Acta Pharmaceutica Sinica B 2026;16(8):5053−5069

Chinese Pharmaceutical Association
Institute of Materia Medica, Chinese Academy of Medical Sciences

Acta Pharmaceutica Sinica B

www.elsevier.com/locate/apsb
www.sciencedirect.com

TOOLS

A novel machine learning framework-based rapid screening of ionizable lipids in LNPs for highly-efficient mRNA expression

Xing Duan[a,†], Jiezhou Chen[b,†], Shanhui Jiang[a,†], Linbo Qing[b,†], Xi He[a], Jun He[a], Yinjie Lei[b], Haixing Shi[a], Tingting Song[a], Xiangyu Jiao[a], Guohong Li[a], Hai Huang[a], Mengran Guo[a], Yongjun Gu[a], Changchun Zhao[a], Xiaoling Yin[a], Shengbin Liu[a], Pingyu Wang[b,*], Xiangrong Song[a,*]

[a]*Department of Critical Care Medicine and Department of Biotherapy, Frontiers Science Center for Disease-related Molecular Network, Cancer Center and State Key Laboratory of Biotherapy, West China Hospital, Sichuan University, Chengdu 610065, China*
[b]*College of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China*

Received 25 April 2025; received in revised form 26 July 2025; accepted 12 August 2025

KEY WORDS

Ionizable lipids;
Lipid nanoparticles;
Machine learning;
mRNA expression;
In vivo;
Expression prediction;
Delivery;
mRNA vaccines

Abstract Ionizable lipids are a pivotal component for lipid nanoparticles (LNPs) to optimize mRNA expression to fulfill the broad demands of various mRNA therapeutics. Traditionally, the screening of ionizable lipids needs extensive synthetic labor and stringent *in vivo* efficacy evaluation, which was often time-consuming, costly, and characterized by a lower success rate. Hence, we pioneered a machine-learning framework named Lipid with Artificial Intelligence (LipidAI) to evaluate the novel ionizable lipids rapidly. In this framework, the Methyl Tail Augmentation (MTA) strategy was first developed to triple the data by precisely adjusting the methyl groups on lipid tail chains. This ground-breaking approach compensated for data paucity in ionizable lipids libraries from the previous research and boosts model accuracy. Subsequently, the Ensemble Stacking Learning (ESL) algorithm was exploited to integrate multiple learning algorithms to surpass the predictive accuracy of a single algorithm used in former studies. Finally, we found that the predicted results of LipidAI were highly consistent with the actual data

This article is part of special issue entitled: Advancing Pharmaceutics and Pharmacology published in Acta Pharmaceutica Sinica B.
*Corresponding authors.
E-mail addresses: songxr@scu.edu.cn (Xiangrong Song), wangpingyu@scu.edu.cn (Pingyu Wang).
[†]These authors made equal contributions to this work.
Peer review under the responsibility of Chinese Pharmaceutical Association and Institute of Materia Medica, Chinese Academy of Medical Sciences.

https://doi.org/10.1016/j.apsb.2025.10.040

according to the *in vivo* expression of Luc-mRNA. Overall, this study highlights the remarkable potential of LipidAI in the rapid screening of ionizable lipids, adeptly avoiding the inherent drawbacks of traditional ionizable lipids development and thereby boosting the progress of LNP-based mRNA nano-drugs.

1. Introduction

Lipid nanoparticles (LNPs)-based mRNA therapy has demonstrated significant potential across various clinical fields, particularly in infectious disease prevention, cancer prevention and treatment, and metabolic disorder regulation[1−4]. The cornerstone of successful mRNA therapy lies in the efficient *in vivo* expression of mRNA[5−8]. Consequently, ionizable lipids, which are the central constituents of LNPs, play a pivotal role in the advancement of mRNA therapeutics[9−12]. However, the ionizable lipids on the market fall short of meeting the extensive demands for mRNA drug development, prompting research to focus on designing novel ionizable lipids that can enhance mRNA expression efficiency[13−17]. However, the traditional methods for developing ionizable lipids are labor-intensive and resource-demanding, necessitating considerable synthetic efforts and rigorous *in vivo* efficacy evaluations[18,19]. This process is not only time-consuming and costly but also characterized by a relatively low success rate. There is an urgent need to establish novel methodologies that can swiftly evaluate the efficacy of ionizable lipids, thereby accelerating the research and development of mRNA therapeutics.

Artificial intelligence (AI) is progressively infiltrating the domain of mRNA drug design, demonstrating substantial potential[20−23]. Wang et al.[20] harnessed data from 325 mRNA vaccine lipid nanoparticles preparations to develop a LightGBM predictive model, thereby significantly enhancing the efficiency of LNPs formulation research and development. For the design of ionizable lipids, the application of AI is still in its early stages. Some promising preliminary results have emerged, indicating that AI has the potential to significantly accelerate the advancement of mRNA therapeutics. Recently, Li et al. employed combinatorial chemistry to construct a dataset encompassing 584 ionizable lipids at the HeLa cell level and subsequently utilized an XGBoost binary classification model to sift through an expanded virtual library comprising 40,000 lipids, thereby facilitating the development of novel lipids[21]. Xu et al.[22] amalgamated 1200 cell-level experimental data points with 60,000 virtual data samples to engineer a platform termed Artificial Intelligence Guided Ionizable Lipid Engineering (AGILE), which incorporates 12,000 virtual data points. This platform has been instrumental in investigating the *in vitro* adaptability of ionizable lipids across various cell lines, which could guide the formulation of lipids tailored to distinct cellular characteristics. Bae et al.[23] deployed a random forest regression model to dissect 314 features of 213 LNPs, subsequently constructing new compounds predicated on these features. Despite the substantial progress in advancing ionizable lipid development, several critical challenges persist. The prevalent single models are susceptible to overfitting and underfitting, which can engender deceptive predictions. Furthermore, models predicated on *in vitro* data may not be entirely congruent with the *in vivo* mRNA expression forecasts, given the marked disparity in mRNA expression efficiency between *in vivo* and *in vitro* conditions. Most notably, the screening of ionizable lipids continues to be heavily dependent on extensive experimental validation, necessitating considerable screening endeavors. Consequently, it is imperative to develop novel strategies capable of rapidly screening novel ionizable lipids to align with the exigencies of *in vivo* applications.

This study proposed an innovative ionizable lipid screening framework named Lipid with Artificial Intelligence (LipidAI), focusing on the association between the chemical structure of ionizable lipids and mRNA expression efficiency of the corresponding LNPs *in vivo*. The predicted mRNA expression levels by LipidAI were found to be highly consistent with the actual results in mice. This high accuracy is attributed to two innovative components of LipidAI: Methyl Tail Augmentation (MTA) and Ensemble Stacking Learning (ESL) algorithm. Firstly, we created an ionizable lipid database called LipidDB, containing two classes of ionizable lipids. The first was derived from ionizable lipids reported in the literature, while the second consisted of preconstructed ionizable lipids created through the MTA method. This method pioneered the expansion of the ionizable lipid library to three times that of the real lipid library. Secondly, an innovative algorithm model, ESL, integrated with multiple algorithms, was developed to alleviate the overfitting and underfitting problems that often occur in a single model, which could reduce the misleading predictions. Thirdly, the performance of the ESL model was evaluated with a series of indicators. SHapley Additive exPlanations (SHAP) analysis was finally applied to reveal the relationship between the chemical structures of ionizable lipids and their mRNA expression efficiency *in vivo*. According to the animal results, the ESL model showed predictive performance and high accuracy, effectively aligning with the requirements of *in vivo* screening, which indeed can reduce the labor and time costs in the development of the newly available ionizable lipids. It also provided a new approach for the rational design of LNPs, thereby promoting the development of mRNA delivery systems with higher mRNA expression. In conclusion, we developed a novel machine-learning framework named Lipid with Artificial Intelligence (LipidAI) to evaluate the novel ionizable lipids rapidly (Fig. 1).

2. Materials and methods

2.1. LipidDB dataset

2.1.1. Data collection

As shown in Table 1[24−49], the LipidDB dataset is collected by retrieving relevant literature from Web of Science and Scopus with some keywords such as “LNP”, “ionizable lipid”, “lipid nanoparticle”, and so on. Following the initial screening process, a total of 47 studies that incorporated the design of novel ionizable lipids were maintained for further analysis[24−40,41−70]. After analyzing

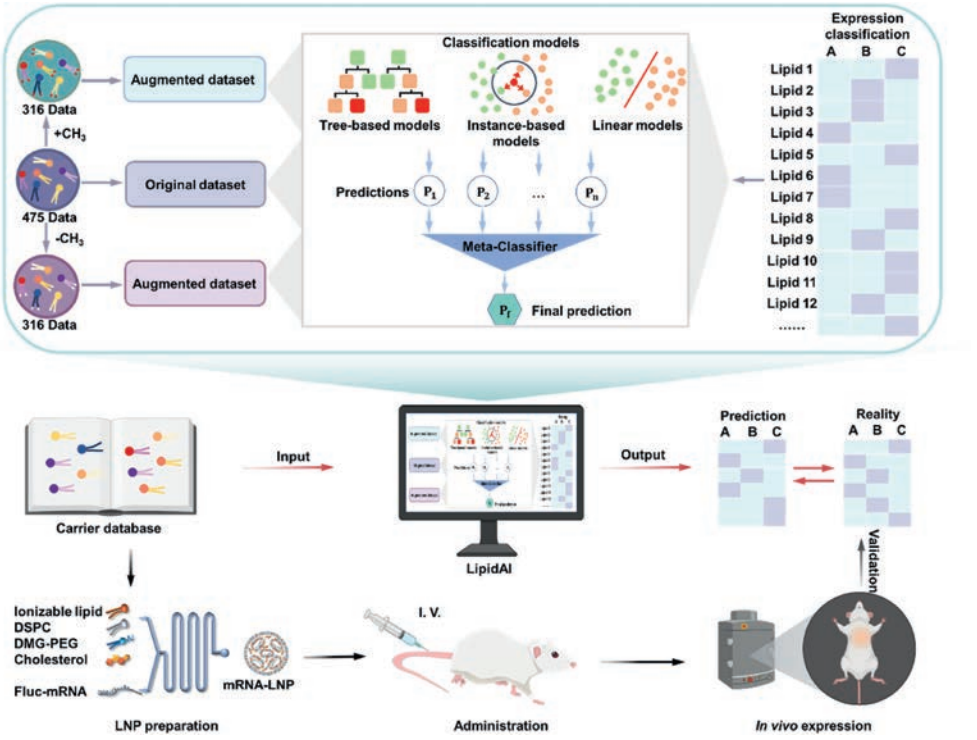


Figure 1 Graphical Abstract. By expanding the LipidDB dataset and employing a stacking learning strategy, the Lipid with Artificial Intelligence (LipidAI) framework is constructed to quickly predict the mRNA expression efficiency of newly designed ionizable lipids. Biological experimental results show that LipidAI has an impressive prediction accuracy, ensuring that the LipidAI model can screen out effective ionizable lipids.

the extracted data from 47 studies, the data are simplified to keep the numerical balance between the data amount and model parameters. All included experiments involved the use of ionizable lipids in combination with mRNA, and the ionizable lipids had corresponding mRNA expression effects *in vivo*. Based on the above standards, we collected the ionizable lipid structures and *in vivo* expression information from the literature and organized the collected data into a database. The final dataset contains the SMILES structures of ionizable lipids, their corresponding *in vivo* expression[71]. Based on the empirical analysis of the expression, we establish the effect classification and classification standards of A, B, and C. Here, we briefly introduce the classification criteria. For example, class A is designated for expressions exceeding 5×10^7, class B for expressions between 5×10^7 and 5×10^6, and class C is for expressions below 5×10^6. Notably, mRNA expression efficiency is regarded as a classification label of ionizable lipid structures.

Finally, the LipidDB dataset has 475 lipid samples, while each sample contains the SMILES structure of the ionizable lipid and its corresponding expression efficiency. As machine learning models cannot directly process chemical structure formulas as input features, the RDKit package is employed to transform the chemical structure of the delivery carrier into a comprehensive set of molecular descriptors[72]. Then, these processed descriptors are input into the machine learning model to predict the mRNA expression efficiency of the corresponding delivery system.

2.1.2. *Data augmentation*

The performance of machine learning models is heavily contingent upon the volume and quality of the training data. Since the original LipidDB is data-restricted, the model trained with the original LipidDB dataset is prone to underfitting, resulting in low prediction performance. In this work, we explore the concept of data augmentation in machine learning to generate additional training samples and facilitate the optimization of model parameters. After reviewing existing works and analyzing experimental data, we observe that the expression efficiency level of ionized lipid molecules, containing long tail chains without functional groups, usually exhibits significant insensitivity to minor variations in carbon chain length[30,34,49,73]. Motivated by this interesting observation, the Methyl Tail Augmentation (MTA) method is proposed to generate new ionizable lipids while maintaining the classification label of mRNA expression efficiency unchanged. Specifically, the MTA method involved the addition or removal of one methyl group from the tail chains of ionizable lipids.

Upon meticulous statistical analysis of the recent works, it has been discerned that the change of expression efficiency in this strategy has scientific significance for data augmentation. Therefore, the dataset is expanded by adding or removing a methyl group to the tail chain structure with a carbon number that is at least 8, while ensuring no interference from functional groups. Based on this empirical method, it can be used to generate a large number of synthesized ionizable lipids to enrich the LipidDB dataset. Finally, we get 1107 valid data points.

2.1.3. *Data splitting strategy*

As machine learning models usually include training, validation, and testing phases, we need to split the LipidDB dataset for each phase. The nested cross-validation strategy is employed to train and evaluate all machine learning models[74]. It includes an inner

Table 1 Ionizable lipid samples of the LipidDB dataset. From the first column to the sixth column, each column denotes serial numbers, chemical structures, SMILES descriptors, luminescence, rates and references, respectively.

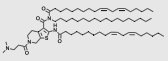
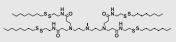
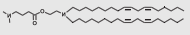
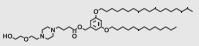
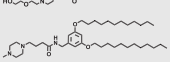
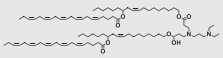
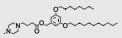
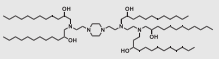
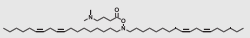
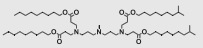
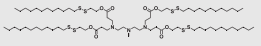
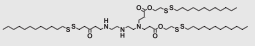
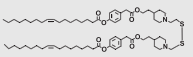
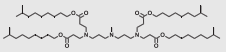
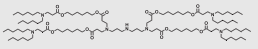
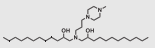
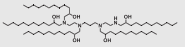
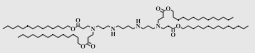
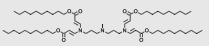
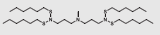
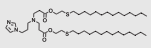
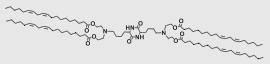
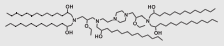
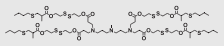
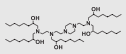
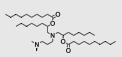
Serial number	Chemical structure	SMILES	Lumines-cence	Class	Ref.
1		O=C(CCCCCCCC/C=C\C/C=C\CCCCC)N(CCCCCCCCCCCCCCCCCC)C(C(C1=C2CN(C(CCN(C)C)=O)CC1)=C(S2)NC(CCCCCCCC/C=C\C/C=C\CCCCC)=O)=O	9000000	B	24
2		CN(CCCN(CCC(NCCSSCCCCCCCC)=O)CCC(NCCSSCCCCCCCC)=O)CCCN(CCC(NCCSSCCCCCCCC)=O)CCC(NCCSSCCCCCCCC)=O	8000000	B	25
3		CN(CCCC(OCCN(CCCCCCCC/C=C\C/C=C\CCCCC)CCCCCCCC/C=C\C/C=C\CCCCC)=O)C	17600000	B	26
4		CN(CCCC(OCC1=CC(OCCCCCCCCCCCCC)=CC(OCCCCCCCCCCCC)=C1)=O)C	31600000	B	27
5		CN1CCN(CCCC(NCC2=CC(OCCCCCCCCCCCCC)=CC(OCCCCCCCCCCCC)=C2)=O)CC1	58000000	A	28
6		CCN(CCCN(CCC(O)OCCCCCCCC/C=C\CC(OC(CCC/C=C\C/C=C\C/C=C\C/C=C\C/C=C\CC)=O)CCCCCC)CCC(OCCCCCCCC/C=C\CC(OC(CCC/C=C\C/C=C\C/C=C\C/C=C\C/C=C\CC)=O)CCCCCC)=O)CC	10000000	B	29
7		CN1CCN(CC1)CCCC(OCC2=CC(OCCCCCCCCC)=CC(OCCCCCCCCC)=C2)=O	4000000	C	30
8		CCCCCCCCCCC(O)CN(CCN1CCN(CCN(CC(O)CCCCCCCCCC)CC(O)CCCCCCCCCC)CC1)CCN(CC(O)CCCCCCCCCC)CC(O)CCCCCCCCCC	1500000000	A	31
9		CCCCC/C=C\C/C=C\CCCCCCCCN(CCOC(CCCN(C)C)=O)CCCCCCCC/C=C\C/C=C/CCCCC	2000000	C	32
10		CCCCCCCCCOC(CCN(CCCN(C)CCCN(CCC(OCCCCCCCC(C)C)=O)CCC(OCCCCCCCC(C)C)=O)CCC(OCCCCCCCC)=O)=O	25000000	B	33
11		O=C(OCCSSCCCCCCCCCCCC)CCN(CCN(C)CCN(CCC(OCCSSCCCCCCCCCCCC)=O)CCC(OCCSSCCCCCCCCCCCC)=O)CCC(OCCSSCCCCCCCCCCCC)=O	50000000	B	34
12		O=C(CCN(CCNCCNCCC(CCSSCCCCCCCCCCCC)=O)CCC(OCCSSCCCCCCCCCCCC)=O)OCCSSCCCCCCCCCCC	2000000	C	35
13		O=C(OC1=CC=C(C=C1)CC(OCCC2CCN(CC2)CCSSCCN3CCC(CC3)CCOC(CC4=CC=C(C=C4)OC(CCCCCCC/C=C\CCCCCCCC)=O)=O)=O)CCCCCCC/C=C\CCCCCCCC	30000000	B	36
14		CC(C)CCCCCCCOC(CCN(CCC(OCCCCCCCC(C)C)=O)CCCN(C)CCCN(CCC(OCCCCCCCC(C)C)=O)CCC(OCCCCCCCC(C)C)=O)=O	1100000000	A	37
15		CCCCCCN(CCCCCC)CCC(OCCCCCCCOC(CCN(CCC(OCCCCCCCOC(CCN(CCCCCC)CCCCCC)=O)=O)CCCNCCCN(CCC(OCCCCCCCOC(CCN(CCCCCC)CCCCCC)=O)=O)CCC(OCCCCCCCOC(CCN(CCCCCC)CCCCCC)=O)=O)=O)=O	200000000	A	38
16		OC(CCCCCCCCCC)CN(CC(CCCCCCCCCC)O)CCCN1CCN(C)CC1	1000000	C	39
17		OC(CN(CCCN(CC(CCCCCCCCCC)O)CCNCC(CCCCCCCCCC)O)CCN(CC(CCCCCCCCCC)O)CC(O)CCCCCCCCCC)CCCCCCCCCC	100000	C	40
18		O=C(CCN(CCCNCCCCNCCCN(CCC(OCCCCCCCCCCCCCC)=O)CCC(OCCCCCCCCCCCCCC)=O)CCC(OCCCCCCCCCCCCCC)=O)OCCCCCCCCCCCCCC	30000000	B	41
19		CN(CCCN(/C=C/C(OCCCCCCCCCC)=O)/C=C/C(OCCCCCCCCCC)=O)CCCN(/C=C/C(OCCCCCCCCCC)=O)/C=C/C(OCCCCCCCCCC)=O	116000000000	A	42
20		CN(CCCN(SCCCCCC)SCCCCCC)CCCN(SCCCCCC)SCCCCCC	22000000	B	43

Table 1 (*continued*)

Serial number	Chemical structure	SMILES	Lumines-cence	Class	Ref.
21		O=C(CCN(CCC(OCCSCCCCCCCCCCCCCCC)=O)CCCN1C=CN=C1)OCCSCCCCCCCCCCCCCCC	9000000	B	44
22		O=C(C(CCCCN(CCOC(CCCCCCCC/C=C\C/C=C\CCCCC)=O)CCOC(CCCCCCCC/C=C\C/C=C\CCCCC)=O)NC1=O)NC1CCCCN(CCOC(CCCCCCCC/C=C\C/C=C\CCCCC)=O)CCOC(CCCCCCCC/C=C\C/C=C\CCCCC)=O	27500000	B	45
23		CCO[C@H](CN(C[C@@H](O)CCCCCCCCCCCCC)CC(O)CCCCCCCCCCCCC)CN(CC1)CCN1CCN(CC(O)CCCCCCCCCCCCC)CC(OCC)CN(C[C@@H](O)CCCCCCCCCCCCC)CC(O)CCCCCCCCCCCCC	450000000	A	46
24		CC(C(OCCSSCCOC(CCN(CCC(OCCSSCCOC(C(C)CSCCCC)=O)=O)CCCN(C)CCCN(CCC(OCCSSCCOC(C(CSCCCC)C)=O)=O)CCC(OCCSSCCOC(C(CSCCCC)C)=O)=O)=O)=O)CSCCCC	300000	C	47
25		OC(CN(CC(O)CCCCCC)CCN(CC1)CCN1CCN(CC(O)CCCCCC)CCN(CC(O)CCCCCC)CC(O)CCCCCC)CCCCCC	50000000	B	48
26		CN(CCCN(CC(OC(CCCCCCCCC)=O)CCCCCC)C[C@H](OC(CCCCCCCCC)=O)CCCCCC)C	1700000000	A	49
27	…	…	…	…	…

loop for model training and an outer loop for model evaluation. For comprehensively evaluating the performance of each machine learning model and optimizing model parameters, a total of 20 rounds of nested cross-validation are conducted for each machine learning model. The detailed process is described as follows.

The train-test-split method from the scikit-learn library in Python is employed within the outer loop to partition the samples, allocating 20% as a test set for evaluating the final model[75]. The test set is diversified by setting different random number seeds for each division, thereby ensuring a comprehensive evaluation. The remaining 80% of the samples are allocated for model training within the inner loop. The subsequent inner loop incorporates a stratified k-fold ($k = 10$) cross-validation strategy. 10% of the training set is randomly selected to evaluate model hyperparameters in each iteration, while the remaining 90% is designated for model training. Moreover, a random grid search technique is employed to expedite the optimization of model hyperparameters. The training process involves evaluating 100 diverse combinations of parameters simultaneously, while systematically recording the optimal model parameters throughout. Subsequently, the model with the best performance is selected to rigorously evaluate model performance with the test set from the outer loop.

2.2. *Ensemble stacking learning*

The challenge of complex datasets often leads to underperformance and susceptibility to interference in single models, prompting researchers to explore various ensemble machine learning techniques. With the motivation of a stacking technique, we put forward an Ensemble Stacking Learning (ESL) model to enhance the generalization capability and achieve higher prediction accuracies[76], the overview of ESL is shown in Fig. 2A. In developing ESL models, we meticulously structure the process into three core phases: base learner construction, meta learner construction, and feature engineering. During the base learner construction phase, we train a variety of models with different types or parameter configurations, selecting those that demonstrate strong performance for the stacking process. The subsequent phase, meta learner construction, involves creating a model designed to integrate the prediction outputs from all base learners, thereby generating the final prediction results. In contrast, the feature engineering phase focuses on simplifying the model structure by removing redundant features, which aims to enhance both the efficiency and performance of the model. In the following sections, we will provide a detailed explanation of the specific construction methods employed at each stage.

2.2.1. *Base learner construction*

In order to learn knowledge from multiple perspectives, we explore eight machine learning models, namely LightGBM, Decision Tree (DT), K-Nearest Neighbour (KNN), Random Forest (RF), Support Vector Machine (SVM), XGBoost, Gradient Boosting Decision Tree (GBDT), and Logistic Regression (LR). In practice, the LightGBM model is constructed with the LightGBM 3.2.1 package[77], while the XGBoost model is built with the XGBoost 1.5.1 package[78]. The remaining models are implemented with the scikit-learn 1.0.2 package. To ensure experimental fairness, all samples are normalized by the preprocessor module of the scikit-learn package. Subsequently, a nested cross-validation strategy in conjunction with random grid search is employed to optimize model hyperparameters. The experimental results indicate that random search tends to outperform manual grid search in terms of efficiency[79]. In the external loop, the optimal models derived from the internal loop are evaluated on the test set. The model with the highest accuracy is saved as the best-optimized model after conducting a thorough comparison over 20 iterations. Additionally, default parameter models are also trained and saved in a similar way. Among the 16 saved models, those with an accuracy exceeding 0.7 are selected as the base learners for the stacking ensemble method, which aims at enhancing the overall model performance.

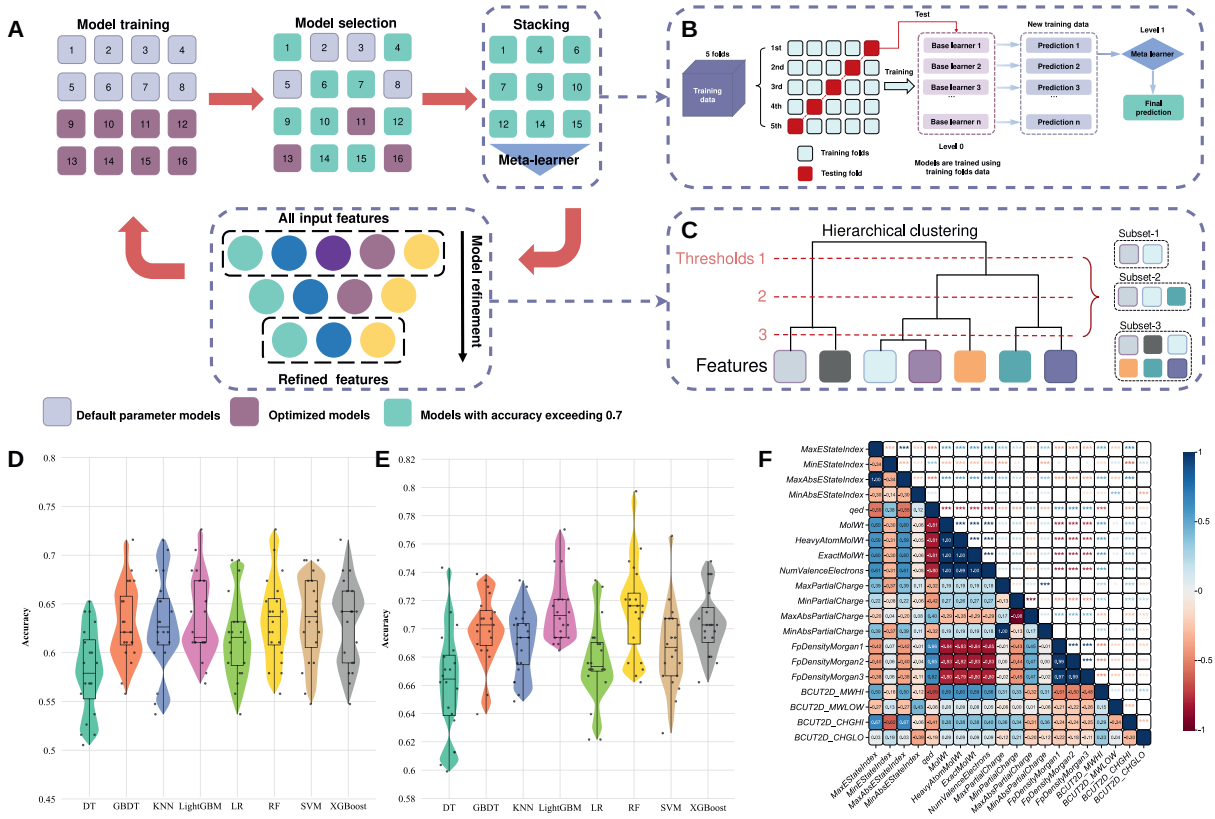


Figure 2 Model construction and summary of the performance of different machine learning models over 20 rounds of external iterations. (A) Overview of the ESL method. Model training: 20 rounds of nested cross-validation are conducted on 8 machine learning models to obtain optimal models, as well as default model training for the 8 machine learning models. Model selection: The models with an accuracy exceeding 0.7 are selected among the 16 models. Stacking: An ensemble learning stacking approach is employed to combine multiple base learners, and LR is utilized as the meta learner. The base learner is dynamically adjusted in each iteration based on the prediction performance in the previous step. Model refinement: Agglomerative hierarchical cluster analysis is implemented to identify the best feature subset. (B) Flowchart of the stacking algorithm. The base learners adopt the cross-validation strategy to create a new dataset for training the meta learner. (C) Flowchart of the feature subset extraction. All input features are organized into a hierarchical clustering structure. The first feature in each cluster is then filtered based on a set threshold and selected as the representative feature for that cluster. (D) Accuracies of models trained with the original 475 samples. (E) Accuracies of models training with the augmented 1107 samples. (F) Heatmap of the Spearman's Rank correlation between the first 20 features. The features with strong correlations are beneficial to model refinement. *** Means extremely significant correlation, $P < 0.001$.

2.2.2. *Meta learner construction*

In order to minimize overfitting, the proposed ESL model adopts the StackingCVClassifier algorithm from the mlxtend 0.23.1 package to incorporate cross-validation. Fig. 2B illustrates the construction of meta learners[80].

Suppose S is a collection of samples $S_i = (x_i, y_i), i = 1,2,3,\ldots, N$, where N is the total number of datasets, x_i represents the input features and y_i denotes the category label of the i-th sample. In this model, $B_j (j = 1, 2, 3, \ldots, M)$ represents base learning algorithms, with M is the total number of algorithms. Initially, the dataset S is divided into two subsets. The first subset is used to train base learners to create level-0 classifiers as shown in Eq. (1):

$$P_{jk} = B_j(S - S_k), \forall k = 1, 2, 3, \ldots \quad (1)$$

where S_k denotes the k-th test set divided in the cross-validation and $S - S_k$ is the training set data used in the k-th cross-validation. The second subset S_k is input into trained classifiers P_{jk} to output prediction results as shown in Eq. (2):

$$w_{jk} = P_{jk}(S_k), \forall j = 1, 2, \ldots, M, \forall k = 1, 2, 3, \ldots \quad (2)$$

In cross-validation, the model P_{jk} is trained with $k - 1$ folds of data, and prediction results are made on the remaining fold of data. Consequently, at the end of the cross-validation process, we obtain prediction results on all k test sets. Given that each partition of the test set is independently distinct, the collective set of all test sets effectively constitutes the original dataset S. We accomplish prediction results for the entire dataset, and then all prediction results are constructed as a new dataset S'. Specifically, this newly constructed dataset $W_j = (w_{j1}, w_{j2}, \ldots, w_{j5})$ is derived from the outputs of the level-0 learner. Therefore, the new input dataset $S' = (W_1, W_2, \ldots, W_M)$, generated by all M base learners, will serve as a fresh input feature to train the level-1 meta learner Q, alongside the category labels y from the original dataset S. Thus, Q can generate the final prediction result by combining the M results from the base learners as shown in Eq. (3):

$$Y_x = Q(B_1(x), B_2(x), \ldots, B_M(x)) \quad (3)$$

where Y_x denotes the final prediction result.

The selection of input attributes and algorithms for the secondary learner plays a crucial role in the stacking ensemble's

overall performance. Given that complex nonlinear transformations have been employed in generating S', it is crucial to select a simple classifier for level-1 to minimize the risk of overfitting. A suitable choice for this purpose is logistic regression within the generalized linear model framework. The meta learner, constructed by LR, utilizes the category probabilities produced by the base learner as input features. Additionally, a random parameter search approach is implemented to further optimize the meta learner, evaluating 100 sets of parameter combinations to identify the most effective configurations. These strategic steps enable the ESL model to effectively leverage the diversity of base learners while mitigating overfitting risks.

2.2.3. Model evaluation

In this study, the prediction performance of ESL is comprehensively evaluated using three key metrics: accuracy (ACC), area under the curve (AUC), and F1 score (F1). When evaluating model performance, these metrics can provide diverse insights to understand its predictive capabilities.

In the context of classification metrics, True Positive (TP) denotes the count of positive instances that are accurately identified as positive. Conversely, False Positive (FP) indicates the number of negative instances that are erroneously classified as positive. True Negative (TN) presents the number of negative instances that are correctly identified as negative. Lastly, False Negative (FN) refers to the number of positive instances that are mistakenly classified as negative. Accuracy measures the proportion of correctly predicted samples to the total samples, as defined in Eq. (4). Throughout the model training process, the accuracy metric is employed to evaluate the effectiveness of model training.

$$\mathrm{ACC}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{4}$$

The F1 score integrates precision and recall metrics to offer a holistic perspective on the model's performance in balancing precision and recall, as defined in Eq. (5):

$$F1=2\cdot\frac{\mathrm{Precision}\cdot\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}} \tag{5}$$

Precision quantifies the proportion of samples predicted as positive that are true positives, reflecting the accuracy of the model's positive identifications, as defined in Eq. (6):

$$\mathrm{Precision}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{6}$$

Conversely, recall gauges the ratio of true positive samples classified by the model, indicating its effectiveness in capturing all positive samples without omission, as defined in Eq. (7). This metric highlights the model's competence in recognizing all positive instances.

$$\mathrm{Recall}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{7}$$

2.2.4. Feature selection

Version 1.7.3 of the scipy library is utilized to conduct agglomerative hierarchical analysis[81]. The initial step involves applying the farthest-neighbor clustering algorithm to organize the input features into a hierarchical cluster structure, which is crucial for feature selection and facilitates grouping features based on similarity. Subsequently, various thresholds are established using the linkage distances to allocate features to their respective clusters. The final step is to extract the first feature from each cluster to represent it and construct a subset of features based on the current threshold. The entire process above is illustrated in Fig. 2C. This approach enables us to extract multiple feature subsets, each offering a distinct perspective of the original dataset. Our evaluation focuses on comparing the performance of stacking models trained with these feature subsets to those trained with the complete feature set. The objective of this comparison is to assess whether model accuracy is impacted by feature reduction. We aim to identify a subset that maintains most of the important information within the full feature set. The discovery of this feature subset suggests that model performance can be effectively enhanced through feature selection, leading to reduced computational complexity. Once confirmed, the refined feature subset will be used to retrain the base learners of the model, followed by repeating the stacking process, as shown in Fig. 2A. This iterative process can be stopped when further refinement of features no longer yields performance improvements. The feature selection is aimed at delivering a highly accurate and efficient model, ensuring effectiveness and scalability in practical applications.

2.2.5. Model interpretation

Inspired by game theory, SHapley Additive exPlanations (SHAP) is used to elucidate the predictions from machine learning models. Renowned for its robust theoretical underpinnings that offer fairness, accuracy, and consistency in interpreting models, SHAP excels at quantifying the impact of each feature on model predictions, thus ensuring transparency in the decision-making process of the model. In this work, Version 0.42.1 of the SHAP library is utilized to conduct a comprehensive analysis on the significance of features within the model[82,83]. By harnessing SHAP values, we can pinpoint the pivotal features influencing the model's predictions, thereby grasping the behavior and potential biases inherent in the model. Such analysis not only enhances model interpretability but also plays a crucial role in enhancing model performance.

2.3. Biological experimental verification

2.3.1. Materials and reagents

As shown in Table 1, Lipid E 1-9 and mRNA encoding luciferase (mLUC) were provided by WestGene Biopharmaceutical Technology Co., Ltd. (Chengdu, China). ALC-0315, SM-102, and MC3 are synthesized in our laboratory. Cholesterol (Chol) was supplied by Shanghai Yuanju Biology Technology Company (Shanghai, China). DMG-PEG-2000 and DSPC were purchased from Avituo Pharmaceutical Technology Co., Ltd. D-Luciferin (potassium salt) was purchased from Biovision.

2.3.2. Animals

BALB/c mice (6−8 weeks) were purchased from Beijing Huafukang Biotechnology Co., Ltd. (Huafukang, Beijing, China). Prior to the commencement of the experiment, the mice were acclimated in the animal facility at Sichuan University for a duration of two weeks. Male animals were used with random grouping. All animal studies were approved by the Sichuan University Institutional Animal Care and Use Committee and performed following the animal care and institutional guidelines. The approval number is 20240226053.

2.3.3. Formulation of mRNA-loaded LNPs

For synthesizing LNPs, a microfluidic device facilitated the mixing of an aqueous phase enriched with mRNA and an ethanolic phase

comprising lipids and cholesterol, thereby enabling the formation of encapsulated RNA within the lipidic matrix. Briefly, the aqueous phase was prepared with 0.01 mol/L citric acid buffer (pH 3) containing mRNA. The ethanolic lipid phase was prepared by blending ionizable lipids, DSPC, cholesterol, and DMG-PEG-2000 in molar ratios of 50%, 10%, 38.5%, and 1.5%, respectively. Then the two distinct phases were mixed in a 3:1 ratio under a microfluidic device, dialyzed using 0.01 mol/L citric acid buffer at pH = 6 for 2 h, and then sterilized with a 0.22 μm filter.

2.3.4. *In vivo biodistribution of Luc mRNA-LNPs*

According to the manufacturer's protocols, we utilized the *in vivo* imaging system from Caliper Life Sciences to track the time-course biodistribution of luciferase activity. Post-vaccination with Luc mRNA-LNPs, photographic documentation of the mice was conducted at 6 h. Subsequently, a luciferase imaging system was employed to assess luciferase activity. Before imaging, the mice received an intraperitoneal injection of d-luciferin potassium salt (MeiLunBio) at a dosage of 150 mg/kg. Furthermore, 6 h post-vaccination with Luc mRNA-LNPs, the mice were euthanized, and the luciferase activity in the heart, liver, spleen, lung, and kidney was determined using the same imaging system. Living Image® 4.3.1 Software (https://www.perkinelmer.com) is used to measure the average radiance of the Region of Interest (ROI).

3. Results and discussion

3.1. *Ensemble stacking learning*

The convergence of machine learning with ionizable lipid research is emerging as a cutting-edge frontier in scientific inquiry, with an escalating number of studies dedicated to harnessing the predictive prowess of AI to revolutionize the development and understanding of lipid-based therapeutics. Daniel G. Anderson is a pioneering figure, being among the first scientists to integrate machine learning techniques with the development of ionizable lipids[21]. The binary classification model he crafted using the XGBoost algorithm has significantly expedited the design process of these lipids, showcasing the power of AI in advancing pharmaceutical research. However, when dealing with intricate and fluctuating datasets, it is insufficient to utilize a single model, especially when the dataset is small. Since the single model struggles to capture all sample features effectively, it may be susceptible to noise and outliers. Previous research on applying machine learning to LNP development has been influenced by this factor[20,84,85]. Differently, the proposed ESL model adopts ensemble stacking learning to enhance overall performance by aggregating predictions from multiple models. As shown in Fig. 2A, this two-level stacking approach involves multiple base learners and a meta learner. Firstly, these base learners undertake the initial learning and prediction tasks. Then, the ESL model deliberately selects diverse base learners, such as LightGBM, XGBoost, Decision Tree algorithm, and so on. Since each base learner has specific advantages for different scenarios, these base learners can provide comprehensive prediction results from various perspectives.

In the proposed ESL model, the logistic regression algorithm serves as the meta learner to improve prediction performance by effectively integrating the prediction results of base learners. To ensure model robustness, a 5-fold cross-validation approach is used to train the base learners, which can prevent overfitting to specific datasets during training. Subsequently, the prediction results from the base learners are utilized as inputs for the meta learner to generate the final predictions. With the help of the meta learner, the proposed ESL model optimally leverages the strengths of each base learner to boost prediction performance and interference resistance.

Initially, we employ a nested cross-validation approach to train and evaluate all selected base learner models. In the process of nested cross-validation, the dataset is divided into a training set and a test set, where 20% of the samples are randomly selected as the test set and the remaining 80% of the samples are used as the training set. The inner loop involves performing hyperparameter optimization for each model using a k-fold ($k = 10$) cross-validation method. Here, the model hyperparameters are fine-tuned by a random search to maximize the prediction performance. We evaluate each model with 20 rounds to obtain the average performance on a randomly generated test set. During the evaluation, it is essential to highlight that we employ the classification accuracy of mRNA expression efficiency as the primary evaluation metric.

The performance of a machine learning model is significantly influenced by both the quantity and quality of the training dataset. Data augmentation is a widely employed technique in machine learning to enhance the dataset. Drawing inspiration from research on the correlation between carbon chain length and the expression efficiency of ionizable lipids, we introduce a novel data augmentation method, MTA, specifically tailored for LNP. Fig. 2D and E summarize the model accuracies obtained on each round of the out-of-loop test set to facilitate a more intuitive comparison of the model performance. Specifically, Fig. 2D illustrates the distribution of prediction accuracies over the 20 rounds of validation, while Fig. 2E presents the distribution of prediction accuracies after MTA. Table 2 presents the mean and standard deviation of the 20-round accuracy distribution for each model, along with the improvement in model accuracy following MTA. With the help of MTA, the accuracy of the decision tree model is significantly improved by 13.8%. Overall, all models achieve an average accuracy improvement of 11%. The significant performance improvement not only provides evidence of the effectiveness of the MTA strategy but also emphasizes the crucial role of appropriate data processing in machine learning model development.

To further assess the stability and superiority of the machine learning model, we apply it once again to the augmented LipidDB dataset with a 10-fold cross-validation approach. Additionally, for comprehensively evaluating the model's performance, we conduct a comparison between the optimal parameter configurations and the default ones. As reported in Table 3, the specific performance evaluation of the model can be found in the rows corresponding to

Table 2 Mean and standard deviation of prediction accuracies for 20 rounds of outer loop, and improvement with MTA.

Strategy	LightGBM	DT	KNN	RF	SVM	XGBoost	GBDT	LR
Without MTA	0.64 ± 0.04	0.58 ± 0.04	0.63 ± 0.05	0.63 ± 0.04	0.63 ± 0.04	0.63 ± 0.04	0.63 ± 0.04	0.62 ± 0.04
MTA	0.71 ± 0.02	0.66 ± 0.04	0.69 ± 0.02	0.71 ± 0.04	0.69 ± 0.03	0.70 ± 0.02	0.70 ± 0.02	0.68 ± 0.03
Improvement	+10.9%	+13.8%	+9.5%	+12.7%	+9.5%	+11.1%	+11.1%	+9.7%

Table 3 Model accuracies following 10-fold cross-validation on the augmented dataset.

Feature set	Parameter setting	LightGBM	DT	KNN	RF	SVM	XGBoost	GBDT	LR
174-set	Default	0.7145	0.6919	0.6893	0.7009	0.6739	0.7154	0.7145	0.6794
	Optimal	0.7235	0.7010	0.6964	0.7199	0.6983	0.7299	0.7190	0.6865
109-subset	Default	0.7136	0.6919	0.6938	0.7037	0.6776	0.7199	0.7217	0.6658
	Optimal	0.7335	0.7146	0.7236	0.7380	0.6983	0.7335	0.7325	0.6793
74-subset	Default	0.7136	0.6901	0.6802	0.7091	0.6685	0.7127	0.7280	0.6640
	Optimal	0.7398	0.7028	0.7200	0.7443	0.6893	0.7407	0.7344	0.6649

the 174-set. According to the selection of base learners, we explore various methods for the 16 models listed in Table 3. The base learners employed in each method and the corresponding accuracy of the stacked models can be found in Table 4.

Specifically, we conduct experiments with various base learner combination strategies to develop more effective stacking models. Initially, we analyze all 16 models presented in Table 3, followed by a detailed examination of all parameter-optimized models. Subsequently, we select the top-performing parametric models with an accuracy threshold exceeding 0.7. As reported in Table 4, we observe that the final strategy obtains the highest prediction accuracy. This strategy exclusively utilizes models with an accuracy greater than 0.7, so the ESL model can achieve the highest prediction performance. This observation not only highlights the effectiveness of this specific approach but also emphasizes a crucial insight regarding the nature of the stacking methodology. The success of stacking models is not a matter of chance but rather arises from the selection and fusion of base learners. In this way, the proposed ESL model not only leverages the most robust prediction capabilities but also preserves the diversity of prediction results for optimal performance.

3.2. Model refinement

We have collected a new lipid dataset, named LipidDB, which contains the chemical structural formulas of ionizable lipids in LNPs and their corresponding mRNA expression efficiencies. However, machine learning models are unable to directly use chemical structural formulas as input. Therefore, the RDKit library in Python is utilized to calculate the chemical structures of these ionizable lipids, and then we transform these ionizable lipids into 209 molecular descriptors. After eliminating feature elements that are consistently zero, the final feature set is reduced to 174 features.

However, introducing too many potentially irrelevant features not only complicates the model training but also hampers the generalization capability of machine learning models. To ensure high prediction performance, it is crucial to maintain the important feature subset of input features for machine learning models.

Table 4 The accuracies of models with different stacking methods. The models in Table 3 are denoted by numbers 1 to 16 in consecutive order.

Number of base learners	Selected base learner	Accuracy
16	1–16	0.7399
8	9–16	0.7236
5	9, 10, 12, 14, 15	0.7181
9	1, 4, 6, 7, 9, 10, 12, 14, 15	0.7479

Hence, we need to eliminate nearly redundant features from the original feature set to refine machine learning models. To underscore the importance of model refinement, Fig. 2F presents the Spearman correlation heatmap for the top 20 features out of 174. Interestingly, the heatmap reveals strong correlations between some features. For example, there is an extremely strong correlation (1.00) among MolWt, HeavyAtomMolWt, ExactMolWt, and NumValenceElectrons, which suggests that they provide almost identical information. The results of Fig. 2F reveal that the previously mentioned examples are not isolated cases. Consequently, it is very crucial to select a feature subset for model refinement.

Specifically, we utilize a farthest neighbor clustering algorithm to organize input features into a hierarchical cluster structure. For assigning features to different clusters, a set of thresholds is established based on linkage distances. As shown in Fig. 2C, the flexibility to adjust these thresholds allows us to divide the original feature set into different feature subsets with varying sizes and characteristics. Following this, the performance of the feature subset is assessed through 10 rounds of cross-validation.

As reported in Table 5, the model trained with the subset of 109 features achieves a prediction accuracy that is comparable to the stacking model using the entire feature set. This observation prompts us to embrace the streamlined 109-subset as the new training feature set. Subsequently, we examine the performance of the base learners trained with this subset in the rows corresponding to the 109-subset in Table 3. Given different feature sets from the original model, a re-implementation of nested cross-validation for all base learner models is conducted to meticulously optimize the model parameters. The optimized KNN model achieves an accuracy of 0.7236. Consistent with the previous approach to select base learners, the KNN model is incorporated into the base learners for stacking. Moreover, to further enhance the model's performance, the meta-learner's parameters are optimized, resulting in an increased prediction accuracy of 0.7488.

After the last refinement, the stacking model trained using the 109-subset is further optimized. Table 6 presents the top 10 feature subsets ranked by accuracy post-refinement. Notably, the subset comprising 74 features has the highest classification accuracies, surpassing the 109-subset while utilizing fewer features. Consequently, we adopt it as the new training feature set for subsequent model training. The 74-subset row in Table 3 depicts the accuracy of the model trained with this subset. The results reveal that, despite feature refinement efforts, there don't exist new models that exceed an accuracy of 0.7. Additionally, despite conducting a new round of random parameter search on the model's base and meta learners, there is no improvement in the stacking model's accuracy. The results of the third and final round reveal that a slight decrease in accuracy even for the most effective subset compared to the preceding iteration. Consequently, this observation leads us to postulate that the model's performance has

Table 5 Summary of each feature subset after the first round of model refinement. The prediction accuracy is calculated through 10-fold cross-validation with the stacking model.

Subset	Number of features	Accuracy	Standard deviation	Wards linkage distance
1	109	0.7479	0.0454	0.1551
2	109	0.7479	0.0454	0.1609
3	98	0.7470	0.0420	0.2126
4	98	0.7470	0.0420	0.2183
5	98	0.7470	0.0420	0.2241
6	98	0.7470	0.0420	0.2298
7	170	0.7460	0.0491	0
8	112	0.7452	0.0399	0.1436
9	104	0.7443	0.0484	0.1781
10	121	0.7443	0.0492	0.0919

Table 6 Summary of each feature subset after the second round of model refinement. Where the accuracy is determined through 10-fold cross-validation using the best stacking model from the prior refinement round.

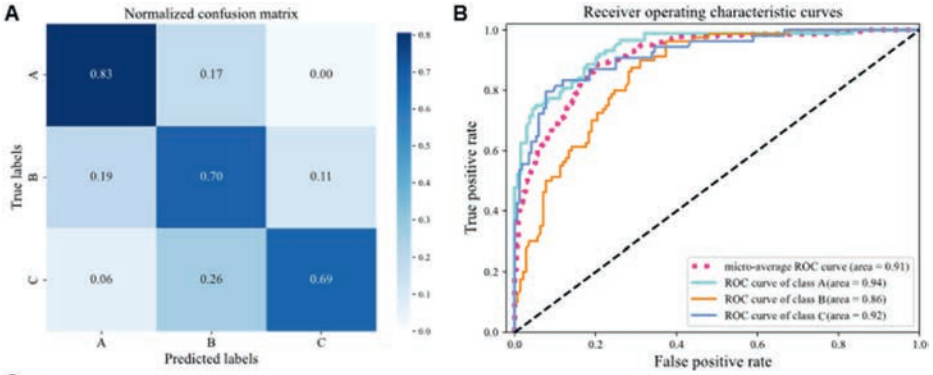
Subset	Number of features	Accuracy	Standard deviation	Wards linkage distance
1	74	0.7579	0.0440	0.3761
2	74	0.7579	0.0440	0.3853
3	99	0.7569	0.0503	0.1926
4	82	0.7515	0.0432	0.3027
5	95	0.7506	0.0433	0.2201
6	67	0.7506	0.0546	0.4403
7	102	0.7506	0.0482	0.1651
8	102	0.7506	0.0482	0.1743
9	102	0.7506	0.0482	0.1834
10	109	0.7488	0.0443	0

plateaued at a relatively stable and optimized level. Accordingly, it is deduced that further reductions in features may not yield the anticipated performance enhancements but could lead to a loss of valuable information.

As illustrated in Fig. 3A and B, the confusion matrix and ROC curves reveal that the proposed model can obtain excellent prediction performance. Specifically, the area under the ROC curve (AUC) is 0.94 for class A and 0.92 for class C, so the proposed ESL model achieves high true positive rates and low false positive rates. The micro-average area under the ROC curve is 0.91, suggesting that the model performs well for all categories. Fig. 3C also shows the values of various metrics obtained from 10-fold cross-validation. These results demonstrate the efficiency and reliability of the proposed ESL model for predicting mRNA expression efficiency.

3.3. *Model interpretation*

Ionizable lipids, with their key features, play a pivotal role in determining the efficacy of LNPs delivery systems, a fact



C

Test set				10-fold cross validation			
Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
0.7477	0.7551	0.7382	0.7448	0.7579	0.7664	0.7518	0.7566

Figure 3 The performance of ESL on a randomly split test set. ESL shows strong prediction performance with a micro-average AUC of 0.91. (A) Confusion matrix. (B) ROC curve. (C) Key performance metrics.

underscored by et al.'s innovative application of a random forest regression model[23]. This model was adeptly employed to analyze an extensive dataset comprising 314 features of 213 LNPs, leading to the construction of novel compounds predicated on these insightful features. The implications of feature assays in the development of ionizable lipids are profound, as they not only enhance our understanding of these lipids' characteristics but also significantly expedite the design process. The model, grounded in this study, offers a wealth of data for analysis, sparking new avenues of thought for the subsequent design of ionizable lipids. SHAP analysis is renowned for its robust interpretative capabilities and is widely applied across various domains. For instance, it is instrumental in deciphering predictions from machine learning models, assessing the pivotal influence of key features on drug release forecasts, and steering the development of innovative long-acting injectable drug formulations[74]. In this study, we utilize the SHAP analysis to investigate the impact of 74 refined features on the prediction capacity for mRNA expression efficacy. To elucidate the significance of individual features, the contribution of each feature to the model prediction is quantified by the SHAP analysis. The SHAP values of all test set samples are projected into the probability space to enhance the interpretability of the model predictions. Since the prediction of mRNA expression efficiency belongs to a three-category problem, SHAP values are calculated separately for each category. This approach facilitates a comprehensive analysis of the model's prediction performance across different categories.

Fig. 4A−C illustrates the 20 most influential features in the ESL prediction process, providing crucial insights into the key variables that affect prediction outcomes. Each subgraph corresponds to one of the three categories in the triple classification. Upon examining these subgraphs, it becomes evident that BCUT2D_LOGPHI is the most significant factor in the ESL prediction process.

Fig. 4D−F presents the SHAP summary plot that offers a comprehensive overview of how each feature contributes to the model prediction. Features are ranked on the vertical axis based on the sum of their SHAP values across all samples, reflecting their importance in prediction. The horizontal axis represents the SHAP value, illustrating how each feature influences the model prediction. Points farther away from the center line (zero) indicate a greater impact on the model prediction. Positive SHAP values signify a positive contribution, while negative values suggest a negative impact. The graph displays the SHAP value for each sample, with the varying color shade indicating the magnitude of the feature's value. This allows for a more precise interpretation of the influence exerted by the feature.

From a macroscopic perspective, we can identify several key features, such as BCUT2D_LOGPHI, PEOE_VSA8, and EState_VSA4. Therefore, we can understand the specific mechanisms by which these features operate in various scenarios. For example, the distribution of BCUT2D_LOGPHI values plays a crucial role in model predictions, as illustrated in Fig. 4D. Notably, the blue points cluster predominantly on the right side of the axis. In contrast, the red points are concentrated on the left side. Simultaneously, it is observed that the red points with brighter colors are distributed further away from the axis. This distribution reveals that elevated BCUT2D_LOGPHI values significantly hinder the model's prediction performance when classifying samples as class A. That is, a higher BCUT2D_LOGPHI value suggests a reduced probability of the sample being classified as class A. Conversely, as depicted in Fig. 4F, an inverse distribution is observed, with blue dots primarily on the left side and red dots on the right side. Here, high values of BCUT2D_LOGPHI positively impact the model's predictions when categorizing samples as class C. This reversal emphasizes the pivotal role of BCUT2D_LOGPHI in determining whether a sample is classified as class A or C, with its influence manifesting distinct effects across various prediction scenarios.

Through SHAP analysis, we can understand the predictive behavior of the model at both macro and micro levels, enabling us to gain insights into how individual features contribute to predicting a specific category for a single sample. The LNPs' chemical structure of the selected test sample, representing a class B drug expression efficacy, is depicted in Fig. 4G. Microscopic interpretation for this sample is achieved *via* the SHAP force plot (Fig. 4H−J). The SHAP force diagram includes a "base_value" reflecting the model's default prediction for the category in the absence of feature information. By mapping SHAP values to a probability space, this base value approximates the category's relative frequency in the test set. Conversely, "f(x)" denotes the model's predicted output, calculated as the sum of the benchmark value and SHAP contributions from all features. In the probability space, "f(x)" signifies the model's probability assessment for the sample's category membership. The force diagram's horizontal bars are color-coded to show features' positive (red) or negative (blue) influences on the model's prediction. Longer bar lengths indicate greater feature impact on predictions. For instance, Fig. 4H highlights BCUT2D_MRLOW's negative influence on prediction, with the most extended bar length among features, making it pivotal in class A determinations. Comparing the force diagrams reveals Fig. 4I with the largest "f(x)" output of 0.82, indicating an 82% likelihood that the sample belongs to class B. This phenomenon confirms the model's accuracy and the effectiveness of the SHAP analysis. By pinpointing influential features and understanding their roles across categories, this analysis offers critical insights for model refinement and interpretation.

Through a systematic investigation that integrates SHAP interpretable machine learning with domain-specific structural insights, this study elucidates how the top five determinants across SHAP hierarchies (classes A−C) govern the biodistribution patterns of ionizable lipids. These computational-structural synergies reveal critical structure-distribution relationships that dictate *in vivo* fate.

Class A lipids exhibit optimized delivery performance due to their synergistic molecular features. The high-charge surface area represented by PEOE_VSA8 indicates that the molecule has both a large surface area available for protonation and a low-charge state group represented by MinEStateIndex, indicating that the molecule has specific functional groups that are easily protonated. The combination of these two descriptors may facilitate the molecule's easy protonation in the acidic endosomal environment, resulting in a positively charged surface and promoting fusion with the negatively charged endosomal membrane. BCUT2D_LOGPHI ensures the stability of the hydrophobic core, and the hydrophobic region also enhances the ability of the molecule to cross the biological membrane, thereby improving the delivery efficiency. The balance between the medium polar region represented by EState_VSA4 and the low polar region represented by BCUT2D_MRLOW maintains the molecule's stability in blood circulation. This balance may help avoid excessive aggregation or degradation during circulation, while maintaining the ability of the molecule to undergo structural changes under low pH conditions and release the loaded mRNA.

Class B lipids demonstrate constrained protonation behavior governed by Balaban-J-quantified topological complexity from intricate ring systems or branched architectures, which impart structural rigidity that reduces conformational adaptability in

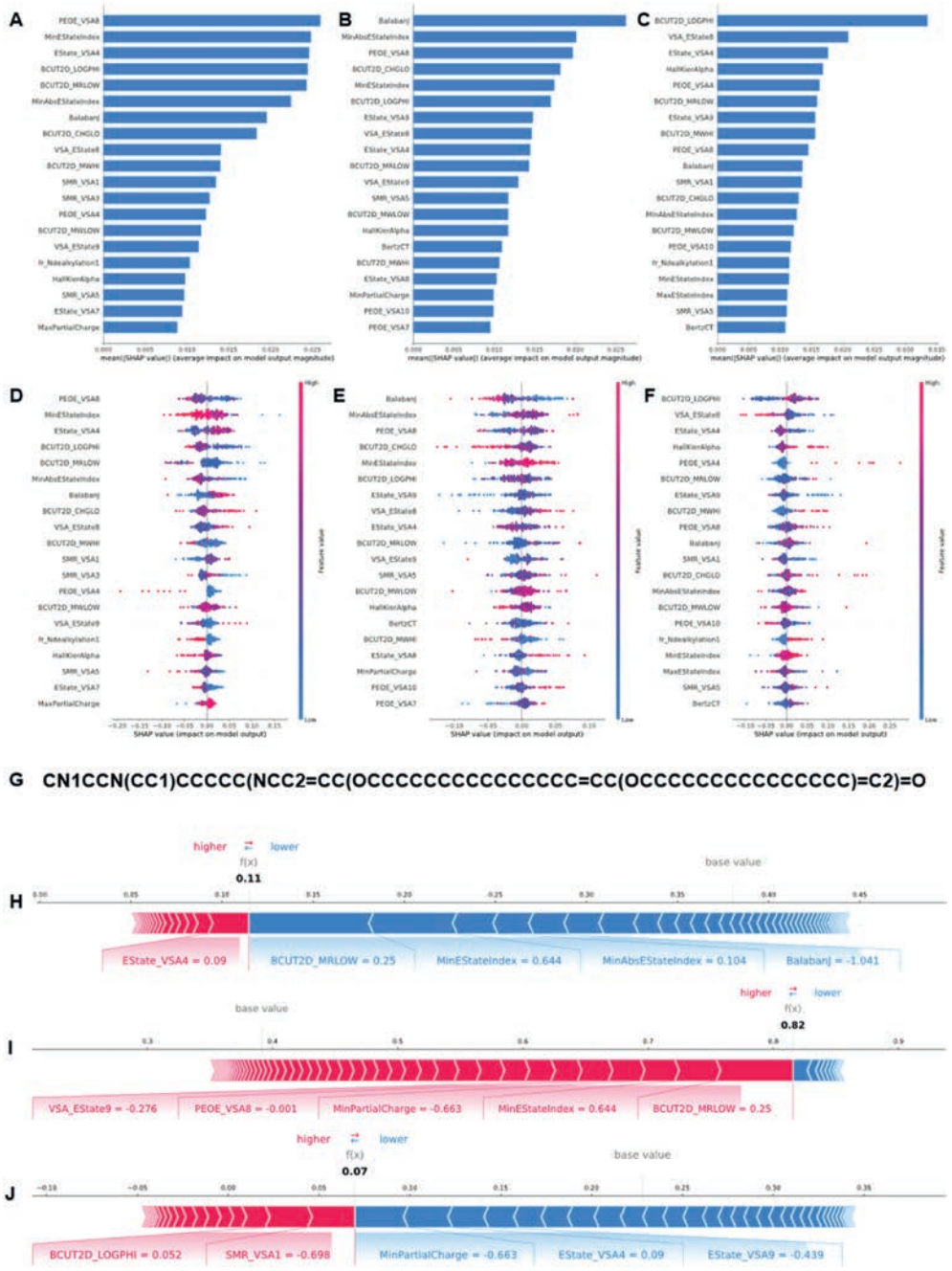


Figure 4 SHAP analysis for the ESL model trained using the 74-subset. (A−C) Analysis of the feature importance plot. (D−F) Analysis of the summary plot. (G) The SMILES code for the initial sample in the test set corresponds to an expression rating of B. (H−J) Analysis of the force plot.

endosomal environments and delays pH-responsive charge conversion. BCUT2D_CHGLO represents regions of low intramolecular charge density, typically hydrophobic or non-polar, while PEOE_VSA8 quantifies regions of high surface charge density, which are polar or ionizable. This combination suggests an uneven charge distribution across the molecule, which may limit the uniformity of protonation. While MinEStateIndex identifies low-electronegativity groups with protonation capability, their limited ionization potential (indicated by MinAbsEStateIndex) and steric occlusion from topological complexity collectively restrict proton accessibility, resulting in suboptimal charge uniformity across the molecular framework. While class B demonstrates equivalent PEOE_VSA8 and MinEStateIndex values to class A, its reduced parameter significance, combined with BCUT2D_CHGLO-associated charge heterogeneity, leads to suboptimal efficacy relative to class A.

Class C lipids face delivery limitations primarily arising from VSA_EState8-quantified high-electronegativity surfaces, which are typically polar or strongly electron-withdrawing. This leads to intense dipole moments and hydrogen bonding capabilities. PEOE_VSA4 represents regions of medium charge density. The combination of high electronegative surface areas and medium

charge regions results in molecules with excessive polarity. This excessive polarity may destabilize the positive charge after protonation under acidic conditions, reducing the molecule's ability to be effectively protonated and hindering membrane fusion and mRNA release. This electrostatic imbalance is compounded by HallKierAlpha-characterized topological complexity, where branched/three-dimensional architectures introduce steric congestion that simultaneously impedes membrane engagement and payload accessibility. Additionally, the conflict between the hydrophobicity of BCUT2D_LOGPHI and the medium polarity as represented by EState_VSA4 prevents the molecule from effectively coordinating the dynamic switching between the hydrophobic core and the polar surface. Together, these factors reduce the molecule's stability, responsiveness, and overall delivery efficiency.

Based on the chemical structural characteristics of ionizable lipids reported in the literature, several key findings have emerged, highlighting the importance of specific structural elements in optimizing delivery efficiency. The squaramide group at the head of the lipid molecule has been shown to interact specifically with mRNA, resulting in the formation of particles with enhanced structural stability and improved expression efficiency[66]. Additionally, the use of amide bonds as connecting units enables selective delivery of mRNA to the lungs of mice, suggesting a potential strategy for tissue-specific targeting[25]. Furthermore, even minor alterations in the tail structure can significantly impact delivery performance, with liposomes containing branched tails demonstrating notably high expression levels *in vivo*[43].

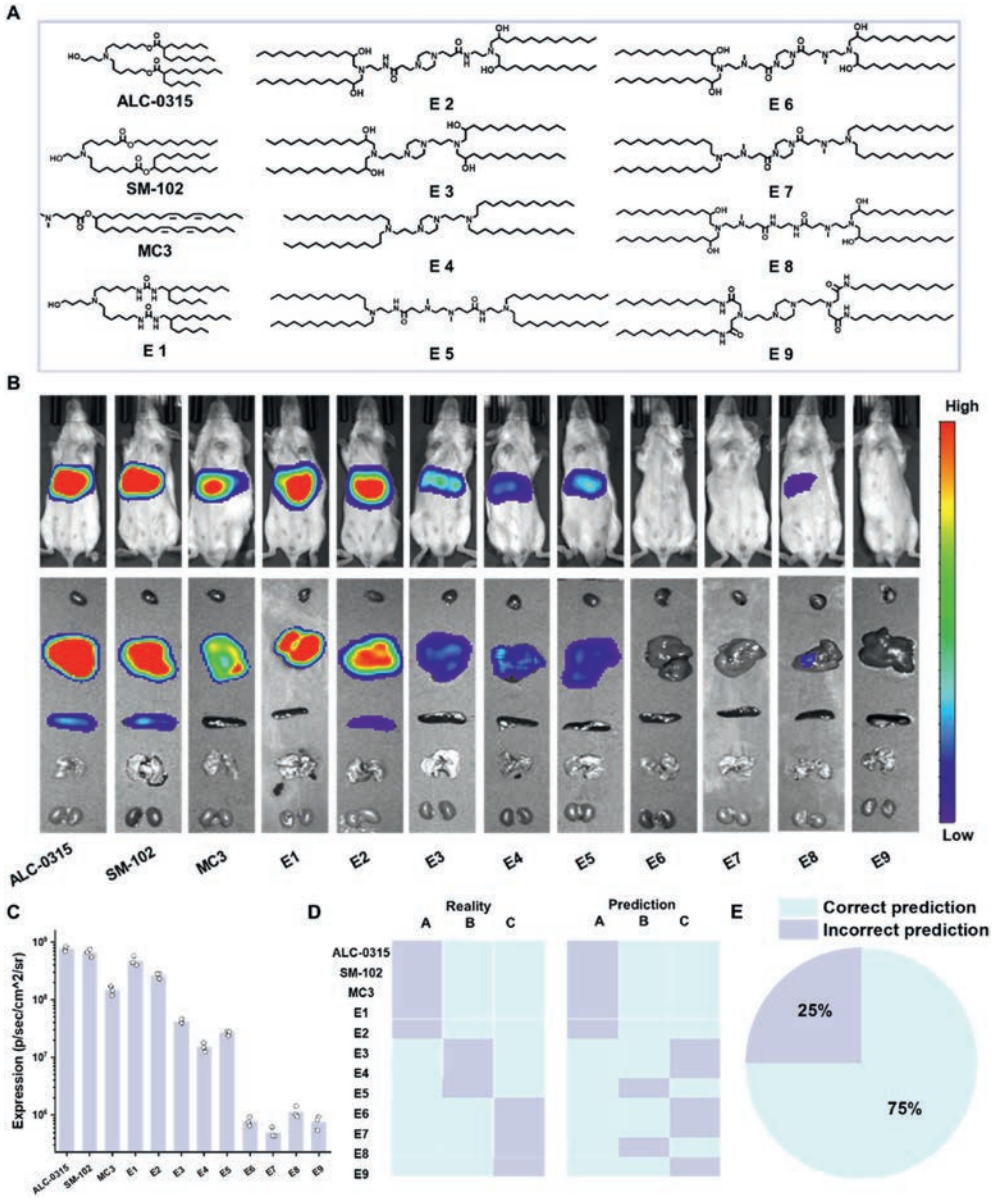


Figure 5 Statistics of structures and validation results of ionizable lipids. (A) Structure of ionizable lipids in intravenously administered LNPs. (B) Quantitative analysis of biodistribution and luciferase expression in mice by IVIS imaging system 6 h after intravenous injection of LNPs. (C) Luciferase expression histogram of ALC-0315, SM-102, MC3, and E1-9. (D) Prediction results and experimental verification of 12 ionizable lipids. (E) Pie chart of prediction results.

Moreover, we also searched the literature that the *in vivo* fate of ionizable lipids may be significantly different in different species, underscoring the importance of species-specific considerations in lipid nanoparticle (LNP) design. Ouyang Defang and colleagues employed advanced modeling techniques, including physiologically based pharmacokinetic (PBPK) and quantum mechanics (QM) approaches, to simulate RNA-LNP delivery in a range of models, from HeLa cells to rats, mice, and humans[86]. Their research revealed that the metabolic fate of ionizable lipids within these systems is primarily governed by the degradation kinetics of the LNP carrier itself, rather than by the hydrolysis of the ionizable lipids.

These comprehensive findings not only deepen our understanding of the structural and mechanistic determinants of ionizable lipid performance but also offer valuable insights for the rational design and optimization of RNA-LNPs. By incorporating these insights into future research and development endeavors, scientists may effectively address current limitations and propel the advancement of nucleic acid-based therapies.

3.4. In vivo biodistribution validation results

Machine learning has emerged as a potent tool for forecasting mRNA expression outcomes, with the majority of models relying on *in vitro* cell-level transfection data for their training datasets[21,22,85,87]. Pioneering researchers like Daniel G. Anderson and Bowen Li have merged the realms of combinatorial chemistry and cellular transfection to devise sophisticated machine learning models, which are poised to inform the design of ionizable lipids[21,19]. Despite the notable strides in advancing the field of ionizable lipids, a critical challenge persists: the discrepancy between *in vitro*-based models and the accuracy of *in vivo* mRNA expression predictions. This inconsistency arises from the pronounced differences in mRNA expression efficiency when transitioning from *in vitro* to *in vivo* conditions. Consequently, the development of a machine learning model capable of accurately predicting *in vivo* mRNA expression effects holds significant practical value.

Twelve ionizable lipids were selected for subsequent verification experiments, and their structures are shown in Fig. 5A. The *in vivo* biodistribution and luciferase expression of LNPs were observed in mice 6 h after tail vein injection (Fig. 5B and C). The predicted results of the 12 ionizable lipids were compared with the experimental results, and the prediction accuracy was intuitively observed from the prediction result pie chart (Fig. 5D and E). In the experimental validation of *in vivo* biodistribution for 12 distinct ionizable lipids, we achieved an overall prediction accuracy of 75 percent. Our results demonstrate a high degree of concordance between the predictive outcomes for class A ionizable lipids and their experimental data. Notably, the validation process involving commercially accessible ionizable lipids has yielded outstanding predictive results, with particular emphasis on ALC-0315, SM-102, and MC3. This preliminary observation validates the model's strong predictive power, particularly for lipids with superior performance. Conversely, the predictive accuracy for class B ionizable lipids is subpar, with only one-third of the predictions being correct. The consistent misclassification from class B lipids to class C highlights a persistent challenge in the model's ability to discern subtle structural differences. To elucidate this phenomenon, we conducted a decorrelation analysis from the perspective of data distribution.

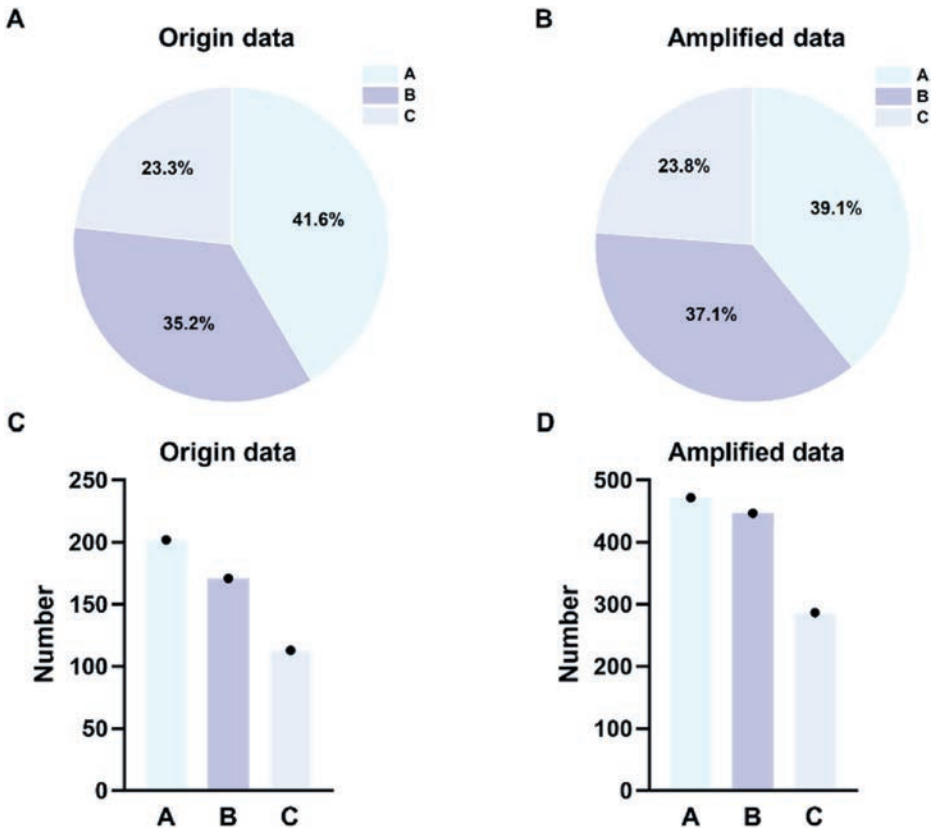


Figure 6 Distribution and data volume of origin and amplified data in the LipidDB database. (A) Distribution of origin data in LipidDB. (B) Distribution of amplified data in LipidDB. (C) Data volume of LipidDB original data. (D) Volume of amplified data in LipidDB.

After data amplification, the dataset comprised 472 class A samples, 447 class B samples, and 287 class C samples, as depicted in Fig. 6A−D. Since class B samples were not the least numerous, their quantity did not negatively impact the overall prediction performance. However, upon examining the features specific to class B, we discovered that the coexistence of low-charge and high-charge regions within class B structures might lead to an uneven distribution of charges and lower protonation efficiency in certain areas. These characteristics likely contributed to the challenges in accurately predicting the behavior of class B ionizable lipids. However, the absence of misclassifications as class A is encouraging, as it suggests that the model can effectively avoid misidentifying high-performing lipids. In the predictive analysis of class C ionizable lipids, the model achieves an excellent prediction accuracy, effectively distinguishing between lipids with less favorable performance characteristics. This result highlights the model's capacity to identify lipids even under suboptimal conditions accurately. The ESL model can accurately classify these lipids, which further confirms the robust prediction capability of the ESL model.

In summary, the model has a robust prediction capacity for the lipid structures of classes A and C. However, when predicting class B structures, there is potential room for optimizing the ESL model. Interestingly, the model can accurately identify the lipid samples of class A; Therefore, potentially high-performing ionizable lipids are not overlooked. These observations highlight the model's potential in guiding the selection and development of ionizable lipids with targeted performance, thereby enhancing its applicability for both academic research and industrial applications.

3.5. Code availability

The codebase data and code used in this study are available in a Github repository (https://github.com/yuhanchenzz/LipidAI) under the MIT License.

4. Conclusions

In this study, we introduced LipidAI, a novel framework that synergistically combined two pivotal components: MTA and ESL algorithms. This integration aimed at elucidating the correlation between the chemical structures of ionizable lipids and the efficiency of mRNA expression *in vivo*. MTA enhanced the predictive accuracy of LipidAI by nearly tripling the dataset size through meticulous modifications to the methyl groups on lipid tail chains. This approach effectively mitigated the previously encountered scarcity of data within ionizable lipid libraries, which was a limiting factor in prior studies. Concurrently, the ESL algorithm leveraged the strengths of multiple learning algorithms, outperforming the predictive accuracy of any single algorithm utilized in previous research. This achievement was underscored by an AUC score of 0.91, thereby establishing a new standard for accuracy within this domain. Furthermore, the efficacy of LipidAI in practical settings was substantiated through biological experiments, which demonstrated a prediction accuracy of up to 75% for ionizable lipids. To enhance the transparency and interpretability of our machine learning methodology, we employed SHAP analysis. This analysis delineated the individual contributions of features to model predictions, enabling us to identify the lipid properties that exerted the most significant influence on mRNA expression efficacy. These insights offered valuable guidance for the rational design of lipids. In summary, the ESL algorithm developed in this study exhibited exceptionally robust predictive performance. It had the potential to significantly reduce the costs, time, and material consumption associated with ionizable lipid research and development. Moreover, it provided a novel technical strategy for the rational design of mRNA delivery vectors, thereby accelerating the advancement of mRNA delivery system research.

Acknowledgments

The work was financially supported by the National Key Research and Development Program of China (Nos. 2023YFC3403200 and 2024YFF1503100).

Author contributions

Xing Duan, Shanhui Jiang and Xiangrong Song designed the research. Jiezhou Chen, Linbo Qing and Pingyu Wang provided the code. Xing Duan, Jiezhou Chen, Shanhui Jiang and Linbo Qing carried out the experiments and performed data analysis. Haixing Shi, Tingting Song, Xiangyu Jiao, Guohong Li, Xi He, Jun He and Yinjie Lei participated part of the experiments. Xiangrong Song, Hai Huang, Changchun Zhao and Yongjun Gu provided experimental drugs and quality control. Xing Duan, Jiezhou Chen, Shanhui Jiang and Linbo Qing wrote the manuscript. Xiangrong Song, Pingyu Wang, Xi He, Shengbin Liu, Xiaoling Yi and Mengran Guo revised the manuscript. All of the authors have read and approved the final manuscript.

Conflicts of interest

The authors declare no conflict of interest.

References

1. Miao L, Zhang Y, Huang L. mRNA vaccine for cancer immunotherapy. *Mol Cancer* 2021;**20**:41.
2. Weng Y, Huang Y. Advances of mRNA vaccines for COVID-19: a new prophylactic revolution begins. *Asian J Pharm Sci* 2021;**16**:263−4.
3. Zhu Y, Li J, Pang Z. Recent insights for the emerging COVID-19: drug discovery, therapeutic options and vaccine development. *Asian J Pharm Sci* 2021;**16**:4−23.
4. Guo J, Lu Z, Zhao R, Li B, Zhang X. Nucleic acid therapy for metabolic-related diseases. *Chin Chem Lett* 2024:109875.
5. Pardi N, Hogan MJ, Porter FW, Weissman D. mRNA vaccines—a new era in vaccinology. *Nat Rev Drug Discov* 2018;**17**:261−79.
6. Sahin U, Karikó K, Türeci Ö. mRNA-based therapeutics—developing a new class of drugs. *Nat Rev Drug Discov* 2014;**13**:759−80.
7. Wang Y, Zhang Z, Luo J, Han X, Wei Y, Wei X. mRNA vaccine: a potential therapeutic strategy. *Mol Cancer* 2021;**20**:33.
8. He Q, Gao H, Tan D, Zhang H, Wang J. mRNA cancer vaccines: advances, trends and challenges. *Acta Pharm Sin B* 2022;**12**:2969−89.
9. Riley RS, Kashyap MV, Billingsley MM, White B, Alameh M-G, Bose SK, et al. Ionizable lipid nanoparticles for in utero mRNA delivery. *Sci Adv* 2021;**7**:eaba1028.
10. Billingsley MM, Singh N, Ravikumar P, Zhang R, June CH, Mitchell MJ. Ionizable lipid nanoparticle-mediated mRNA delivery for human CAR T cell engineering. *Nano Lett* 2020;**20**:1578−89.
11. Li Y, Ye Z, Yang H, Xu Q. Tailoring combinatorial lipid nanoparticles for intracellular delivery of nucleic acids, proteins, and drugs. *Acta Pharm Sin B* 2022;**12**:2624−39.
12. Wang Q, Jiang Q, Li D, Yang Z, Gao L, Liu F, et al. Elaborately engineering of lipid nanoparticle for targeting delivery of siRNA and suppressing acute liver injury. *Chin Chem Lett* 2024;**35**:108683.

13. Karoui KE, Vriese ASD. COVID-19 in dialysis: clinical impact, immune response, prevention, and treatment. *Kidney Int* 2022;**101**: 883−94.
14. Alexander JL, Moran GW, Gaya DR, Raine T, Hart A, Kennedy NA, et al. SARS-CoV-2 vaccination for patients with inflammatory bowel disease: a British Society of Gastroenterology Inflammatory Bowel Disease Section and IBD Clinical Research Group Position Statement. *Lancet Gastroenterol Hepatol* 2021;**6**:218−24.
15. Toyama K, Eto T, Takazawa K, Shimizu S, Nakayama T, Furihata K, et al. DS-5670a, a novel mRNA-encapsulated lipid nanoparticle vaccine against severe acute respiratory syndrome coronavirus 2: results from a phase 2 clinical study. *Vaccine* 2023;**41**:5525−34.
16. Xu K, Lei W, Kang B, Yang H, Wang Y, Lu Y, et al. A novel mRNA vaccine, SYS6006, against SARS-CoV-2. *Front Immunol* 2023;**13**: 1051576.
17. Wilson E, Goswami J, Baqui AH, Doreski PA, Perez-Marc G, Zaman K, et al. Efficacy and safety of an mRNA-based RSV preF vaccine in older adults. *N Engl J Med* 2023;**389**:2233−44.
18. Han X, Zhang H, Butowska K, Swingle KL, Alameh MG, Weissman D, et al. An ionizable lipid toolbox for RNA delivery. *Nat Commun* 2021;**12**:7233.
19. Cheng Q, Wei T, Farbiak L, Johnson LT, Dilliard SA, Siegwart DJ. Selective organ targeting (SORT) nanoparticles for tissue specific mRNA delivery and CRISPR/cas gene editing. *Nat Nanotechnol* 2020; **15**:313−20.
20. Wang W, Feng S, Ye Z, Gao H, Lin J, Ouyang D. Prediction of lipid nanoparticles for mRNA vaccines by the machine learning algorithm. *Acta Pharm Sin B* 2022;**12**:2950−62.
21. Li B, Raji IO, Gordon AGR, Sun L, Raimondo TM, Oladimeji FA, et al. Accelerating ionizable lipid discovery for mRNA delivery using machine learning and combinatorial chemistry. *Nat Mater* 2024;**23**:1002−8.
22. Xu Y, Ma S, Cui H, Chen J, Xu S, Gong F, et al. AGILE platform: a deep learning powered approach to accelerate LNP development for mRNA delivery. *Nat Commun* 2024;**15**:6305.
23. Bae SH, Choi H, Lee J, Kang MH, Ahn SH, Lee YS, et al. Rational design of lipid nanoparticles for enhanced mRNA vaccine delivery *via* machine learning. *Small* 2025;**21**:2405618.
24. Eygeris Y, Gupta M, Kim J, Jozic A, Gautam M, Renner J, et al. Thiophene-based lipids for mRNA delivery to pulmonary and retinal tissues. *Proc Natl Acad Sci U S A* 2024;**121**:e2307813120.
25. Qiu M, Tang Y, Chen J, Muriph R, Ye Z, Huang C, et al. Lung-selective mRNA delivery of synthetic lipid nanoparticles for the treatment of pulmonary lymphangioleiomyomatosis. *Proc Natl Acad Sci* 2022;**119**:e2116271119.
26. Naidu GS, Yong S-B, Ramishetti S, Rampado R, Sharma P, Ezra A, et al. A combinatorial library of lipid nanoparticles for cell type-specific mRNA delivery. *Adv Sci* 2023;**10**:2301929.
27. Sahoo D, Atochina-Vasserman EN, Maurya DS, Arshad M, Chenna SS, Ona N, et al. The constitutional isomerism of one-component ionizable amphiphilic Janus dendrimers orchestrates the total and targeted activities of mRNA delivery. *J Am Chem Soc* 2024; **146**:3627−34.
28. Zhang D, Atochina-Vasserman EN, Maurya DS, Liu M, Xiao Q, Lu J, et al. Targeted delivery of mRNA with one-component ionizable amphiphilic Janus dendrimers. *J Am Chem Soc* 2021;**143**:17975−82.
29. Rhym LH, Manan RS, Koller A, Stephanie G, Anderson DG. Peptide-encoding mRNA barcodes for the high-throughput *in vivo* screening of libraries of lipid nanoparticles for mRNA delivery. *Nat Biomed Eng* 2023;**7**:901−10.
30. Lu J, Atochina-Vasserman EN, Maurya DS, Shalihin MI, Zhang D, Chenna SS, et al. Screening libraries to discover molecular design principles for the targeted delivery of mRNA with one-component ionizable amphiphilic janus dendrimers derived from plant phenolic acids. *Pharmaceutics* 2023;**15**:1572.
31. Kauffman KJ, Dorkin JR, Yang JH, Heartlein MW, DeRosa F, Mir FF, et al. Optimization of lipid nanoparticle formulations for mRNA delivery *in vivo* with fractional factorial and definitive screening designs. *Nano Lett* 2015;**15**:7300−6.
32. Ramishetti S, Hazan-Halevy I, Palakuri R, Chatterjee S, Naidu Gonna S, Dammes N, et al. A combinatorial library of lipid nanoparticles for RNA delivery to leukocytes. *Adv Mater* 2020;**32**:1906128.
33. Hajj KA, Melamed JR, Chaudhary N, Lamson NG, Ball RL, Yerneni SS, et al. A potent branched-tail lipid nanoparticle enables multiplexed mRNA delivery and gene editing *in vivo*. *Nano Lett* 2020; **20**:5167−75.
34. Qiu M, Glass Z, Chen J, Haas M, Jin X, Zhao X, et al. Lipid nanoparticle-mediated codelivery of Cas9 mRNA and single-guide RNA achieves liver-specific *in vivo* genome editing of Angptl3. *Proc Natl Acad Sci* 2021;**118**:e2020401118.
35. Liu J, Chang J, Jiang Y, Meng X, Sun T, Mao L, et al. Fast and efficient CRISPR/Cas9 genome editing *in vivo* enabled by bioreducible lipid and messenger RNA nanoparticles. *Adv Mater* 2019; **31**:e1902575.
36. Tanaka H, Takahashi T, Konishi M, Takata N, Gomi M, Shirane D, et al. Self-degradable lipid-like materials based on "hydrolysis accelerated by the intra-particle enrichment of reactant (HyPER)" for messenger RNA delivery. *Adv Funct Mater* 2020;**30**:1910575.
37. Petersen DMS, Chaudhary N, Arral ML, Weiss RM, Whitehead KA. The mixing method used to formulate lipid nanoparticles affects mRNA delivery efficacy and organ tropism. *Eur J Pharm Biopharm* 2023;**192**:126−35.
38. Yan Y, Liu X, Wang L, Wu C, Shuai Q, Zhang Y, et al. Branched hydrophobic tails in lipid nanoparticles enhance mRNA delivery for cancer immunotherapy. *Biomaterials* 2023;**301**:122279.
39. Kim M, Jeong M, Hur S, Cho Y, Park J, Jung H, et al. Engineered ionizable lipid nanoparticles for targeted delivery of RNA therapeutics into different types of cells in the liver. *Sci Adv* 2021;**7**:eabf4398.
40. Jarzębińska A, Pasewald T, Lambrecht J, Mykhaylyk O, Kümmerling L, Beck P, et al. A single methylene group in oligoalkylamine-based cationic polymers and lipids promotes enhanced mRNA delivery. *Angew Chem Int Ed* 2016;**55**:9591−5.
41. Ding F, Zhang H, Cui J, Li Q, Yang C. Boosting ionizable lipid nanoparticle-mediated *in vivo* mRNA delivery through optimization of lipid amine-head groups. *Biomater Sci* 2021;**9**:7534−46.
42. Lee SM, Cheng Q, Yu X, Liu S, Johnson LT, Siegwart DJ. A systematic study of unsaturation in lipid nanoparticles leads to improved mRNA transfection *in vivo*. *Angew Chem Int Ed* 2021;**60**:5848−53.
43. Hajj KA, Ball RL, Deluty SB, Singh SR, Strelkova D, Knapp CM, et al. Branched-tail lipid nanoparticles potently deliver mRNA *in vivo* due to enhanced ionization at endosomal pH. *Small* 2019;**15**:1805097.
44. Zhao X, Chen J, Qiu M, Li Y, Glass Z, Xu Q. Imidazole-based synthetic lipidoids for *in vivo* mRNA delivery into primary T lymphocytes. *Angew Chem Int Ed* 2020;**59**:20083−9.
45. Fenton OS, Kauffman KJ, Kaczmarek JC, McClellan RL, Jhunjhunwala S, Tibbitt MW, et al. Synthesis and biological evaluation of ionizable lipid materials for the *in vivo* delivery of messenger RNA to B lymphocytes. *Adv Mater* 2017;**29**:1606944.
46. Álvarez-Benedicto E, Farbiak L, Ramírez MM, Wang X, Johnson LT, Mian O, et al. Optimization of phospholipid chemistry for improved lipid nanoparticle (LNP) delivery of messenger RNA (mRNA). *Biomater Sci* 2022;**10**:549−59.
47. Chen Z, Tian Y, Yang J, Wu F, Liu S, Cao W, et al. Modular design of biodegradable ionizable lipids for improved mRNA delivery and precise cancer metastasis delineation *in vivo*. *J Am Chem Soc* 2023; **145**:24302−14.
48. Mrksich K, Padilla MS, Joseph RA, Han EL, Kim D, Palanki R, et al. Influence of ionizable lipid tail length on lipid nanoparticle delivery of mRNA of varying length. *J Biomed Mater Res A* 2024;**112**:1494−505.
49. Han X, Xu J, Xu Y, Alameh M-G, Xue L, Gong N, et al. *In situ* combinatorial synthesis of degradable branched lipidoids for systemic delivery of mRNA therapeutics and gene editors. *Nat Commun* 2024; **15**:1762.
50. Guimaraes PP, Zhang R, Spektor R, Tan M, Chung A, Billingsley MM, et al. Ionizable lipid nanoparticles encapsulating barcoded mRNA for accelerated *in vivo* delivery screening. *J Controlled Release* 2019;**316**:404−17.

51. Maeta M, Miura N, Tanaka H, Nakamura T, Kawanishi R, Nishikawa Y, et al. Vitamin E scaffolds of pH-responsive lipid nanoparticles as DNA vaccines in cancer and protozoan infection. *Mol Pharm* 2020;**17**:1237−47.
52. Miao L, Li L, Huang Y, Delcassian D, Chahal J, Han J, et al. Delivery of mRNA vaccines with heterocyclic lipids increases anti-tumor efficacy by STING-mediated immune cell activation. *Nat Biotechnol* 2019;**37**:1174−85.
53. Yu X, Liu S, Cheng Q, Wei T, Lee S, Zhang D, et al. Lipid-modified aminoglycosides for mRNA delivery to the liver. *Adv Healthc Mater* 2020;**9**:e1901487.
54. Miao L, Lin J, Huang Y, Li L, Delcassian D, Ge Y, et al. Synergistic lipid compositions for albumin receptor mediated delivery of mRNA to the liver. *Nat Commun* 2020;**11**:2424.
55. Lu J, Jiang T, Li X, Tan S, Zhang Y, Gu H, et al. Dual ethanolamine head groups in ionizable lipids facilitate phospholipid-free stable nanoparticle formulation for augmented and safer mRNA delivery 2023 2023;**10.13**:562139.
56. Lam K, Leung A, Martin A, Wood M, Schreiner P, Palmer L, et al. Unsaturated, trialkyl ionizable lipids are versatile lipid-nanoparticle components for therapeutic and vaccine applications. *Adv Mater* 2023;**35**:2209624.
57. Chen J, Ye Z, Huang C, Qiu M, Song D, Li Y, et al. Lipid nanoparticle-mediated lymph node-targeting delivery of mRNA cancer vaccine elicits robust $CD8^+$ T cell response. *Proc Natl Acad Sci* 2022;**119**:e2207841119.
58. Li B, Jiang AY, Raji I, Atyeo C, Raimondo TM, Gordon AGR, et al. Enhancing the immunogenicity of lipid-nanoparticle mRNA vaccines by adjuvanting the ionizable lipid and the mRNA. *Nat Biomed Eng* 2023;**9**:167−84.
59. Han X, Alameh M-G, Butowska K, Knox JJ, Lundgreen K, Ghattas M, et al. Adjuvant lipidoid-substituted lipid nanoparticles augment the immunogenicity of SARS-CoV-2 mRNA vaccines. *Nat Nanotechnol* 2023;**18**:1105−14.
60. Elia U, Ramishetti S, Rosenfeld R, Dammes N, Bar-Haim E, Naidu GS, et al. Design of SARS-CoV-2 hFc-conjugated receptor-binding domain mRNA vaccine delivered *via* lipid nanoparticles. *ACS Nano* 2021;**15**:9627−37.
61. Tilstra G, Couture-Senécal J, Lau YMA, Manning AM, Wong DSM, Janaeska WW, et al. Iterative design of ionizable lipids for intramuscular mRNA delivery. *J Am Chem Soc* 2023;**145**:2294−304.
62. Li Z, Zhang X-Q, Ho W, Li F, Gao M, Bai X, et al. Enzyme-catalyzed one-step synthesis of ionizable cationic lipids for lipid nanoparticle-based mRNA COVID-19 vaccines. *ACS Nano* 2022;**16**:18936−50.
63. Wang J, Zhang Y, Dong S, Zha W, Liu C, Wang Y, et al. Bivalent mRNA vaccines against three SARS-CoV-2 variants mediated by new ionizable lipid nanoparticles. *Int J Pharm* 2023;**642**: 123155.
64. Sabnis S, Kumarasinghe ES, Salerno T, Mihai C, Ketova T, Senn JJ, et al. A novel amino lipid series for mRNA delivery: improved endosomal escape and sustained pharmacology and safety in non-human primates. *Mol Ther* 2018;**26**:1509−19.
65. Xu Y, Hu Y, Xia H, Zhang S, Lei H, Yan B, et al. Delivery of mRNA vaccine with 1, 2-diesters-derived lipids elicits fast liver clearance for safe and effective cancer immunotherapy. *Adv Healthc Mater* 2023;**13**: 2302691.
66. Cornebise M, Narayanan E, Xia Y, Acosta E, Ci L, Koch H, et al. Discovery of a novel amino lipid that improves lipid nanoparticle performance through specific interactions with mRNA. *Adv Funct Mater* 2021;**32**:2106727.
67. Dilliard SA, Sun Y, Brown MO, Sung Y-C, Chatterjee S, Farbiak L, et al. The interplay of quaternary ammonium lipid structure and protein corona on lung-specific mRNA delivery by selective organ targeting (SORT) nanoparticles. *J Controlled Release* 2023;**361**: 361−72.
68. Kim J, Jozic A, Sahay G. Naturally derived membrane lipids impact nanoparticle-based messenger RNA delivery. *Cell Mol Bioeng* 2020; **13**:463−74.
69. Miller JB, Zhang S, Kos P, Xiong H, Zhou K, Perelman SS, et al. Non-viral CRISPR/Cas gene editing *in vitro* and *in vivo* enabled by synthetic nanoparticle co-delivery of Cas9 mRNA and sgRNA. *Angew Chem Int Ed* 2016;**56**:1059−63.
70. Xue L, Gong N, Shepherd SJ, Xiong X, Liao X, Han X, et al. Rational design of bisphosphonate lipid-like materials for mRNA delivery to the bone microenvironment. *J Am Chem Soc* 2022;**144**:9926−37.
71. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 1988;**28**:31−6.
72. Landrum G. *RDKit: a software suite for cheminformatics, computational chemistry, and predictive modeling*. 2013. Available form: https://docplayer.net/11897218-Rdkit-a-software-suite-for-chem-informatics-computational-chemistry-and-predictive-modeling.
73. Zhang D, Atochina-Vasserman EN, Maurya DS, Huang N, Xiao Q, Ona N, et al. One-component multifunctional sequence-defined ionizable amphiphilic janus dendrimer delivery systems for mRNA. *J Am Chem Soc* 2021;**143**:12315−27.
74. Bannigan P, Bao Z, Hickman RJ, Aldeghi M, Häse F, Aspuru-Guzik A, et al. Machine learning models to accelerate the design of polymeric long-acting injectables. *Nat Commun* 2023;**14**:35.
75. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011;**12**:2825−30.
76. Breiman L. Stacked regressions. *Mach Learn* 1996;**24**:49−64.
77. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. In: *Proceedings of the 31st international conference on neural information processing systems*. Red Hook, NY, USA: Curran Associates Inc.; 2017. p. 3149−57.
78. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*; 2016. p. 785−94.
79. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. *J Mach Learn Res* 2012;**13**:281−305.
80. Raschka S. MLxtend: providing machine learning and data science utilities and extensions to Python's scientific computing stack. *J Open Source Softw* 2018;**3**:638.
81. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* 2020;**17**:261−72.
82. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* 2020;**2**:56−67.
83. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan SVN, et al., editors. *Advances in neural information processing systems 30 (NIPS 2017)*. New York: Curran Associates; 2017. p. 4765−74.
84. Ding DY, Zhang Y, Jia Y, Sun J. Machine learning-guided lipid nanoparticle design for mRNA delivery. *ArXiv* 2023. Available from: https://doi.org/10.48550/arXiv.2308.01402.
85. Lewis MM, Beck TJ, Ghosh D. Applying machine learning to identify ionizable lipids for nanoparticle-mediated delivery of mRNA. *BioRxiv* 2023. Available from: https://doi.org/10.1101/2023.11.09.565872.
86. Wang W, Deng S, Lin J, Ouyang D. Modeling on *in vivo* disposition and cellular transportation of RNA lipid nanoparticles *via* quantum mechanics/physiologically-based pharmacokinetic approaches. *Acta Pharm Sin B* 2024;**14**:4591−607.
87. Ding DY, Zhang Y, Jia Y, Sun J. Machine learning-guided lipid nanoparticle design for mRNA delivery. *ArXiv* 2023. Available from: https://doi.org/10.48550/arXiv.2308.01402.