frontiers | Frontiers in Molecular Biosciences

TYPE Original Research
PUBLISHED 21 December 2022
DOI 10.3389/fmolb.2022.1042720

In silico prediction of siRNA ionizable-lipid nanoparticles *In vivo* efficacy: Machine learning modeling based on formulation and molecular descriptors

Abdelkader A. Metwally[1,2]*, Amira A. Nayel[3,4] and Rania M. Hathout[2]

[1]Department of Pharmaceutics, Faculty of Pharmacy, Health Sciences Center, Kuwait University, Kuwait City, Kuwait, [2]Department of Pharmaceutics and Industrial Pharmacy, Faculty of Pharmacy, Ain Shams University, Cairo, Egypt, [3]Clinical Pharmacy Department, Alexandria Ophthalmology Hospital, Alexandria, Egypt, [4]Department of Clinical Pharmacy and Pharmacy Practice, Faculty of Pharmacy, Alexandria University, Alexandria, Egypt

Check for updates

OPEN ACCESS

EDITED BY
Angelina Angelova,
UMR8612 Institut Galien Paris Sud (IGPS), France

REVIEWED BY
Borislav Angelov,
Institute of Physics (ASCR), Czechia
Haiguang Liu,
Microsoft Research Asia (China), China

*CORRESPONDENCE
Abdelkader A. Metwally,
✉ abdelkader.metwally@ku.edu.kw

SPECIALTY SECTION
This article was submitted to Lipids, Membranes and Membranous Organelles,
a section of the journal
Frontiers in Molecular Biosciences

RECEIVED 12 September 2022
ACCEPTED 07 December 2022
PUBLISHED 21 December 2022

CITATION
Metwally AA, Nayel AA and Hathout RM (2022), *In silico* prediction of siRNA ionizable-lipid nanoparticles *In vivo* efficacy: Machine learning modeling based on formulation and molecular descriptors.
Front. Mol. Biosci. 9:1042720.
doi: 10.3389/fmolb.2022.1042720

In silico prediction of the *in vivo* efficacy of siRNA ionizable-lipid nanoparticles is desirable as it can save time and resources dedicated to wet-lab experimentation. This study aims to computationally predict siRNA nanoparticles *in vivo* efficacy. A data set containing 120 entries was prepared by combining molecular descriptors of the ionizable lipids together with two nanoparticles formulation characteristics. Input descriptor combinations were selected by an evolutionary algorithm. Artificial neural networks, support vector machines and partial least squares regression were used for QSAR modeling. Depending on how the data set is split, two training sets and two external validation sets were prepared. Training and validation sets contained 90 and 30 entries respectively. The results showed the successful predictions of validation set log (siRNA dose) with R^2_{val} = 0.86–0.89 and 0.75–80 for validation sets one and two, respectively. Artificial neural networks resulted in the best R^2_{val} for both validation sets. For predictions that have high bias, improvement of R^2_{val} from 0.47 to 0.96 was achieved by selecting the training set lipids lying within the applicability domain. In conclusion, *in vivo* performance of siRNA nanoparticles was successfully predicted by combining cheminformatics with machine learning techniques.

KEYWORDS

siRNA, ionizable lipids, nanoparticles, *in vivo*, QSAR, machine learning

1 Introduction

The process of developing short interfering RNA (siRNA) lipid nanoparticles is lengthy and time consuming because it involves the initial chemical synthesis of a usually large number of ionizable lipids and lipid-like molecules (Jayaraman et al., 2012; Sato et al., 2019; Molla et al., 2020), the formulation of siRNA nanoparticles and the subsequent *in vitro* and *in vivo* evaluation of these nanoparticles, in an attempt to find the best ionizable lipid that is suitable for clinical use in terms of efficacy and safety. Alnylam's small interfering RNA (siRNA) stable nucleic acid lipid nanoparticles, currently marketed as Onpattro™ (Patisiran), obtained FDA approval in 2018. This was followed by FDA approval of Alnylam's Givosiran™ and Lumasiran™ in 2019 and 2020 respectively (Zhang et al., 2021).

Gene silencing by double-stranded RNA (dsRNA) was reported by Fire and Mello in *Caenorhabditis elegans* (Fire et al., 1998) and later siRNA duplexes of length 21-22 nucleotides proved to promote post-transcriptional gene silencing in mammalian cells (Elbashir et al., 2001). Since then, the potential of siRNA as a therapeutic macromolecule against many diseases has been investigated, with more than 40 siRNA based therapies already reaching phases 2, 3 or 4 of clinical trials (Titze-de-Almeida et al., 2017; Dong et al., 2019; ClinicalTrials.gov, 2020). The major barriers against the successful employment of therapeutic siRNA include the lack of stability of the siRNA duplex, the immune response mediated by Toll-like receptors, the rapid renal clearance of naked siRNA, and the difficulty of the intracellular delivery of unmodified siRNA due to its large size and the large number of negative charges on its back-bone (Whitehead et al., 2009; Dowdy, 2017).

One method to overcome the barriers of siRNA delivery is to formulate it as siRNA ionizable lipid nanocomplexes (lipoplexes) or lipidic nanoparticles (Metwally et al., 2012a; Metwally et al., 2012b; Cullis & Hope, 2017; Paunovska et al., 2018). These nanoparticles are multicomponent and may also contain helper lipids, PEG-lipids and phospholipids. An ideal delivery system should ensure response reproducibility, non-immunogenicity, good payload and ease of manufacturing (Cullis & Hope, 2017).

Lipidic nanocarriers for siRNA include liposomes, nanoemulsions, solid lipid nanoparticles, nanostructured lipid carriers, micelles, and liquid crystalline nanoparticles. Since the nature and ratio of ionizable lipids affects the performance of lipid-nucleic acid complexes, the structure of lipid-based self-assembled nucleic acids delivery systems was investigated and was found to tune the supramolecular organization of the complexes (Angelov et al., 2017).

The process of preparing siRNA lipoplexes and nanoparticles involves many steps: the synthesis of the ionizable lipids, their purification and characterization, then the process of preparing the nanoparticles including determining the siRNA to cationic lipid ratio, the cationic lipid to helper lipid (if any) ratio, and nanoparticles characterization in terms of their size, zeta potential, pK_a, stability and *in vivo* evaluation of their safety and silencing efficacy. All of these steps require time and resources and indeed if the *in vivo* efficacy, as measured by either the siRNA dose or knockdown efficiency, could be predicted within reasonable accuracy by using computational means, the process of developing siRNA nanomedicines would be vastly improved in terms of time and costs. Therefore, it is important to attempt to predict the *in vivo* efficacy of siRNA cationic lipid nanoparticles by using machine learning techniques. These techniques can be generally classified into two main groups: supervised and unsupervised learning methods. Supervised learning is used in tasks such as regression and classification, i.e., when there is a dependent variable and one or more independent variables.

In order to extract chemical information from the structures of the molecules under investigation, molecular descriptors, which are important cheminformatics tools, are employed to carry out this task (Hathout et al., 2018; Hathout et al., 2020a). Molecular descriptors are numerical values resulting from either an experimental procedure or from a set of mathematical and/or logical algorithms that are performed on chemical structures (Todeschini & Consonni, 2008). The descriptors can be generally classified as 0D and 1D, when only molecular formula or constitutional properties of a molecule are considered, while 2D descriptors are calculated based on topological properties of a molecule and 3D descriptors depend on geometrical properties of a molecule. Further classifications include 2.5D chiral descriptors and descriptors with more than three dimensions (Consonni & Todeschini, 2010; Valdés-Martiní et al., 2017). Molecular descriptors have been used as predictors of the self-assembly of drug molecules into nanoparticles (Shamay et al., 2018), to model drug binding kinetics (De Benedetti & Fanelli, 2018), in QSAR modeling (Kausar & Falcao, 2018) and in target identification (Reker et al., 2014). Molecular descriptors were also used to successfully predict the binding energy between drug molecules and their nanocarriers and hence predict drug loading onto lipidic and polymeric nanoparticles (Metwally & Hathout, 2015).

Previous QSAR studies on nanoparticles have mostly addressed predicting the cellular uptake and toxicological properties of inorganic nanoparticles, with either unmodified or modified surfaces (Liu et al., 2015; Basant & Gupta, 2017; Wang et al., 2017), however, developing QSAR models for predicting siRNA *in vivo* efficacy has not been achieved before.

In the current work, a data set is prepared using five publications (Jayaraman et al., 2012; Alabi et al., 2013; Kumar et al., 2014; Whitehead et al., 2014; Rajappan et al., 2020). This data set contains the 1D and 2D descriptors of ionizable lipids together with the formulation descriptors (PEG mol%) and the percentage knockdown resulting from a specific siRNA dose. The siRNA nanoparticles *in vivo* efficacy when formulated with these ionizable lipids was included as the response variable; logarithm of the siRNA dose resulting in a specific knockdown percent of the target gene. The data set is split into training and validation sets, where the training set is used to construct the machine learning models, and the validation set is used as an external test set that is used only to evaluate the predictive models constructed by modeling the training set. An

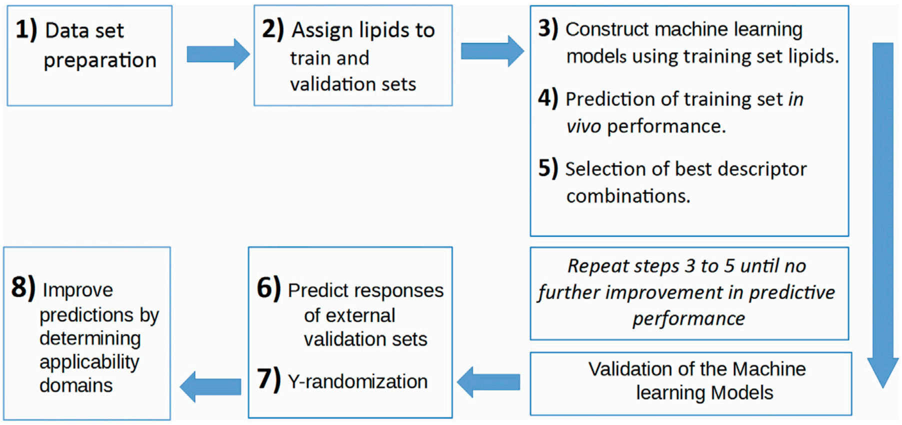
1) Data set preparation
2) Assign lipids to train and validation sets
3) Construct machine learning models using training set lipids.
4) Prediction of training set in vivo performance.
5) Selection of best descriptor combinations.
Repeat steps 3 to 5 until no further improvement in predictive performance
Validation of the Machine learning Models
6) Predict responses of external validation sets
7) Y-randomization
8) Improve predictions by determining applicability domains

FIGURE 1
The workflow of the predictive model building process.

evolutionary algorithm is used to select the best descriptor combinations and is combined with three machine learning techniques; ANN, SVM and PLS regression, to build the predictive models. The performance of the predictive models using the three machine learning techniques and the quality of predictions and how to improve them is presented and discussed. Figure 1 shows the workflow of the modeling and evaluation process. A graphical abstract image is provided in Supplementary Files.

2 Materials and methods

2.1 Data set preparation

2.1.1 Data selection from available literature

For preparing the data set, five publications(Jayaraman et al., 2012; Alabi et al., 2013; Kumar et al., 2014; Whitehead et al., 2014; Rajappan et al., 2020) were retrieved after carrying out online search using PUBMED and Google Scholar servers, where all of them fulfilled the following requirements: siRNA is delivered by means of ionizable lipids, siRNA *in vivo* performance is evaluated *in vivo* against factor FVII expression, all nanoparticles contained the ionizable lipid, DSPC, cholesterol and PEG-lipid (with PEG average molecular weight = 2000), and the PEG-lipid mole % in the formulation is either given or can be calculated. In addition, both the siRNA dose and the percentage knockdown or percentage gene expression resulting from a specific siRNA dose must be provided. Five papers were selected to prepare the data set (Jayaraman et al., 2012; Alabi et al., 2013; Kumar et al., 2014; Whitehead et al., 2014; Rajappan et al., 2020). Wherever the values for the gene expression or dose were not provided numerically, these values were obtained from the relative figures using WebPlotDigitizer v4.2. In case two or more lipids had the same 2D structure, one of them was retained. If an ionizable lipid lacked a well defined *in vivo* efficacy measure, such as a definite dose or knockdown %, it was omitted.

2.1.2 Calculation of the 2D molecular descriptors

The structures of the ionizable lipids were drawn using ACD Chemsketch, and the structures were saved as either individual MDL.mol files or combined together into a single.sdf file using OpenBabel v2.4 (O'Boyle et al., 2011). The following software packages were used for the calculation of the 1D/2D molecular descriptors: Padel Descriptor v2.21 (Yap, 2011), RDKit 2017, and ToMoCoMD QuBiLS-MAS 2020 (Valdés-Martiní et al., 2017). For the calculation of the QuBiLS-MAS descriptors, the following settings were selected: linear algebraic form, atom-based, non-stochastic matrix form, and total groups.

2.1.3 Data set preprocessing

The initial data set containing the descriptors was further processed by removing columns having one or more of either missing or not available (NA) entries. Columns with same-value entries were also removed. If certain columns in the data set showed a high correlation (cutoff r = 0.98) with each other (Racz et al., 2019), all the columns were removed except for one column which has the lowest average correlation with the other descriptor (predictor) columns in the data set. In addition, the formulation descriptor (PEG mol%) and percentage knockdown resulting from a specific siRNA dose were added as predictors. The data set descriptor columns were scaled by calculating the z-scores. The siRNA nanoparticles *in vivo* efficacy was included as the response variable; logarithm of the dose resulting in a specific knockdown percent.

2.2 Principal component analysis of data set

PCA of the scaled data set predictor columns (without response columns) was carried out using ChemometricsWithR package through R software v3.5.

2.3 Splitting the data set into training and validation sets

For modeling purposes, the data set entries were split into a training set (75% of entries) and a validation set (25% of entries). This process was carried out two times separately on the data set where the validation set entries (or observations) were selected either by random selection or by selecting sequentially every fourth entry in the set, with the remainder of the entries in the data set taken as the training set.

2.4 Machine learning models

The modeling process was carried out using either R software version 3.5 or Microsoft Open R v3.5. The following R packages were used for all modeling methods: caret (Deist et al., 2018) and Metrics (Hamner et al., 2018).

Artificial neural networks (ANNs) are a collection of linear and non-linear functions that map an input to an output. These functions can approximate a non-linear complex function. The idea behind the inner working of ANNs is that input data (**x**) are scaled and combined in a linear manner in the form of **Wx** + b, where W is the weights matrix and b is bias, and then the output of this linear combination is fed into a non-linear function (called activation function), the output of which could be used as an input to the next layer and/or to a final output layer (Wesolowski & Suchacz, 2019). For ANN modeling, nnet package was used. The hyperparameters were one hidden layer, two nodes and a weight decay of 0.1 for training and 0.001 for final validation set predictions.

Support vector machines (SVM) are a supervised machine learning technique. For classification, SVM aims to find a hyperplane (decision surface) that can separate two classes of observations with a maximum margin of separation (Maltarollo et al., 2019). Similarly, SVM regression follows the same logic of finding a hyperplane, but with a fixed margin width, epsilon (ε), within which the prediction error is considered zero, and the hyperplane found should minimize the sum of squared error. To enable the formulation of non-linear decision surfaces, a kernel function is applied. The general form of the kernel functions is K (**x1**,**x2**) = <φ(**x1**),φ(**x2**)>, where **x1** and **x2** are two data points. The kernel function thus avoids the actual calculation of the function φ (Heikamp & Bajorath, 2014). SVM regression modeling (epsilon-regression) was carried out using kernlab package (Karatzoglou et al., 2004), with epsilon value of 0.1 and the kernel chosen to be the Gaussian radial basis function kernel defined as $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x_i}) = -\boldsymbol{\sigma}||\boldsymbol{x} - \boldsymbol{x_i}||^2$, where σ is the inverse width parameter and is determined by the package's sigest function.

Partial least squares (PLS) regression is another supervised learning technique (Hathout et al., 2020b). PLS combines dimensionality reduction of the data with a regression model. PLS formulation of the latent variables (scores or components) is carried out with the aim of maximizing the covariance of the components with the response variable, which differentiates PLS from regular principal component analysis (PCA) (Boulesteix & Strimmer, 2007). The response variable in PLS may be univariate or multivariate. For the prediction of a new data point response $\hat{y}_o^{'}$ from a predictor point $x_o^{'}$, the following equation applies: $\hat{y}_o^{'} = (\frac{1}{n})\sum_{i=1}^{n} y_i^{'} + \boldsymbol{B}^T (x_o - (\frac{1}{n})\sum_{i=1}^{n} x_i^{'})$. B is the matrix of regression coefficients, and is defined as: $\mathbf{B} = \mathbf{W}\ (\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\mathbf{T}^{\mathrm{T}}\mathbf{Y}$, where **W** is the matrix of weights and **T** = **XW** (Boulesteix & Strimmer, 2007). PLS modeling was carried out using pls package (Mevik & Wehrens, 2007) with the number of principal components covering 98% of the variance.

2.5 Selection of the molecular descriptors by the evolutionary algorithm

An evolutionary algorithm was written as an R script to select the best descriptors for model building. 400 initial parent combinations of descriptors were randomly selected, and then each one of them was used as an input to construct the machine learning models that are used to predict the training set log (dose) values and their associated RMSEs (training RMSE).

The training RMSE is calculated as follows: the training set is split into three folds, two folds are used to construct the machine learning model, and the third fold is used as a test set to calculate training RMSE. After evaluating the training RMSE for all predictor combinations, the best combinations are kept as parents and are used to construct offspring combinations. The process is repeated until no further improvement in training RMSE for this specific test fold. The whole selection process is repeated for each of the remaining two test folds. The parameters for the evolutionary algorithm are as follows: population size 400, 25% elitism, 20% mutation, number of generations 10-20 and multipoint cross-over.

RMSE is calculated as: $\mathbf{RMSE} = \sqrt{(\frac{\sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{P_i} - \mathbf{A_i})^2}{\mathbf{n}})}$

Bias is calculated as: $\mathbf{Bias} = \mathbf{P_i} - \mathbf{A_i}$where $\mathbf{P_i}$ and $\mathbf{A_i}$ are the predicted and actual log (dose) values of observation (lipid or entry) *i* respectively, and *n* is the number of observations.

2.6 Ensemble learning by averaging of the validation set predictions

The best descriptor combinations that result in the lowest training RMSE were used as inputs for the machine learning

TABLE 1 Summary of data set. The entries represent either distinct lipids or the same lipid but with different PEG mol% and/or knockdown %.

Index of entries	Number of entries per study	Reference
1-30	30	Rajappan et al. (2020)
31-62	32	Alabi et al. (2013)
63-95, 105	34	Jayaraman et al. (2012)
96-104	9	Kumar et al. (2014)
106-120	15	Whitehead et al. (2014)

modeling algorithm that was used in the training; either ANN, SVM or PLS regression. The central tendency of the validation set predictions were calculated as median of these values for each validation set lipid. The validation set RMSE ($RMSE_{val}$) and coefficient of determination (R^2_{val}) were calculated using these median values. The R^2_{val} is calculated as:

$$\mathbf{R^2_{val}} = \frac{\left(\sum_{i=1}^{n}(\mathbf{x_i}-\bar{\mathbf{x}})(\mathbf{y_i}-\bar{\mathbf{y}})\right)^2}{\sum_{i=1}^{n}(\mathbf{x_i}-\bar{\mathbf{x}})^2\sum_{i=1}^{n}(\mathbf{y_i}-\bar{\mathbf{y}})^2} \tag{1}$$

where x_i and y_i are the i^{th} predicted (the median value) and actual responses respectively, $\bar{x}$ and $\bar{y}$ are the mean values of predicted and actual responses respectively.

2.7 Y-randomization of data set

To evaluate the validity of the resulting descriptor combinations, and the possibility that the obtained validation set predictions might be due to random chance, a Y-randomization of the training data set was carried out by randomizing the training set responses (Žuvela et al., 2015). The predictive models were then constructed by using these randomized responses for model training and subsequent validation as described in Section 2.6.

3 Results

3.1 Data set preprocessing and preparation

The number of observations included in the data set after omitting the lipids or entries that fit the omitting criteria explained in section 2.1.1 was 120 entries (rows). The resulting data set contained 438 predictor columns: 436 columns of molecular descriptors, and 2 columns for PEG mol% and knockdown %. In addition, one response column was included; logarithm of siRNA dose that results in a specific knockdown of the target gene. Table 1 provides summary of the data set.

3.2 Splitting the data set into training and validation sets

Two different methods were used to select the validation set entries, with the remainder of the entries in each splitting method being used for training the machine learning models. These selection processes resulted in the following data sets: training set1, validation set 1, training set 2 and validation set 2. These sets are shown in Table 2. Each training and validation set contained 90 and 30 entries, respectively.

PCA is a dimensionality reduction method that transforms dataset features into a smaller number of new features called principal components. PCA scores are the weighted sums of the original features, and they represent the variance in the observations and can be used to detect similarities or dissimilarities among these observations.

The PCA score plots are shown in Figure 2. Principal components 1, 2 and 3 (PC 1, PC 2 and PC 3) contributed to 22%, 19%, and 12% of the total variance, respectively. When points are near each other, this means that they represent observations that share similarities. The observations of validation set 1 and 2, shown as colored triangles, show uniform spread among those of training set 1 and 2 respectively, which infers that the training sets reasonably represent the characteristics of the validation sets.

3.3 Selection of the molecular descriptors by the evolutionary algorithm

When constructing the descriptor combinations to be used as inputs for the machine learning algorithm, the PEG mol% and the knockdown % were always included in the combinations. Any additional molecular descriptors were added and selected by the evolutionary algorithm. Figure 3 shows the top six molecular descriptors with the highest frequencies of appearance in the descriptor combinations that are selected by the evolutionary algorithm. For each machine learning method, ANN, SVM or PLS, the descriptor with highest frequency was considered 100% and the other descriptors frequencies were calculated relative to it. It is evident that each machine learning model resulted in

TABLE 2 Training and validation sets 1 and 2.

Set	Training entries index	Validation entries index
1	3–6, 8–11, 13, 14, 18, 20, 21, 24–30, 33–37, 39, 43–49, 51, 54–60, 62–64, 66–69, 71–73, 75, 78–80, 82–84, 86, 89, 90, 92-108, 110–115, 117–120	1, 2, 7, 12, 15–17, 19, 22, 23, 31, 32, 38, 44, 50, 52, 53, 61, 65, 70, 74, 76, 77, 81, 85, 87, 88, 91, 109, 116
2	1–3, 5–7, 9–11, 13–15, 17-19, 21–23, 25–27, 29–31, 33–35, 37–39, 41–43, 45–47, 49–51, 53–55, 57–59, 61–63, 65–67, 69–71, 73–75, 77–79, 81–83, 85–87, 89–91, 93–95, 97–99, 101–103, 105–107, 109–111, 113–115, 117–119	4, 8, 12, 16, 20, 24, 28, 32, 36, 40, 44, 48, 52, 56, 60, 64, 68, 72, 76, 80, 84, 88, 92, 96, 100, 104, 108, 112, 116, 120

different top descriptors. It is also clear that the training sets one and two resulted in different top descriptors for the same machine learning method. The only common descriptors, taking the two training sets and the three machine learning methods in consideration, were PEOE_VSA9, GATS3m, and GATS8p. PEOE_VSA9 is a Van der Waals surface area descriptor that describes atomic partial charges. GATS3m and GATS8p are Geary autocorrelation - lag three weighted by atomic masses and Geary autocorrelation - lag 8 weighted by atomic polarizabilities respectively. It should be noted that these descriptors are present in combinations of descriptors (predictors) including the PEG mol% and the knockdown %, thus, their direct influence on the *in vivo* performance of the ionizable lipids should be limited to this context.

The number of molecular descriptors in each descriptor combination as selected by the evolutionary algorithm is listed in Table 3. It is to be noted that these molecular descriptors are present in each combination in addition to both PEG mol% and knockdown %, with the latter two being present in each predictor combination. It was noticed that there were repeated combinations in the final selected combinations, as omission of descriptors by the evolutionary algorithm results eventually in offspring combinations of the same descriptors.

The improvement in predictions of the validation set responses at the end of the evolutionary algorithm is shown in Table 4. The $RMSE_{val}$ in the table are calculated as the first quartile of the RMSE of predictions using the initial 400 descriptor combinations and the final 400 descriptor combinations at the end of the evolutionary algorithm iterations. It is clear that there were improvement in the quality of individual predictions for both validation sets and for all methods as evident by the decrease in the $RMSE_{val}$.

The predictive performance of the machine learning models was evaluated by predicting the validation sets responses. The validation sets were neither used in the selection of best descriptor combinations by the evolutionary algorithm nor they were used in the training of the predictive models, thus, the validation sets represent external unkown test samples for the machine learning models. Using the descriptor combinations selected by the evolutionary algorithm, the median (averaged) predictions of the validation sets one and two resulted in R^2_{val} of 0.72–0.89 and $RMSE_{val}$ of 0.23–0.36 (Table 5). The machine learning method used to predict the validation set responses had a strong effect on the predictive performance, with the ANN predictions resulting in the highest R^2_{val} of 0.89 and 0.80 for validation sets one and two respectively. Similarly, ANN resulted in the lowest $RMSE_{val}$ of 0.23 and 0.30 for validation sets one and two respectively. There were also a difference in the predictive performance between validation sets one and two (Table 5), which reflects the effect of both the training set and validation sets compositions. Supplementary Figure S1 shows the structure of a model ANN, with one input layer, 2 nodes in the hidden layer, and one outcome node. The weights are also provided. To investigate if the ANN will perform better even if a different random sampling of training/validation sets was carried out, a third set (set 3) where validation lipids were selected randomly was prepared (Supplementary Table S1). The predictive performance of this set is presented in Supplementary Table S2 where the $RMSE_{val}$ and R^2 of ANN were better than those of SVM and PLS. Taken together, sequential sampling of validation lipids (set 2) as well as random sampling (set 1 and set 3) showed better performance for ANN.

3.4 Evaluation of predictive performance by predicting validation set responses

Figure 4 shows that the three machine learning methods resulted in good validation sets predcitions, as evident from the predicted points being close to the straight lines (shown in red and representing perfect correlation) in the actual vs predicted plots. It is also clear that the different machine learning models were capable of differentiating between the lipids (entries) with low log (dose), which are the desirable lipids (or formulations), and the lipids/formulations with higher doses.

The curated scaled data set together with an example of the resulting predictor combinations (training set 1) after selection by the evolutionary algorithm and ANN is provided as Supplementary Materials. An R script for calculating the median predictions of validation set 1 and the assocciated R^2_{val} and $RMSE_{val}$ using the data set and the descriptor combinations is also provided as Supplementary Material.

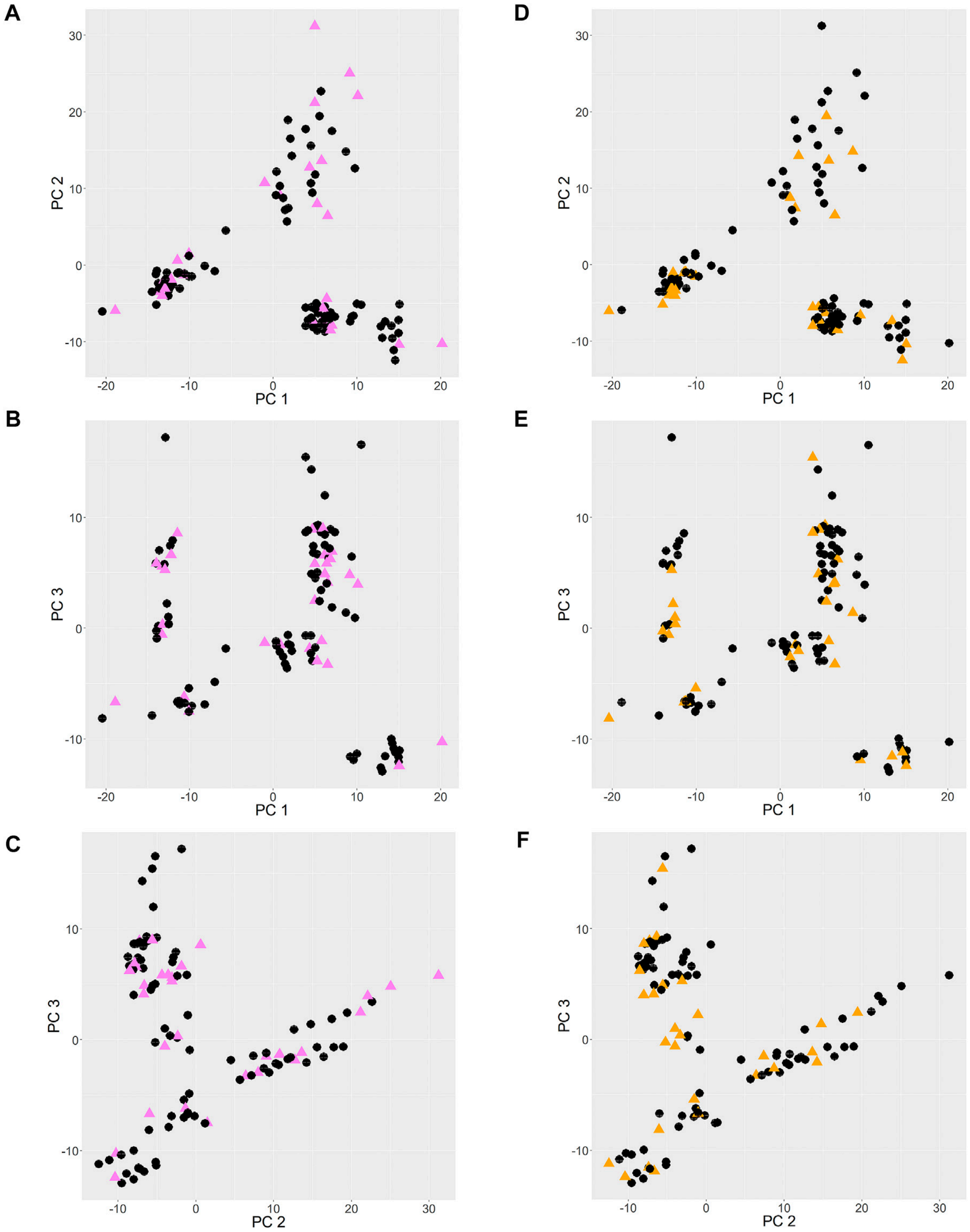

FIGURE 2
PCA score plots. **(A–C)**: training and validation set 1 entries are shown as black circles and pink triangles respectively. **(D–F)**: training and validation set 2 entries are shown as black circles and orange triangles respectively.

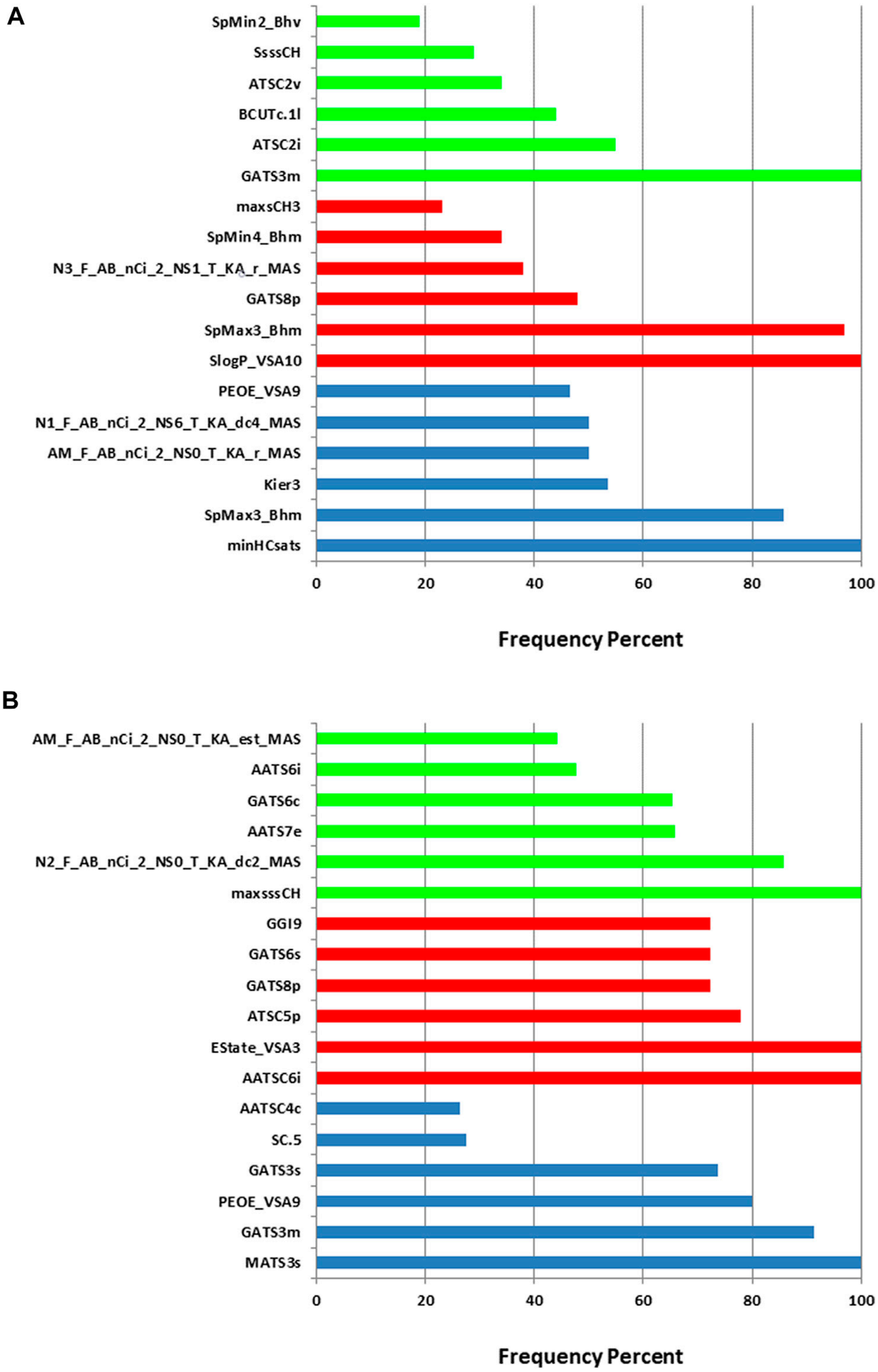

FIGURE 3
Relative frequencies of descriptors in the descriptor combinations selected by the evolutionary algorithm. **(A)** training set 1. **(B)** training set 2. Blue: ANN, red: SVM and green: PLS.

TABLE 3 The minimum, maximum and median number of the molecular descriptors in the final predictor combinations for each training set and machine learning method.

Training set	Machine learning method	min	max	Median
1	ANN	2	7	5
1	SVM	3	7	4
1	PLS	3	7	3
2	ANN	4	9	5
2	SVM	4	9	6
2	PLS	4	9	6

TABLE 4 Improvement of quality of individual validation set predictions by the evolutionary algorithm.

Validation set	Machine learning method	Initial first quartile $RMSE_{val}$	Final first quartile $RMSE_{val}$
1	ANN	0.41	0.33
1	SVM	0.40	0.31
1	PLS	0.41	0.29
2	ANN	0.40	0.35
2	SVM	0.39	0.36
2	PLS	0.44	0.37

TABLE 5 Evaluation of predictive performance of the different machine learning models.

Set	Machine learning model	$RMSE_{val}$	R^2_{val}
1	ANN	0.23	0.89
1	SVM	0.32	0.81
1	PLS	0.26	0.86
2	ANN	0.30	0.80
2	SVM	0.36	0.72
2	PLS	0.34	0.75

3.5 Y-randomization of training set responses

Y-randomization involves randomizing the responses column and then training the predictive models using one of the machine learning methods, with the input descriptors and the responses being mismatched due to the randomization of the responses (Rucker et al., 2007). Y-randomization was carried-out using the final combinations selected by the evolutionary algorithm as inputs. The resulting predictions together with the actual responses are shown in Figure 5. It can be seen that there is no correlation between the predicted and actual responses for both validation sets and for all of the machine learning methods used. The R^2_{val} values ranged from 0.014 to 0.116, with $RMSE_{val}$ values between 0.66 and 0.68. This lack of correlation proves that the results obtained without randomization of the responses (Figure 4 and Table 5) where not due to random chance.

3.6 Effect of setting the formulation descriptor PEG mol% to either the maximum or the minimum value

To examine if the predictive models capture the changes in the formulation descriptor; the PEG mol%, the values of this descriptor were set to either its maximum value or rather its minimum counterpart. It is well known that when using siRNA lipoplexes, there is a certain PEG mol% that results in the maximum *in vivo* efficacy in addition to stabilization of the nanoparticles (Mui et al., 2013; Kumar et al., 2014; Sakurai et al., 2020). The general trend is that increasing the PEG mol% more than a specific mole percent results in decreasing the *in vivo* efficacy. It is generally found that PEG mol% that is equal to 10 decreases efficacy, while values around 1.5% results in good *in vivo* efficacy (Jayaraman et al., 2012; Kumar et al., 2014). Hypothetically, it is assumed that if the PEG mol% descriptor values were set to the maximum (equivalent to 10%), the *in vivo*

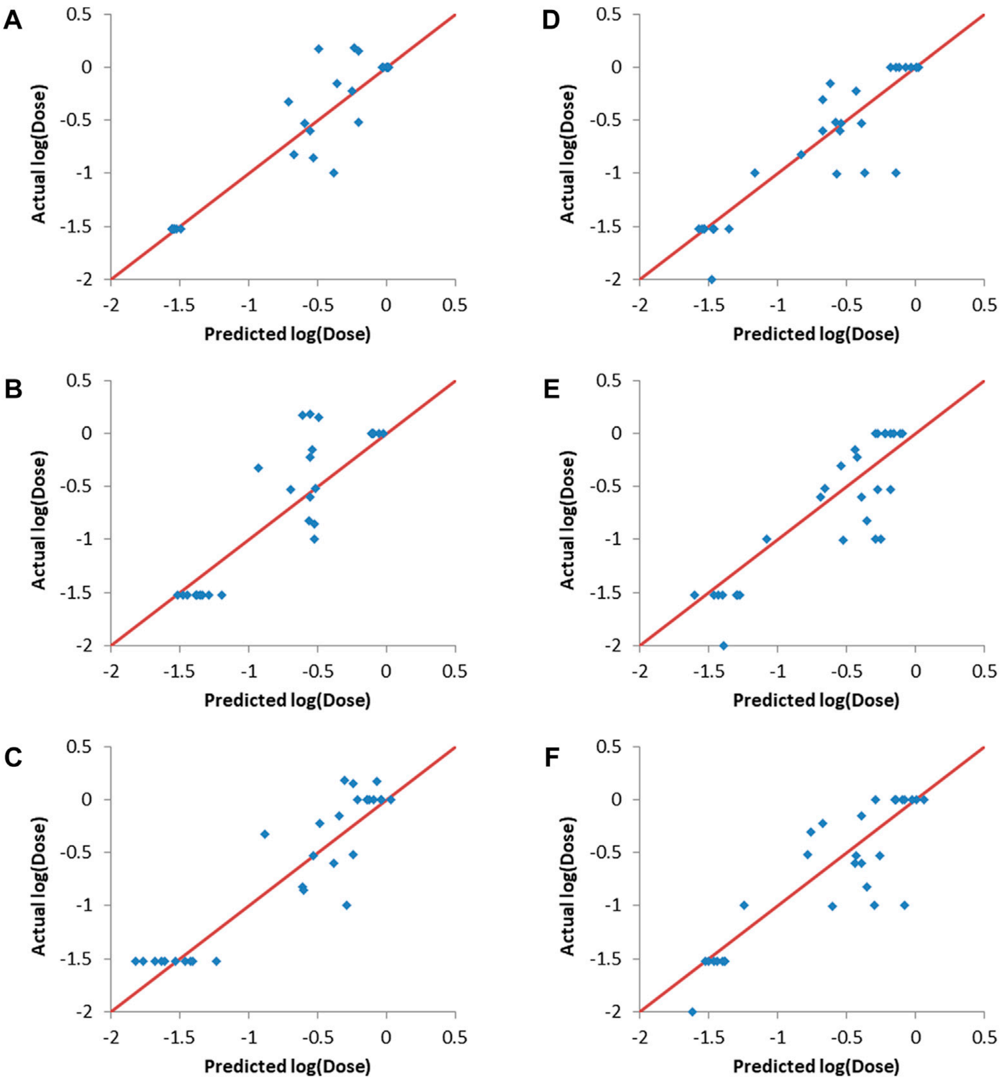

FIGURE 4
Actual vs predicted log (dose) plots. **(A–C)**: Validation set 1, **(A)** ANN, **(B)** SVM and **(C)** PLS. **(D–F)**: Validation set 2, **(D)** ANN, **(E)** SVM and **(F)** PLS.

efficacy should decrease, i.e., the log (dose) should increase. On the other hand, if the PEG mol% values are set to the minimum (equivalent to 1.5%), then the *in vivo* efficacy should generally improve for the validation sets lipids that have PEG mol% higher than 1.5%.

It can be seen in Figures 6A,C that setting the PEG mol% to the minimum values resulted in a decrease in log (dose) as expected, as evident by the shift of the predictions towards the left hand side. Similarly, setting PEG mol% to the maximum value resulted in shifting of the predicted log (dose) towards higher values as it would be expected (Figures 6B,D). These results prove that the predictive models could capture the significance of the formulation descriptor in a correct manner. ANN was the method used to train the models because it resulted in the best predictions as shown in Figure 4 and Table 5. Similar results were obtained with SVM and PLS regression (data not shown).

3.7 Refining the predictions by determining the applicability domain

AD represents a theoretical region in the chemical space of the training set samples. It is expected that predicting the response of unknown samples, e.g., an external validation set, results in more reliable predictions when the unknown samples fall within this region (Weaver & Gleeson, 2008; Tropsha, 2010). One method to determine this region is by

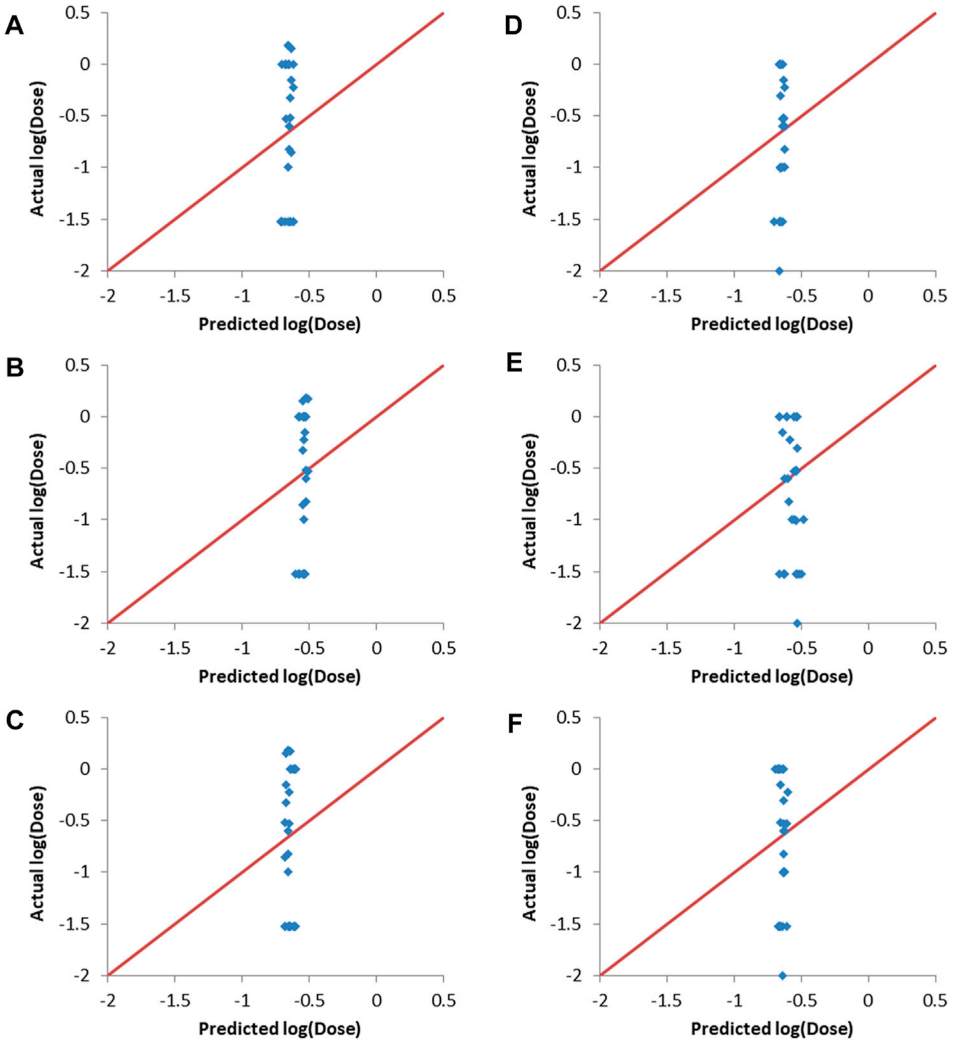

FIGURE 5
Actual vs predicted responses of validation sets after Y-randomization of training sets responses. **(A–C)** validation set one. **(A)** ANN, **(B)** SVM and **(C)** PLS. **(D–F)** validation set two. **(D)** ANN, **(E)** SVM and **(F)** PLS.

applying PCA on the training and validation data, and constructing the region of applicability accordingly (Weaver & Gleeson, 2008). Figure 7A shows the score plot of one fold of training set one and lipid 15 which belongs to validation set one (shown as a red circle). The descriptors combination used to perform PCA were chosen randomly from one of the final combinations selected by the evolutionary algorithm. The region encircled by the blue line is the AD, and it was determined manually by excluding from the training entries under consideration those which are far from lipid 15 in the space generated by plotting PC 1 and PC 2. The first two components capture 66% of the variance in the data. The training lipids selected within the AD were then used by ANN to predict the response of lipid 15. This procedure was repeated for another three lipids from the same validation set. The four lipids selected were chosen based on them having the highest biases in their predicted values (Table 6). It is clear by comparing the predicted responses in Table 6 before and after carrying out the selection of training lipids lying in the AD that there was a vast improvement in the quality of the predictions as seen from the much lower bias values before and after selection. In addition, the R^2 for the four lipids was 0.47 and 0.96 before and after applying AD lipid selection respectively, showing significant improvement in the prediction accuracy of these lipids. The impact of improvement of predictions can be seen in Figures 7B,C, where the predictions lie much closer to the red line in Figure 7C compared to 7B. Since this procedure is carried out manually, we suggest that is should be performed as a refining step for the set of lipids that will be chosen for further wet lab experimentations.

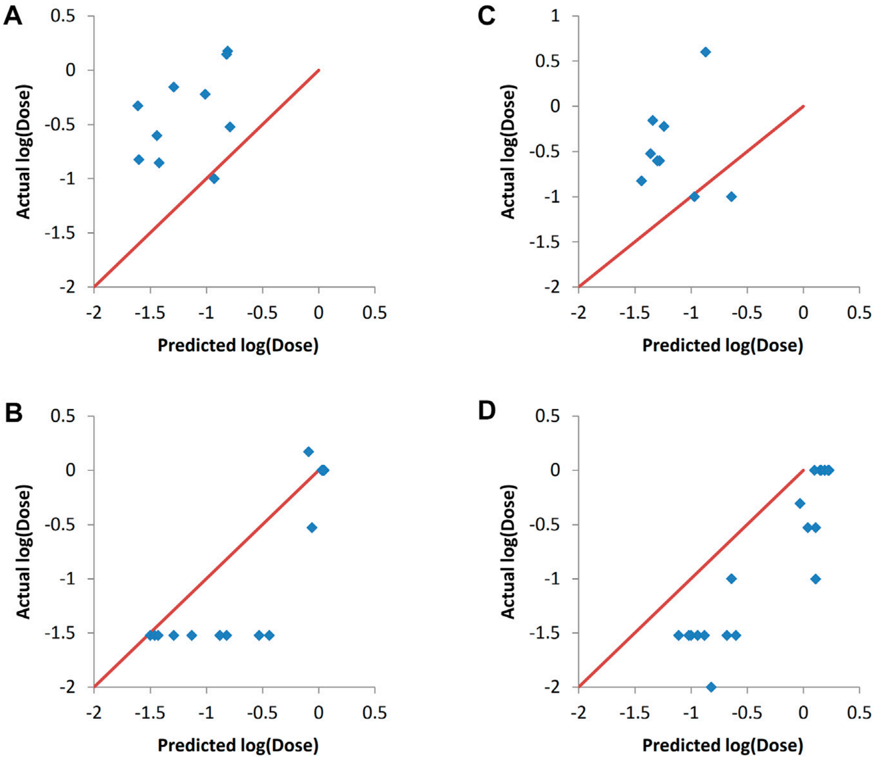

FIGURE 6
Actual vs predicted responses of validation sets after setting the values of the PEG mol% descriptor to either the minimum value **(A, C)** or the maximum value **(B, D)**. The modeling was carried out by ANN. **(A)** and **(B)** validation set one. **(C)** and **(D)** validation set 2. The validation sets entries with the actual PEG mol% being the minimum value were omitted from A and C, while those with the actual PEG mol% being the maximum were omitted from B and D for visualization clarity.

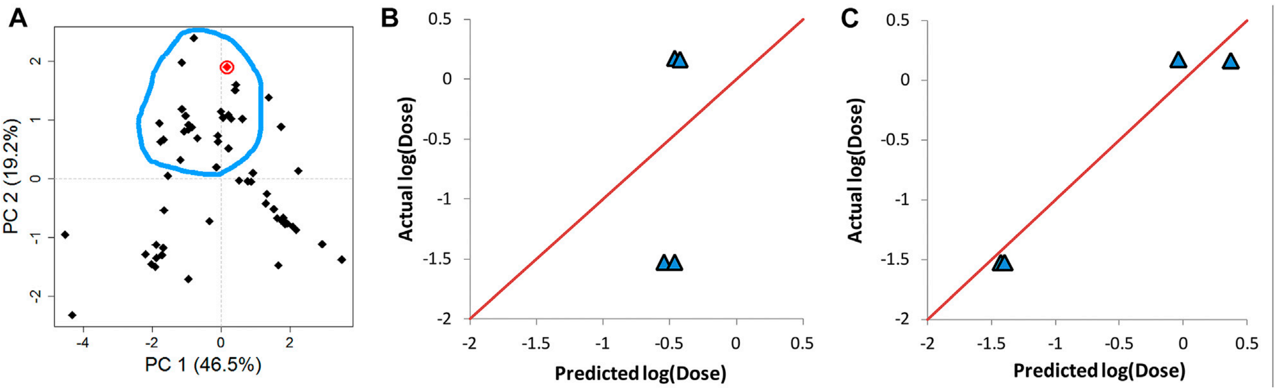

FIGURE 7
Determination of the applicability domain (AD) of four lipids from validation set one. **(A)** PCA of training set together with one of the validation set lipids (lipid 15) shown in red circle. **(B)** The actual vs predicted plot before determining AD. **(C)** The actual vs predicted plot after determining AD. Predictions in C and B are carried out by ANN. The red line in B and C represents perfect correlation between actual and predicted values.

TABLE 6 Refinement of predictions by selecting training lipids within AD.

Lipid index	Actual response Log (dose)	Predicted response before applying AD	Predicted response after applying AD	Bias before applying AD selection	Bias after applying AD selection
15	−1.52	−0.55	−1.43	0.97	0.09
16	−1.52	−0.47	−1.40	1.05	0.12
70	0.18	−0.47	−0.04	−0.65	−0.22
109	0.17	−0.43	0.37	−0.60	0.20

4 Discussion

This study provides a computational framework to predict *in silico* the *in vivo* performance of the siRNA lipid nanoparticles. The main question answered in this manuscript is how to predict the siRNA dose of siRNA lipid nanoparticles given a set of molecular descriptors, formulation characteristics and a required knockdown percent. From the results presented in this work, it is evident that this objective was successfully achieved. In order to produce high quality predictions, the following aspects were carefully considered; 1) The selection of the optimal descriptor combinations 2) The modeling approach 3) Validation of the machine learning models using external validation sets and 4) Improving the predictive outcome of the final models by selecting the training set lipids according to the applicability domain.

When preparing the data set, 2D descriptors were calculated from the ionizable lipid structures rather than 3D descriptors. The reason for avoiding the use of 3D descriptors is that not all the lipids were defined in terms of their stereochemistry. In addition, the optimized 3D structure of a single molecule present in the solution state might differ from the 3D structure of the same molecule if present in close contact with other molecules as in the case of nanoparticles. The effect of the source of the 3D structure and its preparation method and energy minimization in relation to the quality of predictions of three classes of molecules (anilines, carboxylic acids and phenols) has been previously shown (Geidl et al., 2015). There are other potentially important formulation factors that may play a role in the modeling, e.g., particle size and siRNA to lipid ratio, however, they were not included as they were not reported consistently in the selected literature. For example, particle size was reported on occasions as a wide range instead of well defined values. Nanoparticles pK_a was also not included in the descriptors as it is not initially a controllable variable that could be pre-determined compared to the formulation parameters, the lipid structure (by its design) and the required percent knock-down.

As for the descriptor selection, an evolutionary algorithm was used. The evolutionary algorithm comprised: (a) “selection” of the descriptor combinations based on an optimization criterion; the RMSE of the test set after splitting the training set into three folds during training, (b) “crossover” of the selected parent combinations to make new offspring combinations and (c) “mutations” of certain descriptors in offspring combinations. These processes are main elements in any evolutionary algorithm (Sipper et al., 2018). Evolutionary algorithms are suitable for solving the problem of finding optimized solutions of combinations from a set of inputs (descriptors in this case) where an exhaustive search that covers all possible combinations is computationally not feasible (Douguet et al., 2000). In addition, evolutionary algorithms perform better in the presence of noise in data (Arnold & Beyer, 2003). They also offer a set of solutions, which allows for averaging of the predictions of these solutions to get a better predictive performance.

Accordingly, evolutionary algorithms and their variants, such as genetic algorithms, were used to refine the structure of Au nanoparticles (Yu et al., 2016) and to optimize descriptor combinations in counter-propagation artificial neural networks models used to classify drugs as being either hepatotoxic or non-hepatotoxic (Bajželj & Drgan, 2020).

The R software or Microsoft Open R as well as the cheminformatic packages used in this study are available for free, which makes them completely accessible for a wider population of researchers. Using free modeling tools is gaining momentum, for example, additional web-accessible prediction tools and machine-learning based algorithms were successfully utilized to design amphiphilic peptide scaffolds for engineering drug delivery nanoassemblies (Feger et al., 2020).

The modeling approach in the current work involved three machine learning methods: ANN, SVM and PLS. These methods differ in their inner workings. The ANNs are considered a collection of linear and non-linear functions that are governed by the choice of the ANN architecture and activation functions. The SVM belongs to the class of kernel algorithms while PLS regression depends on the construction of latent components (principal components) that result in the best covariance with the response variable. Thus, the difference in their predictive performance could be expected. In order to improve the predictive outcome of the final models, averaging of the predicted response values was carried out. Averaging of predictions belongs to a set of machine learning methods called ensemble learning, and usually results in better prediction outcome (Oprisiu et al., 2012).

Machine learning models require reliable validation to be sure about their ability to successfully predict unknown observations responses. For this purpose, many metrics were suggested and used such as R^2, Q^2 and external validation set R^2. Similarly, RMSE of training set predictions, cross-validation

RMSE and external validation RMSE are used for the same purpose. In addition, techniques such as Y-randomization are used to exclude the possibility of the model predictions being due to random chance. Q^2, the cross-validation coefficient of determination, does not necessarily correlate with good predictive performance for external validation sets (Golbraikh & Tropsha, 2002). Thus, in this work the validation of the final machine learning models was carried out by predicting responses of two external validation sets as well as performing Y-randomization of training set responses, conforming to the best model validation practices (Tropsha, 2010; Maleki et al., 2020). The results showed that the obtained models are reliable.

It is suggested that training set composition and/or the relevant properties of the validation set in relation to the training set governs the predictive performance (Martin et al., 2012; Nalepa & Kawulok, 2019). One way to overcome this is to make sure that the validation set observations are within the applicability domain of the training set (Tropsha, 2010; Maleki et al., 2020). In the current work, rather than selecting the validation set observations that lie within the training set applicability domain, a reverse approach was followed; a subset of the training set elements was selected to be close in the predictor space to the validation element under investigation, i.e., these selected training set elements were used to construct the applicability domain. PCA of the training set and the validation set lipid was carried out to determine this applicability domain visually (Figure 7A). It is evident from the results presented in Figures 7B,C and Table 6 that this protocol resulted in significant improvement in performance.

Recently, *in vitro* cellular uptake of siRNA nanoparticles formulated with hydrophobic derivatives of polyethyleneimine (PEI) was predicted by QSAR modeling using either linear regression, random forests or multilayer perceptron, with the non-linear methods proving to be more efficient than linear regression (Nademi et al., 2021). The R^2 of the external test set ranged between 0.34 and 0.50 depending on the machine learning method used and on the number of input descriptors, with the initial number of 26 descriptors being reduced either by binary encoding or by backward elimination.

Overall, in the current work, *in vivo* performance of siRNA nanoparticles could be predicted accurately by combining machine learning techniques with cheminformatics. This framework will greatly enhance the development of siRNA nanomedicines.

5 Conclusion

The *in vivo* efficacy of siRNA ionizable lipid nanoparticles could be predicted with excellent accuracy provided careful modeling choices. Calculating molecular descriptors of a series of ionizable lipids followed by selecting best descriptor combinations using an evolutionary algorithm in combination with machine learning modeling by ANN, SVM and PLS and then finally making an ensemble of the predictions by calculating the median of validation set predictions resulted in successful predictions of *in vivo* activity of siRNA ionizable lipids nanoparticles. Depending on the machine learning method and the validation set, R^2_{val} of up to 0.89 could be achieved. Further improvement of validation set entries with high bias was achievable by selecting training lipids within the applicability domain, with R^2_{val} improvement from 0.47 to 0.96.

This *in silico* approach allows the evaluation of virtually an endless number of ionizable lipids prior to their actual synthesis and wet lab evaluation and hence saving valuable resources and time while exploring the vast chemical space of these lipids and their formulations.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

Conceptualization, AM; Methodology AM, AN, and RH; Data curation and software, AM; Formal analysis, AM, AN, and RH; Project administration, AM; Writing—original draft preparation, AM, AN, and RH; Writing—review and editing, AM, AN, and RH. All authors have read and agreed to the published version of the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.1042720/full#supplementary-material

References

Alabi, C. A., Love, K. T., Sahay, G., Yin, H., Luly, K. M., Langer, R., et al. (2013). Multiparametric approach for the evaluation of lipid nanoparticles for siRNA delivery. *Proc. Natl. Acad. Sci. U. S. A.* 110 (32), 12881–12886. doi:10.1073/pnas.1306529110

Angelov, B., Garamus, V. M., Drechsler, M., and Angelova, A. (2017). Structural analysis of nanoparticulate carriers for encapsulation of macromolecular drugs. *J. Mol. Liq.* 235, 83–89. doi:10.1016/j.molliq.2016.11.064

Arnold, D. V., and Beyer, H.-G. (2003). A comparison of evolution strategies with other direct search methods in the presence of noise. *Comput. Optim. Appl.* 24 (1), 135–159. doi:10.1023/A:1021810301763

Bajželj, B., and Drgan, V. (2020). Hepatotoxicity modeling using counter-propagation artificial neural networks: Handling an Imbalanced classification problem. *Molecules* 25 (3), 481. doi:10.3390/molecules25030481

Basant, N., and Gupta, S. (2017). Modeling uptake of nanoparticles in multiple human cells using structure–activity relationships and intercellular uptake correlations. *Nanotoxicology* 11 (1), 20–30. doi:10.1080/17435390.2016.1257075

Boulesteix, A. L., and Strimmer, K. (2007). Partial least squares: A versatile tool for the analysis of high-dimensional genomic data. *Brief. Bioinform.* 8 (1), 32–44. bbl016 [pii]. doi:10.1093/bib/bbl016

ClinicalTrials.gov (2020). ClinicalTrials.gov. Available online: https://clinicaltrials.gov/ct2/results?term=siRNA&age_v=&gndr=&type=&rslt=&phase=1&phase=2&phase=3&Search=Apply (Accessed March 2020).

Consonni, V., and Todeschini, R. (2010). "Molecular descriptors," in *Recent Advances in QSAR studies: Methods and Applications. Challenges and Advances in computational Chemistry and PhysicsSpringer*. Editors J. Leszczynski, T. Puzyn, and M. T. D. Cronin, Amsterdam, Netherlands: ScienceDirect, 29–102.

Cullis, P. R., and Hope, M. J. (2017). Lipid nanoparticle systems for enabling gene therapies. *Mol. Ther.* 25 (7), 1467–1475. doi:10.1016/j.ymthe.2017.03.013

De Benedetti, P. G., and Fanelli, F. (2018). Computational modeling approaches to quantitative structure–binding kinetics relationships in drug discovery. *Drug Discov. Today* 23 (7), 1396–1406. doi:10.1016/j.drudis.2018.03.010

Deist, T. M., Dankers, F. J. W. M., Valdes, G., Wijsman, R., Hsu, I. C., Oberije, C., et al. (2018). Machine learning algorithms for outcome prediction in (chemo) radiotherapy: An empirical comparison of classifiers. *Med. Phys.* 45 (7), 3449–3459. doi:10.1002/mp.12967

Dong, Y., Siegwart, D. J., and Anderson, D. G. (2019). Strategies, design, and chemistry in siRNA delivery systems. *Adv. Drug Deliv. Rev.* 144, 133–147. doi:10.1016/j.addr.2019.05.004

Douguet, D., Thoreau, E., and Grassy, G. (2000). A genetic algorithm for the automated generation of small organic molecules: Drug design using an evolutionary algorithm. *J. Comput. Aided. Mol. Des.* 14 (5), 449–466. doi:10.1023/a:1008108423895

Dowdy, S. F. (2017). Overcoming cellular barriers for RNA therapeutics. *Nat. Biotechnol.* 35, 222–229. doi:10.1038/nbt.3802

Elbashir, S. M., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K., and Tuschl, T. (2001). Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* 411 (6836), 494–498. doi:10.1038/35078107

Feger, G., Angelov, B., and Angelova, A. (2020). Prediction of amphiphilic cell-penetrating peptide building blocks from protein-derived amino acid sequences for engineering of drug delivery nanoassemblies. *J. Phys. Chem. B* 124 (20), 4069–4078. doi:10.1021/acs.jpcb.0c01618

Fire, A., Xu, S. Q., Montgomery, M. K., Kostas, S. A., Driver, S. E., and Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in caenorhabditis elegans. *Nature* 391 (6669), 806–811. doi:10.1038/35888

Geidl, S., Svobodová Vařeková, R., Bendová, V., Petrusek, L., Ionescu, C.-M., Jurka, Z., et al. (2015). How does the methodology of 3D structure preparation influence the quality of pKa prediction? *J. Chem. Inf. Model.* 55 (6), 1088–1097. doi:10.1021/ci500758w

Golbraikh, A., and Tropsha, A. (2002). Beware of q2. *J. Mol. Graph. Model.* 20 (4), 269–276. [pii]. doi:10.1016/s1093-3263(01)00123-1

Hamner, B., Frasco, M., and LeDell, E. (2018). *Metrics: Evaluation metrics for machine learning*. Available at: https://cran.r-project.org/package=Metrics

Hathout, R. M., Abdelhamid, S. G., El-Housseiny, G. S., and Metwally, A. A. (2020a). Comparing cefotaxime and ceftriaxone in combating meningitis through nose-to-brain delivery using bio/chemoinformatics tools. *Sci. Rep.* 10 (1), 21250. doi:10.1038/s41598-020-78327-w

Hathout, R. M., El-Ahmady, S. H., and Metwally, A. A. (2018). Curcumin or bisdemethoxycurcumin for nose-to-brain treatment of alzheimer disease? A bio/chemo-informatics case study. *Nat. Prod. Res.* 32 (24), 2873–2881. doi:10.1080/14786419.2017.1385017

Hathout, R. M., Metwally, A. A., Woodman, T. J., and Hardy, J. G. (2020b). Prediction of drug loading in the Gelatin matrix using computational methods. *ACS Omega* 5 (3), 1549–1556. doi:10.1021/acsomega.9b03487

Heikamp, K., and Bajorath, J. (2014). Support vector machines for drug discovery. *Expert Opin. Drug Discov.* 9 (1), 93–104. doi:10.1517/17460441.2014.866943

Jayaraman, M., Ansell, S. M., Mui, B. L., Tam, Y. K., Chen, J., Du, X., et al. (2012). Maximizing the potency of siRNA lipid nanoparticles for hepatic gene silencing *in vivo*. *Angew. Chem. Int. Ed. Engl.* 51 (34), 8529–8533. doi:10.1002/anie.201203263

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). Kernlab - an S4 package for kernel methods in R. *J. Stat. Softw.* 11 (9), 1–20. doi:10.18637/jss.v011.i09

Kausar, S., and Falcao, A. O. (2018). An automated framework for QSAR model building. *J. Cheminform.* 10 (1), 1. doi:10.1186/s13321-017-0256-5

Kumar, V., Qin, J., Jiang, Y., Duncan, R. G., Brigham, B., Fishman, S., et al. (2014). Shielding of lipid nanoparticles for siRNA delivery: Impact on physicochemical properties, cytokine induction, and efficacy. *Mol. Ther. Nucleic Acids* 3, e210. mtna201461 [pii]. doi:10.1038/mtna.2014.61

Liu, R., Rallo, R., Bilal, M., and Cohen, Y. (2015). Quantitative structure-activity relationships for cellular uptake of surface-modified nanoparticles. *Comb. Chem. High. Throughput Screen.* 18 (4), 365–375. CCHTS-EPUB-65712 [pii]. doi:10.2174/1386207318666150306105525

Maleki, F., Muthukrishnan, N., Ovens, K., Reinhold, C., and Forghani, R. (2020). Machine learning algorithm validation: From essentials to advanced applications and implications for regulatory certification and deployment. *Neuroimaging Clin. N. Am.* 30 (4), 433–445. S1052-5149(20)30059-9 [pii]. doi:10.1016/j.nic.2020.08.004

Maltarollo, V. G., Kronenberger, T., Espinoza, G. Z., Oliveira, P. R., and Honorio, K. M. (2019). Advances with support vector machines for novel drug discovery. *Expert Opin. Drug Discov.* 14 (1), 23–33. doi:10.1080/17460441.2019.1549033

Martin, T. M., Harten, P., Young, D. M., Muratov, E. N., Golbraikh, A., Zhu, H., et al. (2012). Does rational selection of training and test sets improve the outcome of QSAR modeling? *J. Chem. Inf. Model.* 52 (10), 2570–2578. doi:10.1021/ci300338w

Metwally, A. A., Blagbrough, I. S., and Mantell, J. M. (2012a). Quantitative silencing of EGFP reporter gene by self-assembled siRNA lipoplexes of LinOS and cholesterol. *Mol. Pharm.* 9 (11), 3384–3395. doi:10.1021/mp300435x

Metwally, A. A., and Hathout, R. M. (2015). Computer-assisted drug formulation design: Novel approach in drug delivery. *Mol. Pharm.* 12 (8), 2800–2810. doi:10.1021/mp500740d

Metwally, A. A., Reelfs, O., Pourzand, C., and Blagbrough, I. S. (2012b). Efficient silencing of EGFP reporter gene with siRNA delivered by asymmetrical N^4, N^9-diacyl spermines. *Mol. Pharm.* 9 (7), 1862–1876. doi:10.1021/mp200429n

Mevik, B.-H., and Wehrens, R. (2007). The**pls**Package: Principal component and partial least squares regression in*R*. *J. Stat. Softw.* 18 (2), 23. doi:10.18637/jss.v018.i02

Molla, M. R., Chakraborty, S., Munoz–Sagredo, L., Drechsler, M., Orian–Rousseau, V., and Levkin, P. A. (2020). Combinatorial synthesis of a lipidoid library by thiolactone chemistry: *In vitro* screening and *in vivo* validation for siRNA delivery. *Bioconjug. Chem.* 31 (3), 852–860. doi:10.1021/acs.bioconjchem.0c00013

Mui, B. L., Tam, Y. K., Jayaraman, M., Ansell, S. M., Du, X., Tam, Y. Y. C., et al. (2013). Influence of polyethylene glycol lipid desorption rates on pharmacokinetics and pharmacodynamics of siRNA lipid nanoparticles. *Mol. Ther. Nucleic Acids* 2 (12), e139. doi:10.1038/mtna.2013.66

Nademi, Y., Tang, T., and Uludağ, H. (2021). Modeling uptake of polyethylenimine/short interfering RNA nanoparticles in beast cancer cells using machine learning. *Adv. NanoBiomed Res.* 1, 2000106. doi:10.1002/anbr.202000106

Nalepa, J., and Kawulok, M. (2019). Selecting training sets for support vector machines: A review. *Artif. Intell. Rev.* 52 (2), 857–900. doi:10.1007/s10462-017-9611-1

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R. (2011). Open Babel: An open chemical toolbox. *J. Cheminform.* 3 (1), 33. doi:10.1186/1758-2946-3-33

Oprisiu, I., Varlamova, E., Muratov, E., Artemenko, A., Marcou, G., Polishchuk, P., et al. (2012). QSPR approach to predict nonadditive properties of mixtures. Application to bubble point temperatures of binary mixtures of liquids. *Mol. Inf.* 31 (6-7), 491–502. doi:10.1002/minf.201200006

Paunovska, K., Gil, C. J., Lokugamage, M. P., Sago, C. D., Sato, M., Lando, G. N., et al. (2018). Analyzing 2000 *in vivo* drug delivery data points reveals cholesterol structure impacts nanoparticle delivery. *ACS Nano* 12 (8), 8341–8349. doi:10.1021/acsnano.8b03640

Racz, A., Bajusz, D., and Heberger, K. (2019). Intercorrelation limits in molecular descriptor Preselection for QSAR/QSPR. *Mol. Inf.* 38 (8-9), e1800154. doi:10.1002/minf.201800154

Rajappan, K., Tanis, S. P., Mukthavaram, R., Roberts, S., Nguyen, M., Tachikawa, K., et al. (2020). Property-driven design and development of lipids for efficient delivery of siRNA. *J. Med. Chem.* 63 (21), 12992–13012. doi:10.1021/acs.jmedchem.0c01407

Reker, D., Perna, A. M., Rodrigues, T., Schneider, P., Reutlinger, M., Mönch, B., et al. (2014). Revealing the macromolecular targets of complex natural products. *Nat. Chem.* 6, 1072–1078. doi:10.1038/nchem.2095

Rucker, C., Rucker, G., and Meringer, M. (2007). y-Randomization and its variants in QSPR/QSAR. *J. Chem. Inf. Model.* 47 (6), 2345–2357. doi:10.1021/ci700157b

Sakurai, Y., Mizumura, W., Ito, K., Iwasaki, K., Katoh, T., Goto, Y., et al. (2020). Improved stability of siRNA-loaded lipid nanoparticles prepared with a PEG-monoacyl fatty acid facilitates ligand-mediated siRNA delivery. *Mol. Pharm.* 17 (4), 1397–1404. doi:10.1021/acs.molpharmaceut.0c00087

Sato, Y., Hashiba, K., Sasaki, K., Maeki, M., Tokeshi, M., and Harashima, H. (2019). Understanding structure-activity relationships of pH-sensitive cationic lipids facilitates the rational identification of promising lipid nanoparticles for delivering siRNAs *in vivo*. *J. Control. Release* 295, 140–152. doi:10.1016/j.jconrel.2019.01.001

Shamay, Y., Shah, J., Işık, M., Mizrachi, A., Leibold, J., Tschaharganeh, D. F., et al. (2018). Quantitative self-assembly prediction yields targeted nanomedicines. *Nat. Mater.* 17 (4), 361–368. doi:10.1038/s41563-017-0007-z

Sipper, M., Fu, W., Ahuja, K., and Moore, J. H. (2018). Investigating the parameter space of evolutionary algorithms. *BioData Min.* 11, 2. doi:10.1186/s13040-018-0164-x

Titze-de-Almeida, R., David, C., and Titze-de-Almeida, S. S. (2017). The race of 10 synthetic RNAi-based drugs to the pharmaceutical market. *Pharm. Res.* 34 (7), 1339–1363. doi:10.1007/s11095-017-2134-2

Todeschini, R., and Consonni, V. (2008). *Handbook of Molecular Descriptors*, Weinheim, Germany: WILEY-VCH Verlag GmbH.

Tropsha, A. (2010). Best practices for QSAR model development, validation, and exploitation. *Mol. Inf.* 29 (6-7), 476–488. doi:10.1002/minf.201000061

Valdés-Martiní, J. R., Marrero-Ponce, Y., García-Jacas, C. R., Martinez-Mayorga, K., Barigye, S. J., Vaz d'Almeida, Y. S., et al. (2017). QuBiLS-MAS, open source multi-platform software for atom- and bond-based topological (2D) and chiral (2.5D) algebraic molecular descriptors computations. *J. Cheminform.* 9 (1), 35. doi:10.1186/s13321-017-0211-5

Wang, W., Sedykh, A., Sun, H., Zhao, L., Russo, D. P., Zhou, H., et al. (2017). Predicting nano–bio interactions by Integrating nanoparticle libraries and quantitative nanostructure activity relationship modeling. *ACS Nano* 11 (12), 12641–12649. doi:10.1021/acsnano.7b07093

Weaver, S., and Gleeson, M. P. (2008). The importance of the domain of applicability in QSAR modeling. *J. Mol. Graph. Model.* 26 (8), 1315–1326. (08) 00003-X [pii]. doi:10.1016/j.jmgm.2008.01.002

Wesolowski, M., and Suchacz, B. (2019). Artificial neural networks: Theoretical background and pharmaceutical applications: A review. *J. AOAC Int.* 95 (3), 652–668. doi:10.5740/jaoacint.SGE_Wesolowski_ANN

Whitehead, K. A., Dorkin, J. R., Vegas, A. J., Chang, P. H., Veiseh, O., Matthews, J., et al. (2014). Degradable lipid nanoparticles with predictable *in vivo* siRNA delivery activity. *Nat. Commun.* 5, 4277. ncomms5277 [pii]. doi:10.1038/ncomms5277

Whitehead, K. A., Langer, R., and Anderson, D. G. (2009). Knocking down barriers: Advances in siRNA delivery. *Nat. Rev. Drug Discov.* 8 (2), 129–138. doi:10.1038/nrd2742

Yap, C. W. (2011). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* 32 (7), 1466–1474. doi:10.1002/jcc.21707

Yu, M., Yankovich, A. B., Kaczmarowski, A., Morgan, D., and Voyles, P. M. (2016). Integrated computational and experimental structure refinement for nanoparticles. *ACS Nano* 10 (4), 4031–4038. doi:10.1021/acsnano.5b05722

Zhang, M. M., Bahal, R., Rasmussen, T. P., Manautou, J. E., and Zhong, X.-b. (2021). The growth of siRNA-based therapeutics: Updated clinical studies. *Biochem. Pharmacol.* 189, 114432. doi:10.1016/j.bcp.2021.114432

Žuvela, P., Liu, J. J., Macur, K., and Bączek, T. (2015). Molecular descriptor subset selection in theoretical peptide quantitative structure–retention relationship model development using nature-Inspired optimization algorithms. *Anal. Chem.* 87 (19), 9876–9883. doi:10.1021/acs.analchem.5b02349