Research Square

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

Chemical-space-aware routing for improved PFAS toxicity prediction: large-scale pretraining, conditional fine-tuning, and leakage control

Zhanting Yang

ztyang1121@163.com

Qinghai University

Guyue Bai

Qinghai University

Junhao Peng

Qinghai University

Xinglei Hao

Qinghai University

Research Article

Keywords: PFAS, Graph neural network, Transfer learning, Data leakage, Chemical-space-aware routing, ADMET-AI

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11160855/v1

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

Additional Declarations: No competing interests reported.

Chemical-space-aware routing for improved PFAS toxicity prediction: large-scale pretraining, conditional fine-tuning, and leakage control

Zhanting Yang[1*†], Guyue Bai[1†], Junhao Peng[2], Xinglei Hao[1]

[1]Medical College, Qinghai University, 810001, China

[2]School of Computer Technology and Applications, Qinghai University, 810001, China

*Correspondence: Zhanting Yang, ztyang1121@163.com.

†Zhanting Yang and Guyue Bai have contributed equally to this work.

Abstract

Predicting the toxicity of per- and polyfluoroalkyl substances (PFAS)—persistent environmental pollutants with bioactivity data available for only a small fraction of inventoried compounds-remains challenging because PFAS-specific training sets are small. Here, we quantify how much domain-specific data is actually required and introduce chemical-space-aware routing, in which a structure-based rule automatically directs PFAS-like molecules to a dedicated prediction panel within an open ADMET platform. We trained directed message-passing neural networks (Chemprop) on a 26-bioassay PFAS panel (n = 1,012) derived from the Cheng–Ng dataset, comparing training from scratch, large-scale multi-task pretraining on ~53,000 non-PFAS compounds from the same assay collection, and PFAS fine-tuning, under random and scaffold (perfluoroalkyl chain-length) splitting regimes with three random seeds and a Morgan-fingerprint random-forest baseline. Canonical-structure overlap analysis revealed that the PFAS panel is a strict subset of the pretraining collection (1,012/1,012 compounds), so pretraining without exclusion fabricates an apparent fine-tuning gain of up to +0.36 AUROC. After decontamination, scratch-trained models showed a large generalization gap (random test AUROC 0.751 vs. scaffold 0.607 ± 0.044) and were outperformed by the random-forest baseline on the scaffold split (0.769); a pretrained model that had never encountered any PFAS achieved 0.798, nearly closing the gap, while fine-tuning added gains only under random splitting (0.864 vs. 0.804) and not on novel chain lengths (0.812 ± 0.032 vs. 0.798). The routing rule detected 82 of 86 held-out PFAS (95.3% sensitivity) and flagged 41 of

368 approved drugs (88.9% specificity), notably, 14 of the 41 flagged drugs are perfluorocarbons or fluorinated inhalational anesthetics chemically akin to PFAS, revealing a previously undocumented overlap between the approved-drug and PFAS chemical spaces. Category-level profiles placed CYP-mediated metabolism as the dominant axis for PFOA and PFOS, both of which appear unremarkable on conventional drug-oriented ADMET summaries. The code, decontaminated data, and model weights are freely available at https://github.com/zhanting-yang/PFAS-GNN (DOI: 10.5281/zenodo.22869943).

Keywords: PFAS; Graph neural network; Transfer learning; Data leakage; Chemical-space-aware routing; ADMET-AI

Introduction

Per- and polyfluoroalkyl substances (PFAS) constitute one of the largest classes of manufactured chemicals, with current inventories listing several thousand compounds [1, 2]. Their carbon-fluorine backbones confer extreme persistence and widespread environmental distribution, and exposure to PFAS such as perfluorooctanoic acid (PFOA) and perfluorooctane sulfonic acid (PFOS) has been linked to immunotoxicity, developmental toxicity and metabolic disruption in toxicological studies[3]. Epidemiological studies have also revealed associations between exposure to specific PFAS and a variety of health effects, including altered immune and thyroid function, liver disease, lipid and insulin dysregulation, kidney disease, adverse reproductive and developmental outcomes, and cancer[4]. Regulatory agencies therefore face a prioritization problem: the overwhelming majority of inventoried PFAS have little or no experimental toxicity data, and filling these gaps by testing alone is infeasible[2, 4].

Machine learning is an obvious complement to experimental prioritization. Cheng and Ng [5] assembled the first PFAS-specific bioactivity dataset-1,012 PFAS annotated across 26 PubChem bioassays-and demonstrated that neural networks trained on this panel achieve high accuracy (mean AUROC $\approx$ 0.92) and can be extrapolated to the thousands of PFAS in the OECD global inventory. This landmark study established feasibility, yet three methodological questions remain unresolved.

First, evaluations in that work and in subsequent molecular machine-learning benchmarks rely overwhelmingly on random train–test splits [6, 7]. For PFAS, whose chemical space is dominated by homologous perfluoroalkyl series, random splitting places near-identical chain-length variants in both training and test sets and thereby rewards memorization rather than generalization, as has been demonstrated for ligand-based classification benchmarks more generally [8]. Second-and to our knowledge unexamined for this dataset-the small PFAS panel is a subset of the much larger PubChem BioAssay collection covering the same 26 assays. Any pretraining on that collection without explicit sample exclusion leaks test-set molecules and their labels into the training data. Third, the practical question of how much PFAS-specific data is truly needed, and whether a generic pretrained model suffices, has not been answered under leakage-controlled conditions.

A further, largely overlooked issue is platform mismatch. The de facto standard tools for ADMET (absorption, distribution, metabolism, excretion, toxicity) prediction are drug-oriented: they were trained on drug-like chemical matter and report endpoints such as Caco-2 permeability, hERG inhibition and drug-induced liver injury [9, 10]. We show below that such platforms are essentially blind to PFAS-relevant bioactivity: PFOA, for example, scores as unremarkable across the standard drug-ADMET spectrum despite its well-documented toxicological profile[4, 11]. An ideal prediction platform would therefore recognize what chemical space an input belongs to and route it to the appropriate expert model.

Here we address all four points in a single study. Using the Cheng–Ng panel and its parent PubChem BioAssay collection, we (i) quantify the generalization gap of from-scratch graph neural networks under homolog-aware scaffold splitting; (ii) detect, document and eliminate pretraining–test leakage, then re-quantify the true contributions of large-scale pretraining and PFAS fine-tuning under identical protocols; (iii) benchmark against a classical Morgan-fingerprint random-forest baseline; and (iv) implement chemical-space-aware routing together with an ADME-Tox category aggregation of the 26 assays, deploying the resulting PFAS panel as a custom model in the open-source ADMET-AI platform[9]. Our central

findings are that large-scale multi-task pretraining transfers across chemical domains and is the primary driver of generalizable PFAS prediction; that PFAS-specific fine-tuning adds value only for compounds within the chemical space covered by the domain training set; and that without leakage control, benchmark design can fabricate an apparent fine-tuning benefit that does not exist.

Methods

Datasets

The PFAS panel of Cheng and Ng [5] comprises 1,012 per- and polyfluoroalkyl substances annotated across 26 bioassays from the Tox21/ToxCast-oriented PubChem BioAssay (PCBA) effort[7, 12], spanning xenobiotic-metabolizing enzymes (CYP2C9, CYP3A4, CYP2D6, ALDH1A1), transcriptional and epigenetic regulators, protein–protein interaction targets, membrane receptors and ion channels, cellular-viability assays, and pathogen-response screens (complete mapping in Table S1). Assay-level positive rates in the panel range from 1.7% to 43%. The large PCBA release covering the same 26 assays contains 60,027 tested compounds[7] and served as the pretraining source.

Leakage detection and decontamination

Because the PFAS panel was assembled by filtering the broader PCBA collection, we tested whether the two datasets overlap. Canonical SMILES matching (RDKit canonicalization, salt removal, retention of the largest fragment) revealed that all 1,012 PFAS of the panel are contained in the large PCBA release. We therefore excluded every panel compound from the pretraining pool before any large-scale training; the decontaminated pretraining set is generated by the released preprocessing script and is part of the archived data.

Model architecture and training

All graph models were directed message-passing neural networks (D-MPNN) as implemented in Chemprop 2.3.1[13, 14], trained in multi-task mode with the 26 assays as prediction heads (≈325K trainable parameters). Default training settings were used unless noted: Adam optimizer, batch size 64, binary cross-entropy loss, up to 100 epochs with early stopping (patience 10) on validation loss, and checkpoint

selection by validation loss. Fine-tuning was performed by initializing from the decontaminated pretrained checkpoint and continuing full-parameter training on the PFAS training pool under an identical schedule.

Evaluation design

Two splitting regimes were constructed. Random splitting assigned molecules to training, validation and test sets at random. Scaffold splitting grouped molecules by Bemis-Murcko scaffold [15], because most PFAS are acyclic and thus lack a Murcko scaffold, acyclic molecules were grouped by heavy-atom count, a proxy for perfluoroalkyl chain length, such that no homologous series straddles the training/test boundary. This regime deliberately withholds entire chain-length series, mimicking the regulatory task of predicting untested PFAS chemistries. All key experiments were repeated with three random seeds (42, 0, 7) and are reported as mean ± standard deviation. Performance was quantified by area under the ROC curve (AUROC) and, given extreme class imbalance[16], area under the precision–recall curve (AUPRC), computed per assay on the held-out test molecules and averaged across assays.

Classical baseline

Random forests (scikit-learn; 300 trees, class-weighted) on 2,048-bit Morgan fingerprints (radius 2)[17, 18] were trained on identical training pools and splits.

Chemical-space-aware routing

Routing is governed by a structural rule implemented with RDKit. A molecule is classified as PFAS if any of three conditions holds: (i) it contains a perfluoroaryl ring-a carbocyclic ring of five or more atoms bearing fluorine substituents on at least all but one ring carbons; (ii) it contains three or more polyfluorinated carbons (any carbon with two or more fluorine neighbors, including hydrogen-bearing -CHF2- groups); or (iii) it contains at least one polyfluorinated carbon and its fluorine fraction (fluorine atoms per heavy atom) reaches 15%. Molecules failing all conditions are routed to the standard drug-oriented ADMET model. The fluorine-fraction gate was calibrated to separate perfluoroalkyl chemistries from organofluorine drugs bearing a single-CF3 group (fluorine fraction typically <14%). Sensitivity was measured on all 86 scaffold-held-out PFAS test molecules. Specificity was measured on the 368

approved drugs of DrugBank[19]: all 268 fluorinated approved drugs plus 100 randomly sampled non-fluorinated drugs (random seed 42). Flagged drugs were manually classified by structural inspection into chemically coherent groups (perfluorocarbon drugs, fluorinated inhalational anesthetics, and borderline fluorinated therapeutics).

ADME-Tox category aggregation

The 26 assays were mapped onto the ADMET taxonomy (full mapping and rationale in Table S1): Metabolism (4 assays: CYP2C9, CYP3A4, CYP2D6, ALDH1A1), Gene regulation (7), Cytotoxicity (3), Membrane targets (3), Protein interaction (3), Cell signaling (2), and Pathogen/DNA stress (4). Within-molecule category means yield a seven-axis ADME-Tox profile. Absorption and excretion are not represented in the PCBA panel and are complemented by the standard ADMET model in the routed workflow.

Platform deployment

The fine-tuned PFAS panel (scaffold- and random-split checkpoints as a two-model ensemble) was registered as a custom model directory (PFAS_Tox21_Panel) within ADMET-AI v2.0.1 [9]. Because ADMET-AI computes DrugBank percentiles for every predicted property[19], its post-processing was extended to skip panel tasks absent from the DrugBank reference; the patch is included in the released code. Predictions are available through the web interface and a command-line API.

Data and code availability

The code, decontaminated datasets and trained model weights are openly available at GitHub (https://github.com/zhanting-yang/PFAS-GNN) and archived with DOI 10.5281/zenodo.22869943 (Zenodo). The original Cheng–Ng dataset is distributed via MoleculeNet[5]. ADMET-AI is available at https://admet.ai.

Results

Pretraining–test leakage in the standard workflow

Canonical-structure matching between the PFAS panel and the large PCBA collection showed that the panel is a strict subset: 1,012 of 1,012 PFAS occur in the pretraining-eligible pool (Fig. 1). Any large-scale pretraining on PCBA followed by

PFAS evaluation therefore exposes the model to test molecules and their labels, either directly or through subsequent fine-tuning. All results below derive from a pretraining pool from which every panel compound was excluded.

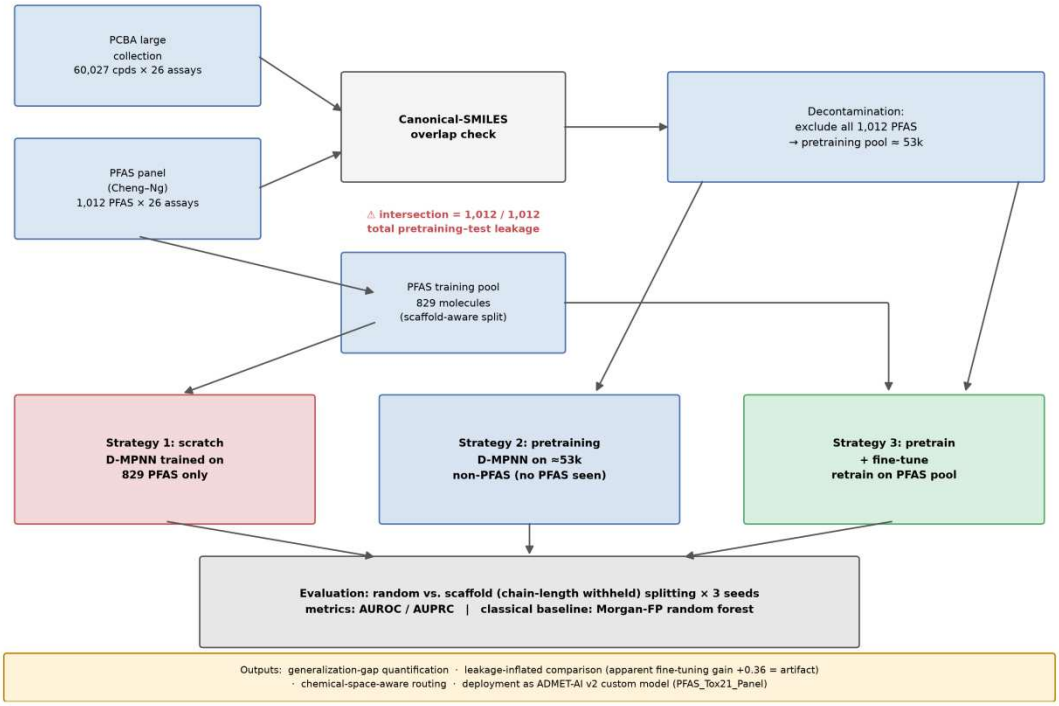


Fig. 1. Study workflow. The Cheng–Ng PFAS panel (1,012 PFAS × 26 bioassays) is a strict subset of the PCBA large collection (60,027 compounds), as revealed by canonical-SMILES overlap analysis (intersection 1,012/1,012); all PFAS were therefore excluded from the pretraining pool. Three learning strategies—from-scratch training on 829 PFAS, large-scale multi-task pretraining on ≈53k non-PFAS compounds, and pretraining followed by PFAS fine-tuning—were compared under random and scaffold (chain-length-withheld) splitting with three random seeds, against a Morgan-fingerprint random-forest baseline. Outputs include the quantification of a leakage-inflated fine-tuning artifact (+0.36 apparent gain under uncontrolled protocols), a chemical-space-aware routing rule, and deployment as a custom model (PFAS_Tox21_Panel) in ADMET-AI v2.

Scratch training exhibits a large generalization gap

Training the D-MPNN from scratch on the PFAS pool yielded a mean test AUROC of 0.751 under random splitting but only 0.607 ± 0.044 (three seeds) under the homolog-aware scaffold regime-a gap of ~0.14–0.20 attributable entirely to withheld chain-length series (Table 1, Fig.2). Training-dynamics analysis (training AUROC 0.761 vs. test AUROC 0.751 under random splitting; smoothly decreasing validation

loss) indicated underfitting rather than overfitting, identifying the information content of ~800 domain training molecules-not model capacity-as the limiting factor.

Table 1. Mean test AUROC of the four model conditions under the two splitting regimes.

Model	**Random split (mean test AUROC)**	**Scaffold split (mean test AUROC, +/- SD, 3 seeds)**
GNN (D-MPNN) from scratch	0.751	0.607 +/- 0.044
Random forest (Morgan FP, class-weighted)	0.760	0.769 (single seed 42)
GNN pretrained on ~53k compounds (no PFAS seen)	0.804	0.798
GNN pretrained + PFAS fine-tuned (proposed)	0.864	0.812 +/- 0.032

Notes: Values are mean test AUROC over the 26 bioassays. GNN results under the scaffold regime are reported as mean ± SD over three random seeds (42, 0, 7); the random forest was evaluated with a single seed (42). The "pretrained" model was trained on the decontaminated PCBA pool (≈53k compounds) from which all 1,012 PFAS were excluded; it therefore never encountered any PFAS molecule. For reference, without decontamination the scaffold-split values inflate to 0.917 (pretrained) and 0.917 (pretrained + fine-tuned).

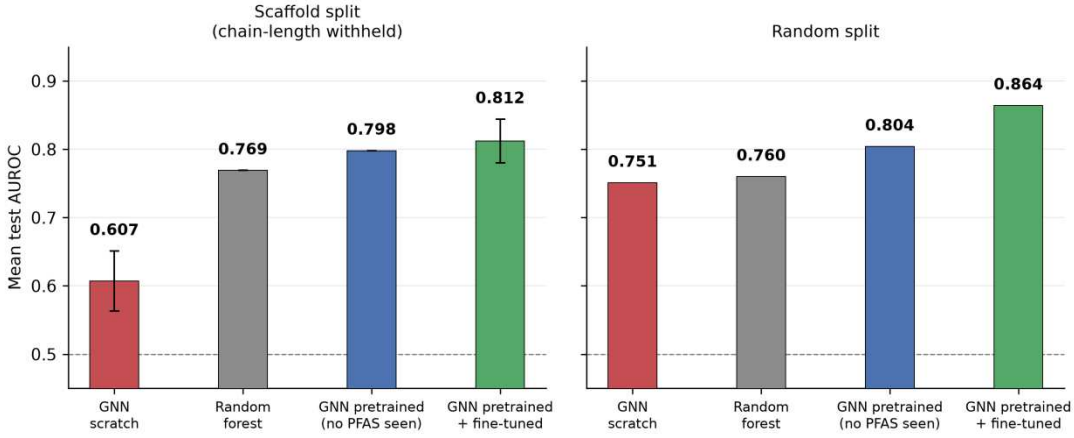


Fig. 2. Mean test AUROC of the four model conditions under the two splitting regimes. (A) Scaffold split with homologous chain-length series withheld (GNN: mean ± SD over three seeds). (B) Random split. The dashed line marks chance level (0.5). Scratch-trained GNNs exhibit a large generalization gap and fall below the random-forest baseline on the scaffold split; the pretrained model—despite never seeing PFAS—nearly closes the gap, and fine-tuning adds gains only under random splitting.

A pretrained model that has never seen PFAS transfers across chemical domains

A single multi-task D-MPNN pretrained on the decontaminated PCBA pool (no PFAS present) achieved a scaffold-test AUROC of 0.798 on PFAS it had never encountered,

a gain of ~0.19 over scratch training, and essentially eliminated the split sensitivity of performance (0.804 random vs. 0.798 scaffold). These results indicate that large-scale multi-task learning captures structure–activity regularities-plausibly the electrostatic and lipophilic features associated with CYP inhibition and related endpoints-that transfer across chemical domains, in line with the view that scaffold identity, rather than domain identity per se, governs generalization in molecular machine learning [8].

Fine-tuning gains are conditional-and can be fabricated by leakage

Fine-tuning the pretrained model on the PFAS pool improved performance under random splitting (0.864 vs. 0.804) but not on the scaffold split (0.812 ± 0.032 vs. 0.798), where the three fine-tuned seeds (0.793, 0.793, 0.849) never exceeded the pretrained baseline beyond seed noise (Table 1). The gain from domain fine-tuning is therefore conditional on test compounds falling within the chemical space covered by the domain training set: positive for PFAS similar to known examples, negligible for genuinely novel chain lengths. Critically, had the leakage documented in third finding not been controlled, the same experiment would have suggested a spurious fine-tuning gain of up to +0.36 (0.556 to 0.917)-an instructive demonstration of how benchmark composition can fabricate conclusions [20].

A strong, split-insensitive classical baseline

Morgan-fingerprint random forests with class weighting were essentially insensitive to the splitting regime (0.769 scaffold vs. 0.760 random) and outperformed the scratch-trained graph model on the scaffold split (Table 1). Two implications follow: fingerprint-based models remain a competitive and calibration-friendly baseline for small, imbalanced toxicology datasets and must not be omitted from comparisons; and the advantage of graph neural networks in this setting materializes only after pretraining. At the per-assay level, pretraining lifts the weakest assays most (e.g., PCBA-504444: 0.10 to 0.92-0.93; Fig. 3, Table S2 and S3).

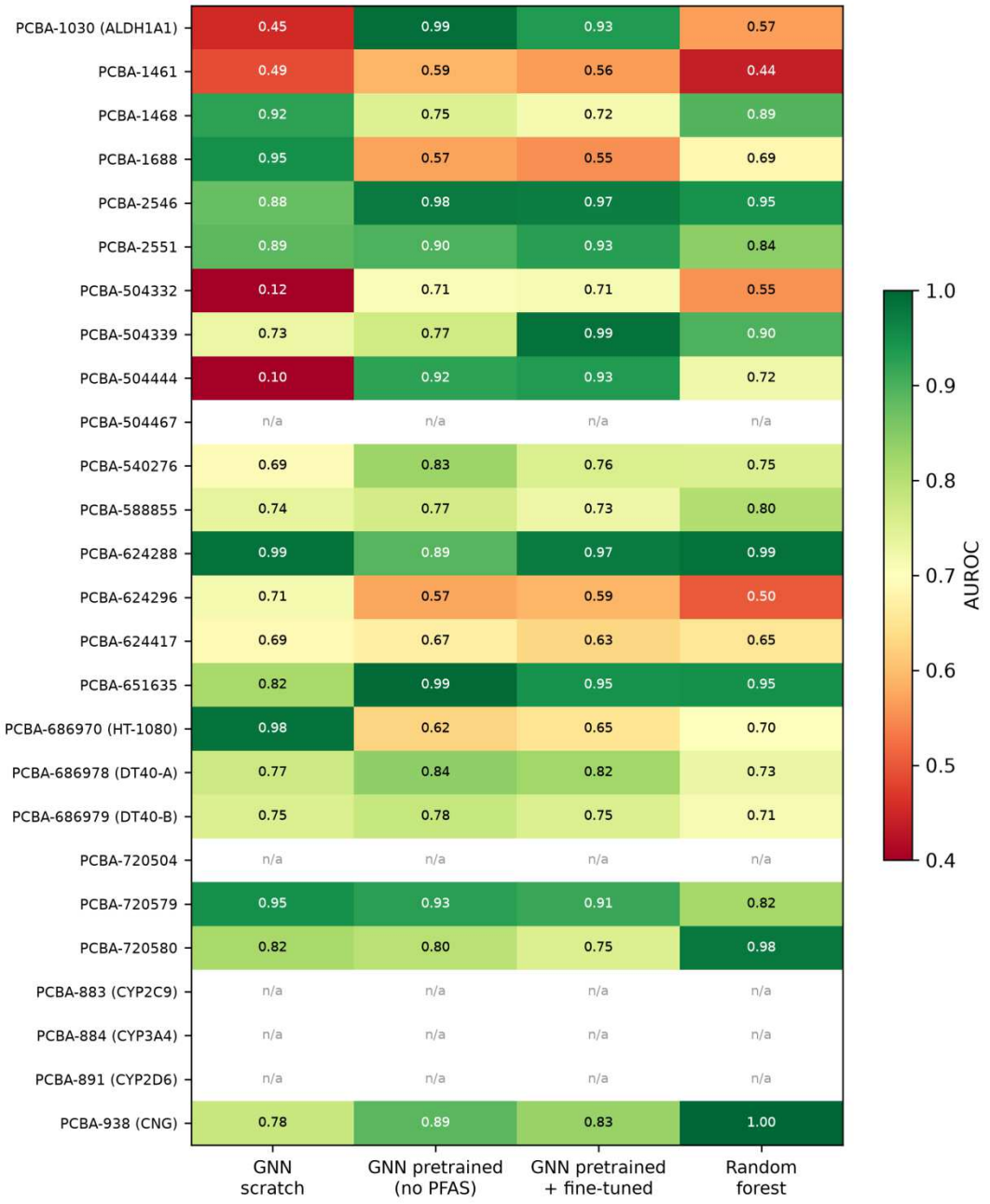


Fig. 3. Per-assay test AUROC (scaffold split, representative checkpoints). Heatmap of 26 bioassays (rows; targets in parentheses where applicable) × four model conditions (columns), evaluated on 86 held-out PFAS. Cells show AUROC; "n/a" denotes assays with no tested molecules in this held-out set, reflecting the sparse assay coverage of the panel. Pretraining lifts the weakest assays most (e.g., PCBA-504444: 0.10 → 0.92–0.93).

Routing validation and the fuzzy drug–PFAS boundary

An initial, stricter rule-requiring hydrogen-free -CF3/-CF2- carbons and a 25% fluorine fraction-detected only 29 of 86 held-out test PFAS (33.7%), revealing that the panel's chemical space extends well beyond classical perfluoroalkyl chains to polyfluorinated (-CHF2--bearing) and perfluoroaryl compounds. Broadening carbon

matching to hydrogen-bearing polyfluorinated carbons and adding the perfluoroaryl clause (final rule) raised sensitivity more than twofold. The routing rule classified 82 of 86 held-out PFAS test molecules as PFAS (sensitivity 95.3%). The four missed compounds are weakly fluorinated members of the panel—two bearing a single -CF3 group in large scaffolds (fluorine fraction ≈ 0.11) and two bearing trifluorophenyl or twin -CF3 groups at fluorine fractions of 0.14-i.e., compounds that sit at the definitional frontier between polyfluorinated chemicals and fluorinated drugs.

On the drug side, the rule flagged 41 of 368 DrugBank-approved drugs (specificity 88.9%; the test set comprised all 268 fluorinated approved drugs plus 100 randomly sampled non-fluorinated drugs, none of the latter flagged). Crucially, manual inspection of the 41 flagged drugs showed that they are not random errors but fall into three chemically coherent groups (Table S4): (i) bona fide perfluorocarbon drugs-perflubutane, perflexane, perflenapent, perflutren, perfluorohexyloctane and pentafluoropropane-which are perfluorinated by any structural definition and for which routing to a PFAS panel is arguably correct; (ii) fluorinated inhalational anesthetics (desflurane, enflurane, isoflurane, sevoflurane, methoxyflurane, norflurane, flurothyl, halothane), which are heavily fluorinated small ethers/halocarbons occupying PFAS-adjacent chemical space; and (iii) 27 heavily fluorinated therapeutic agents (e.g., mefloquine, sitagliptin, dutasteride, aprepitant) that constitute genuine borderline cases. Excluding the chemically defensible groups (i) and (ii), the effective error rate falls to 27 of 368 (92.7%). These results indicate that the apparent false-positive set is, to a substantial degree, a map of the real but poorly documented overlap between the PFAS and approved-drug chemical spaces, rather than noise (Fig. 4).

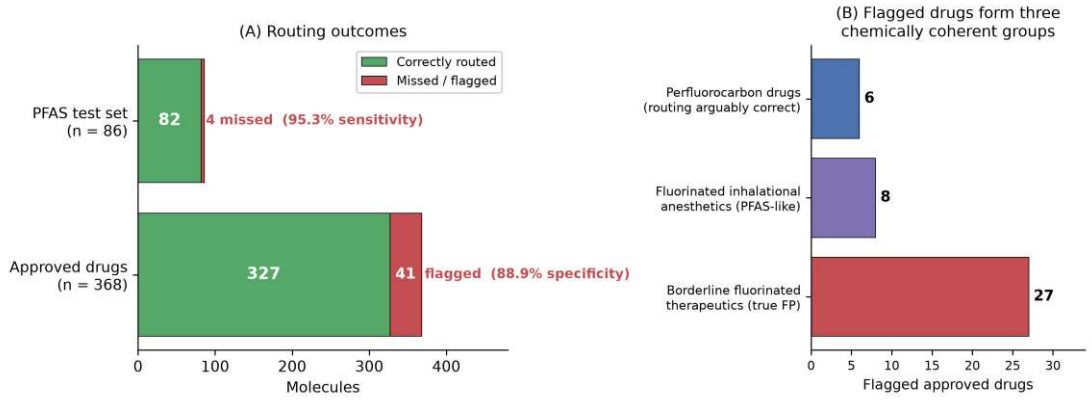


Fig. 4. Validation of the chemical-space-aware routing rule. (A) Routing outcomes on 86 held-out PFAS (sensitivity 95.3%; four missed compounds are weakly fluorinated members at the PFAS definitional frontier) and 368 approved drugs (specificity 88.9%). (B) The 41 flagged approved drugs form three chemically coherent groups: six bona fide perfluorocarbon drugs (routing arguably correct), eight fluorinated inhalational anesthetics with PFAS-like chemistries, and 27 borderline fluorinated therapeutics; excluding the first two groups, the effective error rate is 27/368 (92.7%).

Category-level ADME-Tox profiles

Aggregated seven-axis profiles placed CYP-mediated Metabolism as the dominant axis for both PFOA and PFOS-neither of which occurs in the training panel-with category means of 0.24 and 0.25 respectively and CYP3A4 the single highest endpoint (0.49 for both); Cytotoxicity was the second axis (0.20-0.22), and all remaining categories fell below 0.12. By contrast, the conventional drug-oriented ADMET summary for the same compounds showed no deviation from typical drug behavior (Fig. 5). A non-fluorinated drug (metoprolol) produced a diffuse, mid-level profile without category focus (0.21-0.53 across axes; its Metabolism mean of 0.38 reflects the broad enzymatic promiscuity of a marketed drug rather than a focused signal). The value of the category profile lies in rendering PFAS-relevant bioactivity legible in ADME-Tox vocabulary rather than in discriminating PFAS from drugs-which is the routing layer's role; absolute probabilities for out-of-domain molecules are non-informative, and the panel is intended for applicability-domain-limited use. The full side-by-side comparison of the two radar formats for PFOA is provided in Fig. S1.

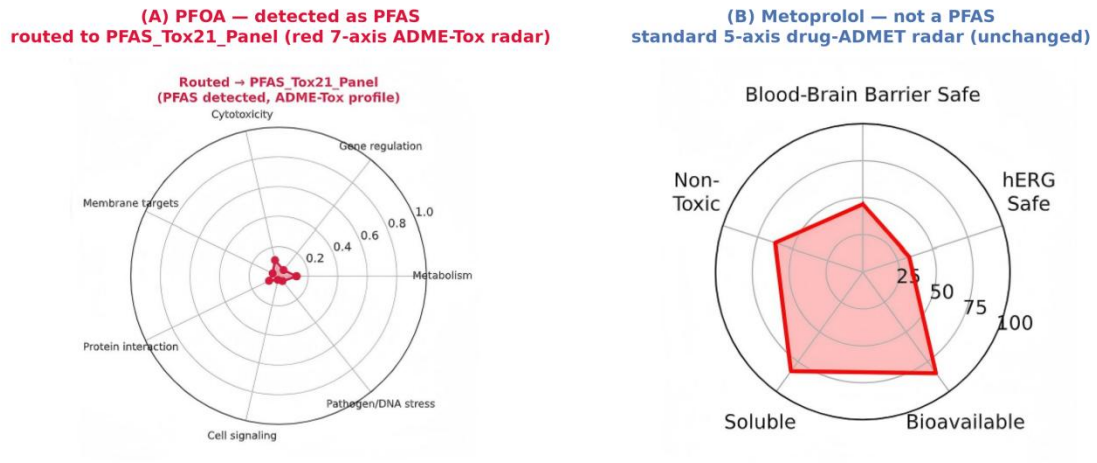


Fig. 5. Chemical-space-aware routing in action on the deployed ADMET-AI v2 platform. Identical input procedure, two molecules. (A) PFOA is detected as PFAS by the structural rule (perfluoroalkyl chain; fluorine fraction ≈ 0.88) and automatically routed to the custom PFAS_Tox21_Panel: the result card displays a red seven-axis ADME-Tox profile (CYP-mediated Metabolism dominant) instead of the default summary. (B) Metoprolol, a non-fluorinated drug, fails all PFAS conditions and receives the standard five-axis drug-oriented ADMET radar. No manual model selection is required; the routing decision is made per molecule at prediction time.

Platform demonstration

The deployed panel returns 26 endpoint probabilities for arbitrary SMILES inputs through both the web interface and the command-line API. PFOA and PFOS receive their highest panel probabilities on the CYP3A4 inhibition assay (0.49 for both), consistent with the prominent CYP-inhibitory activity of PFAS in the underlying screening data, and the full 130-column output (51 standard ADMET properties, 52 DrugBank percentiles, 26 panel endpoints) is downloadable per query.

Discussion

PFAS have become a defining environmental-health challenge of this decade: they are produced and used at large scale[21, 22]. Toxicologically, epidemiological and experimental evidence has linked PFAS exposure to immune, thyroid, hepatic, metabolic, reproductive and developmental effects[23, 24] , yet this evidence remains concentrated in a handful of legacy compounds - PFOA and PFOS above all - while inventories list thousands of PFAS in commerce, the overwhelming majority without any toxicity data[2, 4]. Closing this gap experimentally, compound by compound, is not feasible; computational prioritization is therefore not a convenience but a

necessity, which explains the rapid growth of machine-learning models for PFAS bioactivity in recent years[5, 25].

We build upon ADMET-AI to construct a toxicity assessment platform tailored for PFAS contaminants. Under homolog-aware evaluation, a directed message-passing neural network trained on ~800 PFAS alone generalizes poorly to withheld chain-length series (0.607 ± 0.044), below even a classical Morgan-fingerprint random forest (0.769); a single multi-task model pretrained on ~53,000 non-PFAS compounds-without ever seeing a PFAS-lifts performance to 0.798; and PFAS fine-tuning contributes a further, strictly conditional gain (0.864 vs 0.804 under random splitting; 0.812 ± 0.032 vs 0.798 on novel chain lengths). The chemical-space-aware routing rule built on these findings operates at 95.3% sensitivity and 88.9% specificity on approved drugs, and its deployment in ADMET-AI v2 delivers the panel through a single interface (Fig. 5).

The dominance of pretraining over domain data warrants mechanistic interpretation. We attribute the ~0.19 improvement over scratch training to the breadth of structure–activity regularities acquired across 53,000 compounds and 26 endpoints-regularities that evidently include the electrostatic and lipophilic features underlying CYP inhibition and related readouts, and that therefore transfer across chemical domains. First, the pretrained model is essentially split-insensitive (0.804 random vs 0.798 scaffold), indicating that its knowledge is organized by scaffold-like structural features rather than by dataset-specific idiosyncrasies, consistent with the view that scaffold identity dominates generalization in molecular machine learning[8]. Second, the random-forest baseline-whose features (substructure counts) and inductive bias (axis-aligned partitioning) are far simpler-achieves 0.769 on the scaffold split, outperforming the from-scratch graph model (0.607 ± 0.044) and nearly matching the pretrained one, this ordering demonstrates that the advantage of graph neural networks in this data regime is not architectural but contingent on pretraining, and that omitting such baselines would have overstated the contribution of deep learning. Third, the failure of fine-tuning to improve on novel chain lengths—despite 829 in-domain training molecules—quantifies a general principle that is easy to state

and hard to see without leakage control: the incremental value of domain data is bounded by its chemical-space coverage, and for homolog-dominated chemistry a small domain set teaches the pretrained model little beyond what it already knows. This directly refines recent transfer-learning successes in PFAS modeling [25], which report strong performance but, without overlap audits or homolog-aware splits, cannot separate transferable pretraining signal from domain memorization—an ambiguity our protocol resolves.

The leakage analysis is, in itself, a result worth dwelling on. The PFAS panel is not partially but totally contained in its parent screening collection (1,012/1,012 compounds), so any pretraining on that collection without exclusion leaks the test set in full. The consequences are quantitatively striking: the apparent fine-tuning gain inflates from zero (0.798 to 0.812) to +0.36 (0.556 to 0.917), an artifact large enough to reverse the substantive conclusion of a study. That the naive and decontaminated fine-tuned models even coincide at 0.917 underscores that leaked pretraining saturates the task. Because domain datasets are routinely extracted from public screening collections-and because the leakage is silent, total, and transmitted through fine-tuning even when the domain training split itself is clean-we expect this pattern to generalize well beyond PFAS, to target-family subsets of ChEMBL, pathway subsets of PubChem, and analogous settings in other domains[20]. The two safeguards we propose (canonical-overlap audit before any transfer-learning experiment; homolog-aware splitting for series-dominated chemistry) cost minutes and are released as one-command scripts, and we suggest they become reporting standards.

The routing validation produced a second, unanticipated finding: the boundary between “fluorinated drugs” and “PFAS” is not where database labels place it. Six of the 41 flagged drugs are perfluorocarbons administered as medicines (ultrasound contrast agents, ophthalmic vehicles), and eight are fluorinated inhalational anesthetics whose chemistries overlap the perfluoroether class; only 27 are conventional therapeutics that a domain purist would call false positives. Label-based workflows-treating “approved drug” and “PFAS” as disjoint sets-therefore miss real

chemistry, and the routing rule's residual “errors” double as a map of a genuinely undocumented region of chemical space. This observation also cautions against interpreting the specificity of 88.9% at face value: against chemistry rather than labels, the effective error rate is 27/368 (92.7%). More broadly, the routing pattern-cheap structural arbitration between expert models with explicitly quantified boundaries-is, in our view, a more honest architecture than a single monolithic model as prediction platforms extend beyond drug-likeness, whether to peptides, organometallics, or the multi-component mixtures encountered in traditional medicines.

Therefore, our research delivers an openly deployed prediction capability (ADMET-AI v2 with the PFAS_Tox21_Panel submodel, code and weights archived at DOI 10.5281/zenodo.22869943) that can screen inventoried PFAS at negligible cost, with applicability-domain routing that prevents silent misuse[26]. Methodologically, it argues that two near-zero-cost practices-canonical-overlap audits before any transfer-learning experiment, and homolog-aware splitting for any chemistry dominated by homologous series-should become standard in environmental cheminformatics, and our released scripts make both one-command procedures. Conceptually, routing reframes the one-model paradigm: as prediction platforms extend beyond drug-likeness to peptides, organometallics, natural-product chemotypes, or the multi-component mixtures of traditional medicines, explicit chemical-space arbitration with quantified boundaries is a more honest architecture than a single monolithic model, and the boundary analysis reported here - 95.3% sensitivity, 88.9% specificity, and a chemically coherent residual error set - provides a template for how such boundaries should be documented.

Conclusions

This study shows that for PFAS bioactivity prediction, large-scale multi-task pretraining-not PFAS-specific data is the primary driver of generalizable performance, while the incremental value of domain fine-tuning is bounded by the chemical-space coverage of the domain training set. Under homolog-aware evaluation, a pretrained model that has never encountered PFAS clearly outperforms from-scratch training (0.798 vs. 0.607 AUROC on withheld chain-length series) and matches the best

fine-tuned models on genuinely novel chemistries, with a classical random-forest baseline remaining a strong and under-appreciated reference point. Our results further demonstrate that uncontrolled transfer-learning benchmarks can fabricate apparent gains: because the PFAS panel is entirely contained in its parent screening collection, undetected leakage inflated the apparent fine-tuning benefit to +0.36 AUROC, a magnitude sufficient to reverse substantive conclusions. Chemical-space-aware routing, validated here at 95.3% sensitivity and 88.9% specificity, converts platform blindness to PFAS-relevant chemistry into an explicit, deployable design feature, and its residual errors reveal a real, undocumented overlap between the approved-drug and PFAS chemical spaces. Together with the openly archived code, decontaminated data, and the ADMET-AI-integrated prediction panel, these findings provide both a practical prioritization tool for data-poor PFAS and a reusable methodology-leakage audits and homolog-aware splitting-for evaluating transfer learning in environmental toxicology and beyond.

Acknowledgements

We thank Bingfeng Seng for technical support in software implementation and computational pipeline construction.

Author contributions

Z.Y. and G.B. contributed equally to this work: both participated in study conceptualization, data acquisition, analysis and interpretation, and in drafting and critically revising the manuscript. J.P. developed the prediction software and performed formal analysis. X.H. collected and curated the datasets. All authors read and approved the final manuscript.

Funding

This work was supported by 2025 Postgraduate Supervisor Innovation Team Construction Project of Qinghai University (Public Health and Population Health Team, No.qddt-2511) and Natural Science Foundation of Qinghai Province (NO.2024-ZJ-979).

Availability of data and materials

The code, decontaminated data, and model weights are freely available at https://github.com/zhanting-yang/PFAS-GNN (DOI: 10.5281/zenodo.22869943).

Declarations

Competing interests: The authors declare no competing interests

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

References

1. Gluge J, Scheringer M, Cousins IT et al (2020) An overview of the uses of per- and polyfluoroalkyl substances (PFAS). Environ Sci Process Impacts 22:2345-2373
2. Organisation for Economic Co-operation and Development (OECD) (2018) Toward a New Comprehensive Global Database of Per- and Polyfluoroalkyl Substances (PFASs). Paris. Report No. ENV-JM-MONO(2018)7
3. Lau C, Anitole K, Hodes C et al (2007) Perfluoroalkyl acids: a review of monitoring and toxicological findings. Toxicol Sci 99:366-394
4. Fenton SE, 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. Environ Toxicol Chem 40:606-630
5. Cheng W, Ng CA (2019) Using Machine Learning to Classify Bioactivity for 3486 Per- and Polyfluoroalkyl Substances (PFASs) from the OECD List. Environ Sci Technol 53:13970-13980
6. Wu Z, Ramsundar B, Feinberg EN et al (2018) MoleculeNet: a benchmark for molecular machine learning. Chem Sci 9:513-530
7. Wang Y, Suzek T, Zhang J et al (2014) PubChem BioAssay: 2014 update. Nucleic Acids Res 42:D1075-1082
8. Wallach I, Heifets A (2018) Most Ligand-Based Classification Benchmarks Reward Memorization Rather than Generalization. J Chem Inf Model 58:916-932
9. Swanson K, Walther P, Leitz J et al (2024) ADMET-AI: a machine learning ADMET platform for evaluation of large-scale chemical libraries. Bioinformatics 40:btae416
10. Xiong G, Wu Z, Yi J et al (2021) ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties. Nucleic Acids Res 49:W5-W14
11. Gaudelet T, Day B, Jamasb AR et al (2021) Utilizing graph machine learning within drug discovery and development. Brief Bioinform 22:bbab159
12. Dix DJ, Houck KA, Martin MT et al (2007) The ToxCast program for prioritizing toxicity testing of environmental chemicals. Toxicol Sci 95:5-12
13. Yang K, Swanson K, Jin W et al (2019) Analyzing Learned Molecular Representations for Property Prediction. J Chem Inf Model 59:3370-3388
14. Heid E, Greenman KP, Chung Y et al (2024) Chemprop: A Machine Learning Package for Chemical Property Prediction. J Chem Inf Model 64:9-17
15. Bemis GW, Murcko MA (1996) The properties of known drugs. 1. Molecular frameworks. J Med Chem 39:2887-2893

16. Fernández A, Galar M, García S et al (2018) Learning from Imbalanced Data Sets: Springer, Cham
17. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. J Chem Inf Model 50:742-754
18. Breiman L (2001) Random forests. Machine Learning 45:5-32
19. Wishart DS, Feunang YD, Guo AC et al (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res 46:D1074-D1082
20. Kapoor S, Narayanan A (2023) Leakage and the reproducibility crisis in machine-learning-based science. Patterns 4:100804
21. Sunderland EM, Hu XC, Dassuncao C et al (2019) A review of the pathways of human exposure to poly- and perfluoroalkyl substances (PFASs) and present understanding of health effects. J Expo Sci Environ Epidemiol 29:131-147
22. Kwiatkowski CF, Andrews DQ, Birnbaum LS et al (2020) Scientific Basis for Managing PFAS as a Chemical Class. Environ Sci Technol Lett 7:532-543
23. Costello E, Rock S, Stratakis N et al (2022) Exposure to per- and Polyfluoroalkyl Substances and Markers of Liver Injury: A Systematic Review and Meta-Analysis. Environ Health Perspect 130:046001
24. Steenland K, Winquist A (2021) PFAS and cancer, a scoping review of the epidemiologic evidence. Environ Res 194:110690
25. Wang R, Yi S, Shan G et al (2025) Integration of Transfer Learning and Multitask Learning To Predict the Potential of Per/Polyfluoroalkyl Substances in Activating Multiple Nuclear Receptors Associated with Hepatic Lipotoxicity. Environ Sci Technol 59:23727-23738
26. Netzeva TI, Worth A, Aldenberg T et al (2005) Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52. Altern Lab Anim 33:155-173

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- FigS1blindspot.png
- AdditionalfileTableS1.xlsx
- AdditionalfileTableS2.xlsx
- AdditionalfileTableS3.xlsx
- AdditionalfileTableS4.xlsx