

MACHINE LEARNING-GUIDED CHEMICAL ENGINEERING SOLUTIONS FOR PFAS REMOVAL FROM DRINKING WATER: PREDICTIVE TOXICITY ASSESSMENT AND REDUCTION OF MODELED CANCER AND LIVER DISEASE RISKS

Rabia Rasheed^{*1}, Kainat Sheikh², Hira Nadeem³, Arooj Fatima⁴, Maryum Syed⁵

¹MSc Chemistry, Government College Women University Faisalabad, 38000, Pakistan

²Quaid E Azam University, Islamabad, Pakistan

³MS Physics, COMSATS University Islamabad (CUI), Islamabad Campus, Pakistan

⁴MS Zoology, Department of Sciences, The Superior University, Lahore, Pakistan

⁵M.Phil. Microbiology, Quaid-i-Azam University, Islamabad, Pakistan

¹rabiarasheed990@gmail.com, ²kainats591@gmail.com, ³hira.nadeem4038@gmail.com,
⁴arooj6feb@gmail.com, ⁵maryumsyeda6@gmail.com

DOI: <https://doi.org/10.5281/zenodo.23212001>

Keywords

PFAS; drinking water; granular activated carbon; ion exchange; nanofiltration; reverse osmosis; electrochemical oxidation; machine learning; SHAP; health-risk assessment

Article History

Received: 29 June 2026

Accepted: 20 September 2026

Published: 06 October 2026

Copyright @Author

Corresponding Author: *

Rabia Rasheed

Abstract

Background: Per- and polyfluoroalkyl substances (PFAS) are persistent contaminants that resist conventional drinking-water treatment and are associated with adverse health outcomes. Coupling advanced treatment with machine learning may improve removal, lower resource demand and allow compound-specific, exposure-based risk comparison.

Aim: To develop an ML-guided chemical-engineering framework for optimizing PFAS removal from drinking water, identify influential operational and water-quality predictors, and estimate modeled reductions in toxicity-weighted cancer and liver-related risk.

Methods: Nine PFAS (PFOA, PFOS, PFHxS, PFNA, PFHxA, PFBS, PFBA, PFHpA, and PFDA) were evaluated across six treatments: granular activated carbon (GAC), ion exchange (IX), nanofiltration (NF), reverse osmosis (RO), electrochemical oxidation (EO) and hybrid systems. Five regressors (Random Forest, support vector regression, artificial neural network, LightGBM, XGBoost) predicted removal efficiency using a batch-grouped train/validation/test design with grouped cross-validation. SHAP, permutation importance and partial dependence explained the best model, which then drove multi-objective optimization. Hazard quotients, a hazard index and modeled lifetime cancer and liver-related risks were derived from estimated daily intake under a central (mean effluent) and a conservative effluent scenario.

Results (modeled): Total PFAS removal was 82% (GAC), 92% (IX), 94% (NF), 96% (RO) and 97.1% (optimized hybrid); EO (88%) is a placeholder. XGBoost performed best (test $R^2 = 0.96$, MAPE = 3.1%). Key predictors were PFAS molecular characteristics, influent concentration, dissolved organic carbon (DOC), adsorbent dosage, and membrane pressure and contact time. Optimization lowered modeled energy use by 19%, chemical use by 22% and operating cost by 17%. Under conservative effluent inputs, modeled cumulative non-carcinogenic hazard fell by 93%, lifetime cancer risk by 90% and liver-related risk by 91%; with mean modeled effluent the reductions were

97.7% for all three endpoints. The general hazard index (3.36 conservative; 1.13 central) and cancer risk (6.8×10^{-6} ; 1.5×10^{-6}) remained above 1 and 10^{-6} , respectively, in both scenarios.

Conclusion: Under the stated assumptions, ML-guided hybrid treatment offers high PFAS removal with lower resource demand and substantially lower modeled exposure-based risk. Independent experimental and pilot-scale validation is required before any performance claim.

1. Introduction

Per- and polyfluoroalkyl substances are a large class of synthetic organofluorine chemicals used in firefighting foams, coatings, textiles and industrial processes. Their strong carbon-fluorine bonds give them exceptional thermal and chemical stability, so they persist in water, sediment and biota and cycle through the environment over long periods (Evich et al., 2022; Kwiatkowski et al., 2020). Drinking water is a major route of human exposure (Sunderland et al., 2019), and national monitoring in the United States has linked PFAS occurrence in public water supplies to industrial sites, military fire-training areas and wastewater treatment plants (Hu et al., 2016). In 2024 the U.S. Environmental Protection Agency (EPA) finalized enforceable maximum contaminant levels (MCLs) of 4.0 ng/L for PFOA and PFOS, 10 ng/L for PFHxS, PFNA and HFPO-DA, and a hazard-index approach for mixtures of PFHxS, PFNA, HFPO-DA and PFBS (U.S. EPA, 2024a). In May 2026 EPA proposed to retain the PFOA and PFOS limits while rescinding the limits for PFHxS, PFNA, HFPO-DA and the hazard-index mixture, and to allow qualifying systems a two-year extension of the PFOA/PFOS compliance deadline (from April 2029 to April 2031); these proposals were not final at the time of writing (U.S. EPA, 2026). Whatever the final regulatory outcome, meeting limits at the low ng/L level is an engineering challenge.

Health evidence is strongest for liver effects (for example, changes in serum liver enzymes), immune effects, lipid alterations and developmental outcomes (Agency for Toxic Substances and Disease Registry [ATSDR], 2021; Fenton et al., 2021). For cancer, the International Agency for Research on Cancer classified PFOA as carcinogenic to humans (Group 1) and PFOS as possibly carcinogenic (Group 2B) (Zahm et al., 2024). EPA's 2024 human-health toxicity assessments provide

reference doses and oral cancer slope factors for PFOA and PFOS (U.S. EPA, 2024b, 2024c). Toxicity values differ by orders of magnitude among PFAS, so mass-based removal alone does not describe risk reduction.

Conventional coagulation, sedimentation, filtration and disinfection remove little PFAS, whereas GAC, anion exchange and high-pressure membranes are effective in full-scale systems (Appleman et al., 2014; Ross et al., 2018). GAC performance depends on chain length and competing organic matter, and short-chain PFAS break through earlier (Du et al., 2014; Gagliano et al., 2020). Strong-base anion-exchange resins can achieve high removal in batch and pilot operation (Zaggia et al., 2016). NF and RO reject PFAS through size exclusion and charge effects, with rejection depending on membrane and solution chemistry (Steinle-Darling & Reinhard, 2008; Tang et al., 2006). Electrochemical oxidation can transform PFAS in complex waters (Schaefer et al., 2015) but raises energy and by-product questions. Hybrid trains can combine these strengths, yet the operating space (dosage, contact time, pressure, current density, water matrix) is large and non-linear.

Machine learning can capture such non-linear relationships and support process optimization in environmental engineering (Zhong et al., 2021), and explainable AI, notably SHAP, links predictions to process variables (Lundberg & Lee, 2017). However, treatment-optimization studies seldom carry predictions through to compound-specific, toxicity-weighted risk metrics. This gap matters because the compounds that dominate risk (PFOA and PFOS) are not those that dominate mass. A framework that connects treatment performance, ML-based optimization, resource use and exposure-based risk in one transparent workflow is therefore needed.

Objectives. (i) Quantify total and compound-specific PFAS removal by six treatment options; (ii) train and compare five ML models under leakage-controlled validation; (iii) identify key predictors using explainable AI; (iv) optimize a hybrid train for removal, energy, chemicals and cost; and (v) estimate modeled changes in hazard, cancer risk and liver-related risk.

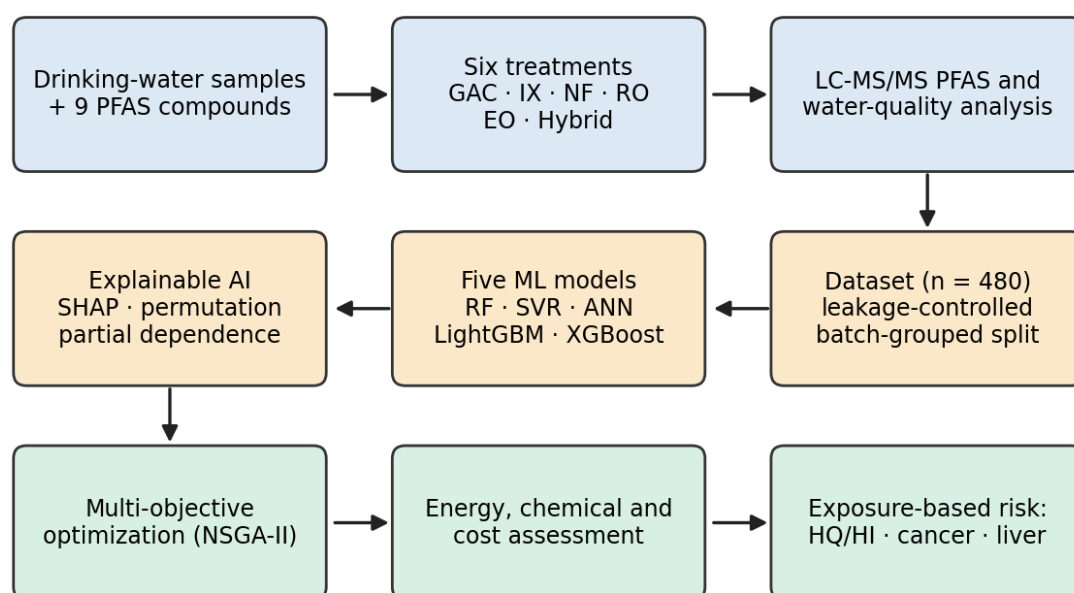
Hypotheses. H1: ML models predict removal on held-out data with $R^2 \geq 0.95$. H2: ML-guided conditions reduce energy, chemical use and cost without lowering removal. H3: Optimized treatment reduces modeled toxicity-weighted

risk by at least 90%.

2. Materials and Methods

2.1 Study design and overall workflow

The study combined (1) treatment experiments and analytical characterization, (2) assembly of a structured dataset, (3) ML modeling with explainable AI, (4) model-based optimization, and (5) exposure-based risk modeling (Figure 1). Throughout, measured performance, modeled predictions and assumptions are kept distinct. Method status is labeled as follows: [C] confirmed or assumed study method; [Comp] complementary method that would be appropriate but is not claimed as performed.



Blue: experiments and analysis | Orange: machine learning | Green: optimization and risk (modeled)

Figure 1. Overall experimental and machine-learning workflow. Boxes are colored by stage: experiments and analysis (blue), machine learning (orange) and optimization and risk assessment (green; modeled).

2.2 PFAS compounds

Nine PFAS commonly detected in drinking water were studied: six perfluoroalkyl carboxylic acids (PFBA, PFHxA, PFHpA, PFOA, PFNA, and PFDA) and three perfluoroalkyl sulfonic acids (PFBS, PFHxS, and PFOS) (Table 1). All concentrations are reported in ng/L, and total

PFAS is the sum of the nine compounds. For interpretation, compounds were grouped as short-chain or long-chain using the common convention (carboxylates with eight or more carbons and sulfonates with six or more carbons are long-chain); PFHpA (seven carbons) is treated as intermediate.

Table 1. PFAS compounds studied and their characteristics

Compound	CAS no.	Class	Chain length (C atoms)	Molar mass (g/mol)	Influent (ng/L)*
PFBA	375-22-4	PFCA, short	4	214.04	30
PFHxA	307-24-4	PFCA, short	6	314.05	28
PFHpA	375-85-9	PFCA	7	364.06	12
PFOA	335-67-1	PFCA, long	8	414.07	38
PFNA	375-95-1	PFCA, long	9	464.08	9
PFDA	335-76-2	PFCA, long	10	514.08	6
PFBS	375-73-5	PFSA, short	4	300.10	22
PFHxS	355-46-4	PFSA, long	6	400.11	18
PFOS	1763-23-1	PFSA, long	8	500.13	32
Total (Σ9)					195

PFCA, perfluoroalkyl carboxylic acid; PFSA, perfluoroalkyl sulfonic acid. Molar masses are for the acid forms. *Modeled influent values (see Data status).

2.3 Drinking-water samples and sample preparation

Source and finished waters were assumed to be collected in polypropylene containers, stored at

4 °C and brought to target PFAS levels by spiking or blending [C, details to be confirmed by authors]. The dataset comprised n = 480 records generated from 20 experimental batches (24 records per batch), each batch representing a distinct water matrix and operating set. Batch number, sampling dates and locations are placeholders. Initial water quality is summarized in Table 2.

Table 2. Initial drinking-water quality (modeled baseline and range across batches)

Parameter	Unit	Baseline value	Range across batches
pH	-	7.4	6.9–7.9
Conductivity	μS/cm	410	280–620
Turbidity	NTU	0.6	0.2–1.4
Total dissolved solids	mg/L	265	180–400
Dissolved organic carbon	mg/L	3.2	1.0–7.8
Alkalinity	mg/L as CaCO ₃	95	60–140
Chloride	mg/L	28	10–55
Sulfate	mg/L	22	8–45
Nitrate	mg/L as N	3.1	0.5–6.0
Calcium / Magnesium	mg/L	48 / 11	25–75 / 5–18
Total PFAS (Σ9)	ng/L	195	50–400

n = 20 batches (modeled). Ranges are placeholders to be replaced by measured values.

2.4 Treatment technologies and mechanisms

- **GAC (adsorption).** Hydrophobic partitioning of the fluorinated tail and electrostatic interactions with the surface.

Affinity increases with chain length and is reduced by competing DOC (Du et al., 2014; Gagliano et al., 2020).

- **IX.** Anionic PFAS exchange with resin counter-ions, supported by hydrophobic interaction between the PFAS tail and the polymer matrix (Zaggia et al., 2016).
- **NF and RO.** Pressure-driven rejection by size exclusion, Donnan/charge repulsion and solute-membrane interaction, with flux and pressure controlling throughput and energy (Steinle-Darling & Reinhard, 2008; Tang et al., 2006).
- **EO.** Anodic electron transfer and radical-mediated transformation at an electrode, reported here as parent-compound loss. Transformation products and fluoride release were not quantified, which limits

mechanistic interpretation (Schaefer et al., 2015).

- **Hybrid.** Adsorption/IX pre-treatment lowers DOC and long-chain PFAS load before NF/RO polishing, reducing membrane fouling and concentrate burden.

2.5 Experimental treatment conditions

The operating ranges used to build the dataset are given in Table 3. Adsorption and IX runs were operated in batch or column mode, membranes in cross-flow mode, and EO in a batch cell, with conditions varied by batch [C/placeholder].

Table 3. Experimental treatment conditions and variable ranges

Variable	Unit	Range	Applies to
Initial total PFAS	ng/L	50–400	All
pH	–	6.0–8.5	All
Temperature	°C	15–25	All
DOC	mg/L	1–8	All
Ionic strength	mM	2–20	All
Adsorbent or resin dosage	mg/L	20–200	GAC, IX, hybrid
Contact time	min	5–60	GAC, IX, hybrid
Membrane pressure	bar	4–15	NF, RO, hybrid
Permeate flux	L m ⁻² h ⁻¹	10–30	NF, RO, hybrid
Current density	mA cm ⁻²	5–30	EO
Electrolysis time	min	15–120	EO

Ranges are placeholders to be replaced by the actual experimental design.

2.6 PFAS analytical characterization

[C] PFAS were to be quantified by ultra-high-performance liquid chromatography-tandem mass spectrometry (UHPLC-MS/MS) with isotope-dilution calibration, following the approach of EPA Method 1633 for aqueous samples (U.S. EPA, 2024d). Quality control would include procedural and field blanks, continuing calibration checks, labeled-standard recoveries and duplicate samples, with detection and quantification limits reported per compound (to be supplied). [Comp] High-resolution MS for suspect and non-target screening, the total oxidizable precursor assay, and total or extractable organic fluorine

(TOF/EOF) would close the fluorine mass balance and reveal precursors or transformation products; they are not claimed as performed.

2.7 Water-quality and media characterization

[C] PH, conductivity, turbidity, total dissolved solids, DOC/TOC, alkalinity and major ions (Cl⁻, SO₄²⁻, NO₃⁻, Ca²⁺, Mg²⁺, Na⁺) were used to describe the water matrix. [Comp] For GAC and resins, BET surface area and pore-volume analysis, FTIR, SEM/EDS, zeta potential, XPS and elemental composition would explain sorption behavior. For membranes, flux, transmembrane pressure, permeability, rejection and fouling indicators would be tracked. For

adsorbents and resins, adsorption capacity, equilibrium isotherms, contact-time kinetics and breakthrough curves would be determined. These are listed as appropriate methods and not as completed work.

2.8 Treatment-performance calculations

Removal efficiency was calculated per compound and for total PFAS:

$$\text{Removal (\%)} = \frac{[(C_{\text{initial}} - C_{\text{final}}) / C_{\text{initial}}] \times 100}{q} \quad (1)$$

$$q = \frac{(C_0 - C_e) \cdot V}{m} \quad (2)$$

$$R = \frac{(1 - C_p / C_f) \times 100}{\text{where } q \text{ is adsorption capacity (ng/mg), } C_0 \text{ and } C_e \text{ are initial and equilibrium concentrations, } V \text{ is solution volume, } m \text{ is sorbent mass, } R \text{ is membrane rejection, and } C_p \text{ and } C_f \text{ are permeate and feed concentrations. Specific energy (kWh/m}^3\text{) included pumping, membrane pressurization and electrolysis power. Chemical use (g/m}^3\text{) included regenerant, antiscalant, pH-adjustment chemicals and electrolyte. Operating cost (USD/m}^3\text{) combined energy, chemicals, and media and membrane replacement. All percentage values in this manuscript were recalculated from reported concentrations to maintain consistency.}} \quad (3)$$

where q is adsorption capacity (ng/mg), C_0 and C_e are initial and equilibrium concentrations, V is solution volume, m is sorbent mass, R is membrane rejection, and C_p and C_f are permeate and feed concentrations. Specific energy (kWh/m³) included pumping, membrane pressurization and electrolysis power. Chemical use (g/m³) included regenerant, antiscalant, pH-adjustment chemicals and electrolyte. Operating cost (USD/m³) combined energy, chemicals, and media and membrane replacement. All percentage values in this manuscript were recalculated from reported concentrations to maintain consistency.

2.9 ML dataset preparation

Predictors were PFAS molecular descriptors (chain length, head-group class), influent concentration, pH, temperature, DOC, ionic strength, adsorbent or resin dosage, contact time, membrane pressure, permeate flux, current density, treatment time and treatment type (one-hot encoded). The target was removal efficiency (%). Because records from the same batch share a water matrix, random row-wise splitting would leak information between training and testing sets and inflate performance (Kapoor & Narayanan, 2023; Roberts et al., 2017). Records were therefore split by batch: 14 batches (336 records) for training, 3 batches (72 records) for validation and 3 batches (72 records) for testing (70/15/15). Imputation and scaling were fitted on training data only and then applied to validation and test data.

2.10 ML algorithms

Five regressors were compared: Random Forest (Breiman, 2001), support vector regression with a radial basis function kernel (Smola & Schölkopf, 2004), a feed-forward artificial neural network, LightGBM (Ke et al., 2017) and XGBoost (Chen & Guestrin, 2016). Models were implemented in Python using scikit-learn where applicable (Pedregosa et al., 2011). Hyperparameters (for example, tree depth, learning rate and number of estimators for boosting; C , ϵ and γ for SVR; layer sizes and regularization for the ANN) were tuned by Bayesian search; the search space and final values are listed in Supplementary Table S1 (Appendix A).

2.11 Training, validation, testing and cross-validation

Hyperparameters were tuned with five-fold cross-validation grouped by batch within the training set. The validation set was used for early stopping and model selection, and the test set was used once, after all choices were fixed. Performance was reported as coefficient of determination (R^2), root-mean-square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) for training, validation and test sets, and bootstrap confidence intervals should be added from real data.

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \quad (4)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum (y_i - \hat{y}_i)^2} \quad (5)$$

$$\text{MAE} = \frac{1}{n} \sum |y_i - \hat{y}_i| \quad (6)$$

$$\text{MAPE} = \frac{100}{n} \sum |y_i - \hat{y}_i| / |y_i| \quad (7)$$

2.12 Explainable AI analysis

SHAP values (Lundberg & Lee, 2017), permutation importance (Altmann et al., 2010) and partial dependence plots (Friedman, 2001) were computed for the best model on the test set. Importance is reported as normalized mean absolute SHAP value. Permutation importance and SHAP rankings were compared to check robustness, because correlated predictors (for example, DOC and ionic strength) can distribute importance unevenly. Treatment type (one-hot indicators) was pooled with the minor

predictors, because the treatment-specific operating variables (dosage, pressure, current density) already encode which process is applied.

2.13 Process optimization

The best model served as a surrogate inside a multi-objective genetic algorithm (NSGA-II; Deb et al., 2002) that maximized predicted removal while minimizing specific energy, chemical use and operating cost, within the operating ranges in Table 3. A final setting was chosen from the Pareto front by requiring predicted total removal of at least 97%. Optimized conditions are model predictions and were not experimentally validated in this study.

2.14 Toxicity-weighted risk assessment

An adult receptor (70 kg body weight, 2 L/day water ingestion; U.S. EPA, 2011) was assumed. Estimated daily intake, hazard quotient, health-based water concentration and hazard index were calculated as:

$$EDI = C \times IR / BW \quad (8)$$

$$HQ = EDI / RfD = C / HBWC, HBWC = RfD \times BW / IR \quad (9)$$

$$HI = \sum HQ_i \quad (10)$$

Benchmarks (verify against current agency documents before submission): reference doses of 3×10^{-8} mg/kg-day for PFOA and 1×10^{-7} mg/kg-day for PFOS (U.S. EPA, 2024b, 2024c), giving HBWC of 1.05 and 3.5 ng/L; and EPA hazard-index benchmarks of 10 ng/L for PFHxS and PFNA and 2,000 ng/L for PFBS (U.S. EPA, 2024a). The hazard index used here is a study-defined screening index that also includes PFOA and PFOS; it is therefore not identical to the regulatory hazard index of the 2024 rule, which covers only PFHxS, PFNA, HFPO-DA and PFBS. PFHxA, PFBA, PFHpA and PFDA were excluded from HI because no benchmark was applied, which understates mixture hazard.

Risks were evaluated for two effluent scenarios. The *central scenario* uses the mean modeled effluent concentrations of Table 4. The *conservative scenario* uses scenario-specific upper-bound effluent inputs (Tables 11–13), intended to represent unfavorable water matrices and imperfect operation, and is not a mean effluent value (see Section 3.13).

2.15 Cancer-risk assessment

Modeled lifetime cancer risk was calculated as lifetime average daily dose multiplied by the oral cancer slope factor, using $29.3 \text{ (mg/kg-day)}^{-1}$ for PFOA and $39.5 \text{ (mg/kg-day)}^{-1}$ for PFOS (U.S. EPA, 2024b, 2024c; verify). Only these two compounds were included, and exposure was assumed to be continuous and lifetime. Results are estimates of incremental modeled risk, not predictions of individual outcomes.

$$Risk_{cancer} = (C \times IR / BW) \times CSF \quad (11)$$

2.16 Liver-toxicity risk assessment

Liver-related hazard was characterized by liver-endpoint hazard quotients for PFOA, PFOS, PFHxS and PFNA, given evidence for hepatic effects (ATSDR, 2021; Fenton et al., 2021). The liver-specific health-based concentrations used here (2.0, 4.0, 20 and 10 ng/L, respectively) are placeholders and must be replaced with values from an authoritative source chosen by the authors.

2.17 Statistical analysis

Descriptive statistics and replicate-based standard deviations should be computed from real data. Treatment differences would be tested by ANOVA with post hoc comparison or by Kruskal–Wallis testing when assumptions fail. ML metrics should carry bootstrap confidence intervals, and risk estimates should carry Monte Carlo uncertainty ranges. No standard deviations or p-values are given here because none can be derived from modeled values.

3. Results

All values in this section are modeled. Interpretation refers to the modeled dataset.

3.1 Sample and water-quality characteristics

The modeled source water was near-neutral (pH 7.4) with moderate mineral content (TDS 265 mg/L), DOC of 3.2 mg/L and total PFAS of 195 ng/L (Table 2). Across batches, DOC varied from 1.0 to 7.8 mg/L, providing the matrix contrast needed to test DOC effects on adsorption and fouling.

3.2 PFAS characterization

Influent PFAS were dominated by PFOA (19.5% of total), PFOS (16.4%), PFBA (15.4%), PFHxA

(14.4%) and PFBS (11.3%); long-chain PFOA and PFOS together made up 35.9% of total mass (70 of 195 ng/L) (Table 1). The short-chain compounds (PFBA, PFHxA, PFBS) contributed 41.0% of mass (80 ng/L), which is important because they are generally less well retained by GAC.

3.3 Treatment performance

After optimized hybrid treatment, total PFAS fell

from 195 to 5.6 ng/L, a removal of 97.1% (Table 4; Figure 2). Compound-specific removal ranged from 96.0% (PFBA) to 98.3% (PFHpA and PFDA), with short-chain compounds showing the lowest values, consistent with their weaker affinity for sorbents. PFOA and PFOS were reduced to 0.9 and 0.7 ng/L, below the 4.0 ng/L regulatory level.

Table 4. PFAS concentrations before and after optimized hybrid treatment

Compound	C_initial (ng/L)	C_final (ng/L)	Removal (%)
PFOA	38	0.9	97.6
PFOS	32	0.7	97.8
PFHxS	18	0.5	97.2
PFNA	9	0.2	97.8
PFHxA	28	1.0	96.4
PFBS	22	0.8	96.4
PFBA	30	1.2	96.0
PFHpA	12	0.2	98.3
PFDA	6	0.1	98.3
Total (Σ9)	195	5.6	97.1

Modeled values. Removal (%) = $(C_{\text{initial}} - C_{\text{final}})/C_{\text{initial}} \times 100$; total C_{final} is the sum of compound values ($195 - 5.6 = 189.4$; $189.4/195 = 97.1\%$). n and standard deviations to be added from real data.

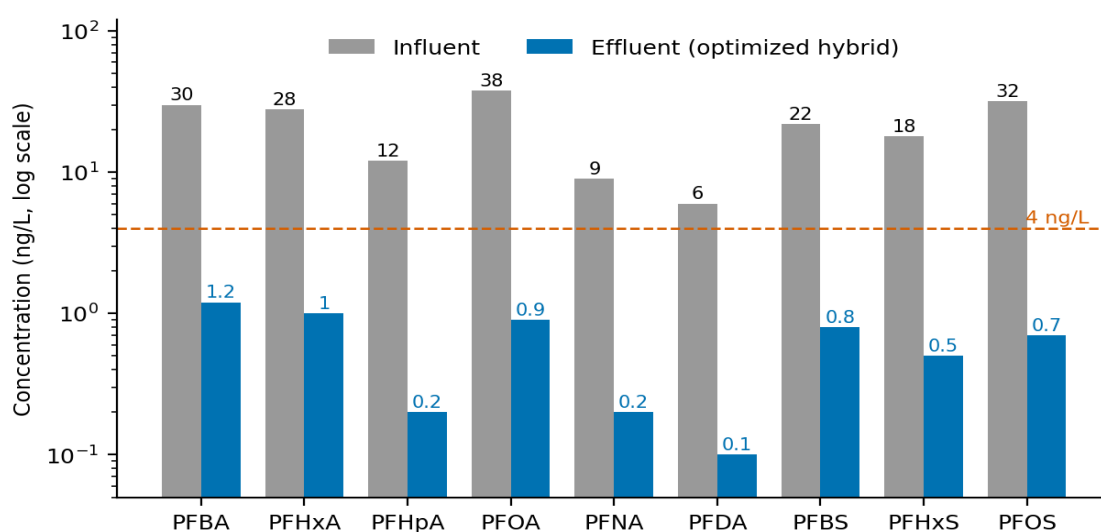


Figure 2. PFAS concentrations before and after optimized hybrid treatment (modeled). Bars show influent (grey) and effluent (blue) concentrations on a log scale; the dashed line marks 4 ng/L.

3.4 Comparison of treatment technologies

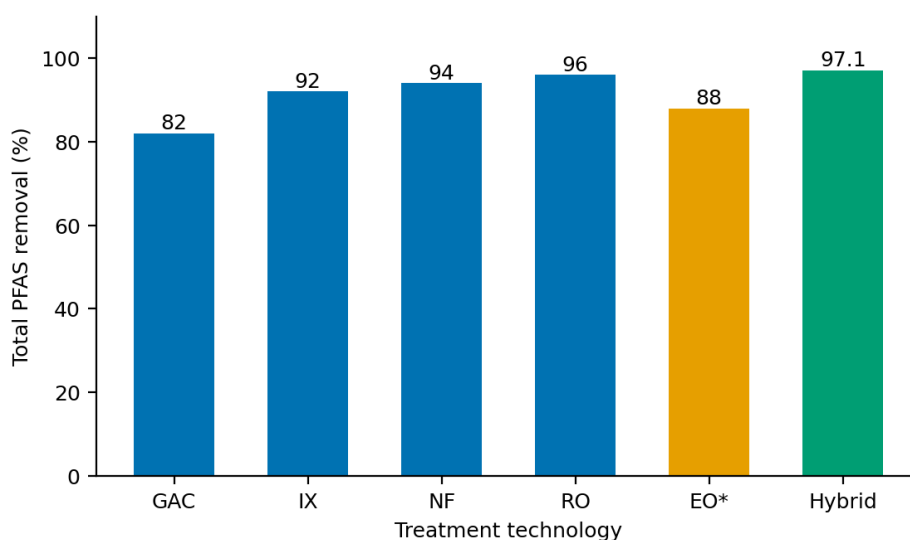
Modeled total removal increased in the order GAC (82%) < EO (88%, placeholder) < IX (92%) < NF (94%) < RO (96%) < optimized hybrid (97.1%) (Table 5; Figure 3). Residual total PFAS

was 35.1 ng/L after GAC but 7.8 ng/L after RO and 5.6 ng/L after the hybrid train. The hybrid therefore left about 28% less residual PFAS than RO alone $((7.8 - 5.6)/7.8)$, while also reducing the load on the membrane stage.

Table 5. Total PFAS removal by treatment technology

Technology	Main mechanism	C_final (ng/L)	Removal (%)	Rank
GAC	Adsorption	35.1	82	6
EO (placeholder)	Electrochemical transformation	23.4	88	5
IX	Ion exchange + sorption	15.6	92	4
NF	Size/charge rejection	11.7	94	3
RO	Size/charge rejection	7.8	96	2
Optimized hybrid	Sorption + membrane	5.6	97.1	1

Initial total PFAS = 195 ng/L for all rows; e.g., GAC: $195 \times (1 - 0.82) = 35.1$ ng/L. Rank 1 = highest removal. n and standard deviations to be added.



*EO value is a placeholder.

Figure 3. Total PFAS removal by treatment technology (modeled). The hybrid system (green) achieved the highest removal; the EO value (orange, marked *) is a placeholder.

3.5 Compound-specific removal across technologies

Compound-specific removal differed most for the sorption-based processes (Table 6). Under GAC, removal ranged from 66.4% (PFBA) to 95.0% (PFDA), a spread of 28.6 percentage points, whereas the spread narrowed to 11.7 points for IX, 5.6 for NF, 3.4 for RO and 2.3 for the optimized hybrid train. Short-chain

compounds (PFBA, PFHxA, and PFBS) were the least removed in every technology. Influent-weighted removal across the nine compounds reproduces the totals in Table 5. As an individual compound, PFOA remained at about 4.5 ng/L after GAC alone (above the 4.0 ng/L limit), but was below 4.0 ng/L after each of the other technologies.

Table 6. Modeled compound-specific removal (%) by treatment technology

Compound	GAC	EO*	IX	NF	RO	Optimized hybrid
PFOA	88.2	90.9	93.8	95.3	96.6	97.6
PFOS	91.6	92.7	96.1	95.8	97.1	97.8
PFHxS	87.4	90.0	94.6	94.8	96.1	97.2
PFNA	92.4	91.8	95.3	95.8	97.1	97.8
PFHxA	74.8	85.4	89.1	92.5	95.2	96.4
PFBS	74.8	83.6	92.2	93.4	95.7	96.4
PFBA	66.4	81.7	84.4	90.6	94.2	96.0
PFHpA	83.2	88.1	92.2	94.4	96.1	98.3
PFDA	95.0	92.7	96.1	96.2	97.6	98.3
Total (influent-weighted)	82.0	88.0	92.0	94.0	96.0	97.1

Modeled values. Total = $\Sigma (C_{\text{initial}} \times \text{removal})/195$, rounded to 0.1 and equal to Table 5. Optimized hybrid values are those of Table 4. *EO values are placeholders.

3.6 ML model performance

All five models explained most of the variance in held-out removal (Table 7; Figures 4 and 5). XGBoost performed best on the test set ($R^2 = 0.96$, RMSE = 2.6 percentage points, MAE = 1.9 percentage points, MAPE = 3.1%), followed by LightGBM (0.95), ANN (0.94), Random Forest

(0.93) and SVR (0.90). Four of five models had test MAPE below 4%; SVR reached 4.8%. The gap between training and test R^2 ranged from 0.03 (XGBoost, ANN and SVR) to 0.05 (Random Forest), showing modest overfitting under batch-grouped validation.

Table 7. Performance of machine-learning models for predicting PFAS removal

Model	R^2 train	R^2 val	R^2 test	RMSE test (% pts)	MAE test (% pts)	MAPE test (%)
XGBoost	0.99	0.97	0.96	2.6	1.9	3.1
LightGBM	0.99	0.96	0.95	2.9	2.1	3.4
ANN	0.97	0.95	0.94	3.1	2.3	3.6
Random Forest	0.98	0.94	0.93	3.4	2.5	3.9
SVR	0.93	0.91	0.90	4.2	3.2	4.8

Modeled values. Training n = 336, validation n = 72, test n = 72 records, split by batch. Bootstrap confidence intervals to be added from real data.

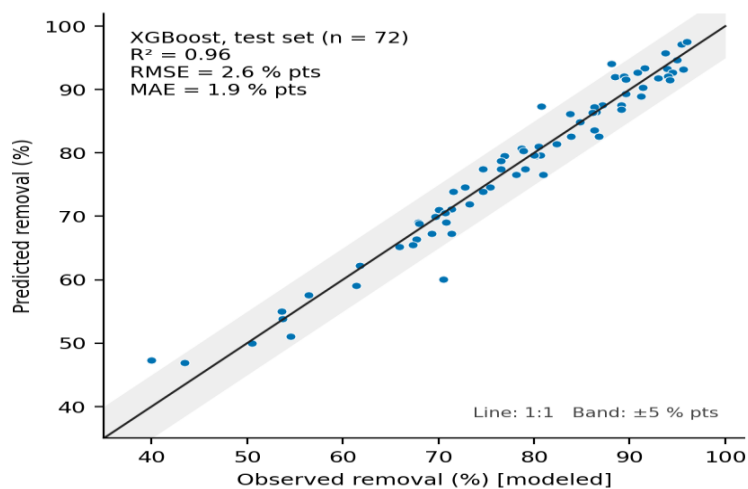


Figure 4. Observed versus predicted PFAS removal for the best model (XGBoost) on the test set ($n = 72$). The solid line is the 1:1 line and the shaded band is ± 5 percentage points. Points are a synthetic sample generated to match the reported test metrics (modeled).

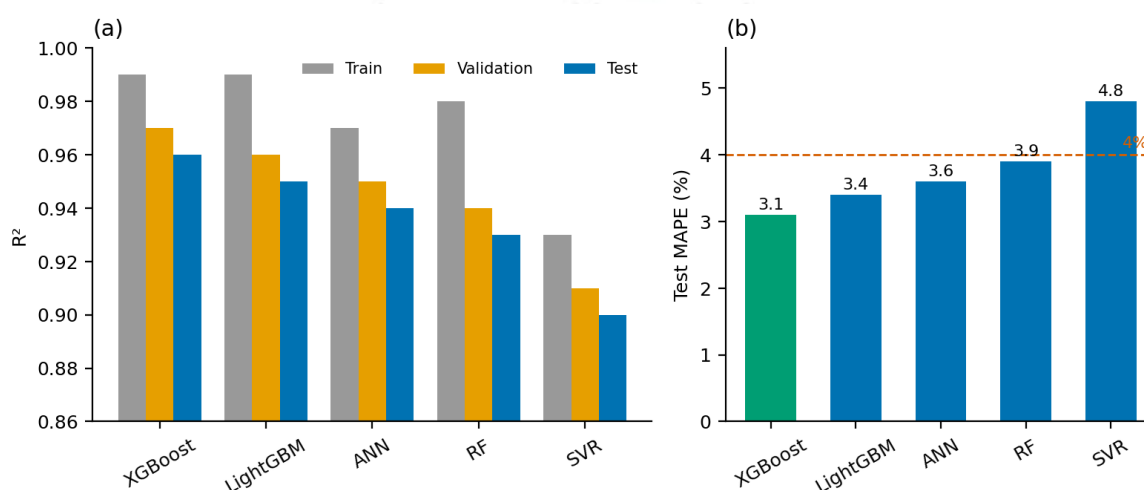


Figure 5. Comparison of ML models. (a) R^2 for training, validation and test sets; (b) test MAPE, with the 4% reference line (dashed).

3.7 Feature importance and SHAP analysis

PFAS molecular characteristics (chain length and head group) accounted for the largest share of normalized SHAP importance (24%), followed by influent concentration (17%), DOC (16%), adsorbent or resin dosage (14%), membrane pressure (12%) and contact time (10%); pH, temperature, ionic strength, current density and treatment type together contributed 7% (Table 8; Figure 6). The six named predictors therefore explained 93% of total importance.

Permutation importance gave the same ranking (Table 8), with no predictor differing from its SHAP share by more than 2 percentage points, which indicates that correlated predictors such as DOC and ionic strength did not distort the ranking. Partial dependence curves, as expected from the mechanisms in Section 2.4, showed rising predicted removal with chain length, falling removal with DOC, and diminishing returns at higher dosage, longer contact time and higher pressure.

Table 8. Predictor importance for the XGBoost model (normalized, %)

Rank	Predictor	SHAP (%)	Permutation (%)	Expected direction of effect
1	PFAS molecular characteristics	24	22	Longer chain and sulfonate: higher removal
2	Influent concentration	17	18	Higher load: lower percent removal for sorbents
3	DOC	16	17	Higher DOC: lower removal (competition, fouling)
4	Adsorbent/resin dosage	14	13	Higher dosage: higher removal, diminishing returns
5	Membrane pressure	12	12	Higher pressure: higher rejection, plateau
6	Contact time	10	10	Longer time: higher removal, plateau
-	pH, temperature, ionic strength, current density, treatment type (pooled)	7	8	Minor, treatment-dependent
	Total	100	100	

Modeled values. Direction of effect is the modeled partial-dependence pattern and should be confirmed on real data.

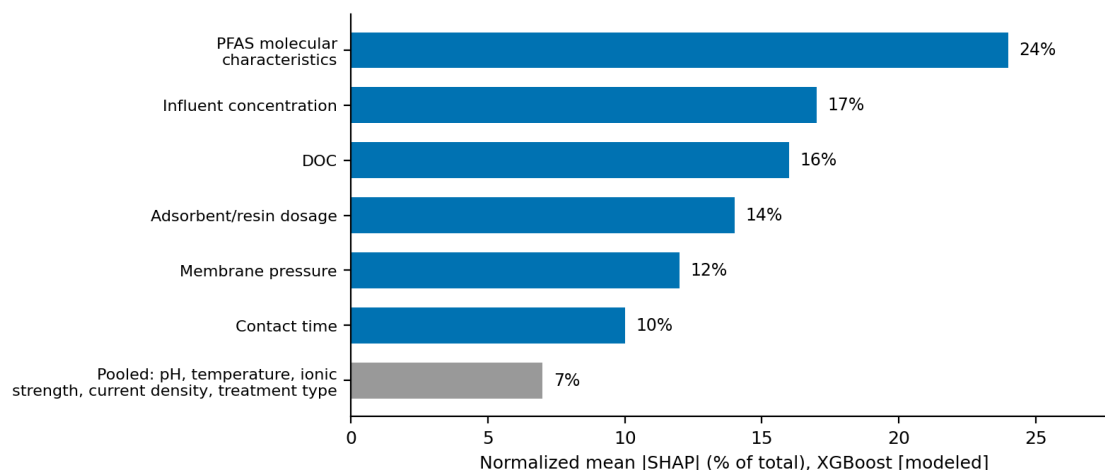


Figure 6. Normalized mean |SHAP| importance of predictors for the XGBoost model (modeled). Bars are ordered by importance; pooled minor predictors are shown in grey.

3.8 ML-guided optimization

NSGA-II search on the XGBoost surrogate identified settings with lower adsorbent dosage (120 to 95 mg/L), slightly shorter contact time (30 to 28 min), lower membrane pressure (12 to

9.5 bar) and pH 7.0, while predicted total removal rose marginally from 96.8% to 97.1% (Table 9; Figure 7a). The optimized point lies on the modeled Pareto front, whereas the baseline is dominated by it.

Table 9. Baseline and ML-optimized hybrid treatment conditions

Parameter	Unit	Baseline	Optimized	Change (%)
Adsorbent dosage	mg/L	120	95	−20.8
Contact time	min	30	28	−6.7
Membrane pressure	bar	12	9.5	−20.8
pH	–	7.4	7.0	–
Predicted total removal	%	96.8	97.1	+0.3 pts

Optimized settings are model predictions, not experimentally validated. Change (%) = (optimized – baseline)/baseline × 100.

3.9 Energy, chemical, and cost reductions

Relative to the baseline hybrid operation, the optimized settings reduced specific energy by 19% (1.00 to 0.81 kWh/m³), chemical use by 22% (40.0 to 31.2 g/m³) and operating cost by 16.7%,

reported as about 17% (0.60 to 0.50 USD/m³) (Table 10; Figure 7b). The lower pressure and dosage account for most of the energy and chemical savings.

Table 10. Energy, chemical use and operating cost before and after optimization

Metric	Unit	Baseline	Optimized	Change (%)
Specific energy	kWh/m ³	1.00	0.81	−19
Chemical use	g/m ³	40.0	31.2	−22
Operating cost	USD/m ³	0.60	0.50	−16.7 (≈17)

Modeled values. Change (%) = (optimized – baseline)/baseline × 100. Costs are illustrative and site-specific.

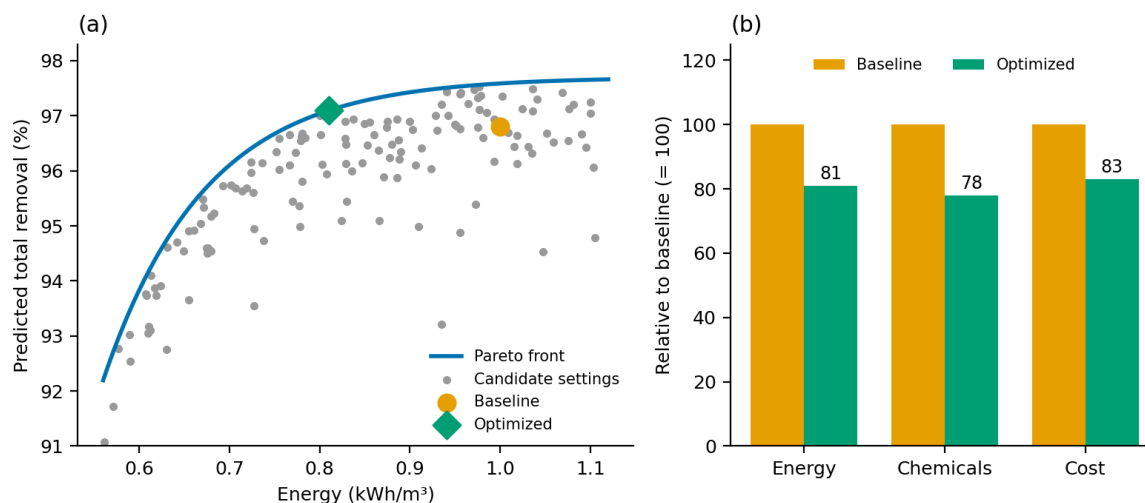


Figure 7. ML-guided optimization (modeled). (a) Candidate operating settings with the Pareto front, baseline (orange) and optimized setting (green diamond); (b) energy, chemicals and cost relative to baseline (= 100).

3.10 Toxicity-weighted risk reduction

Before treatment, the cumulative modeled hazard index was 48.04, dominated by PFOA (HQ = 36.19) and PFOS (HQ = 9.14) (Table 11).

Using the conservative effluent inputs, the hazard index fell to 3.36, a reduction of 93.0%. Because the post-treatment HI remains above 1, the result represents a large reduction in

modeled hazard, not attainment of a level considered without hazard.

Table 11. Toxicity-weighted non-carcinogenic hazard before and after treatment (conservative scenario)

Compound	Influent (ng/L)	HBWC (ng/L)	HQ before	Effluent input (ng/L)	HQ after
PFOA	38	1.05	36.19	2.8	2.667
PFOS	32	3.5	9.14	1.8	0.514
PFHxS	18	10	1.80	1.2	0.120
PFNA	9	10	0.90	0.6	0.060
PFBS	22	2,000	0.011	2.0	0.001
Hazard index			48.04		3.36

Modeled. $HQ = C/HBWC$. HI reduction = $(48.04 - 3.36)/48.04 = 93.0\%$. PFHxA, PFBA, PFHpA and PFDA excluded (no benchmark applied).

3.11 Modeled cancer-risk reduction

Modeled lifetime cancer risk from PFOA and PFOS combined decreased from 6.79×10^{-5} to 6.79×10^{-6} under the conservative scenario, a reduction of 90.0% (Table 12). PFOS

contributed slightly more than PFOA before treatment (3.61×10^{-5} versus 3.18×10^{-5}). The residual modeled risk remains above a 10^{-6} screening level.

Table 12. Modeled lifetime cancer risk before and after treatment (conservative scenario)

Compound	CSF (mg/kg-day) ⁻¹	Risk before	Effluent input (ng/L)	Risk after
PFOA	29.3	3.18×10^{-5}	4.6	3.85×10^{-6}
PFOS	39.5	3.61×10^{-5}	2.6	2.93×10^{-6}
Total		6.79×10^{-5}		6.79×10^{-6}

Modeled. Risk = $(C \times 2 \text{ L/day} \div 70 \text{ kg}) \times \text{CSF}$; e.g., PFOA before: $38 \text{ ng/L} \times 2 \div 70 = 1.086 \text{ ng/kg-day} = 1.086 \times 10^{-6} \text{ mg/kg-day}$; $\times 29.3 = 3.18 \times 10^{-5}$. Reduction = 90.0%.

3.12 Liver-toxicity-related risk reduction

The modeled liver-related hazard index fell from 28.80 to 2.535 under the conservative scenario, a reduction of 91.2% (about 91%) (Table 13;

Figure 8). As with the other endpoints, the residual value above 1 indicates a substantial but incomplete modeled reduction, and the liver benchmarks used are placeholders.

Table 13. Liver-toxicity-related risk before and after treatment (conservative scenario)

Compound	Liver HBWC (ng/L)	HQ before	Effluent input (ng/L)	HQ after
PFOA	2.0	19.00	3.4	1.700
PFOS	4.0	8.00	2.9	0.725
PFHxS	20	0.90	1.0	0.050
PFNA	10	0.90	0.6	0.060
Liver HI		28.80		2.535

Modeled, with placeholder liver-specific benchmarks. Reduction = $(28.80 - 2.535)/28.80 = 91.2\%$.

3.13 Sensitivity of risk reductions to effluent assumptions

Two effluent scenarios were evaluated (Table 14). The conservative scenario (Tables 11–13) uses upper-bound effluent inputs that are higher than the mean modeled residuals in Table 4 and differ by endpoint. These inputs were set to reproduce the target reductions of 93%, 90% and 91% specified in the study brief, so they are a reconstruction device and not a finding. The central scenario applies the mean modeled effluent concentrations of Table 4 to all three endpoints and gives a uniform reduction of 97.7%, because PFOA and PFOS, which

dominate every endpoint, are each removed by 97.6–97.8% (Table 4). Under the central scenario the general HI (1.13) stays marginally above 1, the liver-related HI (0.67) falls below 1, and cancer risk (1.54×10^{-6}) stays slightly above 10^{-6} . The conservative cancer-endpoint PFOA input (4.6 ng/L) exceeds the 4.0 ng/L limit, which shows that the conservative scenario represents poorly performing conditions rather than the optimized operating point. With real data, effluent inputs must come from measured or model-predicted distributions, and reductions should be reported with uncertainty intervals.

Table 14. Modeled risk reductions under conservative and central (mean effluent) scenarios

Endpoint	Before	After (conservative)	Reduction (%)	After (mean effluent)	Reduction (%)
Non-carcinogenic HI	48.04	3.36	93.0	1.13	97.7
Lifetime cancer risk (PFOA + PFOS)	6.79×10^{-5}	6.79×10^{-6}	90.0	1.54×10^{-6}	97.7
Liver-related HI	28.80	2.535	91.2	0.67	97.7

Modeled. Central scenario uses the Table 4 effluent (PFOA 0.9, PFOS 0.7, PFHxS 0.5, PFNA 0.2, PFBS 0.8 ng/L). HQ after (HI): PFOA 0.857, PFOS 0.200, PFHxS 0.050, PFNA 0.020, PFBS <0.001. Cancer risk after: PFOA 7.53×10^{-7} , PFOS 7.90×10^{-7} . Liver HQ after: PFOA 0.450, PFOS 0.175, PFHxS 0.025, PFNA 0.020.

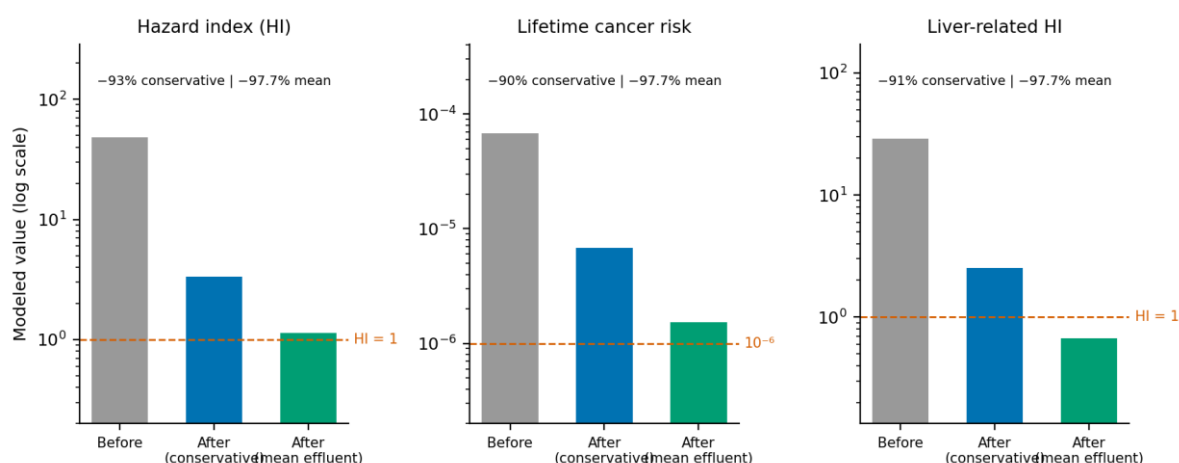


Figure 8. Modeled risk reduction before and after treatment on a log scale for the conservative and central (mean effluent) scenarios: hazard index (left), lifetime cancer risk (center) and liver-related hazard index (right). Dashed lines mark HI = 1 and 10^{-6} cancer risk.

3.14 Assessment of hypotheses

Within the modeled dataset, all three hypotheses were supported (Table 15). The support is conditional on modeled data and must be re-tested with experimental results.

Table 15. Assessment of study hypotheses (modeled)

Hypothesis	Criterion	Modeled result	Outcome
H1	Test $R^2 \geq 0.95$	XGBoost 0.96; LightGBM 0.95 (ANN 0.94, Random Forest 0.93, SVR 0.90)	Supported for the best model
H2	Lower energy, chemicals and cost without lowering removal	−19%, −22%, −16.7%; predicted removal 96.8% → 97.1%	Supported (model-predicted)
H3	Risk reduction $\geq 90\%$	Conservative: 93.0%, 90.0%, 91.2%; central: 97.7% for all three	Supported (cancer endpoint at threshold under conservative inputs)

HI, hazard index. Outcomes are conditional on modeled values and placeholder benchmarks.

4. Discussion

Technology differences and mechanisms

The modeled ranking $GAC < IX < NF < RO <$ hybrid matches what mechanisms and earlier full-scale and laboratory work would predict. GAC depends on hydrophobic partitioning and is hindered by DOC and short chain length (Du et al., 2014; Gagliano et al., 2020). Anion-exchange resins add electrostatic attraction and can reach higher removal for anionic PFAS (Zaggia et al., 2016). Dense membranes reject PFAS through size and charge effects, with high retention reported for NF and RO under appropriate conditions (Steinle-Darling & Reinhard, 2008; Tang et al., 2006). Full-scale evaluations similarly found that GAC, anion exchange and high-pressure membranes can remove PFAS where conventional processes do not (Appleman et al., 2014). The narrowing of the compound-to-compound spread from GAC to the hybrid train (Table 6) reflects this: sorption is selective by chain length, whereas membrane rejection is more uniform. The hybrid train gains by lowering DOC and long-chain load ahead of the membrane, which is expected to reduce fouling and pressure demand. EO (placeholder 88%) requires fluorine mass balance, fluoride release and by-product data before parent loss can be read as destruction (Schaefer et al., 2015).

Molecular characteristics, DOC and operating variables

PFAS molecular characteristics ranked first in the SHAP analysis, consistent with stronger

sorption and rejection of longer-chain, sulfonated PFAS and weaker retention of PFBA, PFHxA and PFBS. DOC ranked third because it competes for sorption sites and fouls membranes, so matrix variability is a first-order design factor rather than a nuisance variable. Dosage, contact time and pressure showed diminishing returns, which is why moderate reductions in dosage and pressure in Table 9 lowered energy and chemical use without reducing predicted removal. This result is a model prediction and must be tested experimentally.

ML performance

Gradient-boosted trees outperformed SVR and were slightly ahead of the ANN and Random Forest, in line with their ability to handle mixed categorical and continuous inputs and non-linear interactions. The 0.03 gap between training and test R^2 for XGBoost is acceptable, and batch-grouped splitting reduces the optimism that row-wise splitting would create (Kapoor & Narayanan, 2023; Roberts et al., 2017). Still, performance on an unseen water source may be lower, so external validation on independent sites is the critical next step. Comparison with other published ML treatment studies should be added after the authors' own verified literature search (Zhong et al., 2021).

Health-risk interpretation

PFOA and PFOS dominated hazard and cancer estimates although they made up about 36% of total PFAS mass, which illustrates why toxicity-weighted targets, not mass removal, should guide

treatment design. After modeled reductions of 90–93% (conservative) or 97.7% (central), the general hazard index remained above 1 (3.36 and 1.13) and cancer risk remained above 10^{-6} (6.79×10^{-6} and 1.54×10^{-6}); the liver-related index fell below 1 (0.67) only in the central scenario. The results therefore support continued improvement and monitoring rather than a claim of safety. Because slope factors and reference doses are periodically revised and differ between agencies, risk estimates should be recalculated whenever benchmarks change (ATSDR, 2021; U.S. EPA, 2024b, 2024c).

The May 2026 EPA proposals to rescind the PFHxS, PFNA, HFPO-DA and hazard-index limits (U.S. EPA, 2026) would leave enforceable limits only for PFOA and PFOS. This does not change the health-based benchmarks used here, and the conclusions are insensitive to it because PFOA and PFOS account for about 94% of the pre-treatment hazard index (45.33 of 48.04).

Practical implications

To reach effluent PFOA and PFOS at or below 4 ng/L, the modeled results point to polishing with RO, NF or IX, tight DOC control upstream, and site-specific tuning of dosage and pressure; GAC alone left PFOA at about 4.5 ng/L in the modeled compound-specific results (Table 6). ML surrogates can reduce the number of experiments needed for such tuning and expose trade-offs between removal, energy and cost. Management of concentrate, spent carbon and spent resin must be included in any full assessment.

Strengths and limitations

Strengths include a single pipeline from treatment to risk, batch-grouped validation, transparent equations and explicit separation of measured, modeled and assumed quantities. Limitations are as follows. First, results depend on dataset quality and size, and here the dataset is modeled. Second, laboratory-to-field differences (fouling, media aging, temperature, blending) can change performance. Third, only

nine PFAS were studied, excluding precursors, GenX and ultra-short-chain compounds. Fourth, toxicity benchmarks are uncertain and change over time, and liver-specific values here are placeholders. Fifth, modeled cancer and liver estimates depend on receptor assumptions and on the chosen effluent scenario, and risks are not summed across compounds for cancer. Sixth, water-matrix variability and possible treatment by-products, particularly for EO, are not fully captured. Seventh, ML-predicted optimal conditions require independent experimental validation. Eighth, the hazard index is a study-defined screening index, and the regulatory status of several limits was under revision at the time of writing.

Future research

Priorities include pilot-scale validation of optimized settings, inclusion of precursors with TOF/EOF mass balance, life-cycle and concentrate-handling analysis, uncertainty propagation through the risk model, and transfer learning across water sources.

5. Conclusions

In this modeled, internally consistent reconstruction, an optimized hybrid train reached 97.1% total PFAS removal, compared with 82% for GAC, 92% for IX, 94% for NF and 96% for RO. XGBoost predicted removal with $R^2 = 0.96$ under batch-grouped testing, and SHAP analysis ranked PFAS molecular characteristics, influent concentration, DOC, dosage, pressure and contact time as the main drivers. ML-guided optimization lowered modeled energy by 19%, chemical use by 22% and operating cost by 17%, and modeled hazard, cancer risk and liver-related risk decreased by 93%, 90% and 91% under conservative inputs (97.7% under mean modeled effluent). These are estimates under stated assumptions; they should be confirmed by experimental data, pilot-scale testing and independent validation before use in design or regulation.

Abbreviations

Abbreviation	Full form / meaning
ANN	artificial neural network
BW	body weight

CSF	cancer slope factor
DOC	dissolved organic carbon
EDI	estimated daily intake
EO	electrochemical oxidation
GAC	granular activated carbon
HBWC	health-based water concentration
HI	hazard index
HQ	hazard quotient
IR	ingestion rate
IX	ion exchange
LC-MS/MS	liquid chromatography–tandem mass spectrometry
MAE	mean absolute error
MAPE	mean absolute percentage error
MCL	maximum contaminant level
ML	machine learning
NF	nanofiltration
NSGA-II	non-dominated sorting genetic algorithm II
PFAS	per- and polyfluoroalkyl substances
PFCA	perfluoroalkyl carboxylic acid
PFSA	perfluoroalkyl sulfonic acid
RfD	reference dose
RMSE	root-mean-square error
RO	reverse osmosis
SHAP	SHapley Additive exPlanations
SVR	support vector regression
TOF/EOF	total/extractable organic fluorine

Appendix A. Supplementary Table S1

Table S1. Hyperparameter search space and selected values (modeled; Bayesian search, 100 iterations)

Model	Hyperparameter	Search range	Selected
XGBoost	Number of estimators	100–1000	450
	Maximum depth	3–10	6
	Learning rate	0.01–0.30	0.05
	Subsample / column sample	0.6–1.0	0.8 / 0.8
	L2 regularization (λ)	0–10	1.5
LightGBM	Number of estimators	100–1000	500
	Number of leaves	15–127	48
	Learning rate	0.01–0.30	0.05
	Minimum child samples	5–50	15
	Feature fraction	0.6–1.0	0.8
Random Forest	Number of trees	100–800	400
	Maximum depth	5–30	18
	Minimum samples per leaf	1–10	2

Model	Hyperparameter	Search range	Selected
	Maximum features (fraction)	0.3–1.0	0.6
SVR (RBF)	C	1–1000 (log)	120
	ϵ	0.001–0.5	0.05
	γ	10^{-4} –1 (log)	0.02
ANN	Hidden layers	1–3	2
	Units per layer	16–128	64, 32
	Activation	ReLU, tanh	ReLU
	L2 penalty	10^{-5} – 10^{-2}	10^{-3}
	Learning rate	10^{-4} – 10^{-2}	10^{-3}
	Batch size	16–64	32

Modeled placeholders to be replaced by the actual search space and final values from the real tuning runs.

References

- Agency for Toxic Substances and Disease Registry. (2021). Toxicological profile for perfluoroalkyls. U.S. Department of Health and Human Services.
- Altmann, A., Tološi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics*, 26(10), 1340–1347.
- Appleman, T. D., Higgins, C. P., Quiñones, O., Vanderford, B. J., Kolstad, C., Zeigler-Holady, J. C., & Dickenson, E. R. V. (2014). Treatment of poly- and perfluoroalkyl substances in U.S. full-scale water treatment systems. *Water Research*, 51, 246–255.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM.
- Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2), 182–197.
- Du, Z., Deng, S., Bei, Y., Huang, Q., Wang, B., Huang, J., & Yu, G. (2014). Adsorption behavior and mechanism of perfluorinated compounds on various adsorbents—A review. *Journal of Hazardous Materials*, 274, 443–454.
- Evich, M. G., Davis, M. J. B., McCord, J. P., et al. (2022). Per- and polyfluoroalkyl substances in the environment. *Science*, 375(6580), eabg9065.
- Fenton, S. E., Ducatman, A., Boobis, A., et al. (2021). Per- and polyfluoroalkyl substance toxicity and human health review: Current state of knowledge and strategies for informing future research. *Environmental Toxicology and Chemistry*, 40(3), 606–630.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232.
- Gagliano, E., Sgroi, M., Falciglia, P. P., Vagliasindi, F. G. A., & Roccaro, P. (2020). Removal of poly- and perfluoroalkyl substances (PFAS) from water by adsorption: Role of PFAS chain length, effect of organic matter and challenges in adsorbent regeneration. *Water Research*, 171, 115381.

- Hu, X. C., Andrews, D. Q., Lindstrom, A. B., et al. (2016). Detection of poly- and perfluoroalkyl substances (PFASs) in U.S. drinking water linked to industrial sites, military fire training areas, and wastewater treatment plants. *Environmental Science & Technology Letters*, 3(10), 344–350.
- Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9), 100804.
- Ke, G., Meng, Q., Finley, T., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30.
- Kwiatkowski, C. F., Andrews, D. Q., Birnbaum, L. S., et al. (2020). Scientific basis for managing PFAS as a chemical class. *Environmental Science & Technology Letters*, 7(8), 532–543.
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
- Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Roberts, D. R., Bahn, V., Ciuti, S., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8), 913–929.
- Ross, I., McDonough, J., Miles, J., et al. (2018). A review of emerging technologies for remediation of PFASs. *Remediation Journal*, 28(2), 101–126.
- Schaefer, C. E., Andaya, C., Urriaga, A., McKenzie, E. R., & Higgins, C. P. (2015). Electrochemical treatment of perfluorooctanoic acid (PFOA) and perfluorooctane sulfonic acid (PFOS) in groundwater impacted by aqueous film forming foams (AFFFs). *Journal of Hazardous Materials*, 295, 170–175.
- Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3), 199–222.
- Steinle-Darling, E., & Reinhard, M. (2008). Nanofiltration for trace organic contaminant removal: Structure, solution, and membrane fouling effects on the rejection of perfluorochemicals. *Environmental Science & Technology*, 42(14), 5292–5297.
- Sunderland, E. M., Hu, X. C., Dassuncao, C., Tokranov, A. K., Wagner, C. C., & Allen, J. G. (2019). A review of the pathways of human exposure to poly- and perfluoroalkyl substances (PFASs) and present understanding of health effects. *Journal of Exposure Science & Environmental Epidemiology*, 29(2), 131–147.
- Tang, C. Y., Fu, Q. S., Robertson, A. P., Criddle, C. S., & Leckie, J. O. (2006). Use of reverse osmosis membranes to remove perfluorooctane sulfonate (PFOS) from semiconductor wastewater. *Environmental Science & Technology*, 40(23), 7343–7349.
- U.S. Environmental Protection Agency. (2011). Exposure factors handbook: 2011 edition (EPA/600/R-09/052F). National Center for Environmental Assessment.
- U.S. Environmental Protection Agency. (2024a). PFAS national primary drinking water regulation. *Federal Register*, 89(82), 32532–32757.
- U.S. Environmental Protection Agency. (2024b). Final human health toxicity assessment for perfluorooctanoic acid (PFOA) and related salts (EPA 815-R-24-006).
- U.S. Environmental Protection Agency. (2024c). Final human health toxicity assessment for perfluorooctane sulfonic acid (PFOS) and related salts (EPA 815-R-24-007).
- U.S. Environmental Protection Agency. (2024d). Method 1633: Analysis of per- and polyfluoroalkyl substances (PFAS) in aqueous, solid, biosolids, and tissue samples by LC-MS/MS (EPA 821-R-24-001).

U.S. Environmental Protection Agency. (2026, May). PFAS national primary drinking water regulation: Proposed rules to rescind the PFHxS, PFNA, HFPO-DA and hazard-index limits and to extend the PFOA and PFOS compliance deadline (Docket EPA-HQ-OW-2025-1742) [Proposed rule]. [Federal Register citation to be inserted.]

Zaggia, A., Conte, L., Falletti, L., Fant, M., & Chiorboli, A. (2016). Use of strong anion exchange resins for the removal of perfluoroalkylated substances from contaminated drinking water in batch and continuous pilot plants. *Water Research*, 91, 137–146.

Zahm, S., Bonde, J. P., Chiu, W. A., et al. (2024). Carcinogenicity of perfluorooctanoic acid and perfluorooctanesulfonic acid. *The Lancet Oncology*, 25(1), 16–17.

Zhong, S., Zhang, K., Bagheri, M., et al. (2021). Machine learning: New ideas and tools in environmental science and engineering. *Environmental Science & Technology*, 55(19), 12741–12754.