IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, September 2026
Impact Factor: 8.2

Data Science in Chemistry and Materials Research: Machine Learning Approaches for Adsorption Performance Prediction and Materials Discovery

Mayuri Jadhav[1] and Aditya Khanna[2]
Department of Chemistry,
Satish Pradhan Dnyanasadhana College, Thane, Maharashtra, India
[2]khannasaditya@gmail.com

Abstract : *The accelerating volume of experimental and computational chemistry data has driven the rapid adoption of data science and machine learning (ML) methods across chemical and materials research. This review surveys the principal ML approaches used in chemistry and materials science — regression, ensemble tree-based methods, and neural networks — together with the general cheminformatics workflow of descriptor generation, feature selection, model training, and validation. Two quantitative case studies are examined in detail: the use of ensemble ML models (Random Forest, Gradient Boosting, XGBoost, and related algorithms) to predict heavy metal adsorption and removal performance from biochar, soil, and metal-organic framework (MOF) datasets; and the use of ML-accelerated high-throughput computational screening (HTCS) to identify high-performing MOF adsorbents from databases containing thousands to tens of thousands of candidate structures. Reported model accuracies (R^2 values ranging from approximately 0.76 to 0.99 depending on algorithm and dataset) and screening scales are compared and visualized. The review also discusses interpretability tools such as SHAP (Shapley Additive Explanations) that are increasingly used alongside predictive models to extract mechanistic insight rather than treating ML as a black box. Persisting challenges — data scarcity and quality, model interpretability, and generalizability across chemical space — are discussed, along with emerging directions including graph neural networks for materials property prediction and closed-loop autonomous experimentation. This synthesis aims to provide a practically oriented, quantitatively grounded reference for applying data science methods in chemistry and materials research, with particular relevance to adsorption-based environmental remediation.*

Key Words: data science, machine learning, cheminformatics, metal-organic frameworks, adsorption prediction, high-throughput screening, materials informatics

1. INTRODUCTION

Chemistry and materials research have historically progressed through iterative cycles of hypothesis, synthesis, and characterization — a process that is reliable but slow relative to the combinatorial size of accessible chemical and materials space. The growth of large experimental and computational databases, combined with advances in machine learning (ML) algorithms and computing power, has given rise to the interdisciplinary field of data science in chemistry, often termed cheminformatics or materials informatics [1,2]. Rather than replacing experimental and computational chemistry, ML methods are increasingly used to accelerate it: predicting properties before synthesis,

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 5, September 2026 Impact Factor: 8.2

prioritizing which candidate materials merit experimental investigation, and extracting structure–property relationships from datasets too large or high-dimensional for manual analysis [2,3].

Two applications of data science are especially relevant to environmental and materials chemistry: predictive modeling of pollutant adsorption performance, where ML models learn the relationship between adsorbent properties (surface area, pH, functional groups) and removal efficiency directly from experimental datasets [4,5]; and machine-learning-accelerated high-throughput computational screening (HTCS), where ML models are trained on a subset of computationally simulated materials and then used to rapidly predict the performance of thousands of additional candidates without requiring a full simulation for each one [6,7].

This review outlines the principal ML algorithms and the general cheminformatics workflow used in chemistry and materials research, and then presents two quantitative case studies grounded in the adsorption and remediation chemistry relevant to this research group: ML-based prediction of heavy metal adsorption/removal performance, and ML-accelerated screening of metal-organic frameworks for adsorption applications. Comparative tables and figures are used to connect algorithm choice and dataset scale to reported predictive accuracy and screening throughput.

II. MACHINE LEARNING METHODS USED IN CHEMISTRY AND MATERIALS RESEARCH

Most applications of ML in chemistry fall into a small number of algorithm families. Linear and multiple linear regression remain widely used as interpretable baseline models but struggle to capture the nonlinear relationships common in adsorption and materials data [8]. Ensemble tree-based methods — Random Forest (RF), Gradient Boosting (GB), Extreme Gradient Boosting (XGBoost), Adaptive Boosting (AdaBoost), and Light Gradient-Boosting Machine (LightGBM) — build many decision trees and combine their predictions, and have become the dominant approach for tabular chemistry datasets owing to their strong performance on nonlinear, moderate-sized data without extensive tuning [4,9]. Deep learning approaches, including artificial neural networks (ANN) and, for structural data, graph neural networks (GNNs) that operate directly on molecular or crystal graphs, are increasingly applied where larger datasets or structural representations are available [10,11].

The general workflow common to these approaches is summarized in Scheme 1: raw experimental or computational data is assembled into a structured dataset, relevant descriptors (features) are generated or selected, a model is trained on a portion of the data and validated on a held-out test set, and the resulting model is used for prediction on new candidates — optionally paired with an interpretability step to extract which features drive the prediction.

Data collection → Feature/descriptor generation → Model training (train/test split) → Validation (R^2, RMSE, MAE) → Prediction on new candidates → Interpretability (SHAP, PDP)

Scheme 1. General cheminformatics/materials-informatics workflow for supervised machine learning in chemistry and materials research.

Model performance is most commonly reported using the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE) on a held-out test set not used during training, which guards against overfitting and gives a realistic estimate of predictive performance on new, unseen materials or conditions [5,9]. Increasingly, interpretability methods such as SHAP (Shapley Additive Explanations) and partial dependence plots (PDPs) are reported alongside accuracy metrics, allowing researchers to identify which input features — for example pH, pyrolysis temperature, or specific surface area — most strongly influence a predicted adsorption outcome, rather than treating the model purely as a black box [9,12].

III. CASE STUDIES: DATA SCIENCE APPLIED TO ADSORPTION AND MATERIALS DISCOVERY

3.1 Case Study I — ML Prediction of Heavy Metal Adsorption and Removal Performance

A growing body of work has applied ensemble ML models to predict heavy metal adsorption or removal efficiency directly from adsorbent and process-condition data, bypassing the need to fit a classical isotherm model for every new adsorbent–pollutant combination. Table 1 and Figure 1 compare reported test-set R^2 values across several such studies, spanning biochar-, soil-, and MOF-based adsorption datasets.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, September 2026
Impact Factor: 8.2

Table 1. Reported predictive accuracy (test-set R^2) of machine learning models for heavy metal adsorption/removal prediction across different adsorbent systems.

Dataset / system	Best-performing model	Dataset size	Test R^2	Ref.
Biochar, multi-metal (Pb, Cd, Ni, Cu, Zn)	XGBoost	353 samples	0.921	[13]
Biochar (heavy metal removal, general)	XGBoost	—	0.99	[14]
Soil adsorption (Cr, Pb, Cd, As)	XGBoost	2112 data points, 22 soil types	0.88 (avg.)	[15]
MOF composites, 6 heavy metals	Combined GBDT (Com-GBDT)	1225 points, 89 MOFs, 44 features	0.92–0.96	[16]
Biomass-derived adsorbents	XGBoost	—	0.76–0.88	[17]
Biochar (RF vs. neural network)	Random Forest	—	0.94 (RF); 0.55 (NN)	[18]
Sulfate-reducing bacteria removal (Cd, Cu)	CatBoost	<1000 points	0.91 (Cd); 0.91 (Cu)	[19]

Figure 1. Reported Predictive Accuracy (R^2) of Machine Learning Models for Heavy Metal Adsorption/Removal Performance

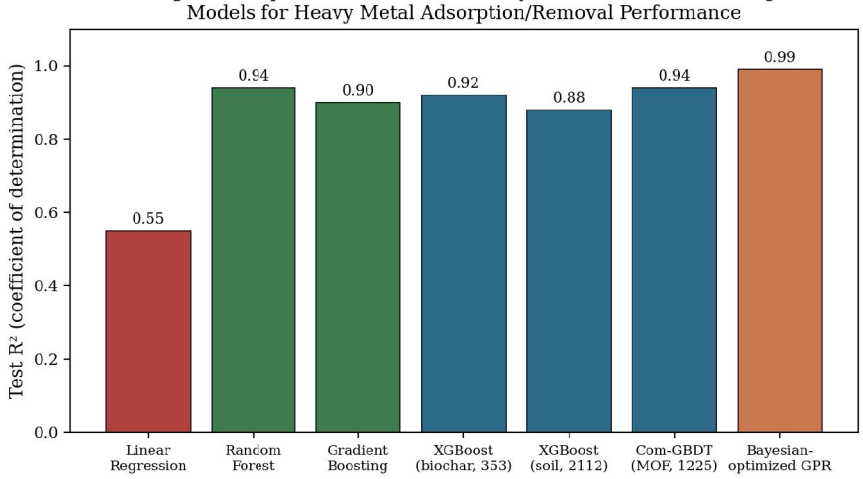


Figure 1. Reported test-set R^2 values for machine learning models predicting heavy metal adsorption/removal performance across biochar, soil, and MOF datasets (data compiled from refs. [13–19]).

Two consistent patterns emerge from Table 1 and Figure 1. First, tree-based ensemble methods (XGBoost, Random Forest, CatBoost, GBDT) consistently outperform simple linear regression and, in at least one direct comparison, outperform a basic neural network on the same dataset (R^2 = 0.94 for Random Forest versus 0.55 for a neural network), reflecting the tabular, moderate-sized nature of most adsorption datasets, for which tree ensembles are generally better suited than deep learning [18]. Second, model accuracy is not simply a function of dataset size: the MOF-based Com-GBDT model achieved R^2 = 0.92–0.96 on a relatively information-rich dataset of 1225 points with 44 engineered features [16], indicating that feature engineering and domain-informed descriptor selection can be as important as raw sample count for achieving high predictive accuracy in adsorption modeling.

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/IJARSCT-40623

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, September 2026
Impact Factor: 8.2

3.2 Case Study II — ML-Accelerated High-Throughput Screening of MOF Adsorbents

A second major application of data science in materials research is high-throughput computational screening (HTCS), in which molecular simulation methods (e.g., Grand Canonical Monte Carlo) are used to predict the adsorption performance of large numbers of candidate MOF structures, with ML models trained on a simulated subset used to extend predictions across much larger databases without simulating every candidate individually [6,20]. In one representative study, HT-GCMC simulation combined with ML screening was applied to the CoRE-MOF database of 10,143 computation-ready experimental MOF structures to identify adsorbents for CF4/N2 separation, narrowing the field to 73 top-performing candidates from the full database and 15 from a randomly sampled subset of 690 structures, based on selectivity, working capacity, and regenerability criteria [21]. Table 2 and Figure 2 summarize this screening funnel.

Table 2. ML-assisted high-throughput screening funnel for MOF adsorbents in a representative CF4/N2 gas-separation study.

Screening stage	Number of MOF structures	Selection criteria applied	Ref.
Full CoRE-MOF database screened (HTCS + ML)	10,143	Computation-ready experimental MOF structures	[21]
Randomly sampled subset	690	Random sample for validation comparison	[21]
Top candidates from full database	73	Selectivity > 60; working capacity > 70 mg/g; regenerability > 70%	[21]
Top candidates from random subset	15	Same criteria applied to the 690-structure subset	[21]

Figure 2. ML-Assisted High-Throughput Screening Funnel for MOF Adsorbents (CF4/N2 Separation Case)

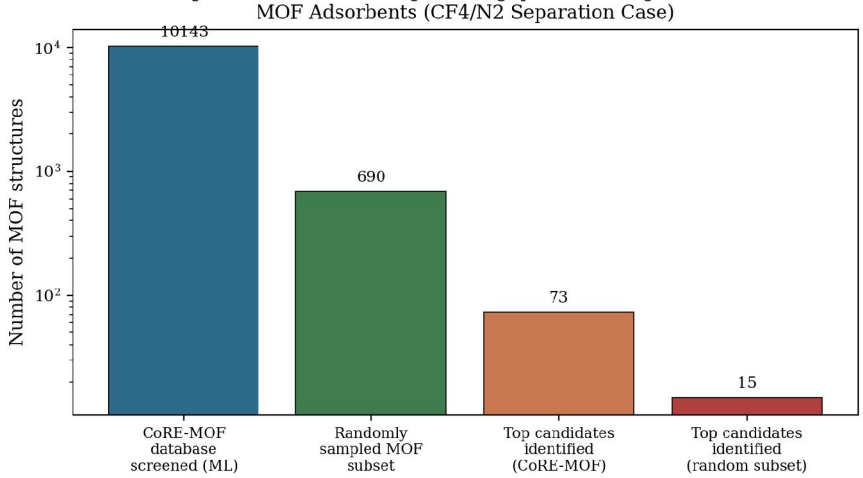


Figure 2. Screening funnel showing the reduction from a full MOF database to a small set of top-performing candidates via combined molecular simulation and machine learning (data from ref. [21]).

This case illustrates the core value proposition of ML-accelerated HTCS: screening over ten thousand candidate structures individually by full molecular simulation would be computationally prohibitive for routine use, whereas an ML model trained on a representative simulated subset can extend predictions across the full database at a small

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 5, September 2026 Impact Factor: 8.2

fraction of the computational cost [6,7]. Related work has applied the same general strategy specifically to heavy-metal-removal MOF discovery, using ML-assisted HTCS to identify optimal MOF structures for Pb(II) removal in aqueous solution and finding that carboxyl coordination environments enhanced adsorption performance for water-stable MOFs [22], directly linking materials-informatics methodology to the adsorption-based remediation chemistry discussed in Section 3.1.

IV. CHALLENGES AND FUTURE OUTLOOK

- Data scarcity and quality: many chemistry and materials datasets remain small (tens to low thousands of points) relative to the datasets typically used in other ML domains, and are compiled from heterogeneous literature sources with inconsistent reporting of experimental conditions, limiting model robustness.
- Interpretability: tree ensembles and neural networks are comparatively opaque; SHAP and partial dependence analysis are increasingly reported alongside accuracy metrics specifically to address this, but interpretability remains an active area of methodological development rather than a solved problem.
- Generalizability: a model trained on one adsorbent–pollutant system or one region of chemical/materials space may not transfer reliably to structurally dissimilar systems, and reported R^2 values should be read as specific to the training distribution rather than universal performance guarantees.
- Reproducibility and benchmarking: as in the adsorption and green-solvent literature discussed elsewhere in this research group's work, standardized benchmarking datasets and consistent reporting practices remain underdeveloped, complicating direct comparison across studies.
- Emerging directions include graph neural networks that operate directly on molecular or crystal-structure graphs rather than hand-engineered descriptors, offering potentially better generalization across structurally diverse materials [10,11]; closed-loop autonomous experimentation, in which ML-guided suggestions are directly synthesized and characterized with minimal human intervention, iteratively refining the model [2]; and continued integration of ML with high-throughput computation for energy materials, gas storage, and separation applications beyond the adsorption-focused case studies presented here [23].

V. CONCLUSION

Data science and machine learning have moved from a peripheral tool to a central methodology in chemistry and materials research, particularly for problems — such as adsorption performance prediction and materials screening — where the space of possible materials or conditions vastly exceeds what can be explored experimentally. The case studies reviewed here demonstrate that ensemble tree-based models can predict heavy metal adsorption performance with test-set R^2 values regularly exceeding 0.90 on datasets of a few hundred to a few thousand points, and that combining molecular simulation with ML can extend rigorous screening from tens of candidate materials to databases of ten thousand or more. Continued progress in data quality, interpretability, and generalizability will be essential to translating these predictive gains into reliable, broadly applicable tools for chemists and materials researchers, including in the adsorption-based environmental remediation applications central to this research group's work.

REFERENCES

[1] Opportunities and Challenges for Machine Learning in Materials Science. Annu. Rev. Mater. Res. / arXiv 2020.
[2] Machine Learning for Accelerating Energy Materials Discovery: Bridging Quantum Accuracy with Computational Efficiency. Adv. Energy Mater. 2026, DOI: 10.1002/aenm.202503356.
[3] Accelerating the discovery and optimization of metal-organic framework materials via machine learning. Comput. Chem. Eng. / ScienceDirect 2025.
[4] Applications of machine learning in heavy metal adsorption modeling: A review. Sep. Purif. Technol. 2025.
[5] Machine learning-driven prediction and mechanistic interpretation of heavy metal adsorption by biomass-derived adsorbents. Chem. Pap. 2026.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, September 2026
Impact Factor: 8.2

[6] Machine Learning Accelerated High-Throughput Computational Screening of Metal-Organic Frameworks. Prog. Chem. 2022.
[7] High-throughput and machine learning approaches for the discovery of metal organic frameworks. APL Mater. 2023, 11 (6), 060901.
[8] Deciphering heavy metal adsorption capacity of soil based on physicochemical properties using machine learning (LR baseline comparison). J. Environ. Manage. / ScienceDirect 2025.
[9] Interpretable machine learning for predicting heavy metal removal and optimizing biochar characteristics. J. Environ. Chem. Eng. 2024.
[10] Machine learning the quantum-chemical properties of metal-organic frameworks for accelerated materials discovery. Nat. Mater. / J. Chem. Phys. 2024.
[11] High-throughput computational screening of MOF adsorbents for efficient propane capture from air and natural gas mixtures. J. Chem. Phys. 2024, 160 (8), 084706.
[12] High-Throughput Computational Screening and Interpretable Machine Learning of Metal-organic Frameworks for Iodine Capture. arXiv 2025.
[13] Heavy metal adsorption efficiency prediction using biochar properties: a comparative analysis for ensemble machine learning models. Sci. Rep. 2025.
[14] Interpretable machine learning for predicting heavy metal removal and optimizing biochar characteristics (XGBoost R2=0.99 result). J. Environ. Chem. Eng. 2024.
[15] Deciphering heavy metal adsorption capacity of soil based on its physicochemical properties and adsorption reaction time using machine learning. J. Environ. Manage. 2025.
[16] Simulation, prediction and optimization of heavy metal adsorption by metal-organic frameworks with machine learning. Chem. Eng. J. 2025.
[17] Machine learning-driven prediction and mechanistic interpretation of heavy metal adsorption by biomass-derived adsorbents. Chem. Pap. 2026.
[18] Predictive Machine Learning Framework for Biochar Performance in Heavy Metals Remediation: A Decision-support Approach. Earth Syst. Environ. 2026.
[19] Applications of machine learning in heavy metal adsorption modeling: A review (sulfate-reducing bacteria CatBoost case). Sep. Purif. Technol. 2025.
[20] Accelerating the discovery and optimization of metal-organic framework materials via machine learning (regression, classification, clustering, deep learning overview). Comput. Chem. Eng. 2025.
[21] Machine learning-assisted high-throughput screening of MOFs for efficient adsorption and separation of CF4/N2. Sep. Purif. Technol. 2024.
[22] Machine learning-assisted screening of metal-organic frameworks (MOFs) for the removal of heavy metals in aqueous solution. Sep. Purif. Technol. 2024.
[23] Machine Learning for Accelerating Energy Materials Discovery: hydrogen storage and CO2 capture screening in MOFs/COFs. Adv. Energy Mater. 2026.