Journal Name

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

Molecular Partition Coefficient from Machine Learning with Polarization and Entropy Embedded Atom-Centered Symmetry Functions†

Qiang Zhu,[a] Qingqing Jia,[a] Ziteng Liu,[a] Yang Ge,[a] Xu Gu,[a] Ziyi Cui,[a] Mengting Fan,[a] and Jing Ma*[a]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

Efficient prediction of the partition coefficient ($\log P$) between polar and non-polar phases could shorten the cycle of drug and materials design. In this work, a descriptor, named $\langle q-ACSFs\rangle_{conf}$, is proposed to take the explicit polarization effects in polar phase and conformation ensemble of energetic and entropic significance in non-polar into considerations. The polarization effects are involved by embedding the partial charge directly derived from force fields or quantum chemistry calculations into the atom-centered symmetry functions (ACSFs), together with the entropy effects which are averaged according to Boltzmann distribution of different conformations taken from similarity matrix. The model was trained with the high-dimensional neural networks (HDNNs) on a public dataset PhysProp (with 41039 samples). Satisfactory $\log P$ prediction performance was achieved on three other datasets, namely, Martel (707 molecules), Star & Non-Star (266) and Huuskonen (1870). The present $\langle q-ACSFs\rangle_{conf}$ model was also applicable to the n-carboxylic acid with the number of carbon ranging from 2 to 14 and the 54 kinds of organic solvents. It is easy to apply the present method to arbitrary sized systems and give a transferable atom-based partition coefficient.

1 Introduction

The partition coefficient (P) is an important parameter which represents the ratio of the solubility between polar and apolar phases, such as the water and n-octanol. (see Figure 1) The logarithm of partition coefficient, i.e., $\log P$, is usually taken as an indicator for screening out promising drug and material candidates in environmental science [1] and pharmacology.[2,3] Among some theoretical models, the $\log P$ parameter is also associated with other molecular properties, such as the aqueous solubility ($\log S$),[4–6] the distribution coefficient ($\log D$),[7,8] and Lipophilic Efficiency (LiPE).[9] It was demonstrated that entropy may contribute to significant changes in the solubility of the nanocrystal-ligands complexes.[10,11] In addition, in the polar phase, the solute polarizability[12] and polarity[13] are much more sensitive to the partition coefficient. The delicate balance between the entropy and the polarity in transporting small drug from phase of water to lipid was also revealed by our molecular dynamics (MD) simulations with both implicit and explicit polarization models .[14] However, the time-consuming MD simulations of the solvation equilibrium in both polar and non-polar phases are impossible to realize the high-throughput screening of promising drugs and material candidates. Thus, an efficient model for $\log P$ prediction is highly desired to take the explicit polarization and entropy into considerations.

Here, we proposed a descriptor encoding the polarization and conformation entropy into the atom-centered symmetry functions (ACSFs), named $\langle q-ACSFs\rangle_{conf}$, and married it with a high-dimensional neural network (HDNN), as shown in Figure 1. (A detailed structure of HDNN could be found in supporting information) Comparing with the atomic-based or fragment-based model[15–22] and molecular descriptors based model,[13,23–30] our model features in the following three aspects: (i) bypassing the laborious jobs in dividing the whole molecules into separate fragments or atoms and constructing descriptors automatically; (ii) no need for the various descriptors which may be difficult and computationally costly to obtain; (iii) adding physically explain-

[a] *Key Laboratory of Mesoscopic Chemistry of Ministry of Education Institute of Theoretical and Computational Chemistry School of Chemistry and Chemical Engineering, Nanjing University, Nanjing, 210023, P. R. China Fax: 86-25-89681772; Tel: 86-25-89681772; E-mail: majing@nju.edu.cn*

† Electronic Supplementary Information (ESI) available: [Additional details in collected datasets, generation of descriptors, computational methods of Molecular Dynamics (MD) simulations and Quantum Mechanisms (QM), hyper-parameter optimization of high-dimensional neural networks, and contribution from distinct elements with different environments]. See DOI: 00.0000/00000000.

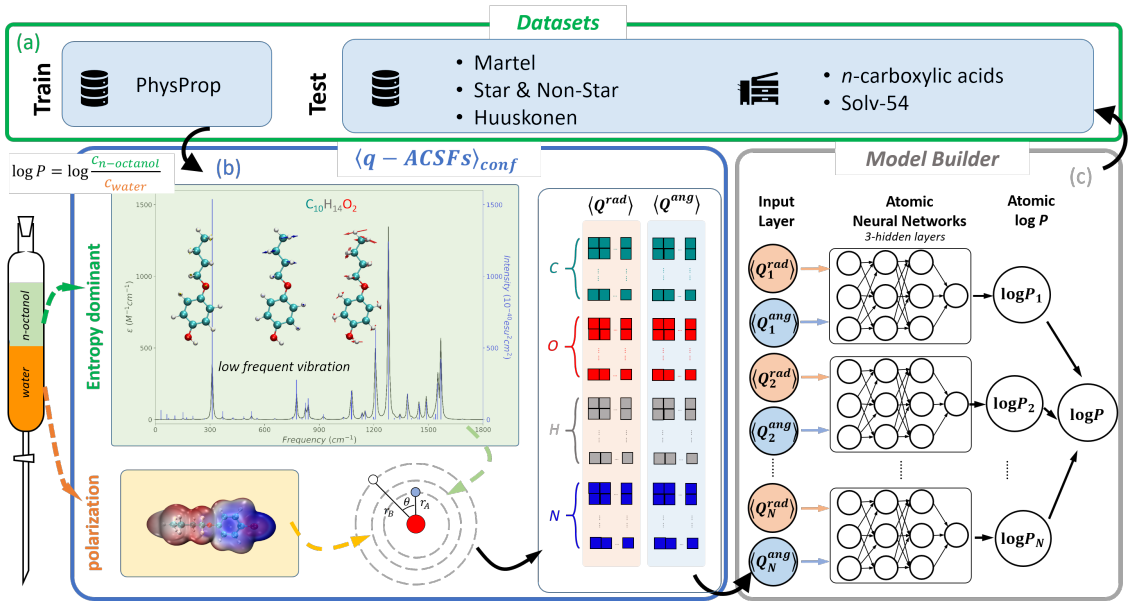

Fig. 1 Schematic illustration of the workflow applied for (a) the collection of datasets, (b) derivation of the polarization and entropy, encoding polarization and entropy into atom-centered symmetry functions ($\langle q-ACSFs\rangle_{conf}$), (c) model building and partition coefficient prediction.

able elements by explicitly taking the polarization effects and conformation entropy into consideration.

2 Methods

Four public datasets (PhysProp,[31] Martel,[32] Star & NonStar,[23] and Huuskonen[33]) were used for training and testing the HDNN model with $\langle q-ACSFs\rangle_{conf}$, as shown in Figure 1 and Figure S1. Two homemade datasets (*n*-carboxylic acids and Solv-54[34]) were adopted as an external test (Table S2). PhysProp database is perhaps the biggest public collection of the experimental $\log P$ data. It contains about 41039 molecular structures as SMILES strings in total, 13553 of which were determined experimentally and the rests were estimated. In Martel, 707 commercial molecules were measured experimentally with high performance liquid chromatography (HPLC) method. Star & Non-Star database is composed of 266 molecules, 223 of which were picked from BioByte StarList[35] for the development of various $\log P$ estimation methods. The other 43 molecules were collected outside of BioByte StarList. Huuskonen database contains a diverse set of 1870 organic molecules.

In the present work, we studied $\log P$ of organic molecules of top four abundance of chemical elements, namely, H, C, N, O. As shown in Figure S1(a), the normalized abundance of H, C, N, O is quit similar and the abundance of element H is the highest. The $\log P$ distribution of the six public databases ranges from -2 to 7, as shown in Figure S1(b).

To show the entropic effect on partition coefficient measured experimentally ($\log P_{exp}$), we fetched experimentally measured entropy and partition coefficient of 14 molecules that commonly utilized as the solvents and ligands coated with CdSe nanocrystals[10,11] (shown in Figure 2(a)). It is interesting to find a relationship between the experimental partition coefficient ($\log P_{exp}$) and the experimentally measured entropy (Figure 2(b)) and calculated entropy by the density functional theory (DFT) at the level of b3lyp/6-31g(d) (Figure 2(c,d)). A closer look at the individual contribution to entropy gives a conclusion that contribution from vibration increased a lot when partition coefficient increases (Figure 2(d)). It is necessary to introduce the entropy effects from the low frequency vibrations into $\log P$ prediction. The conformations of entropic significance were sampled from molecular dynamics (MD) simulations, the simulation details could be found in supporting information. Subsequently, we encoded conformation entropy and polarization into the conventional atom-centered symmetry functions (ACSFs). Starting from ACSFs (see eq. S1 - eq. S3), the resulting radial and angular symmetry functions in polarization weighted ACSFs ($q-ACSFs$) are expressed in eq. 1 and eq. 2, respectively.

$$Q_i^{rad} = \sum_{j\neq i}^{N} g(q_j) e^{-\eta(R_{ij}-R_s)^2} f_c(R_{ij}) \tag{1}$$

$$Q_i^{ang} = 2^{1-\zeta} \sum_{j,k\neq i}^{all} h(q_j,q_k)(1+\lambda\cos(\theta_{ijk}))^{\zeta} \times e^{-\eta(R_{ij}^2+R_{ik}^2+R_{jk}^2)} \times f_c(R_{ij}) f_c(R_{ik}) f_c(R_{jk}) \tag{2}$$

$g(q_j)$ and $h(q_j,q_k)$ are two weighting functions, where both of them are functions of atomic charge (q) of atom j and k. Although the weighting function could take various different definitions, we took the following forms in this work.

$$g(q_j) = q_j \tag{3}$$

$$h(q_j,q_k) = q_j q_k \tag{4}$$

When considering the entropy effects, the atom-centered sym-

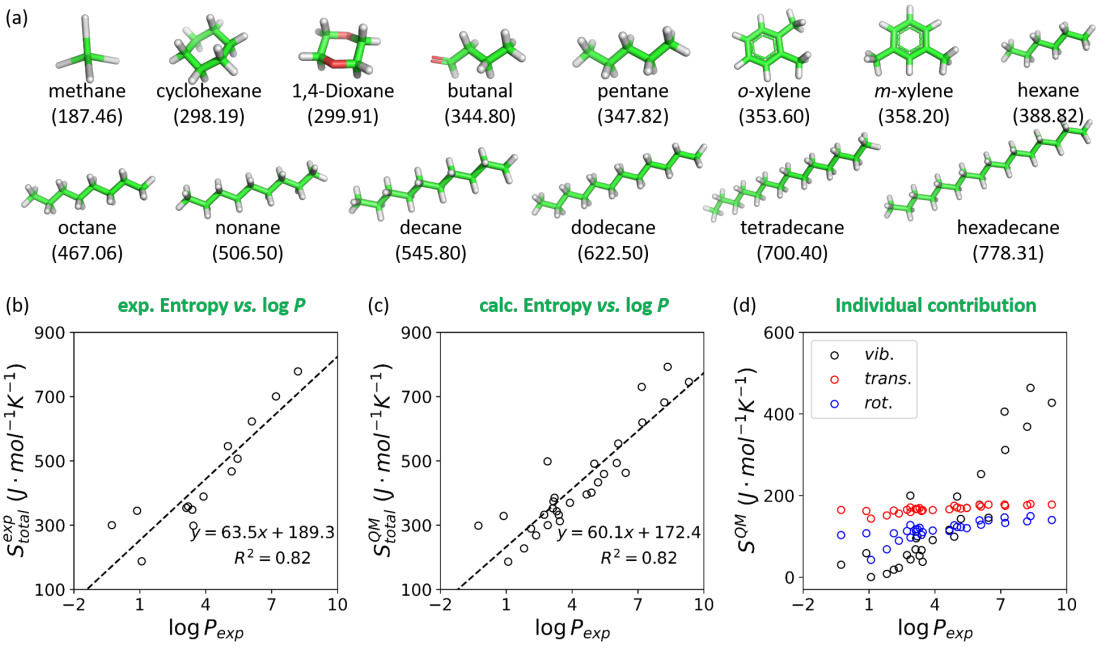

Fig. 2 (a) The selected molecules for building the correlation between the partition coefficient $\log P$ measured experimentally and the entropy (S) extracted from experiments (b), the total entropy (S^{QM}_{total}) (c) calculated by quantum mechanism (QM), and individual entropy (d) came from three distinct contributions namely, vibrational motion (S^{QM}_{vib}), translation motion (S^{QM}_{trans}), and rotational motion (S^{QM}_{rot}) colored in black, red and blue, respectively. Their corresponding experimental entropy were shown in parentheses in the unit of $J \cdot mol^{-1}K^{-1}$. More details could be found in Table S3.

metry functions could be denoted as $\langle q-ACSFs\rangle_{conf}$, where the radial and angular symmetry functions were expressed as below:

$$\langle Q_i^{rad}\rangle = \sum_{a=1}^{N} p_a\{Q_i^{rad}\} \tag{5}$$

$$\langle Q_i^{ang}\rangle = \sum_{a=1}^{N} p_a\{Q_i^{ang}\} \tag{6}$$

where N is the number of conformations we selected, p_a is the Boltzmann distribution probability that conformation a could appear.

With the help of the high-dimensional neural network, the total $\log P$ is the summation over the i-th individual atom, $\log P_i$, and the mathematical form could be expressed as below:

$$\log P = \sum_{i}^{N_{atoms}} \log P_i \tag{7}$$

where N_{atoms} is the total atom numbers of a molecule.

The individual contribution ($\log P_i$) was derived from an atomic neural network, depending on the local chemical environments surrounding the i-th atom with two sets of symmetry functions, $\langle Q_i^{rad}\rangle$ and $\langle Q_i^{ang}\rangle$. As shown in Figure 1, artificial neural network is composed of 3 parts, namely, input layer, hidden layer and output layer. There could be one or more hidden layer in a single network, and the mathematical flexibility between the input and output increased when more hidden layers and more nodes were applied in each hidden layer. A more detailed example and mathematical expression could be found in supporting information (Figure S4).

3 Results and Discussion

To demonstrate the importance of the polarization effects, we firstly generated 100 simple descriptors with RDKit,[36] and evaluated the contributions of each descriptors to the prediction of partition coefficient with two distinct methods, namely, univariate feature selection and mean decrease in impurity (MDI). The full list of the descriptors and details of two feature selection methods could be found in supporting information (Table S1). In Figure S3 (a), top 20 ranked descriptors utilizing the MDI were presented and the importance of single descriptor is reflected by the percentage ratio. We further classified the descriptors into charge related and non-charge related ones. From it, we could draw a conclusion that the partition coefficient are highly related to the charge, as the electrostatic or polarization related descriptors account for 74 % among the top 20 descriptors, and the most important descriptor is $PEOE-VSA_6$, which reflects the direct electrostatic interactions hybrid with the surface area. Same picture was also drawn with the help of univariate feature selection (Figure S3 (b)).

To visualize the effects on the introduction of partial charge into the atom-centered symmetry functions (Figure 3(a)), here, we took the water molecule as an example and utilized Gasteiger partial charge.[38] Firstly, we simply mapped the partial charge into a water molecule, as shown in Figure 3(b), consistent with our chemical intuition, atom O possesses negative values (blue region) and atom H of positive values (red region). In addition, the density around atom O is much more dense than atom H, as

Table 1 Performance of different $\log P$ methods over three datasets

	Martel			Star&Non-Star			Huuskonen		
	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
XLOGP3	0.97	1.60	1.26	0.45	0.36	0.60	0.32	0.20	0.45
MolLogP	1.06	1.93	1.39	0.56	0.46	0.68	0.46	0.36	0.60
ALOGPS 2.1	1.02	1.68	1.30	0.41	0.32	0.56	0.31	0.26	0.51
JPlogP-Coeff	0.93	1.49	1.22	0.57	0.51	0.72	0.40	0.29	0.54
JPlogP-library[a]	0.90	1.42	1.19	-	-	-	-	-	-
$\langle ACSFs \rangle_{conf}$	0.97	1.66	1.29	0.83	1.27	1.13	0.54	0.53	0.73
$ACSFs^{max}$	0.96	1.60	1.27	0.82	1.27	1.13	0.54	0.53	0.73
$\langle q-ACSFs \rangle_{conf}$	0.91	1.50	1.23	0.48	0.44	0.66	0.22	0.12	0.35
$q-ACSFs^{max}$	0.90	1.53	1.23	0.54	0.54	0.74	0.22	0.13	0.37

[a] Results derived from ref,[37] where the estimation shown here was performed over molecules containing element C, H, O and N.

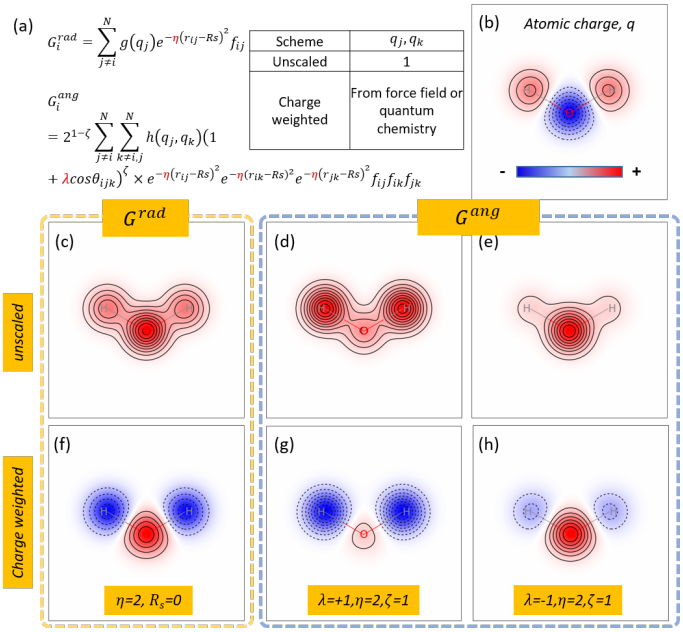

Fig. 3 Importance of polarization effects. (a)Mathematical expression of radial (G^{rad}) and angular related (G^{ang}) descriptors ; (b) Distribution of Gasteiger partial charge mapped onto atom O and H; (c) Radial (c.f., eq. 1) and (d)-(e) angular (c.f., eq. 2) symmetry functions of water with atomic charges all set to be 1; (f) Radial and (g)-(h) radial and (h)-(i) angular symmetry functions of water scaled by each atomic charge q_i. The second and third columns differ in the sign of the phase parameter λ. Blue denotes the negative value and red denotes the positive value.

the absolute partial charge on the atom O is twice larger than atom H. When we ignored the polarization effect (c.f., eq. 3 and eq. 4 where parameter q was set to be 1 for all atoms), the descriptors drawn for atom O and H could not be well separated and the sign of charge information was totally lost, as shown in Figure 3(c) - Figure 3(e). For the angular symmetry functions, with the phase parameter λ switched between $+1$ and -1, the maximum intensity of the descriptors shifted from atom H to O (Figure 3(d) and Figure 3(e)). It is a compensate, hence, we utilized both values to obtain good descriptors at different values of θ_{ijk}. However, when the partial charge was embedded in the generation of the descriptors, significant changes were observed both in the radial (Figure 3(f)) and angular (Figure 3(g) and Figure 3(h)) symmetry functions and signs of descriptors on atom O and H are all opposite. In addition, the value of descriptors on atom O are positive compared with the partial charge of atom O, this phenomenon is resulted from that the value of descriptor on atom O are summation of surrounding environments (Figure 3(a) and eq. 1 and eq. 2). Here, the atom H is embedded into the generation of descriptor on atom O and vice verse for the generation of descriptor for atom H. In Table 1, great improvement was observed with introduction of charge information over three public datasets.

To further disclose the effects of entropy, we proposed another descriptor which only takes the most probable structure into consideration and called it $q-ACSFs^{max}$. As shown in Figure 4, we presented the relationship between the number of rotatable bonds and root-mean-square deviation (RMSD) of trajectories generated from molecular dynamics simulations, from which high correlation was observed. A molecule with the highest number of rotatable bonds among PhysProp was detailed. The conformations generated by MD were grouped into three clusters according to their structural similarity. Details could be found in supporting information. Each cluster is colored separately. From the distribution of the potential energy, we could see that three representative structures, namely, cross, parallel, and ring-like, (the dotted lines and the five point star) could almost cover the whole range. A subsequent principle component analysis also disclosed that 3 clusters may be sufficient as the top 3 principle components account for 65.3 % of the whole systems (Figure S2). The distribution of the number of rotatable bonds and RMSD for four distinct datasets was also drawn, a sharp peak was observed in the datasets PhysProp, Martel and Huuskonen (Figure 4 (a)-(c)). A somehow wide distribution (RMSD ranges from 0 to 0.5 and number of rotatable bonds ranges from 0 to 30) was observed in datasets Star & Non-Star. As a consequence, compared with model $q-ACSFs^{max}$, a distinct improvement of MAE and MSE was achieved over dataset Star & Non-Star using model $\langle q-ACSFs \rangle_{conf}$, while little differences were observed in other two datasets (Table 1). This phenomenon may contribute to the broad space sampled in dataset Star & Non-Star and point out the

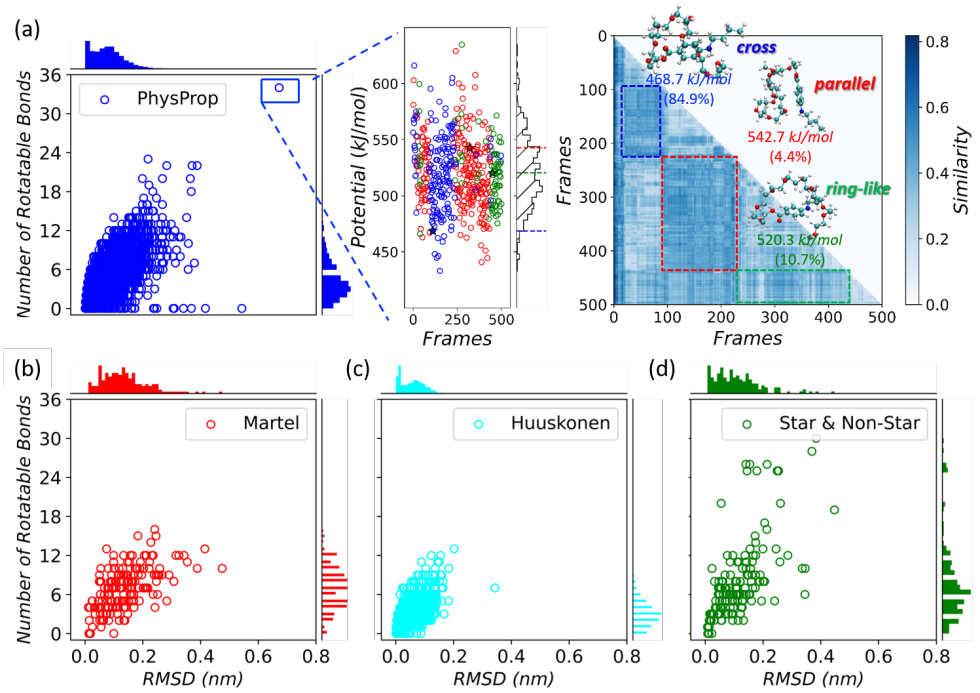

Fig. 4 Correlation between the number of rotatable bonds and the root-mean-square deviations (RMSD) derived from the 1 ns MD trajectories among four public datasets, namely, (a) PhysProp, (b) Martel, (c) Huuskonen, and (d) Star & Non-Star. Three clusters of molecule $C_{30}H_{53}O_{11}N_2$ were grouped from the molecular dynamics simulations with the help of the similarity map. Each cluster is represented in a separate color. Among the trajectories, their representative potentials were drawn in five-pointed star. The representative conformations, namely, cross (colored in blue), parallel (red) and ring-like (green) were shown in the upper right panel together with their potential and probability.

direction for further increasing the model accuracy.

Performance of $\langle q-ACSFs\rangle_{conf}$ was fully asssessed over three distinct datasets through MAE and MSE. Detailed definition of the criterion could be found in supporting information. Five methods, namely, XLOGP3[16], MolLogP[18], ALOGPS 2.1[19], JPlogP-Coeff[37] and JPlogP-library[37] were also applied for comparison. As listed in Table 1, the performances of predicting the partition coefficient in the datasets Martel were less satisfactory over these five methods, as the distribution of partition coefficient is quit different from the one we trained on (Figure S1(b)). However, among them, $\langle q-ACSFs\rangle_{conf}$ outperforms these methods except for JPlogP-library, as the MAE and MSE could decreased to 0.91 and 1.50, respectively. While for the method MolLogP, the MAE and MSE are up to 1.06 and 1.93. Among the datasets Star & Non-Star, the performance of $\langle q-ACSFs\rangle_{conf}$ is almost the same as the XLOGP3. The MAE and MSE predicted by $\langle q-ACSFs\rangle_{conf}$ are only 0.22 and 0.12 over datasets Huuskonen, which is a great improvement with considering both polarization and entropy effects. The present model highlights the advantage of non-need to pre-define types of atom or fragment and applicability to any type of molecules with appropriately formulated data.

To further survey the effect of partial charge and conformations of energetic and entropic significance, we trained different representations with the same procedure and test their performances on the same datasets. As shown in Table 1, introduction of partial charge into the ACSFs, great improvements could be achieved. For example, when scaling all partial charge to 1, the MAE (MSE) of $\langle ACSFs\rangle_{conf}$ is 0.97(1.66), 0.83(1.27) and 0.54(0.53) among datasets Martel, Star & Non-Star, and Huuskonen, respectively. However, when taking the charge effects, MAE (MSE) of the $\langle q-ACSFs\rangle_{conf}$ could be decreased to 0.91(1.50), 0.48(0.44), and 0.22(0.12) for datasets Martel, Star & Non-Star and Huuskonen, respectively. For datasets Star & Non-Star and Huuskonen, the decrease of MAE(MSE) from 0.83(1.27) and 0.54(0.53) to 0.48(0.44) and 0.22(0.12) was observed, respectively. In the term of conformation effects, we could see that the performance differs a little among all three datasets. The vanish of the improvement over dataset Star & Non-Star further demonstrates the significance of both polarization and entropy.

To get a more reliable and robust model, here, we trained it over four public datasets and test it over two homemade datasets, namely, n-carboxylic acids and Solv-54 (Table S2). As shown in Figure 5 (a), with the chain length increasing, more flexible are the molecules which may attribute to entropic significance. Our model could well reproduce the trends. In Solv-54, polar molecules such as the alcohols were highlighted and drawn in five-point stars. As shown in Figure 5 (b), more hydroxyl functional groups molecule contains, stronger interaction with phase of water was expected which results in lower partition coefficient. For example, molecule 3,6,9-trioxa-undecan-1,11-diol (exp: -2.02) and ethane-1,2-diol (exp: -1.36) which contain two polar hydroxyl groups possess much lower partition coefficient among datasets Solv-54. Although the diversity of the compounds, good prediction was achieved especially in the aspect of conformational entropy and polarity.

Benefit from the high-dimensional neural network (HDNN) and

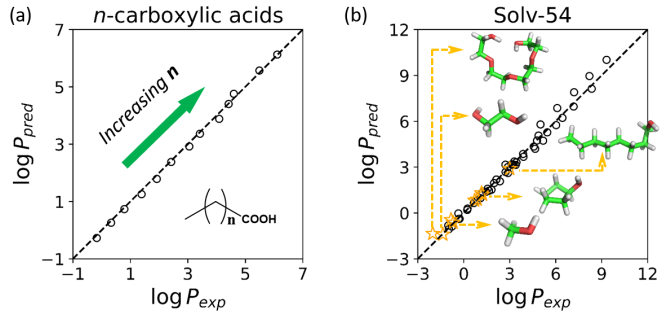

Fig. 5 Performance of the predicted partition coefficient ($\log P_{pred}$) by charge and ensemble weighted atom-center symmetry functions ($\langle q-ACSFs\rangle_{conf}$) over two homemade datasets, namely (a) *n*-carboxylic acids and (b) Solv-54, inserted are polar molecules such as alcohols indicated by the five-pointed star colored in orange.

atom-center symmetry functions (ACSFs), $\langle q-ACSFs\rangle_{conf}$ not only could be able to give a partition coefficient for a single molecule, but also has the ability for deriving the contribution from a single atom, as our input is only dependent on the element and the surrounding environments. As a consequence, we decomposed the partition coefficient of all the molecules into single atoms and analyzed over three datasets. As shown in Figure 6, the values of contribution from single atom C ranges from -0.2 to 0.7, and they mainly concentrated above 0 which is guided by the black dashed line. For atom O and atom N, the values are almost negative. Same conclusion could be drawn from datasets Martel and Huuskonen (middle and right panel of Figure 6). Intriguingly, this phenomenon is consistent with the simple model proposed by Mannhold where the partition coefficient is only related to the number of carbon atoms (NC) and number of hetero atoms ($NHET$) ($\log P = 1.46(\pm0.02)+0.11(\pm0.001)NC-0.11(\pm0.001)NHET$).[23] When digging deep into the contribution from the element H, we found that the peak of the distribution is in accordance with the dashed line at 0, which means in some situations, it increased the partition coefficient, while decreased in the resting situations. Further analyses on dividing the contribution of each atoms according to the surrounding environments could be found in Figure S5-S8 and Table S10. From Figure S5 and Table S10, we could see that atom H prefers to increase the partition coefficient when bonding with atom C independent of hybridization methods. To the contrary, when bonding with hetero atoms, atom H prefers to decrease the partition coefficient.

4 Conclusions and Perspectives

In summary, we designed a new descriptor based on the conventional atom-centered symmetry functions (ACSFs) and we called it $\langle q-ACSFs\rangle_{conf}$. Polarization and entropy effects were treated explicitly by introducing the partial charges derived directly from force field and conformations of energetic and entropic significance sampled from molecular dynamics simulations. Different from atom- or fragment-additive models, our model do not need to pre-define the types of atom or fragment and could bypass the pitfall of missing atoms or fragments. In addition, no prior knowledge was introduced in our model compared with

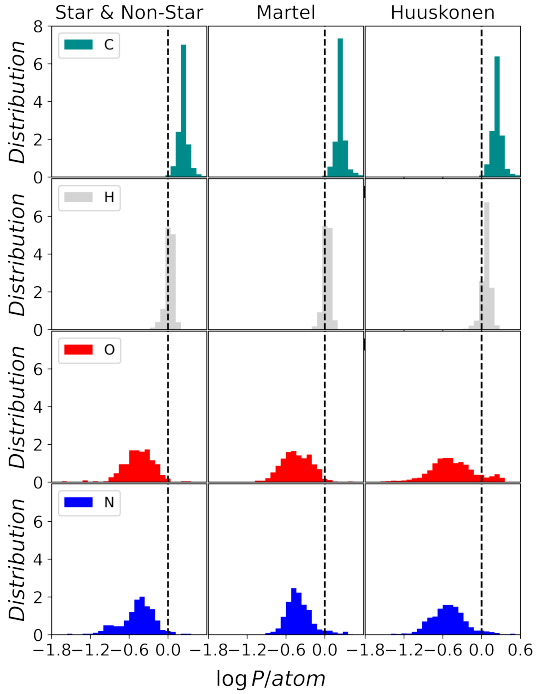

Fig. 6 Partition coefficient decomposed into contribution from each atom of a single molecule. Dashed guide line at 0 is shown here, where the contributions from left are to decrease the partition coefficient, right for increasing it. The colors are in line with the molecule throughout the manuscript.

some models which need pre-calculated molecular descriptors. We further tested the effects of polarization and entropy on model performance, results and feature selection showed that the polarization is important in the prediction of partition coefficient, and significant improvement could be achieved on dataset Star & Non-Star (MAE decreased from 0.83 to 0.48)and Huuskonen (MAE decreased from 0.54 to 0.22). Just a slight improvement was observed for the introduction of different conformations, which may attribute to the conformational space we sampled is insufficient. As a result, to better improve the accuracy of the model, much more emphasis should be laid on the technology of enhanced sampling methods, especially for the target-free methods, such as the metadynamics,[39] umbrella sampling,[40] DAta-Driven acceleration method[41,42] and so on. Some progress is still being made in our lab. In addition, inheriting the advantage of atom-centered symmetry functions, $\langle q-ACSFs\rangle_{conf}$ has the ability to decompose the partition coefficient into each atoms. Interestingly, when we statistically analyzed the contributions from four distinct elements, we found that the contribution from element C are almost positive, while negative for the element O and N, the main conclusion is consistent with the linear model proposed previously.[23] The present study supplies a new strategy for the prediction of partition coefficient and other physical properties, such as the distribution coefficient and aqueous solubility, in the drug and materials design.

Author Contributions

Qiang Zhu: conceptualization, investigation, methodology, visualization, formal analysis, and writing; Qingqing Jia: conceptualization; Ziteng Liu: resources; Yang Ge: resources, and formal analysis; Xu Gu: formal analysis; Ziyi Cui: formal analysis and software; Mengting Fan: investigation; Jing Ma: conceptualization, funding acquisition, methodology, project administration, supervision, and writing.

Conflicts of interest

There are no conflicts to declare

Acknowledgements

This work was supported by the National Key Research and Development Program of China (2019YFC0408303), the National Natural Science Foundation of China (Grant Nos. 21873045, 22033004), Parts of the calculations were performed using computational resources on an IBM Blade cluster system from the High Performance Computing Center (HPCC) of Nanjing University. Prof. Xiaogang Peng (Zhejiang University) and Prof. Congqing Zhu (Nanjing University) were gratefully thanked for fruitful discussions and supports. Qiang Zhu specially thanks Dr. Jeffrey Plante for his kind help with the implementation of JPlogP.

Notes and references

1 D. Mackay, A. K. Celsie and J. M. Parnis, *Environmental reviews*, 2016, **24**, 101–113.
2 H. Van De Waterbeemd and E. Gifford, *Nature reviews Drug discovery*, 2003, **2**, 192–204.
3 T. Barnard, H. Hagan, S. Tseng and G. C. Sosso, *Molecular Systems Design & Engineering*, 2020.
4 Y. Ran and S. H. Yalkowsky, *Journal of chemical information and computer sciences*, 2001, **41**, 354–357.
5 S. H. Yalkowsky and S. C. Valvani, *Journal of pharmaceutical sciences*, 1980, **69**, 912–922.
6 K. Wu, Z. Zhao, R. Wang and G.-W. Wei, *Journal of computational chemistry*, 2018, **39**, 1444–1454.
7 Y. Kwon, *Handbook of essential pharmacokinetics, pharmacodynamics and drug metabolism for industrial scientists*, Springer Science & Business Media, 2001.
8 L. Xing and R. C. Glen, *Journal of chemical information and computer sciences*, 2002, **42**, 796–805.
9 T. Ryckmans, M. P. Edwards, V. A. Horne, A. M. Correia, D. R. Owen, L. R. Thompson, I. Tran, M. F. Tutt and T. Young, *Bioorganic & medicinal chemistry letters*, 2009, **19**, 4406–4409.
10 Y. Yang, H. Qin, M. Jiang, L. Lin, T. Fu, X. Dai, Z. Zhang, Y. Niu, H. Cao, Y. Jin *et al.*, *Nano letters*, 2016, **16**, 2133–2138.
11 Y. Yang, H. Qin and X. Peng, *Nano letters*, 2016, **16**, 2127–2132.
12 G. König, F. C. Pickard, J. Huang, A. C. Simmonett, F. Tofoleanu, J. Lee, P. O. Dral, S. Prasad, M. Jones, Y. Shao *et al.*, *Journal of computer-aided molecular design*, 2016, **30**, 989–1006.
13 O. Fizer, M. Fizer, V. Sidey, Y. Studenyak and R. Mariychuk, *Journal of molecular modeling*, 2018, **24**, 1–12.
14 Q. Zhu, Y. Lu, X. He, T. Liu, H. Chen, F. Wang, D. Zheng, H. Dong and J. Ma, *Scientific reports*, 2017, **7**, 1–10.
15 R. Wang, Y. Fu and L. Lai, *Journal of chemical information and computer sciences*, 1997, **37**, 615–621.
16 T. Cheng, Y. Zhao, X. Li, F. Lin, Y. Xu, X. Zhang, Y. Li, R. Wang and L. Lai, *Journal of chemical information and modeling*, 2007, **47**, 2140–2148.
17 A. K. Ghose, A. Pritchett and G. M. Crippen, *Journal of Computational Chemistry*, 1988, **9**, 80–90.
18 S. A. Wildman and G. M. Crippen, *Journal of chemical information and computer sciences*, 1999, **39**, 868–873.
19 I. V. Tetko and V. Y. Tanchuk, *Journal of chemical information and computer sciences*, 2002, **42**, 1136–1145.
20 A. J. Leo and D. Hoekman, *Perspectives in drug discovery and design*, 2000, **18**, 19–38.
21 A. J. Leo, *Chemical Reviews*, 1993, **93**, 1281–1306.
22 A. A. Petrauskas and E. A. Kolovanov, *Perspectives in drug discovery and design*, 2000, **19**, 99–116.
23 R. Mannhold, G. I. Poda, C. Ostermann and I. V. Tetko, *Journal of pharmaceutical sciences*, 2009, **98**, 861–893.
24 J.-W. Zou, W.-N. Zhao, Z.-C. Shang, M.-L. Huang, M. Guo and Q.-S. Yu, *The Journal of Physical Chemistry A*, 2002, **106**, 11550–11557.
25 S. JALILI, M. TAFAZZOLI and M. JALALI-HERAVI, *Journal of Theoretical and Computational Chemistry*, 2003, **2**, 335–344.
26 N. M. Borges, P. W. Kenny, C. A. Montanari, I. M. Prokopczyk, J. F. Ribeiro, J. R. Rocha and G. R. Sartori, *Journal of computer-aided molecular design*, 2017, **31**, 163–181.
27 C. C. Bannan, G. Calabró, D. Y. Kyu and D. L. Mobley, *Journal of chemical theory and computation*, 2016, **12**, 4015–4024.
28 M. R. Jones, B. R. Brooks and A. K. Wilson, *Journal of computer-aided molecular design*, 2016, **30**, 1129–1138.
29 S. Genheden, *Journal of computer-aided molecular design*, 2017, **31**, 867–876.
30 P. S. Redmill, *Industrial & engineering chemistry research*, 2012, **51**, 4556–4566.
31 *PhysProp Update*, `https://cbec.srcinc.com/interkow/pp1357.html`, Accessed 19 Dec 2017.
32 S. Martel, F. Gillerat, E. Carosati, D. Maiarelli, I. V. Tetko, R. Mannhold and P.-A. Carrupt, *European Journal of Pharmaceutical Sciences*, 2013, **48**, 21–29.
33 J. J. Huuskonen, D. J. Livingstone and I. V. Tetko, *Journal of chemical information and computer sciences*, 2000, **40**, 947–955.
34 Q. Zhu, Y. Gu, L. Hu, T. Gaudin, M. Fan and J. Ma, *The Journal of Chemical Physics*, 2021, **154**, 074502.
35 C. Hansch, A. Leo, D. Hoekman and D. Livingstone, *Exploring QSAR: hydrophobic, electronic, and steric constants*, American Chemical Society Washington, DC, 1995, vol. 2.
36 G. Landrum, *RDKit: Open-source cheminformatics*, `http://www.rdkit.org`.
37 J. Plante and S. Werner, *Journal of cheminformatics*, 2018, **10**, 1–10.
38 J. Gasteiger and M. Marsili, *Tetrahedron*, 1980, **36**, 3219–3228.
39 A. Laio and M. Parrinello, *Proceedings of the National Academy of Sciences*, 2002, **99**, 12562–12566.
40 G. M. Torrie and J. P. Valleau, *Journal of Computational Physics*, 1977, **23**, 187–199.
41 Q. Zhu, Y. Yuan, J. Ma and H. Dong, *Advanced Theory and Simulations*, 2019, **2**, 1800171.
42 Y. Yuan, Q. Zhu, R. Song, J. Ma and H. Dong, *Journal of Chemical Theory and Computation*, 2020, **16**, 4631–4640.