Polyethylenimine (PEI) is one of the most promising nonviral vectors for delivery of short interfering RNA (siRNA) agents into cancer cells. A promising approach that increases the delivery efficiency of PEI is its modification with hydrophobic substitutions. However, the performance of modified PEIs depends on the nature and extent of substitutions. Herein, machine learning algorithms are used on the basis of quantitative structure activity relationship (QSAR) method to predict the cellular uptake of hydrophobically modified PEI/siRNA nanoparticles (NPs) into various cancer cell lines. To this end, 3 different regression models, namely, random forest (RF), multilayer perceptron (MLP), and linear regression (LR), are used. The results show that RF and MLP regression methods have a better performance than the LR method, suggesting that nonlinear models are better estimators when predicting the cellular uptake of PEI/siRNA NPs. Additionally, critical descriptors that have major contributions to cellular uptake are found to be PEI-to-siRNA weight ratio, type of hydrophobic substitution, as well as total numbers of Cs, unsaturated C, and thioester groups on substitutions in each PEI. This study is the first report that predicts cellular uptake with PEI-based carriers, which provides valuable insight into the design of performance-enhancing hydrophobic substituents on PEIs.

1. Introduction

Gene silencing using polynucleotides aims to silence unwanted genes by blocking the translation of target messenger ribonucleic acids (mRNAs).[1] An example of such polynucleotides is short interfering RNAs (siRNAs) that are being explored for the treatment of various diseases including viral infections and cancer. After entry of siRNA into the cytoplasm, the guide strand of the siRNA gets incorporated into RNA-induced silencing complex (RISC), which actively seeks out and degrade the complementary mRNA. For breast cancer, siRNA can be used to downregulate specific mRNAs involved in aberrant signaling, thereby reducing the intracellular levels of targeted proteins at will.[3] Also, by conventionally altering the nucleotide sequence of siRNA molecules, one can silence different mRNAs. Potential silencing targets in breast cancer cells, for example, include Src kinase in MDA-MB-231 cells[4] and anti-apoptotic Mcl-1 in MDA-MB-435.[5] For this action to take place, however, the polynucleotide needs to first enter the cytoplasm of the cells. Without a carrier, the anionic polynucleotides cannot cross anionic lipid bilayers of cell
membranes to induce the gene silencing. Both viral and nonviral carriers can be used to deliver polynucleotides into the cells. Among nonviral carriers, cationic polymers have the advantage of being easily modified with other functional groups, making it feasible to tailor their properties for different applications. Polyethylenimine (PEI) is the most promising and extensively investigated cationic polymer for polynucleotide delivery. The efficacy of PEI was found to increase with its molecular weight. However, high molecular weight (HMW) PEIs (≥25 kDa) exhibited significant toxicity. Low molecular weight (LMW) PEIs (< 2 kDa) had an acceptable level of toxicity, but efficacy of its gene delivery was low. Modification of LMW PEIs with hydrophobic substitutions significantly increased its cellular uptake. Various hydrophobic substitutions were used to enhance efficacy of LMW PEIs including cholesterol, phospholipids, and hydrophilic alkyl groups such as ethyl octyl and aliphatic lipids including caprylic, stearic, and linoleic acids (LAS). As a representative example, Neamnark et al. studied the delivery and transfection efficiency of 2 kDa PEI modified with different hydrophobic substitutions. They observed that these modifications significantly increased cellular uptake compared with unmodified native PEI. The beneficial effect of the hydrophobic modification was shown to depend on both the type and level of the substitution.

A number of studies attempted to address the role of lipids in the delivery efficiency of hydrophobically modified PEIs. Menekesdag-Erol et al. using molecular dynamics (MD) and experimental tools studied LMW PEI modified with short propionic acid (PrA), focusing on the role of level of substitutions. They observed the highest surface hydrophobicity and surface charge density of the PEI/siRNA nanoparticle (NP) at an intermediate substitution ratio; substitution level beyond this optimal value induced migration of PrA toward the NP center and had a deleterious effect on both uptake and silencing. Experimental study of Neamnark et al. showed that at a similar substitution level, PEI modified with LA had a higher transfection efficiency than PEI modified with caprylic acid (CA). These two hydrophobes differ in their length, i.e., number of hydrophobic carbon chains. These studies highlighted the dependence of PEI’s delivery performance on the properties of grafted lipids. However, exhaustive parametric study, experimentally or numerically, on all PEIs and their derivatives is time consuming and impractical. In addition, significant amount of experimental data already exists in the literature on modified PEIs but have not been critically analyzed to explore universal relationships between the physicochemical features of the carriers and their efficiency in delivering polynucleotides.

Machine learning methods can help determine complex relationships between participating factors and desired targets, providing a means to predict the outcome without the need to perform extensive testing. In recent years, chemoinformatics methods including quantitative structure activity/property relationship (QSAR/QSPR) have been utilized to predict the activities/properties of a given compound as functions of its molecular substituents. The core assumption in these methods is that variation in the biological activities of a compound is correlated with changes in its molecular structure. To date, there have been only two studies that utilized QSAR method to correlate molecular properties with cellular uptake of NPs. Both studies examined the cellular uptake data of cross-linked iron oxide (CLIO) NPs from Weissleder et al. and Fourches et al. developed a QSAR model to predict the uptake of CLIO NPs, with a variety of small organic molecules decorating their surface, by human pancreatic cancer cells (PaCa2). The authors used 150 2D descriptors for 109 organic compounds, which included surface area, physicochemical properties (such as the net charge and hydrophilicity), Kier and Hall connectivity indices, kappa shape indices, atom and bond counts, adjacency and distance matrix descriptors, molecular charges, and pharmacophore feature descriptors. A fivefold cross validation k-nearest neighbors (kNN) regression was used as the prediction algorithm. The performance of their model under optimum condition resulted in a R² value of 0.77. It was proposed that higher cellular uptake was associated with higher lipophilicity of the organic molecules bound to the CLIO NP. Using the same data set, Winkler et al. predicted the cellular uptake of CLIO NPs into PaCa2 and human umbilical vein endothelial cells (HUVECs) based on machine learning methods. The data set, containing 108 data points, was separated into training set (87 data points) and test set (21 data points). Two-dimensional DRAGON descriptors were used for the decorated CLIO NPs. Using these descriptors, a linear and nonlinear nano-QSAR model was developed. Eleven DRAGON descriptors were utilized for the uptake of NP into HUVEC cells, and the R² values for the best fits were 0.63 and 0.66, respectively, for the linear and nonlinear models. For the uptake of NPs into PaCa2 cell, 19 DRAGON descriptors were used, which resulted in a R² of 0.79 and 0.54, respectively, for the linear and nonlinear models. Different molecular descriptors were used to predict cellular uptake in different cell lines, which was related to different uptake mechanisms. As these two studies used data set that only involved CLIO NPs, the predictive cellular uptake models can not be applied to more complex systems such as polymeric-nucleotide NPs.

In this work, for the first time, we applied machine learning methods to develop a predictive model where clinical significance (cellular uptake) was used to optimize and improve delivery of PEI/siRNA NPs into breast cancer cells. The data set for this study originated from published studies in our lab since 2011 that utilized different types of breast cancer cells and an evolving pattern of hydrophobically modified LMW PEIs. We selected molecular descriptors that are easy to interpret by chemists, thereby providing informative guidance for the design of effective PEI carriers in gene delivery applications.

2. Experimental Section

2.1. Data Set

The experimental cellular uptake data of PEI/siRNA NPs into various breast cancer cell lines including MDA MB231, MCF7, AU565, MDA 468, MDA 435, and MDA 231 were collected from previous publications by our group. All NPs were tested under similar culture conditions in the same laboratory, thereby enabling direct comparisons across NP formulations. The basic methodology remained the same, where a scrambled FAM-labeled siRNA was formulated with PEI carriers and the uptake determined after 24 h using flow cytometry methodology.
data set was formed from 197 data points where PEI in its native form was the base polymer which was modified with 9 different hydrophobic substitutions including LA, alpha-LA (αLA), thioester linkage LA (tLA), CA, oleic acid (OA), palmitic acid (PA), lauric acid (Lau), stearic acid (StA), and cholesterol (Chol). Molecular structures of aforementioned hydrophobic substitutions are shown in Figure 1. All these structures, except Chol, differed mostly in terms of their length, type of carbon bonds (saturated versus unsaturated), number of unsaturated carbon bonds, and presence of thioester bonds.

2.2. Descriptors

Molecular descriptors were selected to capture the key parameters in the preparation of PEI/siRNA NPs while ensuring easy interpretation by chemists. Considered descriptors and their symbolic notations are shown in Table 1. Three descriptors, namely, $N_c$, $N_{uns}$, and $N_{thio}$ were generated by multiplying two other descriptors as shown in Table 1. The purpose of such an approach was to add complexity and nonlinearity to the descriptors before modeling. The output was logarithmic value of the cellular uptake of PEI/siRNA NPs, hereafter referred to as $y$ ($y = \log$ (cellular uptake)).

2.3. Modeling

The data set was divided into training data (147 data points, 75%) and test data (50 data point, 25%), which was a common ratio when the number of date points was relatively low. Training data were used at the modeling stage whereas the test data were used to evaluate model performance. Three different regression models were used, namely, random forest (RF), multilayer perceptron (MLP), and linear regression (LR). RF was an ensemble of large number of individual decision trees, where the predicted output was the average of predicted values from each decision tree. The MLP model was constructed using one input layer, one hidden layer, and one output layer (see Supporting Information, section S1). The activation function for the hidden layer was the “ReLU” function, whereas the activation function for the output layer was the identity function.[27] The nodes in each layer were connected to the nodes in the next layer by weights, which were interactively adjusted during the training to minimize the network error. The network error was calculated from $\frac{1}{2}||y(\text{predict}) - y(\text{actual})||_2^2 + \sum \alpha ||w||_2^2$, where the subscript 2 represents $l_2$ norm, $\alpha$ is a non-negative hyperparameter that penalize large weights ($\alpha = 0.0001$ used in this work), and $w$ is the weight. LR fitted a linear model between $y$ and the independent descriptors to minimize the model error as calculated by $||y(\text{predict}) - y(\text{actual})||_2^2$. Using both nonlinear (RF and MLP) and linear (LR) models would enable us to map the relation between the descriptors and $y$ with different complexity.

Three approaches were used for the modeling, using different number of descriptors. In Model type 1, all descriptors in Table 1 were directly used, where the two categorical descriptors, $S$ and $C_L$, were transformed using a dummy encoding scheme (see Figure 2). This created a binary column for each category (see Supporting Information, section S2). Using this approach, these two categorical descriptors were translated into

### Table 1. Descriptors used to model cellular uptake of PEI/siRNA NPs.

| Descriptor | Definition | Range |
|------------|------------|-------|
| $M_w$      | MW of PEI [kDa] | 0.6–2 |
| $r$        | PEI to siRNA weight ratio; determines the net charge of NP and the amount of bound siRNA. | 2–10 |
| $BC_{50}$  | PEI to siRNA weight ratio required for 50% binding of a given siRNA amount; siRNA solution without PEI is taken as 0% binding. | 0.1–1.5 |
| $n_{sub}$  | Number of amine groups per PEI modified with hydrophobic substitutions. | 0–6.9 |
| $n_c$      | Number of C on each substitution; represents length of the hydrophobic substitution. | 0–18 |
| $N_c$      | Total number of C on the substitutions in each PEI ($n_c \times n_{sub}$); related to NP hydrobicity. | 0–110 |
| $n_{uns}$  | Number of unsaturated C (C=C bonds) on each substitution. | 0–3 |
| $N_{uns}$  | Total number of unsaturated C on the substitutions in each PEI ($n_{uns} \times n_{sub}$). | 0–13 |
| $n_{thio}$ | Number of thioester groups (R=S=C(O)=R) on each substitution. | 0–1 |
| $N_{thio}$ | Total number of thioester groups on the substitutions in each PEI ($n_{thio} \times n_{sub}$). | 0–7 |
| $S$        | Type of hydrophobic substitution; none (native PEI) or one of the nine types in Figure 1. | – |
| $C_L$      | Type of breast cancer cell line; one of the six types described earlier. | – |
16 descriptors, 10 for S and 6 for C_L. Adding the 10 numerical descriptors in Table 1, the total number of descriptors under Model type 1 was 26. In Model type 2, the total number of descriptors was reduced using a binary encoder, where S and C_L were encoded by 4-digit and 3-digit binary numbers, respectively (see Supporting Information, section S3). The 4-digit binary number for S translated into 4 index descriptors denoted as \( I_{\text{sub}-1}, I_{\text{sub}-2}, I_{\text{sub}-3}, \) and \( I_{\text{sub}-4} \). Similarly, the 3-digit binary number for C_L translated into 3 index descriptors denoted as \( I_{\text{cell}-1}, I_{\text{cell}-2}, \) and \( I_{\text{cell}-3} \). The total number of descriptors under Model type 2 was \( 10 + 4 + 3 = 17 \). Although the smaller number of descriptors lacked the ability to independently describe each type of hydrophobic substitution or cell line, this approach reduced overfitting and potential dimensionality problems. When dimensionality increases, the volume of the descriptor space increases, causing the data points to become sparse, which can be problematic for any models. Using Model type 2, the most informative descriptors were identified as the ones showing more than 10% correlation with cellular uptake, and they were selected for modeling. In Model type 3, the number of descriptors was further reduced from Model type 2 using backward elimination. In this method, independent descriptors were entered into stepwise multilinear regression, followed by the trial deletion of each descriptor using a chosen model fit. The descriptor (if any) whose removal gave statistically the most insignificant deterioration of the model fit was deleted, and this was repeated until no further descriptors could be eliminated without a statistically significant loss of fit. Afterward, the remaining descriptors were used for modeling. Figure 3 shows the selection of final descriptors for Model type 2 and Model type 3.

2.4. Metrics of Model Performance

Among various statistical measures for the performance of regression models, we used two criteria for both the training and test data: squared correlation coefficient usually denoted as \( R^2 \) and root mean square error (RMSE). In addition, accuracy of the model was evaluated using an acceptable difference between the predicted and actual values of \( y \). The overall model accuracy was calculated by dividing the total number of accurate predictions (within 25% of the actual values) by the total number of predictions.

2.5. Statistical Analysis

No transformation was performed on the data set except for the output (\( y \)), which was defined by applying logarithmic transformation on the value of the cellular uptake of PEI/siRNA NPs. The data set was divided into training data (147 data points, 75%) and test data (50 data point, 25%). For the reduction of descriptors using backward elimination method, a significance level of 5% was selected that corresponded to \( p \)-value of 0.05. All modeling and statistical analysis were performed using open source Scikit-learn library.[28] The programming language was Python 3.7. Seaborn python programming package[29] was used for data visualization.

3. Results And Discussion

3.1. Model Type 1

All 26 descriptors (see Figure 2) were directly used for modeling. The regression models are trained and their performances are summarized in Table 2. Better model performance corresponds to higher values of accuracy and \( R^2 \), as well as lower values of RMSE. Among the three methods, performances of RF and MLF were close and both were better estimators than LR.

3.2. Model Type 2

Model type 2 had a reduced number of descriptors (17) compared with Model type 1 (26 descriptors) (see Figure 3). Correlation between these 17 descriptors and \( y \) was calculated using Pearson correlation coefficient (Table 3). The results show that 11 descriptors (highlighted in red in Table 2) had correlation of more than 10% with \( y \). The other six descriptors had negligible correlation and hence were removed from further modeling using Model type 2. The performances of regression models using Model type 2 are shown in Table 2. Similar to Model type 1, RF and MLF were better performers than LR. Compared with
Model type 1 and for RF, $R^2$ and RMSE showed a minor improvement, whereas accuracy slightly decreased using Model type 2. For MLP and LR, the differences in performance metrics between Model type 1 and Model type 2 were insignificant.

### 3.3. Model Type 3

Using lower number of descriptors (Model type 2), no significant decrease in model performance was observed. In Model type 3 where backward elimination was utilized to further remove insignificant descriptors, a significance level of 5% was selected that corresponded to $p$-value of 0.05. LR regressor was fitted using the 11 descriptors identified from Model type 2, and the descriptor with the highest $p$-value was examined. If its $p$-value was greater than the defined significance level of 0.05, the descriptor was removed and LR was implemented again using the remaining descriptors. The process continued until the highest $p$-value from all the remaining descriptors was less than 0.05. The number of descriptors reduced to six after backward elimination, which were $r$, $I_{\text{sub-1}}$, $I_{\text{sub-2}}$, $N_c$, $N_{\text{uns}}$, and $N_{\text{thio}}$. The regression modeling using Model type 3 showed comparable performance to the other two model types (Table 2), suggesting that the most significant descriptors were those in Model type 3. In addition, for all model types, nonlinear models of RF and MLP had a better performance than the linear model of LR. In the following, physical interpretation of the most significant descriptors was presented.

### 3.4. Effect of Most Significant Descriptors

The descriptor $r$, PEI-to-siRNA weight ratio, usually ranges between 2 and 10 in experiments. Higher $r$ leads to higher positive charge on the PEI/siRNA NPs, causing stronger interaction with cell membranes. In addition, higher $r$ increases the presence of free PEIs that are not bound to siRNAs, which might destabilize membrane structure, thereby contributing to the uptake of PEI/nucleic acid NPs. Boeckle et al. attributed the increase in gene expression observed with free PEIs to their ability to facilitate proton sponge effect. They proposed that free PEIs can be internalized into endosomes, which can then merge with vesicles containing PEI/DNA NPs. These free PEIs subsequently assisted in the buffering of vesicular pH and accumulation of $\text{Cl}^{-}$ in the endosome leading to its rupture.

Unlike $r$, the two index descriptors $I_{\text{sub-1}}$ and $I_{\text{sub-2}}$ on a standalone basis are not physical attributes. These descriptors represent a compacted form of the type of hydrophobic substitution, suggesting that type of hydrophobic substitution plays a significant role in the delivery of siRNA into breast cancer cells. As a supporting evidence, Aliabadi et al. explored LMW PEIs substituted by different lipids, as carriers for siRNA-mediated BCRP downregulation. BCRP is an efflux protein whose activity has been connected to multidrug resistance in breast cancer treatment. It was shown that the efficacy of siRNA delivery increased significantly with hydrophobic lipid substitutions ranging from C8 to C18, and that LA- and CA-substituted PEIs were the most effective in both cellular uptake and BCRP downregulation. Sun et al. using MD simulations explored the role of lipid substitutions on PEI/siRNA NPs. It was found that the substituted lipids enhanced the stability of PEI/siRNA NPs via association among them. Such linkages between the lipid side groups formed and broke frequently for the short lipids, whereas the associations between long lipids were more stable.

$N_c$ ($= n_c \times n_{\text{sub}}$) is a measure of hydrophobicity introduced to the PEI molecules. Same level of hydrophobicity can be obtained experimentally by adjusting either the level of substitutions per PEI ($n_{\text{sub}}$) or the number of C on each substitution ($n_c$). It is worth noting that backward elimination in Model type 3 eliminated the descriptor of $n_c$. This indicates that given $N_c$, $n_c$ itself does not play an important role in cellular uptake. Figure 4a shows the relation between $y$ and the two parameters of $N_c$ and $n_c$, based on experimental data. Reading the plot horizontally at a fixed value of $N_c$, the variation in $y$ was small suggesting its insensitivity to $n_c$. Reading the plot vertically, $y$ exhibited an overall increasing trend with $N_c$ for $n_c$ of 8 and 12, whereas no specific trend was observed for $n_c$ of 16 and 18.

The descriptor $N_{\text{uns}}$ was found to be significant to the prediction of cellular uptake, which is a measure of unsaturation level in the lipid substitutions. To further evaluate the role of $N_{\text{uns}}$, the value of $y$ as a function of $N_{\text{uns}}$ and $n_{\text{uns}}$ is shown in Figure 4b. The results suggested that for $n_{\text{uns}} = 2$, between 2 and 3 substitutions per PEI ($N_{\text{uns}} = (2-3) \times n_{\text{uns}} = 4-6$) resulted in higher $y$ value, and substitution level higher than 3 or lower than 2 caused a decrease in $y$. Because $n_{\text{uns}}$ was eliminated in backward elimination, the optimal $N_{\text{uns}}$ was expected to be in the same range for other $n_{\text{uns}}$. This cannot be seen directly from Figure 4b due to the

| Method | No. descriptors | Model type | Accuracy (train) | Accuracy (test) | $R^2$ (train) | $R^2$ (test) | RMSE (train) | RMSE (test) |
|--------|----------------|------------|-----------------|----------------|--------------|--------------|--------------|-------------|
| RF     | 26             | 1          | 0.63            | 0.59           | 0.70         | 0.41         | 0.56         | 0.77        |
| MLP    | 26             | 1          | 0.58            | 0.57           | 0.65         | 0.41         | 0.61         | 0.77        |
| LR     | 26             | 1          | 0.54            | 0.45           | 0.54         | 0.34         | 0.70         | 0.81        |
| RF     | 11             | 2          | 0.58            | 0.47           | 0.71         | 0.50         | 0.56         | 0.70        |
| MLP    | 11             | 2          | 0.52            | 0.55           | 0.51         | 0.43         | 0.72         | 0.75        |
| LR     | 11             | 2          | 0.47            | 0.53           | 0.41         | 0.35         | 0.79         | 0.80        |
| RF     | 6              | 3          | 0.60            | 0.47           | 0.69         | 0.46         | 0.57         | 0.73        |
| MLP    | 6              | 3          | 0.50            | 0.51           | 0.45         | 0.42         | 0.77         | 0.78        |
| LR     | 6              | 3          | 0.45            | 0.53           | 0.41         | 0.39         | 0.80         | 0.78        |
lack of data points to cover wider range of $n_{\text{max}}$ and $N_{\text{max}}$, especially since some data points were not considered during training.

Finally, $N_{\text{thio}}$ is another important descriptor which represents the total number of thioester groups on hydrophobically modified PEIs. This functional group can increase the sensitivity of NPs to dissociation due its labile nature under aqueous conditions. Strong electrostatic interaction between cationic polymers and nucleic acids has been shown to benefit the formation of a stable NP to deliver nucleic acids across cell membranes; however, it was also proposed to limit the efficacy of polymers due to insufficient unpacking of NPs in the cytoplasm. K.C. et al. showed that introducing electronegativity
provide predictions in closed form. This can be useful in practical settings where initial screening of the descriptors is done efficiently with relatively low but acceptable accuracy. Here, closed-form predictions of \( y \) generated from LR regressor are presented for each model type. The coefficient of each descriptor is shown in **Table 4**, and absolute value of the coefficient represents significance of the descriptor. For Model type 1 (26 descriptors in total, notations defined in Supporting Information, section S2), 6 descriptors had a coefficient whose absolute value was greater than 0.5. For Model types 2 (11 descriptors in total), 6 descriptors had a coefficient with absolute value greater than 0.13. For Model type 3 (6 descriptors in total), 3 descriptors had a coefficient with absolute value greater than 0.13. These coefficients are colored red in Table 4. All model types suggested that \( N_{\text{uns}} \) and \( N_{\text{thio}} \) were significant descriptors. This indicates that tLA is an important hydrophobic substitution, being the only substitution that carries thioester group on its structure. Also, for both Model type 2 and Model type 3, \( I_{\text{sub-1}} \) and \( I_{\text{sub-2}} \) were found to be significant.

Although these two descriptors cannot directly distinguish different types of hydrophobic substitutions, some information can be obtained from them. Examining the binary encoding (Table S3, Supporting Information), 7 substitutions had \( I_{\text{sub-1}} = 1 \) or \( I_{\text{sub-2}} = 1 \) including tLA, PA, StA, LA, \( \alpha \)LA, Chol, and OA, suggesting long lipids are better hydrophobic modification than short tail lipids (CA, Lau). Among the 6 cell lines, IAU565 and IMDA 468 had the smallest coefficient (in magnitude) generated by Model type 1, suggesting insensitivity of \( y \) with respect to these two cell lines.

The closed-form LR prediction of \( y \) using Model type 3 is given as follows:

\[
y = 0.10 r + 1.24 I_{\text{sub-1}} - 1 + 0.82 I_{\text{sub-2}} - 2 + 0.02 N_c - 0.05 N_{\text{uns}} - 0.20 N_{\text{thio}}
\]  

Figure 4. a) Experimental data (training and test data) for \( y \) as a function of \( n_c \) and \( n_c \). \( n_c \) values correspond to CA (8), Lau (12), PA (16), and LA (C18). b) Experimental data for \( y \) as a function of \( N_{\text{uns}} \) and \( n_{\text{uns}} \). \( n_{\text{uns}} \) values correspond to OA (1), LA (2), and \( \alpha \)LA (3).
to the aforementioned beneficiary effect of thioester group on cellular uptake. However, it should be pointed out that the descriptors are interconnected; an increase in \( N_{\text{uns}} \) can also be associated with an increase in \( I_{\text{sub-1}} \) or \( I_{\text{sub-2}} \) which are accompanied by positive coefficients. Figure 5 shows representative model prediction for \( y \) by varying \( N_c \), \( N_{\text{uns}} \), and \( N_{\text{thio}} \). In these plots, \( r \) was set to 5, based on its most frequent value in the data set. Index indicators of \( I_{\text{sub-1}} \) and \( I_{\text{sub-2}} \) were assigned 1 and 0 because no hydrophobic substitutions in the data set had value of 1 for both two indices (Table S4, Supporting Information). Figure 3 shows that given \( N_{\text{thio}} \), low \( N_{\text{uns}} \), and high \( N_c \), resulted in higher \( y \) (log (cellular uptake)) into breast cancer cell lines. Such parametric study will help better design more potent PEI carriers for gene delivery applications.

### 3.6. Limitation and Future Perspectives

A common problem in the modeling of biological effects is the limited number of data points and high number of descriptors. This might lead to overfitting, where the model performs well on the training data but poorly on the test data. Many experimental studies are also typically performed under different conditions including cell medium, baseline polymer, polymer concentration, NP size and charge, number of cells, and instrument used to measure cellular uptake, to name a few. This issue can be addressed by performing systematic experimental studies to generate a large data set, and normalizing the data into unitless quantities to minimize differences in measurements. The problem of large number of descriptors can be tackled by proper descriptor selection for model construction as used in our study through Model type 3.

The models generated here can be used to predict the cellular uptake of PEI/siRNA NPs where the parameters of hydrophobic substitutions are new and untested. Reliable predictions can be made within the ranges of the utilized descriptors. The accuracy of utilized algorithms in this work was about 60% which had room for improvement. To enhance the model predictions, new experimental data, especially on the types of hydrophobic substitutions with low number of data points in the current data set, should be added (i.e., wider range of unsaturated lipids with various level of substitutions (Figure 4b)), and then the models need to be retrained. Another approach is to find more informative descriptors that provide additional information regarding uptake of PEI/siRNA NPs into cancer cell lines. An example of such descriptors can be the toxicity level PEI/siRNA NPs. As new descriptors are added to the models, backward elimination should be used to test whether the new descriptors are statistically significant. In addition, our model suggested that cellular uptake of PEI/siRNA NPs was insensitive to AU565 and MDA 468 cell lines which needs to be verified with experiments.

Computational methods that give better model predictions are usually in some form of artificial neural networks, or other nonlinear methods such as the RF used in our study. Selection of methods must take into consideration factors such as the complexity of NPs, number of descriptors affecting cellular uptake, and complexity of the relation between descriptors and cellular uptake. The advantage of nonlinear methods is that they can find more complex relation between descriptors and the target (here, cellular uptake), thereby providing better prediction. However, the disadvantage is that such methods cannot separately describe the contribution of individual descriptors and can only be used as a black box algorithm. The simpler LR can provide direct information on the contribution of each descriptor, at the cost of losing accuracy. To improve the performance of linear methods, more complex descriptors can be generated by imposing nonlinearity. For example, through multiplication of two descriptors as utilized in our study (\( N_c \), \( N_{\text{uns}} \), and \( N_{\text{thio}} \)).

The current data set, originally generated from novel lipid modified PEIs developed in our lab, provided a basis for applying machine learning algorithms into the prediction of cellular uptake of PEI/siRNA NPs. This data set can be expanded to address more complex scientific questions. As a representative example, the purpose of delivering PEI/siRNA NPs is to downregulate specific mRNAs involved in signaling, thereby reducing the intracellular levels of associated proteins. By altering the nucleotide sequence of siRNA molecules, one

---

**Table 4. Coefficients generated from LR using the three model types.**

| Descriptor | Coefficient (Model type 1) | Coefficient (Model type 2) | Coefficient (Model type 3) |
|-----------|---------------------------|---------------------------|---------------------------|
| \( M_{\text{w}} \)  | 0.39 | – | – |
| \( r \) | 0.09 | 0.10 | 0.10 |
| \( I_{\text{sub-1}} \) | – | 0.99 | 1.24 |
| \( I_{\text{sub-2}} \) | – | 0.58 | 0.82 |
| \( BC_{50} \) | 0.52 | 0.14 | – |
| \( n_{\text{uns}} \) | 0.03 | –0.02 | – |
| \( n_{\text{thio}} \) | 0.05 | 0.01 | – |
| \( N_{\text{c}} \) | 0.01 | 0.02 | 0.02 |
| \( N_{\text{uns}} \) | –0.11 | –0.10 | –0.05 |
| \( n_{\text{uns}} \) | 0.79 | 0.16 | – |
| \( n_{\text{thio}} \) | 0.05 | 0.37 | – |
| \( N_{\text{thio}} \) | –0.10 | –0.22 | –0.20 |
| \( I_{\text{hene}} \) | 0.20 | – | – |
| \( I_{\text{CA}} \) | –0.37 | – | – |
| \( I_{\text{au}} \) | –0.12 | – | – |
| \( I_{\text{PA}} \) | 0.33 | – | – |
| \( I_{\text{SA}} \) | 0.57 | – | – |
| \( I_{\text{LA}} \) | –1.15 | – | – |
| \( I_{\text{ID}} \) | –1.20 | – | – |
| \( I_{\text{choi}} \) | –0.10 | – | – |
| \( I_{\text{DA}} \) | –0.68 | – | – |
| \( I_{\text{CA}} \) | 0.047 | – | – |
| \( I_{\text{MDA MB231}} \) | 0.32 | – | – |
| \( I_{\text{MCF7}} \) | –0.50 | – | – |
| \( I_{\text{AUS65}} \) | –0.001 | – | – |
| \( I_{\text{MDA 468}} \) | –0.02 | – | – |
| \( I_{\text{MDA 435}} \) | 0.33 | – | – |
| \( I_{\text{MDA 231}} \) | –0.13 | – | – |
can silence a broad range of mRNAs for specific therapies. In the case of breast cancer cells, a wide range of targets were identified for silencing to control uncontrolled growth of the cells.\cite{35,36} The efficacy of siRNA silencing depends on a functional delivery system that can transport siRNA into cells and release it into the cytoplasm to exert its effect. A new data set can be built with the efficacy of siRNA silencing as a new target. Our current descriptors can continue to be used, with the addition of cellular uptake as a new descriptor. Such a model can provide valuable information on what polymer modifications result in better delivery vector for silencing the target proteins.

### 4. Conclusions

Machine learning algorithms were used to model the uptake of PEI/siRNA NPs into breast cancer cells. A large data set was compiled composed of experimental values for the uptake of LMW PEI/siRNA NPs into six breast cancer cell lines, where the PEIs were modified with different hydrophobic substitutions. Three different regression models, namely, RF, MLP, and LR, as well as various descriptors were used to construct the models. RF and MLP regression methods were found to have a better performance, suggesting that nonlinear models were better estimators than the linear model to predict the cellular uptake of PEI/siRNA NPs. Descriptors that had major contribution to uptake included PEI-to-siRNA weight ratio, type of hydrophobic substitution, as well as the total number of C, thioester groups, and unsaturated C on the substitutions in each PEI. Our results suggested that the type of hydrophobic substitution on PEI molecules plays a crucial role in the delivery of PEI/siRNA NPs into breast cancer cells. Our study was the first report of quantitative modeling and prediction of cellular uptake of PEI/siRNA NPs, using chemically interpretable descriptors, which will facilitate the design of more efficient gene delivery systems that improve cancer therapeutics.

### Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

### Acknowledgements

Compute Canada and Westgrid were gratefully acknowledged for providing the computing resources and technical support. This work was funded by Natural Sciences and Engineering Research Council of Canada (NSERC), Canadian Institutes of Health Research (CIHR), and Alberta Innovates-Technology Futures (AITF).

### Conflict of Interest

The authors declare no conflict of interest.

### Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.
Keywords
breast cancer, gene delivery, machine learning, nanoparticles, polyethylenimine

Received: December 8, 2020
Revised: April 15, 2021
Published online:

[1] H. Uludag, A. Ubeda, A. Ansari, Front. Bioeng. Biotechnol. 2019, 7, 131.
[2] K. A. Whitehead, R. Langer, D. G. Anderson, Nat. Rev.Drug Discov. 2009, 8, 29.
[3] H. M. Aliabadi, R. B. Kc, E. Bousoik, R. Hall, A. Barbarino, B. Thapa, M. Coyle, P. Mahdipoor, H. Uludag, Acta Biomater. 2020, 102, 351.
[4] Y.-D. T. Tseng, P.-F. Liu, J.-Y. Li, L.-F. Liu, S.-Y. Kuo, C.-W. Hsieh, C.-H. Lee, C.-H. Wu, M. Hsiao, H.-T. Chang, C.-W. Shu, Front. Pharmacol. 2018, 9, 1285.
[5] S. M. Garg, A. Falamarzian, M. R. Vakili, H. M. Aliabadi, H. Uludag, A. Lavasanifar, Nanomedicine 2016, 11, 2319.
[6] C. Sun, T. Tang, H. Uludag, Biomacromolecules 2012, 13, 982.
[7] A. P. Pandey, K. K. Sawant, Mater. Sci. Eng. C 2016, 68, 904.
[8] D. Meneksedag-Erol, ACS Appl. Mater. Interfaces 2015, 7, 24822.
[9] Y. Nademi, T. Tang, H. Uludag, Nanoscale 2018, 10, 17671.
[10] V. Incani, A. Lavasanifar, H. Uludag, Soft Matter 2010, 6, 2124.
[11] H. M. Aliabadi, B. Landry, C. Sun, T. Tang, H. Uludag, Biomaterials 2012, 33, 2546.
[12] D. Wang, A. S. Narang, M. Kotb, A. O. Gaber, D. D. Miller, S. W. Kim, R. I. Mahato, Biomacromolecules 2002, 3, 1197.
[13] P. Y. Teo, C. Yang, J. L. Hedrick, A. C. Engler, D. J. Coady, S. Chaem-Maghami, A. J. T. George, Y. Y. Yang, Biomaterials, 34, 7971, 2013.
[14] A. Neammark, O. Suwantong, Mol. Pharm. 2009, 6, 1798.
[15] C. Sun, T. Tang, H. Uludag, Biomaterials 2013, 34, 2822.
[16] C. Sun, T. Tang, H. Uludag, J. Phys. Chem. B 2012, 116, 2405.
[17] R. Perkins, H. Fang, W. Tong, W. J. Welsh, Environ. Toxicol. Chem. An Int. J. 2003, 22, 1666.
[18] R. Weissleder, K. Kelly, E. Y. Sun, T. Shtatland, L. Josephson, Nat. Biotechnol. 2005, 23, 1418.
[19] D. Fourches, D. Pu, C. Tassa, R. Weissleder, S. Y. Shaw, R. J. Mumper, A. Tropsha, ACS Nano 2010, 4, 5703.
[20] D. A. Winkler, F. R. Burden, B. Yan, R. Weissleder, C. Tassa, S. Shaw, V. C. Epa, SAR QSAR Environ. Res. 2014, 25, 161.
[21] H. Montazeri Aliabadi, B. Landry, P. Mahdipoor, H. Uludağ, Mol. Pharm. 2011, 8, 1821.
[22] D. N. Meenakshi Sundaram, C. Kucharski, M. B. Parmar, Macromol. Biosci. 2017, 17, 1600430.
[23] H. M. Aliabadi, R. Maranchuk, C. Kucharski, P. Mahdipoor, J. Hugh, H. Uludag, J. Control. Release 2013, 172, 219.
[24] S. Plianwong, B. Thapa, R. B. Kc, C. Kucharski, T. Rojanarata, H. Uludag, Pharm. Res. 2020, 37, 46.
[25] S. Xiang, H. Tong, Q. Shi, J. C. Fernandes, T. Jin, K. Dai, X. Zhang, J. Control. Release 2012, 158, 371.
[26] B. Thapa, Int. J. Cancer 2018, 142, 597.
[27] X. Zhang, Y. Yu, L. Wang, Q. Gu, in The 22nd International Conf. on Artificial Intelligence and Statistics, 2019, pp. 1524–1534.
[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, J. Mach. Learn. Res. 2011, 12, 2825.
[29] M. L. Waskom, J. Open Source Softw. 2021, 6, 3021.
[30] M. Boeckle, J. Gene Med. 2004, 6, 1102.
[31] H. M. Aliabadi, B. Landry, P. Mahdipoor, C. Y. M. Hsu, H. Uludag, Eur. J. Pharm. Biopharm. 2012, 81, 33.
[32] D. J. Gary, N. Puri, Y.-Y. Won, J. Control. Release 2007, 121, 64.
[33] A.-L. Bolcato-Bellemin, M.-E. Bonnet, G. Creusat, P. Erbacher, J.-P. Behr, in Proc. Natl. Acad. Sci. 2007, 104, 16050–16055.
[34] R. B. Kc, C. Kucharski, H. Uludag, J. Mater. Chem. B 2015, 3, 3972.
[35] S.-X. Lin, J. Chen, M. Mazumdar, D. Poirier, C. Wang, A. Azzi, M. Zhou, Nat. Rev. Endocrinol. 2010, 6, 485.
[36] C. V. Pecot, G. A. Calin, R. L. Coleman, G. Lopez-Berestein, A. K. Sood, Nat. Rev. Cancer 2011, 11, 59.