THE JOURNAL OF PHYSICAL CHEMISTRY C

Article

pubs.acs.org/JPCC

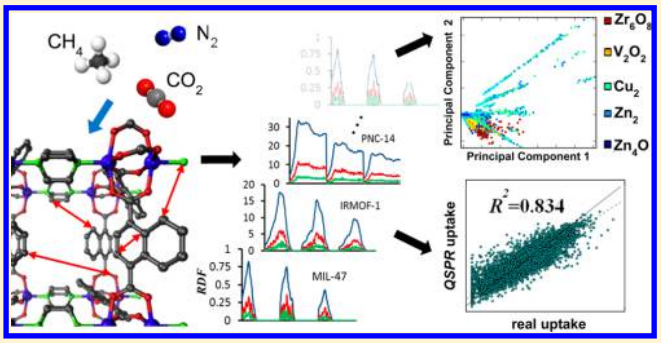
Atomic Property Weighted Radial Distribution Functions Descriptors of Metal−Organic Frameworks for the Prediction of Gas Uptake Capacity

Michael Fernandez, Nicholas R. Trefiak, and Tom K. Woo*

Centre for Catalysis Research and Innovation, Department of Chemistry, University of Ottawa, Ottawa, Canada

Supporting Information

ABSTRACT: Metal−organic frameworks (MOFs) are porous materials with exceptional host−guest properties with huge potential for gas separation. The combinatorial design of MOFs demands the *in silico* screening of the nearly infinite combinations of structural building blocks using efficient computational tools. We report here a novel atomic property weighted radial distribution function (AP-RDF) descriptor tailored for large-scale Quantitative Structure−Property Relationship (QSPR) predictions of gas adsorption of MOFs. A total of ∼58,000 hypothetical MOF structures were used to calibrate correlation models of the methane, N_2, and CO_2 uptake capacities from grand-canonical Monte Carlo (GCMC) simulations. The principal component analysis (PCA) transform of the AP-RDF descriptors exhibited good discrimination of MOF inorganic SBUs, geometrical properties, and more surprisingly gas uptake capacities. While the simulated uptake capacities correlated poorly to the void fraction, surface area, and pore size, the newly introduced AP-RDF scores yielded outstanding QSPR predictions for an external test set of ∼25,000 MOFs with R^2 values in the range from 0.70 to 0.82. The accuracy of the predictions decreased at low pressures, mainly for MOFs with V_2O_2 or Zr_6O_8 inorganic structural building units (SBUs) and organic SBUs with fluorine substituents. The QSPR models can serve as efficient filtering tools to detecting promising high-performing candidates at the early stage of virtual high-throughput screening of novel porous materials. The predictive models of the gas uptake capacities of MOFs are available online via our MOF informatics analysis (MOFIA) tool.



1. INTRODUCTION

Metal−organic frameworks (MOF) are a class of nanoporous solids with exceptional host−guest properties for gas-separation applications. They are formed by the self-assembly of metal ions or clusters and polydentate organic linkers that function as structural building units (SBUs) to create open, crystalline frameworks, some of which possess 'world-record' internal surface areas of >6000 m^2/g.[1] The nearly limitless combinations of metal nodes and linkers can yield a diverse range of structural robustness, flexibility, functionality, pore volume, and surface area which are suitable for gas separation[2] and other applications such as gas storage,[3,4] catalysis,[5−7] nonlinear optics,[8] sensing,[9,10] and light-harvesting.[11]

The building block approach to MOF synthesis makes it possible to target desired properties by linking molecular building blocks with specific features.[12] However, extensive trial-and-error is required to find the precise synthetic conditions (e.g., temperature, pressure, solvent, solute concentration) that result in the organic and inorganic SBU's to assemble into a three-dimensionally networked and crystalline material. In principle, high throughput (HT) experimental screening and combinatorial chemistry techniques widely adopted in the pharmaceutical industry[13] can overcome this to rapidly explore the parameter space. However, there are still many challenges to overcome to develop general HT methods for MOFs, and only a handful of high-throughput (HT) experimental data sets are available on hydrothermal stability[14] and gas adsorption capabilities[15] of MOFs.[16−18]

In contrast, the large scale *in silico* HT screening of MOFs for their gas adsorption properties has been recently realized[15,19] thereby boosting the availability of *in silico* gas adsorption data.[20] For example, Wilmer et al.[20] recently performed large-scale screening of ∼130,000 hypothetical MOF structures where grand canonical Monte Carlo (GCMC) simulations were used to predict methane, N_2, and CO_2 uptake capacities of each MOF under different CO_2 capture conditions. With the large data sets generated in these studies, one might expect that cheminformatic methods would be valuable analytic tools.

A variety of chemoinformatic methods have been implemented to model and retrieve information from HT experiments on bioactive compounds but not on porous functional materials. Structural similarity using 2D and 3D approaches[21] have yet to be widely extended to periodic crystals. Indeed, the similarity analysis of diverse data sets of porous solids using

Received: April 30, 2013
Revised: June 15, 2013
Published: June 17, 2013

   dx.doi.org/10.1021/jp404287t | J. Phys. Chem. C 2013, 117, 14095−14105

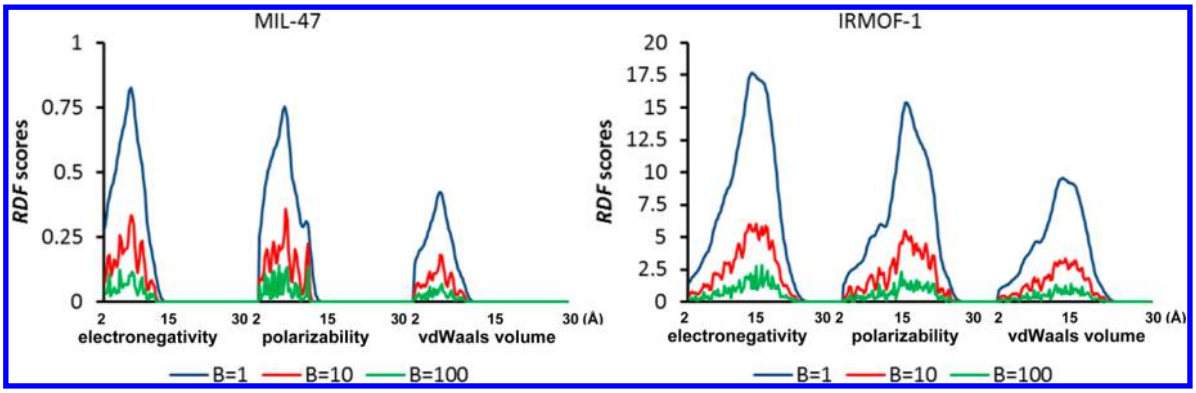


Figure 1. Property weighted RDF profiles of the MOFs MIL-47 and IRMOF-1 for the atomic properties of the electronegativity, polarizability, and van der Waals volume. Three different values of the smoothing parameter B (1, 10, and 100) in the distance range of 2.0 to 30.0 Å with an incremental step of 0.25 Å are shown.

specialized descriptors has been reported just recently on hypothetical zeolites.[22,23] Quantitative structure property relationship (QSPR) models[24,25] would be expected to be relevant since the gas adsorption properties should depend upon a MOF's geometric and chemical structure. The implementation of powerful machine learning algorithms such as Artificial Neural Networks (ANN)[26] and Support Vector Machines (SVM)[27] inside the QSPR framework can increase the accuracy of predictions and extract insights from large compound databases. However, to the best of our knowledge, only a handful of studies report QSPR analyses on MOFs,[28−33] with most of them using less than a dozen structures. Recently, we found that high quality QSPR models could be developed using only geometric descriptors to predict the methane uptake capacity of MOFs under high pressure conditions pertinent to natural gas storage.[34] More specifically, a QSPR model based on the dominant pore size, void fraction, and gravimetric surface area was developed that could predict the methane uptake capacity at 100 bar with a R^2 value of 0.93 on a test set of ∼120,000 MOFs. However, using the same library of MOFs, a preliminary QSPR analysis revealed poor correlation of geometrical features with the CO_2 capture properties at pressures relevant for gas separation applications.

In this paper, we show that the combination of specialized descriptors and sophisticated chemoinformatic methods can provide more comprehensive approaches for structure−property relationships of large-scale data sets and hence further contribute toward the computer-aided design of MOFs and related materials. We introduce an atomic property weighted radial distribution function (AP-RDF) descriptor tailored for the comprehensive QSPR analysis of gas adsorption of porous solids. The simulated methane, CO_2, and N_2 uptake capacities under gas separation conditions of ∼58,000 MOFs from the Northwestern University database were comparatively investigated using geometrical features and the newly introduced AP-RDF scores. We explored the ability of AP-RDF scores to recognize high-performing MOFs using principal component analysis (PCA) and QSPR models. The accuracies of QSPR models trained on the AP-RDF scores were then compared to that of the geometrical features, and the prediction performance was investigated across different building blocks in the data set. Finally, we described a webserver application to predict gas uptake capacities of MOFs.

2. METHODS

2.1. Radial Distribution Functions Descriptors for MOFs. Although simple geometrical features, such as the pore size or surface area, are amenable to interpretation; they are one-dimensional in nature and have been shown to correlate poorly to CO_2 separation parameters.[35−37] Consequently, we need a structure encoding descriptor capable of producing coherent description of the MOF across multiple length scales. In this direction, we found the pair distribution function or radial distribution function (RDF) analysis to be attractive. The RDF analysis is a crystallographic technique sensitive to both short- and long-range structural correlations that is growing in popularity for analysis of both materials and molecular systems.[38,39] To put in simple terms, the RDF is the interatomic separation histogram representing the weighted probability of finding a pair of atoms separated by a given distance.[39] In a crystalline solid, the RDF plot has an infinite number of sharp peaks whose separations and heights are characteristic of the lattice structure. In our case we used the minimum image convention and our RDF scores were uniquely defined inside the crystal unit cell.

The application of RDF scores in chemoinformatics was first proposed by Gasteiger and co-workers[40,41] to encode the 3D structure of molecules and has been used by others to examine different problems.[42,43] The 3D coordinates of the atoms of the system are transformed into a numerical RDF code that is independent of the size of the structure. Our adapted RDF scores in a MOF framework can be interpreted as the weighted probability distribution to find an atom pair in a spherical volume of radius R inside the unit cell according to the eq 1.

$$\mathrm{RDF}^{P}(R) = f \sum_{i,j}^{\text{all atom pairs}} P_i P_j \, e^{-B(r_{ij}-R)^2} \tag{1}$$

In eq 1, the summation is over all atom pairs in the unit cell, and r_{ij} is the minimum image convention distance of these pairs. B is a smoothing parameter, while f is simply a scaling or normalization factor. The RDFs can be weighted to fit the requirements of the chemical information to be represented, by introducing the atomic properties P_i. We weight the radial probabilities by three tabulated atomic properties namely electronegativity, polarizability, and van der Waals volume, whose values are given in Table S1 of the Supporting Information. While a regular RDF function encodes geometric features, the atomic property weighted radial distribution

dx.doi.org/10.1021/jp404287t | J. Phys. Chem. C 2013, 117, 14095−14105

function (AP-RDF) additionally characterizes the chemical features within a MOF.

The AP-RDF profiles for two well-known MOFs, MIL-47[44] and IRMOF-1,[45] are shown in Figure 1 for three atomic properties, the electronegativity, polarizability, and van der Waals volume. The effect of the smoothing parameter B on the AP-RDF profiles is also given in Figure 1. As can be observed, the lower the value of the parameter B, the higher the absolute value of the scores and the smoother the shape of the AP-RDF profile. Consequently, AP-RDF profiles for B = 100 were very noisy and those for B = 1 were featureless. Both extremes hindered the correlation with functional properties and, we found that B = 10 yielded an optimum trade-off between structural information content and smoothness.

A smoothing parameter B = 10 was used for all AP-RDF profiles used in this work. Additionally, a distance range of 2.0 to 30.0 Å in steps of 0.25 Å was used for all calculations. In this way, we computed 113 scores per atomic property, yielding a total of 339 AP-RDF scores or descriptors for each MOF. Both the AP-RDF scores and geometrical features were linearly normalized in the range $[-1, 1]$, where -1 and 1 correspond to the minimum and maximum values of each variable, respectively.

2.2. Data Set and Prior Preparation. The Northwestern University database developed by Wilmer et al. that contains 137,953 hypothetical MOF structures generated from 102 structural building units (SBU) was utilized.[20] Figure 2 depicts a few of those SBUs, while Figure S1 in the Supporting Information shows the complete set. The details of the MOF structure creation algorithm are given in the original work of Wilmer et al.[20] From a total of ∼130,000 MOF structures in the database, only ∼83,000 are unique frameworks. The remaining ∼55,000 MOFs are interpenetrated structures that were generated by arbitrary concatenation of two or more frameworks. In order to reduce the redundancy in the data set, we only included the ∼83,000 non-interpenetrated MOF structures.

MOFs can be used to capture CO_2 from gas mixtures via pressure-swing adsorption (PSA) or vacuum-swing adsorption (VSA), where the material is exposed to impure gas at high pressure and then regenerated at lower pressure. Recently, Bae and Snurr reported five "adsorption evaluation criteria" to assess the performance of porous solids in PSA and VSA processes, namely, the uptake under adsorption conditions, selectivity under adsorption conditions, the working capacity, the regenerability of the sorbent, and the sorbent selection parameter.[35] However, instead of approaching all criteria simultaneously, one can save considerable amount of computational time and resources by first identifying a reduced subset of materials with high uptake capacities at adsorption conditions. This can be followed by further evaluation of other performance criteria such as working capacity, selectivity, and regenerability using more computationally intense simulations. Therefore, we focused on the QSPR modeling of the methane, N_2 and CO_2 uptake capacities of MOFs under PSA and VSA conditions in Table 1. The uptake capacities were computed using GCMC simulations as described by Wilmer et al. and kindly provided by these authors.[46]

A total of six geometrical features, namely, the dominant pore diameter, the maximum pore diameter, the void fraction, the gravimetric surface area, the volumetric surface area, and the density were taken from the Northwestern University database along with the MOF structures. The details of the calculations

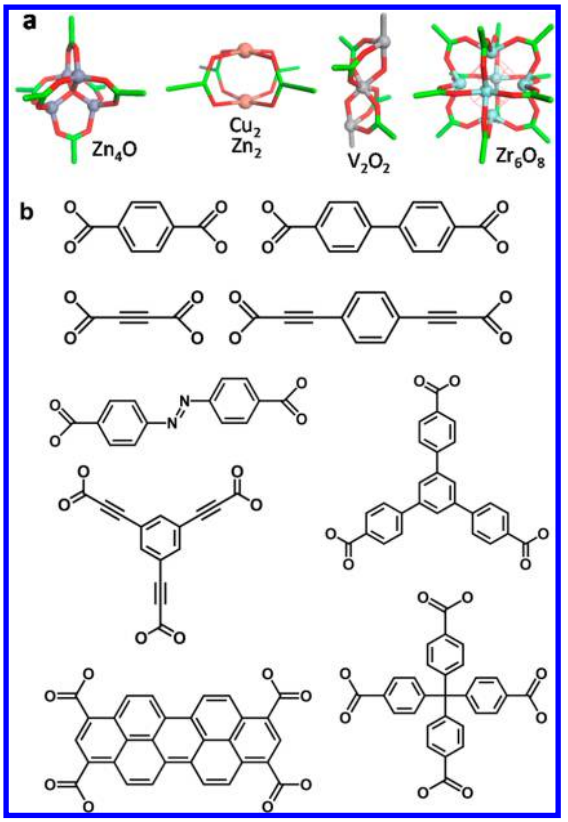


Figure 2. Examples of the inorganic (a) and organic (b) SBUs used to generate the ∼130,000 hypothetical MOFs in the Northwestern University database. The inorganic SBUs are depicted in tube representation, with the metal atoms shown as balls. Shown with the inorganic SBUs are the carboxylate fragments of the organic SBUs, where the carbon atoms of these fragments are shown in green to highlight the coordination geometry. The central part of the Zr_6O_8 SBU is represented in wire representation for clarity. The full list of building blocks appears in Figure S1 in Supporting Information.

Table 1. CO_2 Separation Cases That Correspond to the Adsorption Data Collected from Wilmer et al.[46] Used to Train the QSPR Models[a]

case	application[b]	mixture composition	adsorption pressure (bar)
1	Natural gas purification using PSA	CO_2–CH_4 = 10:90	5
2	Landfill gas separation using PSA	CO_2–CH_4 = 50:50	5
3	Landfill gas separation using VSA	CO_2–CH_4 = 50:50	1
4	Flue gas separation using VSA	CO_2–N_2 = 10:90	1

[a]Temperature is 298 K in all cases. [b]PSA and VSA refer to pressure swing adsorption and vacuum swing adsorption gas separation processes, respectively.

of the geometrical features appear in ref 46. To build the QSPR models, 70% of the data set (∼58,000 MOFs) was randomly selected to train/calibrate the models and the remaining 30% of the data set (∼25,000 MOFs) was used as test set to evaluate the predictive power.

2.3. Optimization of the Regression Models. To correlate the uptake capacities to the descriptors (the AP-RDF scores), we used two regression techniques: multilinear regression (MLR) and support vector machine (SVM). While

dx.doi.org/10.1021/jp404287t | J. Phys. Chem. C 2013, 117, 14095−14105

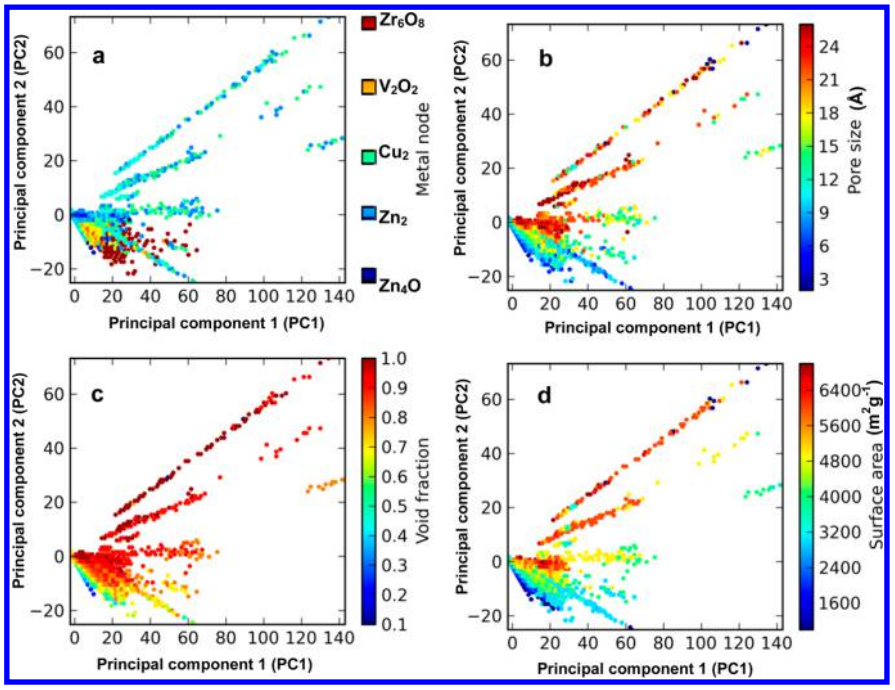


Figure 3. Scatter plots of the first principal component (PC1) vs the second principal component (PC2) of the RDF scores. The dots are colored according to the averaged values of metal type (a), pore size (b), void fraction (c), and surface area (d).

MLR is a simple linear regressor, the SVMs[47] are a machine learning method of broad applicability to many types of pattern recognition problems (see Supporting Information for SVM implementation details). Using only the MOFs in the training set, we developed QSPR models to fit the simulated uptake capacities.

The quality of the fit of the training set (the calibration of the regression model) was measured by its R^2 value according to eq 2. N is the number of compounds, Y_i and Upt_i are the predicted and GCMC uptake capacities, respectively, calculated for MOF i. $\overline{\mathrm{Upt}}$ is the average uptake. Hereafter, we will use the term 'actual' value to refer to the GCMC calculated functional properties for which we are trying to predict. When computed on the training set, R^2 measures the goodness of the calibration fit of the model, while computed for the test set, R^2 is an important measure of the QSPR model's predictive capability.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(Y_i - \mathrm{Upt}_i)^2}{\sum_{i=1}^{N}(Y_i - \overline{\mathrm{Upt}})^2} \quad (2)$$

An internal 3-fold-out (TFO) cross-validation was carried out during the grid search optimization of the SVM hyper-parameters to calculate Q^2 value, an equivalent statistic to R^2. Here we divide the training set into three subsets and then remove one, while the other two are used to fit the regression model. The resulting model is then compared against the simulation data for the left-out subset. This process is repeated until all the subsets have been validated against each other. Equation 2 is used to measure the fit of the model against each left-out subset, except that we denoted this squared correlation coeficient as Q^2 to distinguish from the case where the whole training set is considered. We expect a good QSPR model to have none or very low overfitting with similar R^2 and Q^2 values of training, cross-validation, and test set predictions.

3. RESULTS AND DISCUSSION

3.1. Principal Components Analysis of the AP-RDF Scores. A preliminary structure−performance relationship analysis on this data set using independent correlations of MOFs geometrical properties vs gas uptake was reported by Wilmer et al.,[46] showing poor correlations and rather disperse patterns. Therefore, we initially performed here a qualitative structure−uptake relationship study using principal component analysis (PCA) of the AP-RDF scores. The PCA reduces the dimensionally of the data enabling a preliminary qualitative graphical interpretation. The principle components (PCs) are linear combinations of the original AP-RDF scores such that that the first principal component has the largest possible variance of the AP-RDF scores and each succeeding component in turn has the highest variance possible given the constraint that it be orthogonal with the preceding components (see Supporting Information for PCA details).

Shown in Figure 3 are scatter plots of the second principle component (PC2) of the AP-RDF scores as a function of the first principle component (PC1) of the AP-RDF scores. The same PC1 vs PC2 plot is shown in Figures 3a−d, but with different coloration based on the metal type (3a), pore size (3b), void fraction (3c), and surface area (3d). The first two principle components (PC1 and PC2) were found to have ∼70% and ∼15% of the variance of the original AP-RDF scores, respectively. We found that the PC1 has positive values for all MOFs, while the PC2 was either negative or positive for ∼50% of the data set. It is noteworthy that the PCA clustered the MOFs along straight lines corresponding to different interaction levels of the PCs. Interestingly, Figure 3a reveals that the majority of MOFs with the inorganic paddle wheel SBU (Zn_2 and Cu_2 in Figure 3a) are allocated in the positive region of the PC2 axis. In contrast, MOFs with inorganic SBU Zn_4O showed PC2 values close to zero, while MOFs with more 'complex' inorganic SBUs, V_2O_2 and Zr_6O_8, had negative values. These values can be explained by the fact that at short

dx.doi.org/10.1021/jp404287t | J. Phys. Chem. C 2013, 117, 14095−14105

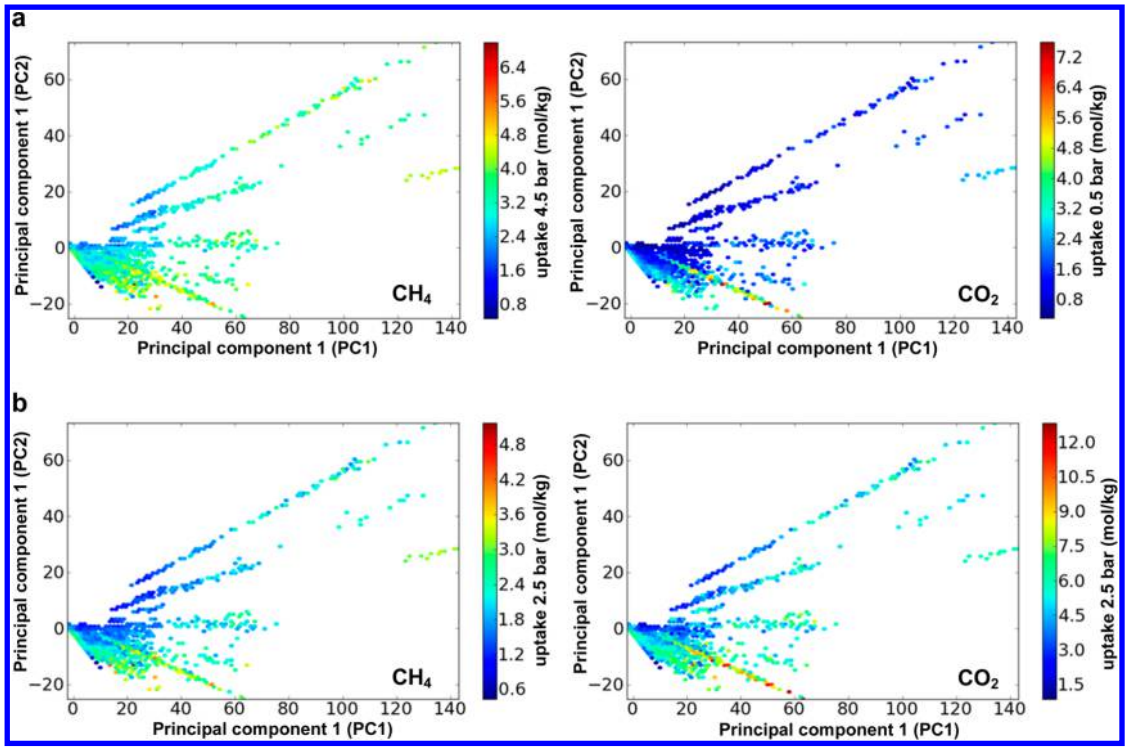


Figure 4. Scatter plots of the first principal component (PC1) vs the second principal component (PC2) of the AP-RDF scores of MOFs. The dots are colored according to the uptake capacities for CH_4 and CO_2 for (a) natural gas purification using PSA (case 1 of Table 1) and (b) for landfill gas separation using PSA (case 2).

distances the PC2 interacts with negative sign with the AP-RDF scores (Figure S2 in Supporting Information) yielding negative PC2 values for MOFs with the highest AP-RDF scores at the shorter distances. This is the case of the inorganic SBUs, V_2O_2 and Zr_6O_8, which form MOFs with complex topologies and mediocre pore size in the range of 4 Å to 12 Å as shown in Figure 3b.

Furthermore, the scatter plots in Figure 3b showed that pore sizes <10 Å had the lowest values of PC2. Meanwhile, MOFs with void fractions >0.8 in Figure 3c corresponded to higher values of the PCs. Finally, Figure 3d depicted a similar pattern where higher values of the PC2 corresponded to higher gravimetric surface areas. In this case, the positive coefficients of the PC2 in the large distance range of 15 to 30 Å in Figure S2 in Supporting Information should yield positive values of the PC2 for large MOFs which usually exhibit high surface area.

The PCs of the AP-RDF scores show some correlation with the uptake capacity of the MOFs. Given in Figure 4 are the PC1 vs PC2 plots color-coded to the methane and CO_2 uptake capacities at PSA conditions for (a) natural gas separation (case 1 of Table 1) and (b) landfill gas separation (case 2). In these plots, the dots on the left-hand-side that correspond to low PC1 values tend to have lower uptakes, while dots near the bottom of the plots corresponding to low PC2 values tend to have higher uptakes. Similar trends were also observed for lower pressure gas adsorption conditions for cases 3 and 4 of Table 1, but these are not shown.

The potential of the AP-RDF approach to uniquely characterize porous structures at a qualitative level was illustrated by the PCA transform where the two main PCs successfully identified structural similarities among MOFs based on inorganic SBUs, geometrical property ranges, and more surprisingly gas uptake capacities.

3.2. Correlation of High Pressure Methane Uptake with AP-RDF Descriptors. The PCA showed that the AP-RDF scores have potential to correlate to the uptakes capacities; thus, the next step in our study was to develop quantitative correlation models of the uptake capacities. We first examine the uptake of methane at high pressure in which we have recently been able to build good QSPR models using 1-D geometric descriptors only.[34] In this previous study, nonlinear SVM models were constructed that could predict the methane storage capacity of a test set of ~120,000 MOFs from the Northwestern database with R^2 values of 0.82 and 0.93 at 35 and 100 bar, respectively. It would be of value to determine if QSPR models of similar quality could also be built using AP-RDF descriptors.

We found that the AP-RDF scores yielded outstanding predictions of the methane uptake at high pressure comparable to that of the 1-D geometric descriptors. A nonlinear SVM model yielded R^2 = 0.83 at 35 bar which emulates our QSPR model at this pressure using pore size, void fraction, and surface area.[34] Meanwhile at 100 bar, the AP-RDF predictions exhibit a higher R^2 = 0.88, which is comparable to but less than the R^2 value of 0.93 of the geometric descriptor predictions.[34] Since the AP-RDF descriptors can generate QSPR models of similar quality to more specific geometric descriptors pore size, void fraction, and surface area, this shows that the AP-RDF descriptors can also capture the geometric features of MOFs.

3.3. Training of AP-RDF models of CO_2, N_2, and CH_4 Uptake. We now turn our attention to examine the uptake of CO_2, N_2, and CH_4 under conditions (Table 1) where preliminary QSPR analysis revealed there was poor correlation of geometrical features with uptake. For the adsorption conditions given in Table 1, we initially built linear MLR models using a set of six geometrical features, namely, the

dx.doi.org/10.1021/jp404287t | J. Phys. Chem. C 2013, 117, 14095−14105

dominant pore diameter, the maximum pore diameter, the void fraction, the gravimetric surface area, the volumetric surface area, and the density. When MLR models yielded very poor Q^2 values <0.5 (data not shown), we further tried to improve the prediction accuracy of the geometrical features using more powerful nonlinear SVM regressions. However, although the accuracies were improved, the Q^2 values were still low in the range of 0.46 to 0.63, as shown in Table 2.

Table 2. Details and Cross-Validation Statistics[a] of the Optimum SVM Models for Gas Uptake Capacity of MOF at 298 K and Different Pressures Using Geometric and AP-RDF Descriptors

			Q^2	
case	gas	adsorption pressure (bar)	geometric	AP-RDF
1	CO_2	0.5	0.62	0.75
	CH_4	4.5	0.46	0.83
2	CO_2	2.5	0.57	0.78
	CH_4	2.5	0.52	0.82
3	CO_2	0.5	0.62	0.75
	CH_4	0.5	0.61	0.77
4	CO_2	0.1	0.63	0.69
	N_2	0.9	0.53	0.70

[a] Q^2 is the squared correlation coefficient of the cross-validation regression performed on a training set of ~58,000 MOFs.

Preliminary linear MLR models using the AP-RDF scores also exhibited poor correlations with Q^2 <0.55 (data not shown). However, nonlinear SVMs model using the AP-RDF scores generally yielded notable improvement over the SVM models using the geometric parameters. The SVM AP-RDF models (AP-RDF models for the rest of the paper) yielded Q^2 values in the range of 0.69 to 0.83 as given in Table 2. The correlation of AP-RDF scores with the uptake capacities of the MOFs was higher at higher pressures. While the AP-RDF model of methane uptake capacity at 4.5 bar in Case 1 exhibited the highest Q^2 value of 0.83, the CO_2 uptake capacity at 0.5 bar in Case 1 was predicted with a lower Q^2 value of 0.75. The accuracy gap is probably related to the fact that a larger fraction of the void space of the MOF can be filled at higher pressure in comparison to at lower pressure. Therefore, it is expected that the uptake capacity at higher pressures correlates more strongly with the global spatial distribution of atoms in the framework that ultimately delimited the void space. In turn, when only a small fraction of the void space can be filled at low pressures the correlation with the global features of the void is weaker because more localized interactions between the gas molecules and the framework can play a more important role in the uptake capacity. These local interactions at lower pressures are probably very diverse in nature and, therefore, more difficult to generalize by the QSPR models to the entire training set.

Similarly, the methane and CO_2 uptake capacities for landfill gas separation by PSA in Case 2 were also predicted with different Q^2 values of 0.82 and 0.78, respectively. Despite the partial pressures of methane and CO_2 in Case 2 being equal, the CO_2 adsorption was predicted with lower accuracy. In this case, the accuracy difference is likely due to the more specific binding of CO_2. More specifically, methane uptake can be accurately modeled without considering electrostatic interactions since the molecule has no net charge, dipole moment, or quadrupole moment. On the other hand, CO_2 has a non-

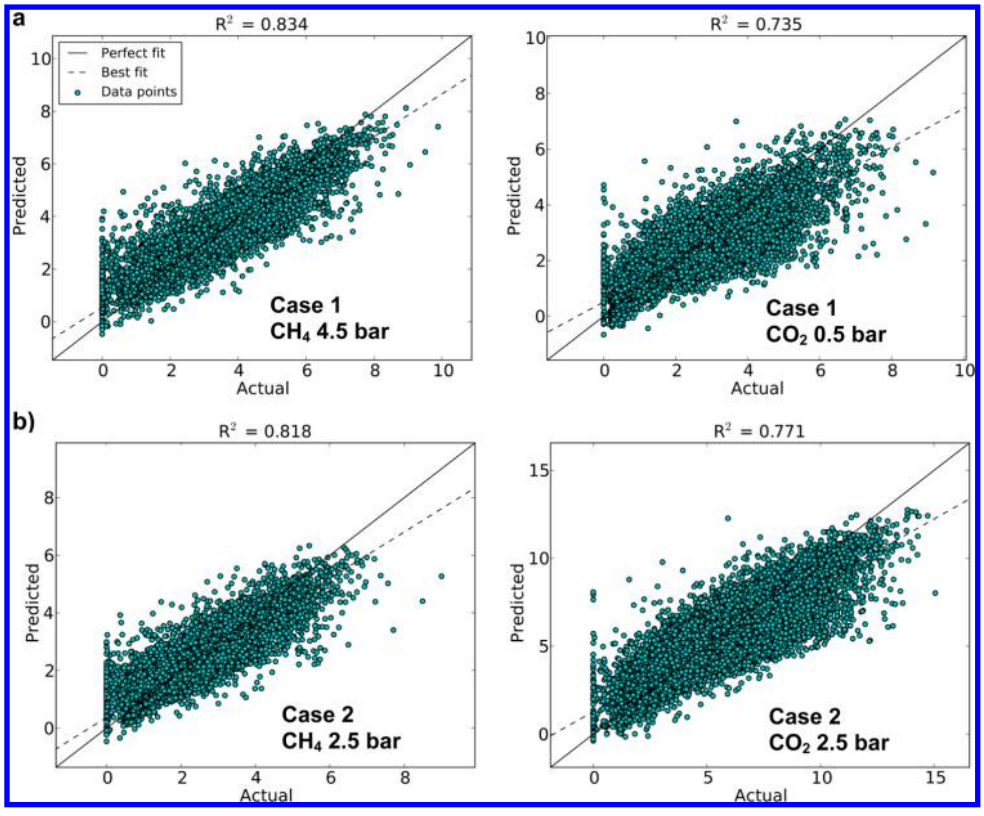


Figure 5. Scatter plots of QSPR predicted vs actual (GCMC) uptake capacities of the ~25,000 MOFs in the test set using AP-RDF scores under PSA conditions relevant to (a) natural gas purification (Case 1) and (b) landfill gas separation (Case 2).

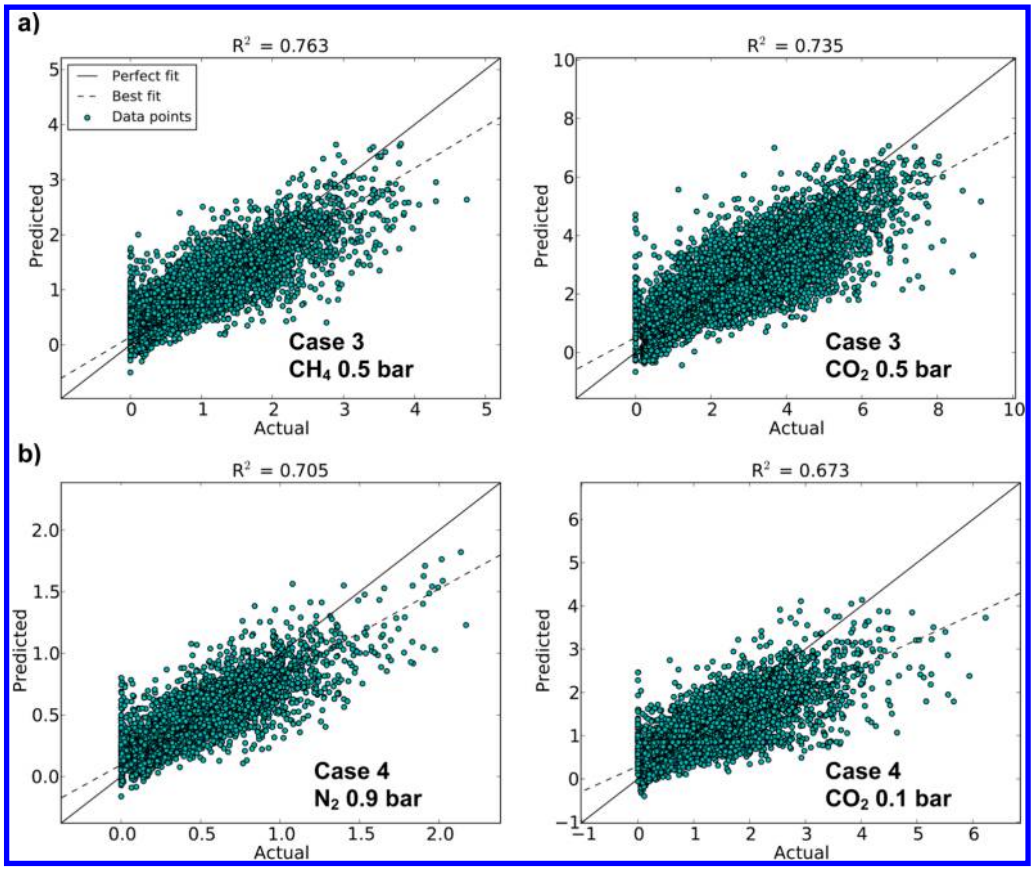


Figure 6. Scatter plots of QSPR predicted vs actual (GCMC) uptake capacities of the ~25,000 MOFs in the test set using AP-RDF scores under VSA conditions relevant to (a) landfill gas separation and (b) flue gas separation.

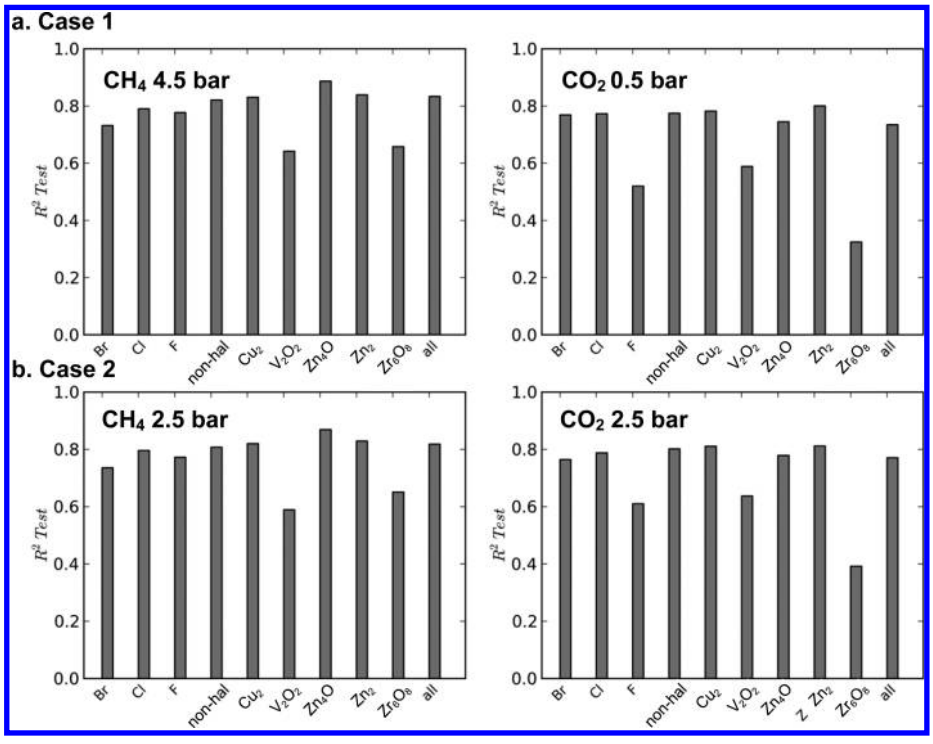


Figure 7. Test set R^2 values of the uptake capacities at adsorption conditions in (a) Case 1 and (b) Case 2 categorized according to different chemical building blocks. 'nonhal' refers to MOFs functionalized with any nonhalogen functional group. Labels for the metal containing inorganic SBUs are defined in Figure 2a.

dx.doi.org/10.1021/jp404287t | J. Phys. Chem. C 2013, 117, 14095−14105

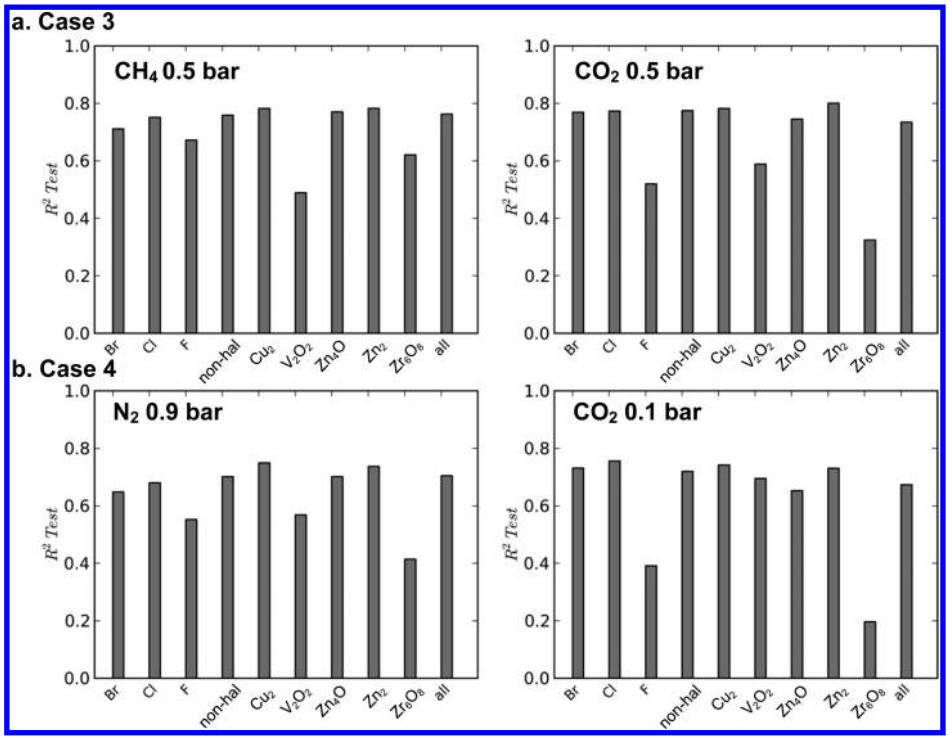


Figure 8. Test set R^2 values of the uptake capacities at adsorption conditions in Case 3 (A) and Case 4 (B) according to different building blocks. 'nonhal' refers to MOFs functionalized with any nonhalogen functional group. Labels for the metal containing SBUs are defined in Figure 2a.

negligible quadrupole moment which can stabilize host−guest interactions via complementary electrostatic contributions.[15] These interactions are likely more sensitive to the local shape and chemistry of the pores that can be somewhat elusive to the AP-RDF scores as previously discussed.

Finally, despite the low partial pressures of VSA separations, the optimum AP-RDF models of landfill gas separation Case 3 also performed adequately well with Q^2 values for methane and CO_2 uptake capacities of 0.77 and 0.75, respectively. Similarly, the SVMs yielded encouraging results for CO_2 capture from postcombustion flue gas in Case 4 with Q^2 values of 0.69 and 0.70 for the uptake of CO_2 and N_2, respectively.

3.4. Predictive Capability of AP-RDF Models of CO_2, N_2, and CH_4 Uptake. The AP-RDF scores yielded accurate SVM models of the uptake capacities of the ∼58,000 MOFs in the training set with cross-validation Q^2 values >0.7. However, the predictions need further validation on an external test set. For doing this, we have selected a test set of ∼25,000 MOFs that were not used during the calibration process. The scatter plots of predicted vs GCMC uptake capacities of the ∼25,000 MOFs in the test set for cases 1 and 2 of Table 1 are provided in Figure 5, where "actual" refers to the values calculated by GCMC simulations and "predicted" corresponds to the QSPR predictions.

In general, the R^2 values of the test set predictions matched the cross-validation accuracies in Table 2. This reveals that the SVM models not only "learned" the structural information from the training set with adequate accuracies but more importantly generalized it to the test set yielding comparable predictions without overfitting. Consequently, the highest prediction accuracy corresponds to the methane uptake capacities in Case 1 (Figure 5a) with R^2 value of 0.83, while the CO_2 adsorption in the same case is predicted with R^2 value of 0.73. For Case 2 in Figure 5b, the scatter plots of the test set also yielded high R^2 values of 0.82 and 0.77 for methane and CO_2, respectively. Noteworthy, the dotted lines of "best fit" are all close to the straight lines of "perfect fit" in Figure 5, which indicates that the QSPR models predicted the test set with very low systematic error along the entire uptake ranges.

Figure 6 shows scatter plots of the QSPR predicted vs actual uptake capacities at low pressures relevant to gas separation using VSA technology. Methane and CO_2 uptake capacities were predicted with R^2 of 0.76 and 0.74, respectively, for landfill gas separation (case 3) as shown in Figure 6a. Figure 6b shows the scatter plots of test set predictions for VSA postcombustion CO_2 capture (case 4). Although this case yielded the lowest accuracies, the R^2 values also matched the Q^2 in Table 2 with agreeable values of 0.71 and 0.67 for N_2 and CO_2, respectively.

3.5. Predictive Capability across SBUs and Functional Groups. Examination of Figures 5 and 6 reveals that the QSPAR models underestimate the uptake capacities of several high-performing MOFs, particularly under low pressure VSA conditions. This brings up the question of how the QSPR models perform for specific SBUs or functional groups. To investigate the prediction accuracies across the chemical diversity of the database, the MOFs were labeled according to the inorganic SBUs and the functionalization of the organic SBUs and the prediction accuracy was measured for each category. Figures 7 and 8 show the R^2 values divided into MOFs in the test set functionalized with Br, Cl, F, and all other nonhalogen functional groups. The MOFs are also divided into those with the metal containing inorganic SBUs shown in Figure 2a.

The uptake capacities under PSA conditions (cases 1 and 2) are given in Figure 7. For methane uptake, similar good accuracies ($R^2 \sim 0.8$) across the different functional groups is observed, with the exception of MOFs containing the inorganic SBUs V_2O_2 and Zr_6O_8, where the uptakes are predicted with lower accuracies ($R^2 < 0.7$). The prediction accuracies of these two inorganic SBUs is also significantly lower for CO_2 uptake,

dx.doi.org/10.1021/jp404287t | J. Phys. Chem. C 2013, 117, 14095−14105

where MOFs with the Zr_6O_8 SBU are predicted the least well ($R^2 < 0.5$). For CO_2 uptake, it is also notable that fluorine functionalization showed lower prediction accuracy with $R^2 < 0.6$.

Figure 8 shows the uptake capacities under lower pressure VSA conditions (Case 3 and 4), divided into different chemical types. Following a similar trend, MOFs containing the inorganic SBUs V_2O_2 and Zr_6O_8 typically showed significantly lower accuracies. This was also observed for N_2, a gas whose uptake is not considered in Cases 1 and 2. For N_2 adsorption at 0.9 atm, the prediction accuracy with the Zr_6O_8 SBU is low with $R^2 < 0.6$. CO_2 uptake at very low pressure (0.1 atm) found in Case 4 also shows particularly poor prediction accuracies for MOFs containing the Zr_6O_8 inorganic SBU and fluorine containing MOFs with R^2 of ∼0.2 and 0.4, respectively.

There are several reasons the QSPR models are systematically less accurate for the MOFs constructed from V_2O_2 or Zr_6O_8 SBUs or functionalized with fluorine. There were generally fewer MOFs in the database with V_2O_2 or Zr_6O_8 metal nodes or functionalized with fluorine, and so the SVMs may not have been as well trained on these MOFs. For the V_2O_2 or Zr_6O_8 containing MOFs it is possible that gas adsorption in more complex frameworks is inherently more challenging to predict. For example, a large number of the Zr_6O_8 containing MOFs in the database are not fully connected through the organic linkers but capped with small organic ligands. This was performed in a variety of ways yielding a rather diverse void space distribution for MOFs with Zr_6O_8 SBUs that is likely difficult to generalize to other materials in the database.

The low prediction accuracy for fluorine-containing MOFs is particularly discouraging considering that they are among the best performing in the database.[46] We found that the SVM models underestimated the CO_2 uptake capacities of the majority of the fluorine-functionalized MOFs mainly at low pressures (also Figure S3 in Supporting Information), possibly due to the unique local chemical environment generated by the fluorine atoms that is not captured by the AP-RDF descriptors. It is also possible that the gas adsorption data from the Northwestern University database is affected by systematic errors. For example, partial atomic charges used in the simulations are derived from a rapid, highly parametrized method and we cannot rule out that the gas adsorption from the GCMC simulations are biased toward fluorine-containing MOFs yielding exceptionally high CO_2 uptake values.

The property weighted AP-RDF models produce adequate estimates of the simulated uptake capacities of MOFs for CO_2, N_2, and CH_4 capture at conditions relevant for PSA and VSA gas separations. Nevertheless, one should pay attention to the applicability domain of the QSPR models defined by the diversity of the training data set. In this regard, although the Northwestern MOF database used is sufficiently diverse, the authors acknowledge that the performance boundaries set in the database are not absolute and unexplored MOFs may exist with lower or higher adsorption values.[46] In any case, we can systematically improve the accuracy of the QSPR predictions and broaden the applicability domain by continuously updating the AP-RDF models upon the availability of new gas adsorption data on MOFs. Additionally, despite the fact that our models were developed using simulated data, there is no reason that the AP-RDF descriptors cannot be applied to experimentally generated data.

3.5. MOF Informatics Analysis (MOFIA) Webserver. We implemented the optimum AP-RDF models summarized in Table 2 into a webserver application so-called MOFIA (MOF Informatics Analysis). The web application takes the structure of a MOF in CIF format as input to compute the AP-RDF scores and then predicts the gas uptake capacities at each pressure. The webserver is freely available online at http://titan.chem.uottawa.ca/woolab/MOFIA.

4. CONCLUSIONS

The tremendous drive for efficient screening of large structural databases confers a steadily increasing role to sophisticated QSPR analysis in the current material discovery process. The availability of large structural databases of MOFs[20] and the efficient estimation of materials properties via computation will allow for QSPR tools to be integrated into studies of metal organic frameworks. In this work we introduce specialized structural descriptors for porous solids in the form of atomic property weighted radial distribution functions (AP-RDF) where the conventional RDF is modulated by tabulated atomic properties such as the electronegativity or van der Waals radii. The PCA transform of the AP-RDF scores successfully caught similarities among MOFs based on the inorganic SBUs, geometrical features, and, somewhat surprisingly, the gas uptake capacities. The AP-RDF descriptors were used to build QSPR models of simulated methane uptake at high pressure, which yielded R^2 values of 0.83 and 0.88 at 35 and 100 bar, respectively, on a test set of ∼120,000 MOFs. This was comparable to QSPR models constructed from 1-D geometric descriptors (pore size, void fraction, surface area), which suggests that the AP-RDF descriptors can capture such geometric features of MOFs. The AP-RDF descriptors were used to build QSPR models of CO_2, N_2, and CH_4 uptake at low pressures (0.1−0.9 bar) relevant to PSA/VSA gas separations applied to natural gas purification, landfill gas purification, and postcombustion CO_2 scrubbing. The introduced AP-RDF scores yielded QSPR models of simulated methane, CO_2, and N_2 uptake capacities that outperformed the combination of simple geometrical features with prediction accuracies ranging from ∼70% to ∼83%. To the best of our knowledge this is the first large-scale QSPR model of gas uptake capacities of MOFs for gas separation applications. The inaccuracies found at low pressures for specific types of MOFs may be related to very unique host−guest interactions or due to errors stemming from the approximate nature of the adsorption model. The QSPR models were implemented as webserver applications freely available online at http://titan.chem.uottawa.ca/woolab/MOFIA.

The atomic property weighted RDF descriptor is extensible in several ways. First, other atomic properties can be integrated into the AP-RDF framework such as the chemical hardness or other definitions of electronegativity. One intriguing possibility for its application to gas-uptake prediction is to use the partial atomic charges or the Lenard-Jones parameters used in the GCMC simulations. Furthermore, the approach should be applicable for the prediction of other properties in MOFs and relevant for other classes of crystalline materials as long as data is available to train the QSPR models. Efforts are currently underway in our lab to predict other gas adsorption parameters such as selectivity and heat of adsorption, as well as the parasitic energy of CO_2 capture.

■ ASSOCIATED CONTENT

Supporting Information

Support vector machines and principal component analysis details and additional figures and tables. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Corresponding author email: tom.woo@uottawa.ca.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors would like to thank Randall Snurr and Christopher Wilmer for the useful discussions and for providing the uptake data from their MOF structure database. Financial support from the Natural Sciences and Engineering Research Council of Canada, Carbon Management Canada, the University of Ottawa, and Canada Research Chairs Program is greatly appreciated, as well as the computing resources provided by Canada Foundation for Innovation and Compute Canada for providing computing resources. Financial support from the Natural Sciences and Engineering Research Council of Canada, Carbon Management Canada, the University of Ottawa, and Canada Research Chairs Program is greatly appreciated, as well as the computing resources provided by Canada Foundation for Innovation and Compute Canada.

■ REFERENCES

(1) Della Rocca, J.; Liu, D.; Lin, W. Nanoscale metal-organic frameworks for biomedical imaging and drug delivery. *Acc. Chem. Res.* **2011**, *44*, 957−968.

(2) Li, J.-R.; Kuppler, R. J.; Zhou, H.-C. Selective gas adsorption and separation in metal-organic frameworks. *Chem. Soc. Rev.* **2009**, *38*, 1477−1504.

(3) Farha, O. K.; Yazaydın, A. Ö.; Eryazici, I.; Malliakas, C. D.; Hauser, B. G.; Kanatzidis, M. G.; Nguyen, S. T.; Snurr, R. Q.; Hupp, J. T. De novo synthesis of a metal-organic framework material featuring ultrahigh surface area and gas storage capacities. *Nat. Chem.* **2010**, *2*, 944−948.

(4) Suh, M. P.; Park, H. J.; Prasad, T. K.; Lim, D.-W. Hydrogen storage in metal-organic frameworks. *Chem. Rev.* **2012**, *112*, 782−835.

(5) Ma, L.; Abney, C.; Lin, W. Enantioselective catalysis with homochiral metal-organic frameworks. *Chem. Soc. Rev.* **2009**, *38*, 1248−1256.

(6) Ma, L.; Falkowski, J. M.; Abney, C.; Lin, W. A series of isoreticular chiral metal-organic frameworks as a tunable platform for asymmetric catalysis. *Nat. Chem.* **2010**, *2*, 838−846.

(7) Lee, J.; Farha, O. K.; Roberts, J.; Scheidt, K. A.; Nguyen, S. T.; Hupp, J. T. Metal-organic framework materials as catalysts. *Chem. Soc. Rev.* **2009**, *38*, 1450−1459.

(8) Evans, O. R.; Lin, W. Crystal Engineering of NLO Materials Based on Metal−Organic Coordination Networks. *Acc. Chem. Res.* **2002**, *35*, 511−522.

(9) Chen, B.; Xiang, S.; Qian, G. Metal-organic frameworks with functional pores for recognition of small molecules. *Acc. Chem. Res.* **2010**, *43*, 1115−1124.

(10) Xie, Z.; Ma, L.; DeKrafft, K. E.; Jin, A.; Lin, W. Porous phosphorescent coordination polymers for oxygen sensing. *J. Am. Chem. Soc.* **2010**, *132*, 922−923.

(11) Kent, C. A.; Mehl, B. P.; Ma, L.; Papanikolas, J. M.; Meyer, T. J.; Lin, W. Energy transfer dynamics in metal-organic frameworks. *J. Am. Chem. Soc.* **2010**, *132*, 12767−12769.

(12) Ockwig, N. W.; Delgado-Friedrichs, O.; O'Keeffe, M.; Yaghi, O. M. Reticular chemistry: occurrence and taxonomy of nets and grammar for the design of frameworks. *Acc. Chem. Res.* **2005**, *38*, 176−182.

(13) Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug. Discovery* **2002**, *1*, 882−894.

(14) Low, J. J.; Benin, A. I.; Jakubczak, P.; Abrahamian, J. F.; Faheem, S. A.; Willis, R. R. Virtual high throughput screening confirmed experimentally: porous coordination polymer hydration. *J. Am. Chem. Soc.* **2009**, *131*, 15834−15842.

(15) Vaidhyanathan, R.; Iremonger, S. S.; Shimizu, G. K. H.; Boyd, P. G.; Alavi, S.; Woo, T. K. Direct observation and quantification of CO_2 binding within an amine-functionalized nanoporous solid. *Science* **2010**, *330*, 650−653.

(16) Banerjee, R.; Phan, A.; Wang, B.; Knobler, C.; Furukawa, H.; O'Keeffe, M.; Yaghi, O. M. High-throughput synthesis of zeolitic imidazolate frameworks and application to CO2 capture. *Science* **2008**, *319*, 939−943.

(17) Sumida, K.; Horike, S.; Kaye, S. S.; Herm, Z. R.; Queen, W. L.; Brown, C. M.; Grandjean, F.; Long, G. J.; Dailly, A.; Long, J. R. Hydrogen storage and carbon dioxide capture in an iron-based sodalite-type metal−organic framework (Fe-BTT) discovered via high-throughput methods. *Chem. Sci.* **2010**, *1*, 184−191.

(18) Wollmann, P.; Leistner, M.; Stoeck, U.; Grünker, R.; Gedrich, K.; Klein, N.; Throl, O.; Grählert, W.; Senkovska, I.; Dreisbach, F.; Kaskel, S. High-throughput screening: speeding up porous materials discovery. *Chem. Commun. (Camb.)* **2011**, *47*, 5151−5153.

(19) Keskin, S.; Van Heest, T. M.; Sholl, D. S. Can metal-organic framework materials play a useful role in large-scale carbon dioxide separations? *ChemSusChem* **2010**, *3*, 879−891.

(20) Wilmer, C. E.; Leaf, M.; Lee, C. Y.; Farha, O. K.; Hauser, B. G.; Hupp, J. T.; Snurr, R. Q. Large-scale screening of hypothetical metal-organic frameworks. *Nat. Chem.* **2012**, *4*, 83−89.

(21) Livingstone, D. The characterization of chemical structures using molecular properties. A survey. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 195−209.

(22) Martin, R. L.; Smit, B.; Haranczyk, M. Addressing challenges of identifying geometrically diverse sets of crystalline porous materials. *J. Chem. Inf. Model.* **2012**, *52*, 308−318.

(23) Martin, R. L.; Willems, T. F.; Lin, L.-C.; Kim, J.; Swisher, J. a; Smit, B.; Haranczyk, M. Similarity-driven discovery of zeolite materials for adsorption-based separations. *ChemPhysChem* **2012**, *13*, 3595−3597.

(24) Sheridan, R. P. Searching for Pharmacophores in Large Coordinate Data Bases and Its Use in Drug Design. *Proc. Natl. Acad. Sci. U.S.A* **1989**, *86*, 8165−8169.

(25) Cramer, R. D.; Poss, M. A.; Hermsmeier, M. A.; Caulfield, T. J.; Kowala, M. C.; Valentine, M. T. Prospective Identification of Biologically Active Structures by Topomer Shape Similarity Searching. *J. Med. Chem.* **1999**, *42*, 3919−3933.

(26) Hall, L.; Kier, L. The E-state as the basis for molecular structure space definition and structure similarity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 784−791.

(27) Urbano-Cuadrado, M.; Carbó, J. J.; Maldonado, A. G.; Bo, C. New quantum mechanics-based three-dimensional molecular descriptors for use in QSSR approaches: application to asymmetric catalysis. *J. Chem. Inf. Model.* **2007**, *47*, 2228−2234.

(28) Kim, D.; Kim, J.; Jung, D. H.; Lee, T. B.; Choi, S. B.; Yoon, J. H.; Kim, J.; Choi, K.; Choi, S.-H. Quantitative structure−uptake relationship of metal-organic frameworks as hydrogen storage material. *Catal. Today* **2007**, *120*, 317−323.

(29) Gaudin, C.; Cunha, D.; Ivanoff, E.; Horcajada, P.; Chevé, G.; Yasri, A.; Loget, O.; Serre, C.; Maurin, G. A quantitative structure activity relationship approach to probe the influence of the functionalization on the drug encapsulation of porous metal-organic frameworks. *Microporous Mesoporous Mater.* **2012**, *157*, 124−130.

(30) Wu, D.; Yang, Q.; Zhong, C.; Liu, D.; Huang, H.; Zhang, W.; Maurin, G. Revealing the structure-property relationships of metal-organic frameworks for CO2 capture from flue gas. *Langmuir* **2012**, *28*, 12094−12099.

(31) Duerinck, T.; Couck, S.; Vermoortele, F.; De Vos, D. E.; Baron, G. V; Denayer, J. F. M. Pulse gas chromatographic study of adsorption of substituted aromatics and heterocyclic molecules on MIL-47 at zero coverage. *Langmuir* **2012**, *28*, 13883−13891.
(32) Amrouche, H.; Creton, B.; Siperstein, F.; Nieto-Draghi, C. Prediction of thermodynamic properties of adsorbed gases in zeolitic imidazolate frameworks. *RSC Adv.* **2012**, *2*, 6028−6035.
(33) Cunha, D.; Gaudin, C.; Colinet, I.; Horcajada, P.; Maurin, G.; Serre, C. Rationalization of the entrapping of bioactive molecules into a series of functionalized porous zirconium terephthalate MOFs. *J. Mater. Chem. B* **2013**, *1*, 1101−1108.
(34) Fernandez, M.; Woo, T. K.; Wilmer, C. E.; Snurr, R. Q. Large-scale Quantitative Structure-Property Relationship (QSPR) Analysis of Methane Storage in Metal-Organic Frameworks. *J. Phys. Chem. C.* **2013**, *117*, 7681−7689.
(35) Bae, Y.-S.; Snurr, R. Q. Development and evaluation of porous materials for carbon dioxide separation and capture. *Angew. Chem., Int. Ed. Engl.* **2011**, *50*, 11586−11596.
(36) Yazaydin, A. O.; Snurr, R. Q.; Park, T.-H.; Koh, K.; Liu, J.; Levan, M. D.; Benin, A. I.; Jakubczak, P.; Lanuza, M.; Galloway, D. B.; Low, J. J.; Willis, R. R. Screening of metal-organic frameworks for carbon dioxide capture from flue gas using a combined experimental and modeling approach. *J. Am. Chem. Soc.* **2009**, *131*, 18198−181199.
(37) Banerjee, R.; Furukawa, H.; Britt, D.; Knobler, C.; O'Keeffe, M.; Yaghi, O. M. Control of pore size and functionality in isoreticular zeolitic imidazolate frameworks and their carbon dioxide selective capture properties. *J. Am. Chem. Soc.* **2009**, *131*, 3875−3877.
(38) Billinge, S. J. L. Nanoscale structural order from the atomic pair distribution function (PDF): There's plenty of room in the middle. *J. Solid State Chem.* **2008**, *181*, 1695−1700.
(39) Young, C. a.; Goodwin, A. L. Applications of pair distribution function methods to contemporary problems in materials chemistry. *J. Mater. Chem.* **2011**, *21*, 6464−6476.
(40) Hemmer, M. C.; Gasteiger, J. Prediction of three-dimensional molecular structures using information from infrared spectra. *Anal. Chim. Acta* **2000**, *420*, 145−154.
(41) Hemmer, M. C.; Steinhauer, V.; Gasteiger, J. Deriving the 3D structure of organic molecules from their infrared spectra. *Vib. Spectrosc.* **1999**, *19*, 151−164.
(42) González, M. P.; Caballero, J.; Tundidor-Camba, A.; Helguera, A. M.; Fernández, M. Modeling of farnesyltransferase inhibition by some thiol and non-thiol peptidomimetic inhibitors using genetic neural networks and RDF approaches. *Bioorg. Med. Chem.* **2006**, *14*, 200−213.
(43) Fernández, M.; Caballero, J.; Fernández, L.; Abreu, J. I.; Garriga, M. Protein radial distribution function (P-RDF) and Bayesian-Regularized Genetic Neural Networks for modeling protein conformational stability: chymotrypsin inhibitor 2 mutants. *J. Mol. Graph. Model.* **2007**, *26*, 748−759.
(44) Barthelet, K.; Marrot, J.; Riou, D.; Férey, G. A breathing hybrid organic-inorganic solid with very large pores and high magnetic characteristics. *Angew. Chem., Int. Ed. Engl.* **2002**, *41*, 281−284.
(45) Li, H.; Eddaoudi, M.; O'Keeffe, M.; Yaghi, O. M. Design and synthesis of an exceptionally stable and highly porous metal-organic framework. *Nature* **1999**, *402*, 276−279.
(46) Wilmer, C. E.; Farha, O. K.; Bae, Y.-S.; Hupp, J. T.; Snurr, R. Q. Structure−property relationships of porous materials for carbon dioxide separation and capture. *Energy Environ. Sci.* **2012**, *5*, 9849−9856.
(47) Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273−297.