Electronic Supplementary Information

A small data-driven predictive model for adsorption properties in polymeric thin film

Uiyoung Han[a] Taegyu Kang[b], Jongho Im[*,b] and Jinkee Hong[*,a]

[a]Department of Chemical & Biomolecular Engineering, College of Engineering, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea

[b]Department of Applied Statistics, Yonsei University, Seoul, 03722, Republic of Korea

1. Experiment detail

The polymer film was formed by self-assembly between negatively charged polyacrylic acid (PAA) and positively charged polyallylamine hydrochloride (PAH). First, the silica substrate was treated by O2 plasma (CUTE-1B, Femtoscience, Yongin, Korea) for 2 min to form a negatively-charged surface. Then, the film was formed via a layer-by-layer dip-coating method. In this procedure, the PAH layer was formed by immersing the substrate in PAH solution (1 mg mL−1) for 10 min. Then, the substrate was rinsed three times (2, 1, and 1 min) with deionized water. Next, the PAA layer was adsorbed onto the substrate in the same manner using 1 mg mL−1 of PAA solution. The pH conditions of each solution were varied from 4 to 11. The amount of adsorbed polymer was measured using a quartz crystal microbalance (QCM; QCM200, Stanford Research Systems, Sunnyvale, CA, USA). The Sauerbrey equation was used to convert the oscillation frequency of the crystal to the mass.

2. Data generation

The first step in converting the experimental results to data involves selecting and defining the variables to be used on the small-data management (SDM) platform (**Fig. S1**). The independent variables denoted as Xi include factors that can be regulated by the researcher, such as the experimental conditions and intrinsic properties of the polymer. The transformed variables denoted as Tj are composed of factors calculated from Xi according to a theoretical relationship. Finally, the response variables denoted as Yk are measured values as the adsorption properties of polymer. In this study, the pH (X1), the number of layers (X2), and the molecular weight (X3) were selected as the independent variables, and the ionization (T1), the adhesion force (T2), and the radius of gyration (T3) were selected as the transformed variables. The amount of adsorbed polymer (Y1) was selected as the response variable influenced by the predictor variables (X, T). The second step in the data conversion process is that of generating the values of the variable X and T. The pH value (X1) of each polyelectrolyte solution (1 mg mL^{-1}) was measured by pH meter (HI 2211, Hanna Instruments). The pH was controlled using 0.1 M HCl and 0.1 M NaOH solutions. The number of layer (X2) expresses the number of PAA/PAH bilayer on substrate, and ranges from 1 to 3. The molecular weight (X3) of each polyelectrolyte indicates weight average molecular weight (Mw) of PAA and PAH. The value of T is calculated from the results or formula demonstrated in previous studies (**Fig. S2**). The degree of ionization (T1) of polyelectrolyte is the capacity of acid/base to ionize itself. It was calculated by substituting the pKa value and pH of the polymers used (PAA and PAH) into formula (1, 2). For PAA, pKa value was assumed to be 6.5 and for PAH, pKa value was assumed to be 8.5.

$$\%\, Ionized\ (weak\ acid) = \frac{1}{1 + 10^{(pKa\, -\, pH)}} \times 100 \tag{1}$$

$$\%\ Ionized\ (weak\ base) = \frac{1}{1 + 10^{(pH - pKa)}} \times 100 \tag{2}$$

Since the adhesion force (T2) between two polyelectrolytes is electrostatic interaction, T2 is directly related to the adsorption properties thereof. It was calculated from the relationship between the degree of ionization of PAA (T1_PAA) and that of PAH (T1_PAH) obtained from Cranford et al.[1] Finally, the radius of gyration of polymer (T3) indicates the configurational properties important in the behavior of a polymer chain in solution, and it is related to the adsorption property. T3 calculated by substituting the molecular weight (degree of polymerization, *N*) and degree of ionization (factor that determines the value of ν) is into formula (3) obtained by Mintis et al.[2]

$$R_g{\sim}N^v \tag{3}$$

Unlike the existing experiment analysis manner in which the number of variables are reduced to investigate the influence of specific variable, the SDM platform designs predictive models by increasing the number and types of variables. In lab-scale experiment generating small dataset, diversifying the experimental factors can be useful to track the effects of complex factors occurring in an actual experiment. So, we expects that the independent variables (X1, X2, X3) will be as suitable for predicting actual experimental results as or more than the dependent variables (T1, T2, T3) calculated assuming the ideal behavior of the polymer.

3. Regression models

The algorithm for predicting the experimental output in the SDM platform is designed based on the conventional probabilistic and statistical methodologies. Among them, regression models and machine learning algorithms that are suitable for supervised learning and facile to add and delete variables are used in the platform. The platform users should design optimized algorithms based on theoretical and experimental knowledge about variable relationships, rather than intactly utilize conventional regression models as a fully automated platform. This is an important strategy of the SDM platform to draw conclusions consistent with scientific facts from small data analysis. So, we selected several regression models suitable for small data analysis of the adsorption property of polymer, then design the optimized algorithms of the predictive model.

In order to design a regression algorithm suitable for the dataset of PAA/PAH complexes, it is necessary to first identify the characteristics of the predictor and response variables. Hence, as shown schematically in **Fig. S3a** the characteristics of the data obtained from our experiment were divided into the following four major Cases: (1) a small amount of data, (2) a non-linear variable relationship, (3) a normal distribution, and (4) a categorical distribution. With respect to Case (1), the majority of manual experiments in which one dataset is created in one experimental cycle have the common feature that the amount of generated data is insufficient for machine learning. In Case (2), the relationship between, for example, the pH (X1) of the solvent and the properties (T1, T2) of the polymer, and that between the molecular weight (X3) and the amount of adsorption (Y1), are non-linear. As an example of Case (3), molecular weight (X3) of synthetic polymers generally display a normal distribution. Also, the measurement error of the adsorption amount (Y1) exhibit approximately normal distributions when considered as the sum of many independent processes. Finally, as an example of Case (4), the data obtained from the adsorption amount (Y1) follow a categorical distribution according to the number of layers (X2).

As shown in **Fig S3b**, the next step is to select the appropriate regression model by considering the data characteristics. For a small dataset (Case 1) where the data are not enough to estimate the model structure, parametric methods may be more suitable, such as a linear model (LM) and a generalized linear mixed model (GLMM).[3] The parametric model also performs well when the data follow a normal distribution (Case 2).[4] On the other hand, non-parametric methods, such as the random forest (RF), the gradient boosting model (GBM), and the multivariate adaptive regression spline (MARS), are suitable for analyzing datasets that do not follow a probability distribution.[5] Although the non-parametric method is prone to overfitting in small data analysis, they often show better results than parametric methods for non-linear data (Case 3) or categorical data (Case 4) analysis.[6, 7] In addition, semi-parametric methods such as the generalized additive model (GAM) and the support vector machine (SVM) can be considered. When estimating the distribution of data that includes both linear and non-linear relationships of variables, the semi-parametric methods have the potential to provide a better performance.[8] These regression algorithms are embodied by using the defined predictor variables, in which a different statistical function can be applied to each variable in consideration of the variable characteristics (**Fig. S3c**). For example, the adsorption properties of PAA/PAH complex follow a categorical distribution according to the number of layers (X2), so a logit link function was applied to the LM, GLMM, and GAM. In GLMM, we applied a random effect to pH of solution (X1), whose value may change during the experiment. We also implemented the GAM as an algorithm with both logit link function and random effect term applied. In order to automate this process, it is necessary to design automated workflow system according to the combination of all variables and statistical functions used for each regression model. Since this study is a proof-of-concept study, we confirmed the validity of the strategy by comparing and evaluating the performance of predictive models with fixed combination of statistical functions and predictor variables.

4. Regression algorithms

Linear model: The model was designed to fit the mean response E(Y|X) to a linear function of X. To keep the fitted values positive, a generalized linear model with a logit link function was used. Specifically, log E(Y|X) was fitted to the linear function and the predicted value was calculated by taking an exponential function of the fitted linear model. The mean response is given by the following expression:

$$\text{Log}\, E(Y \mid X) = \beta_1 X1.PAA + \ \beta_2 X1.PAH + \ \beta_3 X2 + \beta_4 X3.PAA + \ \beta_5 X3.PAH$$

Thus, the corresponding mean response function is:

$$E(Y \mid X) = exp(\beta_1 X1.PAA + \ \beta_2 X1.PAH + \ \beta_3 X2 + \beta_4 X3.PAA + \ \beta_5 X3.PAH)$$

The expressions for the model using the transformed variables T1 from the analysis are:

$$\log E(Y \mid X) = \beta_1 T1.PAA + \ \beta_2 T1.PAH + \ \beta_3 X2 + \beta_4 X3.PAA + \ \beta_5 X3.PAH$$

and

$$E(Y \mid X) = exp(\beta_1 T1.PAA + \ \beta_2 T1.PAH + \ \beta_3 X2 + \beta_4 X3.PAA + \ \beta_5 X3.PAH$$

The generalized linear mixed effect model (GLMM): The generalized linear mixed effect model combines fixed effects and mixed effects into a single model. Fixed effects are parameters that are regarded as constants, whereas random effects are parameters that are regarded as random variables. In the present work, random effects were applied with respect to the pH value of solution, and the model assume the random effects fell on the intercept. The mean response was then fitted as the following:

$$\log E(Y \mid X) = \sum a_i I(X1.PAA) + \ \sum a_j I(X1.PAH) + \beta_1 X2 + \ \beta_2 X3.PAA + \ \beta_3 X3.PAH$$

and

$$E(Y \mid X) = exp(\sum a_i I(X1.PAA) + \sum a_j I(X1.PAH) + \beta_1 X2 + \beta_2 X3.PAA + \beta_3 X3.PAH)$$

The fitted model with the transformed variables T1 is given by:

$$\log E(Y \mid X) = \sum a_i I(T1.PAA) + \sum a_j I(T1.PAH) + \beta_1 X2 + \beta_2 X3.PAA + \beta_3 X3.PAH$$

and

$$E(Y \mid X) = exp(\sum a_i I(T1.PAA) + \sum a_j I(T1.PAH) + \beta_1 X2 + \beta_2 X3.PAA + \beta_3 X3.PAH)$$

where $a_{i,j}$ denotes the random effects on the intercept.

The generalized additive model (GAM): The proposed generalized additive model (GAM) employs the same additive method as the linear model, thus contributing to an explanation of the effects of the variables included in the predictive model.[9] Unlike the linear model, however, the GAM allows much more flexibility for each additive function. More precisely, the GAM fits the mean response function E(Y|X) to the sum of the smooth functions in each predictor. The mean response is given by the following expression:

$$\log E(Y \mid X) = \sum \alpha_i I(X1.PAA) + \sum \alpha_j I(X1.PAH) + s(X2) + s(X3.PAA) + s(X3.PAH)$$

Hence, the mean response function is given by:

$$E(Y \mid X) = exp(\sum \alpha_i I(X1.PAA) + \sum \alpha_j I(X1.PAH) + s(X2) + s(X3.PAA) + s(X3.PAH))$$

and the expressions for the model using the transformed variables from the analysis are:

$$\log E(Y \mid X) = \sum \alpha_i I(T1.PAA) + \sum \alpha_j I(T1.PAH) + s(X2) + s(X3.PAA) + s(X3.PAH)$$

and

$$E(Y \mid X) = exp(\sum \alpha_i I(T1.PAA) + \sum \alpha_j I(T1.PAH) + s(X2) + s(X3.PAA) + s(X3.PAH))$$

The random forest, decision tree, and ensemble learning methods: The random forest method[10] is an ensemble learning method applied to the decision tree method, where the latter is a simple and easy-to-use learning method that iteratively partitions a data set by choosing predictor variables and applying a partitioning rule at each node. Ensemble learning is a machine learning method for improving an otherwise inaccurate learning rule. In the present study, a random forest was constructed via an ensemble method known as bootstrap aggregation or bagging.[11] The algorithm for fitting the random forest is the following:[12]

1) For b = 1 to B:

    A. draw a bootstrap sample Z^* of size N from the training data;

    B. grow a random-forest tree T_b from the bootstrapped data by recursively repeating steps (i) to (iii) for each terminal node of the tree until the minimum node size $n_{\{min\}}$ is reached:

        i. select m variables at random from the p variables,

        ii. pick the best variable/split-point among these, and

        iii. split the node into two daughter nodes.

2) Output the ensemble of trees $\{T_b\}_1^B$

To make a prediction at a new point x, the following regression and classification steps are applied:

Regression: $f_{rf}^{B}(x) = \frac{1}{B}\sum T\_b(x)$

Classification: let $C_b(x)$ be the class prediction of the b^th random-forest tree.

Then: $C_{rf}^{B}(x) = majority\ vote\ \{C_b(x)\}_1^B$

Gradient boosting: Boosting is another ensemble method for updating a weak learning rule into a stronger one.[13-15] AdaBoost proposed a method called gradient boosting, which is a gradient descent method in function space. Usually, a tree model is chosen as a basis weak learner. The algorithm for gradient tree boosting is the following:

```
1) initialize: $f_0 = \arg min \sum L(y_i, \gamma)$;
2) for m = 1 to M:
    A. for i = 1, 2, …, N, compute
$r_{im} = -[\partial L(y_i, f(x_i)) / \partial f(x_i)]$ at $f = f_{\{m-1\}}$;
    B. fit a regression tree to the targets $r_{im}$ to give terminal regions $R_{jm}$, $j = 1, 2, ..., J_m$;
    C. for j = 1, 2, …, $J_m$, compute
$\gamma_{jm} = argmin \sum L(y_i, f_{\{m-1\}}(x_i) + \gamma)$, where the summation ranges over $x_i \in R_{jm}$;
    D. update: $f_m(x) = f_{\{m-1\}}(x) + \sum \gamma_{jm} I(x \in R_{jm})$.
Output: $\hat{f}(x) = f_M(x)$.
```

Multivariate adaptive regression splines (MARS): Multivariate adaptive regression splines (MARS) is an adaptive regression method for constructing a stepwise additive model consisting of splines and the product of two splines.[16] The model fitting strategy of MARS is similar to that of a forward stepwise linear regression, but uses the functions constructed from

$\{(X_j - t)_+, (t - X_j)_+\}$ where $t \in \{x_{1j}, x_{2j}, ...x_{Nj}\}$ $and\ j = 1, 2, ..., p$,

where p is the number of predictors, and pairwise products of the above function (which are also functions). Thus, the fitted model has the form:

$$f(X) = \beta_0 + \Sigma\beta_m h_m(X),$$

where each h_m is a basic function of the type described above.

The support vector machine (SVM): The support vector machine is a method for discriminating the predictor space by fitting an affine linear hyperplane onto the predictor space. Among all possible hyperplanes, the one that generates the maximal margin is selected. A fundamental strategy for fitting the support vector machine model is given by the following:

$\arg\min \frac{1}{2}|w|^2$ subject to $|y_i - w^T x_i - \beta_0| \le \epsilon$,

Where, ϵ is a pre-determined rate of error.

5. References

1. S. W. Cranford, C. Ortiz and M. J. Buehler, *Soft Matter*, 2010, **6**, 4175-4188.
2. D. G. Mintis and V. G. Mavrantzas, *J. Phys. Chem. B*, 2019, **123**, 4204-4219.
3. E. Eide and H. Gish, 1996.

4. A. Azzalini and A. Capitanio, *J. Roy. Stat. Soc. Ser. B. (Stat. Method.)*, 1999, **61**, 579-602.

5. E. L. Lehmann and C. Stein, *Ann. Math. Stat.*, 1949, **20**, 28-45.

6. P. Pérez-Rodríguez, D. Gianola, J. M. González-Camacho, J. Crossa, Y. Manès and S. Dreisigacker, *G3: Genes| Genomes| Genetics*, 2012, **2**, 1595-1605.

7. G. G. Koch, C. M. Tangen, J. W. Jung and I. A. Amara, *Stat. Med.*, 1998, **17**, 1863-1892.

8. R. J. Smith, *Econ. J.*, 1997, **107**, 503-519.

9. T. J. Hastie and R. J. Tibshirani, *Generalized additive models*, CRC press, 1990.

10. L. Breiman, *Mach. Learn.*, 2001, **45**, 5-32.

11. L. Breiman, *Mach. Learn.*, 1996, **24**, 123-140.

12. T. Hastie, R. Tibshirani and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, Springer Science & Business Media, 2009.

13. R. E. Schapire, *Mach. Learn.*, 1990, **5**, 197-227.

14. Y. Freund and R. E. Schapire, 1996.

15. J. H. Friedman, *Ann. Stat.*, 2001, 1189-1232.

16. J. H. Friedman, *Ann. Stat.*, 1991, 1-67.

7. Figures

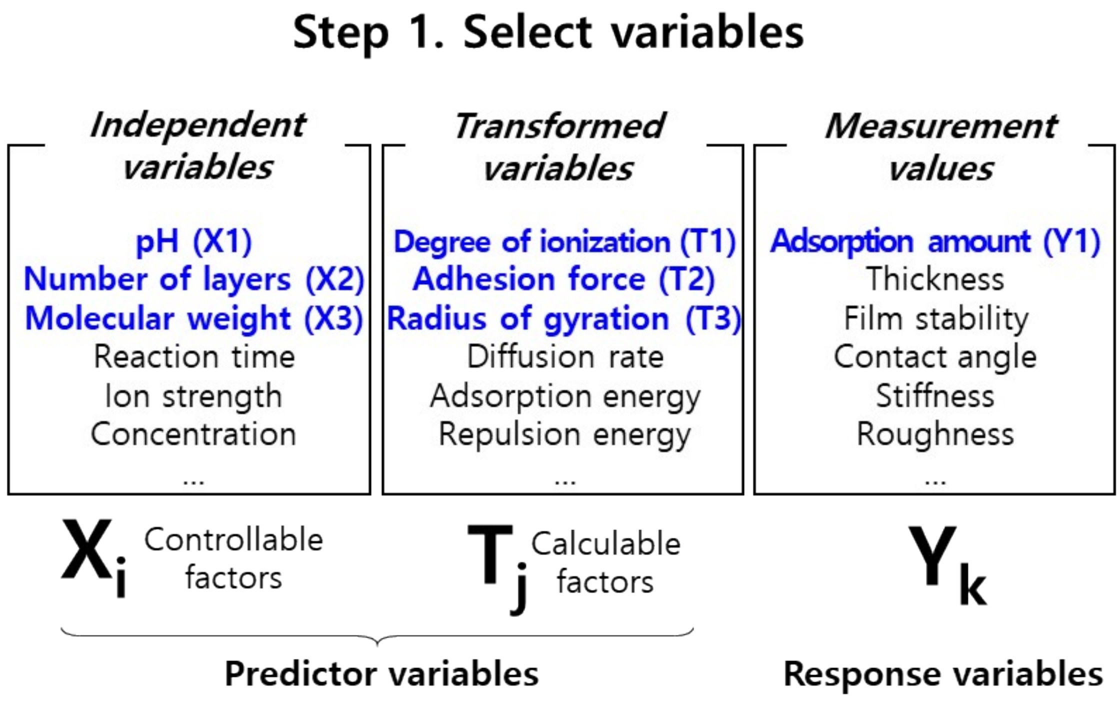


Fig. S1. The predictor variables are composed of controllable factors (independent variables, X), including experimental conditions, and calculable factors (transformed variables, T); response variable (Y1) is selected as property of the polymer complexes measured in the experiment.

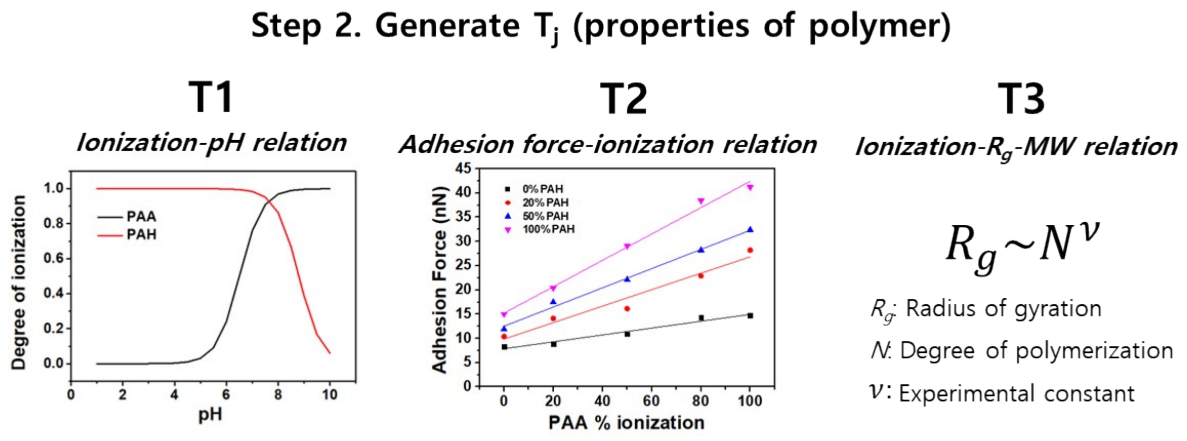


Fig. S2. The transformed variables T1, T2, T3 are obtained from the X-T relationships established using theoretical formulas.

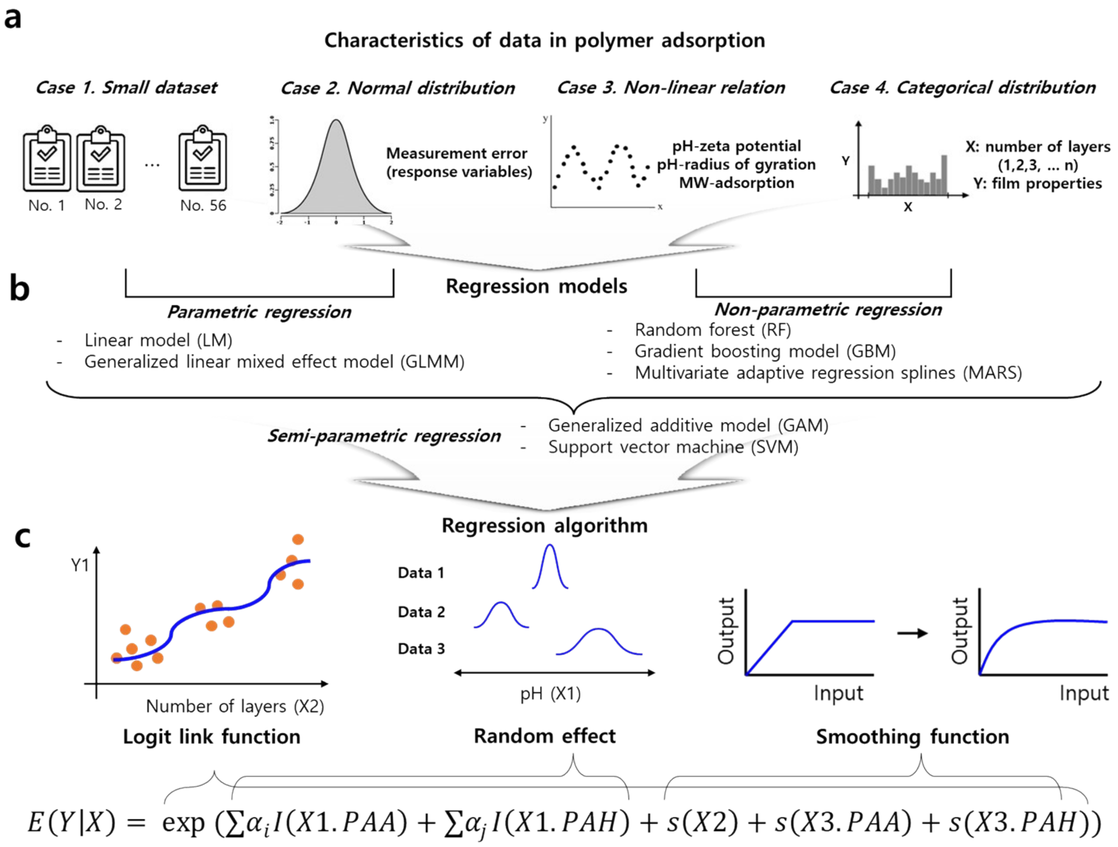
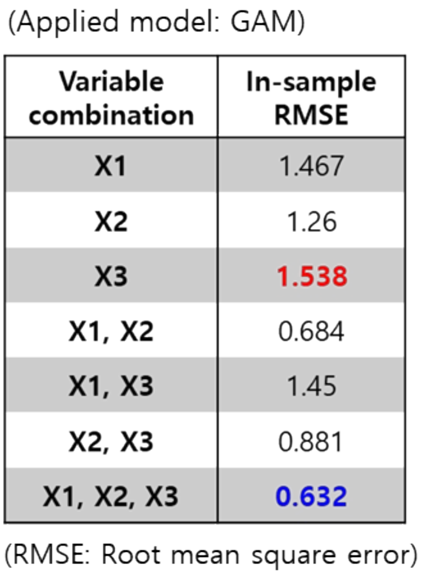

a
Characteristics of data in polymer adsorption
Case 1. Small dataset
No. 1
No. 2
No. 56
Case 2. Normal distribution
Measurement error (response variables)
Case 3. Non-linear relation
pH-zeta potential
pH-radius of gyration
MW-adsorption
Case 4. Categorical distribution
Y
X
X: number of layers (1,2,3, … n)
Y: film properties
b
Regression models
Parametric regression
- Linear model (LM)
- Generalized linear mixed effect model (GLMM)
Non-parametric regression
- Random forest (RF)
- Gradient boosting model (GBM)
- Multivariate adaptive regression splines (MARS)
Semi-parametric regression
- Generalized additive model (GAM)
- Support vector machine (SVM)
c
Regression algorithm
Y1
Number of layers (X2)
Logit link function
Data 1
Data 2
Data 3
pH (X1)
Random effect
Output
Input
Output
Input
Smoothing function
$E(Y|X) = \exp\left(\sum\alpha_i I(X1.PAA) + \sum\alpha_j I(X1.PAH) + s(X2) + s(X3.PAA) + s(X3.PAH)\right)$

Fig. S3. Schematic diagrams showing three preparatory works for regression analysis: (a) determining characteristics of the dataset; (b) screening suitable regression models; (c) designing regression algorithm. The data distribution is divided into four major Cases and three types of regression methods are considered according to their specific characteristics. The regression algorithm is embodied by applying regression model and function to predictor variables individually.

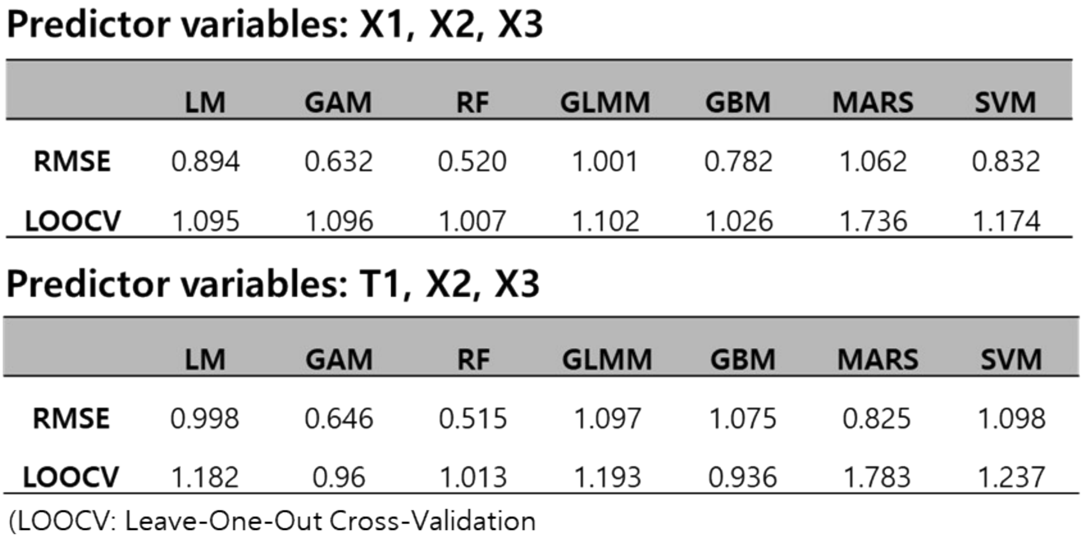


(Applied model: GAM)

Variable combination	In-sample RMSE
X1	1.467
X2	1.26
X3	1.538
X1, X2	0.684
X1, X3	1.45
X2, X3	0.881
X1, X2, X3	0.632

(RMSE: Root mean square error)

Fig. S4. Comparisons of the regression model performance via the in-sample root mean square error (RMSE) with different variable combinations.



Predictor variables: X1, X2, X3

	LM	GAM	RF	GLMM	GBM	MARS	SVM
RMSE	0.894	0.632	0.520	1.001	0.782	1.062	0.832
LOOCV	1.095	1.096	1.007	1.102	1.026	1.736	1.174

Predictor variables: T1, X2, X3

	LM	GAM	RF	GLMM	GBM	MARS	SVM
RMSE	0.998	0.646	0.515	1.097	1.075	0.825	1.098
LOOCV	1.182	0.96	1.013	1.193	0.936	1.783	1.237

(LOOCV: Leave-One-Out Cross-Validation

Fig. S5. Comparisons of the regression model performance via the leave-one-out cross-validation (LOOCV) with different variable combinations.

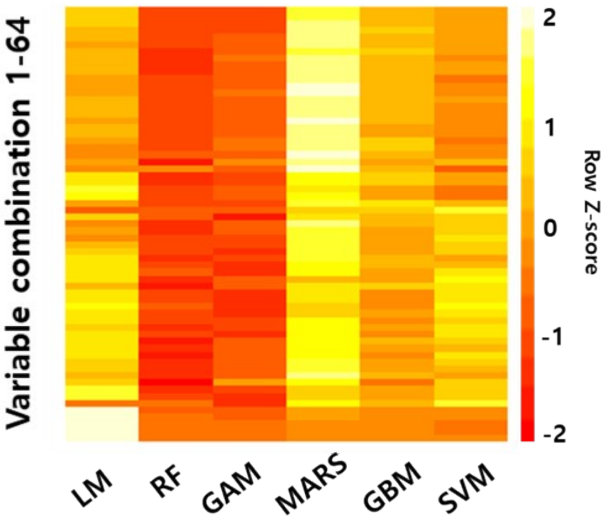


Fig. S6. Comparisons of the regression model performance via heatmapping of the regression model with different variable combinations.