Proceedings of the 2009 International Conference on Artificial Intelligence, ICAI 2009, July 13-16, 2009, Las Vegas Nevada, USA (https://researchr.org/publication/icai-2009)

Machine Learning Identification of Zeolite Framework Types

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, Fairfax, VA, USA
[2]Department of Bioinformatics and Computational Biology, George Mason University, Manassas, VA, USA
[3]Department of Computational and Data Sciences, George Mason University, Manassas, VA, USA

Abstract - *The characteristic framework types of zeolite crystals are routinely determined by calculating coordination sequences and vertex symbols of the 3D crystal structures. This method has limitations and tends to fail when the synthesized crystals are not close to perfect and present some types of crystallographic disorder. A machine learning based Zeolite-Structure-Predictor (ZSP) model is developed to predict framework types for both near perfect and moderately disordered zeolite crystals. The ZSP uses various attributes, including topological descriptors based on a computational geometry approach and relevant physical, chemical properties of the crystals. Trained with 41 framework types, the ZSP can correctly classify zeolite crystals with over 98% accuracy. Additionally, it is shown that the ZSP model is able to predict the framework types for strongly disordered zeolite crystals with reliable success rate.*

Keywords: Data mining, Artificial intelligence, Zeolite, Framework type, Crystal structure.

1 Introduction

Zeolites are best recognized by their role as heterogeneous catalysts in the petrochemical industry [1], where nearly all gasoline is effectively produced. With their intriguing molecular-sized channels and pores, zeolites are also extensively used in adsorption, ion-exchange, and more recently in emerging fields such as health, chemical sensors, solar energy conversion, etc [2].

The primary building unit of a zeolite structure is TO_4, where T is a central atom tetrahedrally coordinated through four oxygen (O) atoms. An underlying framework structure is formed by connecting these TO_4 units together through O-corner sharing, and these rigid networks are the structural signature identifying various types of zeolites. Additionally, in a zeolite crystal there are cations and adsorbent-phase residing within the channels of the rigid T-O framework. For the majority of time since the discovery of these peculiar crystals more than two centuries ago, the term "zeolite" refers to open framework aluminosilicates [3] where T is either a silicon (Si) or an aluminum (Al) atom. With the advent of numerous alike microporous materials that are being synthesized everyday, zeolite science has flourished beyond its original scope. Noteworthy is the extension of the zeolite definition by the International Mineralogical Association [4] to include other elements as T atoms, such as phosphorous, boron, beryllium, gallium, titanium, etc.

Zeolites can be categorized by the topological patterns of their open frameworks [5]. Since 1978, each distinct framework type of zeolites is labeled by a framework type code (FTC) following rules set by the IUPAC Commission on Zeolite Nomenclature [6]. The assignment of FTC is supervised by the Structure Commission of the International Zeolite Association (IZA-SC). New framework types have been gradually recognized by IZA-SC and added to the zeolite family. By the year end of 2008, 186 framework types have been approved showing the intense recent research in the materials science aspects of zeolites (there were 176 recognized frameworks in 2007 and 174 in 2006 correspondingly) [7].

In practice, zeolite framework types are determined with the combined information of the coordination sequences (CS) [8] and vertex symbols (VS) [9] of a particular framework topology [7]. One disadvantage of the CS-VS method is that it is possible for different topologies to have the same CS and VS [10], although this situation is rare. The second disadvantage of the method is its instability. When implemented in different computer codes, the CS-VS method might not agree with each other on some zeolite structures, due to algorithms for connection recognition, cut-off setting, confusion from extra-framework species, etc. Care should be taken when dealing with uncommon zeolite crystals. The major disadvantage of the CS-VS method is its inability to deal with disordered materials. Most synthesized zeolite-like materials are more or less disordered. When the degree of disorder is large enough, the calculated CS and VS might not reflect the actual topology of the framework, and thus can not be used to effectively determine the framework type of the crystals.

2 Related work

Machine learning based methods have been successfully applied to various fields ranging from speech and vision recognition, robot control, business management, to bioinformatics and drug design. Reports have been scarce in the literature to connect machine learning with zeolites [11,12], where attempts were made to use combinatorial methods to facilitate zeolite synthesis and characterization in the experiments.

Delaunay tessellation [13] is a computational geometry approach, which provides an objective, non-arbitrary definition of nearest neighboring points in space. Depending on the motif that the points represent, Delaunay tessellation has been used to model liquids [14,15], proteins [16,17], as well as zeolites [18-20].

Indeed, very recently we combined machine learning and computational geometry to classify zeolites by their mineral names and framework types [19,20]. In this abstract, we present the latest results in this direction. The zeolite structure predictor (ZSP) model is further optimized with improved accuracy based on feature selection from wide range types of attributes.

3 Data preprocessing

Data contained in the Inorganic Crystal Structure Database (ICSD) are used in this study [21]. The ICSD has the most complete collection of zeolite crystallographic information that has been assembled from works published in the literature. There are about 1600 zeolite entries in this database. Along with the crystallographic information, each zeolite crystal entry has its assigned collection code number, chemical formula, mineral name, mineral origin, chemical name, bibliographic reference, and comments. Unfortunately the ICSD provides no information about framework types of zeolites. Based on the contained information, together with the original publication, we were able to assign 99 FTC to 1473 zeolite entries that are in agreement with the online database of zeolite structures of IZA [7]. The remaining ICSD entries do not correspond to an IZA approved zeolite framework, i.e., they are not zeolites by the IZA standard.

The undisputable 1473 zeolite entries were further analyzed by rendering the crystal structure, determining chemical bonding and T-O coordination, and by the comments from the ICSD and the original paper. A set of 103 entries were determined to have strong disorder in the structure. Since their framework types are known, we will use them as disordered test samples to compare the CS-VS method and the artificial intelligence based ZSP model in the identification of zeolite framework types. This leaves us with 1370 more or less regularly structured zeolites.

The above 1370 entries are unevenly distributed among 94 framework types, ranging from 1 entry to 351 entries per framework type. In the machine learning classification, each framework type is deemed as a class, and each entry is an instance. Machine learning studies require a respectful number of instances per class in order to yield meaningful results. Therefore, from the zeolite data representing 94 classes, 53 classes had to be eliminated because they contain only one or two instances.

Summarizing, our machine learning study focused on the more populated 41 classes, each containing at least 3 instances; this reduces for production purpose the total number of curated zeolite crystal instances from 1600 to 1305.

4 Attributes

Along this work different attributes were tested and here we give a description of the thirty most useful for the machine learning classification of zeolite framework types.

Topological descriptors are developed through Delaunay tessellation. Based on the crystallographic information in the ICSD, a large spherical supercell of each zeolite crystal was generated [19,20]. The Computational Crystallography Toolbox [22] was used for this purpose. By filtering out all the oxygen atoms, cations and adsorbent phase, only T atoms were kept in the supercell. Each T atom represents a point in the space for Delaunay tessellation, which was achieved using the Qhull package [23].

Delaunay tessellation generates tens of thousands of tetrahedra (Delaunay simplices) in the supercell. Here the T atoms are of generic type, regardless of their chemistry nature. Therefore, there is only one type of Delaunay simplex in all the zeolites, with one T atom at each of its four vertices. These Delaunay simplices are geometrically different, with most of them quite distorted from a regular tetrahedron. This is in contrast with the TO_4 framework building units (tetrahedra with O at vertices) of zeolites which are always near-equilateral.

For each Delaunay simplex, three characteristic geometrical properties were generated. The first is tetrahedrality (T), which is a quantitative measure of the degree of distortion of a Delaunay simplex from a perfect equilateral tetrahedron [14]. If T is zero the simplex is a perfect tetrahedron, otherwise the simplex is a distorted tetrahedron and a larger value of T means larger degree of distortion. The second geometrical characteristic is the in-sphere volume (V_{in}) of each simplex identifying the volume of the largest sphere that can be enclosed in such a simplex. The third geometrical property is the average edge length (d) of the 6 edges in a simplex.

For each zeolite crystal, there are tens of thousands of Delaunay simplices in the selected spherical supercell. Therefore, the *mean* and *standard deviation (σ)* of the three geometrical properties of a simplex were obtained from all the simplices and taken as geometrical descriptors or geometrical attributes for the machine learning study. This yields six attributes: $<T>$, $<V_{in}>$, $<d>$, σ_T, σ_V_{in}, σ_d.

Additional geometrical attributes were generated with the introduction of secondary simplices. Inspired by the local structure ensemble of Wako and Yamato [24], we adopted secondary simplices corresponding to the second coordination shell in Delaunay space [20]. Similarly as with Delaunay simplices, we generated the following six geometrical attributes for secondary simplices: $<T2>$, $<V2_{in}>$, $<d2>$, σ_T2, σ_V2_{in}, σ_d2, where "2" refers to secondary simplices.

Crystallographic information in each ICSD entry contains the six lattice parameters of the crystal unit cell (three lattice constants and three angles). The average and standard deviation of the edge length of the lattice vectors ($<a>$, σ_a), and average angle deviation from 90 degrees ($<skew>$) were selected as distinctive attributes of the unit cell. The ICSD also provides lattice information for the reduced cell. Therefore, the average values for the reduced cells were also selected: $<a'>$, σ_a', and $<skew'>$.

Chemical composition associated with the zeolite framework is taken into account explicitly by defining attributes for the relative concentrations of the elements in the T positions. Since Si, Al and P are the elements that predominantly populate the T positions in most zeolite structures, the compositional percentage of Si, Al and P were used as attributes: *[Si]*, *[Al]*, and *[P]*.

As an essential property for zeolites, framework density (*FD*) was also calculated as an attribute for all the crystals studied.

Additional attributes were directly taken from information contained in the ICSD: calculated density (ρ), weight of the chemical formula (W), z-value (z), unit cell volume (v), reduced cell volume (v'), crystal system (*sys*), center (c), and space group (*SG*).

This completes the 30-attribute set used for the machine learning process.

5 Attribute selection

Attributes in the 30-attribute set are not all independent. Based on the 41-class zeolite data set, correlation coefficients (cc) between each two of the attributes were calculated. To remove all correlations with $|cc|>0.9$, attributes $<a>$, $<a'>$, $<d>$, $<d2>$ were removed from the attribute set.

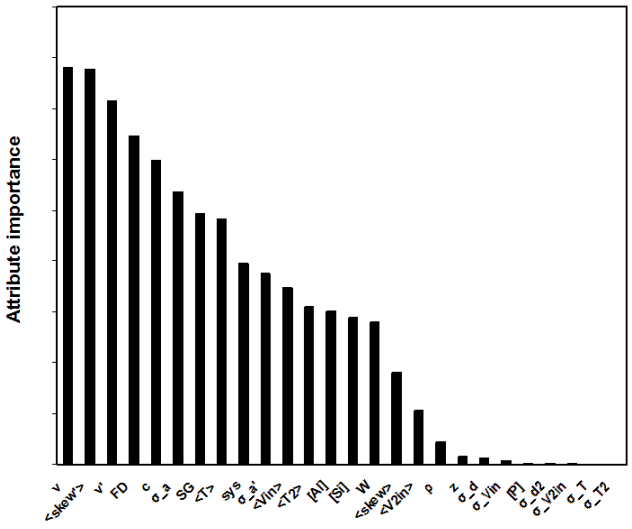

Fig.1 Relative importance of the 26 independent attributes estimated by the Random Forest algorithm.

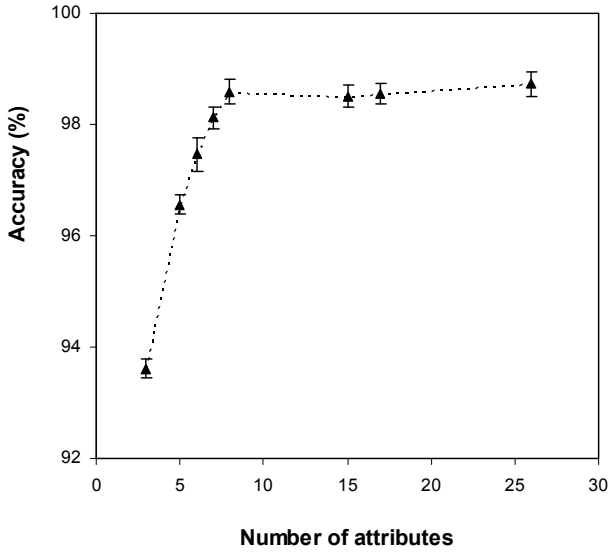

Fig.2 Classification accuracy as a function of the number of most important attributes adopted.

The importance of each of the remaining 26 attributes with respect to the framework classification was estimated by Breiman's Random Forest (RF) algorithm [25] through the R package [26]. The 41-class data set was used in the evaluation. The relative importance scores of the 26 attributes is shown in Fig. 1. The last 9 attributes starting with ρ, are relatively negligible for the classification. Therefore, it is safe to reduce attributes to 17 without losing performance in the classification. The number of less important attributes was then gradually removed from the attribute set, and the resulting classification accuracy was recorded. Accuracy refers to the percentage of correctly classified instances

among all instances, which was obtained by running stratified 10-fold cross validation 20 times with RF algorithm in the WEKA program [27]. One hundred trees in RF were used in the study. The results are shown in Fig.2. It is apparent the machine learning model with the top eight attributes is a transition point. With more attributes added, classification accuracy is not increased significantly. However, if less number of attributes are used in the classification, the model accuracy decreases quickly.

Decision tree classification (C4.5) [28] of the data with top 8 attributes provides a tree with 111 nodes including 59 leaves in which only 7 attributes were used to construct the tree (*v'* was excluded). Therefore, the attribute set can be reduced to a 7-attribute set different from the set using the top 7 attributes in Fig.1 and 2. The final set includes *v, <skew'>, FD, c, σ_a, SG,* and *<T>*. The classification accuracy with this 7-attribute set using RF as classifier is 98.57±0.20%, virtually no change from 98.59±0.23% obtained for the top 8-attribute set.

6 Machine learning results

Throughout this section, machine learning studies were carried out using RF with one hundred trees. The Zeolite Structure Predictor (ZSP) is established based on the above 7-attribute set to predict framework types of zeolite crystals.

Since the performance of the machine learning model is dependent on the size of available zeolite data, the ZSP was tested on four different sized data sets. The first data set includes all instances from 41 classes, with each having 3 or more instances. By gradually increasing the lower limit for number of instances per class, data sets 2 to 4 are constructed by removing small classes populated with less than 9, 19, and 63 instances, respectively. Results are listed in Table I, where OOB refers to out of bag error.

TABLE I
PREDICTION OF ZEOLITE FRAMEWORK TYPES FOR ZEOLITE CRYSTALS FROM THE ICSD USING THE ZSP MODEL. "*x*" REFERS TO NUMBER OF INSTANCES PER CLASS. RESULTS WERE OBTAINED FROM 20 REPEATED RUNS OF STRATIFIED 10-FOLD CROSS VALIDATION.

Data Set	#1	#2	#3	#4
Range of *x*	[3,351]	[9,351]	[19,351]	[63,351]
Total Classes	41	23	15	6
Total Instances	1305	1213	1119	893
OOB (×1000)	15±2	12±3	9±2	4±2
Accuracy (%)	98.6±0.2	98.9±0.1	99.2±0.1	99.8±0.2

With multi-category classification, the ZSP is very accurate for the prediction of zeolite framework types. The prediction power of the ZSP could be further enhanced when there are more instances available per class to train the model. Perfection is reached when each of the classes has more than 63 instances (as in data set 4).

During the data preprocessing process, there were 103 instances determined to be zeolites but with strong disorder in the structure. Among them, 77 instances correspond to a class within the 41-class data set. The ZSP model trained with the 41-class data set was used to classify these 77 zeolites yielding 69 correctly classified (89.6%. accuracy). On the other hand, with the conventional CS-VS method only 20.8% (16 out of 77) of the instances can be correctly predicted. This clearly shows the robustness of the ZSP model for the prediction of imperfect zeolite crystalline structures.

7 Conclusions

The artificial intelligence based zeolite structure predictor (ZSP) model can be used to predict framework types for zeolite crystals with very high accuracy. With new data and new materials added every year to crystal structure databases, the ZSP is expected to perform even better with more accumulated learning experience. With the robustness of the model, the ZSP can be used to reliably predict the framework type of strongly disordered zeolite crystals where the conventional CS-VS approach performs poorly.

The methodology used to build the ZSP model can be potentially applied to other types of materials. Research in this direction is in progress.

8 Acknowledgment

This work was supported under the National Science Foundation grant CHE-0626111. The authors acknowledge the National Institute of Standards and Technology for the ICSD data made available for this work.

9 References

[1] T. F. Degnan, Jr., "Applications of zeolites in petroleum refining," *Topics in Catalysis,* vol. 13, pp. 349-356, 2000.

[2] 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.

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

[4] D. S. Coombs *et al*., "Recommended nomenclature for zeolite minerals: report of the subcommittee on zeolites of the international mineralogical association, commission on new

minerals and mineral names," *The Canadian Mineralogist*, vol. 35, pp. 1571-1606, 1997.

[5] W. M. Meier, and D. H. Olson, "Zeolite frameworks," *Adv. Chem. Ser.*, vol. 101, pp. 155-170, 1971.

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

[7] Online database of zeolite structures from IZA-SC. Available: http://www.iza-structure.org/databases/

[8] 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.

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

[10] M. M. J. Treacy, M. D. Foster, and K. H. Randall, "An efficient method for determining zeolite vertex symbols," *Micropor. Mesopor. Mater.*, vol. 87, pp. 255-260, 2006.

[11] L. A. Baumes, M. Moliner, and A. Corma, "Prediction of ITQ-21 zeolite phase crystallinity: parametric versus non-parametric strategies," *QSAR Comb. Sci.*, vol. 26, pp. 255-272, 2007.

[12] J. M. Serra, L. A. Baumes, M. Moliner, P. Serna, and A. Corma, "Zeolite synthesis modelling with Support Vector Machines: A Combinatorial Approach," *Comb. Chem. High Throughput Screen.*, vol. 10, pp. 13-24, 2007.

[13] B. N. Delaunay, "Sur La Sphere Vide," *Izv. Akad. Nauk. SSSR, Otd Mat Est Nauk,* vol. 7, pp. 793-800, 1934.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] M. D. Foster, I. Rivin, M. M. J. Treacy, and O. D. Friedrichs, "A geometric solution to the largest-free-sphere problem in zeolite frameworks," *Micropor. Mesopor. Mater.*, vol. 90, pp. 32-38, 2006.

[19] D. A. Carr, M. Lach-hab, S. Yang, I. I. Vaisman, and E. Blaisten-Barojas, "Machine Learning Approach for Structure-based Zeolite Classification," *Micropor. Mesopor. Mater.*, vol. 117, pp. 339-349, 2009.

[20] S. Yang, M. Lach-hab, I. I. Vaisman, and E. Blaisten-Barojas, "Machine Learning Approach for Classification of Zeolite Crystals," in *Proc. 2008 Inter. Conf. Data Mining*, Las Vergas, 2008, pp. 702–706.

[21] Inorganic Crystal Structure Database (ICSD). Available: http://www.nist.gov/srd/nist84.htm

[22] Computational Crystallography Toolbox. Available: http://cctbx.sourceforge.net/

[23] Qhull 2003. Available: http://www.qhull.org/

[24] 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.

[25] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5-32, 2001.

[26] The R project for statistical computing version 2.5.0. Available: http://www.r-project.org/

[27] Weka 3: Data Mining Software in Java. Available: http://www.cs.waikato.ac.nz/ml/weka/

[28] Ross Quinlan, *C4.5: Programs for Machine Learning*, San Mateo, CA: Morgan Kaufmann, 1993.