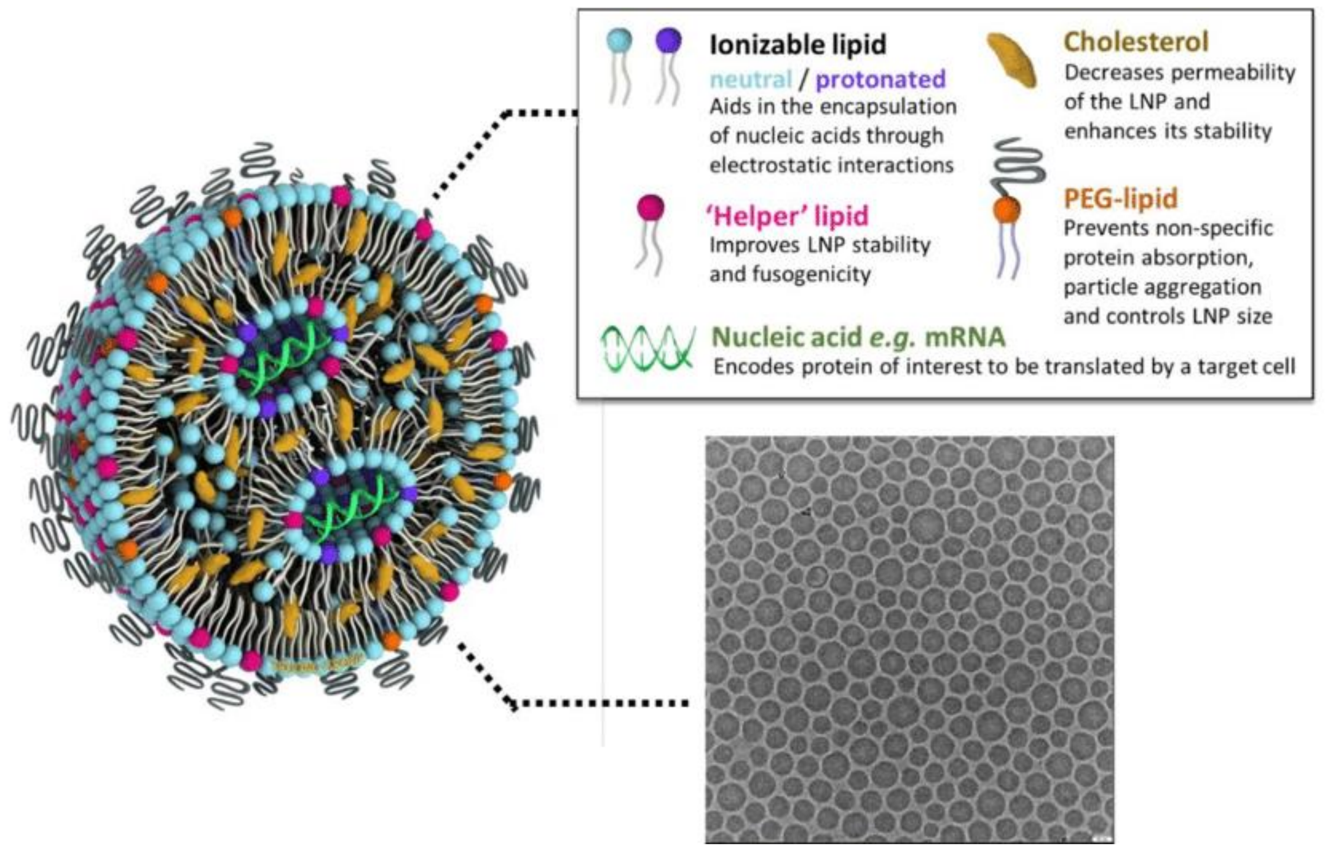
Introduction

Lipid nanoparticles (LNPs) are popular delivery vehicles for a wide range of bioactive agents, particularly nucleic acids, since they enable these molecules to be delivered into cells effectively, by encapsulation, while protecting them from degradation. When optimized, LNPs are biocompatible with diverse capabilities and can be tailored for a variety of therapeutic applications. The steps of LNP delivery consist of the uptake of LNPs by cells followed by endocytosis, endosomal escape, LNP dissociation, and release of payload into the cytosol, and allowing that payload to perform its desired function. The types of lipids used to synthesize LNPs via mixing are known to affect each step of LNP delivery and subsequent transfection efficiency. Using different types and ratios of lipids to make different LNP formulations will result in varying degrees of transfection efficiency, as some formulations yield LNPs with properties that result in better transfection than others. This project aims to analyze and correlate how lipid composition influences transfection efficiency.

Objectives

- Develop a method using machine learning that can reliably predict transfection efficiency from a given DNA-LNP formulation
- Use the results of the machine learning method to correlate how lipid composition affects transfection efficiency

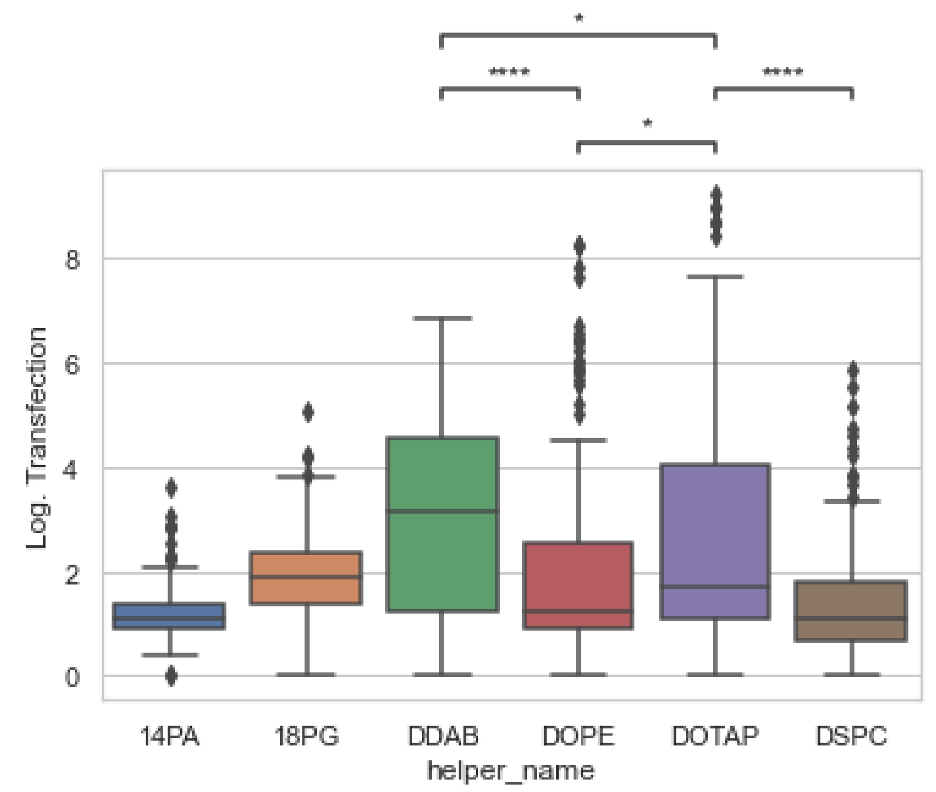
Materials and Methods



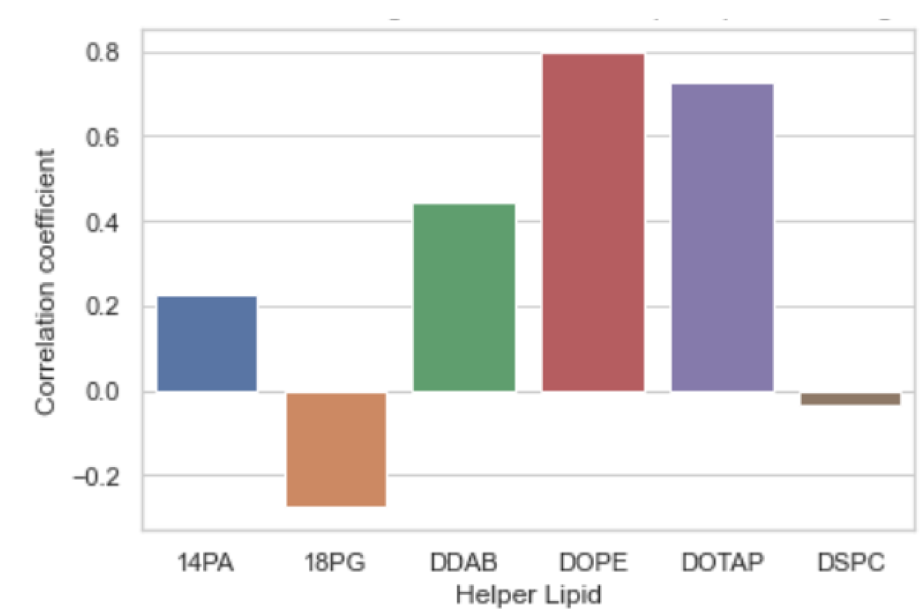
- The first stage of this project involved synthesizing 1000 distinct LNP formulations, each with different ratios of lipid components (shown above), and their transfection data was collected in an in-vitro luciferase assay.
- In the second stage several machine-learning methods were tested, using the ratios of lipid components of each of the 1000 LNP formulations as features and transfection efficiency as the target variable. Such methods include regression and several classification techniques. From this analysis, knowledge on how lipid structure and composition affects DNA-LNP performance in-vitro was obtained.
- Lastly, this analysis was repeated by converting the lipid ratio features to values of lipid properties, such as molecular weight, chain length, and degree of saturation. This way, we are able to predict how certain properties of lipids used in LNPs influence transfection capabilities.

Results

Choice of Helper Lipid Influences Transfection

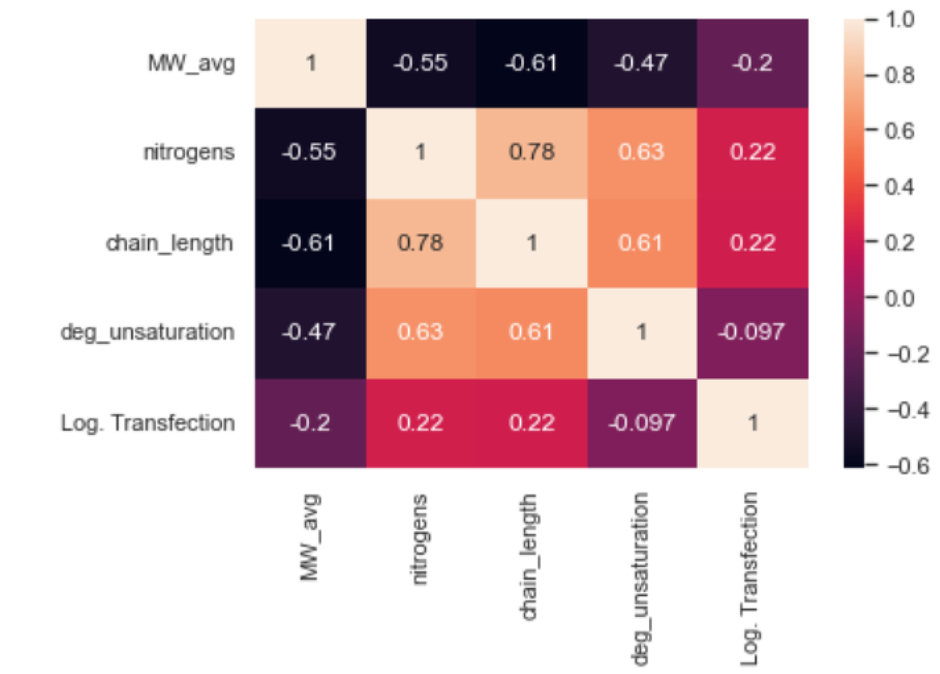


Distribution of Transfection Values By Helper Lipid



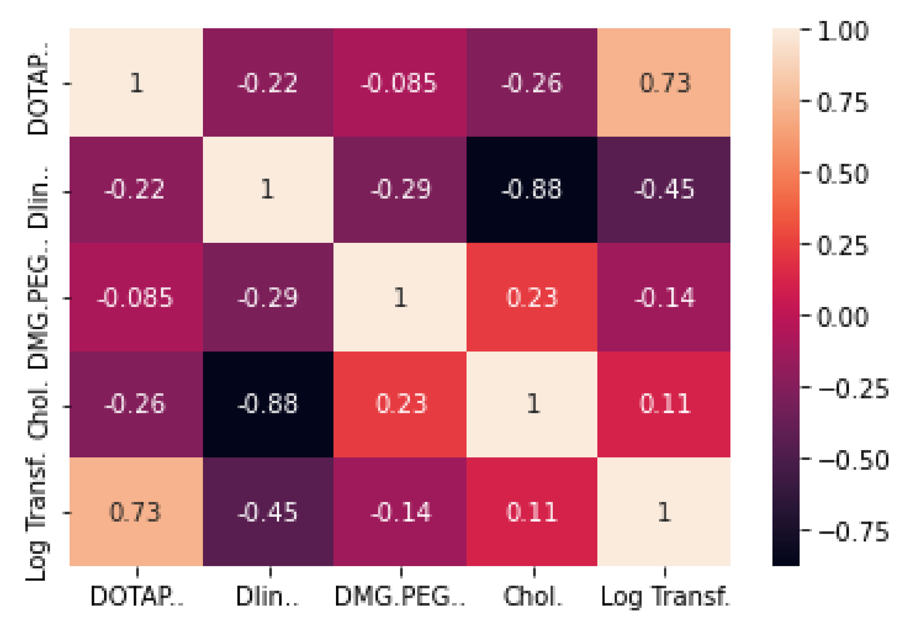
Linear correlation coefficient for Log Transfection vs. Helper Lipid Percentage in Formulation

Analysis of data containing average lipid properties of each formulation

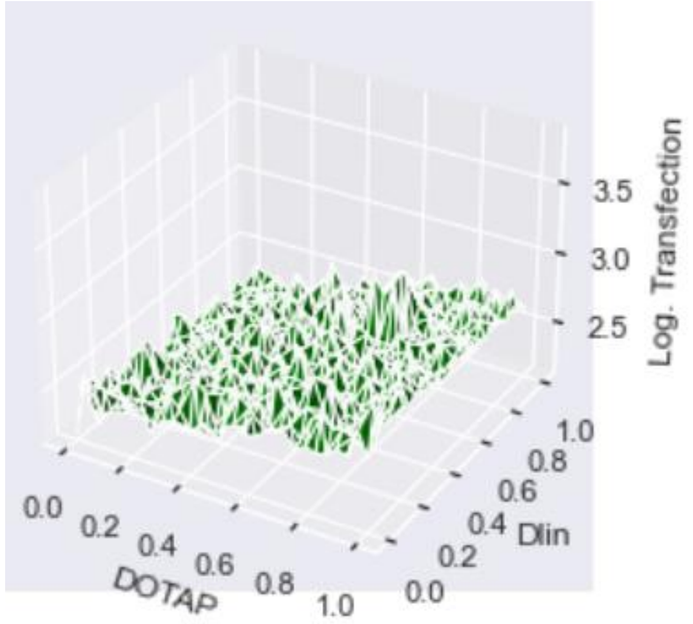


For all 1080 Formulations, the correlation matrix that relates features (average lipid properties) and outcome (transfection).

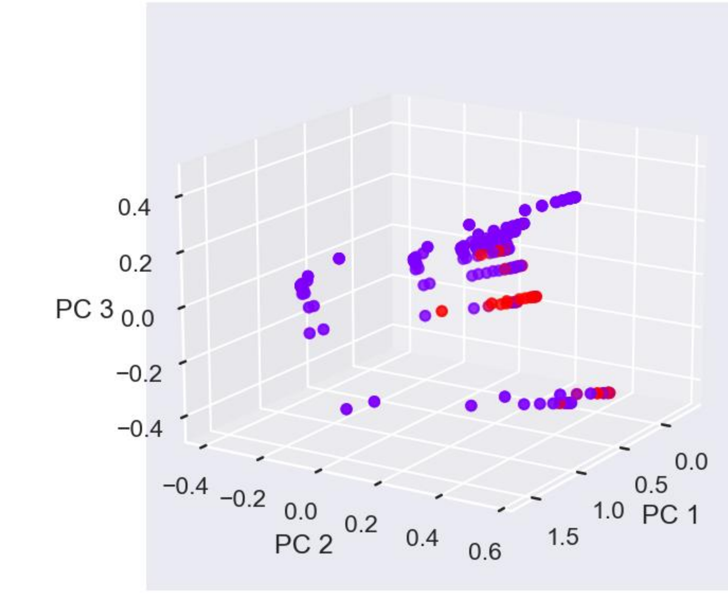
Regression for Predicting the Outcome and Analyzing Associations



For the 180 DOTAP Formulations, the correlation matrix that relates features (lipid component ratios) and outcome (transfection)

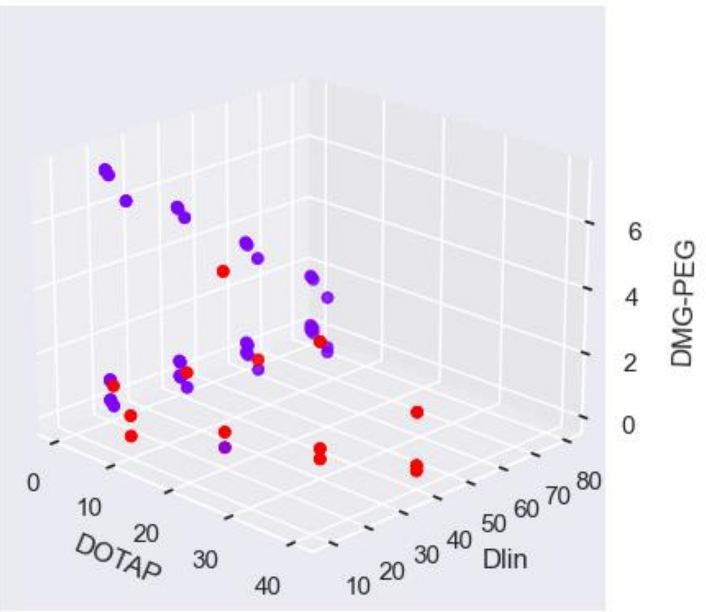


Second-Order Polynomial Regression
Example of a predicted surface is shown above, mean R-Squared value was 0.677.

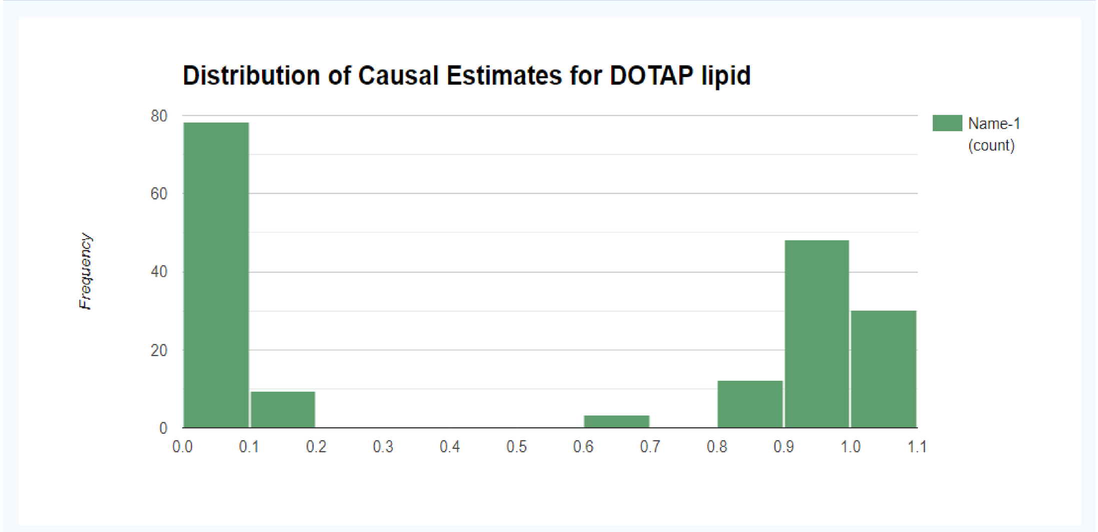


Successful Separation of Data into Classes using PCA
Logistic Regression had highest accuracy among the tested classification techniques (AUC=0.87); predictions are shown above (red=high transfection, purple=low)

Classification for Distinguishing Between High and Low Transfecting Formulations



Successful Separation of Data into Classes
Support Vector Machines (SVM) had highest accuracy among the tested classification techniques (AUC=0.99); predictions are shown above (red=high transfection, purple=low)



Suggests that high DOTAP composition causes high transfection efficiency, but low DOTAP content does not cause low transfection, since formulations are separated into two groups: high causality + high DOTAP content (right cluster) and low causality + low DOTAP content (left cluster).

Future Directions

- Collect data across multiple cell lines and add cell line, and certain details and/or properties of each cell line, as a feature in our models
- Collect and analyze data for in vivo experiments
- Include outcome metrics other than transfection efficiency, such as organ accumulation and biodistribution
- Repeat experiment by changing outcomes to see how lipid composition affects overall LNP physical properties

Conclusions

- By using machine learning methods, transfection efficiency is successfully correlated to LNP compositions.
- The highest-transfecting LNP formulations were the ones with high DOTAP content.
- A combination of low molecular weight, long chain length, and nitrogen-containing lipids contribute to the high transfection efficiency.

Acknowledgements

We thank the Mao Research Group and the entire Institute for NanoBioTechnology for providing us with the resources needed to complete this project and answering our questions. We thank Dr. Orla Wilson for guiding us in every step of the project and giving excellent feedback for our work.