



Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/245144/>

Version: Published Version

Article:

Zarghami Dehaghani, M., Fabiani, T., Tsaftaris, S. et al. (2026) Screening of large covalent organic frameworks databases for urea removal from water using molecular simulations and machine learning. Digital Discovery. ISSN: 2635-098X

<https://doi.org/10.1039/d6dd00159a>

Reuse

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:

<https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



Cite this: DOI: 10.1039/d6dd00159a

Screening of large covalent organic frameworks databases for urea removal from water using molecular simulations and machine learning

Maryam Zarghami Dehaghani,^a Thomas Fabiani,^a Sotirios Tsafaris,^b Simone Dimartino^c and Maria Grazia De Angelis^a

Efficient removal of urea from spent dialysate is a pivotal step toward the realization of wearable artificial kidneys. Covalent organic frameworks (COFs), with their tunable porosity and chemical versatility, represent a promising class of adsorbents for this purpose. In this work, high-throughput molecular simulation insertions were used to rank 993 COFs from the CORE-COF database for their infinite dilution urea binding strength and ideal urea/water selectivity, indicating COF-C6, EB-COF:Cl, COF-F6, and OH-TPB-BPTA-COF as the best candidates. The analysis of CORE-COF descriptors revealed that oxygen-to-carbon ratio (O/C), carbon fraction (C_{frac}), and pore limiting diameter (PLD) are the dominant predictors of urea adsorption. The simulated urea binding strength was used to train and benchmark several ML models, using a nested cross validation. The Random Forest (RF) model was the one with the best performance ($R_{\text{train}}^2 = 0.92$, $R_{\text{test}}^2 = 0.62$), although its accuracy is less satisfactory for the frameworks with high urea adsorption, that are underrepresented in the dataset. Improved performance in this regime will likely require additional training data targeting these rare structures and/or more informative interaction-sensitive features, which are still a challenge for COF structures. Using the best-performing RF model, we rapidly screened a larger dataset of 69 840 hypothetical COFs (hCOFs), which has limited feature-space overlap with CORE-COFs, thus representing a stringent test. Although the model systematically overestimates the hCOFs urea binding strength calculated from simulations, it allows to identify clear chemical patterns that most top hCOF candidates have, such as amide linkers and dia/hcb topologies. The combined simulation–ML strategy here proposed offers a scalable workflow to accelerate the discovery of optimal COFs for urea separation and dialysate regeneration in wearable artificial kidneys.

Received 31st March 2026
Accepted 11th August 2026

DOI: 10.1039/d6dd00159a

rsc.li/digitaldiscovery

1. Introduction

Worldwide, the hemodialysis (HD) population is rising quickly, exceeding 2.5 million in 2020 and projected to reach 5.4 million by 2030.¹ HD sustains patients with kidney failure by removing uremic toxins, balancing electrolytes, and controlling fluid volume.^{2,3} However, the conventional HD process consumes vast amounts of water (280–500 L) and requires lengthy, hospital-based treatment sessions, limiting patient mobility and quality of life.^{4,5} Among the various toxins to be removed,

urea presents the greatest challenge: it is small, uncharged, highly polar, and has chemical properties similar to water, making its selective separation difficult.⁶ Efficient and selective adsorption of urea from dialysate is therefore a pivotal step toward compact, high-performance regeneration systems that can support continuous dialysis to eliminate the need for direct water, potentially revolutionizing renal replacement therapy.⁷

Dialysate regeneration units, which purify and recirculate the spent dialysate, offer a transformative solution by dramatically reducing water usage and enabling portable or wearable artificial kidney (WAK) systems.⁸ Emerging developments in zeolites,^{9,10} MXenes,¹¹ covalent organic frameworks (COFs),¹² and metal organic frameworks (MOFs)¹³ have renewed momentum for sorbent-mediated dialysate purification and are promising to enable the next-generation WAK devices. COFs are an emerging class of crystalline nanoporous materials distinguished by their tunable pore architecture and chemistry with well-defined pore sizes (typically 0.5–4.7 nm), high surface area and exceptional thermal stability.¹⁴ These characteristics make COFs highly promising for adsorption and separation

^aInstitute for Materials and Processes, School of Engineering, University of Edinburgh, King's Buildings, Robert Stevenson Road, EH9 3FB, Edinburgh, Scotland, UK. E-mail: mzargham@ed.ac.uk

^bInstitute for Digital Communications, School of Engineering, University of Edinburgh, Alexander Graham Bell Building, King's Buildings, Thomas Bayes Road, EH9 3FG, Edinburgh, Scotland, UK

^cInstitute for Bioengineering, School of Engineering, University of Edinburgh, King's Buildings, Colin Maclaurin Road, EH9 3DW, Edinburgh, Scotland, UK

^dSchool of Chemical, Materials and Biological Engineering, University of Sheffield, S1 3JD, Sheffield, UK. E-mail: m.zarghamidehaghani@sheffield.ac.uk

processes, offering favorable performance such as selectivity and adsorption capacity.¹⁵ Unlike MOFs, COFs are constructed exclusively from covalent bonds, enabling precise control over functional properties, and imparting superior structural robustness especially in working environment.¹⁶ The pioneering synthesis of COF-1 and COF-5 by Cote *et al.*¹⁷ in 2005 showcased the utility of boronate ester linkages in producing defect-free, two-dimensional frameworks. Since then, advances in dynamic covalent chemistry have expanded the COF toolbox to include linkages such as imine, triazine, and β -ketoenamine, facilitating the fabrication of both two- and three-dimensional architectures.¹⁸ This structural diversity and chemical tunability allow COFs to be tailored toward specific molecular targets, enabling biomedical applications such as drug delivery,¹⁹ biosensing and molecular imaging,²⁰ enzyme immobilization for biocatalysis,²¹ and biomedical separation challenges such as urea removal from dialysate.^{22,23}

To systematically explore the rapidly expanding COF design space, several large-scale, computation-ready COF datasets have been created. For experimentally reported frameworks, the CORE-COF database (1242 structures) provides solvent-free, disorder-free structures, enabling direct use in molecular simulations.²⁴ Complementarily, the CURATED-COFs collection (871 structures) supplies experimentally reported COFs with DFT-optimized geometries and DDEC partial charges, facilitating consistent simulation workflows.²⁵ Notably, these two resources are not independent and contain substantial overlap in the underlying COF structures. Beyond experimentally reported materials, hypothetical COF libraries greatly expand accessible chemical and topological diversity for computational screening. For example, the Berkeley-COFs²⁶ dataset contains 69 840 in silico-assembled COF structures generated for high-throughput evaluation, whereas ReDD-COFFEE²⁷ offers a larger, ready-to-use collection of 268 687 COFs accompanied by system-specific, ab initio-derived force fields, enabling scalable adsorption screening.

While experimental methods are essential for characterizing adsorbents performance, they cannot feasibly capture the vast chemical and structural diversity achievable through combinatorial design and large-scale material databases.^{28,29} Moreover, large-scale COF datasets make systematic materials discovery feasible, as direct adsorption evaluation by molecular simulation across hundreds of thousands of structures remains computationally demanding, motivating high-throughput (HT) simulation workflows. Computational approaches, particularly HT Monte Carlo (MC)-based techniques, enable rapid prediction of adsorption properties and have become indispensable for screening candidate porous materials.^{30–32} Within this framework, Widom insertion simulations are commonly employed to estimate excess chemical potentials and Henry's constants at infinite dilution, providing low-loading adsorption metrics that inform the intrinsic urea uptake capacity.³³ Grand Canonical Monte Carlo (GCMC) simulations can then be used to predict urea adsorption capacity and full adsorption isotherms at finite-concentration conditions. Together, these two computational methods have been applied to identify high-performance organic frameworks for urea-separation

applications. For example, Yıldız *et al.*³⁴ explored adsorption-based separation in bio-MOFs, first identifying hydrophobic frameworks with low Henry's constants for water and then quantifying the urea and water uptakes *via* GCMC at 310 K and 1 bar. Their screening highlighted six promising bio-MOFs including adenine-based structures, bio-MOF-11 (YUVSUE), bio-MOF-12 (BEYSEF), and a dicyanamide-based MOF with the highest urea/water selectivity. Complementarily, Palabıyık *et al.*³⁵ combined GCMC with equilibrium molecular dynamics to investigate both adsorption- and membrane-based mechanisms, further underscoring the potential of bio-MOFs for urea–water separation. In our previous work we employed Widom insertion simulations to evaluate the excess chemical potential at infinite dilution for 560 experimentally reported COFs from the CORE-COFs database and MOFs with respect to water and urea at 310 K.³⁶ Their analysis revealed that urea uptake and urea/water selectivity increased significantly with fluorine content in COFs, whereas other compositional or structural parameters showed little correlation with adsorption performance. Two frameworks, COF-F6 and TF-DH2DPr, were subsequently investigated using molecular dynamics simulations, which elucidated the roles of electrostatic and van der Waals interactions and identified the preferential binding sites governing urea–urea and urea–framework interactions.

The high computational cost of performing MC simulations on large datasets remains a key challenge, limiting scalability to explore large datasets for material discovery. In this context, machine learning (ML) has emerged as a powerful and more efficient predictive strategy capable of capturing complex structure–property relationships and accelerating the identification of high-performance adsorbents.³⁷ To date, ML has been extensively applied to COFs in applications such as gas storage and separation, demonstrating strong predictive capability.^{38,39} Yet, its application to urea adsorption in COFs remains largely unexplored, highlighting the need for systematic benchmarking to establish reliable data-driven frameworks for future development.

In the present work, we first probed the CORE-COF database using Widom insertion simulations to calculate the urea binding strength ($\mu_{\text{urea}}^{\text{ex}}$), water solubility ($\mu_{\text{water}}^{\text{ex}}$) and an indicator of energy-based urea/water selectivity ($\mu_{\text{urea}}^{\text{ex}} - \mu_{\text{water}}^{\text{ex}}$) at 310 K. Next, an initial set of 59 structural and chemical descriptors was considered, followed by feature selection using Pearson correlation, which reduced the set to 23 key features. Several regression machine-learning models were then developed using the urea binding strength as the target variable. Finally, the best-performing model was applied to rapidly screen the Berkeley COFs database.

2. Methodology

2.1. COFs datasets

COF data were sourced from both the latest experimental CORE-COFs database comprising 1242 COFs,^{24,40} as well as the computationally generated Berkeley-COFs database (hCOFs) which includes 69 840 hypothetical COFs.²⁶ Given that both water and urea are present as adsorbates in urea separation, we

focused on COFs with a pore limiting diameter (PLD) greater than 3.8 Å and an available surface area (ASA) greater than 0, ensuring that all target molecules can pass through the pores. Structures containing boron-based linkages (*e.g.*, boroxine, borazine, and boronic esters) were excluded due to their susceptibility to hydrolysis in aqueous environments and poor stability under application conditions. Additionally, COFs containing hazardous and toxin heavy metals were removed to ensure material safety of the application. After applying these criteria, 993 CORE-COFs were selected for molecular simulation and follow up training of the machine learning models. Then, the trained ML models were applied to the predictive screening of the 69 840 hCOFs.

2.2. Molecular simulation

Both the COF crystal structures and the adsorbate molecules of water and urea were treated as rigid entities, an assumption justified by the limited conformational flexibility of the aromatic ring-based COF building blocks and the small, non-rotatable adsorbate molecules. The rigidity of these frameworks, which promotes stacking during synthesis, allows the total energy to be effectively described by intermolecular interactions. Alone non-bonded interactions were modeled using the Lennard-Jones 12–6 (LJ) potential in combination with Ewald summation for electrostatics. A cutoff distance of 12 Å was applied to both van der Waals (VdW) and electrostatic interactions. The Lorentz–Berthelot mixing rules were employed for cross-interactions. COF crystal structures were parameterized using the DREIDING force field,⁴¹ following the protocol adopted in a previous work on COFs.⁴² In contrast to our previous work,³⁶ where partial charges for COFs were assigned using the semi-empirical EQeq method,⁴³ this work uses Partial charges obtained from the more rigorous density-derived electrostatic and chemical (DDEC) partitioning scheme.⁴⁴ The values of DDEC charges were predicted by PACMAN package (PACMAN-charge 1.4.2), a machine-learning atomic-charge predictor for nanoporous materials that uses crystal graph convolutional neural networks. This model achieved a very low test-set mean absolute error of 0.0055 e relative to reference charges and demonstrated consistent performance across DDEC6, Bader, and CM5 charge schemes for diverse nanoporous structures. This benchmarking against DFT-derived charge calculations supports the robustness of PACMAN as a reliable and computationally efficient alternative for assigning partial charges in high-throughput molecular simulations. Nevertheless, as with any machine-learned approximation, the use of PACMAN-predicted charges may introduce a small additional uncertainty compared with performing explicit structure-specific DFT charge calculations for every COF.⁴⁵

The five-site Transferable Interaction Potential (TIP5P) model was employed to represent water molecules.⁴⁶ Periodic boundary conditions were applied, and COFs were assumed as rigid in all simulations. The urea molecule was modelled using parameters developed by Weerasinghe and Smith to accurately reproduce its behavior in aqueous solution.⁴⁷

The Widom insertion method was applied to compute the excess chemical potential at 310 K using the RASPA software (RASPA3 v3.0.30)⁴⁸ according to:

$$\mu_i^{\text{ex}} = -k_{\text{B}}T \ln \langle \exp(-\beta \Delta U) \rangle_{N\Delta S_{N+1}} \quad (1)$$

where index i denotes either water or urea, k_{B} is the Boltzmann constant, T is the absolute temperature, $\beta = \frac{1}{k_{\text{B}}T}$, and ΔU denotes the energy variation upon inserting an adsorbate molecule into the framework. The angular brackets denote an ensemble average over configurations of the system with $N+1$ particles.

For each COF, 5×10^4 random insertions of an adsorbate molecule in the empty framework were performed. This number of Monte Carlo insertions is sufficient to thoroughly sample adsorption sites and achieve convergence of the averaged loading and energy; additional insertions did not produce statistically significant changes in the computed properties (*i.e.*, the results were equilibrated within the selected uncertainty, Table S1).⁴⁹

The Henry's law constants $K_{\text{H},i}^0$ at infinite dilution can be computed from μ_i^{ex} as below:⁴⁸

$$K_{\text{H},i}^0 = \frac{1}{RT\rho_f} \ln \left(\frac{-\mu_i^{\text{ex}}}{RT} \right) \quad (2)$$

where R and ρ_f correspond to the universal gas constant and the framework density, respectively. A more negative μ_i^{ex} increases the Boltzmann-weighted probability that a test molecule can be inserted into the pore without incurring an unfavorable energy penalty. In thermodynamic terms, μ_i^{ex} is the Gibbs free energy change to transfer one molecule from an ideal state into an empty adsorption site at the same T , therefore incorporates both enthalpic and entropic contributions. As μ_i^{ex} becomes more negative, the zero-pressure isotherm becomes steeper and the Henry constant $K_{\text{H,urea}}$ increases.

In the dialysate regeneration process through adsorption, water acts as the primary competitor for the binding of water-soluble toxins like urea. Therefore, in addition to evaluating the intrinsic urea affinity, we define an energy-based ideal urea/water selectivity using the difference between the excess chemical potentials of urea and water, which is equivalent to the logarithmic ratio of their respective Henry's law constants:

$$\mu_{\text{urea}}^{\text{ex}} - \mu_{\text{water}}^{\text{ex}} = RT \ln \left(\frac{K_{\text{H,urea}}^0}{K_{\text{H,water}}^0} \right) \quad (3)$$

A larger $K_{\text{H,urea}}^0$ (*i.e.* a more negative $\mu_{\text{urea}}^{\text{ex}}$) indicates stronger thermodynamic affinity of the framework for urea at infinite dilution. Likewise, a larger $K_{\text{H,water}}^0$ (*i.e.*, a more negative $\mu_{\text{water}}^{\text{ex}}$) indicates stronger framework–water affinity, corresponding to greater infinite dilution solubility of water in the COF.

The term of $\mu_{\text{urea}}^{\text{ex}} - \mu_{\text{water}}^{\text{ex}}$ quantity was used as the energy-based selectivity target in the machine-learning analysis. More negative values of $\mu_{\text{urea}}^{\text{ex}} - \mu_{\text{water}}^{\text{ex}}$ indicate that urea insertion is thermodynamically more favorable than water insertion and interactions between the two molecules do not take place due to

extreme dilution. This descriptor is directly related to the dimensionless ideal Henry's-law selectivity:³⁵

$$S_{\text{urea/water}}^0 = \frac{K_{\text{H,urea}}^0}{K_{\text{H,water}}^0} = \exp\left[-\frac{\mu_{\text{urea}}^{\text{ex}} - \mu_{\text{water}}^{\text{ex}}}{RT}\right] \quad (4)$$

Thus, $\mu_{\text{urea}}^{\text{ex}} - \mu_{\text{water}}^{\text{ex}}$ has units of energy, whereas $S_{\text{urea/water}}^0$ is dimensionless.

2.3. Machine learning

The CORE-COFs dataset contained structural descriptors (void fraction (V_f), porosity, accessible surface area (ASA), density, dimensionality (2D or 3D), topology, largest cavity diameter (LCD), and pore limiting diameter (PLD)) as well as chemical descriptors in terms of elemental composition (carbon (C), nitrogen (N), oxygen (O), sulfur (S), hydrogen (H), and chlorine (Cl) fractions, Bromine (Br), sodium (Na), phosphor (P), Zinc (Zn), silicon (Si), fluoride (F), copper (Cu)). The average value of the Lennard-Jones σ parameter (AVG_SIG) and the average of the square root of the Lennard-Jones ε parameter (AVG_SQRT_EPS) were also included as descriptors,⁵⁰ along a set of hybrid descriptors obtained by combining chemical and structural parameters. A complete list of 59 initial descriptors and their definitions is provided in Table S2. The initial targets were

$\mu_{\text{urea}}^{\text{ex}}$ (urea binding strength), $\mu_{\text{water}}^{\text{ex}}$ (water solubility), and $\mu_{\text{urea}}^{\text{ex}} - \mu_{\text{water}}^{\text{ex}}$ (urea/water selectivity). For feature selection and to reduce redundancy and target interdependence, we computed Pearson correlation coefficients (r) for all pairs: feature–target, feature–feature, and target–target (Fig. S1).⁵¹

Among the feature–feature pairs exhibiting high inter-correlation ($r > 0.8$), the feature with the weaker correlation to the target property was excluded to reduce redundancy. For instance, LCD and PLD were strongly correlated ($r = 0.95$), leading to the removal of LCD from the feature set. Although both descriptors of V_f and porosity are derived from pore volume, they use different denominators: V_f ($\text{cm}^3 \text{ g}^{-1}$) is normalized by framework mass, whereas porosity is normalized by unit-cell volume. As a result, two COFs can have similar porosity but different V_f if their framework densities differ, or similar V_f but different porosity if their cell volumes differ. The moderate correlation (Pearson $r = 0.65$) empirically confirms that they encode complementary geometric information rather than being redundant.

Moreover, the descriptors having almost zero correlation with targets such as topology were removed. Fig. 1 presents the Pearson r coefficients among the 23 selected features. Histogram distribution of these 23 features are demonstrated in Fig. S2.

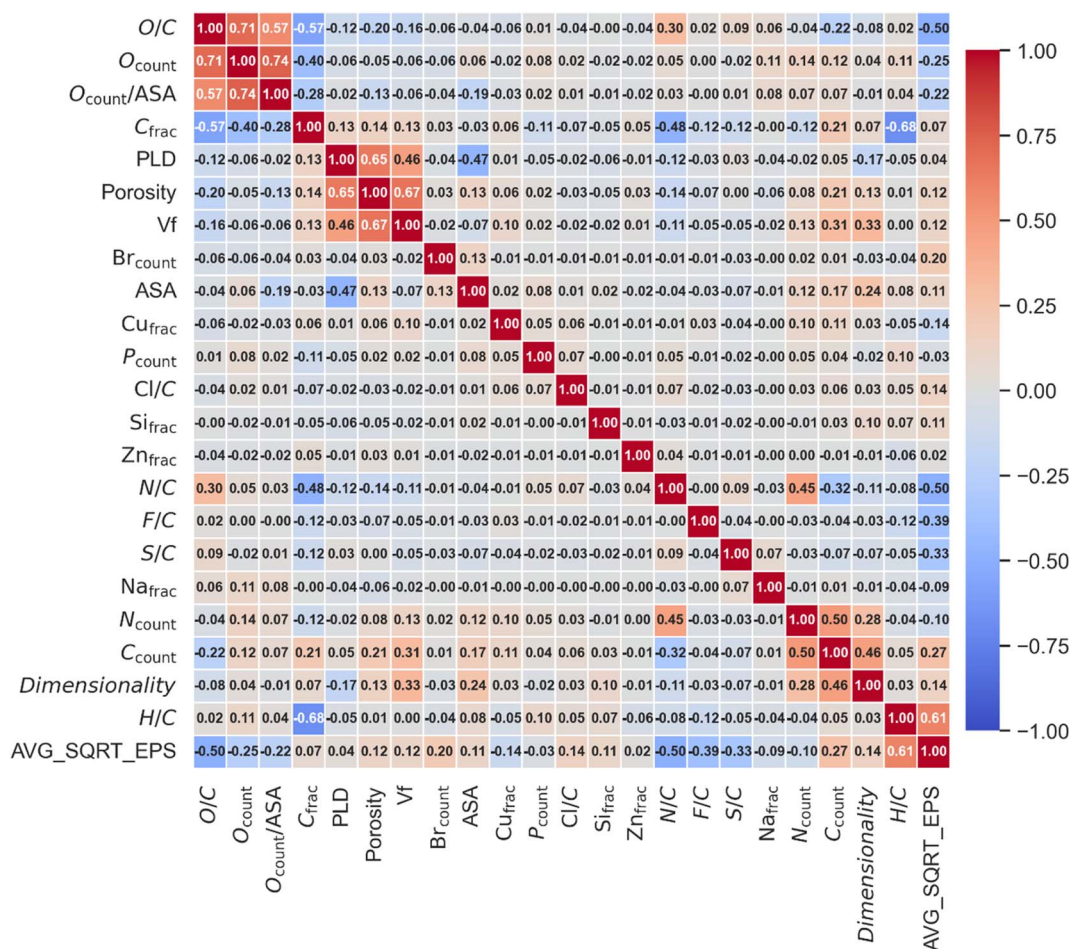


Fig. 1 Pearson correlation matrix of features. The definition of each feature is reported in Table S2.

To assess possible descriptor-level leakage before model training, we performed an additional feature-space audit using only the selected input descriptors. The analysis, including principal component analysis (PCA)^{52,53} visualization of the train/test feature-space distribution and deterministic feature-vector hashing,⁵⁴ showed that exact descriptor-level duplication across the train/test split was negligible (Fig. S3). Therefore, widespread leakage due to identical input-feature vectors was not observed. Details of this analysis are provided in the SI.

ML models of Random Forests (RF), Gradient Boosting Regression (GB), Support Vector regressor (SVR), Extra Trees (ET), and HistGradientBoosting (HGB) were implemented using the *scikit-learn* 1.8.0 (*sklearn*)⁵⁵ library in Python 3.14.5. In addition, a neural network (NN) model was developed using the PyTorch framework.⁵⁶ Hyperparameter tuning for the regression models was performed using Bayesian optimization, implemented through BayesSearchCV from *scikit-optimize* 0.10.2 (*skopt*).⁵⁷ For the neural-network model, Bayesian optimization was also used to tune architecture and training hyperparameters, including the number of layers, hidden dimension, dropout rate, learning rate, weight decay, and batch size. Model training was conducted using a feature set comprising selected structural descriptors, chemical composition features, and their hybrids. To improve model generalisability and minimise overfitting, we employed fivefold nested cross-validation during model development, using an inner loop for hyperparameter optimisation and an outer loop for unbiased performance evaluation.

Therefore, the dataset was first divided into independent training and test subsets using an 80:20 split with a fixed random seed of 43; the test set was kept separate from all model training and hyperparameter optimization steps. Nested cross-validation was then applied only to the training set. In each outer fold, the outer validation fold was held out completely using shuffled five-fold cross-validation with a fixed random seed of 42, while Bayesian hyperparameter optimization was performed only within the corresponding inner folds of the outer-training subset. The inner cross-validation and Bayesian optimization used fixed random seeds of 1 and 0, respectively, to ensure reproducibility. Therefore, the outer-fold validation data were not used for hyperparameter selection. After model selection, the final model was retrained on the full training set and evaluated once on the untouched hold-out test set.⁵⁸

Numerical feature scaling was implemented inside a *scikit-learn* pipeline together with the regression model. Therefore, the scaler was fitted only on the training portion of each fold and then applied to the corresponding validation or test fold, preventing information from the validation/test data from entering the preprocessing step. The same leakage-free procedure was used during the final hold-out test evaluation. The categorical variable dimensionality (2D vs. 3D) was encoded as a binary dummy indicator with 2D = 0 and 3D = 1.⁵⁹ Model accuracy was evaluated using three performance metrics: the coefficient of determination (R^2), the mean absolute error (MAE), and the root mean square error (RMSE).⁶⁰

Regarding feature selection, we have clarified that the final descriptor set was fixed before model benchmarking and then

used consistently across all models. The selected descriptor set was not re-optimized separately for individual models, ensuring that differences in model performance were due to the learning algorithms and hyperparameter settings rather than changes in input features. Numerical scaling of (StandardScaler from *Scikit-learn*) was implemented within the model pipeline, so the scaler was fitted only on the training portion of each fold and then applied to the corresponding validation fold, preventing information leakage from validation or test data into the pre-processing step. To make this point transparent, we explicitly described the descriptor-selection step separately from the nested model-selection procedure and R^2 , RMSE, and MAE values corresponded to model training and hyperparameter optimization performed using the fixed selected descriptor set. Therefore, the reported R^2 , RMSE, and MAE values correspond to model training and hyperparameter optimization performed using the same fixed selected descriptor set.

The Uniform Manifold Approximation and Projection (UMAP) dimensionality reduction method was applied to project the high-dimensional features of CORE-COFS and Berkeley-COFS onto a two-dimensional space. UMAP employs exponential probability distribution to compute the similarity between high dimensional data points:⁶¹

$$P_{ij} = \exp\left(-\frac{d(x_i, x_j) - \rho_i}{\sigma_i}\right) \quad (5)$$

where $d(x_i, x_j)$ denotes the distance between the i -th and j -th data points and ρ shows the distance between i -th data points and its first nearest neighbour, and σ_i is the local smoothness or scaling parameter for point i . In this work, the final selected descriptors were first standardized using a StandardScaler fitted only on the CORE-COF descriptor matrix, and the same transformation was then applied to the Berkeley-COF descriptors.

To further evaluate the proximity of the selected Berkeley-COFS to the CORE-COF training domain, a nearest-neighbor distance analysis was performed in the two-dimensional UMAP space. For each selected Berkeley-COF, the Euclidean distance to the nearest CORE-COF in UMAP space was calculated using a NearestNeighbors search from *Scikit-learn*:⁶²

$$d_j^{\text{UMAP}} = \min_i \sqrt{(\text{UMAP1}_j - \text{UMAP1}_i)^2 + (\text{UMAP2}_j - \text{UMAP2}_i)^2} \quad (6)$$

where (j) represents a selected Berkeley-COF and (i) represents a CORE-COF. The Euclidean nearest-neighbor distance in the standardized selected-descriptor space (with `n_neighbors` = 15, `metric` = Euclidean, `random_state` = 42, `min_dist` = 0.1) was used as the primary quantitative applicability-domain metric.

This procedure produced a distribution of nearest-neighbor distances that describes the internal spread and coverage of the CORE-COF training domain. The 95th percentile of this distribution corresponds to the distance below which 95% of CORE-COFS have another CORE-COF neighbor; therefore, it was used as the boundary of the main applicability domain. The 99th percentile provides a broader and more permissive boundary, including nearly all CORE-COFS except the most isolated

structures. The nearest-CORE distance was then calculated only for the selected Berkeley-COFs validated by Widom insertion simulations. Selected Berkeley-COFs with nearest-CORE distances larger than the 95th percentile threshold were considered outside the main CORE-COF applicability domain, while those below the 99th percentile threshold were interpreted as lying within a broader, lower-confidence region of the training domain. Predictions for selected Berkeley-COFs beyond these thresholds were therefore treated as increasingly extrapolative.

3. Results and discussion

3.1. Molecular simulation

Fig. 2 shows urea selectivity ($\mu_{\text{urea}}^{\text{ex}} - \mu_{\text{water}}^{\text{ex}}$) versus urea binding strength ($\mu_{\text{urea}}^{\text{ex}}$) of the CORE-COF dataset. It can be observed that for moderate values of $\mu_{\text{urea}}^{\text{ex}}$ and $\mu_{\text{urea}}^{\text{ex}} - \mu_{\text{water}}^{\text{ex}}$ are somewhat linearly correlated. However, in the strongly binding regime (very negative ($\mu_{\text{urea}}^{\text{ex}}$)) the data become more dispersed, and the relationship deviates from linearity. In general, by decreasing the values of $\mu_{\text{urea}}^{\text{ex}}$, the values of $\mu_{\text{urea}}^{\text{ex}} - \mu_{\text{water}}^{\text{ex}}$ decreased. According to this, we considered $\mu_{\text{urea}}^{\text{ex}}$ as the primary target for training the ML models, since it captures the dominant driving force for urea uptake. The majority of CORE-COFs showed $\mu_{\text{urea}}^{\text{ex}}$ and $\mu_{\text{urea}}^{\text{ex}} - \mu_{\text{water}}^{\text{ex}}$ less negative than -40 kJ mol^{-1} (93.59%) and -30 kJ mol^{-1} (96.79%), respectively.

3.2. Structure-properties relation

The feature importance on $\mu_{\text{urea}}^{\text{ex}}$ of the whole CORE-COFs dataset according to the RF model is illustrated in Fig. 3, providing valuable insights into the key descriptors governing urea binding strength.⁶³ Before model training, highly correlated descriptors were identified using Pearson correlation analysis, and among strongly correlated feature pairs, the descriptor with weaker relevance to the target was removed.

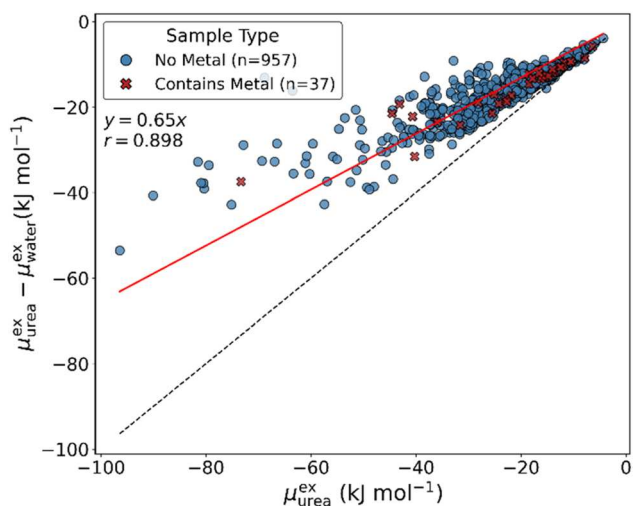


Fig. 2 Selectivity ($\mu_{\text{urea}}^{\text{ex}} - \mu_{\text{water}}^{\text{ex}}$) versus binding strength ($\mu_{\text{urea}}^{\text{ex}}$) of CORE-COFs. Bisect represents no interaction between COF and water (hydrophobic COFs). Red line is linear trend of data points.

This preprocessing step reduced descriptor redundancy and mitigated the influence of correlated variables on RF feature importance. However, descriptor-distribution effects cannot be fully excluded; therefore, the RF feature-importance values should be interpreted as model-derived descriptor associations rather than direct causal measures of urea adsorption. Accordingly, the trends discussed below are used to support qualitative structure–property interpretation within the CORE-COF dataset.

Among the 23 descriptors considered, the O/C ratio emerged as the most influential feature ($\sim 20\%$ of total importance) and showed a negative Pearson correlation with $\mu_{\text{urea}}^{\text{ex}}$ ($r = -0.38$). Because lower $\mu_{\text{urea}}^{\text{ex}}$ indicates more favorable urea insertion (stronger adsorption), this trend implies that oxygen-rich COFs enhance urea affinity. This strong dominance highlights the critical role of framework oxygen content in dictating COF–urea interactions: oxygen-containing groups provide hydrogen-bond acceptor sites, increasing the likelihood of favorable interactions with urea's polar functionalities (carbonyl and amine groups). The second most relevant descriptor is the carbon fraction (C_{frac} , $\sim 18\%$ relative importance), which is positively correlated with $\mu_{\text{urea}}^{\text{ex}}$ ($r = 0.32$). This indicates that increasingly carbon-rich (and typically less polar) pore environments tend to weaken urea–framework interactions, and thus reducing adsorption favorability.

The most relevant structural feature is PLD (10% relative importance), indicating that pore accessibility is a critical geometric parameter. An optimum PLD is therefore expected so that a very small PLD imposes strong excluded-volume constraints and restricts urea's orientational freedom, whereas very large PLD reduces confinement-driven contact with favorable COF interaction sites, leading to weaker binding strength $\mu_{\text{urea}}^{\text{ex}}$.⁶⁴

For Widom insertion (infinite dilution), PLD is often more informative than void fraction and porosity because it captures accessibility and the bottleneck constraint that governs where an inserted molecule can physically fit. PLD represents the minimum restricting aperture along a connected diffusion path.³⁰ Thus, even in frameworks with large cavities and high void fraction, if those cavities are connected through narrow channel, PLD determines whether urea insertions are feasible and how confined the accessible regions are. As a result, PLD strongly influences the $\mu_{\text{urea}}^{\text{ex}}$ through excluded-volume restrictions and reduced configurational freedom in tight bottlenecks. By contrast, void fraction is a global measure of empty volume (often probe-dependent) and is therefore more closely related to how much urea molecules can be stored once access is already guaranteed.⁶⁵ In adsorption isotherms, higher accessible pore volume typically correlates better with high-loading capacity. However, at infinite dilution, void fraction alone can be a weaker structural predictor of affinity because $\mu_{\text{urea}}^{\text{ex}}$ is dominated by the most favorable accessible regions and the ability of the molecule to access them, rather than by the total amount of empty space.

In addition to O/C, other oxygen-related features such as $O_{\text{count}}/\text{ASA}$ and O_{count} showed moderate importance, emphasizing that besides the overall amount of oxygen, the accessibility of oxygen atoms (as captured by $O_{\text{count}}/\text{ASA}$) also is critical

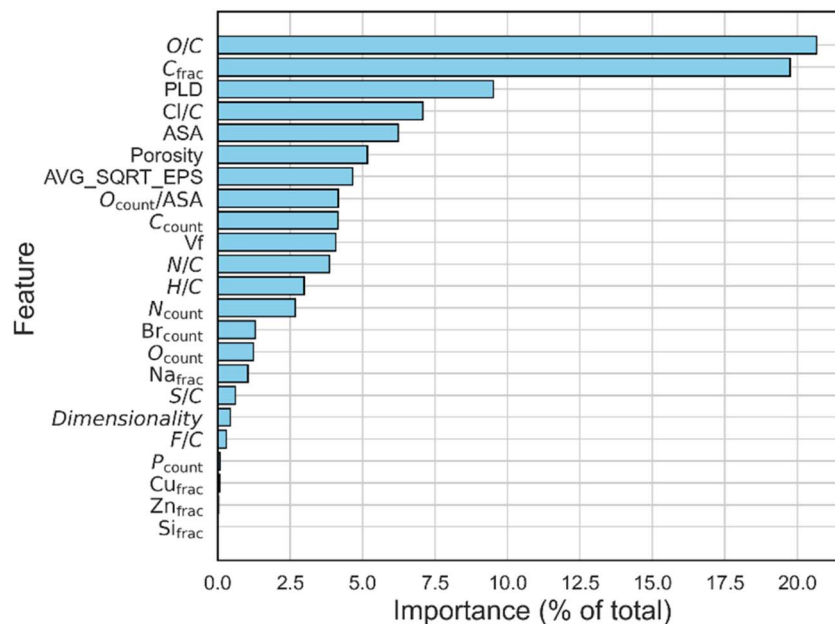


Fig. 3 Feature importance of Random Forest (RF) model for prediction of $\mu_{\text{urea}}^{\text{ex}}$.

for enhancing hydrogen-bonding interactions with urea molecules.

Additional descriptors such as the Cl/C ratio ($\approx 6\%$), accessible surface area (ASA, $\approx 5\%$), and the averaged Lennard-Jones interaction parameter (AVG_SQRT_EPS, $\approx 4\%$) also contributed to the prediction. The role of Cl/C suggests that the presence of chlorine functionalities caused enhancement of urea-framework interactions through noncovalent halogen bonding arises from anisotropic electronic densities (the σ -hole) on a halogen atom interacting with a Lewis base (*e.g.*, O or N in urea), composed of electrostatics (interaction between the σ -hole and a lone pair), induction (polarization), and dispersion contributions.^{66,67} The moderate impact of ASA indicates that adsorption capacity is partially surface-area-driven, while the contribution of AVG_SQRT_EPS highlights the relevance of van der Waals interactions in stabilizing urea adsorption within the pores.

Other compositional features such as H/C, Ncount, and N/C contributed moderately to the model, which is consistent with their established chemical roles in polarity and molecular geometry. For instance, hydrogen contributes indirectly by modulating framework polarity, while nitrogen (both within the framework and in urea) enhances the hydrogen-bonding network due to its electron-rich lone pairs. The contribution of N/C therefore aligns with the fact that nitrogen in urea participates actively in directional hydrogen bonding, thereby influencing the energetics of adsorption. Other chemical features (F/C, S/C, Cucount, Znfrac, P, Nafrac and Sifrac) as well as dimensionality did not significantly affect $\mu_{\text{urea}}^{\text{ex}}$.

The comparatively low RF importance of the fluorine-based descriptor (F/C) relative to Cl- and Br-based descriptors can be rationalized by considering the fundamentally different interaction mechanisms governing halogen-urea interactions

through σ -hole bonding. Although fluorine is the most electronegative halogen (Pauling electronegativity: $\chi_{\text{F}} = 3.98$, $\chi_{\text{Cl}} = 3.16$, $\chi_{\text{Br}} = 2.96$),⁶⁸ it is significantly less polarisable (static dipole polarisability α in $\text{\AA}^3$: $\alpha_{\text{F}} = 0.22$, $\alpha_{\text{Cl}} = 2.16$, $\alpha_{\text{Br}} = 3.29$).⁵⁰ As a result, Cl and Br are approximately one to two orders of magnitude more polarisable than F, which has important consequences for adsorption energetics. In the absence of solvent, urea adsorption in halogen-functionalized COFs is dominated not by purely electrostatic interactions, but by induction and dispersion contributions, both of which scale directly with atomic polarisability. The induction energy scales approximately as *e.g.*, $E_{\text{ind}} \approx -1/2\alpha E^2$,⁶⁹ where E is the electric field generated by the permanent dipole of urea, while the dispersion interaction follows $E_{\text{disp}} \approx -C_6/r^6$,⁷⁰ with the dispersion coefficient C_6 increasing strongly with polarizability α . Consequently, the much larger α values of Cl and Br lead to stronger and more variable attractive interactions with urea, particularly through dispersion-dominated stabilization within confined pore environments. Furthermore, higher polarisability enables σ -hole formation, allowing Cl and Br to participate effectively in halogen bonding with urea's carbonyl oxygen. These σ -hole interactions introduce directional and environment-sensitive binding motifs that vary with local geometry, pore confinement, and functional group orientation. Such variability enhances the statistical importance of Cl/C and Br-related features in the RF model. In contrast, fluorine's extremely low polarisability suppresses σ -hole development, making F a weak halogen-bond donor. As a result, F-urea interactions are dominated by short-range electrostatics that are relatively uniform across structures and less sensitive to local framework geometry. Importantly, while the electronegativity difference across the F-Cl-Br series spans only ~ 1 unit, the corresponding spread in polarisability spans nearly an order

of magnitude ($10\text{--}15\times$ spread in polarisability). This disparity explains why RF-based feature importance tracks polarisability rather than electronegativity. Chlorine and bromine enhance dispersion, induction, and halogen-bonding contributions in a way that varies significantly across different COFs, thereby exerting a stronger influence on $\mu_{\text{urea}}^{\text{ex}}$. Fluorine, despite its strong electronegativity, contributes more uniformly and weakly to urea binding, leading to its near-zero importance in the model. This interpretation aligns with the established trend that halogen-bond strength increases with the polarizability of the σ -hole donor halogen, following $\text{F} < \text{Cl} < \text{Br} < \text{I}$.⁷¹

Although O and S are both chalcogens and can serve as interaction sites for urea adsorption, the relative contributions to the hydrogen-bonding interaction energy differ substantially between O- and S-centred sites. Hydrogen bonding is stabilized by a balance of electrostatic, polarization/charge-transfer (induction), and dispersion contributions (opposed at short range by Pauli repulsion).⁷² Consequently, the higher feature importance assigned by the RF model to O-related descriptors can be rationalized by oxygen's higher electronegativity (Pauling $\chi_{\text{O}} = 3.44$ vs. $\chi_{\text{S}} = 2.58$),⁶⁸ larger permanent dipoles, and a more localized negative electrostatic potential at the lone-pair region. These properties amplify the electrostatic and induction/charge-transfer components of $\text{N}\cdots\text{H}_{\text{urea}}\cdots\text{O}$ contacts, making them strongly directional and highly sensitive to distance and angle, thereby increasing the variance and predictive value of O-centred local descriptors for $\mu_{\text{urea}}^{\text{ex}}$. In contrast, sulfur despite its higher polarizability ($\alpha_{\text{S}} \approx 2.99 \text{ \AA}^3$ vs. $\alpha_{\text{O}} \approx 0.57 \text{ \AA}^3$),⁶⁸ typically exhibits weaker and less localized electrostatic stabilization toward neutral donors and often forms longer, less constrained $\text{N}\cdots\text{H}_{\text{urea}}\cdots\text{S}$ contacts. The net attraction therefore shifts toward relatively larger dispersion and “softer” polarization components that vary more smoothly with local geometry, reducing the distinctiveness (and thus RF-importance) of S-related descriptors. This interpretation is consistent with comparative analyses of O- and S-acceptor hydrogen bonds, which reported weaker binding and different preferred approach geometries for sulfur relative to oxygen, attributable to differences in the underlying multipolar electrostatics of the acceptor atom.⁷³ By

the same reasoning, P-containing descriptors are less informative than N-containing ones, because nitrogen offers stronger permanent dipoles and hydrogen-bond donation/acceptance than phosphorus.

The specific structural and chemical features endowing CORE-COFs with high urea binding ($\mu_{\text{urea}}^{\text{ex}}$ lower than -55 kJ mol^{-1}) were examined. According to the most important feature O/C, we initially focused on categorizing the COFs with $\text{O/C} = 0$ and $\text{O/C} > 0$. Therefore, first, only CORE-COFs lacking oxygen atoms were considered to highlight which other features are essential for strong binding other than O content (Fig. 4a). Only three CORE-COFs exhibited $\mu_{\text{urea}}^{\text{ex}}$ values below -55 kJ mol^{-1} , with all three having Cl/C ratios higher than 0.05, distinguishing them from the rest of the dataset. This is in good agreement with the mentioned earlier feature importance of Cl/C on determining $\mu_{\text{urea}}^{\text{ex}}$. Next, for the subset of oxygen-containing COFs, we further divided them into two groups based on the presence or absence of nitrogen (Fig. 4b). Interestingly, the presence of N is a criterion for $\mu_{\text{urea}}^{\text{ex}}$ lower than -40 kJ mol^{-1} , i.e. strong urea binding. Accordingly, COF with no N were excluded from further analysis, showing the reason for importance of N-related features.

For COFs containing N and O, we focused on finding the pathway to promising COFs among the most important features mentioned above based on RF-model. First, porosity was plotted *versus* C_{frac} as a second chemical important feature (Fig. 5a). The resulting distribution shows that the COFs exhibit $\mu_{\text{urea}}^{\text{ex}}$ lower than -55 kJ mol^{-1} have $0.510 < \text{porosity} < 0.685$ and $0.360 < C_{\text{frac}} < 0.551$, defining the rectangular box in Fig. 5a. To further refine this subset aiming for exclusion of COFs above the line in rectangular box in Fig. 5a, COFs out of this constraints were excluded and the relationship between the composite feature $C_{\text{frac}} \times \text{porosity}^2$ and C_{frac} were examined which revealed that the additional constraint $C_{\text{frac}} \times \text{porosity}^2 < 0.188$ further discriminated the promising COFs (Fig. 5b). For the COFs meeting this requirement, PLD as the most important structural feature was then plotted against C_{count} , yielding a refined range of $6.00 \text{ \AA} < \text{PLD} < 17.800 \text{ \AA}$ (Fig. 5c). Building on this, plotting $C_{\text{frac}}/\text{PLD}$ against O/C showed that two additional

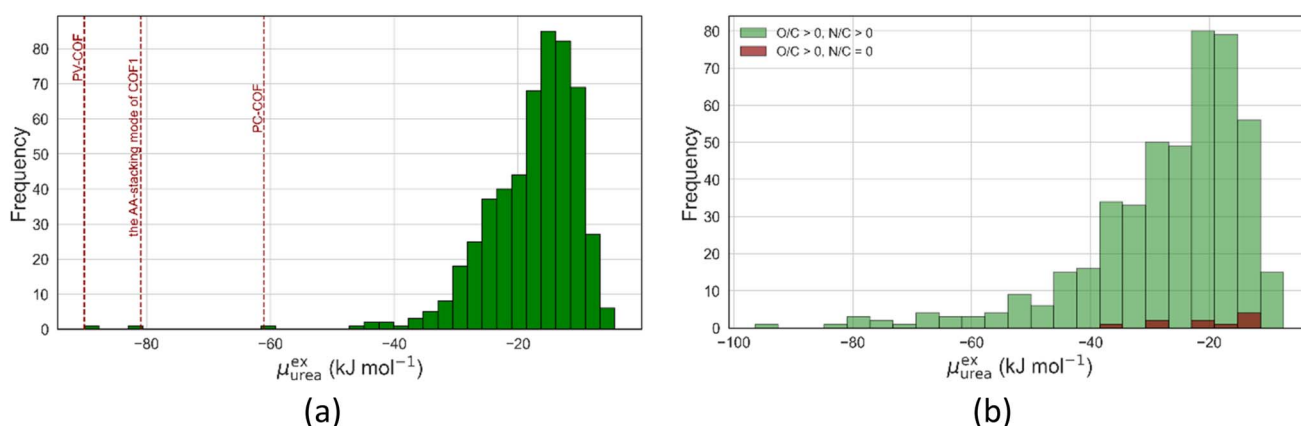


Fig. 4 (a) Histogram distribution of $\mu_{\text{urea}}^{\text{ex}}$ for CORE-COFs without oxygen atoms. Cross markers indicate CORE-COFs with $\mu_{\text{urea}}^{\text{ex}}$ lower than -55 kJ mol^{-1} , (b) histogram distribution of $\mu_{\text{urea}}^{\text{ex}}$ for CORE-COF dataset having $\text{O/C} > 0$, and $\text{N/C} = 0$ (red bars), $\text{O/C} > 0$, and $\text{N/C} > 0$ (green bars).

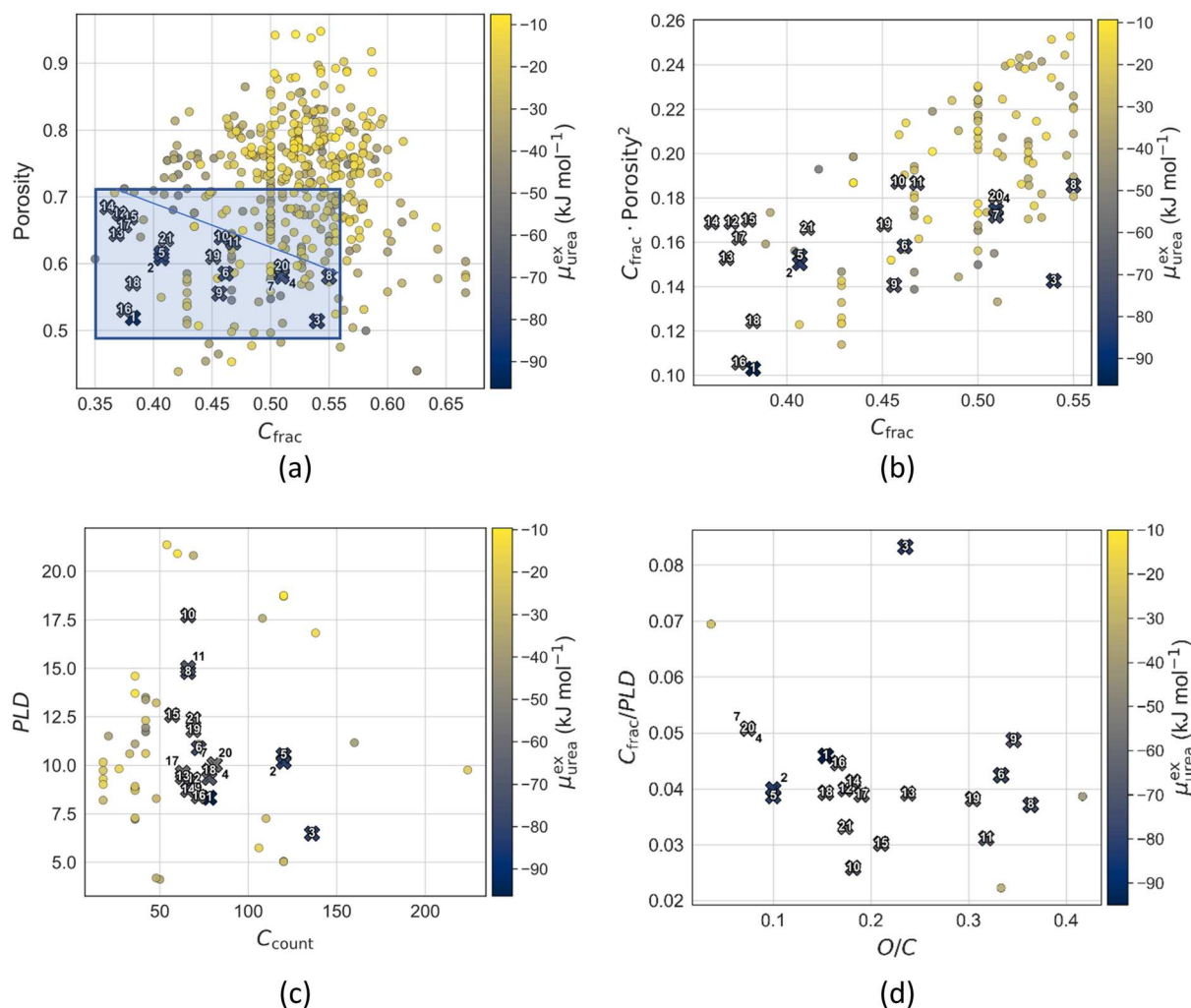


Fig. 5 Structure–property correlations of $\mu_{\text{urea}}^{\text{ex}}$ for CORE-COFs having O and N > 0 under different structural and compositional constraints (a) porosity versus C_{frac} , (b) $C_{\text{frac}} \times \text{porosity}^2$ versus C_{frac} , (c) PLD versus C_{frac} , (d) $C_{\text{count}}/\text{PLD}$ versus O/C. Cross markers indicate COFs with $\mu_{\text{urea}}^{\text{ex}}$ lower than -55 kJ mol^{-1} . The numbers assigned to each CORE-COFs are ID point in Table S2.

filters, $C_{\text{frac}}/\text{PLD} > 0.025$, and $0.050 < \text{O/C} < 0.400$ was necessary to maintain favorable structural characteristics (Fig. 5d). Taken together, these successive filters unraveled a pathway to promising CORE-COFs with enhanced performance so that oxygen and nitrogen containing CORE-COFs satisfying feature constraints of $0.510 < \text{porosity} < 0.685$, $0.360 < C_{\text{frac}} < 0.551$, $C_{\text{frac}} \times \text{porosity}^2 < 0.188$, $6.0 \text{ \AA} < \text{PLD} < 17.80 \text{ \AA}$, $C_{\text{frac}}/\text{PLD} > 0.025$, $0.050 < \text{O/C} < 0.400$ were found to exhibit $\mu_{\text{urea}}^{\text{ex}}$ values lower than -55 kJ mol^{-1} .

The combined interpretation suggests that both framework chemistry (O/C, N/C, and C_{frac} ranges) and geometric constraints (porosity, PLD, and atom-density relationships such as $C_{\text{frac}}/\text{PLD}$) must be balanced to achieve optimal urea adsorption energetics. Oxygen- and nitrogen-containing frameworks provide the necessary polarity and hydrogen-bonding capacity, while the identified structural filters ensure accessible yet confined pore environments that enhance host-guest interactions. Moreover, it was observed that the defining the new combination of structural and chemical features such

as $C_{\text{frac}} \times \text{porosity}^2$, and $C_{\text{frac}}/\text{PLD}$ would be an extra useful features. The list of promising CORE-COFs and their structural features are tabulated in Table S3, with their unit cells structures listed in Table S4.

It should be emphasized that the constraints identified in Fig. 5 are CORE-COF-specific enrichment patterns rather than universal COF design rules. They were derived from the simulated urea binding strengths and descriptor distributions of the CORE-COF dataset and should therefore be used as screening guidelines requiring further validation before being generalized to hypothetical COFs or other COF databases with different linker chemistries, topologies, pore-size distributions, and local functional-group environments.

3.3. Machine learning analysis

3.3.1 Core-COFs. Based on the selected 23 features (shown in Fig. 1), tree-based, SVR and NN models were developed to predict $\mu_{\text{urea}}^{\text{ex}}$ for the urea adsorption on the whole dataset of CORE-COFs. Fig. 6 illustrates the predicted values *versus*

simulated values of $\mu_{\text{urea}}^{\text{ex}}$ for the entire CORE-COFs dataset using six different regression models: Extra Trees (ET), Random Forest (RF), HistGradientBoosting (HG), Gradient Boosting (GB), Support Vector Regression (SVR), and Neural Network (NN).

To compare the different regression models, we employed a nested cross-validation (CV) framework. Model selection was based on the mean RMSE obtained from the outer CV loop, which provides an unbiased estimate of generalization error because the outer test folds are never used during hyper-parameter tuning. This strategy minimizes data leakage and

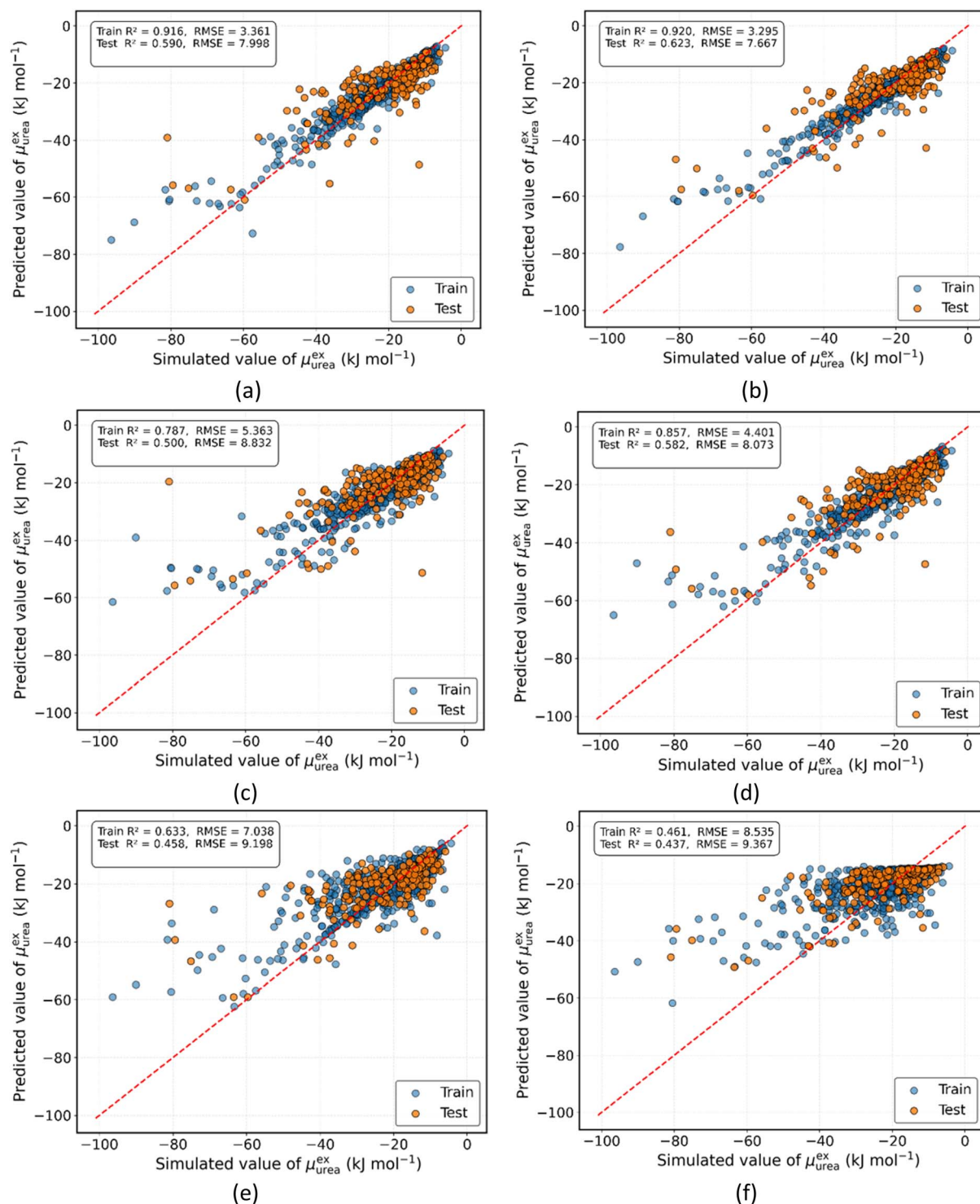


Fig. 6 Predicted versus simulated value of $\mu_{\text{urea}}^{\text{ex}}$ for whole dataset of CORE-COFs and feature set from Fig. 1 from training regression models of (a) Extra Trees, (b) Random Forest, (c) HistGradientBoosting, (d) Gradient Boosting, (e) SVR, (f) Neural Network.

provides a robust estimate of generalization performance, particularly important for relatively small and imbalanced datasets.⁷⁴ After training models and optimizing hyperparameters, we evaluated it on the full training set and an independent hold-out test set. In Fig. 6, we report R^2 and RMSE for both the training and hold-out test sets to illustrate generalization behavior. The obtained optimised hyperparameters, and nested CV mean RMSE of the outer loop for each model are reported in Table S5.

Across all models, a strong linear correlation between predicted and simulated $\mu_{\text{urea}}^{\text{ex}}$ with moderate values is observed, indicating that the selected feature set captures the dominant structure–property relationships governing urea adsorption in COFs. However, deviations from the parity line (red dashed line) are more pronounced at the extreme negative $\mu_{\text{urea}}^{\text{ex}}$ values, which correspond to sparsely populated regions of the dataset, containing the most promising COF structures. One reason for this behavior would be consistent with the imbalanced distribution of the target variable, where the majority of samples cluster around moderate $\mu_{\text{urea}}^{\text{ex}}$ values (histogram distribution of $\mu_{\text{urea}}^{\text{ex}}$ across a dataset of COF structures shown in Fig. S4). Consequently, all models tend to exhibit higher uncertainty and occasional bias when predicting rare, strongly adsorbing COFs.

Another common observation across all regression models is the systematic overprediction of the most strongly negative $\mu_{\text{urea}}^{\text{ex}}$ values. This trend is not model-specific, but instead reflects the combined effects of target imbalance, sparse sampling of the extreme negative regime. Because most COF structures in the dataset exhibit moderate $\mu_{\text{urea}}^{\text{ex}}$ values (less negative values), while the strongly negative region is represented by only a few and structurally heterogeneous samples, models trained with squared-error-based objectives tend to predict values closer to the average response of similar structures. Consequently, predictions in the negative tail are shifted toward less negative values, producing a consistent overprediction of adsorption strength rather than underprediction.

For the ensemble tree-based methods ET (Fig. 6a) and RF (Fig. 6b), the models achieved the lowest mean outer-loop RMSE ($\approx 9.0 \text{ kJ mol}^{-1}$), indicating the best overall performance under nested CV. Consistent with this, they also showed the best predictive capability, with $R_{\text{train}}^2 \approx 0.9$ and $R_{\text{test}}^2 \approx 0.6$. The observed train–test gap is therefore more reasonably attributed to descriptor limitations and target imbalance, particularly the sparsely sampled, strongly negative $\mu_{\text{urea}}^{\text{ex}}$ regime rather than data leakage, because hyperparameter optimization and model evaluation are strictly separated within the nested CV framework.

The boosting-based models HGB (Fig. 6c) and GB (Fig. 6d) showed slightly lower performance than the tree-ensemble bagging methods (ET and RF), with a mean outer-loop RMSE of $\approx 9.4 \text{ kJ mol}^{-1}$ and moderate predictive capability ($R_{\text{train}}^2 \approx 0.8$, $R_{\text{test}}^2 \approx 0.5$). This reduced performance suggests that, for the present dataset, boosting provided less robust generalization than bagged tree ensembles. A consistent trend was discussed by Dietterich⁷⁵ in his comparative study of decision-tree ensembles, where he emphasized that boosting-style methods can become disproportionately influenced by a small subset of

difficult or atypical samples, because they iteratively focus learning on points that are repeatedly mispredicted, whereas bagging stabilizes predictions through variance reduction by averaging over multiple resampled training sets. In our case, the sparsely populated, strongly negative $\mu_{\text{urea}}^{\text{ex}}$ region represents a hard-to-fit portion of the response space, so the variance-reduction effect of ET and RF is expected to yield more stable performance than boosting under such data irregularities.

The SVR model (Fig. 6e) revealed the mean outer-loop RMSE of $\approx 9.43 \text{ kJ mol}^{-1}$ and relatively limited predictive accuracy ($R_{\text{train}}^2 \approx 0.63$, $R_{\text{test}}^2 \approx 0.45$). The parity plot also exhibits the largest dispersion around the 1 : 1 line and a more pronounced underprediction of strongly negative $\mu_{\text{urea}}^{\text{ex}}$ values. This behavior is consistent with the sensitivity of kernel-based models to a non-uniform target distribution, where model fitting is dominated by the densely sampled moderate-response region, making it difficult to achieve comparable accuracy in sparsely sampled extreme regimes.⁷⁶

The trained NN model with early stopping (Fig. 6f) showed the weakest overall performance, yielding the highest mean outer-loop RMSE ($\approx 10.3 \text{ kJ mol}^{-1}$) and the lowest predictive accuracy on both the training and test sets ($R_{\text{train}}^2 \approx 0.46$, $R_{\text{test}}^2 \approx 0.43$). This would be attributed to the fact that the NN is not only less data-efficient for tabular descriptor sets but also more sensitive to uneven coverage of the response space; with few samples in the strongly negative $\mu_{\text{urea}}^{\text{ex}}$ regime compared with tree-based ensemble and boosting-based methods.⁷⁷

It is worth noting that imbalance-aware oversampling methods such as SMOTER/SMOEN were not adopted in this work because they generate synthetic feature–target pairs through interpolation and noise injection in descriptor space, without enforcing physical or chemical constraints. As also noted by Al-Jubouri *et al.*,⁷⁸ SMOEN produces non-unique resampled distributions across runs and can lead models to learn imprecise dependencies between molecular descriptors and target properties, particularly in chemically heterogeneous datasets. Such artificial samples may not correspond to physically realizable structures, which is problematic for nanoporous materials where adsorption properties are tightly coupled to framework geometry and chemistry.

We also examined more conservative options of target-binning stratification and loss reweighting to emphasize the extreme negative $\mu_{\text{urea}}^{\text{ex}}$ regime but these did not improve the outer-loop (nested CV) RMSE and, in some cases, increased variability across folds. This indicates that imbalance alone is not the primary limitation; rather, the key challenge is the sparse and insufficient sampling of the strongly negative $\mu_{\text{urea}}^{\text{ex}}$, where the model has limited empirical support to learn stable patterns. Recent work on imbalanced regression supports this interpretation, showing that degraded tail performance is often driven by data insufficiency (few-shot target regions) and that reweighting or distribution correction strategies may provide limited benefit or even increase variance when rare regions contain very few samples.⁷⁹ Accordingly, we reported nested-CV results using only physically grounded data and explicitly discussed the increased uncertainty in the

sparsely sampled extreme negative $\mu_{\text{urea}}^{\text{ex}}$ regime as a data-coverage limitation rather than an algorithmic artifact.

Moreover, it should be clarified that although the learned feature constraints successfully enrich for strongly binding COFs, the most negative $\mu_{\text{urea}}^{\text{ex}}$ region remains few-shot and intrinsically heterogeneous, providing limited statistical support for learning a stable value-prediction mapping in that regime. Moreover, even within the constrained subset of promising adsorbents, $\mu_{\text{urea}}^{\text{ex}}$ can vary substantially due to finer interaction motifs such as the local functional-group microenvironment, charge distribution, and adsorption-site geometry that are not explicitly encoded by our 23 largely global descriptors. In this sense, identifying constraints is a different and easier task than predicting precise values of $\mu_{\text{urea}}^{\text{ex}}$ so that the descriptors can be sufficient to separate promising structures from the bulk population, yet insufficient to resolve the remaining within-class variability of the extremes. Therefore, the constraints should be viewed as necessary but not sufficient conditions for accurate point predictions in the strongly negative tail, and improved performance in this regime will likely require additional training data and/or more interaction-sensitive, local descriptors.

Regarding the prediction of urea/water selectivity ($\mu_{\text{urea}}^{\text{ex}} - \mu_{\text{water}}^{\text{ex}}$) of CORE-COFs, we adopted RF model as the primary model for downstream prediction because it provided the most reliable generalization for $\mu_{\text{urea}}^{\text{ex}}$ in the nested-CV benchmarking, using the same 23-descriptor feature set. However, $\mu_{\text{urea}}^{\text{ex}} - \mu_{\text{water}}^{\text{ex}}$ is inherently a more complex target than $\mu_{\text{urea}}^{\text{ex}}$ alone, as it reflects the relative affinity of the framework for two competing species and therefore depends on additional, potentially distinct, structure-solvent interactions captured in $\mu_{\text{urea}}^{\text{ex}} - \mu_{\text{water}}^{\text{ex}}$ as well as the correlation between the two terms. Accordingly, we retained the RF model family and retrained it using the urea/water selectivity target while keeping the same descriptors and leakage-free nested CV protocol, ensuring a consistent model complexity and evaluation methodology across both prediction tasks. We emphasize that the $\mu_{\text{urea}}^{\text{ex}}$ RF model is used only to motivate the choice of algorithm; it is not transferred directly, and all selectivity results are obtained from RF models trained specifically on the selectivity target.

As observed in Fig. 7, the model shows lower predictability for $\mu_{\text{urea}}^{\text{ex}} - \mu_{\text{water}}^{\text{ex}}$ than for $\mu_{\text{urea}}^{\text{ex}}$. While the training set exhibits strong agreement ($R_{\text{train}}^2 = 0.91$, $\text{RMSE} = 1.79 \text{ kJ mol}^{-1}$), the test performance drops markedly ($R_{\text{test}}^2 = 0.45$, $\text{RMSE} = 4.12 \text{ kJ mol}^{-1}$) indicating weaker generalization for selectivity. This reduction is expected because selectivity is defined as a difference between two correlated excess chemical potentials, so prediction errors from each component can accumulate and amplify the noise in the final target. Moreover, as mentioned earlier, a richer descriptor set is likely required to more completely capture the urea-framework interactions underlying accurate prediction of $\mu_{\text{urea}}^{\text{ex}}$. This need becomes even more essential for $\mu_{\text{urea}}^{\text{ex}} - \mu_{\text{water}}^{\text{ex}}$ because the model must represent not only urea-framework interactions but also water-framework interactions and, critically, the subtle structural features that determine their relative balance.

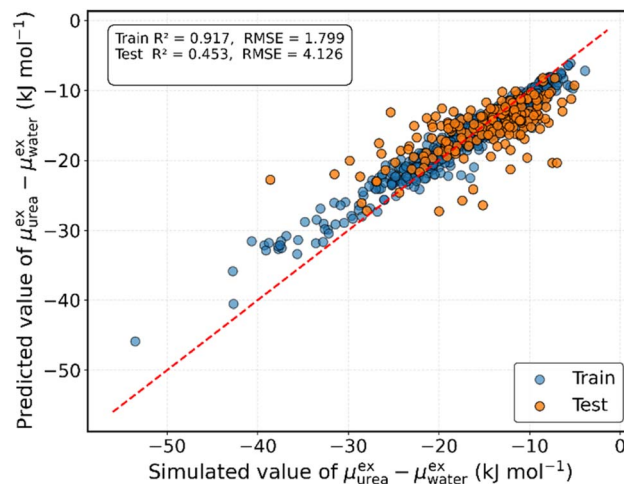


Fig. 7 Predicted versus simulated value of $\mu_{\text{urea}}^{\text{ex}} - \mu_{\text{water}}^{\text{ex}}$ for whole dataset of CORE-COFs and feature set from Fig. 1 from training regression models of RF.

3.3.2 Berkeley h-COFs. Because the RF model trained on the full CORE-COFs dataset showed the best performance and generalization for urea binding strength ($R_{\text{train}}^2 = 0.92$ and $R_{\text{test}}^2 = 0.62$), we selected it for rapid, high-throughput screening of the complete Berkeley-COFs dataset. To assess whether the Berkeley structures occupy a similar descriptor space to the training data, we applied UMAP dimensionality reduction⁸⁰ to project the high-dimensional set of structural and chemical features into two dimensions. As shown in Fig. 8a, CORE-COFs (blue) populate only a limited subset of the broader Berkeley-COFs feature space (orange), indicating that screening the Berkeley library inevitably involves distribution shift (*i.e.*, many Berkeley structures lie outside the densest region spanned by the training data). This observation is important because tree-based regressors such as RF generally interpolate well within the training manifold but can become less reliable when asked to extrapolate into sparsely covered regions. The RF-predicted $\mu_{\text{urea}}^{\text{ex}}$ values for Berkeley-COFs are visualized as a color map over the UMAP projection in Fig. 8b, where darker colors correspond to stronger predicted urea binding. The map reveals that low $\mu_{\text{urea}}^{\text{ex}}$ candidates are not uniformly distributed across the Berkeley chemical space; instead, they appear in localized regions, suggesting that the model is identifying specific combinations of descriptors (*e.g.*, pore environment, polarity, and functional group chemistry) associated with stronger urea-framework interactions.

From this screening, we selected the top ten Berkeley candidates with the lowest predicted $\mu_{\text{urea}}^{\text{ex}}$ (Table 1) and validated them using Widom insertion simulations. Importantly, no Berkeley COFs were used during training, so this validation represents a genuinely external test of the ML screening workflow. While the RF model predicts a narrow band of very favorable values (≈ -46 to -48 kJ mol^{-1}) for the top-ranked candidates, the Widom simulations show that the simulated $\mu_{\text{urea}}^{\text{ex}}$ values are systematically less negative (≈ -16 to -31 kJ mol^{-1}).

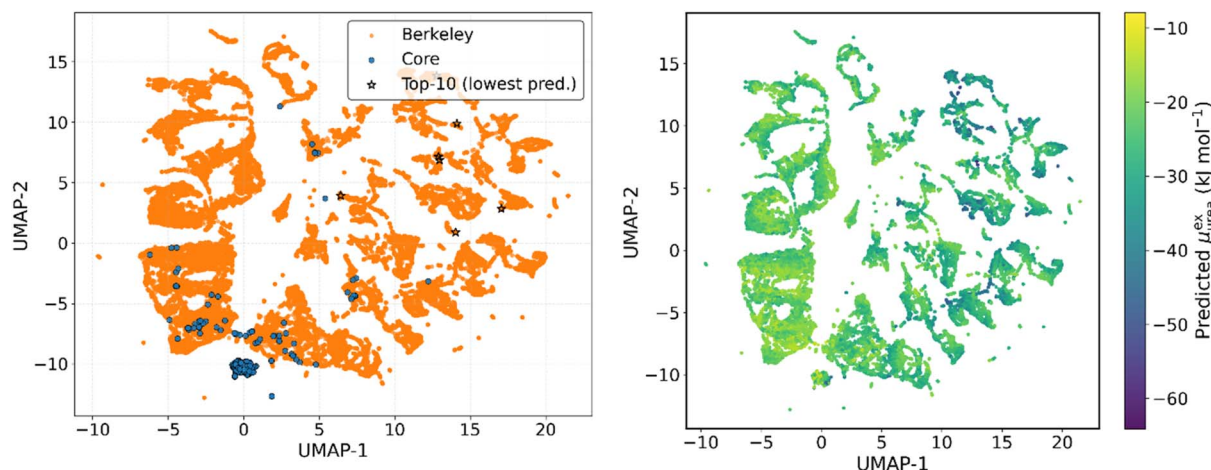


Fig. 8 (a) UMAP projection of the feature space for Berkeley-COFs (orange) and core COFs (blue) with known $\mu_{\text{urea}}^{\text{ex}}$ values. This visualization emphasises the distribution of core COFs within the broader Berkeley dataset, illustrating the regions of chemical space where predictive coverage is strongest, (b) UMAP projection of the Berkeley-COFs dataset coloured by the ML-predicted $\mu_{\text{urea}}^{\text{ex}}$. Lower values (shown in dark blue/purple) correspond to COFs predicted to have stronger urea adsorption. The plot visually reveals regions of the chemical space with higher-performing candidates, assisting in identifying promising materials for further screening.

To assess whether this overestimation was related to extrapolation beyond the CORE-COF training domain, an applicability-domain analysis was performed. Table 1 reports the RF-predicted and Widom-simulated $\mu_{\text{urea}}^{\text{ex}}$ values together with the Euclidean distance from each selected Berkeley-COF to its nearest CORE-COF in the standardized selected-descriptor space used for RF prediction. The 95th and 99th percentile thresholds of leave-one-out nearest-neighbor distances within the CORE-COF dataset were 2.9593 and 6.0923, respectively. All selected Berkeley-COFs were outside the 95% CORE-COF applicability-domain threshold, while 9 out of 10 remained within the broader 99% threshold. These results indicate that the selected Berkeley candidates lie near the edge of, or outside, the main CORE-COF training domain, making their RF predictions partly extrapolative. Therefore, the Berkeley-COF screening should be interpreted primarily as a qualitative prioritization and chemical-pattern discovery step rather than as a quantitatively accurate ranking of Berkeley-COF adsorption performance.

As discussed above, the RF model showed different systematic biases for the CORE-COF test set and the externally screened Berkeley-COFs. In the CORE-COF test set, the strongest binders were generally predicted to have less negative $\mu_{\text{urea}}^{\text{ex}}$ values than those obtained from simulation, indicating underprediction of binding strength and reflecting the tendency of RF regression to smooth extreme values toward the mean in an in-domain setting. In contrast, the Berkeley-COFs were generally assigned more negative $\mu_{\text{urea}}^{\text{ex}}$ values than were later confirmed by simulation, corresponding to an over-prediction of binding strength. This apparent reversal likely arises from the combined effects of distribution shift and selection bias: the Berkeley-COFs occupy a broader and partially out-of-distribution chemical space, and selecting only the most favorable ML-ranked candidates enriches the screened subset in false positives with artificially favorable predictions.

Furthermore, although the Berkeley top hits were predicted to lie near -48 kJ mol^{-1} , their simulated values were mainly between -16 and -31 kJ mol^{-1} , indicating that they do not belong to the same true extreme negative regime as the strongest CORE-COFs. Therefore, the bias observed for the extreme negative region of the CORE-COF test set may not be directly transferable to the Berkeley set. Moreover, the Widom validation of Berkeley-COFs was limited to the top ten candidates identified by the RF screening model and should therefore be interpreted as a proof-of-concept candidate-prioritization test rather than a comprehensive external benchmark of model accuracy across the full Berkeley-COF feature space. Accordingly, a larger subset of Berkeley-COFs including medium-ranked, randomly selected, and uncertainty-selected Berkeley-COFs would need to be validated by simulation to obtain a more reliable assessment of the actual screening performance of the RF model, although this is beyond the scope of the present work.

Despite this bias in absolute values, the screened hits in Table 1 showed that a clear chemical pattern: three top candidates with simulated values for $\mu_{\text{urea}}^{\text{ex}}$ around -30 kJ mol^{-1} are built from linkers containing -NH and -CO functionalities (amide-type chemistry) and are frequently associated with *dia* or *hcb* topologies. These motifs are physically meaningful for urea adsorption because polar functional groups can promote stronger local interactions (*e.g.*, hydrogen bonding and dipole-dipole contacts), while the network topology influences the distribution and accessibility of adsorption sites.

For dialysate-regeneration applications, adsorption performance must be considered together with synthesis feasibility, aqueous/hydrolytic stability, regeneration ability, and biocompatibility. This is particularly important for the Berkeley-COFs, because they are hypothetical structures. Nevertheless, their synthesizability and stability were considered at a theoretical level in the original Berkeley-COF database construction.

Table 1 A list of promising Berkeley-COFs with highest urea bindings strength, together with nearest-CORE applicability-domain distances^a

Berkeley-COF name	linkerA	linkerB	net	Bond type	Chemical formula	ML-predicted $\mu_{\text{urea}}^{\text{ex}}$ (KJ mol ⁻¹)	Simulated $\mu_{\text{urea}}^{\text{ex}}$ (KJ mol ⁻¹)	Distance to nearest CORE-COF	d/d_{95}
linker110_CH2_linker70_NH_dia_relaxed	linker110_CH2	linker75_NH	dia	Amine	C ₂₂₄ H ₂₈₈ N ₆₄ O ₃₂	-48.43	-23.83	4.715	1.593
linker106_NH_linker80_CO_unc_relaxed	linker106_NH	linker80_CO	dia	Amide	C ₂₅₆ H ₂₇₂ N ₆₄ O ₄₈	-48.37	-29.28	5.676	1.918
linker110_NH_linker11_CO_dia_relaxed	linker110_NH	linker11_CO	unc	Amide	C ₂₀₈ H ₂₅₆ N ₆₄ O ₃₂	-48.35	-30.06	6.543	2.211
linker91_CH2_linker29_NH_hcb_relaxed	linker91_CH2	linker29_NH	dia	Amine	C ₁₄₄ H ₁₂₆ N ₇₂ O ₃₆	-48.33	-24.54	5.548	1.874
linker91_NH_linker29_CH2_hcb_relaxed	linker91_NH	linker29_CH2	hcb	Amine	C ₁₄₄ H ₁₂₆ N ₇₂ O ₃₆	-48.33	-24.85	5.562	1.879
linker91_NH_linker18_CH2_hcb_relaxed	linker91_NH	linker18_CH2	hcb	Amine	C ₁₄₄ H ₁₂₆ N ₇₂ O ₃₆	-48.23	-22.73	5.602	1.893
linker110_NH_linker73_CO_dia_relaxed	linker110_NH	linker73_CO	hcb	Amide	C ₂₂₄ H ₂₅₆ N ₆₄ O ₃₂	-47.96	-30.73	5.909	1.996
linker110_CH2_linker67_NH_dia_relaxed	linker110_CH2	linker67_NH	dia	Amine	C ₁₉₂ H ₂₂₄ N ₆₄ O ₃₂	-47.47	-25.41	4.299	1.452
linker101_NH_linker70_CH2_kgm_relaxed	linker101_NH	linker70_CH2	kgm	Amine	C ₂₅₂ H ₂₈₈ N ₆₀ O ₃₆	-47.36	-22.07	5.225	1.765
linker110_CH2_linker75_NH_dia_relaxed	linker110_CH2	linker75_NH	dia	Amine	C ₂₂₄ H ₂₈₈ N ₆₄ O ₃₂	-46.43	-16.48	4.206	1.421

^a d is the Euclidean distance from each selected Berkeley-COF to its nearest CORE-COF in the standardized selected-descriptor space. d_{95} is the 95th percentile of leave-one-out nearest-neighbor distances within the CORE-COF dataset, equal to 2.9593. Values of d/d_{95} greater than 1 indicate that the selected Berkeley candidate lies outside the main 95% CORE-COF applicability domain. The 99th percentile threshold was 6.0923; therefore, 9 out of 10 selected Berkeley-COFs were within the broader 99% threshold, while one candidate was outside even this broader domain.

Specifically, the structures were generated from rigid organic linkers using synthetically motivated COF-forming chemistries, including C-C, imine, amine, amide, and mixed linkages. In addition, chemically unreasonable, unstable, poorly assembled, or non-converged structures were removed during database cleaning, supporting the structural plausibility of the final database. From a stability perspective, C-C-linked frameworks are expected to be the most hydrolytically robust, whereas imine-linked COFs are more synthetically accessible but may be less suitable for long-term aqueous exposure due to the reversibility of the C=N bond. Amine- and amide-linked COFs are particularly relevant because they can combine improved chemical stability with polar functional groups. From a biocompatibility perspective, the Berkeley-COFs are also favorable in that they are fully organic and do not contain metal nodes, reducing concerns related to toxic metal leaching.²⁶ However, theoretical synthesizability and structural stability do not replace experimental validation. Therefore, the Berkeley candidates identified here should be regarded as promising computational leads rather than immediately deployable wearable-artificial-kidney sorbents. Future work should experimentally assess their synthesis feasibility, stability in dialysate-like aqueous media, adsorption/regeneration cycling, leachable degradation products, and biocompatibility, including cytotoxicity and hemocompatibility.

In addition to the studied descriptors for CORE-COFs, the type and number of chemical bonds were also calculated. Their Pearson correlation coefficients with $\mu_{\text{urea}}^{\text{ex}}$ were on the order of 0.01, indicating a negligible relationship (Fig. 9). Therefore, incorporating these descriptors into the feature set would not contribute to improving the performance of the ML models. Improving predictive performance specifically for the rare high-performing materials will therefore likely require alternative featurization that better captures the local adsorption environment and interaction motifs responsible for strong urea binding. We assume that understanding and quantifying functional groups as features could potentially yield a stronger correlation with the target compared to simple bond counts. However, the identification of functional groups in CORE-COFs is not included in the current dataset and cannot be readily computed for such a large number of frameworks. Moreover, the compromising linkers are not separately provided for CORE-COFs dataset to expand the other types of featurization with more accuracy. Future work incorporating automated cheminformatics tools or curated datasets that explicitly encode functional group information may enable more accurate feature engineering and enhance the predictive performance of ML models, particularly given that $\mu_{\text{urea}}^{\text{ex}}$ reflects host-guest interaction strength and adsorption thermodynamics.

Future studies may further improve the quantitative accuracy of the screening workflow by incorporating DFT-validated force-field refinement, and ML-assisted parameterization of COF-urea interaction parameters. Direct use of machine-learning interatomic potentials would also be an attractive long-term direction; however, such potentials are not presently available as a standard input-level capability in the RASPA workflow used here and would require additional

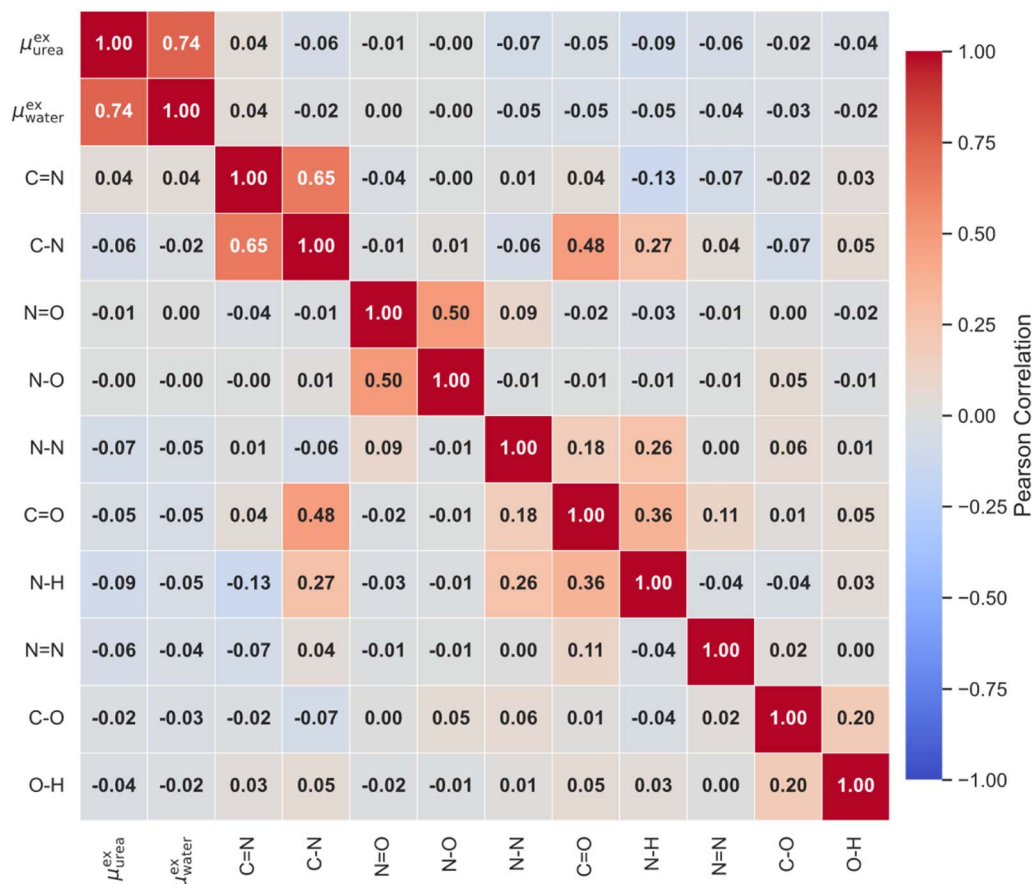


Fig. 9 Pearson correlation matrix of bonds of CORE-COFs and $\mu_{\text{urea}}^{\text{ex}}$.

implementation or coupling to simulation engines that support ML potentials.

Moreover, screening of Berkeley-COFs can be integrated with active learning for incorporating additional training points sampled from the Berkeley feature space.⁸¹

4. Conclusions

We performed a screening of real and hypothetical COFs using a combined Molecular-Machine Learning approach for their application in urea removal from water during dialysate regeneration in a wearable hemodialysis process. The target properties considered were urea/COF binding strength and urea/water ideal selectivity of the COF structures at infinite dilution, with the first property chosen as primary target, as selectivity was shown to be correlated to it. Urea binding strength at infinite dilution simulated *via* Widom insertions on 993 COFs from the CORE-COF dataset was computed, allowing to score the best performing COFs. However, the specific aim of this paper was to use the simulated CORE-COF data to train different ML models and select the one showing the best predictive performance and use it to screen in a computationally efficient way of the Berkeley dataset, containing 69 840 hypothetical COF structures. The analysis of feature importance of the CORE-COF dataset *via* the Random Forest model revealed

that framework composition, particularly oxygen-to-carbon ratio and carbon fraction together with pore limiting diameter are the dominant descriptors governing urea adsorption energetics and allowed to reduce the number of features from the initial 50 to 23. Optimal feature ranges for the structural and chemical features were identified to achieve the best COF performance.

Six ML models including Random Forest, Gradient Boosting, Extra Trees, HistGradientBoosting, Decision Trees, Sport Vector regressor and Neural Networks were then trained on the CORE-COFs dataset. All models predicted well the region of moderate $\mu_{\text{urea}}^{\text{ex}}$ values, confirming that the chosen descriptors capture key structure–property trends in COFs; however, errors increase for the sparsely sampled, highly negative $\mu_{\text{urea}}^{\text{ex}}$ regime, indicating the need for more data in this range and/or more interaction-sensitive features. The RF model trained on CORE-COFs provided the best balance of fit and generalization ($R_{\text{train}}^2 = 0.92$, $R_{\text{test}}^2 = 0.62$) and was therefore used for rapid screening of the Berkeley-COFs dataset. UMAP analysis showed limited feature-space overlap between CORE-COFs and Berkeley-COFs, indicating a dataset shift that leads to systematic over-estimation of $\mu_{\text{urea}}^{\text{ex}}$ compared with Widom simulations. Despite biases in absolute values, the top-ranked candidates exhibit clear chemical trends: they are predominantly built from linkers containing –NH and/or –CO functionalities and often

adopt dia or hcb topologies. These features are consistent with effective urea adsorption, as polar groups enhance local interactions while network topology governs site accessibility. Given the limitations in explicitly defining functional groups and linker compositions in CORE-COFs, further refinement of featurization is still required to achieve improved predictive accuracy for $\mu_{\text{urea}}^{\text{ex}}$, particularly in the strongly negative region that reflects stronger host-guest interactions.

These findings establish design rules for tuning COF chemistry and geometry to enable efficient urea separation, guiding the discovery of optimal materials for urea removal. Moreover, the combined simulation-ML strategy offers a scalable workflow to accelerate COF discovery and guide sorbent-based dialysate regeneration for wearable artificial kidneys.

Author contributions

M. Z. D.: conceptualization, methodology, writing – review and editing. T. F.: methodology. S. D.: supervision, review & editing. S. T.: supervision, review & editing. M. G. D. A.: supervision, writing – review and editing, funding acquisition.

Conflicts of interest

The authors declare no competing interests.

Data availability

The data and code supporting the findings of this study are available in Zenodo at <https://doi.org/10.5281/zenodo.21827298>. The Zenodo record corresponds to the release MaryamZarghami/cof-urea-screening-ml: v1.0.0 – Accepted manuscript version and contains the processed datasets, and machine-learning scripts used in this study. The file Dataset-CORE_COFs contains the processed CORE-COF descriptor matrix together with the corresponding Widom-derived ($\mu_{\text{urea}}^{\text{ex}}$) simulation targets used for the molecular-simulation and machine-learning analyses. The Berkeley-COF dataset contains descriptors calculated and aligned with the final CORE-COF feature set to enable reproduction of the Berkeley-COF screening workflow. The folder ML-Models-supplementary-Digital-discovery contains the scripts and notebooks used for descriptor preprocessing, train/test splitting, hyperparameter optimization, model training, model evaluation, feature-importance analysis, and Berkeley-COF prediction. The code and data are also openly available in the GitHub repository MaryamZarghami/cof-urea-screening-ml under the tagged release v1.0.0 – Accepted manuscript version (<https://github.com/MaryamZarghami/cof-urea-screening-ml>).

Supplementary information (SI): the file Supplementary_Information contains the supplementary information for the manuscript. The file Pearson_correlation_matrix_Fig._S1.csv contains the Pearson correlation matrix used to generate Fig. S1. The file selected_core_cofs_mu_ex_convergence_plots.pdf contains the convergence analysis for the representative CORE-COF Widom insertion simulations. See DOI: <https://doi.org/10.1039/d6dd00159a>.

Acknowledgements

This work was supported by Kidney Research UK through the program Kidney Research UK – Stonegate Research Project Grant SG_KS_RP_012_20221129.

References

- 1 M. Bossola, I. Mariani, C. P. Piccinni, C. Spoliti and E. Di Stasio, Symptom Burden in Patients on Maintenance Hemodialysis: Prevalence and Severity 17 Years Apart, *J. Clin. Med.*, 2024, **13**(18), 5529, DOI: [10.3390/jcm13185529](https://doi.org/10.3390/jcm13185529).
- 2 I. Ledebø and P. J. Blankestijn, Haemodiafiltration—optimal efficiency and safety, *NDT Plus*, 2010, **3**(1), 8–16, DOI: [10.1093/ndtplus/sfp149](https://doi.org/10.1093/ndtplus/sfp149).
- 3 M. Iftikhar, *et al.*, Future research perspectives in hemodialysis membrane technology, *J. Ind. Eng. Chem.*, 2024, **138**, 72–103, DOI: [10.1016/j.jiec.2024.04.024](https://doi.org/10.1016/j.jiec.2024.04.024).
- 4 J. W. M. Agar, Reusing and recycling dialysis reverse osmosis system reject water, *Kidney Int.*, 2015, **88**(4), 653–657, DOI: [10.1038/ki.2015.213](https://doi.org/10.1038/ki.2015.213).
- 5 A wearable artificial kidney for patients with end-stage renal disease - PMC, accessed, Jul. 29, 2024, Online, available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4936831/>.
- 6 T. Niwa, Recent progress in the analysis of uremic toxins by mass spectrometry, *J. Chromatogr. B*, 2009, **877**(25), 2600–2606, DOI: [10.1016/j.jchromb.2008.11.032](https://doi.org/10.1016/j.jchromb.2008.11.032).
- 7 M. Z. Dehaghani, T. Fabiani and M. G. D. Angelis, Chemical engineering contribution to hemodialysis innovation: achieving the wearable artificial kidneys with nanomaterial-based dialysate regeneration, *Phys. Sci. Rev.*, 2025, **10**(3), 279–299, DOI: [10.1515/psr-2024-0055](https://doi.org/10.1515/psr-2024-0055).
- 8 T. Fabiani, M. Zarghamidehaghani, C. Boi, S. Dimartino, S. Kentish and M. G. De Angelis, Sorbent-based dialysate regeneration for the wearable artificial kidney: Advancing material innovation *via* experimental and computational studies, *Sep. Purif. Technol.*, 2025, **360**, 130776, DOI: [10.1016/j.seppur.2024.130776](https://doi.org/10.1016/j.seppur.2024.130776).
- 9 V. Wernert, O. Schäfer, H. Ghobarkar and R. Denoyel, Adsorption properties of zeolites for artificial kidney applications, *Microporous Mesoporous Mater.*, 2005, **83**(1), 101–113, DOI: [10.1016/j.micromeso.2005.03.018](https://doi.org/10.1016/j.micromeso.2005.03.018).
- 10 M. De Pascale, M. G. De Angelis and C. Boi, Mixed Matrix Membranes Adsorbers (MMMAs) for the Removal of Uremic Toxins from Dialysate, *Membranes*, 2022, **12**(2), 203, DOI: [10.3390/membranes12020203](https://doi.org/10.3390/membranes12020203).
- 11 F. Meng, *et al.*, MXene Sorbents for Removal of Urea from Dialysate: A Step toward the Wearable Artificial Kidney, *ACS Nano*, 2018, **12**(10), 10518–10528, DOI: [10.1021/acsnano.8b06494](https://doi.org/10.1021/acsnano.8b06494).
- 12 Y. Zhu, *et al.*, Structure engineering of covalent organic frameworks as metal-free sorbents for urea recovery, *J. Mater. Chem. A*, 2025, **13**(39), 33775–33783, DOI: [10.1039/D5TA04077A](https://doi.org/10.1039/D5TA04077A).
- 13 C.-X. Yang, C. Liu, Y.-M. Cao and X.-P. Yan, Metal-organic framework MIL-100(Fe) for artificial kidney application,

- RSC Adv., 2014, 4(77), 40824–40827, DOI: [10.1039/C4RA05111D](https://doi.org/10.1039/C4RA05111D).
- 14 M. S. Lohse and T. Bein, Covalent Organic Frameworks: Structures, Synthesis, and Applications, *Adv. Funct. Mater.*, 2018, 28(33), 1705553, DOI: [10.1002/adfm.201705553](https://doi.org/10.1002/adfm.201705553).
- 15 Z. Wang, S. Zhang, Y. Chen, Z. Zhang and S. Ma, Covalent organic frameworks for separation applications, *Chem. Soc. Rev.*, 2020, 49(3), 708–735, DOI: [10.1039/C9CS00827F](https://doi.org/10.1039/C9CS00827F).
- 16 P. Mohan, B. Sasikumar, S. A. G. Krishnan and G. Arthanareeswaran, Covalent-organic porous framework (COF) integrated hybrid membranes for energy and environmental applications: Current and future perspectives, *J. Taiwan Inst. Chem. Eng.*, 2025, 166, 105067, DOI: [10.1016/j.jtice.2023.105067](https://doi.org/10.1016/j.jtice.2023.105067).
- 17 A. P. Côté, A. I. Benin, N. W. Ockwig, M. O’Keeffe, A. J. Matzger and O. M. Yaghi, Porous, Crystalline, Covalent Organic Frameworks, *Science*, 2005, 310(5751), 1166–1170, DOI: [10.1126/science.1120411](https://doi.org/10.1126/science.1120411).
- 18 S.-Y. Ding and W. Wang, Covalent organic frameworks (COFs): from design to applications, *Chem. Soc. Rev.*, 2012, 42(2), 548–568, DOI: [10.1039/C2CS35072F](https://doi.org/10.1039/C2CS35072F).
- 19 Covalent Organic Frameworks (COFs): A Necessary Choice For Drug Delivery - Guo - 2022 - ChemistrySelect - Wiley Online Library, accessed, Jan. 13, 2026, Online, available, https://chemistry-europe.onlinelibrary.wiley.com/doi/full/10.1002/slct.202202538?casa_token=Tj44BIB2YNEAAAAA%3ABj6Enz3NsrFZhKTS_vayPxXHMWg2Yk-j5GTGj6D3ajZPqt8hjRPIsLzVCh8nE8eqXlOnJQ1UHH_ofTh.
- 20 Bansidhar, *et al.*, Recent advances in covalent organic frameworks based physico-, chemico- and bio-sensing applications, *J. Macromol. Sci., Part A: Pure Appl. Chem.*, 2025, 62(1), 2–15, DOI: [10.1080/10601325.2024.2437645](https://doi.org/10.1080/10601325.2024.2437645).
- 21 J. Gan, *et al.*, Covalent organic frameworks as emerging host platforms for enzyme immobilization and robust biocatalysis – A review, *Int. J. Biol. Macromol.*, 2021, 167, 502–515, DOI: [10.1016/j.ijbiomac.2020.12.002](https://doi.org/10.1016/j.ijbiomac.2020.12.002).
- 22 Y. Shi, J. Yang, F. Gao and Q. Zhang, Covalent Organic Frameworks: Recent Progress in Biomedical Applications, *ACS Nano*, 2023, 17(3), 1879–1905, DOI: [10.1021/acsnano.2c11346](https://doi.org/10.1021/acsnano.2c11346).
- 23 Z. Xia, Y. Zhao and S. B. Darling, Covalent Organic Frameworks for Water Treatment, *Adv. Mater. Interfaces*, 2021, 8(1), 2001507, DOI: [10.1002/admi.202001507](https://doi.org/10.1002/admi.202001507).
- 24 M. Tong, Y. Lan, Q. Yang and C. Zhong, Exploring the structure-property relationships of covalent organic frameworks for noble gas separations, *Chem. Eng. Sci.*, 2017, 168, 456–464, DOI: [10.1016/j.ces.2017.05.004](https://doi.org/10.1016/j.ces.2017.05.004).
- 25 D. Ongari, A. V. Yakutovich, L. Talirz and B. Smit, Building a Consistent and Reproducible Database for Adsorption Evaluation in Covalent–Organic Frameworks, *ACS Cent. Sci.*, 2019, 5(10), 1663–1675, DOI: [10.1021/acscentsci.9b00619](https://doi.org/10.1021/acscentsci.9b00619).
- 26 R. Mercado, R.-S. Fu, A. V. Yakutovich, L. Talirz, M. Haranczyk and B. Smit, In Silico Design of 2D and 3D Covalent Organic Frameworks for Methane Storage Applications, *Chem. Mater.*, 2018, 30(15), 5069–5086, DOI: [10.1021/acs.chemmater.8b01425](https://doi.org/10.1021/acs.chemmater.8b01425).
- 27 J. S. D. Vos, S. Borgmans, P. V. D. Voort, S. M. J. Rogge and V. V. Speybroeck, ReDD-COFFEE: a ready-to-use database of covalent organic framework structures and accurate force fields to enable high-throughput screenings, *J. Mater. Chem. A*, 2023, 11(14), 7468–7487, DOI: [10.1039/D3TA00470H](https://doi.org/10.1039/D3TA00470H).
- 28 D. N. Ampong, *et al.*, Advances and challenges in covalent organic frameworks as an emerging class of materials for energy and environmental concerns, *Coord. Chem. Rev.*, 2024, 519, 216121, DOI: [10.1016/j.ccr.2024.216121](https://doi.org/10.1016/j.ccr.2024.216121).
- 29 H. Vardhan, G. Rummer, A. Deng and S. Ma, Large-Scale Synthesis of Covalent Organic Frameworks: Challenges and Opportunities, *Membranes*, 2023, 13(8), 696, DOI: [10.3390/membranes13080696](https://doi.org/10.3390/membranes13080696).
- 30 Y. J. Colón and R. Q. Snurr, High-throughput computational screening of metal–organic frameworks, *Chem. Soc. Rev.*, 2014, 43(16), 5735–5749, DOI: [10.1039/C4CS00070F](https://doi.org/10.1039/C4CS00070F).
- 31 E. Ren, P. Guilbaud and F.-X. Coudert, High-throughput computational screening of nanoporous materials in targeted applications, *Digital Discovery*, 2022, 1(4), 355–374, DOI: [10.1039/D2DD00018K](https://doi.org/10.1039/D2DD00018K).
- 32 A. Hamed Mashhadzadeh, *et al.*, Boron nitride nanoslits for water desalination via forward osmosis: A molecular dynamics study, *J. Mol. Graphics Modell.*, 2025, 134, 108905, DOI: [10.1016/j.jmgm.2024.108905](https://doi.org/10.1016/j.jmgm.2024.108905).
- 33 A New Version of the Insertion Particle Method for Determining the Chemical Potential by Monte Carlo Simulation: Molecular Simulation: Vol 5, No 6, accessed, Jan. 14, 2026, Online, available, <https://www.tandfonline.com/doi/abs/10.1080/08927029108022424>.
- 34 T. Yildız and I. Erucar, Revealing the performance of bio-MOFs for adsorption-based uremic toxin separation using molecular simulations, *Chem. Eng. J.*, 2022, 431, 134263, DOI: [10.1016/j.cej.2021.134263](https://doi.org/10.1016/j.cej.2021.134263).
- 35 B. Akkoca Palabıyık, M. Batyrow and I. Erucar, Computational investigations of Bio-MOF membranes for uremic toxin separation, *Sep. Purif. Technol.*, 2022, 281, 119852, DOI: [10.1016/j.seppur.2021.119852](https://doi.org/10.1016/j.seppur.2021.119852).
- 36 T. Fabiani, E. Ricci, C. Boi, S. Dimartino and M. G. D. Angelis, In silico screening of nanoporous materials for urea removal in hemodialysis applications, *Phys. Chem. Chem. Phys.*, 2023, 25(35), 24069–24080, DOI: [10.1039/D3CP01510F](https://doi.org/10.1039/D3CP01510F).
- 37 L. Cao, Recent advances in the application of machine-learning algorithms to predict adsorption energies, *TRECHEM*, 2022, 4(4), 347–360, DOI: [10.1016/j.trechm.2022.01.012](https://doi.org/10.1016/j.trechm.2022.01.012).
- 38 M. Zarghami Dehaghani and M. G. De Angelis, Machine learning-driven computational screening of covalent organic frameworks for gas separation applications, *Sep. Purif. Technol.*, 2025, 377, 134358, DOI: [10.1016/j.seppur.2025.134358](https://doi.org/10.1016/j.seppur.2025.134358).
- 39 H. Wang, Y. Li, X. Xuan, K. Wang, Y. Yao and L. Pan, Machine Learning Accelerated Discovery of Covalent Organic Frameworks for Environmental and Energy

- Applications, *Environ. Sci. Technol.*, 2025, **59**(13), 6361–6378, DOI: [10.1021/acs.est.5c00390](https://doi.org/10.1021/acs.est.5c00390).
- 40 S. C. E. COFs, core-cof/CoRE-COF-Database, 2025, accessed, Nov. 16, 2025, Online, available, <https://github.com/core-cof/CoRE-COF-Database>.
- 41 S. L. Mayo, B. D. Olafson and W. A. Goddard, DREIDING: a generic force field for molecular simulations, *J. Phys. Chem.*, 1990, **94**(26), 8897–8909, DOI: [10.1021/j100389a010](https://doi.org/10.1021/j100389a010).
- 42 T. Fabiani, *Transforming hemodialysis with new materials: an experimental and modelling approach*, PhD thesis, University of Edinburgh, Edinburgh Research Archive, 2025, DOI: [10.7488/era/6475](https://doi.org/10.7488/era/6475).
- 43 C. E. Wilmer, K. C. Kim and R. Q. Snurr, An Extended Charge Equilibration Method, *J. Phys. Chem. Lett.*, 2012, **3**(17), 2506–2511, DOI: [10.1021/jz3008485](https://doi.org/10.1021/jz3008485).
- 44 Chemically Meaningful Atomic Charges That Reproduce the Electrostatic Potential in Periodic and Nonperiodic Materials | *J. Chem. Theory Comput.*, accessed, Mar. 03, 2026, Online, available, https://pubs.acs.org/doi/full/10.1021/ct100125x?casa_token=-mYKTRMMl00AAAAA%3ACvgXG9ZpjVnI-BAoR_4zA9YhmanMMGDpRq-nW0SEBIrRFyzl6SBLIN0v7Ecgs8Hs6yQjGsjNZROSQGDxw.
- 45 G. Zhao and Y. G. Chung, PACMAN: A Robust Partial Atomic Charge Predictor for Nanoporous Materials Based on Crystal Graph Convolution Networks, *J. Chem. Theory Comput.*, 2024, **20**(12), 5368–5380, DOI: [10.1021/acs.jctc.4c00434](https://doi.org/10.1021/acs.jctc.4c00434).
- 46 M. Lísál, J. Kolafa and I. Nezbeda, An examination of the five-site potential (TIP5P) for water, *J. Chem. Phys.*, 2002, **117**(19), 8892–8897, DOI: [10.1063/1.1514572](https://doi.org/10.1063/1.1514572).
- 47 S. Weerasinghe and P. E. Smith, A Kirkwood–Buff Derived Force Field for Mixtures of Urea and Water, *J. Phys. Chem. B*, 2003, **107**(16), 3891–3898, DOI: [10.1021/jp022049s](https://doi.org/10.1021/jp022049s).
- 48 D. Dubbeldam, S. Calero, D. E. Ellis and R. Q. Snurr, RASPA: molecular simulation software for adsorption and diffusion in flexible nanoporous materials, *Mol. Simul.*, 2016, **42**(2), 81–101, DOI: [10.1080/08927022.2015.1010082](https://doi.org/10.1080/08927022.2015.1010082).
- 49 S. Kang and J. Kim, Property-guided inverse design of metal-organic frameworks using quantum natural language processing, *npj Comput. Mater.*, 2025, **11**(1), 321, DOI: [10.1038/s41524-025-01806-z](https://doi.org/10.1038/s41524-025-01806-z).
- 50 X. Peng and H. Wang, Machine learning based screening of organic frameworks for separation of CF₄/N₂, C₂F₆/N₂, and SF₆/N₂, *Chem. Eng. Sci.*, 2024, **296**, 120280, DOI: [10.1016/j.ces.2024.120280](https://doi.org/10.1016/j.ces.2024.120280).
- 51 P. Sedgwick, Pearson's correlation coefficient, *BMJ*, 2012, **345**, e4483, DOI: [10.1136/bmj.e4483](https://doi.org/10.1136/bmj.e4483).
- 52 I. T. Jolliffe and J. Cadima, Principal component analysis: a review and recent developments, accessed, Jun. 02, 2026, Online, available, DOI: [10.1098/rsta.2015.0202](https://doi.org/10.1098/rsta.2015.0202).
- 53 PCA, scikit-learn. accessed, Jun. 02, 2026, Online, available, <https://scikit-learn/stable/modules/generated/sklearn.decomposition.PCA.html>.
- 54 pandas.util.hash_pandas_object — pandas 3.0.3 documentation, accessed, Jun. 02, 2026, Online, available, https://pandas.pydata.org/docs/reference/api/pandas.util.hash_pandas_object.html?utm_source=chatgpt.com.
- 55 O. Kramer, Scikit-Learn, in *Machine Learning for Evolution Strategies*, ed. O. Kramer, Cham, Springer International Publishing, 2016, pp. 45–53, DOI: [10.1007/978-3-319-33383-0_5](https://doi.org/10.1007/978-3-319-33383-0_5).
- 56 PyTorch, *PyTorch*, accessed, Jan. 11, 2026, Online, available: <https://pytorch.org/>.
- 57 T. Head, M. Kumar, H. Nahrstaedt, G. Louppe, and I. Shcherbatyi, *scikit-optimize/scikit-optimize*, Zenodo, 2020, DOI: [10.5281/zenodo.4014775](https://doi.org/10.5281/zenodo.4014775).
- 58 Nested versus non-nested cross-validation, scikit-learn, accessed, Jan. 11, 2026, Online, available: https://scikit-learn/stable/auto_examples/model_selection/plot_nested_cross_validation_iris.html.
- 59 M. M. Ahsan, M. A. P. Mahmud, P. K. Saha, K. D. Gupta and Z. Siddique, Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance, *Technologies*, 2021, **9**(3), 52, DOI: [10.3390/technologies9030052](https://doi.org/10.3390/technologies9030052).
- 60 A. Hamed Mashhadzadeh, *et al.*, Machine Learning-Assisted design of boron and nitrogen doped graphene nanosheets with tailored thermomechanical properties, *Comput. Mater. Sci.*, 2024, **240**, 112998, DOI: [10.1016/j.commatsci.2024.112998](https://doi.org/10.1016/j.commatsci.2024.112998).
- 61 M. Allaoui, M. L. Kherfi, and A. Cheriet, Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study, in *Image and Signal Processing*, ed. A. El Moataz, D. Mammass, A. Mansouri, and F. Nouboud, Cham, Springer International Publishing, 2020, pp. 317–325, DOI: [10.1007/978-3-030-51935-3_34](https://doi.org/10.1007/978-3-030-51935-3_34).
- 62 F. Sahigara, D. Ballabio, R. Todeschini and V. Consonni, Defining a novel k-nearest neighbours approach to assess the applicability domain of a QSAR model for reliable predictions, *J. Cheminf.*, 2013, **5**, 27, DOI: [10.1186/1758-2946-5-27](https://doi.org/10.1186/1758-2946-5-27).
- 63 J. K. Jaiswal and R. Samikannu, Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression, in *2017 World Congress on Computing and Communication Technologies (WCCCT)*, 2017, pp. 65–68, DOI: [10.1109/WCCCT.2016.25](https://doi.org/10.1109/WCCCT.2016.25).
- 64 X. Zhang and W. Shang, Prediction of Rn, N₂, and O₂ Adsorption and Separation in MOFs and MOF@AC Composites via Machine Learning and Graph Neural Networks, *Ind. Eng. Chem. Res.*, 2025, **64**(50), 24227–24239, DOI: [10.1021/acs.iecr.5c04453](https://doi.org/10.1021/acs.iecr.5c04453).
- 65 I.-T. Sung, Y.-H. Cheng, C.-M. Hsieh and L.-C. Lin, Machine Learning for Gas Adsorption in Metal–Organic Frameworks: A Review on Predictive Descriptors, *Ind. Eng. Chem. Res.*, 2025, **64**(4), 1859–1875, DOI: [10.1021/acs.iecr.4c03500](https://doi.org/10.1021/acs.iecr.4c03500).
- 66 P. Politzer, J. S. Murray and T. Clark, Halogen bonding and other σ -hole interactions: a perspective, *Phys. Chem. Chem. Phys.*, 2013, **15**(27), 11178–11189, DOI: [10.1039/c3cp00054k](https://doi.org/10.1039/c3cp00054k).
- 67 M. Z. Dehaghani, *et al.*, Evaluation of water desalination performances of functionalized nanoporous graphene membranes by molecular dynamics simulation, *Comput. Mater. Sci.*, 2023, **228**, 112308, DOI: [10.1016/j.commatsci.2023.112308](https://doi.org/10.1016/j.commatsci.2023.112308).

- 68 R. Bosque and J. Sales, Polarizabilities of Solvents from the Chemical Composition, *J. Chem. Inf. Comput. Sci.*, 2002, **42**(5), 1154–1163, DOI: [10.1021/ci025528x](https://doi.org/10.1021/ci025528x).
- 69 M. R. Karimpour, D. V. Fedorov and A. Tkatchenko, Molecular Interactions Induced by a Static Electric Field in Quantum Mechanics and Quantum Electrodynamics, *J. Phys. Chem. Lett.*, 2022, **13**(9), 2197–2204, DOI: [10.1021/acs.jpclett.1c04222](https://doi.org/10.1021/acs.jpclett.1c04222).
- 70 D. G. Truhlar, Dispersion Forces: Neither Fluctuating Nor Dispersing, *J. Chem. Educ.*, 2019, **96**(8), 1671–1675, DOI: [10.1021/acs.jchemed.8b01044](https://doi.org/10.1021/acs.jchemed.8b01044).
- 71 G. Cavallo, *et al.*, The Halogen Bond, *Chem. Rev.*, 2016, **116**(4), 2478–2601, DOI: [10.1021/acs.chemrev.5b00484](https://doi.org/10.1021/acs.chemrev.5b00484).
- 72 E. Arunan, *et al.*, Definition of the hydrogen bond (IUPAC Recommendations 2011), *Pure Appl. Chem.*, 2011, **83**(8), 1637–1641, DOI: [10.1351/PAC-REC-10-01-02](https://doi.org/10.1351/PAC-REC-10-01-02).
- 73 J. A. Platts, S. T. Howard and B. R. F. Bracke, Directionality of Hydrogen Bonds to Sulfur and Oxygen, *J. Am. Chem. Soc.*, 1996, **118**(11), 2726–2733, DOI: [10.1021/ja952871s](https://doi.org/10.1021/ja952871s).
- 74 F. Gustafsson, Comparing Random Forest, XGBoost and Neural Networks With Hyperparameter Optimization by Nested Cross-Validation, *Bachelor Project*, 2019.
- 75 T. G. Dietterich, An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting and Randomization, *Machine Learning*, 2000, **40**, 139–157, DOI: [10.1023/A:1007607513941](https://doi.org/10.1023/A:1007607513941).
- 76 R. Akbani, S. Kwek, and N. Japkowicz, Applying Support Vector Machines to Imbalanced Datasets, in *Machine Learning: ECML 2004*, ed. J.-F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi, Berlin, Heidelberg, Springer, 2004, pp. 39–50, DOI: [10.1007/978-3-540-30115-8_7](https://doi.org/10.1007/978-3-540-30115-8_7).
- 77 L. Grinsztajn, E. Oyallon, and G. Varoquaux, Why do tree-based models still outperform deep learning on tabular data?, 2022, *arXiv*, arXiv:2207.08815, DOI: [10.48550/arXiv.2207.08815](https://doi.org/10.48550/arXiv.2207.08815).
- 78 An Imbalance Regression Approach to Toxicity Prediction of Chemicals for Potential Use in Environmentally Acceptable Lubricants, *ACS Appl. Mater. Interfaces*, accessed, Jan. 28, 2026, Online, available: <https://pubs.acs.org/doi/full/10.1021/acsami.4c10622>.
- 79 H. Xiong and A. Yao, Deep Imbalanced Regression via Hierarchical Classification Adjustment, 2023, *arXiv*, arXiv:2310.17154, DOI: [10.48550/arXiv.2310.17154](https://doi.org/10.48550/arXiv.2310.17154).
- 80 Prediction of methane adsorption isotherms in metal-organic frameworks by neural network synergistic with classical density functional theory - ScienceDirect, accessed, Apr. 17, 2025, Online, available: <https://www.sciencedirect.com/science/article/pii/S1385894723003431#b0325>.
- 81 H. Tang and J. Jiang, Active learning boosted computational discovery of covalent-organic frameworks for ultrahigh CH₄ storage, *AIChE J.*, 2022, **68**(11), e17856, DOI: [10.1002/aic.17856](https://doi.org/10.1002/aic.17856).