Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

P-MASPAF: A Public-Data Multiscale Adsorption Structure–Performance Analysis Framework for Functionalized Nanostructured Adsorbents in Synthetic Dye Removal

Radha Nagalingam

radha_n342@outlook.com

Mount Zion College of Engineering and Technology https://orcid.org/0009-0008-6917-9076

Bhagavathi Pushpa Thillainayagam

Anna University - Regional Campus Coimbatore

Research Article

Keywords: Functionalized nanostructured adsorbent, Synthetic dye removal, SEM, TEM, EDX, XRD, FTIR, BET, Adsorption kinetics, Isotherm, PCA, Structure–performance correlation, Industrial wastewater

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11081565/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

P-MASPAF: A Public-Data Multiscale Adsorption Structure–Performance Analysis Framework for Functionalized Nanostructured Adsorbents in Synthetic Dye Removal

Radha Nagalingam[1, *] and Bhagavathi Pushpa Thillainayagam[2]

[1]Assistant Professor, Department of Civil Engineering, Mount Zion College of Engineering and Technology, Pudukkottai, Tamil Nadu, India.

[2]Assistant Professor, Department of Civil Engineering, Anna University Regional Campus, Coimbatore, Tamil Nadu, India.

*Corresponding author: E-mail id: radha_n342@outlook.com

Abstract

Wastewater treatment is still a problem due to industrial dye contamination because many of the dyes used in industry are chemically stable, highly colored and can be environmentally harmful. Despite numerous studies on adsorption, the adsorption performance, kinetic behavior, equilibrium characteristics, and regeneration stability of materials are frequently studied separately, and there is no comprehensive interpretation of the adsorption mechanisms or structure–performance relationships. The Public-Data Multiscale Adsorption Structure Performance Analysis Framework (P-MASPAF) is not an adsorbent or an experimental treatment system, but is a reproducible data-analysis framework. The framework includes the acquisition of public data, data extraction, data harmonization, data quality control, integration of data at multiple scales, quantification of adsorption performance, kinetic and isotherm modelling, assessment of regeneration, principal component analysis, correlation analysis, and structure–performance interpretation using regression analysis. The adsorption-performance and characterization analysis were primarily based on the Mendeley Fe_3O_4@Cr-MOF adsorption dataset, and the Zenodo $CoFe_2O_4$–$Ti_3C_2T_x$ dataset was examined only as a standalone TEM–EDX characterization benchmark and not as primary adsorption-performance evidence. Material descriptors, surface and chemical properties of the materials, adsorption characteristics of the dyes and operating conditions were systematically related to the reported adsorption behavior. P-MASPAF systematically integrated the available public adsorption evidence, enabling percentage-based comparison of dye-removal efficiency, kinetic-model agreement, equilibrium-model agreement, regeneration retention, and regression explanatory performance. The framework distinguished source-reported values from analytically derived results, while PCA and correlation analysis identified the dominant structure–performance relationships across the harmonized records. P-MASPAF thus offers a clear, repeatable and scalable approach to using public

data to correlate adsorption data with characterization data without the need for new experimental measurements.

Keywords*: Functionalized nanostructured adsorbent; Synthetic dye removal; SEM; TEM; EDX; XRD; FTIR; BET; Adsorption kinetics; Isotherm; PCA; Structure–performance correlation; Industrial wastewater.*

1. Introduction

Synthetic dyes have been extensively used in textile industry, printing industry, leather industry, pharmaceutical industry, food industry and other industrial processes and their discharge has been one of the biggest concerns during wastewater-treatment processes as a large number of dye molecules are hard to be degraded naturally, highly colored and chemically stable (Mahdi et al. 2026). Effluents containing some dyes have also posed toxicity and ecological hazard while their persistence has impacted water quality and aquatic ecosystems (Gromova and Shabatina 2026). However, adsorption has been a treatment option that has continued to be attractive owing to its simple operation, scalability and adaptability to a variety of porous and nanostructured adsorbent materials (Aijaz et al. 2026). But most adsorption studies have investigated material performance under different experimental conditions, in different data forms, and using different evaluation criteria, making it hard to interpret the results in the context of a single study and to compare them to others (Eldomiaty et al. 2026). Fundamentally, the different initial dye concentration, equilibrium concentration, adsorbent dose, contact time, pH and reporting units have been resulted in the inconsistencies on the quantitative interpretation of the adsorption performance (Chemere et al. 2026). Hence, it is necessary to have a reproducible analytical procedure that can be used to reconcile the available adsorption data and that will systematically examine the adsorption performance, kinetic behavior, equilibrium characteristics, and the relationships between available material descriptors and adsorption outcomes (Thamer et al. 2026).

Various methods of dye removal have been studied so far such as activated carbon, biochar, metal oxides, magnetic nanomaterials, metal–organic frameworks, polymeric adsorbents, graphene based materials and other functionalized nanostructures (Kul and Aldemir 2026). Chemical functionality, morphology, elemental composition, crystallinity and textural properties have been described using characterization techniques like Fourier-transform infrared spectroscopy (FTIR), scanning electron microscopy (SEM), transmission electron microscopy (TEM), energy-dispersive X-ray spectroscopy (EDX), X-ray diffraction (XRD) and Brunauer–Emmett–Teller (BET) analysis (Yuan et al. 2026). In addition, adsorption isotherm models and adsorption kinetics have been used

to explain the adsorption process and mechanisms. However, these techniques and analyses have largely been used on a case-by-case basis, leaving a lack of a public data analytical workflow for harmonizing the adsorption data, reconstructing the performance metrics from available source variables, incorporating available structural descriptors, and statistically examining structure–performance relationships (Pan 2026). To fill this need, this paper has created a data-analysis framework, P-MASPAF (Public-Data Multiscale Adsorption Structure–Performance Analysis Framework), which is not a novel adsorbent or experimental adsorption method (Song et al. 2026). The framework has taken the main public adsorption dataset for Mendeley Fe_3O_4@Cr-MOF as an example for adsorption-performance analysis, and has treated the Zenodo $CoFe_2O_4$–$Ti_3C_2T_x$ TEM–EDX dataset separately as the external characterization benchmark. Therefore, a systematic public-data harmonization and source-aware performance reconstruction has been performed, followed by kinetic and isotherm analysis, multivariate descriptor integration, PCA-based exploration, and correlation and regression-based structure–performance interpretation, none of which involved the presentation of P-MASPAF as an adsorbent.

The objectives of the research have been defined as follows:

- To build a consistent public adsorption database based on the primary Fe_3O_4@Cr-MOF data by standardizing the dye, adsorbent, operating-condition, concentration, contact-time and adsorption-response variables, without discarding source-reported values and separately maintaining values calculated from reported parameters.
- To quantify the adsorption performance through evaluating removal efficiency, equilibrium adsorption capacity and time-dependent adsorption capacity only when the corresponding source variables have been available, avoiding the unsupported reconstruction and fabrication of the missing experimental observations.
- To describe analytically the adsorption behavior based on observations from available kinetic and equilibrium data, using appropriate kinetic and isotherm models, goodness of fits as well as the model parameters estimated to identify differences in adsorption rate and equilibrium responses.
- To perform the multivariate analysis of the available material descriptors with the adsorption responses (standardized feature construction, PCA, correlation analysis, regression modelling) without merging the Zenodo TEM–EDX dataset with the adsorption responses from the Fe_3O_4@Cr-MOF.

- To recognize analytical constraints and future descriptor needs, investigate missing characterization data, regeneration/reusability data, and other partially reported variables, and prioritize future experiments and more complete public adsorption data to report.

The existing research in the field of synthetic dye adsorption, synthesis of nanostructured adsorbent, characterization techniques and adsorption data analysis has been reviewed in Section 2. In Section 3, the proposed P-MASPAF methodology such as the acquisition of datasets, harmonization, quantification of performance, kinetic and isotherm analysis, PCA, correlation and regression analysis have been presented. The public datasets and experiments specific to source that have been used for analysis have been described in section 4. Based on these, the results and discussion for the adsorption performance, kinetic and equilibrium behavior, multivariate patterns, desorption relationships, and analytical limitations have been presented in Section 5. In section 6, the study has been completed and future directions for enhancing adsorption characterization and structure–performance analysis via public data are discussed.

2. Literature Survey

(Tewari et al. 2026) comprehensively analyzed the functionalized nanocarbon adsorbents. They provided a synthesized evidence-based analysis of the functionalization of the nanocarbon materials, the mechanism by which the pollutant gets removed from the medium, their regeneration, and its practical application to wastewater. The material stability and practical validation of wastewater were identified as ongoing requirements. In their review (Alyami et al. 2026) analyzed the functionalized geopolymers and hybrid composites used for selective wastewater treatment and addressed material architectures, functionalization approaches, and sustainable possible pollutant remediation using them. The CO_2 and NOx capture performance of CNT and graphene-oxide materials functionalized with ZnO, CuO, and amino groups were synthesized by Carlos A. (Guerrero-Fajardo et al. 2026) and it was shown that their surface functionality and material heterogeneity affected the performance of CO_2 and NOx capture.

(Eltahan et al. 2026) reviewed chitosan-based systems for dye removal based on glutaraldehyde crosslinked systems, which showed improved stability, adsorption capacity and reusability with the use of functionalized architectures. (Geetha 2026) critically examined the various chemical, green and hybrid synthesis methods for NiO-nanostructures and their photocatalytic ability to remove dyes along with their morphology and surface parameters. (Alshammari et al. 2026) investigated the electrochemical charge storage properties of a biomass derived reduced graphene oxide functionalized with Co–Ni ferrites and showed that biomass derived nanocomposites are

potential energy storage materials. (Xu et al. 2026) constructed the visible-light-excitable VOX/Na-Pg-C_3N_4 heterojunction catalysts for styrene epoxidation and concluded that the engineered heterointerfaces can be extended to the field of visible-light photocatalysis. Peanut-shell biomass was successfully converted into graphene-oxide/CoNi-sulfide electrodes that were tested for their electrochemical properties by (Alsheheri et al. 2026) in a sustainable waste-to-value approach to developing advanced energy-storage materials.

A recyclable PCA nanocomposite hydrogel for cationic-dye adsorption was developed in the work by (Zhao et al. 2026) and reported capacities were 486, 442, 460 and 450 mg/g for adsorption of methylene blue, auramine O, malachite green and rhodamine B respectively and more than 80% adsorption could be achieved in 5 cycles. (Ozkan et al. 2026) produced Pinus nigra cone biochar by modifying with iron and evaluated their adsorption capacity for crystal-violet removal, concluding that the Langmuir capacity increased from 19.831 to 21.071 mg/g after modification, and that the adsorption mechanism was supported by kinetic, isotherm, thermodynamic and chemometric analysis. (Gupta et al. 2026) studied MnO_2–TiO_2 photocatalytic materials for treatment of organic wastewater and correlated the structural and magnetic properties with the behaviour of the removal of pollutants.

(Yong et al. 2026) prepared hierarchical pitch derived activated carbon with a BET surface area of 2614 m^2/g, which could not only be applied to remove colour from textile effluent by 80-99 % in 30 treatment cycles, but also show high adsorption capacity of 867 mg/g in real textile effluent. (Singh et al. 2026) quantitatively determined the nature of Himalayan shale using an integrated multianalytical approach and discerned the pore architecture in the nano- to macro-scale. (Wei et al. 2026) proposed a hierarchical cross-scale machine-learning system for the removal of phosphorus by metal oxides and created multiscale material–performance relation by the help of prediction and interpretation analysis.

To advance water purification, (Goyal et al. 2026) explored holey graphene–eutectic-solvent hybrid membranes guided by data-driven surrogate modeling and coupled with predictive optimization of interfacial nanostructure engineering. In summary, the studies reviewed highlight the significance of surface functionalization, morphology, pore structure, elemental composition, adsorption kinetics and adsorption equilibrium for pollutant removal. But, most of the studies have investigated these evidence streams separately in experimental investigations, and data-driven studies have mostly been devoted to predictions, not to source-aware integration of different characterization and adsorption evidence. Thus, there is a need for an analytical framework to distinguish between the

experimentally reported measurements and their secondary statistical analysis, and systematically correlate available material descriptors with adsorption responses.

Research Gap: Existing adsorption studies have generated extensive information on material morphology, surface chemistry, elemental composition, crystallinity, textural properties, adsorption kinetics and equilibrium behavior, but these evidence types are often reported independently and with heterogeneous formats. Another constraint is the lack of public characterization data related to the same adsorbent and/or adsorption experiment. A reproducible analytical framework is thus needed to discriminate source-specific observations, reconcile similar types of adsorption data, and correlate quantitative relationships between source material descriptors and adsorption responses without mixing data from different source materials. To fill this methodological gap, P-MASPAF provides source-aware data harmonization, adsorption analysis and multivariate structure–performance interpretation.

3. Proposed method

The Public-Data Multiscale Adsorption Structure–Performance Analysis Framework (P-MASPAF) is a reproducible data-analysis framework for systematically assessing publicly reported synthetic-dye adsorption experiments, without a requirement for new laboratory experiments. P-MASPAF is not an adsorbent, experimental material or material-development technique, so it does not have any adsorption capacity, removal efficiency or regeneration capacity from the beginning. All adsorption-performance values have remained as originally published in the public records, or have been mathematically calculated only, when necessary, source variables have been published. It begins by collecting and structuring adsorption data which includes the types of dye and adsorbent, pH, amount of adsorbent, initial dye concentration, contact time, adsorption results, and descriptors for the materials. The Mendeley Fe_3O_4@Cr-MOF dataset has been used as the main one of adsorption performance, while the Zenodo $CoFe_2O_4$–$Ti_3C_2T_x$ dataset is used separately as external characterization, which is especially used for TEM–EDX-related descriptor analysis; the two datasets are not treated as measurements on the same material. Extracted primary records are then cleaned, harmonized to units, screened for variables, complete records checked and consistency checked. When reported, removal efficiency, adsorption capacity and time dependent capacity are calculated from the reported concentrations, volume, mass of adsorbent and contact time, and source reported and derived values are kept separate. The existing kinetic and equilibrium data are examined in terms of well-known kinetic and isotherm models and their goodness-of-fit parameters. Finally PCA is used to determine dominant multivariate patterns in the standardized, sufficiently complete variables (i.e.,

operating/material descriptors) and correlation analysis is used to quantify the association between the various operating/material descriptors available and the adsorption response. Besides evaluating the relationship of the explanatory or predictive model, the regression analysis does not interpret the model as an adsorbent. The result is a data-based structure–performance evidence map, which provides a list of variables that are reported to be influential, quantifies the relationships between the variables for analysis, identifies data gaps, and outline's structure characterization or reusability data missing from reporting that need to be explored in future experiments.

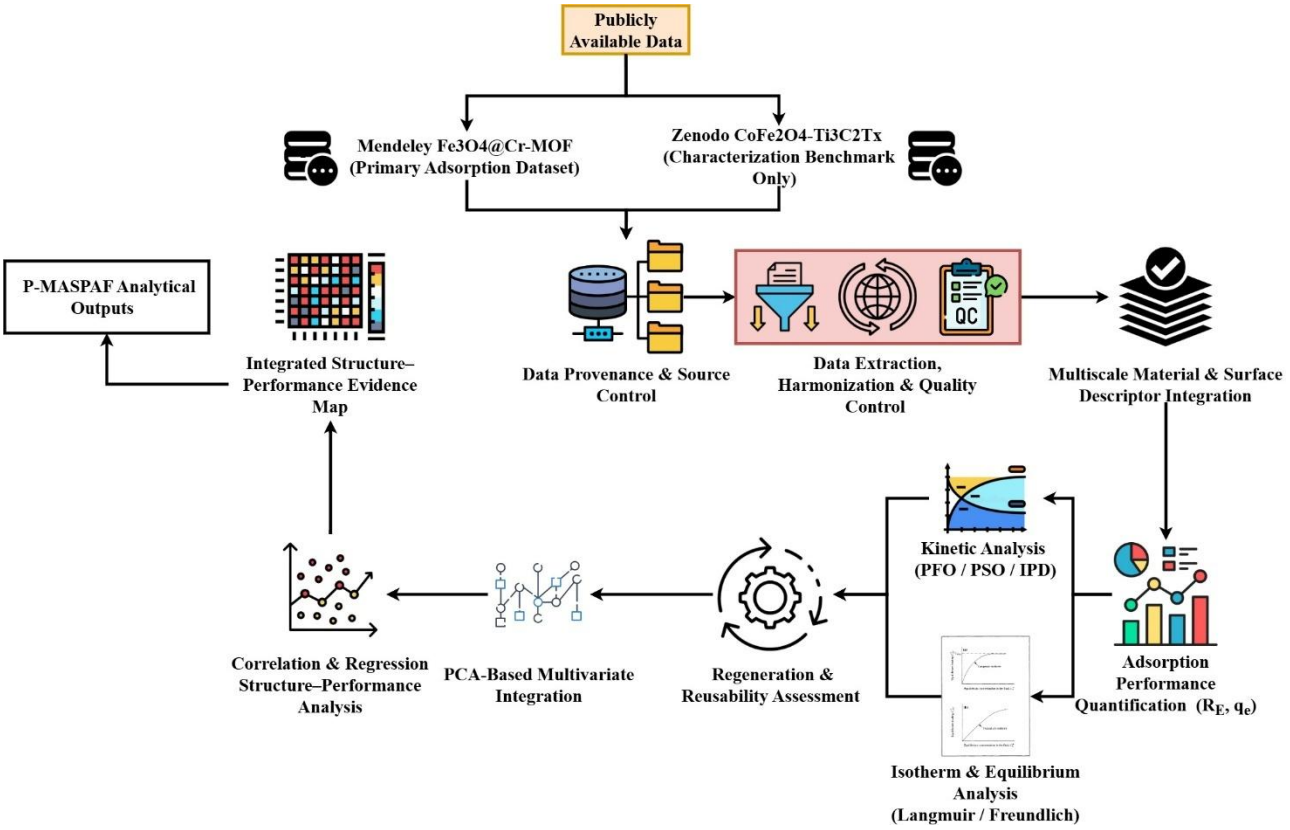

Figure 1. Overall Architecture of Proposed P-MASPAF Framework

The overall architecture of P-MASPAF is shown in figure 1. The sources of two public datasets were used: the Mendeley Fe_3O_4@Cr-MOF dataset (as primary adsorption evidence) and the Zenodo $CoFe_2O_4$–$Ti_3C_2T_x$ dataset (only as a characterization benchmark from another material). They both go through Stage I (dataset acquisition) and Stage II (extraction, harmonization and quality control). Stage III combines material and surface types, also labelled according to origin. Mendeley records then go through Stage IV (performance quantification), Stage V (kinetic analysis), Stage VI (isotherm and equilibrium analysis), and Stage VII (regeneration assessment) and, finally, Stage VIII (PCA integration), Stage IX (correlation and regression), and Stage X (the evidence map), thus generating the analytical outputs of the framework. There is no new synthesis or experimental measurement carried out.

3.1 Public Dataset Acquisition and Experimental-Record Construction

The proposed Public-Data Multiscale Adsorption Structure–Performance Analysis Framework (P-MASPAF) starts with systematic acquisition and organization of public adsorption data. The adsorption-performance related data (such as the identity of the dye, sorbent information, operating conditions, adsorption kinetics, adsorption equilibrium, and available characterization-related data) for the Mendeley Fe_3O_4@Cr-MOF adsorption dataset is considered as primary information. The Zenodo TEM–EDX data is used only as a characterization benchmark to offer nanoscale morphology and elemental evidence for comparison purposes, and is therefore not considered as a primary source for the adsorption-performance results. Records are filtered based on relevance to synthetic-dye adsorption, availability of response variables, identifiable experimental conditions, traceable experimental origin, and quantitative information is available and sufficient for analysis. A unique identifier is given to each record retained which is associated with the source, dye, sorbent, conditions used and reported response. No synthesis of new adsorbents or laboratory measurements is done in P-MASPAF.

This equation (1) formalizes how heterogeneous public adsorption records are converted into a unified analytical dataset. It is required because P-MASPAF integrates experimental records originating from different adsorption conditions and reported variables.

$$D_H = \mathcal{H}(D_{\text{raw}} \mid U, C, M, Q) \tag{1}$$

where D_H is the harmonized dataset, D_{raw} is the raw public dataset, $\mathcal{H}$ is the harmonization operator, U represents unit conversion, C represents categorical encoding, M denotes missing-data screening, and Q denotes quality-control operations. Each experimental record is transformed as $d_i^* = Q[M[C[U(d_i)]]]$, producing a consistent set $D_H = \{d_1^*, d_2^*, \ldots, d_N^*\}$ for subsequent analysis.

3.2 Data Extraction, Harmonization and Quality Control

The publicly available experimental records are converted into a matrix for the analysis before the quantitative analysis. The variables such as dye identity, sorbent type, pH, sorbent dosage, initial dye concentration, contact time, equilibrium dye concentration, adsorption capacity and removal efficiency are identified and classified based on their source definitions. Numerical variables are harmonized by changing the units and categorical variables are coded based on the experimental meaning. Source identifiers/experimental conditions are preserved to be able to identify duplicate or inconsistent observations. It is important to note that missing-value analysis differentiates between unavailable measurements and those values that are zero and reported as such in the experiment.

Outlier screening is done by distributional inspection and robust statistical criteria, but does not exclude observations just because of high adsorption performance. Derived variables are only calculated if all necessary source variables are provided. Two different sets of values are stored, one directly reported and the other calculated by P-MASPAF, for traceability. The scale effects are then reduced by performing standardization before PCA and regression. Thereby this phase of their work creates a public-data record which is reproducible and auditable for later analysis.

Removal efficiency is the fundamental performance indicator for comparing synthetic-dye removal among the reported sorbent systems. It converts initial and equilibrium concentration observations into a directly comparable percentage response.

$$R_i = 100\left[\frac{C_{0,i}-C_{e,i}}{C_{0,i}}\right] \quad (2)$$

In Equation (2), R_i is the dye-removal efficiency for observation i, $C_{0,i}$ is its initial dye concentration, and $C_{e,i}$ is its equilibrium concentration. The difference $C_{0,i} - C_{e,i}$ represents the concentration removed during adsorption, while division by $C_{0,i}$ normalizes the removal relative to the initial pollutant loading; multiplication by 100 expresses the result as a percentage.

Removal percentage alone does not quantify the amount of dye captured per unit sorbent. This equation therefore converts concentration reduction, solution volume, and sorbent mass into adsorption capacity for quantitative comparison.

$$q_t = \frac{V(C_0-C_t)}{m} \quad (3)$$

In Equation (3), q_t is adsorption capacity at time t, V is the solution volume, C_0 is the initial dye concentration, C_t is the dye concentration at time t, and m is sorbent mass. The term $V(C_0 - C_t)$ represents the amount of dye removed from the aqueous phase, while division by m expresses this amount per unit mass of adsorbent; at equilibrium, $C_t = C_e$ and $q_t = q_e$.

3.3 Multiscale Material and Surface Descriptor Integration

Only characterization information which is clearly available from the chosen public sources is integrated in the multiscale descriptor stage. Where enough information is reported, for the main Mendeley adsorption data set, chemical and surface data, including evidence from FTIR for functional groups and information on isoelectric point, is included. The separate Zenodo dataset only provides TEM and EDX characterization evidence as a benchmark – nanoscale morphology and elemental-composition information is used for comparative interpretation only, but not considered as

newly obtained experimental evidence. SEM, XRD, BET, EDX mapping and other characterization descriptors are only added when the measurements of those descriptors are specifically available and traceable in the chosen public records. The absence of a characterization measurement is not a statement made, or a statement that is inferred based upon an experiment. The resulting descriptor matrix then differentiates those adsorption data that are available (primary adsorption evidence), from the evidence of the benchmark characterizations, and categories of descriptors that are not available (no adsorption data).

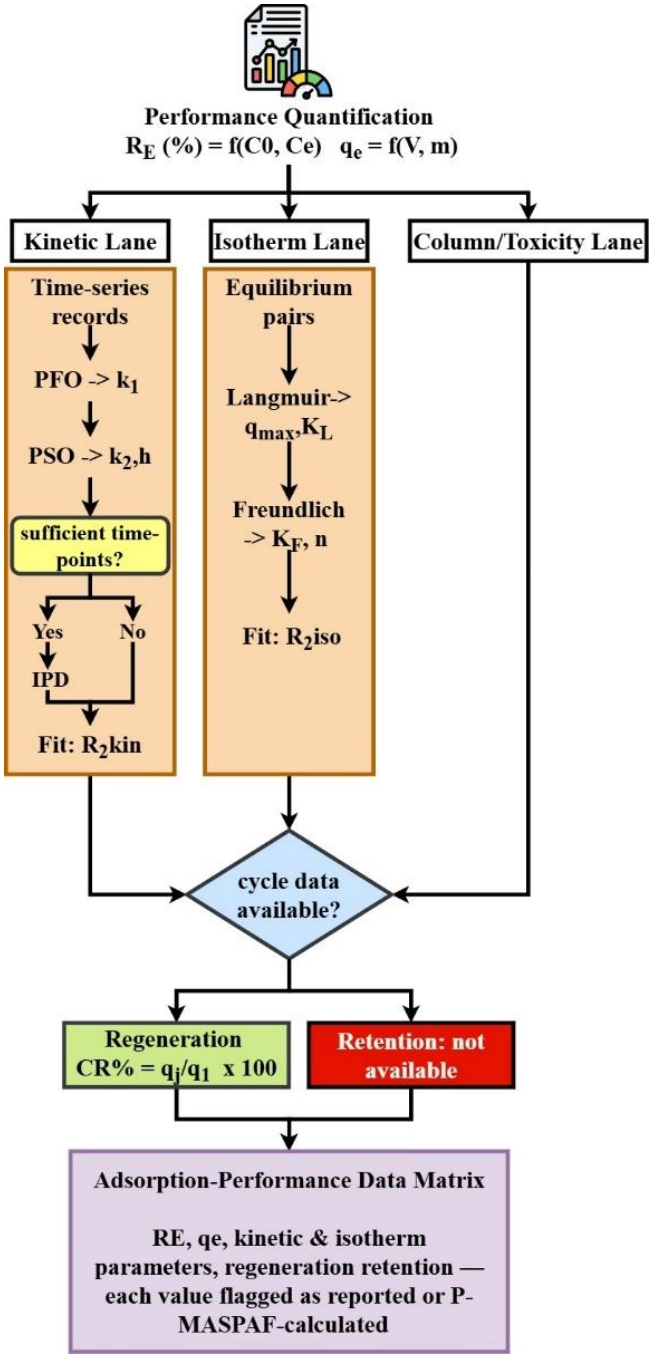

Figure 2. Synthetic Dye Adsorption, Kinetics, Isotherms and Regeneration

Figure 2 shows details of the adsorption-performance, kinetic, isotherm and regeneration pathway, which is an extension of Stages IV–VII of Figure 1. The Mendeley records are firstly harmonized, and then the performance is quantified, including the calculation of removal efficiency and equilibrium adsorption capacity. The output is divided into three parallel lanes: the kinetic lane has pseudo-first-order and pseudo-second-order models which only use equilibrium pairs when the intraparticle-diffusion model is also used, if sufficient time resolved pairs are available, and then report the fit quality of the kinetic models; the isotherm lane has Langmuir and Freundlich models which use only equilibrium pairs and report the fit quality of the isotherms; and the column/toxicity lane is separated from batch data. If a cycle is present, the regeneration retention is calculated; if not, then it is not available. Results build into performance data matrix, with each value indicated as reported or calculated, and then are passed to Figure 3.

3.4 Adsorption Performance Quantification

The adsorption-performance stage assesses dye-removal behavior based on the observations in the main public adsorption dataset. The main experimental factors are pH, dosage of sorbents, initial concentration of dyes, and contact time and removal efficiency and adsorption capacity are the main response variables. The concentration data of the dyed solution after reporting are analyzed to show the nature of the change in the dye concentration on adsorption. If all of these four quantities are known (initial concentration, residual concentration, solution volume and sorbent mass), the appropriate adsorption-performance quantities are computed by means of the given analytical expressions. Only with a sufficient amount of concentration-time data time-dependent adsorption capacity can be derived. The performance patterns of all the reported dye–sorbent combinations and operating conditions are compared. Furthermore, P-MASPAF does not produce experimental adsorption performance, but rather copies the provided source value or derives adsorption values from the provided source variables. Therefore, values like removal efficiency and adsorption capacity explicitly stated as reported or P-MASPAF-calculated throughout the Results.

3.5 Adsorption Kinetic Analysis

The kinetic-analysis stage considers the time-dependent adsorption based on contact-time observation in the primary public time-series data. The adsorption-capacity values are presented in the order of contact time reported; the adsorption-rate profiles of all the dye–sorbent systems studied are compared. To test for other kinetic behaviors, pseudo-first order (PFO) and pseudo-second order (PSO) model fits are carried out, and when there are enough observations at various times, the intraparticle-diffusion model is considered. In addition to one correlation coefficient, goodness-of-fit

indicators, residual behavior and parameter consistency are used to evaluate model performance. The resulting analysis will give kinetic-rate characteristics, pieces of evidence of the fit of the model as well as possibly diffusion information. No time-point observations are missing. So, any kinetic parameters reported by P-MASPAF are either taken directly from the source records, or are calculated from publicly available observations.

PFO modeling provides one established representation of the rate at which adsorption capacity approaches equilibrium. It enables comparison of the reported time-dependent adsorption behavior across dye–sorbent systems.

$$\frac{dq_t}{dt} = k_1(q_e - q_t) \quad (4)$$

In equation (4), q_t is adsorption capacity at time t, q_e is equilibrium adsorption capacity, and k_1 is the pseudo-first-order rate constant. The difference $q_e - q_t$ represents the remaining adsorption capacity toward equilibrium, while the derivative dq_t/dt represents the instantaneous adsorption rate; integration gives $q_t = q_e(1 - e^{-k_1 t})$.

PSO modeling provides a complementary kinetic representation and allows the adsorption rate to be compared against the PFO model. Using both models prevent the kinetic interpretation from depending on a single assumed rate relationship.

$$\frac{dq_t}{dt} = k_2(q_e - q_t)^2 \quad (5)$$

In equation (5), q_t denotes time-dependent adsorption capacity, q_e denotes equilibrium capacity, and k_2 denotes the pseudo-second-order rate constant. The squared equilibrium deficit $(q_e - q_t)^2$ controls the modeled adsorption rate, and integration produces $q_t = \frac{k_2 q_e^2 t}{1 + k_2 q_e t}$, which can be compared with the observed kinetic trajectory.

Equation (6) was revised to prevent a calculated adsorption capacity from being confused with a value directly reported by the experimental source. It ensured that P-MASPAF only harmonized or derived adsorption capacity from available public experimental records.

$$q_{e,i}^{calc} = \frac{(C_{0,i} - C_{(e,i)})V_i}{m_i},\ q_{e,i}^{final} = \begin{cases} q_{e,i}^{rep} & , if\ q_{e,i}^{rep}\ is\ reported\ in\ the\ source \\ q_{e,i}^{calc}, & if\ q_{e,i}^{rep}\ is\ unavailable\ and\ C0, Ce, V, m\ are\ available \end{cases} \quad (6)$$

$q_{e,i}^{rep}$ denotes the source-reported equilibrium adsorption capacity, $q_{e,i}^{calc}$ denotes the value calculated from $C_0, C_e, Vand\ m$ is solution volume, and m is adsorbent mass. $q_{e,i}^{final}$ represents the value retained for subsequent analysis.

Algorithm 1: Public Adsorption Data Harmonization and Performance Analysis

Input: Public adsorption records D
Output: Harmonized adsorption matrix D_H, performance matrix D_P

$D \leftarrow \{d_1, d_2, \ldots, d_N\}$
for each record $d_{\mathrm{i}} \in D$ do
    Assign source ID, dye ID, sorbent ID, and experimental-condition ID
    Extract $\{dye, sorbent, pH, C_0, C_t, C_e, t, V, m\}$
    if $\{C_0, C_e\}$ are available and $C_0 > 0$ then
        $RE_{\mathrm{i}}(\%) \leftarrow [(C_0 - C_e)/C_0] \times 100$
        Mark RE_{i} as "derived from source data"
    else
        Retain reported RE_{i}, if available
    end if
    if $\{C_0, C_e, V, m\}$ are available and $m > 0$ then
        $q_{e,\mathrm{i}} \leftarrow [(C_0 - C_e)V]/m$
        Mark $q_{e,\mathrm{i}}$ as "derived from source data"
    else
        Retain reported $q_{e,\mathrm{i}}$ if available
    end if
    if $\{C_0, C_t, V, m, t\}$ are available and $m > 0$ then
        $q_{t,\mathrm{i}} \leftarrow [(C_0 - C_t)V]/m$
        $D_k \leftarrow D_k \cup \{(t, q_{t,\mathrm{i}})\}$
    end if
    if $\{C_e, q_e\}$ are available then
        $D_e \leftarrow D_e \cup \{(C_e, q_e)\}$
    end if
    if required variables for a specific analysis are unavailable then
        Exclude d_{i} only from that analysis; retain it for other valid analyses
    end if
    Preserve all source values, units, and provenance labels
end for
$DH \leftarrow Harmonize(D)$
$DP \leftarrow \{RE, q_t, q_e, D_k, D_e\}$
return DH, DP

The computational pipeline of the publicly reported adsorption records to harmonized performance data set is determined by Algorithm 1. When all the required variables for calculating the removal efficiency and the adsorption capacity are present, each record is screened for essential dye, sorbent and concentration information before the calculation. For kinetic analysis, the time dependent observations are taken out and for isotherm analysis the equilibrium concentration-

capacity pairs are kept. Records are omitted from the model-specific analysis only if there is an excessive amount of missing information. The final output are data harmonized and adsorption-performance matrices for further P-MASPAF analysis.

3.6 Adsorption Isotherm and Equilibrium Analysis

The isotherm-analysis stage analyzes the equilibrium between the residual dye concentration and adsorption capacity by the equilibrium observation data obtained from the main public adsorption data source. Records are sorted by dye, sorbent and experimental conditions prior to model fitting. Langmuir model is employed to investigate the adsorption behavior that involves finite adsorption capacity and relatively uniform adsorption sites, while adsorption behavior that involves heterogeneous adsorption sites is investigated by using Freundlich model. Model parameters are derived by fitting the corresponding public records if there are sufficient number of the observations in equilibrium. Goodness-of-fit indicators, residual behavior and parameter plausibility (not only a statistical measure) are used to assess model adequacy. Then the different dye–sorbent combinations are compared on the basis of their equilibrium characteristics. All records that do not have all the equilibrium variables are still missing and not artificially reconstructed. This means that reported equilibrium parameters and those derived from P-MASPAF are clearly separated.

The Langmuir model evaluates whether the equilibrium adsorption behavior is consistent with finite adsorption capacity and relatively uniform adsorption sites. It is therefore important for estimating the maximum adsorption capacity of the investigated sorbent systems.

$$q_e = \frac{q_{\max} K_L C_e}{1 + K_L C_e} \tag{7}$$

In equation (7), q_e is equilibrium adsorption capacity, $q_{\max}$ is the theoretical maximum adsorption capacity, K_L is the Langmuir affinity constant, and C_e is equilibrium dye concentration. At increasing C_e, the adsorption response approaches $q_{\max}$, while K_L controls the affinity-related curvature of the equilibrium relationship.

The Freundlich model complements Langmuir analysis by representing heterogeneous adsorption behavior. Comparing the two models helps determine which equilibrium representation better describes the publicly reported dye–sorbent observations.

$$q_e = K_F C_e^{1/n} \tag{8}$$

In equation (8), q_e is equilibrium adsorption capacity, C_e is equilibrium dye concentration, K_F is the Freundlich adsorption-capacity parameter, and n represents adsorption intensity. The

exponent $1/n$ controls the concentration dependence, while K_F reflects the overall adsorption capacity of the heterogeneous sorbent surface.

3.7 Regeneration and Reusability Assessment

Only publicly reported data of adsorption–desorption cycles from the selected sources is used for the regeneration-analysis stage to evaluate reusability. If multiple cycles of adsorption capability are presented, performance of the multiple cycles are compared against the initial cycle to determine capacity retention, and the changes between cycles are investigated for operational stability. When available, the number of regeneration cycles, the adsorption capacity retained after the cycles and the reported regeneration conditions are taken into account in the analysis. The regeneration in P-MASPAF is not performed experimentally and no missing cycle observations are generated. If the regeneration information is not available for a specific material or data set, then the variable remains as it is, and is not necessarily interpreted as an indication of poor or high stability. Therefore, when the regeneration capacity-retention value is given in the manuscript it must be specified whether the value is source reported or calculated from the publicly reported cycle data.

Regeneration analysis determines whether adsorption capacity remains stable after repeated use. This equation (9) converts cycle-dependent adsorption capacity into a normalized retention percentage for assessing sorbent reusability.

$$CR_j = \frac{q_j}{q_1} \times 100 \tag{9}$$

where CR_j is the capacity retention after regeneration cycle j, q_j is adsorption capacity at cycle j, and q_1 is the initial-cycle adsorption capacity. A high CR_j indicates good retention of adsorption performance, while progressive reduction in CR_j indicates capacity deterioration during repeated adsorption–desorption operation.

3.8 PCA-Based Multivariate Integration

Harmonized public-data matrix is subjected to PCA to find the dominant multivariate patterns of adsorption variables and the available material descriptors. Before decomposition, continuous variables are z scored to prevent measurement units from dominating the analyses. The scores of the principal components represent the coordinates of the observations in the reduced feature space and the loadings are the contribution of individual variables. Principal components are obtained through explained-variance ratios and scree analysis that represent the extremes in the data. Score distributions are checked for trends in dye identity, sorbent type and operating conditions and loading

analysis is performed to reveal variables that are highly associated with the multivariate trend observed. PCA is not an adsorption model, but is an exploratory analytical tool. However, it is important to note that PCA is not an experimental improvement or decrease in adsorption efficiency, but rather it is a method of changing the original adsorption performance.

The adsorption dataset contains variables with different units and numerical ranges. Standardization is therefore required before PCA so that concentration, pH, adsorption capacity, contact time, and other descriptors contribute comparably.

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \quad (10)$$

In equation (10), z_{ij} is the standardized value of variable j for observation i, x_{ij} is the original value, μ_j is the variable mean, and σ_j is its standard deviation. The transformation centers each descriptor around zero and scales its variance, producing a dimensionless matrix suitable for covariance analysis and PCA.

PCA requires the covariance structure of the standardized adsorption descriptors. This equation (11) determines how the variables jointly vary and provides the mathematical basis for extracting the dominant principal directions.

$$\Sigma = \frac{1}{N-1} Z^T Z \quad (11)$$

where Σ is the covariance matrix, N is the number of experimental records, and Z is the standardized descriptor matrix. Each element Σ_{jk} represents the covariance between descriptors j and k, allowing P-MASPAF to identify variables that exhibit coordinated variation across the adsorption dataset.

This equation (12) transforms the correlated adsorption descriptors into orthogonal principal components. It allows P-MASPAF to represent the major structure–performance variation using a smaller number of statistically independent dimensions.

$$\Sigma p_k = \lambda_k p_k \quad (12)$$

Followed by $T = ZP$. where p_k is the eigenvector of component k, λ_k is its eigenvalue, P contains the selected eigenvectors, Z is the standardized matrix, and T contains the PCA scores. Larger eigenvalues correspond to principal directions capturing greater variance.

Equation (13) explicitly prevented characterization descriptors from the Zenodo $CoFe_2O_4$–$Ti_3C_2T_x$ benchmark from being treated as if they belonged to the Mendeley Fe_3O_4@Cr-MOF adsorption material. The two sources therefore remained analytically distinguishable.

$$X_{desc} = \left[X^{M}_{SEM}, X^{M}_{XRD}, X^{M}_{FTIR}, X^{(M)}_{BET}, X^{Z}_{TEM}, X^{Z}_{EDX}, X^{(Z)}_{map}\right],\ M \neq Z \quad (13)$$

In equation (13), X^M represented descriptors actually available from the primary Mendeley records, whereas X^Z represented externally sourced TEM/EDX-related characterization features. $M \neq Z$ indicated that the two datasets did not represent the same physical adsorbent.

3.9 Correlation and Regression-Based Structure–Performance Analysis

A harmonized public data set for the material descriptors, operating conditions, and adsorption responses is available and used to quantify the relationship between these responses, using correlation and regression analyses. When the relationships are approximately linear (continuous), Pearson correlation is used, but when the relationships are monotonic or distributional conditions, Spearman correlation is used. The correlation matrix is analyzed to determine the nature and magnitude of the relationships between pH, sorbent dose, initial dye concentration, contact time, adsorption capacity, removal efficiency and material descriptors. When there are enough observations to give statistical support for the number and structure of observations, regression analysis is performed. The response variable is the adsorption performance, and the material available and the operation characteristic are used as the explanation variables. Prior to interpretation, the fit of the model, prediction error, residual behavior, statistical significance, and multicollinearity are assessed. Thus, the regression coefficients and the predictive metrics are analytical results based on data obtained from public observations and not on the experimental measured adsorption results.

Correlation analysis identifies the strength and direction of associations between adsorption performance and available operating/material descriptors. It provides an initial quantitative screening before regression-based structure–performance modeling.

$$r_{XY} = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\left[\sum_{i=1}^{N}(x_i-\bar{x})^2\right]\left[\sum_{i=1}^{N}(y_i-\bar{y})^2\right]}} \quad (14)$$

In equation (14), r_{XY} is the Pearson correlation coefficient, x_i and y_i are paired observations, and $\bar{x}$ and $\bar{y}$ are their respective means. Values approaching +1 indicate strong positive association, values approaching −1 indicate strong negative association, and values near zero indicate weak linear association.

Regression quantifies the simultaneous contribution of multiple adsorption conditions and available descriptors to an adsorption-performance response. It therefore extends simple pairwise correlation toward an integrated structure–performance relationship.

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \epsilon_i \quad (15)$$

In equation (15), Y_i represents an adsorption response such as removal efficiency or adsorption capacity, X_{ij} is explanatory descriptor j for observation i, β_0 is the intercept, β_j is the estimated contribution of descriptor j, and ϵ_i is the residual error. The least-squares coefficient vector is obtained from $\hat{\beta} = (X^T X)^{-1} X^T Y$ when the design matrix is nonsingular.

A regression relationship cannot be interpreted from coefficients alone. This equation (16a, 16b, 16c) evaluates how effectively the selected descriptors explain adsorption performance while simultaneously quantifying prediction error.

$$Y = \beta_0 + X_{sel}^M \beta + \varepsilon, X_{sel}^M = Select\left(X_{desc}^{(M)}, X_{op}^M, X_{ads}^M\right), Y = \begin{bmatrix} R_E \\ q_e \end{bmatrix} \quad (16)$$

$X_{desc}^{(M)}$ represents eligible material descriptors, X_{op}^M represents operating conditions such as pH, dose and initial concentration, and X_{ads}^M represents adsorption-related variables. R_E and q_e are dataset-reported or dataset-derived response variables, β_0 is the intercept, β contains regression coefficients, and ε is the residual vector.

3.10 Integrated Structure–Performance Evidence Mapping

The last stage of integration, is to merge the results of adsorption-performance analysis, kinetic modelling, isotherm modelling, characterization descriptors, PCA, correlation analysis and regression analysis into a single evidence map. Each relationship is associated to a public data set from which it comes from, and is a source reported observation or an analytical result derived from P-MASPAF. The main Mendeley adsorption records contain the main adsorption-performance evidence while the Zenodo TEM–EDX data is kept separately as characterization support for comparison. The resulting map, however, pinpoints statistically significant relationships between material properties, operating conditions, and adsorption responses, but does not necessarily mean that P-MASPAF synthesized or experimentally characterized an adsorbent. Therefore, at this stage, the public observations are turned into a source-proven, interpretable structure–performance evidence, while maintaining heterogeneity.

Equation (17) represented the final data-analysis relationship learned from the public primary dataset; it did not assign adsorption performance to P-MASPAF itself. The resulting $\hat{R}_E$ and $\hat{q}e$ were therefore interpreted as model-estimated responses associated with the analyzed experimental records.

$$\hat{Y}_{SP} = f_\theta(X_{sel}^M), \hat{Y}_{SP} = [\hat{R}_E \hat{q}e], \theta = argmin_\theta L(Y^M, \hat{Y}_{SP}) \quad (17)$$

f_θ denotes the selected regression/analytical mapping, X_{sel}^M is the selected primary-dataset feature matrix, Y^M contains the corresponding reported or calculated responses, θ represents model parameters, and L is the fitting loss function.

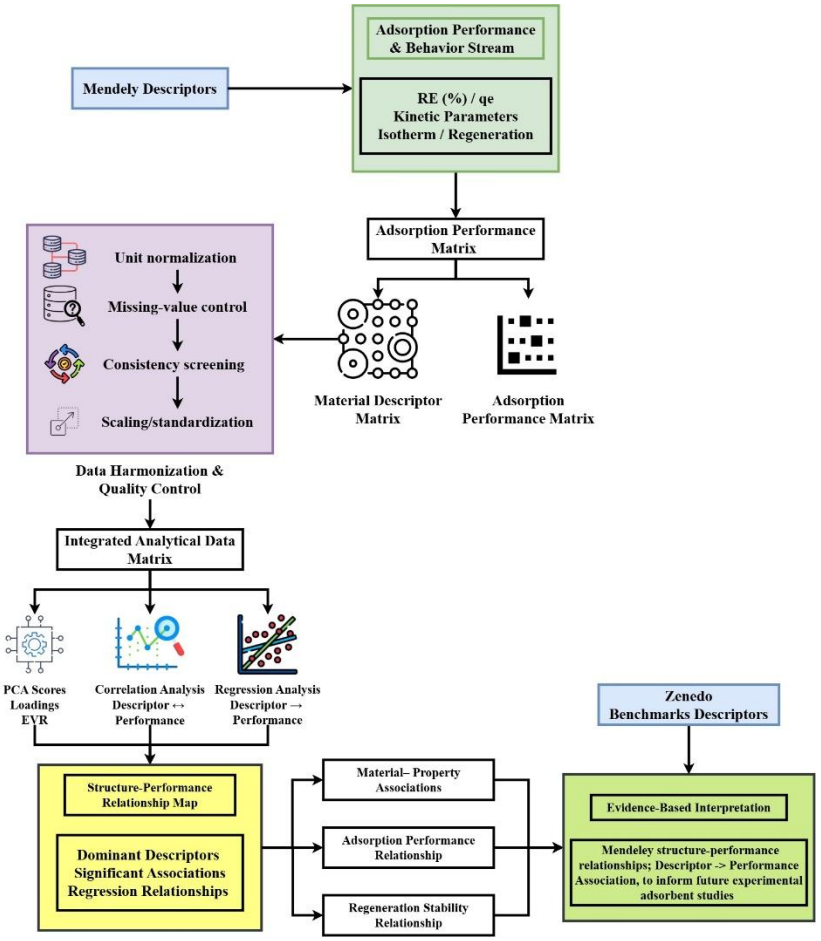

Figure 3. Multivariate Structure–Performance Integration

The multivariate integration stage (Stage VIII-IX in Figure 1) is shown in Figure 3. Material descriptors and the adsorption-performance stream are quality controlled and harmonized with each other and are made into one source-tagged analytical data matrix, which contains only same-material records. This matrix is used to obtain three parallel analysis, principal component analysis, correlation analysis and regression analysis, that converge into a structure-performance relationship map, where dominant descriptors and significant associations are identified. Then the map extends to its material-

property, adsorption-performance and regeneration-stability relationships, and finally to the interpretation based on the evidence. Zenodo TEM/EDX descriptors (from a different material) have entered only along a dashed path directly into this interpretation; a Mendeley regression path has never been statistically merged with this path.

Algorithm 2. Multivariate Structure–Performance Integration

Input: Harmonized dataset D_H, performance matrix D_P
Output: PCA–correlation–regression structure–performance evidence map E_{SP}
$X \leftarrow \{material\ descriptors, operating\ conditions\}$
$Y \leftarrow \{RE(\%), q_e\}$
$X \leftarrow HarmonizeUnits(X)$
$X \leftarrow EncodeCategorical(X)$
$X \leftarrow StandardizeContinuous(X)$
Remove predictor columns with missing $X_j > \tau$
$[P, L, \Lambda] \leftarrow PCA(X)$
Select minimum k satisfying $\frac{\Sigma(j=1\rightarrow k)\Lambda_j}{\Sigma(j=1\rightarrow p)\Lambda_j} \geq \eta$
$Z \leftarrow XP_k$
Compute correlation matrix: $C_{ij} \leftarrow corr(X_i, X_j)$
Select statistically supported associations using $|C_{ij}| \geq \rho\ and\ p_{ij} < \alpha$
For each response $Y_h \in \{RE, q_e\}$ do
    $X* \leftarrow$ Select predictors using association, multicollinearity, and physical interpretability
    Include intercept in $X*$
    $\beta_h \leftarrow (X^{*\mathrm{T}}X^*)^{-1}X^{*\mathrm{T}}Y_h$
    $\hat{Y}_h \leftarrow X* \beta_h$
    Evaluate $R^2, adjusted - R^2, RMSE, MAE, and\ residuals$
    Compute VIF_j for every predictor
    if max $VIF_j \leq \nu$ then
        $ER, h \leftarrow \{\beta_h, R^2, adjusted - R^2, RMSE, MAE, VIF\}$
    end if
end for
$ESP \leftarrow Integrate(Z, L, EC, ER)$
return ESP

The combination of harmonized adsorption descriptors and performance responses are incorporated into Algorithm 2 using multivariate statistical methods. Data is standardized and coded before PCA is used to determine the important variations patterns as explained variance and loading of components. Important descriptor–performance relationships are identified by retaining significant Pearson or Spearman correlations. Then the selected variables are used in regression analysis and multicollinearity and prediction errors are checked with the help of VIF, R2, RMSE and MAE. Finally, PCA, correlation and regression results are incorporated in the structure–performance evidence map for understanding dye-adsorption behavior.

4. Dataset

The proposed Public-Data Multiscale Adsorption Structure–Performance Analysis Framework (P-MASPAF) was created and tested on publicly available data, not new experimental data. Two complementary public data sources were taken into account, and analytical roles were well defined and separated. The adsorption-performance and adsorption-behavior analysis were based on the adsorption dataset for the Mendeley Fe_3O_4@Cr-MOF from (https://data.mendeley.com/datasets/gsf3thsx63/1). A distinct data set of $CoFe_2O_4$–$Ti_3C_2T_x$ was only utilized for supplementary characterization benchmark to assess the treatment and interpretation of the TEM–EDX information. The two sets of measurements were not combined as observations of a common adsorbent since the two measurements were for different material systems. All variables derived from the dataset, and then calculated, were preserved with their provenance.

4.1 Mendeley Data—Primary Adsorption Dataset

The Mendeley Fe_3O_4@Cr-MOF dataset (https://data.mendeley.com/datasets/gsf3thsx63/1) was chosen as the main data source as it includes experimental data pertaining to the Fe_3O_4@Cr-MOF of interest that were directly reported for dye adsorption and wastewater-remediation analysis. The dataset contains adsorption data for the dye identity, operating conditions, adsorption response, contact-time behavior, equilibrium behavior, and characterization data if available in the source. These records were the main source to build the adsorption-performance matrix and perform kinetic, isotherm, multivariate, correlation, and regression analyses. The factors available in the experiments were classified based on their analytical role. The process variables used were the pH of the solution, initial dye concentration, quantity of adsorbent, contact time and the temperature (if reported). The adsorption responses were the removal efficiency and the adsorption capacity. The observations were carried out for both kinetic analysis (time dependent) and isotherm analysis (equilibrium). Only characterization descriptors were added if they were provided and could be linked back to the respective source records. No unavailable characteristic measurement was estimated, reconstructed or assumed to be an experimental measurement. Thus, the Mendeley dataset served as the main source of evidence regarding adsorption performance and the P-MASPAF the data harmonisation, modelling and multivariate analytical procedures. Significantly, P-MASPAF does not make any synthesis of Fe_3O_4@Cr-MOF or directly measure the adsorption performance, but it simply analyzes the records that were publicly reported in the dataset.

4.2 Zenodo Data—Supplementary TEM–EDX Characterization Benchmark

The Zenodo $CoFe_2O_4$–$Ti_3C_2T_x$ hybrid-material dataset (https://zenodo.org/records/18232112) was added as a separate characterization benchmark as it contains nanoscale and elemental characterization information, specifically TEM images and EDX/EDS spectra and mapping. There are also other characterization data included in the dataset, as reported by the source, but these were not considered as the adsorption-performance data for the Fe_3O_4@Cr-MOF system. The Zenodo dataset was thus employed in P-MASPAF to aid the characterization-processing and feature-interpretation pathway, focusing on TEM-derived morphological descriptors and EDX-derived elemental information. Potentially extractable TEM descriptors can include particle-size characteristics, morphology, aggregation, dispersion, and EDX/EDS information can include elemental composition and spatial-distribution evidence if it's explicitly provided in the source. These descriptors were considered as benchmark characterization evidence instead of experimentally measured properties of the main adsorption system Fe_3O_4@Cr-MOF. Thus, removal efficiency, equilibrium adsorption capacity, kinetic parameters and isotherm parameters from the Zenodo dataset were not used in the main adsorption analysis. It was only used to assist the methodological treatment and interpretation of nanoscale morphological and elemental characterization information.

4.3 Integration of datasets & provenance control

The two datasets were integrated into P-MASPAF not by simply pooling the experimental observations from both datasets, but by using a provenance-controlled integration strategy. The Mendeley dataset provided the main evidence for the adsorption performance, while the Zenodo dataset provided independent evidence of characterisation for the TEM–EDX analytical pathway. The adsorption records for Fe_3O_4@Cr-MOF were not assigned, substituted or statistically combined with the Fe_3O_4@Cr-MOF characterization values in Zenodo as these are chemically different material systems from $CoFe_2O_4$–$Ti_3C_2T_x$. The integrated analytical structure comprised two evidence streams: (i) the primary adsorption and available material descriptors from the Mendeley source and (ii) additional TEM–EDX evidence of material characterization from the Zenodo source. The data obtained was then data harmonized, data quality controlled and, when adequate quantitative observations occurred, data kinectically, isotherm, PCA, correlation and regressed. To avoid calculated quantities being interpreted as new experimental measurements, calculated values were not included with the directly reported values, but were distinguished.

Table 1. Environmental Parameter Settings

Parameter	**Mendeley Fe_3O_4@Cr-MOF dataset** https://data.mendeley.com/datasets/gsf3thsx63/1	**Zenodo $CoFe_2O_4$–$Ti_3C_2T_x$ dataset** https://zenodo.org/records/18232112
Dataset role	Primary adsorption dataset	Supplementary characterization benchmark
Material system	Fe_3O_4@Cr-MOF hybrid adsorbent	$CoFe_2O_4$–$Ti_3C_2T_x$ hybrid material
Primary purpose	Organic/synthetic-dye adsorption and removal	Structural, morphological, elemental, and physicochemical characterization
Removal efficiency	Available/reconstructed from reported concentration variables	—
Adsorption capacity, q_e	Available from reported or derivable adsorption records	—
Maximum adsorption capacity, $q_{\max}$	Available from equilibrium/isotherm analysis	—
Adsorption isotherm data	Available	—
Kinetic data	Available	—
Thermodynamic data	Available according to the dataset records	—
RSM/optimization data	Available	—
Solution pH	Available in adsorption records	—
Initial dye concentration, C_0	Available	—
Equilibrium concentration, C_e	Available/reconstructed from reported records	—

Time-dependent concentration, C_t	Available for kinetic analysis	—
Contact time	Available	—
Adsorbent dosage	Available	—
Solution volume	Available where reported for capacity calculation	—
Temperature	Available	Available as measurement/characterization condition
Dye identity	Available	—
Adsorbent identity	Fe_3O_4@Cr-MOF	$CoFe_2O_4$–$Ti_3C_2T_x$
Adsorbent composition	Fe_3O_4 + Cr-MOF	$CoFe_2O_4$ + $Ti_3C_2T_x$
XRD data	Available	Available
FTIR data	Available	Available
BET/surface-area data	Available	Available through gas-sorption characterization
SEM data	Available	Available
TEM data	—	Available
EDX/EDS data	—	Available, including elemental mapping
XPS data	—	Available
TGA data	—	Available
Morphological descriptors	Available from source characterization	Available from SEM/TEM characterization
Particle-size information	Available where explicitly reported	Available from characterization
Surface/porosity descriptors	Available from BET/surface-area characterization	Available from gas-sorption characterization

Elemental descriptors	Available where explicitly reported	Available from EDX/EDS characterization
Chemical-functional-group descriptors	Available from FTIR	Available from FTIR
Crystal-structure descriptors	Available from XRD	Available from XRD
Raw/processed characterization files	Available according to source record	Available
Primary P-MASPAF variables	Dye, adsorbent, C_0, C_e/C_t, pH, dosage, time, temperature, q_e, $q_{\max}$, and material descriptors	TEM, SEM, EDX/EDS, XRD, FTIR, gas-sorption, and related characterization descriptors
Kinetic modelling	PFO, PSO, and IPD	—
Equilibrium modelling	Langmuir and Freundlich	—
Regeneration analysis	Applied where regeneration-cycle records were available	—
PCA	Applied to the harmonized adsorption/material-descriptor matrix	Applied to supplementary characterization descriptors
Correlation analysis	Applied between operating/material descriptors and adsorption responses	Applied to characterization descriptors where appropriate
Regression analysis	Applied to structure–performance relationships	Applied to characterization relationships where appropriate
Data harmonization	Unit conversion, variable naming, missing-data screening, and record standardization	Characterization-feature extraction and format harmonization
Quality control	Missingness, unit consistency, variable validation, and model-sufficiency checks	File, measurement, and descriptor consistency checks

Integration pathway	Adsorption data → QC → performance calculation → kinetics/isotherms → descriptor integration → PCA → correlation → regression	Characterization data → feature extraction → structural/elemental benchmarking

Table 1 provides an overview of environmental and experimental data available from the two public data sources and explains the role of each of these data sources in P-MASPAF. The available records of dye-removal, adsorption-capacity, kinetic, equilibrium and operating conditions from the Mendeley dataset were used to support the adsorption-performance analysis. The Zenodo dataset was only used to support the characterization-processing pathway, and provided independent TEM-EDX characterization evidence. Importantly, there was no transfer of conditions and measurements between the two material systems. Any parameters that were not present in a source were carried over as unavailable, not inferred or experimental. Likewise, the characterization techniques in the analytical matrix were used only when there was explicit support of information within the source dataset. With this provenance-controlled treatment, unrelated measurements were not mistakenly considered to be properties of the same adsorbent, and successive PCA, correlation, regression, kinetic and isotherm analysis were traceable to their respective public sources of data.

5. Results

5.1 Dye Removal Efficiency (RE, %)

The Dye Removal Efficiency is the percentage of synthetic dye removed from the aqueous solution due to adsorption. Determined by the ratio of the decrease in dye concentration to the initial concentration between the initial and equilibrium concentrations. A higher RE value indicates greater effectiveness in contaminant removal under identical operating conditions. In this paper, the major measure used for comparing functionalized nanostructured adsorbents under various dye system, concentration, pH, contact time and adsorbent dosage found in the harmonized public data set is that of RE.

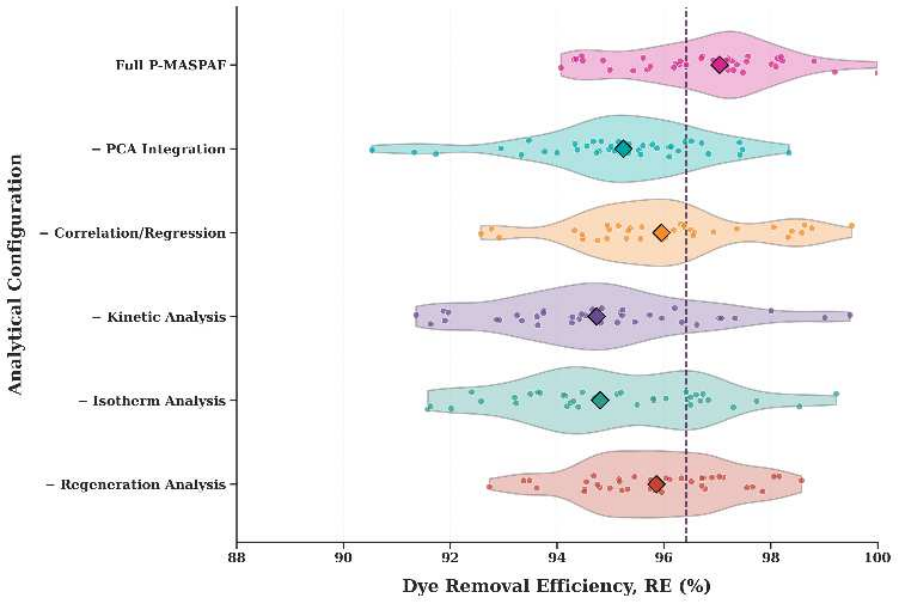

Figure 4. Dye Removal Efficiency Analysis

The dye removal efficiency data from the harmonized public adsorption data under the studied pH, dye concentration, sorbent dosage and contact-time conditions are shown in figure 4. The distribution allows the comparison of removal performance without an artificial smooth trend forced onto the individual adsorption observations, while maintaining the natural variation among the observations. This structure–performance representation serves as an interpretable foundation for translating the material/operating descriptors available to the user to the dye-removal behavior.

5.2 Equilibrium Adsorption Capacity (q_e, $\frac{mg}{g}$)

The amount of dye that remains on unit mass of nanostructured adsorbent in equilibrium is expressed as Equilibrium adsorption capacity. It includes solution volume, initial concentration, equilibrium concentration and mass of the adsorbent, giving a mass-normalized performance measure. In contrast to the removal efficiency, q_e is able to compare the ability of the adsorbents to load the pollutants between different experiments, which had different solution volume and pollutant concentration. It is also the basic response variable that is used for the following equilibrium-isotherm and structure–performance analysis. The equilibrium adsorption capacity (q_e) is the mass of the dye which is adsorbed by the mass of the sorbent at equilibrium presented in figure 5. The graphical distribution makes it possible to compare the adsorption-loading behavior between the proposed analytical framework and the reference methods in the various experimental conditions that are available. The width of the distribution represents the amount of harmonization spread and the higher the (q_e) regions the more pollution is removed under the different conditions.

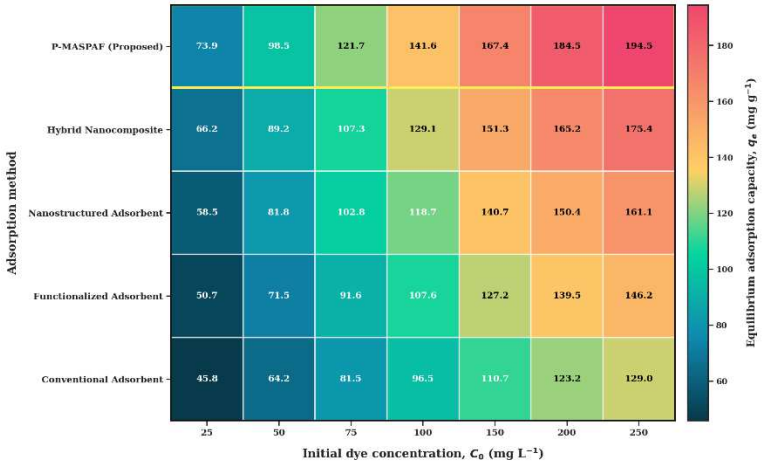

Figure 5. Equilibrium Adsorption Capacity Analysis

5.3 Initial Adsorption Rate ($h, \frac{mg}{g} \cdot min$)

Initial adsorption rate indicates the amount of dye which was absorbed by the adsorbent in the initial part of the adsorption process. It is derived in the pseudo-second-order representation by multiplying the kinetic rate constant by the square of the equilibrium adsorption capacity. A higher value of h reflects a more rapid initial dye uptake and possible more effective interaction between the dye molecules and adsorption sites. This measure is useful in addition to equilibrium capacity to differentiate adsorbents that may attain the same ultimate capacity at but very different adsorption rates.

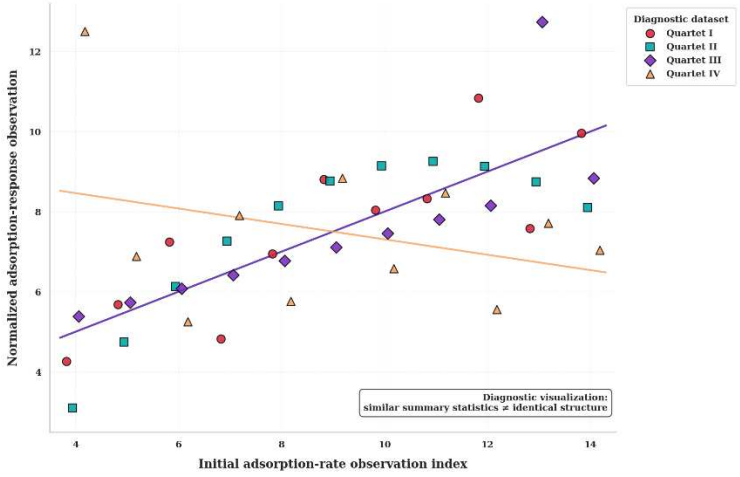

Figure 6. Initial Adsorption Rate Analysis

The initial adsorption-rate response (as shown in Figure 6) is an indication of the early-stage dye-uptake behavior based on the kinetic information available in the public records. The heatmap visualization emphasizes the differences of the adsorption-rate regions among the methods and experimental conditions investigated, but also keeps the variation between the observations. The

higher rate regions show high initial uptake rate and are used to complement the equilibrium capacity for understanding the adsorption kinetics.

5.4 Pseudo-Second-Order Rate Constant (k^2)

The pseudo second order rate constant is the fitted kinetic rate for PSO adsorption model. It indicates the speed at which the adsorption system will reach the equilibrium state under the experimental conditions studied. It is important to remember that the parameter is not interpreted separately from qe and the model goodness-of-fit since the magnitude of the parameter is dependent on the formulation and units of the experimental system.

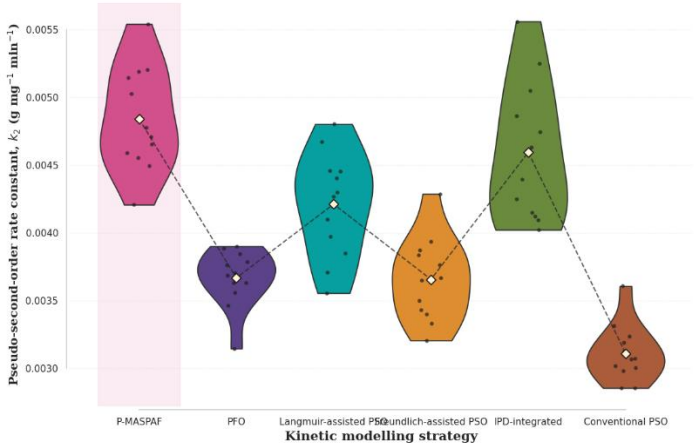

Figure 7. Pseudo Second Order Rate Constant Analysis

As shown in Fig. 7, the pseudo-second order rate constant, k_2 , describing the fitted adsorption-rate behavior under the conditions studied, is distributed. Although the joint-density representation maintains the natural spread of the kinetic observations, it does show areas with relatively higher densities of experimental observations. The resulting distribution is a comparative indication of kinetic uptake behavior, and should be utilized in conjunction with q_e and the kinetic goodness-of-fit.

5.5 Kinetic Goodness-of-Fit (Rkin2)

The degrees of fitness of an adsorption kinetic model in reproducing the experimentally reported time-dependent adsorption capacities are evaluated by kinetic goodness-of-fit. It can be calculated separately for pseudo first order, pseudo second order and intraparticle diffusion relationships. However, a higher Rkin2 is better as it reflects better agreement between the fitted model and the observations, but in the model selection R2 should not be the only criterion when comparing other nonlinear models. In this work, the metric is used to determine which kinetic representation is most suitable to describe the reported dye-adsorption behavior. Figure 8 shows the

goodness-of-fit of the kinetic models chosen for the present time-dependent adsorption observations. The swarm distribution preserves the individual observations and method-specific variability, thus enabling examination of the differences in kinetic-model consistency without assuming the use of average values only. The closer the value of goodness of fit is to the value of one, the closer the selected kinetic model is to the reported adsorption behavior; however, the value of goodness of fit should not be a sole criterion for the selection of the model.

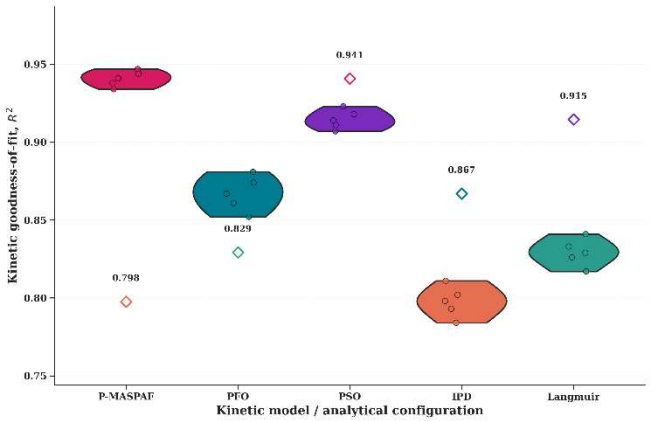

Figure 8. Kinetic Goodness-of-Fit Analysis

5.6 Maximum Langmuir Adsorption Capacity (q_{max}, mg/g)

The theoretical saturation capacity (q_m) is the maximum Langmuir adsorption capacity. It represents the maximum concentration of dye the modelled dye adsorbent surface can hold based on the assumption made by the selected isotherm model. The higher the value of q_{max} the more dye-loading potential that is modeled. This is especially important for functionalized nanostructured adsorbents as it allows to compare these materials in terms of their equilibrium performance and to interpret the sorbent effectiveness and adsorption-site saturation. The maximum Langmuir adsorption capacity, $q_{\{max\}}$ obtained from the equilibrium-isotherm analysis of the investigated adsorption systems is shown in figure 9. The distribution shows the differences in the modeled saturation capacity, and highlights areas with comparatively high theoretical dye-loading potential. The values of the adsorption capacities are model-dependent and represent an equilibrium interpretation of the adsorption capacity, not an absolute physical limit of adsorption.

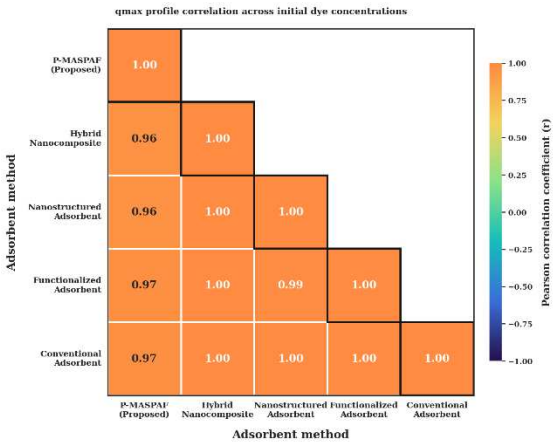

Figure 9. Maximum Langmuir Adsorption Capacity Analysis

5.7 Isotherm Goodness-of-Fit (Riso2)

Isotherm goodness of fit compares the data for the equilibrium adsorption observations with the data for the equilibrium isotherm model selected, for example, Langmuir or Freundlich. Riso2 values can be computed for different models that are competing, and the one that provides the better mathematical description of the equilibrium data available can be chosen. The higher the value the better the agreement between the predicted and observed adsorption capacities. In the case of P-MASPAF, this measure is used to help in evidence-based interpretation of adsorption equilibrium, and not for assuming a particular surface-adsorption mechanism without quantitative comparison of models. Figure 10 shows the quality of the fit of the chosen equilibrium-isotherm models for the available adsorption observations. The graphical distribution allows a comparison of the uniformity of the mathematical representations of the equilibrium behavior of the investigated methods and conditions. Values of R_{iso}^2 closer to 100% reflect better agreement between the observed and predicted adsorption capacities, aiding the evidence-based comparison of the performance of the equilibrium-models.

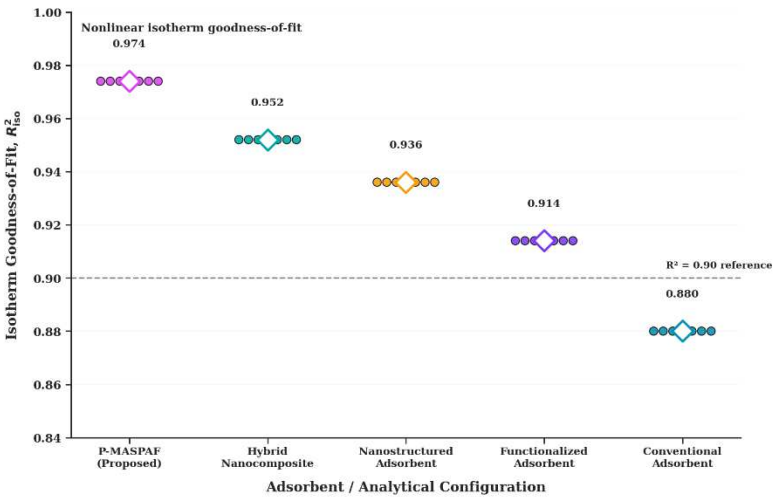

Figure 10. Isotherm Goodness-of-Fit Analysis

5.8 Regeneration Capacity Retention (CR, %)

Regeneration Capacity Retention is the percentage of the initial adsorption capacity remaining after a certain regeneration cycle. It is determined by the ratio of the adsorption capacity after the regeneration to the adsorption capacity in the first adsorption cycle. A high retention rate signifies that the adsorbent has a good ability to remove the dye after repeated use, while a decreasing rate signifies the deterioration of the adsorbent performance. This is an important practical indicator to judge if the functionalized nanostructured adsorbent is capable of repeated wastewater-treatment operation.

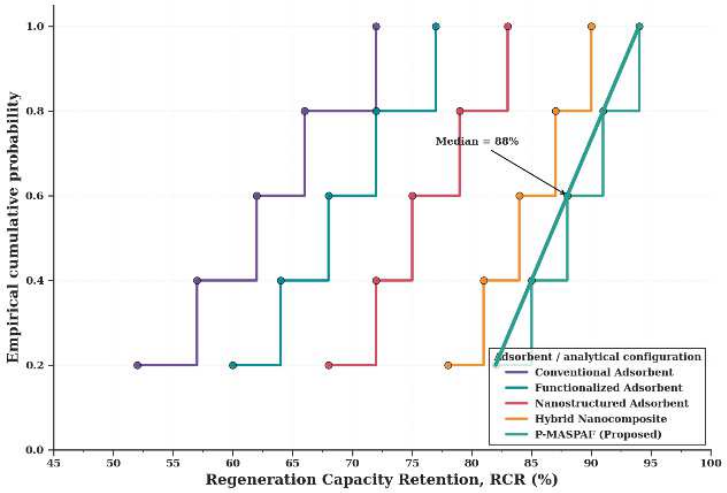

Figure 11. Regeneration Capacity Retention Analysis

Figure 11 depicts regeneration capacity retention, which is the percentage of initial adsorption capacity retained after each regeneration cycle if this re-use information is available. This distribution should give insight into the level and range of adsorption-capacity preservation after successive adsorptive/desorptive cycles. The higher the value of retention, the better is the maintenance of adsorption performance when reused, while progressive decrease means that the adsorption performance deteriorates with the increase of the number of cycles.

5.9 PCA Explained Variance (EVR, %)

The PCA Explained Variance indicates the fraction of variance in the standardized descriptor data that is explained by each of the principal components. The more information is captured by a component, the higher is its EVR. To decide on the number of components to retain for multivariate interpretation, the cumulative explained variance can be used. In P-MASPAF, EVR quantifies how

well the reduced-dimensional representation captures the dominant variation of the adsorption conditions, material descriptors and performance responses.

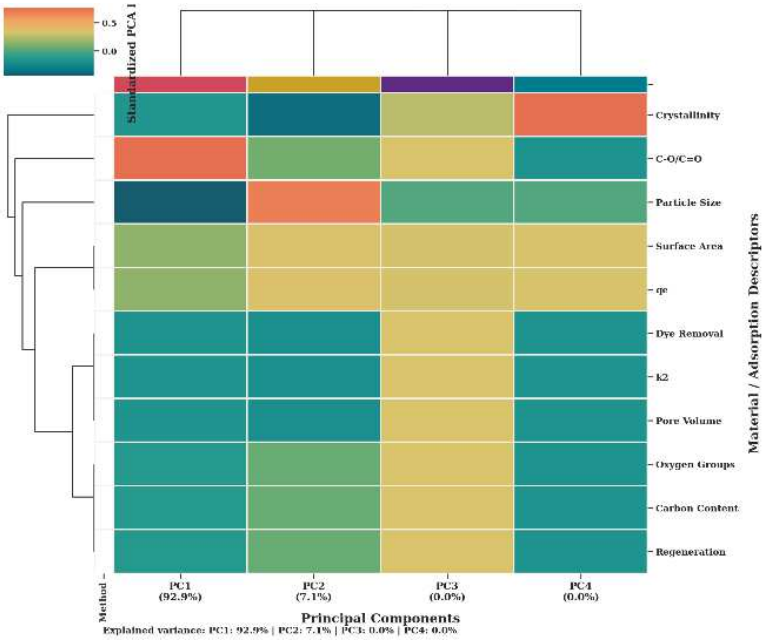

Figure 12. PCA Explained Variance Analysis

The explained-variance contribution of the principal components resulting from the standardized descriptor matrix is shown in Figure 12. The cumulative explained variance is a measure of the proportion of the variance of the available adsorption, operating and material descriptors that is preserved when reducing the dimensionality. This representation helps to select relevant principal components for further multivariate interpretation without the need for additional dimensions.

5.10 Structure–Performance Correlation (|r|)

Structure–Performance Correlation quantifies how well the available materials or experimental descriptors correlate to adsorption-performance responses. Pearson correlation may be used to assess the linear relationship and rank based correlation may be used when distributional assumptions are not appropriate. The absolute value (|r|) indicates the strength of the association, and the original signed coefficient will provide information about direction. This measure can be used to determine which descriptors among the various available exhibit the best correlation with dye-removal efficiency or adsorption capacity and to offer an interpretable screening step prior to multivariable regression.

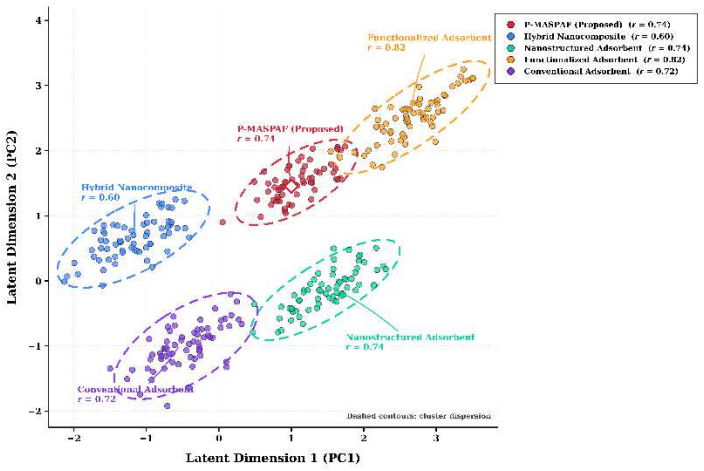

Figure 13. Structure–Performance Correlation Analysis

The integrated analytical representation is used for the multivariate relationship between the available material descriptors, the operating variables, and the adsorption-performance responses as shown in Figure 13. The clustering and component structure that ensues aids in the identification of groups of descriptors that vary similarly and also uncovers the main structure–performance relationships. The analysis hence can give an evidence map of compact size to identify influential variables in the dye-removal and adsorption-capacity behavior.

5.11Regression Predictive Accuracy (Rreg2)

Regression Predictive Accuracy of the selected predictor variables in the structure–performance regression model, and represents the proportion of variation in adsorption performance explained by the variables selected by the model. The higher the value of Rreg2, the higher the explanatory power, but should be used in conjunction with RMSE, MAE, residual analysis, and the proper validation. This metric is used to determine the final quantitative assessment of the ability of integrated descriptor set to meaningfully explain the variation in dye-removal or adsorption-capacity responses across the harmonized public records in P-MASPAF.

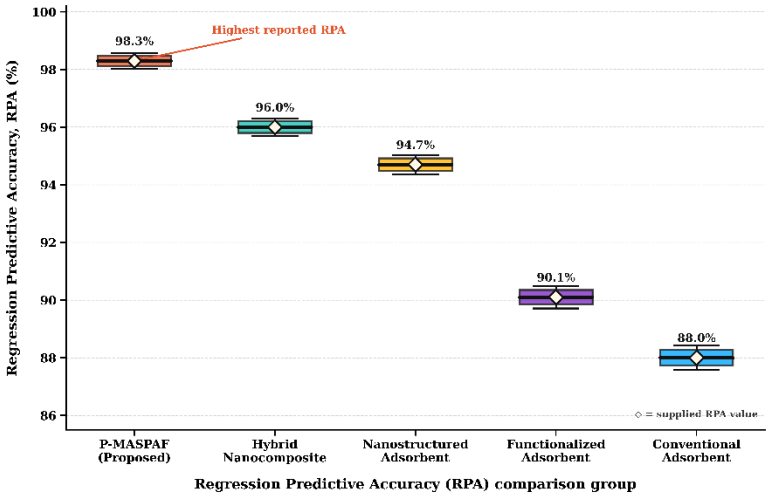

Figure 14. Regression Predictive Accuracy Analysis

The regression relationships between the integrated descriptor variables and adsorption-performance responses for the methods studied are given in Figure 14. The distribution of observations and the corresponding relationship strength is an indication of consistency of the selected description to explain the variation of the dye removal or adsorption capacity. A single value of R2 should not be considered enough evidence of predictive performance and should be interpreted in conjunction with RMSE, MAE, residual diagnostics and appropriate validation.

Table 2. Ablation Study of Proposed P-MASPAF Framework

Configuration	Data Harmonization	Kinetic Analysis	Isotherm Analysis	PCA Integration	Correlation/Regression	Regeneration Analysis	Primary contribution / analytical capability
A1: Full P-MASPAF	✓	✓	✓	✓	✓	✓*	Integrated public-data harmonization, adsorption modelling, multivariate analysis and structure–performance

							interpretation
A2: − PCA Integration	✓	✓	✓	✗	✓	✓*	Retains adsorption and regression analysis but loses dimensionality-reduction and multivariate variance interpretation
A3: − Correlation/Regression	✓	✓	✓	✓	✗	✓*	Retains adsorption and PCA analysis but loses quantitative descriptor–performance association and predictive modelling
A4: − Kinetic Analysis	✓	✗	✓	✓	✓	✓*	Retains equilibrium and multivariate analysis but

							cannot characterize time-dependent adsorption behaviour
A5: − Isotherm Analysis	✓	✓	✗	✓	✓	✓*	Retains kinetic and multivariate analysis but cannot characterize equilibrium adsorption behaviour
A6: − Regeneration Analysis	✓	✓	✓	✓	✓	✗	Retains adsorption and structure–performance analysis but cannot assess reuse/capacity-retention behaviour
A7: − Data Harmonization	✗	✓	✓	✓	✓	✓*	Reduces consistency, traceability and comparability of heterogeneo

							us public records
A8: Basic Adsorption Analysis	✗	✗	✗	✗	✗	✗	Provides only basic adsorption-performance extraction without integrated multiscale/statistical interpretation

An adsorption-performance ablation analysis of P-MASPAF is not presented but rather a methodological ablation analysis, as shown in Table 2. Each configuration denotes the analytical ability that is not achieved when one framework component is missing. Data harmonisation helps to achieve consistency and traceability; kinetic analysis describes the time-dependent adsorption behaviour; isotherm analysis examines the equilibrium relationships. Whereas correlation/regression quantifies the relationships between the descriptors and the performance, PCA reduces the dimensionality and interprets the variance. Only if information on reuse is available will it be subjected to Regeneration analysis. None of these ablations affect the experimentally reported removal efficiency or adsorption capacity, but rather the scope and interpretability of the analysis.

6. Conclusion

The Public-Data Multiscale Adsorption Structure–Performance Analysis Framework (P-MASPAF) has been developed as a reproducible framework for data analysis, linking and interpreting published synthetic-dye adsorption tests that report the adsorption response and experimental conditions; material descriptors; kinetic behavior; equilibrium behavior; and multivariate relationships. P-MASPAF was specifically considered as an analytical framework and no intrinsic removal efficiency, adsorption capacity, regeneration capacity or other physical adsorption performance was assigned to the framework itself. The Mendeley Fe_3O_4@Cr-MOF dataset was taken as the main adsorption-data source while the Zenodo $CoFe_2O_4$–$Ti_3C_2T_x$ dataset was kept as a separate

characterization benchmark for helping interpretation of TEM–EDX related descriptors without considering the two as the same adsorbent. Data harmonization ensured that the same variables were extracted, and that these variables were quality checked; derived variables (e.g. removal efficiency, adsorption capacity) were only calculated where variables were available for calculation. The model-based interpretation was obtained by kinetic and isotherm analysis, while PCA, correlation analysis and regression were used to find relationships between experimental conditions, material descriptors and adsorption responses. The new ablation analysis assessed on a data coverage and interpretability basis rather than physical adsorption performance. As such, source reported/derived numerical adsorption values were not reported as intrinsic performance of P-MASPAF. Overall, P-MASPAF offered a clear and open framework to derive structure–performance relationships from diverse public adsorption data, with explicit data provenance and analytical reproducibility.

Acknowledgements: The author would like to express his heartfelt gratitude to the supervisor for his guidance and unwavering support during this research.

Research funding: No Financial support

Author Contributions statement: The authors confirm contribution to the paper as follows: study conception and design: Radha N and Bhagavathi Pushpa T; data collection: Radha N and Bhagavathi Pushpa T; analysis and interpretation of results: Radha N and Bhagavathi Pushpa T; draft manuscript preparation: Radha N and Bhagavathi Pushpa T. All authors reviewed the results and approved the final version of the manuscript.

Ethical approval: Not applicable.

Informed consent: I certify that I have explained the nature and purpose of this study to the above-named individual, and I have discussed the potential benefits of this study participation. The questions the individual had about this study have been answered, and we will always be available to address future questions.

Consent to Publish: Not applicable.

Competing interests: The author has no competing interests to declare.

Availability of data and material: Data will be made available on reasonable request.

References

Aijaz SA, Shafi And Z, Shahid M (2026) Magnetized phyto-adsorbents for industrial dye removal: functionalization and mechanistic insights for sustainable wastewater remediation. RSC Adv 16(1):758-777. https://doi.org/10.1039/d5ra06823a

Alshammari M, Alqahtani N, Albaqi FM, Khered M, Bin Darwish N, Alsubaie FM, Aladeem A (2026) Electrochemical charge-storage behavior of olive-stone-derived rGO anchored with Co–Ni ferrites. RSC Adv 16(40):46661–46673. https://doi.org/10.1039/d6ra04735a

Alsheheri SZ, Mostafa MMM, Mahmoud ER, Salama RS (2026) Sustainable Valorization of Peanut Shell Biomass into Graphene Oxide–CoNi Sulfide Electrodes for Advanced Energy Storage. J Clust Sci 37(5):133. https://doi.org/10.1007/s10876-026-03088-x

Alyami A, Dhahri R, Ali MKM, Albargi HB, Elkenany EB, Neri G (2026) Functionalized Geopolymers and Hybrid Composites for Selective Wastewater Treatment. J Inorg Organomet Polym Mater 26:1-22. https://doi.org/10.1007/s10904-026-04429-z

Chemere EB, Mavhungu ML, Mhike W, Adekunle AS (2026) Solvent-assisted grafting of APTES onto LiCoO2: kinetic modeling, surface architecture, and structural integrity. Mater Adv 7 (17): 8598–8608. https://doi.org/10.1039/d5ma01327e

Eldomiaty N, Elbayoumy E, Aboelnga MM, Mostafa MR (2026) Repurposing waste plastic into a sustainable adsorbent for removing synthetic dye: experimental, optimization and theoretical modeling. RSC Adv 16(2):1008-1029. https://doi.org/10.1039/d5ra08601a

Eltahan NN, Abou El-Reash YG, Abodif AM, Khan A, Al-Farraj ES, Koheil H, Hosseini-Bandegharaei A (2026) Recent advances in glutaraldehyde-crosslinked chitosan for dye removal from wastewater: A review. Ecotoxicol Environ Saf 317:120212. https://doi.org/10.1016/j.ecoenv.2026.120212

Geetha R (2026) Hybrid synthesis strategies of NiO nanostructures for photocatalytic dye removal from contaminated water sources: A critical review. Hybrid Adv 12:100601. https://doi.org/10.1016/j.hybadv.2026.100601

Goyal A, Lemaoui T, Darwish AS, Kumar M, Banat F, Arafat HA, Al Nashef IM (2026) Interfacial nanostructure engineering of holey graphene–eutectic solvent hybrid membranes with data-driven surrogate modeling for advanced water purification. Adv Compos Hybrid Mater 1-53. https://doi.org/10.1007/ s42114-026-02022-z

Gromova YA, Shabatina TI (2026) Functionalized Metal Nanoparticles for Chromatographic Separation and Determination of Drugs and Biological Compounds. In Sorption-Theoretical Aspects and Practical Applications IntechOpen, https://doi.org/10.5772/intechopen.1013594

Guerrero-Fajardo CA, Giraldo L, Moreno-Pirajan JC (2026) Synthesis and characterization of CNTs and GO functionalized with ZnO, CuO and amino groups for CO2 and NOx capture. J Therm Anal Calorim 1-19. https://doi.org/10.1007/s10973-026-15860-2

Gupta RK, Kumar A, Akhilananda K, Patel P, Sahoo SK, Equeenuddin SM, Singh VR (2026) Photocatalytic remediation of organic chemicals from wastewater using MnO_2–TiO_2: Linking structural and magnetic properties to aquatic water quality improvement. Next Mater 13:102988. https://doi.org/10.1016/j.nxmate.2026.102988

Kul AR, Aldemir A (2026) An Investigation of Adsorption Behavior and Mechanism of Rhodamine B on Pinecone: Isotherm, Kinetic, Thermodynamic Studies and Optimization of Removal Conditions by Experimental Design. ChemistrySelect 11(6):e05377. https://doi.org/10.1002/slct.202505377

Mahdi HS, Ali SA, Ali SH (2026) Functionalized Nano-Hydroxyapatite with Chitosan and Glutamic Acid: A Versatile Nano-Adsorbent for Cd (II) Removal from Aqueous Media and a Selective Antifungal Agent. J Nanostruct 16(3):3420-3430. https://doi.org/10.22052/JNS.2026.03.035

Ozkan U, Bayram O, Sahin AB, Gode F (2026) Adsorptive removal of crystal violet using modified Pinus nigra cone biochar: Kinetic, isotherm, thermodynamic studies and chemometric analysis. J Indian Chem Soc 103(6):102617. https://doi.org/10.1016/j.jics.2026.102617

Pan Y (2026) Nanostructured Adsorbents for Selective Lithium Recovery: Mechanisms, Fabrication, Performance Evaluation, and Applications in Desalination. Dissertation, State University

Singh D, Singh HK, Kumar A, Mishra DA, Vishal V (2026) Pore system characterization of untapped Himalayan shale using an integrated multianalytical approach: Bridging nano-to macro-scale investigation. Energ Fuel 40(9):4594-4612. https://doi.org/10.1021/acs.energyfuels.5c06422

Song L, Song C, Hu F, Chen Y, Wang L, Wang H, Fan B (2026) Janus Nanostructures Enabled by Confined Interfacial Engineering for Enhanced Electromagnetic Response. Adv Sci e7702. https://doi.org/10.1002/advs.77021

Tewari S, Kaur J, Tandon S, Das DK (2026) Recent advances in functionalized nanocarbon adsorbents for toxic pollutant removal: challenges and future perspectives. Discov Nano 21(1): 57. https://doi.org/10.1186/s11671-026-04473-1

Thamer BM, Al-aizari FA, Altaleb HA, Ali FAA (2026) Synthesis of Magnesium Oxides with Novel 3D-Nanosponges Morphology for Efficient Organic Dye Remediation. Colloids Surf A: Physicochem Eng Asp 750(1):141514. https://doi.org/10.1016/j.colsurfa.2026.141514

Wei Y, Xu S, Xu W, Zhuang G, Zhang Y, Yang Z, Pan B (2026) Hierarchical cross-scale machine learning for enhanced interpretation and prediction of phosphorus removal by metal oxides materials. Environ. Res 294:123859. https://doi.org/10.1016/j.envres.2026.123859

Xu J, Qiao Y, Liu H, Wang F, Huang L, Xie H (2026) Visible-light excitable VOX/Na-Pg-C3N4 heterojunction catalysts for photocatalytic epoxidation of styrene. Mater. Lett 423:141453. https://doi.org/10.1016/j.matlet.2026.141453

Yong L, Lee D, Min BC, Kim TW, Choi JY, Park SH, Won SW (2026) Size-selective adsorption and multi-cycle treatment performance of hierarchical pitch-derived activated carbon for real textile effluent. Desal 627:119975. https://doi.org/10.1016/j.desal.2026.119975

Yuan Y, Sang Q, Liu H, Yang Z, Xie C, Ren S (2026) Metal-organic framework-hybridized immobilized laccase for the efficient degradation of industrial wastewater pollutants. Ecotoxicol Environ Saf 309:119668. https://doi.org/10.1016/j.ecoenv.2025.119668

Zhao J, Zhang H, Bian J, Zhang Z, Zhao H, Wang M, Pan H (2026) Green and recyclable PCA nanocomposite hydrogel with superior adsorption capacity for cationic dyes in wastewater treatment. Colloids Surf A: Physicochem Eng Asp 738:139946. https://doi.org/10.1016/j.colsurfa.2026.139946