Fast Prediction of Lipophilicity of Organofluorine Molecules: 
Deep Learning Derived Polarity Characters and 
Experimental Tests

Journal: *Journal of Chemical Information and Modeling*

Manuscript ID: ci-2022-01201t.R1

Manuscript Type: Article

Date Submitted by the Author: 04-Oct-2022

Complete List of Authors: Jia, Qingqing; Nanjing University, School of Chemistry and Chemical Engineering Ni, Yifan; Nanjing University, School of Chemistry and Chemical Engineering Liu, Ziteng; Nanjing University, School of Chemistry and Chemical Engineering Gu, Xu; Nanjing University Cui, Ziyi; Nanjing University Fan, Mengting; Nanjing University Zhu, Qiang; Nanjing University Wang, Yi; Nanjing University, School of Chemistry and Chemical Engineering Ma, Jing; Nanjing University, School of Chemistry and Chemical Engineering

Qingqing Jia,¹ Yifan Ni,² Ziteng Liu,¹ Xu Gu,¹ Ziyi Cui,¹ Mengting Fan,¹ Qiang Zhu,¹,* Yi Wang,²,* Jing Ma¹,²,*

¹Key Laboratory of Mesoscopic Chemistry of Ministry of Education, School of Chemistry and Chemical Engineering, Nanjing University, Nanjing 210023, P. R. China
²Jiangsu Key Laboratory of Advanced Organic Materials, School of Chemistry and Chemical Engineering, Nanjing University, Nanjing 210023, P. R. China

E-mail: majing@nju.edu.cn; yiwang@nju.edu.cn; csu1505110121@163.com;

Abstract

Fast and accurate estimation of lipophilicity for organofluorine molecules is of great demand in accelerating drug and materials discovery. A lipophilicity dataset of organofluorine molecules (OFL dataset), containing 1907 samples, is constructed through density functional theory (DFT) calculations and experimental measurements. An efficient and interpretable model, called PoLogP, is developed to predict the n-octanol/water partition coefficient, log \( P_{o/w} \), of organofluorine molecules on basis of the descriptors of polarization which is a combination of polarity descriptors including the molecular polarity index (MPI) and molecular polarizability (\( \alpha \)), and hydrogen bonds (HBs) index, consisting of the number of donors (\( N_{HB}^D \)) and acceptors (\( N_{HB}^A, N_{HB}^{A-F} \)). The present PoLogP with a combination of polarity descriptors is demonstrated to perform better than the dipole moment (\( \mu \)) alone for the F-contained molecules. With the aid of a multilevel attention graph convolutional neural network model, the fast generation of polarity descriptors of organofluorine molecules could be achieved with the DFT accuracy based only on topological molecular graph structure. The performance of PoLogP is further validated on synthesized organofluorine molecules and 2626 non-fluorinated molecules with satisfactory accuracy, highlighting the potential usage of PoLogP in high-throughput screening of the functional molecules with the desired solubility in various solvent media.

Introduction

Introducing fluorine atom(s) is an important strategy for improving the biological activities in drug design and materials discovery.¹⁻⁴ Substitution of H by F atom can profoundly change the physicochemical properties of a specific molecule, such as the enhanced metabolic stability⁵ and
absorption in the body. Recently, the increased metabolic stability by the introduction of -CF$_3$ group has also been found in the discovery of the novel COVID-19 oral antiviral drug, PAXLOVID™ (PF-07321332) released by Pfizer.

The F substitution could also tune the lipophilicity, which is commonly quantified by the partition coefficient, the ratio of solubility between two different phases, such as the $n$-octanol (called o) and water (w). The partition coefficient is often represented in the log unit and denoted as log $P_{o/w}$. The accurate and reliable determination of log $P_{o/w}$ for organic molecules has been a highly active field in the drug and materials discovery. Diverse experimental methods and computational methods have been developed. Such as, several models armed with 2D molecular fingerprints perform well and is comparable with those used 3D fingerprints in the prediction of log $P_{o/w}$. However, for organofluorine molecules, either a quantitative and interpretable partition coefficient prediction model is scarce or the effect on introducing perfluoroalkyl group is elusive. For example, CH$_3$/CF$_3$ exchange in indole derivatives and alkanols could lead to the decrease and increase of partition coefficient, respectively. A simple bond vector analysis was proposed so as to explain the change in log $P_{o/w}$ when introducing fluorine. However, it could only be applied to molecules with similar structures and give qualitative assessments. Thus, developing a method which could rapidly and accurately estimate the partition coefficient of organofluorine molecules is in demand.

The ideal descriptors to predict the log $P_{o/w}$ should simultaneously describe non-covalent solute-solvent interactions in two phases of $n$-octanol and water, especially for the electrostatic interaction and the hydrogen bonding interaction in water solvents. The information of electrostatic potential on van der Waals (vdW) surface, named as GIPF (general interaction properties function), was demonstrated to be important in predicting log $P_{o/w}$. The $\Pi$ descriptor was introduced to measure the local polarity, in correlation with two molecular polarity descriptors (i.e., dielectric constant and dipolarity/polarizability). In analogy to the $\Pi$ descriptor, the molecular polarity index (MPI) as equipped in academic software, is applied in this study to reflect the polarity of organofluorine molecules. In addition, the molecular polarizability ($\alpha$) was reported to correlate with the solubility of molecules. As exemplified in Figure 1a, the increase in lipophilicity upon F-substitution from C$_4$H$_4$F$_5$OH to C$_4$H$_2$F$_7$OH could not be solely predicted by one descriptor. The question of how to reasonably describe the trade-off of polarization parameters such as $\alpha$ and MPI is a key for building a predictive model for organofluorine molecules. It will be shown in this work that the effects of $\alpha$ and MPI on the log $P_{o/w}$ are different and even opposite in some cases.
In this work, an efficient model, called PoLogP, is developed with polarity descriptors of $\alpha$ and $MPI$ and hydrogen bonds (HBs) by combining density functional theory (DFT) calculations and the machine learning (ML) techniques (Figure 1b). A specific dataset for organofluorine molecules named OFL was built and conformational effects were taken into consideration. The $\alpha$ and $MPI$ descriptors with DFT accuracy were achieved by multilevel attention graph convolutional neural network which automatically learn features from the molecular graphs. The performance of PoLogP was further tested by our lipophilicity measurements over 11 synthesized molecules. The transferability of PoLogP was demonstrated by four non-fluorinated molecules datasets with totally 2626 samples. The quantification, efficiency, and interpretability of PoLogP may aid to reveal the mystery of polarization effect in lipophilicity so as to shed light on the fields of drug design and materials discovery.

Figure 1. (a) The illustration of electrostatic potential on vdW surface for C$_4$H$_4$F$_5$OH and C$_4$H$_2$F$_7$OH molecules along with $\alpha$, $MPI$ and HBs descriptors as well as $\log P_{exp}^{\text{logP}}$. (b) Flowchart including the construction of OFL datasets together with the conformation sampling; the development of PoLogP, including the selection of HBs and $\alpha$, $MPI$ descriptors and the fast generation of $\alpha$ and $MPI$ by multilevel attention graph convolutional neural network (called DMN)
as well as the development of interpretable machine learning models with generalization tests and transferability tests.

Materials and Methods

**Dataset Construction.** There are several public databases, such as, Star,22 PhysProp,37 Martel,38 Huuskonen,39 ChEMBL40 and so on. To compare the performance of different algorithms, a standard benchmark, MoleculeNet41, which curates multiple public datasets, was built. However, to the best of our knowledge, the dataset for organofluorine is absent. Here, we built a database targeted at organofluorine molecules and named it as OFL dataset. Among the OFL dataset (Table S1), different conformations were taken into considerations and the computational details was shown in Section S1 (Figure S1). The whole dataset consists of four subsets (OFL-I, OFL-II, OFL-III, and OFL-IV) with partition coefficients determined by experiments or DFT calculations (Figure 1b). Besides, four non-fluorined molecules datasets (nonF-datasets), including nonF-SAMPL6, nonF-Nonstar, nonF-Aol and nonF-ChEMBL data set, were selected to test the transferability of the PoLogP model.

The first subset, OFL-I, is composed of 1542 organofluorine molecules collected from the above mentioned 5 public databases (Table S2). The selected F-containing molecules mainly contain C, H, O, N, F, S, Cl, P, and Br elements with wide distributions of the number of total atoms ($N_a = 5 \sim 128$), rotatable bonds ($N_b = 0 \sim 21$) and partition coefficient that measured experimentally ($\log P_{o/w}^{exp} = -1.38 \sim 6.06$). The details could be found in Section S1 (Figure S2 and S3) of Supporting Information. About 53% of molecules in OFL-I fall in the desirable window of $\log P_{o/w}$ for oral drugs. The OFL-I dataset was utilized to carefully pick out key descriptors and develop the robust and effective quantitative structure property relationship models for the estimation of $\log P_{o/w}$ of organofluorine molecules.

The second subset, OFL-II, a virtual library, contains 350 conformations which are derived from 44 organofluorine molecules dominant of fluorinated alcohol and borate derivatives (Table S3 and S4). The $\log P_{o/w}$ values (ranging from -0.01 to 2.83) for OFL-II were calculated in silico using the n-octanol/water phase-transfer free energy42-44 (Section S1). In order to distinguish with the experimentally measured partition coefficient ($\log P_{o/w}^{exp}$), herein, we nominated the partition coefficient using the transfer free energy as $\log P_{o/w}^{DFT}$.

Two external test sets, namely, OFL-III (Table S5) and OFL-IV (Table S6), were built for the validation of the generalization of the PoLogP model. The OFL-III dataset consists of F-containing molecules from SAMPL645 challenge as well as marked drugs, e.g., 5-fluorouracil, Fleroxacin, PF-
07321332. The OFL-IV dataset contains 11 synthesized organofluorine molecules, whose log $P_{o/w}^{exp}$ data were measured in this work.

As for the non-fluorinated molecules, nonF-SAMPL6 (Table S7), is composed of 8 uncharged non-fluorinated drug-like compounds which were taken from SAMPL6 log $P$ dataset. The nonF-Nonstar contains 36 diverse molecules (log $P_{o/w} = -2 \sim 7$ log units) (Table S8) collected from Nonstar data set. The nonF-Aol consists of 40 uncharged alcohol derivatives (Table S9) collected from Star data set. The last one, nonF-ChEMBL, consists of 2542 uncharged molecules (Table S10) that are part of the ChEMBL data set and contain C, H, O, N, S, Cl, P, and Br elements. The log $P_{o/w}^{exp}$ value of nonF-ChEMBL falls in the range of -1.48 ~ 4.5 log units.

**Data Preprocessing and Model Building.** The OFL-IV data set was manually stratified as four groups (Table S11), and then 90% of data from each group was randomly selected and combined as training set, in which the hyper-parameter optimization and cross-validation were performed. The remaining 10% data served as test set which is unseen throughout the development of ML models. All descriptors were pre-processed by scaling between 0 and 1, using the MinMax scaler protocol. The 10-fold cross-validation was performed on ML models based on different descriptors sets, using three evaluation metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Coefficient of determination ($R^2$) (Supporting Information, Section S2). Seven machine learning algorithms, including Extra Trees Regressor (ETR), Support Vector Regressor (SVR), Random Forest Regressor (RFR), Gradient Boosting Regressor (GBRT), Ridge, Decision Tree Regressor (DTR), and K-nearest Neighbours Regressor (KNR) were applied with the optimized hyper-parameters given in Supporting Information (Section S2, Table S12). All ML models were implemented with Python scripts using the scikit-learn package.

**Experimental Details.** Several methods for the determination of lipophilicity have been reported. They are classified as direct measurement methods, such as shake-flask method, and indirect measurement method, such as chromatographic method. The shake-flask method is a gold standard for the experimental determination of log $P_{o/w}$. The chromatographic method is widely used in high-throughput determination of log $P_{o/w}$, such as applied in the construction of the Martel benchmark dataset. It should be mentioned that a good linear correlation was found between the log $P_{o/w}$ gauged by chromatographic method and shake-flask method in 51 compounds including neutral, acidic, basic, amphoteric and zwitterionic drugs, with the $R^2$ of 0.90 (Section S3, Figure S14). The reversed-phase high performance liquid chromatography (RP-HPLC) method, which is one of the most widely used chromatographic method in the
measurement of lipophilicity, was used in this work (Section S3). The details of the general reaction procedure together with the characterizations could be found in Supporting Information (Section S4).

Results and Discussions

**Feature Selection and the Construction of PoLogP Model.** The controversial issue about whether or not the conformation has a great influence on the $\log P_{o/w}$ has aroused intensive interests. Taking the 3,3,3-trifluoropropan-1-ol molecule as an example, the lowest-energy conformation in $n$-octanol and in water is conformation 2 ($\log P_{DFT}^{o/w} = 1.23$ log unit) with an intramolecular organofluorine hydrogen bonds (C─F … H─O) and conformation 1 ($\log P_{DFT}^{o/w} = 1.19$ log unit) without intramolecular organofluorine hydrogen bonds, respectively. (Figure S4) In tandem with the different distribution of 3,3,3-trifluoropropan-1-ol conformations in $n$-octanol and water, the significant differences in lipophilicity between conformers demonstrate the importance of conformation sampling to establish ML models in the view of ensuring the quality of data. The important effect of conformation on the accuracy of lipophilicity prediction was further substantiated by the significantly better performance of ML models developed based on descriptors calculated on the lowest-energy conformations of the OFL-I dataset than those calculated on the conformers without conformation sampling (Figure S5), consistent with Muehlbacher et al’s work and our recent work. Two-step conformation samplings were performed for OFL-I, OFL-II and OFL-III with different theoretical levels (Figure 1b and Section S1). As trade-off between the cost and accuracy of data, the lowest-energy conformations of OFL-I and OFL-III subsets as well as the low-energy conformers of OFL-II subset were calculated by conformation sampling (see Section S1, Figure S6) in conformational space composed of 3,180,000 conformers and eventually generated OFL dataset. For filtering conformations, thresholds including the deviation of energy and root-mean-square deviation (RMSD) were applied. Similarly, these thresholds for filtering conformations of energetic significance could be replaced by the scoring function utilized in docking or the pair distance between any two specified conformations. Although ignoring conformers would sacrifice the accuracy of ML models to some extent. To save computational cost, all calculations for non-fluorinated molecules in nonF-datasets were performed on randomly constructed conformers instead of using the representative conformer from conformer sampling.

Two polarity descriptors, i.e., $\alpha$ and MPI, together with three HBs descriptors ($N_{HB}^{D}, N_{HB}^{A}, N_{HB \cdot F}^{A}$) were selected from gradually feature elimination from the initial 10-descriptor scheme by
the correlation analysis and the feature importance analysis (details in Section S2, Table S13, Figure S7 and S8).

According to the definition of the partition coefficient, the concentration ratio between n-octanol/water phases could be reflected by the transfer free energy (differences between solvation free energy of water ($\Delta G_w$) and n-octanol ($\Delta G_o$)). Intriguingly, the solvation free energy ($R^2 = 0.97$, Figure 2a) and DFT descriptors (Pearson’s $r > 0.99$, Figure S9) in n-octanol highly correlates well with those in water. The organofluorine molecules in polar protic solvent are more susceptible than those in apolar solvent, as shown by the negative values of $\Delta \alpha_{(o-w)}$ in Figure 2b. Such a significant difference of polarization may give rise to the more uneven charge distribution in water than that in n-octanol. Taking the conformation 1 of the 3,3,3-trifluoropropan-1-ol molecule as an example, the Mulliken charge of O atom ($q_O$) and H atom ($q_H$) of hydroxyl(-OH) is significantly higher in water ($q_O = -0.475 \text{ e}$ and $q_H = +0.286 \text{ e}$) than that in n-octanol ($q_O = -0.450 \text{ e}$ and $q_H = +0.269 \text{ e}$), implying a stronger tendency to form intermolecular hydrogen bonds in water than in n-octanol (Table S14).

Figure 2. Polarization effect between two distinct phases. (a) Correlation between the solvation free energy calculated in water ($\Delta G_w$) and n-octanol ($\Delta G_o$); (b) The distribution of the difference of polarizability, $\Delta \alpha_{(o-w)}$, between n-octanol and water.

The polarity gauged by overall molecular dipole moment ($\mu$)$^{29}$ was applied to describe the change of lipophilicity for various compounds (including 3-substituted indole derivatives, $\alpha$-difluoromethylated ethers, etc.) induced by fluorination,$^{28, 63}$ sometimes, in conjunction with molecular size (i.e. hydrophobic surface$^{29}$ or volume$^{64}$) and/or the HB-acidity$^{65}$ determined by $^1$H NMR analysis proposed by Abraham.$^{66}$ Different from those schemes, the present polarity descriptors of $\alpha$ and MPI, performed better than $\mu$ alone for evaluating the polarity of
organofluorine molecules. Thus, we applied $\alpha$, MPI and HB features to qualitatively estimate the change of lipophilicity induced by fluorination.

The distinct effects of $\alpha$ and MPI on lipophilicity are revealed by 1542 samples in Figure S10, in which the remarkable lipophilicity with high log $P_{\text{o/w}}$ values (marked by blue color) is featured by the large $\alpha$ value and small MPI values. Taking the fluorinated alkanols derivatives as an example, the significant higher lipophilicity of heptafluorinated butanol (L1, log $P_{\text{o/w}}^{\text{exp}}$ = 1.98) than that (L2, log $P_{\text{o/w}}^{\text{exp}}$ = 1.30) of pentfluorinated butanol could be ascribed to the simultaneous increase in $\alpha$ ($\Delta \alpha$ = +0.002, which favors the phase of $n$-octanol) and the decrease in MPI ($\Delta$MPI = -0.086, lessening the interaction with the phase of water for L1), as shown in Figure 3a. On the contrary, the decrease in lipophilicity for L3 (log $P_{\text{o/w}}^{\text{exp}}$ = 0.87) than that of L2 attributes to the decrease in $\alpha$ ($\Delta \alpha$ = -0.007) and the increase in MPI ($\Delta$MPI = +0.032). The similar results were found in fluorinated 5-propyloxyindoles derivatives (Figure 3b). All these results indicate that the competition between $\alpha$ and MPI would be a qualitative rule in evaluating the change of lipophilicity caused by fluorination. It is still worth noting that there is no simple linear relationship between $\alpha$ and MPI, as mentioned in the feature selection section (Section S2). Therefore, it is inadequate to simply measure the influence of fluorination on the lipophilicity by the change of relative magnitudes in $\alpha$ and MPI, in the case where the $\alpha$ and MPI simultaneous decrease or increase caused by fluorination. An example was shown in Figure 3c. For the local anesthetics (fluorinated ropivacaine (Rp1, Rp3) and fluorinated levobupivacaine (Rp2)), both decreases in $\alpha$ and MPI result in a higher lipophilicity. Interestingly, the dipole moment ($\mu$) commonly used in literature is insufficient to characterize polarity of organofluorine molecules. The increase in lipophilicity of Rp1 (log $P_{\text{o/w}}^{\text{exp}}$ = 3.0) was not consistent with its significant decrease in polarizability ($\Delta \alpha$ = -0.035) and increase in $\mu$ ($\Delta \mu$ = +0.154) compared to that (log $P_{\text{o/w}}^{\text{exp}}$ = 2.80) of Rp2 (Figure 3c).
Figure 3. The normalized values for the $\alpha$, MPI and $\mu$ estimated in the phase of water as well as the experimentally determined partition coefficient ($\log P_{\text{exp}}^{\text{o/w}}$) for fluorinated derivatives of (a) butanol, (b) 5-propyloxyindoles, (c) ropivacaine and levobupivacaine.
Seven machine learning algorithms (ETR, SVR, RFR, GBRT, Ridge, DTR, and KNR) were applied to the prediction of $\log P_{\text{exp}}^\text{o/w}$ of organofluorine molecules based on $\alpha$ and $\text{MPI}$ descriptors in water and HBS descriptors ($\text{N}_{\text{HB}}^D, \text{N}_{\text{HB}}^A, \text{N}_{\text{HB}^{-F}}^A$) calculated on the lowest-energy conformers. Among 7 machine learning algorithms, ETR gives the best performance with the $R^2$ of 0.80 and RMSE of 0.540 on the cross validation set (Figure S11, Section S2). The performance of above ETR model is comparable with that based on the full descriptor set and better than that using dipole moment ($\alpha + \mu + \text{HBs}$) for evaluating the polarity of organofluorine molecules (see Section S2).

**Performance of PoLogP Model.** The performance was further accessed over the test set of OFL-I. The PoLogP performed well against the popular academic tools (XLOGP3, MolLogP, and JPlogP-Coeff) with the RMSE of 0.473 log units and $R^2$ of 0.85, respectively (Table 1). The data with the absolute error no more than 0.5 log unit accounts for 71 % among the 157 test samples (Figure S12).

**Table 1.** The performance of PoLogP, XLOGP3, MolLogP, and JPlogP-Coeff over the test set of OFL-I dataset, using descriptors of $\alpha$, $\text{MPI}$ and HBS descriptors calculated by DFT ($\alpha_{\text{DFT}}$ and $\text{MPI}_{\text{DFT}}$) or DeepMoleNet ($\alpha_{\text{DMN}}$ and $\text{MPI}_{\text{DMN}}$, results shown in parentheses)

<table>
<thead>
<tr>
<th></th>
<th>PoLogP</th>
<th>XLOGP3</th>
<th>MolLogP</th>
<th>JPlogP-Coeff</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MAE</strong></td>
<td>0.375</td>
<td>0.643</td>
<td>0.760</td>
<td>0.753</td>
</tr>
<tr>
<td></td>
<td>(0.390)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>RMSE</strong></td>
<td>0.473</td>
<td>0.990</td>
<td>1.070</td>
<td>1.066</td>
</tr>
<tr>
<td></td>
<td>(0.497)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>$R^2$</strong></td>
<td>0.85</td>
<td>0.34</td>
<td>0.23</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>(0.83)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The PoLogP method was further tested on the F-containing molecule from SAMPL6 log $P$ challenge (SM02) as well as marketed drugs (5-fluorouracil and Fleroxacin), with the smallest absolute error against XLOGP3, MolLogP and JPlogP-Coeff method (Figure 4). Finally, PoLogP was successfully extended to the lipophilicity prediction of COVID-19 oral antiviral drug, PF-07321332, with the value of 2.42 log units, falling in the desirable window 1~3 log units of oral drugs.
Figure 4. The partition coefficient measured by experiment as well as predicted via PoLogP, XLOGP3, MolLogP and JPlogP-Coeff, respectively, for the SM02 molecules, 5-fluorouracil, Fleroxacin and PAXLOVID™ (PF-07321332).

A multilevel attention graph convolutional neural network (DeepMoleNet, called DMN in short) developed by our group was utilized here so as to reduce the time cost for the calculation of polarity descriptors ($\alpha$ and $MPI$) of organofluorine molecules (see Section S2). The graph convolutional neural network automatically learns features from the molecular graphs and could simultaneously predict the 12 properties related to electronic structure, including polarizability, dipole moment, HOMO (highest occupied molecular orbital), LUMO (lowest unoccupied molecular orbital), etc. The detailed hyper-parameters of DeepMoleNet could be found in Table S15. Here, the OFL-I dataset was split into train set, validation set and test set with the ratio of 8:1:1. The predicted values for polarizability ($\alpha^{DMN}$) as well as molecular polarity index ($MPI^{DMN}$) via DeepMoleNet correlated well with the results from direct DFT calculations, with the $R^2$ of 1.00 and 0.77, $MAE$ of 0.10 a.u./atom and 1.31 kcal/mol (Figure 5a, 5b), respectively. With these descriptors derived from DeepMoleNet together with the HBs features, the excellent performance of PoLogP against the test set of OFL-I was retained (Table 1).

The ability of PoLogP using descriptors ($\alpha^{DMN}$ and $MPI^{DMN}$) derived from multilevel attention graph convolutional neural network was further validated on 11 synthesized fluorine-containing
molecules (OFL-IV). As shown in Figure 5c, the values predicted by PoLogP$^{DMN}$ correlated well with the values measured experimentally ($R^2 = 0.74$ and RMSE = 0.375). It should be mentioned that armed with graph convolutional neural network, the above prediction is fast since no DFT calculation was required. Interestingly, our method could well capture the trend observed by experiment. For example, when the methyl group of OFL-IV-2 was replaced by methoxy group (OFL-IV-3), the partition coefficient increased. The increase in lipophilicity of OFL-IV-3 is consistent with the simultaneous increase of $\alpha^{DMN}$ and the decrease of $MPI^{DMN}$ than those in OFL-IV-2 (Table S6).
Figure 5. (a) Correlation of polarizability ($\alpha$) and (b) molecular polarity index ($MPI$) calculated by DFT vs predicted by graph convolutional neural network ($DMN$), with $MAE$ for $\alpha$ in a.u./atom as well as for $MPI$ in kcal/mol, respectively. (c) The comparison between the partition coefficient estimated by PoLogP using $\alpha^{DMN}$, $MPI^{DMN}$ and HBs descriptors ($log P^{DMN}_{o/w}$) and values determined
experimentally (log $P_{o/w}^{exp}$) over 11 synthesized fluorine-containing molecules (OFL-IV). Together shown are their corresponding chemical structures.

**Transferability of PoLogP Model.** The transferability of PoLogP was tested over four non-fluorined molecules datasets i.e., nonF-SAMPL6, nonF-Nonstar, nonF-Aol, and nonF-ChEMBL data sets (Tables S16-19). For the smallest nonF-datasets, nonF-SAMPL6, the PoLogP performed well and was comparable with those predicted by XLOGP3, MolLogP and JPlogP-Coeff method (Table S16). For the nonF-Nonstar data set, which is far beyond the scope of our training set, the RMSE of PoLogP is slightly higher than those obtained by MolLogP and JPlogP-Coeff method (Table S17). The present PoLogP armed with $\alpha_{DFT}$ and $MPI_{DFT}$ could be transferred to the prediction of log $P_{o/w}$ for nonF-Aol dataset including 40 non-fluorinated alcohol derivatives with reasonable and comparable accuracy against the popular academic tools, with the RMSE of 0.424 log units and $R^2$ of 0.79, respectively (Figure 6, Table S18). Furthermore, the graph convolutional neural network model was successfully transferred to the fast generation of $\alpha$ and $MPI$ for the non-fluorinated molecules with DFT accuracy, evaluated by the $R^2$ of 1.00 and 0.75, $MAE$ of 0.144 a.u./atom and 1.476 kcal/mol (Figure S13), respectively. The accuracy of PoLogP was retained using HBs, $\alpha_{DMN}$ and $MPI_{DMN}$ descriptors (Table S18), which indicates that the HBs descriptors and polarity characters are also important factors affecting the log $P_{o/w}$ for non-fluorinated molecules. Moreover, the test on a larger dataset, nonF-ChEMBL, which is part of ChEMBL benchmark dataset, shows satisfactory accuracy for PoLogP based on both schemes, PoLogP$_{DFT}$ and PoLogP$_{DMN}$, with RMSE of 0.754 and 0.766 (Table S19), respectively. The satisfactory transferability of $\alpha$, $MPI$ and HBs descriptors as well as the graph convolutional neural network model highlights the potential usage of PoLogP in high-throughput screening of the functional molecules with and without F-substitution.
Figure 6. The correlation of partition coefficient determined experimentally ($\log P_{o/w}^{exp}$) vs estimated by PoLogP armed with $\alpha^{DFT}$, $MPI^{DFT}$ and HBs descriptors over 40 alcohol derivatives (nonF-Aol data set).

Conclusion

A dataset named OFL was built for the partition coefficient prediction of organofluorine molecules. An interpretable machine learning model (PoLogP) armed with descriptors of $\alpha$, $MPI$ and HBs is developed to efficiently predict the values of lipophilicity for organofluorine molecules. The good performance of PoLogP was also shown in 11 synthesized molecules (OFL-IV) and the lipophilicity prediction of drug and 2626 non-fluorinated molecules. With the aid of a multilevel attention graph convolutional neural network model, fast generation of theoretical descriptors was achieved with the DFT accuracy but without time consuming DFT calculations. The organofluorine molecule dataset and the predictive ML model are useful to accelerate the fluoro-pharmaceutical discovery and design of new materials, with and without F atom(s).

Data and Software Availability
All data were collected into a database, which is open for all academic usages at http://106.15.196.160:5662/ upon the request for a license.

Associated Content

Supporting Information: DFT datasets; computational, machine learning and deep learning details, including the creation of training and testing sets, model evaluation metrics, hyper-parameters, feature selection and the performance of interpretable ML models; experimental details, materials, and methods, including the $^1$H NMR, $^{13}$C NMR, $^{19}$F NMR and HPLC spectra

Author Information

Corresponding Author

Jing Ma – Key Laboratory of Mesoscopic Chemistry of Ministry of Education, School of Chemistry and Chemical Engineering, Nanjing University, Nanjing 210023, P. R. China; orcid.org/0000-0001-5848-9775; Email: majing@nju.edu.cn

Yi Wang – Jiangsu Key Laboratory of Advanced Organic Materials, School of Chemistry and Chemical Engineering, Nanjing University, Nanjing 210023, P. R. China; Email: yiwang@nju.edu.cn

Qiang Zhu – Key Laboratory of Mesoscopic Chemistry of Ministry of Education, School of Chemistry and Chemical Engineering, Nanjing University, Nanjing 210023, P. R. China; orcid.org/0000-0002-5612-0728; Email: csu1505110121@163.com

Authors

Qingqing Jia – Key Laboratory of Mesoscopic Chemistry of Ministry of Education, School of Chemistry and Chemical Engineering, Nanjing University, Nanjing 210023, P. R. China

Yifan Ni – Jiangsu Key Laboratory of Advanced Organic Materials, School of Chemistry and Chemical Engineering, Nanjing University, Nanjing 210023, P. R. China

Ziteng Liu – Key Laboratory of Mesoscopic Chemistry of Ministry of Education, School of Chemistry and Chemical Engineering, Nanjing University, Nanjing 210023, P. R. China
Xu Gu – Key Laboratory of Mesoscopic Chemistry of Ministry of Education, School of Chemistry and Chemical Engineering, Nanjing University, Nanjing 210023, P. R. China

Ziyi Cui – Key Laboratory of Mesoscopic Chemistry of Ministry of Education, School of Chemistry and Chemical Engineering, Nanjing University, Nanjing 210023, P. R. China

Mengting Fan – Key Laboratory of Mesoscopic Chemistry of Ministry of Education, School of Chemistry and Chemical Engineering, Nanjing University, Nanjing 210023, P. R. China

Notes

The authors declare no competing financial interest.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (2019YFC0408303), and the National Natural Science Foundation of China (grant nos. 22033004, 21873045). We are grateful to the High Performance Computing Centre of Nanjing University for providing the IBM Blade cluster system.
Reference


