Identifying Zeolite Frameworks with a Machine Learning Approach

Shujiang Yang,[†] Mohammed Lach-hab,[†] Iosif I. Vaisman,[†,‡] and Estela Blaisten-Barojas*[,†,§]

Computational Materials Science Center, and Department of Computational and Data Sciences, George Mason University, MSN 6A2, Fairfax, Virginia 22030, and Department of Bioinformatics and Computational Biology, George Mason University, MSN 5B3, Manassas, Virginia 20110

Received: July 23, 2009; Revised Manuscript Received: October 14, 2009

Zeolites are microporous crystalline materials with highly regular framework structures consisting of molecular-sized pores and channels. The characteristic framework type of a zeolite is conventionally defined by combining information on its coordination sequences, vertex symbols, tiling, and transitivity information. Here we present a novel knowledge-based approach for zeolite framework type classification. We show the predicting abilities of a machine learning model that uses a nine-dimensional feature vector including novel topological descriptors obtained by computational geometry techniques, together with selected physical and chemical properties of zeolite crystals. Trained on the crystallographic structures of known zeolites, this model predicts the framework types of zeolite crystals with very high accuracy.

1. Introduction

Zeolites, with the diversity of their natural forms, are among the most abundant mineral species on earth. In addition to about forty species occurring naturally, hundreds of other zeolites have been synthesized. They are widely used for adsorption, ion-exchange, and heterogeneous catalysis (a substantial portion of gasoline is produced with zeolites as catalysts) and in a number of emerging areas such as biomedical technology, sensors, and solar energy conversion.[1] These applications capitalize on the unique microporous crystalline structure of zeolites, characterized by uniformly distributed pores and channels of molecular size that give a topology signature to each zeolite type. As a consequence, proper identification of the zeolite topology pattern is crucial to a specific application.

The structure of a zeolite is composed of a three-dimensional supporting network filled with loosely bound exchangeable cations and adsorbent phase. The building blocks of the underlying network are TO_4 chemical groups where the central T atom (most commonly a Si, Al, or P atom) is tetrahedrally coordinated by four oxygen atoms. The backbone structure is constructed by linking TO_4 tetrahedral units through oxygen-corner sharing, yielding a network-like pattern. This pattern replicates periodically giving rise to well-organized arrays of channels that comprise topological characteristics specific to the zeolites.[2] Such an atomic backbone constitutes the framework of a zeolite, which gives a topological signature for identifying the network connectivity of the TO_4 building units. Frameworks do not depend on specific cations, adsorbent phase, chemical composition, or physical and mechanical properties of the zeolite crystals. Following the rules set up by the Commission on Zeolite Nomenclature of the International Union of Pure and Applied Chemistry,[3] a distinct framework type is labeled by a framework type code (FTC) denoted by three capital letters. FTCs are assigned and curated by the Structure Commission of the International Zeolite Association (IZA).[4] Search for novel type of zeolites has been in the past and continues to be today an actively pursued research area. Currently, 191 distinct framework types have been approved by IZA, including 5 frameworks approved in the first half of 2009 and 10 others in 2008.[4] The FTC of a zeolite is normally determined unambiguously using the standard approach relying on the combined determination of zeolite framework coordination sequences[5] and vertex symbols.[6] More recently IZA has included the symbolic tiling and transitivity descriptions[7,8] in the characterization of 189 ideal framework types. Delaney symbols (D-sym) can be determined from the tiling information, indicating the complexity of the atomic network associated to each framework type.[9] Complexity is important but not unique to each framework type. Albeit rare, it may happen that two real zeolite crystals belonging to different framework types would have identical coordination sequences and vertex symbols.[10] The latter is possible because FTCs are backed-up by theoretically built perfect framework structures. Therefore, a mechanism for univocally identifying FTCs from known natural and synthetic zeolite crystals, which are never perfect, is highly desirable.

Machine learning algorithms are used to discover complex patterns embedded within large amounts of data. They have been successfully applied in fields ranging from speech and vision recognition, robot control, and business management to bioinformatics and drug design. However, in materials science, and for the analysis of zeolite structure in particular, machine learning methods are practically unexplored as evident from the very limited body of literature on this subject.[11,12] Spatial patterns in the condensed phases of materials can be identified using Delaunay tesselation[13] of the point set associated to the site-location of atoms in such materials. This computational geometry approach provides an objective, nonarbitrary definition of nearest neighbor points in space that has been successfully applied for structural and topological characterization of a variety of condensed matter systems including liquids,[14] proteins,[15,16] and zeolites.[17]

In this work we combine the computational geometry techniques for generating essential topological descriptors characterizing the structure of zeolite crystals with the machine learning methods affording a novel classification approach of their framework types. Previously we have explored a similar

* Corresponding author. E-mail: blaisten@gmu.edu.
[†] Computational Materials Science Center.
[‡] Department of Bioinformatics and Computational Biology.
[§] Department of Computational and Data Sciences.

10.1021/jp907017u
Published on Web 12/07/2009

methodology for the classification of zeolites as mineral species relying on a small data set.[18] In this article we present for the first time the novel machine learning model that allows prediction of zeolite structure type based on topological, physical, and chemical descriptors.

This paper is organized as follows. Section 2 gives a description of the machine learning preprocess analysis of the data to be used and provides the methodology employed in determining the novel descriptors used to build the knowledge-based models. Section 3 develops and gives performance results for the Framework Type Predictor (FTP), which is a novel supervised classification model based on descriptors derived from the resolved crystallographic structure of zeolites archived in the FIZ/NIST Inorganic Crystal Structure Database (ICSD).[19,20] The FTP employs the Random Forest algorithm.[21] Section 4 presents the conclusions from this study.

2. Analysis of the Zeolite Data

Mapping the terminology of chemistry to that of machine learning, a zeolite framework type corresponds to a *class*, a zeolite crystal entry in the ICSD corresponds to an *instance*, and a zeolite property corresponds to a *feature*. The ICSD is the most comprehensive collection of zeolite crystal structure data, containing about 1600 entries queried as zeolites. This work utilizes the ICSD data.

2.1. Data Cleaning. Building machine learning models relies on a training process with a portion of the available data, and better training is achievable the larger the amount of data used for that purpose. Therefore, the FTP needs to be trained with known structural data of as many crystals as possible. Additionally, machine learning models rely on clean data, data that are reliably and effectively curated. After a thorough analysis of the information contained in the ICSD, we identified 1473 entries as zeolite crystals out of the 1600 available. The remaining entries do not correspond to any of the IZA-approved rules and thus should not be included in the training set. Furthermore, to assess the reliability of the 1473 entries, we evaluated the T−T bond length, T−O bond length, and TO_4 coordination and check their consistency with accepted zeolite structure. This analysis cleaned the data further by excluding 103 entries with excessive geometrical distortions. Finally, we assigned a framework type to the remaining 1370 entries because that information is not included in the ICSD. Recurring to the conventional method, we calculated both the coordination sequences and the vertex symbols with the help of zeoTsites[22] and TOPOS[23] programs and compare them with the standard table of framework types.[4] As a result, we confirmed that the 1370 entries populate 94 different framework types of zeolites.[24] The distribution of the 1370 entries among these framework types ranges from 1 to 351 entries per framework type.

Besides the data cleaning, machine learning studies require as many instances as possible within each class for a proposed model to be trained effectively. Therefore, classes populated with one or two instances are clearly inadequate for developing a reliable machine learning classifier. Consequently, we excluded 53 classes from consideration due to their small population. As a result, the proposed FTP model can be built based on the remaining 41 classes containing at least three instances each. The distribution of 1305 instances among these 41 retained classes is shown in Figure 1. Additionally, the horizontal-top scale of this figure depicts the D-sym of each framework obtained from the ideal IZA framework data employing the TOPOS and Systre/3dt[25] programs.

2.2. Feature Generation and Selection. A feature, also called attribute or descriptor, is a property of an instance that provides input to the machine learning process. In developing the FTP we initially generated more than 30 features[26−28] to characterize each of the zeolite crystals in the ICSD. From those, some were eliminated, based on a correlation analysis because only one of the correlated features was retained. Then, the remaining features were evaluated based on their importance for the FTP and the Random Forest classifier.[21,29] As a result, the nine most important features were retained containing topological, chemical, and physical information as described below.

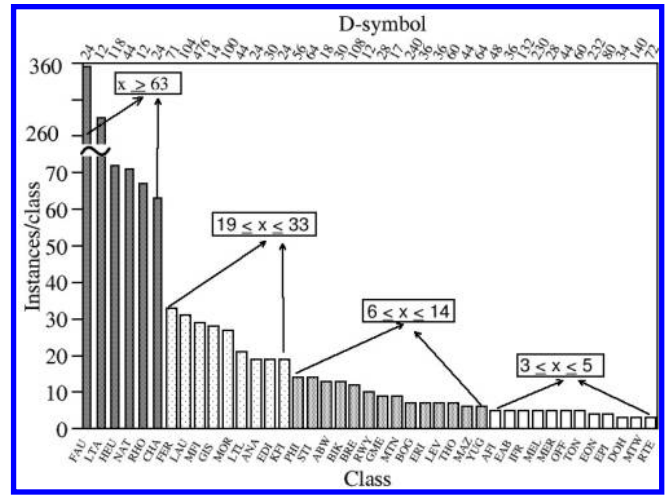

Figure 1. Distribution of zeolite instances in the 41 framework classes. Here x is the number of instances/class.

The ICSD contains symmetry operators and coordinates of the asymmetric cell for each crystal entry. It is then possible to span the unit cell for each zeolite crystal and subsequently build a supercell by translating such a cell in each of the three dimensions. It should be mentioned that the unit cells in zeolites are very large, containing up to 3000 atoms among the crystals analyzed here. In order to compare properties of different zeolites, three steps were applied to all of them. As a first step, all supercells were built to a size such that a spherical cut of radius 35.32 Å could be carved out of them. Typically, for those zeolites with huge unit cells, the spherical cut included the central unit cell and at least one neighboring unit cell in each direction in 3D. The next step was to filter out all the oxygen atoms, cations, and adsorbent phase within the 1305 spherical cuts. This process leaves only naked T-atoms within the sphere with the purpose of focusing more directly on the framework structure. In the third step, Delaunay tesselation was performed on the points occupied by the T-atoms inside the sphere, yielding tens of thousands of Delaunay simplices per spherical supercell. Just as a reminder, a Delaunay simplex is a tetrahedron (perfect or distorted) with T-atoms occupying its four vertices. The majority of Delaunay simplices obtained on the T-atoms are quite distorted tetrahedra contrary to the almost perfect TO_4 tetrahedra (primary building units) composing the network structure of the zeolites. This three-step procedure is schematically illustrated in Figure 2a,b,c.

The family of topological descriptors that sustain the FTP model was generated from the Delaunay simplices associated to each crystal. Two geometrical properties were considered for each Delaunay simplex: tetrahedrality and volume of the sphere inscribed in a simplex. Tetrahedrality is a quantitative measure of the simplex degree of distortion from a regular tetrahedron defined as:

$$T = \sum_{i}^{5} \sum_{j>i}^{6} \frac{(d_i - d_j)^2}{15\bar{d}^2} \quad (1)$$

where d_i is the length of the ith edge and $\bar{d}$ is the mean length of the six edges. When $T = 0$, the simplex is a perfect tetrahedron. The volume of the largest sphere inscribed in each simplex is referred to as in-sphere volume. For each zeolite crystal, there are tens of thousands of Delaunay simplices in the supercell spherical cut of radius 35.32 Å and thus mean and standard deviation of both tetrahedrality ($\langle T_1 \rangle$, σ_{T_1}) and in-sphere volume ($\langle V_1 \rangle$, σ_{V_1}) were generated and constitute four of the geometrical descriptors (features) entering in the FTP model.

Consideration of the second coordination shell in Delaunay space[30] is an important topology indicator that we incorporated in the model. Because each face of a Delaunay simplex is shared by a neighboring simplex, these neighboring simplices introduce four new points in space that form a larger tetrahedron. These are secondary simplices. Four additional geometrical descriptors were generated based on secondary simplices by considering the mean and standard deviation of both tetrahedrality ($\langle T_2 \rangle$, σ_{T_2}) and in-sphere volume ($\langle V_2 \rangle$, σ_{V_2}).

In addition to the geometrical descriptors, we selected six physical and chemical descriptors to complete the desired feature vector. Framework density (FD) is an essential property of zeolites defined as number of T-atoms per 1000 Å^3. Three other descriptors are the concentration of Si, Al, and P ([Si], [Al], [P]) selected because these elements occur in most zeolites in our data. The last two descriptors are properties of the normalized reduced cell.[31] One is the volume (v) and the other is the skewness (s) defined as the deviation of the mean of the lattice angles from 90°.

The importance of the 14 features for the FTP classification was evaluated with the Random Forest module in the R statistical package[29] The evaluation was based on the classification of the 41-class data set and the importance scores of the futures as illustrated in Figure 3. Consequently, the nine most important features v, s, FD, $\langle T_1 \rangle$, $\langle V_1 \rangle$, [Al],[Si], $\langle T_2 \rangle$, $\langle V_2 \rangle$ are retained in the model, thus including topological, chemical, and physical information. The other five features are of considerably less importance for the classification as clearly shown in Figure 3. In summary, four geometrical and five physical and chemical descriptors were generated for each crystal sample to serve as the nine-feature vector in the machine learning process.

3. Framework Type Predictor (FTP)

The FTP is a supervised classification model based on the nine-dimensional feature vector described in the previous section. The FTP employs the Random Forest classification algorithm[21] that outperforms many other machine learning techniques for our data. Machine learning classification with

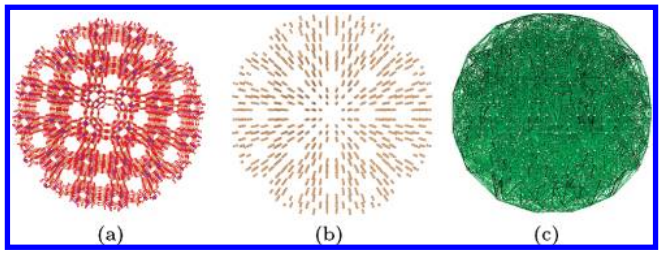

Figure 2. Procedure for obtaining the Delaunay simplices from one zeolite entry in the ICSD: (a) supercell spherical cut containing all atoms, (b) representation of the points in space occupied by the T-atoms, (c) all simplices obtained with Delaunay tesselation.

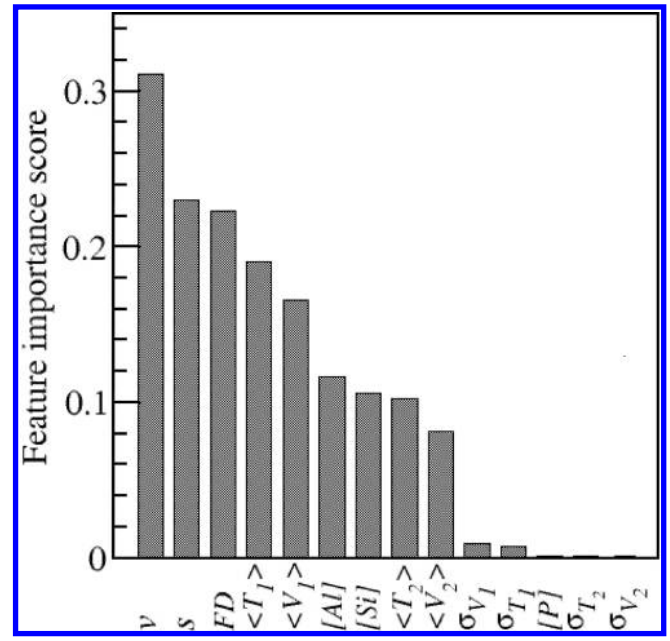

Figure 3. Importance score of the 14 features based on classification of the data set of 1305 instances in 41 classes.

the FTP model was performed using the WEKA package[32,33] and 100 trees in the forest of the classification algorithm.

What can the FTP model do? First, we find that the model performance for supervised binary classification (classification of data sets including two classes) is excellent. To measure the FTP performance in this area, data distributed in 41 classes were subdivided into subsets containing only data belonging to any pair of classes. This exercise gives 820 different two-class subsets of data. Evaluation of the FTP on these data subsets with leave-one-out cross validation shows that 704 out of all the 820 data sets are classified with 100% accuracy, 72 data sets are classified with 95−100% accuracy, 28 data sets are classified with 90−95% accuracy, and only 16 data sets (2%) are classified with less than 90% accuracy. Accuracy is defined as the percentage of instances correctly predicted by the model. The nonperfect classification cases involve one or both classes populated with a small number of instances. On the other hand, when both classes are well populated with instances, the classification is perfect or close-to-perfect. This exploratory test shows that the nine features on which the FTP is built are very capable of distinguishing different framework types of zeolites.

Apart from binary classification where machine learning models usually perform well, multiclass classification is a more difficult task rarely used in systems with complex patterns. The data for training the FTP are distributed in a highly unbalanced manner as shown in Figure 1. It is then worthwhile to investigate the relationship between data set size and the accuracy of FTP predictions in multiclass situations. For that purpose, the learning curve for the model was constructed based on a balanced data set of 360 instances equally distributed in six classes. Therefore, there is a maximum of 60 instances populating each of the six classes, which allow subdivision into training and testing sets of variable size. The model learning ability is then measured by analyzing how well the model performs as the size of the training set increases. Worth noting is that instances not selected for training constitute the testing data set. The learning curve for the FTP is shown in Figure 4. Results were obtained based on training sets containing instances drawn at random and without replacement from the full balanced data set of 360. Dots and error bars in Figure 4 are averages and corresponding

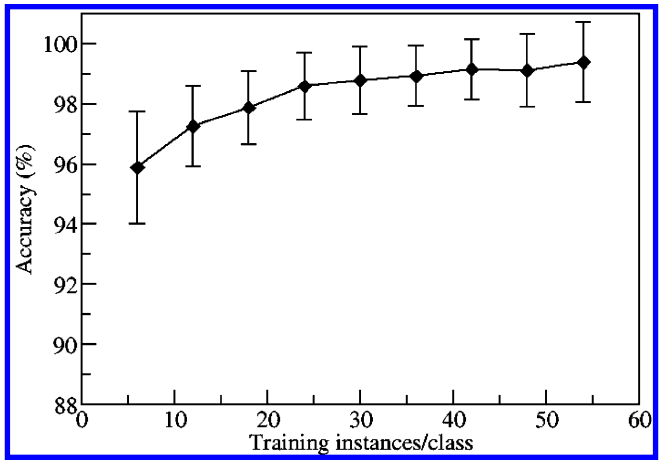

Figure 4. Learning curve of the FTP for classification into six classes. Shown is the mean ± SD.

TABLE 1: FTP Classification of Three Data Subsets Composed of Different Instances Populating Classes with Variable Population[a]

	set-12-3 $3 \le x < 5$	set-14-6 $6 \le x < 14$	set-15-19 $x \le 19$
number of classes	12	14	15
number of instances	52	134	1119
OOB (%)	7.6 ± 1.3	5.3 ± 0.9	1.0 ± 0.1
accuracy (%)	94.0 ± 0.9	96.1 ± 0.5	98.9 ± 0.1

[a] Results are mean ± SD and x = instances/class.

standard deviations obtained on 100 training/test splits of the data. It is shown in Figure 4 that the FTP model improves rapidly in the beginning, improves more gradually as the training data increases over 40% of the data set size, and reaches a plateau for training data sizes over 70% of the full data set. Clearly shown is the perfect performance of the FTP when trained with 90% of the data set. Basically all the test instances are classified correctly as belonging to one of the six considered classes.

Effects and influence of data set size in multiple-class classifiers may manifest itself in a variety of situations. For example, we investigated the FTP performance for classification of data sets in about 12−15 classes that have significantly different populations. By grouping the 1305 instances into subsets according to population of the classes, we formed three subsets as follows: SET-12-3 composed of 12 low population classes with 3 to 5 instances per class; SET-14-6 composed of 14 moderately populated classes with 6 to 14 instances per class; SET-15-19 composed of 15 well populated classes with 19 or more instances per class. Results of this experiment are shown in Table 1, where accuracy is obtained with stratified 10-fold cross validation[33] and repeated 20 times. The out of bag error (OOB) is an alternative estimate of the model performance in which the error on the testing subset of data is performed for each bootstrap of samples in Random Forest.[33] It is evident that the more instances available in each class, the higher accuracy and lower OOB the model can achieve. More importantly, both accuracy and OOB results cast in Table 1 show that the FTP performance for multicategory classification with the FTP model is highly reliable.

Based on the performance of the FTP for classifying data into multiple classes, the model was additionally built on the full 1305 instances nonuniformly distributed between the 41 classes with stratified 10-fold cross validation which was repeated 20 times. Results are summarized as case-1 in Table 2 showing a very high accuracy of 98.5% and very low OOB. Moreover, inspection of a second case (case-2 in Table 2) in which the data set is depleted by eliminating instances belonging to classes populated with less than 19, shows an expected improvement. Inspection of a third case (case-3 in Table 2) in which the criterion of reducing the number of classes with poor population is exploited further, shows that the FTP can achieve perfect classification (about 100% accuracy).

TABLE 2: FTP Classification on Three Selected Cases of Unbalanced Data Sets[a]

	case-1 $x \ge 3$	case-2 $x \ge 19$	case-3 $x \ge 63$
number of classes	41	15 6	
number of instances	1305	1119	893
OOB (%)	1.7 ± 0.2	1.0 ± 0.1	0.3 ± 0.1
accuracy (%)	98.5 ± 0.1	98.9 ± 0.1	99.8 ± 0.1

[a] Results are mean ± SD and x = instances/class.

Comparison of the FTP performance with a random model gives an estimate of the significance of our results. Indeed, randomly shuffling instances between the 41 classes and keeping the same feature vector builds such random model. Classification with Random Forest obtained from this random model yields only an accuracy of 2.4 ± 0.6% (mean ± SD) on 20 runs, which is consistent with a random guessing for 41 class labels. This affirms the effectiveness of the FTP model.

Geometry distortions from the ideal zeolite structure were detected in the ICSD data during the data cleaning process. We have researched thoroughly the structure of the 103 instances with geometric distortions, checked all associated publications, and tentatively proposed a framework type for them.[34] On the basis of this study, 77 instances belong to the 41 classes included in the FTP. The conventional methodology of coordination sequences/vertex symbols (using geoTsites,TOPOS, Systre/3dt) was able to spit sequences and symbols in some cases but failed in producing the adequate tiling and transitivity information. The D-symbol could not be calculated either. One example is ICSD entry no. 37148 for which sequences and symbols can be determined, but they do not match any IZA framework type. For this particular entry, tiling, transitivity, and D-symbol could not be obtained. In summary, with the conventional method only 39 instances out of 77 (50.6% accuracy) matched the IZA framework types. On the other hand, the FTP predicts 64 of the instances correctly with accuracy of 83% when trained with the data set of case-1 reported in Table 2.

In recent years, machine learning and data mining successes have been transformative in fields where huge amounts of data are being accrued at unprecedented speeds. Although the amount of data used for the development of the FTP model is currently limited, the FTP is expandable when more data becomes available as new crystal data are added to the ICSD and new hypothetical zeolites are generated at an increasing pace. The FTP is not only an alternative to the coordination sequences and vertex symbols determination and tiling/transitivity but is also more robust for handling data containing crystal distortions than the conventional method. The FTP model offers a novel way for predicting zeolite framework types based on crystallographic data.

4. Conclusions

We developed and presented here the first knowledge-based model capable of classifying and predicting topological types of zeolite structures with very high accuracy in a completely

automated fashion. Such a model may prove extremely useful for the curation of large amount of structural data submitted to the crystallographic structure repositories and for various applications related to the design of new zeolites and other materials. Furthermore, the combination of computational geometry and machine learning methodologies employed to build FTP is not limited to zeolites and will substantially advance the field of materials and chemical informatics because most mineral groups are not well classified. The FTP marks a step forward for learning from data and opens a horizon for utilizing the hidden power of over 100 000 crystal entries collected in the ICSD. FTP is potentially adaptable for unsupervised learning and for classifying other families of inorganic crystals, enhancing the aesthetics of data manipulation and other crystallographic databases.

Acknowledgment. This work was supported under the National Science Foundation grant CHE-0626111. Authors gratefully acknowledge the Standard Reference Data Program of the National Institute of Standards and Technology for making available the zeolite data set from the ICSD. TERAGRID grant PHY050026T is acknowledged for the computer time allocation.

References and Notes

(1) Payra, P.; Dutta, P. K. In *Handbook of Zeolite Science and Technology*; Auerbach, S. M.; Carrado, K. A.; Dutta, P. K., Eds.; Marcel Dekker: New York, 2003.

(2) McCusker, L. B.; Liebau, F.; Engelhardt, G. *Pure Appl. Chem.* **2001**, *73*, 381–394.

(3) Barrer, R. M. *Pure Appl. Chem.* **1979**, *51*, 1091–1100.

(4) IZA-SC database of ideal zeolite structures. http://www.iza-structure.org/databases, 2009.

(5) Meier, W. M.; Moeck, H. J. J. *Solid State Chem.* **1979**, *27*, 349–355.

(6) OÕKeeffe, M.; Hyde, S. T. *Zeolites* **1997**, *19*, 370–374.

(7) Delgado-Friedrichs, O. *Theor. Comput. Sci.* **2003**, *303*, 431–445.

(8) Delgado-Friedrichs, O.; Dress, A. W. M.; Huson, D. H.; Klinowski, J.; Mackay, A. L. *Nature* **1999**, *400*, 644–647.

(9) Dress, A. W. M. *Adv. Math.* **1987**, *63*, 196–212.

(10) Treacy, M. M. J.; Foster, M. D.; Randall, K. H. *Microporous Mesoporous Mater.* **2006**, *87*, 255–260.

(11) Serra, J. M.; Baumes, L. A.; Moliner, M.; Serna, P.; Corma, A. *Comb. Chem. High Throughput Screen.* **2007**, *10*, 13–24.

(12) Baumes, L. A.; Moliner, M.; Corma, A. *QSAR Comb. Sci.* **2007**, *26*, 255–272.

(13) Delaunay, B. N. *Izv. Akad. Nauk. SSSR, Otd Mat Est Nauk* **1934**, *7*, 793–800.

(14) Vaisman, I. I.; Brown, F. K.; Tropsha, A. *J. Phys. Chem.* **1994**, *98*, 5559–5564.

(15) Singh, R. K.; Tropsha, A.; Vaisman, I. I. *J. Comput. Biol.* **1996**, *3*, 213–221.

(16) Vaisman, I. I. In *Handbook of Computational Statistics*; Gentle, J. E.; Hrdle, W.; Mori, Y., Eds.; Springer: New York, 2004.

(17) Foster, M. D.; Rivin, I.; Treacy, M. M. J.; Friedrichs, O. D. *Microporous Mesoporous Mater.* **2006**, *90*, 32–38.

(18) Carr, D. A.; Lach-hab, M.; Yang, S.; Vaisman, I. I.; Blaisten-Barojas, E. *Microporous Mesoporous Mater.* **2009**, *117*, 339–349.

(19) Inorganic Crystal Structure Database (ICSD). http://www.nist.gov/srd/nist84.htm (2007).

(20) Belsky, A.; Hellenbrandt, M.; Karen, V. L.; Luksch, P. *Acta Crystallogr.* **2002**, *B58*, 364–369.

(21) Breiman, L. *Mach. Learn.* **2001**, *45*, 5–32.

(22) Sastre, G.; Gale, J. D. *Microporous Mesoporous Mater.* **2001**, *43*, 27–40.

(23) Blatov, V. *IUCr Comput. Commun. Newsl.* **2006**, *7*, 4–38.

(24) Lach-hab, M.; Yang, S.; Vaisman, I. I.; Blaisten-Barojas, E. Assignment of Framework Types to the Zeolite Crystals in the Inorganic Crystal Structure Database. arXiv:0904.2597V1. http://arxiv.org/0904.2597 (2009).

(25) Delgado-Friedrichs, O.; O'Keeffe, M. *J. Solid State Chem.* **2005**, *178*, 2480–2485.

(26) Yang, S.; Lach-hab, M.; Vaisman, I. I.; Blaisten-Barojas, E. *Proceedings of 2008 International Conference on Data Mining*, Stahlbock, R.; Crone, S. F.; Lessmann, S., Eds.; CSREA: Las Vegas, NV, 2008, pp 702−706.

(27) Yang, S.; Lach-hab, M.; Vaisman, I. I.; Blaisten-Barojas, E. *Lect. Notes Comput. Sci.* **2009**, *5545*, 160–168.

(28) Yang, S.; Lach-hab, M.; Vaisman, I. I.; Blaisten-Barojas, E. *Proceedings of the 2009 International Conference on Artificial Intelligence*, Arabnia, H. R.; de la Fuente, D.; Olivas, J. A., Eds.; CSREA: Las Vegas, NV, 2009, pp 340−344.

(29) The R project for statistical computing version 2.5.0, http://www.r-project.org (2007).

(30) Wako, H.; Yamato, T. *Protein Eng.* **1998**, *11*, 981–990.

(31) Mighell, A. D. J. *Res. Natl. Inst. Stand. Technol.* **2003**, *108*, 447–452.

(32) Weka 3: Data Mining Software in Java. http://www.cs.waikato.ac.nz/ml/weka (2009).

(33) Witten, I. H.; Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed.; Morgan Kaufmann: San Francisco, CA, 2004; Ch. 5.

(34) Yang, S.; Lach-hab, M.; Vaisman, I. I.; Blaisten-Barojas, E. J. Phys. Chem. Ref. Data, submitted.

JP907017U