Chemical Engineering Science 246 (2021) 116995

Contents lists available at ScienceDirect

Chemical Engineering Science

journal homepage: www.elsevier.com/locate/ces

Anomalous diffusion in zeolites

Pan Huang [a,1], Zhijian Yin [b,1], Yun Tian [c,1], Jie Yang [d], Wei Zhong [e], Chunzhong Li [a], Cheng Lian [a,d,*], Li Yang [b,*], Honglai Liu [a,d,*]

[a] State Key Laboratory of Chemical Engineering, Shanghai Engineering Research Center of Hierarchical Nanomaterials, School of Chemical Engineering, East China University of Science and Technology, Shanghai 200237, PR China
[b] Key Laboratory of Green Chemical Process of Ministry of Education, Key Laboratory of Novel Reactor and Green Chemical Technology of Hubei Province, School of Chemical Engineering and Pharmacy, Wuhan Institute of Technology, Wuhan 430205, PR China
[c] Engineering Research Center of Advanced Functional Material Manufacturing of Ministry of Education, School of Chemical Engineering, Zhengzhou University, Zhengzhou 450001, PR China
[d] School of Chemistry and Molecular Engineering, East China University of Science and Technology, Shanghai 200237, PR China
[e] MinJiang Collaborative Center for Theoretical Physics, Department of Physics and Electronic Information Engineering, Minjiang University, Fuzhou 350108, PR China

HIGHLIGHTS

- A universal approach is established to predict anomalous diffusion in zeolites.
- A database of anomalous diffusion in zeolites is first constructed.
- The relation between anomalous diffusion and structure is systematically analyzed.
- The structural parameters are ranked in order of importance on anomalous diffusion.
- 200,000 hypothetical zeolites are predicted by the universal approach.

ARTICLE INFO

Article history:
Received 14 May 2021
Received in revised form 10 July 2021
Accepted 30 July 2021
Available online 03 August 2021

Keywords:
Anomalous diffusion
Zeolite
Molecular dynamics simulations
Machine learning
Structure–property relationship

ABSTRACT

Anomalous diffusion plays an important role in many pivotal chemical engineering processes involving zeolites. However, the structure–property relationships of anomalous diffusion remain unclear, and fast prediction of anomalous diffusion properties is still challenging. Herein, the anomalous diffusion behaviors of light alkanes (methane, ethane and propane) in zeolites are investigated by combining molecular dynamics (MD) simulations with machine learning (ML) method. The Gradient Boosted Regression Trees (GBRT) algorithm is utilized to construct the structure–property relationship from 2200 groups of anomalous diffusion exponent α and anomalous diffusion coefficient D_α calculated by MD simulations. Furthermore, the structural parameters are ranked in order of importance and it is identified that the largest free sphere is the key factor governing anomalous diffusion phenomena. Finally, the method is employed to predict the diffusion behaviors of 200,000 hypothetical zeolites, which provides in-depth understanding of the anomalous diffusion trends in porous materials.

1. Introduction

Zeolites, as a group of crystalline aluminosilicates with microporous structures, have been widely employed as catalysts (Dusselier and Davis, 2018; Shamzhy et al., 2019), ion-exchangers (Wen et al., 2018), and adsorbents (Wang and Peng, 2010) in various practical chemical engineering processes, such as petroleum refining (Primo and Garcia, 2014), petrochemistry (Vogt and Weckhuysen, 2015), fuel cells (Yu et al., 2013), air-pollution remediation (Zhang et al., 2016), biomass conversion (Ennaert et al., 2016) and wastewater treatment (Jiang et al., 2018). Diffusion in the micropores of zeolites is a crucial factor to determine the catalytic efficiency, side reactions, catalyst deactivation and product distribution (Pérez-Ramírez et al., 2008; Corma, 1997) in these chemical engineering processes. Therefore, thoroughly understanding and accurately predicting the distribution of particles in zeolites play a significant role in industrial

* Corresponding authors at: State Key Laboratory of Chemical Engineering, Shanghai Engineering Research Center of Hierarchical Nanomaterials, School of Chemical Engineering, East China University of Science and Technology, Shanghai 200237, PR China (C. Lian and H. Liu).
E-mail addresses: liancheng@ecust.edu.cn (C. Lian), liyang@wit.edu.cn (L. Yang), hlliu@ecust.edu.cn (H. Liu).
[1] Equally contributing authors.

https://doi.org/10.1016/j.ces.2021.116995
0009-2509/

production. Fortunately, given the self-diffusion coefficient D (m^2/s) and the corresponding initial and boundary conditions, the spatial distribution of particles at any time can be calculated by the second Fick's law (Smit, 2008).

At present, the Einstein's relation, a simple linear relationship between the mean squared displacement (MSD, m^2) and time, has been extensively applied to calculate (Smit, 2008) (see Section S1 in Supplementary Material for details). Nevertheless, the growth of MSD with time sometimes anomalously features a power-law trend in heterogeneous and anisotropic diffusion mediums, such as in zeolites (de Azevedo et al., 2006; de Azevedo et al., 2006; Hahn et al., 1996), and can be generalized as (Bo et al., 2019):

$$\mathrm{MSD} = \left\langle (\Delta r)^2 \right\rangle = 2dD_\alpha t^\alpha \quad (1)$$

where D_α is anomalous diffusion coefficient (m^2/s^α), α is anomalous diffusion exponent, Δr is the displacements of particles in a given time interval Δt (s), $\langle\cdot\rangle$ is the ensemble average, and d is the spatial dimension. Anomalous diffusion can be divided into four categories by α, including sub-diffusion ($0 < \alpha < 1$) (Tan et al., 2018; Zaks and Nepomnyashchy, 2019), normal diffusion ($\alpha = 1$, satisfying the Einstein's relation), super-diffusion ($1 < \alpha < 2$) (Zaks and Nepomnyashchy, 2019; Ilievski et al., 2018), and ballistic diffusion ($\alpha = 2$) (Livorati et al., 2018). The relevance of α with particle trajectories is visually displayed in Fig. S1. In anomalous diffusion, with D_α and α, the density profiles of particles in zeolites can be accurately predicted by the fractional diffusion equation (de Azevedo et al., 2006) (see Section S2 in Supplementary Material for details).

Compared with experiments (Hahn et al., 1996; Golding and Cox, 2006), molecular dynamics (MD) simulations (Voigtmann and Horbach, 2009; Satija et al., 2017) provide an ideal alternative to calculate D_α and α both efficiently and accurately. However, it's still challenging to screen the system with specific D_α and α by MD simulations with hundreds of thousands of zeolites in a short time. Therefore, it's necessary to identify the relationship between the diffusion behaviors and the structural parameters of zeolites. Unfortunately, current theoretical model, such as continues time random walk (CTRW) and fractal method, cannot provide this structure–property relationship. Recently, machine learning (ML) method has emerged as a powerful tool to efficiently correlate anomalous diffusion with the structural properties of materials (Bo et al., 2019; Janczura). For example, Muñoz-Gil *et al.* (Muñoz-Gil et al., 2020) classified a given trajectory into one of several anomalous diffusion models and estimated the anomalous diffusion exponent by using a Random Forest (RF) algorithm; Kowalek *et al.* (Kowalek et al., 2019) utilized Convolutional Neural Network (CNN) to identify different modes of diffusion from known trajectories. However, almost all of the reported works focused on the prediction of single-particle trajectories (SPT) in cells and the training set was generated by limited experimental data. Prediction of the anomalous diffusion behaviors of a large number of particles in zeolites by ML method is rarely reported. Furthermore, a database of anomalous diffusion in zeolites at the MD simulation level has not been reported.

Herein, we constructed a universal approach by combining MD simulations with ML method to predict anomalous diffusion of methane, ethane, and propane in zeolites. The procedure of our hybrid method is shown in Fig. 1. Specifically, the trajectories of three light alkanes at four different concentrations in 185 zeolites from the international zeolite association (IZA) database, listed in Table S6, were first calculated by MD simulations. Then, D_α and α were obtained by logarithmic linear fitting Equation (1). Further-

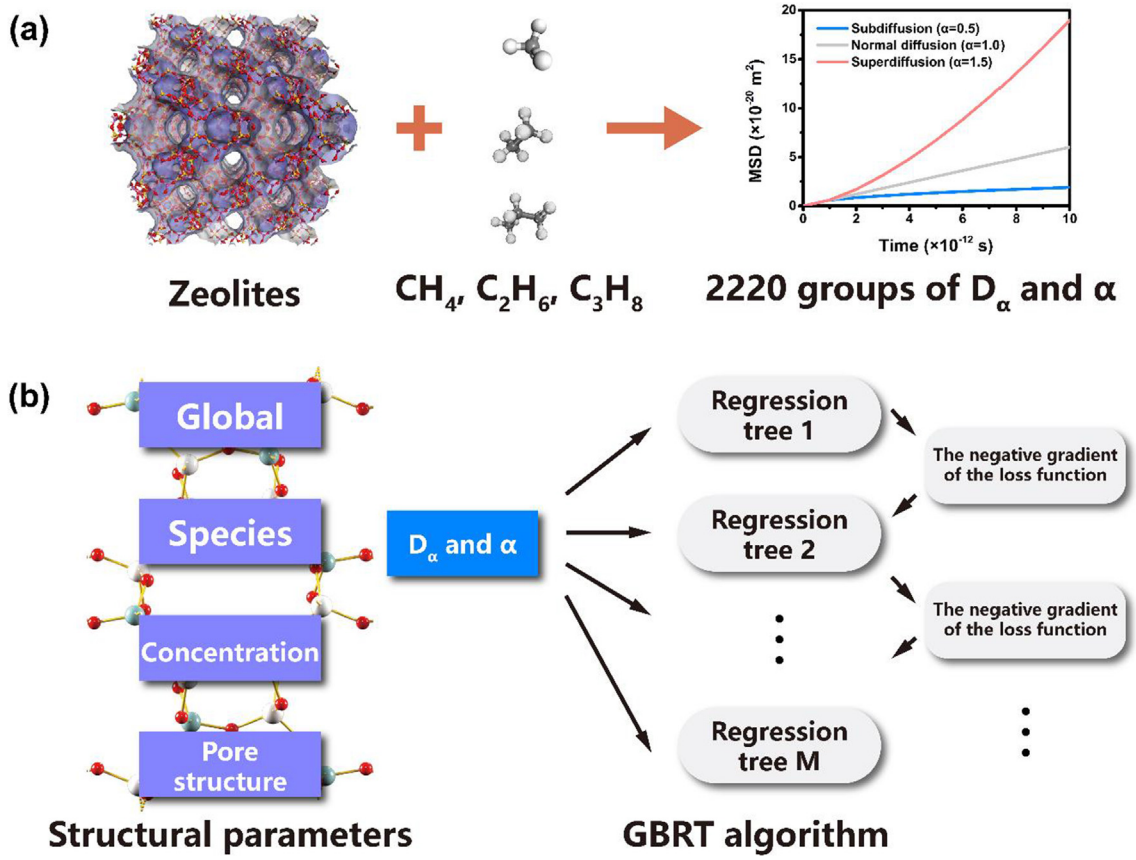

Fig. 1. The flow diagram of the hybrid approach combining (a) molecular dynamics (MD) simulations and (b) machine learning (ML) method to predict anomalous diffusion properties in zeolites.

more, the simulation data was screened by the goodness of fit R^2 and the distributions of D_α and α were also analyzed. The relationships between D_α, α and structure parameters, such as molecular species and porosity features, were qualitatively analyzed. Notably, a *combined structure number* (CSN) for rough prediction of α was identified. Subsequently, the Gradient Boosted Regression Trees (GBRT) algorithm was applied to predict D_α and α in zeolites and to rank different structural features in the order of importance. Finally, to further validate this method, the diffusion properties of 200,000 hypothetical zeolite structures were predicted with a computational cost of only a couple of seconds. This study is aimed to demonstrate the efficiency and accuracy of our hybrid procedure for fast prediction of anomalous diffusion properties, and provide comprehensive insights into the relationship between anomalous diffusion and structure in complex media.

2. Methodology

2.1. Molecular dynamics simulations

In this work, all zeolites were simulated with rigid structures to efficiently perform large scale calculations and to sufficiently capture the essence. The three light alkanes (methane, ethane and propane) were modeled with TraPPE force field (Häse et al., 2017; Martin and Siepmann), and zeolites were modeled with Universal force field (UFF) force field.

GROMACS 4.6.7 (GROMACS 4, 2020; GROMACS, 2005) was used for MD simulations. The relevant parameters were set as follows: NVT ensemble was used; the temperature was controlled at 298.15 K with Nose-Hoover thermostat bath; the leap-frog algorithm was used to solve the classical Newton's equation of motion under periodic boundary conditions; the Switch method was used to compute the Lennard-Jones potential with a cutoff length of 1.1 nm; the PME method was used to compute the electrostatic interactions with a real-space cutoff length of 1.1 nm; the bonds with H-atoms were constrained using LINCS algorithm (Hess et al.); the simulated time step was 1 fs; the overall simulation time was 1 ns; and the trajectory was generated every 5 ps for the analysis.

The ensemble average $\langle\cdot\rangle$ in Equation (1) requires the simulation system as large as possible and the simulation time as long as possible. The repeated structures of zeolites and simple structures of light alkanes makes it possible to get accurate α and D_α with reasonably smaller simulation box and less structural parameters. Therefore, each simulation box in this work was constructed with a $10 \times 10 \times 10$ supercell, where 10^0 to 10^2 light alkane molecules were simulated for 1 ns.

2.2. Calculation of D_α and α

Suppose M steps of trajectories are generated and the position vectors of each step are $r(0)$, $r(1)$, …, and $r(M)$. The MSD of step $i(i \le M)$ for a single particle is calculated by:

$$\mathrm{MSD}(i) = |r(i) - r(0)|^2 \tag{2}$$

Then the MSD of each particle is averaged to calculate the MSD of all particles in the system. Notably, only trajectories of the first 800 ps were used to calculate MSD in order to obtain intrinsic D_α and α. When the guest molecules are adsorbed on the surface of zeolites, the logarithm of MSD is no longer linear with the logarithm of time, as shown in Fig. S3. Therefore, many relative articles (Wang and Hou, 2011) just use the straight part (the first 800 ps in this work) to calculate the diffusion coefficient D. Finally, after taking logarithm of both sides of Equation (1) and performing linear fitting for $\log_{10}(\mathrm{MSD})$ and $\log_{10}(t)$, D_α and α were finally obtained. The value of R^2 is used to evaluate the accuracy of linear fitting:

$$R^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(y_i - \hat{y}_i)}{N\sigma_y\sigma_{\hat{y}}} \tag{3}$$

where y_i is the observed value, $\bar{y}$ is the average value, $\hat{y}_i$ is the predicted value, N is the number of samples, δ_y is variance of the observed value, and $\delta_{\hat{y}}$ is variance of the predicted value.

As shown in Fig. S2, unreliable data are firstly removed according to R^2, and then the key parameters determining R^2 are analyzed (see Section S3 in Supplementary Material for details). Furthermore, Fig. S3 and Table S1 show the time evolutions of MSD and some other simulation information for maximum and minimum D_α and α, respectively.

2.3. Machine learning method

In this work, Gradient Boosted Regression Tree (GBRT) algorithm is selected to predict D_α and α of methane, ethane, and propane in zeolites. GBRT algorithm relies on a regression tree built with a relatively shallow depth and then matching a subsequent regression tree with the ascendent (Ivatt and Evans). This step is repeated until an appropriate level of complexity is achieved, where the model generalizes the data set without over-fitting. GBRT algorithm has many advantages including: (i) GBRT can capture non-linear relationships which features anomalous diffusion; (ii) the regression-tree-based machine learning technique is more interpretable and robust than neural-net-based models (Kingsford and Salzberg, 2008); (iii) GBRT algorithm costs a relatively less training time, allowing more efficient cross validation to tune the hyper-parameters; (iv) and it is applicable for the relatively small datasets as we did in this work (Gaillac et al., 2020). To quickly construct the GBRT algorithm, we take advantage of the *scikit-learn package* (Pedregosa et al.) in Python 3.7.

Descriptors (*i.e.*, structural parameters) aim to accurately describe the key structures of zeolites closely related to the diffusion properties. The descriptors in this work can be categorized into four different groups: (i) *global* descriptors (directly obtained from the CIF file (Hall et al., 1991), *e.g.*, framework density and space group; (ii) *species* descriptors (depending on the species of diffusing particles), *e.g.*, carbon number; (iii) descriptors related to *porosity* (calculated from the CIF file using Zeo++ (Willems et al., 2012), *e.g.*, largest free sphere and accessible volume; and (iv) *particle concentration* descriptor (characterized by numbers of light alkane molecules in zeolites), *e.g.*, number density. The complete list of descriptors is given in Table S2, and Table S3 shows their statistics.

Hyper-parameters are the parameter set before data training and represents the complexity required by ML model (Bergstra and Bengio). The criterion for choosing hyper-parameters is to provide high prediction accuracy and flexibility to minimize overfitting. To obtain optimal GBRT algorithm, grid searching technique is adopted to tune hyper-parameters and ten k-fold cross validation is utilized to reliably and stably evaluate performance of model in specific hyper-parameters. The evaluation criterion of GBRT algorithm is the root mean squared error (RMSE) between predicted value and the observed value. RMSE is defined as:

$$\mathrm{RMSE}(y, \hat{y}) = \left[\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2\right]^{\frac{1}{2}} \tag{4}$$

The optimized hyper-parameters are given in Table S4, and other hyper-parameters are set to default values in the GBRT function.

3. Results and discussion

3.1. The frequency distribution of D_α and α

Frequency distribution refers to the correspondence of a set of frequencies with the set of categories, intervals, or values into which a statistical population is classified. The frequency distributions of D_α and α at $R^2 \geq 0.8$ are shown in Fig. 2 to illustrate the general characteristics of diffusion behaviors of light alkanes in zeolites. Interestingly, the frequency distribution of α satisfies the normal distribution and the center of distribution is at $\alpha = 0.94$. Considering the numerical error of fitting and 3σ criteria, samples where α is greater than 0.95 and less than 1.05 are considered as normal diffusion, and the rest are identified as sub-diffusion and super-diffusion, respectively. Sub-diffusion accounts for the majority (63.8%) and super-diffusion takes up 8.1%. Interesting, super-diffusion phenomenon in zeolite (Thomas and Subramanian, 2018) has rarely been reported. However, the proportion of normal diffusion turns out to be about only 28.1%, indicating that Equation (1) is more reliable for general prediction of diffusion coefficients rather than the simple linear relation of Einstein's equation. Although the largest α is 1.46 and the smallest is 0.01, the distribu-

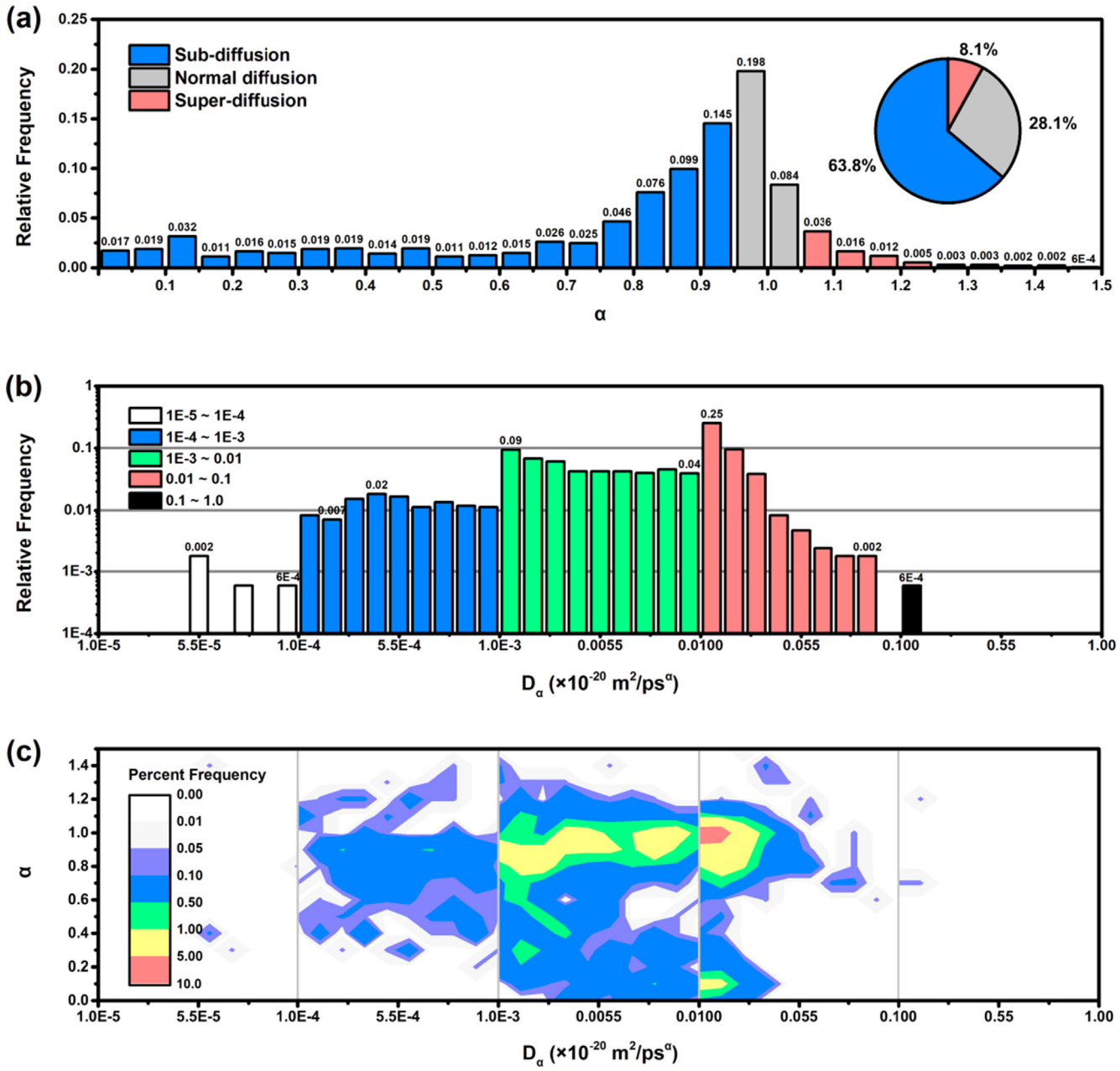

Fig. 2. The frequency distributions of D_α and α at $R^2 \geq 0.8$. (a) The frequency distribution of α satisfies the normal distribution, where the average $\mu = 0.945 \pm 0.005$, standard deviation $\sigma = 0.173 \pm 0.014$ and goodness of fit $R^2 = 0.913$. Data with α is greater than 0.95 and less than 1.05 are considered as normal diffusion (grey pattern), and others are sub-diffusion (blue pattern) and super-diffusion (red pattern), respectively. (b) The frequency distribution of D_α. The order of magnitude of D_α is mainly between 10^{-3} and 10^{-2} $Å^2/ps^\alpha$. (c) The two-dimension frequency distribution for D_α and α on different order of magnitude, percent frequency shown as color scale. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

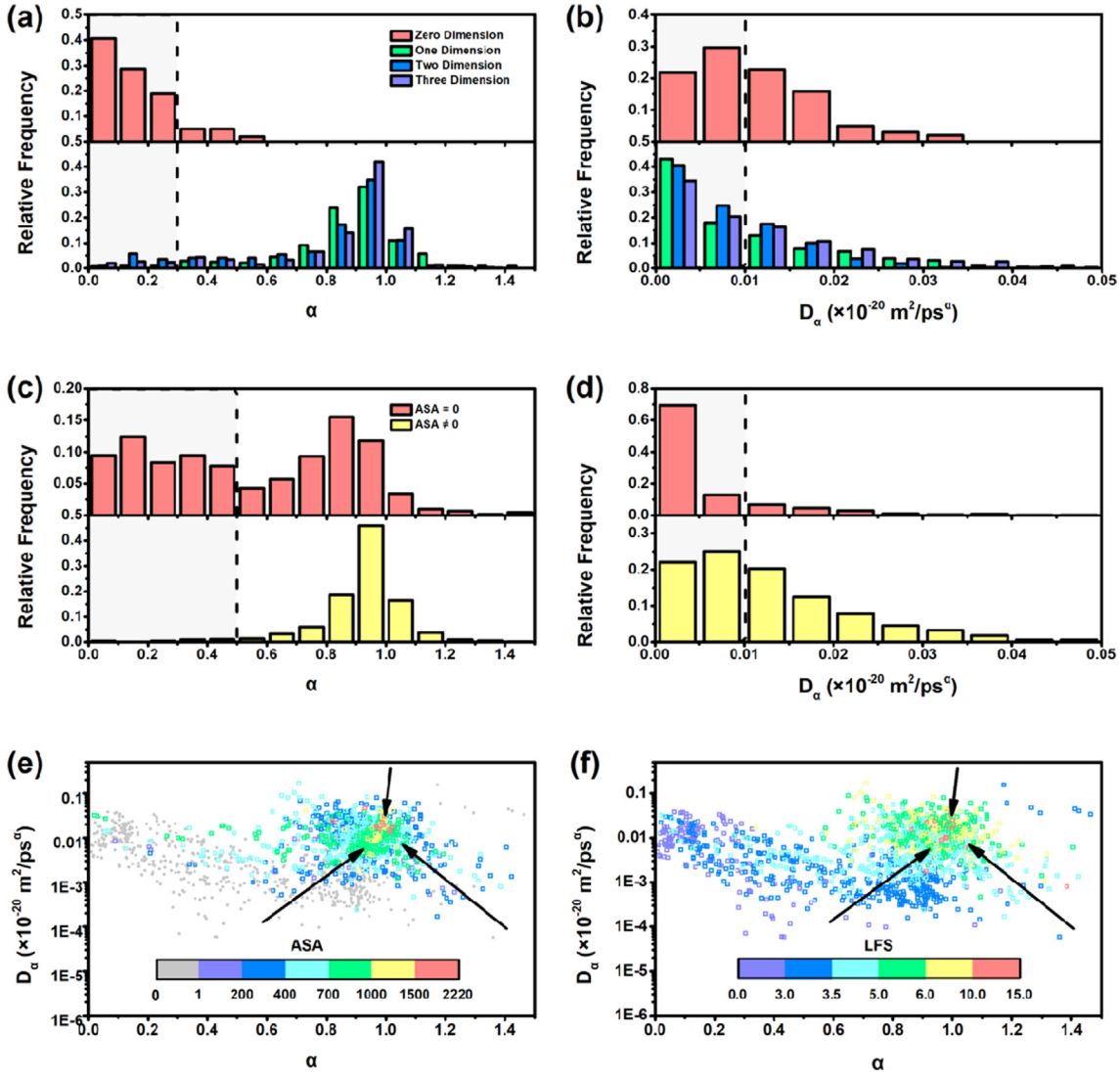

Fig. 3. Overall influence of structure of zeolites on anomalous diffusion. (a), (b) The frequency distribution of α andD_α in different channel dimensionality (CD). The variation of frequency distribution for zero dimension is opposite to that for other dimensionalities. Notably, most of data gather at $\alpha < 0.3$ in channel with zero dimension and with CD increasing, the center of distribution gradually moves to the right in other CDs. The variation of D_α in the range of 0 to 1×10^{-2}Å^2/ps$^\alpha$ in channel with zero dimension is also opposite to that in others. (c), (d) The frequency distribution of α and D_α, respectively, in zeolites with available surface area (ASA, m^2/g) equals to zero and non-zero. (e), (f) The scatter diagram of D_α vs. α with ASA and LFS as color scale, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

tion range of D_a value is even wider, covering from 5.8×10^{-5} to 1.7×10^{-1}Å^2/ps$^\alpha$. Furthermore, the majority of D_α are concentrated within the region from 10^{-3} to 10^{-2}Å^2/ps$^\alpha$, especially close to10^{-2}Å^2/ps$^\alpha$, which is not consistent with empirical value for normal diffusion (Catlow et al., 1991). Finally, the two-dimensional frequency distribution of D_α and α indicates that percent frequencies of D_α in the range of 1×10^{-2} to 2×10^{-2}Å^2/ps$^\alpha$ and α in the range of 0.95 to 1.05 (normal diffusion) are the highest. In addition, smaller α is more likely to correspond with larger D_α.

3.2. Qualitative relations between anomalous diffusion and structure

In order to accurately and efficiently predict D_α and α by ML method, the qualitative relations between structural parameters and anomalous diffusion (D_α and α) need to be understood firstly, and then relative important structural parameters are to be determined. In this work, structures are divided into structure of light alkanes and zeolites. Specially, we neglect the intra-molecular structural factors of light alkanes including bond length and bond angle to avoid data redundancy. Therefore, carbon number (CN) is selected as the only structural parameter to reflect the influence of species of light alkanes on anomalous diffusion. The influences of CN on anomalous diffusion in zeolites are clearly shown in Fig. S4. Three parallel lines are found along the lower edge of the shape of distribution between D_α and ND and move up as the alkanes becomes heavier. However, the difference of α values for different alkanes is very small. (see Section S4 in Supplementary Material for details).

The influences of the structures of zeolites on anomalous diffusion are much more complicated than those of species of light alkanes due to the more complex structural parameters of zeolites. Inspired by Fig. S2.e and Fig. S2.f, channel dimensionalities (CD) and accessible surface area (ASA) are naturally considered as the important structural parameters. Because the distributions of α and D_α at ASA = 0 are not equivalent with those in channel with zero-dimension, we consider both ASA and CD as independent

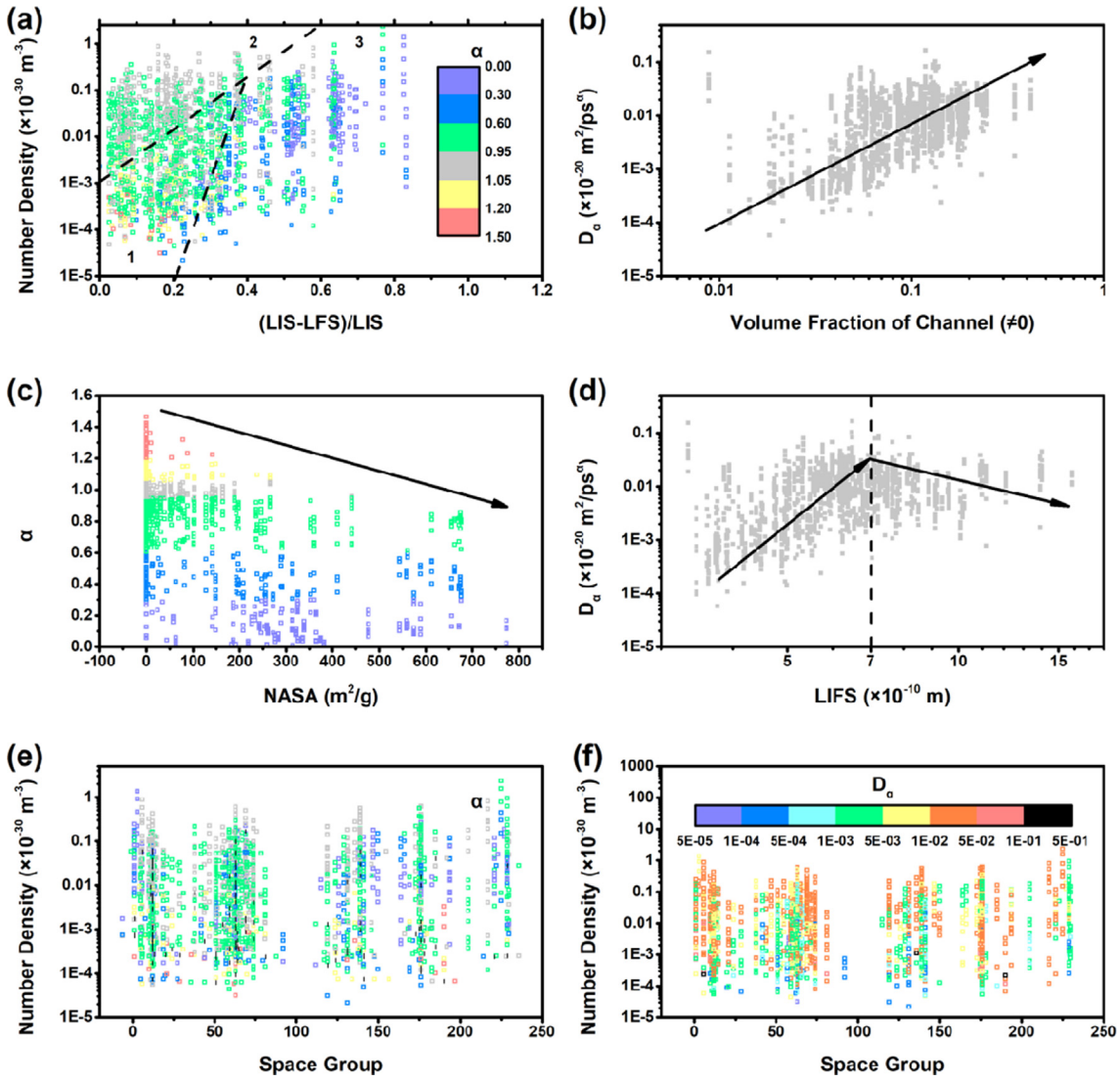

Fig. 4. Influence of structure of zeolites on anomalous diffusion with respect to different ND of light alkanes. (a) The scatter diagram of ND vs. (LIS-LFS)/LIS, where LIS is largest included sphere (Å). (b) The scatter diagram of D_α vs. none-zero volume fraction of channel (VFC), showing a strong positive linear relation between them. (c) The scatter diagram of α vs. not available surface area (NASA, m^2/g). Only when NASA is less than $300 m^2/g$, can super-diffusion take place. (d) The scatter diagram of D_α vs. largest included free sphere (LIFS, Å). D_α does not increase constantly, but enlarges until LISF = 7.0Å then reduces. (e), (f) The scatter diagram of ND vs. space group (SG) with α and D_α as color scale, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

parameters. From Fig. 3.a and Fig. 3.b, we can see that the variation of frequency distribution of α and D_α for zero-dimension is opposite to that for other dimensionalities, which is consistent with those of R^2 in different CDs. Notably, most of data gather at $\alpha < 0.3$ in channel with zero-dimension and with increasing CD, the center of distribution gradually moves to the right side in other CDs. The variation of D_α in the range of 0 to $1 \times 10^{-2} Å^2/ps^\alpha$ in channel with zero-dimension is also opposite to that in others. Except the zero-dimension channels, the distributions of α follow the law of normal distribution. Overall, light alkanes in zeolites with zero-dimension channels tend to diffuse with greater D_α and smaller α. The distribution of α at ASA≠0 also shows a normal distribution while that at ASA = 0 is relatively uniform. Interestingly, although CD = 0 and ASA = 0 both represent the extreme narrow and anomalous pore structures, their distributions of D_α are exactly opposite with those of normal pore structure when D_α is less than $1 \times 10^{-2} Å^2/ps^\alpha$. Fig. 3.e further illustrates the opposite influences of ASA on both D_α and α. The result indicates that data with ASA = 0 are randomly distributed with different D_α and α, causing the uniform distribution of α at ASA = 0. However, with increasing ASA, more and more data concentrates in this region $D_\alpha = 1 \times 10^{-2} Å^2/ps^\alpha$ and $\alpha = 1.0$, where the percent frequency of data is the highest, as shown in Fig. 2.c. Those results demonstrate that zeolites with a smaller ASA is more likely to result in anomalous diffusion ($\alpha < 0.95$ or $\alpha > 1.05$) and the D_α beyond empirical estimations. Interesting, LFS and ASA have similar effects on α and D_α, as shown in Fig. 3.f.

Number density (ND) should also be simultaneously taken into account when discussing the effects of zeolite structures on anomalous diffusion (de Azevedo et al., 2006; Khalifi et al., 2020). The ratio of the difference between largest included sphere (LIS, Å) and LFS to LIS, (LIS − LFS)/LIS, represents the proportion of the pore of zeolites where guest molecules cannot move freely. Fig. 4.a shows that the data regarding different (LIS − LFS)/LIS versus ND can be divided into three regions with respect to the distribution of α. Super-diffusion takes place in region 1 and sub-diffusion with $\alpha \leq 0.6$ occurs in region 3. Although α can not be simply determined by not available surface area (NASA, m^2/g with

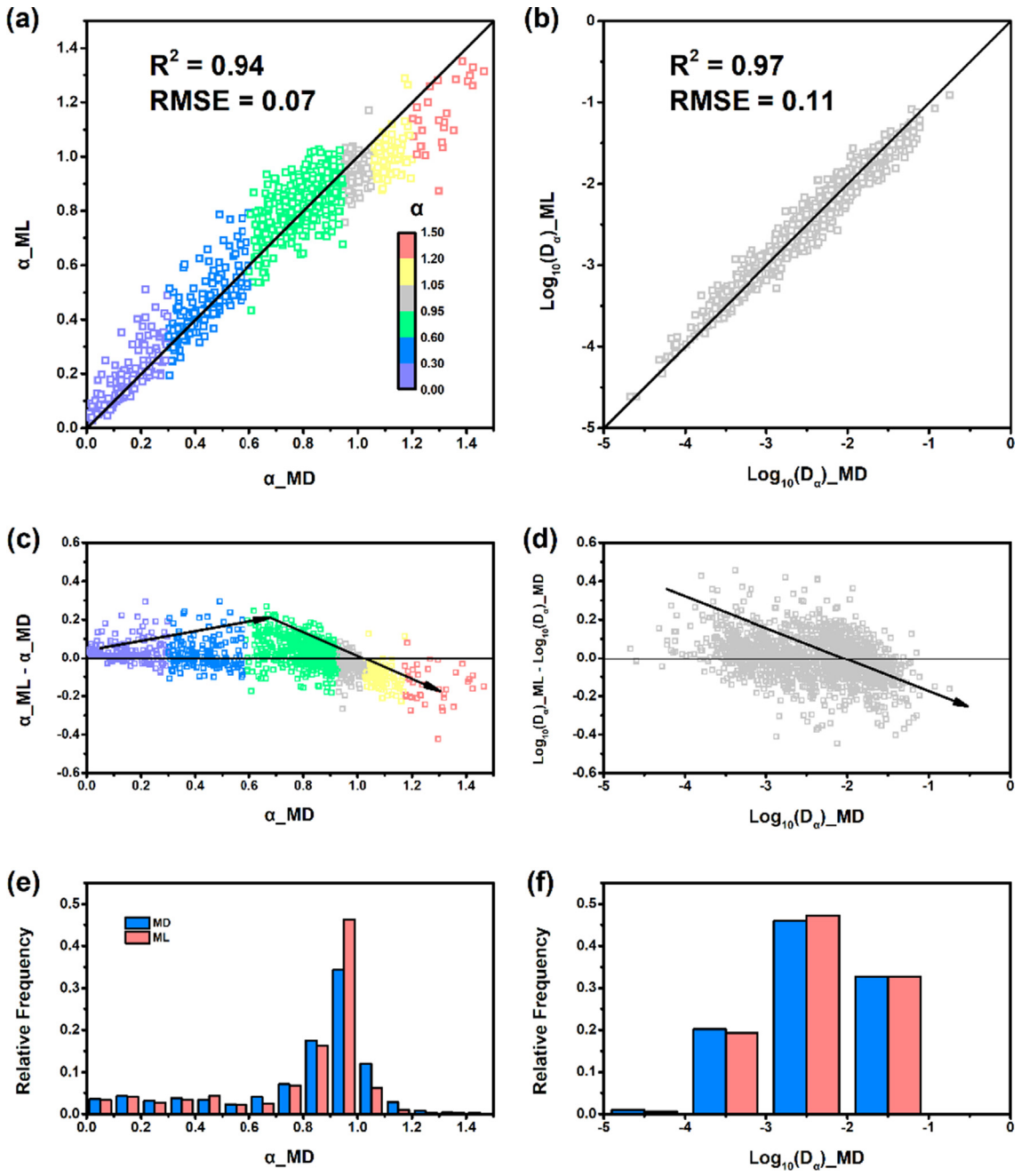

Fig. 5. The accuracy of ML model implemented with GBRT algorithm. (a), (b) Comparison of ML predictions with MD training set for anomalous diffusion exponent α and diffusion coefficient D_α of 185 zeolites with four different particle concentration. This model works well for predicting α and $\log_{10}(D_\alpha)$ with R^2 as greater as 0.94 and 0.97, respectively. (c), (d) The scatter diagram of the error for prediction of α and $\log_{10}(D_\alpha)$. (e), (f) Comparison of predictions of frequency distribution of α and D_α by ML method with training data calculated by MD simulations.

a probe radius of 1.82 Å), zeolites with larger NASA have smaller values of α, as plotted in Fig. 4.c. And the super-diffusion phenomenon corresponds to the region where NASA is less than 300m^2/g. Fig. 4.b illustrates a strong positive linear relationship between D_α and none-zero volume fraction of channel (VFC), while Fig. 4.d indicates that the largest D_α corresponds with LIFS = 7.0Å, instead of the maximum value of largest included free sphere (LIFS, Å). Space group (SG) is the combination of all the symmetry elements in a unit cell, and is also identified as a key factor. As shown in Fig. 4.e and Fig. 4.f, SG naturally exhibits superior performance as a descriptor to represent the global property of different zeolites.

In order to find the best structural parameters to reveal the inherent relations with anomalous diffusion, it is an effective way to reconstruct the structural parameters to obtain a series of modified parameters, especially those exhibiting linear correlations with D_α and α. The process of constructing *combined structure number* ($CSN = [\log_{10}(100/ND)/((LIS - LFS)/LIS)]^{0.01} \times \exp(VFC)^{0.05}$), as a dimensionless number considering the structure of zeolites and the number density of guest molecules, can be found in Section S5 in Supplementary Material. A positive linear relationship between α and CSN can be found in Fig. S5.a, which can be used to demonstrate the reliability of the predictions with ML method and even

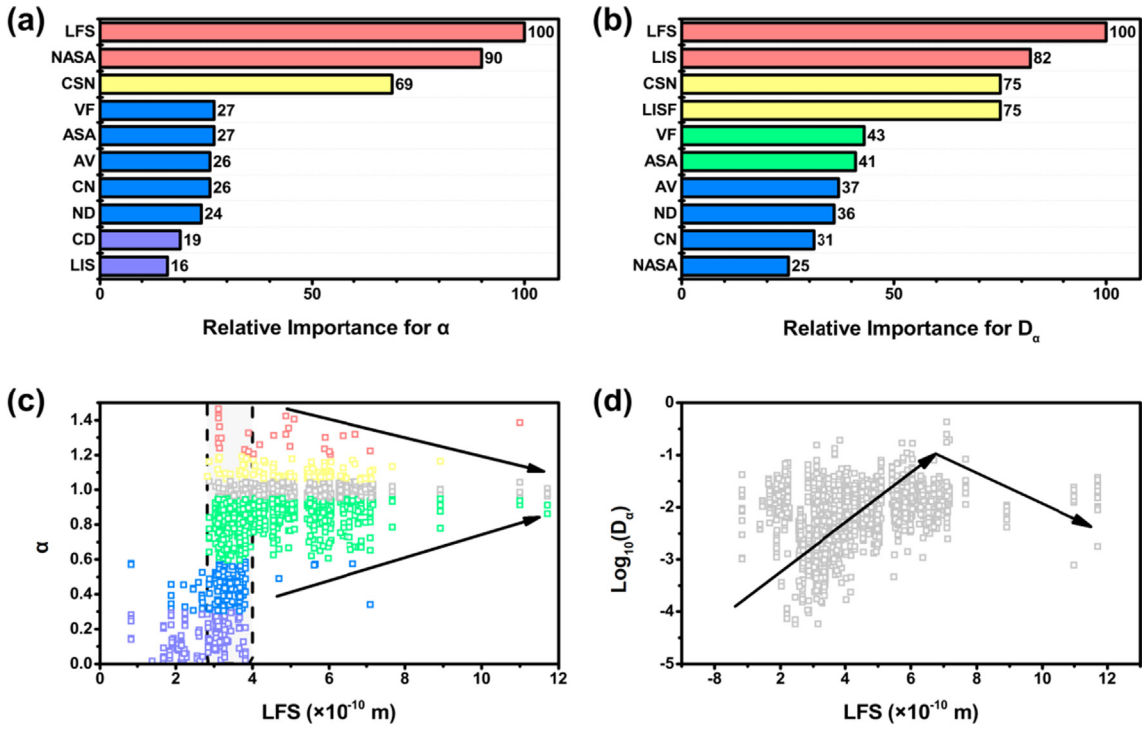

Fig. 6. Relative importance of the ten most important descriptors for (a) Anomalous diffusion exponent α and (b) Anomalous diffusion coefficient D_α, where LFS, CSN, ASA and NASA represent largest free sphere, *combined structure number*, available surface area and not available surface area, respectively. The detailed information of the ten key features is listed in Table S2. (c), (d) Correlations between α, $\log_{10}(D_\alpha)$ and LFS from the training data calculated by MD simulations.

roughly predicting α. The linear trend between 697 α of CO_2 calculated by MD simulations and corresponding CSN is shown in Fig. S5.b to prove the universality of CSN.

In summary, although the structures of light alkanes and zeolites both have significant effects on D_α and α, some of the currently selected structural parameters show no obvious quantitative relationship with D_α and α, as shown in Fig. S4, Fig. 3, and Fig. 4. Besides, the rest of the selected structural parameters seem even irrelevant with anomalous diffusion being masked by other factors. Furthermore, the quantitative relation between structure and anomalous diffusion could not be completely determined, because a lot of the considered structural factors have certain degree of nonlinear correlation with anomalous diffusion, as shown in Fig. S5. Herein, to deal with this problem, machine learning, as a novel and efficient tool, is utilized to discover the inherent relation between structural parameters and anomalous diffusion.

3.3. Accuracy of machine learning method on anomalous diffusion prediction

The GBRT algorithm can capture the non-linear relationships between anomalous diffusion and the structural parameters and it is more interpretable than Convolutional Neural Net (CNN) algorithm. Notably, $\log_{10}(D_\alpha)$ is adopted as the predicting target, which avoids overestimating of D_α with the algorithm.

Fig. 5.a and Fig. 5.b demonstrate the relatively high accuracy and low variance of ML method, where the MD training data is plotted to compare with the predictions of ML implemented with GBRT algorithm. Particularly, RMSE for α and $\log_{10}(D_\alpha)$ are 0.07 and 0.11, and R^2 are 0.94 and 0.97, respectively. This is extremely accurate given that the GBRT algorithm is constructed from a relatively small training data. It can be seen from Fig. 5.c that GBRT algorithm tends to systematically underestimate super-diffusion and overestimate sub-diffusion, while for normal diffusion, the predictions are quite reasonable. The errors in predicting $\log_{10}(D_\alpha)$ decreases with increasing values of D_α, as plotted in Fig. 5.d. The frequency distributions for the predicted α and $\log_{10}(D_\alpha)$ indicate that the distributions are generally consistent while locally different with training data, especially in the regions where $0.9 \leq \alpha \leq 1.1$ and $1 \times 10^{-3} < D_\alpha < 1 \times 10^{-2}$ Å^2/ps$^\alpha$. Finally, Fig. S6 is plotted to show the relative error of percent frequency on distribution of $\log_{10}(D_\alpha)$ and α. The results indicate that in most regions, the relative errors are small and GBRT algorithm exhibits excellent predictive power for both D_α and α.

3.4. Ten key features identified for anomalous diffusion in zeolites

The regression trees adopted by GBRT algorithm have a unique advantage that the construction of the trees is substantially the selection of features. This information can be interpreted simultaneously to obtain the relative importance of each of the chosen structural parameters (descriptors), as plotted in Fig. 6, showing the weights of the ten most contributing descriptors in the GBRT algorithm. The results indicate that LFS contributes significantly to the model for both α and D_α, where the relative importance values are both 100.

The influence of LFS on both D_α and α has been discussed in Fig. 3.f and the individual effects are shown in Fig. 6.c and Fig. 6.d. Interestingly, the relationship between LFS and α is non-monotonic. When LFS is less than 2.8Å, the data belongs to sub-diffusion category. When LFS is around 2.8Å, normal diffusion and super-diffusion begin to appear and α increases sharply, which leads to a huge difference of α in the regions where LFS is less than 2.8Å and larger than4.0Å. When LFS covers from 4.0 to 12.0 Å, sub-diffusion is still the dominate type of anomalous diffusion though most of normal diffusion and super-diffusion also occur in this region. Besides, the shape of the region is like a triangle and α gradually gets closer to 1.0 with increasing LFS. In addition, with

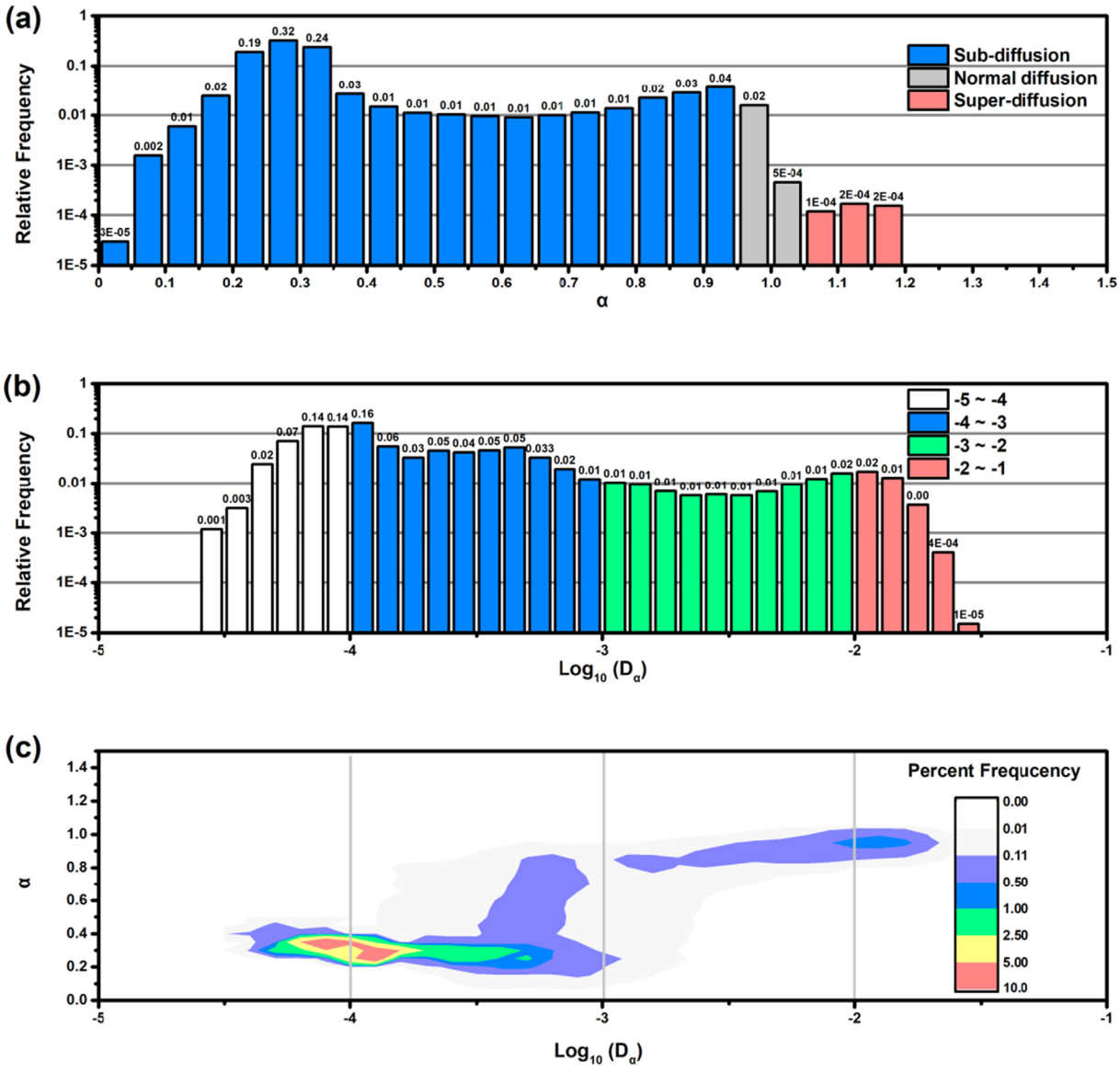

Fig. 7. The frequency distributions of α and $\log_{10}(D_\alpha)$ in PCOD2 database. (a) The frequency distribution of α dose not satisfy normal distribution. Sub-diffusion, normal diffusion and super diffusion account for 98%, 2%, and 0.05%, respectively. (b) The frequency distribution of D_α is also not consistent with that in IZA database. (c) The two-dimension frequency distribution for $\log_{10}(D_\alpha)$ and α, percent frequency as color scale. Data predicted by ML clusters at $-4.2 < \log_{10}(D_\alpha) < -3.8$ and $0.2 < \alpha < 0.4$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

increasing LFS values,$\log_{10}(D_\alpha)$ increases until LFS = 6.0Å and then remains constant. The importance of LFS as a key structural parameter of zeolites to correlate with diffusion properties has previously been reported (Foster et al., 2006). In this work, we further demonstrate that a fairly strong and nonlinear correlation between LFS and anomalous diffusion exists in a large number of zeolites.

3.5. Application to a hypothetical database

The GBRT algorithm constructed in this work is capable to predict the diffusion properties in a zeolite in less than a millisecond which allows researchers to perform large-scale predictions. To prove the effectiveness of our predictive model, this method is employed to predict α and D_α for a hypothetical all-silica zeolite database built by Deem and co-workers, PCOD2 (Earl and Deem, 2006; Deem et al., 2009). This database contains about four million new zeolite topologies obtained by a method that combines simulated annealing, Monte Carlo simulations, and refinement using interatomic potentials. We take advantage of the structural information available in CIF files in the PCOD2 database and apply GBRT algorithm to predict D_α and α with excellent accuracy.

The optimal GBRT algorithm with hyperparameters described in Table S4 is utilized to predict the anomalous diffusion properties of a total of 200,000 zeolites from the PCOD2 database in seconds. ND and CN is randomly generated by Python 3.7, as shown in Fig. S10. The frequency distribution of α and $\log_{10}(D_\alpha)$ predicted by GBRT algorithm is plotted in Fig. 7. The distribution of α has two peaks near $\alpha = 0.3$ and $\alpha = 0.95$, which does not satisfy normal distribution. Besides, the proportions of normal diffusion and super-diffusion are only 2% and 0.05%, due to a mass of structural parameters of zeolites in PCOD2 database equal to zero as shown in Figs. S7 and S9. α value is located in the regions of 0.03–1.2, which is narrower than that of IZA database. Furthermore, the majority of values of $\log_{10}(D_\alpha)$ are concentrated from −5 to −3, especially near −4, which is also not consistent with Fig. (S3.b). Finally, the two-

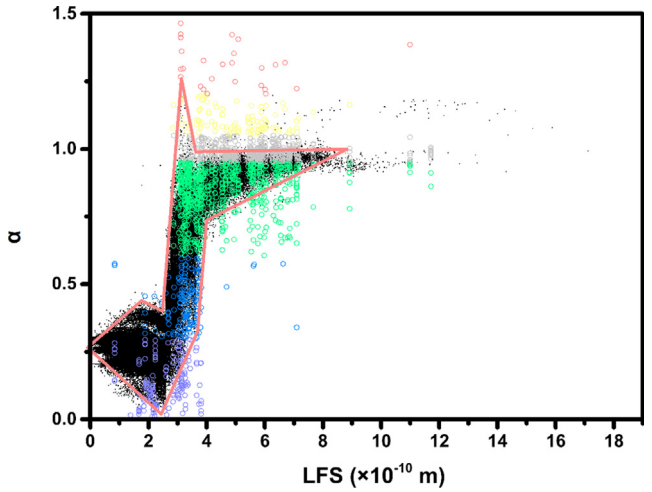

Fig. 8. The distribution of α in the PCOD2 database predicted by ML method (black points) and α in IZA database calculated by MD simulations (colorful points), both plotted against LFS, agree with each other. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

dimension frequency distribution for $\log_{10}(D_\alpha)$ and α indicates that percent frequencies of data with $\log_{10}(D_\alpha)$ in the range from −4.2 to −2.8 and α in the range from 0.2 to 0.4 are the highest (about 50%).

Finally, the distributions of α predicted by ML method, plotted against LFS, is illustrated in Fig. 8. The distribution of black points predicted in PCOD2 database is similar to that of colorful points calculated by MD simulations, and they both have a step region where α increases sharply. The step regions of PCOD2 database and IZA database are very close, which further proves the accuracy of GBRT algorithm.

4. Conclusion

D_α and α are important indicators for understanding diffusion phenomena, since in practical applications, the internal diffusion of reactant, product or intermediates in zeolites tends to be the rate-limiting step in many industrial processes. To identify the relationship between the diffusion behaviors and the structural parameters of zeolite is the key to understand anomalous diffusion and realize rapid prediction of diffusion coefficients in a more general and accurate way. Therefore, we constructed a novel hybrid methodology combining molecular dynamics (MD) and machine learning (ML) for large-scale prediction of diffusion coefficient D_α and anomalous diffusion exponent α in zeolites.

First, a database on anomalous diffusion of light alkanes (methane, ethane and propane) in 185 zeolites, about 2200 groups of D_α and α, established by massive MD simulation data in this work, enables us with abundant kinetic information to study anomalous diffusion behaviors. It is identified that sub-diffusion is the dominate form of diffusion behavior in zeolites and super-diffusion also takes a relatively small proportion, however, the Einstein equation's linear relation and the second Fick's law fail to describe either of them. Furthermore, the qualitative relationships between numerous structural parameters and anomalous diffusion are analyzed for the first time, and a *combined structure number* (CSN) for rough prediction of α is identified. Subsequently, a model for predicting D_α and α by structures with great accuracy ($R^2 = 0.97$ and $R^2 = 0.94$) is established by GBRT algorithm. Besides, the structural parameters are ranked in order of importance and largest free sphere (LFS) is recognized as the most important descriptor to interpret anomalous diffusion. The region where the value of LFS is located in can be considered as the criterion for classification of anomalous diffusion in zeolite. Besides, the ranking open up new strategies for optimizing the structure of zeolites when anomalous diffusion should be restrained or enhanced in specific system. Finally, the anomalous diffusion properties of 200,000 hypothetical zeolites are predicted by the model, and we expect these results could provide theoretical guidance for experimentalists to design and synthesize high-performance zeolites.

More structural parameters should be analyzed and considered as the descriptors of the GBRT algorithm in the next work to improve prediction accuracy in the future. Besides, light alkanes will be adsorbed, even react in zeolites, which have great influence on anomalous diffusion and should be considered. Future studies will examine more kinetic features to provide a model for prediction of the diffusion behaviors of hydrocarbon compounds in SiO_2 polymorphs for applications as catalysts and adsorbents.

CRediT authorship contribution statement

Pan Huang: Conceptualization, Investigation, Methodology, Software, Validation, Formal analysis, Data curation, Visualization, Writing – original draft, Writing – review & editing. **Zhijian Yin:** Investigation, Methodology, Software, Validation. **Yun Tian:** Writing – original draft. **Jie Yang:** Writing – review & editing. **Wei Zhong:** Writing – review & editing. **Chunzhong Li:** Writing – review & editing. **Cheng Lian:** Conceptualization, Methodology, Formal analysis, Project administration, Writing – review & editing. **Li Yang:** Conceptualization, Methodology, Formal analysis, Project administration, Writing – review & editing. **Honglai Liu:** Funding acquisition, Writing – review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was sponsored by the National Natural Science Foundation of China (No. 91834301, 22078088) and National Natural Science Foundation of China for Innovative Research Groups (No. 51621002). We kindly thank Haiping Su and Shengwei Deng for helpful discussions.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ultsonch.2019.104640.

References

Bergstra, J., Bengio, Y. Random search for hyper-parameter optimization. 25.
Bo, S., Schmidt, F., Eichhorn, R., Volpe, G., 2019. Measurement of anomalous diffusion using recurrent neural networks. Phys. Rev. E 100, (1) 010102.
Catlow, C.R.A., Freeman, C.M., Vessal, B., Tomlinson, S.M., Leslie, M., 1991. Molecular dynamics studies of hydrocarbon diffusion in zeolites. J. Chem. Soc., Faraday Trans. 87 (13), 1947.
Corma, A., 1997. From microporous to mesoporous molecular sieve materials and their use in catalysis. Chem. Rev. 97 (6), 2373–2420.
de Azevedo, E.N., da Silva, D.V., de Souza, R.E., Engelsberg, M., 2006. Water ingress in Y-type zeolite: Anomalous moisture-dependent transport diffusivity. Phys. Rev. E 74, (4) 041108.
de Azevedo, E.N., de Sousa, P.L., de Souza, R.E., Engelsberg, M., Miranda, M. de N. do N., Silva, M.A., 2006. Concentration-dependent diffusivity and anomalous diffusion: a magnetic resonance imaging study of water ingress in porous zeolite. Phys. Rev. E. 73 (1).
Deem, M.W., Pophale, R., Cheeseman, P.A., Earl, D.J., 2009. Computational discovery of new zeolite-like materials. J. Phys. Chem. C 113 (51), 21353–21360.

Dusselier, M., Davis, M.E., 2018. Small-pore zeolites: synthesis and catalysis. Chem. Rev. 118 (11), 5265–5329.
Earl, D.J., Deem, M.W., 2006. Toward a database of hypothetical zeolite structures. Ind. Eng. Chem. Res. 45 (16), 5449–5454.
Ennaert, T., Van Aelst, J., Dijkmans, J., et al., 2016. Potential and challenges of zeolite chemistry in the catalytic conversion of biomass. Chem. Soc. Rev. 45 (3), 584–611.
Foster, M.D., Rivin, I., Treacy, M.M.J., Delgado, Friedrichs O., 2006. A geometric solution to the largest-free-sphere problem in zeolite frameworks. Microporous Mesoporous Mater. 90 (1), 32–38.
Gaillac, R., Chibani, S., Coudert, F.-X., 2020. Speeding up discovery of auxetic zeolite frameworks by machine learning. Chem. Mater. 32 (6), 2653–2663.
Golding, I., Cox, E.C., 2006. Physical nature of bacterial cytoplasm. Phys. Rev. Lett. 96, (9) 098102.
GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation | Journal of Chemical Theory and Computation. Accessed November 26, 2020.
GROMACS: Fast, flexible, and free - Van Der Spoel - 2005 - Journal of Computational Chemistry - Wiley Online Library. Accessed November 26, 2020.
Hahn, K., Kärger, J., Kukla, V., 1996. Single-file diffusion observation. Phys. Rev. Lett. 76 (15), 2762–2765.
Hall, S.R., Allen, F.H., Brown, I.D., 1991. The crystallographic information file (CIF): a new standard archive file for crystallography. Acta Crystallogr. A 47 (6), 655–685.
Häse, F., Kreisbeck, C., Aspuru-Guzik, A., 2017. Machine learning for quantum dynamics: deep learning of excitation energy transfer properties. Chem. Sci. 8 (12), 8419–8426.
Hess, B., Bekker, H., Berendsen, H.J.C. LINCS: A linear constraint solver for molecular simulations. J. Comput. Chem. 18(12), 10.
Ilievski, E., De Nardis, J., Medenjak, M., Prosen, T., 2018. Superdiffusion in one-dimensional quantum lattice models. Phys. Rev. Lett. 121, (23) 230602.
Ivatt, P., Evans, M.J. Improving the prediction of an atmospheric chemistry transport model using gradient boosted regression trees. Atmos. Chem. Phys. Discuss.
Janczura, J. Classification of particle trajectories in living cells: Machine learning versus statistical testing hypothesis for fractional anomalous diffusion. Phys. Rev. E.
Jiang, N., Shang, R., Heijman, S.G.J., Rietveld, L.C., 2018. High-silica zeolites for adsorption of organic micro-pollutants in water treatment: a review. Water Res. 144, 145–161.
Khalifi, M., Sabet, N., Zirrahi, M., Hassanzadeh, H., Abedi, J., 2020. Concentration-dependent molecular diffusion coefficient of gaseous ethane in liquid toluene. AIChE J. 66 (6).
Kingsford, C., Salzberg, S.L., 2008. What are decision trees? Nat. Biotechnol. 26 (9), 1011–1013.
Kowalek, P., Loch-Olszewska, H., Szwabiński, J., 2019. Classification of diffusion modes in single-particle tracking data: feature-based versus deep-learning approach. Phys. Rev. E 100, (3) 032410.
Livorati, A.L.P., Kroetz, T., Dettmann, C.P., Caldas, I.L., Leonel, E.D., 2018. Transition from normal to ballistic diffusion in a one-dimensional impact system. Phys. Rev. E 97, (3) 032205.
Martin, M.G., Siepmann, J.I. Transferable potentials for phase equilibria. 1. United-atom description of n-alkanes 9.
Muñoz-Gil, G., Garcia-March, M.A., Manzo, C., Martín-Guerrero, J.D., Lewenstein, M., 2020. Machine learning method for single trajectory characterization. New J. Phys. 22, (1) 013010.
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. Scikit-learn: machine learning in python. Mach. Learn. PYTHON. 6.
Pérez-Ramírez, J., Christensen, C.H., Egeblad, K., Christensen, C.H., Groen, J.C., 2008. Hierarchical zeolites: enhanced utilisation of microporous crystals in catalysis by advances in materials design. Chem. Soc. Rev. 37 (11), 2530–2542.
Primo, A., Garcia, H., 2014. Zeolites as catalysts in oil refining. Chem. Soc. Rev. 43 (22), 7548–7561.
Satija, R., Das, A., Makarov, D.E., 2017. Transition path times reveal memory effects and anomalous diffusion in the dynamics of protein folding. J. Chem. Phys. 147, (15) 152707.
Shamzhy, M., Opanasenko, M., Concepción, P., Martínez, A., 2019. New trends in tailoring active sites in zeolite-based catalysts. Chem. Soc. Rev. 48 (4), 1095–1149.
Smit, B., 2008. Molecular simulations of zeolites: adsorption, diffusion, and shape selectivity. Chem. Rev. 108 (10), 4125–4184.
Tan, P., Liang, Y., Xu, Q., et al., 2018. Gradual crossover from subdiffusion to normal diffusion: a many-body effect in protein surface water. Phys. Rev. Lett. 120, (24) 248101.
Thomas, A.M., Subramanian, Y., 2018. Diffusion processes in a poly-crystalline zeolitic material: a molecular dynamics study. J. Chem. Phys. 149, (6) 064702.
Vogt, E.T.C., Weckhuysen, M.B., 2015. Fluid catalytic cracking: recent developments on the grand old lady of zeolite catalysis. Chem. Soc. Rev. 44 (20), 7342–7370.
Voigtmann, Th., Horbach, J., 2009. Double transition scenario for anomalous diffusion in glass-forming mixtures. Phys. Rev. Lett. 103, (20) 205901.
Wang, J., Hou, T., 2011. Application of molecular dynamics simulations in molecular property prediction II: Diffusion coefficient. J. Comput. Chem. 32 (16), 3505–3519.
Wang, S., Peng, Y., 2010. Natural zeolites as effective adsorbents in water and wastewater treatment. Chem. Eng. J. 156 (1), 11–24.
Wen, J., Dong, H., Zeng, G., 2018. Application of zeolite in removing salinity/sodicity from wastewater: a review of mechanisms, challenges and opportunities. J Clean Prod. 197, 1435–1446.
Willems, T.F., Rycroft, C.H., Kazi, M., Meza, J.C., Haranczyk, M., 2012. Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. Microporous Mesoporous Mater. 149 (1), 134–141.
Yu, N., Wang, R.Z., Wang, L.W., 2013. Sorption thermal storage for solar energy. Prog. Energy Combust. Sci. 39 (5), 489–514.
Zaks, M.A., Nepomnyashchy, A., 2019. Subdiffusive and superdiffusive transport in plane steady viscous flows. Proc. Natl. Acad. Sci. 116 (37), 18245–18250.
Zhang, R., Liu, N., Lei, Z., Chen, B., 2016. Selective transformation of various nitrogen-containing exhaust gases toward N_2 over zeolite catalysts. Chem. Rev. 116 (6), 3658–3721.