Thermodynamics-Based Model Construction for the Accurate Prediction of Molecular Properties From Partition Coefficients

Deliang Chen*, Xiaoqing Huang and Yulan Fan

Jiangxi Key Laboratory of Organo-Pharmaceutical Chemistry, Chemistry and Chemical Engineering College, Gannan Normal University, Ganzhou, China

Developing models for predicting molecular properties of organic compounds is imperative for drug development and environmental safety; however, development of such models that have high predictive power and are independent of the compounds used is challenging. To overcome the challenges, we used a thermodynamics-based theoretical derivation to construct models for accurately predicting molecular properties. The free energy change that determines a property equals the sum of the free energy changes (ΔGf-s) caused by the factors affecting the property. By developing or selecting molecular descriptors that are directly proportional to ΔGf-s, we built a general linear free energy relationship (LFER) for predicting the property with the molecular descriptors as predictive variables. The LFER can be used to construct models for predicting various specific properties from partition coefficients. Validations show that the models constructed according to the LFER have high predictive power and their performance is independent of the compounds used, including the models for the properties having little correlation with partition coefficients. The findings in this study are highly useful for applications in drug development and environmental safety.

Keywords: computational chemistry, linear free energy relationships, molecular properties, partition coefficient, noncovalent interactions, quantitative structure-property relationships

INTRODUCTION

The rapid development of new organic compounds in various chemical-related laboratories and industries has increased the difficulty in measuring the physicochemical and absorption, distribution, metabolism, excretion, and toxicity (ADME/Tox) properties of all possible compounds. Therefore, the development of techniques for predicting these properties via computational tools is imperative (Sarkar et al., 2012; Li et al., 2019; Sun et al., 2019; Suay-Garcia et al., 2020). Quantitative structure–property relationships (QSPRs) with multiple predictive variables are widely used for predicting various properties of organic compounds. QSPR employs regression statistics using algorithms, such as artificial neural networks, (Deeb et al., 2011; Song et al., 2017), machine learning (Bushdid et al., 2018; Cheng and Ng, 2019; Zhong et al., 2019), and partial least square (Deeb et al., 2011; T. Stanton, 2012), with predictive variables usually selected from a few thousand molecular descriptors based on mathematical and statistical tools (Mansouri et al., 2018; Lee et al., 2019; Fiorese et al., 2020). A large number of articles related to QSPR were published per year and QSPR has gained importance in a wide range of fields, such as drug design, pesticide design,
and environmental toxicology (Roy et al., 2018; Yang et al., 2018; Zhu et al., 2018; He et al., 2019; Khan et al., 2019; Zhu et al., 2020a). For example, predicting the ADME/Tox of drug candidates before synthesis can significantly reduce the cost and time of drug development and increase the success rate (Cheng et al., 2013; Dickson et al., 2017; Zhu et al., 2018). Predicting soil/water partition coefficients and the toxicities of organic compounds is vital for environmental risk assessments (Freitas et al., 2014; Sabour et al., 2017; Khan et al., 2019). Some properties can be predicted accurately with hydrophobicity (logPoct, the logarithm of the partition coefficient between n-octanol and water) and/or other commonly used molecular descriptors, e.g., electrophilicity index (ω) (Raevsky, 2004; Pal et al., 2019; Jana et al., 2020). For example, logPoct has been used to predict the water solubility with high accuracy, (Raevsky, 2004). Robust multiple linear regression (MLR) models for toxicity prediction can be constructed by using the combinations of electronic factor (ω, ω², or ω³) and hydrophobicity factor [logPocn, (logPoct)²] as predictors (Pal et al., 2019; Jana et al., 2020). The robustness of the models were ascertained by neural networks. However, for many properties, constructing QSPR models with high predictive accuracy and reliability remains a challenge. The performance of QSPR models greatly depends on the compounds used for investigation, quality of the data, and modelling methodology employed (Song et al., 2017; Mansouri et al., 2018; Zhang et al., 2020). For a given property, the predictive variables would be different if the data in the training set are different. In addition, QSAR models usually work well only for the compounds within their applicability domains and do not have good predictive accuracy for other compounds (Kaneo, 2017; Liu and Wallqvist, 2019). However, it is difficult to define the accurate applicability domains for QSPR models because there is no general agreement for quantifying compound similarity (Carrió et al., 2016). It is thus important to develop a new methodology for constructing models that have high predictive power and the performance of the models is independent of the compounds used.

The quantitative formula and quantitative relationships that are developed via theoretical derivation in physical chemistry are absolutely correct and are independent of the compounds used. For example, the partition coefficient between water and an organic solvent (logPoct) for a solute is directly proportional to the standard free energy change for transferring the solute from water to the organic solvent (ΔGtr). The ΔGtr in turn depends on the standard enthalpy change (ΔHtr) and entropy change (ΔStr) of the phase-transferring process. Thus, at a given temperature, the model logPoct = b1ΔHtr + b2ΔStr + c (b1, b2, and c are constants) is absolutely correct and has high predictive power for predicting logPoct. This example indicates that the models developed via thermodynamics-based theoretical derivations may overcome the shortages of the models developed by using mathematical and statistical tools. A large number of physicochemical properties, ADME/Tox qualities, and many other properties of organic compounds depend on the changes in free energy caused by the intermolecular noncovalent interactions of the compounds with their environments. The enormous catalytic power of many enzymes depends on the noncovalent interactions between substrates and enzymes (Warshel et al., 2006; Chen et al., 2019). It is thus expected that models with high predictive power for many properties can be developed by considering the free energy changes related to the properties. In this study, we used a thermodynamics-based theoretical derivation to develop a general linear free energy relationship (LFER) for predicting various properties of organic compounds. The LFER can be used to construct models for many specific properties. Validation shows that the models for specific properties have high predictive power and their performance is independent of the compounds used.

### COMPUTATIONAL METHODS

#### Data set selection

In this study, all experimental data of logPoct, logP16 (the logarithm of the partition coefficient between hexadecane and water), logPch (the logarithm of the partition coefficient between chloroform and water), logPd (the logarithm of the partition coefficient between aniline and water), logKbrain (the logarithm of the partition coefficient from air to human brain) and logKp (logarithm of experimental human skin permeability) are collected from literatures (Abraham et al., 1994; Abraham et al., 1999; Abraham and Martins, 2004; Abraham et al., 2006; Abraham et al., 2015; Zhang et al., 2017). Hydrogen bond acceptors (HBAs) include very weak H-bond acceptors. For example, the sp2 carbon atoms from carbon-carbon double bonds and aromatic rings are weak HBAs. Hydrogen bond donors (HBDs) include very weak H-bond donors. For example, the hydrogen atoms in CHCl3 and CH3NO2 are weak HBDs.

#### Calculation of S_m

S_m is a molecular descriptor developed in this study. The S_m values of organic compounds were calculated based on the formula of the compounds. Assume the formula of a neutral organic compound is C_Hx_Oy_Nz_S_F_Cl_Br_I, the S_m of this compound is

\[
S_m = c + 0.3h + o + n + 2s + 0.6f^+ + 1.8c + 2.2br + 2.6i - 0.2Nc3 - 0.6Nc4. \tag{1}
\]

In Eq. 1, c, h, o, n, s, f, cl, br and i are the numbers of carbon, hydrogen, oxygen, nitrogen, sulfur, fluoride, chloride, bromide and iodide atoms in the solute, Nc3 is the number of sp3 carbons connecting three heavy atoms (fluoride is not included), Nc4 is the number of sp3 carbons connecting four heavy atoms (fluoride is not included).

#### Calculation of H_M_HBD

H_M_HBD values of solutes were calculated based on the approach reported in a previous study (Chen et al., 2020).

#### Calculation of Flexibility

In this study, the flexibility of a solute is calculated by summarizing the flexibilities of the bonds of the solute. If a bond is not rotatable or if the rotation of a bond does not change the conformation of the solute, the flexibility of the bond is set to
zero (note: hydrogen atoms are not included for determining conformations). The flexibility of the C–C bond in R’CH2—CH2R2 is set to one. If the energy barrier for rotating a bond is obviously higher than that for rotating the R’CH2—CH2R bond, the flex value is set to zero. For example, the C—N bond in RCO—NH and the C–C bond in Ar—CO are set to zero. If the energy barrier for rotating a bond is obviously lower than that for rotating the R’CH2—CH2R2 bond, the flex value is set to 1.5. For example, the energy barrier for rotating the R’O—CH2R bond is lower than that for R’CH2—CH2R (Sun et al., 2019) and thus the flex value of the C—O bond is set to 1.5. Also, the flexibility of C—C in R’CH2—C6H5 is set to 0.5 because of the symmetry of phenyl ring.

Calculation of the effects of HBAs on the logPoct/logPchl

The free energy changes for transferring depolarized solutes from water to hexadecane (ΔGtr_depol) were calculated based on the method reported in previous study (Chen et al., 2020). Based on the logPoct (or logPchl) and ΔGtr_depol values of nonpolar compounds, the model for the regression of logPoct (or logPchl) against ΔGtr_depol was developed. This model was then used to calculate the logPoct (or logPchl) values for depolarized solutes. For a solute containing HBAs but no HBDs, the difference between the calculated logPoct (or logPchl) for the depolarized solute and the experimental logPoct (or logPchl) of the solute is the effect of HBAs on the logPoct (or logPchl) of the solute.

Model development

All the models and the statistical reliabilities of the models were obtained by performing the multiple linear regressions implemented in Excel.

RESULTS AND DISCUSSION

Thermodynamics-Based Theoretical Derivation for Generating a Linear Free Energy Relationship

In the theoretical derivation, we used “Y” to represent a property and the symbol “ΔGY” to represent the free energy change that determines Y. The ΔGY values for many properties are not easy to be calculated directly. Thus, we decomposed ΔGY into the free energy changes that are caused by the factors affecting Y. The free energy change caused by a factor is denoted by “ΔGF”. Thus, ΔGY equals the summarization of the ΔGFs for all the factors affecting Y.

\[
ΔGY = ΣΔGF
\] (2)

For the properties depending on the noncovalent interactions of solutes, they are affected by the molecular sizes, hydrogen-bond acceptors (HBAs), and hydrogen-bond donors (HBDs) of the solutes, which was demonstrated in a previous study (Chen et al., 2020). Many properties are also affected by the flexibilities of solutes. For example, the partition coefficients of organic compounds between a flexible environment (e.g., blood) and a much less flexible environment (e.g., muscle) are obviously affected by the flexibilities of the compounds. It is challenging to accurately quantify the ΔGY for various properties. However, it is possible to develop molecular descriptors that are directly proportional to ΔGY. We used Df to represent the molecular descriptor that is directly proportional to ΔGY. Then, ΔGY can be expressed as:

\[
ΔGY = ΣkFDf
\] (3)

The Df values are constant for a given property. Theoretically, if the molecular descriptors apply to various properties, Eq. 3 can be used to construct models with high predictive power for the properties. Many properties are mainly affected by the molecular sizes, HBAs, HBDs and flexibilities of solutes. Thus, in this study, we developed or selected molecular descriptors for quantifying the effects of molecular size, HBAs, HBDs and flexibility on the properties.

The molecular descriptor we developed for quantifying the effects of molecular size on properties is denoted by “Sm”. The Sm values of organic compounds represent the relative molecular sizes of the compounds and can be easily calculated from their molecular formulas (see Computational Methods). For example, the Sm for catechol (formula: C6H4O2) is 9.8 (num. for C + 0.3x num. for H + num. for O). To illustrate whether Sm is an ideal molecular descriptor for molecular size, we first explored the linear associations between logP16 and Sm and between logPoct and Sm for a series of alkane compounds (Figure 1A). The logP16 and logPoct values for alkane compounds are affected merely by the sizes of the compounds. The robust linear associations in Figure 1A support that Sm is directly proportional to the effects of molecular size on logP16 and logPoct. We next explore whether Sm is also an ideal molecular descriptor of molecular size for the properties that have little correlation with logP16 or logPoct. As reported in a previous study, logKbrain has little correlation with logP16 and logPoct (Chen et al., 2020). We thus explored the linear association between logKbrain and Sm for nonpolar solutes (Figure 1B). The R2 and SD values indicate that there is a strong linear association between logKbrain and Sm. In Figure 1C, we plotted the free energy changes for transferring the depolarized compounds from water to hexadecane (ΔGtr_depol) against the Sm values for the compounds from Supplementary Table S1 of a previous study (Chen et al., 2020). The high statistical reliability for the regression of ΔGtr_depol against Sm further supports that Sm is an ideal molecular descriptor for quantifying the effect of molecular size on the properties depending on noncovalent interactions. Thus, Sm is an ideal molecular descriptor for molecular size and applies to various properties.

In previous studies, we defined the water to hexadecane phase transferring free energy contributed by the electrostatic interactions of the HBAs of a solute as the overall H-bonding capability of the HBAs of the solute (Chen et al., 2019; Chen et al.,...
and this overall H-bonding capability is donated by “HM_HBA.” The definition indicates that HM_HBA is an ideal molecular descriptor for quantifying the effects of HBAs on logP_{16}. We next explored whether HM_HBA is an ideal molecular descriptor for logP_{oct} and logP_{chl}. The strong linear associations between the effect of HBAs on logP_{oct} and HM_HBA (Figure 2A) and between the effect of HBAs on logP_{chl} and H_{M_HBA} (Figure 2B) suggest that HM_HBA is an ideal molecular descriptor for quantifying the effects of HBAs on various properties. Similarly, we defined the water to hexadecane phase-transferring free energy contributed by the electrostatic interactions of the HBDs of a solute as the overall H-bonding capability of the HBDs of the solute (HM_HBD) (Chen et al., 2016; Chen et al., 2019; Chen et al., 2020). In a previous study, we revealed that the contribution of a protein-ligand H-bond to the protein-ligand binding free energy is directly proportional to the H-bonding capability of the HBA and the H-bonding capability of the HBD (Chen et al., 2016). We also found that the effect of an enzyme-substrate H-bond interaction on the free energy barrier of the enzymatic reaction is directly proportional to the H-bonding capability of the atom from the enzyme (Chen et al., 2019). Thus, we believe that the effects of HBAs and HBDs of solutes on the properties related to noncovalent interactions are directly proportional to the HM_HBA and HM_HBD values of the solutes. HM_HBA and HM_HBD are ideal molecular descriptors for quantifying the effect of HBAs and HBDs on the properties related to noncovalent interactions.

The molecular descriptor for quantifying the effect of molecular flexibility on properties is denoted by “Flex.” The effects of molecular flexibility on properties mainly result from rotatable bonds of the solutes because the rotatable bonds of the solutes can rotate more freely in some environments than in other
environments. The flexibilities of solutes are calculated from the rotatable bonds of the solutes, especially the rotatable bonds that change the conformations of solutes (see Computational Methods). Thus, for many properties that are affected by molecular size, HBAs, HBDs and flexibility, they can be quantified with the following equation

$$Y = k_1S_m + k_2HM_{HBA} + k_3HM_{HBD} + k_4\text{Flex} + c \quad (4)$$

where $k_1$, $k_2$, $k_3$, $k_4$ and $c$ are constants for a given property. Organic compounds usually contain multiple HBAs and the HBAs affect each other. The accurate calculation of $HM_{HBA}$ for many organic compounds is not easy. Eq. 4 would become simpler and easier to use if $HM_{HBA}$ is replaced by $\logP_{oct}$ because $\logP_{oct}$ is a well-known molecular descriptor for predicting properties (Liu et al., 2019; Zhu et al., 2020b) and can be obtained accurately via experimental and/or computational approaches. Based on the fact that $\logP_{oct}$ is a property and Eq. 4 also applies to $\logP_{oct}$, we can convert Eqs. 4 to 5 (see Supplementary Text S1 for the detail of the process of the conversion).

$$Y = b_1\logP_{oct} + b_2S_m + b_3HM_{HBD} + b_4\text{Flex} + c \quad (5)$$

where $b_1$, $b_2$, $b_3$, and $b_4$ are constants, $\logP_{oct}$ is $\logP_{16}$ or $\logP_{ch}$. Eq. 5 is identical to Eq. 4. Both equations are correct for the properties that are determined by the noncovalent interactions of solutes with flexible environments. All the factors related to effects of noncovalent interactions on phase-transferring free energies, including electrostatic interaction, desorption, van der Waals interactions, entropy change, etc. are considered in Eq. 5. Eq. 5 is the general LFER we developed for predicting the properties that depend on the noncovalent interactions of solutes with flexible environments. Although $S_m$ and $HM_{HBD}$ may be strongly correlated with $\logP_{oct}$ for some properties, none of the molecular descriptors can be omitted because Eq. 4 is a general LFER for various different properties.

Validation of the General LFER: Model Construction for Specific Properties

Prediction of Various Organic Solvent/Water Partition Coefficients

To prove that this general LFER can be used to construct models with high predictive power for various specific properties, we first demonstrated that it can be used to predict an organic solvent/water partition coefficient of organic compounds (Fig. 3). The LFER was used to predict the partition coefficients of selected compounds with experimental data. The model parameters were estimated using the least squares method. The $R^2$ values of the simple regressions of $\logP_{16}$ against $\logP_{oct}$ and of the corresponding models constructed according to the LFER are shown in Fig. 3A and B, respectively. The predicted partition coefficients were compared with the experimental data. The results show that the LFER can be used to predict the partition coefficients of organic compounds with high accuracy.
water partition coefficient from another organic solvent/water partition coefficient with high accuracy. Eighty-nine compounds with experimental logP_{oct}, logP_{16}, and logP_{chl} values (Abraham et al., 1994; Abraham et al., 1999) (Supplementary Table S1) were used for this investigation. Among the compounds, 45 compounds contain HBAs but no HBDs and 41 compounds containing HBDs. The equations and statistical results of the simple regressions of logP_{16} against logP_{oct} and logP_{chl} against logP_{oct} for various types of compounds are shown in Supplementary Text S2. The R^2 (squared correlation coefficient) values of the regressions range from 0.501 to 0.972 (gray columns, Figures 3A,B) and the SD (standard deviation) values of the regressions range from 0.241 to 0.965, indicating that the strength of the linear associations between two partition coefficients largely depends on compounds used for investigation. Then the same data for constructing the simple regressions were used to construct models according to the general LFER and the results are also shown in Supplementary Text S2 (note: the model descriptor Flex is not used because Flex has little effect on logP_{oct}). The R^2 values of the models range from 0.947 to 0.992 (black columns, Figures 3A,B) and the SD values range from 0.183 to 0.248. The results indicate that the models constructed according to the LFER have a high statistical reliability and the performance of the models is independent of the compounds for investigation.

To demonstrate whether the models have high predictive power, we compared the experimental logP_{16} values of 200 organic compounds [from Supplementary Table S1 of a previous study (Chen et al., 2020)] and the logP_{16} values calculated from the model constructed according to the LFER by using an external validation approach (Figure 3C and Supplementary Text S3). The result shows that the model has high predictive power. For comparison, the predictive power of the corresponding simple regression was also investigated (Figure 3D), which is much worse than that for the model constructed according to the LFER. Thus, the LFER is powerful for constructing models with high predictive power.

Prediction of the Human Skin Permeability

We next used the LFER to construct a model for predicting the human skin permeability of neutral organic molecules. Supplementary Table S2 shows the logK_p (Abraham and Martins, 2004; Zhang et al., 2017) values of 51 organic compounds. Thirty-two of the compounds were used as training set to develop the model with logP_{oct} as a predictive valuable and the other 19 compounds as a test set to validate the model. The model constructed according to the LFER is shown below.

\[
\log K_p = 0.6157 \log P_{oct} + 0.0156S_m - 0.0626H + 0.0988\text{Flex} - 5.646; \\
N = 32, R^2 = 0.953, Q^2 (\text{ext}) = 0.966, SD = 0.178, F = 136.7
\]

(6)

This model is characterized by high statistical reliability according to the R^2 and SD values. It is used to calculate the logK_p values of the 19 solutes in the test set. The plot of the experimental logK_p values versus the calculated logK_p values is characterized by statistically robust linearity (Figure 4A). The accurate prediction of logK_p can provide a rapid and accurate prediction of human skin permeability of organic compounds, which is very useful for evaluating environmental risks due to contact with skin.

Prediction of Air to Human Brain Partition Coefficient

To further illustrate the reliability and accuracy of the LFER, we used the LFER to construct models for the properties that have little correlation with logP_{oct}. The strength of the linear association between logK_{brain} and logP_{oct} (or logP_{16}) is weak (R^2 < 0.1) (Chen et al., 2020). Supplementary Table S3 lists the compounds that were used to demonstrate the weak linear association between logK_{brain} and logP_{16} (or logP_{oct}) in a previous study (Chen et al., 2020). Based on the experimental
log\(K_{\text{brain}}\), log\(P_{16}\) and log\(P_{\text{oct}}\) data of the compounds, we constructed two models according to the LFER.

\[
\log K_{\text{brain}} = -0.5129 \log P_{16} + 0.5006 S_m + 0.1009 H_{M, \text{HBD}} - 0.1893 \text{Flex} - 1.64; \\
N = 34, R^2 = 0.964, Q^2_{\text{LOO}} = 0.947; SD = 0.265; F = 195.7
\]

(7)

\[
\log K_{\text{brain}} = -0.7755 \log P_{\text{oct}} + 0.5473 S_m + 0.1790 H_{M, \text{HBD}} - 0.0886 \text{Flex} - 1.386 \\
N = 31, R^2 = 0.931, Q^2_{\text{LOO}} = 0.914; SD = 0.368; F = 87.4
\]

(8)

Results indicate that both models have high predictive power and the model with log\(P_{16}\) as predictive variable is better than the model with log\(P_{\text{oct}}\) as predictive variable. The predicted log\(K_{\text{brain}}\) obtained from the LOO cross-validation (log\(P_{16}\) is used) and the observed log\(K_{\text{brain}}\) show a robust linear association (Figure 4B). Thus, the general LFER works well for the properties that have little correlation with partition coefficients, supporting the reliability and efficacy of the general LFER in the accurate prediction of the properties related to noncovalent interactions. We believe that the thermodynamics-based theoretical derivation is a powerful methodology for developing robust models and will be useful in many fields, including drug design, environmental safety and human health.

Model Simplification
In some cases, not all the molecular descriptors in the LFER are required for specific models with high predictive power. Some models still have high predictive power without using the molecular descriptor H\(_{M, \text{HBD}}\). For example, if the HBAs and HBDS of an organic solvent are obviously weaker than the HBAs and HBDS of water, the partition coefficient between water and the organic solvent can be predicted accurately from the model with log\(P_{16}\) and \(S_m\) as predictive variables. Eq. 9 is the model for predicting the aniline/water partition coefficient (log\(P_{\text{an}}\)) with log\(P_{16}\) and \(S_m\) as predictive variables (see Supplementary Table S4 for the data). Its statistical reliability is high and is obviously better than that for the simple regression (Eq. 10).

\[
\log P_{\text{an}} = 0.4695 \log P_{16} + 0.1506 S_m + 0.010; N = 54, R^2 = 0.975, SD = 0.208, F = 1008.
\]

(9)

\[
\log P_{\text{an}} = 0.6416 \log P_{16} + 0.726; N = 54; R^2 = 0.910; SD = 0.394, F = 524.9.
\]

(10)

Without using H\(_{M, \text{HBD}}\), the model for predicting log\(K_{\text{brain}}\) from log\(P_{16}\), \(S_m\) and Flex still has high predictive power.

\[
\log K_{\text{brain}} = -0.6194 \log P_{16} + 0.5446 S_m - 0.1928 \text{Flex} - 1.637; \\
N = 34, R^2 = 0.954, Q^2_{\text{LOO}} = 0.938; SD = 0.295; F = 207.7
\]

(11)

Because the calculation of \(S_m\) and Flex is easy, the accurate prediction of some properties from log\(P_{16}\) or log\(P_{\text{oct}}\) is easy for the researchers across various fields. For example, log\(K_{\text{brain}}\) can be accurately predicted from log\(P_{16}\) without the need for additional experimental data or complicated calculations. Without using the LFER, the accurate prediction of log\(K_{\text{brain}}\) from log\(P_{16}\) or another organic solvent/water partition coefficient is difficult because there is little correlation between log\(K_{\text{brain}}\) and organic solvent/water partition coefficients. For the models containing H\(_{M, \text{HBD}}\), the H\(_{M, \text{HBD}}\) values of solutes are calculated with computer software. All the molecular descriptors in the LFER are easy to be understood and used by the researchers in various research fields. However, when constructing QSRR models by using mathematical and statistical tools, the predictive variables are usually selected from a few thousand molecular descriptors. The meanings of many predictive variables, e.g., the 3D-MoRSE descriptors, (Zapadka et al., 2019), are not easy to be understood or used by many researchers in various research fields.

Performance of Models With all Molecular Descriptors Calculated From Solute Structures
Because log\(P_{\text{oct}}\) and log\(P_{16}\) can be calculated accurately from the structures of solutes (Chen et al., 2020), it is expected that this method still performs well when all of the molecular descriptors in the LFER are calculated from solute structures. For example, the R\(^2\), Q\(^2\) and SD of the model for predicting human skin permeability, in which all predictive variables are calculated from solute structures, are 0.940, 0.957, and 0.202 (see Supplementary Text S4). Thus, the general LFER developed in this study has obvious advantages in predicting many properties related to noncovalent interactions.

Importance of Thermodynamics-Based Theoretical Derivation
Above examples indicate that the models constructed according to the LFER for many specific properties have high predictive power. Moreover, the performance of the models is independent of the compounds for investigation, suggesting that the models can provide guidance for improving properties of organic compounds and designing compounds with optimal properties. The merits of the LFER result from the theoretical derivation, which ensures that the quantitative relationships in the models constructed according to the LFER are correct in the aspect of thermodynamics. For the QSRR models developed using mathematical and statistical tools, the predictive variables are selected from a few thousand molecular descriptors based on the data of the compounds in training sets. The relationships between the properties and molecular descriptors in the QSRR models are statistical relationships for the compounds in training sets. The QSRR models usually work well only for the compounds in the training set and similar compounds, but may do not work well for other compounds. Thus, for the properties determined by the noncovalent interactions of solutes with flexible environments, the models developed according to the proposed LFER performs better than the QSRR models developed by using mathematical and statistical tools, including robust artificial neural networks. Developing models according to the proposed LFER is faster and computationally cheap than developing traditional QSRR models because the process of the variable selection is not required. Moreover, the proposed LFER is quite simple and can be easily used by the researchers across various fields, while expert knowledge is required for developing robust artificial neural networks, such as the knowledge in choosing the most appropriate approach. Thus, the method developed in study has obvious advantages over the traditional QSRR construction.
method. Thermodynamics-based theoretical derivation can be used to solve many problems that are hard to be solved by using mathematical and statistical tools. In addition, results in this study demonstrate that there are quantitative relationships between the properties related to thermodynamics, suggesting that many properties can be accurately predicted from other properties.

**Future Works**
The theoretical derivation in this study is based on the assumption that solutes have similar interactions with their environments, which requires that the environments for the properties are flexible or the properties have little relationship with the conformation or orientations of solutes. Thus, the present LFER may not work well in predicting the binding affinities of ligands because the binding sites of proteins are not flexible. If the environment for a property is rigid (e.g., the binding sites of proteins), the model for predicting the property should consider H-bond interactions individually, rather than the overall H-bond interactions. In our further study, we will explore how to develop models for the properties related to rigid environments, which can be used to develop scoring functions for predicting protein-ligand binding affinities and develop QSAR models for screening databases of ligands. Furthermore, in this study, we demonstrated to effectiveness of the LFER for predicting the properties of neutral organic compounds. If a dataset contains ionizable compounds, it will be necessary to include molecular descriptors for the ionized forms. Although several approaches currently exist for considering the effects of ionization on various molecular properties (Li et al., 2006; Zhang et al., 2017), our future work will involve attempts to adapt the proposed LFER for use in these situations.

**CONCLUSION**
In this study, we used a thermodynamics-based theoretical derivation to develop a general LFER for accurately predicting various properties from partition coefficients. The theoretical derivation ensures that many specific properties can be accurately quantified with the molecular descriptors in the LFER. It overcomes the shortages of constructing QSPR models by using mathematical and statistical tools. It is expected that the thermodynamics-based theoretical derivation can be used to solve many difficult problems, including the accurate prediction of protein-ligand binding affinities.

**DATA AVAILABILITY STATEMENT**
The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

**AUTHOR CONTRIBUTIONS**
DC and XH conceived and designed the study. YF and XH built their models, validated the models. DC and YF wrote the first draft of the manuscript. All authors read, revised and approved the submitted version.

**FUNDING**
This work is supported by grants from the National Natural Science Foundation of China (21763002, 21473041) and the Natural Science Foundation of Jiangxi Province (20202ACBL203008).

**SUPPLEMENTARY MATERIAL**
The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2021.737579/full#supplementary-material
