JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

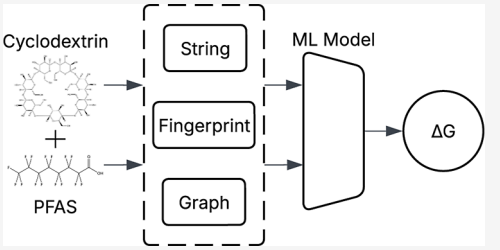
Evaluating Molecular Representations for Predicting Cyclodextrin-PFAS Binding Energy with Machine Learning: Domain Transfer and Data Limitations

Cole Brzakala,* Othonas A. Moultos, Jan Peter van der Hoek, and Riccardo Taormina

Cite This: https://doi.org/10.1021/acs.jcim.5c03121 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Per- and polyfluoroalkyl substances (PFAS) persist in water systems and resist conventional removal methods such as activated carbon, which shows reduced efficiency with short-chain PFAS and in the presence of dissolved organic matter. Cyclodextrin-based polymers (CDPs) have emerged as sustainable alternatives, with competitive and selective PFAS adsorption capabilities. These polymers consist of glucose-based cyclodextrin (CD) units that can form host−guest inclusion complexes with PFAS pollutants. However, these binding interactions are not fully understood or quantified. We conducted an evaluation of machine learning approaches to model these host−guest interactions, providing insights into predictive capabilities for later CDP design. This study systematically compares molecular representations (Mordred, ECFP, ChemBERTa, UniMol2, etc.) across several machine learning architectures to predict CD-PFAS binding energies. First, we generated molecular embeddings of 3459 experimental host−guest pairs in the OpenCycloDB data set and 63 external CD-PFAS pairs. We then compared these embeddings via AlignedUMAP visualizations and nearest neighbor analyses. Next, we trained and evaluated predictive models using these embeddings on the OpenCycloDB data set, exploring the effectiveness of transfer learning and finetuning techniques. We finally tested model generalizability on two external experimental CD-PFAS binding data sets. All embeddings captured relevant chemical features, where UniMol2 differed most from other methods in embedding space analysis. Predictive models performed variably based on embedding choice and architecture, with the best-performing combination achieving moderate accuracy on the OpenCycloDB data set. Embeddings pretrained on large molecular data sets and finetuning the ChemBERTa embeddings both showed predictive improvements. However, external validation revealed limited generalizability to CD-PFAS complexes, highlighting domain shift challenges. Notably, leave-one-out cross-validation on the external PFAS-specific data indicated that training on in-domain data improved predictive performance at the cost of generalizability. This work demonstrates that molecular representation choice is critical for small-data host−guest binding prediction. However, domain shift between general CD data and specialized CD−PFAS applications remains a fundamental challenge, for which transfer learning and finetuning may offer potential solutions for future data-driven pipelines for CDP design and sustainable PFAS removal.



INTRODUCTION

Modern industrial and consumer practices have introduced complex pollutants to water systems globally. Among these pollutants, per- and polyfluoroalkyl substances (PFAS) have emerged as a significant concern due to associated health risks, even at low (ng/L) concentrations, and their persistence in the environment.[1−3] A diverse set of over 12,000 compounds characterized by their fluorinated carbon chains, PFAS can vary in size and structure. The combination of this heterogeneity, low concentrations, and competition with dissolved organic matter (DOM) of PFAS makes removal of all species from water systems challenging.[4,5]

In humans, PFAS exposure links to adverse health effects, including immune and endocrine system dysfunction, reproductive and development dysfunction, and increased risk of certain cancers, but the full extent of their impact is still being studied.[6−9] As a result, the European Food Safety Authority established a tolerable weekly PFAS intake of 4.4 ng/kg body weight per week.[10] PFAS commonly occur in water sources, stemming from their widespread use in various applications, including firefighting foams, nonstick coatings, and water-repellent fabrics.[11] This has led to the establishment of regulatory limits on PFAS concentrations in drinking water: 500 ng/L for total PFAS and 100 ng/L for the sum of 20 specific PFAS in the European Union.[12] In the United States, individual PFAS limits are as low as 4 ng/L for six PFAS species.[13] While conventional adsorption methods for PFAS removal, such as activated carbon and ion-exchange resins, can remove PFAS, they have limitations in effectiveness.

Received: January 13, 2026
Revised: June 10, 2026
Accepted: June 12, 2026

Specifically, these methods struggle in the presence of DOM and with PFAS of diverse chain lengths, particularly short-chain PFAS. In addition, these technologies can have expensive regeneration and disposal costs, making them less suitable as a long-term solution.[14,15] Other membrane-based methods, including reverse osmosis, can filter PFAS but produce highly concentrated waste streams.[16] These challenges necessitate the development of more effective and sustainable remediation strategies.

Recently, cyclodextrin-based polymers (CDPs) have emerged as a promising alternative for PFAS removal. These adsorbents consist of cyclodextrins (CD), cyclic oligosaccharides composed of glucose units connected via cross-linkers. CDs can contain six (α-CD), seven (β-CD), or eight (γ-CD) glucose units, shown in Figure 1A–C. CDPs bind to other

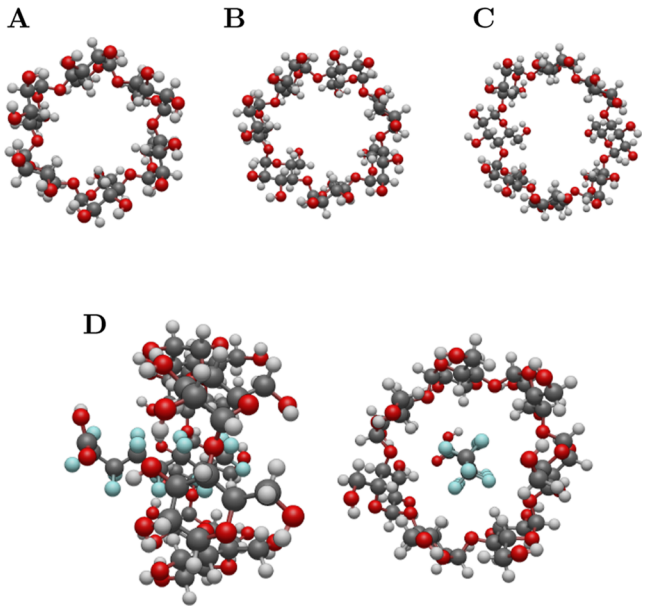


Figure 1. Molecular structures of cyclodextrins and host–guest complexation mechanism. (A–C) show the cyclic structures of α-cyclodextrin (α-CD, 6 glucose units), β-cyclodextrin (β-CD, 7 glucose units), and γ-cyclodextrin (γ-CD, 8 glucose units). (D) shows two perspectives of the host–guest complexation mechanism where guest molecules (PFOA shown) are encapsulated within the hydrophobic CD cavity. Here, carbon (dark gray), hydrogen (light gray), oxygen (red), and fluorine (light blue) atoms are shown.

molecules, including PFAS, in many ways. Of these, host–guest complexation is one of the major capture mechanisms, where pollutants are restricted in the CD cavity, shown in Figure 1D.[17,18] These sustainable polymers have already been shown to surpass traditional treatment methods in PFAS removal in realistic water matrices.[19,20] Through tuning of the cross-linker chemistry, CDPs can be further optimized to selectively bind PFAS. Albaiee et al. demonstrated that rigid aromatic group CDP cross-linkers enhance organic micropollutant binding with adsorption rate constants 15 to 200 times greater than activated carbon and previous CDP designs.[21] Reduction of the nitrile groups in tetrafluoroterephthalonitrile-CDPs improved binding affinity toward anionic PFAS species.[22] Targeted adsorption of PFOA was achieved by Xiao et al. (2017) by cross-linking β-CD with decafluorobiphenyl with similar removal metrics to activated carbon.[23] However, the full effects of CDP cross-linker chemistry on binding affinity are not yet well understood and quantified, which limits the efficient design of CDPs for PFAS removal.[24] This cross-linker design space is vast, and laboratory experiments incur high financial and time costs for few results. Fluorine-19 nuclear magnetic resonance (19F NMR) spectroscopy or isothermal titration calorimetry (ITC) are typically used to examine PFAS adsorption by CDs, but these methods are low throughput and require specialized equipment and expertise.[25] In addition, experimental results rarely include all relevant removal criteria, increasing difficulties in comparison.[26] Therefore, traditional experimental screening approaches are inefficient for exploring the CDP design space.

Machine learning (ML) methods can accelerate the design process and significantly enhance our understanding of the molecular interactions involved in CD-PFAS binding to narrow down the vast chemical space to a few high-potential candidates. Aiming to model these interactions, ML can provide insights into the factors that influence binding behavior, which is critical for the design of effective CDPs. Researchers have successfully applied ML methods to similar domains, including drug discovery and materials science. These methods predict molecular properties, including drug toxicity and potency in Quantitative Structure–Activity Relationship (QSAR) research and polymer science.[27−29] In particular, graph neural networks (GNNs) have been used to represent chemicals in different contexts, such as predicting combustion-related properties and enzyme screening, often outperforming previous architectures.[30−33] Still, ML implementations studying host–guest inclusion are limited.

Ma et al. (2022) developed three ML models (artificial neural network, support vector machine, and logistic regression) to predict CD inclusion complex formation using 200 compounds, resulting in the discovery of three new inclusion complexes.[34] Jeschke et al. (2019) used 3D spectrophore descriptors with a random forest model to predict binding constants of CD complexes, which was competitive with commercial models (R^2 = 0.95).[35] Using five ML methods, Di et al. (2020) developed consensus models to predict binding free energies of CD complexes with small organic molecules, further explaining interactions with molecular docking and free energy calculations.[36] Cai et al. (2025) used a LightGBM with 2D and 3D molecular descriptors to predict binding free energies (R^2 = 0.80) of CD complexes with volatile terpenes.[37] Zhao et al. (2019) used LightGBM, random forest, and deep learning models to predict complexation free energy (R^2 = 0.86) between CD and guest molecules, emphasizing the synergy between ML and molecular simulation.[38]

However, few studies have focused on PFAS as guest molecules, which have unique properties due to the fluorinated carbon chains that may affect interactions with CDs. The scarcity of public CD-PFAS data limits these ML models' size and performance. Ling et al. (2019) analyzed different molecular descriptors using 200 experimental micropollutant data points, with 3656 descriptors for each, drawing mechanistic conclusions that support CDP functionalization for targeted adsorption.[39] Other studies have focused on more broad adsorbent classes or specific PFAS targets. For example, Zhang et al. (2026) employed machine learning-assisted computational molecular modeling to investigate the removal of perfluorobutanoic acid (PFBA), a short-chain PFAS, using covalent organic frameworks (COFs), proposing triazine-based COFs as the best PFBA adsorbent.[40] Recent advancements in

https://doi.org/10.1021/acs.jcim.5c03121
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

molecular representations and transfer learning techniques offer promising avenues to overcome the data limitation that challenges CD-PFAS ML modeling.[41,42] A crucial step toward efficient CDP design is understanding the molecular representations that can effectively capture the features of PFAS and CDPs and how transfer learning could enhance model performance under data scarcity.

This study evaluates molecular representations including descriptor-based methods (OpenCycloDB Enriched and Mordred[43,44]), fingerprints (ECFP and ECFP+[45]), learned graph embeddings (UniMol2,[46] GROVER,[47] and Chemeleon[48]), and SMILES-based encoders (ChemBERTa[49]) for encoding the features of PFAS and CDs. We first generated molecular embeddings for host−guest pairs in the recently published OpenCycloDB data set,[50] which contains 3459 experimental CD-guest binding energies. We then created embedding UMAP visualizations and calculated cosine and Tanimoto similarity metrics to evaluate embedding differences. We used these embeddings to train and evaluate ML models predicting equilibrium binding affinities. After hyperparameter tuning, we tested the models on external experimental CD-PFAS binding data. Using the two external test sets, we tested the models' out-of-domain performance and generalizability. The results emphasized the role of molecular representations on prediction capability in low-data CD−PFAS host−guest systems, where data limitations led to poor generalizability to CD-PFAS systems. However, pretrained models and finetuning improved CD-guest prediction, which should be validated in subsequent polymer-focused ML studies. Ultimately, this research aims to guide future data-driven CDP design to enable effective and sustainable PFAS removal.

METHODS

Data Pipeline

We established the following pipeline to evaluate the most effective encoding of CD and guest molecules and binding predictions. First, we compiled data sets containing the host and guest molecules and the associated binding energy values. Next, we converted these data sets into various molecular representations suitable for machine learning: strings, molecular fingerprints, and graphs. Then, we trained predictive models on these representations to learn the mapping between the molecular features and the binding energies. Finally, we compared these predictions to the true values to evaluate model performance. Figure 2 shows the high-level architecture of the modeling approach.

Data Sources

To train and validate the ML models, we utilized data sets from three different sources. (1) We used the OpenCycloDB set to train and test all predictive models using all embeddings.[50] We split this data set of 3459 pairs into training ($OCDB_{train}$), validation ($OCDB_{val}$), and test ($OCDB_{test}$) sets, using the validation set for hyperparameter tuning and the test set for final evaluation. We collected two additional external test sets from existing experimental literature to test the models on the PFAS/CD domain. (2) The external PFAS test set (E-$PFAS_{test}$) consists of 21 $\alpha/\beta/\gamma$-CD and diverse PFAS pairs,[51] and (3) the external β-CD test set (E-β-CD_{test}) consists of 42 β-CD derivatives (aminated and thiolated) and core PFAS pairs.[52] We then combined these external test sets (n = 63) and used them to train the LOOCV models.

OpenCycloDB. The OpenCycloDB data set is a comprehensive collection of CD-guest binding energies.[50] It includes experimental binding energies for various CD derivatives and their interactions with different guest molecules. The data set provides a relatively diverse resource for training machine learning models to predict binding affinities given the historical lack of large, public data sets. The data includes information for CD type, guest molecule, experimental pH and temperature, and the corresponding binding energy, along with source data. The data set includes 1767 guest molecules and 15 CDs and CD derivatives, resulting in 3459 total molecule pairs and associated binding energies. While ML-based analyses with comparably sized data sets have been performed, the OpenCycloDB data set is one of the first publicly available CD binding data sets of this size. This data set covers a domain distinct from that used in the application, as it includes only 96 host−guest pairs in which the guest molecule is fluorinated. OpenCycloDB researchers curated the experimental results from literature, after which the authors performed data preprocessing to ensure consistency and quality. Figure 3A shows the distribution of binding energies (ΔG), indicating a range of binding affinities across different CD-guest pairs. Figure 3B shows the composition of CD hosts in the data set, where the base α, β, and γ appear along with their respective derivatives. These derivatives include alkyl, hydroxyalkyl, acetyl, carboxyl, and sulfate functional groups.

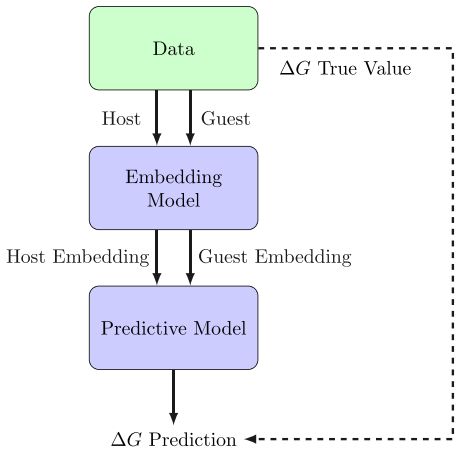


Figure 2. High-level overview of the model architecture for predicting CD-guest binding energies. The pipeline begins with molecular data (host and guest structures), which are processed through an embedding model to generate numerical feature representations. These embeddings are then concatenated and fed into a predictive model that outputs the predicted Gibbs free energy of binding (ΔG), which are compared to the true values from the data set.

The OpenCycloDB data set was carefully curated to exclude outliers, duplicate entries, nonstandard pH and temperature conditions, nonwater solvents, and measurements obtained with less accurate techniques. Even so, its experimental origin inevitably introduces some variability. The data set includes measurements from various sources, which used different experimental conditions and techniques, contributing to variability in the reported binding energies. Additionally, the data set includes only standard pH and temperature values, which may not capture the full range of conditions under which CD-guest interactions can occur. Therefore, there is inherent label noise, which can limit model performance and generalizability. However, the lack of large, standardized, and public data sets in this domain necessitates the use of such data for model training and evaluation, and the OpenCycloDB data set represents a significant step forward in providing a resource for ML applications in CD binding prediction.

External Test Data. We sourced two sets of external test data from recent experimental studies on CD-PFAS interactions. The E-$PFAS_{test}$ includes binding pairs of α, β, and γ-CDs with emerging PFAS,[51] summarized in Table S1. Only one PFAS guest in $OCDB_{train}$ is also present in E-$PFAS_{test}$, accounting for 9.1% of the guests. The E-β-CD_{test} includes binding energies for β-CD derivatives with core PFAS compounds,[52] listed in Table S2. Of the four core PFAS in E-β-CD_{test}, one is present in $OCDB_{train}$, accounting for 25% of the guests. However, none of the β-CD derivatives in this set are present in the training data. These include β-CD with amine and thiol functional

https://doi.org/10.1021/acs.jcim.5c03121
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

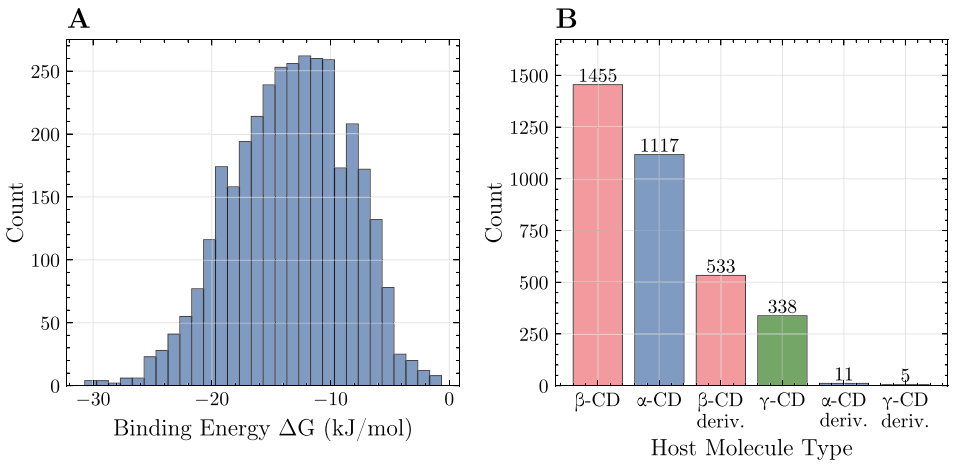


Figure 3. Distribution of binding energies (ΔG) in the OpenCycloDB data set (A) showing the range and frequency of experimental binding affinities across 3459 CD-guest pairs, and composition of CD hosts (B) illustrating the relative abundance of α, β, and γ-CDs and their derivatives in the data set. The distribution in (A) has the following statistics: mean = −13.37 kJ/mol, median = −13.13 kJ/mol, standard deviation = 4.89 kJ/mol.

groups. While these data sets also have noise inherent to the measurements, their conditions match those in the OpenCycloDB. These external test sets provide a challenging evaluation of the models' ability to generalize to new chemistries.

Molecular Embedding Generation

OpenCycloDB Enriched. The OpenCycloDB data set includes molecular embeddings for each host−guest pair.[50] This OpenCycloDB Enriched embedding includes molecular descriptors for the host and guest molecules. These descriptors include molecular weight, charge, topological polar surface area, and other relevant chemical properties, generated using RDKit, a popular cheminformatics toolkit.[53] In addition, the authors converted the SMILES strings representing the CD and guest molecules to 10-value vectors generated via a variant of the word2vec approach for ease of implementation in machine learning methods.[54] The resulting data set contains 40 numerical features for each CD-guest pair.

Mordred. Mordred is a cheminformatics tool that calculates physicochemical descriptors from molecular structures.[43,44] These descriptors capture fundamental chemical properties such as molecular weight, hydrophobicity, aromaticity, and various topological indices, which are useful for characterizing molecular interactions. Mordred descriptors have been used in many applications, such as polyimide screening.[55] Mordred generates a set of 1613 numerical features (combination of continuous and categorical) for each molecule. We concatenated the Mordred descriptors from both the host and guest molecules, resulting in a 3226-dimensional vector for each CD-guest pair. This representation provides a complementary approach to fingerprints and learned embeddings, emphasizing explicit chemical properties.

ECFP and ECFP+. We generated Extended-Connectivity Fingerprints (ECFP) using RDKit. ECFP is a type of molecular fingerprint that captures the local environment of each atom in a molecule, allowing for the representation of complex molecular structures.[45] ECFP is widely used in cheminformatics and has been shown to be effective in capturing molecular features relevant for tasks like polymer property prediction and protein binding affinity prediction.[56−58] We generated the fingerprints using Morgan's algorithm[59] with a radius of 2 and a length of 1024 bits, resulting in a 2048-length vector for each CD-guest pair. While ECFP are fingerprints by definition, we will refer to them as embeddings throughout this work for simplicity.

The ECFP+ representation extends this by incorporating additional experimental variables, i.e., temperature and pH, which can influence binding interactions, resulting in a total of 2050 numerical features for each CD-guest pair. We created this representation to evaluate the influence of these variables' inclusion on binding affinity predictions. However, pH and temperature values are not always recorded in experimental data sets, including the external test sets used in this study.[26]

UniMol2. We created the UniMol2 embeddings with the pretrained UniMol2 model, a transformer-based architecture designed for molecular representation learning that mirrors scaling law developments in natural language processing and computer vision.[46] This method utilizes a 3D molecular graph representation, which aims to integrate features at the atomic, graph, and geometry structure levels. Graph representations treat a molecule's atoms as the nodes and the bonds as edges in a computational graph. UniMol2 authors pretrained the model on a large data set of molecular structures, allowing it to learn generalizable features that can be applied to various downstream molecular prediction tasks, including enzyme prediction.[60] Multiple model sizes are available; we used the smallest model in this study to reduce computational requirements at 84 M parameters. We generated the UniMol2 embeddings by passing the molecular graph through the transformer architecture, resulting in a 768-dimensional embedding for each molecule in a CD-guest pair for a final vector of 1536 dimensions.

GROVER. The graph-based GROVER pretrained model uses a message-passing mechanism to update node representations based on their neighbors, effectively capturing the local and global structure of 2D molecular graphs.[47] The authors pretrained the model on a data set of 10 M unlabeled molecules using self-supervised learning, allowing it to learn generalizable features that can be applied to various downstream tasks. In addition to their ability to learn from nonstructured data, graph-based methods have shown potential in areas of molecular discovery with little data.[61,62] GROVER generates 4800-dimensional embedding vectors for each molecule in a CD-guest pair, resulting in a total size of 9600 dimensions.

Chemeleon. Chemeleon is a graph neural network (GNN) model pretrained on a large corpus of molecules with their physicochemical properties, enabling it to learn representations that capture chemistry-relevant features.[48] Unlike methods that learn primarily from structure alone, Chemeleon integrates physicochemical information into its pretraining objective, allowing it to develop embeddings that encode property-aware relationships between molecules. The model generates 2048-dimensional embeddings for each molecule in a CD-guest pair, resulting in 4096-dimensional pair representations.

ChemBERTa. The ChemBERTa model encodes molecular structures as sequences of tokens. While the architecture is based on RoBERTa (a Transformer encoder variant), the authors pretrained the model on SMILES strings using a custom BPE tokenizer, following the RoBERTa pretraining strategy.[49] Researchers then trained the transformer using 10 M SMILES from the PubChem database. By using a pretrained model to generate ChemBERTa embeddings, the model can leverage the learned representations of

https://doi.org/10.1021/acs.jcim.5c03121
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

molecular structures to improve performance on this prediction application. Researchers have applied similar SMILES-based approaches in other domains, including de novo molecular design.[63,64] The ChemBERTa model generates a 768-dimensional embedding for each of the host and guest in a CD-guest pair, generating a 1536-dimensional pair representation.

Representation Summary. Table 1 summarizes the different molecular representations used in this study, including their types, dimensions, and representation basis.

Table 1. Summary Comparison of the Eight Molecular Representations Evaluated in this Study

Representation	Type	Dimension	Representation Basis
OpenCycloDB Enriched	descriptors	40	engineered descriptors + SMILES word2vec
Mordred	descriptors	3226	engineered physicochemical descriptors
ECFP	fingerprint	2048	Morgan fingerprint from 2D connectivity
ECFP+	fingerprint	2050	Morgan fingerprint + pH and temperature
UniMol2	transformer	1536	3D molecular encoder
GROVER	GNN	9600	structural 2D graph encoder
Chemeleon	GNN	1536	physicochemical 2D graph encoder
ChemBERTa	transformer	1536	SMILES transformer encoder

Embedding Comparison

To visualize the distribution of molecular embeddings across training and test sets, we used Aligned Uniform Manifold Approximation and Projection (AlignedUMAP) to jointly embed all data sets into a shared 2D coordinate space. This approach enables direct visual comparison of embedding distributions across different molecular types and data sets. AlignedUMAP is particularly effective for high-dimensional data, preserving both local and global structure while maintaining consistency across the aligned coordinate system.[65] Crucially, this method allows for the visualization of the high dimensional embedding space, but it must not be used to draw conclusions about the relationships between molecules. The dimensionality reduction process can distort distances and relationships in the original space.

To assess the quality and consistency of the generated molecular embeddings, we calculated similarity metrics both for individual embedding comparisons (used in nearest neighbor analysis) and for data set-level comparisons between training and validation sets. We employed cosine similarity for dense embeddings, while we used Tanimoto similarity for sparse count-based representations. These metrics have been shown to capture differences in molecular representations for continuous and discrete cases.[66,67] We calculated average values for data set-level comparisons to provide an overall measure of similarity between the training and validation sets.

For individual comparisons of nonfingerprint embeddings (OpenCycloDB Enriched, Mordred, UniMol2, GROVER, Chemeleon, and ChemBERTa/ChemBERTa Finetuned), cosine similarity (CS) was calculated using eq 1

$$\mathrm{CS} = \frac{\mathbf{e}_i \cdot \mathbf{e}_j}{\|\mathbf{e}_i\| \cdot \|\mathbf{e}_j\|} \tag{1}$$

where $\mathbf{e}_i$ and $\mathbf{e}_j$ denote individual embedding vectors, $\|\cdot\|$ denotes the L2 norm of the vector, and $\cdot$ denotes the dot product. Cosine similarity values range from −1 to 1, where 1 indicates identical direction, 0 indicates orthogonality, and −1 indicates opposite direction. Though Mordred descriptors have categorical features, cosine similarity captured the variations in molecular properties enough to observe trends between the data sets. For individual comparisons of fingerprint embeddings (ECFP and ECFP+), Tanimoto similarity (TS) was used due to its appropriateness for binary and count-based fingerprint representations, as shown in eq 2

$$\mathrm{TS} = \frac{\mathbf{e}_i \cdot \mathbf{e}_j}{\|\mathbf{e}_i\|^2 + \|\mathbf{e}_j\|^2 - \mathbf{e}_i \cdot \mathbf{e}_j} \tag{2}$$

Tanimoto similarity values range from 0 to 1, where 0 indicates no shared features and 1 indicates identical feature sets. For data set-level comparisons between training and validation embedding sets, average cosine similarity ($\overline{\mathrm{CS}}$) is computed according to eq 3

$$\overline{\mathrm{CS}} = \frac{1}{|V| \cdot |T|} \sum_{v \in V} \sum_{t \in T} \frac{\mathbf{e}_v \cdot \mathbf{e}_t}{\|\mathbf{e}_v\| \cdot \|\mathbf{e}_t\|} \tag{3}$$

where V is the set of validation embeddings and T is the set of training embeddings. Similarly, for data set-level comparisons of fingerprint representations (ECFP and ECFP+) between training and validation sets, average Tanimoto similarity ($\overline{\mathrm{TS}}$) is calculated using eq 4

$$\overline{\mathrm{TS}} = \frac{1}{|V| \cdot |T|} \sum_{v \in V} \sum_{t \in T} \frac{\mathbf{e}_v \cdot \mathbf{e}_t}{\|\mathbf{e}_v\|^2 + \|\mathbf{e}_t\|^2 - \mathbf{e}_v \cdot \mathbf{e}_t} \tag{4}$$

These metrics provide insight into both individual molecular similarity (for nearest neighbor identification) and overall representational consistency across different data sets.

In the nearest neighbor analysis, we identified the top k most similar training molecules for each validation molecule. For dense embeddings, average validation cosine similarities were computed in eq 5

$$\overline{\mathrm{CS}}_{\mathrm{V}} = \frac{1}{n_k} \sum_{t=1}^{n_k} \frac{\mathbf{e}_V \cdot \mathbf{e}_t}{\|\mathbf{e}_V\| \cdot \|\mathbf{e}_t\|} \tag{5}$$

where n_k is the number of neighbors, $\mathbf{e}_V$ is the validation embedding, and $\mathbf{e}_t$ are training embeddings. For fingerprint embeddings (ECFP and ECFP+, when pH and temperature features were available), average validation Tanimoto similarities are defined according to eq 6

$$\overline{\mathrm{TS}}_{\mathrm{V}} = \frac{1}{n_k} \sum_{t=1}^{n_k} \frac{\mathbf{e}_V \cdot \mathbf{e}_t}{\|\mathbf{e}_V\|^2 + \|\mathbf{e}_t\|^2 - \mathbf{e}_V \cdot \mathbf{e}_t} \tag{6}$$

Predictive Models

We selected two types of predictive models to evaluate the effectiveness of different molecular embeddings in predicting CD-guest binding energies: Light Gradient Boosting Machine (LGBM) and Feedforward Neural Network (FNN). Both predictive models used two molecular embeddings (host and guest molecules) as input to output the pair's binding energy. The first predictive model, LGBM, is a gradient-boosting framework that uses tree-based learning algorithms.[68] LGBM is known for its efficiency and scalability, making it suitable for data sets of varying size and high-dimensional feature spaces. Researchers have widely used it in various machine learning tasks, including regression and classification problems in chemical domains such as toxicity prediction.[69] In addition, the LGBM model is the only publicly available model trained on the OpenCycloDB data set, allowing for direct comparison with existing work.[50]

The second predictive model type we chose was the FNN, a type of artificial neural network that consists of an input layer, one or more hidden layers, and an output layer. FNNs are known for their ability to model complex nonlinear relationships and have been widely used in various machine learning tasks, including virtual screening and activity prediction in drug discovery.[70,71] However, these networks can require significant time and computational resources and data for training if not properly sized and optimized. They can also suffer in performance if not properly regularized, especially in low data settings. In this work, we employed the FNN to enable joint, end-to-end optimization with the Deep Learning-based embedding models, ensuring that molecular representations can be directly tuned for the binding prediction task.

https://doi.org/10.1021/acs.jcim.5c03121
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

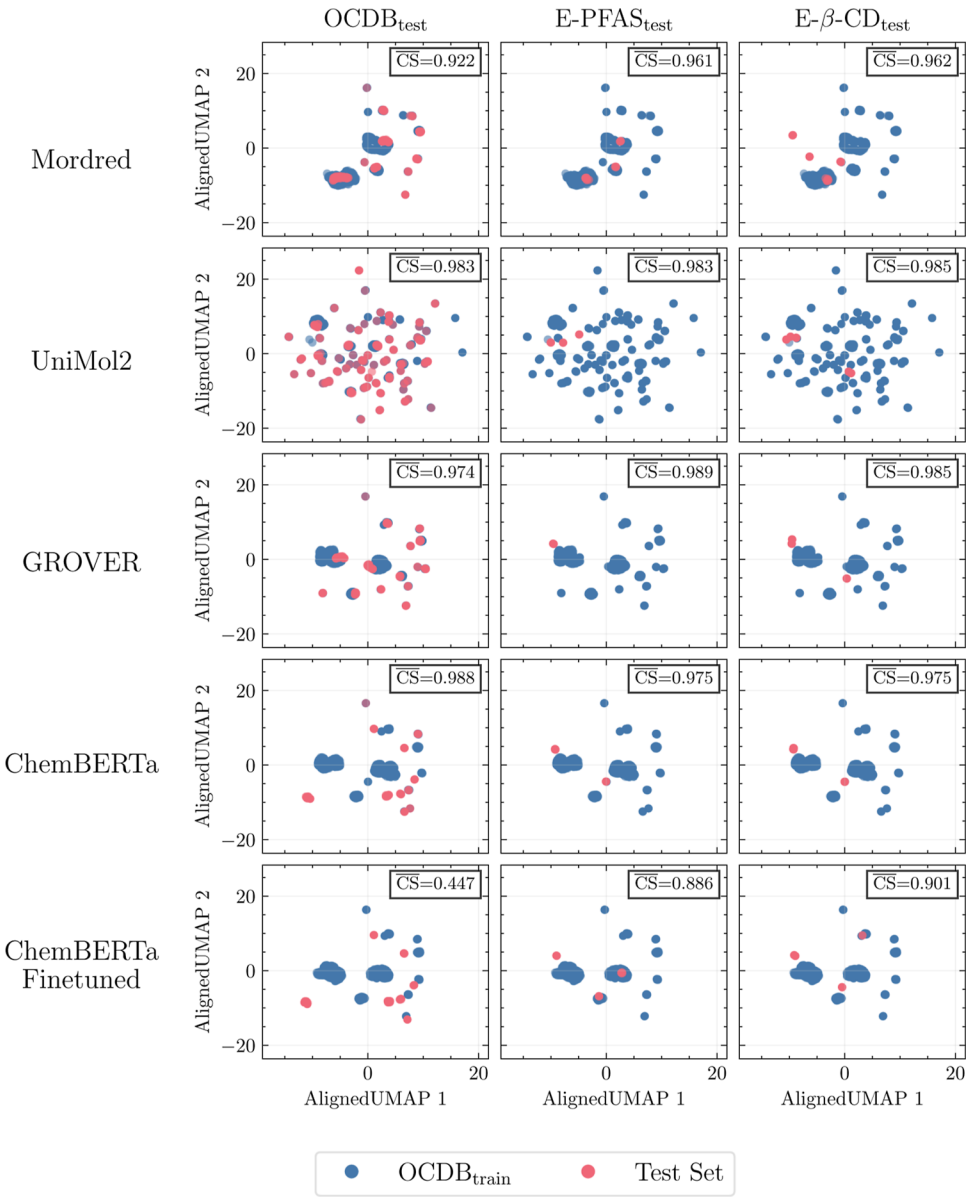


Figure 4. AlignedUMAP visualization of host molecular embeddings across different representation methods. The blue points represent the OpenCycloDB training data (OCDB$_{train}$). The pink points represent the test set. The three columns show the OpenCycloDB test set (OCDB$_{test}$), the external PFAS test set (E-PFAS$_{test}$), the external β-CD test set (E-β-CD$_{test}$). All embeddings were jointly reduced to a shared 2D coordinate space, enabling direct visual comparison of how training and test embeddings distribute relative to each other. The average cosine similarity ($\overline{CS}$) or average Tanimoto similarity ($\overline{TS}$) between training and test embeddings is displayed for each panel. All embeddings show tight cluster-like distributions except UniMol2, which shows cloud-like spread.

Training and Model Selection

For model training, we split the OpenCycloDB data set into OCDB$_{train}$ (80%), OCDB$_{val}$ (10%), and OCDB$_{test}$ (10%). We stratified the split by CD type (α, β, and γ) to ensure that each set contained a representative distribution of CD derivatives. There is a significant limitation with this method of stratified splitting, as there are host−guest pairs with the same host and similar guests across the training, validation, and test sets. This leads to performance metrics that are not fully representative of the model's ability to generalize to new chemistries. Scaffold-based splits or leave-one-out splits by host or guest would be more appropriate for evaluating generalization, but we chose this method to enable direct comparison with existing work on the OpenCycloDB data set, which used a similar splitting strategy.[50] We therefore rely on the external test sets for evaluating generalization to the PFAS domain, using the LOOCV models for domain-specific performance comparison.

All input embeddings were scaled by each feature before training for FNN models. We used OCDB$_{train}$ for model training and the OCDB$_{val}$ for hyperparameter tuning. Due to limited data availability, the final models were retrained on the joint OCDB$_{train}$-OCDB$_{val}$ data set. Overfitting is unlikely, as model complexity and regularization were determined beforehand on a smaller data set and kept fixed. Additionally, early stopping limits identified from model training were applied during final model training to limit overfitting. Thus, each final model serves as the criterion for model comparison.

We conducted training on a machine with an AMD Ryzen 9 7950X 16-Core CPU, 64GB RAM, and an NVIDIA GeForce RTX 4090 GPU. The training times varied depending on the model and embedding type, with LGBM models typically training faster than FNNs. We monitored the training process using Weights & Biases (wandb) for experiment tracking and hyperparameter optimization.[72]

https://doi.org/10.1021/acs.jcim.5c03121
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

We evaluated model performance using standard regression metrics, including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and coefficient of determination (R^2). While none of these metrics capture every aspect of prediction performance individually, RMSE and MAE measure prediction error in the original units, while R^2 provides a measure of how well the model explains the variance in the data.[73] The combination of these metrics, which is commonly used in ML applications, provides a comprehensive view of model performance.[74] These metrics are defined in eqs 7−9

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{7}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{8}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{9}$$

Here, y_i is the true binding energy, $\hat{y}_i$ is the predicted binding energy, $\bar{y}$ is the mean of the true values, and n is the number of samples.

To complement these standard regression metrics, we evaluated Spearman's rank correlation coefficient ρ (eq 10) and Kendall's rank correlation coefficient τ (eq 11), which assess monotonic agreement between observed and predicted values.

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{10}$$

$$\tau = \frac{(n_c - n_d)}{n(n-1)/2} \tag{11}$$

In eq 10, d_i is the difference between the ranks of y_i and $\hat{y}_i$ for sample i.[75] In eq 11, n_c is the number of concordant prediction pairs (where the ordering of both predictions agrees) and n_d is the number of discordant prediction pairs (where the ordering disagrees).[76] Both metrics range from −1 to +1, where +1 indicates perfect positive monotonic agreement, 0 indicates no monotonic relationship, and −1 indicates perfect negative monotonic agreement. While Spearman's ρ is computationally simpler, Kendall's τ provides a direct probabilistic interpretation of the ranking agreement.

We trained each model for a maximum of 1000 epochs with early stopping based on validation loss (RMSE) to prevent overfitting. The early stopping criteria are defined in Table S3. We selected the best-performing model from 200 runs for each embedding type based on $\mathrm{OCDB_{val}}$ performance.

Embedding Model Finetuning

Due to computational constraints and the large size of pretrained models, we focused finetuning on ChemBERTa. We selected the best-performing ChemBERTa FNN model from the initial phase and unfroze the embedding weights so they could be updated together with the FNN during training. We then trained the combined model, applying the same hyperparameter tuning and early stopping criteria as before, and repeated this process for 30 runs. The details of the finetuning hyperparameter tuning are defined in Table S3. This approach allows the embedding model to adapt specifically to the CD-guest binding prediction task, potentially improving performance by learning task-specific features.

Leave-One-Out Cross-Validation

We performed Leave-One-Out Cross-Validation (LOOCV) to evaluate model performance under extreme data scarcity conditions that are representative of emerging contaminant domains where limited experimental data exists.[77] By training on the combined external test data ($n = 63$), this approach simulates the scenario where only a small number of CD-PFAS binding measurements are available for model development. In LOOCV, each sample in the data set is held out once as a test sample while the remaining $n - 1$ samples are used for training. This process is repeated n times, where n is the total number of samples in the data set. The LOOCV framework ensures robust performance estimation by maximizing the use of available data while providing unbiased evaluation of each embedding's effectiveness under low-data regimes, at the cost of increased computational demand of training n models. We performed LOOCV with an LGBM predictive model directly on the combined $\mathrm{E\text{-}PFAS_{test}}$ and $\mathrm{E\text{-}\beta\text{-}CD_{test}}$ ($n = 63$).

For a given performance metric M (such as MAE, RMSE, or R^2), the LOOCV score is calculated according to eq 12

$$M_{\mathrm{LOOCV}} = \frac{1}{n}\sum_{i=1}^{n} M_i \tag{12}$$

where M_i is the metric calculated between the predicted and true value for the single held-out sample i, using a model trained on the remaining $n - 1$ samples. This approach provides a robust estimate of model performance by ensuring that every sample serves as both training and test data exactly once, thereby maximizing the use of available data for both training and validation. We conducted LOOCV on each embedding for 200 runs for hyperparameter tuning, and we selected the model with the lowest RMSE.

RESULTS AND DISCUSSION

Host Embedding Analysis

We first used AlignedUMAP to jointly project the learned embeddings into a shared 2D coordinate space and evaluated their effectiveness. Figure 4 shows the AlignedUMAP visualizations for all host embeddings across the $\mathrm{OCDB_{test}}$, $\mathrm{E\text{-}PFAS_{test}}$, and $\mathrm{E\text{-}\beta\text{-}CD_{test}}$ compared to the $\mathrm{OCDB_{train}}$. Since all data sets share the same coordinate system, the plots allow direct visual comparison of how training and test embeddings align. While these plots show trends across the data sets, choice of hyperparameters can influence the visualization. Each plot shows the first two AlignedUMAP components for the training and test embeddings, with the average similarity metric between the training and test embeddings displayed in the corner of each panel. Additional host AlignedUMAP plots for the ECFP and Chemeleon representations are provided in Figure S1.

In Figure 4, cluster-like distributions can be observed for all AlignedUMAP visualizations except for the UniMol2 embeddings, which show cloud-like spread. The Chemeleon embeddings in Figure S1 also shows larger spread, but this is likely due to the visualization parameters, not the embedding itself. The aligned coordinate space is particularly valuable for identifying the distinctive nature of UniMol2 embeddings, as the distinctiveness is not due to different embedding orientations but rather fundamental differences in how this representation organizes the host chemical space. This is likely due to the generation of 3D conformers in the UniMol2 embedding generation, leading to similar embeddings for the same molecule with different conformers.

The host plots in the first column show that all embeddings of the $\mathrm{OCDB_{test}}$ overlap the $\mathrm{OCDB_{train}}$, the expected behavior from being drawn from the same distribution containing the same host molecules. The ChemBERTa and ChemBERTa Finetuned models show similar behavior for the $\mathrm{OCDB_{test}}$ set. For the external test sets ($\mathrm{E\text{-}PFAS_{test}}$ and $\mathrm{E\text{-}\beta\text{-}CD_{test}}$) in the middle and right columns, the embeddings cover much less of the space. This is similarly expected, as the OpenCycloDB host domain shares a limited number of compounds with the external test domain. Again, the UniMol2 test embeddings show the most spread of the embeddings, likely due to the 3D

https://doi.org/10.1021/acs.jcim.5c03121
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

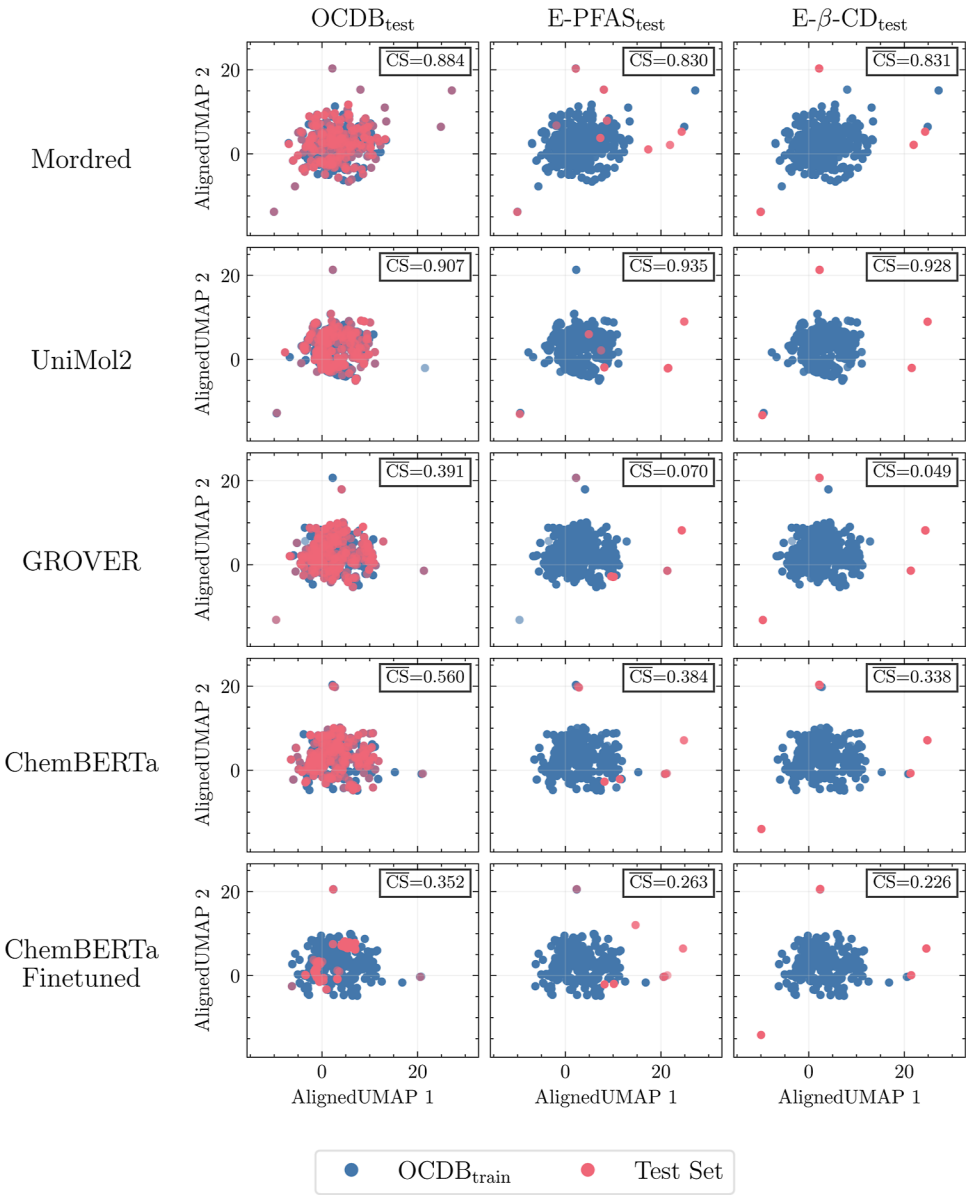


Figure 5. AlignedUMAP visualization of guest molecular embeddings across different representation methods. The blue points represent the OCDB$_{train}$ data. The pink points represent the test set. The three columns show the OpenCycloDB validation set (OCDB$_{val}$), the external PFAS test set (E-PFAS$_{test}$), and the external β-CD test set (E-β-CD$_{test}$). All embeddings were jointly embedded into the same 2D coordinate space, enabling direct visual comparison of how different molecular representations organize the guest chemical space. The average cosine similarity ($\overline{CS}$) or average Tanimoto similarity ($\overline{TS}$) between training and test embeddings is displayed for each panel. Cluster distributions are more spread out than in the host embeddings, reflecting the greater chemical diversity of guest molecules.

conformational information they contain. The other embeddings show more clustered test points, indicating that the embeddings in each test set are similar to each other. This behavior could affect model performance, as the models may struggle to differentiate between the different test points' binding interaction differences if their molecular embeddings are too similar.

The average similarity metrics, $\overline{TS}$ and $\overline{CS}$, provide further insight into the performance of the host embeddings. These metrics capture only broad trends in the data, as they average the differences between all pairs of training and test embeddings. The similarity metric of the OpenCycloDB Test data set can be considered the baseline embedding spread, as it is drawn from the same distribution as the OCDB$_{train}$ that it is being compared to. Overall, the similarity metrics for most host embeddings are relatively high (above 0.9), indicating little diversity in the host domain. Additionally, the similarity metrics remain high for the external test sets, indicating that the host embeddings in these sets are similar to those in the training set. The ECFP, UniMol2, GROVER, and ChemBERTa embeddings show this trend, with similarity metrics above 0.9 for all test sets.

Examination of the ChemBERTa Finetuned model reveals two notable trends. First, the OCDB$_{test}$ similarity metric drops dramatically from $\overline{CS}$ = 0.988 for the base ChemBERTa model to $\overline{CS}$ = 0.447 for the finetuned model. This indicates that the finetuned model has learned an embedding space more specific to CDs, making differences between hosts more pronounced. Second, the similarity metrics for the external test sets also drop between the base and finetuned models. This is

https://doi.org/10.1021/acs.jcim.5c03121
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

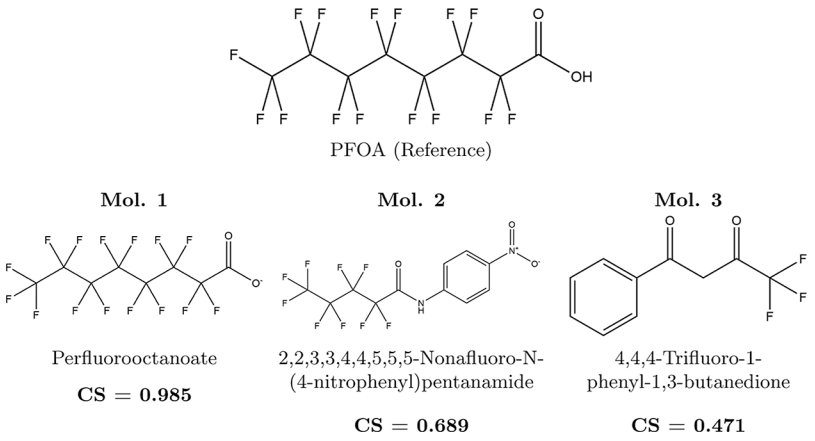


Figure 6. Molecular similarity comparison between PFOA (reference) and three structurally diverse molecules. Cosine similarity (CS) values indicate the degree of structural resemblance to PFOA, with higher values representing greater similarity. Three molecules with varying functional groups and chain lengths are shown to illustrate the range of similarity captured by the GROVER embeddings.

consistent with the idea that the finetuned model has learned a more specific representation of the data, increasing the spread of the host embedding distribution. However, the values drop much less than for the $OCDB_{test}$, likely because the external test sets contain a smaller, less diverse set of hosts. In addition, the ChemBERTa Finetuned embeddings show different clustering behavior for the external test sets than in the base case. There is an increased number of test clusters in the external test sets for the finetuned model compared to the base model. Further, these clusters differ between the $E\text{-}PFAS_{test}$ and $E\text{-}\beta\text{-}CD_{test}$, whereas they match in the ChemBERTa plots. Apart from these trends, the other embeddings show consistent similarity metrics across all test sets, indicating that the host embeddings are not the limiting factor in the generalizability of the models to the validation sets.

Guest Embedding Analysis

The guest AlignedUMAP visualizations in Figure 5 reveal more varied clustering behavior between embeddings and test sets, with all embeddings exhibiting cloud-like distributions rather than the tight clusters seen in the host plots. Similar to the host visualizations, some AlignedUMAP plots show all external test guests clustered together, indicated by the pink points in the second and third columns. Additional guest AlignedUMAP plots for the ECFP and Chemeleon representations are provided in Figure S2.

All representations show more cloud-like spread in the guest embeddings than that of the hosts. This is expected, as the guest molecules in OpenCycloDB are much more diverse than the hosts. The ChemBERTa and ChemBERTa Finetuned models show a notable difference in clustering behavior of the $OCDB_{test}$ set. After finetuning the embedding model, the test points show tighter clustering behavior than in the base case. This could be due to embeddings becoming more specific to the OCDB guest domain and the predictive task, instead of the more general chemical space captured by the base ChemBERTa model. For the $E\text{-}PFAS_{test}$ set, the guest embeddings follow similar spatial behaviors, where the Mordred and UniMol2 embeddings display more spread. However, for the $E\text{-}\beta\text{-}CD_{test}$ set, all test points are mapped roughly to the same regions of the embedding space. This is likely due to the small number of guest molecules in this set, which are also relatively similar to each other.

The $OCDB_{test}$ similarity metrics drop significantly (0.56 or lower) for the guest embeddings compared to the host embeddings, except for the UniMol2 embeddings, which remain relatively high (0.983). Figure S4 shows the distributions of the UniMol2 features, where nearly all features are centered around zero with relatively low standard deviations, except for a few features with much higher value means and standard deviations. This indicates that these differences in relative magnitudes cause the larger guest similarity values. As UniMol2 embeddings were pretrained on 10 million diverse molecules, the feature magnitudes could well represent meaningful chemical information, but cause the similarity metrics to be less informative. All other embeddings show much lower similarity metrics relative to the host values, indicating that the guest embeddings are more spread out in the embedding space. This is expected, as the guest molecules in OpenCycloDB are much more diverse than the hosts.

GROVER shows the largest drop in similarity between the $OCDB_{test}$ and the external test sets, as shown in the third row of Figure 5, indicating that the embeddings of the PFAS guests are distinctly different from the majority of guests in the training set. This alone does not mean that these embeddings will be better or worse than the others at predicting the binding affinities between CDs and PFAS. Still, it does demonstrate how the different embedding models represent the chemical space of the guests. Similar to what is observed with the host embeddings, the ChemBERTa Finetuned embeddings show lower similarity scores than the base ChemBERTa embeddings, indicating that the finetuning process has increased differences in the embedding space between all guests. This could help the model better differentiate between guests, potentially improving predictive performance.

While these plots and metrics visualize differences between the embeddings, they cannot be used to draw conclusions about the relationships between molecules. Neighbor analysis was performed on the guest embeddings to further evaluate whether models can effectively embed guest molecules sharing similar features between the training and test sets.

Nearest Neighbor Analysis

To further investigate the guest embedding space, we analyzed the most similar molecules from the training set to molecules in the external test sets. We first illustrate this with GROVER

https://doi.org/10.1021/acs.jcim.5c03121
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

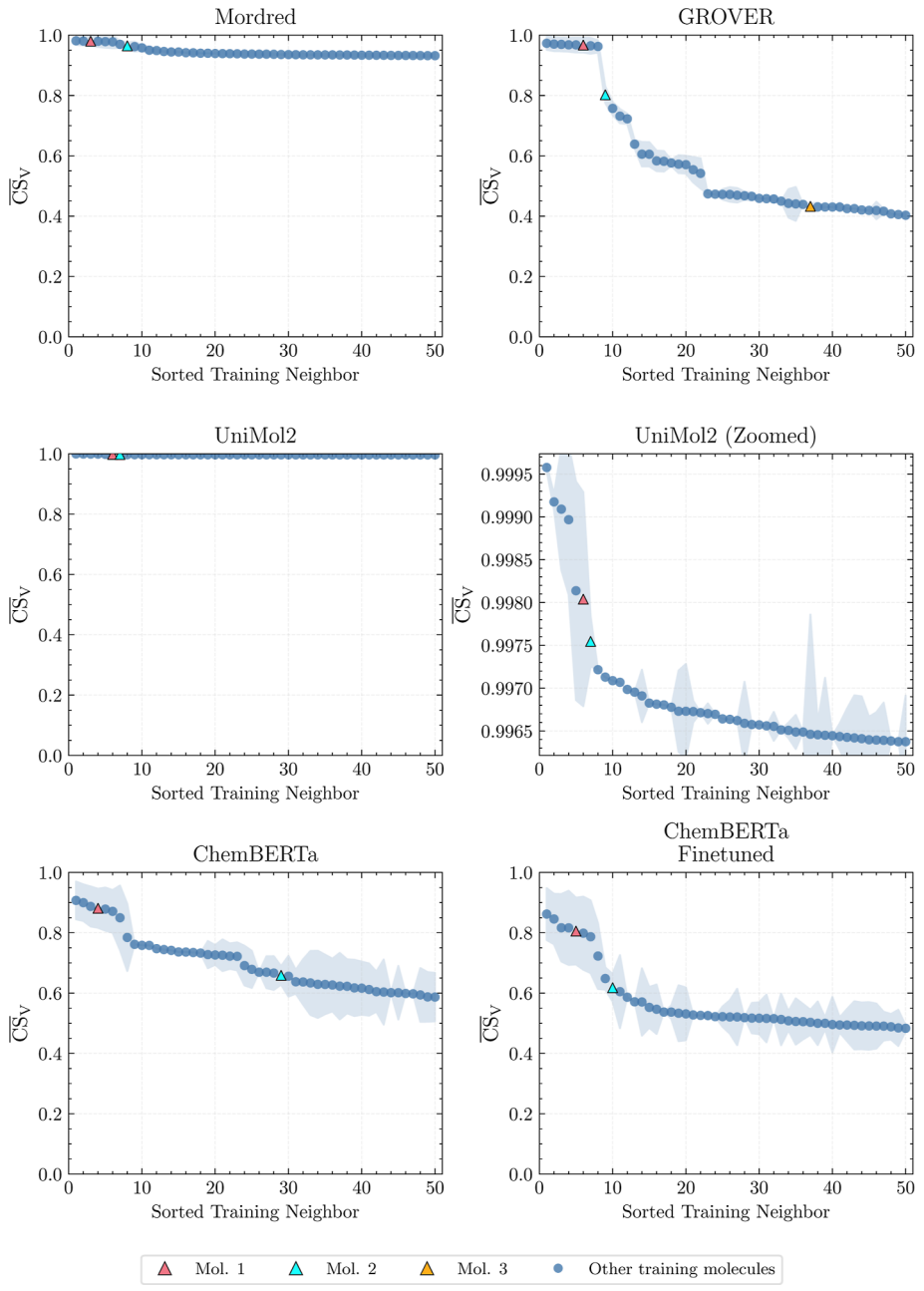


Figure 7. Ranked distribution of cosine similarity values averaged across all external test molecules. The top 50 most similar guest molecules from $OCDB_{train}$ are displayed across different embedding methods. The ECFP representation uses Tanimoto similarity, while all other methods use cosine similarity. Three molecules of varying similarity to PFOA are highlighted for reference (see Figure 6). The dominant trend of approximately 10 high similarity molecules holds across all embeddings, but specific values and other behaviors vary.

embeddings to examine a specific reference molecule, perfluorooctanoic acid (PFOA), which appears in both the training and test sets. Figure 6 shows the three most similar molecules to PFOA in the GROVER embedding space, along with their cosine similarity scores. Additional host and guest UMAP visualizations used in the neighbor analysis are provided in Figure S3 in the Supporting Information

In Figure 6, perfluorooctanoate (Mol. 1) is the most similar molecule to PFOA, as it is the conjugate base of the reference molecule. As expected, they share a high cosine similarity value of 0.985. The second most similar molecule is 2,2,3,3,4,4,5,5,5-Nonafluoro-N-(4-nitrophenyl)pentanamide (Mol. 2), with a cosine similarity of 0.689. This molecule contains a shorter perfluorinated chain and an amide functional group, making it structurally similar but distinct from PFOA. The third most similar molecule is 4,4,4-Trifluoro-1-phenyl-1,3-butanedione (Mol. 3), with a cosine similarity of 0.471. This molecule has a trifluoromethyl group and a diketone functional group, representing a more significant structural deviation from PFOA. These results demonstrate how one embedding method (GROVER) captures structural differences among molecules with varying degrees of similarity to PFOA. We then extended this analysis to all external test molecules and compared the different embedding methods. Figure 7 shows the ranked distribution of the top 50 training molecules with the highest average similarity across all external test molecules for each embedding method. Additional neighbor analysis

https://doi.org/10.1021/acs.jcim.5c03121
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 2. Performance Comparison of Different Model Types and Molecular Representations on the OpenCycloDB Dataset[a]

Model	Representation	$OCDB_{train}$	$OCDB_{val}$	$OCDB_{test}$				
		RMSE (kJ/mol)	RMSE (kJ/mol)	RMSE (kJ/mol)	MAE (kJ/mol)	R^2	ρ	τ
LGBM	OpenCycloDB Enriched	0.80	2.72	2.58	1.79	0.71	0.84	0.66
	Mordred	0.61	2.51	2.44	**1.64**	0.73	**0.86**	**0.69**
	ECFP	1.14	2.60	2.55	1.75	0.71	0.85	0.68
	ECFP+	0.75	2.56	2.48	1.69	0.73	0.85	0.69
	UniMol2	**0.18**	2.89	2.89	2.10	0.63	0.80	0.61
	GROVER	0.38	**2.48**	**2.42**	1.67	**0.74**	**0.86**	**0.69**
	Chemeleon	2.56	3.28	3.40	2.56	0.49	0.71	0.52
	ChemBERTa	0.30	2.82	2.79	1.95	0.65	0.83	0.64
FNN	OpenCycloDB Enriched	1.24	2.73	3.22	2.47	0.54	0.76	0.58
	Mordred	1.85	2.86	2.85	2.13	0.64	0.80	0.62
	ECFP	0.92	2.88	2.66	1.84	0.69	0.83	0.66
	ECFP+	**0.88**	2.88	3.22	2.24	0.54	0.78	0.59
	UniMol2	1.82	2.83	2.67	1.92	0.68	0.83	0.65
	GROVER	1.18	3.06	3.11	2.17	0.57	0.77	0.59
	Chemeleon	3.60	3.56	3.92	3.12	0.32	0.61	0.43
	ChemBERTa	2.36	2.78	2.71	1.98	0.67	0.83	0.64
	ChemBERTa FT	1.94	**2.62**	**2.49**	**1.78**	**0.73**	**0.85**	**0.68**

[a]Models were trained on the $OCDB_{train}$ (n = 2767) and evaluated on the $OCDB_{val}$ (n = 346) and $OCDB_{test}$ (n = 346). RMSE and MAE are reported in kJ/mol. Bold values indicate the best performance for each model type. LGBM tends to outperform FNN across most representations, but finetuning ChemBERTa with FNN yields competitive results. For correlation metrics (ρ and τ), an asterisk (*) denotes $p \geq 0.0001$; unmarked values satisfy $p < 0.0001$.

plots for the ECFP and Chemeleon representations are provided in Figure S3 in the Supporting Information

In Figure 7, similarity value distributions vary significantly across embedding methods. In general, each model identifies approximately 10 training molecules highly similar to the external test molecules. After this group, all models show a drop in similarity, though the extent and rate vary across models. The GROVER and ChemBERTa embeddings show similar behavior, with a gradual decrease in similarity values after the initial group of highly similar molecules. Again, the UniMol2 embeddings show a different trend, with consistently high similarity values across all top 50 molecules relative to the other embedding types. However, when examining this high-similarity region, the general trend still holds at a much different scale. The Mordred embeddings show similarly high similarity scores, though to a lesser extent. The ChemBERTa Finetuned embeddings show a more gradual drop in similarity values compared to the base ChemBERTa embeddings. In addition, the similarity values for the 40 lowest ranked molecules are lower than those of the base ChemBERTa embeddings. This indicates that the finetuning process has increased the spread of the guest embeddings, similar to what was observed in the AlignedUMAP visualizations.

The three molecules from Figure 6 are highlighted in Figure 7 to provide context for the similarity values. All three molecules appear in the top 50 most similar molecules in the GROVER embedding, with Mol. Three not appearing in the top 50 lists for the remaining embeddings. While the order of the molecules remains the same across all embeddings, the relative similarity values differ significantly. For example, the finetuning process on the ChemBERTa embeddings causes an increase in similarity value to the external test molecules for Mol. 2. This indicates that finetuning caused ChemBERTa to represent this molecule as more similar to the external test molecules than the base ChemBERTa model, as expected. Overall, this analysis demonstrates how different embedding methods capture molecular similarity in distinct ways, which can impact the performance of predictive models trained on these embeddings.

OpenCycloDB Model Performance

Table 2 summarizes the evaluation results for embeddings and predictive models on the OpenCycloDB data set. The results show the training RMSE, validation RMSE, test RMSE, test MAE, and test R^2 scores across different combinations of model architectures and molecular representations. Both the LGBM and FNN models show signs of overfitting to the training data, indicated by the lower training RMSE compared to the validation RMSE. While overfitting often implies poor generalization, in this case, the validation and test performances remain comparable, indicating that the models still generalize. This outcome is consistent with the use of regularization during hyperparameter tuning and an extensive search (200 runs) of the hyperparameter space. The apparent overfitting may therefore reflect the OpenCycloDB data set's characteristics, such as limited size or measurement noise, rather than a true failure of model generalization.

The LGBM model trained on GROVER embeddings performs best on both validation and test sets, followed closely by the LGBM models trained on ECFP+ and Mordred embeddings and the FNN model trained on ChemBERTa Finetuned embeddings. While this result is positive, the performance remains relatively modest, such that these models cannot replace experimental or molecular dynamics simulation results without additional improvement. Still, these models can provide insight into the effectiveness of different embeddings in capturing information relevant to CD/guest interactions. However, the performance of the models on the OpenCycloDB test set does not necessarily indicate how well they will generalize to PFAS-specific external test sets.

As shown in Table 2, for all embeddings except ChemBERTa, LGBM outperforms FNN, suggesting that

https://doi.org/10.1021/acs.jcim.5c03121
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 3. External Test Set Performance Comparison Across Different Model Architectures and Molecular Representations[a]

Model	Representation	E-PFAS$_{test}$ (n = 21)					E-β-CD$_{test}$ (n = 42)				
		RMSE (kJ/mol)	MAE (kJ/mol)	R^2	P	τ	RMSE (kJ/mol)	MAE (kJ/mol)	R^2	ρ	τ
LGBM	ECFP	6.63	5.32	0.02	0.84	0.68	4.73	3.68	−0.03	0.62	0.46
	Mordred	**5.02**	**3.78**	**0.44**	**0.88**	0.70	4.66	**3.46**	0.00	**0.70**	**0.51**
	UniMol2	5.42	3.92	0.35	**0.88**	**0.76**	**4.62**	3.48	**0.02**	0.53	0.37
	GROVER	10.12	7.89	−1.28	0.11*	0.06*	7.64	6.52	−1.68	0.24*	0.17*
	Chemeleon	7.50	5.80	−0.25	0.26*	0.15*	5.95	4.79	−0.62	0.34*	0.23*
	ChemBERTa	6.23	4.67	0.14	0.82	0.52	5.03	4.09	−0.16	0.53	0.39
FNN	ECFP	8.60	7.36	−0.65	**0.76**	**0.52**	6.95	5.57	−1.22	0.41*	0.30*
	Mordred	6.98	5.43	−0.09	0.69	0.48*	6.33	5.42	−0.84	**0.62**	**0.43**
	UniMol2	11.44	8.96	−1.91	0.10*	0.17*	9.57	8.15	−3.20	0.19*	0.16*
	GROVER	16.40	14.96	−4.99	−0.57*	−0.34*	17.22	16.61	−12.62	−0.23*	−0.19*
	Chemeleon	8.14	6.28	−0.48	0.25*	0.18*	**4.92**	**3.77**	−0.11	0.02*	0.00*
	ChemBERTa	6.37	5.01	0.10	0.63	0.44	6.84	5.68	−1.15	0.59	0.42
	ChemBERTa FT	**6.20**	**4.79**	**0.15**	0.62	0.44	7.02	5.98	−1.26	**0.62**	0.42

[a]Models trained on OCDB$_{train}$ (n = 2767) were evaluated on E-PFAS$_{test}$ (n = 21) containing CD-PFAS binding data and E-β-CD$_{test}$ (n = 42) containing β-CD complexes with various guests. RMSE and MAE are reported in kJ/mol. Bold values indicate the best performance on each test set. Most models show little to no generalization to these external datasets. For correlation metrics (ρ and τ), an asterisk (*) denotes $p \geq 0.0001$; unmarked values satisfy $p < 0.0001$

tree-based models are more effective for this task. This likely reflects the limited size of OpenCycloDB (n = 3459), which may be insufficient for training deep neural networks to their full potential. In contrast, gradient boosting methods like LGBM are well-suited to smaller data sets and can effectively capture complex relationships. The ECFP + representation shows a slight performance increase with the LGBM model compared to the ECFP representation. This indicates that the additional features, pH and temperature, provide information relevant to the binding affinity prediction task, but not enough to make a significant difference in performance. In addition, many experiments were performed at similar standard conditions (pH 7 and 25 °C), which could limit the influence of these features and their ability to generalize to a wider range of conditions.

In the FNN model, performance drops markedly between the ECFP and ECFP + representations on the test set, suggesting that the additional features may introduce noise the FNN cannot handle as effectively as LGBM. A similar decline is observed for the OpenCycloDB Enriched embedding. These patterns highlight the greater sensitivity of FNNs to limited data, where small validation and test splits (n = 346) may differ in distribution from the training set (n = 2767). In contrast, LGBM exhibits more stable performance across data splits.

Overall, the representations that use pretrained embedding models (UniMol2, GROVER, Chemeleon, and ChemBERTa) show a wide range of performances in the prediction task. The GROVER embeddings perform well with the LGBM predictive model but poorly with the FNN predictive model, indicating the effect of both embedding and predictive model choices on performance. Still, this result motivates further exploration of transfer learning for this limited data domain. The pretrained ChemBERTa FNN model performs at an average level relative to the other models, but finetuning yields a substantial performance gain. All metrics improve, indicating that finetuning adapts the ChemBERTa embeddings to better encode information relevant for host−guest binding affinity prediction. This result underscores the value of finetuning pretrained models for domain-specific tasks, consistent with applications in areas such as brain tumor classification in MR images[78,79] and interatomic potential prediction in materials science.[80,81] However, it is possible that finetuning does not have the same impact on all pretrained models. While these models achieve reasonable accuracy on OpenCycloDB, they require further improvement for practical applications. This baseline performance enables meaningful comparison with external test sets relevant to CDP design for PFAS removal.

External Test Set Performance

We then evaluated the trained models on external test sets to measure their generalization to the domain of interest, CD-PFAS binding. Table 3 presents the performance of different model architectures and molecular representations on external test sets. In E-PFAS$_{test}$, models perform poorly, with only a few model-embedding combinations achieving performance metrics that indicate any generalization. Overall, LGBM models perform better on the test sets. The best-performing model is the Mordred embeddings with the LGBM model, which achieves an R^2 score of 0.44 and high correlation metrics. While the error metrics show modest performance, the correlation metrics are promising, indicating that even when the models fail to predict accurately, they can still capture order relations between the predicted and actual values. The remaining models exhibit even lower performance, with some displaying negative R^2 scores, which can be interpreted as worse performance than simply predicting the data set mean. This poor performance can be attributed to several factors. First, E-PFAS$_{test}$ contains PFAS compounds that are structurally distinct from those in the training set, leading to a significant domain shift that challenges the model's ability to generalize. Additionally, the limited size of the test set (n = 21) covers a narrow representation of the chemical space, further hindering the model's predictive capabilities. In this new domain, the Mordred descriptors perform better than the other embeddings. This could be due to the inability of the pretrained embedding models to effectively learn the relevant chemical features for PFAS compounds, whereas the physicochemical information in the Mordred descriptors provides better generalization to this new chemical space. This result highlights the importance of chemical knowledge in

https://doi.org/10.1021/acs.jcim.5c03121
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

low-data settings. When examining the ChemBERTa models on the E-PFAS$_{test}$ set, finetuning increases the performance slightly, indicating that it has helped the model adapt to the specific task of predicting CD/PFAS binding affinities, even though the overall performance remains limited.

Similar trends occur in E-β-CD$_{test}$, where no model shows significant generalization and nearly all models achieve negative R^2 scores. Only UniMol2 with an LGBM predictor achieves a positive R^2 score of 0.02. These results indicate that no model effectively captures the effect of these functional groups in combination with the PFAS guests on binding affinity. Notably, the Mordred descriptors achieve moderate correlation metrics using both the LGBM and FNN predictors on the external test sets. Unlike the other test sets, both the hosts and guests in this data set are absent from OCDB$_{train}$, which likely contributes to the poor performance. The hosts in this data set are CDs with functional groups much different from those present in the training set. In this test set, the ChemBERTa Finetuned embeddings perform worse than the base ChemBERTa embeddings, indicating that finetuning has made the embeddings more specific to the training data, which hinders their ability to generalize to unseen data, highlighting the need for not only sufficient data sources for training, but also large, diverse data sets to improve finetuning effectiveness. These results indicate that while the models can perform reasonably well on the OpenCycloDB data set, they struggle to generalize to unseen chemical space, particularly when both hosts and guests are outside the training domain. This highlights the challenges of applying machine learning models to predict CD-PFAS binding affinities, especially given the limited availability of experimental data in this area.

LOOCV Models

LOOCV was performed using only the LGBM predictor to evaluate model performance under extreme data scarcity conditions that are representative of emerging contaminant domains where limited experimental data exists. By training on the combined external test data sets (n = 63), this approach simulates the realistic scenario where only limited CD-PFAS binding measurements are available for LGBM model development. The results in Table 4 show that Mordred achieves the best LOOCV performance with an R^2 score of 0.47, followed by UniMol2 (0.45). The ChemBERTa Finetuned model shows an increase in performance compared to the base ChemBERTa model, indicating the effectiveness of finetuning with a more general data set in improving model performance under data scarcity conditions. These results demonstrate that when training data is limited to the target domain of CD-PFAS interactions, the models can achieve modest cross-validation performance, suggesting that the embeddings capture some relevant chemical features even with minimal training data. However, these predictive results are not enough to justify using these models directly in practical applications. These LOOCV results are not directly comparable to the zero-shot results of the previous predictive models, but they provide insight into the models' predictive accuracy within the limited training domain. However, these models are expected to perform poorly on unseen data, even if relatively similar to the training data, due to the small size of the training set.

Comparing the LOOCV results to the previous predictive models highlights the challenges of domain shift and data scarcity. The LOOCV models are trained directly on the target domain, allowing them to learn specific features relevant to CD-PFAS binding, while the previous models are trained on a larger, but more general, data set that does not adequately represent the target domain. This comparison underscores the importance of having training data that closely aligns with the intended application domain, especially in scenarios where data is limited. These findings suggest the need for more targeted data collection efforts to expand the availability of cyclodextrin-PFAS binding measurements, which would enable the development of more robust and generalizable predictive models.

Table 4. Leave-One-Out Cross-Validation (LOOCV) Performance Comparison of Selected Molecular Representations on Combined External Test Datasets (n = 63)[a]

Model	Representation	RMSE	MAE	R^2	ρ	τ
LGBM	ECFP	4.38	3.37	0.36	0.66	0.46
	Mordred	**4.02**	**3.08**	**0.47**	**0.71**	**0.51**
	UniMol2	4.07	3.16	0.45	0.67	0.49
	GROVER	4.51	3.52	0.33	0.63	0.43
	Chemeleon	4.40	3.47	0.36	0.64	0.46
	ChemBERTa	4.42	3.36	0.35	0.65	0.45
	ChemBERTa Finetuned	4.17	3.23	0.42	0.67	0.48

[a]Models were trained using LGBM predictor with LOOCV to evaluate performance under extreme data scarcity conditions. RMSE and MAE are reported in kJ/mol. Bold values Indicate the best performance. Results improve upon predictive models trained on OCDB$_{train}$ (n = 2767) when evaluated on the same external test sets, but are limited in practical applicability due to small training size. For correlation metrics (ρ and τ), an asterisk (*) denotes $p \geq 0.0001$; unmarked values satisfy $p < 0.0001$.

CONCLUSION

This study systematically evaluated various molecular representations with widely different characteristics, including molecular fingerprints, learned graph methods, and pretrained transformer and geometric models, to predict CD-PFAS binding affinities using machine learning. The analysis of molecular embeddings through AlignedUMAP visualizations and nearest neighbor analysis showed that while all embeddings were capable of capturing relevant chemical features, they exhibited different clustering behaviors and sensitivities to molecular variations. After generating these embeddings, two predictive models, LGBM and a feedforward neural network, were trained on the OpenCycloDB data set using different combinations of host and guest embeddings. The best-performing model, a LightGBM model using GROVER embeddings, achieved an R^2 of 0.74 and RMSE of 2.42 kJ/mol. However, this model and others showed limited generalizability to external test sets of CD-PFAS complexes. With FNN models showing more sensitivity to the choice of molecular representation in the low-data setting than LGBM models, the performance results on the external test data sets motivate the development of more targeted data collection efforts to expand the domain availability for CD-PFAS complexes. Leave-one-out cross-validation experiments further support this, demonstrating that models trained directly on CD-PFAS data could achieve better performance on this specific domain at the cost of generalizability to unseen host and guest chemistries.

https://doi.org/10.1021/acs.jcim.5c03121
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

One main limitation of this work is the lack of rigorous uncertainty quantification in the model predictions. While the models provide point estimates of binding affinities, they do not capture the uncertainty associated with these predictions, which is crucial for guiding experimental validation and decision-making. Future work could incorporate techniques for uncertainty quantification, utilizing statistical tests to assess the reliability of performance comparisons. Additionally, the interpretability of the models is limited, making it difficult to understand which molecular features are driving the predictions. Developing interpretable models or using posthoc interpretability techniques could help explain the key chemical characteristics that influence CD-PFAS binding, providing insights for rational design of cyclodextrin-based polymers.

There are several avenues for future work to improve the predictive performance and generalizability of models for CD-PFAS binding. First, expanding the data set of experimentally measured binding affinities for CD-PFAS complexes would provide more training data to capture the relevant chemical space. These experiments can be expensive, motivating the development of molecular dynamics simulations capable of generating high-quality binding affinity data at a lower cost. Second, exploring more advanced predictive model architectures, such as graph neural networks that can directly model the interactions between host and guest molecules, may better capture the complex nature of host−guest binding. Full training of a local GNN may capture relevant chemical information better than the pretrained graph-based techniques explored in this work. We show these embeddings are capable of capturing individual molecular characteristics, but modeling the entire interaction space for CD-PFAS complexes remains untested. This work also explores finetuning for the ChemBERTa representation, but others could likely improve with finetuning as well. Our analysis uses the raw embeddings of each representation, though dimensionality reduction techniques, such as PCA, could improve the performance of models with larger latent embeddings. Lastly, investigating transfer learning approaches that leverage knowledge from related domains, such as drug discovery or materials science, in combination with finetuning strategies could help overcome data scarcity in this specific application. The pretrained models incorporated in this work were competitive with traditional embedding methods, and finetuning these models was shown to improve their performance.

These findings point toward a data-driven pipeline for modeling CDP-PFAS interactions, where targeted data collection, advanced model architectures, and transfer learning can be combined to develop robust predictive models. Current or improved embedding backbones can be pretrained on public polymer data or CDP/PFAS-specific MD simulation data. This architecture could learn host−guest interactions directly via cross-attention or pair-encoder methods, rather than concatenating separate embeddings. We could then use semisupervised learning to generate pseudolabels on unlabeled CD-PFAS pairs, retraining and iterating to improve the model. Physics-based auxiliary targets, such as 3D conformers or docking scores, could also help to regularize models toward realistic interactions. In addition, active learning strategies can guide experiments or simulations where uncertainty is highest or domain gaps are largest. This pipeline should improve generalization while focusing experiments on the most informative areas. Using these strategies, we can develop predictive models that can effectively guide the design of cyclodextrin-based polymers for PFAS remediation, ultimately contributing to more effective and sustainable water treatment solutions.

■ ASSOCIATED CONTENT

Data Availability Statement

All data sets used in this study are publicly available from the related sources listed in the Materials and Methods section and the Supporting Information. The code and all data needed to reproduce the results are available at the following links: https://github.com/colebrzakala/CDPFAS_HG_Analysis (code) and https://doi.org/10.4121/a5051137-a93d-433e-9cfe-50980247930c (data, models, and reports). Further information can be requested from the corresponding author.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.5c03121.

Supporting Information for Publication.pdf: external test data compositions, hyperparameter tuning configurations, and UniMol2 embedding distribution (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Cole Brzakala − *Water Management Department, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Stevinweg 1 2628CN Delft, Netherlands;* orcid.org/0009-0007-9970-8111; Email: C.Brzakala@tudelft.nl

Authors

Othonas A. Moultos − *Engineering Thermodynamics, Process and Energy Department, Faculty of Mechanical Engineering, Delft University of Technology, 2628CB Delft, Netherlands;* orcid.org/0000-0001-7477-9684

Jan Peter van der Hoek − *Water Management Department, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Stevinweg 1 2628CN Delft, Netherlands; Waternet, Research and Innovation, 1096AC Amsterdam, Netherlands;* orcid.org/0000-0002-0674-388X

Riccardo Taormina − *Water Management Department, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Stevinweg 1 2628CN Delft, Netherlands*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.5c03121

Author Contributions

C.B., R.T., and O.M. conceptualized the project. C.B. and R.T. designed the study. C.B. collected the data, developed the models, and performed the training and analysis. C.B. wrote the manuscript, with input and revisions from R.T. and O.M., who also provided supervision throughout the project. R.T., O.M., and J.P.H. acquired funding for the project.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This publication is part of the SYROP project with file number 20023 of the Open Technology Programme which is (partly) financed by the Dutch Research Council (NWO) with correspondence number 2023/TTW/01457390. The authors

also acknowledge support from the SYROP consortium partners, including Cyclopure, Waternet, and Witteveen + Bos.

■ ABBREVIATIONS AND NOMENCLATURE

Datasets

OCDB	OpenCycloDB
$OCDB_{train}$	OpenCycloDB training set
$OCDB_{val}$	OpenCycloDB validation set
$OCDB_{test}$	OpenCycloDB test set
$E\text{-}PFAS_{test}$	External PFAS test set
$E\text{-}\beta\text{-}CD_{test}$	External β-cyclodextrin test set

Molecular Representations

ECFP	Extended-Connectivity Fingerprints
ECFP+	ECFP with temperature and pH features
Mordred	Mordred physicochemical descriptors
GROVER	Graph Representation from self-supervised MP transformer
UniMol2	Universal Molecular Representation Model (version 2)
Chemeleon	Chemeleon pretrained graph representation
ChemBERTa	Chemistry-specific BERT architecture
ChemBERTa FT	ChemBERTa Finetuned
SMILES	Simplified Molecular Input Line Entry System
UMAP	Uniform Manifold Approximation and Projection

Predictive Model Training

LGBM	Light Gradient-Boosting Machine
FNN	Feedforward Neural Network
LOOCV	Leave-One-Out Cross-Validation

Chemical Names

PFAS	Per- and polyfluoroalkyl substances
CD	Cyclodextrin
CDP	Cyclodextrin-based polymer
α-CD	Alpha-cyclodextrin
β-CD	Beta-cyclodextrin
γ-CD	Gamma-cyclodextrin
DOM	Dissolved organic matter

■ REFERENCES

(1) Trudel, D.; Horowitz, L.; Wormuth, M.; Scheringer, M.; Cousins, I. T.; Hungerbühler, K. Estimating Consumer Exposure to PFOS and PFOA. *Risk Anal.* **2008**, *28*, 251−269.

(2) Brunn, H.; Arnold, G.; Körner, W.; Rippen, G.; Steinhäuser, K. G.; Valentin, I. PFAS: Forever Chemicals−Persistent, Bioaccumulative and Mobile. Reviewing the Status and the Need for Their Phase out and Remediation of Contaminated Sites. *Environ. Sci. Eur.* **2023**, *35*, 20.

(3) Evich, M. G.; Davis, M. J. B.; McCord, J. P.; Acrey, B.; Awkerman, J. A.; Knappe, D. R. U.; Lindstrom, A. B.; Speth, T. F.; Tebes-Stevens, C.; Strynar, M. J.; Wang, Z.; Weber, E. J.; Henderson, W. M.; Washington, J. W. Per- and polyfluoroalkyl substances in the environment. *Science* **2022**, *375*, No. eabg9065.

(4) CompTox Chemicals Dashboard. 2025; https://comptox.epa.gov/dashboard/chemical-lists/PFASMASTER, Website; (accessed Nov 09, 2025).

(5) Buck, R. C.; Korzeniowski, S. H.; Laganis, E.; Adamsky, F. Identification and Classification of Commercially Relevant Per- and Poly-fluoroalkyl Substances (PFAS). *Integr. Environ. Assess. Manage.* **2021**, *17*, 1045−1055.

(6) Fenton, S. E.; Ducatman, A.; Boobis, A.; DeWitt, J. C.; Lau, C.; Ng, C.; Smith, J. S.; Roberts, S. M. Per- and Polyfluoroalkyl Substance Toxicity and Human Health Review: Current State of Knowledge and Strategies for Informing Future Research. *Environ. Toxicol. Chem.* **2020**, *40*, 606−630.

(7) Stahl, T.; Mattern, D.; Brunn, H. Toxicology of Perfluorinated Compounds. *Environ. Sci. Eur.* **2011**, *23*, 38.

(8) Sunderland, E. M.; Hu, X. C.; Dassuncao, C.; Tokranov, A. K.; Wagner, C. C.; Allen, J. G. A Review of the Pathways of Human Exposure to Poly- and Perfluoroalkyl Substances (PFASs) and Present Understanding of Health Effects. *J. Exposure Sci. Environ. Epidemiol.* **2019**, *29*, 131−147.

(9) Ding, N.; Harlow, S. D.; Randolph, Jr. J. F.; Loch-Caruso, R.; Park, S. K. Perfluoroalkyl and Polyfluoroalkyl Substances (PFAS) and Their Effects on the Ovary. *Hum. Reprod. Update* **2020**, *26*, 724−752.

(10) EFSA Panel on Contaminants in the Food Chain EFSA CONTAM Panel; Schrenk, D.; Bignami, M.; Bodin, L.; Chipman, J. K.; del Mazo, J.; Grasl-Kraupp, B.; Hogstrand, C.; Hoogenboom, L. R.; Leblanc, J.; et al Risk to human health related to the presence of perfluoroalkyl substances in food. *EFSA J.* **2020**, *18*..

(11) Glüge, J.; Scheringer, M.; Cousins, I. T.; DeWitt, J. C.; Goldenman, G.; Herzke, D.; Lohmann, R.; Ng, C. A.; Trier, X.; Wang, Z. An Overview of the Uses of Per- and Polyfluoroalkyl Substances (PFAS). *Environ. Sci.:Processes Impacts* **2020**, *22*, 2345−2373.

(12) PFAS Pollution in European Waters. 2024; https://www.eea.europa.eu/en/analysis/publications/pfas-pollution-in-european-waters, Website; (accessed Dec 04, 2025).

(13) PFAS National Primary Drinking Water Regulation. 2024; https://www.federalregister.gov/documents/2024/04/26/2024-07773/pfas-national-primary-drinking-water-regulation, Website; (accessed Dec 04, 2025).

(14) Ling, Y.; Klemes, M. J.; Xiao, L.; Alsbaiee, A.; Dichtel, W. R.; Helbling, D. E. Benchmarking Micropollutant Removal by Activated Carbon and Porous β -Cyclodextrin Polymers under Environmentally Relevant Scenarios. *Environ. Sci. Technol.* **2017**, *51*, 7590−7598.

(15) Quinlivan, P. A.; Li, L.; Knappe, D. R. U. Effects of Activated Carbon Characteristics on the Simultaneous Adsorption of Aqueous Organic Micropollutants and Natural Organic Matter. *Water Res.* **2005**, *39*, 1663−1673.

(16) Ebrahimzadeh, S.; Wols, B.; Azzellino, A.; Martijn, B. J.; van der Hoek, J. P. Quantification and modelling of organic micropollutant removal by reverse osmosis (RO) drinking water treatment. *J. Water Proc. Eng.* **2021**, *42*, 102164.

(17) Amato, M. E.; Lipkowitz, K. B.; Lombardo, G. M.; Pappalardo, G. C. High-Field NMR Spectroscopic Techniques Combined with Molecular Dynamics Simulations for the Study of the Inclusion Complexes of α- and β-Cyclodextrins with the Cognition Activator 3-Phenoxypyridine Sulphate (CI-844). *Magn. Reson. Chem.* **1998**, *36*, 693−705.

(18) Tang, Z.; Chang, C.-e. A. Binding Thermodynamics and Kinetics Calculations Using Chemical Host and Guest: A Comprehensive Picture of Molecular Recognition. *J. Chem. Theory Comput.* **2018**, *14*, 303−318.

(19) Wang, R.; Lin, Z.-W.; Klemes, M. J.; Ateia, M.; Trang, B.; Wang, J.; Ching, C.; Helbling, D. E.; Dichtel, W. R. A Tunable Porous β -Cyclodextrin Polymer Platform to Understand and Improve Anionic PFAS Removal. *ACS Cent. Sci.* **2022**, *8*, 663−669.

(20) Ching, C.; Klemes, M. J.; Trang, B.; Dichtel, W. R.; Helbling, D. E. β -Cyclodextrin Polymers with Different Cross-Linkers and Ion-Exchange Resins Exhibit Variable Adsorption of Anionic, Zwitterionic, and Nonionic PFASs. *Environ. Sci. Technol.* **2020**, *54*, 12693−12702.

(21) Alsbaiee, A.; Smith, B. J.; Xiao, L.; Ling, Y.; Helbling, D. E.; Dichtel, W. R. Rapid Removal of Organic Micropollutants from Water by a Porous β -Cyclodextrin Polymer. *Nature* **2016**, *529*, 190−194.

(22) Klemes, M. J.; Ling, Y.; Ching, C.; Wu, C.; Xiao, L.; Helbling, D. E.; Dichtel, W. R. Reduction of a Tetrafluoroterephthalonitrile-β -Cyclodextrin Polymer to Remove Anionic Micropollutants and Perfluorinated Alkyl Substances from Water. *Angew. Chem., Int. Ed.* **2019**, *58*, 12049−12053.

(23) Xiao, L.; Ling, Y.; Alsbaiee, A.; Li, C.; Helbling, D. E.; Dichtel, W. R. β-Cyclodextrin Polymer Network Sequesters Perfluorooctanoic

https://doi.org/10.1021/acs.jcim.5c03121
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Acid at Environmentally Relevant Concentrations. *J. Am. Chem. Soc.* **2017**, *139*, 7689−7692.
(24) Lacalamita, D.; Mongioví, C.; Crini, G.; Morin-Crini, N. Cyclodextrins for the Removal of Per- and Polyfluoroalkyl Substances: A Review. *Environ. Chem. Lett.* **2025**, *23*, 1713−1743.
(25) Musuc, A. M. C. Advances in Chemistry, Toxicology, and Multifaceted Applications. *Molecules* **2024**, *29*, 5319.
(26) Fischer, A.; van Wezel, A. P.; Hollender, J.; Cornelissen, E.; Hofman, R.; van der Hoek, J. P. Development and Application of Relevance and Reliability Criteria for Water Treatment Removal Efficiencies of Chemicals of Emerging Concern. *Water Res.* **2019**, *161*, 274−287.
(27) Mansouri, K.; Cariello, N. F.; Korotcov, A.; Tkachenko, V.; Grulke, C. M.; Sprankle, C. S.; Allen, D.; Casey, W. M.; Kleinstreuer, N. C.; Williams, A. J. Open-Source QSAR Models for pKa Prediction Using Multiple Machine Learning Approaches. *J.—Cheminf.* **2019**, *11*, 60.
(28) Nguyen, D.; Tao, L.; Li, Y. Integration of Machine Learning and Coarse-Grained Molecular Simulations for Polymer Materials: Physical Understandings and Molecular Design. *Frontiers in Chemistry* **2022**, *9*, 820417.
(29) *Advances in QSAR Modeling; Challenges and Advances in Computational Chemistry and Physics*, Roy, K., Ed.; Springer International Publishing: Cham, 2017; Vol. *24*.
(30) Schweidtmann, A. M.; Rittig, J. G.; König, A.; Grohe, M.; Mitsos, A.; Dahmen, M. Graph Neural Networks for Prediction of Fuel Ignition Quality. *Energy Fuels* **2020**, *34*, 11395−11407.
(31) Jiang, S.; Webb, M. A. Physics-Guided Neural Networks for Transferable Property Prediction in Architecturally Diverse Copolymers. *Macromolecules* **2025**, *58*, 4971−4984.
(32) Ramírez-Palacios, C.; Marrink, S. J. Super High-Throughput Screening of Enzyme Variants by Spectral Graph Convolutional Neural Networks. *J. Chem. Theory Comput.* **2023**, *19*, 4668−4677.
(33) Liu, S.; Guo, H.; Tang, J. Molecular Geometry Pretraining with SE(3)-Invariant Denoising Distance Matching. In *The Eleventh International Conference on Learning Representations*, 2023.
(34) Ma, Y.; Niu, Y.; Yang, H.; Dai, J.; Lin, J.; Wang, H.; Wu, S.; Yin, Q.; Zhou, L.; Gong, J. Prediction and Design of Cyclodextrin Inclusion Complexes Formation via Machine Learning-Based Strategies. *Chem. Eng. Sci.* **2022**, *261*, 117946.
(35) Jeschke, S.; Cole, I. S. 3D-QSAR for Binding Constants of β -Cyclodextrin Host-Guest Complexes by Utilising Spectrophores as Molecular Descriptors. *Chemosphere* **2019**, *225*, 135−138.
(36) Di, P.; Chen, J.; Liu, L.; Li, W.; Tang, Y.; Liu, G. *In Silico* Prediction of Binding Capacity and Interaction Forces of Organic Compounds with α - and β -Cyclodextrins. *J. Mol. Liq.* **2020**, *302*, 112585.
(37) Cai, S.; Chen, D.; Cai, J.; Tan, A.; Zhou, J.; Zhuo, M.; Liu, M.; Zhu, C.; Li, S. Machine Learning-Guided Selection of Cyclodextrins for Enhanced Biosynthesis and Capture of Volatile Terpenes. *J. Agric. Food Chem.* **2025**, *73*, 3602−3610.
(38) Zhao, Q.; Ye, Z.; Su, Y.; Ouyang, D. Predicting Complexation Performance between Cyclodextrins and Guest Molecules by Integrated Machine Learning and Molecular Modeling Techniques. *Acta Pharm. Sin. B* **2019**, *9*, 1241−1252.
(39) Ling, Y.; Klemes, M. J.; Steinschneider, S.; Dichtel, W. R.; Helbling, D. E. QSARs to Predict Adsorption Affinity of Organic Micropollutants for Activated Carbon and β -Cyclodextrin Polymer Adsorbents. *Water Res.* **2019**, *154*, 217−226.
(40) Zhang, M.; Ma, T.; Ozlek, M.; Yazaydin, A. O. Machine Learning-Assisted Exploration of Covalent Organic Frameworks for Short-Chain per- and Polyfluoroalkyl Substances (PFAS) Removal from Water. *J. Colloid Interface Sci.* **2026**, *702*, 138970.
(41) Baptista, D.; Correia, J.; Pereira, B.; Rocha, M. Evaluating Molecular Representations in Machine Learning Models for Drug Response Prediction and Interpretability. *Journal of Integrative Bioinformatics* **2022**, *19*, 20220006.
(42) Buterez, D.; Janet, J. P.; Kiddle, S. J.; Oglic, D.; Lió, P. Transfer Learning with Graph Neural Networks for Improved Molecular Property Prediction in the Multi-Fidelity Setting. *Nat. Commun.* **2024**, *15*, 1517.
(43) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: a molecular descriptor calculator. *J.—Cheminf.* **2018**, *10*, 4.
(44) Burns, J. JacksonBurns/mordred-community. Software 2026. https://github.com/JacksonBurns/mordred-community.
(45) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(46) Ji, X.; Wang, Z.; Gao, Z.; Zheng, H.; Zhang, L.; Ke, G.; E, W. *Proceedings of the 38th International Conference on Neural Information Processing Systems*; Exploring molecular pretraining model at scale: Red Hook, NY, USA, 2024.
(47) Rong, Y.; Bian, Y.; Xu, T.; Xie, W.; Wei, Y.; Huang, W.; Huang, J. Self-Supervised Graph Transformer on Large-Scale Molecular Data. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 12559−12571.
(48) Burns, J. W.; Zalte, A. S.; Abreu, C. R. A.; Sieg, J.; Feldmann, C.; Mathea, M.; Green, W. H. Deep Learning Foundation Models from Classical Molecular Descriptors. *arXiv* **2025**.
(49) Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**. accessed April 24, 2026
(50) Tahıl, G.; Delorme, F.; Le Berre, D.; Monflier, E.; Sayede, A.; Tilloy, S. Curated Dataset of Association Constants between a Cyclodextrin and a Guest for Machine Learning. *Chem. Data Collect.* **2023**, *45*, 101022.
(51) Weiss-Errico, M. J.; Ghiviriga, I.; O'Shea, K. E. 19F NMR Characterization of the Encapsulation of Emerging Perfluoroether-carboxylic Acids by Cyclodextrins. *J. Phys. Chem. B* **2017**, *121*, 8359−8366.
(52) Restrepo-Osorio, R. A.; O'Shea, K. E. Control of host-guest complexation and release of PFAS with pH changes employing ionizable β -cyclodextrin derivatives. *J. Hazard. Mater. Adv.* **2025**, *20*, 100904.
(53) *RDKit: Open-source Chem—informatics*, Software; 2025.
(54) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(55) Zhang, B.; Li, X.; Xu, X.; Cao, J.; Zeng, M.; Zhang, W. Multi-property prediction and high-throughput screening of polyimides: An application case for interpretable machine learning. *Polymer* **2024**, *312*, 127603.
(56) Wójcikowski, M.; Kukiełka, M.; Stepniewska-Dziubinska, M. M.; Siedlecki, P. Development of a Protein−Ligand Extended Connectivity (PLEC) Fingerprint and Its Application for Binding Affinity Predictions. *Bioinformatics* **2019**, *35*, 1334−1341.
(57) Minami, T.; Okuno, Y. Number Density Descriptor on Extended-Connectivity Fingerprints Combined with Machine Learning Approaches for Predicting Polymer Properties. *MRS Adv.* **2018**, *3*, 2975−2980.
(58) Sánchez-Cruz, N.; Medina-Franco, J. L.; Mestres, J.; Barril, X. Extended Connectivity Interaction Features: Improving Binding Affinity Prediction through Chemical Description. *Bioinformatics* **2021**, *37*, 1376−1382.
(59) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107−113.
(60) Qiu, J.-H.; Lin, Z.; Chen, K.-W.; Sun, T.-Y.; Zhang, X.; Yuan, L.; Tian, Y.; Wu, Y.-D. SAKPE: A Site Attention Kinetic Parameters Prediction Method for Enzyme Engineering. *bioRxiv: the preprint server for biology 2025, Preprint; repository: bioRxiv*, **2026**; submitted 2025−04−30.
(61) Zhou, J.; Cui, G.; Hu, S.; Zhang, Z.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; Sun, M. Graph Neural Networks: A Review of Methods and Applications. *AI Open* **2020**, *1*, 57−81.
(62) Dou, B.; Zhu, Z.; Merkurjev, E.; Ke, L.; Chen, L.; Jiang, J.; Zhu, Y.; Liu, J.; Zhang, B.; Wei, G.-W. Machine Learning Methods for Small Data Challenges in Molecular Science. *Chem. Rev.* **2023**, *123*, 8736−8780.

https://doi.org/10.1021/acs.jcim.5c03121
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

(63) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268−276.
(64) Bjerrum, E. J.; Sattarov, B. Improving Chemical Autoencoder Latent Space and Molecular De Novo Generation Diversity with Heteroencoders. *Biomolecules* **2018**, *8*, 131.
(65) McInnes, L.; Healy, J.; Saul, N.; Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **2018**, *3*, 861.
(66) Bero, S. A.; Muda, A. K.; Choo, Y. H.; Muda, N. A.; Pratama, S. F. Similarity Measure for Molecular Structure: A Brief Review. *J. Phys., Conf. Ser.* **2017**, *892*, 012015.
(67) Bajusz, D.; Rácz, A.; Héberger, K. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *J.—Cheminf.* **2015**, *7*, 20.
(68) Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems* **2017**.
(69) Zhang, J.; Mucs, D.; Norinder, U.; Svensson, F. LightGBM: An Effective and Scalable Algorithm for Prediction of Chemical Toxicity−Application to the Tox21 and Mutagenicity Data Sets. *J. Chem. Inf. Model.* **2019**, *59*, 4150−4158.
(70) Rosenblatt, F. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review* **1958**, *65*, 386−408.
(71) Agatonovic-Kustrin, S.; Beresford, R. Basic Concepts of Artificial Neural Network (ANN) Modeling and Its Application in Pharmaceutical Research. *J. Pharm. Biomed. Anal.* **2000**, *22*, 717−727.
(72) Biewald, L. Experiment Tracking with Weights and Biases. Software 2020. https://www.wandb.com/.
(73) Chai, T.; Draxler, R. R. Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? − Arguments against Avoiding RMSE in the Literature. *Geosci. Model Dev.* **2014**, *7*, 1247−1250.
(74) Liu, Y.; He, X.; Mo, Y. Discrepancies and Error Evaluation Metrics for Machine Learning Interatomic Potentials. *npj Comput. Mater.* **2023**, *9*, 174.
(75) Spearman, C. The proof and measurement of association between two things. *Am. J. Psychol.* **1904**, *15*, 72−101.
(76) Kendall, M. G. A New Measure of Rank Correlation. *Biometrika* **1938**, *30*, 81−93.
(77) Airola, A.; Pahikkala, T.; Waegeman, W.; Baets, B. D.; Salakoski, T. A Comparison of AUC Estimators in Small-Sample Studies. In *Proceedings of the Third International Workshop on Machine Learning in Systems Biology*, 2009; pp 3−13.
(78) Swati, Z. N. K.; Zhao, Q.; Kabir, M.; Ali, F.; Ali, Z.; Ahmed, S.; Lu, J. Brain Tumor Classification for MR Images Using Transfer Learning and Fine-Tuning. *Computerized Medical Imaging and Graphics* **2019**, *75*, 34−46.
(79) Valverde, J. M.; Imani, V.; Abdollahzadeh, A.; De Feo, R.; Prakash, M.; Ciszek, R.; Tohka, J. Transfer Learning in Magnetic Resonance Brain Imaging: A Systematic Review. *Journal of Imaging* **2021**, *7*, 66.
(80) Zaverkin, V.; Holzmüller, D.; Bonfirraro, L.; Kästner, J. Transfer Learning for Chemically Accurate Interatomic Neural Network Potentials. *Phys. Chem. Chem. Phys.* **2023**, *25*, 5383−5396.
(81) Radova, M.; Stark, W. G.; Allen, C. S.; Maurer, R. J.; Bartók, A. P. Fine-Tuning Foundation Models of Materials Interatomic Potentials with Frozen Transfer Learning. *npj Comput. Mater.* **2025**, *11*, 237.

https://doi.org/10.1021/acs.jcim.5c03121
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX