Artificial Intelligence in the Life Sciences 2 (2022) 100047

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

journal homepage: www.elsevier.com/locate/ailsci

Research Article

Modeling bioconcentration factors in fish with explainable deep learning

Linlin Zhao[a], Floriane Montanari[b], Henry Heberle[a], Sebastian Schmidt[c,*]

[a] *Bayer AG, Crop Science Division, Field Solutions, Monheim 40789, Germany*
[b] *Bayer AG, Pharmaceuticals Division, Machine Learning Research, Berlin 13353, Germany*
[c] *Bayer AG, Crop Science Division, Regulatory Science, Monheim 40789, Germany*

ARTICLE INFO

Keywords:
Bioaccumulation
Bioconcentration factor
BCF
QSAR
Multitask learning
Explainable AI
Environmental safety

ABSTRACT

The Bioconcentration Factor (BCF) is an important parameter in the environmental risk assessment of chemicals, relevant for industrial and academic research as well as required in many regulatory contexts. It represents the potential of a substance to accumulate in organic tissues or whole animals and is most frequently measured in fish. However, animal welfare reasons, throughput limitations, and costs push the need for alternative methods that allow accurate and reliable estimations of BCF in silico. We present a new deep learning model to predict BCF values from chemical structures, that outperforms currently available models (R^2 of 0.68 and RMSE of 0.59 log units on an external test set; R^2 of 0.70 and RMSE of 0.74 log units in a demanding cluster split validation). The model is based on molecular representations encoded as CDDD descriptors and exploits a large in-house dataset with measured logD values as an auxiliary task.

Additionally, we developed a post-hoc explainability method based on SMILES character substitutions to accompany our predictions with atom-level interpretations. These sensitivity scores highlight the most influential moieties in the molecule and can help to understand the predictions better and design new molecules.

1. Introduction

The bioconcentration factor (BCF) represents the tendency of a chemical to accumulate in organic tissues or living animals - typically fish. It is defined as the ratio between the concentration in the organism and the concentration in the surrounding water phase at steady-state conditions in the laboratory [1]. Oftentimes, this ratio is normalized to a fixed lipid content (5%) of the test organism. Common test species are fathead minnow, bluegill sunfish, or rainbow trout.

BCF values are required for environmental hazard and risk assessments around the globe and used as optimization criterion to select or design new chemicals with more benign environmental profiles. However, for a BCF study according to OECD Guideline 305 [1] and fulfilling global regulatory requirements, more than 200 test animals are needed per test molecule. Even screening-level studies still require around 30 test animals to be reliable. Obviously, there is a strong desire to minimize vertebrate testing, and many stakeholders are joining forces. A prominent example is the 3R strategy followed by the European Commission [2] and US Environmental Protection Agency [3] to reduce, refine, and replace animal tests. In silico models are a fundamental building block in these endeavors, but they need to be accurate, reliable, and ideally provide mechanistic insights and indicate their limitations or uncertainties.

Multiple quantitative structure-property/activity relationships (QSAR models) have already been developed for predicting BCF from molecular structures (see two recent comparative studies [4,5]). Many of these models rely on the water-octanol partition coefficient (Kow) and additional features of the molecular structures with varying complexity. One of the most well-known models is the BCF model by Meylan [6] as implemented in the US-EPA's EPISuite [7], which is among the best performers and applicable to a relatively broad range of chemicals [4,5]. It is a linear regression model based on logKow and a list of correction factors for several molecular substructures.

The Kow has been demonstrated to be a good substitute for the partitioning of the substance between water and lipids of the organisms [8], but this approach also has some limitations, especially for hydrogen-bond donors, polar, or ionic compounds, which have stronger interactions with membrane lipids compared to storage lipids or octanol [4,8,9]. Thus, additional descriptors are required to cover these interactions. Besides the water-lipid partitioning, other processes like kinetic restrictions of uptake or elimination, metabolism, and specific interactions like binding to proteins are important for a complete mechanistic understanding and better predictions of bioaccumulation [4].

Besides mechanistic aspects, the predictivity of BCF models is generally limited by the small amount of experimental studies, the varying reliability and heterogeneity of the data, and imbalances in coverage of the chemical space and BCF ranges. In addition, trade-offs often have

[*] Corresponding author.
E-mail address: sebastian.schmidt1@bayer.com (S. Schmidt).

https://doi.org/10.1016/j.ailsci.2022.100047
Received 28 September 2022; Received in revised form 2 November 2022; Accepted 16 November 2022
Available online 2 December 2022

to be made between predictive performance and interpretability. The former is often achieved with more complex models, while the latter is easier for less complex models. Evaluating the reliability of individual predictions of complex "black box" models is a challenge.

In recent years, Deep Learning models have been successfully employed for different molecular property predictions [10,11]. Common modeling strategies involve fully connected networks from pre-computed molecular fingerprints [12] or latent representations [13] and, when enough training data is available, end-to-end training with graph convolutional networks (GCNs) [14,15]. In low data regimes, as is often the case in toxicology, unsupervised pretraining on molecular representations like SMILES or InChI codes can provide powerful representations that can be fine-tuned on the available labeled data [16,17].

The complexity of Deep Learning models for QSAR has increased the need for equally powerful interpretability methods [18–20]. These are crucial to gain trust in the model, understand its limitations, and support the user's knowledge and intuition in the process of property optimization and molecular design. Explanations for QSAR models can take the form of atomic attributions [21,22] (particularly for GCNs [23,24]), feature importances [25] (using packages such as SHAP [26]), or substructure importances [27]. Counterfactuals [28] and activity cliff [29] approaches have also been proposed, which explain a particular prediction by selecting other examples with similar structure but different outcome.

In this work, we present a new model for predicting BCF, which addresses some of the difficulties mentioned above. In particular, we make use of recent advances in the fields of deep learning and explainability. The model is built on the CDDD descriptors [13], which are a learned representation of an autoencoder model for the translation of molecular structures in SMILES format. This provides a continuous representation of the chemical space and avoids the manual definition of molecular features or substructures. We use a multitask learning approach to predict simultaneously logBCF and logKow. This approach addresses the small size of available BCF datasets by exploiting large available logKow datasets as an auxiliary task. The choice of logKow as auxiliary task is based on data availability, known correlation with BCF, and previous work in the literature demonstrating BCF prediction improvement upon transfer learning of logKow [30].

To overcome the mentioned trade-off between predictivity and interpretability, we developed an explainability module on top of our deep learning model. It is based on SMILES string substitutions and is particularly suitable for models built on CDDD descriptors, regardless of the machine learning algorithm used. The output quantifies the influence of each character in the input SMILES on the predicted BCF or Kow values. We name the resulting character scores *sensitivity scores* in the rest of the work. We explore the agreement between the sensitivity scores for Kow predictions with basic chemistry knowledge and check the robustness of the scores for symmetric molecules. We compare the sensitivity scores summed up over entire functional groups with known correction factors for lipophilicity from the literature. We show that our model is able to recognize the molecular context, which is a significant advantage over classical group contribution methods that employ a fixed contribution for given functional groups in all molecules. Finally, by comparing explanations for the Kow and BCF endpoints, we can retrieve known mechanistic effects such as metabolism.

2. Methods

2.1. Data

Experimental BCF data for model training was taken from Grisoni et al. [4]. They compiled BCF values for 1056 compounds by merging three datasets, which had been used for the development of BCF QSAR models, and 45 compounds with special metabolism or known Kow-BCF relationships (e.g., pyrethroids, organophosphorus compounds, perfluorinated and polychlorinated compounds). In terms of chemical space, this set covers a broad range of chemical classes, including aliphatic and aromatic hydrocarbons, alcohols, amines, esters, amides, ethers, phenols, anilines, heteroaromatics, nitroaromatics, organochlorines, organophosphates, sulfonic acids, and thiols. In terms of use classes, this set contains industrial chemicals, agrochemicals, pharmaceuticals, plant secondary metabolites, and pollutants. All BCF values refer to the wet weight of the whole fish and were not normalized on lipid content, which is the most frequent form in scientific literature and the basis for calibration of most available QSAR models. The set contains two duplicate entries for glyphosate and 2,2-dichloropropionic acid. The logBCF values for the latter were very similar and we kept the average. However, the BCF values for glyphosate differed significantly. We therefore kept only the entry which agreed with the value given in the Pesticide Properties DataBase (PPDB [31]).

To build the external test set, we searched for BCF values in the PPDB [31] (as of 08-03-2022) that were neither part of our training dataset nor part of the Meylan model [6,7] training set. Only entries with an assigned quality level of 4 or 5 were kept, which means the endpoints have been reviewed and curated by the PPDB providers and were used for regulatory purposes in the majority of the cases. To maintain feasibility with our modelling approach, we excluded molecules containing metal atoms and stereoisomers with more than 0.8 log units difference in BCF between the isomers. Furthermore, we cross-checked the reliability of the data based on original study reports or public list of endpoints. This led to the correction of flufenoxuron, spirodiclofen, metaflumizone, camphechlor and removal of aldicarb, vinclozolin, 1-dodecanol, and acrylonitrile from the set. However, we could not check ca. one third of the molecules because the data origin was not referenced precisely. In total, the final curated test set contains 80 molecules.

Experimental data for water-octanol distribution coefficients (logD) of 63,127 compounds at neutral pH was taken from an in-house dataset, based on HPLC screening methods. In contrast to logP or logKow, which refer to the neutral molecule, logD refers to the total sum of the neutral molecule and all (de-)protonated species. Therefore, it is directly applicable also to molecules with acidic or basic properties. 36 molecules of the logD dataset overlap with the BCF dataset. Missing logD values for the remaining compounds from the BCF dataset were imputed with an in-house model for logD values at neutral pH (RMSE=0.3, for a detailed description of the methodology see Montanari et al. [32]).

2.2. Molecular structures and descriptors

All molecular structures were normalized with routine workflows that standardize charges, keep only the largest fragment (strip off salt counter ions), clear the stereo information, deprotonate bases, protonate acids, and canonicalize tautomers. Finally, canonical non-isomeric SMILES codes were generated as structure representation for each molecule.

All canonical SMILES strings were encoded by the CDDD encoder [13] to 512 dimensional descriptors without any further pre-processing by CDDD. Note that this circumvents the applicability domain restrictions imposed by the CDDD pipeline (logP -5 to 7; molecular weight 12 to 600 g/mol; 3 to 50 heavy atoms) and affected around 3% of the molecules from our datasets. For a discussion of the impact of this decision, see subsection Applicability Domain. Nine molecules had to be excluded from the training set at this stage, because they contain Sn, which is not part of the CDDD vocabulary.

Additionally, we generated ten randomized SMILES strings for each molecule using RDKit [33] and calculated the CDDD descriptors the same way as described above for the canonical SMILES.

2.3. Training, validation and external test sets

To avoid potential information leakage or biases towards subsets during model training, the compounds from the combined BCF and logD

Table 1
Cluster splits: number of compounds from each task in each cluster.

Cluster	0	1	2	3	4	5	6
#compounds (BCF task)	118	19	63	61	105	280	399
#compounds (logD task)	8687	5376	21,359	2710	10,662	8255	6078

datasets were clustered by similarity. For this, circular fingerprints of radius 6, folded to a 2048 vector length were calculated, and k-means clustering was applied with k=10. Smaller clusters were merged to ensure that enough compounds with a BCF label are present in each cluster. The resulting partition is shown in Table 1. The smallest cluster with 19 BCF compounds comprising polyfluorinated molecules with high weights (>400 g/mol) was considered an outlier and excluded from our cross-validation processes.

All our cross-validation routines were conducted on these cluster splits, for both single-task and multitask models. For training processes involving both hyperparameter optimizations and model selections, nested cross-validation [34] was employed to avoid optimistic biases when evaluating model performances. Nested cross-validation uses the data effectively to create a series of data splitting (training, validation and test). Simple leave-one-cluster-out cross-validation was used when only tuning hyperparameters for fixed model structures.

The external test set containing 80 molecules from PPDB (see Data section) was compiled and used only after the model training had been completed. Both BCF datasets are made available as supplementary information.

2.4. Modeling approaches

2.4.1. Single task models

Support vector regressor (SVR), random forest (RF), and XGBoost were trained on the BCF data using the CDDD embeddings as input features. Nested cluster cross-validation was employed to perform hyperparameter optimizations and model selection. Our cross-validated training results showed that the CDDD descriptors work best in combination with SVR, as previously reported [13]. The optimal SVR has radial basis function as kernel, where the kernel coefficient is 1/512, ϵ is 0.05, and the optimal regularization strength is 1.

In addition, a simple linear regression model (ordinary least squares with default settings from scikit-learn) with in-house logD values as the only independent variable, and a null model taking the mean logBCF value as prediction were built on the BCF dataset. These two models are very simple baselines to benchmark the improvement of more sophisticated approaches.

2.4.2. Multitask learning model

Because logD is known to be tightly related to BCF (Pearson's correlation coefficient of 0.752 in our data), and we have a large logD dataset internally available, we developed a multitask model combining both endpoints. This model is built upon CDDD embeddings and is referred to as *xBCF* in the remaining text. The model architecture was kept very simple, with two fully-connected hidden layers of dimensions 600 and 50 (see Fig. 1) predicting logD and logBCF values simultaneously. Rectified Linear Units (ReLU) are used in the dense layers as nonlinear transform functions.

A simple coefficient weighting strategy was used to balance the individual task losses in the loss function which can be formulated as

$$\mathcal{L}_{total} = \alpha\mathcal{L}_b + (1-\alpha)\mathcal{L}_d, \qquad (1)$$

where $\mathcal{L}_b = \sum(\hat{y}_b - y_b)^2/n_b$ and $\mathcal{L}_d$ are the mean squared errors (MSE) of logBCF and logD, respectively; $\alpha \in (0,1)$ is the weighting coefficient for the loss from logBCF to balance the learning between two tasks. Moreover, dropout regularization was applied to both 512-dimensional descriptors (optimal rate 0.2) and the first hidden layer (optimal rate 0.4). The BCF dataset was given higher weight by sampling with replacement to match the logD dataset size. The loss weighting coefficient α was optimized during hyperparameter search looking only at logBCF performance; the final optimal value was 0.01. The downstream network was optimized with leave-one-cluster-out cross validation by stochastic gradient descent with momentum of 0.9 and learning rate of 0.005. In addition, early stopping was employed to avoid overfitting, where the patience and epoch number are finetuned during optimization. All hyperparameter values specified in the text were obtained by optimizing the loss function on an extensive hyperparameter grid. The training scripts and all optimal hyperparameters can be found in the supporting information.

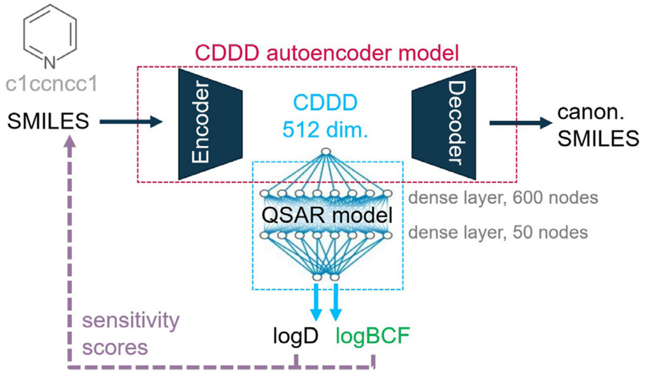

Fig. 1. Model structure of the xBCF multitask model based on CDDD descriptors.

2.4.3. Literature models

For comparison, we employed several publicly available and well-known BCF QSAR models. The Meylan [6] and Arnot-Gobas [35] models were run using the BCFBAF module from the EPI Suite [7] in batch mode. In this setup the Meylan model estimates the logKow internally with the KOWWIN module [36] for use in its regression-based approach with correction terms depending on the functional groups of the molecule. The Arnot-Gobas model is a physiology-based model that uses the same logKow estimation from KOWWIN and includes biotransformation estimates based on functional groups, logKow, and molecular weight. We used the parameterization for the upper trophic level. In addition, the CAESAR [37,38] model (version 2.1.15) and KNN-Read-Across model (version 1.1.1) from the VEGA in silico platform [39] were used. CAESAR is a consensus model combining two radial basis function neural networks based on 2D structural and physicochemical descriptors. The KNN-Read-Across model calculates a similarity index [40] and selects the three most similar molecules from the training set. The final BCF prediction is a weighted average of the selected molecules. Molecules for which VEGA reported low reliability were considered outside the applicability domain of the model. OPERA [41] (v2.8) is a QSAR model suite for physico-chemical and environmental fate properties. We used its weighted k-nearest-neighbour model for logBCF (LogBCFv2.6, QMRF:Q17-24a-0023) including the internal structure standardization pre-processing steps. Additional recent articles [42,43] report results but do not share their models and therefore could not be compared with our approaches.

2.5. Performance metrics

Coefficient of determination (R^2) and root mean squared error (RMSE) were used as goodness-of-fit metrics for evaluating model performances. In combination with Leave-One-Cluster-Out cross validation, R^2 and $RMSE$ were calculated on the entire BCF and logD datasets. For instance, when cluster i was chosen as validation set, the model trained on the rest of the clusters predicted compounds in cluster i. This proce-

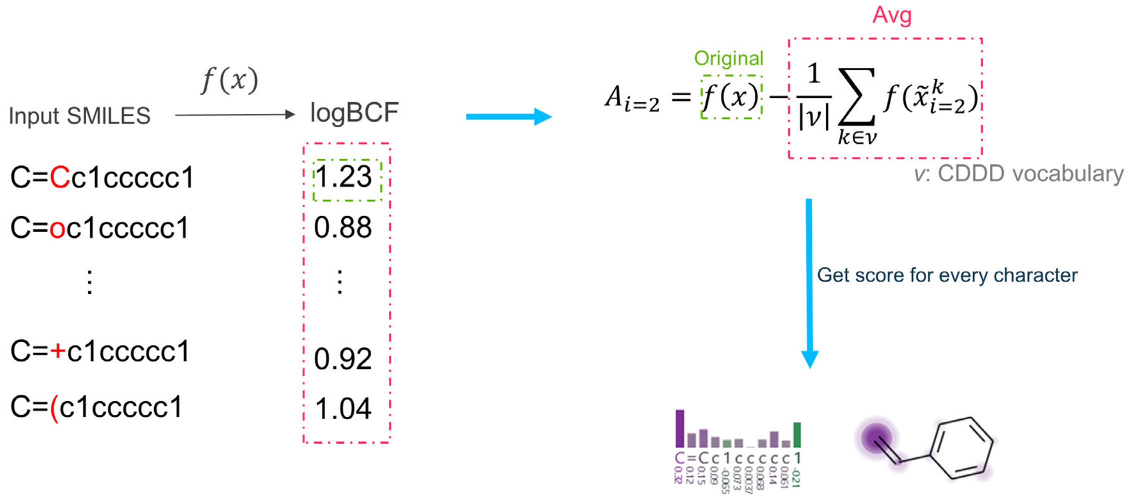

Fig. 2. Illustrating the character substitution method.

dure was repeated for each cluster until all compounds were predicted once, then the metrics were calculated on the predictions of all clusters.

2.6. *Explanation methods*

To explain our BCF model, we borrowed ideas from post-hoc perturbation methods, feature ablation and occlusion [44,45]. The aim is to establish a gradient-free method for assessing the importance of individual SMILES characters. In general, such methods replace the original input features with a baseline value to create a perturbed input. It is expected that the more important a feature is, the greater the change in prediction its replacement will lead to. To put it formally in the context of BCF modeling, let $f : \mathbb{R}^{v\times n} \rightarrow \mathbb{R}^2$ be the model that maps the one-hot encoded SMILES strings (vocabulary size of v and length of n) to the two target endpoints (here logBCF and logD). Supposing $\tilde{x}_i \in \mathbb{R}^{v\times n}$ is the perturbed version of x with the ith character replaced with a baseline character, the attribution (or sensitivity score) A^j_i of the character in target j can be written as

$$A^j_i = f_j(x) - f_j(\tilde{x}_i), \quad i \in \{0, 1, \dots, n-1\}, \ j \in \{0, 1\}. \tag{2}$$

Positive A_i values indicate a drop in predicted values when character i is replaced with the baseline; negative A_i values indicate an increase in predicted value when character i is replaced with the baseline. Because it is not immediately obvious what a good baseline character would be, we tested and compared two distinct approaches.

Character deletion In analogy to the black images commonly used as reference in attribution methods for natural images [46], we look at the effect of deleting a particular input character. In practice, we replace each character in turn with a dummy character “A” that is not part of the allowed vocabulary for CDDD. The CDDD feature encoding pipeline ignores such unsupported characters and effectively removes the character of interest from the input SMILES.

Character substitution Alternatively, we propose to define the character-wise sensitivity score of the input SMILES as the expectation of predictions when a particular position is occupied by any character in the vocabulary (see Fig. 2). Formally, for a character at position i of a given SMILES x of length n, its attribution to predicting target j is defined by

$$A^j_i = f_j(x) - \frac{1}{\mid v \mid}\sum_{k\in v} f_j(\tilde{x}^k_i), \quad i \in \{0, 1, \dots, n\}, \ j \in \{0, 1\} \tag{3}$$

with v the vocabulary set. $\tilde{x}^k_i$ is the perturbed SMILES with position i replaced by character k.

2.7. *Visualizations*

The raw output of the explainability method is a sensitivity score per SMILES character. To facilitate visualization and result analysis, we used XSMILES [47], which has been developed in parallel with this work. XSMILES is an interactive visualization library that combines 2D molecular depictions with bar charts. The two are interactively connected, allowing a user to make the link between a particular SMILES character (and associated score) with a particular atom or substructure of the molecule. Figures were created using XSMILES v0.6.3 with the following settings for logBCF, logD, and Diff, respectively: color palettes BayerBlRd9/RdBu_9_reverse/PuOr_9_reverse, color domains [-0.7, 0, 0.7]/[-1, 0, 1]/[-0.7, 0, 0.7], thresholds [0.25, 0.5, 0.75], highlight true.

3. Results and discussion

3.1. *BCF prediction models*

In this section, we evaluate our new xBCF multitask model and demonstrate its increased predictive power in predicting logBCF compared to baseline models, single task models and widely used models from the literature. We evaluated our models with leave-one-cluster-out cross-validation (LOCO), where clusters of molecules are predicted by models trained on the remaining subset of the training set. With this approach, predictions are not influenced by very similar molecules in the training set. Hence, we can measure the model’s ability to generalize (extrapolate). LOCO usually leads to less optimistic performance metrics compared to random split or leave-one-out cross-validation techniques. Performance results on the training set are summarized in Table 2. As expected, the null model performs poorly, but can provide a baseline reference point. A linear regression on logD as sole descriptor can already provide a reasonable estimation of logBCF, with an RMSE of 0.911 log units. Among the simple machine learning methods we tested, SVR in combination with CDDDs showed the best performance in nested cross-validation experiments and is the one kept here as our best single-task model. Both SVR and the new xBCF model outperform the baseline models by a comfortable margin. Data augmentation with randomized SMILES for xBCF did not lead to improved predictivity (xBCF randSML in Table 2). xBCF additionally achieves an impressive predictivity on logD (R^2 score of 0.908) and its goodness-of-fit, as evaluated on the entire training data, is very high (R^2 scores of 0.958 and 0.993 for logBCF and logD, respectively, see also Fig. 3). It is worth mentioning that molecular representations from Transformer models [21,48] and pre-

Table 2
Performance metrics evaluated on the BCF training set.

Model	Evaluation	#mol	R^2 logBCF	RMSE logBCF	R^2 logD	RMSE logD
Null model (mean)	LOCO	1026	-0.093	1.413	-0.066	1.884
Linear (LogD)	LOCO	1026	0.546	0.911	-	-
SVR	LOCO	1026	0.701	0.739	-	-
xBCF	fitted on entire set[a]	1045	0.956	0.276	0.993	0.147
xBCF	**LOCO**	**1026**	**0.703**	**0.736**	**0.908**	**0.554**
xBCF randSML	LOCO	1026	0.691	0.751	0.908	0.550
Meylan	entire BCF training set	1026	0.735	0.695	-	-
Meylan	excl. Meylan training set	535	0.662	0.829	-	-
Arnot-Gobas	entire BCF training set	1026	0.448	1.005	-	-
Vega Read-across	entire BCF training set[b]	1018	0.928	0.362	-	-
Vega Read-across	excl. Vega training set	184	0.704	0.784	-	-
Vega Read-across	inside AD[c]	102	0.718	0.680	-	-
CAESAR	entire BCF training set	1026	0.589	0.867	-	-
CAESAR	excl. training set	656	0.455	1.001	-	-
CAESAR	inside AD[c]	162	0.727	0.795	-	-
OPERA	entire BCF training set	1026	0.783	0.630	-	-
OPERA	excl. training set	576	0.672	0.800	-	-
OPERA	inside AD[c]	523	0.704	0.747	-	-

[a] model was trained on the entire data set including cluster 1 with 19 polyfluorinated molecules.

[b] model did not predict 8 molecules.

[c] excluding molecules outside the applicability domain (AD) or present in the training set of the model.

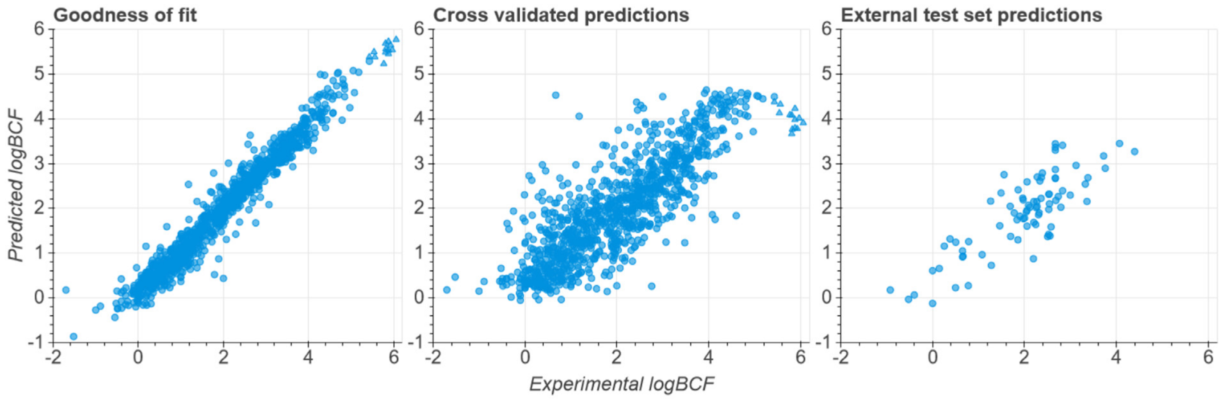

Fig. 3. Correlation between xBCF predictions and experimental values (left: goodness-of-fit on training set, middle: cross-validated training set, right: external test set). Triangles represent a subset of polychlorinated biphenyls for which xBCF underestimates logBCF in the LOCO setting.

trained graph neural nets [23] did not perform as well as CDDD descriptors in our experiments.

For comparison, we employed various publicly available models on our BCF training set. Note that we cannot retrain those models and we do not necessarily know the composition of their original training sets. Where possible, we provide the performance metrics on the subset of molecules from our BCF training set that were not part of the original training sets of the employed models. If information on the applicability domain is provided with the predictions, we additionally tested the effect of excluding predictions outside the applicability domain or marked with low reliability. In general, we observed that a good proportion of our dataset was used to train those models. This is not surprising since experimental BCF data is scarce and we took a dataset from the literature that had already been used for model development. As expected, all models tested performed significantly better on the entire set compared to the subset of molecules that were new to the respective models. An extreme example is the Vega Read-across model with more than 80% of the molecules being part of the training set. Moreover, this model simply takes the experimental value as prediction output leading to seemingly excellent performance (RMSE of 0.36). However, performance drops drastically for *new* molecules (RMSE of 0.78). Restricting evaluation to molecules inside the applicability domain can increase performance and is usually recommended. Indeed, the Vega Read-across and CAESAR models perform better on molecules with medium or high reliability.

From the models tested, only Vega Read-across was able to outperform our xBCF model, but only when considering molecules inside the applicability domain, i.e. when very similar molecules exist in the training set. This restricted the applicability to 102 out of 184 molecules. Our xBCF model on the other hand, was evaluated on the full set of molecules. Moreover, the LOCO evaluation process involves a certain degree of extrapolation, which means the model is useful even for new chemical classes.

We also compared the performance of xBCF, trained on the entire training data, with the other models on an external test set (Table 3 and Fig. 3). There is almost no overlap with the training sets of the models, which allows direct comparison. An interactive plot showing individual predictions on the test set for different models is available in the Supplementary Information.

On the test set, xBCF outperforms all other models. The performance metrics (R^2 of 0.68 and RMSE of 0.59) are even better compared to the cross-validation results due to higher similarity to the training set. For the other models, the external test set is a more demanding task compared to the training data and many models fail to properly predict these molecules. Some of the models even struggle to outperform the linear logD model, which might indicate overfitting. The generally inferior performance can have several reasons. First, the data we collected, although carefully checked, might have undergone a less rigorous curation than the training set, which had been checked and used in multiple works already. Therefore the experimental BCF values might be slightly

Table 3
Performance metrics on the independent logBCF test set with 80 compounds.

Models	#mol	R^2	RMSE
Null model (mean)	80	-1.457	1.632
Linear (LogD)	80	0.500	0.736
SVR	80	0.539	0.707
xBCF	80	**0.676**	**0.593**
Meylan	80	0.312	0.864
Arnot-Gobas	80	-0.036	1.060
Vega Read-across[a]	79	0.325	0.850
Vega Read-across (inside AD)	29	0.248	0.773
CAESAR[b]	80	-0.523	1.285
CAESAR (inside AD)	7	0.723	0.636
OPERA[b]	80	0.221	0.919
OPERA (inside AD)	66	0.272	0.916

[a] 2 molecules present in the training set; 1 molecule could not be predicted.
[b] 1 molecule present in the training set.

more noisy. Second, the more recently tested molecules are less similar to the molecules in the training sets of the models. Third, the molecules in the test set are larger and, we can assume, more difficult to predict. We expect that models that can generalize well will perform reasonably well on this set. In this case, the multitask approach used in xBCF brings an edge over other models, since its training set contains more diverse molecules and since multitask learning usually helps with generalization [49,50].

Figure 3, panel 2 displays a group of molecules with very high experimental logBCF values, which our CV model underestimates by 1–2 log units. This group consists exclusively of polychlorinated biphenyls (PCB) with 6 to 9 chlorine atoms. Other PCBs with fewer chlorine atoms are only slightly underestimated by our model, within the expected error range. Other polychlorinated molecules that are not biphenyls are well predicted, too. The data for these PCBs is originating from a study by Fox et al. [51], who reported only the kinetic BCF determined from the uptake and depuration rates, and unfortunately not the steady-state BCF like the majority of our dataset. Therefore, we cannot conclude whether this behaviour is indeed a special feature of these PCBs or an issue with the adequacy of that data. A general search for PCBs in the EPA-ECOTOX [52] database and the OECD QSAR Toolbox [53] revealed that the majority of PCBs have logBCF values between 3 and 5 (and not between 5 and 6), which would be more in line with our model predictions. It is also interesting to note that the underestimation almost disappears in the final model (where PCBs are included in the training set), which indicates that the model architecture is flexible enough to account for such complex features.

3.2. Applicability domain of the xBCF model

Our model's applicability domain is mainly defined by CDDD [13], which can operate on nearly all organic molecules consisting of the elements H, B, C, N, O, F, P, S, Cl, Br, and I. Stereoinformation cannot be taken into account and is removed during the pre-processing step. CDDD was originally designed for a molecular weight range up to 600 g/mol, but our model predicted logBCF for molecules with a weight of up to 886 g/mol without notable decrease in accuracy. The logBCF values in the training set range from -1.7 to 6.1. In our dataset, we have not identified any particular type of molecule for which our model fails or performs significantly worse. Our checks include ionic, large, lipophilic, and molecules with multiple hydrogen bond donor or acceptor moieties. Thus, we assume that xBCF's applicability domain covers the whole space of organic molecules with a molecular weight up to 900 g/mol. Note that xBCF's applicability domain benefits from the additional 60k molecules for which we have logD measurements in-house. In case of doubt, the xBCF prediction of logD can be compared to experimental data to cross-check that the model captured the lipophilicity of the molecule correctly.

3.3. Uncertainty of the xBCF model

The expected prediction error for new molecules can be derived from the performance metrics on the external test set and the cross-validation results. While the RMSE of 0.59 log units on the test set can be seen indicative for represented organic molecules, the cluster-split cross-validated RMSE of 0.74 log units is representative for new types of organic chemistry that were not in the training set of the xBCF model.

Additionally, we provide confirmatory checks that can detect model failures and thus increase the reliability of the successful predictions. We trained another version of xBCF on the augmented BCF set, where each compound was represented by 10 different randomized SMILES generated by RDKit. The resulting *xBCF randSML* model showed similar performance as the original xBCF model during cross validation (see Table 2). We computed the standard deviation of predictions from the ten SMILES of each compound and identified a slight correlation with prediction error (see Supporting Information). We conclude that a high spread in predictions from randomized SMILES indicates increased uncertainty. This can be due to uncertainty in the CDDD encoder or the downstream BCF prediction model.

As second confirmatory check, we can compare the predicted logD value to the experimental logD value. In case of high discrepancy (e.g. >2 log units), there is probably a problem with the CDDD encoder for that molecule. However, no strong correlation between logD and logBCF prediction errors was observed (see Supporting Information).

Finally, the sensitivity scores (see below) for logD and logBCF can be compared to expert expectations.

3.4. SMILES character sensitivity scores as a local explanation method

Our aim with the new predictive model for BCF is two-fold: first, we want to build the best possible model that generalizes well in chemical space and can be used in research projects to prioritize ideas and discover potential problems with chemical series. Second, we want to provide reliable alternatives to costly animal studies for environmental risk assessments. For this, good predictions are critical, but additional supporting information might be needed to increase the trust in our method. We therefore propose to accompany the predictions with local explanations that provide a qualitative or semi-quantitative interpretation of the predicted logD and logBCF values and help identify the most influential moieties and features in the molecules. Such *sensitivity scores* displayed onto the input chemical structures have several advantages. They can be seen as a plausibility check for the model's behavior, therefore allowing users to gain trust in the model or identify potential flaws and biases. When used in a more exploratory fashion, explanations can help discover unsuspected mechanistic effects or can guide the rational design of new molecules.

With these goals in mind, we introduce two input perturbation methods to obtain signed atom-level sensitivity scores. Together with smart visualizations, these methods allow inspecting the model predictions for each input molecule of interest. Because they are gradient-free, the methods can be widely applied, independently of the downstream modeling algorithm used.

Our explainability method is conceptually inspired by the one proposed by Sheridan [18] in 2019. In his approach, atoms in the query molecule are in turn replaced by the atom “Na” (which typically does not occur with covalent bonds in an organic molecule), and the atom importance is obtained by subtracting the predicted activity in presence of the Na atom to the predicted activity in presence of the original atom.

Instead of replacing an atom in a molecular structure, we work on the input SMILES strings and either remove each character (character deletion method) or replace it by each possible SMILES token included in the CDDD vocabulary (substitution method), and then calculate the difference in prediction to the original input SMILES. In contrast to atom-based methods, our approach allows to extract information for non-atom characters in the SMILES string, like ring openings/closings

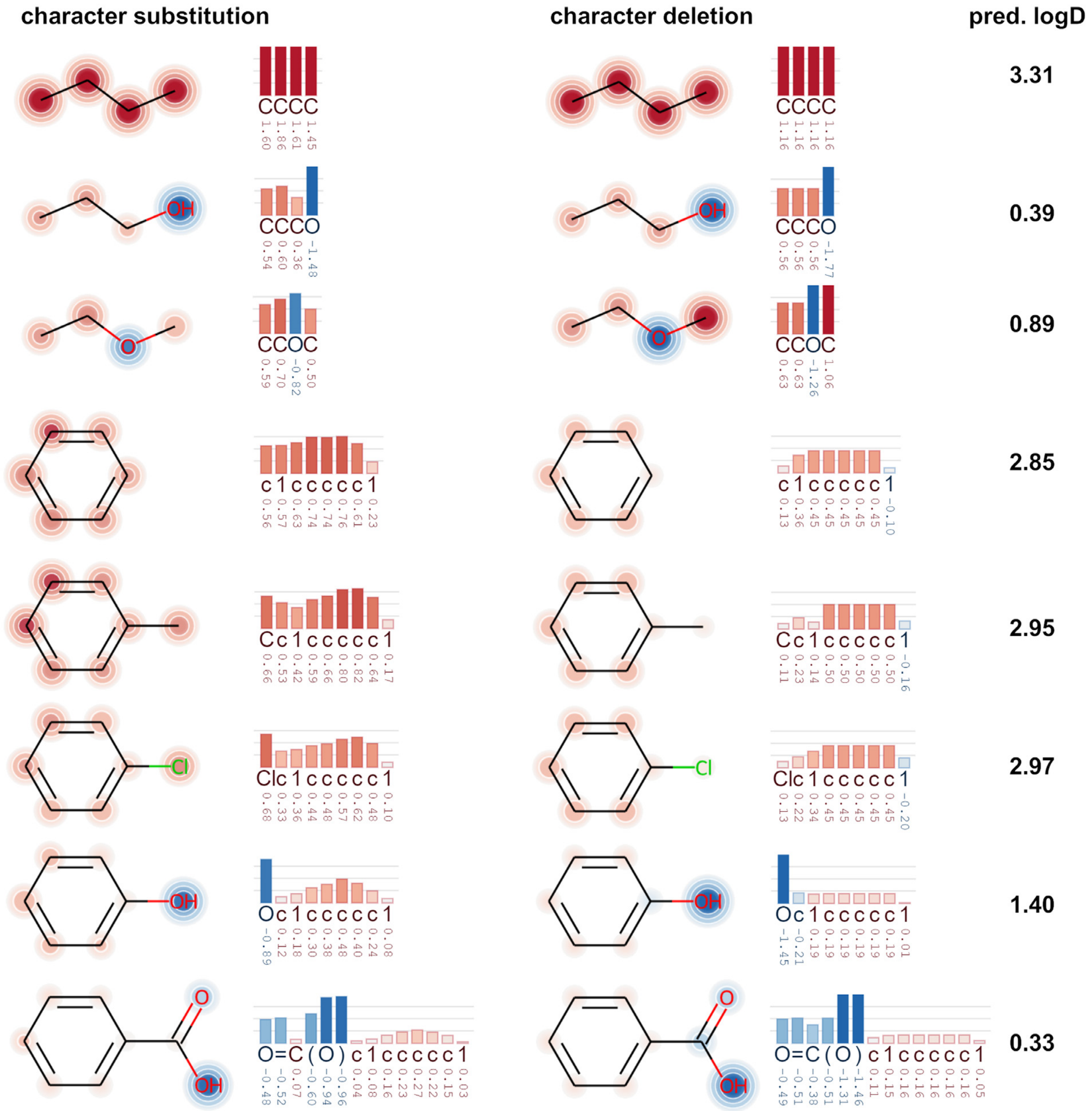
character substitution
character deletion
pred. logD
CCCC 1.60 1.86 1.61 1.45
CCCC 1.16 1.16 1.16 1.16
3.31
OH
CCCO 0.54 0.60 0.36 -1.48
OH
CCCO 0.56 0.56 0.56 -1.77
0.39
O
CCOC 0.59 0.70 -0.82 0.50
O
CCOC 0.63 0.63 -1.26 1.06
0.89
c1ccccc1 0.56 0.57 0.63 0.74 0.74 0.76 0.61 0.23
c1ccccc1 0.13 0.36 0.45 0.45 0.45 0.45 0.45 -0.10
2.85
Cc1ccccc1 0.66 0.53 0.42 0.59 0.66 0.80 0.82 0.64 0.17
Cc1ccccc1 0.11 0.23 0.14 0.50 0.50 0.50 0.50 0.50 -0.16
2.95
Cl
Clc1ccccc1 0.68 0.33 0.36 0.44 0.48 0.57 0.62 0.48 0.10
Cl
Clc1ccccc1 0.13 0.22 0.34 0.45 0.45 0.45 0.45 0.45 -0.20
2.97
OH
Oc1ccccc1 -0.89 0.12 0.18 0.30 0.38 0.48 0.40 0.24 0.08
OH
Oc1ccccc1 -1.45 -0.21 0.19 0.19 0.19 0.19 0.19 0.01
1.40
O
OH
O=C(O)c1ccccc1 -0.48 -0.52 0.07 -0.60 -0.94 -0.96 0.04 0.08 0.16 0.23 0.27 0.22 0.15 0.03
O
OH
O=C(O)c1ccccc1 -0.49 -0.51 -0.38 -0.51 -1.31 -1.46 0.11 0.15 0.16 0.16 0.16 0.16 0.16 0.05
0.33

Fig. 4. LogD sensitivity score visualizations for small molecules with single-character, single-atom or single-group changes.

or branching points. On the other hand, this might make interpretation more difficult. For example, a high sensitivity score might actually hint to the non-impact of an absent atom, that would be in that position in the string if present. It is important to note that these methods always reflect changes in the input string and do not represent absolute contributions to the desired property. We therefore prefer to call the resulting numeric values *sensitivity scores*, because they represent the degree of sensitivity of the predicted value with respect to changes of one particular character in a given molecule.

One immediate concern of applying character substitution or deletion approaches, is the chemical meaning of such modifications for the input molecule (note that this concern also applies to Sheridan's approach). Replacing characters (or atoms) by others will often lead to syntactically invalid SMILES or unstable molecules. Nevertheless, the CDDD encoder is able to generate continuous embeddings even for invalid perturbed input SMILES. These embeddings can then be used to generate similar (valid) molecules via the CDDD decoder. Although artificial from a chemical point of view, these small perturbations allow us to probe for the local behavior of the model around points that are very close to the original input, a concept that is also used in the well-known LIME approach [54].

The resulting sensitivity scores are context-dependent, as opposed to rule-based systems which would assign the same meaning to a particular substructure in all molecules, regardless of the context. The context of course refers to the input SMILES, such that different SMILES notations lead to small variations in the sensitivity scores. We provide examples based on randomized SMILES in the Supporting Information that show the qualitative stability of the explanations.

3.5. Interpretation of character sensitivity scores

A major difficulty we faced during development of our methods was the lack of a ground truth for the sensitivity scores. Although it is possible to calculate (and measure) the differences in logD or BCF between distinct molecules, it is not possible to map these differences to atoms or characters. They are not real physical entities and can neither be

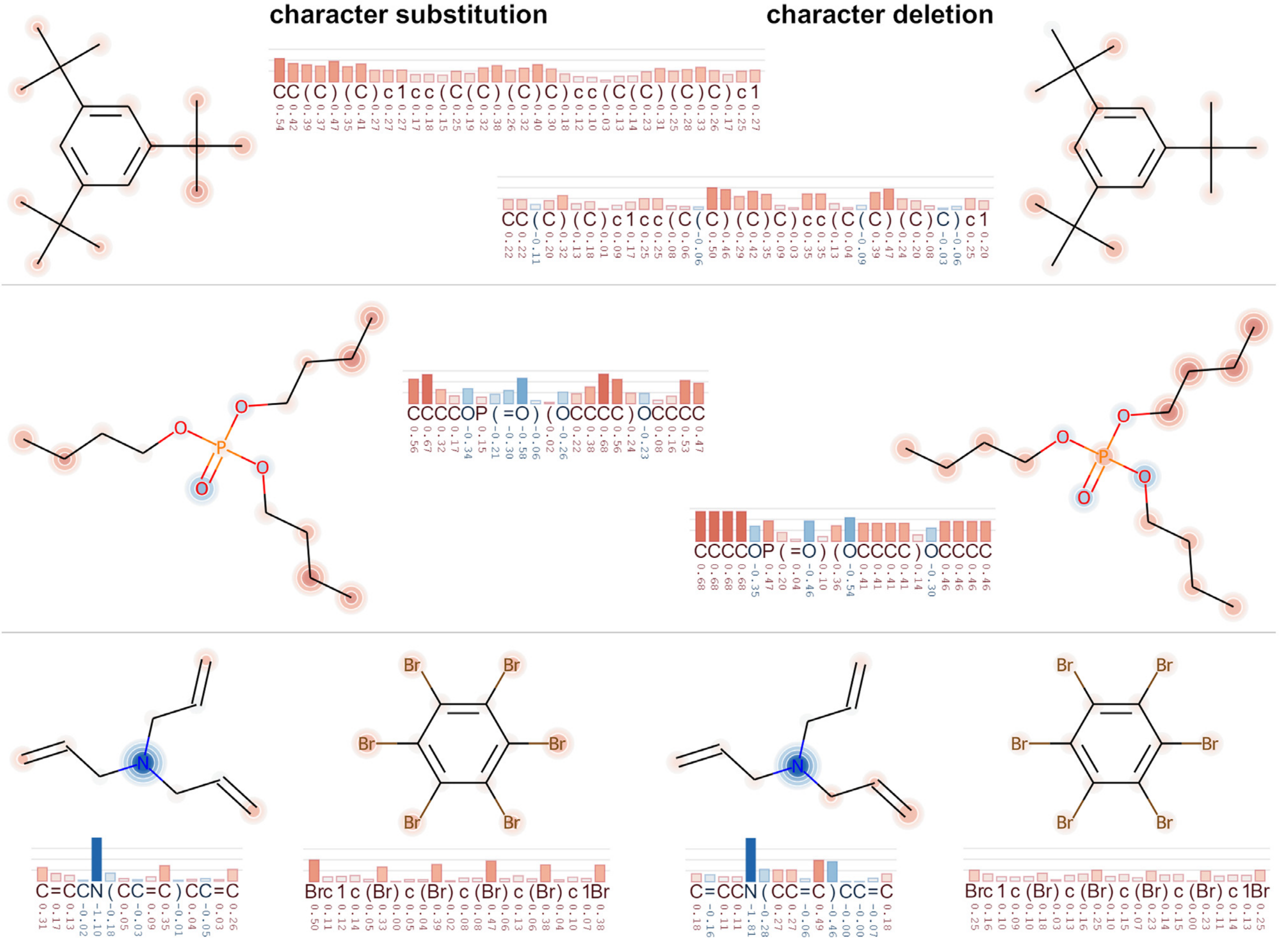

Fig. 5. LogD sensitivity visualizations for symmetric molecules. Left: character substitution method. Right: character deletion method.

measured nor be strictly derived from theoretical concepts. This particularly applies to SMILES characters but is also true for atomic input features or fingerprints, unless a synthetic atom-based additive ground truth like molecular weight or Crippen's logP [20] are modeled. However, in our case we are interested in a real physico-chemical property in a biological system. Therefore, in our analysis we rely mainly on expectations of chemistry experts and qualitative models from the field of physical organic chemistry [55], which have proven useful in explaining reactivity and chemical properties during the last decades. Additionally, we provide exemplary guidelines to help users interpret the sensitivity scores. We note that a more rigorous development of these explainability concepts is an important and promising endeavor for future research.

3.5.1. Comparison of explanations methods

As a first step, we provide examples that the explanations are in line with the general (*common sense*) expectations of trained physical-organic chemists and highlight some perks that can guide the interpretation. We focus on logD because the prediction model performs better on this task and the thermodynamics underlying the octanol-water distribution are well-understood and less complex compared to BCF.

We start by changing single atoms in small organic molecules (Fig. 4). The two methods (deletion and substitution) provide qualitatively similar explanations, in line with expert expectations. However, they show some notable differences due to their different design. In general, the score from the substitution method represents the change in the prediction upon change of a particular character, while the deletion method reflects removal of that character. An illustrative example is toluene: The substitution method assigns a medium sensitivity score to the aliphatic carbon (CH_3 group, character "C") because a replacement of this character would lead among others to phenol, which has a significantly lower logD. On the other hand, the deletion method assigns a value of almost zero because removal of the C character leads to a benzene molecule, which has an almost identical predicted logD value.

Another outcome from the depicted examples is that carbon atoms typically lead to positive sensitivity scores, while oxygen atoms lead to negative scores. We attribute this to their ability to form hydrogen bonds which strongly increase hydrophilicity. An even stronger effect can be observed for carboxylic acids, which are negatively charged at neutral pH, leading to very low logD values. For benzoic acid, the sensitivity scores for the characters encoding the COOH group have large negative values, including the non-atom characters like the equal sign and the parentheses. The sensitivity scores for the carbon atoms of the benzene ring are in this case very small because their influence on logD is minor in this molecule.

To explore the robustness of the approach and the ability of our model to correctly recognize patterns, i.e., functional groups, we performed tests for symmetric molecules. Figure 5 displays some of the resulting sensitivities. These molecules contain symmetry-equivalent functional groups, which share identical surroundings in the molecule and therefore should have identical properties. However, the input SMILES strings do not reflect the symmetry, and therefore our model has to infer this knowledge from its own internal representation. Overall, the symmetry of these molecules is qualitatively reflected in the attribution values, which means that the model is able to recognize functional groups and their molecular surroundings independent of their representation as characters in the SMILES string. The character substitution method seems more robust in this test compared to the character deletion method. Interestingly, the character deletion method attributed very high values to several terminal atoms with a double bond and closing brackets, which appear in non-terminal positions in the SMILES string (see e.g., triallylamine example). Removal of these characters would lead to the formal introduction of new bonds between remote parts of the molecule, and thus a more drastic change to the molecule than removal of a single atom.

Overall, we consider the character substitution method slightly more useful for users who want to focus on the visualization of the molecular structures, because of its robustness with respect to symmetry. The character deletion method can provide additional insights for terminal atoms, brackets, or other characters that have pronounced importance due to their particular position in the SMILES string. In the following sections, we focus on the character substitution method.

3.5.2. Functional group analysis

By training, many organic chemists are used to think in terms of functional groups to derive conceptual models and explain molecular properties. Not surprisingly, there is a large variety of QSAR approaches in the literature building on functional groups or substructure contributions, e.g., classical group contribution methods, atom-centered fragments, fingerprints, and many more [56]. An ideal explainable AI framework for molecular property prediction would assign attributions to both individual atoms and larger substructures. We were curious whether our data-driven approach could be used to explore effects of functional groups, keeping in mind that our sensitivity scores are conceptually different from additive contributions. For this analysis, we summed up the sensitivity scores for a set of common functional groups (including bracket and bond characters) for the molecules in our BCF training set.

We compare these group sensitivity scores to the correction factors of the KOWWIN model [7,36] as a typical representative for group contribution models. KOWWIN predicts logKow as a linear combination of group contributions, i.e., fitted parameters for a set of pre-defined atoms and functional groups. Additionally, we compared the QSARpy modulators [57] that were derived from differences in measured logKow upon addition or removal of molecular fragments. When multiple modulators matched the KOWWIN description we selected only the most general definition.

Figure 6 illustrates the comparison for several influential functional groups. Overall, the three approaches are in qualitative agreement, i.e. lipophilic functional groups like halogen atoms have small positive values whereas hydrophilic groups like ring heteroatoms, hydroxy groups or carboxylic acids have negative values. This indicates that our sensitivity scores can be used intuitively to interpret the effect of functional group changes on lipophilicity. However, there are notable differences between these metrics. Our sensitivity scores display a much broader spread and can be different for each individual molecule and even individual atoms/groups within the same molecule. In other words, they depend on the molecular context.

In the case of carboxylic acids, the sensitivity scores can span extreme ranges from around zero to -8. We investigated these molecules in more detail. Generally, carboxylic acids are acidic and therefore negatively charged under neutral pH conditions, resulting in very high hydrophilicity and low logD values. Molecules with an extreme group sensitivity score of around -8 are relatively small and contain several functional groups that usually lead to hydrophobic molecules. However, due to the negative charge, they become hydrophilic. Therefore, removal of the COOH group (charge) would drastically increase the logD value. In contrast, molecules with a group sensitivity score of around -3 are typically large lipophilic molecules where the carboxylic acid (charge) has only a minor effect. Molecules with sensitivity scores around zero contain a second acidic group. That means that the removal of one acidic group would not change the fact that these molecules are charged and therefore has only a minor effect. This effect is further illustrated in the molecule series benzene - benzoic acid - terephtalic acid (Fig. 7). Benzene is not charged and relatively lipophilic with a logD value of 2.13. Going to benzoic acid, the logD value is significantly decreased due to the negative charge. However, introducing the second COOH group leaves the logD mostly unchanged. This fact has been recognized before, e.g., by the introduction of an expert rule in the KOWWIN model that diminishes the effect of additional COOH groups to half of its original value [36]. Such rules were designed by experts and apply only to very few selected functional groups. Our model can correctly account for this effect and demonstrates that it is able to recognize the molecular context, which is a significant advantage over classical group contribution methods that employ a fixed contribution for all molecules. This also means that our prediction model seems more suitable for larger and more complex molecules, which are typically difficult for group contribution methods because of the many effects these functional groups can exert on each other [58]. The example of terephthalic acid also clearly demonstrates how not to interpret the sensitivity scores: These values are not linear contribution values that sum up to the prediction for the whole molecule. Instead, they highlight the effect of changes in the molecule. In general, the concept of linear additivity of functional groups seems not to hold. The molecular context is critical and interactions between functional groups can play an important role.

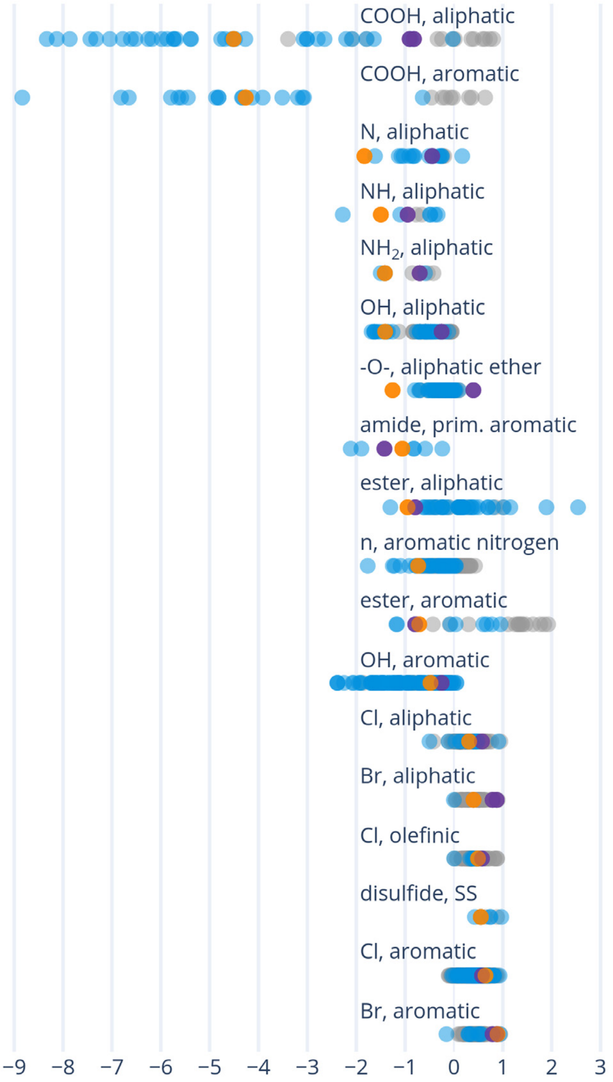

Fig. 6. Distribution of group sensitivity scores (blue for single and grey for multiple occurrences in a single molecule) together with the corresponding contribution factors from KOWWIN (orange) and QSARpy modulators (purple). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

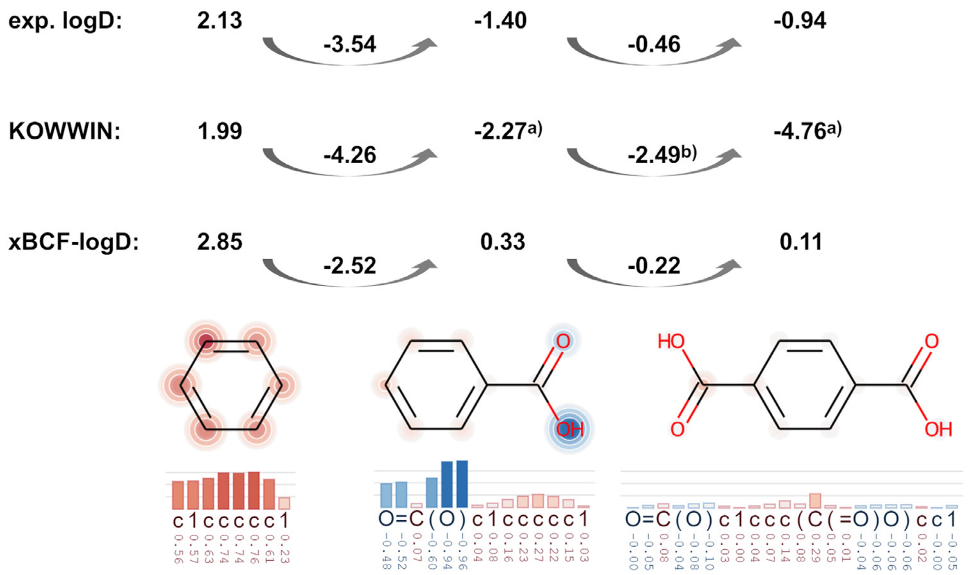

Fig. 7. Comparison of logD predictions and sensitivity scores for the series benzene - benzoic acid - terephtalic acid. a) calculated as sodium salt, b) KOWWIN employs a specific expert rule for carboxylic acids to reduce the correction term for each additional COOH group.

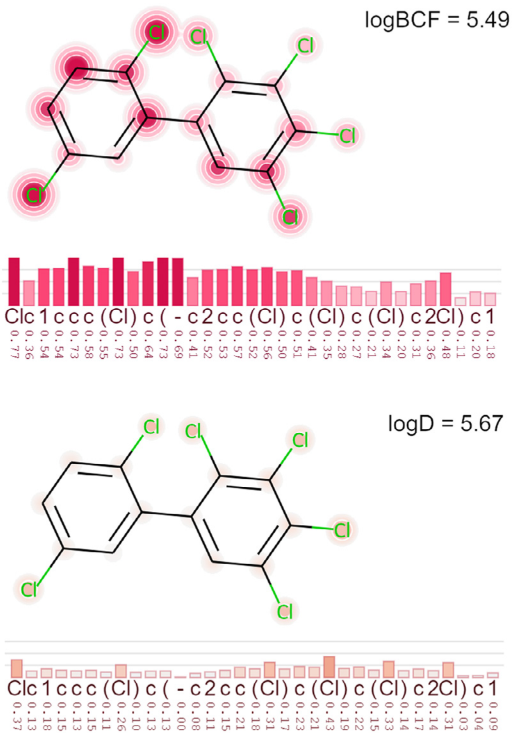

Fig. 8. Sensitivity scores for a PCB molecule.

Looking for more complex structural patterns, we revisited the PCBs. When included in the training set, our model has learned to correctly predict PCBs with more than 5 chlorine atoms. Figure 8 illustrates such an example. The whole biphenyl moiety and the chlorine atoms receive high scores for logBCF but not for logD, demonstrating that our model indeed recognizes this complex pattern. The two chlorine atoms in the left ring show higher scores because substitution of those atoms would result in a phenyl ring with very low degree of chlorination and presumably lower BCF. Chlorinated phenyl and biphenyl moieties have also been proposed as structural alerts for bioaccumulation by Valsecchi et al. [59], which confirms the relevance of this pattern. Another structural alert we were able to confirm are CF_3 groups (see below).

Whether our group sensitivity scores can be interpreted quantitatively, requires further analysis. However, we conclude that the group sensitivity scores can provide a first qualitative interpretation of moieties with high influence. We recommend these scores as starting points for a deeper analysis of the model and its predictions.

Additional insights can be gained from the relationship between logD and logBCF. Frequently, a linear relationship is assumed [4,5], which holds for the majority of molecules, typically with a fitted coefficient around 0.6 for logD plus correction terms for defined functional groups. We are particularly interested in these corrections and the deviation from the linear relationship because these deviations can give valuable insights into mechanistic processes. Since we have a multitask model that predicts both logBCF and logD simultaneously, we can exploit this relationship and calculate sensitivity scores for the deviation, i.e.

$$A_i^{Diff} = A_i^{logBCF} - 0.6 * A_i^{logD} \tag{4}$$

These values correspond to sensitivities for the part of the BCF prediction that cannot be explained by lipophilicity alone. Please note that although we calculate quantitative metrics, their derivation is mostly empirical and they should only be interpreted qualitatively.

We plotted the sensitivity differences for various functional groups in Fig. 9. For many functional groups, the differences are indeed close to zero, as expected. However, for several functional groups, we see significant deviation from the linear relationship. On the negative side, functional groups like disulfides, esters, amides, and ketones reduce the BCF. These are known to be readily metabolized in fish and other organisms. On the contrary, carboxylic acids and trifluoromethyl (CF_3) have positive sensitivity differences. The latter is often deliberately introduced in active ingredients to block metabolism in a specific position. It is a remarkable observation that our data-driven approach can highlight such mechanistic insights. This does not mean that our model can predict metabolism, but it can provide hypotheses based on the parts of the molecule that change the deviation from the expected logD-logBCF relationship. Many physiology-based BCF prediction models, e.g., the Arnot-Gobas model [35], explicitly model metabolization as an important depuration and detoxification mechanism. The Arnot-Gobas model predicts the biotransformation rate constant from functional group coefficients. For the highlighted groups mentioned above the Arnot-Gobas model employs high biotransformation coefficients (see Supplementary Information), which are in qualitative agreement with our observations (with the exception of disulfides, which are not covered in the Arnot-Gobas model).

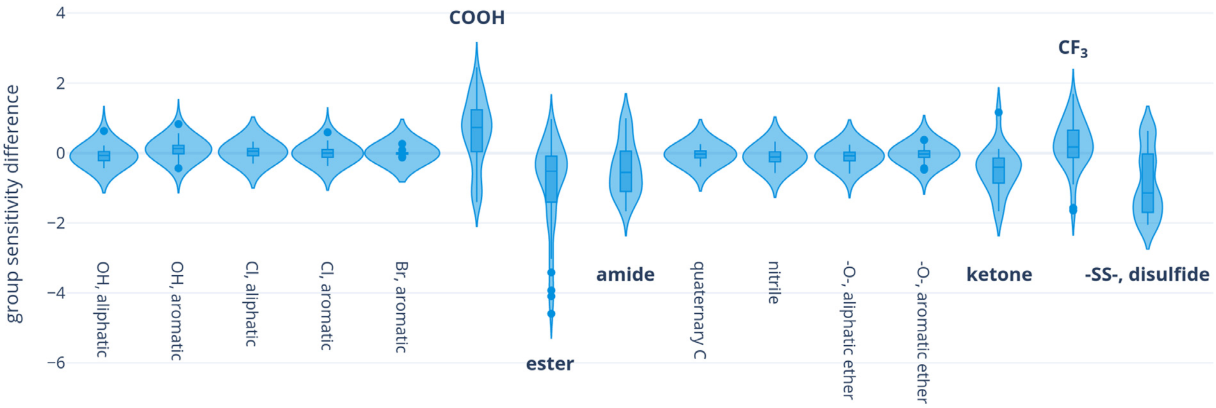

Fig. 9. Sensitivity Difference between logBCF and logD for selected functional groups.

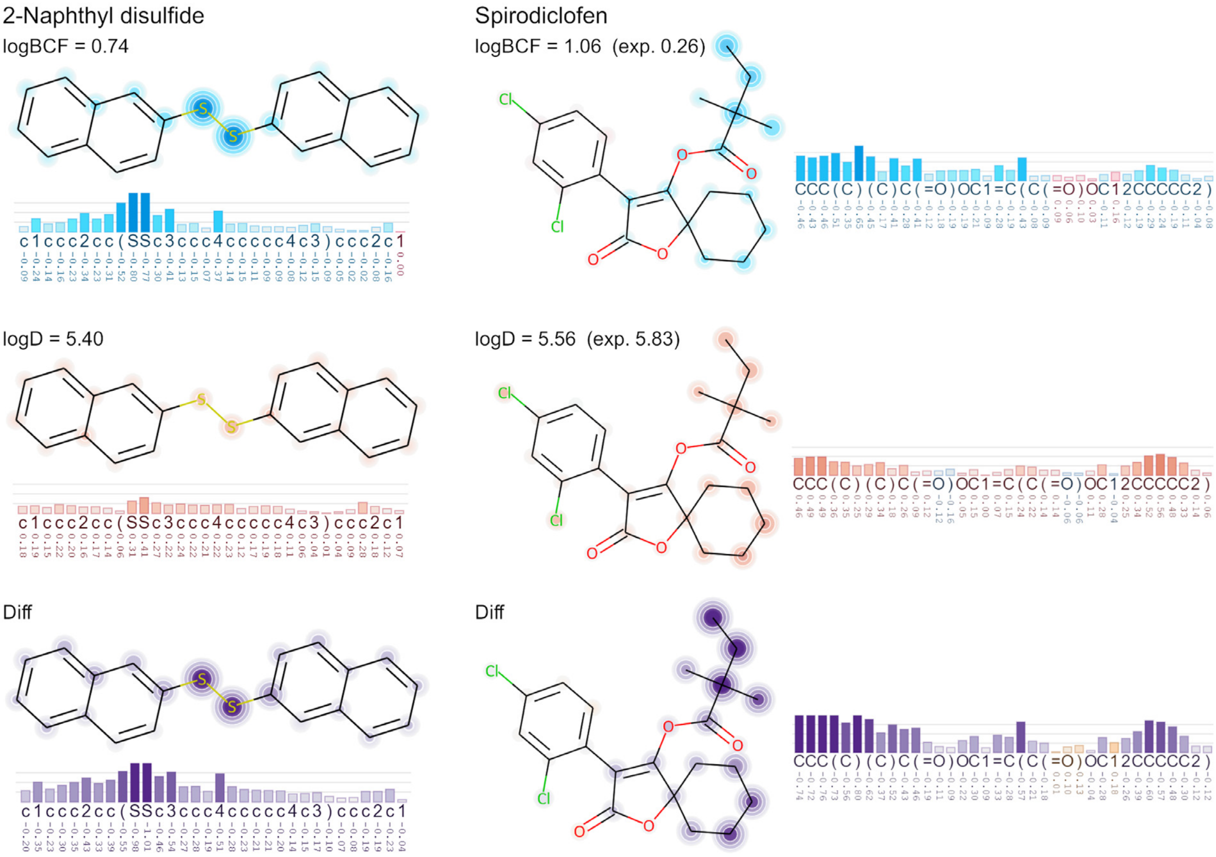
2-Naphthyl disulfide
logBCF = 0.74
c1ccc2cc(SSc3ccc4ccccc4c3)ccc2c1
logD = 5.40
c1ccc2cc(SSc3ccc4ccccc4c3)ccc2c1
Diff
c1ccc2cc(SSc3ccc4ccccc4c3)ccc2c1
Spirodiclofen
logBCF = 1.06 (exp. 0.26)
CCC(C)(C)C(=O)OC1=C(C(=O)OC12CCCCC2)
logD = 5.56 (exp. 5.83)
CCC(C)(C)C(=O)OC1=C(C(=O)OC12CCCCC2)
Diff
CCC(C)(C)C(=O)OC1=C(C(=O)OC12CCCCC2)
Cl
Cl
O
O
O
O
S
S

Fig. 10. Sensitivity scores and differences for example molecules known to be readily metabolized.

Figure 10 highlights examples with high logD and low BCF values due to rapid metabolization. Disulfides show high reactivity and are generally known to be rapidly metabolized by a wide range of organisms. Our data-driven model clearly identified the two sulfur atoms. The second example is Spirodiclofen, which is known to be rapidly metabolized with a depuration half-life of 1–2 days [60]. The main metabolic routes are ester hydrolysis to the enol and subsequently hydroxylation in the 4-, and 3-position of the cyclohexyl ring. Our model highlights both of these moieties. We are currently exploring more examples in a more systematic way.

4. Conclusion

In this work, we propose a new model to predict the bioconcentration factor of small organic molecules. We use a multitask learning approach, where both lipophilicity (logD) and BCF are predicted simultaneously. Based on a rigorous evaluation, we show that our model is more accurate than previously published models, and generalizes well on unknown regions of the chemical space. It is purely data-driven and therefore independent from expert-crafted rules or the definition of functional groups. More importantly, we provide SMILES token sensitivity scores which serve as local explanations for each prediction. This new explanation method relies on SMILES token substitutions or deletions, and can be applied to downstream models based on CDDD descriptors or other string-based models learning directly from SMILES strings. We hope that together with a robust prediction model, the explainability layer will increase the value and trustworthiness of the predictions. It allows users to poke inside the inner workings of the model and could help with molecular design and mechanistic understanding of processes involved in bioaccumulation. We discussed various examples to spotlight important perks of the explanation methods, and to give guidance for their interpretation. We could show that the two methods highlight, how the prediction changes upon changing (substitution approach) or removing (deletion approach) a character from the input SMILES. We emphasize that this approach provides useful means to understand the *predictions* of the model based on its internal representation. However, the sensitivity scores should not be confused with incremental contribution factors that sum up to the total prediction for the molecule.

Overall, with its accuracy (RMSE = 0.59 log units), the broad applicability, its ability to generalize, and the explainability aspects, our model satisfies the criteria for many use cases and could be considered an alternative for animal testing.

5. Data and model availability

We share the logBCF training and test sets along with the associated logD values (measured or predicted), cluster splits, and BCF predictions. The logD training set is a proprietary in-house dataset and cannot be shared. We provide the sensitivity scores as a JSON file that can be used with XSMILES. The python code repository to calculate sensitivity scores is available via https://github.com/Bayer-Group/xBCF.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

The authors are employees of Bayer AG, a manufacturer of pharmaceuticals, agricultural, and consumer health chemicals. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Data availability

Data will be made available on request.

Acknowledgments

The authors acknowledge funding from the Bayer AG Life Science Collaboration (Explainable AI). The authors want to thank Marco Bertolini, Julian Heinrich, Ryan Henderson, Jacopo Negroni, Robin Winter for valuable discussions, contributions, and support in the context of this project.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ailsci.2022.100047.

References

[1] OECD. Test No. 305: bioaccumulation in fish: aqueous and dietary exposure. 2012. 10.1787/9789264185296-en
[2] Halder M, Kienzler A, Whelan M, Worth A. EURL ECVAM strategy to replace, reduce and refine the use of fish in aquatic toxicity and bioaccumulation testing. Publications Office; 2014.
[3] U.S. Environmental Protection Agency, Office of Chemical Safety and Pollution Prevention. Strategic plan to promote the development and implementation of alternative test methods within the TSCA program. Washington, DC2018;.
[4] Grisoni F, Consonni V, Villa S, Vighi M, Todeschini R. QSAR models for bioconcentration: is the increase in the complexity justified by more accurate predictions? Chemosphere 2015;127:171–9. doi:10.1016/j.chemosphere.2015.01.047.
[5] Gissi A, Lombardo A, Roncaglioni A, Gadaleta D, Mangiatordi GF, Nicolotti O, et al. Evaluation and comparison of benchmark QSAR models to predict a relevant reach endpoint: the bioconcentration factor (bcf). Environ Res 2015;137:398–409. doi:10.1016/j.envres.2014.12.019.
[6] Meylan WM, Howard PH, Boethling RS, Aronson D, Printup H, Gouchie S. Improved method for estimating bioconcentration/bioaccumulation factor from octanol/water partition coefficient. Environ Toxicol Chem 1999;18(4):664–72. doi:10.1002/etc.5620180412.
[7] United States Environmental Protection Agency. Epi suite (estimation programs interface suite), version 4.11 (november 2012), the software can be obtained free of charge from https://www.epa.gov/tsca-screening-tools/download-epi-suitetm-estimation-program-interface-v411. 2012.
[8] Hermens JL, de Bruijn JH, Brooke DN. The octanol-water partition coefficient: strengths and limitations. Environ Toxicol Chem 2013;32(4):732–3. doi:10.1002/etc.2141.
[9] Endo S, Escher BI, Goss K-U. Capacities of membrane lipids to accumulate neutral organic chemicals. Environ Sci Technol 2011;45(14):5912–21. doi:10.1021/es200855w.
[10] Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, et al. Analyzing learned molecular representations for property prediction. J Chem Inf Model 2019;59(8):3370–88.
[11] Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: a benchmark for molecular machine learning. Chem Sci 2018;9(2):513–30.
[12] Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model 2010;50(5):742–54.
[13] Winter R, Montanari F, No F, Clevert D-A. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. Chem Sci 2019;10:1692–701. doi:10.1039/C8SC04175J.
[14] Montanari F, Kuhnke L, Ter Laak A, Clevert D-A. Modeling physico-chemical ADMET endpoints with multitask graph convolutional networks. Molecules 2019;25(1):44.
[15] Duvenaud DK, Maclaurin D, Iparraguirre J, Bombarell R, Hirzel T, Aspuru-Guzik A, et al. Convolutional networks on graphs for learning molecular fingerprints. Adv Neural Inf Process Syst 2015;28.
[16] Bouhedjar K, Boukelia A, Khorief Nacereddine A, Boucheham A, Belaidi A, Djerourou A. A natural language processing approach based on embedding deep learning from heterogeneous compounds for quantitative structure–activity relationship modeling. Chem Biol Drug Des 2020;96(3):961–72. doi:10.1111/cbdd.13742.
[17] Chithrananda S., Grand G., Ramsundar B.. ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. arXiv preprint arXiv:2010098852020. 10.48550/ARXIV.2010.09885
[18] Sheridan RP. Interpretation of QSAR models by coloring atoms according to changes in predicted activity: how robust is it? J Chem Inf Model 2019;59:1324–37. doi:10.1021/acs.jcim.8b00825.
[19] Matveieva M, Polishchuk P. Benchmarks for interpretation of QSAR models. J Cheminform 2021;13:41. doi:10.1186/s13321-021-00519-x.
[20] Rasmussen M.H., Christensen D.S., Jensen J.H.. Do machines dream of atoms? A quantitative molecular benchmark for explainable ai heatmaps. ChemRxiv preprint 2022-gnq3w2022. 10.26434/chemrxiv-2022-gnq3w
[21] Karpov P, Godin G, Tetko IV. Transformer-CNN: Swiss knife for QSAR modeling and interpretation. J Cheminform 2020;12(1):1–12.
[22] McCloskey K, Taly A, Monti F, Brenner MP, Colwell LJ. Using attribution to decode binding mechanism in neural network models for chemistry. Proc Natl Acad Sci 2019;116(24):11624–9. doi:10.1073/pnas.1820657116.
[23] Henderson R, Clevert D-A, Montanari F. Improving molecular graph neural network explainability with orthonormalization and induced sparsity. In: International conference on machine learning. PMLR; 2021. p. 4203–13.
[24] Xie S., Lu M.. Interpreting and understanding graph convolutional neural network using gradient-based attribution method. arXiv preprint arXiv:1903037682019;.
[25] Rodríguez-Pérez R, Bajorath J. Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. J Comput Aided Mol Des 2020;34(10):1013–26. doi:10.1007/s10822-020-00314-0.
[26] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 2017;30.
[27] Preuer K., Klambauer G., Rippmann F., Hochreiter S., Unterthiner T.. Interpretable deep learning in drug discovery. arXiv preprint arXiv:1903027882019.
[28] Wellawatte GP, Seshadri A, White AD. Model agnostic generation of counterfactual explanations for molecules. Chem Sci 2022;13:3697–705. doi:10.1039/D1SC05259D.
[29] Jimnez-Luna J, Skalic M, Weskamp N. Benchmarking molecular feature attribution methods with activity cliffs. J Chem Inf Model 2022;62:274–83.
[30] Karpov K, Mitrofanov A, Korolev V, Tkachenko V. Size doesn't matter: predicting physico- or biochemical properties based on dozens of molecules. J Phys Chem Lett 2021;12(38):9213–19. doi:10.1021/acs.jpclett.1c02477.
[31] Lewis KA, Tzilivakis J, Warner DJ, Green A. An international database for pesticide risk assessments and management. Hum Ecol Risk AssessInt J 2016;22(4):1050–64. doi:10.1080/10807039.2015.1133242.
[32] Montanari F, Kuhnke L, Ter Laak A, Clevert D-A. Modeling physico-chemical admet endpoints with multitask graph convolutional networks. Molecules 2020;25(1). doi:10.3390/molecules25010044.
[33] Rdkit: Open-source cheminformatics; http://www.rdkit.org. 2021.
[34] Cawley GC, Talbot NL. On over-fitting in model selection and subsequent selection bias in performance evaluation. J Mach Learn Res 2010;11:2079–107.
[35] Arnot J, Gobas F. A generic QSAR for assessing the bioaccumulation potential of organic chemicals in aquatic food webs. QSAR Comb Sci 2003;22(3):337–45. doi:10.1002/qsar.200390023.
[36] Meylan WM, Howard PH. Atom/fragment contribution method for estimating octanol–water partition coefficients. J Pharm Sci 1995;84(1):83–92. doi:10.1002/jps.2600840120.
[37] Lombardo A, Roncaglioni A, Boriani E, Milan C, Benfenati E. Assessment and validation of the caesar predictive model for bioconcentration factor (BCF) in fish. Chem Cent J 2010;4:S1. doi:10.1186/1752-153X-4-S1-S1.
[38] Zhao C, Boriani E, Chana A, Roncaglioni A, Benfenati E. A new hybrid system of QSAR models for predicting bioconcentration factors (BCF). Chemosphere 2008;73:1701–7. doi:10.1016/j.chemosphere.2008.09.033.

[39] Vega in silico platform, version 1.2.0, available from www.vega-qsar.eu. 2021.

[40] Floris M, Manganaro A, Nicolotti O, Medda R, Mangiatordi GF, Benfenati E. A generalizable definition of chemical similarity for read-across. J Cheminform 2014;6:39. doi:10.1186/s13321-014-0039-1.

[41] Mansouri K, Grulke CM, Judson RS, Williams AJ. Opera models for predicting physicochemical properties and environmental fate endpoints. J Cheminform 2018;10:10. doi:10.1186/s13321-018-0263-1.

[42] Miller TH, Gallidabino MD, MacRae JI, Owen SF, Bury NR, Barron LP. Prediction of bioconcentration factors in fish and invertebrates using machine learning. Sci Total Environ 2019;648:80–9. doi:10.1016/j.scitotenv.2018.08.122.

[43] Kobayashi Y, Yoshida K. Development of QSAR models for prediction of fish bioconcentration factors using physicochemical properties and molecular descriptors with machine learning algorithms. Ecol Inform 2021;63:101285. doi:10.1016/j.ecoinf.2021.101285.

[44] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: European conference on computer vision. Springer; 2014. p. 818–33.

[45] Ivanovs M, Kadikis R, Ozols K. Perturbation-based methods for explaining deep neural networks: a survey. Pattern Recognit Lett 2021;150:228–34. doi:10.1016/j.patrec.2021.06.030.

[46] Ancona M., Ceolini E., Öztireli C., Gross M.. Towards better understanding of gradient-based attribution methods for deep neural networks. arXiv preprint arXiv:1711061042017;.

[47] Heberle H, Zhao L, Schmidt S, Wolf T, Heinrich J. XSMILES: interactive visualization for molecules, SMILES and XAI scores. Research Square preprint; 2022. doi:1021203/rs3rs-2121152/v1.

[48] Honda S., Shi S., Ueda H.R.. Smiles transformer: pre-trained molecular fingerprint for low data drug discovery. arXiv preprint arXiv:1911047382019;.

[49] Caruana R. Multitask learning. Mach Learn 1997;28(1):41–75.

[50] Ruder S.. An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706050982017;.

[51] Fox K, Zauke GP, Butte W. Kinetics of bioconcentration and clearance of 28 polychlorinated biphenyl congeners in zebrafish (Brachydanio rerio). Ecotoxicol Environ Saf 1994;28:99–109. doi:10.1006/eesa.1994.1038.

[52] Olker JH, Elonen CM, Pilli A, Anderson A, Kinziger B, Erickson S, et al. The ecotoxicology knowledgebase: a curated database of ecologically relevant toxicity tests to support environmental research and risk assessment. Environ Toxicol Chem 2022;41:1520–39. doi:10.1002/etc.5324.

[53] Dimitrov SD, Diderich R, Sobanski T, Pavlov TS, Chankov GV, Chapkanov AS, et al. QSAR toolbox – workflow and major functionalities. SAR QSAR Environ Res 2016;27:203–19. doi:10.1080/1062936X.2015.1136680.

[54] Ribeiro MT, Singh S, Guestrin C. "Why should i trust you?": explaining the predictions of any classifier. In: KDD '16: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. New York, NY, USA: Association for Computing Machinery; 2016. p. 1135–44. doi:10.1145/2939672.2939778. ISBN 9781450342322

[55] Anslyn EV, Dougherty DA. Modern physical organic chemistry. University Science Books; 2006. ISBN 9781891389313

[56] Todeschini R, Consonni V. Molecular descriptors for chemoinformatics, vol 41. Wiley-VCH; 2009. doi:101002/9783527628766. ISBN 9783527318520

[57] Ferrari T, Lombardo A, Benfenati E. QSARpy: a new flexible algorithm to generate QSAR models based on dissimilarities. the log Kow case study. Sci Total Environ 2018;637–638:1158–65. doi:10.1016/j.scitotenv.2018.05.072.

[58] Schneider N, Klein R, Lange G, Rarey M. Nearly no scoring function without a hansch-analysis. Mol Inform 2012;31(6–7):503–7. doi:10.1002/minf.201200022.

[59] Valsecchi C, Grisoni F, Consonni V, Ballabio D. Structural alerts for the identification of bioaccumulative compounds. Integr Environ Assess Manag 2019;15:19–28. doi:10.1002/ieam.4085.

[60] . [14c]-baj2740-bioconcentration in bluegill (lepomis macrochirus) under flow-through conditionsGLP study report. Bayer AG; 2000. Full study report available upon request via Bayer's transparency website or cropscience-transparency@bayer.com