Supplementary Materials

AI-Driven Prediction of Uranium Adsorption on Acid-Modified Biochar: Integrating Large Language Models with Interpretable Machine Learning

Jingyang Sun [1], Sut Ian Chan [2], Shuting Zhuang [3]*

1 School of Ecology & Environment, Renmin University of China, Beijing 100872, P. R. China

2 School of Information, Renmin University of China, Beijing 100872, P. R. China

3 School of Chemistry & Life Resources, Renmin University of China, Beijing 100872, P. R. China

This file includes:

Texts S1-S2: page 2

Tables S1-S5: page 3-6

Figures S1-S4: page 7-10

∗ Corresponding author: Email: zst@ruc.edu.cn;

Text S1 | Predefined criteria for literature screening

In the first stage of automatic screening, DeepSeek-R1 filtered out irrelevant publications based on four predefined, data-specific criteria to ensure the initial dataset focused exclusively on high-relevance studies. These criteria were explicitly designed to target studies with direct implications for biochar's uranium adsorption performance, and included: (1) primary focus on biomass-derived pyrolysis biochar (excluding biochar composite materials); (2) application of acid modification processes (e.g., HCl, HNO_3, or other acid treatments, with clear description of modification steps); (3) clear reporting of uranium adsorption capacity (quantitative values under specified conditions, e.g., equilibrium adsorption capacity, maximum uptake, or kinetic parameters); (4) availability of key experimental parameters (including biochar dosage, initial uranium concentration, pH, and temperature).

Text S2 | System Prompt

"You are a specialized assistant for extracting scientific data from research papers. Please extract the following information from the provided text section. For numerical values, include both the value and unit. If information is not present, return "Not reported". Extraction Targets: 1. Pyrolysis temperature (°C); 2. Hold duration (h); 3. Heating rate (°C/min); 4. Acid type; 5. Acid concentration (mol/L); 6. Acid treatment time (h); 7. Acid treatment temperature (°C); 8. Sequence of pyrolysis and modification; 9. Specific surface area (m^2/g); 10. Average pore size (nm); 11. Total pore volume (cm^3/g); 12. C (%); 13. O (%); 14. N (%); 15. H (%); 16. C/N; 17. O/C; 18. H/C; 19. (O+N)/C; 20. Point of zero charge (pH_{pzc})."

Table S1 | Quantitative scoring rule for data extraction using LLMs.

Paradigm	Consistency level	Definition (for each feature)	Score (0~100)
Single-LLM	–	A single LLM generated results consistent with those verified manually.	100
	–	A single LLM generated results that are inconsistent with those verified manually.	0
Multi-LLM	Full consensus	All three LLMs output the same results, which were consistent with those verified manually.	100
	Majority consensus	Two out of three LLMs output the same results, which were consistent with those verified manually.	80
	Partial consensus	Only one of the three LLMs output the same results as those verified manually.	40
	No consensus	All three LLMs output distinct results, and none of them were consistent with those verified manually.	0

Table S2 | Four tree-based ensemble ML model prediction results under different division ratios.

Division ratio	Model	Training R^2	Training RMSE	Test R^2	Test RMSE
70:30	RF	0.991754	14.920804	0.955069	42.409025
	XGBoost	0.999936	1.315045	0.965583	37.117187
	CatBoost	0.998345	6.683698	0.971989	33.485218
	LightGBM	0.968726	29.056867	0.939691	49.133607
80:20	RF	0.993175	13.977987	0.975148	31.475375
	XGBoost	0.999901	1.680214	0.977844	29.718954
	CatBoost	0.998590	6.354147	0.981672	27.029917
	LightGBM	0.971777	28.423664	0.961339	39.257546
85:15	RF	0.993716	13.639480	0.967174	35.432152
	XGBoost	0.999905	1.680211	0.982154	26.124649
	CatBoost	0.998517	6.626181	0.980428	27.358985
	LightGBM	0.973759	27.871694	0.963690	37.264718

Table S3 | The hyperparameters and their ranges of the model used in this study.

Model	Hyperparameter*	Value Range
Random Forest	n_estimators	100-500 (step = 50)
	max_depth	[None, 10, 20, 30]
	min_samples_split	2–10 (integer)
	min_samples_leaf	1–5 (integer)
XGBoost	max_depth	3–10 (integer)
	learning_rate	0.01–0.3 (log scale)
	subsample	0.6–1.0 (step = 0.1)
	reg_lambda	0–5 (step = 0.5)
CatBoost	max_depth	3–10 (integer)
	learning_rate	0.01–0.3 (log scale)
	l2_leaf_reg	1–10 (integer)
LightGBM	n_estimators	100–200 (step = 50)
	max_depth	3–10 (integer)
	num_leaves	30–100 (integer)
	learning_rate	0.01–0.3 (log scale)

* For the XGBoost and CatBoost models, the default value of the hyperparameter "n_estimators" was used during the training process.

Table S4 | Sample counts for the two preparation sequences used to define Seq_P_M in the compiled dataset.

Seq_P_M	Seq_P_M = 0	Seq_P_M = 1
Records	319	270

* Seq_P_M is a binary indicator describing the relative order of modification and pyrolysis: Seq_P_M = 0 denotes modification before pyrolysis (pre-modification), and Seq_P_M = 1 denotes pyrolysis before modification (post-modification).
* Records refer to individual adsorption experiments extracted from the literature (i.e., one record corresponds to one set of reported experimental conditions and adsorption outcome). If multiple experiments were reported within the same article, each experiment was counted as one record.

Table S5 | Seed sensitivity analysis of the CatBoost model performance.

Metric	Mean ± SD	Min–Max
Test R^2	0.9827 ± 0.0007	0.9809–0.9838
Test RMSE (mg/g)	25.69 ± 0.53	24.86–27.00

* The train/test split was fixed (test size = 0.15; random_state = 42), and preprocessing, feature set, and hyperparameters were kept unchanged across runs. Only the CatBoost random seed was varied (seeds = 0–19). Metrics were computed on the same held-out test set.

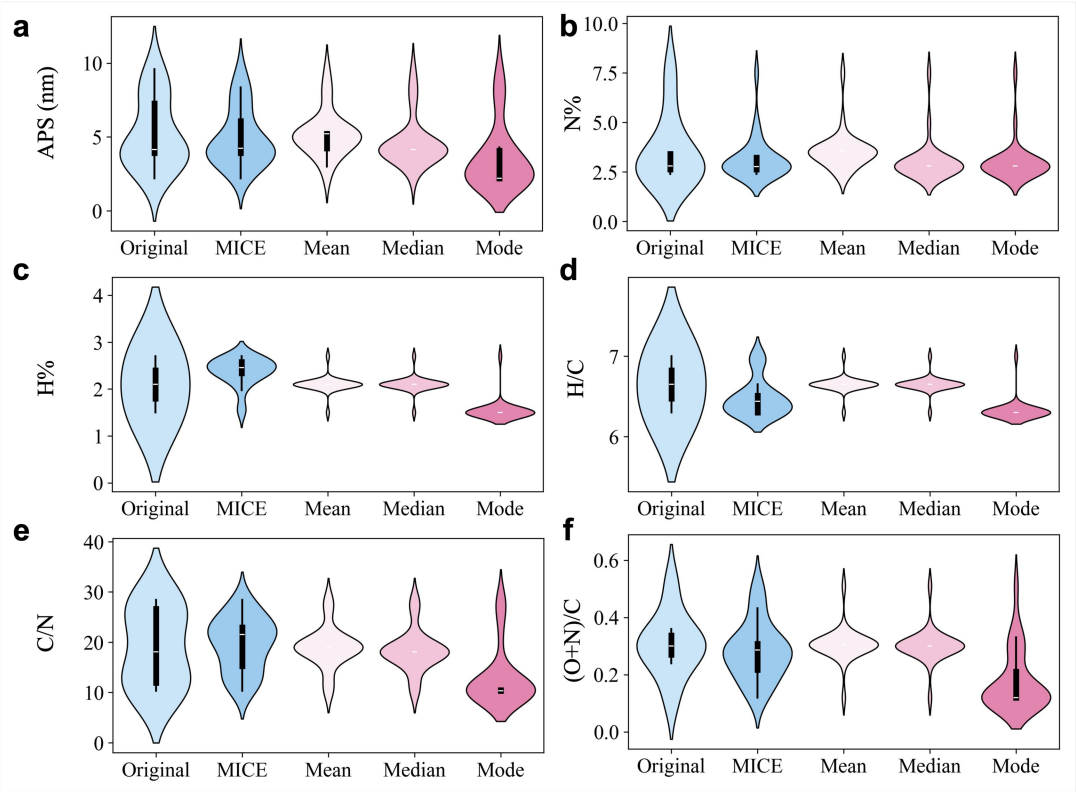

Figure S1 | Performance comparison of MICE, Mean, Median and Mode in feature missing value imputation with missing ratios exceeding 50%.

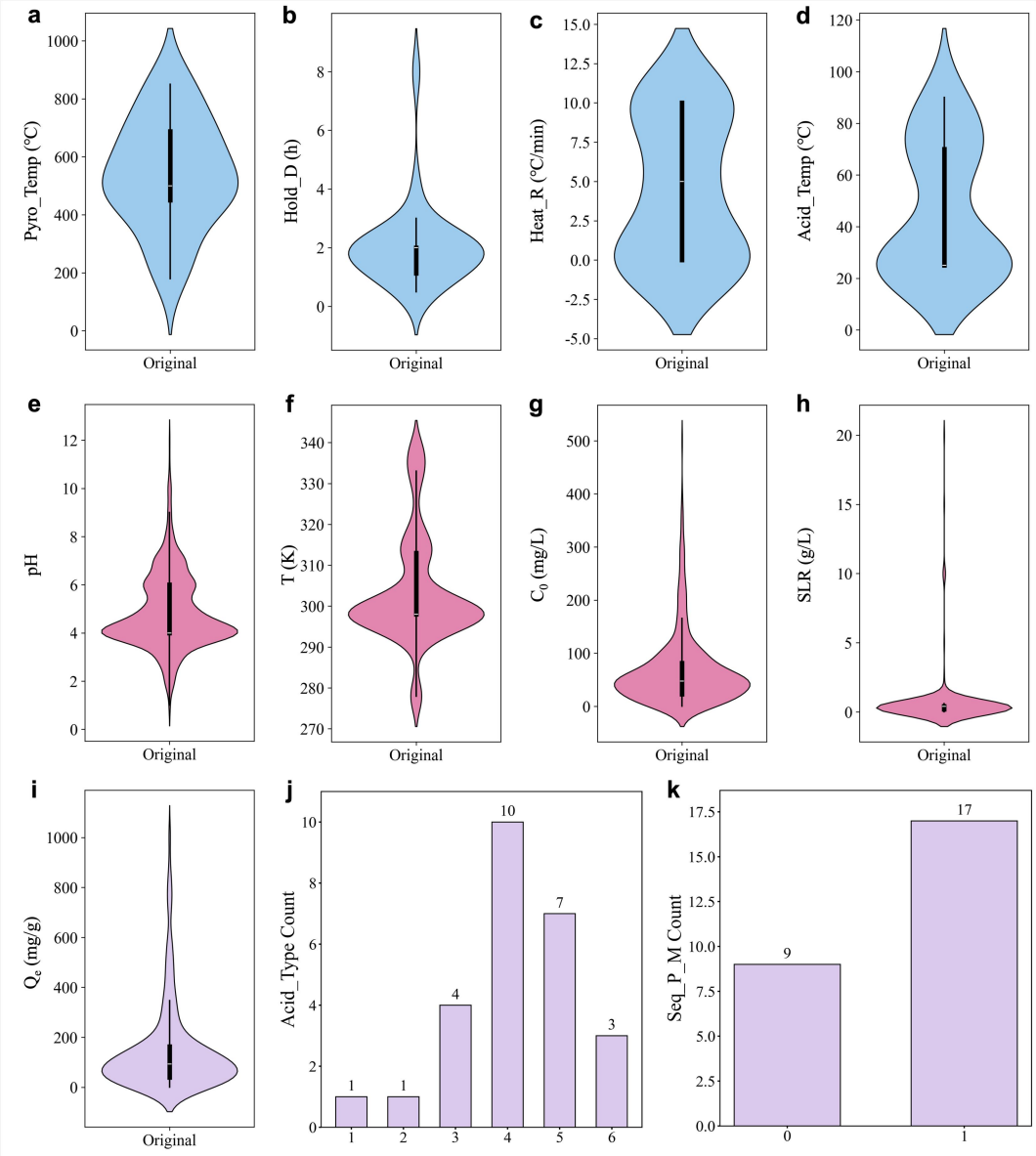

Figure S2 | The distribution of the remaining features. The histogram represents the distribution of category features.

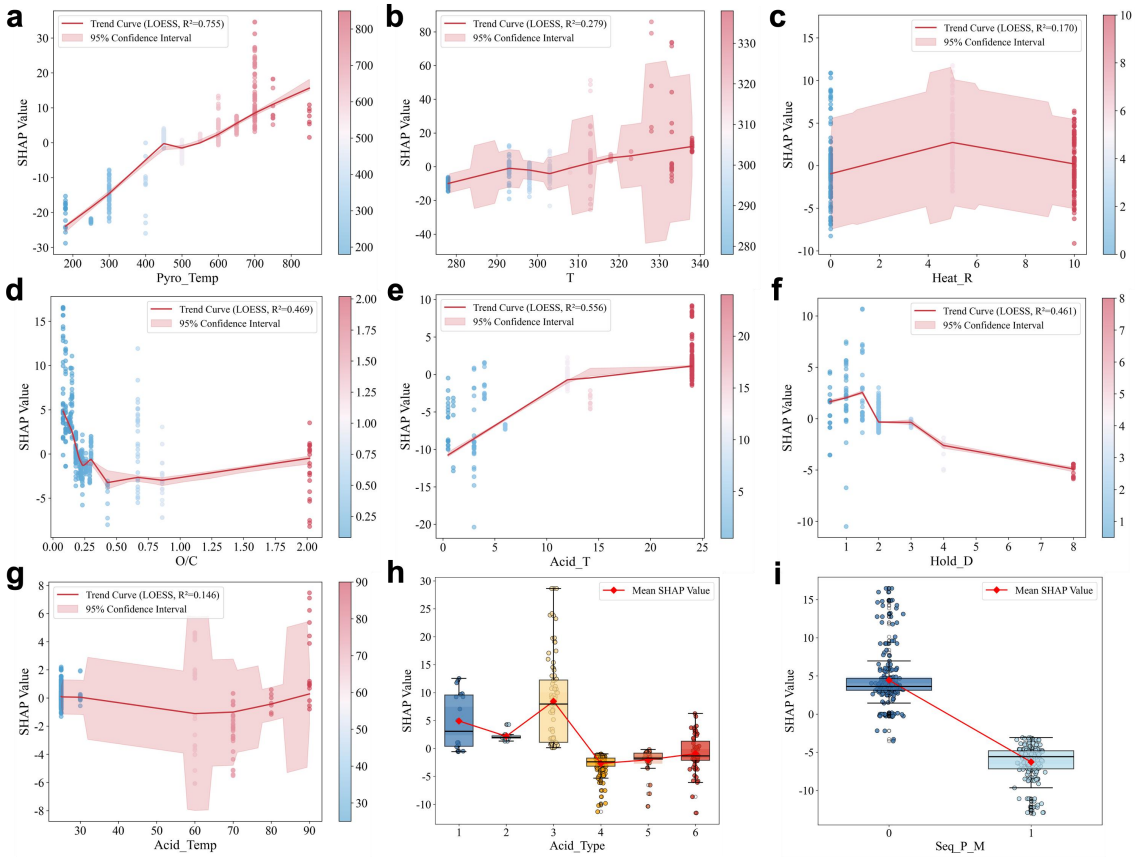

Figure S3 | The SHAP dependence plots of the remaining features.

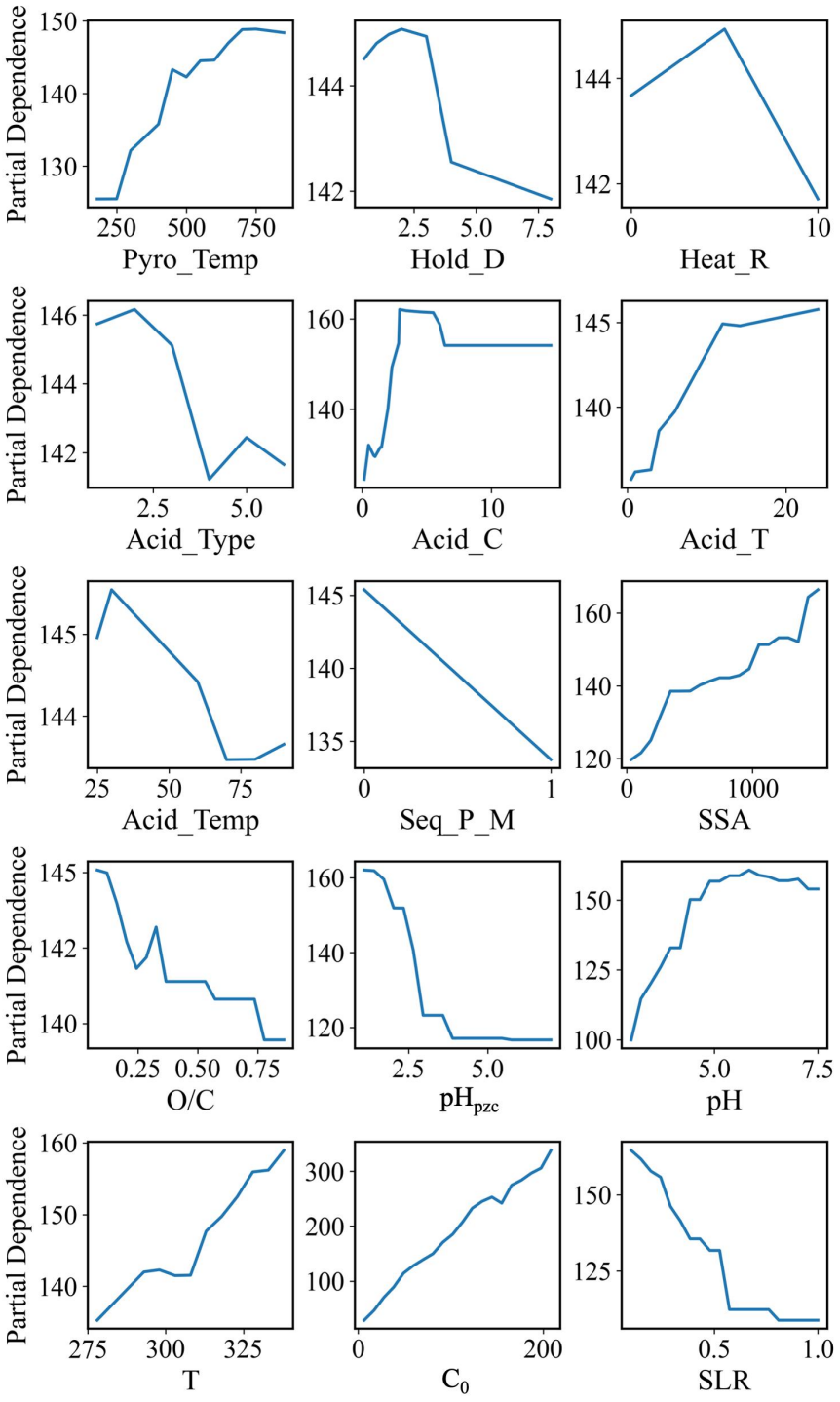

Figure S4 | Model interpretation based on partial dependence plots of uranium adsorption capacity on input features.