

***Chlorella vulgaris* and *Arthrospira platensis* show distinct removal dynamics for multiple heavy metals in real water matrices**

Lamprini Malletzidou, Eleni Kyrtzopoulou, Evangelos Nerantzis and Nikolaos A. Kazakis*

Laboratory of Archaeometry and Physicochemical Measurements, Athena- Research and Innovation Center in Information, Communication and Knowledge Technologies, Kimmeria University Campus, PO Box 159, GR-67100 Xanthi, Greece

*Corresponding author

S1. Supplementary Methods

S1.1. Machine learning (ML) dataset and feature configurations

The complete experimental dataset consisted of 48 observations obtained from two microalgal species (*Chlorella vulgaris*, *C. vulgaris* and *Arthrospira platensis*, *A. platensis*), three water matrices (fish processing facility wastewater, FPFW, river water, RW, and lake water, LW) and eight sampling days (0-7). Each observation represents a specific combination of them, i.e. the microalgal species, the water matrix, and the cultivation day, resulting in 24 observations per microalgal species.

The target variable was the removal efficiency (R, %) of each metal. The complete input feature set consisted of 21 input variables, including:

1. Water matrix (FPFW, RW, LW)
2. Cultivation day (0-7)
3. pH
4. Electrical conductivity (EC)
5. Biomass concentration
6. Biomass proxy, microalga-specific: normalized cell count measurements for *C. vulgaris*, and diffuse reflectance NIR signal at 1062.6 nm for *A. platensis*
7. Initial concentration (C_0) of each of the five metals
8. Measured concentration (C_t) at time t (days) of each of the five metals
9. Daily uptake of the five metals, defined as $(C_{t-1} - C_t)/\text{biomass concentration}$

Four input feature configurations were evaluated independently for each target metal and microalgal species, as a structured sensitivity analysis to determine how model performance changes when direct target-metal related information was retained, excluded, or temporally restricted.

Configuration 1. Full feature configuration. This configuration included all 21 input features, including contemporaneous residual metal concentrations (C_t) and daily uptakes. Because the contemporaneous concentration of the target metal is directly involved in calculating removal efficiency, this configuration was treated only as a full-information diagnostic reference and not as evidence of independent predictive performance.

Configuration 2. Reduced baseline configuration. All contemporaneous metal-concentration and daily concentration-change variables were excluded. The model therefore contained 11 variables describing cultivation time, water matrix, physicochemical conditions, biomass and initial metal concentrations.

Configuration 3. Cross-metal configuration. The target metal concentration and uptake variables were excluded. The C_t and uptake variables of the four coexisting metals were kept, resulting in 19 features that were used as input variables.

Configuration 4. Previous time point configuration. All uptake variables were excluded, and the daily metal concentration variables C_t were replaced by the

corresponding concentrations measured at the preceding time point C_{t-1} , resulting in 16 input features.

The complete feature inclusion for the four configurations is presented in Table S1.

Table S1. The features used as input variables for each one of the four ML configurations. ✓: included, X: excluded.

Feature or Setting	Configuration 1	Configuration 2	Configuration 3	Configuration 4
Water matrix	✓	✓	✓	✓
Cultivation day	✓	✓	✓	✓
pH	✓	✓	✓	✓
EC	✓	✓	✓	✓
Biomass concentration	✓	✓	✓	✓
Biomass proxy	✓	✓	✓	✓
C_0 of all metals	✓	✓	✓	✓
C_t	✓	X	Non-target metals	Replaced by C_{t-1}
Metal uptake	✓	X	Non-target metals	X
Total number of features	21	11	19	16
Sampling days used	0-7	1-7	1-7	1-7
Observations per species-metal model	24	21	21	21
LODO folds	8	7	7	7
Training observations per LODO fold	21	18	18	18
Held-out observations per LODO fold	3	3	3	3

S1.2. Feature preprocessing and model implementation

Two regression algorithms were evaluated, the Gaussian process regressor (GPR), and extreme gradient boosting (XGBoost) analyses. For GPR, all continuous input variables were standardized to zero mean and unit variance using the *StandardScaler* implementation of scikit-learn. The scaler was fitted exclusively to the training partition of each cross-validation fold and subsequently applied to the corresponding held-out partition. Consequently, data from the held-out sampling day did not contribute to the estimation of the scaling parameters. Features were not standardized for XGBoost because tree-based models do not require feature scaling.

GPR was implemented using the scikit-learn *GaussianProcessRegressor*. The kernel consisted of a constant kernel multiplied by a Matérn kernel and combined additively with a white-noise kernel. The Matérn kernel was initialized with a length scale of 1.0 and $\nu=1.5$, while the white-noise component was initialized with a noise level of 0.1. Kernel hyperparameters were optimized separately within each training fold by maximizing the log-marginal likelihood, using five optimizer restarts. Target normalization was applied internally using *normalize_y = True*, and the random seed was set to 42.

XGBoost was implemented using *XGBRegressor* with 200 estimators, a maximum tree depth of 3, a learning rate of 0.05, subsampling of 0.8, column subsampling of 0.8, L_1 regularization of 0.5 and L_2 regularization of 1.0. Features were used without standardization, and the random seed was set to 42. The same parameter settings were used for all target metals, microalgal species and input configurations. Because of the limited dataset size, no nested hyperparameter search was performed for XGBoost; instead, the parameters were

fixed across all analyses to avoid selecting model settings on the basis of held-out test performance.

S1.3. Cross-validation and performance evaluation

Model performance was assessed using leave-one-day-out cross-validation (LODO-CV). For this reason, the *LeaveOneGroupOut* implementation was used in scikit-learn, with the cultivation day being applied as the grouping variable. During each iteration, all observations from one sampling day across the three water matrices were used as the held-out test set, while observations from the remaining sampling days were used for model development. All preprocessing steps, including GPR feature standardization and kernel optimization, were performed independently within the training partition of each fold. Model performance was evaluated using the coefficient of determination (R^2), root-mean-square error (RMSE) and mean absolute error (MAE). Because removal efficiency was represented on a 0–100% scale, RMSE and MAE are expressed in percentage points. The exact days and number of folds used in each configuration are reported in Supplementary Table S1. The performance values should be interpreted as exploratory estimates of within-dataset model behavior rather than precise estimates of generalization to other species, water matrices or experimental conditions.

S1.4. Shapley Additive exPlanations (SHAP) analysis

SHAP analysis was applied to the XGBoost models using *TreeExplainer*. For each cross-validation fold, SHAP values were calculated for the held-out observations and subsequently combined across folds. Mean absolute SHAP values were then used to quantify the average contribution of each feature to the corresponding model predictions. Configuration 1 was retained only as a full-information reference because the expected prominence of contemporaneous target-metal concentration reflects its direct mathematical relationship with removal efficiency. Interpretation therefore focused primarily on variables that remained influential in Configurations 2–4, after contemporaneous target-related information had been excluded or restructured. Furthermore, SHAP results were considered only where the corresponding XGBoost model presented adequate predictive performance. SHAP values describe associations within the fitted XGBoost models and do not establish causal relationships, physiological regulation or specific metal-removal mechanisms.

S1.5. Software

All ML analyses were implemented in Python. Data handling was performed using NumPy and pandas. GPR, feature preprocessing, cross-validation, encoding, and model performance calculations were done using scikit-learn. XGBoost was performed using the XGBoost package and SHAP analysis was performed using the SHAP package. The analysis was performed with the following versions:

- Python 3.12
- scikit-learn 1.8.0
- XGBoost 3.3.0
- SHAP 0.52.0
- NumPy 2.4.4
- pandas 3.0.2
- Matplotlib 3.10.8
- Seaborn 0.13.2

S2. GPR and XGBoost performance metrics

The performance metrics of the GPR and XGBoost models under the four input configurations are presented in the following tables.

Table S2. Configuration 1: Full feature configuration. Model performance.

Microalga	Metal	GPR			XGB		
		R^2	RMSE	MAE	R^2	RMSE	MAE
<i>C. vulgaris</i>	Pb ²⁺	0.952	6.00	3.25	0.674	15.67	9.54
	Zn ²⁺	0.943	5.81	3.50	-0.276	27.56	15.8
	Cd ²⁺	0.968	4.42	2.77	0.010	24.47	13.39
	Ni ²⁺	0.988	1.79	1.04	0.434	12.09	5.92
	Cu ²⁺	0.996	1.53	0.79	0.331	20.11	11.18
<i>A. platensis</i>	Pb ²⁺	0.997	1.46	0.99	-0.372	29.75	16.51
	Zn ²⁺	0.902	9.56	5.60	0.824	12.78	7.45
	Cd ²⁺	0.97	4.74	3.03	-0.230	30.12	16.26
	Ni ²⁺	0.533	14.11	9	0.606	12.96	8.71
	Cu ²⁺	0.99	2.82	1.67	0.652	16.8	10.68

Table S3. Configuration 2: Reduced baseline configuration. Model performance.

Microalga	Metal	GPR			XGB		
		R^2	RMSE	MAE	R^2	RMSE	MAE
<i>C. vulgaris</i>	Pb ²⁺	0.772	8.65	7.06	0.549	12.18	9.46
	Zn ²⁺	0.634	6.53	5.21	0.507	7.58	5.75
	Cd ²⁺	0.790	4.95	3.66	0.551	7.25	5.52
	Ni ²⁺	0.961	2.29	1.70	0.926	3.16	2.54
	Cu ²⁺	0.628	9.97	5.00	0.489	11.68	6.99
<i>A. platensis</i>	Pb ²⁺	0.517	6.06	4.39	-0.430	10.42	8.26
	Zn ²⁺	0.628	13.78	6.48	0.496	16.05	7.82
	Cd ²⁺	0.931	2.89	2.32	0.873	3.91	3.05
	Ni ²⁺	0.508	9.87	7.31	0.168	12.84	9.40
	Cu ²⁺	0.350	13.78	7.01	-0.073	17.71	10.36

Table S4. Configuration 3: Cross-metal configuration. Model performance.

Microalga	Metal	GPR			XGB		
		R^2	RMSE	MAE	R^2	RMSE	MAE
<i>C. vulgaris</i>	Pb ²⁺	0.487	12.99	8.97	0.627	11.07	8.58
	Zn ²⁺	0.384	8.47	6.27	0.562	7.14	5.20
	Cd ²⁺	0.923	3.00	2.15	0.653	6.37	4.98
	Ni ²⁺	0.918	3.32	2.34	0.910	3.47	2.77
	Cu ²⁺	0.599	10.35	5.60	0.443	12.20	6.86
<i>A. platensis</i>	Pb ²⁺	0.389	6.82	5.10	0.155	8.01	6.27
	Zn ²⁺	0.645	13.46	7.46	0.560	14.98	7.56
	Cd ²⁺	0.886	3.70	2.72	0.868	3.99	3.00
	Ni ²⁺	0.295	11.81	8.15	0.060	13.64	10.38
	Cu ²⁺	0.510	11.97	6.84	0.155	15.72	9.27

Table S5. Configuration 4: Previous time point configuration. Model performance.

Microalga	Metal	GPR			XGB		
		R^2	RMSE	MAE	R^2	RMSE	MAE
<i>C. vulgaris</i>	Pb ²⁺	0.678	10.29	7.91	0.812	7.87	6.98
	Zn ²⁺	0.416	8.25	6.07	0.521	7.47	5.55
	Cd ²⁺	0.957	2.25	1.70	0.749	5.42	4.05
	Ni ²⁺	0.936	2.93	2.00	0.903	3.62	2.95
	Cu ²⁺	0.659	9.54	5.33	0.602	10.31	6.13
<i>A. platensis</i>	Pb ²⁺	0.288	7.35	5.42	-0.017	8.79	7.30
	Zn ²⁺	0.546	15.22	8.01	0.625	13.84	7.42
	Cd ²⁺	0.828	4.55	3.53	0.918	3.14	2.61
	Ni ²⁺	-0.134	14.98	11.54	0.115	13.24	9.77
	Cu ²⁺	0.526	11.78	6.30	-0.008	17.16	9.83

Table S6. Performance of the simple persistence baseline used for comparison with Configuration 4.

Microalga	Metal	R^2	RMSE	MAE
<i>C. vulgaris</i>	Pb ²⁺	-0.132	19.31	13.60
	Zn ²⁺	-4.144	24.48	13.35
	Cd ²⁺	-2.732	20.90	11.11
	Ni ²⁺	-0.513	14.26	7.33
	Cu ²⁺	-1.167	24.08	12.88
<i>A. platensis</i>	Pb ²⁺	-8.462	26.82	14.93
	Zn ²⁺	-0.153	24.27	11.50
	Cd ²⁺	-4.825	26.48	12.19
	Ni ²⁺	-1.407	21.83	14.47
	Cu ²⁺	-1.694	28.06	14.75

S3. pH evolution of the two microalgal species

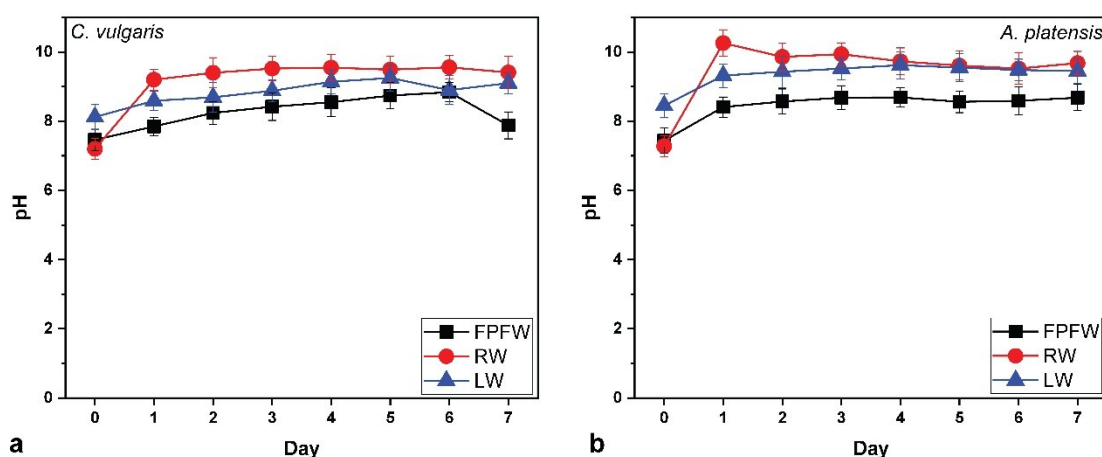


Fig. S1. Temporal evolution of pH during cultivation of *C. vulgaris* (a) and *A. platensis* (b), under exposure to the five-metal mixture in fish-processing facility wastewater (FPFW), river water (RW), and lake water (LW). Values represent the means of three independent biological replicates (n=3) and error bars indicate standard deviations.