A Cheminformatics Approach for Zeolite Framework Determination

Shujiang Yang[1], Mohammed Lach-hab[1], Iosif I. Vaisman[1,2], and Estela Blaisten-Barojas[1,3]

[1] Computational Materials Science Center, George Mason University, MSN 6A2, Fairfax, Virginia 22030, USA
[2] Department of Computational Biology and Bioinformatics, George Mason University, MSN 5B3, Manassas, Virginia 20110, USA
[3] Department of Computational and Data Sciences, George Mason University, MSN 6A2, Fairfax, Virginia 22030, USA
blaisten@gmu.edu

Abstract. Knowledge of the framework topology of zeolites is essential for multiple applications. Framework type determination relying on the combined information of coordination sequences and vertex symbols is appropriate for crystals with no defects. In this work we present an alternative machine learning model to classify zeolite crystals according to their framework types. The model is based on an eighteen-dimensional feature vector generated from the crystallographic data of zeolite crystals that contains topological, physical-chemical and statistical descriptors. Trained with sufficient known data, this model predicts the framework types of unknown zeolite crystals within 1-2 % error and shows to be better suited when dealing with real zeolite crystals, all of which always have geometrical defects even when the structure is resolved by crystallography.

1 Introduction

Zeolites are crystalline materials with regular structures consisting of molecular-sized pores and channels. These crystals are widely used in the field of adsorption, ion-exchange, heterogeneous catalysis (basically all gasoline production employs zeolite catalysts), as well as in health applications, sensors, solar energy conversion [1]. There are hundreds of zeolite species occurring naturally and/or synthetically, and millions more have been hypothetically proposed [2]. Zeolite crystals are constructed from an underlying three-dimensional network of TO_4 building block units. Within this network there are loosely bonded exchangeable cations, adsorbent phases and the building block central atom is tetrahedrally coordinated with four oxygen atoms. Predominantly, Si, Al or P is the element in the center of the tetrahedral building blocks and is referred as the T-atom. Zeolite networks are constructed by spatially accommodating the TO_4 building blocks by corner sharing the oxygens located at their vertices. These networks span a certain length and then repeat periodically along the crystal. Thus, these underlying networks depend directly on the connectivity of TO_4 units. Other zeolite crystal components such as cations, adsorbent phase,

G. Allen et al. (Eds.): ICCS 2009, Part II, LNCS 5545, pp. 160–168, 2009.

chemical composition, and observed crystallography properties are irrelevant in the determination of the underlying network. There are topological differences between networks in different crystals. Once a crystal network possesses a recognized topology, the Structure Commission of the International Zeolite Association (IZA-SC) approves it as an established framework type [3] and crystals displaying one of the approved framework types are then cataloged as zeolites. Non-approved network topologies fall into the category of *hypothetical* zeolites or *are not* zeolites. The IZA-SC currently recognizes 186 unique framework topologies [3]. The IUPAC Commission on Zeolite Nomenclature assigns a three-capital-letter acronym, the framework type codes (FTC), to each framework topology [4]. Known zeolite crystals belong to one of these topological categories. Crystals suspected to be zeolites that do not meet the established framework types cannot be cataloged as zeolites.

The FTC is conventionally determined with the combined information of the coordination sequences (CS) [5] and vertex symbols (VS) [6] of a zeolite crystal. Although it is not excluded that different framework types would have identical coordination sequences and vertex symbols, these cases are infrequent [7]. The conventional CS-VS method has limitations when applied to real zeolite crystals. Indeed, we noticed that multiple zeolite crystals in the Inorganic Crystal Structure Database (ICSD) [8] are distorted in various ways or might not contain complete crystal information. Coordination sequences or vertex symbols calculated for these crystals are erroneous and as a consequence their FTC cannot be predicted.

Exploitation of data mining and machine learning approaches is emerging in recent years in the field of chemical and materials informatics as a powerful approach for designing models that are developed based on data archived in databases [9,10]. The challenging task is to turn such models uniquely based on data analysis into novel applications. Similar approaches have been successful in a diversity of fields ranging from speech and vision recognition, robot control, business management, to bioinformatics and drug design.

In this work, we introduce a machine learning methodology for classifying zeolite crystals according to their framework type. The model presented here is the second-generation Zeolite-Structure-Predictor (ZSP2) developed on a data set of around 1300 zeolite crystals contained in the ICSD. ZSP2 is an extension of the original ZSP model [10] in which different topological descriptors are considered. The methodology used for building this model can be easily ported for the structural analysis of other families of crystals.

2 Methodology

The ZSP2 uses Breiman's Random Forest (RF) algorithm [11], which consists of an ensemble of decision trees trained on a bootstrap sample of the training data. The algorithm considers random groups of attributes for creating many trees rather than using all attributes to build one tree. Classification predictions are made by majority vote of all the trees. In this work the forest contains 100 trees and the WEKA [12] implementation of RF is used throughout.

2.1 Data Preprocessing

The process of cleaning the data is of paramount importance in data mining. When queried for zeolite crystals, the ICSD gives about 1600 crystal entries. Data in these entries are collected from published literature and do not include the framework type information. Based on the structure content in each crystal entry, we were able to assign the CS and VS to 1473 crystal entries by means of the zeoTsites package [13]. The remaining crystal entries have spurious geometry disorder or insufficient information in the database and no CS-VS could be determined precisely [14]. The CS and VS of these 1473 crystals were compared with the IZA-SC table, confirming that 1370 crystals can be referred as zeolites belonging to 94 framework types. The conventional CS-VS method proves incapable for identifying a framework type of the remaining 103 entries. Therefore the CS-VS method fails in 7.5 % of the cases to assign a framework type to a suspected zeolite crystal.

In machine learning terminology a framework type is referred as a *class* and each zeolite is referred as an *instance*. Machine learning models are more robust when classes are populated by a large number of instances. The 1370 zeolite entries are unevenly distributed among the 94 framework types. Indeed, class population ranges from 1 to 351 instances. There are 53 classes populated with only one or two instances, which are clearly inadequate for developing a data-based model. Because of this limitation, our machine learning study focuses on the 41 classes populated with at least 3 instances. However, the model described in this work can be easily extended to classify according to 186 classes in the zeolite case, and can be used for other crystal families as well.

Although the ICSD is the largest and most comprehensive database of inorganic crystals, it presents constrains for building models based on data contained in its repository. There is hope that the ICSD will continue adding crystals to the existing portfolio, which would then allow for further informatics approaches based on the crystal data to become useful to the materials and solid state chemistry community.

2.2 Feature Generation

An *attribute* is a descriptor of a certain crystal property. A *feature* is the specification of an attribute. The ZSP2 model for classification of zeolites into framework types includes categorical and quantitative features of topological, chemical and statistical nature.

The topological descriptors in the ZSP2 are based on a statistical geometry approach based on the Delaunay tessellation [15] of a supercell of each zeolite crystal [10]. Delaunay tessellation provides an objective, non-arbitrary definition of nearest neighboring points in space. Depending on the motif of the points, such tessellation has been used to characterize liquids [16,17], proteins [18], as well as zeolites [19]. The ICSD crystal entries provide the asymmetric unit cell of the crystal resolved from X-ray experiments. With this information, it is possible to generate the unit cell of a given crystal and once the unit cell is known, a supercell containing several unit cells can be generated numerically [20]. In the zeolites analyzed, the unit cells span a wide range of sizes and contain between 20 and 3040 atoms (excluding the hydrogens).

With the purpose of proposing topological descriptors, large supercells of all 1370 zeolites are generated such that a spherical cut of fixed radius 35.32 Å could be carved out of each supercell. The sphere radius is chosen to ensure that the carved sphere in crystals with huge unit cells encompasses a central unit cell and at least one neighboring unit cell in each of the three directions. The next step is to remove from the supercell sphere all oxygen atoms, all cations, and the full adsorbent phase. Thus, only T-atoms are retained inside the sphere. These T-atoms constitute the backbone of the zeolite framework and the Delaunay tessellation is performed on points in space coinciding with their location. This procedure yields the Delaunay simplices (distorted tetrahedra), which contain T-atoms at their vertices. Tens of thousands of Delaunay simplices are obtained per zeolite spherical supercell. Most simplices are distorted tetrahedra with edges that can be very long. In contrast, the TO_4 units that sustain the zeolite framework are near-to-perfect tetrahedra with edge lengths consistent with small variations around the oxygen-oxygen bond length.

In this work, the proposed topological descriptors are based on three geometrical properties of the Delaunay simplices: i) mean edge length ($\overline{d}$) of the six edges of each simplex; ii) in-sphere volume (*iV*) of the largest sphere inscribed in a simplex; iii) tetrahedrality (*T*) defined as the degree of distortion of a simplex from a regular tetrahedron:

$$T = \sum_{i=1}^{5} \sum_{j=i+1}^{6} \frac{(d_i - d_j)^2}{15\overline{d}^2} ,$$

where d_i is the *i*-th edge length of the simplex. Mean and standard deviation (σ) of these three properties are calculated for all simplices within each zeolite. The six topological descriptors are *mean_*$\overline{d}$, σ_$\overline{d}$, *mean_iV*, σ_*iV*, *mean_T*, and σ_*T*.

Additional geometrical descriptors were adopted by considering secondary simplices corresponding to a second coordination shell in Delaunay space [21]. Because each simplex has four adjacent tetrahedra that share one of its faces, the four new vertices can be linked into a larger tetrahedron defining the secondary simplex. Six geometrical descriptors: *mean_*$\overline{d}_2$, σ_$\overline{d}_2$, *mean_iV*$_2$, σ_*iV*$_2$, *mean_T*$_2$, and σ_*T*$_2$, based on secondary simplices were adopted.

Finally, six physical and chemical properties of a crystal are considered as descriptors: framework density (ρ), unit cell volume (V_o), space group (*SG*), and the chemical composition of T-atoms Si, Al and P (*[Si], [Al], [P]*). Among them, *SG* is the only nominal feature.

In summary, the zeolite classifier model ZSP2 is based on an 18-feature vector composed of twelve topological/statistical descriptors and six physical-chemical descriptors.

3 Results and Discussion

The performance of the ZSP2 model depends on the size of the feature vector used to create it. The performance is measured in terms of *accuracy*, which is defined as the percentage of instances that the model classifies correctly. Traditional classifiers in data mining contain very few classes, and typically the classification is reduced to two classes, which can then be addressed in binary language. Considering the dataset of

1370 instances and 41 classes, classification accuracy increases progressively as additional relevant features increase the dimensionality of the feature vector. This effect is shown in Table 1, where the reported accuracy is calculated with stratified 10-fold cross validation and averaged over ten trials. Classification accuracy is 90.6% with a feature set containing the six topological descriptors based on the first Delaunay shell (*mean_*$\bar{d}$, σ_$\bar{d}$, *mean_iV*, σ_*iV*, *mean_T*, and σ_*T*). By gradually adding features from secondary simplices, T-atom composition, ρ, *Vo*, and *SG*, the classification performance is progressively improved. Finally, with the 18 features included in the model an impressive accuracy of 97.9% is reached. Consistently, the out of bag error (OOB) decreases as classification accuracy increases.

The worth of features for this classifier was investigated by evaluating their information gain with respect to the classes considered. This analysis determines that *mean* _$\bar{d}$, *mean_iV*, *mean_T* , ρ, V_o, *SG, [Si], [Al]* are the nine most significant features. However, the ZSP2 built with the18-feature set is computationally very fast and thus all 18 features are kept throughout this study.

To analyze the effect of the population size of each class on the ZSP2, seven different models were built each of them classifies into a similar number of classes, but these classes are populated with different number of instances per class (*x*). Figure 1 illustrates the performance of the seven classifications. The plotted accuracy is a result of using stratified 10-fold cross validation and averaging over ten trials. Although the classification process is dependent on the motif of classes involved in each data group, it is evident that the ZSP2 model is more accurate when trained with well populated classes. In fact, the ZSP2 yields 100% accuracy when built with six classes that have more than 63 instances.

Table 1. ZSP2 classification with various feature sets, 1370 instances and 41 classes

Feature set	*mean_*$\bar{d}$, σ_$\bar{d}$, *mean_iV*, σ_*iV*, *mean_T*, σ_*T*	*mean_*$\bar{d}$, σ_$\bar{d}$, *mean_iV*, σ_*iV*, *mean_T*, σ_*T*, *mean_*$\bar{d}$ *2*,σ_$\bar{d}$ *2*, *mean_iV2*,σ_*iV2*, *mean_T2*, σ_*T2*	*mean_*$\bar{d}$, σ_$\bar{d}$, *mean_iV*, σ_*iV*, *mean_T*, σ_*T*, *mean_*$\bar{d}$ *2*,σ_$\bar{d}$ *2*, *mean_iV2*,σ_*iV2*, *mean_T2*, σ_*T2* *[Si], [Al], [P]*	*mean_*$\bar{d}$, σ_$\bar{d}$, *mean_iV*, σ_*iV*, *mean_T*, σ_*T*, *mean_*$\bar{d}$ *2*,σ_$\bar{d}$ *2*, *mean_iV2*,σ_*iV2*, *mean_T2*, σ_*T2* *[Si], [Al], [P],* ρ, V_o, *SG*
OOB	0.024±0.001	0.099±0.004	0.054±0.003	0.024±0.001
Accuracy (%)	90.6±0.2	92.8±0.2	94.6±0.3	97.9±0.1

The model improvement through experience is demonstrated through its learning curve. For this experiment, a balanced dataset is constructed such that each class is populated equally. Figure 2 shows the learning curve of the ZSP2 when the model classifies into six classes, each of them populated with 60 instances. In this figure the accuracy (vertical axis) pertains to instances correctly classified from split of the total number of instances (*test*-set) once the model was *trained* with the remaining split of

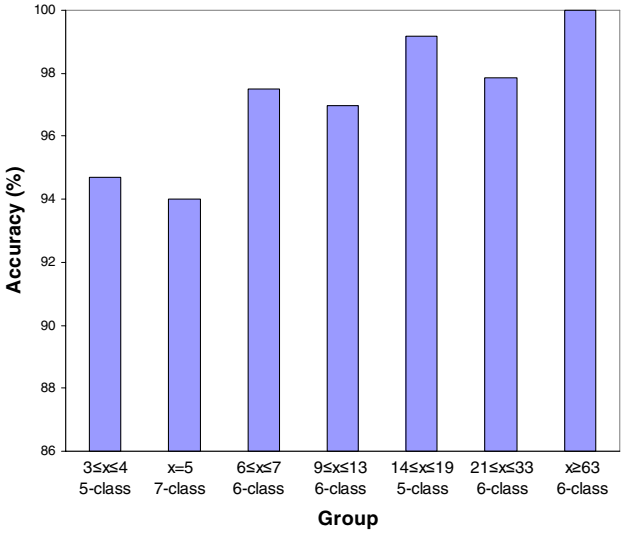

Fig. 1. Classification performance of the ZSP2 obtained for seven data groupings with about constant number of classes. "*x*" is number of instances per class.

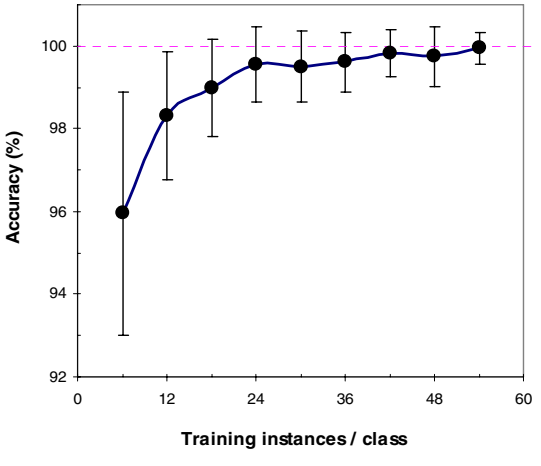

Fig. 2. The learning curve of the ZSP2 built for a balanced dataset of 360 instances and six classes

available instances (plotted on the abscissa). For each training/test split, the training instances are drawn at random from the available data. Next, the test instances are randomly drawn from the remaining instances. The split is repeated 100 times for each point. Both mean and standard deviation of this process are shown in Figure 2. It is clearly shown that the ZSP2 improves fast when the training set is small, then more smoothly when the training set exceeds 24 instances/class, and reaches a plateau at

about 42 instances/class. Finally, the ZSP2 model performs perfectly when trained with 54 instances per class. Therefore, the ZSP2 yields perfect classification into six classes for datasets containing 60 instances per class.

Among the 1370 available instances, 1041 of them are distributed in 11 classes with more than 27 instances/class. The ZSP2 model for this 11-class dataset, using stratified 10-fold cross validation and repeated ten times, classifies 99.3% of the instances correctly, which is not perfect but excellent.

The ZSP2 could be tested with larger datasets and more than 41 classes. However, we have been limited to the content of the ICSD, which currently allows for as much as 41-class classification as shown in Table 1. To predict the classification of instances falling within a class not included in the 41 trained classes, a *bag* class was defined containing the 53 frameworks poorly populated in the ICSD. Now ZSP2 is built with 1370 instances to classify into 42 classes including the bag-class. With ten times 10-fold cross validation, the ZSP2 correctly classifies 95.3% of all instances. If the number of classes is reduced by keeping in the ZSP2 only those classes populated with x or more instances, and placing the rest into the bag-class, the predictive power of the model improves as the number of classes decreases as shown in Table 2.

Table 2. The ZSP2 classification of four datasets including a bag-class

Dataset	Dataset1	Dataset2	Dataset3	Dataset4
Number of instances per class=x	x≥3	x≥9	x≥19	x≥63
Size of the bag class	65	157	251	477
OOB	0.052±0.003	0.037±0.003	0.027±0.002	0.009±0.001
Accuracy (%)	95.3±0.1	96.5±0.1	97.3±0.1	99.0±0.2

During the data cleaning procedure 103 instances with determined CS-VS were removed from the analysis because they were not consistent with any framework type. The ZSP2 for a 41-class model predicts that 60 of these instances belong to these 41 classes. This finding was further corroborated by details given in the published literature that originated the entries in the ICSD of these 60 crystals.

Compared with the conventional method to assign zeolite framework types, the ZSP2 machine learning model is more robust to geometry disorder and occasional errors in the data. By examining the zeolite entries in the ICSD, it was noticed that measured structural crystal data may differ substantially from perfect crystals. As a consequence, erroneous FTC would be determined for these cases. On the other hand, these types of errors are more likely to be tolerated by the ZSP2 model.

4 Conclusion

In this work we present the ZSP2, a machine learning model for classifying zeolites crystals according to their framework type. The approach requires as input the resolved crystallographic data of each crystal only for T-atoms in the framework. The ZSP2 is then a more efficient model than the ZSP where the crystallographic resolution had to contain all atoms. The complete crystallographic information is also required to assign a framework type to a zeolite crystal using the conventional coordination sequence and vertex symbol approach. The ZSP2 performance is highly accurate. Indeed, the model is able to predict correct classification with up to 100% accuracy when enough data are available. The novel approach is considerably more robust than the conventional identification method and can potentially be used to study other families of crystals. There are over 100,000 crystal entries in the ICSD and the ZSP2 model can be tailored for clustering and classifying with a variety of different objectives. Work is in progress in this direction.

Acknowledgments. This work was supported by the National Science Foundation grant CHE-0626111. Authors acknowledge the NIST Standard Reference Data Program for making available the ICSD zeolite data set.

References

1. Payra, P., Dutta, P.K.: Zeolites: a Primer. In: Auerbach, S.M., Carrado, K.A., Dutta, P.K. (eds.) Handbook of Zeolite Science and Technology, pp. 1–19. Marcell Dekker, New York (2003)
2. Foster, M.D., Treacy, M.M.J.: A Database of Hypothetical Zeolite Structures, http://www.hypotheticalzeolites.net
3. IZA-SC and its Standard Database of Zeolite Frameworks, http://www.iza-structure.org/databases/
4. Barrer, R.M.: Chemical Nomenclature and Formulation of Compositions of Synthetic and Natural Zeolites. Pure Appl. Chem. 51, 1091–1100 (1979)
5. Meier, W.M., Moeck, H.J.: The Topology of Three-dimensional 4-Connected Nets: Classification of Zeolite Framework Types Using Coordination Sequences. J. Solid State Chem. 27, 349–355 (1979)
6. O'Keeffe, M., Hyde, S.T.: Vertex Symbols for Zeolite Nets. Zeolites 19, 370–374 (1997)
7. Treacy, M.M.J., Foster, M.D., Randall, K.H.: An Efficient Method for Determining Zeolite Vertex Symbols. Micropor. Mesopor. Mater. 87, 255–260 (2006)
8. Inorganic Crystal Structure Database (ICSD), http://www.nist.gov/srd/nist84.htm
9. Fischer, C.C., Tibbetts, K.J., Morgan, D., Ceder, G.: Predicting Crystal Structure by Merging Data Mining with Quantum Mechanics. Nature Mater. 5, 641–646 (2006)
10. Carr, D.A., Lach-hab, M., Yang, S., Vaisman, I.I., Blaisten-Barojas, E.: Machine Learning Approach for Structure-based Zeolite Classification. Micorpor. Mesopor. Mater. 117, 339–349 (2009)
11. Breiman, L.: Random Forests. Machine Learning 45, 5–32 (2001)
12. Weka 3: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/

13. Sastre, G., Gale, J.D.: Zeotsites: A Code for Topological and Crystallographic Tetrahedral Sites Analysis in Zeolites and Zeotypes. Micorpor. Mesopor. Mater. 43, 27–40 (2001)
14. Yang, S., Blaisten-Barojas, E.: Internal communication to the ICSD
15. Delaunay, B.N.: Sur La Sphere Vide. Bull. Acad. Sci. USSR (in Russian) 7, 793–800 (1934)
16. Medvedev, N.N., Naberukhin, Y.I.: Analysis of Structure of Simple Liquids and Amorphous Solids by Method of Statistical Geometry. Zh. Strukt. Khimii 28, 117–132 (1987)
17. Vaisman, I.I., Brown, F.K., Tropsha, A.: Distance Dependence of Water Structure around Model Solutes. J. Phys. Chem. 98, 5559–5564 (1994)
18. Vaisman, I.I.: Statistical and Computational Geometry of Biomolecular Structure. In: Gentle, J.E., Härdle, W., Mori, Y. (eds.) Handbook of Computational Statistics, pp. 981–1000. Springer, New York (2004)
19. Foster, M.D., Rivin, I., Treacy, M.M.J., Friedrichs, O.D.: A Geometric Solution to the Largest-free-sphere Problem in Zeolite Frameworks. Micropor. Mesopor. Mater. 90, 32–38 (2006)
20. Computational Crystallography Toolbox, `http://cctbx.sourceforge.net/`
21. Wako, H., Yamato, T.: Novel Method to Detect a Motif of Local Structures in Different Protein Conformations. Protein Engineering 11, 981–990 (1998)