Catalysis Today 159 (2011) 74–83

Contents lists available at ScienceDirect

Catalysis Today

journal homepage: www.elsevier.com/locate/cattod

Novel graph machine based QSAR approach for the prediction of the adsorption enthalpies of alkanes on zeolites

A. Goulon[a], A. Faraj[a], G. Pirngruber[b], M. Jacquin[b], F. Porcheron[b], P. Leflaive[b,*], P. Martin[c], G.V. Baron[c], J.F.M. Denayer[c]

[a] Applied Mathematics Department, IFP, 1&4, Avenue de Bois-Préau, 92852 Rueil-Malmaison Cedex, France
[b] Separation Department, IFP-Lyon, Rond-point de l'échangeur de Solaize, BP3, 69360 Solaize, France
[c] Department of Chemical Engineering, Vrije UniVersiteit Brussel, Pleinlaan 2, B-1050 Brussel, Belgium

ARTICLE INFO

Article history:
Available online 25 September 2010

Keywords:
Alkanes
Zeolites
QSAR
QSPR
Graph machines

ABSTRACT

This paper presents the application of a new machine learning approach called Graph Machines to the prediction of the adsorption enthalpies of linear and branched alkanes on various zeolites. In this approach, the molecules are considered as structured data and are represented by graphs. For each individual of the data set, a mathematical function (graph machine) is built, which structure reflects the one of the molecule under consideration. This approach differs from classical quantitative structure–activity relationship (QSAR) methods where molecules are generally described using vectors composed of descriptors. Since no molecular descriptors are used, the collection, computation and selection of these descriptors, which is often a major issue in QSAR applications, is no longer required.

The models developed using this approach allowed to satisfactorily predict adsorption enthalpies of all zeolites, even using a very limited training set of 10 molecules. The efficiency of such models, allowing the modelling of complex adsorption behaviour like zeolite ZSM-22 using only few experimental data, illustrates the potential of this new approach in the screening of zeolites for catalytic or adsorption based separation applications.

© 2010 Published by Elsevier B.V.

1. Introduction

Adsorption enthalpies of linear and branched alkanes on zeolite NaY, Na-USY, mordenite, beta, ZSM-5 and ZSM-22, originally published in [1], were examined by statistical data analysis in a previously published work [2]. Partial least squares (PLS) regression was used to build quantitative structure–activity relationship (QSAR) models which were applied to predict the adsorption enthalpy. PLS regression together with descriptors carrying topological and structural information about the adsorbates was found to be efficient to predict the adsorption behaviour of the zeolites. It was nevertheless found that the predictions for zeolite ZSM-22 were always less efficient than for the other zeolites. It was concluded that, for this specific solid, a non-linear model could probably provide better results than the linear PLS based models. Recent works proposed the use of artificial neural networks (ANN) to establish non-linear models to predict the adsorption of different molecules mainly on activated carbons, and compared the results to those of linear models such as multiple linear regression (MLR) or PLS.

Brasquet and Le Cloirec [3] studied the adsorption properties of 368 aliphatic and aromatic compounds on an activated carbon while the adsorption of 55 compounds onto activated carbon fibers was considered in [4]. In both cases, experimental data were fitted with the Freundlich equation. In these works, the authors compared two statistical tools, MLR and neural networks (NN), the major difference between the two models being that the first one uses a well-defined function to fit the data while the neural network performs a model-free mapping of the molecular structure descriptors to predict the adsorbability. The models were first compared for their ability to represent training data. Both studies showed the good ability of neural network to describe the adsorption data, the statistical quality of the NN model being higher than that of the multiple linear regression model. Models were also compared for their ability to predict the adsorption data of compounds out of the training database. The prediction performance of neural network was found to be low, even below that of MLR especially when test data are very different from training data in terms of compound structure.

More recently the same group tested a multi-linear regression and a neural network (NN) model to predict the integral adsorption enthalpies from zero coverage to saturation of volatile organic compounds (VOCs) on activated carbons [5]. The authors used a two-step procedure, using first the MLR model to discriminate the

* Corresponding author. Tel.: +33 4 78 02 28 34; fax: +33 4 78 02 20 66.
E-mail address: philibert.leflaive@ifp.fr (P. Leflaive).

0920-5861/$ – see front matter © 2010 Published by Elsevier B.V.
doi:10.1016/j.cattod.2010.07.029

significant input variables out of a set of eight for the molecules and of a set of five for the activated carbons, and then a NN approach to improve the MLR model by introducing non-linear relationships. Almost no improvement was obtained with neural networks. On the opposite, Timofei et al. [6], using the same procedure (i.e. descriptor selection with MLR and implementation of these descriptors as input in a NN model) applied to the adsorption of disazo anionic dyes on cellulose, found a significantly improved fitting of the NN model over the MLR model.

In the field of zeolites, a neural network model was used to predict the ternary adsorption equilibria of 2,6- and 2,7-dimethylnaphtalene isomers dissolved in supercritical carbon dioxide on NaY-type zeolite [7]. It was found that despite a limited number of data points available for the training of the network, the model was capable of predicting the adsorption equilibria very precisely. Ravikumar et al. [8] also used artificial neural networks to correlate the isosteric heat of adsorption of organics molecules over zeolites with various calculated descriptors such as equalized electronegativities, chemical hardness, Lennard–Jones parameters. When comparing the results to those of multi-linear regression models based on the same input parameters, the authors found a better representation of the data by ANN compared to that of MLR.

Based on all these literature results, it can be expected that the use of non-linear models (especially using ANN) would provide a better representation of the data as compared to linear models such as MLR or PLS. The model accuracy is expected to be especially improved when linear models fail to accurately fit the experimental data.

The purpose of this work is to use a novel approach called graph machines, exploiting artificial neural networks (ANN). All the models used in the literature for adsorption are based on vector machines, i.e. that the adsorbate molecules have to be first transformed into a vector of variables called descriptors. The machine learning then performs a mapping of a set of input vectors (i.e. the characteristics or the properties of the molecules under consideration) to a set of output vectors (i.e. their adsorption properties). Recently, Dreyfus group [9–12] showed that when the inputs of a system can be described as structured data (e.g. applications in chemistry where inputs are molecules), this structure can directly and efficiently be used to model the outputs of the system. In this approach, called graph machines, the molecules are considered as structured data and are represented by graphs. For each individual of the data set, a mathematical function (graph machine) is built, which structure reflects the one of the molecule under consideration. This approach was successfully applied to the prediction of boiling point or toxicity of organic molecules [10,11] or their anti-HIV activities [11,12]. When compared to MLR [10,11], PLS [12] or descriptor based neural networks [10,12], graph machines based models were always found to be more efficient in terms of modelling and prediction.

This new graph machine/QSAR method was applied to predict the adsorption enthalpy at zero coverage of C5–C8 alkanes (17 molecules) on Na-Y, Na-USY, beta, mordenite, ZSM-5, ZSM-22. The predicted data are compared to both the experimental data published by Denayer et al. [1] and to the values predicted by PLS models [2].

2. Graph machines

Statistical learning consists in building, from empirical data, mathematical models which copy the behaviour of a process, so that the values of the outputs of this process can be predicted from its inputs. Modelling the relationships between molecular structures and their properties and activities (such as here the adsorption enthalpy on a given solid) is an important field of research. Classical modelling techniques draw non-linear mappings between the studied properties and structural features or other properties of the molecules, called descriptors. Graph machines [9–12] circumvents the main drawbacks of the descriptor approach (difficulty of the choice of relevant descriptors and their computation), by drawing a direct relationship between the structure of the data and the modelled property. Molecules are considered as structured data, and represented as graphs. Each graph of the dataset (i.e. each molecule considered in the study) is associated to a mathematical function (graph machine) with the same structure, which is intended to encode the structure of the graph and to provide a prediction of the studied property. This graph machine function is obtained by composing identical parameterized functions (called base or node function), for example polynomials or neural networks. In our case, the node function is implemented by a feedforward neural network. Modelling the properties associated to the graphs (e.g. the properties or activities of molecules) then consists in estimating the parameters of the node function so that the values of the graph machine functions are as close as possible to the values of these properties (best fit).

2.1. Mathematical structures of the graph machines

In a first step, the molecules, described by their SMILES (simplified molecular input line entry specification), are converted into labelled graphs by the association of each non-hydrogen atom to a vertex, and each bond to an edge. The vertices are also assigned labels describing the atoms they are related to (e.g. their natures, degrees or stereoisomeries). Then, the adjacency matrices associated to these labelled graphs are generated. These matrices are put into a canonical form, by the use of an algorithm ranking the nodes, according to criteria such as their degree or their belonging to a cycle [13]. This canonical form allows the choice of the root nodes, and the conversion of the graphs into directed acyclic graphs. As many edges as there are cycles in the graphs are selected and cut, and finally the edges are given a direction, from the roots of the trees to their terminal nodes. Although the cut edges are no longer present in the directed acyclic graph formed in this way, the information on their presence is saved due to the labels of the nodes. Fig. 1 illustrates an example of conversion of a molecule from its SMILES representation into a directed acyclic graph.

Then, for each graph G_i, a mathematical function is built the following way: each node of G_i is associated to a parameterized function called "node function" $f_{\boldsymbol{\theta}}$, $\boldsymbol{\theta}$ being the vector of parameters, which is the same for all the functions. These functions $f_{\boldsymbol{\theta}}$ are then

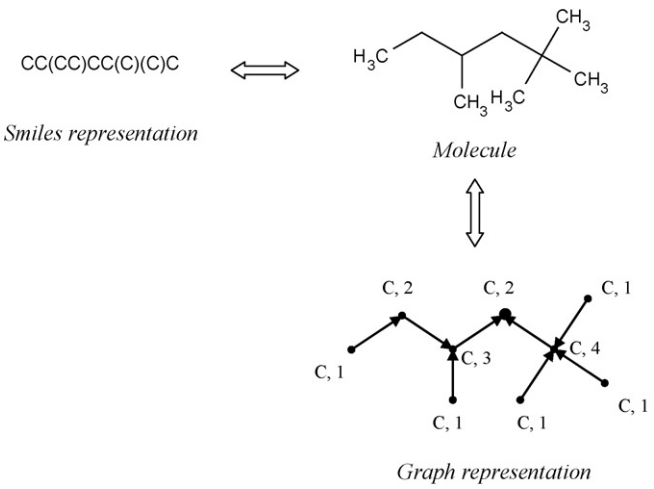
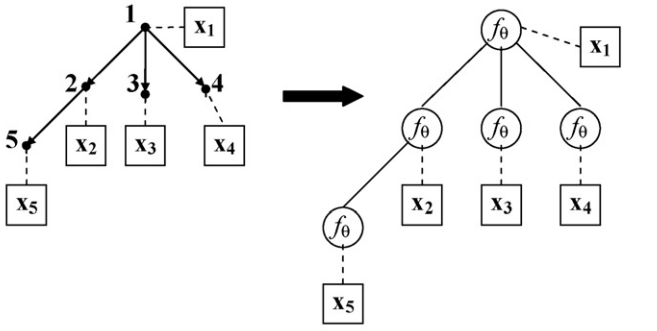


Fig. 1. Encoding of 2,2,4-trimethylpentane molecule into a directed acyclic graph.



$$G_{\theta} = f_{\theta}\left(f_{\theta}\left(f_{\theta}\left(0,0,0,\mathbf{x}_5\right),0,0,\mathbf{x}_2\right),f_{\theta}\left(0,0,0,\mathbf{x}_3\right),f_{\theta}\left(0,0,0,\mathbf{x}_4\right),\mathbf{x}_1\right)$$

Fig. 2. Building of the graph machine function associated with the directed acyclic graph of a molecule.

composed so that the global function reflects the structure of the graph: if s_a and s_b are two nodes of G_i, so that an edge comes from s_a and ends to s_b (i.e. s_a is the child of s_b), then the result of the function associated to s_a is an input of that associated to s_b. The node function corresponding to the node s_b is then:

$$f_{\boldsymbol{\theta}}(\mathbf{z}_b) = f_{\boldsymbol{\theta}}(\mathbf{v}_b, \mathbf{x}_b)$$

where $\mathbf{v}_b$ is a vector which components are equal to the outputs of the children nodes of s_b. If the node has no child, this vector is null. $\mathbf{x}_b$ is an optional vector which conveys information about the nodes. This information can be related to the nature of the atom the node represents, its degree in the original graph, or its stereoisomery.

The parameterized function, called graph machine, related to the graph G_i, is then:

$$g^i_{\boldsymbol{\theta}} = f_{\boldsymbol{\theta}}(\mathbf{z}_r)$$

where $\mathbf{z}_r$ is the input vector of the function associated to the root node.

When such functions are built from a set of graphs $G = \{G_i\}$, the node functions $f_{\boldsymbol{\theta}}$ are identical within each function g^i and across all those functions.

Fig. 2 illustrates the processing of a molecule from its graph representation into a graph machine.

In this graph, node 1 is the root node, and parent of the nodes 2, 3 and 4; node 2 is the parent of node 5; nodes 5, 3 and 4 have no child. If we denote x_j the value of x for node j and v_j the output of $f_{\boldsymbol{\theta}}$ for this node, then $v_2 = f_{\boldsymbol{\theta}}(v_5, 0, 0, x_2)$ for node 2 has only one child (node 5), $v_5 = f_{\boldsymbol{\theta}}(0, 0, 0, x_5)$, $v_3 = f_{\boldsymbol{\theta}}(0, 0, 0, x_3)$, $v_4 = f_{\boldsymbol{\theta}}(0, 0, 0, x_4)$ and $v_1 = f_{\boldsymbol{\theta}}(v_2, v_3, v_4, x_1)$. Finally, $G_{\boldsymbol{\theta}} = v_1$.

2.2. The training of graph machines

Training the graph machines consists in searching the parameters $\boldsymbol{\theta}$ which lead to the best approximation of the regression function, with the help of the pairs of inputs/outputs of the training set. In the framework of classical learning methods, the parameters of a model $g_{\boldsymbol{\theta}}$ (for instance a polynomial or a neural network...) are estimated with a set of examples, formed by N pairs $\{(\mathbf{x}^i, y^i), i = 1, \ldots, N\}$, where the vectors $\boldsymbol{x}^i$ are the inputs of the model, and y^i the measured values of the property to be modelled. The model is identical for all examples, and the cost function to minimize is:

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{N} (y^i - g(x^i, \boldsymbol{\theta}))^2$$

where $g(\mathbf{x}^i, \boldsymbol{\theta})$ is the value of the model for example i.

During the training of graph machines, the training set is composed of N structures/outputs pairs $\{(G_i, y^i), i = 1, \ldots, N\}$, where G_i is the parameterised function associated to graph i, and y^i is the value of the modelled property for this graph. The model is not unique: each example i corresponds to a model defined by a singular function $g^i_{\boldsymbol{\theta}}$, which is the composition of several functions $f_{\boldsymbol{\theta}}$, so that $g^i_{\boldsymbol{\theta}}$ reflects the structure of the graph. A cost function, similar to the traditional least square function, can be defined. This cost function takes into account the discrepancy between the predictions of the models and the observations present in the training set:

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{N} (y^i - g^i_{\boldsymbol{\theta}})^2$$

Graph machines are trained in the usual framework of empirical risk minimisation. Similarly to classical modelling techniques, usual model selection techniques such as hold-out, K-fold cross-validation, leave-one-out, virtual leave-one-out, can be applied to recursive networks and to graph machines. In the following application to QSAR, cross-validation and leave-one-out were used.

2.3. Neural networks (NN)

Neural networks are statistical tools, derived from a simplified concept of the brain which enable a non-linear relationship between a dependent and some independent variables to be determined. A neural network contains some nodes, called neurons, which are interconnected in a netlike structure generally composed of three layers: one input layer, one output layer, and one intermediate layer, the hidden layer. The required number of hidden neurons is optimized by an iterative process. The degree of influence between interconnected neurons is represented by numerical weights called connection weights. The overall behaviour of the system is modified by adjusting the connection weight values through the repeated application of the back-propagation algorithm. Neural network training is achieved when the error function, which measures the difference between calculated and desired output values, is minimized.

2.4. Comparison of graph machine and artificial neural network approach

A brief summary of the artificial neural network based graph machine approach (i.e. graph machine with ANN as the node function) as compared to classical ANN based QSAR method is given in Table 1.

In graph machines, as chemical information is brought in the model using the graph of the molecules, the node or base $f_{\boldsymbol{\theta}}$ function can have a simpler form than the $F(\boldsymbol{\theta})$ used in the ANN approach. Especially, when the base function of the graph machine is a neural network, the number of hidden neurons and thus the number of parameters to be adjusted during training is much smaller in the

Table 1
Comparison of graph machine and classical ANN based QSAR methods.

Approach	Chemical information introduced in the model	Mathematical form of the model	Function adjusted during training
Graph machine	Graph of the molecules	One function $G(f_{\boldsymbol{\theta}})$ for *each* molecule	$f_{\boldsymbol{\theta}}$
Artificial neural network	Molecular descriptors	One function $F(\boldsymbol{\theta})$ for all molecules	$F(\boldsymbol{\theta})$

graph machines as compared to classical ANN. Therefore, the number of inputs required for an efficient training of the model, which often a major burden in the ANN approach, is also smaller.

3. Results and discussion

Graph machines based QSAR approach is used to analyse and predict the adsorption properties of hydrocarbons on zeolites with different topologies. In particular, the models are used to predict the zero coverage adsorption enthalpies obtained from the temperature dependence of the Henry's constants of 17 alkanes and iso-alkanes on zeolites NaY, Na-USY, beta, mordenite, ZSM-5 and ZSM-22. The experimental data were originally published in [1] and are presented in Tables 3–5. The adsorption enthalpies are considered as the coordinates of the 17 individuals (17 molecules) in a six-dimensional space where the dimensions are vectors corresponding to the zeolites. A comprehensive statistical analysis of the data is provided in [2].

Graph machines were built for each individuals of the whole set (i.e. each molecule). The f_{θ} functions were implemented by feedforward neural networks with one to three hidden neurons. As previously described, each neural network is associated to a node in the graph. Its inputs are equal to the outputs of the neural networks associated to the parent nodes, whereas its output is the input of its children nodes. The only exception is the central node: its output is the output of the function associated to the graph, i.e. the prediction of the model. The complexity of the model was chosen with the leave-one-out technique: for each model (1, 2 and 3 hidden neurons), the leave-one-out score, which measures the generalization abilities of the model, was computed. The best score was obtained with two hidden neurons, and the corresponding f_{θ} function was selected to model the adsorption enthalpy of the molecules.

3.1. Model validation

The model validation was carried out using a classical leave-one-out cross-validation procedure [14]. As the name suggests, leave-one-out (LOO) cross-validation [15] involves the use of a single observation (here the zero coverage adsorption enthalpy of a molecule on a given zeolite) from the original data set as a validation data, and the other observations (i.e. the zero coverage adsorption enthalpy of the other molecules on the same zeolite) as the training data. This procedure is repeated such that each observation in the whole set is used once as the validation data. For each zeolite, the zero coverage adsorption enthalpy of a given molecule is then predicted with a model set with the adsorption enthalpies of all the other molecules on that zeolite. A correlation is then made between the predicted and the experimental values of the adsorption enthalpies.

Results of leave-one-out procedure are indicated in Fig. 3 for all zeolites. The correlation coefficient R and a mean error (ME) are indicated in Table 2.

We can see in Fig. 3 and Table 2 that regardless of the zeolite considered, there is a very good correlation between the predicted and the experimental values of the adsorption enthalpies. This indicates that model robustness is excellent whatever the nature of the zeolite in terms of pore size and the amount of cations in the zeolite. One can nevertheless see that for both ZSM-5 and beta, LOO procedure reveals one outlier value in each case, namely 2,2-dimethylbutane on ZSM-5 and 2,2,4-trimethylpentane on beta. In both cases, the predicted values much exceed the experimental values (respectively 71.9 kJ mol^{-1} instead of 63.9 kJ mol^{-1} for 2,2-dimethylbutane on ZSM-5 and 79.4 kJ mol^{-1} instead of 73.4 kJ mol^{-1} for 2,2,4-trimethylpentane). These results can be attributed to the increased contribution of repulsion forces at pore sizes smaller than the diameter of the molecules, leading to a decrease of the sorption enthalpies as reported by Eder and Lercher [16]. As 2,2-dimethylbutane can hardly enter the ZSM-5 zeolite [17] this can also lead to a much higher uncertainty on the adsorption enthalpy. Accordingly, 2,2-dimethylbutane on ZSM-5 is taken out from the data for the prediction study. In the case of 2,2,4-trimethylpentane on beta, the molecule can normally enter the zeolite, especially at high temperature where original experiments were carried out. The large error found with the LOO procedure can thus be attributed to the lack of information in the training set (composed of the other molecules), 2,2,4-trimethylpentane being the only tri-branched molecule in the all set and 2,2-dimethylbutane being the only other molecule containing a quaternary carbon atom. 2,2,4-Trimethylpentane on beta data will therefore be kept for the prediction study.

Concerning ZSM-22, the overall mean error (1.40) is larger than for all the other zeolites. This can be attributed to the fact that in gas phase, branched alkanes do not enter the micropores of zeolite ZSM-22. Indeed the pore mouths being the active adsorption sites, it induces a very low adsorption enthalpy for the bulkiest molecules such as 2,2-dimethylbutane and 2,2,4-trimethylpentane [18]. The much wider range of adsorption enthalpies measured on ZSM-22 as compared to the other zeolite (about twice as much) is also due to narrow pore width. The increase of adsorption enthalpy for linear compounds being around 12.8 kJ mol^{-1} per CH_2 group added to the molecule compared to around 6 kJ mol^{-1} per CH_2 group in the case of faujasites. The overall mean error of the LOO model has also to be compared to the experimental standard error estimated around 1.9 kJ mol^{-1} for ZSM-22. The LOO mean error is higher for this zeolite as compared to the other but the ratio LOO error/experimental error remains almost the same.

Preliminary analysis of the overall data set using leave-one-out cross-validation procedure shows that the QSAR approach based on graph machines can efficiently be used for calculating alkanes/zeolites zero coverage adsorption enthalpies. This indicates that this property can be estimated by only taking into account the structure of the molecule and the values of the adsorption enthalpy of other molecules on the same solid. In the following, in order to avoid unduly high bias, 2,2-dimethylbutane on ZSM-5 value was considered as an outlier and was then excluded from the set.

3.2. Prediction results and discussion

A second model was then build with a very limited training set of 10 selected molecules and applied to predict the properties of the other molecules forming the test set. The 10 molecules constituting the training set were taken among those which adsorption data are available and were chosen to be well separated in the molecular space to maximize training information.

Table 2
Correlation coefficient, max and mean error values between the LOO predicted adsorption enthalpy and the experimental data.

	Na-Y	Na-USY	Beta	Beta without 2,2,4-TMP	Mordenite	ZSM-5	ZSM-5 without 2,2-DMB	ZSM-22
Correlation coefficient	0.998	0.997	0.981	0.998	0.993	0.969	0.9998	0.978
Max. error (kJ mol^{-1})	0.93	0.78	5.99	2.22	2.63	8.04	0.43	6.42
Mean error (kJ mol^{-1})	0.17	0.27	0.62	0.24	0.48	0.62	0.10	1.40

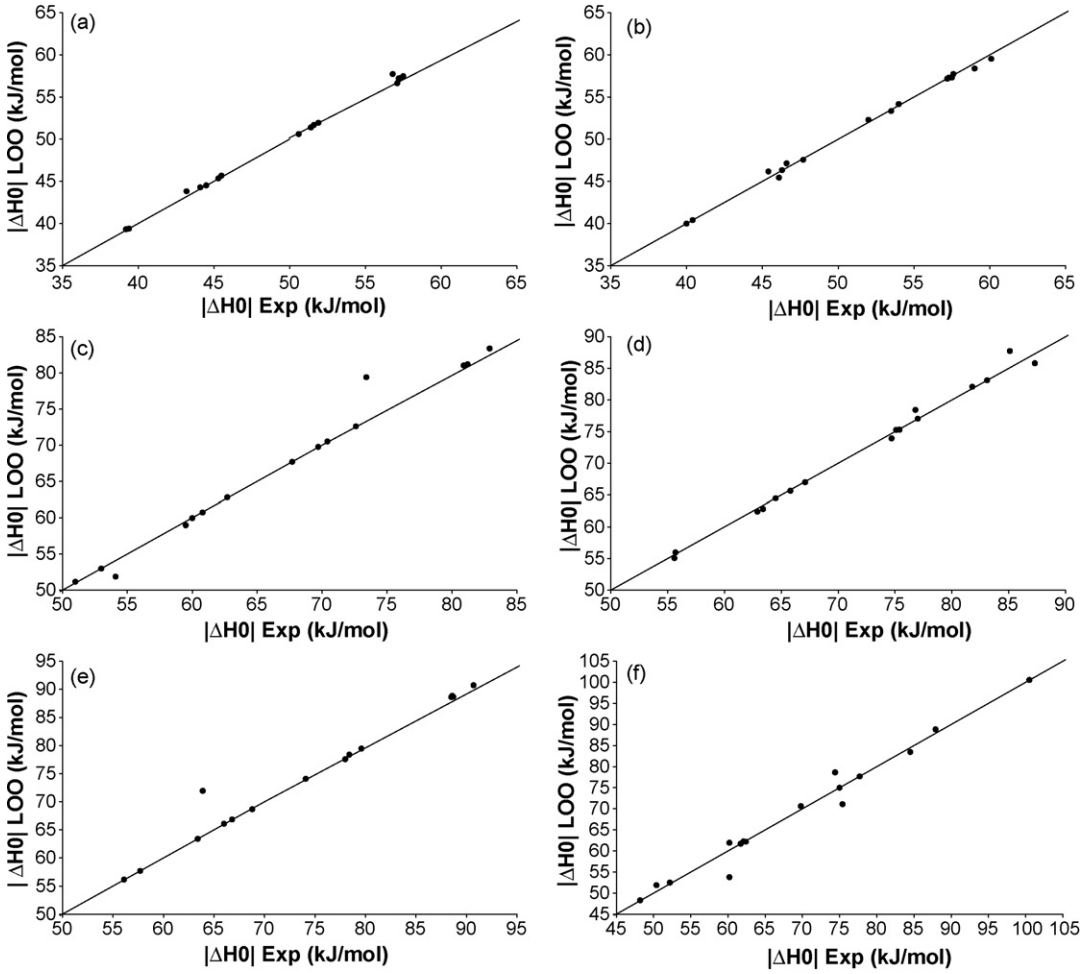


Fig. 3. Comparison between predicted and experimental zero coverage adsorption enthalpy data using leave-one-out cross-validation. (a) Na-Y, (b) Na-USY, (c) beta, (d) mordenite, (e) ZSM-5 and (f) ZSM-22.

For Na-Y, beta, mordenite and ZSM-22, the 10 molecules forming the test set are n-pentane, n-hexane, 2-methylpentane, 3-methylpentane, 2,3-dimethylbutane, n-heptane, 2-methylhexane, 2,3-dimethylpentane, 4-methylheptane and 2,2,4-trimethylpentane. The remaining molecules, namely 2-methylbutane, 2,2-dimethylbutane, 3-methylhexane, n-octane, 2-methylheptane, 3-methylheptane, 2,5-dimethylhexane compose the prediction set.

As the experimental adsorption enthalpy were not available for 2-methylhexane on Na-USY and for 2,2,4-trimethylpentane on ZSM-5, the training and prediction sets were re-arranged for these zeolites as indicated in Tables 4 and 5.

Training and prediction results are compiled in Tables 3–5 and are represented in Fig. 4. Not surprisingly, the mean error between calculated and experimental values for the training set is always lower than the mean error for the prediction set. Nevertheless the difference between mean errors of the training and of the test set remains low indicating the absence of overtraining with two hidden neurons in the model.

3.2.1. Na-Y, Na-USY

The alkanes and iso-alkanes studied can easily be accommodated in the supercage of these faujasites and the influence of the number of cations for the two zeolites considered is small. Adsorption enthalpies of the isomers are then found very close one to the other, the main contribution to the enthalpy being the total number of carbons contained in the molecule.

For both zeolites, considering the very low number of molecules in the training set, there is a very good agreement between experimental and calculated values both for the training and the test sets, the mean error value being below 1.2 kJ mol^{-1} compared to a mean experimental error of 0.3 kJ mol^{-1}. The error between experimental and calculated values is almost the same for linear and monobranched alkanes (between 0.6 kJ mol^{-1} and 0.8 kJ mol^{-1}). A higher absolute error between experimental and predicted values is observed for some multi-branched molecules. This last result is mainly attributed to the lack of iso-alkanes containing quaternary carbon molecules in the training set.

3.2.2. Beta, mordenite and ZSM-5

Contrarily to large pore zeolites such as Faujasites for which adsorption enthalpies of monobranched isomers are very close to the corresponding n-alkane, beta, mordenite and ZSM-5 exhibit a lower value for 2- and 3-methylalkanes as compared to the corresponding n-alkane attributed to steric hindrance. This behaviour is well represented by the model, the adsorption enthalpies of monobranched isomers being always lower than the adsorption enthalpy of the corresponding linear alkane.

In the case of zeolite beta, the compounds having a quaternary carbon atom have a very specific adsorption behaviour, these molecules being trapped in the channels intersections [2,19]. This can be seen on the experimental data where the bulkiest molecules studied (2,2-dimethylbutane and 2,2,4-trimethylpentane) exhibit a much lower adsorption enthalpy compared to the linear or

Table 3
Experimental [1] and predicted adsorption enthalpies and mean and max. error values of C5–C8 alkanes on Na-Y, beta, mordenite and ZSM-22 zeolites. The training set is composed of the first ten molecules, the others composing the test set.

	Na-Y			Beta			Mordenite			ZSM-22		
Molecule	$\lvert\Delta H0\rvert$ Exp (kJ mol^{-1})	$\lvert\Delta H0\rvert$ Predict. (kJ mol^{-1})	Error (kJ mol^{-1})	$\lvert\Delta H0\rvert$ Exp (kJ mol^{-1})	$\lvert\Delta H0\rvert$ Predict. (kJ mol^{-1})	Error (kJ mol^{-1})	$\lvert\Delta H0\rvert$ Exp (kJ mol^{-1})	$\lvert\Delta H0\rvert$ Predict. (kJ mol^{-1})	Error (kJ mol^{-1})	$\lvert\Delta H0\rvert$ Exp (kJ mol^{-1})	$\lvert\Delta H0\rvert$ Predict. (kJ mol^{-1})	Error (kJ mol^{-1})
n-Pentane	39.4 ± 0.35	38.10	1.30	53.0 ± 0.49	52.68	0.32	55.7 ± 0.64	55.69	0.01	62.1 ± 2.0	63.89	1.79
n-Hexane	45.5 ± 0.32	44.84	0.66	62.7 ± 0.46	62.36	0.34	67.1 ± 0.80	67.08	0.02	75.0 ± 1.3	75.12	0.12
2-Methylpentane	45.3 ± 0.23	44.44	0.86	60.8 ± 0.51	61.79	0.99	65.8 ± 0.82	68.33	2.53	62.4 ± 1.2	62.23	0.17
3-Methylpentane	44.5 ± 0.42	45.00	0.50	60.0 ± 0.49	59.68	0.32	64.5 ± 0.82	65.61	1.11	61.7 ± 2.1	61.66	0.04
2,3-Dimethylbutane	44.1 ± 0.28	44.83	0.73	59.5 ± 0.41	58.64	0.86	63.4 ± 0.72	64.92	1.52	52.2 ± 2.9	53.19	0.99
n-Heptane	51.9 ± 0.39	52.68	0.78	72.6 ± 0.78	72.89	0.29	77.0 ± 1.16	76.63	0.37	87.9 ± 0.7	88.83	0.93
2-Methylhexane	51.6 ± 0.31	52.47	0.87	70.4 ± 0.25	70.99	0.58	75.1 ± 1.41	75.10	0.00	75.4 ± 1.3	75.83	0.43
2,3-Dimethylpentane	50.6 ± 0.21	49.89	0.71	67.7 ± 0.53	66.75	0.95	74.7 ± 1.08	71.93	2.77	60.2 ± 2.9	60.19	0.01
4-Methylheptane	57.2 ± 0.40	57.89	0.69	81.1 ± 0.41	78.98	2.12	81.8 ± 1.37	80.18	1.63	77.7 ± 1.8	82.95	5.25
2,2,4-Trimethylpentane	56.8 ± 0.22	54.67	2.13	73.4 ± 0.54	71.60	1.80	76.8 ± 1.32	76.44	0.36	48.2 ± 3.1	49.40	1.20
2-Methylbutane	39.2 ± 0.24	38.77	0.43	51.0 ± 0.52	53.04	2.04	55.6 ± 0.67	57.15	1.55	50.4 ± 1.4	54.88	4.48
2,2-Dimethylbutane	43.2 ± 0.38	40.34	2.86	54.1 ± 0.41	55.68	1.58	62.9 ± 0.49	60.27	2.63	38.2 ± 3.1	41.65	3.45
3-Methylhexane	51.4 ± 0.23	51.48	0.08	69.7 ± 0.39	69.95	0.25	75.4 ± 1.24	72.70	2.70	69.8 ± 2.4	67.69	2.11
n-Octane	57.5 ± 0.23	57.39	0.11	82.9 ± 0.55	80.56	2.34	87.3 ± 1.48	85.57	1.73	100.5 ± 0.8	98.54	1.96
2-Methylheptane	57.2 ± 0.36	55.81	1.49	81.2 ± 0.73	75.39	5.81	85.1 ± 0.61	84.18	0.92	84.5 ± 1.6	87.21	2.71
3-methylheptane	57.3 ± 0.30	57.10	0.10	80.9 ± 0.62	77.28	3.62	83.1 ± 1.05	79.92	3.18	74.4 ± 1.7	81.87	7.47
2,5-Dimethylhexane	57.1 ± 0.28	55.42	1.68	–	74.06	–	–	–	–	60.2 ± 1.9	66.40	6.20
Mean error training set			0.92			0.86			1.03			1.09
Mean error test set			0.96			2.61			2.12			4.05
Overall mean error			0.94			1.51			1.44			2.31
Max. error			2.86			5.81			3.18			7.47
Mean error n-alkanes			0.71			0.82			0.53			1.20
Mean error monobranched-alkanes			0.63			1.97			1.70			2.83
ME di-branched alkanes			1.50			1.13			2.31			2.66

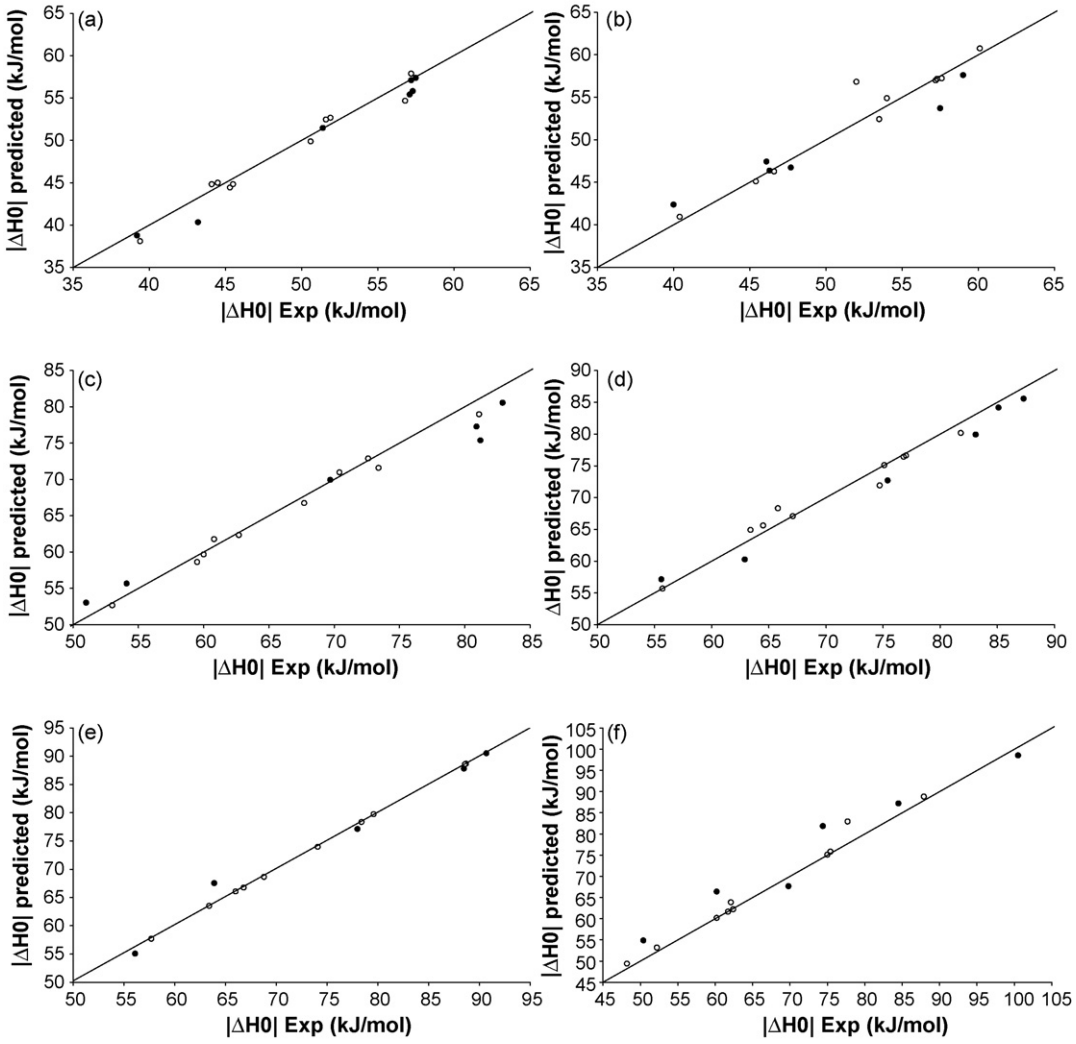


Fig. 4. Comparison between predicted and experimental zero coverage adsorption enthalpy data for both training (open symbols) and test (full symbols) sets. (a) Na-Y, (b) Na-USY, (c) beta, (d) mordenite, (e) ZSM-5 and (f) ZSM-22.

monobranched isomers or even 2,3-dimethyl compounds. This specific adsorption behaviour of compounds having a quaternary carbon is fairly well represented by the model. The errors between experimental and calculated values are 1.58 kJ mol^{-1} for 2,2-dimethylbutane and 1.80 kJ mol^{-1} for 2,2,4-trimethylpentane respectively. Despite the relative lack of information about this specific adsorption behaviour in the training set, only one quaternary alkane molecule being in this training set, the mean error value between experimental and calculated values for quaternary alkanes is lower than the mean error found for monobranched isomers (1.69 kJ mol^{-1} vs 1.97 kJ mol^{-1} respectively). The prediction can then be considered as good for all compounds including those having a quaternary carbon atom, the model predicting much lower values for these compounds compared to the other isomers.

For zeolite beta also, the presence of only two compounds with 8 carbon atoms in the training set, namely 4-methylheptane and 2,2,4-trimethylpentane, combined with the much lower adsorption enthalpy of the latter compared to the other C8 molecules have a clear influence on the determination of the adsorption enthalpies of C8 compounds in the prediction set. Indeed, as carbon number has a strong influence on adsorption enthalpy, the presence of only one C8 molecule beside 2,2,4-trimethylpentane in the training set causes a clear underestimation of the adsorption enthalpy of C8 compounds in the prediction set (n-octane, 2-methylheptane, 3-methylheptane and 2,5-dimethylhexane) which explains the relatively high error values between experimental and predicted enthalpies for the compounds (respectively 2.34 kJ mol^{-1} for n-octane, 5.81 kJ mol^{-1} for 2-methylheptane and 3.62 kJ mol^{-1} for 3-methylheptane). This underestimation was not encountered in the leave-one-out procedure which indicates that even if the results can be considered as good, the very low number of molecules in the training set causes a lack of accuracy for predictions.

3.2.3. ZSM-22

Contrarily to large pore zeolites such as faujasites or medium pore zeolites such as beta, mordenite or ZSM-5, adsorption enthalpies of branched isomers on ZSM-22 are much lower than the values of the corresponding n-alkanes. 2-Methyl-branched alkanes show about 12.5–16 kJ mol^{-1} lower values, whereas 3-methyl-branched alkanes have an adsorption enthalpy which is 13–26 kJ mol^{-1} lower than those with a linear chain. 2-Methyl-branched and the 3-methyl-branched isomers also exhibit differences in adsorption enthalpies (0.7 for methylpentanes, 5.6 for methylhexanes and 10.1 for methylheptanes respectively). All these features are due to the very narrow pore structure of the zeolite ZSM-22, which induces that the branched alkanes do not enter the micropores while the corresponding n-alkanes can enter the pores. Branched alkanes adsorb only on pore mouths which leads to

Table 4
Experimental [1] and predicted adsorption enthalpies and mean and max. error values of C5–C8 alkanes on Na-USY. The training set is composed of the first ten molecules, the others composing the test set.

Molecule	Na-USY		
	$\|\Delta H0\|$ Exp (kJ mol^{-1})	$\|\Delta H0\|$ Predict. (kJ mol^{-1})	Error (kJ mol^{-1})
n-Pentane	40.4 ± 0.42	40.95	0.55
2-Methylpentane	46.6 ± 0.54	46.28	0.32
2,2-Dimethylbutane	45.4 ± 0.46	45.12	0.28
n-Heptane	54.0 ± 0.44	54.90	0.90
3-Methylhexane	53.5 ± 0.33	52.43	1.07
2,3-Dimethylpentane	52.0 ± 0.27	56.83	4.83
n-Octane	60.1 ± 0.24	60.75	0.65
3-Methylheptane	57.3 ± 0.27	57.17	0.13
4-methylheptane	57.2 ± 0.29	56.99	0.21
2,5-Dimethylhexane	57.6 ± 0.42	57.24	0.36
2-Methylbutane	40.0 ± 0.41	42.38	2.38
n-Hexane	47.7 ± 0.35	46.74	0.96
3-Methylpentane	46.3 ± 0.25	46.38	0.08
2,3-Dimethylbutane	46.1 ± 0.53	47.44	1.34
2-Methylhexane	–	55.59	–
2-Methylheptane	59.0 ± 0.30	57.62	1.38
2,2,4-Trimethylpentane	57.5 ± 0.28	53.72	3.78
Mean error training set			0.93
Mean error test set			1.66
Overall mean error			1.20
Max. error			3.78
Mean error n-alkanes			0.61
Mean error monobranched-alkanes			0.80
Mean error di-branched alkanes			1.70

much lower adsorption enthalpies [18,20]. Very interestingly, the model can efficiently predict this specific behaviour, the linear, the 2-methyl and the 3-methyl isomers always being in the right order for the adsorption enthalpies, i.e. linear » 2-methyl ≥ 3-methyl.

When compared to other models established to predict the adsorption enthalpies of linear and branched alkanes on zeolite ZSM-22 [2,21,22] the new GM based QSAR model provide better

Table 5
Experimental [1] and predicted adsorption enthalpies and mean and max. error values of C5–C8 alkanes on ZSM-5. The training set is composed of the first ten molecules, the others composing the test set.

Molecule	ZSM-5		
	$\|\Delta H0\|$ Exp (kJ mol^{-1})	$\|\Delta H0\|$ Predict. (kJ mol^{-1})	Error (kJ mol^{-1})
n-Pentane	57.7 ± 0.42	57.69	0.01
n-Hexane	68.8 ± 0.26	68.65	0.15
2-Methylpentane	66.8 ± 0.22	66.75	0.05
3-Methylpentane	66.0 ± 0.18	66.08	0.08
2,3-Dimethylbutane	63.4 ± 1.49	63.51	0.11
n-Heptane	79.6 ± 0.26	79.76	0.16
2-Methylhexane	78.4 ± 0.44	78.37	0.03
2,3-Dimethylpentane	74.1 ± 0.96	73.99	0.11
4-Methylheptane	88.7 ± 0.45	88.69	0.01
2-Methylheptane	88.6 ± 0.34	88.60	0.00
2-Methylbutane	56.1 ± 1.4	55.1	1.03
2,2-Dimethylbutane	63.9 ± 1.22	67.6	3.65
3-Methylhexane	78.0 ± 1.25	77.1	0.88
n-Octane	90.7 ± 0.28	90.5	0.16
3-Methylheptane	88.5 ± 0.37	87.8	0.68
Mean error training set			0.07
Mean error test set			1.28
Overall mean error			0.48
Max. error			3.65
Mean error n-alkanes			0.12
Mean error monobranched-alkanes			0.35
Mean error di-branched alkanes			0.97

Table 6
Experimental [18,20] and predicted adsorption enthalpies and mean and max. error values of C5–C8 alkanes on ZSM-22. The training set is composed of the first ten molecules, the others composing the test set.

Molecule	ZSM-22		
	$\|\Delta H0\|$ Exp (kJ mol^{-1})	$\|\Delta H0\|$ Predict. (kJ mol^{-1})	Error (kJ mol^{-1})
n-Pentane	63.3 ± 1.6	63.34	0.04
n-Hexane	77.1 ± 0.9	77.07	0.03
n-Heptane	89.4 ± 0.7	89.36	0.04
n-Octane	100.6 ± 1.3	100.57	0.03
2-Methylbutane	49.3 ± 1.3	49.55	0.25
2-Methylpentane	60.0 ± 1.9	60.07	0.07
2-Methylhexane	76.1 ± 1.7	76.05	0.05
2-Methylheptane	87.3 ± 1.3	87.35	0.05
3-Methylpentane	59.4 ± 0.4	59.38	0.02
3-Methylhexane	72 ± 1.7	71.70	0.30
3-Methylheptane	84.5 ± 1.5	84.56	0.06
n-Nonane	112.5 ± 0.9	112.73	0.23
2-Methyloctane	97.8 ± 0.9	100.26	2.46
3-methlyoctane	97.8 ± 1.3	90.53	7.27
2-Methylnonane	107	108.03	1.03
Mean error training set			0.09
Mean error test set			2.75
Overall mean error			0.79
Max. error			7.27
Mean error n-alkanes			0.07
Mean error monobranched-alkanes			1.16

results. Especially, the new model has an overall mean error as low as 2.31 kJ mol^{-1} which is below the one of the two PLS models reported in [2] for this zeolite (the most efficient models reported so far), with an overall mean error of 2.76 kJ mol^{-1} and 3.38 kJ mol^{-1} respectively.

In order to confirm the prediction ability of the GM/QSAR model developed, prediction of adsorption enthalpies of some molecules out of the initial C5–C8 range were carried out. C9–C10 alkanes adsorption enthalpies on ZSM-22 were then predicted and compared to the experimental values reported by Ocakoglu et al. [18,20]. As those values originate from a different series of experiments sometimes exhibiting large difference in the experimental values (e.g. 84.5 kJ mol^{-1} instead of 74.4 kJ mol^{-1} for 3-methylheptane) the model was reset with the new C5–C8 experimental values indicated in these references. This difference in the experimental values is attributed to differences between the surface of the two zeolites considered in Ref. [1] and the one used in Refs. [20,22] respectively. This results in different adsorption enthalpies for the branched alkanes that adsorb on the pore mouth, the adsorption enthalpies remaining the same for the n-alkanes which enter the micropores. 11 molecules were taken for the training set (nC5, 2-methylbutane and the linear, 2-methyl- and 3-methyl-branched isomers from C6 to C8) and 4 molecules were considered out of this initial C5-C8 range as the prediction set: n-nonane, 2-methyloctane, 3-methyloctane and 2-methylnonane. Results are resumed in Table 6 and presented in Fig. 5.

The mean errors are as low as 0.09 kJ mol^{-1} for the training set and 2.75 kJ mol^{-1} for the prediction set. This clearly indicates the very good ability of the model to predict adsorption enthalpies out of the initial C5–C8 range used for the training. When looking into more details, predictions of adsorption properties of C9–C10 molecules are good except for 3-methyloctane for which the calculated value is much lower than the experimental one. This large underprediction can be attributed to the systematic difference between adsorption enthalpies in the training set for 2-methyl and 3-methyl alkanes that is not measured for 2- and 3-methyloctanes. Still, predictions which are extrapolations to higher adsorption enthalpies can be considered as very good and are better than previously reported model calculations [21,22].

Table 7
Comparison of the mean and max. error values of the different models: PLS#1 (topological descriptors) [1], PLS#2 (custom-made descriptors) [1] and graph machines (this study).

	Na-Y			Na-USY			Beta			Mordenite			ZSM-5			ZSM-22		
	PLS#1	PLS#2	GM	PLS#1	PLS#2	GM	PLS#1	PLS#2	GM	PLS#1	PLS#2	GM	PLS#1	PLS#2	GM	PLS #1	PLS#2	GM
Overall mean error	0.32	0.35	0.94	0.53	0.53	1.20	0.32	0.91	1.51	0.60	0.73	1.44	0.25	0.39	0.48	2.76	3.38	2.31
Max. error	1.24	0.92	2.86	2.30	1.30	3.78	0.73	2.25	5.81	1.75	2.63	3.18	0.68	0.81	3.65	16.32	6.93	7.47

When compared to the above reported values (see Table 3) for C5–C8 sets containing also multi-branched molecules, the error value obtained for the test set which was 4.05 kJ mol^{-1} is now reduced to 2.75 kJ mol^{-1} for the prediction set. This result is attributed to a greater homogeneity of the molecules present in the whole set, multi-branched molecules being absent from both training and prediction set. The training set is also slightly larger (11 molecules instead of 10).

3.3. Comparison of PLS and graph machine based models for the prediction of alkanes and iso-alkanes adsorption enthalpies on zeolites

The mean and max. error values of the different models: PLS#1 (topological descriptors) [2], PLS#2 (custom-made descriptors) [2] and graph machines (this study) are reported in Table 7.

From Table 7, it can be seen that the best results were obtained from the PLS approach for all studied zeolites except for the zeolite ZSM-22 for which the graph machines model gives a better prediction. Hence, for large and medium pore zeolites, for which the adsorption properties can be well represented by a linear model, the use of neural networks does not provide a better prediction. The PLS models can also provide insights on the adsorption mechanisms which cannot be provided by the GM neural networks models which are "blackbox models". Nevertheless the GM/QSAR approach reported in this study still provides good results for all zeolites, especially when considering the very low number of molecules in the training sets.

For zeolite ZSM-22 which exhibits a much more complex adsorption behaviour as compared to the other studied zeolites, the new non-linear model reported in this study provides better results indicating that this new model can better handle this complex behaviour. Moreover, one can see from Tables 3–5 that except for zeolite NaY for which mean error values are comparable for the training and the test sets, the mean error on the training set is below the one of the test set for all zeolites. This is especially true for zeolite ZSM-22 for which the values are 1.09 and 4.05 kJ mol^{-1} respectively. This indicates that in this case overtraining is probably occurring. The use of less than two neurons in the hidden layer in the base function of the graph machine resulting in bad prediction results, this indicates that for this zeolite the use of a slightly larger training set avoiding overtraining would certainly result in an more efficient GM model whereas the prediction ability of PLS models would not be greatly increased. This is confirmed by the leave-one-out cross-validation which indicates that the mean error can be reduced to a value as low as 1.40 kJ mol^{-1} when using 16 molecules in the training set.

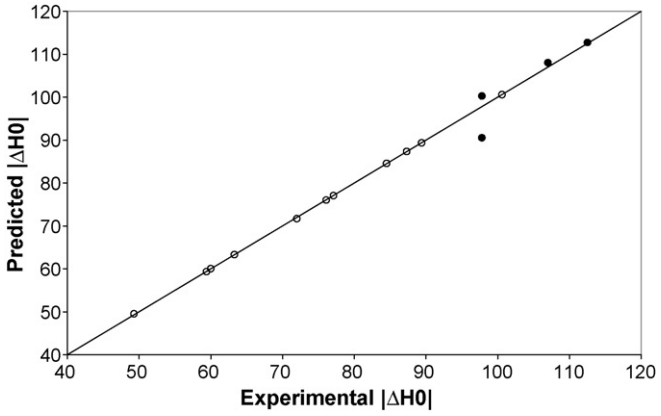


Fig. 5. Parity diagram of the zero coverage adsorption enthalpy of C5–C8 alkanes (training set – open symbols) and C9–C10 alkanes (prediction set – full symbols) on ZSM-22. Experimental data are taken from [20,22].

4. Conclusion

This study describes the use of neural network based models to predict the adsorption enthalpies of n- and iso-alkanes on various zeolites using a new approach called graph machines. In these models, the adsorbates were considered as structured data, represented and encoded as graphs. This GM/QSAR approach provided good results for all zeolites, especially when considering the very low number of molecules in the training set.

When compared to previously reported PLS models [2] on the same set of data, the new models resulted in slightly less efficient predictions for all zeolites except for ZSM-22 for which the graph machines model gives a better prediction. In particular, the specific behaviour of this zeolite is well predicted, the adsorption enthalpies of the different isomers (i.e. linear, 2-methyl and the 3-methyl) always being in the experimental order. The difference in the GM and PLS based models prediction efficiency can also be probably enhanced using a larger array of experimental data.

References

[1] J.F.M. Denayer, W. Souverijns, P.A. Jacobs, J.A. Martens, G.V. Baron, J. Phys. Chem. B 102 (1998) 4588–4597.
[2] P. Leflaive, G. Pirngruber, A. Faraj, P. Martin, G.V. Baron, J.F.M. Denayer, Micropor. Mesopor. Mater. 132 (1–2) (2010) 246–257.
[3] C. Brasquet, P. Le Cloirec, Wat. Res. 33 (1999) 3603–3608.
[4] C. Brasquet, B. Bourges, P. Le Cloirec, Environ. Sci. Technol. 33 (1999) 4226–4231.
[5] S. Giraudet, P. Pré, H. Tezel, P. Le Cloirec, Carbon 44 (2006) 1413–2421.
[6] S. Timofei, L. Kurunczi, T. Suzuki, W.M.F. Fabian, S. Muresan, Dyes Pigments 34 (3) (1997) 181–193.
[7] S.K. Jha, G. Madras, Ind. Eng. Chem. Res. 44 (2005) 7038–7041.
[8] N.V.K. Dutt, S.J. Kulkarni, Y.V.L. Ravikumar, B.S.N. Murthy, J. Chem. Sci. 118 (4) (2006) 345–349.
[9] A. Goulon-Sigwalt-Abram, A. Duprat, D. Dreyfus, Theor. Comput. Sci. 344 (2005) 298–344.
[10] A. Goulon, A. Duprat, D. Dreyfus, Proceedings of the Applied Stochastic Models and Data Analysis (ASMDA 2005) Conference, 2005, available on the web from http://conferences.telecom-bretagne.eu/asmda2005/IMG/pdf/proceedings/552.pdf.
[11] A. Goulon, A. Duprat, D. Dreyfus, Lect. Notes Theor. Comput. Sci. 4135 (2006) 1–19.
[12] A. Goulon, T. Picot, A. Duprat, D. Dreyfus, SAR QSAR Environ. Res. 18 (2007) 141–153.
[13] C. Jochum, J. Gasteiger, J. Chem. Inf. Comput. Sci. 17 (1977) 113–117.
[14] K.L. Priddy, P.E. Keller (Eds.), Artificial Neural Networks: An Introduction, SPIE Press, 2005.
[15] G. Rothenberg, Catal. Today 137 (2008) 2–10.
[16] F. Eder, J.A. Lercher, J. Phys. Chem. B 101 (1997) 1273–1278.

[17] C.L. Cavalcante, D.M. Ruthven, Ind. Eng. Chem. Res. 34 (1995) 185–191.
[18] J.F. Denayer, A.R. Ocakoglu, W. Huybrechts, J.A. Martens, J.W. Thybaut, G.B. Marin, G.V. Baron, Chem. Commun. (2003) 1880–1881.
[19] P.S. Bárcia, J.A.C. Silva, A.E. Rodrigues, Micropor. Mesopor. Mater. 79 (2005) 145–163.
[20] R.A. Ocakoglu, J.F.M. Denayer, G.B. Marin, J.A. Martens, G.V. Baron, J. Phys. Chem. B 107 (2003) 398–406.
[21] C.S.L. Narasimham, J.W. Thybaut, G.B. Marin, J.A. Martens, J.F.M. Denayer, G.V. Baron, J. Catal. 218 (2003) 135–147.
[22] T.L.M. Maesen, M. Schenk, T.J.H. Vlugt, J.P. de Jonge, B. Smit, J. Catal. 188 (1999) 403–412.