Journal of Chemical&Engineering Data

pubs.acs.org/jced Article

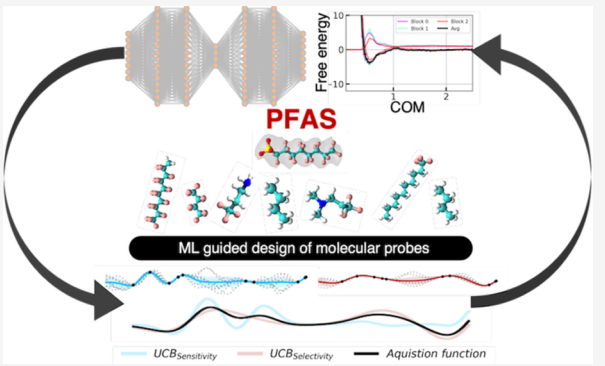
Data-Driven Discovery of Linear Molecular Probes with Optimal Selective Affinity for PFAS in Water

Siva Dasetty,[○] Maximilian Topel,[○] Yifeng Oliver Tang, Yuqin Wang, Eric Jonas, Seth B. Darling, Junhong Chen, and Andrew L. Ferguson*

Cite This: https://doi.org/10.1021/acs.jced.3c00404 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Approaches to tackle the wide and growing variety of highly persistent per- and polyfluoroalkyl substances (PFAS) are of pressing global need because of their detrimental human health effects, such as cancer, birth defects, and hormone imbalance. Sensitive, selective, and easy-to-use real-time sensors to monitor and detect PFAS and sorbents to extract them are critical to meeting government-mandated environmental concentrations. In this work, we combine all-atom molecular dynamics simulations, enhanced sampling, deep representational learning, and Bayesian optimization to perform high-throughput virtual screening for highly sensitive and selective molecular probes. Our molecular design space consists of 3850 linear hydrocarbon chains with varying degrees of halogenation with and without amine- and phosphine-based headgroups. By employing a data-driven search process, we efficiently explore the molecular design space to optimize the sensitivity to perfluorooctanesulfonic acid (PFOS) as a prototypical PFAS analyte and selectivity relative to a sodium dodecyl sulfate (SDS) interferent. We calculate 504 Gibbs free energies of probe-analyte and probe-interferent interactions and identify probes with PFOS association free energies of up to $(-\Delta G_{PFOS}) = 9.8 \pm 0.2$ kJ/mol and selectivities relative to SDS of $(-\Delta\Delta G_{PFOS-SDS}) = 3.1 \pm 1.5$ kJ/mol. A $C_{11}Br_{23}P(CH_3)_2$ probe containing 11 backbone brominated carbons and a tertiary phosphine headgroup possesses the most sensitive binding constant to PFOS within the defined search space of $K_b^{PFOS} = 177.4 \pm 12.7$, and a semibrominated probe $C_5H_{11}C_7Br_{14}N(CH_3)_2$ containing 12 backbone carbons and a tertiary amine headgroup possesses the highest selectivity relative to SDS of $K_b^{PFOS}/K_b^{SDS} = 4.6 \pm 1.7$. A retrospective analysis of our data to extract interpretable design rules reveals that the sensitivity of linear hydrogenated probes increases by approximately 1 kJ/mol per C−C bond. The addition or removal of halogen atoms and amine or phosphine headgroups produces nonmonotonic changes in both sensitivity and selectivity with changes to the sensitivity of up to 2.5 kJ/mol. This work places empirical limitations on the performance of a wide range of linear probes for PFOS detection and offers a generic strategy for high-throughput computational screening to promote selective and sensitive binding.



1. INTRODUCTION

Per- and polyfluoroalkyl substances (PFAS) represent a large group of synthetic chemicals defined by the U.S. Environmental Protection Agency (EPA) as compounds containing R−(CF_2)−C(F)(R′)R″ units, where CF_2 and CF groups are saturated carbons and the R groups (R, R′, or R″) are not hydrogens.[1,2] Owing to their excellent properties including heat resistance, hydrophobicity, and oleophobicity, PFAS have been extensively used in numerous commercial applications such as nonstick cookware, firefighting foam and flame retardants, stain-resistant fabrics, and fast food packaging.[3−6] PFAS compounds enter the environment during the production, use, or waste of these consumer products and, due to the highly stable carbon−fluorine covalent bonds—one of the strongest single bonds with a dissociation energy of ∼440 kJ/mol due to the high electronegativity of fluorine[7]—are extremely persistent, resulting in PFAS accumulation in both the environment and living organisms and leading to them often being referred to as "forever chemicals."[8−11] PFAS exposure can cause adverse health effects in humans, such as increased cholesterol levels, birth defects, disruption in thyroid hormone balance, and potentially a high risk of cancer.[12] Owing to these detrimental properties, PFAS attracted broad interest from various regulatory agencies worldwide.[13] It is one of the pressing global challenges to address the growing quantities of PFAS in the environment, especially in water resources from which these molecules are readily disseminated

Received: June 28, 2023
Accepted: October 6, 2023

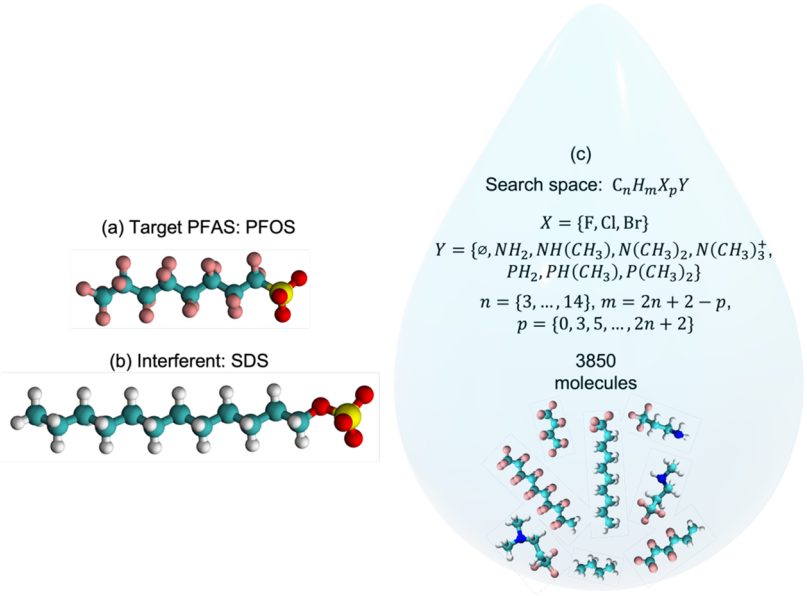


Figure 1. Illustrations of the target PFAS analyte, interferent, and molecular search space over which the probes with optimal sensitivity and selectivity are discovered. (a) Three-dimensional (3D) structure of the target PFAS variant (perfluorooctanesulfonic acid, PFOS), (b) 3D structure of the template interferent (sodium dodecyl sulfate, SDS), and (c) search space comprising linear hydrogenated or halogenated chains of 3−14 carbons with and without amine- or phosphine-based headgroups. All semihalogenated probes considered contain contiguous blocks of halogenated methylene groups for experimental characterizability. We explored primary, secondary, tertiary, and quaternary states for the amine-based headgroup and the primary, secondary, and tertiary states of the phosphine headgroups. In total, our search space comprises 3850 molecules, over which we employ our computational active learning and high-throughput screening approach to discover optimal molecular probes. The 3D structures are rendered using VMD,[25] where carbon, hydrogen, oxygen, nitrogen, sulfur, and fluorine are colored in cyan, white, red, blue, yellow, and pink, respectively.

to living organisms.[7,8,10] This requires an approach to effectively and efficiently detect and eliminate the growing number (>4700 as per the EPA[14]) of PFAS variants and to do so at or below the very dilute concentrations of approximately 4 ng/L mandated by EPA guidelines for safe drinking water.[15]

Current standard approaches available for regulatory or guidance activities of PFAS in water sources by the EPA utilize solid-phase extraction (SPE)- and liquid chromatography/tandem mass spectrometry (LC-MS/MS)-based analytical tools.[7,16] These approaches—formally referred to as Method 537.1 and 533—can measure 18 and 25 types of PFAS, respectively, with a common total of 29 PFAS molecules in potable water sources.[17,18] The lowest concentration minimum reporting level (LCMR) for Method 537.1 and 533 ranges from 0.53 to 6.3 ng/L and 1.4−16 ng/L, respectively, depending on the PFAS variant.[17,18] Both methods offer LCMR well below the established health advisory levels −70 ng/L (70 ppt) for lifetime exposure set by the USEPA in 2016 for the most commonly used and studied PFAS—perfluorooctanoic acid (PFOA) and perfluorooctanesulfonic acid (PFOS). However, the proposed legal maximum contaminant levels (MCL) and nonenforceable maximum contaminant level goals (MCLG) of PFAS in drinking water by EPA in early 2023[15] for both PFOA and PFOS (MCL = 4 ng/L and MCLG = 0 ng/L) are approaching or below the limits of LC-MS/MS. In addition, LC-MS/MS requires expensive instruments intended for skilled analysts with complex and time-consuming procedures for sample preparation, data collection, and interpretation.[7] In contrast to LC-MS/MS-based tools, portable sensors engineered with molecular probes for detecting PFAS are a promising alternative to standard methods with potential advantages such as lower cost, simpler operation, and real-time analysis.[7,19,20] However, given the vast and growing number of PFAS variants, it is a challenge to design molecular probes that can sensitively and selectively detect PFAS molecules in water.[7]

In this work, we engage this challenge by developing a high-throughput machine learning (ML)-guided computational screening protocol to efficiently navigate the chemical design space to find probes with optimal sensitivity and selectivity. We target perfluorooctanesulfonic acid (PFOS) as one of the major PFAS molecules of concern due to its wide initial production and adverse effects such as low birth weight, tissue damage, and cancer.[21−23] PFOS is a linear surfactant molecule with eight fluorinated carbon groups and a sulfonate headgroup (Figure 1a) that remains in an anionic state at the pH values in typical environmental water courses.[7,22,24] While this work focuses on PFOS, our approach is equally applicable to any PFAS variant. The particular choice of PFOS is made because it is one of the earliest PFAS known for its toxic properties.[21] The fluorinated PFOS tail is hydrophobic while the anionic sulfonate headgroup is hydrophilic and elevates solubility in water.[22] We search for optimal PFOS probes over a design space consisting of 3850 linear hydrogenated and halogenated molecules of length ranging from 3 to 14 carbons with and without an amine- or phosphine-based headgroup (Figure 1c). We are motivated to define our design space as such based on the synthetic accessibility of these molecules and their similar chemical nature to PFOS, wherein we can exploit hydrophobic binding to the aliphatic tail and potentially favorable electrostatic interactions with a headgroup. For synthesizability reasons, we consider only semihalogenated probes wherein the halogenated groups form contiguous blocks of fluorinated/chlori-

https://doi.org/10.1021/acs.jced.3c00404
J. Chem. Eng. Data XXXX, XXX, XXX−XXX

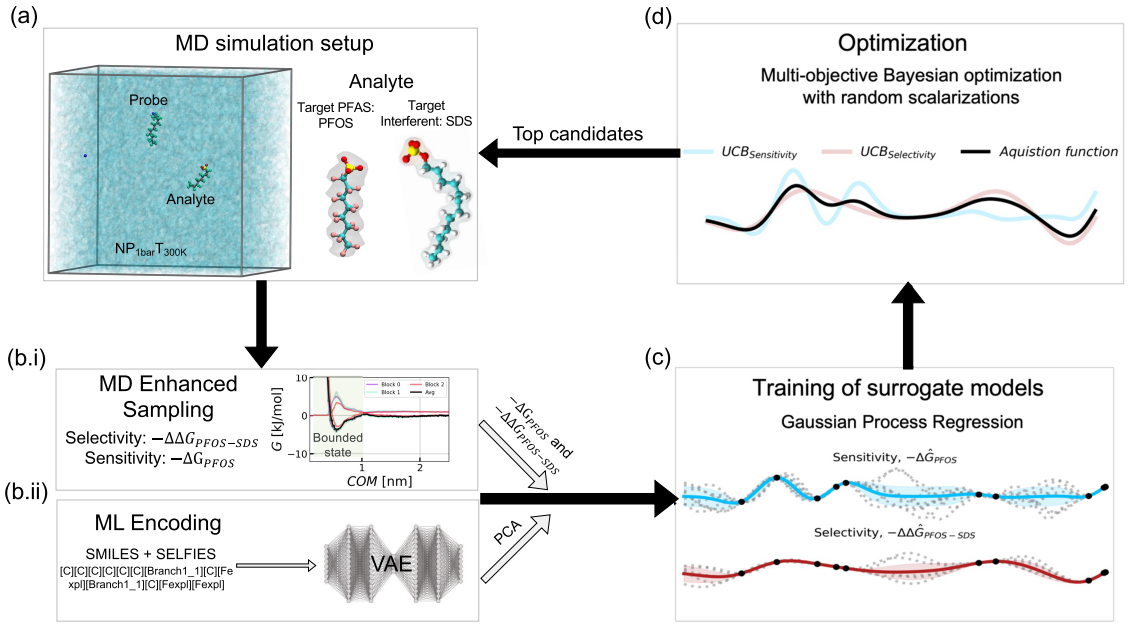


Figure 2. Illustration of the computational active learning approach employed to identify linear molecular probes with optimal sensitivity and selectivity. (a) Set up of all-atom molecular dynamics simulations and enhanced sampling using parallel bias metadynamics to estimate sensitivity ($-\Delta G_{\mathrm{PFOS}}$) and selectivity ($-\Delta\Delta G_{\mathrm{PFOS-SDS}}$), (b) embedding sensitivity and selectivity information generated through (b.i) the results of these enhanced sampling simulations and (b.ii) an encoding of molecular probes into a smooth low-dimensional latent space using a pretrained variational autoencoder, (c) construction of Gaussian process regression (GPR) surrogate models over the latent space to predict probe sensitivity and selectivity, and (d) multiobjective Bayesian optimization (BO) using random scalarizations to guide selection of the next molecular probes within the design space for molecular simulation.

nated/brominated methylenes. While some of these probe molecules show a strong resemblance to PFAS, they are designed to be tethered to surfaces and deployed at such small scales as not to represent any significant threat to the PFAS contamination challenge that they are designed to help solve in their potential future uses within portable sensors.

Having specified PFOS as our target analyte, we can compute the sensitivity of candidate molecular probes in our design space by measuring the PFOS-probe binding free energy, ΔG_{PFOS}. Additionally, we adopted sodium dodecyl sulfate (SDS) as an interferent molecule representative of the diversity of other potential absorbates that may be present in a water sample (Figure 1b). This allows us to quantify the selectivity of candidate probes via the relative PFOS-probe and SDS-probe binding free energies, $\Delta\Delta G_{\mathrm{PFOS-SDS}} = \Delta G_{\mathrm{PFOS}} - \Delta G_{\mathrm{SDS}}$. It is computationally intractable to measure the selectivity of a candidate probe against all possible competing interferents, so we selected SDS as a representative interferent that is often found in waterways with PFAS. The chemical similarity of PFOS and SDS means that it is expected to be a challenging task for a molecular probe to discriminate the PFOS analyte from the SDS interferent based on differential thermodynamic binding affinity.

It is the goal of our computational search to simultaneously minimize ΔG_{PFOS} and $\Delta\Delta G_{\mathrm{PFOS-SDS}}$ or, equivalently, simultaneously maximize $(-\Delta G_{\mathrm{PFOS}})$ and $(-\Delta\Delta G_{\mathrm{PFOS-SDS}})$ to discover highly sensitive and selective probe candidates for experimental testing. To achieve this, we employ a computational active learning approach involving deep representational learning of the molecular probes, training of surrogate Gaussian process regression models, and the application of a multiobjective Bayesian optimization with random scalarizations to identify Pareto optimal probes with respect to sensitivity and selectivity. We train our active learning platform over 10 rounds and ∼250 probe molecules (∼6% of the design space) to identify probes on the sensitivity−selectivity Pareto frontier with sensitivities as high as $-\Delta G_{\mathrm{PFOS}} = 9.8 \pm 0.2$ kJ/mol, with a corresponding binding constant of $K_{\mathrm{b}}^{\mathrm{PFOS}} = 177.4 \pm 12.7$, and selectivities as high as $-\Delta\Delta G_{\mathrm{PFOS-SDS}} = 3.1 \pm 1.5$ kJ/mol, with a corresponding binding constant ratio of $K_{\mathrm{b}}^{\mathrm{PFOS}}/K_{\mathrm{b}}^{\mathrm{SDS}} = 4.6 \pm 1.7$. A retrospective analysis of our data reveals that marginal increases in chain length provide a significant increase in sensitivity, whereas the incorporation of amine headgroups and halogenation can improve the sensitivity and selectivity of probes. However, there is no clear trend as to which headgroup and halogenation combination results in better sensitivity and selectivity for a given length of hydrogenated or fluorinated probe. A $C_{11}Br_{23}P(CH3)_2$ probe containing 11 backbone brominated carbons and a tertiary phosphine headgroup possesses the highest computed sensitivity to PFOS within the design space, and a semi-brominated probe $C_5H_{11}C_7Br_{14}N(CH_3)_2$ containing 12 backbone carbons and a tertiary amine headgroup possesses the highest selectivity.

2. METHODS

We employ a machine learning-guided approach to efficiently navigate the molecular design space to discover probes with high sensitivity and selectivity. This approach is similar to the active learning frameworks we have employed in recent high-throughput screening campaigns for self-assembling π-conjugated peptides,[26] switchable nanostructured materials,[27] and small organic molecules to selectively permeate cardiolipin membranes.[28] Figure 2 illustrates our employed active learning framework that involves four key components: (i) all-atom molecular dynamics (MD) simulations and enhanced sampling

https://doi.org/10.1021/acs.jced.3c00404
J. Chem. Eng. Data XXXX, XXX, XXX−XXX

using parallel bias metadynamics to estimate sensitivity ($-\Delta G_{PFOS}$) and selectivity ($-\Delta\Delta G_{PFOS-SDS}$), (ii) encoding of probes in the search space using a pretrained variational autoencoder employing a Self-Referencing Embedded Strings (SELFIES)[29] representation, (iii) construction of Gaussian process regression (GPR) surrogate models over the collected simulation data to predict sensitivity and selectivity for the encoded probes in the search space, and (iv) multiobjective Bayesian optimization (BO) using random scalarizations and the trained GPR surrogate models to identify the next round of probes in the search space with optimal sensitivity and selectivity.

2.1. Estimation of Sensitivity ($-\Delta G_{PFOS}$) and Selectivity ($-\Delta\Delta G_{PFOS-SDS}$). *2.1.1. All-Atom Molecular Dynamics Simulations of PFOS-Probe and SDS-Probe Systems.* We perform all-atom MD simulations of PFOS-probe and SDS-probe interactions accelerated with parallel bias metadynamics[30] (Figure 2a). Although constant pH MD simulations[31] could dynamically capture the protonation state of a molecule, we fix the protonation state and perform standard MD simulations because of their lower computational cost and higher computational efficiency that are of paramount importance for our high-throughput virtual screening. We model PFOS in its anionic state[32,33] because of its extremely low pK_a[34] (<1) and consider a target pH range of 6.5−8.5 corresponding to typical tap and drinking water sources.[35−37] Similarly, we model SDS in the anionic state that has a low pK_a[38] of 3.3. Details on the generation of initial structures in the PDB format of PFOS, SDS, and the probes are provided in the Supporting Information. Calculations to determine partial charges are performed using the restrained electrostatic potential (RESP) method[39] with Gaussian 16RevA.03[40] and force-field parameters are taken from General Amber Force Field (GAFF)[41] by antechamber.[42] Specifically, we performed geometry optimization and the partial charge calculations in vacuum using density functional theory (DFT) at B3LYP/6-31G(d) basis level following the conventional RESP procedure.[39] A more computationally burdensome but more physically accurate scheme for partial charge assignments would employ a solvation model in the charge calculation.[43−45] System topology generation in GROMACS[46] format is facilitated by ACPYPE.[47] Each simulation system comprises a single probe and a single target analyte in water, which can be interpreted as a system at an infinite dilution concentration limit. This follows the primary objective of this work in designing probes for detecting target PFAS at extremely low target concentrations enforced by EPA. The number of water molecules and length of the cubic simulation box used for each probe-PFOS and probe-SDS system are provided in Tables S2−S12 in the Supporting Information. Each system was first energy minimized and then equilibrated for 1 ns in an NVT ensemble at 300 K followed by a 1 ns equilibration in an NPT ensemble at 300 K and 1 bar. MD simulation parameters applied during energy minimization and the equilibration runs are provided in the Supporting Information.

2.1.2. Enhanced Sampling Using Parallel Bias Well-Tempered Metadynamics. Accelerated sampling of probe-SDS and probe-PFOS interactions makes high-throughput simulation of the targeted design space computationally accessible by reducing simulation costs while ensuring good sampling of the intermolecular free energy landscape. In this work, we employ the parallel bias metadynamics (PBMetaD) method[30] to accelerate the sampling of probe-SDS and probe-PFOS interactions and estimate their equilibrium-binding free energies and binding constants. Details on the PBMetaD procedure, the choice of collective variables (CVs), calculation of the potential of mean force curves as a function of the center of mass separation $PMF(r)$, and binding free energies ΔG between the probes and PFOS or SDS are provided in the Supporting Information.

All the molecular simulations with parallel bias metadynamics are performed using the GROMACS-2018.6 suite[46] with PLUMED-2.5.2[48] library. These enhanced sampling calculations took ∼3−5 days for each PFOS or SDS and probe system depending on the system size on a shared supercomputing node utilizing a 1 × V100 GPU and 20 × Intel Skylake CPU cores.

2.2. Encoding of Probe Molecules Using a Pretrained Variational Autoencoder Using SELFIES Representations. *2.2.1. Molecular Design Space.* In this work, we focus our search for optimal probes by considering linear hydrogenated and halogenated molecules with the number of carbons ranging from 3 to 14 (Figure 1). We consider three halogens, fluorine, chlorine, and bromine, for the halogenated molecules with a constraint that the halogenated methyl and methylene groups are contiguous along the backbone of the probe molecule. The halogenated methylene and methyl groups were primarily considered to take advantage of the halogen-mediated interactions between the probe and the target PFAS.[49−51] For each of the hydrogenated or halogenated molecules, we also include amine- and phosphine-based headgroups to utilize electrostatic interactions for binding the headgroup of the probe with the anionic sulfonate headgroup of PFOS. These headgroups replace terminal hydrogenated or halogenated methyl groups in a probe. For amine headgroups, we include primary (NH_2), secondary ($NH(CH_3)$), tertiary ($N(CH_3)_2$), and quaternary ($N(CH_3)_3^+$) states. For phosphine headgroups, primary (PH_2), secondary ($PH(CH_3)$), and tertiary ($P(CH_3)_2$) states are considered. Collectively, the headgroup combinations together with the linear hydrogenated and halogenated tails, there are 3850 molecules in our search space (Figure 1).

2.2.2. Molecular Featurization and Embedding Using a Pretrained Small-Molecule Variational Autoencoder (VAE). The variational autoencoder (VAE)[52,53] model utilized in this study was previously trained on more than 1.2 M small molecules contained within the ZINC data set[54−56] plus a number of common chemical screening libraries.[57] It was our anticipation that this pretrained VAE over a large class of small molecules would provide good representations of the 3850 linear probe molecules considered in this work as rich but interpretable featurization appropriate for our active learning search. The training process first requires a unique representation of the molecules using SELFIES[29] and optimization of a loss function to learn a continuous representation of these molecules within a low-dimensional latent space constituting the information bottleneck layer between the encoder and the decoder. This latent space provides a smooth and low-dimensional representation of the molecular design space that is well suited to the construction of surrogate structure−property models and enables a Bayesian-guided traversal and optimization of molecules within the design space.[26,28,53,57] We provide a brief description of the VAE model construction and training in the Supporting Information; full details are available in Tang et al.[57]

https://doi.org/10.1021/acs.jced.3c00404
J. Chem. Eng. Data XXXX, XXX, XXX−XXX

The 3850 candidate probe molecules in the design space were created and stored as SMILES strings using RDKit.[58] We use selfies 1.0.4[29] to encode each molecule into a SELFIES representation using the generated SMILES strings. The molecules in SELFIES representation are then projected into the molecular latent space of the pretrained VAE model (Figure 2b.ii). For this, we apply the same one-hot encoding SELFIES representation used during the training process of the VAE model. To reduce the 100-dimensional latent space representation of each molecule, we then apply principal component analysis (PCA).[59] The top five principal components (PC) capture more than 95% of the cumulative variance, allowing us to further compress the latent space representation for the purpose of GPR training and BO-active learning with limited loss of information (Figure S1). We also verify that the 3850 linear probe molecules are smoothly embedded into the five leading PCs of the latent space, exhibiting smooth transitions in key molecular properties such as molecular weight and number of carbon or halogen atoms in the probe (Figures S2−S6). This implies that the molecules in a given neighborhood of latent space have similar properties and might therefore also be expected to have similar observables, such as binding affinity to a target analyte. The present work considers a finite design space of 3850 candidate molecules that we project into the VAE latent space to provide a featurization of these molecules suitable for our active learning search. As such, we do not exploit the generative capacity of the VAE decoder to produce novel molecules conditioned on a particular location in the latent space, although we note that this capability could be useful for expanding the search into new regions of molecular design space.

2.3. Construction of Gaussian Process Regression (GPR) Surrogate Models. We train the GPR surrogate models[60] to predict $(-\Delta G_{\mathrm{PFOS}})$ and $(-\Delta\Delta G_{\mathrm{PFOS-SDS}})$ as a function of the 5D vector x specifying the projection of each candidate probe molecule into the leading five principal components of the VAE latent space (Figure 2c). We construct our GPR kernel using the widely used and infinitely differentiable radial basis function (RBF) for the GPR covariance function. The RBF length scale parameter l that sets the distance for two points to be correlated is treated as hyperparameter that is optimized separately for each of the five dimensions of x during training by maximizing the log marginal likelihood.[61] The bounds of RBF length scale parameter l during optimization were set to 0.001 and 200, which are close to the minimum and maximum distances between the 3850 candidate molecules along the top five principal components. The GPR models are trained using scikit-learn.[62] The models are trained over the $(-\Delta G_{\mathrm{PFOS}})$ and $(-\Delta\Delta G_{\mathrm{PFOS-SDS}})$ values collected for all probes to date and return predictions of an estimated mean μ and standard deviation σ for $(-\Delta G_{\mathrm{PFOS}})$ and $(-\Delta\Delta G_{\mathrm{PFOS-SDS}})$ for any vector x. In this manner, GPR serves as a surrogate model for the sensitivity and selectivity of new probe molecules that have not yet been simulated.

2.4. Multiobjective Bayesian Optimization (BO). The trained GPR surrogate models are passed to a multiobjective Bayesian Optimization (BO) protocol using random scalarizations[63] (Figure 2d). The aim of the BO-guided search is to prospectively identify probe molecules residing on the high-sensitivity−high-selectivity Pareto frontier. A single acquisition function for the sensitivity (i.e., maximize $(-\Delta G_{\mathrm{PFOS}})$) and selectivity (i.e., maximize $(-\Delta\Delta G_{\mathrm{PFOS-SDS}})$) objectives is constructed from a random scalarization of the two GPR surrogate models.[63,64] In this approach, two independent acquisition functions are first employed for each of the objectives and then merged to obtain a single scalarized acquisition function.

As is standard practice when employing random scalarizations, we employ upper confidence bound (UCB) acquisition functions $\alpha_{\mathrm{UCB,sens}}(x) = \mu(-\Delta G_{\mathrm{PFOS}}(x)) - \beta\sigma(-\Delta G_{\mathrm{PFOS}}(x))$ and $\alpha_{\mathrm{UCB,select}}(x) = \mu(-\Delta\Delta G_{\mathrm{PFOS-SDS}}(x)) - \beta\sigma(-\Delta\Delta G_{\mathrm{PFOS-SDS}}(x))$[65] for each of the two design objectives,[63,66] where the mean and standard deviations in the sensitivity and selectivity at a particular x are extracted from the most recently trained GPR models and β is a random variable drawn from a log uniform distribution on the range $[\log(10^{-4}), \log(10^{4})]$ that controls the degree of exploitation vs exploration.[64] The two UCB acquisition functions are merged via a randomly weighted linear combination into a single scalarized acquisition function $\alpha(x) = \gamma\alpha_{\mathrm{UCB,sens}}(x) + (1 - \gamma)\alpha_{\mathrm{UCB,select}}(x)$ (Figure 2d). Colloquially, $\alpha(x)$ provides a measure of desirability for each potential candidate molecule with embedding vector x under a random sample of the relative weighting $\gamma \sim U(0, 1)$ between the two objectives, where $U(0, 1)$ is a uniform distribution over $(0, 1)$. Under sufficient number of trials, γ will explore various weightings of the sensitivity−selectivity design objectives and guide sampling to explore the entire Pareto frontier. We calculate $\alpha(x)$ for all as-yet-unsampled candidate molecules in the design space and employ the Kriging believer[67] to perform a batched selection of molecules for the enhanced sampling calculations.

We seed the active learning search process with 45 manually selected candidates from the design space designed to comprise a diversity of molecules consisting of hydrogenated and halogenated probes of different lengths with and without amine or phosphine headgroups (Table S2). Rather than distributing the initial probes over the search space using a diversity maximizing strategy,[68] we instead adopted a hypothesis-driven approach motivated by experimental insights in an attempt to inject this prior knowledge into the initialization of the search. Specifically, we hypothesized that the fluorophilic interactions can be critical in binding the probe with the target PFOS containing a fluorinated backbone relative to the interferent SDS that has a hydrogenated backbone. For this reason, we initially selected probes of different lengths ranging from 3 to 14 backbone carbons with different degrees of contiguous blocks of fluorination. In addition, we added amine headgroups to the probes to explore the utility of electrostatic interactions with the head group in PFOS and SDS. For a selected few probes that are found to be optimal with respect to sensitivity and selectivity in the fluorinated probes, we studied chlorinated variants of the fluorinated probes. We performed a total of 10 rounds of active learning over the course of which we simulated a total of 252 probe molecules (∼6% of the 3850 candidate design space) at a cost of ∼25,000 GPU-h and ∼5000 CPU-h. Approximately $N = 20$ molecules were sampled within each active learning cycle, but this number varied in each round in order to make efficient use of parallel computing resources. Specifically, molecules whose 1 μs simulation completed within the ∼2 week time horizon of a single active learning round participated in the active learning cycle. The remaining molecules whose simulations were incomplete at the beginning of the new cycle were added to the next cycle. In addition, we added five hand-selected probes with amine headgroup in active learning

https://doi.org/10.1021/acs.jced.3c00404
J. Chem. Eng. Data XXXX, XXX, XXX−XXX

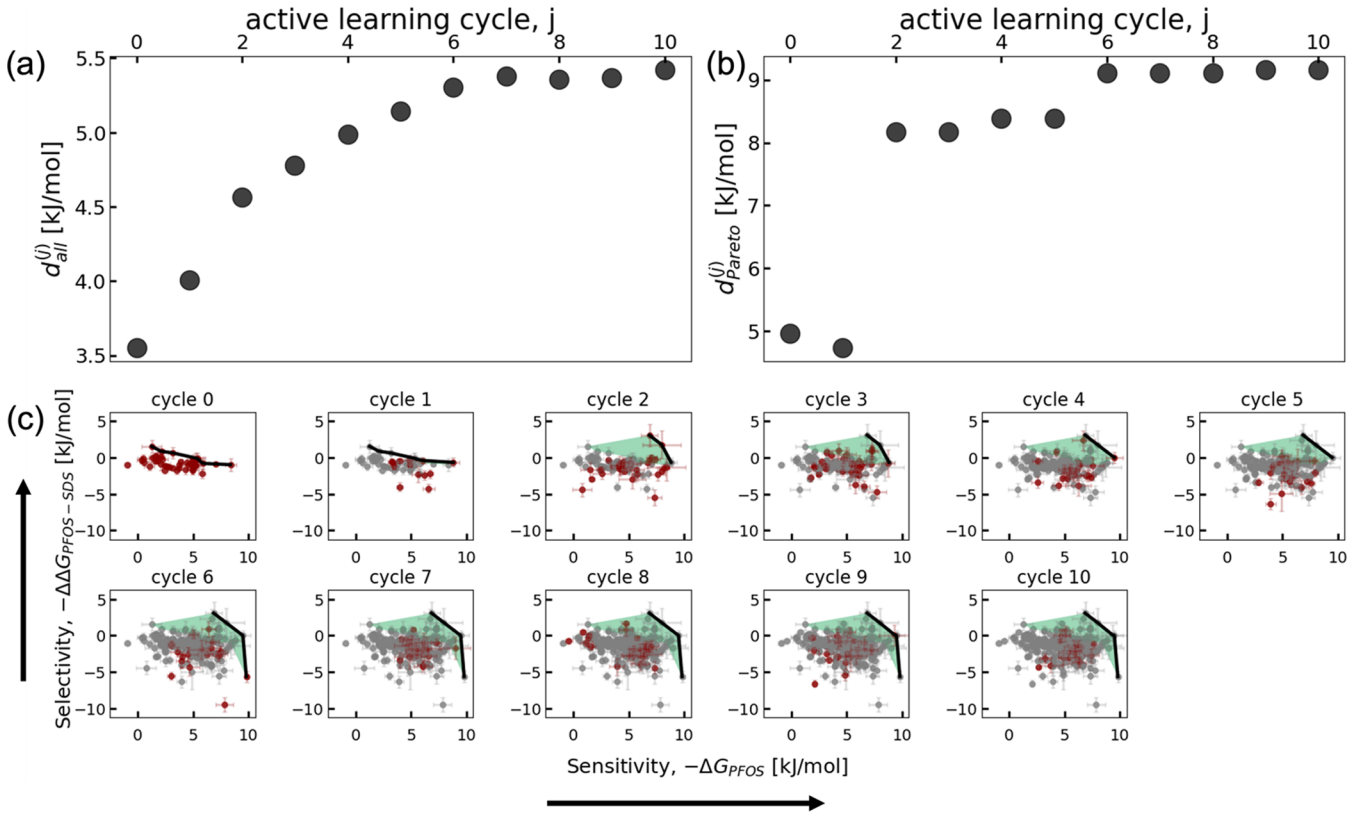


Figure 3. Progress of the active learning campaign in advancing the sensitivity−selectivity frontier of PFOS probes within each active learning cycle. (a) Mean distance $d^{(j)}_{all}$ reporting the cumulative average over the d_i values of all n probes sampled so far up to and including cycle j. (b) Pareto distance $d^{(j)}_{Pareto}$ reporting an average over the d_i values of those probes constituting the Pareto frontier at the end of round j, including contributions to the frontier from all prior rounds. (c) Scatter plots of the sampled probes within the sensitivity−selectivity design space over the course of the active learning campaign. The sensitivity and selectivity increase in the positive x- and y-directions, respectively. The probes sampled in the present round are indicated by red markers and those accumulated from all previous rounds by gray markers. The Pareto frontier is indicated by a black line. The green-shaded region highlights the shift in the Pareto frontier relative to cycle 0. Cycle 0 comprises the 45 molecules used to train the initial GPR model.

cycle 8 within a human-in-the-loop intervention into the otherwise automated search protocol. Specifically, we were interested in understanding the role of amine headgroups upon probe performance and appreciated that such molecules had heretofore been undersampled in the active learning screen. Incorporating these molecules within our screen was valuable for our retrospective analyses of the role of probe length, degree of fluorination, and presence of amine headgroups upon probe performance.

A notebook containing the computational active learning process with multiobjective BO is available at https://github.com/Ferg-Lab/activeLearningPFASLinear.git. Sensitivities ($-\Delta G_{PFOS}$), selectivities ($-\Delta\Delta G_{PFOS-SDS}$), and equilibrium-binding constants of the probes explored in each round of the active learning search are reported in Tables S2−S12, along with their two-dimensional (2D) structures, molecular weights, and IUPAC names. In addition, the SMILES strings, sensitivities, selectivities, and equilibrium-binding constants (K_b^{PFOS}, K_b^{SDS}) of each cycle are available in machine-readable format at https://github.com/Ferg-Lab/activeLearningPFASLinear.git along with helper notebooks and scripts to help facilitate ready uptake and use by the community.

2.5. Training of LASSO Regression Models to Infer Important Chemical Substructures of Probes. We train and analyze LASSO (Least Absolute Shrinkage and Selection Operator) regression models to extract important chemical substructures of probes that correlate with their sensitivity and selectivity to PFOS. In contrast to the relatively opaque GPR models, these simple and interpretable linear models are commonly applied for feature selection and to understand substructures or subgraphs that play an important role in predicting observed molecular properties and responses.[28,69] LASSO regression employs standard least-squares regressions with an L1 regularization over the weights to minimize the cost function $l = \sum_i^N (y_{true,i} - \sum_j^K w_j x_{ij})^2 + \lambda \sum_j^K ||w_j||_1$, where $y_{true,i}$ corresponds to the measured response, x_{ij} corresponds to the value of the jth feature of sample i, w_j is the linear regression coefficient accorded to feature j within the LASSO model, N is the number of samples, K is the number of features, and λ is a hyperparameter controlling the strength of the L1 regularization. The hyperparameter λ is typically tuned by cross-validation. The effect of the regularization term is to shrink the regression coefficients of the least important features to zero, thereby inducing sparsity in the model to include only those features with the greatest explanatory power. The magnitude and sign of the weights w_j lend themselves to simple interpretability as to the strength and direction of the effect of that feature on the response.

We train separate LASSO regression models to predict the measured negative sensitivity ($y_{true,i} = \Delta G^i_{PFOS}$) or negative selectivity ($y_{true,i} = \Delta\Delta G^i_{PFOS-SDS}$) of the 252 simulated probe

https://doi.org/10.1021/acs.jced.3c00404
J. Chem. Eng. Data XXXX, XXX, XXX−XXX

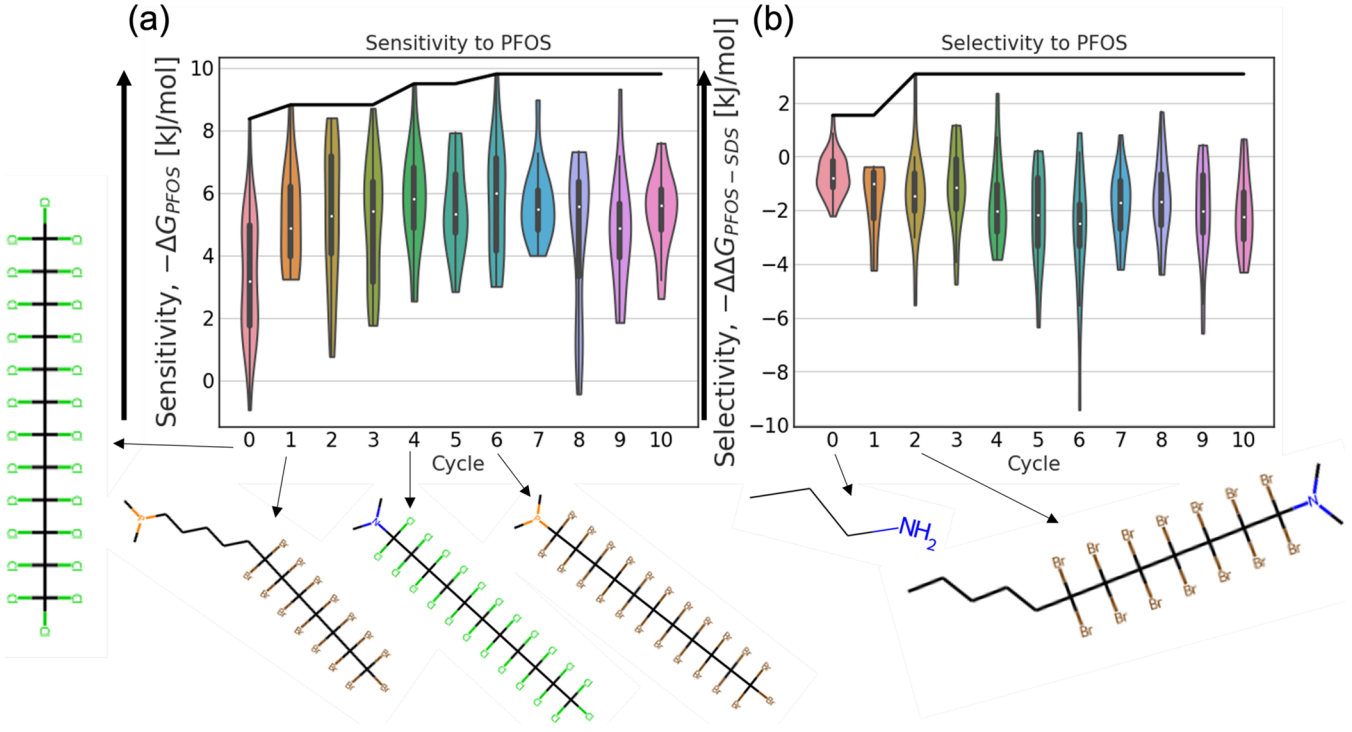


Figure 4. Change in the distribution of sensitivity and selectivity of probes to PFOS with each active learning cycle. Violin plots of the (a) sensitivity ($-\Delta G_{PFOS}$) and (b) selectivity ($-\Delta\Delta G_{PFOS-SDS}$) as functions of the active learning cycle. Chemical structures of molecules with maximum sensitivity and selectivity among all of the probes studied up to a given cycle are shown at the cycles in which they were discovered. The black solid line shows the change in the maximum sensitivity or selectivity of probes to PFOS with each active learning cycle. Both sensitivity and selectivity increase with an increase in the positive direction of the *y*-axis.

molecules using an input featurization enumerating the presence or absence of particular chemical substructures (i.e., molecular subgraphs) within each molecular probe. We follow the procedure described by Bhattacharjee and Vlachos[69] to extract the substructures of the molecular probes explored in the active learning process using RDKit[58] by setting the minimum and maximum edge lengths as 1 and 8, respectively. Hydrogen atoms are not considered to be part of the substructures. We note that the substructures mined consider only edge connectivity between atoms but not the atom properties, such as their charge state. A matrix with each molecule and the count of each unique substructure mined from RDKit[58] as rows and columns, respectively, are used as features for training the LASSO regression models. Standard scalers in scikit-learn[62] are applied to normalize the features or matrix representing substructure count in each molecule prior to training. The hyperparameter λ in the LASSO regression model is tuned by 10-fold cross-validation to minimize the validation prediction error on the validation set (Figure S7). The trained models are evaluated on a randomly selected 20% hold-out test set that was not exposed to the model during any part of training (Figure S8).

3. RESULTS AND DISCUSSION

3.1. Discovery of Optimal Linear Molecular Probes Using Computational Active Learning. Our goal is to discover probes in the design space of 3850 linear molecules possessing optimal sensitivity and selectivity to PFOS (Figure 1). We efficiently traverse the design space using our computational active learning pipeline (Figure 2) that we seed with an initial batch of 45 probes and execute 10 cycles of active learning sampling with N = 13−29 molecules per round (Tables S2−S12). To assess the convergence of the active learning campaign, we calculate

$$d_i = \sqrt{(\text{sensitivity}_i)^2 + (\text{selectivity}_i)^2} = \sqrt{(-\Delta G^i_{\text{PFOS}})^2 + (-\Delta\Delta G^i_{\text{PFOS-SDS}})^2} \quad (1)$$

as the distance of each probe i from the origin in the 2D sensitivity−selectivity design space. Recalling that it is our objective to simultaneously maximize sensitivity ($-\Delta G_{\text{PFOS}}$) and selectivity ($-\Delta\Delta G_{\text{PFOS-SDS}}$), this presents a quantitative measure of the advancement of the active learning screen in discovering desirable molecular probes. We further compute

$$d^{(j)}_{\text{all}} = \overline{d}_i,\ i \in \{(0), \ldots, (j)\} \quad (2)$$

$$d^{(j)}_{\text{Pareto}} = \overline{d}_i,\ i \in \text{Pareto}_{\{(0),\ldots,(j)\}} \quad (3)$$

where $d^{(j)}_{\text{all}}$ reports a cumulative average over the d_i values of all probes sampled in the campaign so far up to and including cycle j, and $d^{(j)}_{\text{Pareto}}$ reports an average over the d_i values of those probes constituting the Pareto frontier at the end of round j. We present in Figure 3c scatter plots showing the location of the sampled probes in each cycle of the campaign within the sensitivity−selectivity design space and the advancement of the Pareto frontier. The trends in Figure 3 indicate a convergence of our active learning search by approximately cycle 6, beyond which we do not observe any further improvements in $d^{(j)}_{\text{all}}$ or $d^{(j)}_{\text{Pareto}}$, or further advancement of the Pareto frontier. This suggests that the search has identified the top-performing molecules within the molecular design space and that we may not expect the discovery of significantly superior candidates with additional cycles of the search.

https://doi.org/10.1021/acs.jced.3c00404
J. Chem. Eng. Data XXXX, XXX, XXX−XXX

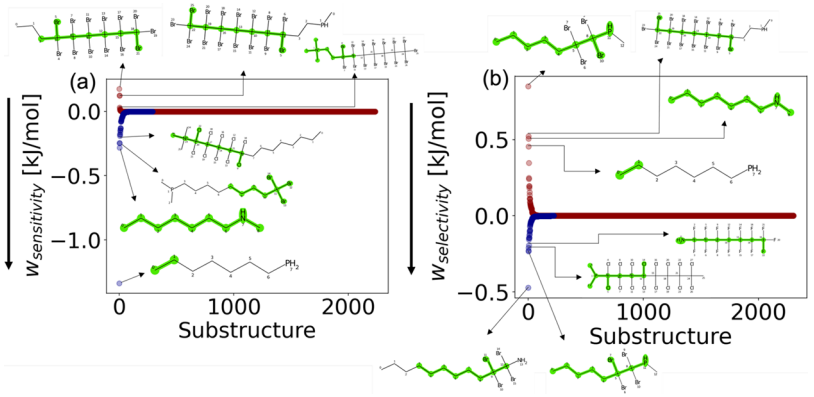


Figure 5. Chemical substructures within the 252 molecular probes considered within our active learning search identified by LASSO regression to be the leading determinants of PFOS (a) sensitivity and (b) selectivity. Large LASSO weights, w, identify substructures that are the leading determinants of sensitivity or selectivity. The chemical substructures possessing the largest LASSO weights are illustrated in green on representative probe molecules containing these substructures. The vertical arrows on the y-axis indicate the direction of favorable LASSO weights.

We present in Figure 4 violin plots illustrating the cumulative sensitivity and selectivity distributions of all probes sampled over the course of the active learning campaign. In terms of sensitivity, we see round-on-round improvements in the top candidate identified up to cycle 6, beyond which no further advances are observed. In terms of selectivity, we see improvements only through cycle 2, where the top candidate is identified. The initial 45 probes (cycle 0) have a broad range of sensitivities——0.9 ± 0.3 to 8.4 ± 0.5 kJ/mol (Figure 4a)—but their selectivities span a relatively narrow range −2.2 ± 0.6 to 1.5 ± 0.9 kJ/mol (Figure 4b). A fully chlorinated probe $C_{12}Cl_{26}$ with 12 backbone carbons has the highest sensitivity to PFOS among all of the studied probes in cycle 0. However, this probe has lower selectivity to PFOS than probe $C_3H_7NH_2$ that has three backbone carbons and one primary amine headgroup, which has the highest selectivity among all of the probes in cycle 0. We discovered three more probes with improved sensitivity in cycles 1, 4, and 6. These include a semibrominated probe $C_7Br_{15}C_5H_{10}P(CH_3)_2$ with 12 backbone carbons and a tertiary phosphine headgroup, a completely chlorinated probe $C_{11}Cl_{23}N(CH_3)_2$ with 11 backbone carbons and a tertiary amine headgroup, and a completely brominated probe $C_{11}Br_{23}P(CH3)_2$ with 11 backbone carbons and a tertiary phosphine headgroup. A semibrominated probe $C_5H_{11}C_7Br_{14}N(CH_3)_2$ with 12 backbone carbons and a tertiary amine headgroup has the highest selectivity among all of the probes discovered in the active learning cycles.

To facilitate contact with experimental measurements, we compute the equilibrium-binding constant K_b from our calculated free energy profiles by numerically evaluating the following integral[70,71]

$$K_b = C^0 \int_0^{r_{bound}} 4\pi r^2 \exp(-\mathrm{PMF}(r)/k_BT)\mathrm{d}r \quad (4)$$

where $C^0 = 1/1661$ Å^{-3} is the standard state concentration at 1 mol/L, and the cutoff of $r_{bound} = 1$ nm delimits the bound and unbound states, motivated by the 1 nm cutoff of the van der Waals and real-space electrostatic interactions implemented in our simulations. A derivation of this expression is provided in the Supporting Information. We report in Table S1 the values of K_b^{PFOS}, K_b^{SDS}, and K_b^{PFOS}/K_b^{SDS} for the five top-performing probes residing on the sensitivity−selectivity Pareto frontier in the terminal round of our active learning campaign. The K_b^{PFOS} and K_b^{SDS} values for all molecules considered within our screen are reported in Tables S2−S12.

How are the top-performing probes identified in our screen predicted to perform in practice? With regard to sensitivity, adopting a target PFOS concentration of 4 ng/L ≈ 8 × 10^{-12} mol/L mandated by EPA guidelines for safe drinking water[15] and assuming a 10-fold higher initial free PFOS concentration and negligible interferents present, we can invert the relationship $K_b^{PFOS} = C^0 C_{probe-PFOS}/C_{probe}C_{PFOS}$ to estimate that the most sensitive probe discovered in our screen with $K_b^{PFOS} = (177.4 \pm 12.7)$ would require to be present at a concentration of $C_{probe} \approx 0.05$ mol/L. It is conceivable that such local probe concentrations may be achievable by surface immobilization within a sensing device, but substantially lower concentrations may be expected for previously reported high-affinity molecular host−guest systems such as β-cyclodextrin ($K_b \approx 10^4-10^5$, $C_{probe} \approx 0.1-1$ mmol/L)[72,73] or guanidinocalix[5]arenes ($K_b \approx 10^7$, $C_{probe} \approx 1$ μmol/L).[73] Assuming that our active learning search has identified the top-performing candidates within the molecular design space, these results indicate that this class of linear molecules does not contain highly sensitive molecular probes for PFOS detection. With regard to selectivity, assuming equimolar initial free concentrations of PFOS and SDS of 8 × 10^{-11} mol/L, the most selective probe in our screen with $K_b^{PFOS} = 62.7 \pm 15.3$, $K_b^{SDS} = 13.6 \pm 3.8$, and $K_b^{PFOS}/K_b^{SDS} = 4.6 \pm 1.7$ would require a concentration of $C_{probe} \approx 0.1$ mol/L to achieve the EPA target PFOS concentration and would produce a bound PFOS to SDS relative enrichment of $C_{probe-PFOS}/C_{probe-SDS} \approx 1.4$. This suggests that the linear molecular candidate space does not contain highly selective molecular probes, although we observe that even modest selectivities can be exploited within multiplexed sensing schemes.[74]

3.2. Identification of Chemical Substructures that Are the Principal Determinants of Sensitivity and Selectivity. Having efficiently navigated the search space using our active learning protocol to discover the candidates populating the sensitivity−selectivity Pareto frontier, we now seek to infer the chemical features underpinning the calculated sensitivity and selectivity of the molecular probes to PFOS. This understanding can both help rationalize the trends observed within our active learning search and guide and inform the subsequent design and exploration of augmented libraries engineered to be enriched with molecules with promising

https://doi.org/10.1021/acs.jced.3c00404
J. Chem. Eng. Data XXXX, XXX, XXX−XXX

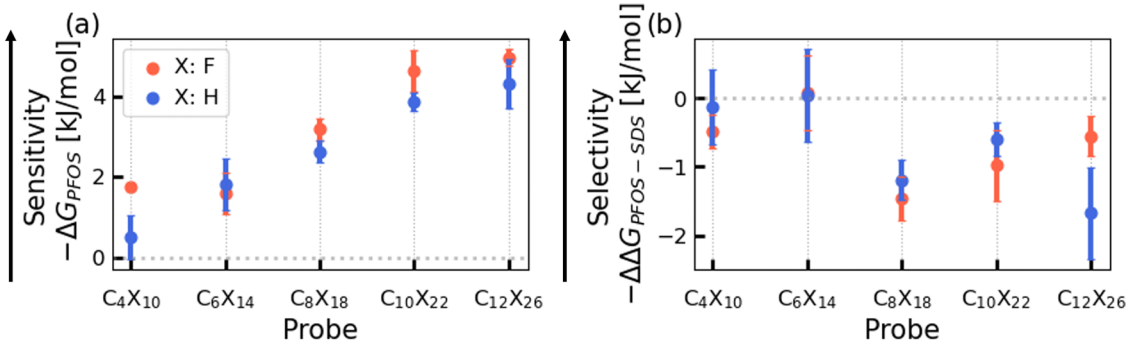


Figure 6. Effect of probe length upon PFOS (a) sensitivity and (b) selectivity for fully hydrogenated (blue) and fully fluorinated (red) probes. Markers and error bars represent means and standard errors estimated by a threefold block averaging over a 900 ns production run. Standard propagation of errors is employed to obtain the standard errors on selectivity. The vertical arrows on the *y*-axis indicate the direction of favorable sensitivity and selectivity.

chemical features. To do so, we trained LASSO regression models as interpretable "glass box" predictive models to identify the chemical substructures within the molecular probes that are the principal determinants of PFOS sensitivity and selectivity. We present in Figure 5 the LASSO regression weights w associated with each chemical substructure. The magnitude of these weights can be interpreted, under the linear LASSO model, as the relative importance of that substructure in determining the sensitivity or selectivity. Negative weights indicate molecular substructures that tend to promote elevated sensitivity ($-\Delta G^{i}_{\mathrm{PFOS}}$) or selectivity ($-\Delta\Delta G_{\mathrm{PFOS-SDS}}$), whereas positive weights indicate substructures that tend to be detrimental to sensitivity or selectivity.

Analysis of the sensitivity LASSO model (Figure 5a) indicates that the substructure with the most negative LASSO weight and therefore the chemical motif that most promotes high PFOS sensitivity is the C−C bond. The next three most favorable substructures are a hydrogenated probe with seven backbone carbons and a secondary amine headgroup, a five backbone carbon substructure with one brominated methyl group, and a chlorinated substructure with an amine-based headgroup. The top 20 substructures that promote favorable sensitivity are presented in Figure S9 and include hydrogenated and halogenated substructures containing amine and phosphine headgroups. Taken together, these results indicate that the chain length (i.e., number of C−C bonds) and the presence of chlorinated or fluorinated heads or tails, and brominated headgroup appear to elevate PFOS sensitivity. Turning to the positive LASSO weights, we identify a rather small number of significant chemical substructures that are detrimental to sensitivity, including two brominated substructures with seven backbone carbons and a semibrominated substructure with a quaternary amine headgroup. The remaining top seven substructures that promote unfavorable sensitivity are presented in Figure S10 and include substructures with chlorinated methyl groups and primary phosphine headgroups. These results appear to indicate that bromination of the tail and the inclusion of phosphine headgroups are detrimental to probe sensitivity.

Analysis of the selectivity LASSO model (Figure 5b) reveals a semibrominated substructure with eight branched backbone carbons to be the most favorable in promoting selectivity of probes for the PFOS target over the SDS interferent. This is followed by a similar semibrominated substructure but seven branched backbone carbons and a phosphine-based headgroup. Chlorinated, fluorinated substructures with phosphine- and amine-based headgroups are the next most favorable substructures. The top 20 substructures that promote favorable selectivity are presented in Figure S11, from which we identify the importance of halogenation and headgroups in improving the selectivity of a probe. In contrast to sensitivity, we identify a relatively large number of unfavorable substructures to selectivity. These include a semibrominated substructure with seven branched backbone carbons and phosphine headgroup, a branched brominated substructure with seven backbone carbons, a hydrogenated substructure with seven backbone carbons and a secondary amine headgroup, and the C−C bond. The top 20 substructures unfavorable to selectivity are presented in Figure S12. This indicates that bromination and increasing chain length can be unfavorable to promoting the selectivity of the molecular probes.

Considering now both sensitivity and selectivity, we observe that the third most unfavorable substructure to selectivity is the same as the second most favorable substructure to sensitivity. Furthermore, the fourth most unfavorable substructure to selectivity, the C−C bond, is the same as the most favorable substructure to sensitivity. This analysis suggests the competing nature of both chain length and certain substructures for the simultaneous optimization of both PFOS sensitivity and selectivity and exposes an inherent challenge for this two-dimensional optimization by controlling the presence of particular chemical groups.

3.3. Correlation Analysis of Important Identified Chemical Substructures. Our substructure analysis exposed C−C groups, halogenation, and headgroups as chemical substructures as important determinants of sensitivity and selectivity. Informed by this analysis, we now perform correlation analysis to quantify the role of these features upon sensitivity ($-\Delta G_{\mathrm{PFOS}}$) and selectivity ($-\Delta\Delta G_{\mathrm{PFOS-SDS}}$). In Figure 6, we plot the sensitivity and selectivity dependence upon the chain length for fully hydrogenated and fully fluorinated probes. For hydrogenated probes, we observed ∼0.4−1.3 kJ/mol increase on average in sensitivity with each additional C−C bond (Figure 6a). Fluorinated probes appear to be marginally more sensitive than their hydrogenated counterparts, but the corresponding values generally do not lie outside standard errors. Contrariwise, there is a degradation in selectivity with increasing probe length of ∼1−2 kJ/mol upon elongating the hydrogenated and fluorinated probes from 4 to 12 C−C bonds (Figure 6b), but this trend is not monotonic. In all cases, the selectivity for SDS is higher than or equal to that for PFOS, indicating that these probes are not selective

https://doi.org/10.1021/acs.jced.3c00404
J. Chem. Eng. Data XXXX, XXX, XXX−XXX

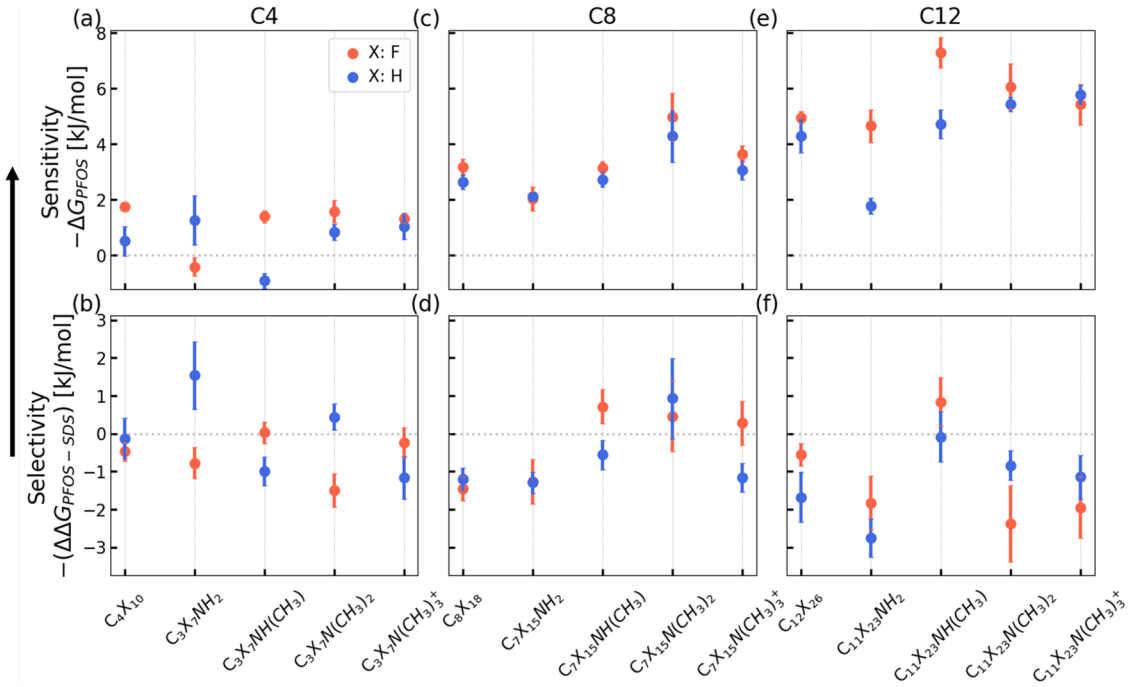


Figure 7. Effect of amine headgroup upon PFOS sensitivity and selectivity for fully hydrogenated (blue) and fully fluorinated (red) probes with a backbone length of (a, b) 4, (c, d) 8, and (e, f) 12 single bonds. Markers and error bars represent means and standard errors estimated by a threefold block averaging over a 900 ns production run. Standard propagation of errors is employed to obtain the standard errors on selectivity. The vertical arrows on the *y*-axis indicate the direction of favorable sensitivity and selectivity.

toward the desired target molecule. Again, there tends not to be statistically meaningful differences in selectivity between fully fluorinated and hydrogenated probes of the same length.

Although the PFOS target has a fluorinated tail, whereas the SDS interferent has a hydrogenated tail, our results indicate both fluorinated and hydrogenated probes have largely similar sensitivities and selectivities. This indicates the relatively weak role of fluorophilic interactions in modulating probe performance and is consistent with the DFT-SAPT-based study by Tsuzuki and Uchimaru,[49] which indicates that the dispersion forces that dominate intermolecular interactions are nearly identical among CF_4−CF_4, CF_4−CH_4, and CH_4−CH_4. It is only the electrostatic interactions that arise due to electronegativity effects that are higher in unlike pairs relative to like pairs, but their net contributions are smaller than dispersion forces.[49] Furthermore, the interactions between the probes and PFOS and SDS can be largely entropic, with a significant role played by the solvent. For example, fluorination of hydrocarbons can increase hydrophobicity[75] and alter the hydration shell structure of fluorinated methyl groups.[50] Complementary work shows that these changes are nonmonotonic in methyl groups with increasing fluorination.[51] In the context of the present work, we find similarly nonmonotonic trends in interaction strengths as a function of the degree of fluorination.

In Figure 7, we present the effect of different amine headgroups on the sensitivity and selectivity of fully hydrogenated and fluorinated probes of various lengths. Specifically, we consider probe molecules C_4X_{10}, C_8X_{18}, and $C_{12}X_{26}$, where X = H or F, and consider replacing the terminal CH_3 group with a primary NH_2, secondary $NH(CH_3)$, tertiary $N(CH_3)_2$, and charged quaternary $N(CH_3)_3^+$ amines. Overall, we see relatively muted and nonmonotonic trends in the influence of headgroup upon sensitivity and selectivity, although there is evidence for a generally favorable influence of substituting in a primary, secondary, or tertiary amine upon the sensitivity and selectivity at the C_8 probe length. There are also no clear trends in the sensitivity and selectivity difference of fully hydrogenated or fully fluorinated probes; however, the addition or removal of halogen atoms and amine or phosphine headgroups produces nonmonotonic changes in both sensitivity and selectivity with changes to sensitivity of up to 2.5 kJ/mol. The addition of charge to the probe via the charged quaternary amine did not tend to lead to statistically meaningful improvements in sensitivity or selectivity.

4. CONCLUSIONS

The detection and elimination of PFAS "forever chemicals" within environmental water courses represent a pressing challenge for the design of highly sensitive and selective molecular probes. In this work, we employed a computational active learning pipeline combining enhanced sampling calculations, deep representational learning, Gaussian process regression, and multiobjective Bayesian optimization to efficiently screen molecular candidate libraries for highly sensitive and selective molecular probes. We chose to focus on a design space comprising 3850 linear hydrogenated and halogenated molecules of length 3−14 carbons with and without an amine- or phosphine-based headgroup. The design of the library was motivated by the ready synthetic accessibility of these candidate probes, their similar chemical nature to PFAS, and their capacity to exploit hydrophobic binding to the aliphatic tail and electrostatic binding to the head. After conducting 10 rounds of our active learning screen during which we simulated a total of 252 probe molecules (~6% of the 3850 candidate design space) at a cost of ~25,000 GPU-h and ~5000 CPU-h, we identified five candidate molecules lying on the sensitivity−selectivity Pareto frontier with sensitivities spanning $(-\Delta G_{PFOS})$ = 6.9−9.8 kJ/mol and

https://doi.org/10.1021/acs.jced.3c00404
J. Chem. Eng. Data XXXX, XXX, XXX−XXX

selectivities spanning $-\Delta\Delta G_{PFOS-SDS} = -5.6-3.1$ kJ/mol. The most sensitive probe identified in our screen, a $C_{11}Br_{23}P(CH3)_2$ molecule containing 11 backbone brominated carbons and a tertiary phosphine headgroup, possesses a binding constant of $K_b^{PFOS} = 177.4 \pm 12.7$ and the most selective probe, a semibrominated molecule $C_5H_{11}C_7Br_{14}N(CH_3)_2$ containing 12 backbone carbons and a tertiary amine headgroup, possesses a binding constant ratio of $K_b^{PFOS}/K_b^{SDS} = 4.6 \pm 1.7$. An analysis of the influence of particular chemical motifs within the probes that are the primary determinants of the observed sensitivity and selectivity exposes delicate and relatively weak trends in probe length, halogenation, and headgroup in modulating their behaviors.

The relative paucity of highly performant linear probe molecules identified by our ML-enabled workflow militates for an expansion of the molecular search space to molecular libraries containing a richer chemical diversity of candidate molecules including branched and cyclic hydrocarbons, as well as more exotic molecules such as cyclodextrin-based probes that have garnered some attention and success in PFAS sequestration.[7,76−78] Further enlargements of the design space can make the computational screening platform established in this work even more valuable in performing an efficient filtration of the enlarged design space to identify the most promising molecular probes to be explored experimentally and also in excavating a molecular-level understanding of probe performance that can be coupled with experimental intuition to help inform rational molecular design. The performance of our active learning pipeline in achieving convergence after considering only 6% of the molecular space provides support for its capacity to efficiently navigate and traverse large search spaces and quickly focus on the most promising candidates.

While we focused on only one interferent in this work, it may be desirable to find probes that are selective to a given target PFAS in the presence of various types of interferents for practical applications. To this end, one may take the top-performing probes identified in the primary screen into a lower-throughput secondary screen against a broader panel of potential interferents to better evaluate the breadth of their selectivity. In the future, it is of interest to also extend our approach for discovering molecular probes for some of the other commonly known PFAS variants[79] such as perfluorooctanoic acid (PFOA), perfluorohexanesulfonate (PFHxS), and 6:2 fluorotelomer sulfonate (6:2FTS). Similarly, it may be of interest to evaluate the performance of the top probes over a wider range of operating temperatures and pH conditions. We would also like to couple our computational search to hybrid computational/experimental active learning campaigns, in which we conduct asynchronous but simultaneous computational and experimental screens in order to integrate high-throughput computation and low-throughput experimentation to minimize the experimental costs in time and labor to identify the top-performing probes for applications in PFAS sensors.[68]

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jced.3c00404.

Initial structure generation and MD simulations details during energy minimization and equilibration, production MD simulations details with PBMetaD, a brief description of VAE model training, and derivation of expression for the calculation of binding constants from free energy profiles; fraction of variance explained by principal components of latent space of molecular probes; variation of molecular weight, number of carbons, number of fluorines, number of chlorines, and number of bromines along top five principal components of latent space of molecular probes; LASSO model hyperparameter tuning and sensitivity−selectivity parity plots; leading chemical substructures identified by LASSO models; sensitivities, selectivities, and K_b values of top five performing probes residing on the sensitivity−selectivity Pareto frontier in the terminal round of our active learning campaign; SMILES strings, molecular weights, CAS (Chemical Abstracts Service) registration number (RN), IUPAC names, 2D chemical structures, sensitivity, selectivity, and K_b values for molecular probes studied in each active learning cycle (PDF)

Special Issue Paper

Published as part of the *Journal of Chemical & Engineering Data* virtual special issue "Machine Learning for Thermophysical Properties."

■ AUTHOR INFORMATION

Corresponding Author

Andrew L. Ferguson − *Pritzker School of Molecular Engineering, University of Chicago, Chicago, Illinois 60637, United States;* orcid.org/0000-0002-8829-9726; Email: andrewferguson@uchicago.edu

Authors

Siva Dasetty − *Pritzker School of Molecular Engineering, University of Chicago, Chicago, Illinois 60637, United States;* orcid.org/0000-0002-1666-7980

Maximilian Topel − *Department of Physics, University of Chicago, Chicago, Illinois 60637, United States*

Yifeng Oliver Tang − *Pritzker School of Molecular Engineering, University of Chicago, Chicago, Illinois 60637, United States;* orcid.org/0000-0003-4247-6712

Yuqin Wang − *Pritzker School of Molecular Engineering, University of Chicago, Chicago, Illinois 60637, United States; Chemical Sciences and Engineering Division, Argonne National Laboratory, Lemont, Illinois 60439, United States; Advanced Materials for Energy-Water Systems Center, Argonne National Laboratory, Lemont, Illinois 60439, United States;* orcid.org/0000-0003-4444-2487

Eric Jonas − *Department of Computer Science, University of Chicago, Chicago, Illinois 60637, United States*

Seth B. Darling − *Pritzker School of Molecular Engineering, University of Chicago, Chicago, Illinois 60637, United States; Chemical Sciences and Engineering Division, Argonne National Laboratory, Lemont, Illinois 60439, United States; Advanced Materials for Energy-Water Systems Center and Advanced Energy Technologies Directorate, Argonne National Laboratory, Lemont, Illinois 60439, United States;* orcid.org/0000-0002-5461-6965

Junhong Chen − *Pritzker School of Molecular Engineering, University of Chicago, Chicago, Illinois 60637, United States; Chemical Sciences and Engineering Division, Argonne National Laboratory, Lemont, Illinois 60439, United States; Physical Sciences and Engineering Directorate, Argonne*

https://doi.org/10.1021/acs.jced.3c00404
J. Chem. Eng. Data XXXX, XXX, XXX−XXX

National Laboratory, Lemont, Illinois 60439, United States; orcid.org/0000-0002-2615-1347

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jced.3c00404

Author Contributions

○S.D. and M.T. contributed equally to this work.

Notes

The authors declare the following competing financial interest(s): A.L.F. is a co-founder and consultant of Evozyne, Inc. and a co-author of US Patent Applications 16/887,710 and 17/642,582, US Provisional Patent Applications 62/853,919, 62/900,420, 63/314,898, 63/479,378, and 63/521,617, and International Patent Applications PCT/US2020/035206 and PCT/US2020/050466. J.H.C. is a founder of NanoAffix Science LLC.

■ ACKNOWLEDGMENTS

This material is based on the work supported by the National Science Foundation under Grant No. DGE-2022023. This work was supported in part with funding by the University of Chicago Data Science Institute (DSI). This work was completed in part with resources provided by the University of Chicago Research Computing Center. The authors gratefully acknowledge the computing time on the University of Chicago's high-performance GPU-based cyberinfrastructure supported by the National Science Foundation under Grant No. DMR-1828629. The authors gratefully acknowledge the computing time on Frontera hosted by the Texas Advanced Computing Center (TACC) at The University of Texas at Austin. Work by S.B.D. and Y.W. at Argonne National Laboratory was supported by the AMEWS EFRC funded by DOE, Office of Science, BES under contract DE-AC02-06CH11357.

■ REFERENCES

(1) Wallington, T. J.; Andersen, M. S.; Nielsen, O. The case for a more precise definition of regulated PFAS. *Environ. Sci.: Processes Impacts* **2021**, *23*, 1834−1838.

(2) U.S. Environmental Protection Agency, TSCA Section 8(a)(7) Reporting and Recordkeeping Requirements for Perfluoroalkyl and Polyfluoroalkyl Substances. https://www.federalregister.gov/documents/2021/06/28/2021-13180/tsca-section-8a7-reporting-and-recordkeeping-requirements-for-perfluoroalkyl-and-polyfluoroalkyl, (accessed September 2022).

(3) Baker, E. S.; Knappe, D. R. U. Per- and polyfluoroalkyl substances (PFAS)—contaminants of emerging concern. *Anal. Bioanal. Chem.* **2022**, *414*, 1187−1188.

(4) Glüge, J.; Scheringer, M.; Cousins, I. T.; DeWitt, J. C.; Goldenman, G.; Herzke, D.; Lohmann, R.; Ng, C. A.; Trier, X.; Wang, Z. An overview of the uses of per-and polyfluoroalkyl substances (PFAS). *Environ. Sci.: Processes Impacts* **2020**, *22*, 2345−2373.

(5) Xia, C.; Diamond, M. L.; Peaslee, G. F.; Peng, H.; Blum, A.; Wang, Z.; Shalin, A.; Whitehead, H. D.; Green, M.; Schwartz-Narbonne, H.; Yang, D.; Venier, M. Per- and Polyfluoroalkyl Substances in North American School Uniforms. *Environ. Sci. Technol.* **2022**, *56*, 13845−13857.

(6) Sim, W.; Choi, S.; Choo, G.; Yang, M.; Park, J.-H.; Oh, J.-E. Organophosphate Flame Retardants and Perfluoroalkyl Substances in Drinking Water Treatment Plants from Korea: Occurrence and Human Exposure. *Int. J. Environ. Res. Public Health* **2021**, *18*, 2645.

(7) Wang, Y.; Darling, S. B.; Chen, J. Selectivity of Per-and Polyfluoroalkyl Substance Sensors and Sorbents in Water. *ACS Appl. Mater. Interfaces* **2021**, *13*, 60789−60814.

(8) Ghisi, R.; Vamerali, T.; Manzetti, S. Accumulation of perfluorinated alkyl substances (PFAS) in agricultural plants: A review. *Environ. Res.* **2019**, *169*, 326−341.

(9) De Silva, A. O.; Armitage, J. M.; Bruton, T. A.; Dassuncao, C.; Heiger-Bernays, W.; Hu, X. C.; Kärrman, A.; Kelly, B.; Ng, C.; Robuck, A.; Sun, M.; Webster, T. F.; Sunderland, E. M. PFAS exposure pathways for humans and wildlife: a synthesis of current knowledge and key gaps in understanding. *Environ. Toxicol. Chem.* **2021**, *40*, 631−657.

(10) Cousins, I. T.; Johansson, J. H.; Salter, M. E.; Sha, B.; Scheringer, M. Outside the Safe Operating Space of a New Planetary Boundary for Per-and Polyfluoroalkyl Substances (PFAS). *Environ. Sci. Technol.* **2022**, *56*, 11172−11179.

(11) Kwiatkowski, C. F.; Andrews, D. Q.; Birnbaum, L. S.; Bruton, T. A.; DeWitt, J. C.; Knappe, D. R.; Maffini, M. V.; Miller, M. F.; Pelch, K. E.; Reade, A.; Soehl, A.; Trier, X.; Venier, M.; et al. Scientific basis for managing PFAS as a chemical class. *Environ. Sci. Technol. Lett.* **2020**, *7*, 532−543.

(12) Fenton, S. E.; Ducatman, A.; Boobis, A.; DeWitt, J. C.; Lau, C.; Ng, C.; Smith, J. S.; Roberts, S. M. Per-and polyfluoroalkyl substance toxicity and human health review: Current state of knowledge and strategies for informing future research. *Environ. Toxicol. Chem.* **2021**, *40*, 606−630.

(13) Brennan, N. M.; Evans, A. T.; Fritz, M. K.; Peak, S. A.; von Holst, H. E. Trends in the Regulation of Per- and Polyfluoroalkyl Substances (PFAS): A Scoping Review. *Int. J. Environ. Res. Public Health* **2021**, *18*, 10900.

(14) CompTox Chemicals Dashboard, U.S. Environmental Protection Agency, PFAS: Listed in OECD Global Database. https://comptox.epa.gov/dashboard/chemical-lists/pfasoecd, (accessed September 2022).

(15) U.S. Environmental Protection Agency, Proposed PFAS National Primary Drinking Water Regulation. https://www.epa.gov/sdwa/and-polyfluoroalkyl-substances-pfas, (accessed March 2023).

(16) U.S. Environmental Protection Agency, PFAS Analytical Methods Development and Sampling Research. https://www.epa.gov/water-research/pfas-analytical-methods-development-and-sampling-research, (accessed September 2022).

(17) U.S. Environmental Protection Agency, Method 537.1 Determination of Selected Per- and Polyflourinated Alkyl Substances in Drinking Water by Solid Phase Extraction and Liquid Chromatography/Tandem Mass Spectrometry (LC/MS/MS). https://cfpub.epa.gov/si/si_public_record_report.cfm?dirEntryId=348508&Lab=CESER&simpleSearch=0&showCriteria=2&searchAll=537.1&TIMSType=&dateBeginPublishedPresented=03%2F24%2F2018 (accessed September 2022).

(18) U.S. Environmental Protection Agency, Method 533: Determination of Per- and Polyfluoroalkyl Substances in Drinking Water by Isotope Dilution Anion Exchange Solid Phase Extraction and Liquid Chromatography/Tandem Mass Spectrometry. https://www.epa.gov/dwanalyticalmethods/method-533-determination-and-polyfluoroalkyl-substances-drinking-water-isotope, (accessed September 2022).

(19) Cheng, Y. H.; Barpaga, D.; Soltis, J. A.; Shutthanandan, V.; Kargupta, R.; Han, K. S.; McGrail, B. P.; Motkuri, R. K.; Basuray, S.; Chatterjee, S. Metal-organic framework-based microfluidic impedance sensor platform for ultrasensitive detection of perfluorooctanesulfonate. *ACS Appl. Mater. Interfaces* **2020**, *12*, 10503−10514.

(20) Allonia, Allonnia's protein breakthrough brings commercial PFAS sensor a step closer. https://allonnia.com/wp-content/uploads/2022/08/Aug22-GWI-SWW-Allonnia.pdf, (accessed September 2022).

(21) Agency for Toxic Substances and Disease Registry, United States Department of Health and Human Services, Per- and Polyfluoroalkyl Substances (PFAS) and Your Health. https://www.atsdr.cdc.gov/pfas/health-effects/index.html, (accessed September 2022).

(22) Meegoda, J. N.; Kewalramani, J. A.; Li, B.; Marsh, R. W. A review of the applications, environmental release, and remediation

technologies of per-and polyfluoroalkyl substances. *Int. J. Environ. Res. Public Health* **2020**, *17*, 8117.
(23) Buck, R. C.; Franklin, J.; Berger, U.; Conder, J. M.; Cousins, I. T.; De Voogt, P.; Jensen, A. A.; Kannan, K.; Mabury, S. A.; van Leeuwen, S. P. Perfluoroalkyl and polyfluoroalkyl substances in the environment: terminology, classification, and origins. *Integr. Environ. Assess. Manage.* **2011**, *7*, 513−541.
(24) U.S. Environmental Protection Agency, Aquatic Life Criteria - Perfluorooctane Sulfonate (PFOS). https://www.epa.gov/wqc/aquatic-life-criteria-perfluorooctane-sulfonate-pfos, (accessed September 2022).
(25) Humphrey, W.; Dalke, A.; Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graphics* **1996**, *14*, 33−38.
(26) Shmilovich, K.; Mansbach, R. A.; Sidky, H.; Dunne, O. E.; Panda, S. S.; Tovar, J. D.; Ferguson, A. L. Discovery of self-assembling π-conjugated peptides by active learning-directed coarse-grained molecular simulation. *J. Phys. Chem. B* **2020**, *124*, 3873−3891.
(27) Dasetty, S.; Coropceanu, I.; Portner, J.; Li, J.; de Pablo, J. J.; Talapin, D.; Ferguson, A. L. Active learning of polarizable nanoparticle phase diagrams for the guided design of triggerable self-assembling superlattices. *Mol. Syst. Des. Eng.* **2022**, *7*, 350−363.
(28) Mohr, B.; Shmilovich, K.; Kleinwächter, I. S.; Schneider, D.; Ferguson, A. L.; Bereau, T. Data-driven discovery of cardiolipin-selective small molecules by computational active learning. *Chem. Sci.* **2022**, *13*, 4498−4511.
(29) Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 045024.
(30) Pfaendtner, J.; Bonomi, M. Efficient Sampling of High-Dimensional Free-Energy Landscapes with Parallel Bias Metadynamics. *J. Chem. Theory Comput.* **2015**, *11*, 5062−5067.
(31) Martins de Oliveira, V.; Liu, R.; Shen, J. Constant pH molecular dynamics simulations: current status and recent applications. *Curr. Opin. Struct. Biol.* **2022**, *77*, 102498.
(32) Cheng, J.; Psillakis, E.; Hoffmann, M.; Colussi, A. Acid dissociation versus molecular association of perfluoroalkyl oxoacids: environmental implications. *J. Phys. Chem. A* **2009**, *113*, 8152−8156.
(33) Lindim, C.; Van Gils, J.; Cousins, I. T. Europe-wide estuarine export and surface water concentrations of PFOS and PFOA. *Water Res.* **2016**, *103*, 124−132.
(34) Lampert, D. J.; Frisch, M. A.; Speitel, G. E., Jr Removal of perfluorooctanoic acid and perfluorooctane sulfonate from wastewater by ion exchange. *Pract. Period. Hazard., Toxic, Radioact. Waste Manage.* **2007**, *11*, 60−68.
(35) Yin, S.; López, J. F.; Solís, J. J. C.; Wong, M. S.; Villagrán, D. Enhanced adsorption of PFOA with nano MgAl2O4@ CNTs: Influence of pH and dosage, and environmental conditions. *J. Hazard. Mater. Adv.* **2023**, *9*, 100252.
(36) City of Chicago, Department of Water Management - Bureau of Water Supply, 2023 Q1-Q2 Comprehensive Chemical Analysis. https://www.chicago.gov/city/en/depts/water/supp_info/water_quality_resultsandreports/comprehensive_chemicalanalysis.html, (accessed August 2023).
(37) *Guidelines for Drinking-Water Quality: Fourth ed. Incorporating the First and Second Addenda*; World Health Organization: Genève, Switzerland, 2022.
(38) Moosavi-Movahedi, A.; Gharanfoli, M.; Nazari, K.; Shamsipur, M.; Chamani, J.; Hemmateenejad, B.; Alavi, M.; Shokrollahi, A.; Habibi-Rezaei, M.; Sorenson, C.; Sheibani, N. A distinct intermediate of RNase A is induced by sodium dodecyl sulfate at its pKa. *Colloids Surf., B* **2005**, *43*, 150−157.
(39) Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem. A* **1993**, *97*, 10269−10280.
(40) Frisch, M. J.et al. *Gaussian 16 Revision C.01*; Gaussian Inc.: Wallingford CT, 2016.
(41) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157−1174.
(42) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graphics Modell.* **2006**, *25*, 247−260.
(43) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *J. Phys. Chem. B* **2009**, *113*, 6378−6396.
(44) Ehlert, S.; Stahn, M.; Spicher, S.; Grimme, S. Robust and efficient implicit solvation model for fast semiempirical methods. *J. Chem. Theory Comput.* **2021**, *17*, 4250−4261.
(45) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Generalized born solvation model SM12. *J. Chem. Theory Comput.* **2013**, *9*, 609−620.
(46) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1−2*, 19−25.
(47) Sousa da Silva, A. W.; Vranken, W. F. ACPYPE - AnteChamber PYthon Parser interfacE. *BMC Res. Notes* **2012**, *5*, No. 367.
(48) Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R. A.; Parrinello, M. PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Comput. Phys. Commun.* **2009**, *180*, 1961−1972.
(49) Tsuzuki, S.; Uchimaru, T. Magnitude of attraction in CF4-CH4 interactions: Are CF4-CH4 interactions weaker than average of CF4-CF4 and CH4-CH4 interactions? *J. Fluorine Chem.* **2020**, *231*, 109468.
(50) Robalo, J. R.; Streacker, L. M.; Mendes de Oliveira, D.; Imhof, P.; Ben-Amotz, D.; Verde, A. V. Hydrophobic but Water-Friendly: Favorable Water−Perfluoromethyl Interactions Promote Hydration Shell Defects. *J. Am. Chem. Soc.* **2019**, *141*, 15856−15868.
(51) Robalo, J. R.; Mendes de Oliveira, D.; Imhof, P.; Ben-Amotz, D.; Vila Verde, A. Quantifying how step-wise fluorination tunes local solute hydrophobicity, hydration shell thermodynamics and the quantum mechanical contributions of solute−water interactions. *Phys. Chem. Chem. Phys.* **2020**, *22*, 22997−23008.
(52) Kingma, D. P.; Welling, M. Auto-Encoding Variational Bayes. 2022, arXiv:1312.6114. arXiv.org e-Print archive. https://arxiv.org/abs/1312.6114.
(53) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268−276.
(54) Irwin, J. J.; Shoichet, B. K. ZINC−a Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.
(55) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757−1768.
(56) Irwin, J. J.; Tang, K. G.; Young, J.; Dandarchuluun, C.; Wong, B. R.; Khurelbaatar, M.; Moroz, Y. S.; Mayfield, J.; Sayle, R. A. ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *J. Chem. Inf. Model.* **2020**, *60*, 6065−6073.
(57) Tang, Y.; Kim, J. Y.; IP, C. K.; Bahmani, A.; Chen, Q.; Rosenberger, M. G.; Esser-Kahn, A. P.; Ferguson, A. L. Data-driven discovery of innate immunomodulators via a closed-loop system combining machine learning and high throughput screening. *Chem. Sci.* **2023**, DOI: 10.1039/D3SC03613H.
(58) Landrum, G.et al. *rdkit/rdkit: 2022_09_5 (Q3 2022) Release*, 2023.
(59) Jolliffe, I. T.; Cadima, J. Principal component analysis: a review and recent developments. *Philos. Trans. R. Soc., A* **2016**, *374*, 20150202.

https://doi.org/10.1021/acs.jced.3c00404
J. Chem. Eng. Data XXXX, XXX, XXX−XXX

(60) Deringer, V. L.; Bartók, A. P.; Bernstein, N.; Wilkins, D. M.; Ceriotti, M.; Csányi, G. Gaussian Process Regression for Materials and Molecules. *Chem. Rev.* **2021**, *121*, 10073−10141.
(61) Rasmussen, C. E.; Williams, C. K. *Gaussian Processes for Machine Learning; Adaptive Computation and Machine Learning*; MIT Press: Cambridge, MA, 2005.
(62) Pedregosa, F.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(63) Paria, B.; Kandasamy, K.; Póczos, B. In *A Flexible Framework for Multi-objective Bayesian Optimization Using Random Scalarizations*, Proceedings of The 35th Uncertainty in Artificial Intelligence Conference, 2020; pp 766−776.
(64) Shmilovich, K.; Yao, Y.; Tovar, J. D.; Katz, H. E.; Schleife, A.; Ferguson, A. L. Computational discovery of high charge mobility self-assembling π-conjugated peptides. *Mol. Syst. Des. Eng.* **2022**, *7*, 447−459.
(65) Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.* **2002**, *3*, 397−422.
(66) Cox, D. D.; John, S. In *SDO: AStatistical Method for Global Optimization*, IEEE International Conference on Systems, Man, and Cybernetics, 1997; pp 315−329.
(67) Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R. P.; De Freitas, N. Taking the human out of the loop: A review of Bayesian optimization. *Proc. IEEE* **2016**, *104*, 148−175.
(68) Shmilovich, K.; Panda, S.; Stouffer, A.; Tovar, J. D.; Ferguson, A. L. Hybrid computational-experimental data-driven design of self-assembling π-conjugated peptides. *Digital Discovery* **2022**, *1*, 448−462.
(69) Bhattacharjee, H.; Vlachos, D. G. Thermochemical Data Fusion Using Graph Representation Learning. *J. Chem. Inf. Model.* **2020**, *60*, 4673−4683.
(70) Li, C.; Voth, G. A. Accurate and Transferable Reactive Molecular Dynamics Models from Constrained Density Functional Theory. *J. Phys. Chem. B* **2021**, *125*, 10471−10480.
(71) Raniolo, S.; Limongelli, V. Ligand binding free-energy calculations with funnel metadynamics. *Nat. Protoc.* **2020**, *15*, 2837−2866.
(72) Weiss-Errico, M. J.; O'Shea, K. E. Detailed NMR investigation of cyclodextrin-perfluorinated surfactant interactions in aqueous media. *J. Hazard. Mater.* **2017**, *329*, 57−65.
(73) Zheng, Z.; Yu, H.; Geng, W.-C.; Hu, X.-Y.; Wang, Y.-Y.; Li, Z.; Wang, Y.; Guo, D.-S. Guanidinocalix [5] arene for sensitive fluorescence detection and magnetic removal of perfluorinated pollutants. *Nat. Commun.* **2019**, *10*, No. 5762.
(74) Alzate-Carvajal, N.; Park, J.; Bargaoui, I.; Rautela, R.; Comeau, Z. J.; Scarfe, L.; Menard, J.-M.; Darling, S. B.; Lessard, B. H.; Luican-Mayer, A. Arrays of Functionalized Graphene Chemiresistors for Selective Sensing of Volatile Organic Compounds. *ACS Appl. Electron. Mater.* **2023**, *5*, 1514−1520.
(75) Wilhelm, E.; Battino, R.; Wilcock, R. J. Low-pressure solubility of gases in liquid water. *Chem. Rev.* **1977**, *77*, 219−262.
(76) Xiao, L.; Ling, Y.; Alsbaiee, A.; Li, C.; Helbling, D. E.; Dichtel, W. R. β-Cyclodextrin polymer network sequesters perfluorooctanoic acid at environmentally relevant concentrations. *J. Am. Chem. Soc.* **2017**, *139*, 7689−7692.
(77) Ching, C.; Klemes, M. J.; Trang, B.; Dichtel, W. R.; Helbling, D. E. β-cyclodextrin polymers with different cross-linkers and ion-exchange resins exhibit variable adsorption of anionic, zwitterionic, and nonionic PFASs. *Environ. Sci. Technol.* **2020**, *54*, 12693−12702.
(78) Ching, C.; Lin, Z.-W.; Dichtel, W. R.; Helbling, D. E. Evaluating the Performance of Novel Cyclodextrin Polymer Granules to Remove Perfluoroalkyl Acids (PFAAs) from Water. *ACS ES&T Engg.* **2023**, *3*, 661−670.
(79) United States Environmental Protection Agency, Working List of PFAS Chemicals With Research Interest and Ongoing Work by EPA. https://www.epa.gov/chemical-research/working-list-pfas-chemicals-research-interest-and-ongoing-work-epa, (accessed August 2023).

https://doi.org/10.1021/acs.jced.3c00404
J. Chem. Eng. Data XXXX, XXX, XXX−XXX