

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

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

*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*

E-mail: majing@nju.edu.cn

Abstract

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

(HDNN) 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.

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 science¹ and pharmacology.^{2,3} 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} ionization constant (pKa),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 is much more sensitive to the partition coefficient.^{12,13} 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. Comparing with the atomic-based or fragment-based model¹⁵⁻²² and molecular descriptors based model,^{13,23-30} our model features in the following three aspects: (i) bypassing the fussy jobs in dividing

the whole molecules into separate fragments or atoms and no experienced chemical intuition is assumed for building the machine learning (ML) model; (ii) no need for the various descriptors which may be difficult and computationally costly to obtain; (iii) adding physically explainable elements by explicitly taking the polarization effects and conformation entropy into consideration.

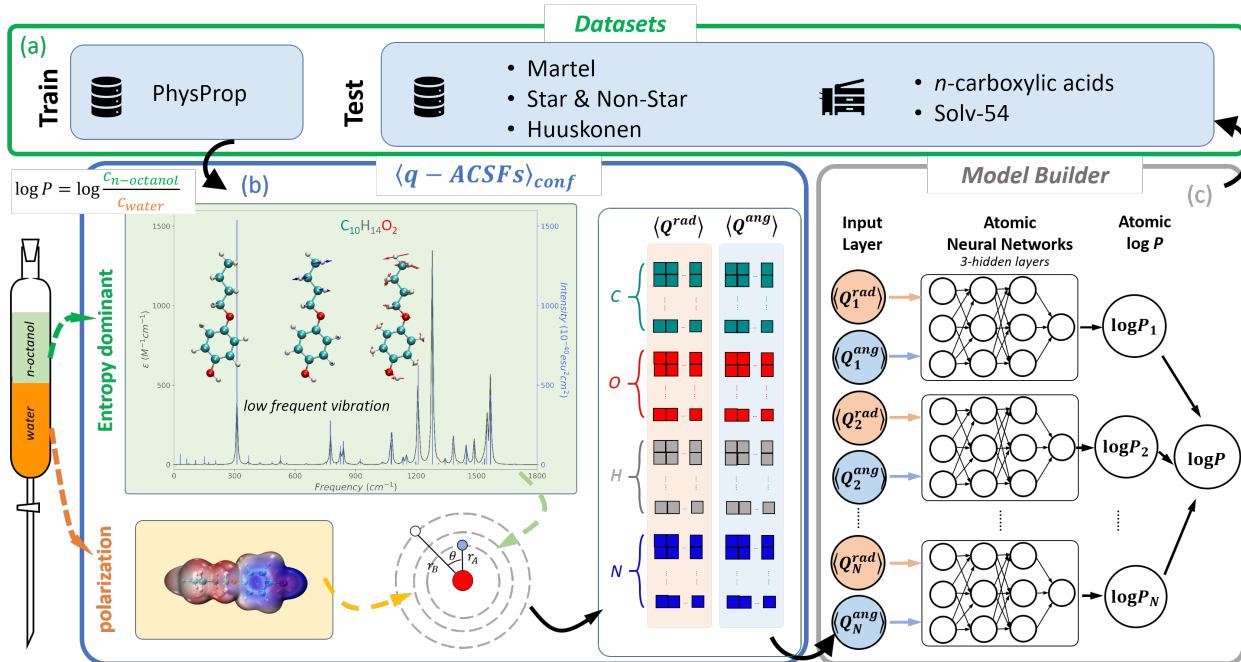


Figure 1: Schematic illustration of the workflow applied for (a) the collection of datasets, (b) derivation of the polarization and entropy, encoding polarization and entropy into atom-centered symmetry functions ($\langle q - \text{ACSFs} \rangle_{\text{conf}}$), (c) model building and partition coefficient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

Four public datasets (PhysProp,³¹ Martel,³² Star & NonStar,²³ and Huuskonen³³) were used for training and testing the HDNN model with $\langle q - \text{ACSFs} \rangle_{\text{conf}}$, as shown in Figure 1 and Figure S1. Two homemade datasets (*n*-carboxylic acids and Solv-54³⁴) 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 StarList³⁵ 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. From Figure S1(b), we could conclude that the $\log P$ distribution of the six public databases ranges from -2 to 7 , which is a broad range.

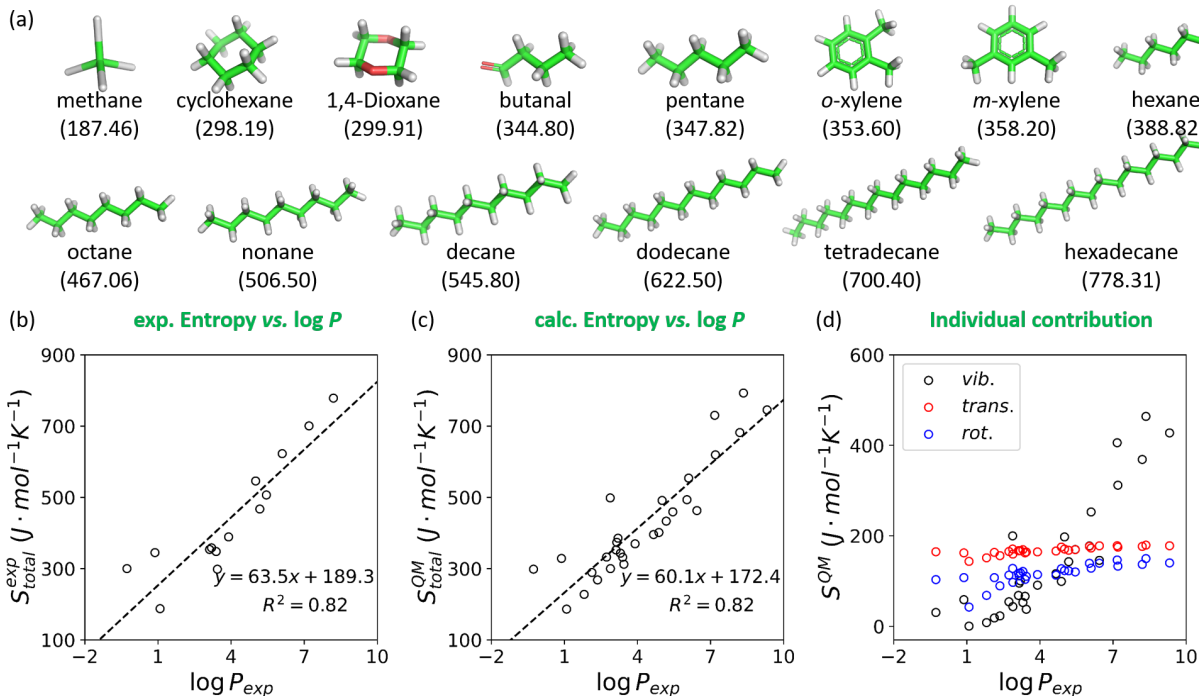


Figure 2: (a) The selected molecules for building the correlation between the partition coefficient $\log P$ measured experimentally and the entropy (S) extracted from experiments (b), the total entropy (S_{total}^{QM}) (c) calculated by quantum mechanics (QM), and individual entropy (d) came from three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

To show the entropic effect on $\log P$, we fetched experimentally measured entropy and partition coefficient of 14 molecules shown in Figure 2(a). It is interesting to find a relation-

ship between the experimental partition coefficient ($\log P$) 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.

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

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

$g(q_j)$ 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.

$$g(q_j) = q_j \quad (3)$$

$$h(q_j, q_k) = q_j q_k \quad (4)$$

When considering the entropy effects, the atom-centered symmetry functions could be denoted as $\langle q-ACSFs \rangle_{conf}$, where the radial and angular symmetry functions were expressed

as below:

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

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

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:

$$\log P = \sum_i^{N_{atoms}} \log P_i \quad (7)$$

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).

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) As shown in Figure S3 (a), top 20 ranked descriptors utilizing the MDI were presented and the

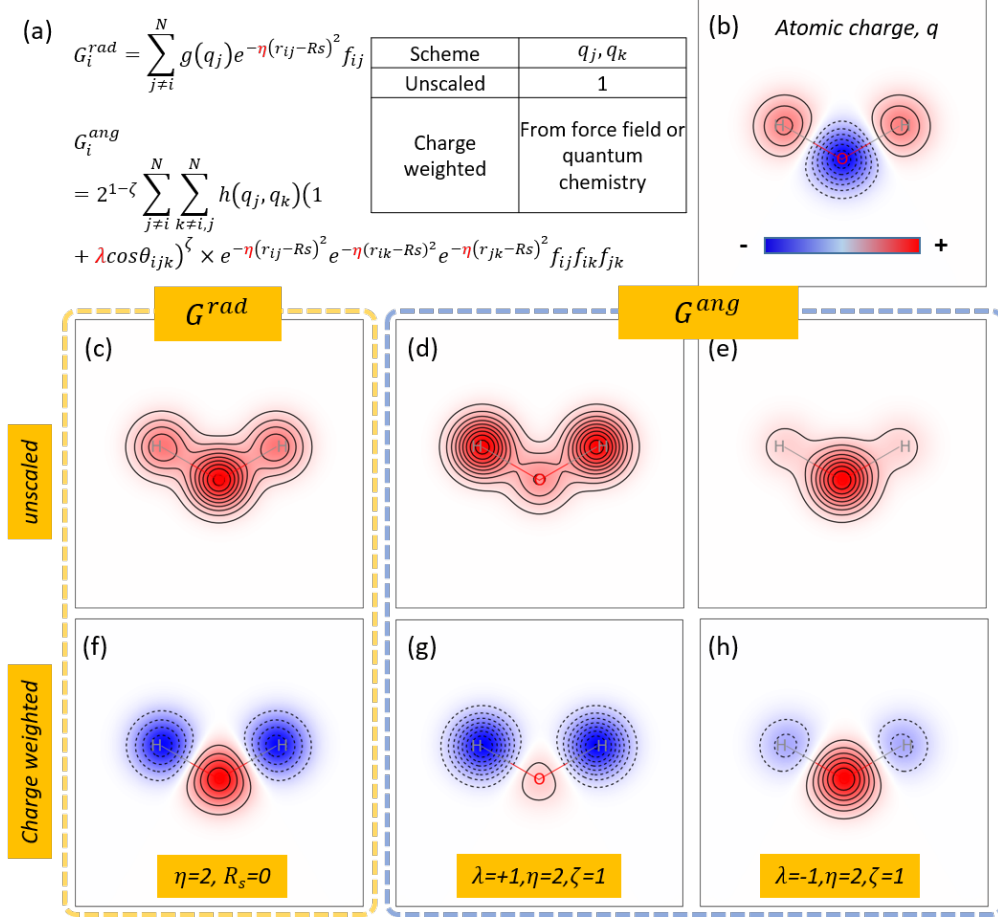


Figure 3: Importance of polarization effects. (a) Mathematical expression of radial (G^{rad}) and angular related (G^{ang}) descriptors; (b) Distribution of Gasteiger partial charge mapped onto atom O and H; (c) Radial (c.f. eq. 1) and (d)-(e) angular (c.f. eq. 2) symmetry functions of water with atomic 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.

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 and polarizability, as the electrostatic or polarizability 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.³⁷ 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 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 versa 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 disclosed 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

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,³⁸ where the estimation shown here was performed over molecules containing element C, H, O and N.

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, Section S7.1), each cluster is colored separately. From the distribution of the potential, we could see that three representative structures (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 was observed in datasets Star & Non-Star, which may account for the performance that the $\langle q - ACSFs \rangle_{conf}$ outperform the $q - ACSFs^{max}$ over datasets Star & Non-Star where little differences were observed in other two datasets (Table 1).

Performance of $\langle q - ACSFs \rangle_{conf}$ was fully assessed over three distinct datasets through MAE and MSE . (Detailed definition of the criterion could be found in supporting information, Section S8) Five methods, namely, XLOGP3,¹⁶ MolLogP,¹⁸ ALOGPS 2.1,¹⁹ JPlogP-Coeff³⁸ and JPlogP-library³⁸ were also applied for comparison. As listed in Table 1, it should

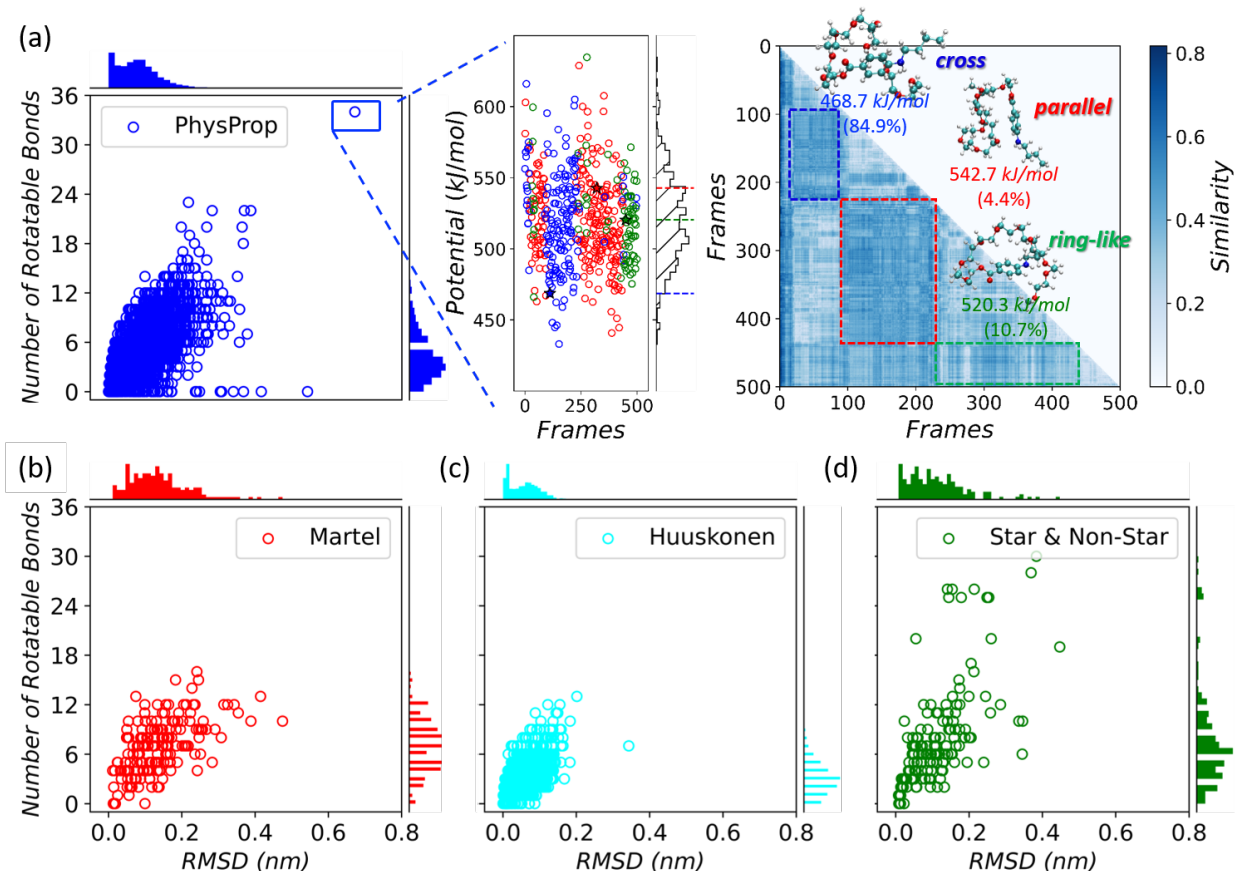


Figure 4: Correlation between the number of rotatable bonds and the root-mean-square deviations (RMSD) derived from the 1 ns MD trajectories among four public datasets, namely, (a) PhysProp, (b) Martel, (c) Huuskonen, and (d) Star & Non-Star. Three clusters of molecule $C_{30}H_{53}O_{11}N_2$ were grouped 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.

be a hard task for these five methods to predict the partition coefficient in the datasets Martel, as the distribution of partition coefficient is quite 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 be 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.

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. Almost twice improvement was observed on the datasets Star & Non-Star and Huuskonen. In the term of conformation effects, we could see that the performance differs a little among datasets Martel and Huuskonen with the introduction of different conformations, which may arise from the little difference among different conformations. However, an improvement was observed for datasets Star & Non-Star, which may be attributed to the broad distribution of conformations. 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. 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 atom-center symmetry

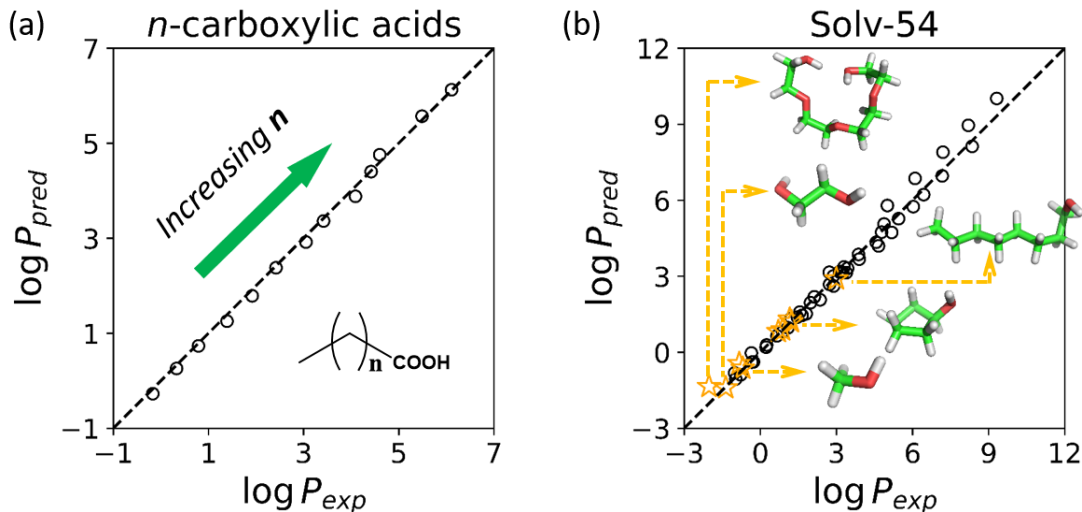


Figure 5: Performance of the predicted partition coefficient ($\log P$) by charge and ensemble weighted atom-center symmetry functions ($\langle q - ACSFs \rangle_{conf}$) over two homemade datasets, namely (a) *n*-carboxylic acids and (b) Solv-54, inserted are polar molecules such as alcohols indicated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

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 ($NHET$) ($\log P = 1.46(\pm 0.02) + 0.11(\pm 0.001)NC - 0.11(\pm 0.001)NHET$).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.

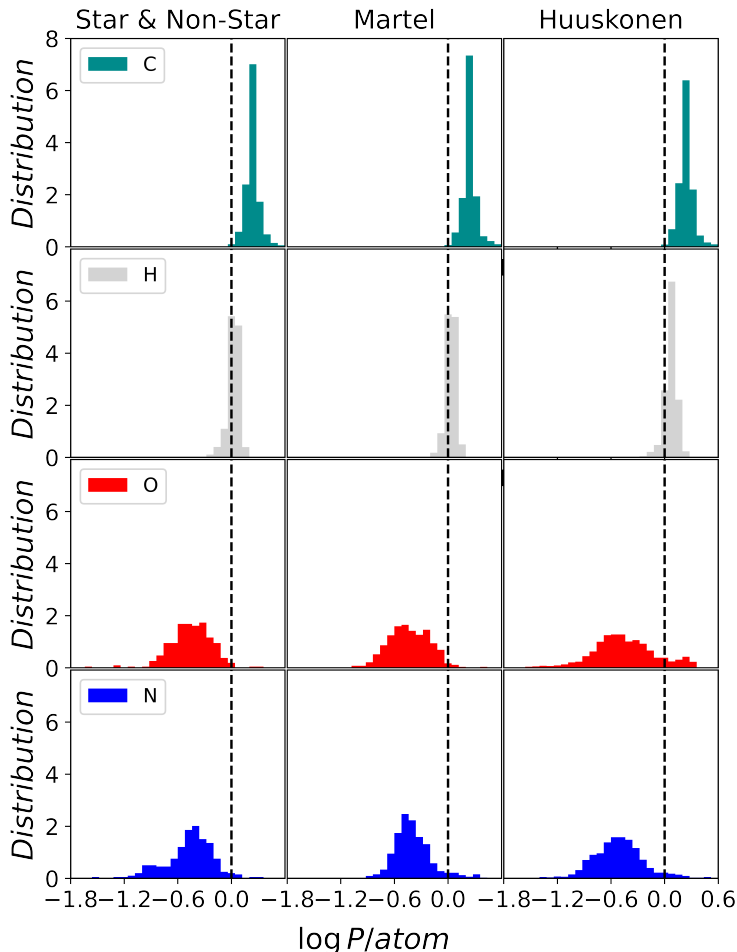


Figure 6: Partition coefficient decomposed into contribution from each atom of a single molecule. Dashed guide line at 0 is shown here, where the contributions from left are to decrease the partition coefficient, right for increasing it. The colors are in line with the molecule throughout the manuscript.

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 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 increased from 0.83 to 0.48) and Huuskonen (MAE increased from 0.54 to 0.22). However, 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.

Acknowledgement

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.

Supporting Information Available

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

- (1) Mackay, D.; Celsie, A. K.; Parnis, J. M. The evolution and future of environmental partition coefficients. *Environmental reviews* **2016**, *24*, 101–113.
- (2) Van De Waterbeemd, H.; Gifford, E. ADMET in silico modelling: towards prediction paradise? *Nature reviews Drug discovery* **2003**, *2*, 192–204.
- (3) Barnard, T.; Hagan, H.; Tseng, S.; Sosso, G. C. Less may be more: an informed reflection on molecular descriptors for drug design and discovery. *Molecular Systems Design & Engineering* **2020**,
- (4) Ran, Y.; Yalkowsky, S. H. Prediction of drug solubility by the general solubility equation (GSE). *Journal of chemical information and computer sciences* **2001**, *41*, 354–357.
- (5) Yalkowsky, S. H.; Valvani, S. C. Solubility and partitioning I: solubility of nonelectrolytes in water. *Journal of pharmaceutical sciences* **1980**, *69*, 912–922.
- (6) Wu, K.; Zhao, Z.; Wang, R.; Wei, G.-W. TopP-S: Persistent homology-based multi-task deep neural networks for simultaneous predictions of partition coefficient and aqueous solubility. *Journal of computational chemistry* **2018**, *39*, 1444–1454.

- (7) Kwon, Y. *Handbook of essential pharmacokinetics, pharmacodynamics and drug metabolism for industrial scientists*; Springer Science & Business Media, 2001.
- (8) Xing, L.; Glen, R. C. Novel methods for the prediction of logP, p K a, and logD. *Journal of chemical information and computer sciences* **2002**, *42*, 796–805.
- (9) Ryckmans, T.; Edwards, M. P.; Horne, V. A.; Correia, A. M.; Owen, D. R.; Thompson, L. R.; Tran, I.; Tutt, M. F.; Young, T. Rapid assessment of a novel series of selective CB2 agonists using parallel synthesis protocols: A Lipophilic Efficiency (LipE) analysis. *Bioorganic & medicinal chemistry letters* **2009**, *19*, 4406–4409.
- (10) Yang, Y.; Qin, H.; Jiang, M.; Lin, L.; Fu, T.; Dai, X.; Zhang, Z.; Niu, Y.; Cao, H.; Jin, Y., et al. Entropic ligands for nanocrystals: from unexpected solution properties to outstanding processability. *Nano letters* **2016**, *16*, 2133–2138.
- (11) Yang, Y.; Qin, H.; Peng, X. Intramolecular entropy and size-dependent solution properties of nanocrystal–ligands complexes. *Nano letters* **2016**, *16*, 2127–2132.
- (12) König, G.; Pickard, F. C.; Huang, J.; Simmonett, A. C.; Tofoleanu, F.; Lee, J.; Dral, P. O.; Prasad, S.; Jones, M.; Shao, Y., et al. Calculating distribution coefficients based on multi-scale free energy simulations: an evaluation of MM and QM/MM explicit solvent simulations of water-cyclohexane transfer in the SAMPL5 challenge. *Journal of computer-aided molecular design* **2016**, *30*, 989–1006.
- (13) Fizer, O.; Fizer, M.; Sidey, V.; Studenyak, Y.; Mariychuk, R. Benchmark of different charges for prediction of the partitioning coefficient through the hydrophilic/lipophilic index. *Journal of molecular modeling* **2018**, *24*, 1–12.
- (14) Zhu, Q.; Lu, Y.; He, X.; Liu, T.; Chen, H.; Wang, F.; Zheng, D.; Dong, H.; Ma, J. Entropy and polarity control the partition and transportation of drug-like molecules in biological membrane. *Scientific reports* **2017**, *7*, 1–10.

- (15) Wang, R.; Fu, Y.; Lai, L. A new atom-additive method for calculating partition coefficients. *Journal of chemical information and computer sciences* **1997**, *37*, 615–621.
- (16) Cheng, T.; Zhao, Y.; Li, X.; Lin, F.; Xu, Y.; Zhang, X.; Li, Y.; Wang, R.; Lai, L. Computation of octanol- water partition coefficients by guiding an additive model with knowledge. *Journal of chemical information and modeling* **2007**, *47*, 2140–2148.
- (17) Ghose, A. K.; Pritchett, A.; Crippen, G. M. Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships III: Modeling hydrophobic interactions. *Journal of Computational Chemistry* **1988**, *9*, 80–90.
- (18) Wildman, S. A.; Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *Journal of chemical information and computer sciences* **1999**, *39*, 868–873.
- (19) Tetko, I. V.; Tanchuk, V. Y. Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program. *Journal of chemical information and computer sciences* **2002**, *42*, 1136–1145.
- (20) Leo, A. J.; Hoekman, D. Calculating log P (oct) with no missing fragments; The problem of estimating new interaction parameters. *Perspectives in drug discovery and design* **2000**, *18*, 19–38.
- (21) Leo, A. J. Calculating log Poct from structures. *Chemical Reviews* **1993**, *93*, 1281–1306.
- (22) Petrauskas, A. A.; Kolovanov, E. A. ACD/Log P method description. *Perspectives in drug discovery and design* **2000**, *19*, 99–116.
- (23) Mannhold, R.; Poda, G. I.; Ostermann, C.; Tetko, I. V. Calculation of molecular lipophilicity: State-of-the-art and comparison of log P methods on more than 96,000 compounds. *Journal of pharmaceutical sciences* **2009**, *98*, 861–893.

- (24) Zou, J.-W.; Zhao, W.-N.; Shang, Z.-C.; Huang, M.-L.; Guo, M.; Yu, Q.-S. A Quantitative Structure- Property Relationship Analysis of logP for Disubstituted Benzenes. *The Journal of Physical Chemistry A* **2002**, *106*, 11550–11557.
- (25) JALILI, S.; TAFAZZOLI, M.; JALALI-HERAVI, M. Comparison of Multiple Linear Regression and Artificial Neural Networks in Predicting Octanol/Water Partition Coefficient of a Variety of Organic Molecules. *Journal of Theoretical and Computational Chemistry* **2003**, *2*, 335–344.
- (26) Borges, N. M.; Kenny, P. W.; Montanari, C. A.; Prokopczyk, I. M.; Ribeiro, J. F.; Rocha, J. R.; Sartori, G. R. The influence of hydrogen bonding on partition coefficients. *Journal of computer-aided molecular design* **2017**, *31*, 163–181.
- (27) Bannan, C. C.; Calabró, G.; Kyu, D. Y.; Mobley, D. L. Calculating partition coefficients of small molecules in octanol/water and cyclohexane/water. *Journal of chemical theory and computation* **2016**, *12*, 4015–4024.
- (28) Jones, M. R.; Brooks, B. R.; Wilson, A. K. Partition coefficients for the SAMPL5 challenge using transfer free energies. *Journal of computer-aided molecular design* **2016**, *30*, 1129–1138.
- (29) Genheden, S. Solvation free energies and partition coefficients with the coarse-grained and hybrid all-atom/coarse-grained MARTINI models. *Journal of computer-aided molecular design* **2017**, *31*, 867–876.
- (30) Redmill, P. S. Estimating octanol–water partition coefficients for selected nanoscale building blocks using the COSMO-SAC segment contribution method. *Industrial & engineering chemistry research* **2012**, *51*, 4556–4566.
- (31) PhysProp Update. <https://cbec.srcinc.com/interkow/pp1357.html>, Accessed 19 Dec 2017.

- (32) Martel, S.; Gillerat, F.; Carosati, E.; Maiarelli, D.; Tetko, I. V.; Mannhold, R.; Carrupt, P.-A. Large, chemically diverse dataset of log P measurements for benchmarking studies. *European Journal of Pharmaceutical Sciences* **2013**, *48*, 21–29.
- (33) Huuskonen, J. J.; Livingstone, D. J.; Tetko, I. V. Neural network modeling for estimation of partition coefficient based on atom-type electrotopological state indices. *Journal of chemical information and computer sciences* **2000**, *40*, 947–955.
- (34) Zhu, Q.; Gu, Y.; Hu, L.; Gaudin, T.; Fan, M.; Ma, J. Shear viscosity prediction of alcohols, hydrocarbons, halogenated, carbonyl, nitrogen-containing, and sulfur compounds using the variable force fields. *The Journal of Chemical Physics* **2021**, *154*, 074502.
- (35) Hansch, C.; Leo, A.; Hoekman, D.; Livingstone, D. *Exploring QSAR: hydrophobic, electronic, and steric constants*; American Chemical Society Washington, DC, 1995; Vol. 2.
- (36) Landrum, G. RDKit: Open-source cheminformatics. <http://www.rdkit.org>.
- (37) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219–3228.
- (38) Plante, J.; Werner, S. JPlogP: an improved logP predictor trained using predicted data. *Journal of cheminformatics* **2018**, *10*, 1–10.
- (39) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proceedings of the National Academy of Sciences* **2002**, *99*, 12562–12566.
- (40) Torrie, G. M.; Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics* **1977**, *23*, 187–199.
- (41) Zhu, Q.; Yuan, Y.; Ma, J.; Dong, H. A Data-Driven Accelerated Sampling Method for

Searching Functional States of Proteins. *Advanced Theory and Simulations* **2019**, *2*, 1800171.

- (42) Yuan, Y.; Zhu, Q.; Song, R.; Ma, J.; Dong, H. A two-ended data-driven accelerated sampling method for exploring the transition pathways between two known states of protein. *Journal of Chemical Theory and Computation* **2020**, *16*, 4631–4640.

Graphical TOC Entry

