Machine Learning Approach for Classification of Zeolite Crystals

Shujiang Yang, Mohammed Lach-hab, Iosif I. Vaisman, and Estela Blaisten-Barojas

***Abstract*—A machine learning approach is applied to classify zeolite crystals according to their framework type. The Zeolite-Structure-Predictor is introduced based on the Random Forest algorithm. Zeolites structural data from the Inorganic Crystal Structure Database (ICSD) are used to train the model. The ZSP uses sixteen attributes including topological descriptors obtained with statistical geometry and physical and chemical properties of individual zeolites. Trained with 40 framework types containing at least 5 instances per class, the ZSP can correctly classify zeolites with over 95% accuracy. The performance is shown to improve when more zeolite instances per class are available.**

Keywords: Zeolite framework, Zeolite classification, Data mining, Delaunay tessellation, Random forest

I. Introduction

Zeolites are microporous crystalline materials with a 3-dimensional underlying framework structure. Each framework type gives a signature of the channels and cages contained in the structure. In general, zeolite crystals contain uniformly distributed pores of molecular sizes in the range of 3 to 20 Å. Such characteristics are essential for the major applications of zeolites in adsorption, ion-exchange, heterogeneous catalysis, and other fields such as health, sensors, solar energy conversion just to mention a few [1].

Traditionally, zeolites are aluminosilicate crystals, each of them belonging to one of the unique framework types containing mostly Si, Al and O atoms. The cavities may be filled with large ions and water molecules, both of which have considerable freedom of movement permitting ion-exchange and reversible dehydration [2]. Each Si or Al atom in the framework is conventionally labeled as a T-atom because of its tetrahedral chemical bonding to four O atoms. Despite this comprehensive definition, the term "zeolite" has been over-used in the literature for describing other microporous and/or mesoporous substances.

Classification of zeolitic materials by framework type is widely accepted in the zeolite community. Framework types identify different connectivity of T-atoms in a topological network. Framework type codes (FTC) consist of three capital Roman letters are used for distinguishing different framework types [3] and, thus for identifying different zeolites. With the combined information of the coordination sequence [4] and the vertex symbol [5] for a particular framework topology, different framework types can be distinguished [6]. Albeit rare, situations exist where two frameworks with the same coordination sequences and vertex symbols belong to different topologies [7].

In this work, we explore a machine learning approach for determining framework types based on the structural information of zeolites embedded in the data collected in the Inorganic Crystal Structure Database (ICSD). This novel Zeolite-Structure-Predictor (ZSP) model shows that a machine learning methodology based on data analysis can reliably predict framework types of zeolite crystals. This model becomes an alternative classifier to the framework type classification.

II. Database

Structural data used in this study is contained in the ICSD. This database has a collection of more than 100,000 entries of inorganic crystal structures, all collected from crystallographic information contained in published journals. In this work we analyzed the 1436 zeolite crystals contained in the ICSD, which were graciously made available to us by the National Institute of Standards and Technology. Among these data entries there are 822 different chemical names for the different zeolite crystals and 289 different mineral names were assigned to them. FTCs are not contained in the ICSD; therefore we identified the corresponding framework type [8] for each data entry. We find that 96 FTCs are represented in the zeolite data set. It is to be noted that 179 framework types are currently approved [8]. Therefore the data set analyzed represents approximately half of the different topologies that exist in the realm of zeolites.

The distribution of the 96 observed framework types in the ICSD is illustrated in Fig. 1. There are forty framework types occurring at least five times in the database. Twenty-one of them occurred more than ten times, and fourteen of them occurred more than twenty times.

This work is supported under the National Science Foundation grant CHE-0626111.

S. Yang is with the Computational Materials Science Center, George Mason University, MSN 6A2, Fairfax, Virginia 22030 USA (phone: 703-993-3614; fax: 703-993-9300; e-mail: syangf@gmu.edu).

M. Lach-hab is with the Computational Materials Science Center, George Mason University, MSN 6A2, Fairfax, Virginia 22030 USA (phone: 703-993-9325; fax: 703-993-9300; e-mail: mlachhab@gmu.edu).

I. Vaisman is with the Computational Materials Science Center and the Department of Computational Biology and Bioinformatics, George Mason University, MSN 5B3, Manassas, Virginia 20110 USA (phone: 703-993-8431; fax: 703-993-8401; e-mail: ivaisman @gmu.edu).

E. Blaisten-Barojas is with the Computational Materials Science Center and the Department of Computational and Data Sciences, George Mason University, MSN 6A2, Fairfax, Virginia 22030 USA (phone: 703-993-1988; fax: 703-993-9300; e-mail: blaisten@gmu.edu).

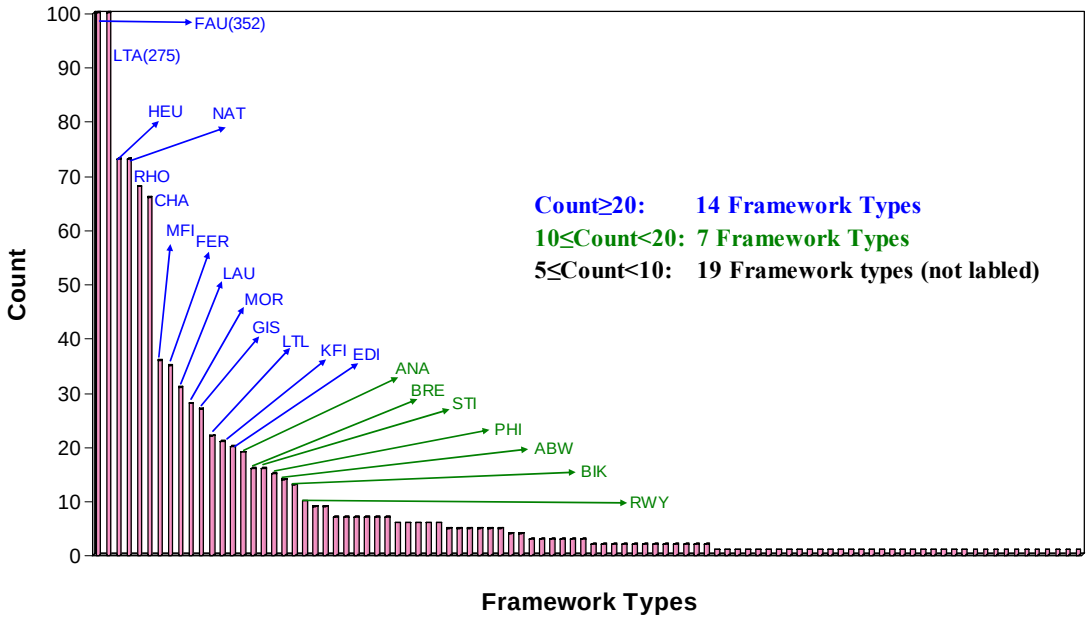

Fig. 1 Distribution of the 96 determined framework types of zeolites occurring in the ICSD. Count refers to the number of instances within a framework type.

III. Descriptor Generation

Topological descriptors are generated through a statistical geometry approach using Delaunay tessellation. Delaunay tessellation provides a non-arbitrary identification of neighboring points in molecular systems represented by extended sets of points in space [9]. It has been applied to simple liquids [10], [11], water and aqueous solutions [12], [13], as well as protein structures [14], [15]. Recently, it has been used to determine the largest-free-sphere volume of zeolites [16].

First, the unit cell of each zeolite crystal is generated based on the asymmetric unit and space group information contained in the ICSD. The Computational Crystallography Toolbox (cctbx) [17] is used for this purpose. Secondly, the unit cell is replicated periodically in 3-dimentions to yield a supercell of each zeolite. A sphere of material is carved out of each zeolite supercell. This method ensures that all zeolite cuts have the same volume: a sphere with fixed radius of 35.32 Å. Thirdly, framework T-atoms in the sphere are kept as active points for Delaunay tessellation. This filters out from our analysis all other atoms, including cations and adsorbent phase. Lastly, each sphere containing only T- atoms is tessellated using the Qhull algorithm [18]. The Delaunay tessellation on a representative zeolite crystal is visualized in Fig. 2.

Delaunay tessellation generates a series of 4-point simplices such that the sphere touching them does not enclose any other points. Then, in 3D a Delaunay simplex is a deformed tetrahedron with its four vertices coinciding with four of the available points. In our case, each simplex determines direct contacts between T-atoms in a given zeolite. In this study, any framework atom is treated equally as a T-atom, although in reality these locations may be occupied by several elements such as Si, Al, P, etc. Therefore each simplex is composed of four T-atoms.

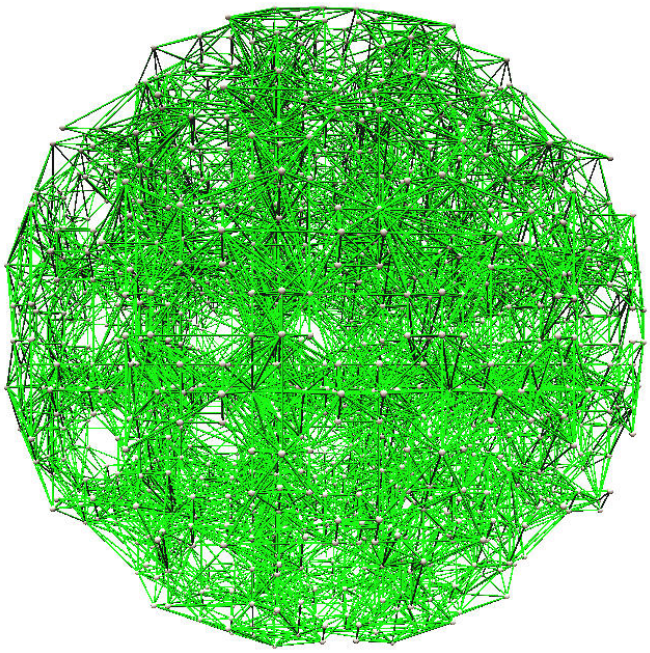

Fig.2 Delaunay tessellation of points inside a spherical cut of a zeolite where only framework T-atoms were kept.

Simplices are characterized by three attributes. The first is tetrahedrality (*tetra*), which is a quantitative measure of the degree of distortion of a Delaunay simplex from the perfect tetrahedron:

$$Tetra = \sum_{i=1}^{6} \sum_{j>i}^{6} \frac{(d_i - d_j)^2}{15\bar{d}^2} \quad ,$$

where d_i is one of the six simplex edges. *Tetra* is a distortion index, which is zero for the perfect tetrahedron. The second attribute is the simplex in-sphere volume (*in-V*), which is the volume of the largest sphere that can be inscribed in the tetrahedron. The third attribute is the mean edge length (*d*) of the simplex. All three attributes are averages over tens of thousands of simplices obtained for each zeolite structure after Delaunay tessellation.

In addition to the three primary topological descriptors discussed above, three secondary topological descriptors are added with the introduction of secondary simplices. A secondary simplex is constructed by joining the four furthest vertices of the four simplices surrounding a central primary simplex [19]. Averages over simplices are taken to obtain three additional attributes (*tetra2, in-V2, d2*) for each zeolite.

To keep statistical tracking of these properties, we included six additional attributes associated with the standard deviation of the six above described topological attributes.

Finally, the chemical composition of each zeolite crystal is taken into account by using as attributes the percentages of different elements contained in the spherical cut. Since Si, Al and P are the overwhelmingly populated T-atoms in the data set of 1438 zeolites, the compositional percentage of Si, Al and P (*comp-Si, comp-Al, comp-P*) are used as attributes.

Framework density (*fd*) is defined as the number of T-atoms per 1000 $Å^3$. This is an important property of a zeolite, which sets a criterion for distinguishing zeolites and zeolite-like materials from denser tectosilicates [6]. This property is the 16th attribute considered in this work.

All together a vector of dimension 16 is then used for mining the date as described in the next section.

IV Machine Learning Analysis

Breiman's Random Forest (RF) algorithm [20] is applied as the classifier of zeolites. RF is an ensemble of unpruned trees, each trained on a bootstrap sample of the training data. Classification predictions are made by majority vote of the trees. RF differs from Bagging in that at each node of each tree, the algorithm considers as splitting candidates a random sample of the variables rather than all variables. WEKA [21] is used as the tool for carrying out the classification analysis and 100 trees are used in the forest.

Our model of zeolites is the Zeolite-Structure-Predictor (ZSP) defined on the 16-dimension space, as described in the previous section. For classification, each class is a different framework type. We have available 96 possible classes in the data set. However, the available data set is inhomogeneous, as shown in Fig. 1, not all classes have enough number of zeolites for this study. It is then useful to analyze the dependence of the model on sample size.

Zeolite classification considered in this work pertain to three subsets of classes: i) 14 classes that are well populated all above 20 instances per class; ii) 21 classes that have population of at least 10; and iii) 40 classes that are populated with at least 5 zeolites.

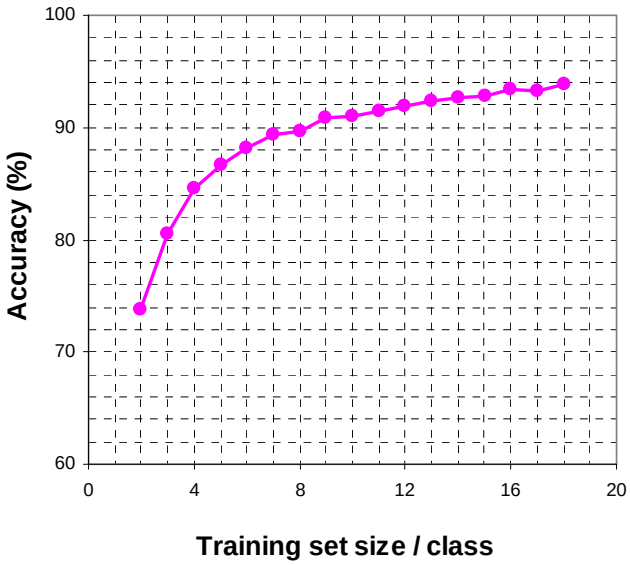

Fig. 3 The learning curve for the 14 most abundant framework type classes. The total size for each training/test split is 20/class. Each point is an average of 100 repeated runs.

Size of samples is an important parameter. To study the effect of sampling size on the ZSP a learning curve is constructed with the model containing 14 classes. This case corresponds to framework types most abundantly populated. For the learning process, a homogeneous set of data is considered such that each class contains 20 instances. Therefore, the total number of zeolites used for this experiment is 280. Next, 100 different sets of 280 instances are generated at random. In the starting training/test set, 2 instances per framework class are randomly selected as the training set, and remaining 18 instances are selected as the test set. Then, the training set size is gradually increased, with the corresponding test set gradually decreased to keep the total instances per class at 20. At the end of the process we have 18 instances per class for training, and 2 instances per class for testing. This process is repeated 100 times for each training/test split. An average is taken for each training/test split, as shown in Fig. 3. In this figure, accuracy refers to the percentage of correctly classified zeolites.

As seen in Fig. 3, when the ZSP model is built with a small number of instances, the prediction is only fair. However, the model gains accuracy very rapidly as the number of instances is larger when building the model. However, as training size grows, the predicting accuracy levels off. This effect ensures that at-or-about 8/12 training/test split the ZSP will already perform with more than 90% accuracy. Evidently more data would improve the accuracy of the ZSP model gradually. With the available data, the model can already correctly predict 94% instances at the end of the learning process.

Additionally the ZSP was tested with as much of the available zeolite data as possible. Three inhomogeneous implementations of the model were considered. Sample-1 is a set of 1127 zeolites belonging to 14 classes. Sample-2 is a set of 1230 zeolites belonging to 21 classes, and sample-3 is a set of 1350 zeolites with 40 classes represented in the set. Stratified 10-fold cross-validation is repeated 10 times for each sampling. The results are listed in Table I.

TABLE I
PERFORMANCE OF THE ZSP ON THE CLASSIFICATION OF ZEOLITE FRAMEWORK TYPES. "X" IS THE NUMBER OF INSTANCES PER CLASS. ACCURACY VALUES AND OUT OF BAG ERRORS (OOB) REFER TO THE AVERAGE OF 10 REPEATED RUNS OF STRATIFIED 10-FOLD CROSS-VALIDATION, WITH STANDARD DEVIATION SHOWN AT THE END.

Sample	Sample-1	Sample-2	Sample-3
total instances	1127	1230	1350
instances/class	$x \geq 20$	$x \geq 10$	$x \geq 5$
number of FTC	14	21	40
OOB	0.025±0.003	0.034±0.002	0.050±0.002
accuracy (%)	97.5±1.5	96.6±1.4	95.5±1.1

Therefore, when trained with at least 20 instances per class, the ZSP can correctly predict the classification of an unknown zeolite with 97.5% accuracy, as long as the unknown crystal is within one of the 14 training classes. When the ZSP is trained with t40 framework type classes consisted of at least 5 instances per class, the ZSP performs with 95.5% accuracy. The out-of-bag error (OOB) decreases as the number of instances per class is larger. As expected, the OOB improves if more zeolite data are available.

Further analysis of the above result on Sample-3 is shown in Table II, which summarizes average of the area under the curve (AUC) and true-positive (TP) rate. Sample-3 with 40 framework types can be divided into three groups based on the population of instances per class (x), $x \geq 20$, $10 \leq x < 20$, and $5 \leq x < 10$. It is clear that the average TP rate for the three sized groupings increases significantly as the size increases. The AUC score is not sensitive to the sample size.

TABLE II
THE BEHAVIORS OF THREE DIFFERENT SIZED CLASS GROUPS IN THE CLASSIFICATION OF THE SAMPLE WITH 40 FRAMEWORK TYPES. "X" IS THE NUMBER OF INSTANCES PER CLASS. AUC REFERS TO THE AREA UNDER THE CURVE SCORE. TP RATE REFERS TO THE AVERAGE TRUE POSITIVE RATES FOR EACH GROUP. AUC AND TP RATE VALUES ARE ALSO AVERAGED OVER 10 REPEATED RUNS OF STRATIFIED 10-FOLD CROSS-VALIDATION, WITH STANDARD DEVIATION SHOWN AT THE END.

Sample	40 frameworks with x≥5		
Group	**Group-1**	**Group-2**	**Group-3**
total instances / no. of classes	1127 / 14	103 / 7	120 / 19
instances/class	$x \geq 20$	$10 \leq x < 10$	$5 \leq x < 10$
AUC	0.997±0.002	0.993±0.007	0.976±0.009
TP rate (%)	95.3±0.9	87.8±2.2	79.7±3.3

V CONCLUSION

The Zeolite-Structure Predictor (ZSP) is developed in this work for classifying zeolite structures based on the framework types. This machine learning model is novel for predicting zeolite classification with high reliability. As new zeolites are discovered, the ZSP becomes a useful tool for their classification. The methodology used for developing the ZSP can potentially be applied to other types of materials. Research toward this goal is in progress.

ACKNOWLEDGMENTS

The authors would like to thank the National Institute of Standards and Technology for making available the ICSD data for this study. Figure 2 is a courtesy of Dr. Yanlin Luo using the Glisten software developed at George Mason University by Prof. D. Carr.

REFERENCES

[1] P. Payra, and P. K. Dutta, "Zeolites: a primer," in *Handbook of Zeolite Science and Technology*, S. M. Auerbach, K. A. Carrado, and P. K. Dutta, Eds. New York: Marcell Dekker, 2003, pp. 1-19.

[2] J. V. Smith, "Topochemistry of zeolites and related materials. 1. topology and geometry," *Chem. Rev.*, vol. 88, pp. 149-182, Jan. 1988.

[3] R. M. Barrer, "Chemical nomenclature and formulation of compositions of synthetic and natural zeolites," *Pure Appl. Chem.*, vol. 51, pp. 1091-1100, May 1979.

[4] W. M,. Meier, and H. J. Moeck, "The topology of three-dimensional 4-connected nets: Classification of zeolite framework types using coordination sequences," *J. Solid State Chem.*, vol. 27, pp. 349-355, Mar. 1979.

[5] M. O'Keeffe, and S. T. Hyde, "Vertex symbols for zeolite nets," *Zeolites*, vol. 19, pp. 370-374, 1997.

[6] Ch. Baerlocher, W. M. Meier, and D. H. Olson, *Atlas of Zeolite Framework Types*, 5th ed. Amsterdam: Elsevier, 2001.

[7] M. M. J. Treacy, M. D. Foster, and K. H. Randall, "An efficient method for determining zeolite vertex symbols", *Microporous and Mesoporous Materials*, vol. 87, pp. 255-260, 2006.

[8] Database of Zeolite Structures. Available: http://www.iza-structure.org/databases/, 2008.

[9] B. N. Delaunay, *Izv. Akad. Nauk. SSSR, Otd Mat Est Nauk,* vol. 7, pp. 793-800, 1934

[10] V. P. Voloshin, Y. I. Naberukhin, and N. N. Medvedev, "Can various classes of atomic configurations (Delaunay simplices) be distinguished in random dense packings of spherical particles?" *Molecular Simulation*, vol. 4, pp. 209-227, 1989.

[11] N. N. Medvedev, Y. I. Naberukhin, "Analysis of structure of simple liquids and amophous solids by method of statistical geometry," *Zh. Strukt. Khimii*, vol. 28, pp. 117-132, 1987.

[12] I. I. Vaisman, M. L. Berkowitz, "Local structural order and molecular associations in water-DMSO mixtures. Molecular dynamics study," *J. Am. Chem. Soc.*, vol. 114, pp. 7889-7896, 1992.

[13] I. I. Vaisman, F. K. Brown, A. Tropsha, "Distance dependence of water structure around model solutes," *J. Phys. Chem.*, vol. 98, pp. 5559-5564, 1994.

[14] R. K. Singh, A. Tropsha, I. I. Vaisman, "Delaunay Tessellation of Proteins: Four Body Nearest Neighbor Propensities of Amino Acid Residues," *J. Comput. Biol.*, vol. 3, pp. 213-221, 1996.

[15] I. I. Vaisman, "Statistical and Computational Geometry of Biomolecular Structure," in *Handbook of Computational Statistics*. J. E. Gentle, W. Härdle, Y. Mori, Eds. New York: Springer, 2004, pp. 981-1000.

[16] M. D. Foster, I. Rivin, M. M. J. Treacy, O. D. Friedrichs, "A geometric solution to the largest-free-sphere problem in zeolite frameworks," *Microporous and Mesoporous Materials*, vol. 90, pp. 32-38, 2006.
[17] Computational Crystallography Toolbox. Available: http://cctbx.sourceforge.net/, 2007.
[18] Qhull. Available: http://www.qhull.org/, 2003.
[19] H. Wako, and T. Yamato, "Novel method to detect a motif of local structures in different protein conformations," *Protein Engineering*, vol. 11, pp.981-990, 1998.
[20] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5-32, 2001.
[21] Weka 3: Data Mining Software in Java. Available: http://www.cs.waikato.ac.nz/ml/weka/, Dec. 2007.