REVIEW ARTICLE

UPI JOURNAL OF CHEMICAL AND LIFE SCIENCES (UPI-JCLS)

ISSN: 2581-4648 (An International online Peer Reviewed, Open Access Journal)

www.uniquepubinternational.com Published by Unique Pub International (UPI)

DATA-DRIVEN DESIGN OF HYDROGEL ADSORBENTS: MACHINE-LEARNING DATASETS, DESCRIPTORS AND INVERSE DESIGN FOR WATER TREATMENT – A CRITICAL REVIEW

MAKARIM A. MAHDI*

Department of Chemistry, College of Education, University of Al-Qadisiyah, Al-Diwaniyah, Iraq.

***Corresponding Author**

Makarim A. Mahdi*

makarim.ali@qu.edu.iq

Received: 19 09 2026 Revised: 24 10 2026 Accepted: 14 11 2026

Abstract

Hydrogel adsorbents are attractive for water treatment. They can be made from inexpensive biopolymers, are water swellable and have heavy functional groups. But most new hydrogels are still discovered by trial and mistake, one formulation at a time. Machine learning (ML) has the potential to abbreviate this cycle and ML studies on adsorption have increased in number over the last 5 years. This review critically evaluates what the literature can and cannot support when the adsorbent is a hydrogel. To begin, we discuss why hydrogels are a difficult target: they are adjusted at multiple length scales simultaneously, from binding groups to mesh size and bead porosity. Then we look at the datasets that are underlying existing models. The majority are modest, single laboratory studies in which operating variables such as dose, pH, time and concentration dominate while the material itself is rarely specified. Literature-mined sets are bigger but have unit errors, linearised fits and missing metadata. Next, we compare descriptor families from synthesis recipes and network parameters to polymer line notations and adsorbate solvation descriptors. We demonstrate that random point-level splitting, still the default, inflates test scores and hides poor transfer to new materials. Finally, we consider inverse design. Bayesian optimisation has already yielded improved hydrogels from modest datasets, although generative polymer models have yet to be tried for adsorbents. We conclude with a minimum reporting set and a roadmap linking FAIR data, grouped validation and closed-loop synthesis. Our reading is that today development is limited by data quality, rather than choice of algorithm.

Keywords: *Hydrogel adsorbents; Machine learning; Materials descriptors; Data leakage; Inverse design; Bayesian optimization; Water treatment.*

Copyright:© 2026 The author(s). This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

1. INTRODUCTION

Adsorption is still one of the simplest ways to take dyes, metal ions and organic micropollutants out of water. Among the many sorbents studied, hydrogels hold a special place. They are three-dimensional, water-swollen polymer networks, and their chains can be decorated with carboxyl, amine, sulfonate or hydroxyl groups almost at will [1–3]. Many are made from chitosan, alginate, cellulose or other inexpensive biopolymers, and capacities above 500 mg g^{-1} have been reported for several organic and inorganic pollutants [4].

The trick is the size of the design space. A typical graft-copolymer hydrogel is made up of a backbone, one or two monomers, a crosslinker, an initiator, sometimes a filler, and a set of reaction conditions. Each option affects the others, and the performance of the final gel depends not only on its ingredients, but also on the functions of its network [5]. Researchers often adjust one item at a time or do a tiny response-surface design. The finest recipe found in this approach seldom transfers to a new contaminant.

Machine learning (ML) offers an alternative method. Models trained on past data can be used to estimate uptake, rate candidates and suggest future experiment. ML for adsorption has been reviewed at a consistent rate [6-8], however most of the instances reviewed have been for biochar, activated carbon and other carbon sorbents [9]. Hydrogels are mentioned, if at all, in passing. This is important, as hydrogels behave differently. They swell, their networks are pH and ionic strength responsive and their binding sites are in a three dimensional gel rather than on a solid surface [10, 11].

This review asks a single focused question. What would it take to design a hydrogel adsorbent from data, rather than describe one after it is made? We tackle the subject in four stages: what datasets are out there, what descriptors are utilised to describe a hydrogel, how are models verified, and the inverse-design methods that turn predictions into recipes. Throughout, we try to keep apart what has been shown and what has only been claimed.

2. WHY HYDROGEL ADSORBENTS ARE A DIFFICULT TARGET

Hard materials such as zeolites or metal–organic frameworks can be written down as a crystal structure. A hydrogel cannot. It is an ensemble of chains with a spread of lengths, crosslink points and defects, and two gels made from the same recipe in two laboratories may differ in gel fraction, swelling and porosity. This is one reason why data-driven methods, so successful for hard materials, have been slow to reach soft matter [12].

Performance is also set at several length scales at the same time (Fig. 1). At the molecular scale, the type and density of functional groups decide which interactions are possible; electrostatic attraction, ion exchange, hydrogen bonding, π–π stacking and chelation are all reported, often together [10]. At the network scale, crosslink density and mesh size control how far a solute can diffuse and how many sites it can actually reach. The swollen polymer network models that go back to Flory and Rehner tie these quantities to the equilibrium swelling ratio and the polymer volume fraction [11, 13]. At the bead or monolith scale, porosity and particle size set the rate of uptake. And the water itself matters. pH changes the ionisation of both gel and pollutant, salts screen charges, and dissolved organic matter competes for sites; in activated-carbon beds, for example, the dissolved organic carbon of the influent was one of the three strongest predictors of breakthrough [14].

So a hydrogel adsorbent has more hidden variables than a typical carbon sorbent, and many of them are never measured. A model that sees only the recipe has to learn the network structure implicitly. A model that sees only the operating conditions learns nothing about the material at all. Richbourg et al. showed for synthetic hydrogels that synthesis variables map onto swelling through simple, independent linear relations, but that methacrylamide-modified gelatin breaks these relations [15]. Most adsorbent hydrogels are biopolymer-based. We should therefore expect recipe-to-property maps to be noisy and system-specific, and plan datasets and descriptors with that in mind.

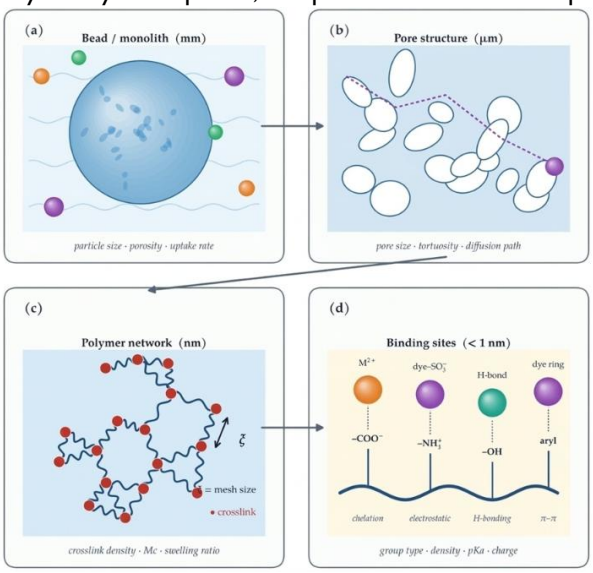


Figure 01: The hydrogel adsorbent as a multiscale design object. (a) Bead or monolith in water (mm); (b) internal pore structure and diffusion path (µm); (c) crosslinked network with mesh size ξ (nm); (d) binding groups and their typical interactions with metal ions, sulfonated dyes and aromatic pollutants (< 1 nm). Italic lines give the descriptors that belong to each scale. Illustrative drawing, not to scale.

3. DATASETS: WHERE THE NUMBERS COME FROM

3.1. Single-study experimental datasets

The bulk of the ML work on hydrogel adsorbents falls into one pattern. A group prepares one hydrogel, or a small family of them, runs batch experiments over pH, dose, time, temperature and initial concentration, and then fits a neural network or a tree ensemble to the results (Table 1). Pauletto et al. modelled single and binary dye uptake on five chitosan-based hybrid hydrogels with an artificial neural network (ANN) and reported R = 0.9987 [16]. Similar exercises have been published for sodium alginate–kaolin beads [17], alginate–cobalt ferrite magsorbents [18], cellulose nanocrystal blends for Ni(II) and Cr(VI) [19, 20], polyethylenimine sponges [21], gelatin/carboxymethyl cellulose cryogels [22], chitosan–PVA beads [23] and karaya gum/PEI gels [24]. A few studies pooled literature data for one gel family, such as PVA/CMC composites [25], or used ensembles on a polymer–clay nanocomposite [26].

Table 01: Selected machine-learning studies on hydrogel and polymer-based adsorbents. Values are those reported in the original abstracts; n.s. = not stated in the abstract.

Ref.	Adsorbent	Pollutant(s)	Main inputs	Algorithm(s)	Reported fit	Design use
[16]	Five chitosan-based hybrid hydrogels (with AC or CNT)	Acid Blue 9, Allura Red AC (single and binary)	C_0 of each dye, time, porosity, carbon content	ANN 5-10-10-10-2 (LM)	R = 0.9987; RMSE = 0.0119	Process model
[17]	Sodium alginate–kaolin beads	Methylene blue	C_0, time, pH, dose, T, agitation	Regression, genetic programming, ANN	ANN: R^2 = 0.97; RMSE = 3.59	Process model
[18]	Cross-linked alginate/$CoFe_2O_4$magsorbent	Cationic organic pollutant	Batch conditions	ANN + docking/MD	96.54% removal at model optimum	Process optimization
[19]	Cellulose nanocrystal/corn starch blend hydrogel	Ni(II)	C_0, pH, time, dose	ANN, ANFIS	Training R^2 = 0.986 (ANN), 0.995 (ANFIS)	Process model
[20]	Modified CNC–sodium alginate	Cr(VI)	Time, dose, pH, C_0 (CCD)	ANFIS, ANN, RSM	R = 0.997, 0.990, 0.989	Process model
[21]	Ice-templated PEI macroporous sponge	Congo red	Dose, time, C_0; column flow rate	SVR, ANN, MR	Batch R = 0.9955–0.9972 (SVR/ANN); column R ≥ 0.9991	Batch and column process model
[22]	Magnetic gelatin/CMC cryogel	Congo red	Batch conditions	Six models; XGBoost best; SHAP	Test R^2 = 0.9862; RMSE = 10.39 mg g^{-1}	Process model; C_0 ranked first
[23]	Chitosan–PVA beads (2.1–3.2 mm)	Tartrazine	Bead size, T, time	ANN vs PFO, PSO, Elovich	R^2 > 0.94	Kinetic surrogate
[24]	Karaya gum/PEI hydrogel	Methyl orange	pH, time, C_0, T	Several ML models; gradient boosting best	n.s.	Process model
[25]	PVA/CMC hydrogels with GO and bentonite	Methylene blue	Adsorbent type, T, C_0, pH, time (literature data)	Cascade-correlation NN	R^2 = 0.9980; AARD = 2.4%	Process model on pooled data
[26]	PEI/GO/bentonite/MgFeAl-LTH composite	Methyl orange, crystal violet	Batch data	Bagging and boosting; Extra Trees best	n.s.	Process model
[27]	SA-g-poly(AAm-co-IA), chosen among 14 DFT clusters	Methylene blue (also sulfasalazine, DBT)	DFT E_{ads}; pH, time, C_0	DFT screening + statistical/ML optimization	q_e = 3244 mg g^{-1} at optimum	Material screening + process optimization
[28]	Hydrogel sorbents (synthesis space)	Cu(II), Pb(II)	Reaction T, time, initiator and crosslinker contents	XGBoost + Bayesian optimization	log K_d(Pb) 2.76 → 3.37; error 0.025–0.172	Recipe design, tested on 10 new gels
[29]	30 polymeric resins within 149 sorbents	165 organic compounds	pp-LFER descriptors, C_e, BET area, pore volume	NN-LFER, group-selection split	RMSE = 0.23–0.31 log units	Cross-material prediction
[30]	30 polymer resins and 50 biochars (1750	73 organic compounds	Abraham descriptors, V_t, BET area, log C_e	15 algorithms; XGBoost	R^2 = 0.974 (XGBoost)	Cross-material prediction

	isotherms)					
[31]	Polymeric adsorbents	Oxyanions	RDKit descriptors; own + literature data	ML with SHAP	n.s.	Mechanism and targeted design

These studies are useful as process models. They interpolate within one material and one set of conditions, and they often fit the data better than pseudo-second-order or Langmuir equations do. But they are not design tools. Their inputs are mostly operating variables, so the model has no way to say what a different crosslinker or monomer would do. The SHAP analysis of Cui et al. is telling here: initial dye concentration dominated the prediction [22]. A cross-study synthesis of SHAP results for emerging contaminants found the same pattern, with initial concentration, surface area and pH at the top [32]. Such rankings mostly restate the isotherm.

Two exceptions point the way forward. Umar et al. screened 14 candidate monomer–alginate clusters by DFT adsorption energy before any synthesis, and only then used statistical and ML tools to tune the selected gel [27]. Zhang et al. trained XGBoost on synthesis variables, namely reaction temperature, time, initiator and crosslinker contents, to predict the distribution coefficient K_d of metal ions, and then let Bayesian optimization choose new recipes [28]. In both works the hydrogel itself, not the batch condition, is the variable being learned.

3.2. Literature-mined and multi-source datasets

Larger datasets can be assembled from published papers (Fig. 2). For carbonaceous sorbents this has worked reasonably well. Zhang et al. compiled aqueous adsorption data for 165 organic compounds on 50 biochars, 34 carbon nanotubes, 35 activated carbons and 30 polymeric resins, and combined neural networks with polyparameter linear free energy relationships [29]. Hassan and Kazemi later used 1750 isotherms that again included 30 polymer resins [30]. Resins are the closest relatives of hydrogels in these sets, which suggests that a hydrogel database is feasible.

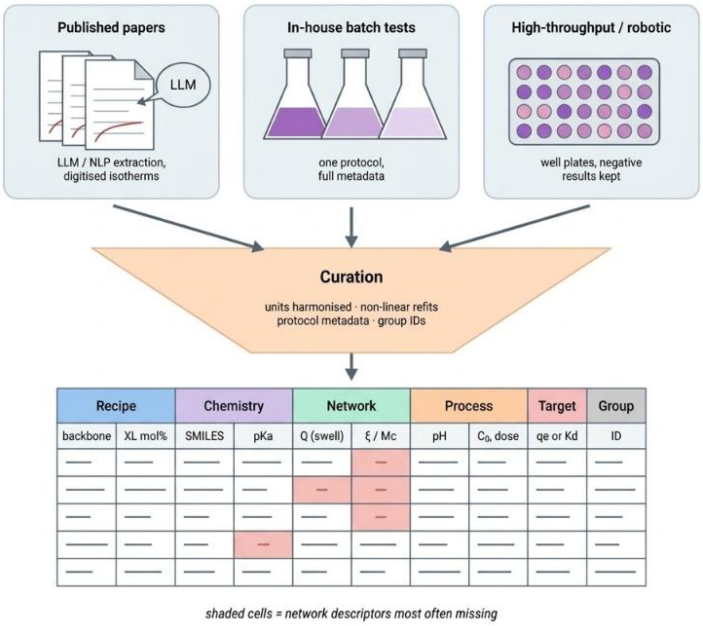


Figure 02: Building a machine-readable hydrogel adsorbent dataset. Three data sources (published papers mined with language models, in-house batch tests and high-throughput plates) pass through a curation step before entering a table whose columns are grouped by descriptor tier. Shaded cells mark the network descriptors that are most often missing from published reports. Illustrative drawing.

Manual extraction is slow, though, and language models now do much of it. Shetty et al. extracted about 300,000 property records from polymer abstracts [33], and Gupta et al. moved to full texts, obtaining more than one million records for over 106,000 polymers [34]. Conversational models with follow-up verification prompts reached precision and recall close to 90% on materials data [35], and fine-tuned models can return structured JSON records directly [36, 37]. Wang et al. compiled a literature database of synthesis parameters to predict the swelling states of PNIPAAm gels in a similar spirit [38].

The difficulty is less in extraction than in meaning. Adsorption papers are notorious for inconsistent reporting [39]. Capacities may be given as Langmuir q_{max} from linearised plots, as experimental q_e, or as removal percentages at an arbitrary dose. Linearisation alone can shift fitted parameters and favour the wrong model [40, 41], and q_{max} is a fitted parameter that should be compared between adsorbents only with caution [42]. Yunus et al. tried to build a dataset for magnetite nanoparticle adsorbents from 50 papers. Only about 37% of the 137 observations reported q_m, 23% reported BET area and 14% reported magnetization [43]. Their conclusion, that descriptor completeness rather than algorithm

choice limits prediction, applies with even more force to hydrogels, because swelling ratio, gel fraction and crosslink density are reported even less often. Yuan et al. add a second warning. A model trained on literature data alone gave biased mechanistic conclusions and failed in targeted adsorbent design, whereas mixing in consistent in-house data fixed the problem [31].

3.3. High-throughput and automated datasets

A third source is purpose-built experimentation. Seifermann et al. screened more than 900 hydrogel pads in a miniaturised format and used the data to drive Bayesian optimization of photodegradation [44]. Cheong et al. used a liquid-handling robot to prepare 120 alginate formulations as a seed library for active learning of drug release [45]. Neither study targeted adsorption. Still, batch uptake is easy to read in well plates with UV–vis or fluorescence detection, so the step is a small one. Such data have qualities that literature sets lack: one protocol, full metadata and the negative results that groups running self-driving laboratories regard as one of their main assets [46].

Table 2 lists the main data problems and the practice we recommend for each. Most items are not new. They restate long-standing advice from the adsorption community [39, 47] in a form that also suits machine reading under the FAIR principles [48].

Table 02: Common data problems in adsorption datasets, their effect on machine learning, and the practice recommended in this review.

Problem	Typical symptom	Effect on the model	Recommended practice	Refs.
Linearised isotherm or kinetic fits	q_{max}, k_2 taken from linear plots	Biased targets; wrong model favoured	Report raw (C_e, q_e) pairs; non-linear fits with error statistics	[39–41]
q_{max} used as the target	Capacities compared across unlike conditions	Target is not comparable between studies	Report q_e at stated C_e, or K_d at defined conditions, plus the full isotherm	[28, 42]
Removal (%) at an arbitrary dose	Removal depends on dose, volume and C_0	Spurious correlations; poor transfer	State dose, volume, C_0 and C_e explicitly	[39]
Missing material descriptors	No swelling ratio, gel fraction, pH_{pzc} or BET area	Model learns conditions only	Adopt the minimum descriptor set of Table 3	[43]
Only successful gels published	Failed formulations absent	Biased dataset; over-optimistic optimization	Report all tested formulations in the supplementary data	[46]
Mixed sources and protocols	Different equilibration times, agitation, units	Noise and batch effects	Record protocol metadata; harmonise units; add consistent in-house data	[31, 48]
Grouped data split at random	Points of one isotherm in both train and test sets	Inflated R^2; hidden failure on new materials	Split by adsorbent or by study	[29, 49]

4. DESCRIPTORS: HOW TO WRITE A HYDROGEL DOWN

A descriptor is any number that is fed to the model. For hydrogel adsorbents we find it useful to separate four tiers, summarised in Fig. 3 and Table 3.

4.1. Recipe descriptors

The simplest tier is the recipe: backbone type, monomer feed ratios, crosslinker and initiator contents, reaction temperature and time. These are always known to the authors and easy to report. Zhang et al. showed that such variables alone can predict K_d well enough to guide synthesis [28]. Their weakness is that categorical entries such as "chitosan" or "MBA" carry no chemistry. A model cannot generalise from acrylic acid to itaconic acid unless the monomers are encoded by structure.

4.2. Chemical descriptors of chains and pollutants

Polymer informatics offers several ways to encode structure. BigSMILES writes stochastic polymers as a line notation [50]. Fingerprint schemes such as Polymer Genome map repeat units onto hierarchical features [51, 52], and language models such as polyBERT learn fingerprints directly from polymer strings [53]. Aldeghi and Coley went a step further and encoded monomer stoichiometry and chain architecture in a graph, which suits copolymer gels well [54]. To our knowledge none of these has yet been applied to a hydrogel adsorbent. On the pollutant side, Abraham solvation descriptors and related free-energy terms work well for organic compounds [14, 29, 30], while metal ions are usually described by charge, ionic radius, electronegativity and hydrolysis constants. Quantum descriptors sit in between. DFT adsorption energies ranked candidate alginate gels before any were made [27], and a transfer-learning model reached DFT-level accuracy for site-level adsorption energies with one tenth of the data needed to train from scratch [55]. An elegant short-cut was found for sticky gels by mapping protein sequence statistics onto random copolymer compositions [12] by Liao et al.

Biopolymer backbones require special care. Chitosan is not one single substance. The density of amine groups and the solubility of the chain are modified by its degree of deacetylation and molecular weight. Alginate varies in the ratio of mannuronate to guluronate and this affects how firmly it will bind calcium. Carboxymethyl cellulose varies in its degree of substitution. These numbers are generally displayed on the supplier's certificate and seldom appear in the techniques section. In the absence of these, two entries of "chitosan" in a dataset may in fact be very distinct polymers that the model cannot distinguish between.

4.3. Network and morphology descriptors

This is the tier that distinguishes a hydrogel from a resin. Equilibrium swelling ratio, gel fraction, average molecular weight between crosslinks (M_c), mesh size, pore size and charge density govern the accessibility of sites [11, 13]. They can be measured with simple gravimetry, rheology and titration. The equilibrium swelling ratio gives the polymer volume fraction; through the Flory–Rehner relation this yields M_c, and scaling laws then give an estimate of mesh size [11, 13]. Such estimates are rough for polyelectrolyte and biopolymer gels, but a rough network descriptor still tells the model more than no descriptor at all. Yet these quantities appear in only a small fraction of adsorption papers. We suggest treating them as compulsory outputs of every new hydrogel adsorbent study. They also open a bridge between chemistry and performance: a two-stage model can predict network descriptors from the recipe, and uptake from the network, which is easier to check against theory than a direct black box (Fig. 3).

4.4. Process and matrix descriptors

Last come the batch or column conditions and the water matrix: pH relative to the pKa of gel and pollutant, ionic strength, competing ions, dissolved organic carbon, contact time and flow rate. Including them is unavoidable. They should, however, be kept apart from the material descriptors when importance is analysed; otherwise the model's insight is just the isotherm again. Shen et al. showed for biochar that engineered material features, such as molar elemental ratios, can carry more physical meaning than raw inputs [56]. No comparable effort has yet been made for hydrogels.

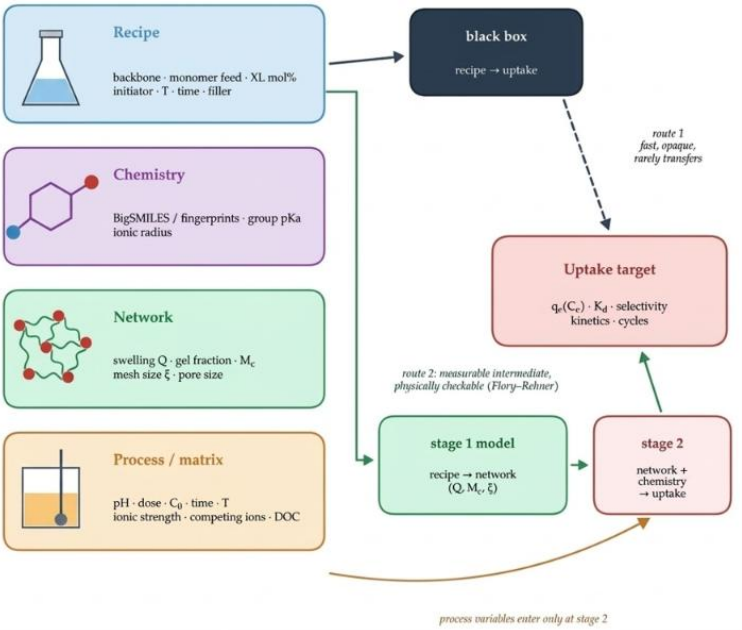


Figure 03: Four descriptor tiers for hydrogel adsorbents and two routes from recipe to uptake. Route 1 maps the recipe directly onto uptake; route 2 passes through measurable network descriptors (swelling ratio Q, M_c, mesh size ξ) that can be checked against swollen-network theory. Process variables enter only at the second stage. Conceptual scheme.

Table 03: Descriptor tiers for hydrogel adsorbents. Typical availability is the authors' assessment of how often each tier is reported in adsorption papers.

Tier	Example descriptors	How obtained	Typical availability	Main caveat	Refs.
Recipe	Backbone, monomer feed ratio, crosslinker and initiator mol%, T, time, filler wt%	Laboratory record	High	Categorical entries carry no chemistry unless encoded by structure	[28]
Chemistry (polymer)	BigSMILES, hierarchical fingerprints, polyBERT embeddings, ensemble graphs, group density	Computed from structure	Rare in adsorption work	Copolymer stoichiometry and architecture must be encoded	[50, 51, 53, 54]
Chemistry	Abraham descriptors, charge,	Computed or	Moderate	Well developed for	[14, 29,

(pollutant)	ionic radius, pKa, molecular volume, RDKit descriptors	tabulated		organics, thinner for metal ions	31]
Quantum	DFT adsorption energy, frontier orbitals, charge distribution	Computed	Rare	Costly; transfer learning lowers the cost	[27, 55]
Network and morphology	Swelling ratio, gel fraction, M_c, mesh size, pore size, porosity, pH_{pzc}, charge density	Gravimetry, rheology, titration, SEM, BET	Rare	The hydrogel-specific tier; seldom measured	[11, 13]
Process and matrix	pH, dose, C_0, time, T, ionic strength, competing ions, DOC, flow rate	Experiment	Very high	Dominates importance rankings; keep separate from material tiers	[14, 32]

5. MODEL BUILDING, VALIDATION AND INTERPRETATION

The algorithm option has garnered the greatest attention and perhaps deserves the least. Hatami et al. reported that the performance of 87 heavy-metal investigations was dependent on dataset size, data quality, and feature selection, not the method [9]. Tree ensembles like as random forest and XGBoost are being adopted and easier to interpret on tabular data [32], but ANNs are still the most used option. The advantage of Gaussian process regression for small data sets is that it provides an uncertainty on each prediction, which is what an optimiser needs [57, 58].

Validation is the Achilles heel. Data of adsorption are grouped: dozens of points belonging to one isotherm, many isotherms belonging to one adsorbent, several adsorbents presented in one paper. If the points are split at random, near-copies of each test point sit in the training set (Fig. 4). The model then looks excellent and still fails on a new material. This is a form of data leakage, a problem that has distorted results in many fields that adopted ML [59]. In biochar research, Jang et al. found that 98% of the 64 most-cited ML papers used random point-level splits, and their re-analysis showed accuracy inflated by up to 86% [49]. We see no reason to think hydrogel studies fare better. None of the abstracts behind Table 1 mentions a split that holds out whole materials; the Bayesian-optimization study of Zhang et al. is the one partial exception, because its predictions were checked on newly synthesised gels [28]. The remedy is well known. Zhang et al. used a group-selection split for adsorbent–adsorbate pairs as early as 2020 [29].

The choice of target matters as much as the split. Removal percentage depends on dose and volume as much as on the material, and q_{max} inherits every weakness of the fitted isotherm. Two better options exist. One is to predict the equilibrium loading at a stated equilibrium concentration, so that the model learns the whole isotherm; the NN-LFER models of Zhang et al. worked in this way and could be applied at any concentration, whatever isotherm equation the original authors had used [29]. The other is a distribution coefficient on a log scale, as used for metal ions by Zhang et al. and for organic compounds by Hassan and Kazemi [28, 30]. Both targets have a clear physical meaning and can be compared across studies.

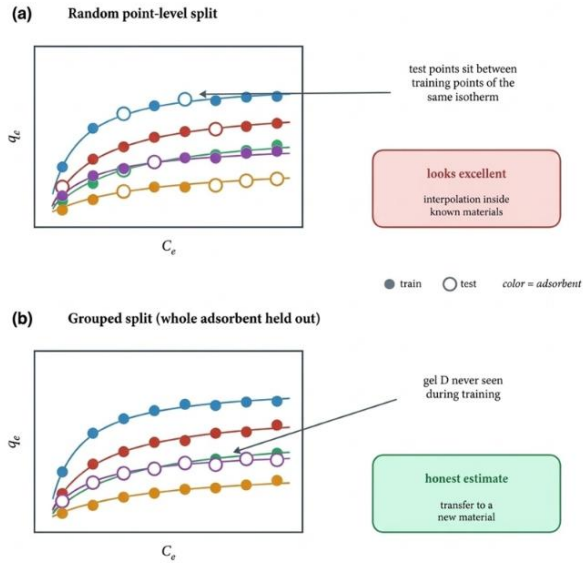


Figure 04: Data leakage in adsorption datasets. (a) A random point-level split leaves test points (open circles) between training points (filled circles) of the same isotherm, so the model only interpolates within known materials. (b) A grouped split holds out a whole adsorbent (gel D) and gives an honest estimate of transfer to a new material. Schematic Langmuir-type curves; colours denote adsorbents.

Three other practices are to be mentioned. First, determine the scope of application. Distance-based metrics in feature space can identify candidates that are outside the training data, where both predictions and their uncertainty are

incorrect [60, 61]. This is important for inverse design, as optimisers tend to wander out of domain. Second, deliberately test out-of-distribution performance, e.g. by holding out a full polymer family. Many of the apparent extrapolation tests in materials science were actually interpolation [62]. Third, report uncertainty and compare the model with simple baselines, as recommended in the general recommendations for ML in chemistry and materials science [63, 64].

Interpretation requires caution as well. SHAP values are now standard [65]. However, the adsorption inputs are highly associated (dose vs. % removal, pH vs. surface charge) and Shapley estimations can be misleading if the features are dependent [66]. SHAP discusses the model , not the chemistry . Mechanistic arguments should be validated by independent evidence, as Yuan et al. achieved with theoretical calculations and targeted synthesis [31]. Hybrid models that integrate an isotherm or a surface-complexation scheme inside the learner are another way to provide physically compatible predictions . [67, 68].

6. FROM PREDICTION TO INVERSE DESIGN

Forward models answer the query "what will this gel do ? " Inverse design asks "what gel will do this?" [69]. For polymers, approaches fall into three families: screening of enumerated candidates, global optimisation over a design space, and generative models [70]. Table 4 compares them for hydrogel adsorbents.

6.1. Optimization over formulation space

For a hydrogel adsorbent the design space is usually a recipe with five to fifteen continuous variables. Bayesian optimization (BO) suits this well. It fits a probabilistic surrogate, often a Gaussian process, and picks the next experiment by trading predicted performance against uncertainty [57] (Fig. 5). Zhang et al. used this approach to raise log K_d for Pb from 2.76 to 3.37 and for Cu from 2.70 to 3.06; the gap between predicted and measured log K_d for Pb was only 0.025–0.172 [28]. Outside adsorption the record is longer. Kumar et al. tuned the cloud point of poly(2-oxazoline)s by gradient boosting plus particle-swarm inverse design and synthesised 17 polymers at four target temperatures [71]. Liao et al. started from 180 bioinspired hydrogels and used ML to reach underwater adhesive strengths above 1 MPa [12]. Genetic algorithms coupled with property predictors have also proposed polymers with extreme property combinations [72].

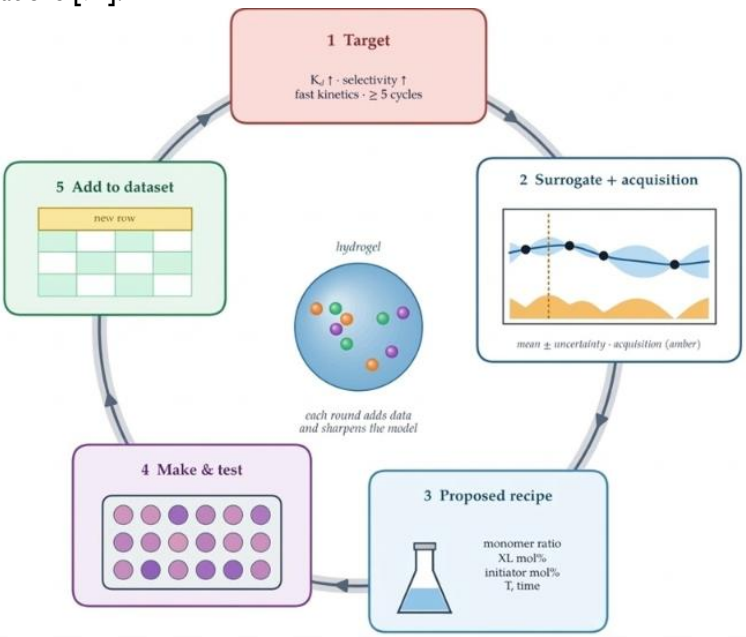


Figure 05: Closed-loop inverse design of a hydrogel adsorbent. A multi-property target (1) is passed to a probabilistic surrogate whose acquisition function (amber) balances predicted performance against uncertainty (2); the proposed recipe (3) is made and tested in parallel (4), and the new results are added to the dataset (5) before the next round. The surrogate panel is schematic.

A real adsorbent has to meet several targets at once: capacity, selectivity against competing ions, fast kinetics, mechanical stability and regenerability. Multi-objective BO and Pareto-based active learning handle this naturally [73]. We think this is where data-driven design can most clearly beat intuition, since human designers tend to optimise one property and check the others afterwards. Feasibility is a hidden objective as well. Many recipes simply fail to gel, or give gels that crumble on swelling. A practical loop therefore needs a gel/no-gel classifier next to the uptake regressor, so that the optimizer does not waste rounds in impossible regions. Xu et al. met the same problem with peptides and handled it by iterating between gelation experiments and a predictive score [74].

6.2. Generative models

Generative models, such as variational autoencoders, adversarial networks and reinforcement-learning agents, can propose new repeat units or copolymer compositions. Benchmarks on polymer datasets show that several

architectures produce valid and novel structures [75], and recent models handle stoichiometry and chain architecture [76]. The Open Macromolecular Genome restricts generation to polymers that known reactions and commercial reagents can make, which is a sensible constraint [77]. For hydrogels, Hu et al. used a conditional Wasserstein GAN to generate ionically conductive formulations from 233 measured samples [78].

Our assessment is cautious. Generative design needs a reliable forward model, and forward models for hydrogel uptake do not yet transfer across materials. Generating thousands of new monomers is of little use if the property predictor has seen five. For now, generation over formulations, that is, mixtures of known monomers and crosslinkers, is more realistic than generation of new chemistry.

6.3. Closed-loop experimentation

The natural end point is a loop in which an algorithm proposes recipes, a robot makes and tests them, and the results update the model. Self-driving laboratories have done this for polymer synthesis [79] and for many other systems [80, 81]. Peptide hydrogels have been designed with a human-in-the-loop version of the same idea [74]. A loop for hydrogel adsorbents is technically simple: redox- or photo-initiated gelation in well plates, a wash step, and an optical uptake assay. The obstacles are practical ones, mainly the handling of viscous precursors and the slow approach to swelling equilibrium. Low-cost platforms of this kind have been reported for thermoresponsive polymers and could be adapted [82].

Table 04: Inverse-design strategies and their fit for hydrogel adsorbents.

Strategy	Principle	Data need	Examples	Fit for hydrogel adsorbents
Virtual screening	Rank an enumerated candidate list with a forward model or quantum calculation	Model valid over all candidates	DFT screening of 14 alginate clusters [27]	Good for small, well-defined libraries
Bayesian optimization	Probabilistic surrogate plus acquisition function selects each next experiment	Tens to hundreds of points	Hydrogel K_d [28]; photodegradable gels [44]; release profiles [45]	Best current choice for recipe spaces
Evolutionary and swarm search	Genetic algorithm or particle swarm over a trained predictor	Trained predictor	Cloud-point tuning [71]; polymers with extreme properties [72]	Works; no uncertainty unless added
Multi-objective active learning	Search for the Pareto front of competing targets	As for Bayesian optimization	Dispersant polymers [73]	Needed for capacity–selectivity–regeneration trade-offs
Generative models	VAE, GAN, RL or language models propose structures or formulations	Large datasets and a reliable predictor	Polymer benchmarks [75]; copolymers [76]; OMG [77]; conductive gels [78]	Promising; untested for adsorbents
Closed-loop self-driving lab	Robot executes the optimizer's choices and feeds results back	Generates its own data	Autonomous polymerisation [79]; reviews [80, 81]	Feasible with well-plate gelation and optical assays

7. CRITICAL OUTLOOK AND ROADMAP

Read as a whole, this literature supports five conclusions.

First, the field is limited by data, not by algorithms. Most hydrogel models rest on a few hundred points or fewer, taken from one material. They are good interpolators and poor designers. A shared, curated database of hydrogel adsorbents, with recipe, network and process descriptors kept in separate fields, would do more than any new architecture. The surface-complexation community has shown that such a FAIR database can be built from literature data [68]. Small-data tools such as transfer learning and data augmentation can help [58, 83], but they cannot create information that was never measured.

Second, reporting has to change. If every new hydrogel paper reported swelling ratio, gel fraction, an estimate of crosslink density, the point of zero charge, the full isotherm in tabular form and non-linear fit statistics, the literature would become machine-usable within a few years. Table 2 gives our minimum set. Journals could ask for it as supplementary data at little cost.

Third, validation must be grouped and honest. Every model should state its split strategy, its applicability domain and its performance on held-out materials, not only on held-out points [49, 59].

Fourth, the target must be realistic. Single-solute batch capacity in deionised water is a poor proxy for performance in effluents. Models should be trained, at least in part, on data from real or synthetic wastewater with competing ions and organic matter, and should include regeneration cycles. Column data, which ML has already handled for polymer sponges and activated carbon [14, 21], are the obvious next step.

Fifth, cost and sustainability belong in the objective. A hydrogel that doubles capacity but needs an expensive monomer, or loses half its capacity after three cycles, is not progress for a treatment plant. Reagent cost, regeneration efficiency and, where possible, simple life-cycle indicators can be added as extra targets or constraints at almost no computational cost.

We would add one cultural point. ML is now often appended to adsorption papers as a final section that reports a high R^2. This adds length but rarely knowledge. A model earns its place when it predicts something that is then made and tested, as in the Bayesian-optimization studies discussed above.

8. CONCLUSION

Machine learning can change how hydrogel adsorbents are designed, but only if the data allow it. At present most models describe one material under varying conditions, and their high R^2 values reflect interpolation and, often, leakage between correlated points. The few studies that put the hydrogel recipe itself into the model, and then made the suggested gels, show what is possible: better K_d values reached in a handful of rounds. Three changes are needed to scale this up. Hydrogel papers should report network descriptors and raw isotherm data in machine-readable form. Models should be validated with grouped splits and should declare their applicability domain. And inverse design should be multi-objective, with capacity weighed against selectivity, regeneration and cost in real water matrices. Generative chemistry for new monomers is an exciting prospect, yet it will stay speculative until forward models generalise across polymer families. A community database and simple closed-loop synthesis platforms are the most practical next steps.

9. ACKNOWLEDGEMENT

Not Declared.

10. CONFLICT OF INTEREST STATEMENT

Not Applicable.

11. INFORM CONSENT AND ETHICAL STATMNET

Not Applicable

12. FUNDING

Nil

13. REFERENCES

1. M. Khan, I.M.C. Lo, A holistic review of hydrogel applications in the adsorptive removal of aqueous pollutants: Recent progress, challenges, and perspectives, Water Research 106 (2016) 259–271. https://doi.org/10.1016/j.watres.2016.10.008
2. V.V. Tran, D. Park, Y.-C. Lee, Hydrogel applications for adsorption of contaminants in water and wastewater treatment, Environmental Science and Pollution Research 25 (2018) 24569–24599. https://doi.org/10.1007/s11356-018-2605-y
3. Z. Darban, S. Shahabuddin, R. Gaur, I. Ahmad, N. Sridewi, Hydrogel-Based Adsorbent Material for the Effective Removal of Heavy Metals from Wastewater: A Comprehensive Review, Gels 8 (2022) 263. https://doi.org/10.3390/gels8050263
4. Md. Ahmaruzzaman, P. Roy, A. Bonilla-Petriciolet, M. Badawi, S.V. Ganachari, N.P. Shetti, T.M. Aminabhavi, Polymeric hydrogels-based materials for wastewater treatment, Chemosphere 331 (2023) 138743. https://doi.org/10.1016/j.chemosphere.2023.138743
5. Y. Seida, H. Tokuyama, Hydrogel Adsorbents for the Removal of Hazardous Pollutants—Requirements and Available Functions as Adsorbent, Gels 8 (2022) 220. https://doi.org/10.3390/gels8040220
6. W. Zhang, W. Huang, J. Tan, D. Huang, J. Ma, B. Wu, Modeling, optimization and understanding of adsorption process for pollutant removal via machine learning: Recent progress and future perspectives, Chemosphere 311 (2023) 137044. https://doi.org/10.1016/j.chemosphere.2022.137044
7. H.E. Reynel-Ávila, I.A. Aguayo-Villarreal, L.L. Diaz-Muñoz, J. Moreno-Pérez, F.J. Sánchez-Ruiz, C.K. Rojas-Mayorga, D.I. Mendoza-Castillo, A. Bonilla-Petriciolet, A Review of the Modeling of Adsorption of Organic and Inorganic Pollutants from Water Using Artificial Neural Networks, Adsorption Science & Technology 2022 (2022) 9384871. https://doi.org/10.1155/2022/9384871
8. Z. Wang, Q. Wang, F. Yang, C. Wang, M. Yang, J. Yu, How machine learning boosts the understanding of organic pollutant adsorption on carbonaceous materials: A comprehensive review with statistical insights, Separation and Purification Technology 350 (2024) 127790. https://doi.org/10.1016/j.seppur.2024.127790
9. R. Hatami, A. Dalvand, O. YousefianZadeh, Machine Learning Approaches for Heavy Metal Adsorption in Water Treatment Systems: A Systematic Review, Journal of Environmental Health and Sustainable Development 11 (2026). https://doi.org/10.18502/jehsd.v11i3.22834

10. H. Zhu, S. Chen, Y. Luo, Adsorption mechanisms of hydrogels for heavy metal and organic dyes removal: A short review, Journal of Agriculture and Food Research 12 (2023) 100552. https://doi.org/10.1016/j.jafr.2023.100552
11. N.R. Richbourg, N.A. Peppas, The swollen polymer network hypothesis: Quantitative models of hydrogel swelling, stiffness, and solute transport, Progress in Polymer Science 105 (2020) 101243. https://doi.org/10.1016/j.progpolymsci.2020.101243
12. H. Liao, S. Hu, H. Yang, L. Wang, S. Tanaka, I. Takigawa, W. Li, H. Fan, J.P. Gong, Data-driven de novo design of super-adhesive hydrogels, Nature 644 (2025) 89–95. https://doi.org/10.1038/s41586-025-09269-4
13. [T. Canal, N.A. Peppas, Correlation between mesh size and equilibrium degree of swelling of polymeric networks, Journal of Biomedical Materials Research 23 (1989) 1183–1193. https://doi.org/10.1002/jbm.820231007
14. Y. Koyama, M.A.K. Fasaee, E.Z. Berglund, D.R.U. Knappe, Machine Learning Models to Predict Early Breakthrough of Recalcitrant Organic Micropollutants in Granular Activated Carbon Adsorbers, Environmental Science & Technology 58 (2024) 17114–17124. https://doi.org/10.1021/acs.est.4c01316
15. N.R. Richbourg, M. Wancura, A.E. Gilchrist, S. Toubbeh, B.A.C. Harley, E. Cosgriff-Hernandez, N.A. Peppas, Precise control of synthetic hydrogel network structure via linear, independent synthesis-swelling relationships, Science Advances 7 (2021) eabe3245. https://doi.org/10.1126/sciadv.abe3245
16. P.S. Pauletto, J.O. Gonçalves, L.A.A. Pinto, G.L. Dotto, N.P.G. Salau, Single and competitive dye adsorption onto chitosan–based hybrid hydrogels using artificial neural network modeling, Journal of Colloid and Interface Science 560 (2020) 722–729. https://doi.org/10.1016/j.jcis.2019.10.106
17. N. Marzban, A. Moheb, S. Filonenko, S.H. Hosseini, M.J. Nouri, J.A. Libra, G. Farru, Intelligent modeling and experimental study on methylene blue adsorption by sodium alginate-kaolin beads, International Journal of Biological Macromolecules 186 (2021) 79–91. https://doi.org/10.1016/j.ijbiomac.2021.07.006
18. C. Cojocaru, A.C. Humelnicu, P. Pascariu, P. Samoila, Artificial neural network and molecular modeling for assessing the adsorption performance of a hybrid alginate-based magsorbent, Journal of Molecular Liquids 337 (2021) 116406. https://doi.org/10.1016/j.molliq.2021.116406
19. M. Banza, H. Rutto, Modelling of adsorption of nickel (II) by blend hydrogels (cellulose nanocrystals and corn starch) from aqueous solution using adaptive neuro-fuzzy inference systems (ANFIS) and artificial neural networks (ANN), The Canadian Journal of Chemical Engineering 101 (2023) 1906–1918. https://doi.org/10.1002/cjce.24603
20. M. Banza, T. Seodigeng, H. Rutto, Comparison Study of ANFIS, ANN, and RSM and Mechanistic Modeling for Chromium(VI) Removal Using Modified Cellulose Nanocrystals–Sodium Alginate (CNC–Alg), Arabian Journal for Science and Engineering 48 (2023) 16067–16085. https://doi.org/10.1007/s13369-023-07968-6
21. R.A. Aftab, S. Zaidi, A.A.P. Khan, M.A. Usman, A.Y. Khan, M. Danish, K.B. Ansari, M. Danish, A.M. Asiri, Machine Learning Models for Efficient Adsorption of Congo Red Dye on High-Performance Polyethyleneimine Macroporous Sponge, Arabian Journal for Science and Engineering 49 (2024) 7945–7960. https://doi.org/10.1007/s13369-023-08604-z
22. C. Cui, W. Qiao, D. Li, L.-J. Wang, Dual cross-linked magnetic gelatin/carboxymethyl cellulose cryogels for enhanced Congo red adsorption: Experimental studies and machine learning modelling, Journal of Colloid and Interface Science 678 (2025) 619–635. https://doi.org/10.1016/j.jcis.2024.09.136
23. S. Domínguez Beltrán, G. Miranda Piña, E.E. Granda Gutiérrez, R. Alejo Eleuterio, J.L. García Rivas, A. Reyes García, ANN-Based Prediction of Tartrazine Adsorption on Chitosan–Polyvinyl Alcohol Hydrogel Beads: A Comparison with Kinetic Models, Modelling 6 (2025) 149. https://doi.org/10.3390/modelling604014
24. Z. Hamden, M.R. Abdulbaqi, T.A.G. Al Hujran, Y.A. Bin Jardan, A Sustainable Karaya Gum/Polyethylenimine Crosslinked Hydrogel for Anionic Dye Removal From Water: Adsorption Behavior and Machine Learning Prediction, Journal of Applied Polymer Science 143 (2026) e71138. https://doi.org/10.1002/app.71138
25. A.H. Alibak, M. Khodarahmi, P. Fayyazsanavi, S.M. Alizadeh, A.J. Hadi, E. Aminzadehsarikhanbeglou, Simulation the adsorption capacity of polyvinyl alcohol/carboxymethyl cellulose based hydrogels towards methylene blue in aqueous solutions using cascade correlation neural network (CCNN) technique, Journal of Cleaner Production 337 (2022) 130509. https://doi.org/10.1016/j.jclepro.2022.130509
26. G.M. BinMakhashen, S.A. Bahadi, H.A. Al-Jamimi, S.A. Onaizi, Ensemble meta machine learning for predicting the adsorption of anionic and cationic dyes from aqueous solutions using Polymer/graphene/clay/MgFeAl-LTH nanocomposite, Chemosphere 349 (2024) 140861. https://doi.org/10.1016/j.chemosphere.2023.140861
27. M. Umar, H. Khan, S. Hussain, M. Arshad, H. Choi, E.C. Lima, Integrating DFT and machine learning for the design and optimization of sodium alginate-based hydrogel adsorbents: Efficient removal of pollutants from wastewater, Environmental Research 247 (2024) 118219. https://doi.org/10.1016/j.envres.2024.118219
28. J. Zhang, K. Fu, D. Wang, S. Zhou, J. Luo, Refining hydrogel-based sorbent design for efficient toxic metal removal using machine learning-Bayesian optimization, Journal of Hazardous Materials 479 (2024) 135688. https://doi.org/10.1016/j.jhazmat.2024.135688
29. K. Zhang, S. Zhong, H. Zhang, Predicting Aqueous Adsorption of Organic Compounds onto Biochars, Carbon Nanotubes, Granular Activated Carbons, and Resins with Machine Learning, Environmental Science & Technology 54 (2020) 7008–7018. https://doi.org/10.1021/acs.est.0c02526

30. R. Hassan, M.R. Kazemi, Machine learning frameworks to accurately estimate the adsorption of organic materials onto resin and biochar, Scientific Reports 15 (2025) 15157. https://doi.org/10.1038/s41598-025-99759-2
31. L. Yuan, H. Zhang, H. Yu, R. Xu, W. Zhang, Y. Zhang, M. Hua, L. Lv, B. Pan, Machine Learning Reveals Key Adsorption Mechanisms for Oxyanions Based on Combination of Experimental and Published Literature Data, Environmental Science & Technology 59 (2025) 11401–11413. https://doi.org/10.1021/acs.est.5c03992
32. T. Taher, S.A. Muhtar, D.A. Hapidin, A. Rianjanu, K. Khairurrijal, A Decade of Machine Learning for Adsorption of Emerging Contaminants: A Systematic Review of Algorithms, Performance, and Methodological Transparency, Green and Sustainable Environments 1 (2026) 06. https://doi.org/10.62755/gsenv.2026.06
33. P. Shetty, A. Chitteth Rajan, C. Kuenneth, S. Gupta, L.P. Panchumarti, L. Holm, C. Zhang, R. Ramprasad, A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing, npj Computational Materials 9 (2023) 52. https://doi.org/10.1038/s41524-023-01003-w
34. S. Gupta, A. Mahmood, P. Shetty, A. Adeboye, R. Ramprasad, Data extraction from polymer literature using large language models, Communications Materials 5 (2024) 269. https://doi.org/10.1038/s43246-024-00708-9
35. M.P. Polak, D. Morgan, Extracting accurate materials data from research papers with conversational language models and prompt engineering, Nature Communications 15 (2024) 1569. https://doi.org/10.1038/s41467-024-45914-8
36. J. Dagdelen, A. Dunn, S. Lee, N. Walker, A.S. Rosen, G. Ceder, K.A. Persson, A. Jain, Structured information extraction from scientific text with large language models, Nature Communications 15 (2024) 1418. https://doi.org/10.1038/s41467-024-45563-x
37. M. Schilling-Wilhelmi, M. Ríos-García, S. Shabih, M.V. Gil, S. Miret, C.T. Koch, J.A. Márquez, K.M. Jablonka, From text to insight: large language models for chemical data extraction, Chemical Society Reviews 54 (2025) 1125–1150. https://doi.org/10.1039/d4cs00913d
38. Y. Wang, T. Wallmersperger, A. Ehrenhofer, Prediction of hydrogel swelling states using machine learning methods, Engineering Reports 6 (2024) e12893. https://doi.org/10.1002/eng2.12893
39. H.N. Tran, S.-J. You, A. Hosseini-Bandegharaei, H.-P. Chao, Mistakes and inconsistencies regarding adsorption of contaminants from aqueous solutions: A critical review, Water Research 120 (2017) 88–116. https://doi.org/10.1016/j.watres.2017.04.014
40. M.E. González-López, C.M. Laureano-Anzaldo, A.A. Pérez-Fonseca, M. Arellano, J.R. Robledo-Ortíz, A Critical Overview of Adsorption Models Linearization: Methodological and Statistical Inconsistencies, Separation & Purification Reviews 51 (2021) 358–372. https://doi.org/10.1080/15422119.2021.1951757
41. J.-P. Simonin, On the comparison of pseudo-first order and pseudo-second order rate laws in the modeling of adsorption kinetics, Chemical Engineering Journal 300 (2016) 254–263. https://doi.org/10.1016/j.cej.2016.04.079
42. Q. Hu, R. Lan, L. He, H. Liu, X. Pei, A critical review of adsorption isotherm models for aqueous contaminants: Curve characteristics, site energy distribution and common controversies, Journal of Environmental Management 329 (2023) 117104. https://doi.org/10.1016/j.jenvman.2022.117104
43. A.I. Yunus, J. Song, S.A. Darko, Y. Chen, J.F. Bozeman III, Machine Learning for Magnetite Nanoparticles Contaminant Adsorption under Uncertainty: Revealing Data Quality Limits, ACS ES&T Water 6 (2026) 6064–6076. https://doi.org/10.1021/acsestwater.6c00772
44. M. Seifermann, P. Reiser, P. Friederich, P.A. Levkin, High-Throughput Synthesis and Machine Learning Assisted Design of Photodegradable Hydrogels, Small Methods 7 (2023) 2300553. https://doi.org/10.1002/smtd.202300553
45. [45] E. Cheong, D.C. Radford, A.J. Gormley, Automated active learning to optimize hydrogel drug release profiles, Journal of Controlled Release 391 (2026) 114602. https://doi.org/10.1016/j.jconrel.2026.114602
46. M. Seifrid, R. Pollice, A. Aguilar-Granda, Z. Morgan Chan, K. Hotta, C.T. Ser, J. Vestfrid, T.C. Wu, A. Aspuru-Guzik, Autonomous Chemical Experiments: Challenges and Perspectives on Establishing a Self-Driving Lab, Accounts of Chemical Research 55 (2022) 2454–2466. https://doi.org/10.1021/acs.accounts.2c00220
47. Y. Wang, C. Wang, X. Huang, Q. Zhang, T. Wang, X. Guo, Guideline for modeling solid-liquid adsorption: Kinetics, isotherm, fixed bed, and thermodynamics, Chemosphere 349 (2024) 140736. https://doi.org/10.1016/j.chemosphere.2023.140736
48. M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. Bonino da Silva Santos, P.E. Bourne, et al., The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data 3 (2016) 160018. https://doi.org/10.1038/sdata.2016.18
49. H.J. Jang, W. Ng, S. Chen, R. Setia, B. Minasny, Illusions of Accuracy: Widespread Data Leakage Distorts Machine Learning Results in Biochar Research, GCB Bioenergy 18 (2026) e70173. https://doi.org/10.1111/gcbb.70173
50. [50] T.-S. Lin, C.W. Coley, H. Mochigase, H.K. Beech, W. Wang, Z. Wang, E. Woods, S.L. Craig, J.A. Johnson, J.A. Kalow, K.F. Jensen, B.D. Olsen, BigSMILES: A Structurally-Based Line Notation for Describing Macromolecules, ACS Central Science 5 (2019) 1523–1531. https://doi.org/10.1021/acscentsci.9b00476
51. C. Kim, A. Chandrasekaran, T.D. Huan, D. Das, R. Ramprasad, Polymer Genome: A Data-Powered Polymer Informatics Platform for Property Predictions, The Journal of Physical Chemistry C 122 (2018) 17575–17585. https://doi.org/10.1021/acs.jpcc.8b02913

52. L. Chen, G. Pilania, R. Batra, T.D. Huan, C. Kim, C. Kuenneth, R. Ramprasad, Polymer informatics: Current status and critical next steps, Materials Science and Engineering: R: Reports 144 (2021) 100595. https://doi.org/10.1016/j.mser.2020.100595
53. C. Kuenneth, R. Ramprasad, polyBERT: a chemical language model to enable fully machine-driven ultrafast polymer informatics, Nature Communications 14 (2023) 4099. https://doi.org/10.1038/s41467-023-39868-6
54. M. Aldeghi, C.W. Coley, A graph representation of molecular ensembles for polymer property prediction, Chemical Science 13 (2022) 10486–10498. https://doi.org/10.1039/d2sc02839e
55. Z. Wang, H. Zhang, J. Ren, X. Lin, T. Han, J. Liu, J. Li, Predicting adsorption ability of adsorbents at arbitrary sites for pollutants using deep transfer learning, npj Computational Materials 7 (2021) 19. https://doi.org/10.1038/s41524-021-00494-9
56. T. Shen, H. Peng, X. Yuan, Y. Liang, S. Liu, Z. Wu, L. Leng, P. Qin, Feature engineering for improved machine-learning-aided studying heavy metal adsorption on biochar, Journal of Hazardous Materials 466 (2024) 133442. https://doi.org/10.1016/j.jhazmat.2024.133442
57. B. Shahriari, K. Swersky, Z. Wang, R.P. Adams, N. de Freitas, Taking the Human Out of the Loop: A Review of Bayesian Optimization, Proceedings of the IEEE 104 (2016) 148–175. https://doi.org/10.1109/jproc.2015.2494218
58. P. Xu, X. Ji, M. Li, W. Lu, Small data machine learning in materials science, npj Computational Materials 9 (2023) 42. https://doi.org/10.1038/s41524-023-01000-z
59. S. Kapoor, A. Narayanan, Leakage and the reproducibility crisis in machine-learning-based science, Patterns 4 (2023) 100804. https://doi.org/10.1016/j.patter.2023.100804
60. C. Sutton, M. Boley, L.M. Ghiringhelli, M. Rupp, J. Vreeken, M. Scheffler, Identifying domains of applicability of machine learning models for materials science, Nature Communications 11 (2020) 4428. https://doi.org/10.1038/s41467-020-17112-9
61. L.E. Schultz, Y. Wang, R. Jacobs, D. Morgan, A general approach for determining applicability domain of machine learning models, npj Computational Materials 11 (2025) 95. https://doi.org/10.1038/s41524-025-01573-x
62. K. Li, A. Niyongabo Rubungo, X. Lei, D. Persaud, K. Choudhary, B. DeCost, A.B. Dieng, J. Hattrick-Simpers, Probing out-of-distribution generalization in machine learning for materials, Communications Materials 6 (2025) 9. https://doi.org/10.1038/s43246-024-00731-w
63. N. Artrith, K.T. Butler, F.-X. Coudert, S. Han, O. Isayev, A. Jain, A. Walsh, Best practices in machine learning for chemistry, Nature Chemistry 13 (2021) 505–508. https://doi.org/10.1038/s41557-021-00716-z
64. A.Y.-T. Wang, R.J. Murdock, S.K. Kauwe, A.O. Oliynyk, A. Gurlo, J. Brgoch, K.A. Persson, T.D. Sparks, Machine Learning for Materials Scientists: An Introductory Guide toward Best Practices, Chemistry of Materials 32 (2020) 4954–4965. https://doi.org/10.1021/acs.chemmater.0c01907
65. S.M. Lundberg, G. Erion, H. Chen, A. DeGrave, J.M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable AI for trees, Nature Machine Intelligence 2 (2020) 56–67. https://doi.org/10.1038/s42256-019-0138-9
66. K. Aas, M. Jullum, A. Løland, Explaining individual predictions when features are dependent: More accurate approximations to Shapley values, Artificial Intelligence 298 (2021) 103502. https://doi.org/10.1016/j.artint.2021.103502
67. A.A. Alimin, K. Srasamran, W. Yuenyong, A. Charoensaeng, B.-J. Shiau, U. Suriyapraphadilok, Surfactant representation using COSMO screened charge density for adsorption isotherm prediction using Physics-Informed Neural Network (PINN), Journal of Cheminformatics 17 (2025) 84. https://doi.org/10.1186/s13321-025-01027-y
68. E. Chang, M. Zavarin, L. Beverly, H. Wainwright, A chemistry-informed hybrid machine learning approach to predict metal adsorption onto mineral surfaces, Applied Geochemistry 155 (2023) 105731. https://doi.org/10.1016/j.apgeochem.2023.105731
69. B. Sanchez-Lengeling, A. Aspuru-Guzik, Inverse molecular design using machine learning: Generative models for matter engineering, Science 361 (2018) 360–365. https://doi.org/10.1126/science.aat2663
70. K. Sattari, Y. Xie, J. Lin, Data-driven algorithms for inverse design of polymers, Soft Matter 17 (2021) 7607–7622. https://doi.org/10.1039/d1sm00725d
71. J.N. Kumar, Q. Li, K.Y.T. Tang, T. Buonassisi, A.L. Gonzalez-Oyarce, J. Ye, Machine learning enables polymer cloud-point engineering via inverse design, npj Computational Materials 5 (2019) 73. https://doi.org/10.1038/s41524-019-0209-9
72. C. Kim, R. Batra, L. Chen, H. Tran, R. Ramprasad, Polymer design using genetic algorithm and machine learning, Computational Materials Science 186 (2021) 110067. https://doi.org/10.1016/j.commatsci.2020.110067
73. K.M. Jablonka, G. Melpatti Jothiappan, S. Wang, B. Smit, B. Yoo, Bias free multiobjective active learning for materials design and discovery, Nature Communications 12 (2021) 2312. https://doi.org/10.1038/s41467-021-22437-0
74. T. Xu, J. Wang, S. Zhao, D. Chen, H. Zhang, Y. Fang, N. Kong, Z. Zhou, W. Li, H. Wang, Accelerating the prediction and discovery of peptide hydrogels with human-in-the-loop, Nature Communications 14 (2023) 3880. https://doi.org/10.1038/s41467-023-39648-2
75. T. Yue, L. Tao, V. Varshney, Y. Li, Benchmarking study of deep generative models for inverse polymer design, Digital Discovery 4 (2025) 910–926. https://doi.org/10.1039/d4dd00395k

76. G. Vogel, J.M. Weber, Inverse design of copolymers including stoichiometry and chain architecture, Chemical Science 16 (2025) 1161–1178. https://doi.org/10.1039/d4sc05900j
77. S. Kim, C.M. Schroeder, N.E. Jackson, Open Macromolecular Genome: Generative Design of Synthetically Accessible Polymers, ACS Polymers Au 3 (2023) 318–330. https://doi.org/10.1021/acspolymersau.3c00003
78. Q. Hu, L. Xiao, H. Jiang, Data-Driven Inverse Design of Ionically Conductive Hydrogel Formulations via Generative and Graph-Based Learning, Chem & Bio Engineering (2026). https://doi.org/10.1021/cbe.6c00128
79. S.T. Knox, S.J. Parkinson, C.Y.P. Wilding, R.A. Bourne, N.J. Warren, Autonomous polymer synthesis delivered by multi-objective closed-loop optimisation, Polymer Chemistry 13 (2022) 1576–1585. https://doi.org/10.1039/d2py00040g
80. M. Abolhasani, E. Kumacheva, The rise of self-driving labs in chemical and materials sciences, Nature Synthesis 2 (2023) 483–492. https://doi.org/10.1038/s44160-022-00231-0
81. G. Tom, S.P. Schmid, S.G. Baird, Y. Cao, K. Darvish, H. Hao, S. Lo, S. Pablo-García, E.M. Rajaonson, M. Skreta, N. Yoshikawa, S. Corapi, G.D. Akkoc, F. Strieth-Kalthoff, M. Seifrid, A. Aspuru-Guzik, Self-Driving Laboratories for Chemistry and Materials Science, Chemical Reviews 124 (2024) 9633–9732. https://doi.org/10.1021/acs.chemrev.4c00055
82. G. Xu, R. Zhang, T. Luo, Self-Driving Laboratory Optimizes the Lower Critical Solution Temperature of Thermoresponsive Polymers, Advanced Intelligent Discovery (2026) e202500177. https://doi.org/10.1002/aidi.202500177
83. C. Liu, J. An, X.C. Nguyen, P. Balasubramanian, Machine learning prediction of hydrochar adsorption capacity for methylene blue with limited data: Inspired by generative adversarial network-based augmentation, Energy & Environmental Sustainability 1 (2025) 100043. https://doi.org/10.1016/j.eesus.2025.100043