Graph Machine Based-QSAR Approach for Modeling Thermodynamic Properties of Amines: Application to CO2 Capture in Postcombustion.

Fabien Porcheron, Marc Jacquin, Nabil El Hadri, Diego Saldana-Miranda, Aurélie Goulon, Abdelaziz Faraj

Oil & Gas Science and Technology – Rev. IFP Energies nouvelles, Vol. 68 (2013), No. 3, pp. 469-486

DOI: 10.2516/ogst/2012025

Dossier

Discovery and Optimization of Catalysts and Adsorbents Using High Throughput Experimentation
Découverte et optimisation de catalyseurs et d'adsorbants par expérimentations haut débit

Graph Machine Based-QSAR Approach for Modeling Thermodynamic Properties of Amines: Application to CO_2 Capture in Postcombustion

F. Porcheron*, M. Jacquin, N. El Hadri, D.A. Saldana, A. Goulon and A. Faraj

IFP Energies nouvelles, Rond-point de l'échangeur de Solaize, BP 3, 69360 Solaize - France
e-mail: fabien.porcheron@ifpen.fr - marc.jacquin@ifpen.fr - nabil.el-hadri@ifpen.fr - diego.saldana-miranda@ifpen.fr
aurelie.goulon@ifpen.fr - abdelaziz.faraj@ifpen.fr

* Corresponding author

Résumé — Approche QSAR *Graph Machines* pour la modélisation des propriétés thermodynamiques des amines : application au captage du CO_2 en postcombustion — Le procédé d'absorption aux amines est considéré comme la technologie la plus efficace pour limiter les rejets de CO_2 dans le cadre du captage en postcombustion puis du stockage du CO_2. Cependant, l'optimisation des propriétés du solvant nécessite d'évaluer un grand nombre de candidats potentiels et donc de collecter une quantité importante de propriétés expérimentales. Dans ce contexte, l'utilisation de méthodes de modélisation statistique de type QSAR (*Quantitative Structure Activity Relationship*) s'avère être un outil très précieux puisqu'elles permettent d'établir une relation entre un ensemble de vecteurs d'entrées (*i.e.* les caractéristiques ou les propriétés des molécules étudiées) et un ensemble de vecteurs de sorties (*i.e.* les propriétés ciblées). Dans ce travail, nous avons utilisé un équipement d'expérimentation à haut débit pour mesurer la solubilité du CO_2 dans un ensemble de 46 solutions aqueuses d'amines. Les isothermes d'absorption sont modélisées en utilisant une approche thermodynamique basée sur l'évaluation de deux constantes d'équilibres, pKa^* et pKc^* caractéristiques des principales réactions chimiques intervenant dans la phase liquide. Nous avons ensuite utilisé une approche statistique baptisée *graph machines* à la fois pour classifier les molécules et modéliser la variation de la constante d'acidité pKa^* en fonction de la structure moléculaire. L'originalité de notre approche réside dans l'utilisation des graphes associés aux molécules afin de les représenter dans des espaces multidimensionnels et construire, en même temps, un modèle prédictif de leurs propriétés physico-chimiques. Cette approche est appliquée dans cet article pour prédire les propriétés thermodynamiques d'un ensemble de 5 nouvelles molécules.

Abstract* — *Graph Machine Based-QSAR Approach for Modeling Thermodynamic Properties of Amines: Application to CO_2 Capture in Postcombustion — *Amine scrubbing is usually considered as the most efficient technology for CO_2 mitigation through postcombustion Carbon Capture and Storage (CCS). However, optimization of the amine structure to improve the solvent properties requires to sample a large number of possible candidates and hence to gather a large amount of experimental data. In this context, the use of QSAR (Quantitative Structure Activity Relationship) statistical modeling is a powerful tool as it 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 targeted properties). In this work, we used a high throughput screening experimental device to measure CO_2 solubility data on a set of 46 amine aqueous solutions. Absorption isotherms are represented using a thermodynamic model based on two thermodynamic constants, pKa^* and pKc^*, accounting for the main chemical reactions occurring in the liquid phase between amine and CO_2. Then, we used a statistical approach named Graph Machines at the*

same time to cluster the molecules and to model the variation of the acidity constant pKa as a function of the molecular structure. The originality of our approach is the use of graphs to represent molecules in multidimensional spaces and simultaneously construct predictive models of their physicochemical properties based on these graphs. This approach is applied in this paper to predict the thermodynamic properties of a set of 5 new molecules.*

INTRODUCTION

The control of CO_2 emissions to the atmosphere has become a worldwide issue over the last few years as a direct correlation between greenhouse gas emissions and climate change is now commonly accepted. An important amount of carbon dioxide is generated by coal-fired power stations where the flue gas at atmospheric pressure is predominantly composed of N_2 (around 90%) with a small fraction of CO_2 (around 10%). Although some controversy has arisen in recent literature [1], postcombustion Carbon Capture and Storage (CCS) technology is one of the solution considered on a short-term schedule as it does not require deep modifications of existing power stations [2]. In amine scrubbing plants, the flue gas is usually contacted with an aqueous amine solution within an absorption tower (or absorber) at temperatures around $T = 313$ K. The solvent selectively captures CO_2 molecules thereby yielding a targeted removal (usually 90%) of carbon dioxide contained in the gas stream. At the bottom of the absorber the rich solvent is directed towards a regeneration column (or stripper) at higher temperature around $T = 393$ K, where water reflux is used to strip the CO_2 from the liquid solution. The lean absorbent is then cycled back to the absorber while carbon dioxide is pressurized prior to its transport and storage. The benchmark amine is MonoEthanolAmine (MEA), a primary amine that displays a high reactivity towards CO_2 absorption even at low partial pressure. However, the 30 mass% MEA process usually suffers from high energy requirement, corrosion and degradation [3]. To evaluate the potential of new absorbents for CCS, one has to characterize for each candidate molecule an extensive list of properties like the thermodynamic and kinetics of absorption in aqueous solution, the rate of degradation in the process or the toxicity.

Thermodynamic of absorption remains a primary criterion for estimating the potential of a novel absorbent for carbon dioxide capture. This property is mostly characterized from CO_2 absorption isotherms (or Vapor Liquid Equilibrium, VLE) where carbon dioxide equilibrium partial pressure (P_{CO_2}) is computed as a function of solvent loading α (number of moles of CO_2 per mole of amine in the liquid phase). The measurement of absorption isotherms enables the calculation of the rich loading (α_{rich}) characterizing the overall absorption capacity of the solvent or the cyclic capacity ($\Delta\alpha$) *i.e.* the loading difference between the rich and the lean solvent (α_{lean}) at the top of the absorber. These properties will deeply impact the performance of the solvent in term of energy requirement (*i.e.* reboiler heat duty) for regenerating the solution in the stripper [4].

Therefore, many works have focused on measuring absorption isotherms or calorimetric properties of CO_2 absorption in aqueous amine solutions [5-16]. In addition, numerical thermodynamic models have also been developed to calculate resulting species partitions in the solution and to model the experimental data [17-20]. More recently, systematic screenings of amine properties have appeared in the literature [21-23]. In a recent work [24], we performed a thermodynamic screening of mono-amines using a High Throughput Screening (HTS) device which was designed to measure CO_2 absorption isotherms in aqueous amine solutions. This kind of device generates enough experimental data to establish a Quantitative Structure Activity Relationship (QSAR) and thus to optimize the molecular structure for a specific targeted activity.

QSAR methods are based on the principle that the physicochemical properties (or activities) of molecules depend strongly on the structure thereof. These methods rely mostly on the decomposition of the molecular structure into molecular descriptors [25, 26], usually generated by molecular modeling techniques [27-29]. Then, statistical learning techniques are used to identify the best mathematical function which for all the molecules would link their set of molecular descriptors (represented by an input vector) to their properties. A potential criticism of these methods is that these models are not directly related to the molecular structure but are based on new variables, molecular descriptors, which are in fact vectorial representations of the structure.

Graphs are a mode of representation increasingly used to directly take into account the complex structure of such data. One major advantage of graphs, is that they describe in an adequate formalism the objects and relationships between objects. This is the case in several areas, such as bioinformatics, molecular chemistry, social network analysis or spatial or textual data processing [30-33], where the data analyzed are in the form of complex structures (social networks, arrangements of atoms, spatial contiguity relationship, grammatical construction of sentences, etc.). Graph are especially suited when processes input features are molecules for which one seeks to predict the physicochemical properties, by building models from experimental data.

Recently, Goulon *et al.* [34-36] showed that the structure of molecules can directly and efficiently be used for QSAR modeling. In this approach, called Graph Machines (GM), the molecules are considered as structured data and are represented by graphs. For each molecule, a mathematical function (Graph Machine) is built, which structure reflects the one of the molecule under consideration. Statistical learning

methods are then applied to estimate the parameters of such functions.

This approach was successfully applied to model boiling point or toxicity of organic molecules [35], anti-HIV activities [36] and more recently to model the adsorption enthalpy of alkanes in zeolites [37]. GM are especially efficient when linear models such as Partial Least Square regression (PLS) or Multiple Linear Regression (MLR) fail to correctly establish a relationship between the structure of the molecule and the response of the system.

GM and QSAR are intrinsically different from each other by the fact that GM is a structural approach while QSAR is a statistical one. Objects, in QSAR, are formally presented as points in a multi-dimensional space, where the dimensions are constituted by the molecular descriptors. Hence algebraic properties – like sum, product and distance of vectors – can be naturally applied. The drawback of graphs arising from the lack of a possible representation of patterns in an algebraic space can however be overcome by graph-based representations [38]. Many approaches dealing with the problem of definition of distance between graphs are developed, for example, for the purpose of clustering a document collection into categories for handling document queries [39], or to find the most similar chemical product to a molecule exhibiting a certain activity in large databases [40].

In this work, CO_2 solubility measurements are performed on a set of 46 mono-amines using a 6-reactors High Throughput Screening (HTS) experimental device. Automated CO_2 injections are performed within each reactor and we compute the resulting transient pressure curves in order to obtain absorption isotherms. In addition, a simplified thermodynamic model with two adjustable parameters (pKa^* and pKc^* that account for thermodynamic constants of amine acidity and carbamate stability, respectively) is used to represent the behavior of CO_2 absorption within aqueous solutions of mono-amines. Then, we propose an original representation of graphs in multi-dimensional algebraic space based on GM method. This representation is then used for the clustering of amine molecules. GM are then applied on a set of 40 of the experimented molecules (training set) to build statistical models of pKa^* variations as a function of the molecular structure. The most efficient model is chosen using a validation set containing the 6 remaining molecules. The resulting model is then applied to predict pKa^* values of 5 new mono-amines (prediction set) and compared to the values obtained experimentally afterward, to check the prediction ability of the model.

1 EXPERIMENTAL SECTION

1.1 Materials

Amines were purchased from *Sigma-Aldrich* with the highest purity available (*i.e.* > 97%). Samples of amine at 30 mass% in aqueous solutions were subsequently prepared using deionized water. The list of amine molecules screened in this work is reported in Table 1, together with their corresponding SMILES (Simplified Molecular Input Line Entry Specification) which provides a description of the graph structure of the molecule as a character string [41, 42] and their corresponding set (training, validation, prediction) for the statistical modeling.

1.2 High Throughput Screening Apparatus

The experimental apparatus used for measuring CO_2 solubility is shown in Figure 1 and schematically represented in Figure 2. It consists of six stirred cell reactors designed to operate at pressures ranging from vacuum up to 10^3 kPa and at temperatures up to 393.15 K. Each reactor can be operated independently and at different temperatures.

A regulating device comprising a heating resistance surrounding the lower part of the reactor and a compressed air injection system to cool down the reactor is used to keep the temperature constant with variations of ± 0.5 K. Temperatures sensors are placed on the upper and lower part of the reactor, to record temperature variations within the gas (T_G) and liquid (T_L) phase. A *Keller* PAA35XHTT pressure sensor is used to detect pressure variations in the gas phase (P_G). The pressure sensor uncertainty is estimated to 0.3 kPa. Pressures and temperatures are recorded every 200 ms on a computer.

Prior to each experiment, vacuum is made in the reactors and the pressure drift should not exceed 0.03 kPa/h. The solvent is then introduced inside the reactor using a syringe. The density of the solution at the experimental temperature is determined using an *Anton Paar* DMA4100 densimeter and the weight loss of the syringe is measured to determine

Figure 1

High throughput screening experimental device used to measure CO_2 solubility in aqueous amine solutions.

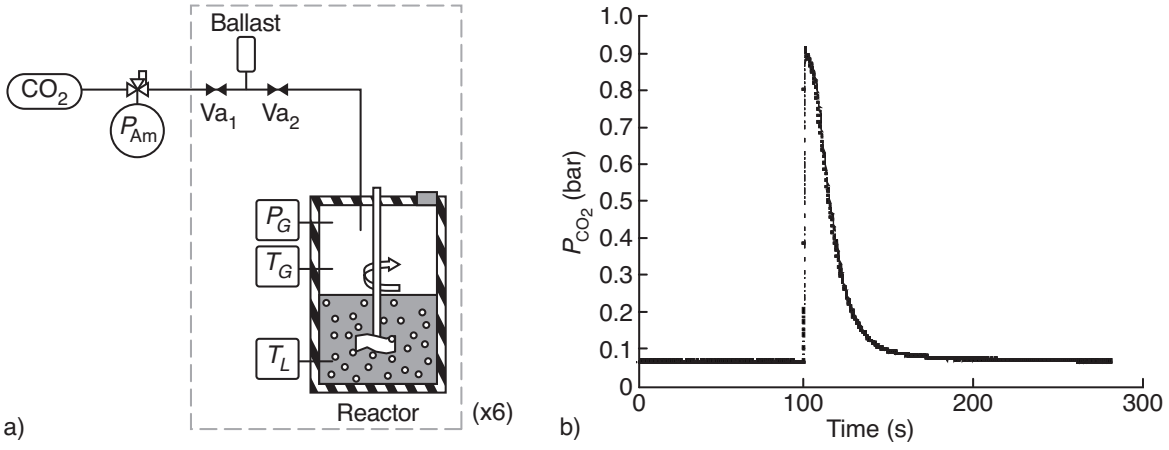

Figure 2

a) Schematic representation of the HTS device and b) transient pressure curve following a CO_2 injection obtained during a HTS experiment.

TABLE 1

List of mono-amines screened in this work

Set	Name	Smiles
Training	1-(dimethylamino)-2-propanol	CN(C)CC(C)O
Training	1-diethylamino-3-butanone	CC(=O)CCN(CC)CC
Training	1-methyl-2-pyrrolidine	O=C1CCCN1C
Training	1-propyl-4-piperidone	CCCN1CCC(=O)CC1
Training	2-(ethylamino)-ethanol	CCNCCO
Training	2,6-dimethylmorpholine	CC1CNCC(C)O1
Training	2-amino-1-butanol	CCC(N)CO
Training	2-amino-1-propanol	CC(N)CO
Training	2-amino-2-methyl-1,3-propanediol	CC(N)(CO)CO
Training	2-ethyl-2-oxazoline	CC\C1=N\CCO1
Training	2-hydroxymethyl-N-methylpiperidine	OCC1CCCCN1C
Training	2-methoxyethyl-amine	NCCOC
Training	2-methylamino-ethanol	CNCCO
Training	2-morpholino-ethylamine	NCCN1CCOCC1
Training	3-amino-1-propanol	NCCCO
Training	3-amino-propionitrile	N#CCCN
Training	3-dimethylamino-propionitrile	CN(C)CCC#N
Training	3-hydroxy-methylpiperidine	OCC1CCCNC1
Training	3-hydroxypiperidine	OC1CCCNC1
Training	3-methoxypropylamine	NCCCOC
Training	3-morpholino-1,2-propanediol	OCC(O)CN1CCOCC1
Training	4-amino-1-butanol	NCCCCO
Training	4-hydroxy-N-methylpiperidine	OC1CCN(C)CC1
Training	4-hydroxypiperidine	C1C(O)CCNC1
Training	2-amino-2-methylpropanol	NC(CO)(C)C
Training	2-(butylamino)ethanol	CCCCNCCO
Training	Bis-(2-methoxyethyl)-amine	COCCNCCOC
Training	Diethanolamine (DEA)	OCCNCCO
Training	Diisopropanolamine	OC(CNCC(O)C)C
Training	Methyldiethanolamine (MDEA)	OCCN(CCO)C
Training	Methylaminoacetaldehyde dimethylacetal	CNCC(OC)OC
Training	Morpholine	C1CNCCO1
Training	N-tertbutyl-diethanolamine	CC(C)(C)N(CCO)CCO
Training	Sarcosine	CNCC(=O)O
Training	Hydroxyethylpiperazine	OCCN1CCNCC1
Training	3-dimethylamino-1-propanol	N(CCCO)(C)C
Training	N-methylmorpholine	C1OCCN(C)C1
Training	2-(2-dimethylaminoethoxy)ethanol	CN(C)CCOCCO
Training	4-ethylmorpholine	C1OCCN(CC)C1
Training	3-dimethylamino-1,2-propanediol	N(CC(CO)O)(C)C
Validation	6-amino-1-hexanol	NCCCCCCO
Validation	1-amino-2-propanol	CC(O)CN
Validation	2-amino-(2-hydroxymethyl)-1,3-propanediol	OCC(N)(CO)CO
Validation	Monoethanolamine (MEA)	NCCO
Validation	Trans-4-aminocyclohexanol	NC1CCC(O)CC1
Validation	Triethanolamine	N(CCO)(CCO)CCO
Prediction	2-(dimethylamino)-ethanol	CN(C)CCO
Prediction	N-ethyldiethanolamine	CCN(CCO)CCO
Prediction	2-(2-aminoethoxy)ethanol	NCCOCCO
Prediction	3-amino-1,2-propanediol	NCC(O)CO
Prediction	2-pyrrolidinone	O=C1CCCN1

the volume of solvent introduced in the reactor (V_L). Knowing the total reactor volume (V_R) one can easily determine the volume available to the gas phase (V_G). Usually, about half of the reactor volume is filled with the solvent. Stirring of the solution is done by gas-inducing agitators. CO_2 is pumped from the gas phase to the liquid phase where it is dispersed in the solution through a set of perforations punched in the blades of the agitator. Using this kind of device allows minimizing resistance to mass transfer in the gas phase.

CO_2 injections within the reactors are then performed using fixed volume ballast connected to the tubing upstream from the reactor. Each ballast, of a known volume (V_B), is surrounded by two pneumatic valves Va_1 and Va_2. A CO_2 feed tank is connected to the HTS device and a gas regulator is used to impose a constant pressure (P_B) at the outlet of the tank. The carbon dioxide is then feed to the ballasts by opening the corresponding Va_1 valves. Closing Va_1, the ballast volume is now filled with CO_2 at a pressure P_B. Upstream from the reactor, no temperature regulation device was considered in this apparatus so the ballasts are at room temperature T_R, which is recorded every second using a temperature sensor.

The second valve Va_2 is then opened and closed rapidly, leading to the injection of carbon dioxide within the reactor. The time t_{inj} between opening and closing Va_2 is set by the user. Following the injection, the pressure increases sharply up to a maximum before a continuous decrease corresponding to the absorption of CO_2 by the liquid solution *(Fig. 2b)*. In order to obtain the best resolution of the maximum transient pressure, the stirring of the solution is stopped a few seconds (t_{stop} s) before opening Va_2. Following the injection, the stirring is then started again a few seconds (t_{start} s) after closing Va_2.

1.3 High Throughput Screening Procedure

We use a S7-300 *Siemens* automaton which can be sequentially programmed to control different thermodynamic conditions for each reactor. The program starts with the solvent loaded into the reactor after vacuum has been made. The reactor is then at room temperature T_R and the pressure corresponds to the vapor pressure P_{vap} of the liquid solution at T_R. The automaton now proceeds to reach the targeted temperature (T) at which the isotherm will be measured, inducing an increase of the bubble pressure of the solvent up to P_0. During the course of an experiment, the gas and liquid phase temperatures may slightly differ in a reactor due to the heat of reaction. However, this difference is always lower than the estimated temperature uncertainty, hence we consider that $T_G = T_L = T$.

CO_2 injections are now performed within the reactor. Prior to the experiment, the user defines a number of N_s equilibrium total pressure steps (P_i, $i = 1, N_s$) to be reached by the system. Following the first injection, the solution absorbs CO_2 for t_1 s after which the gas phase transient pressure is P_G. If $P_G < P_i$, the system immediately proceeds to another injection in the reactor and the same process is repeated. If $P_G > P_i$, the relaxation of the system is pursued for another t_2 s after which a new comparison of the actual pressure P_G with the target pressure P_i is made. If $P_G > 0.9P_i$ the system is considered to have reached the targeted pressure, if not the automaton proceeds to another injection and the algorithm is repeated until the pressure reaches the targeted pressure. The algorithm cycles through the different pressure steps defined by the user. Once the final step is reached, the reactor is cooled down to room temperature and the program is completed.

1.4 Computation of the Absorption Isotherm

A transient pressure curve $P_G = f(t)$ is then obtained at the end of the experiment and can be transformed into an absorption isotherm curve following mass and volume balance calculations within the system.

The most straightforward way to transform the $P_G = f(t)$ curve into an absorption isotherm is to perform a follow-up of the gas phase transient pressure in each reactor. Following a CO_2 injection, the difference between the transient pressure before the injection and the maximum transient pressure (P_P) reached in the gas phase allows to calculate the number of CO_2 moles injected within the reactor. Once the system reaches the thermodynamic equilibrium state, it becomes straightforward to add up the number of carbon dioxide moles absorbed in the solution from all the gas injections and to calculate the resulting solvent loading. The absorption isotherm can be subsequently calculated using this methodology. However, this procedure might encounter severe limitations for highly reactive systems like primary amines, *e.g.* MEA. Indeed, if the kinetics of absorption is faster than the time of injection, a part of the CO_2 volume may have already been transferred within the solvent before the maximum pressure is reached by the system. The value of P_P is then underestimated, inducing an error in the mass balance calculation. In order to overcome this problem, we used a methodology modeling the CO_2 injection process by considering mass balance calculations on the ballast (synthetic method).

For this procedure, we need to estimate the resulting CO_2 pressure in the cell if the liquid solution was non-absorptive. We focus on the evolution of the system before the injection and right after the opening of the valve Va_2, where the pressure reaches its maximum in the reactor. During the course of an experiment, prior to an injection, the system is in a thermodynamic state where (P_B, V_B, $T_B = T_R$) for the ballast and (P_G, V_G, T) for the gas phase in the reactor. In our experiments, the pressure in the ballast is set to P_B = 550 kPa, so the CO_2 gas phase can be modeled as ideal. The number of CO_2 moles contained in the ballast ($n^B_{CO_2}$) is then equal to:

$$n^R_{CO_2} = \frac{P_B V_B}{R T_B} \tag{1}$$

and in the reactor, the number of moles of (CO_2 + solvent) in the gas phase n_G^R is given by:

$$n_G^R = \frac{P_G V_G}{RT} \tag{2}$$

with R the ideal gas constant. When we mix these two systems by opening Va_2 and assuming an absence of CO_2 consumption by chemical reaction or physical solubility, the theoretical total pressure P_P reached in the gas phase of the reactor is deduced from:

$$P_P = \frac{\left(n_{CO_2}^B + n_G^R\right)RT}{V_B + V_G} = \frac{P_B V_B + P_G V_G \dfrac{T_B}{T}}{V_B + V_G} \frac{T}{T_B} \tag{3}$$

This theoretical transient total gas pressure would be the one reached by the system if the liquid was non-absorptive. This calculation can be repeated for every injection performed in each reactor during an experiment.

Starting from a thermodynamic equilibrium in the reactor (P_i), the CO_2 equilibrium partial pressure ($P_{CO_2}(i)$) is expressed as:

$$P_{CO_2}(i) = P_i - P_0 \tag{4}$$

The system then automatically operates N_{i+1}^j injections before reaching the next equilibrium state (P_{i+1}). Following an injection j within this sequence, the actual total pressure P_G in the reactor sharply increases up to P_P. The amount of CO_2 introduced in the reactor, $n_{CO_2}^{R,j}$, is then:

$$n_{CO_2}^{R,j} = \frac{(P_P - P_G)V_G}{RT} \tag{5}$$

After N_{i+1}^j injections, the total amount of CO_2 injected in the reactor is then:

$$n_{CO_2}^T = \sum_{j=1}^{N_{i+1}^j} n_{CO_2}^{R,j} \tag{6}$$

The system then reaches the next thermodynamic equilibrium state where the number of moles of CO_2 in the gas phase of the reactor is expressed as:

$$n_{CO_2}^R = \frac{(P_{i+1} - P_0)V_G}{RT} = \frac{P_{CO_2}(i+1)V_G}{RT} \tag{7}$$

In this work, we assume that the activity of the solvent (*i.e.* water) does not change with the loading so the vapor pressure of the solvent, P_0, is considered to be independent of the CO_2 injections and therefore constant. We checked that this approximation does not strongly impact the absorption isotherms in the concentration and temperature range sampled in this work. The incremental amount of CO_2 transferred in the solvent (n_L^{i+1}) is then deduced from:

$$n_L^{i+1} = n_{CO_2}^T - n_{CO_2}^R \tag{8}$$

and the loading of the solvent α_{i+1}:

$$\alpha_{i+1} = \alpha_i + \frac{n_L^{i+1}}{n_S} \tag{9}$$

with n_S the number of moles of amine introduced in the reactor. The absorption isotherm ($P_{CO_2}(i) = f(\alpha_i)$) can be subsequently calculated using this methodology.

2 THERMODYNAMIC MODELING

As carbon dioxide is a solute in the aqueous solution of amine, its partial pressure is expressed by the product of the Henry constant (H) and the molality of CO_2 (m_{CO_2}) in the solution:

$$P_{CO_2} = H m_{CO_2} \tag{10}$$

The molality of CO_2 in the liquid phase requires the computation of the speciation, which is calculated using mass action laws associated to the following system of chemical equilibrium:

$$2H_2O \leftrightarrow HO^- + H_3O^+ \tag{11}$$

$$CO_2 + 2H_2O \leftrightarrow HCO_3^- + H_3O^+ \tag{12}$$

$$HCO_3^- + H_2O \leftrightarrow CO_3^{2-} + H_3O^+ \tag{13}$$

$$R_1R_2R_3NH^+ + H_2O \leftrightarrow R_1R_2R_3N + H_3O^+ \tag{14}$$

where $R_1R_2R_3N$ is the amine chemical formula. If R_3 = H, another chemical equilibrium has to be considered, due to carbamate formation:

$$R_1R_2R_3NCOO^- + H_2O \leftrightarrow R_1R_2NH + HCO_3^- \tag{15}$$

We assume that the molar fraction of water in the liquid phase does not vary with the CO_2 absorption and we have yet 8 molar compositions to determine. To obtain these values, we solve numerically:

- the 5 mass action laws corresponding to each chemical reaction;
- the 2 mass balances which are related to amine and carbon dioxide respectively;
- the electro-neutrality of the solution.

Taking into account physical solubility of CO_2 and chemical reactions with amine as proposed, is not sufficient to obtain a good representation of the experimental data and non ideality of the solution has to be implemented. In this model, we assume that the activity of water is equal to its molar fraction and the activity coefficients of molecular solutes (amine and carbon dioxide) are set to unity. However, the activities of the ionic species are given by the extended Pitzer Debye – Hückel approach:

$$\log[\gamma] = \frac{-Az_i\sqrt{I}}{1 + BrI} + CI \tag{16}$$

where I is the ionic strength and z_i is the ionic charge of the solute. A and B are two constants that account for solvent effects (*i.e.* water) and depend only of temperature. The parameter r is the closer approach diameter and C is an empirical parameter.

The Henry constant of CO_2 into the solvent ($H_{solvent}$) is deduced from the one in water (H_{water}) by a Sechenov approach:

$$\log\left(\frac{H_{solvent}}{H_{water}}\right) = k_1 I + k_2 m_{CO_2} \quad (17)$$

The expressions of equilibrium constants corresponding to the chemical equilibrium (11) to (13) and the Debye-Hückel parameters as a function of the temperature can be found in other work [43]. Thus, the thermodynamic model used in this work remains with six adjustable parameters: C and r for the activity model, k_1 and k_2 for the Sechenov approach, and the thermodynamic constant Ka and Kc corresponding to the chemical equilibrium (14) and (15) respectively. Then, values of C, r, k_1 and k_2 have been regressed on several isotherms of classical alkanolamine such as MEA, DEA and MDEA, having known values of Ka and Kc. These four parameters are then set to the regressed values for all the others amines studied in this work. Finally, the model remains with two adjustable parameters for a new amine: Ka^* and Kc^*.

These two adjustable parameters are not rigorously the equilibrium constants Ka and Kc, corresponding to the chemical equilibrium (14) and (15) and are tarnished with small variations of activity coefficient and Henry constant. However, this choice has been made considering the fact that we are in a screening phase and we do not acquire enough equilibrium data to correctly determine each parameter of the thermodynamic model for a new amine. Therefore, we choose to fix the activity model and to regress the pseudo-equilibrium constants Ka^* and Kc^* which have a stronger effect on the phase behavior.

3 GRAPH MACHINES

Statistical learning consists in building, from empirical data, mathematical models which reproduce the behavior of a process, so that the values of the outputs (here the thermodynamic properties of molecules) of this process can be predicted from its inputs (here the molecular structures). Classical modeling techniques – *i.e.* QSAR – draw linear or non-linear functions between the studied properties and structural features or other properties of the molecules, such as descriptors. The main drawbacks of these methods are the difficulty to choose the relevant descriptors and to perform their computation.

A new modeling technique developed by Goulon *et al.* [34-36] circumvents these problems, by drawing a direct relationship between the structure of the molecule and the property to be modeled. In this approach, called GM, molecules are considered as structured data and represented as graphs.

Then, the method consists in associating to each graph of the dataset (*i.e.* each molecule considered in the study) a mathematical function with the same structure, which will provide a prediction of the studied property. This function is obtained by composing identical parameterized functions, here neural networks. Modeling the properties consists in estimating the parameters shared by all the parameterized functions (*e.g.* the weights on the connections between the hidden units and the input and output variables) so that the values taken by functions associated to the graphs (*i.e.* molecules) are as close as possible to the values of their corresponding properties.

This method is based on the principle that two molecules with similar structures (and thus similar functions associated to their graphs) will have similar properties (and thus similar values taken by the functions).

3.1 Mathematical Structures of the Graph Machines

In a first step, the molecules, described by their SMILES, are converted into labeled graphs by the association of each non-hydrogen atom to a vertex and each bond to an edge. Labels describing the atoms (*e.g.* at least their natures, degrees and eventually other informations like stereoisomeries) are also assigned to the vertices. Then, the adjacency matrices associated to these labeled 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 [44]. 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 external nodes of the graphs to their terminal nodes. Even if 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. Figure 3 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 in 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 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. Then, the node function corresponding to the node s_b is:

$$f_{\boldsymbol{\theta}}(\mathbf{z}_b) = f_{\boldsymbol{\theta}}(\mathbf{v}_b, \mathbf{x}_b) \quad (18)$$

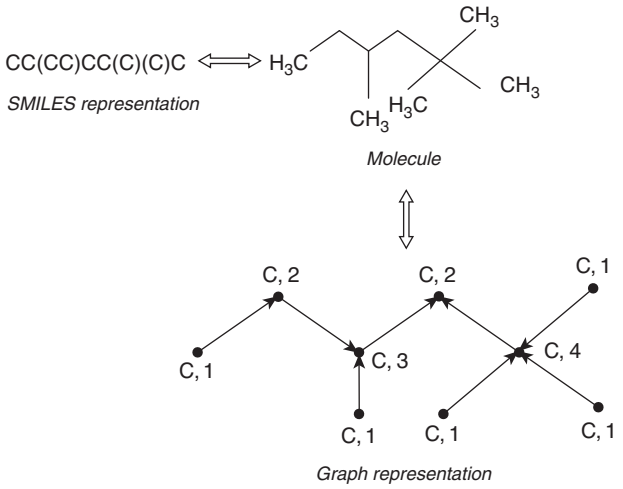

Figure 3

Conversion of a molecule from its SMILES representation into a directed acyclic graph.

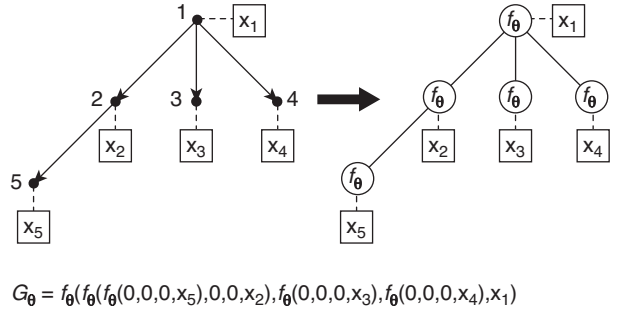

Figure 4

Processing of a molecule from its graph representation into a Graph Machine.

where:

- $\mathbf{v}_b$ is a vector whose 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, given in the labels.

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

$$g^i_{\boldsymbol{\theta}} = f_{\boldsymbol{\theta}}(\mathbf{z}_r) \tag{19}$$

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 $g^i_{\boldsymbol{\theta}}$ and across all those functions and share the same parameters $\boldsymbol{\theta}$. Figure 4 illustrates the processing of a molecule from its graph representation into a GM.

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, $g_{\boldsymbol{\theta}} = v_1$.

3.2 The Training of Graph Machines and Neural Networks (NN)

Training the GM consists in estimating 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. During the training of GM, the training set is composed of N molecules characterized by their structures/outputs pairs $\{(G_i, y^i),\ i = 1, \ldots, N\}$, where G_i is the parameterized function associated to graph of molecule i and y^i is the value of the modeled property for this molecule. 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 molecules present in the training set:

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{N} \left(y^i - g^i_{\boldsymbol{\theta}} \right)^2 \tag{20}$$

GM are trained in the usual framework of empirical risk minimization similarly to classical statistical learning techniques.

Neural networks are statistical tools [45] which enable to compute an output variable as a nonlinear function of input variables. A neural network is constituted by nodes, called neurons, which are interconnected in a netlike structure generally composed of three layers: one input layer (associated with the input variables), one output layer (associated to the output variables) and one intermediate layer, the hidden layer connected to the two other layers. The degree of influence between interconnected neurons is represented by numerical weights called connection weights. The required number of hidden neurons, N_h and the weights are optimized by an iterative process. The overall behavior of the system is modified by adjusting the connection weight values through the repeated application of the back-propagation algorithm. Neural network training is terminated when the cost function defined by Equation (20), which measures the difference between calculated and actual output values, is minimized.

Experimental data will be divided into three sets: a first set, called training set, is used to build several neural models with different hidden neurons (varying from 1 to 4), the second one, called validation set, is used to select among all those models the one with the best predictive ability and finally a third set, called prediction set, on which the best model is applied on.

3.3 Molecule Selection in the GM Multi-Dimensional Space

Models computed by the GM approach permit to express directly the structure of molecules and provide good predictive capabilities of their physicochemical activities and properties. However, they have the disadvantage of being black boxes for not allowing to clarify the possible links between the properties and the structure of molecules, or to represent the molecules as cloud of points in a multidimensional space making their typology accessible by clustering in homogeneous groups, as this is possible with molecular descriptors.

Indeed, the advantage of having a set of molecular descriptors is that they allow a representation of N molecules as points in the multidimensional space that they induce. In this space, provided with an appropriate metric, one can have a notion of similarity between molecules so that two points close in this space are likely to be associated with two similar molecules, while conversely two distant points would likely be associated with two dissimilar molecules in terms of all the descriptors characterizing them.

In this work, we overcome the disadvantage of the GM approach, as we develop a way to represent molecules in a multidimensional space. For this purpose, we used the functions $g^i_{\boldsymbol{\theta}}$ defined above. Let N be the number of possible candidate molecules (both those for which experimental measurements are available and those for which we do not have experimental measurements and would like to predict their properties by GM). M vectors are drawn randomly $\boldsymbol{\theta}^{*1}$, $\boldsymbol{\theta}^{*2}$, ..., $\boldsymbol{\theta}^{*M}$ in the space of model parameters. M is selected generally quite high[1] so that M parameters $\boldsymbol{\theta}^{*m}$ ($m = 1, \ldots, M$) are representative of the entire space of parameters. For each molecule i among the N candidate molecules, we associate the value $g^i_{\boldsymbol{\theta}}$ so that we obtain an array of size $N \times M$ where the lines are associated with the N molecules and columns are associated with the M parameter vectors $\boldsymbol{\theta}^{*m}$. The N molecules can then be represented by a cloud of N points in the multidimensional space of size M constructed. If we provide this space with an appropriate metric (*e.g.* the Euclidean metric), two similar molecules (respectively dissimilar) in terms of structure would be close (respectively distant) in this space. The distance between two candidate molecules i and i' can naturally be defined as:

$$d2(i,i') = \sum_{m=1}^{M} \left(g^i_{\boldsymbol{\theta}_m} - g^{i'}_{\boldsymbol{\theta}_m} \right)^2 \qquad (21)$$

1 The number M must be large enough to fill the space of vectors $\boldsymbol{\theta}$ constituted of neural network parameters. It depends, in fact, on the number of these parameters. However, the larger it is the more it increases the time of calculations in the space $\mathbf{R}^M$. We tested several values of M ranging from 100 to 10 000, for different numbers of parameters and have observed that $M = 1\,000$ offered the best compromise because there was little effect on the results beyond this value while allowing time quite acceptable for calculations.

Then, we use a factorial method, Principal Components Analysis (PCA) [46], to reduce the dimension of this space and have a suitable synthetic representation of the cloud of molecules in a lower dimensional space. Molecules belonging to the training, validation and prediction set are then chosen manually from this synthetic representation. The details of molecules belonging to the three sets are reported in Table 1 and a graphical representation of these molecules in the factorial plane is shown in Figure 5.

4 RESULTS AND DISCUSSION

First, we identify a set of 142 candidate molecules for CO_2 capture (see *Tab. A1* in the supporting information section). In this work, we focus on monoamine molecules but we have also considered a few polyamine candidates containing two nitrogen atoms. However, for these specific structures, one of the amine functions is practically not reactive in the sampled conditions (*e.g.* a tertiary amine having a pKa lower than the one of CO_2). Figure 5 shows the cloud of 142 candidate molecules projected in the plane formed by the two main principal components. This graph shows that the candidate molecules are arranged in three groups well enough separated from each other. Among all these molecules, we then choose several sets:

– 40 molecules are used for the training of the GM *i.e.* for the generation of statistical models;
– 6 molecules are used to choose the best model among those previously generated (validation);

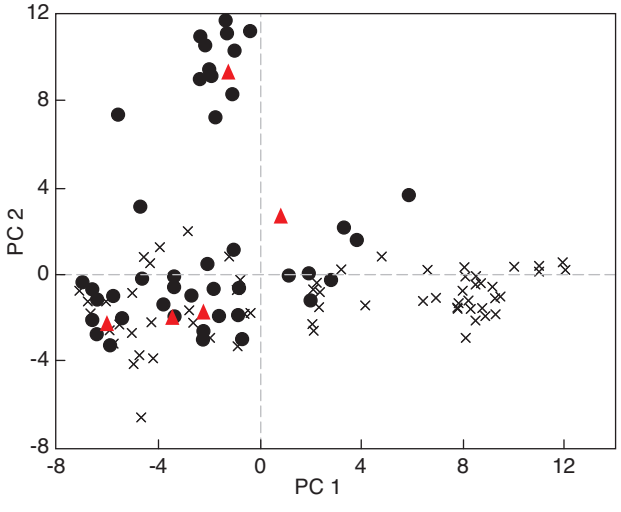

Figure 5

Projection of the 142 mono-amines on the main principal components. (×) Candidate, (●) training and validation and (▲) prediction set.

– 5 molecules are used for estimating the prediction ability of the model.

In this work, 46 molecules will then be used to build a Graph Machine model. Then, the ability of the model to predict the property of interest will be evaluated on 5 molecules. In each set, we choose the molecules so that they are equally divided within the different groups of molecules and they are representative of the 142 molecules ensemble. However, we observe that a whole group of molecules displaying high values of the first principal component is not represented. A closer inspection at the molecular structure reveals that most of these molecules are in fact aromatics amines. Therefore, they have limited solubility in water and were not viable candidates for our HTS experiments.

4.1 Experimental Absorption Isotherms and Thermodynamic Modeling

For each molecule considered in this work, solubility measurements are performed on 30 mass% amine aqueous solutions at T = 313.15 K. A standard comparison of performance for carbon dioxide capture would require to measure absorption isotherms at the same molar concentration of amine in each solvent. However, for high molar mass amines, this would induce large mass% of amine in the solution which would not be realistic for industrial applications due to degradation or mass transfer limitations induced by a large viscosity.

The same sequence of targeted pressure steps is used within each reactor that is 2.5, 5.0, 7.5, 10.0, 25.0, 50.0, 100.0, 200.0 and 300.0 kPa and the operating conditions used during the experiments are reported in Table 2.

TABLE 2

Operating conditions for HTS experiments performed in this work

Algorithm time		Experimental conditions	
τ (ms)	200	V_R (mL)	100
t_1 (s)	5	V_B (mL)	10
t_2 (s)	10	v (rpm)	1 500
τ_1 (s)	900	P (bar)	0-3
τ_2 (s)	10 800	T (°C)	40

The thermodynamic model is then used to fit the experimental HTS data by resorting to a modified Levenberg-Marquardt algorithm [47]. The objective function F_{obj} is defined as:

$$F_{obj} = \sum_{i=1}^{N_S} \left[P_{CO_2}(i) - P^T_{CO_2}(i) \right]^2 \quad (22)$$

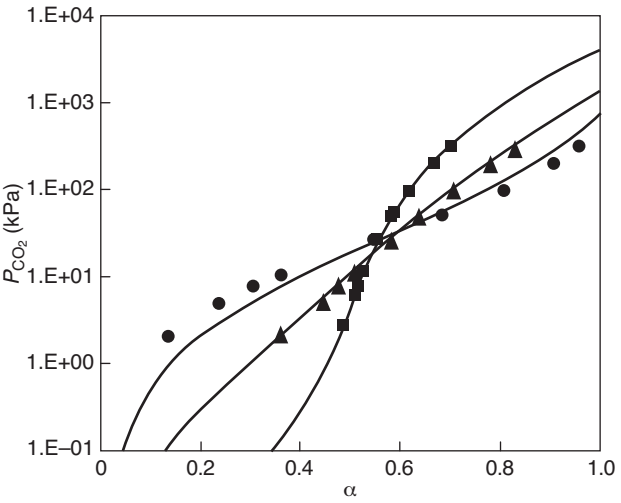

Figure 6

Experimental CO_2 absorption isotherm in 30.0 mass% MDEA (●), DEA (▲) and MEA (■) aqueous solution at T = 313.15 K. Smooth fitting using the thermodynamic model at T = 313.15 K (solid line).

where $P^T_{CO_2}$ refers to the theoretical CO_2 equilibrium partial pressure, that is the one calculated from the thermodynamic model. CO_2 absorption isotherms in 30 mass% MEA, DEA and MDEA aqueous solutions are measured at T = 313.15 K and experimental CO_2 solubility data are plotted in Figure 6, together with regressed thermodynamic model used in this work. The fitted absorption isotherms represent quite accurately the evolution of solvent loadings with pressure for the three kind of amine as the average deviation between the model and the experimental loadings does not exceed σ = 0.04. A small discrepancy is observed for MDEA aqueous solution at low loadings but it should be emphasized once again that the same parameters of the thermodynamic model are used to fit the loading variations on the three mono-amines.

In a previous work [24], we showed that CO_2 absorption isotherms display a wide variety of behavior as some isotherms are rather flat whereas some other display a high slope with increasing loadings. These behaviors are directly connected to the nature of the amine and more specifically to the value of the (pKa^*, pKc^*) couple. In the next section, we focus on the statistical modeling of the pseudo-constant of acidity pKa^*, as this thermodynamic property is common to the whole set of molecules whether they are primary, secondary or tertiary amines. Hence, from the modeling of pKa^* we will be able to reconstruct the whole absorption isotherms for tertiary amines for which the absence of formation of carbamate species ($Kc^* = 0$) usually induces lower energy of regeneration than those of primary or secondary amines [4].

4.2 Statistical Modeling

One important aspect of the GM approach is the degree of complexity of the node function used. On one hand, if we choose a simple node function, the degree of complexity of the resulting statistical models may not be sufficient to correctly establish a relationship between the structure of the molecules and the property of interest. On the other hand, using a complex node function may lead to overtraining. The resulting models will then be able to represent the slightest unphysical variation of the targeted property and will therefore be unfit to predict the behavior of a new molecule. In this work, the node functions are neural network, whose complexity are controlled by the number of hidden neurons. Therefore, we perform a series of simulations with an increasing number N_h, varying from 1 to 4, of hidden neurons. For each simulation, we compute the averaged absolute deviation of pKa^* ($<|\delta pKa^*|>$) both for the training and the validation sets *(Tab. 3, Fig. 7)*. When the number of hidden neurons increases, the quality of the modeling of the training set increases as $<|\delta pKa^*|>$ decreases almost linearly. For $N_h = 4$, we obtain a perfect modeling as $<|\delta pKa^*|> = 0$, indicating that the model is probably overfitting the data. Indeed, the evolution of the averaged pKa^* deviation for the validation set shows that we first obtain a poor modeling of the data as $<|\delta pKa^*|> = 0.4$ for $N_h = 1$; but as the number of hidden neurons increases the quality of the model improves to a point where $<|\delta pKa^*|> = 0.1$ for $N_h = 3$. The averaged pKa^* deviation is then similar for both the training and the prediction set when using three hidden neurons. Beyond this value, we clearly observe that the complexity of the model leads to overtraining as $<|\delta pKa^*|>$ increases for the validation set. Therefore, in the remainder of this paper we set the number of hidden neurons to $N_h = 3$ to obtain a good predictive model from the experimental data. The detailed modeling results obtained on the validation set are reported in the scatter plot of Figure 8 and in Table 4. The GM technique provides a good modeling of the data as the maximum absolute deviation is $|\delta pKa^*| = 0.4$ and $|\delta pKa^*| = 0.3$ for the training and the validation set respectively.

TABLE 3

Correlation coefficient (r), maximum and mean absolute deviation values between the predicted acidity constant and the experimental data

Hidden neurons	Training set			Validation set		
	r	mean	max	r	mean	max
1	0.894	0.45	1.75	0.868	0.41	0.63
2	0.926	0.35	2.36	0.982	0.13	0.25
3	**0.993**	**0.12**	**0.43**	**0.986**	**0.10**	**0.30**
4	1.000	0.00	0.00	0.958	0.26	0.54

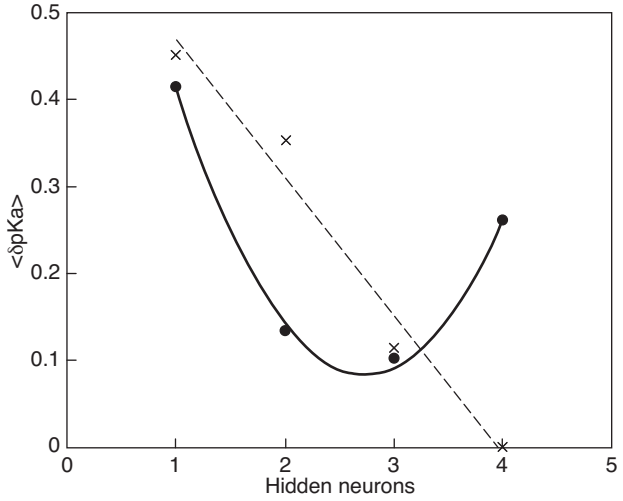

Figure 7

Average absolute deviation of pKa with the number of hidden neurons used in the statistical model. (×) Training set and (●) validation set. The lines are intended to guide the eye.

TABLE 4

Predicted and experimental pKa values of the validation set. Model: 3 hidden neurons

Molecule	pKa*	pKa* GM
Trans-4-aminocyclohexanol	9.37	9.34
1-amino-2-propanol	9.12	9.14
Triethanolamine	7.43	7.50
6-amino-1-hexanol	9.23	9.13
MEA	9.06	9.36
2-amino-(2-hydroxymethyl)-1,3-propanediol	7.79	7.91

The ability of the model to predict the pseudo-acidity constant pKa^* of new molecules is assessed by computing pKa^* for the five molecules of the prediction set and comparing these results to the experimental data measured afterwards. Details of the modeling results are reported in Table 5 and highlight the very good performance of the model to

TABLE 5

Predicted and experimental pKa* values of the prediction set. Model: 3 hidden neurons

Molecule	pKa*	pKa* GM
2-(dimethylamino)-ethanol	8.88	9.02
N-ethyldiethanolamine	8.41	8.37
2-(2-aminoethoxy)ethanol	8.74	8.61
3-amino-1,2-propanediol	8.89	8.98
2-pyrrolidinone	3.66	4.02

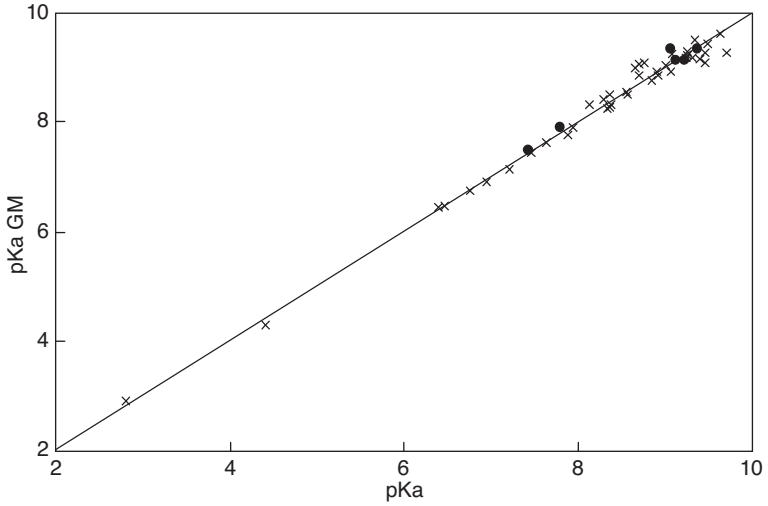

Figure 8

Comparison between predicted and experimental pKa (T = 40°C) for (×) training set and (●) validation set. Model: 3 hidden neurons.

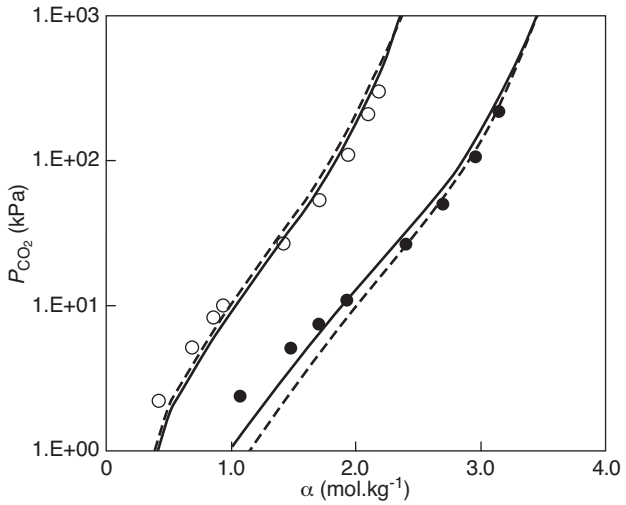

Figure 9

Predicted and experimental CO_2 absorption isotherms in 30 mass% 2-(dimethylamino)ethanol (●) and N-ethyldiethanolamine (○). (symbols) Experimental solubility data, (solid line) thermodynamic modeling and (dashed lines) QSAR GM predicted absorption isotherm.

predict acidity constant for a large variety of molecular structure. The maximum absolute pKa* deviation is $|\delta pKa^*| = 0.4$ and is obtained for a molecule displaying a low pKa* for which very few data are available in the training set.

Moreover, for the two tertiary amines of the prediction set (*i.e.* 2-(dimethylamino)ethanol and N-ethyldiethanolamine) we can rebuild the complete absorption isotherm and compare the solubility data for both the experimental values obtained with the HTS device and the theoretical value obtained from the thermodynamic model using the predicted pKa* *(Fig. 9)*. We observe a very good agreement for the two amines along the whole domain of loadings sampled in this work. Therefore, the QSAR GM modeling allows the computation of virtual absorption isotherms for tertiary amine molecules which have not yet been tested with the HTS device. Then, this method represents a powerful tool to identify the most suited structures which will display the most efficient thermodynamic properties.

CONCLUSIONS

GM based QSAR approach is used to build a relationship between the properties of amines for CO_2 capture and their molecular structures. We first introduce an original way to represent graph of molecules as cloud of points in a multidimensional space. A set of M vectors is drawn randomly in the space of model parameters and we compute the values $g^i_{\boldsymbol{\theta}}$ for each molecule. Then, the N possible candidate molecules can be represented by a cloud of N points in the multidimensional space of size M constructed. Two similar molecules (respectively dissimilar) in terms of structure would be close (respectively distant) in this space. Then, we used a factorial method, as Principal Components Analysis (PCA) to reduce the dimension of this space and have a suitable synthetic representation of the cloud of molecules in a lower dimensional space. This allowed us to cluster molecules in homogeneous and separable classes on the basis of the principal components and to select, in a representative way, training, validation and prediction sets for the estimate of the best neural model.

We performed CO_2 solubility measurements on a set of 46 amine molecules, chosen over a set of 142 possible candidate molecules, using a HTS experimental device. The evolution of loading with the CO_2 partial pressure for each molecule is then represented using a dedicated thermodynamic model resorting to two thermodynamic constants, pKa^* and pKc^*. The evolution of pKa^* with the molecular structure is then modeled using GM. The mathematical function associated to each graph of the dataset is obtained by composing identical parameterized functions, here neural networks.

The set of 46 molecules is divided in two sets (training, validation) used to build the statistical model and we used 5 other molecular structures to evaluate its prediction ability. Several models are generated by increasing the number of hidden neurons contained in the model and the most efficient one is chosen by checking that the data are accurately modeled on one hand but that we avoid overtraining on the other hand. An optimum number of $N_h = 3$ hidden neurons is found, which provides a good modeling of the data. The efficiency of this model is then verified by predicting the pKa^* for the prediction set. The results showed that the statistical model is able to accurately predict the evolution of pKa^* for a large variety of molecular structure. For tertiary amines, the computation of these thermodynamic parameters can be used to predict CO_2 solubility data and QSAR GM modeling therefore allows the computation of virtual absorption isotherms for amine molecules which have not yet been tested with the HTS device. Work is in progress to extend this methodology to the statistical modeling of polyamines candidates for which several molecular species may be forming from reactions with carbon dioxide.

REFERENCES

1 de Coninck H. (2010) Advocacy for carbon capture and storage could arouse distrust, *Nature* **463**, 293.

2 Rao A.B., Rubin E.S. (2002) A Technical, economic and environmental assessment of amine-based CO_2 capture technology for power plant greenhouse gas control, *Environ. Sci. Technol.* **36**, 4467-4475.

3 Rochelle G.T. (2009) Amine scrubbing for CO_2 capture, *Science* **325**, 1652-1654.

4 Porcheron F., Gibert A., Jacquin M., Mougin P., Faraj A., Goulon A., Bouillon P.-A., Delfort B., Le Pennec D., Raynal L. (2011) High Throughput Screening of amine thermodynamic properties applied to postcombustion CO_2 capture process evaluation, *Energy Procedia* **4**, 15-22.

5 Versteeg G.F., van Swaaij W.P.M. (1988) Solubility and diffusivity of acid gases (CO_2, N_2O) in aqueous alkanolamine solutions, *J. Chem. Eng. Data* **33**, 29-34.

6 Haji-Sulaiman M.Z., Aroua M.K., Illyas Pervez Md. (1996) Equilibrium concentration profiles of species in CO_2-alkanolamine-water systems, *Gas Sep. Purif.* **10**, 13-18.

7 Chauhan R.K., Yoon S.J., Lee H., Yoon J.-H., Shim J.-G., Song G.-C., Eum H.-M. (2003) Solubilities of carbon dioxide in aqueous solutions of triisopropanolamine, *Fluid Phase Equilib.* **208**, 239-245.

8 Seo D.-J., Hong W.-H. (1996) Solubilities of carbon dioxide in aqueous mixtures of diethanolamine and 2-amino-2-methyl-1-propanol, *J. Chem. Eng. Data* **41**, 258-260.

9 Ma'mun S., Jakobsen J.P., Svendsen H.F., Juliussen O. (2006) Experimental and modeling study of the solubility of carbon dioxide in aqueous 30 mass % 2-((2-aminoethyl)amino)ethanol Solution, *Ind. Eng. Chem. Res.* **45**, 2505-2512.

10 Ermatchkov V., Pérez-Salado Kamps A., Maurer G. (2006) Solubility of carbon dioxide in aqueous solutions of N-methyldiethanolamine in the low gas loading region, *Ind. Eng. Chem. Res.* **45**, 6081-6091.

11 Jou F.-Y., Mather A.E., Otto F.D. (1982) Solubility of H_2S and CO_2 in aqueous methyldiethanolamine solutions, *Ind. Eng. Chem. Process. Des. Dev.* **21**, 539-544.

12 Rho S.-W., Yoo K.-P., Lee J.S., Nam S.C., Son J.E., Min B.-M. (1997) Solubility of CO_2 in aqueous methyldiethanolamine solutions, *J. Chem. Eng. Data* **42**, 1161-1164.

13 Shen K.-P., Li M.-H. (1992) Solubility of carbon dioxide in aqueous mixtures of monoethanolamine with methyldiethanolamine, *J. Chem. Eng. Data* **37**, 96-100.

14 Ma'mun S., Nilsen R., Svendsen H.F., Juliussen O. (2005) Solubility of carbon dioxide in 30 mass % monoethanolamine and 50 mass % methyldiethanolamine solutions, *J. Chem. Eng. Data* **50**, 630-634.

15 Mathonat C., Majer V., Mather A.E., Grolier J.-P.E. (1998) Use of flow calorimetry for determining enthalpies of absorption and the solubility of CO_2 in aqueous monoethanolamine solutions, *Ind. Eng. Chem. Res.* **37**, 4136-4141.

16 Jou F.-Y., Otto F.D., Mather A.E. (1994) Vapor-Liquid Equilibrium of carbon dioxide in aqueous mixtures of monoethanolamine and methyldiethanolamine, *Ind. Eng. Chem. Res.* **33**, 2002-2005.

17 Kent R., Eisenberg B. (1976) Better data for amine treating, *Hydrocarbon Proc.* **55**, 87-90.

18 Sartori G., Savage D.W. (1983) Sterically hindered amines for CO_2 removal from gases, *Ind. Eng. Chem. Fundam.* **22**, 239-249.

19 Austgen D.M., Rochelle G.T., Peng X., Chen C-C. (1989) Model of vapor-liquid equilibria for aqueous acid gas-alkanolamine systems using the Eletrolyte-NRTL equation, *Ind. Eng. Chem. Res.* **28**, 1060-1073.

20 Benamor A., Aroua M.K. (2005) Modeling of CO_2 Solubility and carbamate concentration in DEA, MDEA and their mixtures using the Deshmukh–Mather model, *Fluid Phase Equilib.* **231**, 150-162.

21 Ma'mun S., Svendsen H.F., Hoff K.A., Juliussen O. (2007) Selection of new absorbents for carbon dioxide capture, *Energy Convers. Manage.* **48**, 251-258.

22 Bonenfant D., Mimeault M., Hausler R. (2003) Determination of the structural features of distinct amines important for the absorption of CO_2 and regeneration in aqueous solution, *Ind. Eng. Chem. Res.* **42**, 3179-3184.

23 Puxty G., Rowland R., Allport A., Yang Q., Bown M., Burns R., Maeder M., Attalla M. (2009) Carbon dioxide postcombustion capture: A novel screening study of the carbon dioxide absorption performance of 76 amines, *Environ. Sci. Technol.* **43**, 6427-6433.

24 Porcheron F., Gibert A., Mougin P., Wender A. (2011) High Throughput Screening of CO_2 solubility in aqueous monoamine solutions, *Environ. Sci. Technol.* **45**, 2486-2492.

25 Hansch C., Leo A., Hoekman D. (1995) *Exploring QSAR – Hydrophobic, electronic and steric constants*, American Chemical Society, Washington, D.C.

26 Wold S. (1991) Validation of QSARs, *QSAR* **10**, 191-193.

27 Friesner R.A. (1991) New methods for electronic structure calculations on large molecules, *Ann. Rev. Phys. Chem.* **42**, 341-367.

28 Marten B., Kim K., Cortis C., Friesner R.A., Murphy R.B., Ringnalda M.N., Sitkoff D., Honig B. (1996) New model for calculation of solvation free energies: corrections of self-consistent reaction field continuum dielectric theory for short-range hydrogen-bonding Effects, *J. Phys. Chem.* **100**, 11775-11788.

29 Tannor D.J., Marten B., Murphy R., Friesner R.A., Sitkoff D., Nicholls A., Ringnalda M., Goddard W.A., Honig B. (1994) Accurate first principles calculation of molecular charge distributions and solvation energies from Ab initio quantum mechanics and continuum dielectric theory, *J. Am. Chem. Soc.* **116**, 11875-11882.

30 Abbaci K., Hadjali A., Lietard L., Rocacher D. (2011) A similarity skyline approach for handling graph queries – A preliminary report, *2011 IEEE 27th International Conference on Data Engineering Workshops (ICDEW)*, Hannover, Germany, 11-16 April.

31 Conte D., Foggia P., Sansone C., Vento M. (2004) Thirty years of graph matching in pattern recognition, *Int. J. Pattern Recogn. Artif. Intell.* **18**, 265-298.

32 Hu H., Hang Y., Han J., Zhou X. (2005) Mining Coherent dense subgraphs across massive biological network for functional discovery, *Bioinformatics* **1**, 1-9.

33 Tian Y., McEachin R., Santos C., States D.J., Patel J.M. (2007) Saga: A subgraph matching tool for biological graphs, *Bioinformatics* **23**, 232-239.

34 Goulon-Sigwalt-Abram A., Duprat A., Dreyfus D. (2005) From hopfield nets to recursive networks to graph machines: Numerical machine learning for structured data, *Theor. Comput. Sci.* **344**, 298-344.

35 Goulon A., Duprat A., Dreyfus D. (2006) Graph machines and their applications to computer-aided drug design: A new approach to learning from structured data, Lecture Notes in *Comput. Sci.* **4135**, 1-19.

36 Goulon A., Picot T., Duprat A., Dreyfus D. (2007) Predicting activities without computing descriptors: graph machines for QSAR, *SAR QSAR Environ. Res.* **18**, 141-153.

37 Goulon A., Faraj A., Pirngruber G., Jacquin M., Porcheron F., Leflaive P., Martin P., Baron G.V., Denayer J.F.M. (2011) Novel graph machine based QSAR approach for the prediction of the adsorption enthalpies of alkanes on zeolites, *Catal. Today* **159**, 74-83.

38 Bunke H., Riesen K. (2011) Recent advances in graph-based pattern recognition with application in document analysis, *Pattern Recogn.* **44**, 1057-1067.

39 Schenker A., Bunke H., Last M., Kandel A. (2005) *Graph-theoretic techniques for web content mining*, World Scientific.

40 Klinger S., Austin J. (2005) Chemical similarity searching using a neural graph matcher, in *Proc. of 13th European Symposium on Artificial Neural Networks (ESANN)*, Bruges, Belgium, 27-29 April, pp. 479-484

41 Weininger D. (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, *J. Chem. Inf. Comput. Sci.* **28**, 31-36.

42 Weininger D., Weininger A., Weininger J.L. (1989) SMILES, a chemical language and information system. 2. Algorithm for generation of unique SMILES notation, *J. Chem. Inf. Comput. Sci.* **29**, 97-101.

43 Blanchon Le Bouhelec E., Mougin P., Barreau A., Solimando R. (2007) Rigorous modelling of the acid gas heat of absorption in alkanolamine solutions, *Energy Fuels* **21**, 2044-2055.

44 Jochum C., Gasteiger J. (1977) *J. Chem. Inf. Comput. Sci.* **17**, 113-117.

45 Hastie T., Tibshirani R., Friedman J. (2009) *The elements of statistical learning*, Springer, 2nd Ed.

46 Livingstone D. (2002) *Data Analysis for Chemists*, Oxford University Press.

47 Aarnink W.A.M., Weishaupt A., Vansilfhout A. (1990) Angle-resolved X-ray photoelectron-spectroscopy (ARXPS) and a modified Levenberg-Marquardt fit procedure – A new combination for modelling thin-layers, *Appl. Surf. Sci.* **45**, 37-48.

Final manuscript received in April 2012
Published online in February 2013

APPENDIX

Supporting information description

The IUPAC name, SMILES and CAS number of the 142 candidate molecules for CO_2 capture.

TABLE A1 - PART I

List of candidate molecules

IUPAC	SMILES	CAS Number
2-aminoethan-1-ol	NCCO	141-43-5
2-(methylamino)ethan-1-ol	CNCCO	109-83-1
2-methoxyethan-1-amine	NCCOC	109-85-3
3-aminopropan-1-ol	NCCCO	156-87-6
1,2-oxazole	c1ccno1	288-14-2
1,3-oxazole	c1cnco1	288-42-6
1-aminopropan-2-ol	NCC(O)C	78-96-6
2-(dimethylamino)ethan-1-ol	CN(C)CCO	108-01-0
2-(ethylamino)ethan-1-ol	CCNCCO	110-73-6
Morpholine	C1OCCNC1	110-91-8
2-amino-2-methylpropan-1-ol	NC(CO)(C)C	124-68-5
4-aminobutan-1-ol	NCCCCO	13325-10-5
N,N-diethylhydroxylamine	ON(CC)CC	3710-84-7
Methoxy(propan-2-yl)amine	NC(C)COC	37143-54-7
3-methoxypropan-1-amine	NCCCOC	5332-73-0
Pyrrolidin-2-one	O=C1CCCN1	616-45-5
2-aminobutan-1-ol	OCC(CC)N	96-20-8
2-ethyl-4,5-dihydro-1,3-oxazole	CC\C1=N\CCO1	10431-98-8
1-(dimethylamino)propan-2-ol	N(CC(O)C)(C)C	108-16-7
4-methylmorpholine	C1OCCN(C)C1	109-02-4
2-(propan-2-ylamino)ethan-1-ol	CC(C)NCCO	109-56-8
1-(dimethylamino)propan-2-ol	CN(C)CC(O)C	203-556-4
3-(dimethylamino)propan-1-ol	N(CCCO)(C)C	3179-63-3
Piperidin-4-ol	C1C(O)CCNC1	5382-16-1
Piperidin-3-ol	OC1CCCNC1	6859-99-0
1-methylpyrrolidin-2-one	O=C1CCCN1C	872-50-4
4-(dimethylamino)butan-2-one	CN(C)CCC(C)=O	2543-57-9
2-(diethylamino)ethan-1-ol	C(N(CCO)CC)C	100-37-8

TABLE A1 - PART II

Pyridin-3-ylmethanol	C1(CO)=CC=CN=C1	100-55-0
4-ethylmorpholine	C1OCCN(CC)C1	100-74-3
1-methylpiperidin-4-ol	C1C(O)CCN(C)C1	106-52-5
IUPAC	SMILES	CAS Number
2-(butylamino)ethan-1-ol	CCCCNCCO	111-75-1
4-methoxypyridine	O(c1ccncc1)C	1122-96-9
4-aminophenol	C1=C(O)C=CC(N)=C1	123-30-8
2-(diethylamino)ethan-1-ol	OCCN(CC)CC	100-37-8
2,6-dimethylmorpholine	CC1CNCC(C)O1	141-91-3
2-methoxypyridine	O(c1ncccc1)C	1628-89-3
2-(pyrrolidin-1-yl)ethan-1-ol	OCCN1CCCC1	2955-88-6
2-Piperidinylmethanol	C1CCCNC1CO	3433-37-2
1-methylpiperidin-3-ol	C1(O)CN(C)CCC1	3554-74-3
6-aminohexan-1-ol	NCCCCCCO	4048-33-3
Piperidin-3-ylmethanol	C1(CO)CCCNC1	4606-65-9
2-(tert-butylamino)ethan-1-ol	C(NCCO)(C)(C)C	4620-70-6
3-aminophenol	C1(O)=CC=CC(N)=C1	591-27-5
Piperidin-4-ylmethanol	C1C(CO)CCNC1	6457-49-4
2,6-dimethylmorpholine	N1CC(C)OC(C)C1	141-91-3
2-(dimethylamino)-2-methylpropan-1-ol	CN(C)C(C)(C)CO	7005-47-2
3-methoxypyridine	O(c1cccnc1)C	7295-76-3
2-aminophenol	C1=CC=CC(N)=C1O	95-55-6
1-methylpiperidin-4-ol	OC1CCN(C)CC1	106-52-5
Piperidin-3-ylmethanol	OCC1CCCNC1	4606-65-9
1-ethylpiperidin-3-ol	C1(O)CN(CC)CCC1	13444-24-1
2-(piperidin-2-yl)ethan-1-ol	C1CCCNC1CCO	1484-84-0
(1-methylpiperidin-2-yl)methanol	C1C(CO)N(C)CCC1	20845-34-5
2-amino-5-methylphenol	Cc1cc(O)c(N)cc1	2835-98-5
2-(piperidin-1-yl)ethan-1-ol	C1CN(CCO)CCC1	3040-44-6
1-ethylpiperidin-4-one	C1N(CC)CCC(=O)C1	3612-18-8
2-amino-4-methylphenol	Cc1ccc(O)c(N)c1	95-84-1
3-amino-4-methylphenol	Cc1ccc(O)cc1N	2836-00-2
(1-methylpiperidin-2-yl)methanol	OCC1CCCCN1C	20845-34-5
8-methyl-8-azabicyclo[3.2.1]octan-3-ol	CN2C1CCC2CC(O)C1	120-29-6

TABLE A1 - PART III

2-methoxy-5-methylaniline	Nc1cc(C)ccc1OC	120-71-8
4-ethoxyaniline	CCOc1ccc(N)cc1	156-43-4
1-propylpiperidin-4-one	CCCN1CCC(=O)CC1	23133-37-1
IUPAC	SMILES	CAS Number
(4-methoxyphenyl)methanamine	COc1ccc(CN)cc1	2393-23-9
3-(pyridin-3-yl)propan-1-ol	OCCCc1cccnc1	2859-67-8
4-(diethylamino)butan-2-one	CC(CCN(CC)CC)=O	3299-38-5
2-[bis(propan-2-yl)amino]ethan-1-ol	CC(C)N(CCO)C(C)C	96-80-0
3-(dimethylamino)phenol	C1(O)=CC=CC(N(C)C)=C1	99-07-0
4-(dimethylamino)benzaldehyde	CN(C)c1ccc(C=O)cc1	100-10-7
2-(benzylamino)ethan-1-ol	C1(CNCCO)=CC=CC=C1	104-63-2
5-(diethylamino)pentan-2-one	CC(CCCN(CC)CC)=O	105-14-6
Quinolin-8-ol	Oc1cccc2cccnc12	148-24-3
2,2,6,6-tetramethylpiperidin-4-ol	OC1CC(C)(C)NC(C)(C)C1	2403-88-5
1-(propan-2-yl)piperidin-4-one	C1N(CC(C)C)CCC(=O)C1	5355-68-0
2-(4-methoxyphenyl)ethan-1-amine	COc1ccc(CCN)cc1	55-81-2
2-[methyl(phenyl)amino]ethan-1-ol	C1=CC=CC(N(CCO)C)=C1	93-90-3
2-[benzyl(methyl)amino]ethan-1-ol	C1(CN(CCO)C)=CC=CC=C1	101-98-4
4-tetrahydro-2H-pyran-4-ylpyridine	n1ccc(cc1)C2CCOCC2	26684-56-0
1,2,2,6,6-pentamethylpiperidin-4-one	C1(C)(C)N(C)C(C)(C)CC(=O)C1	5554-54-1
4-cyclohexylmorpholine	O1CCN(C2CCCCC2)CC1	6425-41-8
5-aminonaphthalen-1-ol	Nc2cccc1c2cccc1O	83-55-6
3-(diethylamino)phenol	CCN(CC)c1cccc(O)c1	91-68-9
2-[ethyl(phenyl)amino]ethan-1-ol	OCCN(CC)c1ccccc1	92-50-2
1-benzylpiperidin-4-ol	OC2CCN(Cc1ccccc1)CC2	4727-72-4
1-benzylpiperidin-4-ol	C2(CN1CCC(O)CC1)=CC=CC=C2	4727-72-4
3-butyl-2-(heptan-3-yl)-1,3-oxazolidine	CCC(CCCC)C1OCCN1CCCC	165101-57-5
2-aminoacetic acid	NCC(=O)O	56-40-6
2-(methylamino)acetic acid	CNCC(O)=O	107-97-1
3-aminopropane-1,2-diol	NCC(O)CO	616-30-8
2-[(2-hydroxyethyl)amino]ethan-1-ol	N(CCO)CCO	111-42-2
2-(dimethylamino)acetic acid	CN(C)CC(O)=O	1118-68-9
2-amino-2-methylpropane-1,3-diol	CC(N)(CO)CO	115-69-5

TABLE A1 - PART IV

3-(methylamino)propane-1,2-diol	N(CC(CO)O)C	40137-22-2
2-(2-aminoethoxy)ethan-1-ol	NCCOCCO	929-06-6
2-[(2-hydroxyethyl)(methyl)amino]ethan-1-ol	OCCN(CCO)C	105-59-9
IUPAC	SMILES	CAS Number
(2,2-dimethoxyethyl)(methyl)amine	CNCC(OC)OC	122-07-6
Morpholine-4-carbaldehyde	O=CN1CCOCC1	4394-85-8
3-(dimethylamino)propane-1,2-diol	N(CC(CO)O)(C)C	623-57-4
1-[(2-hydroxypropyl)amino]propan-2-ol	OC(CNCC(O)C)C	110-97-4
Bis(2-methoxyethyl)amine	COCCNCCOC	111-95-5
2-[ethyl(2-hydroxyethyl)amino]ethan-1-ol	CCN(CCO)CCO	139-87-7
2-[2-(dimethylamino)ethoxy]ethan-1-ol	CN(C)CCOCCO	1704-62-7
2-(morpholin-4-yl)ethan-1-ol	C1N(CCO)CCOC1	622-40-2
2,4-dimethoxyaniline	Nc1ccc(cc1OC)OC	2735-04-8
2-[butyl(2-hydroxyethyl)amino]ethan-1-ol	CCCCN(CCO)CCO	102-79-4
2-[tert-butyl(2-hydroxyethyl)amino]ethan-1-ol	CC(C)(C)CN(CCO)CCO	2160-93-2
(2,2-diethoxyethyl)dimethylamine	CN(C)CC(OCC)OCC	3616-56-6
2-amino-3-phenylpropanoic acid	OC(=O)NCc1ccccc1	150-30-1
3-[bis(propan-2-yl)amino]propane-1,2-diol	OCC(O)CN(C(C)C)C(C)C	85721-30-8
2-[(2-hydroxyethyl)(phenyl)amino]ethan-1-ol	C1=CC=CC(N(CCO)CCO)=C1	120-07-0
2-(3,4-dimethoxyphenyl)ethan-1-amine	COc1cc(ccc1OC)CCN	120-20-7
1-(3,3-dimethoxypropyl)piperidine	COC(CCN1CCCCC1)OC	31007-28-0
2-amino-2-(hydroxymethyl)propane-1,3-diol	NC(CO)(CO)CO	77-86-1
2-[bis(2-hydroxyethyl)amino]ethan-1-ol	N(CCO)(CCO)CCO	102-71-6
3-(morpholin-4-yl)propane-1,2-diol	OCC(O)CN1CCOCC1	6425-32-7
1-[bis(2-hydroxypropyl)amino]propan-2-ol	CC(O)CN(CC(C)O)CC(O)C	122-20-3
2-amino-3-(4-hydroxyphenyl)propanoic acid	Oc1ccc(CC(N)C(O)=O)cc1	556-03-6
1-[bis(2-hydroxypropyl)amino]propan-2-ol	N(CC(O)C)(CC(O)C)CC(O)C	122-20-3
Bis(2,2-diethoxyethyl)(methyl)amine	CCOC(CN(C)CC(OCC)OCC)OCC	6948-86-3
3-aminopropanenitrile	N#CCCN	151-18-8
3-(dimethylamino)propanenitrile	CN(C)CCC#N	1738-25-6
2-(piperazin-1-yl)ethan-1-ol	OCCN1CCNCC1	103-76-4
2-aminopropan-1-ol	CC(N)CO	6168-72-5
2-(morpholin-4-yl)ethan-1-amine	NCCN1CCOCC1	2038-03-1
4-aminocyclohexan-1-ol	NC1CCC(O)CC1	27489-62-9