Supplementary Information

Data-Augmented Response Surface Methodology-Machine Learning Hybrid Model for Predicting Polyvinyl Butyral Synthesis

Yingying Liu[1], Chi Ying Vanessa Li[3*], Liang Gao[1, 2*]

1. School of Chemical Engineering and Light Industry, Guangdong University of Technology, Guangzhou, 510006, P. R. China.
2. Jieyang Branch of Chemistry and Chemical Engineering Guangdong Laboratory (Rongjiang Laboratory), Jieyang 515200, P. R. China.
3. Technological and Higher Education Institute of Hong Kong

[*]**Corresponding Authors:** gaoliang@gdut.edu.cn; cyvli@thei.edu.hk

Supplementary Note I. Central Composite Design

By employing Design Expert 10 software, we delineated experimental variables along with their respective ranges, thereby obtaining a CCD experimental protocol**(Table S1).**

Table S1. AD and D_{43} of PVB under Different Acetalization Conditions.

NO.	A: PVA (wt%)	B: Ratio	C: HNO_3 (mol/L)	D: Time (min)
1	5	0.32	0.05	120
2	10	0.32	0.05	120
3	5	0.6	0.05	120
4	10	0.6	0.05	120
5	5	0.32	0.1	120

6	10	0.32	0.1	120
7	5	0.6	0.1	120
8	10	0.6	0.1	120
9	5	0.32	0.05	180
10	10	0.32	0.05	180
11	5	0.6	0.05	180
12	10	0.6	0.05	180
13	5	0.32	0.1	180
14	10	0.32	0.1	180
15	5	0.6	0.1	180
16	10	0.6	0.1	180
17	2.5	0.46	0.075	150
18	12.5	0.46	0.075	150
19	7.5	0.18	0.075	150
20	7.5	0.74	0.075	150
21	7.5	0.46	0.025	150
22	7.5	0.46	0.125	150
23	7.5	0.46	0.075	90
24	7.5	0.46	0.075	210
25	7.5	0.46	0.075	150
26	7.5	0.46	0.075	150
27	7.5	0.46	0.075	150
28	7.5	0.46	0.075	150
29	7.5	0.46	0.075	150
30	7.5	0.46	0.075	150

Supplementary Note II. Synthesis and Characterization of PVB

Polyvinyl alcohol (PVA) with the degree of alcoholysis of 98-99% was provided by Sinopec Chongqing SVW Chemical Co., Ltd., China. Analytical grade *n*-butanal, sodium dodecyl

sulfate (SDS), concentrated nitric acid (HNO_3) (98%), and sodium hydroxide (NaOH) were provided by Shanghai Macklin Biochemical Co., Ltd.

PVB was synthesized through a co-precipitation method involving a series of sequential procedures to obtain the target product. The synthesis process consists of a continuous acetalization reaction between PVA and n-butyraldehyde under acidic catalytic conditions, which evolves progressively over time. Taking the sixth experimental formulation **(Table 2)** as an example, the detailed synthesis procedure of PVB is outlined as follows.

(1) Dissolution of 10% PVA: Weigh 120 g PVA and transfer to a 2000 mL round-bottom flask. Add 1080 mL deionized water to the container, heat to 90°C, and maintain heating for 2 hours to form a homogeneous solution.

(2) Preparation of emulsifier: Place 0.4 g SDS in a test tube, add 10 mL deionized water, and dissolve completely in a 50-60°C oven for 30 min.

(3) PVA pretreatment: Place 200 mL cooled room-temperature PVA (10%) in a 500 mL round-bottom flask. Add SDS emulsifier, corresponding proportion of n-butyraldehyde (12.95 mL), control temperature at 28°C (monitored by temperature probe inserted into the solvent), and mix thoroughly with magnetic stirring at 300 rpm for 30 min.

(4) PVB synthesis reaction: Add concentrated nitric acid (1.34 mL) and allow sufficient reaction for 120 min.

(5) Neutralization reaction: Stop magnetic stirring, transfer the sample to a 500 mL beaker, add appropriate amount of prepared NaOH solution to the beaker, and allow solid-liquid separation after standing.

(6) Sample drying: Perform multiple vacuum filtrations to wash away impurities, label the samples, and dry in an oven at 45°C for 1-2 days.

Before measuring and characterizing the two target features of PVB acetalization degree and particle size, we need to obtain pure PVB samples to ensure that the measurement results are not interfered by impurities. Prior to the measurement and characterization of the two target

parameters—the degree of acetalization (AD) and the particle size (D_{43})—pure PVB samples must be obtained to eliminate potential interference from impurities during analysis. Specifically, we need to pretreat the mother liquor containing PVB products by adding a certain amount of NaOH solution to neutralize the mother liquor to pH 7, and then thoroughly wash the mother liquor 2-3 times by vacuum filtration and dry it for 1-2 days.

Through the above means, the possible impurities in the PVB products can be eliminated, and the treated PVB products can be used for characterization and analysis. Therefore, we can avoid the influence of non-target characteristic peaks in ^{1}H nuclear magnetic resonance spectrum analysis and the influence of impurity particle size in PVB particle size measurement.

The degree of acetalization (AD) measurement and data characterization of PVB were obtained through ^{1}H-NMR spectral analysis. Using d_6-DMSO as solvent, ^{1}H-NMR spectra of PVB were recorded on a nuclear magnetic resonance spectrometer and analyzed using MestraNova x64 software. AD of PVB was calculated through peak area integration. The ^{1}H-NMR chemical shift range is 0-5 ppm, and spectral data should not exceed this shift range. **Fig. S1** presents a representative ^{1}H-NMR spectral analysis of PVB. Based on the spectral interpretation, an AD value of 0.283 is observed. **Fig. S2** presents the NMR measurement data for AD of 30 PVB samples from the CCD design experiments. The integral calculation is shown as follows:

$$\bar{AD} = \frac{2}{3} \times \frac{A_8}{A_{2+4}} \quad (1)$$

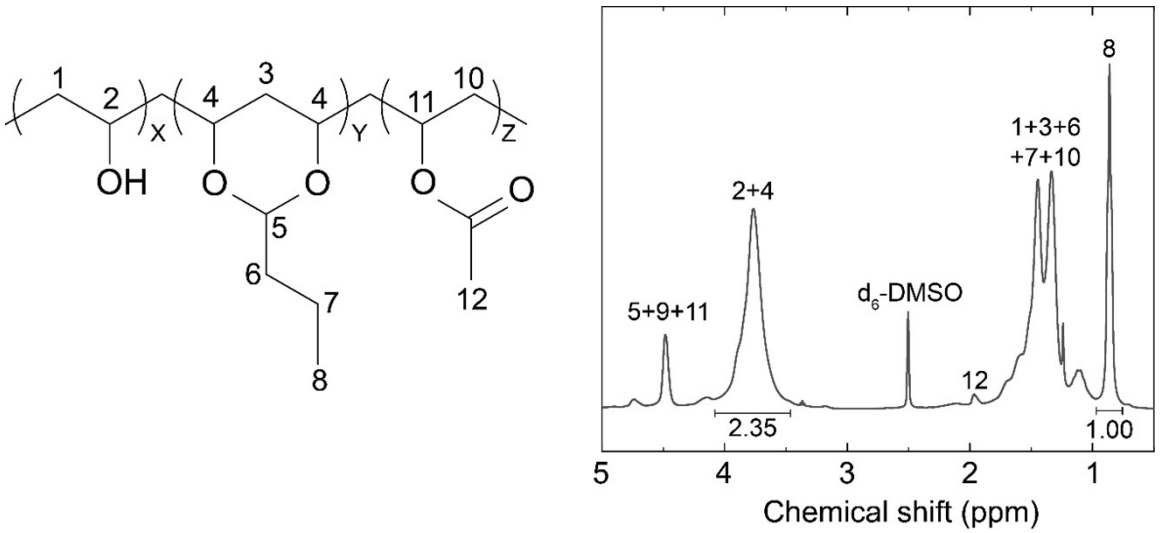


Figure S1. Typical ^{1}H-NMR spectrum of PVB. AD is defined as the molar ratio of substituted OH groups to original OH groups. Based on the segment ratios shown in **Fig. S1**, AD is defined as: AD = 2y/(x+2y+z).

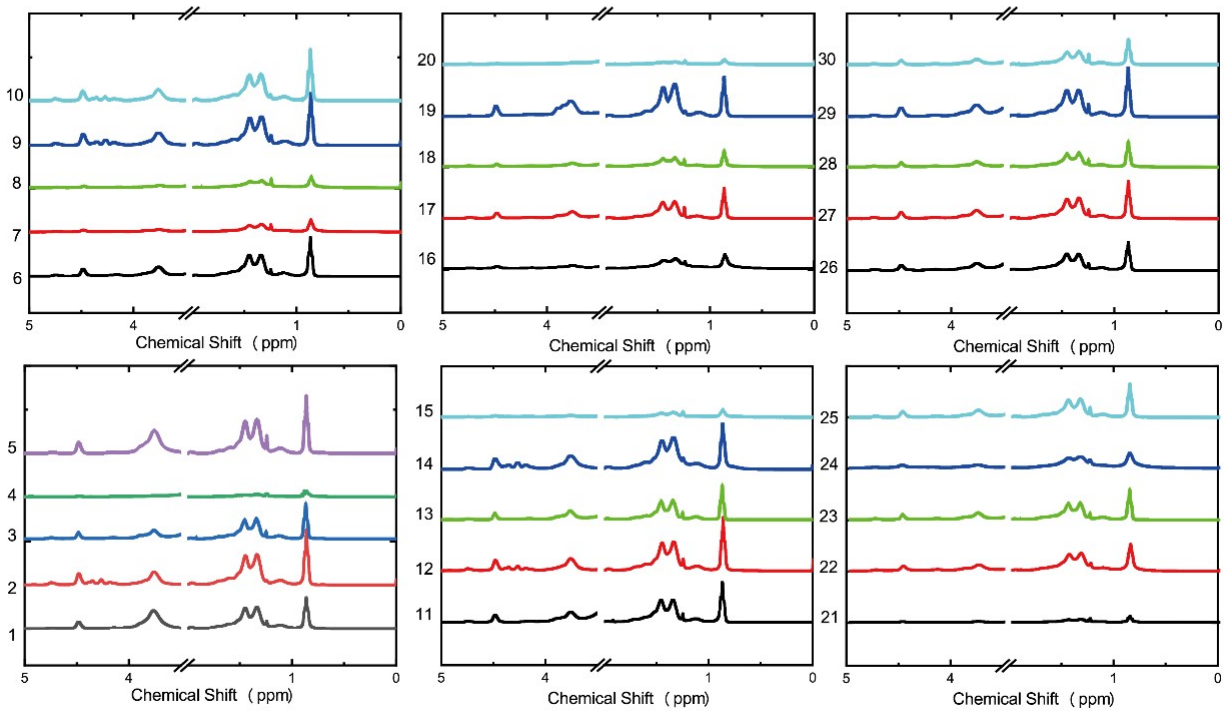


Figure S2. NMR spectral data for 30 samples from CCD experimental design.

In the ^{1}H NMR spectrum data related to the analysis of PVB acetalization degree, proton signals from different functional groups can be clearly distinguished by chemical shifts (δ).

Among these, PVB spectrum analysis indicates that the acetal proton (-O-C-O-) has a δ value of 3.5–4.2 ppm (multiple peaks, corresponding to the protons adjacent to the methylene group in the acetal structure); unreacted hydroxyl (-OH) has a δ value of 1.5–2.0 ppm (broad peak, affected by hydrogen bonding); the methylene group (-CH_2) of the PVA main chain has a δ value of 1.2–1.4 ppm (single peak); and the butyraldehyde residue (-$CH_2CH_2CH_2CH_3$) has a δ value of 0.8–1.0 ppm (triple peak, terminal methyl group). After removing the DMSO-d6 peak and water peak influences from all PVB spectra, the acetalization proton peaks and unreacted hydroxyl peaks were identified. The ^{1}H NMR spectrum analysis results fully confirm that PVA and n-butanal underwent an acetalization reaction. Additionally, AD of PVB and its distribution pattern were determined through analysis of CCD experiments under various conditions.

According to Mie scattering theory (applicable to arbitrary particle size to wavelength ratios) or Fraunhofer approximation (applicable when particle size is much larger than light wavelength), particle size analysis is performed by detecting scattered light patterns to reverse-calculate particle size. Based on the principle of laser diffraction, we initialized the Malvern 9000+ laser particle size analyzer. A small amount of the purified PVB product was placed in the sample cell to achieve a preset obscuration of 6% to 10%. The equipment was then activated to record and analyze the data, with the volume average particle size (D_{43}) serving as the standard for particle size measurement. **Table S2** presents the particle size analysis results, clearly showing various indicators of PVB product particle size data analysis. **Fig. S3** shows the particle size distribution of this PVB product, while **Fig. S4** presents the measurement data of 30 PVB samples from the CCD design experiments. The mathematical formula is shown as follows:

$$D_{43} = \frac{\sum\left(n_i \cdot d_i^4\right)}{\sum\left(n_i \cdot d_i^3\right)} \tag{2}$$

Where n_i is the number of particles in the i-th particle size interval, and d_i is the representative diameter of the i-th particle size interval.

Table S2. PVB Particle Size Analysis.

Particle Size Characteristics	Data Results
Concentration (%)	0.0881
Span	0.764
Uniformity	0.238
Specific surface area (m^2kg^{-1})	40.43
D_{43} (μm)	139
D_{43} (μm)	150

*The particle size analysis data presented is one component of the data analysis content, derived from NO.1 of the CCD experimental results. Detailed data can be referenced in the RSM model construction section of the main text.

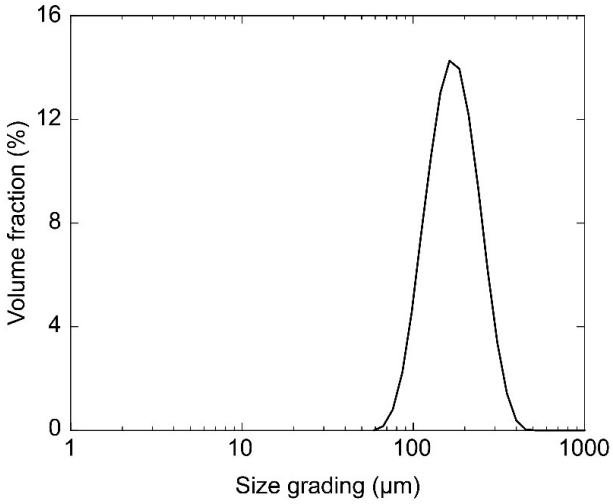


Figure S3. PVB product particle size distribution data.

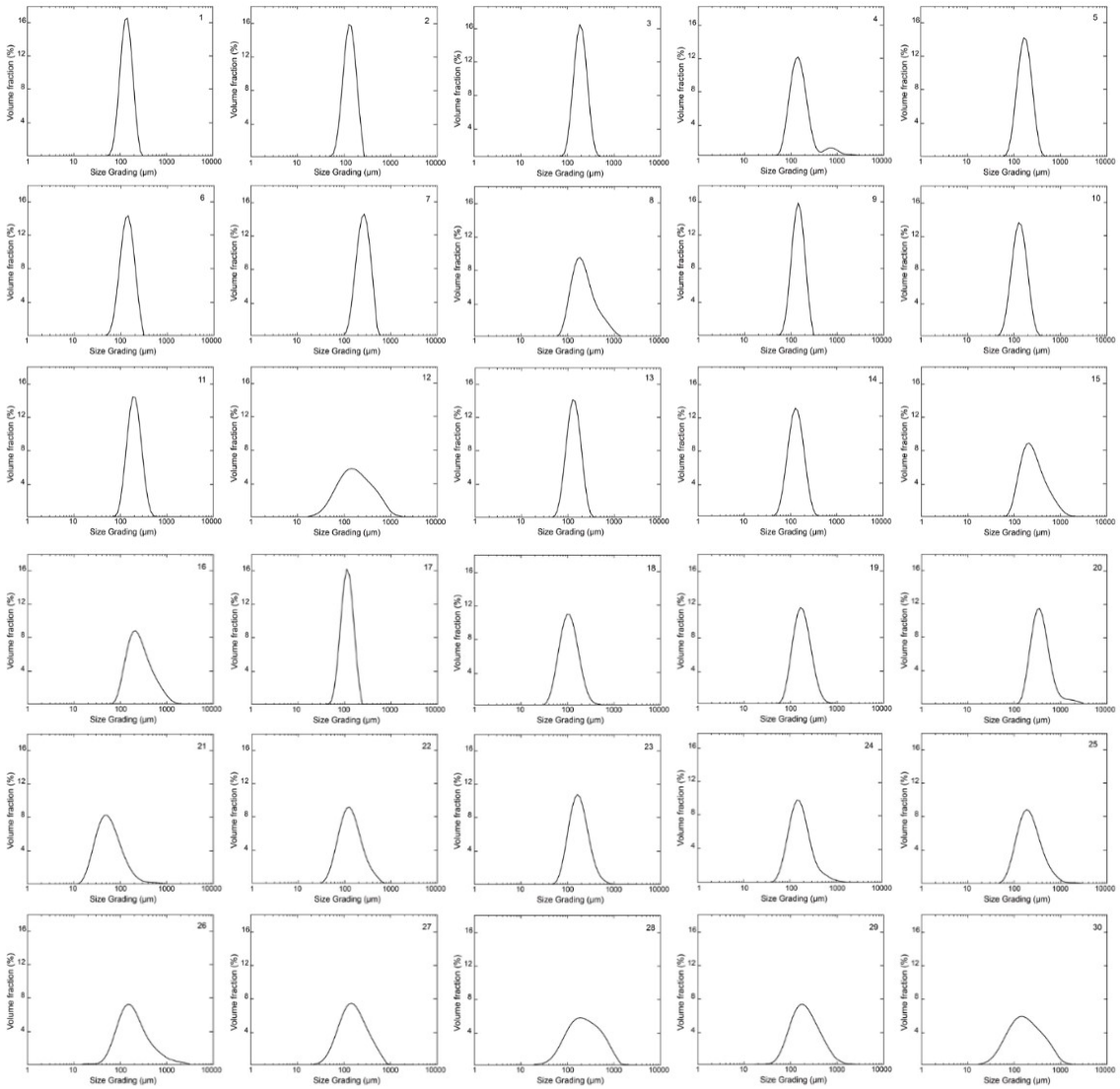


Figure S4. Particle size distribution data for 30 samples from CCD experimental design.

Supplementary Note III. The ANOVA results of RSM

During the construction and optimization of the RSM model, a second-order polynomial RSM model was built based on the CCD experimental data points, and variance analysis was conducted on the target features AD and D_{43} of PVB to further evaluate the performance of the RSM model. The results are summarized in **Tables S3-S5**. **Fig. S5** shows the 3D surface results representing the interaction among the experimental variables in the RSM model.

Table S3. Estimated Coefficients of the Second-Order Polynomial Models for AD and Particle size. **Table S3** reveals that the interaction term between n-butanal/PVA ratio (B) and HNO_3

concentration (C) exhibits the largest positive coefficient (18.67857) for AD, consistent with the proton-catalyzed nature of acetalization—higher HNO_3 concentrations (C) activate n-butanal aldehyde groups, enhancing reactivity with PVA hydroxyl groups. In contrast, the quadratic term for C (41.03333) indicates a non-monotonic effect: HNO_3 concentrations exceeding 0.075 mol/L accelerate PVB hydrolysis (as noted in Section 1), explaining the constraint of C to 0.05–0.10 mol/L in Section 2.3. These coefficients validate the RSM model's alignment with fundamental reaction chemistry, rather than purely mathematical fitting.

Term	**Coefficient for AD**	**Coefficient for D_{43}**
Intercept (β_0)	-1.90914	-375.97160
Linear Terms		
A (β_1)	0.24735	69.22048
B (β_2)	5.20599	-1286.06293
C (β_3)	-7.23881	8000.14286
D (β_4)	6.34901E-03	3.28790
Interaction Terms		
AB (β_{12})	-0.15750	16.07143
AC (β_{13})	-0.42000	-50.00000
AD (β_{14})	-6.15000E-04	0.07167
BC (β_{23})	18.67857	5071.42857
BD (β_{24})	-9.52381E-03	2.55952
CD (β_{34})	-0.02583	-2.66667
Quadratic Terms		
A^2 (β_{11})	-2.41667E-03	-5.59200

B^2 (β_{22})	-3.94026	889.03061
C^2 (β_{33})	41.03333	-57800.00000
D^2 (β_{44})	2.04398E-05	-0.01550

Note: A: PVA concentration; B: n-butanal feed ratio; C: HNO_3 concentration; D: Reaction time. All coefficients are based on coded factor levels.

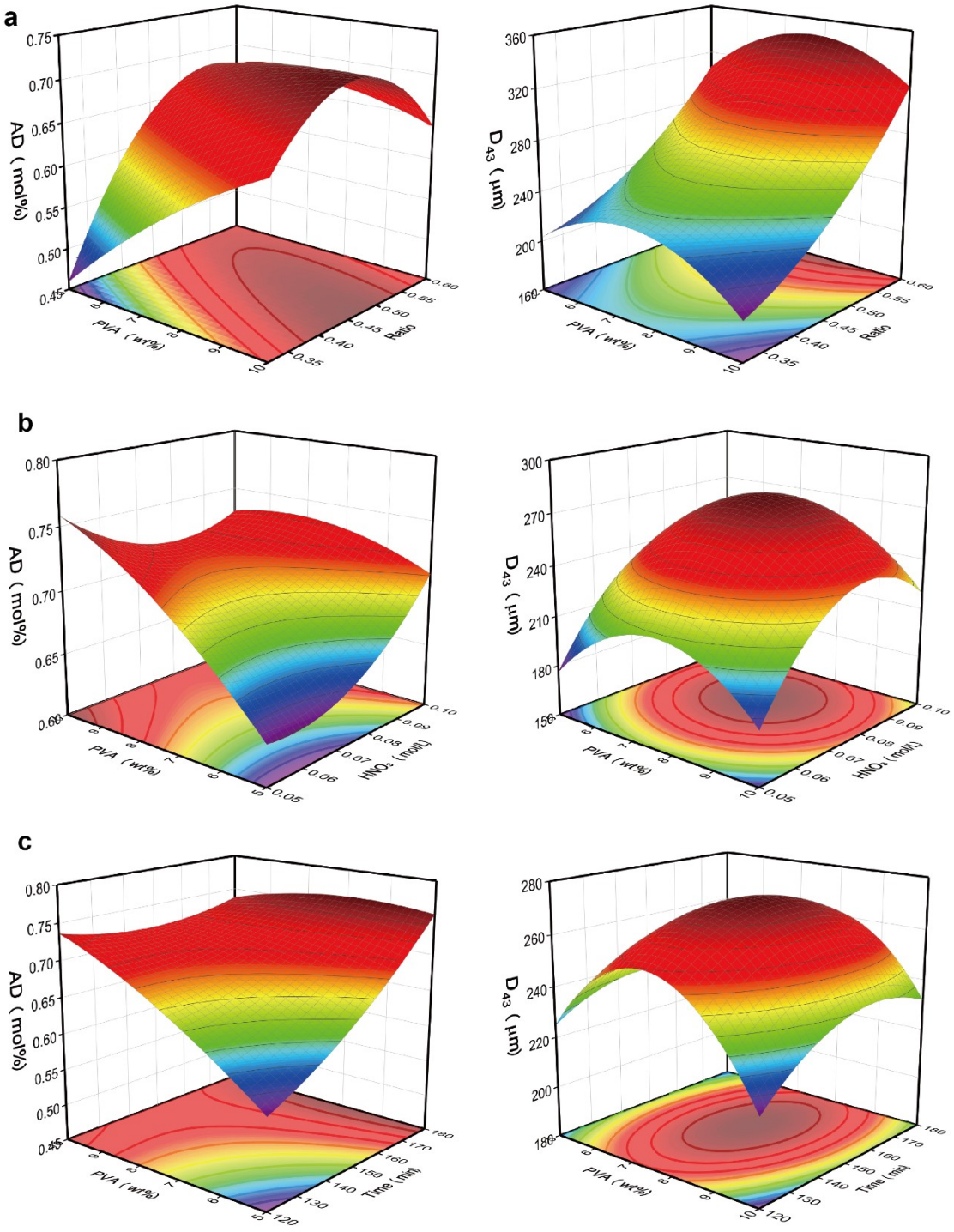
Table S4. ANOVA for the AD value.

Source	Sum of Squares	df	Mean Square	F-Value	p-value Prob > F	
Model	0.57	14	0.041	10.11	<0.0001	significant
A	0.033	1	0.033	8.27	0.0116	
B	0.065	1	0.065	16.17	0.0011	
C	$3.504E^{-003}$	1	$3.504E^{-003}$	0.87	0.3657	
D	0.052	1	0.052	12.89	0.0027	
AB	0.049	1	0.049	12.07	0.0034	
AC	0.011	1	0.011	2.74	0.1188	
AD	0.034	1	0.034	8.45	0.0108	
BC	0.068	1	0.068	16.98	0.0009	
BD	0.025	1	0.026	6.36	0.0235	
CD	$6.006E^{-003}$	1	$6.006E^{-003}$	1.49	0.2409	
A^2	$6.257E^{-003}$	1	$6.257E^{-003}$	1.55	0.2317	
B^2	0.16	1	0.16	10.62	<0.0001	
C^2	0.018	1	0.018	4.48	0.0514	
D^2	$9.282E^{-003}$	1	$9.282E^{-003}$	2.30	0.1498	
Residual	0.060	15	$4.027E^{-003}$			
Lack of Fit	0.054	10	$5.361E^{-003}$	3.94	0.0716	not significant
Pure Error	$6.799E^{-003}$	5	$1.360E^{-003}$			
Cor Total	0.53	29				

R^2 = 0.9042 Adj.R^2= 0.8147 Adeq. Precision=12.917 C.V.%=9.65

Table S5. ANOVA for the average size.

Source	Sum of Squares	df	Mean Square	F-Value	p-value Prob > F	
Model	$1.843E^{+005}$	14	13161.23	52.89	< 0.0001	significant
A	10.67	1	10.67	0.043	0.8388	
B	81666.67	1	81666.67	328.21	< 0.0001	
C	11828.16	1	11828.16	47.54	< 0.0001	
D	504.17	1	504.17	2.03	0.1751	
AB	506.25	1	506.25	2.03	0.1742	
AC	156.25	1	156.25	0.63	0.4405	
AD	462.25	1	462.25	1.86	0.1930	
BC	5041.00	1	5041.00	20.26	0.0004	
BD	1849.00	1	1849.00	7.43	0.0156	
CD	64.00	1	64.00	0.26	0.6194	
A^2	33504.07	1	33504.07	134.65	< 0.0001	
B^2	8328.15	1	8328.15	33.47	< 0.0001	
C^2	35794.71	1	35794.71	143.86	< 0.0001	
D^2	5337.67	1	5337.67	21.45	0.0003	
Residual	3732.35	15	248.82			
Lack of Fit	2300.35	10	230.04	0.80	0.6420	not significant
Pure Error	1432.00	5	286.40			
Cor Total	$1.880E^{+005}$	29				
	$R^2 = 0.9801$ Adj.$R^2 = 0.9616$ Adeq. Precision = 33.644 C.V.%=7.17					


a
AD (mol%)
0.75
0.70
0.65
0.60
0.55
0.50
0.45
PVA (wt%)
Ratio
0.60
0.55
0.50
0.45
0.40
0.35
D43 (μm)
360
320
280
240
200
160
b
AD (mol%)
0.80
0.75
0.70
0.65
0.60
PVA (wt%)
HNO3 (mol/L)
0.10
0.09
0.08
0.07
0.06
0.05
D43 (μm)
300
270
240
210
180
150
c
AD (mol%)
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
PVA (wt%)
Time (min)
190
180
170
160
150
140
130
120
D43 (μm)
280
260
240
220
200
180

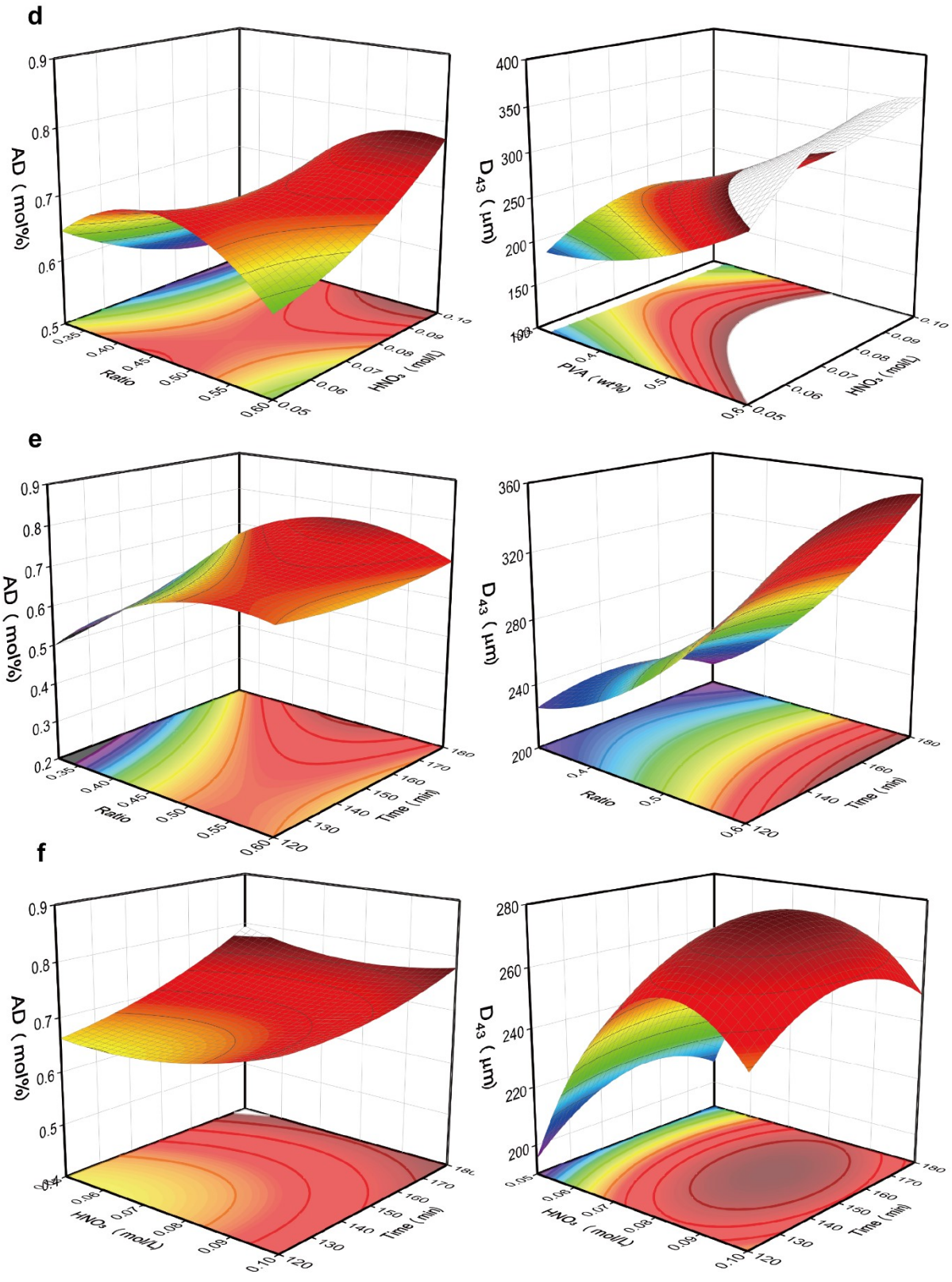


Figure S5. 3D surface plot of response surface methodology.

Supplementary Note IV. Model Exploration and Selection

To effectively construct the key predictive model of this study, the data characteristics and algorithm advantages were comprehensively considered. Three representative machine learning algorithms, namely artificial neural network (ANN), support vector regression (SVR), and random forest (RF), were selected for comparative research. **Fig. S6** presents schematic diagrams of three machine learning methods.

Support vector regression (SVR) is a supervised learning algorithm based on the support vector machine (SVM) framework, aiming to map nonlinear regression data to a high-dimensional space to make it linearly separable[1]**(Fig.S7)**. It fits data by constructing an "ε-insensitive tube" that allows most data points to fall within the margin while minimizing model complexity and prediction error. SVR utilizes kernel functions (Kernel Trick) and structural risk minimization (SRM) principles, offering stronger theoretical guarantees and generalization capabilities in modeling small sample, high-dimensional nonlinear data[2]. Furthermore, artificial neural networks (ANN) are better suited for large data volumes ($n>10^4$) with deep nonlinear relationships between features[3]; however, due to their complex hyperparameter systems (number of hidden layers, neurons, activation function selection, etc.), parameter tuning is challenging. Random forests (RF), which rely on decision tree splitting criteria, have advantages in feature importance ranking or when processing categorical variables[4], but they are more sensitive to outliers and require more features or data training for identification. Therefore, for a small sample, few feature PVB process optimization and model prediction-related research, SVR performs better.

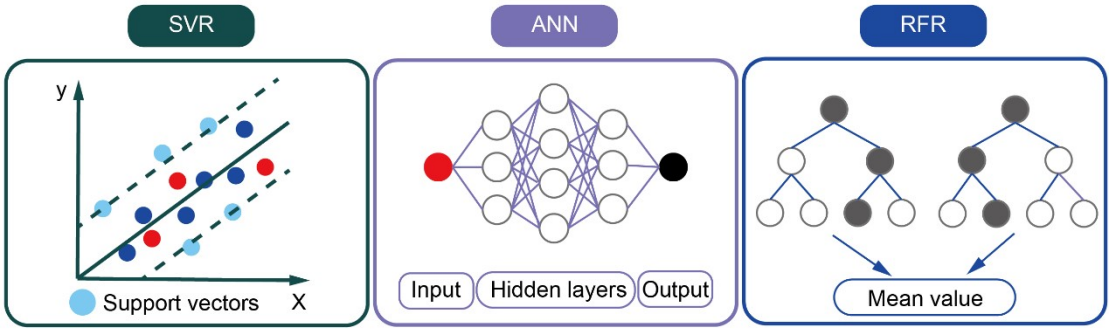


Figure S6. schematic diagrams of three machine learning methods.

*The figure shows the Support Vector Regression (SVR), Random Forest Regression (RFR), and Artificial Neural Network (ANN) methods. For SVR, the blue circles represent the training samples, and the light blue support vectors are used to generate the pipeline around the derived function (hyperplane). For SVR and ANN methods, the red circles indicate the test samples. For RFR, DT represents the decision tree, and the path from the root node to the leaf is highlighted with gray circles. Unlike the other methods, RFR is an ensemble method that relies on independently derived decision tree models. RFR and SVR represent the adaptive improvements of the original classification algorithms for predicting numerical values. ANN uses nonlinear activation functions to map the input values to the corresponding output through the calculation of neuron layers. More method details on RFR, SVR, and ANN are provided by Breiman[5], Drucker et al.[6], and Khamparia and Singh[7], respectively.

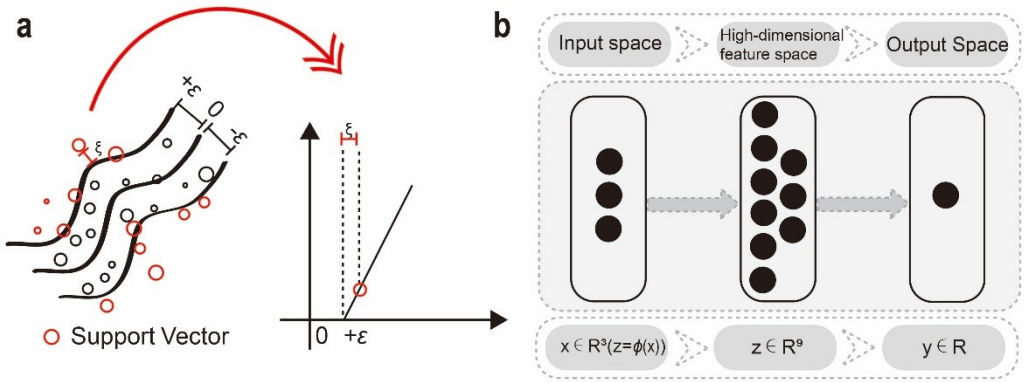


Figure S7. Schematic diagram of SVR solving nonlinear regression problems. (a) Support vector machine model showing insensitive parameter (ξ), support vectors (red circles), and soft margin loss defined by ξ[8]. The black area represents the margin with tolerance ε, and only samples outside the margin (support vectors) participate in model parameter optimization. (b) SVR maps input space to high-dimensional feature space through kernel functions and constructs linear regression models for sample data[2].

SVM has multiple types of kernel functions, including linear, nonlinear, polynomial, radial basis function (RBF), and sigmoid SVM. Different kernel function choices determine the effectiveness of handling classification and regression problems[9]. SVM performance largely depends on parameter selection[10], including penalty coefficient C, tolerance ε, kernel function type, and kernel function parameter σ. Optimization of different parameters can be achieved through grid search[10], random search[11], and Bayesian optimization. Support vector machine regression (SVR) specifically refers to SVM algorithms applied to regression

problems[8]. In this experimental investigation, SVR combined with RSM is used for data learning and model training. **Fig. S8** presents the general learning and training process for machine learning models, with emphasis on the parameter optimization scheme for SVR models.

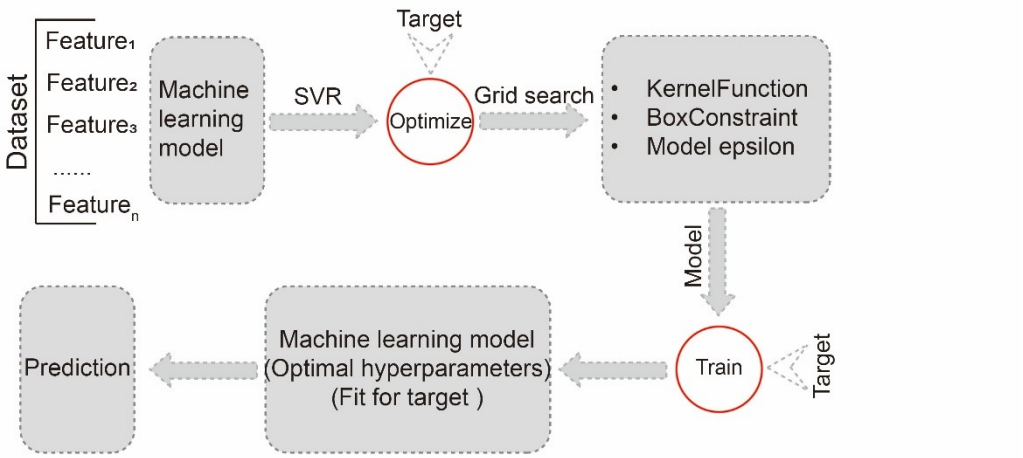


Figure S8. Machine learning model training and parameter optimization. Using SVR as an example, grid search is performed on SVR hyperparameters to obtain the optimal SVR hyperparameter model, and training set data is used for training. When the training set model effect reaches the set target, model training is completed, and this model can be used to predict the test set.

Supplementary Note V. Model performance and Validation dataset

To comprehensively evaluate the performance of the RSM-ML model, we also conducted five sets of completely independent experimental schemes **(Table S6)** for subsequent assessment and analysis verification.

Table S6. Independent validation set data table.

NO.	A: PVA (wt%)	B: Ratio	C: HNO_3 (mol/L)	D: Time (min)	Size (μm)	AD (mol%)
#1	2.5	0.32	0.025	90	24.1	0.529

#2	5	0.46	0.05	120	110	0.733
#3	5	0.6	0.075	120	166	0.702
#4	5	0.18	0.1	120	262	0.404
#5	10	0.74	0.1	150	742	0.654

References

1. Roy, A.; Chakraborty, S., Support vector machine in structural reliability analysis: A review. *Reliability Engineering & System Safety* **2023,** *233*, 109126.

2. Roy, A.; Manna, R.; Chakraborty, S., Support vector regression based metamodeling for structural reliability analysis. *Probabilistic Engineering Mechanics* **2019,** *55*, 78-89.

3. Yoo, S.-D.; Kim, J. Y.; Han, S.-K.; Lee, B.-H.; Choi, D. H.; Park, E.-S., Development of prediction model with machine learning in continuous twin screw granulation. *Journal of Pharmaceutical Investigation* **2023,** *53* (5), 707-722.

4. Kumar, K. S.; Razak, A.; Yadav, A.; Raghavendra Rao, P. S.; Majdi, H. S.; Khan, T. M. Y.; Almakayeel, N.; Singh, K., Experimental analysis of cycle tire pyrolysis oil doped with 1-decanol + TiO2 additives in compression ignition engine using RSM optimization and machine learning approach. *Case Studies in Thermal Engineering* **2024,** *61*, 104863.

5. Breiman, L., Random Forests. *Machine Learning* **2001,** *45* (1), 5-32.

6. Drucker, H.; Burges, C. J. C.; Kaufman, L.; Smola, A.; Vapnik, V., Support vector regression machines. In *Proceedings of the 10th International Conference on Neural Information Processing Systems*, MIT Press: Denver, Colorado, 1996; pp 155–161.

7. Khamparia, A.; Singh, K. M., A systematic review on deep learning architectures and applications. *Expert Systems* **2019,** *36* (3), e12400.

8. Smola, A. J.; Schölkopf, B., A tutorial on support vector regression. *Statistics and Computing* **2004,** *14* (3), 199-222.

9. Schölkopf, B.; Smola, A. J., *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press: 2001.

10. Hsu, C.-w.; Chang, C.-c.; Lin, C.-J., A Practical Guide to Support Vector Classification Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. **2003**.

11. Bergstra, J.; Bengio, Y., Random search for hyper-parameter optimization. **2012,** *13* (null

%J J. Mach. Learn. Res.), 281–305.