SCIENTIFIC CULTURE
www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 12, No. 4, (2026), pp. 11071-11086
Open Access. Online & Print

DOI: 10.5281/zenodo.124261060

MACHINE LEARNING MODELS FOR PREDICTING DRUG DELIVERY EFFICIENCY IN PERSONALIZED MEDICINE

Dr. Asma Rani[1]*, Dr. Nageswari Ponnusamy Thangarasu[2], Dr. Amandeep Kaur[3], Shweta Tiwari[4], Dr. Dinesh Senapati[5], Priyanka K[6], Ranjan Banerjee[7]

[1]Assistant Professor, Department of Computer Science and Engineering, Dr. B. R. Ambedkar Institute of Technology, Sri Vijayapuram, Andaman and Nicobar Islands-744103 ORCID ID: 0000-0001-6087-6248 Email ID: asma.sags@gmail.com
[2]Senior Resident Department of Pharmacology, Specialization in MBBS MD Pharmacology Sri Venkateswaraa Medical College Hospital and Research Institute, Sri Venkateswaraa University, Redhills, Chennai-600067
ORCID ID: 0009-0005-2304-9266 Email ID: nageswari1896@gmail.com
[3]Assistant Professor Department of Computer Science & Engineering, I.K. Gujral Punjab Technical University, Jalandhar, Punjab, India. ORCID ID: 0009-0009-8727-0161 Email ID: akdhaliwal361@gmail.com
[4]Assistant Professor Specialisation in NLP, AI, ML, DL in computer science and exploring fields/areas Department of Information Technology Rajkiya Engineering College Ambedkar Nagar, Akbarpur-224122 ORCID ID: https://orcid.org/0009-0005-4015-8421 Email ID: shwetatiwari08@recabn.ac.in
[5]Postgraduate, Department of Public Health Dentistry, Specialization in Public Health Dentistry, Kalinga Institute of Dental Sciences (KIDS), Kalinga Institute of Industrial Technology (KIIT) Deemed to be University, Bhubaneswar -751024, Odisha, India, ORCID ID : 0009-0005-4732-3039 Email ID: dineshsenapati.07@gmail.com
[6]Nursing Tutor, Dr. M.G.R. Educational and Research Institute, Maduravoyal, Chennai - 600095, Tamil Nadu, India.
ORCID ID:0009-0001-5933-2523 Email ID: priyanka.physioaoc@drmgrdu.ac.in
[7]Assistant Professor, Department of Computer Science & Engineering Brainware University, Barasat, Kolkata – 125, West Bengal, India-700125 ORCID ID: 0009-0003-1950-7530 Email ID: rbkpccst@gmail.com

Corresponding Author: Dr. Asma Rani
(asma.sags@gmail.com)

ABSTRACT

Lipid nanoparticles (LNPs) are critical delivery systems for nucleic-acid therapeutics, but their formulation optimization remains experimentally intensive because delivery efficiency depends on complex molecular and formulation-level interactions. This study developed machine learning models to predict LNP delivery efficiency using molecular representations of ionizable lipids and formulation-level information. A curated LNP transfection-efficiency dataset containing 572 formulations was analyzed. The primary endpoint was a four-class delivery-efficiency label (y1), while a secondary binary endpoint (y2) distinguished high-efficiency formulations from low/moderate-efficiency formulations. Molecular representations, including circular fingerprints, graph convolutional features, expert descriptors, and learned molecular embeddings, were evaluated with multiple supervised classifiers. Five-fold stratified cross-validation and an independent 80:20

train-test split was used for model assessment. In multiclass prediction, Histogram Gradient Boosting achieved the strongest independent test-set performance, with 0.6957 accuracy, 0.6009 balanced accuracy, 0.6076 macro-F1, and 0.6952 weighted-F1. In binary high-efficiency prediction, Logistic Regression achieved the best F1-score of 0.7692, with 0.9217 accuracy, 0.8857 balanced accuracy, 0.8333 recall, and 0.9450 ROC-AUC. Learned molecular embeddings, particularly MMB-FT features, provided stronger predictive performance than conventional representations. Permutation importance analysis indicated that predictive information was distributed across multiple latent molecular dimensions. These findings demonstrate that machine learning can support early-stage LNP screening and contribute to data-driven personalized nucleic-acid delivery design.

KEYWORDS: Lipid nanoparticles; Machine learning; Drug delivery efficiency; Personalized medicine; Transfection efficiency.

1. Introduction

Lipid nanoparticles (LNPs) are one of the most significant non-viral delivery vehicles for nucleic-acid-based drugs, such as mRNA, siRNA and gene-regulatory molecules. LNP performance has been predicted using machine learning models trained on molecular representations and labels for transfection efficiency [1]. Machine-learning models with representation learning, including molecular embeddings derived from large language models, have opened opportunities for predicting delivery efficiency from lipid structures, rather than relying on empirical trial and error. Likewise, machine learning-based LNP design has been suggested as an approach to accelerate the discovery of mRNA delivery systems by connecting LNP structure, composition and function [2]. Open-access resources like LipoBART also help ensure model reproducibility by offering standardised data, molecular representations and code for LNP transfection-efficiency prediction [3].

LNPs are important for biomedical applications due to the rapid growth of mRNA-based therapies and precision medicine. LNPs shield nucleic acids against degradation, enhance cellular internalization, and enhance endosomal escape, and are thus key components of RNA delivery platforms [4]. The mRNA vaccines have shown that nucleic-acid drugs can be quickly developed and administered with the help of delivery systems [5]. Over time, LNP technology has progressed from traditional liposomes to sophisticated nanocarriers with variable ionizable lipids, helper lipids, cholesterol and polyethylene glycol-lipids [6]. Such developments have seen LNPs emerge as delivery technologies for gene therapies, in which active delivery into cells is a critical translational need [7].

Increasing clinical translation has also bolstered the importance of LNP delivery. The success and development of nucleic-acid nanomedicines, including clinically translated siRNA formulations, demonstrate LNPs can transition from research platforms to therapeutic formulations [8]. LNP technology has been critical for in vivo delivery to the liver for gene regulation because liver delivery takes advantage of nanoparticle biodistribution and uptake mechanisms [9]. But mRNA delivery is complicated because delivery efficiency is dependent on a complex interplay between lipid chemistry, particle structure, formulation, biological milieu and intracellular processes [10]. Recent experiments also indicated that LNP deconvolution is necessary to understand the role of formulation components on efficient mRNA delivery [11].

Although significant advances have been made, there are challenges associated with formulation. LNP stability, storage, structure and biological performance are interrelated, particularly for mRNA-LNP vaccines [12]. More generally, the translation of nanoparticles is still challenging due to biological variability, safety, reproducibility and scalability [13]. In cancer nanomedicine, successful delivery and reliable therapeutic outcomes do not always translate [14]. Complex lipid and dendrimer-based nanoparticles have shown promise for therapeutic applications in disease models, such as protein replacement using mRNA [15]. Lipid-like nanoparticles have also been used to deliver nucleic acids in vivo, suggesting the need for systematic approaches to predict the best formulations [16].

Machine learning provides a means to tackle this design challenge. Machine learning has been applied for prediction, prioritization and decision making in chemically and biologically complex data in drug discovery and development [17]. Previous research on machine-learning approaches in drug discovery also demonstrated the utility of computational models for predicting drug properties, virtual screening, and knowledge-driven drug design [18]. In terms of molecular representation, extended-connectivity fingerprints continue to be a critical method for converting chemical structures to numerical representations [19]. Newer graph-based models enable molecular features to be learned from chemical graph representations, enabling more flexible representations [20]. The development of transformer-based pretrained molecular models, such as Chemformer, also showcase the usefulness of learned representations for molecular modeling in computational chemistry [21].

Hence, this work explores and compares machine learning models with molecular representations of ionizable lipids and formulation-level descriptors for LNP delivery prediction. The work explores multiclass and binary prediction of transfection efficiency, compares several supervised machine learning models, assesses molecular representations, and employs model interpretability tools to discover structure-activity relationships for LNP delivery. This study provides a reproducible computational approach for initial LNP screening and contributes to the goal of data-driven LNP design for personalized nucleic-acid delivery.

2. Literature Review

2.1 Lipid Nanoparticles as Platforms for Nucleic-Acid Delivery

Lipid nanoparticles (LNPs) are one of the most

successful non-viral delivery systems for the delivery of nucleic-acid-based therapies. Their importance has grown significantly with the advent of mRNA vaccines, siRNA drugs and gene-regulation products. LNPs offer a safe nanocarrier environment to protect vulnerable RNA molecules, prevent degradation by enzymes, enhance cell uptake and promote endosomal escape to release therapeutic nucleic acids into the cytoplasm of cells [4]. The broader success of mRNA vaccines has also shown that nucleic-acid therapies can be swiftly developed with an effective delivery system [5]. Previous liposomal formulations laid the groundwork for lipid-based nanomedicines but current LNPs include ionizable lipids, helper phospholipids, cholesterol and PEG-lipids to control stability, delivery and transfection [6]. This versatility makes LNPs a promising delivery vehicle for personalized medicine, where delivery vehicles may need to be tailored for specific therapeutic cargo, target organ, disease state, and individual patient treatment needs.

2.2 Clinical Translation and Therapeutic Relevance of LNPs

LNPs are critical for therapeutic success because they can deliver genetic medicines without viruses. Cullis and Hope concluded that LNP systems have enabled significant advances in gene therapies because they can encapsulate nucleic acids and facilitate their release [7]. The development of patisiran, discussed in the Onpattro case study, demonstrated that LNP-based nucleic-acid nanomedicines can be successfully translated into the clinic if their delivery efficiency, safety and quality are optimised [8]. LNPs have been particularly important for the delivery of nucleic acids to the liver, as liver uptake pathways promote the uptake of nucleic-acid-loaded nanoparticles into the liver [9]. But the same complexity of biological events that facilitates tissue targeting also confounds the experimental prediction of LNP performance, especially when changes in lipid structure and composition, as well as biological conditions, have nonlinear effects on transfection efficiency.

2.3 Formulation and Biological Challenges in LNP Delivery

One of the main challenges in LNP formulation is that delivery is regulated by a complex interplay of formulation and molecular characteristics. Kauffman, Webber and Anderson observed that non-viral mRNA delivery requires materials that can overcome stability challenges in the extracellular environment, cell membrane entry, and lysosomal (endosomal) entrapment [10]. Eygeris et al. also demonstrated that the structure of LNP needs to be deconvoluted to determine how different LNP formulation components affect mRNA delivery efficiency [11]. Stability is also another critical consideration for mRNA-LNP vaccines as particle integrity, storage conditions, and formulation components can impact biological activity and translational potential [12]. The problems are not unique to vaccines. There are still challenges with reproducibility, biodistribution, safety, scale-up and efficacy for nanoparticle delivery systems in general [13]. For instance, in cancer nanomedicine, there is still a significant gap between successful delivery in pre-clinical studies and clinical efficacy [14]. These results suggest that experimental formulation screening may be suboptimal, and that computational methods are required to prioritise formulations for screening.

2.4 Experimental Evidence for Advanced Lipid and Lipid-Like Nanoparticles

Several experimental studies have shown the clinical potential of new lipid and lipid-like nanoparticles. For example, Cheng et al. demonstrated the use of dendrimer-based lipid nanoparticles to successfully deliver therapeutic mRNA of the enzyme FAH and increase survival in a disease model, demonstrating that rationally designed lipid systems can mediate restoration of clinically relevant protein functions [15]. Lipid-like nanoparticles have also been studied for nucleic acid delivery in vivo, further supporting the notion that chemical design of lipid materials is important for delivery [16]. However, these studies also demonstrate a key problem: while high-performance formulations can be found experimentally, the structure-performance relationships are hard to predict. This calls for machine learning tools that can learn from the structure of molecules, families of lipids, and transfection efficiencies.

2.5 Machine Learning in Drug Discovery and Formulation Prediction

Machine learning is increasingly being used in drug discovery, formulation, and nanomedicine as it can capture complex nonlinear relationships in multidimensional chemical and drug datasets. Vamathevan et al. highlighted the increasing use of machine learning in drug discovery and development for prediction, prioritization and decision-making in chemical complex data [17]. Lavecchia also highlighted that machine-learning methods can be used to aid virtual screening,

molecular property prediction, and drug design [18]. This is particularly useful for LNPs because their performance is dependent on several variables rather than a single physicochemical attribute. This means machine learning can help save time and money by prioritising formulations for experimental study.

2.6 Molecular Representation Methods for Predictive Modeling

One challenge in using machine learning for molecular delivery is transforming chemical structures into feature representations. One of the most popular molecular representations is extended-connectivity fingerprints, which describe local atomic environments in fixed-size vectors that can be used for predictive modeling [19]. But fingerprints may fail to capture higher-order molecular interactions. Graph neural networks overcome this shortcoming by encoding molecular fingerprints from graph representations of atoms and bonds as nodes and edges, respectively [20]. More recently, transformer-based chemical language models (like Chemformer) have demonstrated that pretrained molecular representations can learn useful chemical features from SMILES strings and be used for downstream molecular prediction [21]. These representation learning methods are very important for LNP design because the structures of ionizable lipids are diverse and potentially contain latent features that can determine transfection efficiency.

2.7 Machine Learning for LNP Transfection-Efficiency Prediction

Recent machine learning-based studies specific to LNP have tackled this challenge. Moayedpour et al. proposed LipoBART, showing that embeddings of lipid nanoparticles from large language models can be used to predict transfection efficiency from lipid structures [1]. Ding et al. also reported machine learning-driven LNP design for mRNA delivery, demonstrating that machine learning models can aid in the discovery of optimal LNP formulations [2]. The release of the LipoBART public repository also enhances reproducibility by offering datasets and tools for LNP design [3]. Overall, these studies show that machine learning can transform LNP development from trial and error to predictive design.

2.8 Research Gap and Study Rationale

While this is promising, there is still more work to be done. Many previous works focus either on experimental LNP design, or on generic molecular machine learning, with fewer studies offering a holistic framework that includes the comparison of several molecular representations, the evaluation of multiple supervised machine learning classifiers, multiclass and binary prediction of LNP delivery efficiency, and interpretability analysis. And the importance of identifying formulations with high efficiency is especially relevant to personalized medicine, where nucleic-acid therapies may need to be rapidly adapted. Thus, this study aims to fill this gap by implementing and comparing machine learning models of LNP delivery efficiency prediction based on molecular embeddings, fingerprints and formulation descriptors. This study offers a reproducible data-driven LNP formulation screening workflow for personalized nucleic-acid delivery, which combines multiclass transfection-efficiency prediction, binary prediction of high efficiency, representation comparison, and explainability analysis via permutation.

3. Methodology

3.1 Research Design

The study used a quantitative, computational, and predictive research design to build machine learning models for predicting the efficiency of lipid nanoparticle (LNP) delivery. The study was based on a supervised learning problem, where molecular and formulation descriptors were applied to predict experimentally determined classes of transfection efficiency. The main objective of the methodology was to compare several supervised classifiers under identical data preprocessing, model validation, and model evaluation procedures, to identify machine learning models that can consistently classify low-, moderate-, and high-efficiency LNP formulations. This involved data collection, data cleaning, molecular descriptor representation, model-building, cross-validation, independent test-set evaluation, and model interpretability.

3.2 Data Source and Data Collection

The present study collected secondary experimental data from an ionizable phospholipid LNP dataset provided in CSV, TXT, and JSON formats. The main modeling file, iphos_multiclass.csv, contained 572 LNP formulation records with formulation identifiers, molecular-family labels, molar composition variables, molecular structures, and delivery-efficiency labels. The variables included name, family, y1, y2, p1, p2, p3, p4, and four molecular components represented as SMILES strings: m1, m2, m3, and m4. The m1 variable represented the variable ionizable lipid component

and was therefore selected as the principal molecular input, because ionizable lipids play a central role in LNP-mediated nucleic-acid delivery. The composition variables p1, p2, p3, and p4 represented the relative molar proportions of the LNP components.

The primary outcome variable was y1, a multiclass delivery-efficiency label derived from transfection performance. The secondary outcome variable was y2, a binary label used to distinguish high-efficiency formulations from lower-efficiency formulations. Supporting files were used to verify and enrich the modeling dataset. The full_iPhos_lipids.csv file provided structural information for ionizable phospholipid components, iphos_targets.csv contained the target-label matrix, and iphos_smiles.txt listed molecular SMILES strings. Precomputed molecular representation files in JSON format mapped SMILES strings to numerical feature vectors, including circular fingerprints, graph convolutional representations, expert descriptors, and molecular embedding vectors. These representation files enabled direct transformation of chemical structures into machine-learning-compatible inputs.

3.3 Population and Sampling

The population of interest was LNP formulations with ionizable lipids for nucleic-acid delivery. The researcher accessed a curated, existing data set and did not conduct any additional experimental, animal, clinical or human-subject sampling. The analytical sample was all formulations with complete target labels and molecular representations for the chosen ionizable lipid SMILES. A complete-case approach was adopted after examining the integrity of the files, missingness, target consistency, class balance, and coverage of molecular representations. Stratified sampling was applied when splitting data into train-test and cross-validation sets to ensure the same class distribution of delivery efficiency in the model-development and -evaluation data.

3.4 Data Preprocessing and Feature Engineering

The dataset was first examined for structure, variable types, missing values, duplicate records, class imbalance, and consistency between iphos_multiclass.csv and iphos_targets.csv. The multiclass label y1 was used for the primary classification task, while y2 was retained for a secondary binary high-efficiency prediction task. The m1 SMILES string was matched with the selected SMILES-to-vector mappings to generate molecular feature matrices. Several molecular representations were assessed, including circular fingerprints, graph convolutional features, expert-engineered descriptors, and pretrained molecular embeddings. The principal model was developed using molecular embedding features because these representations capture distributed structural information that may not be fully represented by sparse binary fingerprints.

The final predictor matrix was constructed by combining the molecular representation of m1 with the molecular-family variable. The family feature was retained because it provides chemically meaningful grouping information that may influence delivery-efficiency patterns. Linear and kernel-based classifiers were implemented within standardization pipelines to ensure scale-consistent optimization, whereas tree-based models were trained without feature scaling because they are not dependent on feature magnitude. Target variables and any directly target-derived information were excluded from the predictor matrix to minimize information leakage.

3.5 Machine Learning Model Development

The main analysis was conducted as a multiclass classification problem to predict y1. A binary classification task was also performed to predict y2, allowing assessment of the model's capacity to predict efficient LNP formulations. The data was split into training and testing sets in an 80:20 ratio with a stratified split and a set random seed for reproducibility. Moreover, five-fold stratified cross-validation was performed to assess the model's performance on different data partitions.

A range of supervised learning models were employed to offer both simple and sophisticated model comparisons. The Dummy Classifier was implemented as a naïve majority-class model. A linear Logistic Regression model was used. We used the Linear Support Vector Machine and radial basis function Support Vector Machine to test margin-based classification. Random Forest and Extra Trees were used as ensemble tree methods able to model complex feature interactions. Histogram Gradient Boosting was used as a boosting algorithm for tabular data. Imbalanced class distributions were handled by stratified folds and balanced class weights.

3.6 Model Evaluation and Data Analysis

Accuracy, balanced accuracy, macro-averaged F1-score, weighted F1-score, precision and recall were used assessing model performance. Macro-F1 was

highlighted as the main comparative measure as it treats all classes equally and avoids the bias towards the majority class that is present in the standard accuracy measure. We also reported balanced accuracy to consider class imbalance. For the binary classification experiment, receiver operating characteristic area under the curve was also reported for models that estimated probabilities.

Confusion matrices were also produced for the best-performing multiclass and binary classification models to assess class-wise misclassifications. Performance tables were generated for cross-validation and independent test results. We also performed model-representation comparisons to determine whether different molecular feature encodings affected classification performance. For interpretability, we used permutation importance for the best-performing model. This involved calculating the decrease in macro-F1 score following random shuffling of each individual feature, thus determining the most important dimensions of the molecular representation and family-based descriptors contributing to model performance.

3.7 Ethical Considerations

The data used in this study were publicly available, secondary, non-identifiable formulation-level data and the researcher did not recruit human participants, access patient medical records, collect human biological samples, or conduct animal experiments. As a result, human-subjects ethical review was not necessary. But the researcher adhered to principles of responsible computational research by preserving the integrity of the data set, refraining from fabricating data, ensuring reproducibility by setting fixed random seeds and storing model artifacts, and interpreting the model outputs within the scope of the data set. Given that the models produce formulation-level rather than patient-specific predictions, the results should be interpreted as evidence for decision support, early screening and hypothesis generation of LNPs. Experimental validation is still required prior to translational, regulatory or clinical use.

4. Results

4.1 Dataset Characteristics and Endpoint Distribution

The final analytical dataset consisted of **572 lipid nanoparticle (LNP) formulations**, each containing formulation identifiers, molecular-family labels, lipid composition variables, SMILES-based molecular structures, and delivery-efficiency labels. No missing values were detected in the core modeling variables, indicating that the dataset was suitable for complete-case supervised learning. The primary endpoint, y1, represented a four-class transfection-efficiency outcome, while the secondary endpoint, y2, represented a binary high-efficiency classification outcome. The multiclass endpoint was imbalanced, with class counts of **189, 294, 76,** and **13** for classes 0, 1, 2, and 3, respectively. This imbalance was methodologically important because it justified the use of stratified sampling, balanced class-weight settings where applicable, and macro-averaged evaluation metrics.

Table 1. Dataset and model summary

Item	Value
Dataset	LipoBART / iPhos lipid nanoparticle transfection-efficiency dataset
Number of formulations	572
Primary target	y1 multiclass delivery-efficiency class
Secondary target	y2 binary high-efficiency class
Main molecular input	m1 ionizable-lipid SMILES
Best multiclass model	Hist_Gradient_Boosting
Best binary model	Logistic_Regression
Best multiclass test macro-F1	0.607621978
Best multiclass test accuracy	0.695652174
Best binary test F1	0.769230769
Best binary test accuracy	0.92173913

This table summarizes the dataset size, modeling targets, molecular input, best-performing models, and key performance metrics.

The class distribution of the primary multiclass endpoint is presented in **Figure 1.**

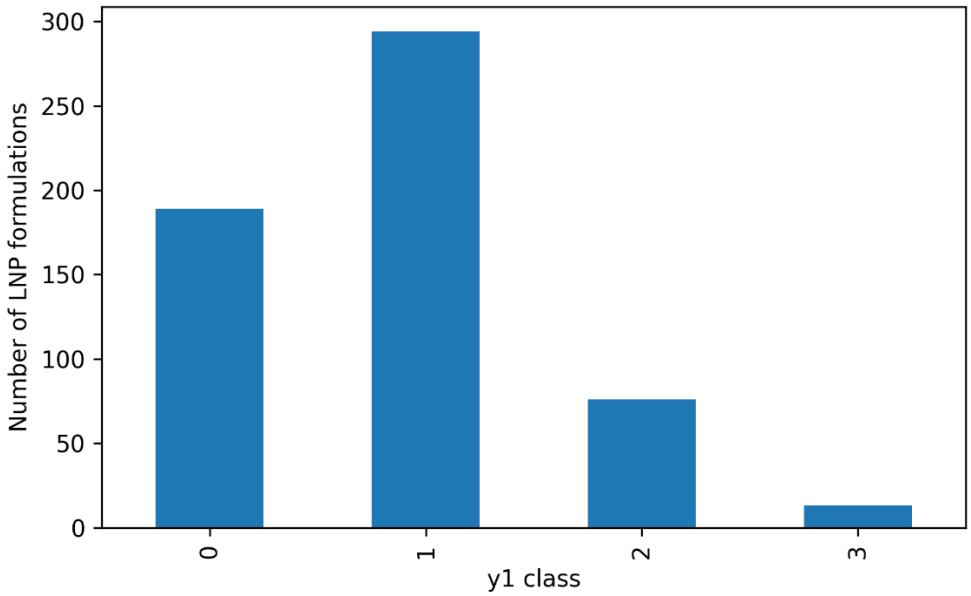

Figure 1. Distribution of multiclass delivery-efficiency labels

This figure displays the distribution of the four delivery-efficiency classes and highlights the imbalance in the highest-efficiency class.
The binary endpoint grouped classes 0 and 1 as low/moderate efficiency and classes 2 and 3 as high efficiency. This produced **483 low/moderate-efficiency** and **89 high-efficiency** formulations, allowing the model to be evaluated for its ability to identify promising LNP candidates. The binary endpoint distribution is shown in **Figure 2.**

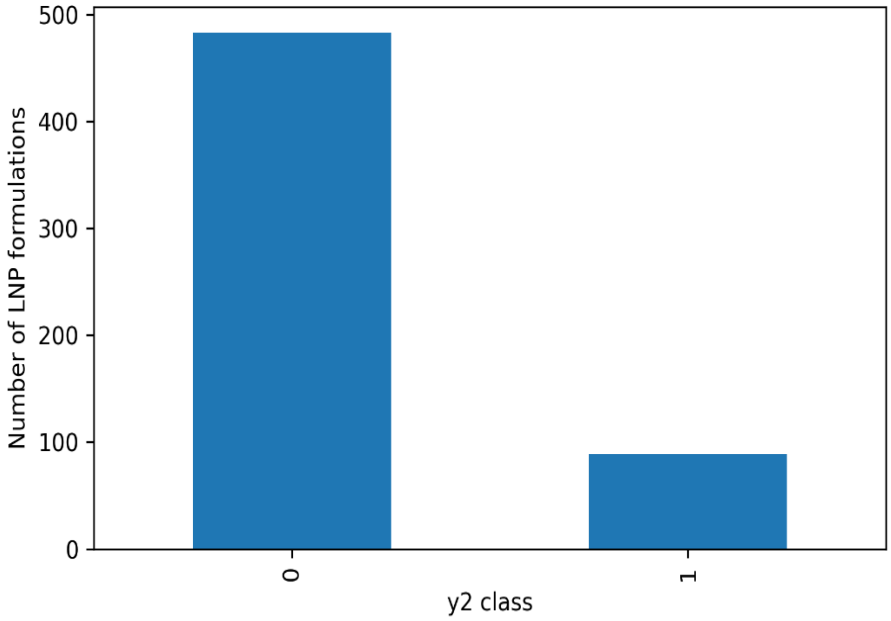

Figure 2. Distribution of binary high-efficiency labels

This figure shows the class balance for the binary high-efficiency prediction task.
The molecular family variable was unevenly distributed across seven families, with family-specific counts ranging from **13 to 130**. This heterogeneity suggested that chemical family membership may contain useful structure-performance information and supported its inclusion as an auxiliary feature. The family distribution is presented in **Figure 3.**

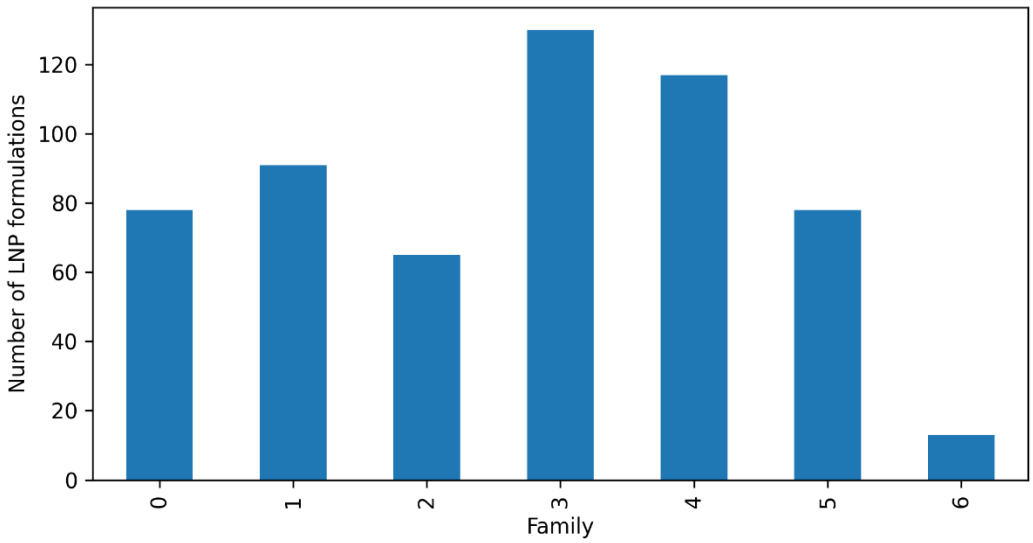

Figure 3. Distribution of molecular families

This figure illustrates the representation of different molecular families in the analytical dataset.

4.2 Molecular Representation Coverage and Feature Construction

The principal molecular representation files showed complete coverage for the m1 ionizable lipid SMILES. Specifically, CFP_2048, GCN_32, Expert_200, MMB_FT_512, and MMB_512 matched **572 of 572 formulations**, corresponding to **100% exact SMILES coverage**. In contrast, the all-data MegaMB and CFP files showed no exact matches with the m1 strings, most likely because of differences in SMILES canonicalization. Therefore, only representations with complete exact coverage were retained for formal model comparison. The molecular representation coverage is reported in **Table 2.**

Table 2. Molecular representation coverage

representation	file	json_molecules	vector_dimension	exact_m1_matches	coverage_percent
CFP_2048	mol2fp_CFP.json	575	2048	572	100
GCN_32	mol2fp_gcn.json	572	32	572	100
Expert_200	mol2fp_Expert.json	575	200	572	100
MMB_FT_512	mol2fp_MMB-FT.json	575	512	572	100
MMB_512	mol2fp_MMB.json	575	512	572	100
CFP_all_data_2048	mol2fp_cfp_all_data.json	604	2048	0	0
MegaMB_base_all_data_512	mol2fp_MegaMB_base_all_data.json	591	512	0	0
MegaMB_finetuned_all_data_512	mol2fp_MegaMB_finetuned_all_data.json	591	512	0	0

This table reports the number of molecules, vector dimensions, exact SMILES matches, and coverage percentage for each molecular representation file.

The main feature matrix was constructed using the **MMB-FT molecular embedding** with the family feature appended, yielding **572 observations and 513 predictors**. The stratified 80:20 split generated **457 training samples** and **115 independent test samples**, while preserving the original class structure.

4.3 Cross-Validation Performance

Five-fold stratified cross-validation was conducted to estimate model robustness across alternative data

partitions. The Histogram Gradient Boosting classifier achieved the strongest cross-validation performance, with mean accuracy of **0.6765**, balanced accuracy of **0.6426**, macro-F1 of **0.6467**, and weighted-F1 of **0.6734**. Linear SVM also performed competitively, with macro-F1 of **0.6295**. In contrast, the Dummy Classifier achieved a macro-F1 of only **0.1697**, confirming that the supervised models learned meaningful structure-performance patterns beyond majority-class prediction. The full cross-validation comparison is presented in **Table 3**.

Table 3. Five-fold cross-validation performance of multiclass models

model	accuracy_mean	accuracy_std	balanced_accuracy_mean	balanced_accuracy_std	f1_macro_mean	f1_macro_std	f1_weighted_mean	f1_weighted_std
Hist_Gradient_Boosting	0.677	0.037	0.643	0.047	0.647	0.040	0.673	0.036
Linear_SVM	0.656	0.030	0.641	0.036	0.630	0.029	0.656	0.028
Extra_Trees	0.661	0.041	0.593	0.065	0.599	0.072	0.653	0.040
Logistic_Regression	0.645	0.031	0.629	0.040	0.591	0.029	0.646	0.030
Random_Forest	0.671	0.028	0.569	0.064	0.578	0.041	0.661	0.026
RBF_SVM	0.600	0.057	0.643	0.070	0.574	0.076	0.602	0.055
Dummy_Most_Frequent	0.514	0.003	0.25	0	0.170	0.001	0.349	0.004

This table compares cross-validation accuracy, balanced accuracy, macro-F1, and weighted-F1 across all multiclass models.

4.4 Independent Test-Set Performance

Independent test-set evaluation confirmed the superiority of Histogram Gradient Boosting, which achieved **0.6957** accuracy, **0.6009** balanced accuracy, **0.6076** macro-F1, and **0.6952** weighted-F1. Random Forest achieved the same overall accuracy of **0.6957**, but its lower macro-F1 score of **0.5786** indicated weaker class-balanced performance. The large gap between the Dummy Classifier and the trained models demonstrated that the molecular embeddings contained predictive information relevant to LNP delivery efficiency. The independent test-set results are reported in **Table 4.**

Table 4. Independent test-set performance of multiclass models

model	accuracy	balanced_accuracy	precision_macro	recall_macro	f1_macro	f1_weighted
Hist_Gradient_Boosting	0.696	0.601	0.626	0.601	0.608	0.695
Random_Forest	0.696	0.574	0.588	0.574	0.579	0.694
Linear_SVM	0.643	0.570	0.583	0.570	0.568	0.644
Logistic_Regression	0.652	0.602	0.556	0.602	0.566	0.655
Extra_Trees	0.652	0.555	0.544	0.555	0.547	0.656
RBF_SVM	0.565	0.554	0.490	0.554	0.491	0.574
Dummy_Most_Frequent	0.513	0.25	0.12826087	0.25	0.16954023	0.347926037

This table reports independent test-set performance for each classifier using accuracy, balanced accuracy, precision, recall, macro-F1, and weighted-F1.

The best model achieved strongest class-wise performance for classes 0 and 1, with F1-scores of **0.6849** and **0.7395**, respectively. Class 2 was predicted with moderate reliability, with an F1-score of **0.6061**. Class 3 showed weaker performance, with an F1-score of **0.4000**, which is expected because the test set contained only three class-3 samples. Thus, the main limitation of the multiclass task was restricted representation of the rare highest-efficiency class. The class-wise classification report is provided in **Table 5.**

Table 5. Class-wise classification report for the best multiclass model

	precision	recall	f1-score	support
0	0.714285714	0.657894737	0.684931507	38
1	0.733333333	0.745762712	0.739495798	59
2	0.555555556	0.666666667	0.606060606	15
3	0.5	0.333333333	0.4	3
accuracy	0.695652174	0.695652174	0.695652174	0.695652174
macro avg	0.625793651	0.600914362	0.607621978	115
weighted avg	0.697763975	0.695652174	0.695204856	115

This table provides class-wise precision, recall, F1-score, and support for the best multiclass classifier. The corresponding confusion matrix demonstrates that most prediction errors occurred between adjacent delivery-efficiency classes rather than between the lowest and highest classes, suggesting that the model captured a meaningful ordinal structure in the delivery-efficiency labels. The confusion matrix is shown in **Figure 4**.

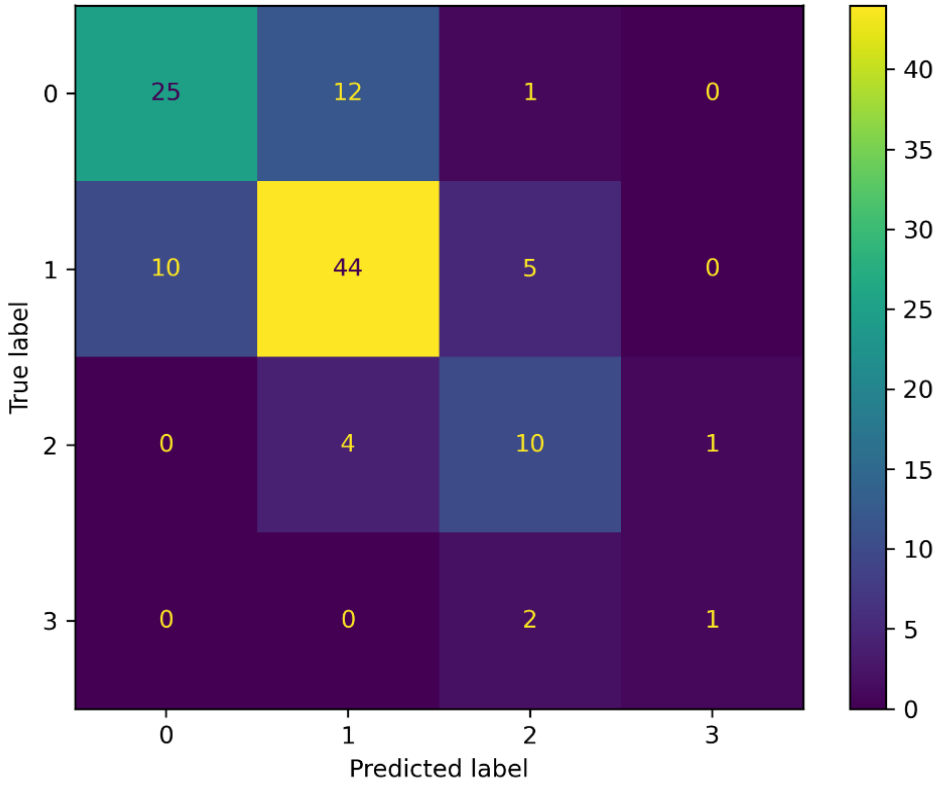

Figure 4. Confusion matrix for the best multiclass model

This figure visualizes class-wise prediction patterns and misclassification behavior for the best multiclass model.

4.5 Molecular Representation and Model Comparison

The comparison of molecular representations showed that learned molecular embeddings were more informative than sparse or lower-dimensional representations. The strongest representation-model combination was **MMB_FT_512 with Linear SVM**, which achieved a mean macro-F1 of **0.6295**. MMB-based models consistently ranked among the strongest combinations, indicating that pretrained molecular embeddings captured delivery-relevant chemical structure more effectively than conventional circular fingerprints or compact graph convolutional features. The full representation-level comparison is reported in **Table 6**.

Table 6. Model performance across molecular representations

representation	model	n_rows	n_features	accuracy_mean	balanced_accuracy_mean	f1_macro_mean	f1_weighted_mean
MMB_FT_512	Linear_SVM	572	513	0.65553013	0.640844076	0.629547599	0.655618895
MMB_512	Logistic_Regression	572	513	0.613470633	0.666872051	0.59979648	0.617108318
MMB_512	Extra_Trees	572	513	0.646773455	0.589412774	0.598946668	0.641094841
MMB_512	Linear_SVM	572	513	0.623981693	0.634883142	0.596767888	0.62682723
MMB_FT_512	Logistic_Regression	572	513	0.645064836	0.629246639	0.59096102	0.646084266
MMB_FT_512	Extra_Trees	572	513	0.657391304	0.574098096	0.583748446	0.649430904
MMB_512	Random_Forest	572	513	0.639786423	0.569881111	0.581361188	0.631864508
MMB_FT_512	Random_Forest	572	513	0.669565217	0.565531019	0.575943686	0.658615817
Expert_200	Linear_SVM	572	201	0.543676583	0.585243865	0.50377113	0.552057432
Expert_200	Random_Forest	572	201	0.587215866	0.489240978	0.486497149	0.580914312
Expert_200	Extra_Trees	572	201	0.578627002	0.478770456	0.465927908	0.573947536
CFP_2048	Random_Forest	572	2049	0.543554539	0.498063072	0.45916196	0.544333531
CFP_2048	Extra_Trees	572	2049	0.540091533	0.498283385	0.458697608	0.540927589
Expert_200	Logistic_Regression	572	201	0.515667429	0.585725814	0.457353632	0.527551288
CFP_2048	Linear_SVM	572	2049	0.522639207	0.534857797	0.438721899	0.529058733
CFP_2048	Logistic_Regression	572	2049	0.496369184	0.562816034	0.413325084	0.518254814
GCN_32	Linear_SVM	572	33	0.41610984	0.323526975	0.288771965	0.418858426
GCN_32	Random_Forest	572	33	0.461540809	0.288985028	0.287944427	0.437632488
GCN_32	Logistic_Regression	572	33	0.35136537	0.342779837	0.281554906	0.359120136
GCN_32	Extra_Trees	572	33	0.447490465	0.282636325	0.280487334	0.433451521

This table compares classifier performance across different molecular representations.

4.6 Binary High-Efficiency Prediction

The secondary binary task produced stronger discrimination than the four-class task. Logistic Regression achieved the best binary F1-score of **0.7692**, with **0.9217 a**ccuracy, **0.8857** balanced accuracy, **0.7143** precision, **0.8333** recall, and **0.9450** ROC-AUC. Although Random Forest produced slightly higher accuracy, Logistic Regression was preferable because it achieved stronger recall for the high-efficiency class, which is more relevant for screening promising LNP formulations. The binary model comparison is presented in **Table 7**.

Table 7. Binary high-efficiency classification performance

model	accuracy	balanced_accuracy	precision	recall	f1	roc_auc
Logistic_Regression	0.92	0.89	0.71	0.83	0.77	0.95
RBF_SVM	0.90	0.90	0.64	0.89	0.74	0.95
Random_Forest	0.93	0.78	1.00	0.56	0.71	0.95
Extra_Trees	0.91	0.77	0.83	0.56	0.67	0.95
Dummy_Most_Frequent	0.84	0.50	0.00	0.00	0.00	0.50

This table reports performance metrics for binary classification of low/moderate versus high-efficiency formulations.

The binary confusion matrix confirmed that the best model retained strong discrimination while correctly identifying most high-efficiency formulations. The matrix is shown in **Figure 6**.

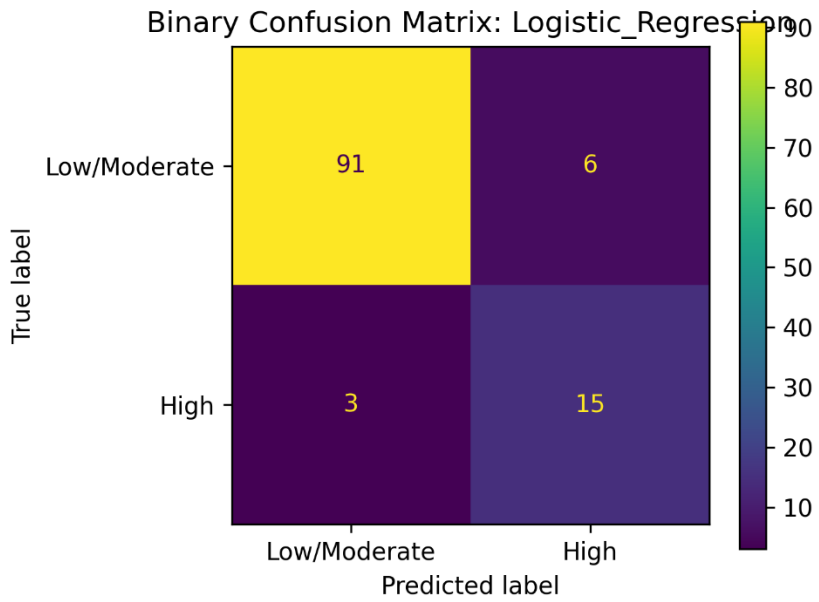

Figure 6. Confusion matrix for binary high-efficiency prediction

4.7 Model Explainability

Permutation importance analysis showed that predictive information was distributed across multiple latent embedding dimensions rather than concentrated in a single descriptor. The most influential features included fp_27, fp_457, fp_153, fp_295, and fp_490, indicating that several learned molecular dimensions contributed to delivery-efficiency classification. The top-ranked features are visualized in **Figure 7**.

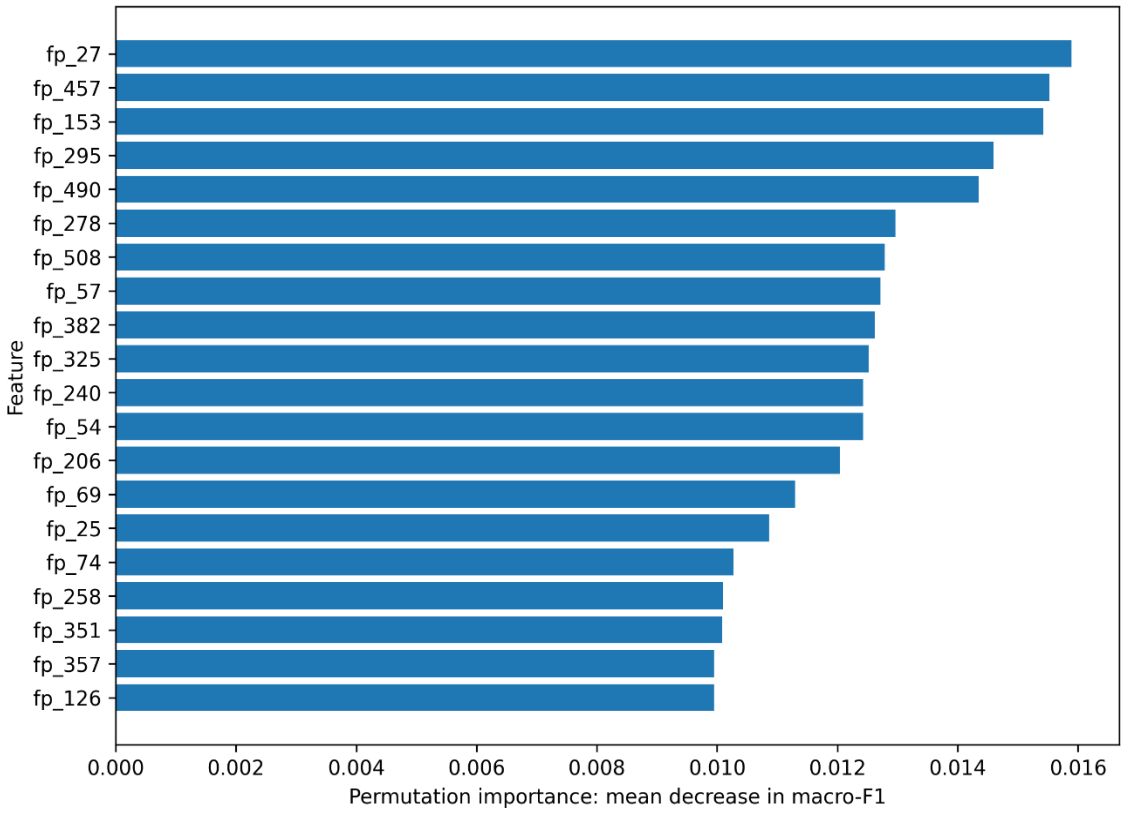

Figure 7. Top molecular features based on permutation importance

This figure presents the most influential molecular embedding features for the best multiclass model.

5. Discussion

5.1 Interpretation of Results

The results show that machine learning models can predict the efficiency of lipid nanoparticle (LNP) delivery with reasonable accuracy using molecular fingerprints of ionizable lipids and formulation-level data. The four-class prediction task demonstrated moderate but valuable performance, with Histogram Gradient Boosting outperforming other models on the independent test-set, with an accuracy of 0.6957 and macro-F1 score of 0.6076. The gap between accuracy and macro-F1 shows that the model was affected by class imbalance, especially for the least abundant highest-efficiency class. The model was more successful in distinguishing classes 0 and 1, for which it was more strongly represented in the test set, than class 3. This suggests that the predictive model learned structure-activity relationships, but its ability to predict rare high-efficiency subclasses is limited. The binary task showed better performance, with Logistic Regression reaching an F1-score of 0.7692 and ROC-AUC of 0.9450, suggesting that the data are more informative for high-efficiency vs low/moderate-efficiency predictions than for fine-scale four-class predictions.

5.2 Comparison with Existing Literature

The findings agree with recent reports that molecular representations can inform data-driven LNP design and prediction of LNP transfection efficiency. Moayedpour et al. [1] showed that molecular representations derived from large language models can be used to predict LNP transfection efficiency, and Ding et al. [2] highlighted the importance of designing LNPs for mRNA delivery using machine learning. Our results endorse this by demonstrating that learned molecular representations, such as MMB-FT, were superior to more traditional or compact representations. This is consistent with other studies in molecular machine learning that have found that learned molecular representations are able to capture complex structure-property relationships better than hand-crafted descriptors in certain cases [17], [21]. The influence of LNP structure on mRNA delivery is also consistent with experimental reports that ionizable lipid chemistry, nanoparticle composition and formulation structure/architecture have a significant effect on mRNA delivery efficiency [4], [10], [11]. Hence, the current findings support the idea that LNP design can transition from trial-and-error to predictive and representation-based LNP design.

5.3 Implications

The research has methodological and translational implications. First, it offers a replicable computational approach for molecular representation, supervised learning, and multiclass and binary classification evaluation of delivery efficiency, as well as feature interpretation. From a translational perspective, the binary classification performance indicates that the models may be helpful to screen LNP candidates prior to experimentation. This is an important consideration for LNP development, which typically requires a lot of experimental testing, and by pre-screening computationally, we can cut down on the number of experiments we may need to perform by focusing on formulations with a higher likely delivery efficiency. In terms of patient-level prediction, the model does not directly predict patient response, but rather supports the design of delivery systems that can be further customized for disease-specific, cargo-specific, or tissue-specific therapeutic applications. This is especially important for nucleic-acid-based therapies, where the mRNA or siRNA cargo can be rapidly designed to treat a particular disease, but equally requires efficient delivery.

5.4 Limitations

There are several caveats. First, the data is formulary-level rather than patient-level, so our results should not be viewed as clinical predictions. Second, the multiclass category with the highest efficiency was underrepresented, which led to less confidence in class-3 prediction and a decrease in macro-F1. Third, molecular embeddings are predictive but not necessarily interpretable, particularly when key features correspond to the latent rather than physicochemical space. Fourth, the models are trained on the available labels, so potential noise, bias or variability in the original transfection-efficiency data could affect model performance. Finally, this was a retrospective secondary analysis, without external experimental validation, limiting our claims to generalizability across all LNP chemistries, biological systems, or therapeutic cargos.

5.5 Future Directions

Future studies should include more LNP formulations, particularly high-efficiency formulations, to better balance classes and enhance prediction of rare classes. External validation with new LNP datasets would help assess generalizability. Future research should also include explicit physicochemical properties, such as particle

size, zeta potential, encapsulation efficiency, pKa, and ratios of lipids, in addition to molecular embeddings for better model performance and interpretability. Further improvements in performance may be achieved with more sophisticated models, such as graph neural networks and transformers for molecular encoding, if more data is available. Lastly, combining LNP formulation predictions with biological information (such as target tissue, cell type, disease model and cargo type) would take the approach closer to personalized LNP design and enable better prediction of delivery efficiency in a clinical context.

6. Conclusion

Machine-learning models were developed and tested for predicting the efficiency of lipid nanoparticles (LNP) for delivery using molecular descriptors of ionizable lipids and formulation-level data. The results demonstrate that supervised learning models can discover structure-activity relationships in LNP transfection-efficiency data. In the multiclass prediction scenario, the best performance on the independent test set was achieved with Histogram Gradient Boosting, suggesting that it is well suited for predicting delivery-efficiency levels under imbalanced class distributions. However, the prediction of the lowest-frequency highest-efficiency class was hindered by the lack of samples. On the other hand, binary prediction of high-efficiency LNP formulations achieved better performance, with Logistic Regression exhibiting high discriminative power, suggesting that the dataset is well suited for screening potentially efficient LNP formulations. The molecular representation analysis showed that the learned molecular embeddings were more informative than traditional or reduced-dimensionality molecular descriptors, particularly the MMB-FT features. Permutation importance analysis also suggested that the predictive information was spread across multiple latent molecular features, highlighting the importance of representation learning in LNP design. In conclusion, this work offers a computational approach to early-stage LNP screening, model comparison and prediction of LNP delivery efficiency. While the model does not explicitly predict patient-level clinical outcomes, it helps to achieve the larger aim of personalized nucleic-acid delivery by facilitating data-driven LNP prioritization. Future models should include larger and more balanced datasets, external validation, descriptors of the physiochemical properties of formulations, and biological context such as cell type, tissue target and cargo type. This would enhance interpretability, generalizability and translational potential for design of personalized nanomedicine.

REFERENCES

[1] S. Moayedpour *et al*., "Representations of lipid nanoparticles using large language models for transfection efficiency prediction," *Bioinformatics*, vol. 40, no. 7, Art. no. btae342, 2024, doi: 10.1093/bioinformatics/btae342.

[2] D. Y. Ding, Y. Zhang, Y. Jia, and J. Sun, "Machine learning-guided lipid nanoparticle design for mRNA delivery," *arXiv preprint arXiv:2308.01402*, 2023, doi: 10.48550/arXiv.2308.01402.

[3] Sanofi-Public, "LipoBART: Lipid nanoparticle design repository," *GitHub*, 2024.

[4] X. Hou, T. Zaks, R. Langer, and Y. Dong, "Lipid nanoparticles for mRNA delivery," *Nature Reviews Materials*, vol. 6, no. 12, pp. 1078–1094, 2021, doi: 10.1038/s41578-021-00358-0.

[5] N. Pardi, M. J. Hogan, F. W. Porter, and D. Weissman, "mRNA vaccines: A new era in vaccinology," *Nature Reviews Drug Discovery*, vol. 17, no. 4, pp. 261–279, 2018.

[6] R. Tenchov, R. Bird, A. E. Curtze, and Q. Zhou, "Lipid nanoparticles: From liposomes to mRNA vaccine delivery, a landscape of research diversity and advancement," *ACS Nano*, vol. 15, no. 11, pp. 16982–17015, 2021, doi: 10.1021/acsnano.1c04996.

[7] P. R. Cullis and M. J. Hope, "Lipid nanoparticle systems for enabling gene therapies," *Molecular Therapy*, vol. 25, no. 7, pp. 1467–1475, 2017, doi: 10.1016/j.ymthe.2017.03.013.

[8] A. Akinc *et al*., "The Onpattro story and the clinical translation of nanomedicines containing nucleic acid-based drugs," *Nature Nanotechnology*, vol. 14, no. 12, pp. 1084–1087, 2019, doi: 10.1038/s41565-019-0591-y.

[9] D. Witzigmann, J. A. Kulkarni, J. Leung, S. Chen, P. R. Cullis, and R. van der Meel, "Lipid nanoparticle technology for therapeutic gene regulation in the liver," *Advanced Drug Delivery Reviews*, vol. 159, pp. 344–363, 2020.

[10] K. J. Kauffman, M. J. Webber, and D. G. Anderson, "Materials for non-viral intracellular delivery of messenger RNA therapeutics," *Journal of Controlled Release*, vol. 240, pp. 227–234, 2016, doi: 10.1016/j.jconrel.2015.12.032.

[11] Y. Eygeris, S. Patel, A. Jozic, and G. Sahay, "Deconvoluting lipid nanoparticle structure for messenger RNA delivery," *Nano Letters*, vol. 20, no. 6, pp. 4543–4549, 2020, doi: 10.1021/acs.nanolett.0c01386.
[12] L. Schoenmaker *et al*., "mRNA-lipid nanoparticle COVID-19 vaccines: Structure and stability," *International Journal of Pharmaceutics*, vol. 601, Art. no. 120586, 2021.
[13] A. C. Anselmo and S. Mitragotri, "Nanoparticles in the clinic: An update," *Bioengineering & Translational Medicine*, vol. 4, no. 3, Art. no. e10143, 2019, doi: 10.1002/btm2.10143.
[14] J. Shi, P. W. Kantoff, R. Wooster, and O. C. Farokhzad, "Cancer nanomedicine: Progress, challenges and opportunities," *Nature Reviews Cancer*, vol. 17, no. 1, pp. 20–37, 2017.
[15] Q. Cheng, T. Wei, Y. Jia, L. Farbiak, K. Zhou, S. Zhang, Y. Wei, H. Zhu, and D. J. Siegwart, "Dendrimer-based lipid nanoparticles deliver therapeutic FAH mRNA to normalize liver function and extend survival in a mouse model of hepatorenal tyrosinemia type I," *Advanced Materials*, vol. 30, no. 52, Art. no. 1805308, 2018, doi: 10.1002/adma.201805308.
[16] B. Li, X. Zhang, Y. Dong, *et al*., "In vivo delivery of nucleic acid therapeutics using lipid-like nanoparticles," *Nano Letters*, vol. 17, pp. 4590–4598, 2017.
[17] J. Vamathevan *et al*., "Applications of machine learning in drug discovery and development," *Nature Reviews Drug Discovery*, vol. 18, no. 6, pp. 463–477, 2019, doi: 10.1038/s41573-019-0024-5.
[18] A. Lavecchia, "Machine-learning approaches in drug discovery: Methods and applications," *Drug Discovery Today*, vol. 20, no. 3, pp. 318–331, 2015, doi: 10.1016/j.drudis.2014.10.012.
[19] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *Journal of Chemical Information and Modeling*, vol. 50, no. 5, pp. 742–754, 2010.
[20] D. K. Duvenaud *et al*., "Convolutional networks on graphs for learning molecular fingerprints," in *Advances in Neural Information Processing Systems 28*, 2015, pp. 2224–2232.
[21] R. Irwin, S. Dimitriadis, J. He, and E. J. Bjerrum, "Chemformer: A pre-trained transformer for computational chemistry," *Machine Learning: Science and Technology*, vol. 3, no. 1, Art. no. 015022, 2022.