Supplementary Information (SI) for Environmental Science: Processes & Impacts.

Appendix:

1. Adsorption kinetics results

(1)The mathematical expression of the pseudo-first-order kinetic model is as follows:

$$\frac{dq_t}{dt} = k_1(q_e - q_t) \qquad (1)$$

The conditions for integration are t = 0 to t and q_t = 0 to q_t.

$$q_t = q_e(1 - e \qquad (2)$$

Taking the logarithm of both sides Ln simultaneously yields the following formula.

$$In(q_e - q_t) = lnq_e - K_1 t \qquad (3)$$

In the formula, K_1 (min^{-1}) is the rate constant of the pseudo-first-order kinetic model, and q_t ($mg \cdot g^{-1}$) and q_e ($mg \cdot g^{-1}$) are the adsorption amounts at time t and equilibrium, respectively.

(2) The mathematical expression of the differential form of the pseudo-second-order kinetic model is as follows:

$$\frac{dq_t}{dt} = k_1(q_e - q_t)^2 \qquad (4)$$

The conditions for integration are t = 0 to t and q_t = 0 to q_t.

$$q_t = q_2^2 k_2 t/(1 + q_2 k_2 t) \qquad (5)$$

Transforming it into a linear relationship yields the following formula:

$$\frac{t}{q_t} = \frac{1}{k_2 q_e^2} + \frac{t}{q_e} \qquad (6)$$

In the formula, K2 ($g \cdot mg^{-1} \cdot min^{-1}$) is the rate constant of the pseudo-second-order kinetic model, and qt ($mg \cdot g^{-1}$) and qe ($mg \cdot g^{-1}$) are the adsorption amounts at time t and equilibrium, respectively.

（2）(3) The mathematical derivation of the Elovich equation is as follows:

	$$\frac{dq_t}{dt} = \alpha exp(-\beta q_t)$$	（7）

Setting t and q_t to both 0, we get

	$$q_t = \frac{1}{\beta}\ln(t + t_0) - \frac{1}{\beta}\ln(t_0)$$	（8）

Let t0 = 1/αβ, and t >> t_0, then we get

	$$q_t = \frac{1}{\beta}\ln(\alpha\beta) - \frac{1}{\beta}\ln(t)$$	（9）

Let A = -1/β, B = 1/βLn(αβ), the above equation can be simplified to A

	$$q_t = Alnt + B$$	（10）

Parameters A and B are the fitting parameters of the Elovich model, qt is the adsorption amount ($mg \cdot g^{-1}$) at adsorption time t, and t is the reaction time (h).

(4) The intraparticle diffusion model describes the adsorption process by plotting the relationship between q_t and $t^{0.5}$. Its mathematical expression is:

	$$q_t = K_3\, t^{0.5} + C$$	（11）

In the formula, K3 ($g \cdot mg^{-1} \cdot min^{-1}$) is the diffusion rate constant within the particle, C ($mg \cdot g^{-1}$) is a constant related to the adsorption boundary thickness, qt is the adsorption ($mg \cdot g^{-1}$) at adsorption time t, and t is the reaction time (h).

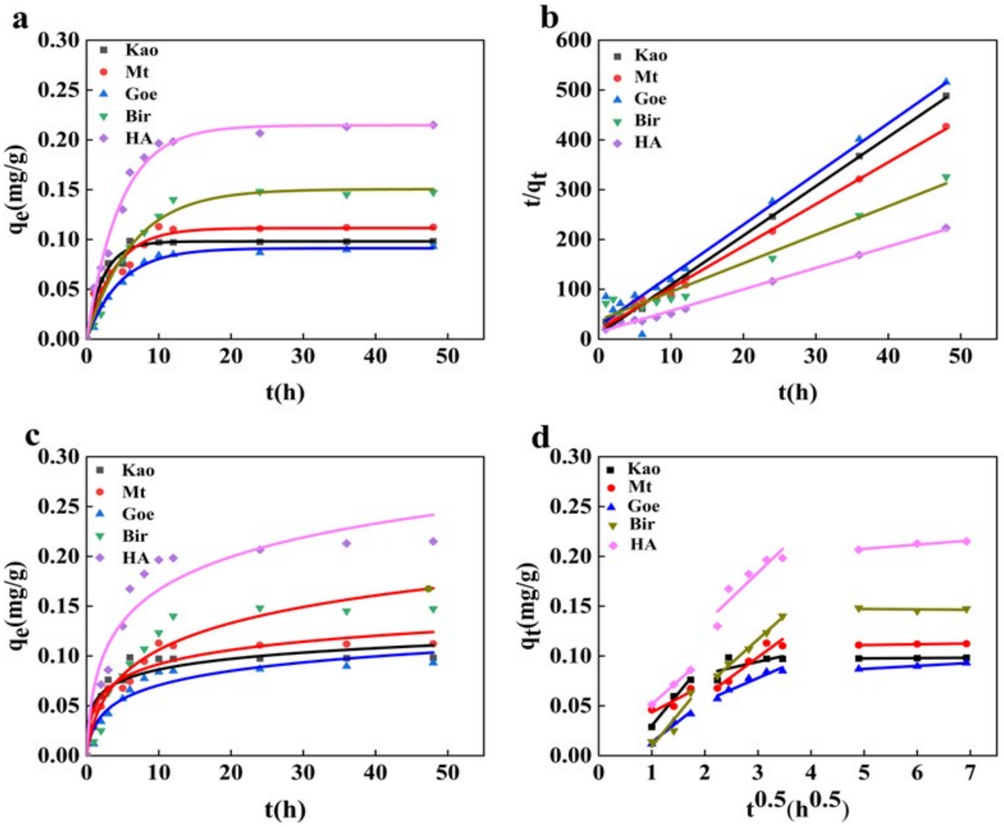


Figure S1 Fitting diagram of benzene adsorption kinetic model under different treatment conditions: pseudo-first-order kinetic model (a), pseudo-second-order kinetic model (b), Elovich model (c), intragranular diffusion model (d)

Table S1 Fitting parameters of benzene adsorption kinetic model under different treatment conditions : pseudo-first-order kinetic model, pseudo-second-order kinetic model, Elovich model

Adsorbate	Pseudo-first-order kinetic model			Pseudo-second-order kinetic model			Elovich model		
	Q_e ($mg \cdot g^{-1}$)	K_1 (min^{-1})	R^2	Q_e ($mg \cdot g^{-1}$)	K_2 ($g \cdot mg^{-1} \cdot min^{-1}$)	R^2	A	B	R^2
Kao	0.092	0.43	0.9879	0.093	0.016	0.9777	2.46	8.05	0.8756
Mt	0.11	0.26	0.9315	0.12	0.021	0.9953	5.83	12.32	0.9189
Goe	0.098	0.21	0.9520	0.10	0.017	0.9998	17.66	17.27	0.9173
Bir	0.16	0.15	0.9830	0.17	0.040	0.9996	19.86	4.05	0.9148
HA	0.21	0.21	0.9872	0.23	0.063	0.9952	6.99	7.52	0.9146

Table S2 Fitting parameters of the intragranular diffusion model of benzene under different treatment conditions

Adsorbate	K_{P1}	C_1	R_1^2	K_{P2}	C_2	R_2^2	K_{P3}	C_3	R_3^2
Kao	0.065	-0.035	0.9908	0.012	0.058	0.9824	0.0003	0.096	0.9987
Mt	0.039	-0.019	0.9368	0.028	0.015	0.9332	0.0007	0.11	0.9677
Goe	0.042	-0.029	0.9606	0.023	0.0086	0.9291	0.0030	0.072	0.9903
Bir	0.065	-0.056	0.9598	0.048	0.026	0.9967	0.0004	0.15	0.9379
HA	0.048	0.0040	0.9599	0.031	0.031	0.9371	0.0002	0.19	0.9423

Adsorption kinetics results showed that the adsorption of benzene by different minerals and humic acid proceeded rapidly within 0–4 h, gradually slowed down from 4–12 h, and reached equilibrium at approximately 12 h. The equilibrium adsorption capacity followed the order: humic acid (0.22 $mg \cdot g^{-1}$) > manganese hydrate (0.15 $mg \cdot g^{-1}$) > montmorillonite (0.11 $mg \cdot g^{-1}$) > goethite (0.098 $mg \cdot g^{-1}$) > kaolinite (0.093 $mg \cdot g^{-1}$), indicating that humic acid had the strongest adsorption capacity. Kinetic model fitting showed that, except for kaolinite, the pseudo-second-order model best fit the adsorption systems ($R^2 \approx 0.995$–0.999), and the fitted Qe was consistent with the measured value, indicating that the adsorption process was mainly chemisorption.

Humic acid had the highest rate constant K_2 (0.063 g·mg^{-1} min^{-1}), indicating that it had a stronger affinity for benzene. Kaolinite better fit the pseudo-first-order model, indicating that its adsorption was mainly a diffusion process. The intraparticle diffusion model (R^2>0.93) shows that the adsorption process can be divided into three stages: rapid surface adsorption, pore/interlayer diffusion, and equilibrium. The fitted line does not pass through the origin, indicating that the adsorption process is jointly controlled by surface adsorption and intraparticle diffusion. By fitting the kinetic curves, we found that the adsorption of benzene by minerals and humic acid reached equilibrium after 12 h, which can provide a time reference for subsequent adsorption experiments.

2. Relevant Formulas and Schematic Diagrams of Machine Learning Models

When configuring model parameters, the random forest (RF) model was set with n _estimators (number of trees) = 300, max _depth (maximum depth of a tree) = 10, min _samples _split (minimum samples required for node splitting) = 2, and random _state (random seed) = 42. The Artificial Neural Network (ANN) model was set with hidden_size (dimension of the hidden layer) = 14, ReLU (rectified linear unit) activation, learning_rate = 0.01, epochs (number of full training iterations) = 500, and random_state (random seed) = 42. The Support Vector Machine (SVM) model was set with max_iter (maximum training iterations) = 10000, C (regularization parameter) = 300, kernel = poly (polynomial kernel), epsilon (tolerance interval) = 0.1, and random_state (random seed) = 15000.

The RF model calculates feature importance using the Gini Impurity Decrease method, which evaluates how much each feature reduces Gini impurity when splitting decision tree nodes. The formula is as follows:

	$I(t) = 1 - \sum_{K=1}^{K} p_{kt}^2$	（12）
	$\Delta I(j,t) = I(t) - \left(\frac{N_{left}}{N_t} I(t_{left}) + \frac{N_{right}}{N_t} I(t_{right})\right)$	（13）

	$$Importance_T(j) = \sum_{t \in T} \Delta I(j,t) \cdot \mathrm{II}\,(j,t)$$	(14)
	$$Importance(j) = \frac{1}{N_{trees}} \sum_{T=1}^{N_{trees}} Importance_T(j)$$	(15)
	$$Normalized\ Importance(j) = \frac{Importance(j)}{\sum_{i=1}^{d} Importance(j)}$$	(16)

In these formulas, K represents the total number of categories; p_{kt} denotes the proportion of samples of the k-th category in node t; $I(t)$ signifies the Gini impurity before splitting the node; t_{left} and t_{right} denote the left and right child nodes after splitting; N_t, N_{left}, and N_{right} stand for the number of samples in node t, the left child node, and the right child node; N_{trees} represents the number of trees in the forest, and T denotes a single decision tree; $II(\cdot)$ is an indicator function (1 if feature j is split at node t, otherwise 0); and *Normalized Importance* (j) denotes the normalized feature importance.

The ANN model calculates feature importance using the Garson algorithm (Fernandes et al., 2020), which evaluates the relative contribution of each input variable based on the connection weights between layers. The specific calculation formulas are as follows:

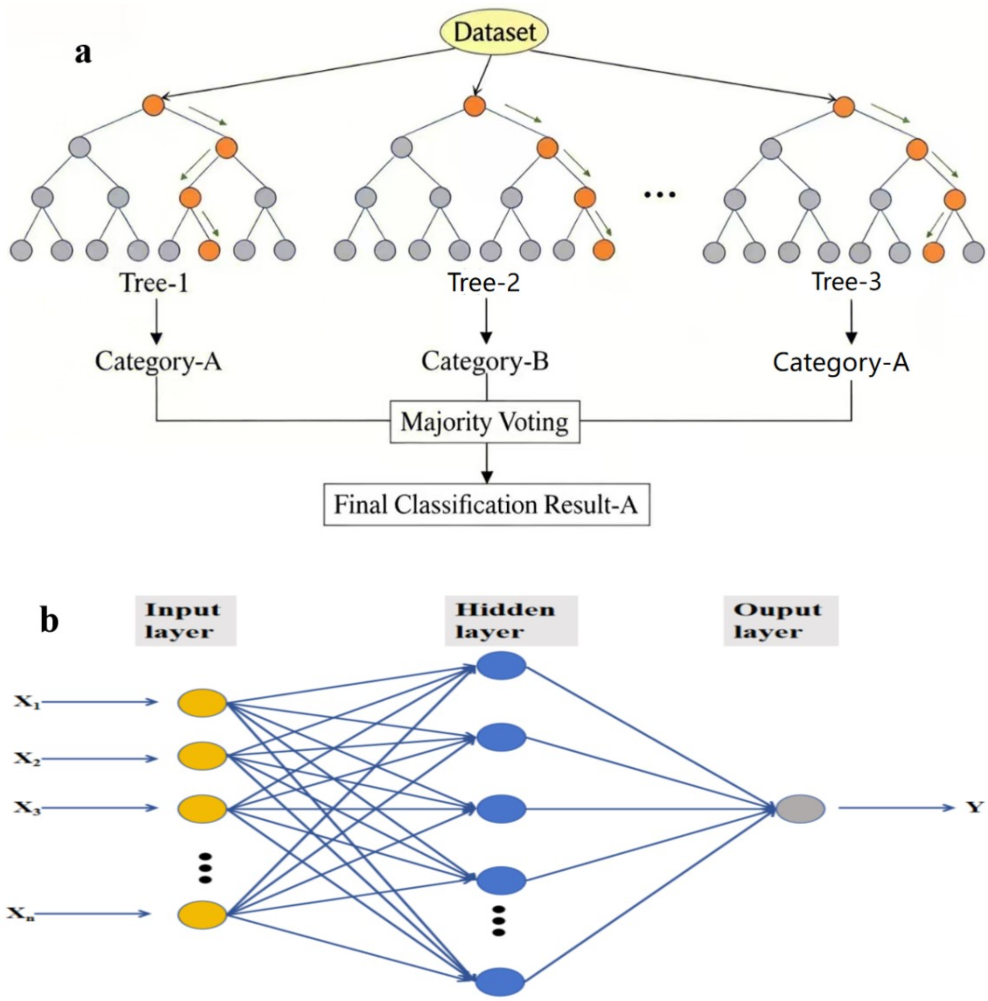
	$$I_j = \frac{\sum_{m=1}^{m=Nh}\left[\left(\frac{\lvert W_{jm}^{ih}\rvert}{\sum_{k=1}^{N_i}\lvert W_{km}^{ih}\rvert}\right)\times\lvert W_{mn}^{ho}\rvert\right]}{\sum_{k=1}^{k=Ni}\left[\sum_{m=1}^{m=Nh}\left[\left(\frac{\lvert W_{jm}^{ih}\rvert}{\sum_{k=1}^{N_i}\lvert W_{km}^{ih}\rvert}\right)\times\lvert W_{mn}^{ho}\rvert\right]\right]}$$	(17)

In the formula, I_j represents the relative importance of an input variable in the output variable, and N_i and N_h denote the number of neurons in the input and hidden layers; W signifies the synaptic weights connecting the layers of the neural network, and I, h, and o stand for the input, hidden, and output layers, respectively; K, m, and n refer to the number of input, hidden, and output neurons, respectively.

The SVM model evaluates the importance of features through the absolute value of the weight coefficient. For linear kernel SVM, the calculation formulas are as follows:

	$f(x) = W^T x + b$	(18)
	$Importance_i = \frac{\lvert w_i \rvert}{\lVert W \rVert_1}$	(19)

In the formulas, W represents the weight vector; T denotes the transpose operation (which converts the weight vector from a column vector to a row vector); x stands for the input feature vector; b is the bias term; |w_i| signifies the absolute value of the i-th weight coefficient, and |(|W|)|_1 indicates the normalized weight vector.



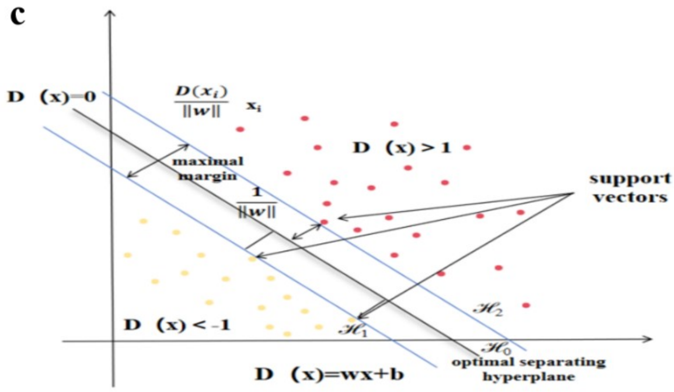


Figure S2 Schematic diagram of the Random Forest (a), Artificial Neural Network (b), and Support Vector Machine (c) models

3. Relevant Charts and Graphs in the Materials and Methods Section

3.1 Characterization of Experimental Materials

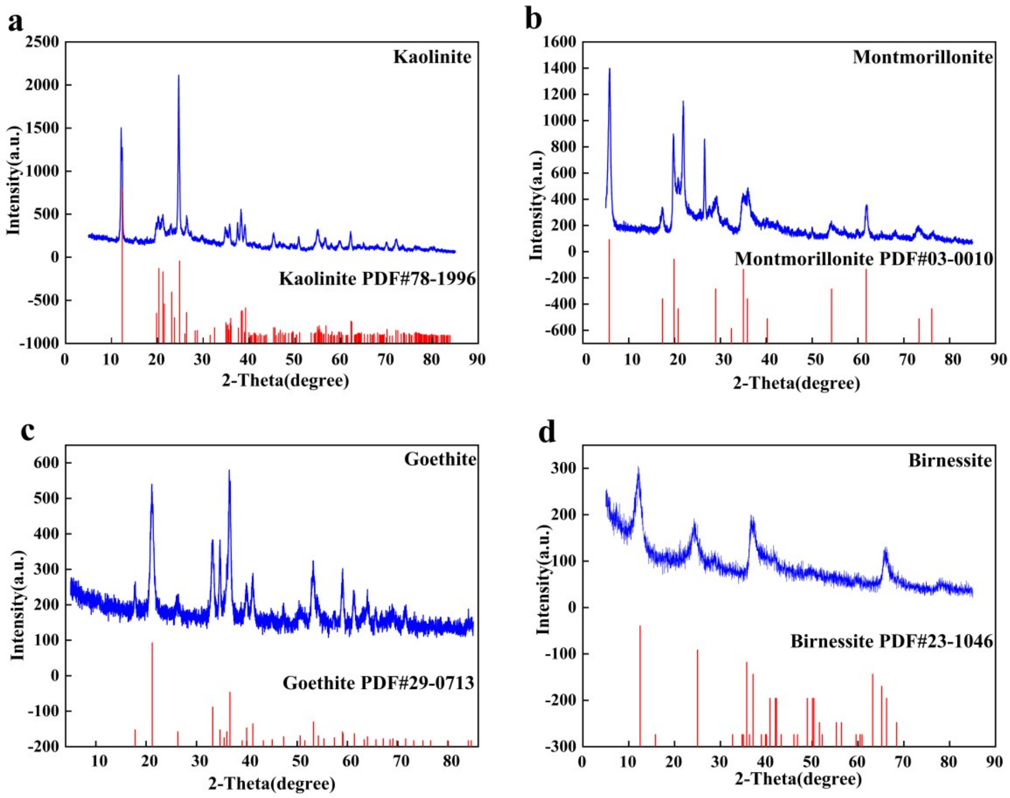


Figure S3 Comparison spectra of Kao (a), Mt (b), Goe (c), Bir (d) and corresponding standard card XRD

3.2 Machine Learning Modeling

The formulas for calculating relevant indicators are as follows:

	$R^2 = \frac{\left[\sum_{i=1}^{n}(Y_{e,i}-\bar{Y}_{e,i})(Y_{p,i}-\bar{Y}_{p,i})\right]^2}{\sum_{i=1}^{n}(Y_{e,i}-\bar{Y}_{e,i})^2(Y_{p,i}-\bar{Y}_{p,i})^2}$	（20）
	$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_{p,i}-Y_{e,i})^2$	（21）
	$RMSE = \left[\frac{1}{n}\sum_{i=1}^{n}(Y_{p,i}-Y_{e,i})^2\right]^{\frac{1}{2}}$	（22）
	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert Y_{e,i}-Y_{p,i}\rvert$	（23）
	$MAPE = \frac{100\%}{n}\left(\sum_{i=1}^{n}\frac{\lvert Y_{p,i}-Y_{e,i}\rvert}{\lvert Y_{e,i}\rvert}\right)$	（24）

In these formulas, $Y_{e,i}$ represents the experimental value, $\bar{Y}_{e,i}$ denotes the average experimental value, $Y_{p,i}$ signifies the predicted value, $\bar{Y}_{p,i}$ indicates the average predicted value, and n stands for the sample size.

4. Relevant Charts and Graphs in the Results and Analysis section

4.1 The Interactive Effects of Environmental Factors on Benzene Adsorption

4.1.1 Response Surface Evaluation

Table S3 Box-Behnken test design and measurement results（mg g^{-1}）

Groups	T(℃)	pH	Pb^{2+}(mol L^{-1}）	Kao	Mt	Goe	Bir	HA
1	15	7.0	0	10.2	12.9	11.0	18.0	46.0
2	15	9.0	0.005	9.6	11.3	8.7	13.2	30.6
3	15	7.0	0.01	6.8	9.7	8.2	13.2	31.0
4	15	5.0	0.005	7.1	9.1	10.0	15.1	47.5
5	15	9.0	0.01	6.8	7.9	5.1	8.6	31.0
6	25	7.0	0.005	5.9	8.1	6.4	9.6	31.7
7	25	5.0	0	7.1	9.6	10.1	14.4	45.2
8	25	9.0	0	9.7	11.8	7.8	12.4	33.0
9	25	7.0	0.005	5.9	7.4	6.2	8.9	30.5
10	25	7.0	0.005	6.8	7.9	7.1	9.7	30.6
11	25	7.0	0.005	6.3	7.9	6.9	10.8	27.2
12	25	5.0	0.01	4.7	5.9	7.3	11.3	19.3
13	25	7.0	0.005	6.7	7.4	6.7	10.0	27.4
14	35	9.0	0.005	4.6	5.9	3.3	2.8	8.0
15	35	5.0	0.005	3.3	3.6	4.7	7.7	20.3
16	35	7.0	0.01	3.1	3.5	2.8	4.7	9.8
17	35	7.0	0	5.6	7.3	5.3	8.9	24.6

Notes: 1. Kao = Kaolinite; Mt = montmorillonite; Goe = Goethite; Bir = Birnessite; HA = humic acid. 2. T and Pb^{2+}represents temperature (°C) and coexisting Pb^{2+} concentration (mol/L)

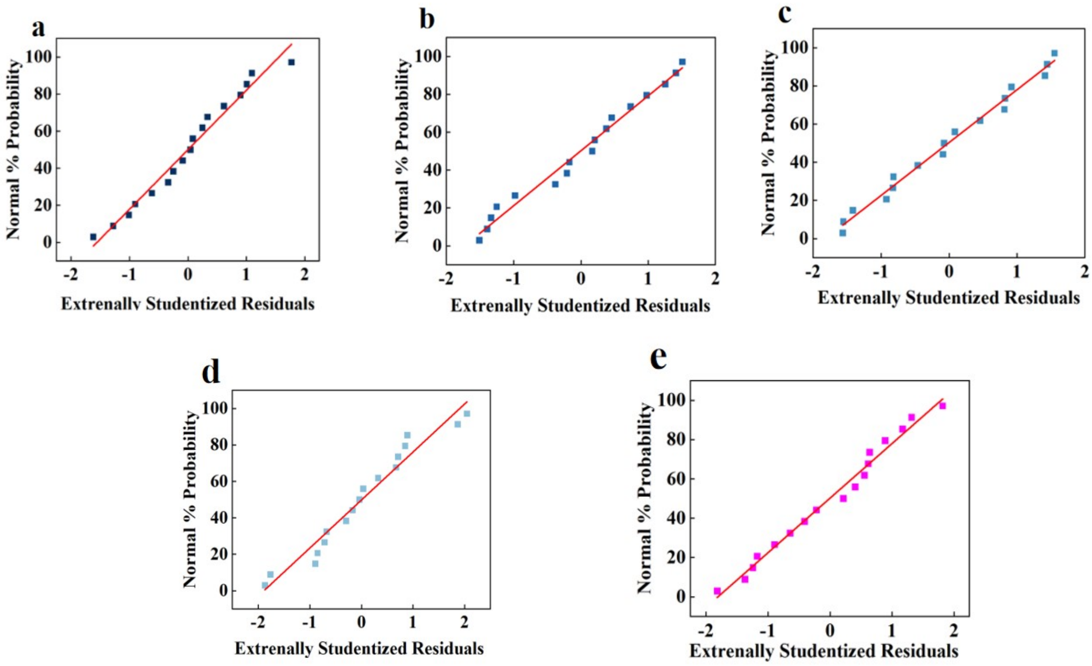


Figure S4 Normal probability distribution plots of 3D response surface model fittings under different treatments: Kao (a), Mt (b), Goe (c), Bir (d), HA (e)

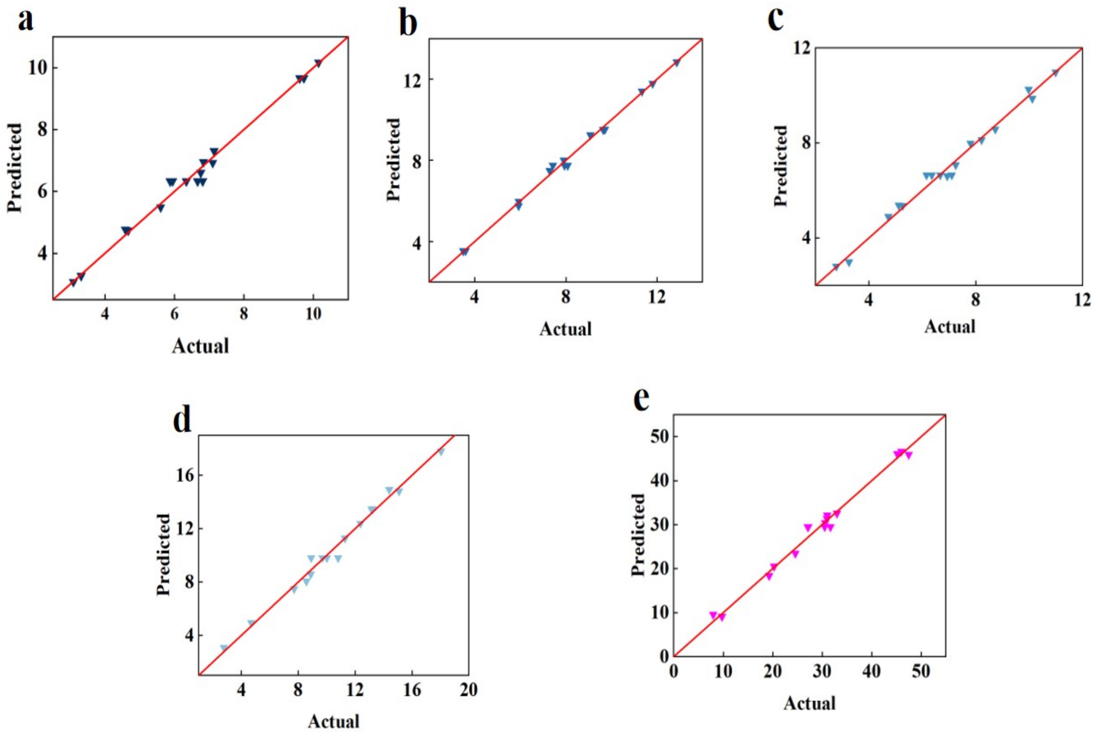


Figure S5 Scatter distribution plots of predicted Vs. actual values fitted to the 3D esponse surface model under different treatments: Kao (a), Mt (b), Goe (c), Bir (d), HA (e)

Table S4 Response surface fitting quadratic model ANOVA table for benzene adsorption by Kao, Mt, Goe, Bir and HA

Factor item		Model	X_1	X_2	X_3	X_1X_2	X_1X_3	X_2X_3	X_1^2	X_2^2	X_3^2	Residual	Deviance	Pure error	Sum	R^2	R^2adj	CV(%)	Precision
Degree of freedom		9	1	1	1	1	1	1	1	1	1	7	3	4	16				
Sum of squares	Kao	63.92	36.56	8.92	15.86	0.38	0.1	0.06	0.72	0.24	1.08	0.85	0.15	0.70	64.77	0.9869	0.9700	5.38	26.52
	Mt	104.96	64.36	9.55	26.48	0.0008	0.10	0.0095	0.58	0.051	3.93	0.55	0.18	0.37	105.51	0.9948	0.9781	3.47	43.25
	Goe	83.36	59.53	6.42	14.59	0.013	0.021	0.0083	0.56	0.66	1.20	1.02	0.41	0.61	84.38	0.9879	0.9723	5.53	27.91
	Bir	220.56	157.30	16.57	31.94	2.36	0.13	0.12	0.34	0.13	11.70	3.01	1.11	1.90	3.01	0.9865	0.9692	6.22	29.31
	HA	1911.2	1069.26	351.94	416.59	5.27	0.0044	0.067	53.66	2.03	15.82	27.18	10.30	16.88	1938.45	0.9860	0.9679	6.79	24.85
Mean square	Kao	7.10	36.56	8.92	15.86	0.38	0.16	0.06	0.72	0.24	1.08	0.12	0.05	0.18					
	Mt	11.66	64.36	9.55	26.48	0.01	0.10	0.01	0.59	0.050	3.93	0.08	0.06	0.09					
	Goe	9.26	59.53	6.42	14.59	0.01	0.02	0.01	0.56	0.66	1.20	0.15	0.14	0.15					
	Bir	24.51	157.30	16.57	31.94	2.36	0.13	0.12	0.34	0.13	11.70	0.43	0.37	0.47					
	HA	212.36	1069.26	351.94	416.59	5.27	0.01	0.07	53.66	2.03	15.82	3.88	3.43						
F	Kao	58.40	300.66	73.39	130.44	3.12	1.32	0.50	5.96	1.98	8.88		0.29						
	Mt	148.00	816.81	121.16	336.09	0.01	1.29	0.12	7.39	0.64	49.90		0.65						
	Goe	63.50	410.86	44.04	100.06	0.09	0.10	0.06	3.83	4.50	8.22		0.88						
	Bir	54.69	275.35	90.63	107.28	1.36	0.01	0.02	13.82	0.52	4.07		0.81						
	HA	54.69	275.35	90.63	107.28	1.36	0.01	0.02	13.82	0.52	4.07		0.81						
P	Kao	0.0001**	0.0001**	0.0001**	0.0001**	0.1209	0.2878	0.5009	0.0447*	0.2021*	0.0205*		0.8339						
	Mt	0.0001**	0.0001**	0.0001**	0.0001**	0.9213	0.2933	0.7392	0.0298*	0.4500*	0.0002**		0.6229						
	Goe	0.0001**	0.0001**	0.0003**	0.0001**	0.7754	0.7151	0.8188	0.0910	0.0715	0.0241*		0.5220						
	Bir	0.0001**	0.0001**	0.0004**	0.0001**	0.0516	0.6057	0.6137	0.4016	0.5940	0.0012*		0.5622						
	HA	0.0001**	0.0001**	0.0001**	0.0001**	0.2821	0.9741	0.8996	0.0075	0.4936	0.0834		0.5492						

Notes: X1, X2, and X3 represent temperature (℃), pH, and coexisting Pb^{2+} concentration (mol/L), respectively; * indicates significant difference ($p < 0.05$), and ** indicates extremely significant difference ($p < 0.01$)

4.1.2 Interactive Analysis of 3D Response Surfaces

1. The second-order polynomial regression equation for the interaction of environmental factors on kaolinite's adsorption of benzene:

$$Y_1 = 7.60191+0.081251X_1+0.137502X_2-497.90948X_3-0.015388X_1X_2+4.01167X_1X_3-12.37408X_2X_3-0.0004148X_1^2+0.059798X_2^2+20260.35211X_3^2 \quad (25)$$

2. The second-order polynomial regression equation for the interaction between montmorillonite and environmental factors on benzene adsorption is:

$$Y_2 = 12.37422+0.086756X_1+0.1695282X_2-636.65897X_3-0.000719X_1X_2-3.18884X_1X_3-4.86240X_2X_3-0.003719X_1^2+0.027360X_2^2+38653.31830X_3^2 \quad (26)$$

3. The second-order polynomial regression equation for the interaction of environmental factors on benzene adsorption by goethite:

$$Y_3 = 20.90894-0.078891X_1-1.78229X_2-551.73454X_3-0.002832X_1X_2+1.45188X_1X_3+4.54165X_2X_3-0.003645X_1^2+0.098737X_2^2+21350.85161X_3^2 \quad (27)$$

4. The second-order polynomial regression equation for the interaction of environmental factors on benzene adsorption by birnessite:

$$Y_4 = 23.11461-0.049645X1-0.297188X_2-1033.78770X_3-0.038408X_1X_2+3.54373X_1X_3-17.31797X_2X_3-0.002853X_1^2+0.044602X_2^2+66676.57467X_3^2 \quad (28)$$

5. The second-order polynomial regression equation for the interaction of environmental factors on benzene adsorption by humic acid:

$$Y_5 = 87.49769+0.223697X_1-7.24328X_2-2325.32259X_3+0.057399X_1X_2+0.662989X_1X_3+12.89474X_2X_3-0.035698X_1^2+0.173393X_2^2+77529.10295X_3^2 \quad (29)$$

In the above equations, X1, X2, and X3 represent temperature, pH, and coexisting Pb^{2+} concentration, respectively, while Y denotes the adsorption amount of benzene

4.2 Construction and Validation of Predictive Model

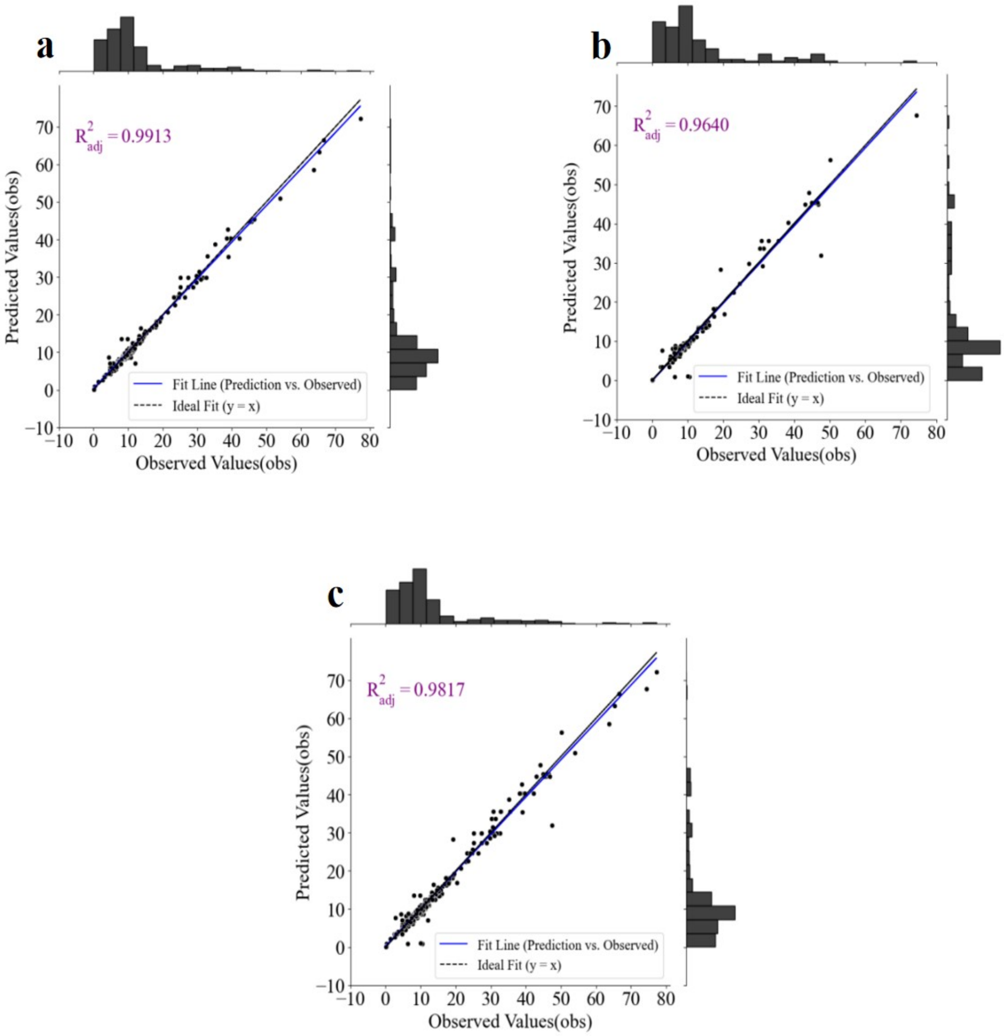


Figure S6 Marginal histogram of the RF model dataset: training set (a), test set (b), overall dataset (c)

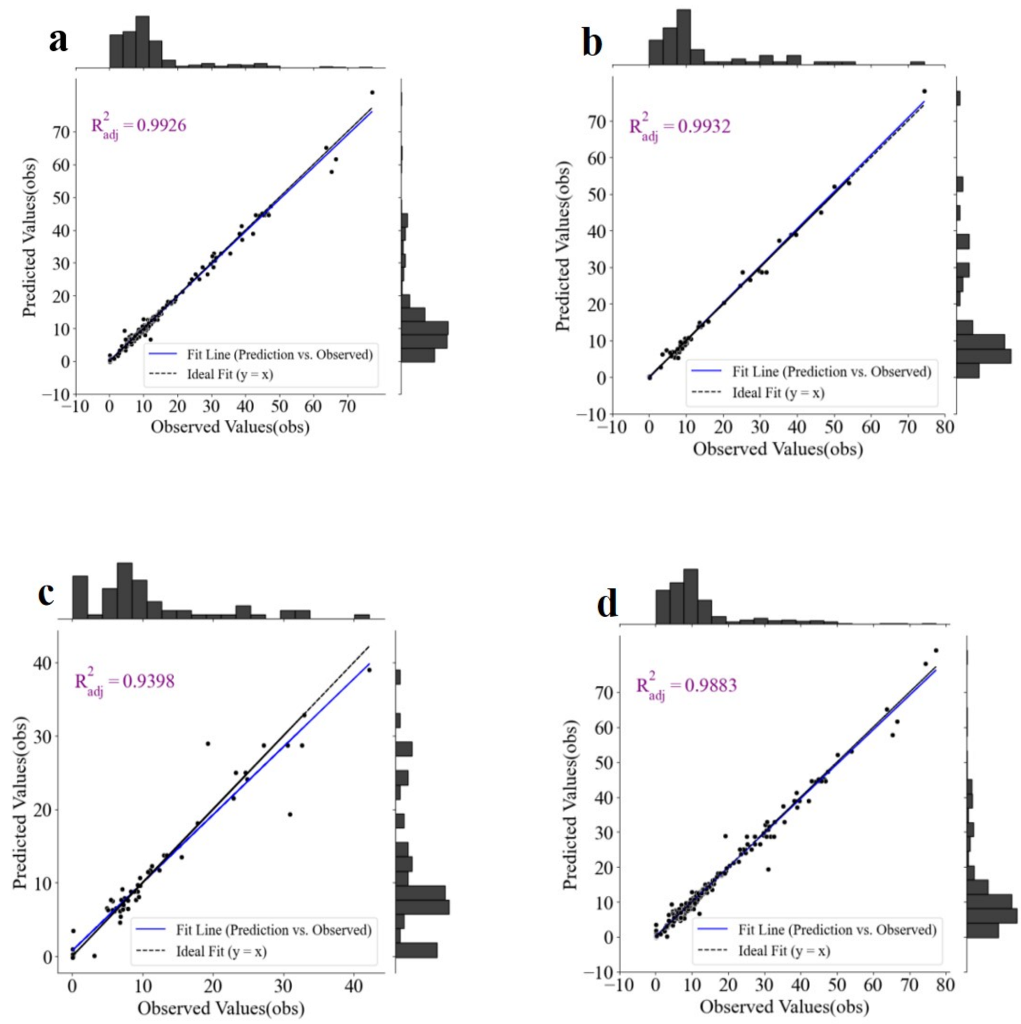


Figure S7 Marginal histograms of different datasets of the ANN models: training set (a), validation set (b), test set (c), overall dataset (d)

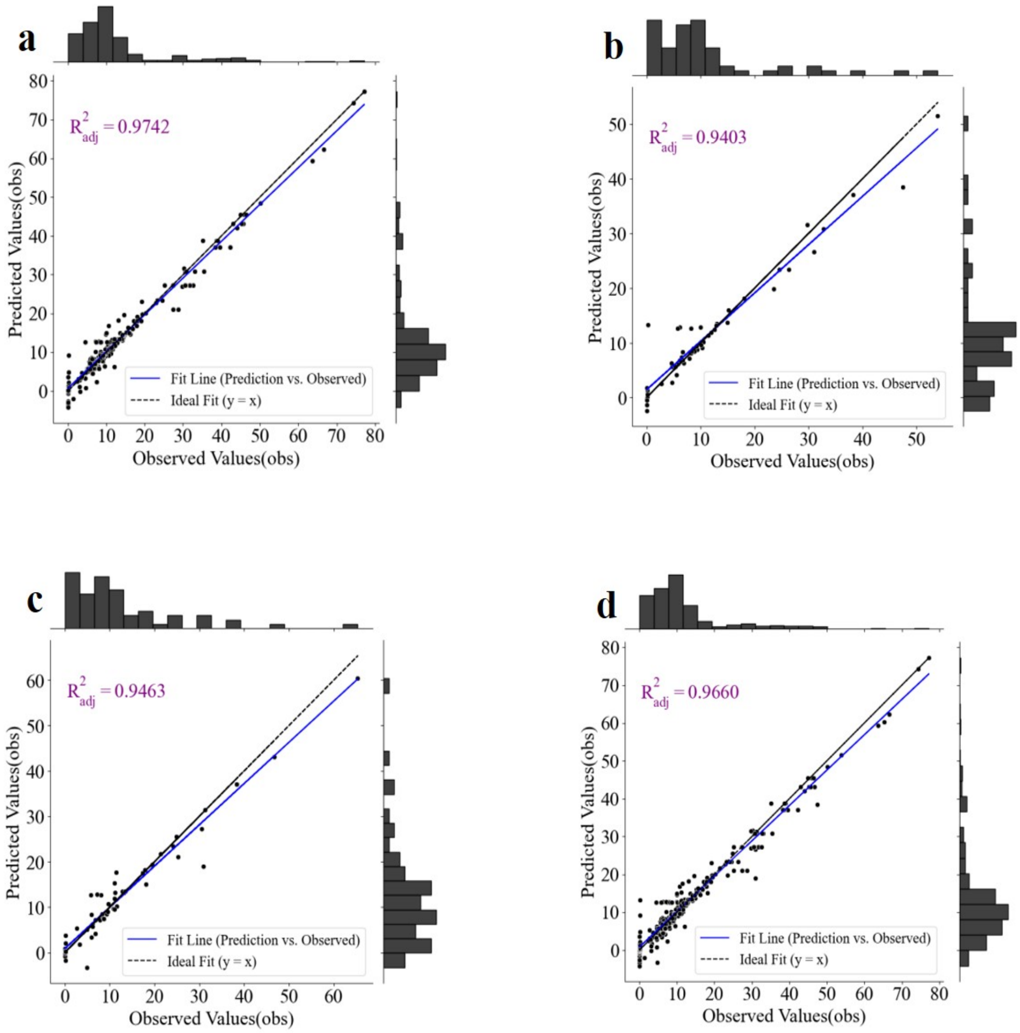


Figure S8 Marginal histograms of different datasets of the SVM models: training set (a), validation set (b), test set (c), overall dataset (d)