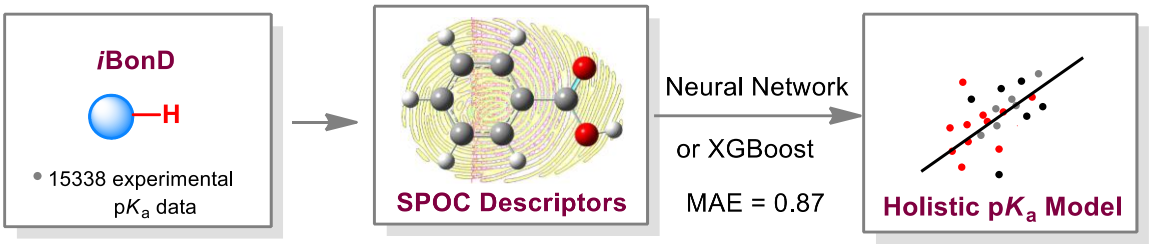
Holistic Prediction of pK_a in Diverse Solvents Based on Machine Learning Approach

Qi Yang, Yao Li, Jin-Dong Yang, Yidi Liu, Long Zhang*, Sanzhong Luo,* Jin-Pei Cheng

Center of Basic Molecular Science, Department of Chemistry, Tsinghua University, Beijing, China, 100084

E-mail: luosz@tsinghua.edu.cn, zhanglong@tsinghua.edu.cn



ABSTRACT: The acid dissociation constant pK_a dictates a molecule's ionic status, and is a critical physicochemical property in rationalizing acid-base chemistry in solution and in many biological contexts. Although numerous theoretic approaches have been developed for predicating aqueous pK_a, fast and accurate prediction of non-aqueous pK_as has remained a major challenge. On the basis of *i*BonD experimental pK_a database curated across 39 solvents, a holistic pK_a prediction model was established by using machine learning approach. Structural and physical organic parameters combined descriptors (SPOC) were introduced to represent the electronic and structural features of molecules. With SPOC and ionic status labelling (ISL), the holistic models trained with neural network or XGBoost algorithm showed the best prediction performance with MAE value as low as 0.87 pK_a unit. The holistic model showed better performance than all the tested single-solvent models (SSMs), verifying the transfer learning features. The capability of prediction in diverse solvents allows for a comprehensive mapping of all the possible pK_a correlations between different solvents. The *i*BonD holistic model was validated by prediction of aqueous pK_a and micro-pK_a of pharmaceutical molecules and pK_as of organocatalysts in DMSO and MeCN with high accuracy. An on-line prediction platform (http://pka.luoszgroup.com) was constructed based on the current model.

KEYWORDS: pK_a prediction, machine learning, iBond, XGBoost, neural network, organocatalyst

INTRODUCTION

The acid dissociation constant, pK_a, dictates the extent to which a proton dissociates from a molecule. As a fundamental physicochemical parameter defining heterolytic X-H bond (X = C or other heteroatoms) cleavage free energies, pK_a is also the primary constant for the derivation of other bond energies such as BDE, BDFE, pK_a ($HA^{+\cdot}$), hydride affinity, which provides largely the quantitative basis for the rational evolution of chemistry.[1] Hence, accurate measurement and prediction of pK_a have been actively pursued in chemical and medical science and continue to be research focus echoing the increasing demand on rational design and development.[1c-1e] Historically, pK_a was first measured under aqueous media and large quantity of data along this line has been accumulated ever since largely driven by its fundamental importance in assessing the ionic status of biomolecules and pharmaceutical molecules.[2] As heterolytic constant, pK_a is highly sensitive to solvents. Non-aqueous pK_as, critical for rationalization of acid/base catalysis, energy materials and ADMET of drug molecules in solution, are profoundly different from their aqueous counterparts, as a result of the varied solvation behaviors between a molecule and its derived ions. However, data on non-aqueous pK_as are rather scarce. Analysis of the most extensive public pK_a database[3] revealed less than 30% non-aqueous experimental pK_as distributing across more than 40 non-aqueous solvents (Figure 1). Many efforts have been devoted to rescale or transfer the aqueous pK_a to its organic solvent counterparts,[4] however, the successes along this line were unfortunately very limited and only applicable to those closely analogue compounds.

In the pursuit of highly accurate pK_a prediction, quantum mechanical (QM) computation have been extensively explored and could now reach prediction accuracy of MAE <1 pK_a unit in organic solvents,[5] however, the QM method is time and resource exhausting process particularly for large molecules with more than 50 atoms. On the basis of traditional QSAR (quantitative structure activity/property relationship) strategy,[6] machine learning (ML) algorithms such as random forest, extreme gradient boosting (XGBoost), support vector machine (SVR), neural network (NN), etc. have recently been explored in the predication of pK_a using either public or industrial data as training sets. [7] The state-of-the-art ML models such as Bayer's "S+pK_a" (incorporated in the commercial ADMET Predictor software[8]) could reach a mean average error (MAE) below 1 pK_a unit.[9] The Bayer's model was developed using the so-called Artificial Neural Network Ensembles (ANNEs) on 10 pre-classified libraries of compounds. However, all these ML models are only applicable for aqueous pK_as and ML models for non-aqueous pK_a prediction with reasonable accuracy remain virtually underdeveloped. Very recently, Grzybowski and coworkers developed a prediction model for pK_a in DMSO with MAE 2.1 pK_a units by using graph convolutional neural networks (GCNNs), the model was trained with a small data set (817 pK_as) composed with half experimental data and half DFT calculated ones and is limited to only C-H acidity prediction in DMSO.[10]

In the past decade, we have engaged in collecting and curating accurate bond energies. In 2016, we established a user-friendly internet-based databank of pK_a and BDE data: *i*BonD,[3] which is freely available at http://ibond.nankai.edu.cn. The *i*BonD covers more than 30000 experimental pK_a data for about 20000 compounds in 46 different solvents. With these data in hand, we have developed a holistic model for predicating both aqueous and nonaqueous pK_a using machine learning algorithm without pre-classification of compounds. The *i*BonD model could provide an accurate prediction over 39 solvents with MAE 0.87 pK_a unit. The model could be successfully applied in the prediction of microstate pK_as of pharmaceutical compounds in water and also in the scaling of organocatalysts in DMSO and acetonitrile. We'll present the details on construction and applications of this model in this full article. The model is also incorporated into a Web version freely available at http://pka.luoszgroup.com, which we hope will become a useful tool for rationalization of acid-based equilibration and reactions in solution.

RESULTS AND DISCUSSION

1. The dataset

Our holistic model was constructed based on experimental data collected in *i*BonD database. Over the years, the data in *i*BonD has underwent rigorous quality and consistency check: duplicates were removed, errors in structures were corrected and doubtful values were either double-checked with original sources or verified by independent experimental measures. In this study, the pK_a dataset in *i*BonD was further cleaned and curated according to the following principles: 1) According to the solvent leveling effect,[11] pK_a data out of each solvent's pK_a window are excluded from the dataset, and so as for those data were mainly obtained by extrapolation of the experimental data, hence containing large errors; 2) since experimental data extracted from different resources may vary as recorded in *i*BonD, for each molecule with multiple recorded entries, an average pK_a was used as the output if the variation is less than 2.0 pK_a unit; otherwise the molecule was removed from the training set; and 3) outliers were picked for further scrutiny during the process of model training and in this way we have identified more than 100 errors, mostly arisen from erroneous drawings of structures or typos in pK_a values. For cases that no clear-cut rationalization on the large pK_a deviation could be made, the molecules were excluded from the training dataset. The extensive and rigorous vetting process ensures high quality and credibility of the experimental data in *i*BonD to be used for model training.

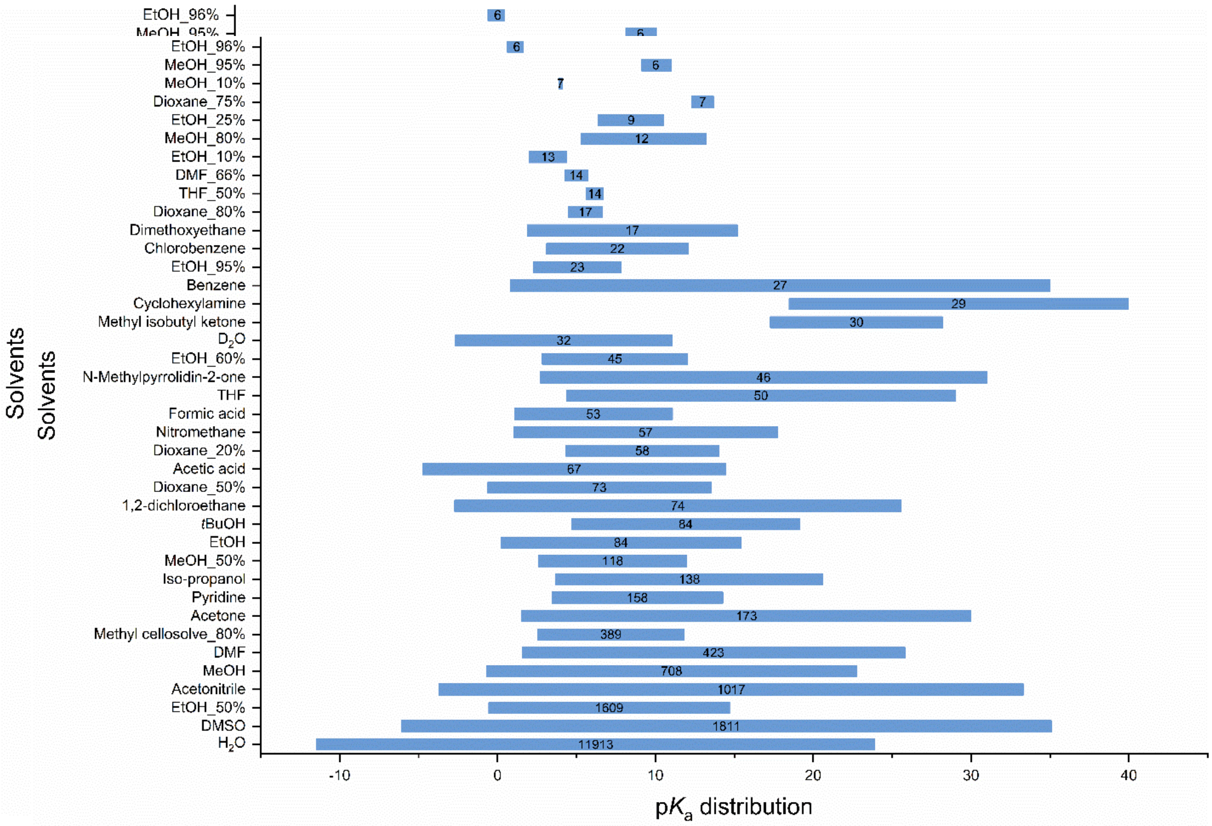


Figure 1. The pK_a distribution in different solvents. The blue bar denotes the spanning pK_a range in each solvent and the amount of molecules are shown in the blue bar.

Eventually, we reached a curated dataset containing 15338 chemical compounds with a total of 19397 pK_a values in 39 solvents. The most applied solvents with more than 400 pK_a data include water (61.7%), DMSO (9.3 %), EtOH/H_2O (1:1) (8.2 %), CH_3OH (3.6 %), CH_3CN (5.2 %) and DMF (2.2 %) and the aqueous pK_a data are dominant (Figure 1). There are 487 compounds with pK_a values available in both of the top 2 solvents H_2O and DMSO, 187 compounds in H_2O, DMSO and MeOH, and meanly 3 compounds (acetic acid, protonated pyridine and 4-chloro-3-nitrobenzoic acid) with data available in all the top 6 solvents, indicating the severe scarcity of non-aqueous pK_as.

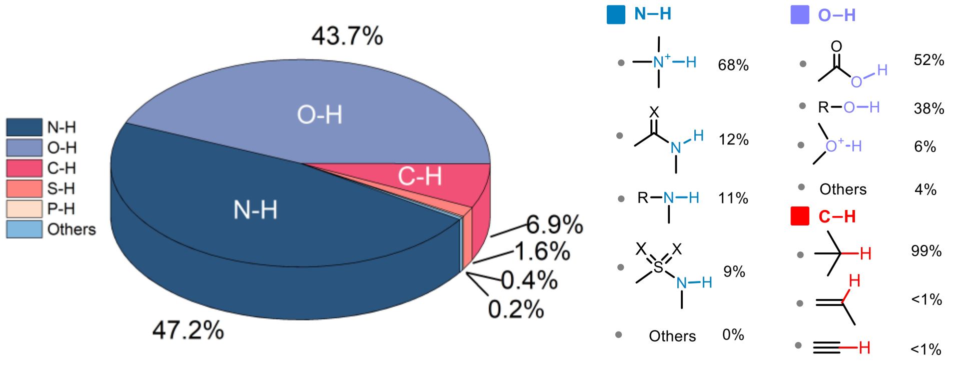


Figure 2. Category distribution of the database.

Structurally, the curated *i*BonD dataset consists of mainly N-H (47%), O-H (44%) and C-H (7%) pK_as and also S-H, P-H and minor other constants (2% in total) (**Figure 2**). The sub-categories of N-H acidity include the pK_as of protonated amines (68%), sulfonamides (12%), amines (11%) and amides (9%). As for the O-H acidity, the carboxylic acids are dominant (52%) together with 38% for alcohols and phenols as well as minor protonated O-H (6%). For C-H pK_as, the most experimentally determined ones are for C(sp^3)-H (99.3%) with minor C(sp^2)-H (0.4%) and C(sp)-H (0.3%).

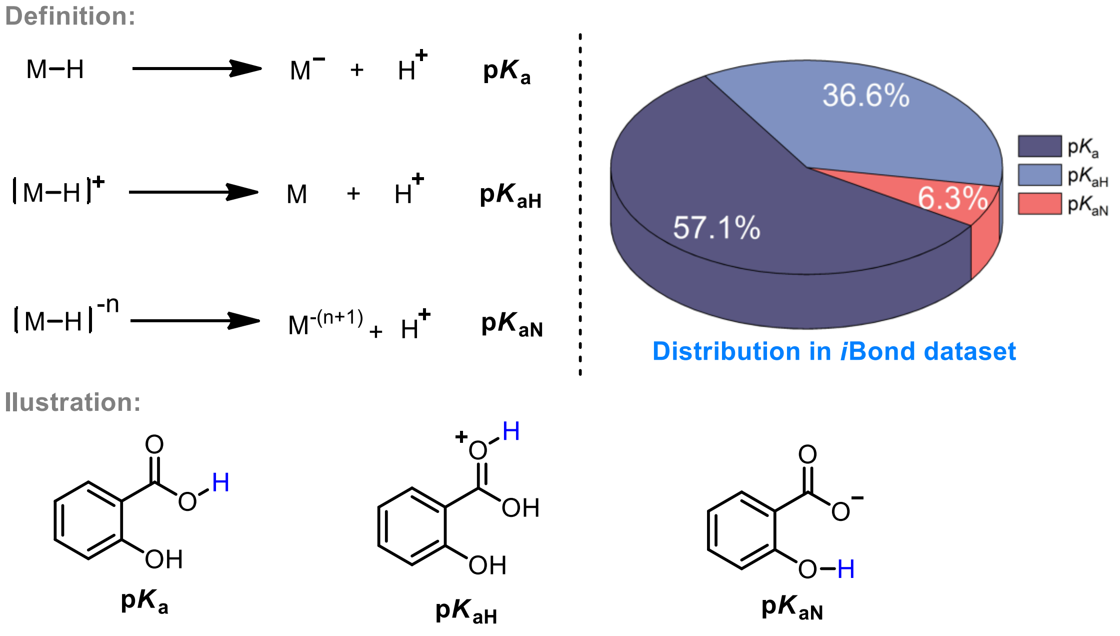
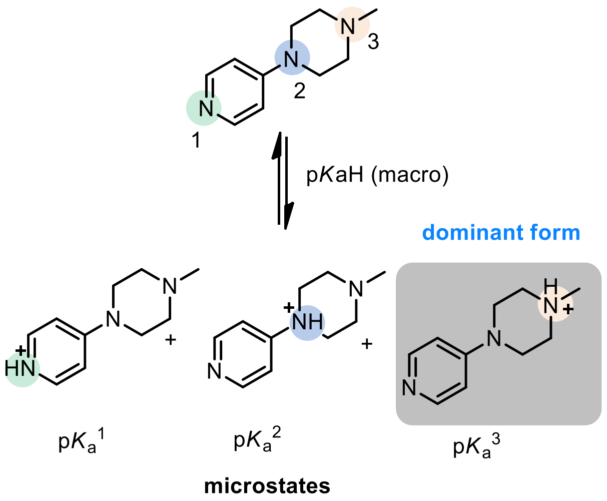


Figure 3. Definition of pK_a subtype on the basis of ionic status.

In *i*BonD, each experimental pK_a value was assigned to the major responsible ionization site when possible, hence the resulted library of compounds in the curated *i*BonD pK_a dataset is composed of >50% neutral molecules (ca. 11105 pK_a values), >1/3 protonated molecules (ca. 7104 pK_a values), and a small portion of negatively charged molecules (1224 pK_a values). The latter two types of ionic molecules were annotated as pK_{aH} and pK_{aN}, respectively (Figure 3). Particularly, pK_{aH}, an indication of basicity of a molecule, refers to the dissociation constant of its conjugate acid. A primary concern is the appropriate identification of the site of protonation/deprotonation in regard to the pK_a value. For charged molecules, additional challenge arises when the molecule contains multiple ionization (or titratable) sites. In these cases, the recorded experimental data are macroscopic pK_a values, *i.e.* the acid dissociation constant of a given molecule regardless of individual titratable site (Scheme 1). The acidity of each of the titratable group is known as microscopic pK_a, and the experimentally determined macroscopic pK_a is the net results of equilibration of each microstate according to equation 1 (Scheme 1). For these multi-ionizable charged molecules (both positively and negatively charged ones), we restrict our structural designation to the dominant microstate, of which the microscopic pK_a can serve as a good approximation of the macroscopic constant. We have designated more than 2000 such multi-ionizable compounds in the curated *i*BonD dataset. In a few cases that dominant microstates cannot be unambiguously assigned, the molecules were excluded from the dataset.

In addition, the average Tanimoto similarity was also tested based on comparisons of Morgan Fingerprints[12] with radii = 2 of all possible pairs of molecules, only 0.18 was found for the entire set, verifying its structural diversity.



$$\mathrm{p}K_{\mathrm{aH}}(\mathrm{macro}) = \log\left(\sum_{1}^{i} 10^{\mathrm{p}K_{a}^{i}}\right) \qquad eq.1$$

Scheme 1. The relationship between macro- and micro-pK_a

2. Strategies and methods

A workflow for model training is depicted in Figure 4. Firstly, the suitable molecular descriptors were collected to represent the features of the molecules. The acidity of a given compound was mainly determined by its structure and consequently physicochemical properties. Therefore, we introduced in this study Structure and PhysicoChemical (denote also Physical Organic Chemistry) properties combined descriptors (**SPOC**). The **SPOC** was generated by taking into consideration of their general applicability as well as the computation cost. Readily available molecular fingerprints such as MACCS[13], Estate[X] and Morgan fingerprints[X] , and physicochemical descriptors extracted from the RDKit were screened to account for the electronic and fragment features of molecules. By doing so, the selection-bias can be largely minimized and the expected **SPOC** for each molecule can be generated in milliseconds. The molecules were also annotated regarding their ionic status with respective to neutral (pK_a), positively charged (pK_{aH}) or negatively charged (pK_{aN}) (named as ionic status labeling (**ISL**)) (Figure 3).

A holistic model (HM) was trained over the entire curated *i*BonD dataset in all the 39 solvents. The HM would address the issues associated with the scarcity of data in organic solvents in model training. For comparison, single solvent models (SSM) for the top six most used solvents (H_2O, DMSO, EtOH/H_2O(1:1), acetonitrile, MeOH and DMF) were also constructed using the individual pK_a sub-dataset in each of the six solvents. The following model training was performed with a range of algorithms such as Tree Regression, Random Forest, Gaussian Process, Gradient Boosting, Support Vector Machine (SVM), K-nearest Neighbors (KNN), Ada Boost, Bagging Tree and Extra Tree. The Neural Network (NN) was also examined considering its powerful ability in identifying non-linear patterns. In addition, the XGBoost algorithm, which is known for its high efficiency and accuracy, was also tested. Once the optimal algorithms were identified, data curation was first conducted in the preliminary rounds of model training. The final HM model was reached on the curated dataset.

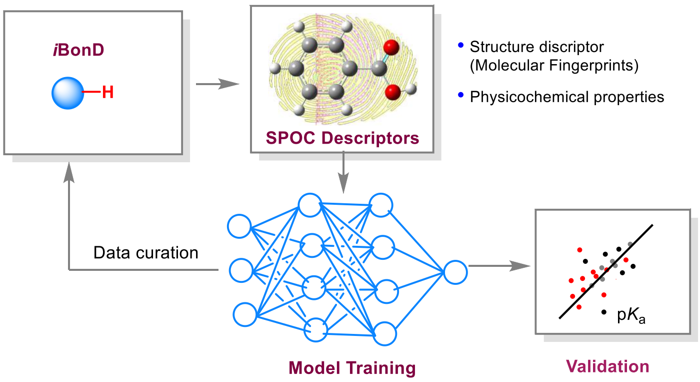

*i*BonD
H
SPOC Descriptors
• Structure discriptor (Molecular Fingerprints)
• Physicochemical properties
Data curation
Model Training
pK_a
Validation

Figure 4. Work flow for the construction of *i*BonD pK_a model.

3. Model Training

Selections of Descriptors: Selecting suitable descriptors is very critical for the pK_a prediction task. in this study, we tried to use several molecular fingerprints (MF) such as Morgan fingerprints with 2 or 3 radii, estate and MACC to account for the structural information. In addition, the molecular physicochemical descriptors were generated using RDkit to account for the physical organic chemical (POC) information. The XGBoost algorithm was used to evaluate their prediction performance in a holist model. It is found that the sole use of either MF or POC descriptor led to RMSE = 1.91 and 1.82, respectively, with the latter performed slightly better (**Figure 5** and SI). Among all the other molecular fingerprints screened, MACC fingerprints showed the best performance. Delightfully, the combination of MACC with POC, the **SPOC**, could improve the prediction with RMSE = 1.70, R^2 = 0.92. Inclusion of the ionic status label **(ISL)** in this **SPOC** descriptor could further improve the results to RMSE = 1.50, R^2 = 0.94. The following model construction were performed with **SPOC-ISL** descriptors.

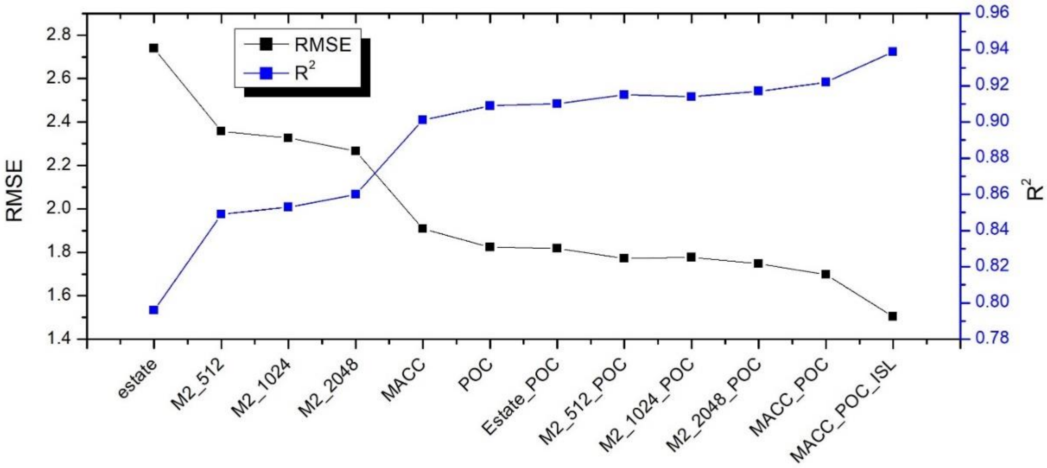


Figure 5. The selection of descriptors based on holistic model training with XGBoost.

Screening of algorithms. To find the most suitable ML algorithm for the pK_a prediction task of HM, a 5-fold cross validation (CV) was performed based on the whole dataset. The average results of the five runs were used as the CV statistic. Among different algorithms examined (**Figure 6A**), full collected neutral network (NN) with three hidden layers was found to give the best results with RMSE = 1.41, MAE = 0.87 and R^2 = 0.95 **(Figure 6C**). To avoid overfitting, the early stopping was used to halt the training of neural network. The XGBoost algorithm also gave comparable results with RMSE = 1.50, MAE = 0.88 and R^2 = 0.94 (**Figure 6B**). Other boosting methods such as Gradient boosting and Ada boost showed slightly poorer performance, The Gaussian process was also a good choice, with RMSE = 1.59, MAE = 0.92 and R^2 = 0.93. Based on these results, the following model training and validation were performed with NN and XGBoost algorithm.

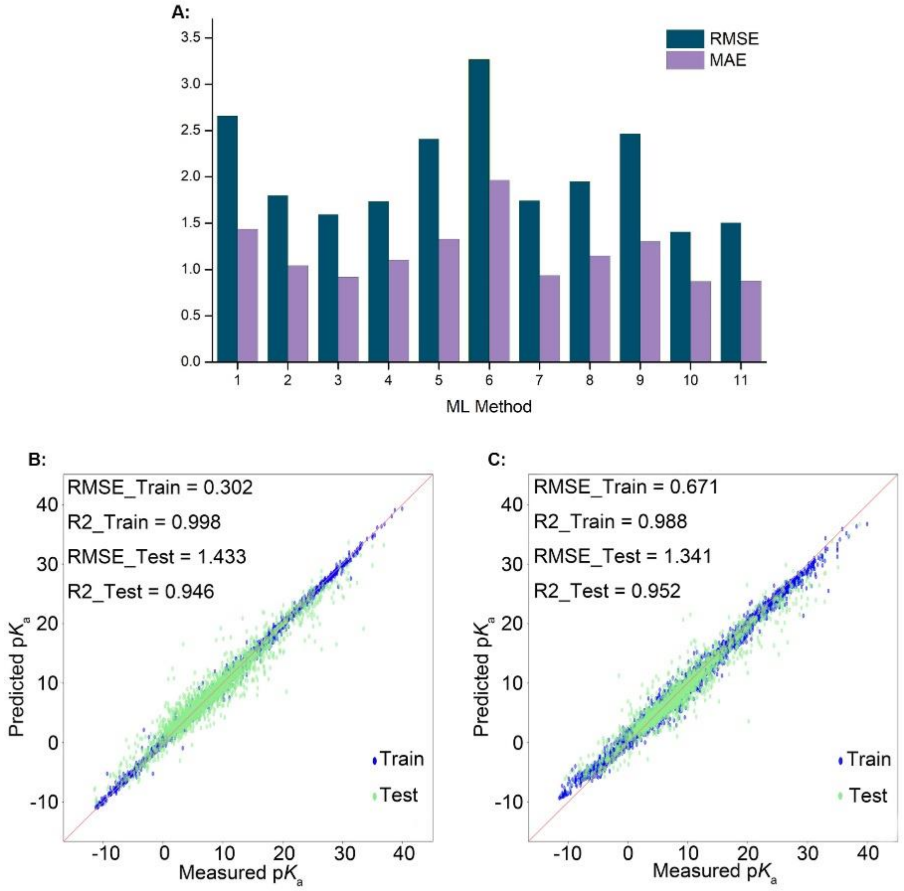


Figure 6. The 5-fold cross-validation of **HM** with different ML methods (**A**): 1-Tree regression; 2-Random Forest; 3- Gaussian Process; 4-Gradient Boosting; 5-SVM; 6-KNN; 7-Ada Boost; 8-Bagging Tree; 9-Extra Tree; 10-NN; 11-XGBoost; the performance of XGBoost (**B**) and Neutral Network (**C**) model.

Comparison of SSM and HM. As experimental pK_a values in organic solvent are only sparsely available, the training of SMM in these solvents would suffer from lacking of data with high risk of overfitting. For the holistic model in 39 solvents, it is expected that the hidden relationship between pK_a values in different solvents could be gleaned during the model training. To verify this hypothesis, we reorganized the pK_a data in the most widely used six solvents by randomly splitting into 80:20 ratio of training and testing subsets, and the individual training and testing subsets were then separately combined to train a holistic model **HM-**

6S (**Figure** 7, **A**). As clearly illustrated in Figure 7B, **HM-6S** performed better than all the six **SMM**s with the same testing set. **HM-6S**, with MAE = 0.89, was comparable with the holistic model for 39 solvents. These results indicated that the **HM** was more robust than the **SSM**, and some structural and physiochemical features among molecules in different solvents were indeed learned during the model training process.

It is worthy to note that MAE of **SMM** models varies dramatically, and exceptional large MAE was observed in DMSO and acetonitrile. This can be rationalized by considering the varied range of pK_a values between different solvents due to the solvent-leveling effect. For example, the recorded pK_a ranges of DMSO and acetonitrile in *i*BonD are -6.08~35.1 and -3.7~33.3, respectively. In comparison, those for EtOH/H_2O (1:1) and MeOH were much narrow as -0.54~14.7 and -0.67~22.74, respectively. Indeed, the lowest MAE was observed in EtOH/H_2O with the narrowest pK_a distributions.

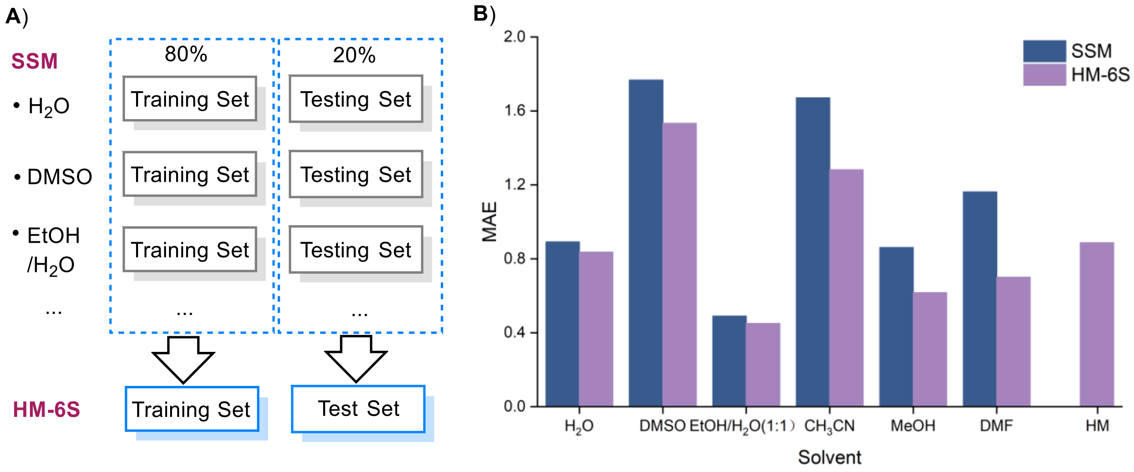


Figure 7. The construction of **SSM** and **HM-6S** model (**A**) and their comparison (**B**).

The pK_a correlation between different solvents. As experimental pK_a values are dominant in aqueous media and to a much less extent also DMSO, the direct scaling and transfer of these data to other solvents have been extensively explored according to the free energy relationship pK_a (solvent 1) = a* pK_a (solvent 2) + b.[14] However, good linear correlations have only been observed in structurally related analogue compounds.[15] The current holistic model **HM** allows us to correlate pK_as between different solvents over a broad range of molecular structures. Hence, pK_as in top six solvents for the entire *i*BonD compounds (15338 compounds) were predicted and analyzed for the intrinsic free-energy relationship between any pair among the six solvents (Table S6, SI). The obtained correlation was also compared with the experimental one if sufficient experimental pK_as are available., A survey of the correlation data revealed the following general trends (**Figure 8**, see also Table S6 in SI): 1) Good linear correlations (with $R^2 > 0.9$) were generally not observed. Protic solvent pairs such as H_2O/EtOH-H_2O (1:1) and MeOH/EtOH-H_2O (1:1) showed good correlations due to their similar solvation behaviors. This similarity has been used to estimate the pK_a value in pure water *via* Yasuda-Shedlovsky extrapolation.[16] As for the aprotic polar DMSO-DMF pair, their properties are extremely similar and the correlation with experimental data also give a perfect linear relationship (R^2 =0.98, Figure S11 in SI); 2) like-solvents correlate better than the like-unlike pairs do. Among the six examined solvents (protic: H_2O, EtOH/H_2O and MeOH; aprotic: DMSO, MeCN and DMF), DMSO-H_2O combination shows a rather poor correlation with $R^2 = 0.66$. In comparison, the R^2 of DMSO-MeCN is 0.74. The H_2O-DMSO-MeCN 3D-plot shows much scattered distribution than that of H_2O-MeOH-EtOH/H_2O plot (**Figure 8**, **A** vs **B**). Similar pattern was also observed between DMSO-MeCN-EtOH/H_2O and DMSO-DMF-MeCN (**Figure 8**, **C** vs **D**); 3) Generally, the predicated correlations are in good consistence with the experimentally determined ones with slightly better R^2. Large deviation was observed for the DMSO-EtOH/H_2O, MeCN-DMF and MeOH-DMF solvent pairs (Table S6, entries 8, 13 and 15). These exceptions were mainly caused by the few experimental data found in the dataset and its irregular distribution (See Figure S9, 14 and 16, SI). In all, these observations clearly showed that the inherent link and different solvation behaviors have been learned during the training of the holistic model.

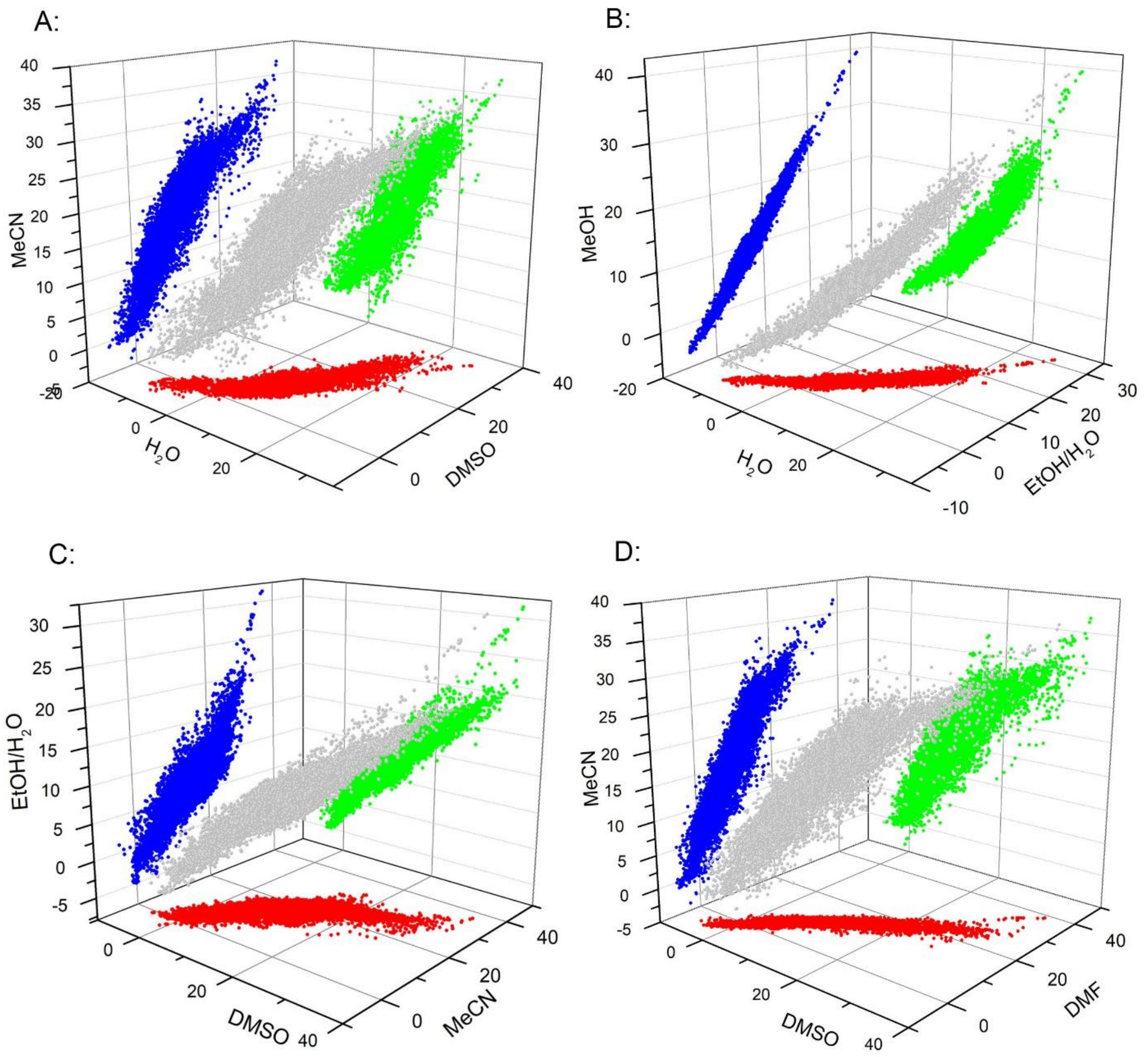


Figure 8. The 3D-scatter plot of pK_a data in different solvents, **A**: MeCN-H_2O-DMSO; **B**: MeOH-H_2O-EtOH/H_2O (1:1); **C**: EtOH/H_2O-DMSO; **D**: EtOH/H_2O-DMSO-DMF (Grey: Scatter plot in 3D-space; Red: Projection on XY-plane; Green: Projection on XZ-plane; Blue: Projection on YZ-plane).

4. Verification and application of the model

The obtained holistic model (**HM**) was further tested in out-of-sample predications and three representative applications are shown herein.

Macro and micro-pK_a prediction of pharmaceutical molecules in H_2O. In the year of 2017, a blind pK_a prediction challenge named SAMPL6 has been established by the Drug Design Data Resource Community,[17] which consists of predicting microscopic and macroscopic pK_a of 24 drug-like small molecules (17 drug-fragment-like and 7 drug-like). The submitted prediction strategies included quantum-chemistry based calculation,[18] EC-RISM theory,[19] QM/MM approach,[20] ab initio quantum mechanical prediction[21] as well as machine learning.[22] The machine learning methods were built with the general Gaussian process with unfortunately only moderate accuracy. Based on the HM-XGBoost model, we obtained a prediction with MAE = 0.80 and RMSE = 1.07 (Figure 9 and Figure S17A). The prediction with holistic NN model also gave comparable results (See SI, Figure S17B), with slightly higher MAE and RMSE. In contrast, the prediction with the **SSM-H₂O** model gave a higher MAE of 0.92, further confirming the robustness of the **HM** model. Significant overestimate for SM14 (ΔpK_a = 2.45) and SM24 (ΔpK_a = 1.58) were observed, it was deduced that the large deviations come from the doubly protonated status for which the examples in the training data are extremely rare in *i*BonD. Large deviations were also observed for SM03 (+2.0), SM05 (+1.99) and SM18a (-3.12) and the reason are unclear, likely related to their strong intra- or intermolecular H-bonding features.

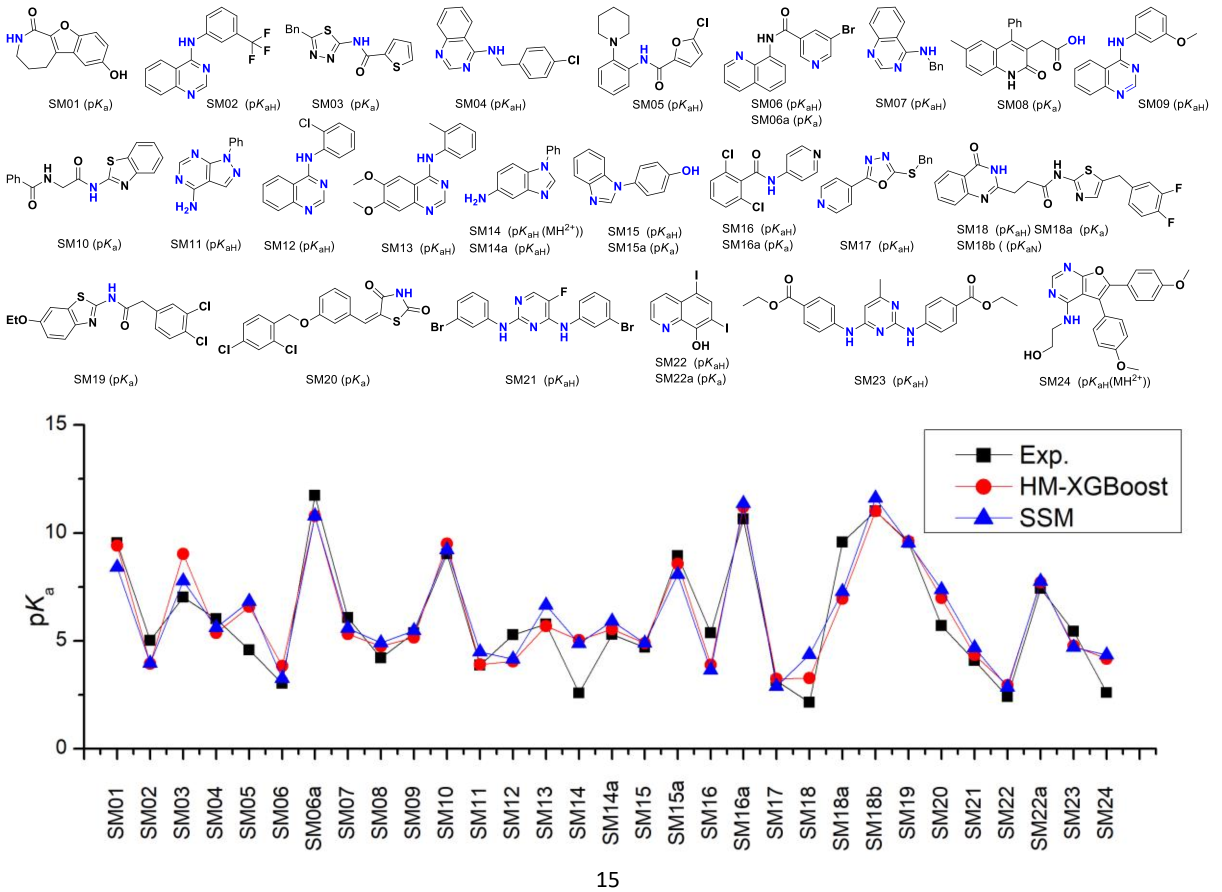


Figure 9. The prediction of SAMPL6 puzzles with the HM-XGboost model

Most biologically active molecules contain multiple protonation/deprotonation sites, their site-specific micro-pK_as are critical in the rationalization of their pharmaceutical profiles. To verify the capability of our holistic model in predicting micro-pK_as, 17 pharmaceutical molecules with multi-dissociation sites, which were not included in the original data set, were selected as external validation set.[23] As shown in **Figure** 10, the results with HM-XGboost are extremely good with the average MAE = 0.44. In these cases, the micro-pK_a of all dissociable sites were predicted independently and the macro-pK_a were calculated according to equation 1 (Scheme 1). The micro-pK_a provides a definitive assignment of the protonation sites as well as the determination of the equilibrium distributions. For example, for Thenalidine, Aprindine and Methaphenilene, protonation would mainly occur on the tertiary amine moieties with few on the aromatic nitrogen (< 0.05%) under physiologic pH. While for Metoclopramide, though protonation onto primary amine is still dominant, protonation on aromatic nitrogen is unneglectable and constitutes about 35%.

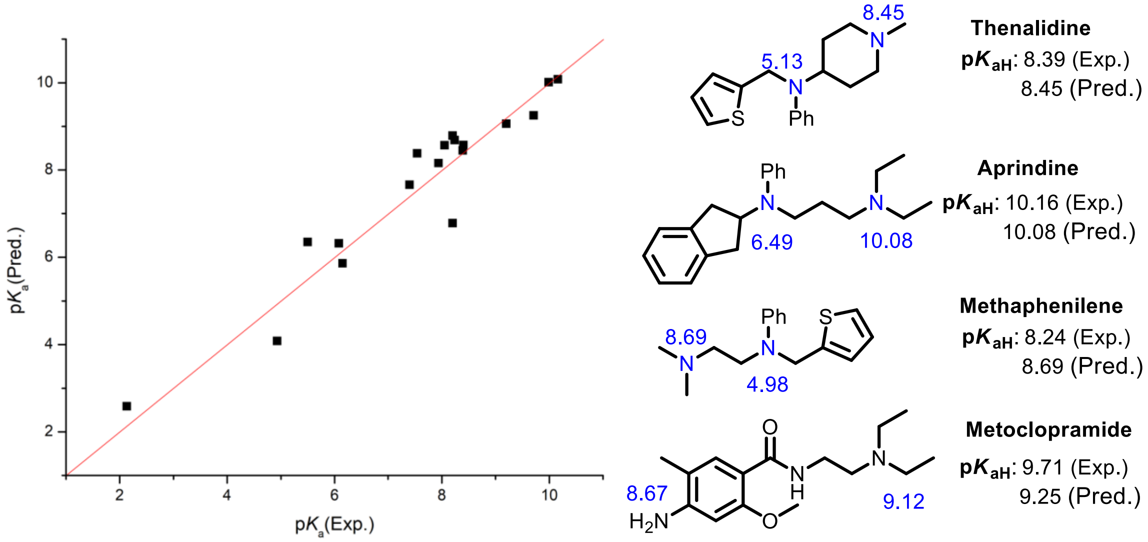


Figure 10. The pK_a prediction of pharmaceutical molecules with HM-XGboost, micro-pK_as were shown in blue.

Prediction and scaling of organocatalysts acidity in DMSO. Hydrogen-bonding (HB) moieties, ubiquitous structural units in chiral organocatalysts, are critical in tuning both activity and stereoselectivity via hydrogen-bonding interaction or acid catalysis to a broad sense. Acidity, represented by pK_as, is one of the key physicochemical parameters that define HB catalytic properties. We and others have experimentally determined the pK_as for a series of organocatalysts including thiourea, squaramide, BINOL, prolinamides and 6'-hydrogen bonding cinchona alkaloids (6'-HBCA) in DMSO. [24] In total, an external set involving 84 organocatalysts' pK_as, not included in the training set, were tested with **HM-XGBoost**. As shown in Figure 11, the prediction of thioureas and prolinamides gave extremely accurate results, with MAE = 0.16 and 0.23 respectively. The prediction of squaramides also give good results, with MAE = 0.64. These predictions even surpassed the results obtained with DFT calculations.[25] In contrast, the prediction of Binol-type hydrogen bond catalysts showed relatively poor accuracy and most of the pK_a value were overestimated. It should be noted that the internal H-bonding between the two phenol groups may significantly enhance the acidity of the molecule, however, the HM model likely failed to identify this feature due to lack of examples in the training set. As for the 6'-HBCA, the prediction is slightly worse than mean accuracy, possibly caused by the partially zwitterionic form for these molecules. The overall MAE for the five types of organocatalysts was 1.01.

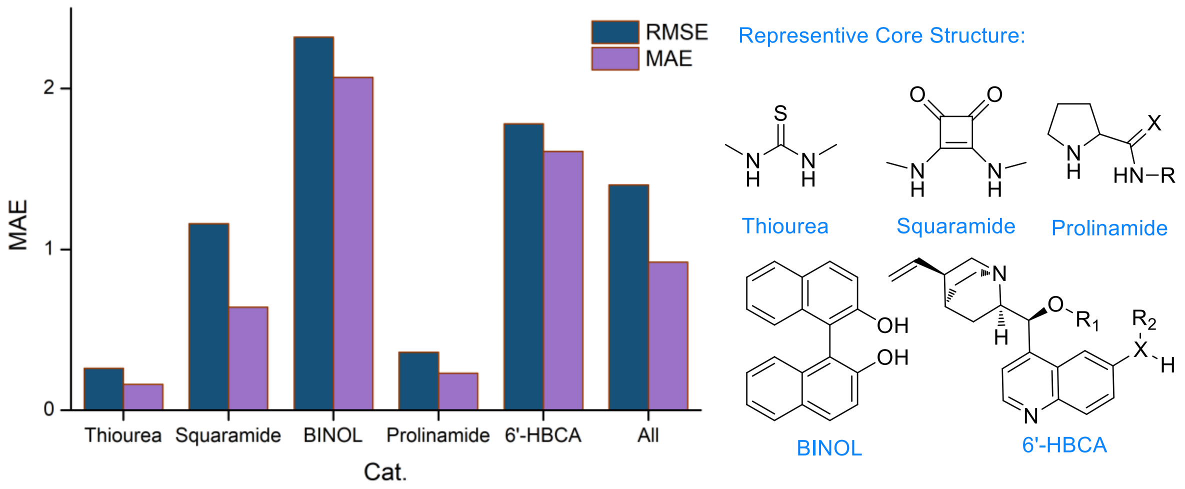


Figure 11. The pK_a prediction of organocatalysts.

Prediction and scaling of aminocatalysts basicity in MeCN. The nucleophilicity of aminocatalysts is a key parameter that defines their catalysis in enamine or iminium-based transformations. The pK_{aH} of aminocatalyst is frequently employed as a rough estimation of its nucleophilicity. Recently, Mayr and coworkers have experimentally determined the pK_{aH} of a series of commonly used secondary aminocatalysts, including those privileged ones. [26] The acidity of these protonated molecules ranges from 10.54 to 24.02. These data were used as out-of-sample test for our model. To our delight, the **HM** performed quite well (**Figure** 12) and the overall MAE was found to be 1.57. Considering that only 1017 pK_a records in MeCN can be found in the original datasets, the slightly large MAE was reasonable and acceptable. The imidazolinones and silyl substituted pyrrolidines showed large deviations, possibly caused by the lack of examples in training set. It should be noted that most of the predicted values in these cases were underestimated, indicating the influence of the appended substituents to the pyrrolidinyl core was overestimated.

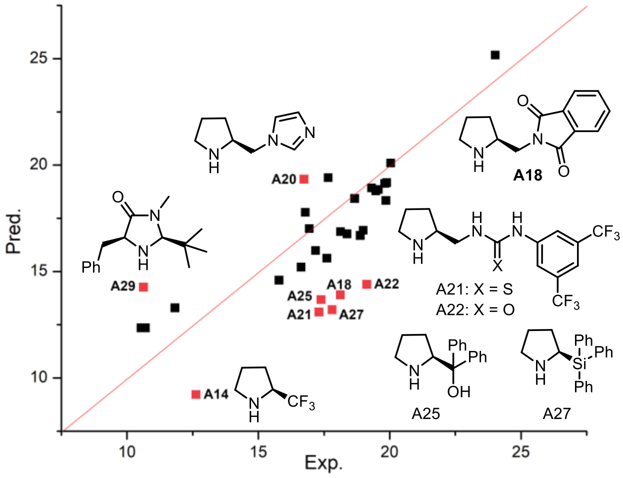


Figure 12. The pK_{aH} prediction of aminocatalysts in MeCN

CONCLUSION

In summary, we have developed a holistic pK_a prediction model (**HM**) based on the *i*BonD experimental pK_a dataset. Structural and physicochemical combined descriptors (**SPOC**) were introduced to represent molecular features and the optimal model was identified with Neural Network and XGboost algorithm, showing accuracy of MAE = 0.87 and 0.88 pK_a unit, respectively. The *i*BonD **HM** model provides so far the best accuracy in prediction non-aqueous pK_as and is equally applied for aqueous and micro- pK_a prediction. The superior performance of **HM** over **SSM** verified the transfer learning from one solvent to another. The capability in predicting pK_as in diverse solvents allows for a comprehensive mapping of the pK_a relationships between different solvents. Finally, the robustness of this HM was validated in out-of-sample predictions of aqueous pK_as of pharmaceuticals, pK_as and pK_{aH} of organocatalysts in organic solvents (DMSO and MeCN) with good accuracy. We also provide a website interface for this pK_a prediction model (http://pka.luoszgroup.com), which we hope will become a useful tool for the scientific community in rationalizing acid-base chemistry.

AUTHOR INFORMATION

Corresponding Author

Luosz@tsinghua.edu.cn

zhanglong@tsinghua.edu.cn

Notes

Any additional relevant notes should be placed here.

ACKNOWLEDGMENT

We thank the National Science & Technology Fundamental Resource Investigation Program of China (2018FY201200), Tsinghua University Initiative Scientific Research Program (2019Z07L01005) and the Natural Science Foundation of China (21672217, 21572232 and 21933008) for financial support. S.L. is supported by the National Program of Top-notch Young Professionals.

REFERENCES

(1) (a) Bell, R. P. The proton in chemistry; Springer Science & Business Media: **2013**. (b) Stewart, R. The proton: Applications to organic chemistry; Academic Press: New York, **2012**. (c) Yang, J.-D.; Ji, P.-J.; Xue, X.-S.; Cheng, J.-P. Recent Advances and Advisable Applications of Bond Energetics in Organic Chemistry. *J. Am. Chem. Soc.* **2018**, *140*, 8611-8623. (d) Xue, X.-S.; Ji, P.; Zhou, B.; Cheng, J.-P. The Essential Role of Bond Energetics in C-H Activation/Functionalization. *Chem. Rev.* **2017**, *117*, 8622-8648. (e) Yang, J.-D.; Xue, J.; Cheng, J.-P. Understanding the role of thermodynamics in catalytic imine reductions. *Chem. Soc. Rev.* **2019**, *48*, 2913-2926.

(2) Manallack, D. T.; Prankerd, R. J.; Yuriev, E.; Oprea, T. I.; Chalmers, D. K. The Significance of Acid/Base Properties in Drug Discovery. *Chem. Soc. Rev*. **2013**, *42*, 485−496. (b) Manallack, D. T. The pK_a Distribution of drugs: Application to drug discovery. *Perspect. Med. Chem.* **2007**, *1*, 25−38. (e) Charifson, P. S.; Walters, W. P. Acidic and Basic Drugs in Medicinal Chemistry: A Perspective. *J. Med. Chem.,* **2014**, *57*, 9701 − 9717. (f) Manallack, D. T.; Prankerd, R. J.; Nassta, G. C.; Ursu, O.; Oprea, T. I.; Chalmers, D. K. A Chemogenomic Analysis of Ionization Constants-Implications for Drug Discovery. *ChemMedChem*, **2013**, *8,* 242-255. (h) Manallack, D. T. The Acid-Base Profile of a Contemporary Set of Drugs: Implications for Drug Discovery. *SAR QSAR Environ Res.* **2009**, *20*, 611-655. (i) Schultz, T.W. The Use of the Ionization Constant (pK_a) in Selecting Models of Toxicity in Phenols. *Ecotox. Environ. Safe.* **1987**, *14*, 178-183. (j) Jeschke, P. Propesticides and Their Use as Agrochemicals. *Pest. Manag. Sci.* **2016**, *72*, 210-225. (k) Clark, R. D. Predicting Mammalian Metabolism and Toxicity of Pesticides in Silico. *Pest Manag Sci.* **2018**, *74*, 1992-2003.

(3) Yang, J.-D.; Xue, X.-S.; Ji, P.; Li, X.; Cheng, J.-P. Internet Bond-energy Databank (pK_a and BDE): *i*BonD Home Page. http://ibond.chem.tsinghua.edu.cn or http://ibond.nankai.edu.cn.

(4) (a) Rossini, E.; Knapp, E. W. Proton Solvation in Protic and Aprotic Solvents. *J. Comput. Chem.* **2016**,

37, 1082−1091. (b) Rossini, E.; Bochevarov, A. D.; Knapp, E. W. Empirical Conversion of pK_a Values between Different Solvents and Interpretation of the Parameters: Application to Water, Acetonitrile, Dimethyl Sulfoxide, and Methanol. *ACS Omega* **2018**, *3*, 1653−1662. (c) Rossini, E.; Netz, R. R.; Knapp, E. W. Computing pKa Values in Different Solvents by Electrostatic Transformation. *J. Chem. Theory Comput.* **2016**, *12*, 3360−3369.

(5) (a) Fu, Y.; Liu, L.; Li R.-Q.; Liu, R.; Guo, Q.-X. First-Principle Predictions of Absolute pK_a's of Organic Acids in Dimethyl Sulfoxide Solution. *J. Am. Chem. Soc.* **2004**, *126*, 814-822. (b) Shen, K.; Fu, Y.; Li, J.-N.; Liu, L.; Guo, Q.-X. What are the pK_a values of C-H bonds in aromatic heterocyclic compounds in DMSO? *Tetrahedron* **2007**, *63*, 1568-1576. (c) Alongi, K. S.; Shields, G. C. Chapter 8 - Theoretical Calculations of Acid Dissociation Constants: A Review Article. *Annu. Rep. Comput. Chem.* **2010**, *6*, 113−138. (d) Ho, J.; Coote, M. L. First-principles prediction of acidities in the gas and solution phase. Wiley Interdiscip. Rev.: *Comput. Mol. Sci.* **2011**, *1*, 649−660. (e) Shields, G. C.; Seybold, P. G. Computational approaches for the prediction of pK_a values; CRC Press: Boca Raton, London, New York, **2013**. (f) Ho, J. Predicting pKa in implicit solvents: Current status and future directions. *Aust. J. Chem.* **2014**, *67*, 1441−1460. (g) Casasnovas, R.; Ortega-Castro, J.; Frau, J.; Donoso, J.; Mun˜oz, F. Theoretical pK_a calculations with continuum model solvents, alternative protocols to thermodynamic cycles. Int. *J. Quantum Chem.* **2014**, *114*, 1350−1363. (h) Seybold, P. G.; Shields, G. C. Computational estimation of pKa values. Wiley Interdiscip. Rev.: *Comput. Mol. Sci.* **2015**, *5*, 290−297. (i) Philipp, D. M.; Watson, M. A.; Yu, H. S.; Steinbrecher, T. B.; Bochevarov, A. D. Quantum chemical pK_a prediction for complex organic molecules. *Int J Quantum Chem.* **2018**, *118*, e25561

(6) (a) Clark, J.; Perrin, D. Prediction of the Strength of Organic Bases. *Q. Rev. Chem. Soc.* **1964**, *18*, 295 – 320. (b) Perrin, D. D.; Dempsey, B.; Serjeant, E. P. pKa Prediction for Organic Acids and Bases. Chapman and Hall/CRC Press, Boca Raton, **1981**. (c) Harris, J. C.; Hayes, M. J. Acid dissociation constant. In

Handbook of Chemical Property Estimation Methods; Lyman, W. J., Reehl, W. F., Rosenblatt, D. H., Eds.; McGraw-Hill, Inc.: New York, **1982**; pp 6.1-6.28. (d) Livingstone, D. J. Theoretical property predictions. *Curr. Top. Med. Chem.* **2003**, *3*, 1171－1192. (e) Rupp, M.; Körner, R.; Tetko, I. V. Predicting the pK_a of small molecule. *Comb. chem. high throughput screen.* **2011**, *14*, 307-327. (f) Lee, A. C.; Crippen, G. M. Predicting pKa. *J. Chem. Inf. Model.* **2009**, *49*, 2013-2033. (g) Fraczkiewicz, R. In Silico Prediction of Ionization. In Reference Module in Chemistry, Molecular Sciences and Chemical Engineering. Elsevier, **2013**.

(7) (a) Jover, J.; Bosque, R.; Sales, J. Neural network based QSPR study for predicting pK_a of phenols in different solvents. *QSAR Comb. Sci.* **2007**, *26*, 385-397. (b) Harding, A. P.; Wedge, D. C.; Popelier, P. L. A. pK_a prediction from “Quantum Chemical Topology” Descriptors. *J. Chem. Inf. Model.* **2009**, *49*, 1914-1924. (c) Jover, J.; Bosque, R.; Sales, J. QSPR Prediction of pK_a for Aliphatic Carboxylic Acids and Anilines in Different Solvents. *QSAR Comb. Sci.* **2008**, *27*, 1204-1215. (d) Chen, B.; Zhang, H.; Li, M. Prediction of pK_a values of neutral and alkaline drugs with particle swarm optimization algorithm and artificial neural network. *Neu. Comput. App.* **2019**, *31*, 8297-8304. (e) Milletti, F.; Storchi, L.; Goracci, L.; Bendels, S.; Wagner, B.; Kansy, M.; Cruciani G. Extending pK_a prediction accuracy: High-throughput pK_a measurements to understand pK_a modulation of new chemical series. *Eur. J. Med. Chem.* **2010**, *45*, 4270-4279. (f) Mansouri, K.; Cariello, N. F.; Korotcov, A.; Tkachenko, V.; Grulke, C. M.; Sprankle, C. S.; Allen, D.; Casey, W. M.; Kleinstreuer, N. C.; Williams, A. J. Open-source QSAR models for pK_a prediction using multiple machine learning approaches. *J. Cheminform.* **2019**, *11*, 60 (g) Zhou T.; Jhamb, S.; Liang, X.; Sundmacher, K.; Gani, R. Prediction of acid dissociation constants of organic compounds using group contribution methods. *Chem. Eng. Sci.* **2018**, *183*, 95-105. (h) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological Torsion: A New Molecular Descriptor for SAR Applications. Comparison with Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82-85. (i) Lu, Y.; Anand, S.; Shirley, W.; Gedeck, P.; Kelley,

B. P.; Skolnik, S.; Rodde, S.; Nguyen, M.; Lindvall, M.; Jia, W. Prediction of pKa Using Machine Learning Methods with Rooted Topological Torsion Fingerprints: Application to Aliphatic Amines. *J. Chem. Inf. Model.* 2019, *59*, 4706−4719.

(8) ADMET Predictor, Simulations Plus, 42505 10th Street West, Lancaster, CA 93534.

(9) Fraczkiewicz R.; Lobell M.; Go¨ller A. H.; Krenz U.; Schoenneis R.; Clark R. D.; Hillisch, A. Best of Both Worlds: Combining Pharma Data and State of the Art Modeling Technology To Improve in Silico pK_a Prediction. *J. Chem. Inf. Model.* **2015**, *55*, 389−397.

(10) Roszak, R.; Beker, W.; Molga, K.; Grzybowski, B. A. Rapid and Accurate Prediction of pK_a Values of C−H Acids Using Graph Convolutional Neural Networks. *J. Am. Chem. Soc.* **2019**, *141*, 17142-17149.

(11) Anslyn, E. V.; Dougherty, D. A. *Modern Physical Organic Chemistry.* **2006,** University Science Books: Sausalito, CA.

(12) Morgan, H. L. The generation of a unique machine description for chemical structures—a technique developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107–113.

(13) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comp. Sci.* **2002**, *42*, 1273−1280.

(14) (a) Cox, B. G. *Acid and Bases, Solvent Effects on Acid-Base Strength.* Oxford University Press, **2013**.

(15) (a) Nicoleti, C. R.; Marini, V. G.; Zimmermann, L. M.; Machado, V. G. Anionic Chromogenic Chemosensors Highly Selective for Fluoride or Cyanide Based on 4-(4-Nitrobenzylideneamine)phenol. *J. Braz. Chem. Soc.* **2012**, *23*, 1488−1500. (b) Klicic´, J. J.; Friesner, R. A.; Liu, S.-Y.; Guida, W. C. Accurate Prediction of Acidity Constants in Aqueous Solution via Density Functional Theory and Self-Consistent Reaction Field Methods. *J. Phys. Chem. A* **2002**, *106*, 1327−1335. (c) Bochevarov, A. D.; Watson, M. A.; Greenwood, J. R.; Philipp, D. M. Multiconformation, Density Functional Theory-Based pKa Prediction in

Application to Large, Flexible Organic Molecules with Diverse Functional Groups. *J. Chem. Theory Comput.* **2016**, *12*, 6001−6019. (d) Klamt, A.; Eckert, F.; Diedenhofen, M.; Beck, M. E. First Principles Calculations of Aqueous p*K*a Values for Organic and Inorganic Acids Using COSMO-RS Reveal an Inconsistency in the Slope of the pKa Scale. *J. Phys. Chem. A* **2003**, *107*, 9380−9386. (e) Eckert, F.; Klamt, A. Accurate Prediction of Basicity in Aqueous Solution with COSMO-RS. *J. Comput. Chem.* **2006**, *27*, 11−19. (f) Muckerman, J. T.; Skone, J. H.; Ning, M.; Wasada-Tsutsui, Y. Toward the Accurate Calculation of pKa Values in Water and Acetonitrile. *Biochim. Biophys. Acta* **2013**, *1827*, 882−891.

(16) (a) Yasuda, M. Dissociation constants of some carboxylic acids in mixed aqueous solvents. *Bull. Chem. Soc. Japan* **1959**, *32*, 429-432. (b) Shedlovsky, T. *The behaviour of carboxylic acids in mixed solvents.* In: Pesce B (ed) Electrolytes. Pergamon Press, New York, pp 146-151, **1962**.

(17) Drug Design Data Resource: https://drugdesigndata.org/about/sampl6

(18) (a) Pracht, P.; Wilcken, R.; Udvarhelyi, A.; Rodde, S.; Grimme, S. High accuracy quantum-chemistry-based calculation and blind prediction of macroscopic pK_a values in the context of the SAMPL6 challenge. *J. Comput. Aided Mol. Des.* 2018, *32*, 1139-1149. (b) Zeng, Q.; Jones, M. R.; Brooks, B. R. Absolute and relative pK_a predictions *via* a DFT approach applied to the SAMPL6 blind challenge. *J. Comput. Aided Mol. Des.* **2018**, *32*, 1179-1189.

(19) Tielker, N.; Eberlein, L.; Güssregen, S.; Kast S. M. The SAMPL6 challenge on predicting aqueous pK_a values from EC-RISM theory. *J. Comput. Aided Mol. Des.* **2018**, *32*, 1151-1163.

(20) Prasad, S.; Huang, J.; Zeng, Q.; Brooks B. R. An explicit-solvent hybrid QM and MM approach for predicting pK_a of small molecules in SAMPL6 challenge. *J. Comput. Aided Mol. Des.* **2018**, *32*, 1191-1201.

(21) Selwa, E.; Kenney, I. M.; Beckstein, O.; Iorga, B. I. SAMPL6: calculation of macroscopic pK_a values from ab initio quantum mechanical free energies. *J. Comput. Aided Mol. Des.* **2018**, *32*, 1203-1216.

(22) Bannan, C. C.; Mobley, D. L.; Skillman, A. G. SAMPL6 challenge results from pK_a predictions based on a general Gaussian process model. *J. Comput. Aided Mol. Des.* **2018**, *32*, 1165-1177.

(23) Prankerd, R. J. *Profiles of Drug Substances, Excipients, and Related Methodology*. Elsevier Academic Press, **2007**, Vol. 33.

(24) (a) Ni, X.; Li, X.; Li, Z.; Cheng, J.-P. Equilibrium acidities of BINOL type chiral phenolic hydrogen bonding donors in DMSO. *Org. Chem. Front.* **2016**, *3*, 1154-1158. (b) Ni, X.; Li, X.; Cheng, J.-P. Equilibrium acidities of cinchona alkaloid organocatalysts bearing 6 '-hydrogen bonding donors in DMSO. *Org. Chem. Front.* **2016**, *3*, 170-176. (c) Li, Z.; Li, X.; Ni, X.; Cheng, J.-P. Equilibrium Acidities of Proline Derived Organocatalysts in DMSO. *Org. Lett.* **2015**, *17*, 1196-1199. (d) Ni, X.; Li, X.; Wang, Z.; Cheng, J.-P.Squaramide Equilibrium Acidities in DMSO. *Org. Lett.* **2014**, *16*, 1786-1789.

(25) Ho, J.; Zwicker, V. E.; Yuen, K. K. Y.; Jolliffe, K. A. Quantum Chemical Prediction of Equilibrium Acidities of Ureas, Deltamides, Squaramides, and Croconamides. *J. Org. Chem.* **2017**, *82*, 10732.

(26) An, F.; Maji, B.; Min, E.; Ofial, A. R.; Mayr, H. Basicities and Nucleophilicities of Pyrrolidines and Imidazolidinones Used as Organocatalysts. *J. Am. Chem. Soc.* **2020**, *142*, 3, 1526-1547.