



Computational screening methodology identifies effective solvents for CO₂ capture

Alexey A. Orlov¹, Alain Valtz², Christophe Coquelet², Xavier Rozanska³, Erich Wimmer³, Gilles Marcou¹, Dragos Horvath¹, Bénédicte Poulain⁴, Alexandre Varnek¹  [✉] & Frédérick de Meyer^{2,4}  [✉]

Carbon capture and storage technologies are projected to increasingly contribute to cleaner energy transitions by significantly reducing CO₂ emissions from fossil fuel-driven power and industrial plants. The industry standard technology for CO₂ capture is chemical absorption with aqueous alkanolamines, which are often being mixed with an activator, piperazine, to increase the overall CO₂ absorption rate. Inefficiency of the process due to the parasitic energy required for thermal regeneration of the solvent drives the search for new tertiary amines with better kinetics. Improving the efficiency of experimental screening using computational tools is challenging due to the complex nature of chemical absorption. We have developed a novel computational approach that combines kinetic experiments, molecular simulations and machine learning for the *in silico* screening of hundreds of prospective candidates and identify a class of tertiary amines that absorbs CO₂ faster than a typical commercial solvent when mixed with piperazine, which was confirmed experimentally.

¹Laboratory of Chemoinformatics, Faculty of Chemistry, University of Strasbourg, 67081 Strasbourg, France. ²MINES ParisTech, PSL University, Centre of Thermodynamics of Processes (CTP), 35 rue St Honoré, 77300 Fontainebleau, France. ³Materials Design SARL, 42 avenue Verdier, 92120 Montrouge, France. ⁴TOTALEnergies S.E., OneTech, Gas & Low Carbon Entity, CCUS R&D Program, 2 Place Jean Millier, 92078 Paris, France. [✉]email: varnek@unistra.fr; frederick.de-meyer@totalenergies.com

Numerous technologies exist for capturing CO₂ including chemical absorption, cryogenic separation, removal with membranes, and adsorption with zeolites or metal-organic frameworks^{1–6}. The cyclic chemical absorption and regeneration process based on common primary and secondary amines such as monoethanolamine (MEA) and diethanolamine (DEA) is the most mature in industrial applications^{3,5}. Unhindered primary and secondary amines react rapidly with CO₂ to form very stable carbamates. The amount of energy required for the regeneration of these solvents is large. Carbon capture applied to a coal-fired power plant may reduce the net output of the plant by 30%⁶. With sterically hindered amines or tertiary amines like the standard methyldiethanolamine (MDEA), CO₂ is captured as bicarbonate, which has a much smaller heat of reaction than carbamate formation, resulting in regeneration energy savings⁷. Moreover, their CO₂ absorption capacity is much higher. Tertiary amines are therefore increasingly used in the high-pressure natural gas treatment industry to remove acid gases like CO₂. However, in general, the rate of direct bicarbonate formation is much lower than that of carbamate formation resulting in much slower CO₂ absorption rates with tertiary amines and thus in unacceptable large equipment for low pressure, anthropogenic (flue gas), CO₂ capture applications^{5,7}. To tackle this problem, several approaches were suggested. Several studies reported that the usage of a catalyst allows one to speed up the absorption of CO₂ and/or to lower the energetic cost of solvent regeneration⁸. Another option, which is currently followed by the industry, consists in adding an activator, piperazine, significantly boosting the overall CO₂ absorption rate without increasing the regeneration energy too much⁹. A more straightforward strategy would be the identification of new tertiary amines with much higher absorption rates with respect to standard MDEA and to which piperazine can eventually be added. Since experimental measurement of CO₂ absorption kinetics is a time and labor-intensive process, the rational approach to the design of tertiary amines that can rapidly absorb CO₂ requires a quantitative model enabling to select only the best candidates for experimental measurements.

Concerning alternative processes based on adsorption in porous solids (still under development), a lower theoretical energy consumption is expected due to the weaker physical adsorption. Molecular simulations and machine learning have already been extensively used to perform virtual screening of hundreds of thousands of structures to identify potentially better materials for CO₂ adsorption^{10,11}. Until now it was not possible to apply a similar methodology for amines, because of the difficulty related to the computation of chemical reactions. Amines were rationally designed based on physical and thermodynamic properties and the CO₂ absorption rates were measured experimentally for only the most promising candidates^{7,12}. Previously, machine-learning algorithms were tentatively applied for modeling quantitative structure–property relationship (QSPR) of alkanolamines' CO₂ absorption-related properties^{13–18}. However, the availability of only a very small amount of data points limited the applicability domain of the models. Hence, to address this issue, we developed and applied a methodology for the identification of tertiary amines effectively absorbing CO₂ based on the combination of molecular simulations¹⁹ and machine learning. In parallel, an experimental setup for the measurement of CO₂ absorption rates has been specifically designed and put in place to validate the approach.

Results and discussion

Design of the methodology for CO₂ absorbents screening. The workflow of the methodology is presented in Fig. 1. Chowdhury et al.²⁰ published a consistent experimental dataset of the absorption rates of CO₂ for 24 aqueous tertiary amines (313 K, 30 wt% amine). In the absence of a clear relationship between the

structure or the chemical properties (e.g., the basicity) of the amines and the CO₂ absorption rates, we developed a molecular dynamics (MD) based model that can accurately predict those experimental CO₂ absorption rates¹⁹. It was found that, while the basicity of the amine (quantified by the pK_a) is important, the key to the precision of molecular simulations is the inclusion of subtle but important solvation effects in the calculation of the activation Gibbs free energy of the reaction with an accuracy better than 1 kJ mol^{−1}. One of the important features of the MD model¹⁹ is the robustness to reasonable changes in the concentration of amine and in temperature, enabling to apply it to a rather wide range of experimental setups. Hence, the model was applied to predict the rates at 13 mol% of amines and at 323 K, because these conditions are more representative of industrial absorption⁵.

Being much less resource- and cost-demanding, molecular simulations can thus be used instead of the experiments to get enough data for building a reliable QSPR model with a wide applicability domain.

Molecular simulations of CO₂ absorption process. A dataset containing 100 structurally diverse tertiary amines was composed based on the in-house TotalEnergies's dataset of amines with known experimental properties, complemented with tertiary amines extracted from literature and public databases (PubChem^{21,22}, ZINC^{23,24}). The selected compounds comprise diverse chemotypes, including linear and cyclic amines, diamines, amines containing thiol and thioether groups. Molecular simulations (see “Methods”) were performed for the initial set of 24 amines and for the selected set of 100 amines at 323 K and using a 13 mol% concentration of amine. From MD simulations absorption rates (R_{MD}) and free energies of absorption (ΔG_{MD}) were obtained. Notably, the R_{MD} values calculated at 313 and 323 K are highly correlated (Fig. 2a, Spearman rank correlation coefficient (ρ) 0.99).

As shown in Fig. 2b, the most rapidly absorbing compound according to the MD calculations and the data from Chowdhury et al.²⁰ was 3-(Diethylamino)-1,2-propanediol (DEA-1,2-PD). However, most of the other compounds with the largest predicted rates of absorption (R_{MD}) contained either piperidine or pyrrolidine cycles. This is in line with the data from Chowdhury et al.²⁰, who showed that 3-piperidino-1,2-propanediol (3PP-1,2-PD) and 1-methyl-2-piperidineethanol were significantly faster than the industrially used methyldiethanolamine (MDEA). Figure 2c illustrates that the computed CO₂ absorption Gibbs free energies ΔG_{MD} are almost perfectly correlated with the CO₂ absorption rates, R_{MD} (Spearman ρ −0.98): the slower the CO₂ absorption, the higher the absorption Gibbs free energy. The correlation is not linear, and the decrease of ΔG_{MD} slows down significantly at higher CO₂ absorption rates.

Virtual screening of tertiary amines and experimental validation. Machine-learning algorithms were applied to establish quantitative structure–property relationships and screen a set of tertiary amines from a public dataset. The values of pK_a predicted by the OPERA model²⁵ can be used as a rather good predictor for ΔG_{MD}. Indeed, the fitting of linear regression with the pK_a values as the only predictor leads to a reasonable predictive performance in cross-validation (Supplementary Table 2). For modeling both end-points (ΔG_{MD} and R_{MD}), we implemented a machine-learning workflow combining several machine-learning algorithms and various descriptors of molecular structures. Thus, predicted pK_a values were complemented with other descriptor types: physicochemical descriptors from OPERA and various types of molecular fragments calculated using ISIDA-Fragmentor^{26,27}. Finally, we used a consensus of several individual models built with the help of random forest (RF)²⁸ and

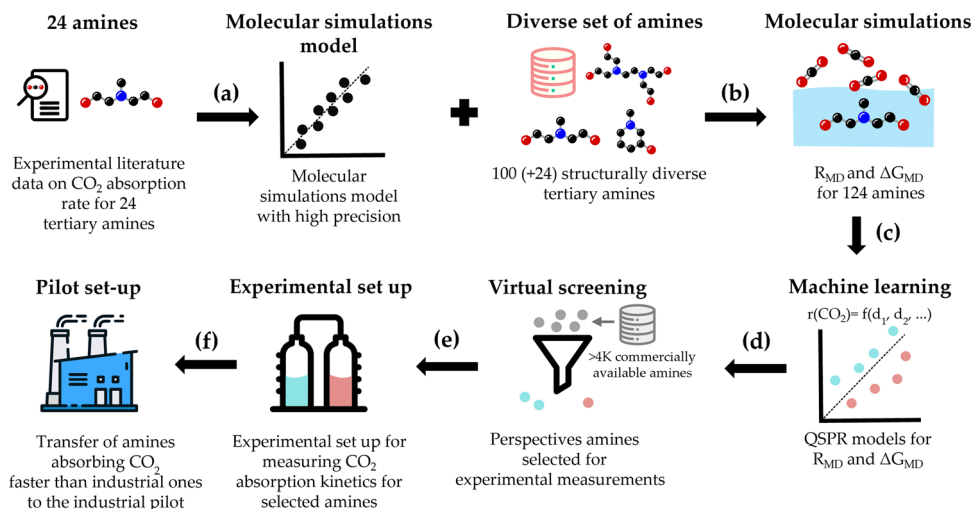


Fig. 1 Workflow of the methodology suggested in this paper. **a** A high precision molecular simulation-based model for absorption rate prediction is developed¹⁹ and validated with experimental data on CO₂ absorption rates for 24 tertiary amines²⁰. The accuracy of the Gibbs free energies of absorption is better than 1 kJ mol⁻¹ in comparison to experimental values¹⁹. **b** This model is applied to a diverse dataset containing 100 tertiary amine structures to calculate the CO₂ absorption rate (R_{MD}) and free energy of absorption (ΔG_{MD}) (see “Methods”). **c** QSPR models were built for R_{MD} and ΔG_{MD}. **d** QSPR models were used to select perspective commercially available amines from public datasets. **e** Experimental measurement of CO₂ absorption rates for selected amines. **f** The most selective ones can be further studied and eventually tested in a pilot unit.

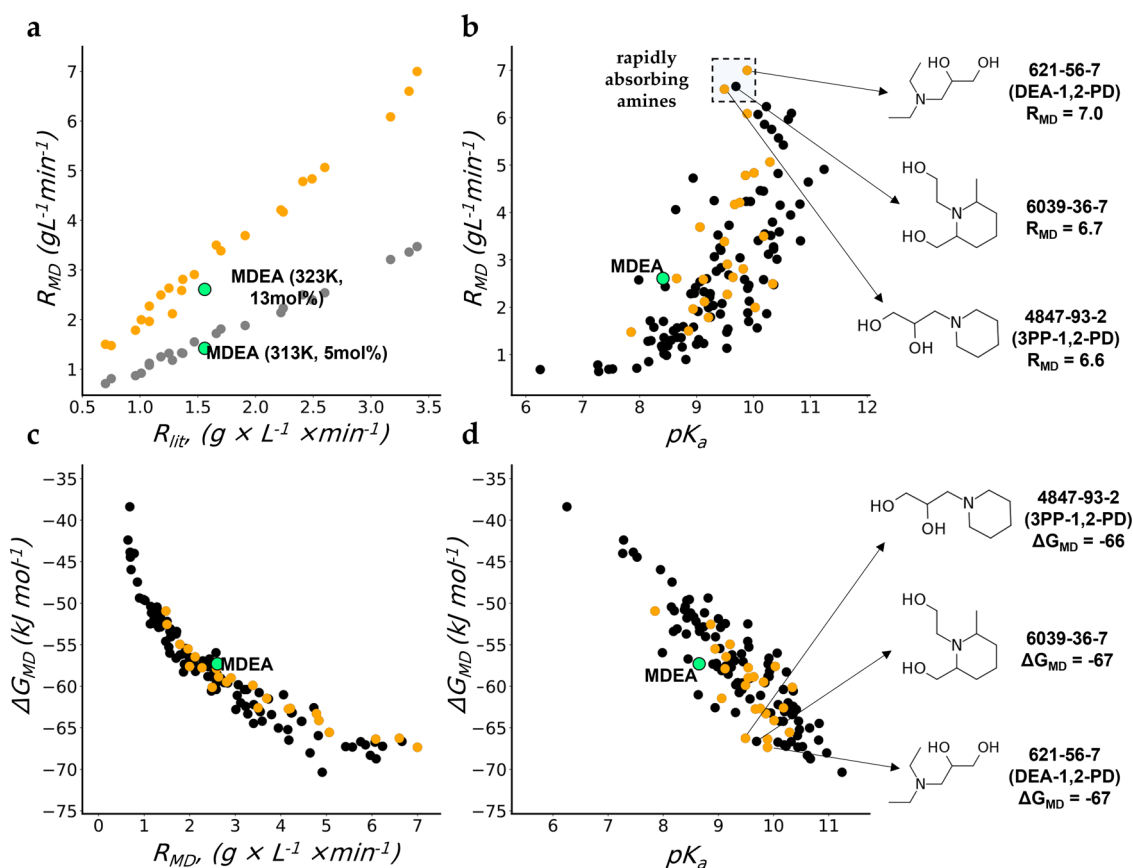


Fig. 2 Results of molecular dynamics simulations of the CO₂ absorption process. **a** CO₂ absorption rates (R_{MD}) at 313 K (gray) and 323 K (orange) predicted using MD and the experimental absorption rates (R_{ite}) at 313 K. **b** R_{MD} vs predicted pK_a values (pK_a). **c** energy of absorption (ΔG_{MD}) predicted by MD vs R_{MD}. **d** ΔG_{MD} vs predicted pK_a. The 24 amines from Chowdhury et al.²⁰ are shown in orange. The 100 amines for which MD simulations were performed are shown in black. Industrially used reference compound (MDEA) is shown in green.

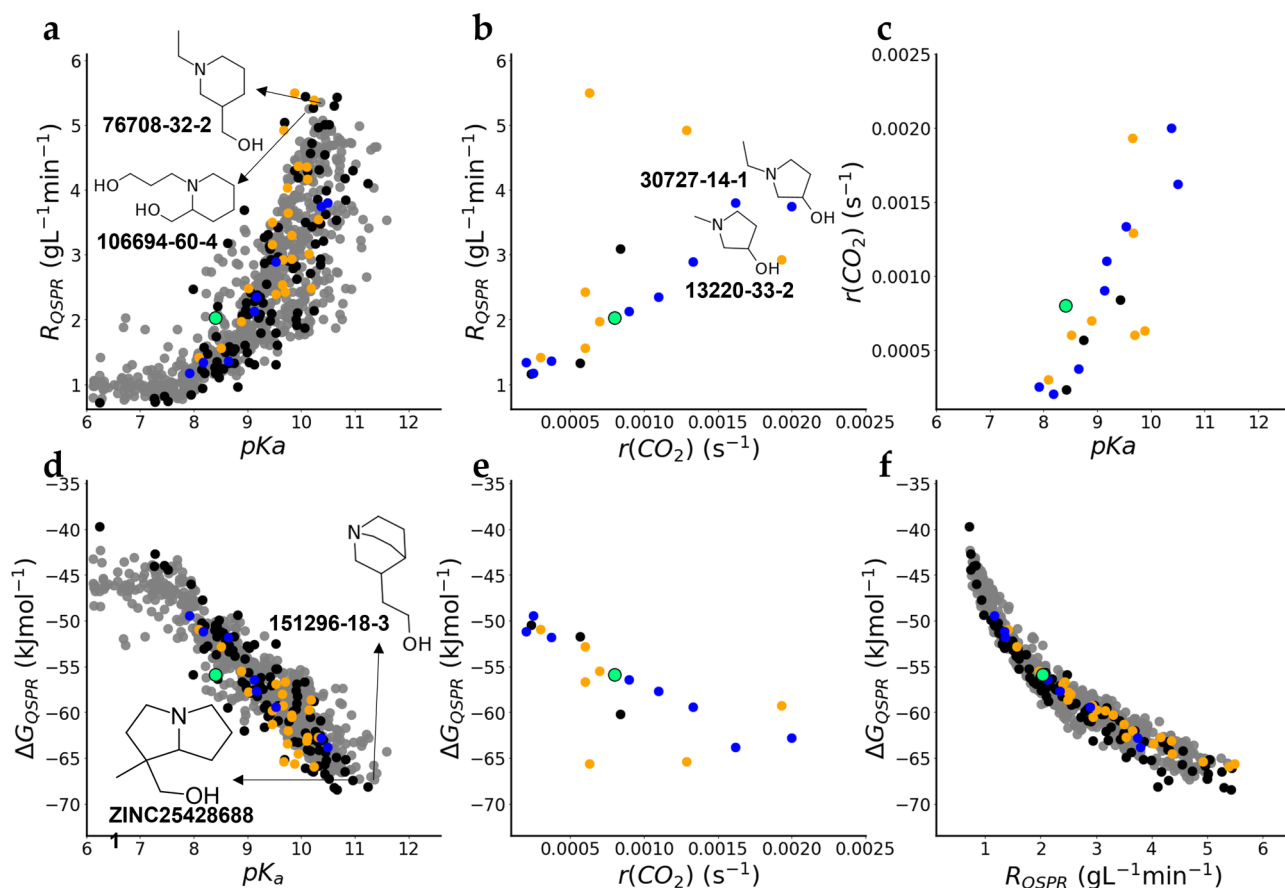


Fig. 3 Virtual screening of tertiary amines and experimental validation. **a** Absorption rates predicted by the QSPR model (R_{QSPR}) vs predicted pK_a values. **b** R_{QSPR} vs experimentally measured absorption rate ($r(\text{CO}_2)$). **c** $r(\text{CO}_2)$ vs predicted pK_a . **d** Free energies of absorption predicted by QSPR model (ΔG_{QSPR}) vs predicted pK_a . **e** ΔG_{QSPR} vs $r(\text{CO}_2)$. **f** ΔG_{QSPR} vs R_{QSPR} . Amines present in the initial dataset from Chowdhury et al.^{19,20} are shown in orange. Amines selected for MD simulations in the present work are shown in black. The industrially used reference compound (MDEA) is shown in green. Eight novel amines which were not present in the training set are shown in blue. The CAS numbers of the most perspective compounds are shown.

eXtreme Gradient Boosting (XGBoost)²⁹ machine-learning algorithms on a merged subset of ISIDA fragments and descriptors generated with the OPERA tool. Although the predictive accuracy in terms of RMSE is of the same order of magnitude as in Kuenemann et al.¹³ for absorption rates (Supplementary Table 2 and Supplementary Fig. 1), the applicability domain of our models is much larger, since the training set contained three times more compounds. It is worth noting that a QSPR model which did not allow one to achieve an excellent accuracy can still be useful for ranking the amines from the large compounds databases^{13,30}. Therefore, we retrieved from the public database ZINC²³ the tertiary amines which were not too large ($M_w \leq 250 \text{ gmol}^{-1}$), not too lipophilic ($-1 \leq \text{clogP} \leq 1$), and readily available from suppliers. In total, more than 800 amines were screened virtually. Numerous amines outranking MDEA in terms of the predicted absorption rates (R_{QSPR}) were identified (Fig. 3a). For example, various substituted piperidines were among the compounds with the largest R_{QSPR} (Fig. 3a).

Experimental measurement of the CO_2 absorption kinetics. An experimental setup was put in place to measure and compare the rate of CO_2 absorption in aqueous tertiary amines. For each experiment, the same initial amount of CO_2 was set in contact with the solvent and the evolution toward equilibrium of the partial pressure of CO_2 in the gas phase was measured over time. The slope of the absorption curve at the time at which 50% of the

CO_2 was absorbed (with respect to the equilibrium value) was calculated ($r(\text{CO}_2)$). It is a measure of the rate of CO_2 absorption. Eighteen amines comprising 7 amines from the initial set of 24 amines from Chowdhury et al.²⁰, 3 amines from the diverse dataset of 100 amines, and 8 novel amines that were never present in the training set were purchased and an assessment of their absorption rate was performed (Fig. 3b, c, e and Supplementary Tables 3 and 4). Both ΔG_{QSPR} and absorption rates R_{QSPR} were highly correlated with $r(\text{CO}_2)$ for eight novel amines (Spearman ρ 0.93) as well as the predicted pK_a values. Five out of eight purchased amines absorbed CO_2 faster than MDEA. Two amines: 1-methyl- and 1-ethyl-3-pyrrolidinol (EPOL) were especially effective. These compounds represent an interesting class of the tertiary amines, which to our knowledge have not been explored yet.

While tertiary amines like the standard MDEA are often used for high-pressure natural gas treatment, they are not suitable for low-pressure anthropogenic CO_2 removal due to the low CO_2 absorption rate. Activators such as piperazine can be added to enhance the CO_2 absorption rate. The impact of piperazine is shown in Fig. 4 for two amines, namely MDEA and EPOL. The latter is a tertiary amine that has been selected for its fast CO_2 absorption rate following the virtual screening. In the absence of piperazine EPOL absorbs CO_2 much faster than MDEA. The addition of piperazine significantly enhances the CO_2 absorption rates with EPOL + PZ showing the fastest absorption.

In conclusion, a methodology for computer-aided design of tertiary amines effectively absorbing CO_2 was suggested in this

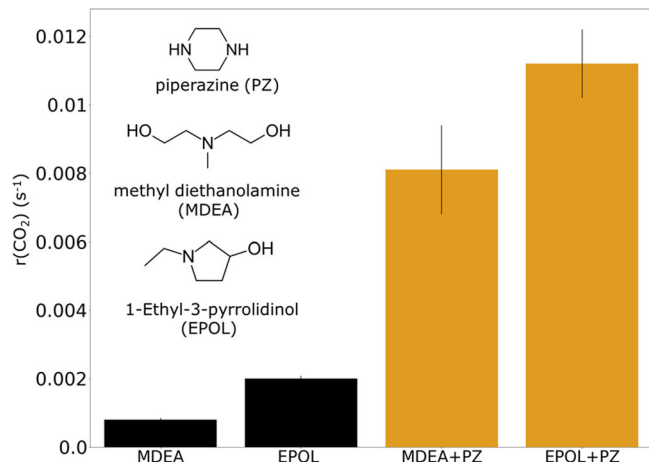


Fig. 4 Experimental kinetic measurements with piperazine.

Experimentally measured CO_2 absorption rate ($r(\text{CO}_2)$) of standard MDEA and EPOL, a new amine suggested in this work, and their mixtures with piperazine (+PZ). Aqueous alkanolamine mixtures contain 13 mol% amine and water and mixtures with PZ contain 11 mol% amine, 2.5 mol% PZ and water. Standard deviations of the values are shown as error bars.

paper. The methodology is based on the combination of state-of-the-art molecular dynamics simulations that generate a sufficiently large dataset that are used as an input for machine-learning modelling followed by large-scale virtual screening. In parallel, the approach is experimentally validated. It allowed the identification of amines that absorb CO_2 faster than those currently used in the industry. Since the development of an optimal solvent is a multi-objective task, we believe that the proposed methodology can be provisionally repurposed to application for modeling of other industrially important properties of alkanolamine-based solvents.

Methods

Molecular simulations. The approach developed recently and described in Rozanska et al.¹⁹, was used to compute the rates of CO_2 absorption in aqueous amine solvents (see Supplementary Methods), which relies primarily on the solvation properties of OH^- , CO_2 , and HCO_3^- . In this model, the tertiary amine solely acts as a base.

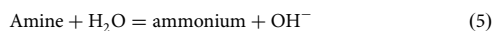
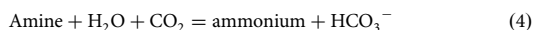
$$R_{\text{MD}} = A(T) \times \exp\left(\frac{-\Delta G^\ddagger}{RT}\right) \times [\text{CO}_2][\text{OH}^-] \quad (1)$$

The rates are obtained from Eq. (1) where R_{MD} is the absorption rate, $[\text{CO}_2]$ and $[\text{OH}^-]$ are the concentrations of carbon dioxide molecules and hydroxyl anions, respectively, ΔG^\ddagger is the Gibbs free energy barrier of the reaction $\text{CO}_2 + \text{OH}^-$ to HCO_3^- , RT is the macroscopic thermodynamic energy unit, where R is the universal gas constant and T the absolute temperature, and $A(T)$ is a temperature-dependent pre-exponential factor. In Eq. (1), ΔG^\ddagger is obtained from a Polanyi–Evans relation with as input the energy differences of solvation of $\text{OH}^- + \text{CO}_2$ (reactants) and HCO_3^- (product) computed in the 124 aqueous amine solvents. The concentrations $[\text{CO}_2][\text{OH}^-]$ are obtained numerically solving pH equations, and $A(T)$ is fitted using the experimental rates of the reaction $\text{CO}_2 + \text{OH}^-$ in ten aqueous amine solvents. The Polanyi–Evans relation between ΔG^\ddagger and energy differences of solvation, ΔG , of $\text{OH}^- + \text{CO}_2$ and HCO_3^- is given by Eq. (2).

$$\Delta G^\ddagger = a\Delta G(T) + b \quad (2)$$

where a and b are fitted to reproduce the experimental rates in pure water and ten aqueous amine solvents and $\Delta G(T)$ is the energy difference of solvation of $\text{OH}^- + \text{CO}_2$ and HCO_3^- obtained from molecular dynamics simulations. Additional details and the values for $A(T)$, a , and b can be found in Rozanska et al.¹⁹.

For the calculation of the regeneration energy, the following three reactions are considered:



The free energy of absorption is ΔG_4 ($=\Delta G_{\text{MD}}$ in Fig. 2) $=\Delta G_3 + \Delta G_5$ with ΔG_3

calculated from the molecular simulations ($\Delta G(T)$ in Eq. (2)) in every aqueous amine and ΔG_5 calculated from the amine pK_a .

Quantitative structure–property relationship modeling. All compound structures were standardized using RDKit³¹ nodes in KNIME³². The standardization procedure included aromatization, stereochemistry depletion, removal of salts/solvents, neutralization, removal of explicit hydrogens. Standardized structures for 124 amines are given in Supplementary Table 1 and at https://github.com/AxelRolvov/CO2_chemical_solvents.

In all, 193 different ISIDA fragment descriptors were generated using the Fragmentor17 software^{26,27}. These fragments represent either sequences (the shortest topological paths with an explicit presentation of all atoms and bonds), atom pairs, or triplets (all the possible combinations of three atoms in a graph with the topological distance between each pair indicated).

Various physicochemical properties (pK_a , $\log P$, melting and boiling points, vapor pressure, water solubility, etc.) and several substructural fragments counts (ring count, heavy atom count, etc.) used as descriptors, were calculated using OPERA v.2.6²⁵.

All descriptors used in this work are available at https://github.com/AxelRolvov/CO2_chemical_solvents.

Prior to the application of machine-learning algorithms R_{MD} and ΔG_{MD} values were transformed to a logarithmic scale, i.e., the negative value of decimal logarithm was taken ($-\log_{10}R_{\text{MD}}$, $-\log_{10}(-\Delta G_{\text{MD}})$).

Random forest (RF): RF algorithm²⁸ implemented in sci-kit learn library (v. 0.22.1)³³, was used. The following hyperparameters were optimized (grid search): number of trees (100, 300, 1000), number of features (all features, one-third of all features, \log_2 of the number of features), the maximum depth of the tree (10, 30, full tree), bootstrapping (with and without the usage of bootstrap samples for building the tree).

XGBoost (XGB): XGBoost algorithm²⁹ as implemented in XGBoost python module (v.1.2.0; https://xgboost.readthedocs.io/en/latest/python/python_intro.html) was used. The following hyperparameters were tuned during optimization (grid search): number of trees (50, 100, 300, 500), number of features (all features, 70% of all features), number of samples (all samples, 70% of all samples), the maximum depth of the tree (5, 20, full tree), learning rate (0.3, 0.1, 0.5, 0.05). All other parameters were left as default.

Support vector regression (SVR): SVR algorithm³⁴ implemented in sci-kit learn library (v. 0.22.1), was used. The descriptors were scaled to the [0,1] range before applying the algorithm. The following hyperparameters were tuned during optimization (grid search): kernel (linear, rbf, poly, sigmoid), kernel coefficient (1, 0.1, 0.01, 0.001, 0.0001), regularization parameter (0.1, 1, 10, 100, 1000).

The modeling workflow was implemented using the sci-kit learn library (v. 0.22.1) in Python 3.7 scripting language (Supplementary Fig. 2). Identical modeling workflows were used for modeling absorption rates (R_{MD}) and energies of absorption (ΔG_{MD}). The values were expressed as negative logarithms of base 10. At the first stage of the modeling, a machine-learning algorithm: RF, SVR, and XGB were tested in fivefold cross-validation, which was repeated five times. For each descriptor set, the model's measures of performance were calculated and several models with a coefficient of determination $Q^2_{\text{CV}} \geq 0.6$ for (R_{MD}) and $Q^2_{\text{CV}} \geq 0.7$ for (ΔG_{MD}) were selected for consensus modeling. Consensus models were built for each descriptor type separately. In order to assess a propensity to predict data never seen during the training of the model, a nested cross-validation procedure³⁵ has been implemented. Here the method hyperparameters were found by optimizing the model performance in the fivefold cross-validation inner loop, while prediction was made for the test set from the outer loop, which represent a fold of the outer fivefold cross-validation cycle. To avoid a bias with the compounds numbering in the parent set, this procedure was repeated five times after reshuffling of the compounds. In such a way, the overall performance of the model (Q^2_{NCV} , $RMSE_{\text{NCV}}$, MAE_{NCV}) were estimated as an average of related statistical parameters obtained for each (out of 5) individual cross-validation loop.

Equations (6–8) were used to calculate the measures of the model's performance in cross-validation:

$$Q^2_{\text{CV}} = \frac{\sum_{j=1}^5 \left(1 - \frac{\sum_{i=1}^n (y_{i,\text{exp}} - y_{i,\text{pred}})^2}{\sum_{i=1}^n (y_{i,\text{exp}} - \bar{y})^2}\right)}{5} \quad (6)$$

$$RMSE_{\text{CV}} = \frac{\sum_{j=1}^5 \sqrt{\frac{\sum_{i=1}^n (y_{i,\text{exp}} - y_{i,\text{pred}})^2}{n}}}{5} \quad (7)$$

$$MAE_{\text{CV}} = \frac{\sum_{j=1}^5 \sum_{i=1}^n \frac{|y_{i,\text{exp}} - y_{i,\text{pred}}|}{n}}{5} \quad (8)$$

Above, n is the number of compounds in the learning set, $y_{i,\text{exp}}$, $y_{i,\text{pred}}$ experimental and values predicted in fivefold cross-validation for compound i from the learning set, j is the index of the repetition of the tenfold cross-validation procedure.

Each of the selected models was then associated with an Applicability Domain (AD), defined as a boundary box. The pool of selected models extracted from the given dataset can now be used as a consensus predictor, returning for each input solvent candidate a mean value of solubility estimates and its standard deviation,

taken over the predictions returned by each model in the pool or, alternatively, over the predictions returned by only those models having the candidate within their AD.

Outlying data points were defined as the data points, for which absolute errors ($|\chi_{\text{exp}} - \chi_{\text{pred}}|$) from cross-validation were larger than $2 \times \text{RMSE}_{\text{CV}}$ threshold.

The absence of chance correlation was checked through the Y-randomization procedure. A Y-randomization test was performed in the following way: $-\log_{10} \chi$ values (y values) were shuffled, models were built using shuffled values and the values from the corresponding cross-validation test set were calculated. This procedure was repeated 100 times for each fold and the maximum values of the out-of-bag coefficient of determination were reported.

A library for virtual screening was performed in the following way. At first, all compounds from ZINC database with molecular weight no larger than 250 g/mol and calculated logP in the range of $(-1, 1)$ were retrieved. Structures were standardized and then filtered. All compounds which did not contain tertiary amines, compounds, containing double bonds, aromatic rings, primary or secondary amine groups, ketones and sulfur-containing compounds except for thiols and thioethers were removed. Structures of screened compounds as well as predicted values are available at https://github.com/AxelRolov/CO2_chemical_solvents.

Experimental measurement of CO₂ absorption rates. To measure the kinetics of absorption and desorption of acid gases in aqueous amine solutions, a thermoregulated constant interfacial area Lewis-type reactor cell was used³⁶. The reactor (Supplementary Figs. 3–6) is equipped with an internal stirring system (magnetic stirrer) with the external motor. The operator needs to take care to select the speed of stirring without disturbing the interface (interface must be flat). Temperature is given by two platinum probes located at the upper and lower flanges (with the possibility to determine the gradient of temperature). The cell is immersed in a liquid bath. An electric resistor is introduced into the upper flange to control the gradient of temperature and avoid condensation of water and amine. Two capillary samplers are adapted to sample the vapor phase. The capillary samplers (ROLSI®, Armines' patent) are capable of withdrawing and sending micro samples to a gas chromatograph without perturbing the equilibrium conditions over numerous samplings, thus leading to repeatable and reliable results. Analytical work was carried out using a gas chromatograph (PERICROM model PR2100, France) equipped with a thermal conductivity detector (TCD) connected to a data software system. Helium is used as the carrier gas in this experiment. The model of the GC column is Porapak R (Porapak R 80/100 mesh, 1 m × 2 mm ID Silcosteel). Each ROLSI® sampler is connected to a TCD. A tube allows either to evacuate or to introduce CO₂ from or into the cell. The kinetics of gas absorption are determined by recording the pressure drop through a calibrated pressure transducer. A computer equipped with data acquisition system records the pressure as a function of time.

The experimental procedure is the following:

The desired amount of solvent is introduced into the cell. The density obtained using a low-pressure vibrating tube densitometer (Anton Paar DSA 5000) is used to determine the exact mole number of solvent.

At least 5 bar of methane is added. We add methane because with this configuration, it is not possible to sample at pressures lower than GC carrier gas pressure.

CO₂ is added from the thermal press. We record pressure and temperature before and after the loading (see Supplementary Fig. 7 as an example). It permits to calculate very accurately the mole number of CO₂ introduced and so, we can estimate very accurately the loadings of CO₂.

The experimental method³⁶ is similar to the one used to calculate the solubility of CO₂ in alkanolamine amine solution at equilibrium. The method considered is based on the "static-synthetic method". More details concerning the method are presented in the Supplementary Methods.

During the absorption of the CO₂, we take samples to follow the evolution of the vapor composition (and so CO₂ partial pressure) as a function of time. When the equilibrium is reached (constant pressure and constant temperature), the vapor phase composition is determined.

We have used the GERG 2008 Equation of state³⁷ implemented in REFPROP 10.0³⁸ to estimate the densities of the vapor phase which is a mixture of CO₂ and CH₄.

The calculation of the acid gas solubility in the solvent is based on mass balance. The volume of the liquid phase is obtained by considering the mole number of solvent introduced and its density at the temperature of measurement.

$$V^{\text{L}} = \frac{n_{\text{solvent}}}{\rho(T_{\text{cell}})} \quad (9)$$

Consequently, the volume of the vapor phase is calculated by difference between the total volume and the volume of the liquid phase.

$$V^{\text{V}} = V^{\text{T}} - V^{\text{L}} \quad (10)$$

If the introduction of the solute doesn't modify the level of the liquid interface in the equilibrium cell, we can consider Eq. (11).

$$V^{\text{L}} = \pi r_{\text{cell}}^2 h_{\text{liq}} \quad (11)$$

Where r_{cell} is the radius of the equilibrium cell, h_{liq} the level of the vapor liquid interface.

The mole number of solute in the vapor phase is calculated by considering the density of the gas at the temperature and pressure of solute ($P_{\text{solute}} = P_{\text{cell}} - P_{\text{solvent}}^{\text{sat}}$). REFPROP v10.0 is used to calculate this density $\rho^{\text{V}}(T_{\text{cell}}, P_{\text{solute}})$. In the case of a mixture, the global composition needs to be considered $\rho^{\text{V}}(T_{\text{cell}}, P_{\text{solute}}, y)$.

The volume of the vapor phase is used to calculate the mole number of solute in the vapor phase (Eq. (12)).

$$n^{\text{V}} = V^{\text{V}} \rho^{\text{V}}(T_{\text{cell}}, P_{\text{solute}}) \quad (12)$$

In the case of a mixture, the same equation is used to calculate the total mole number of solute in the vapor phase.

So, the mole number of solute in the liquid phase is determined by considering Eq. (13).

$$n^{\text{L}} = n^{\text{T}} - n^{\text{V}} \quad (13)$$

In the case of the mixture, the mole number of each species is calculated by considering the global composition of the mixture (z) and the composition of the vapor phase (y), Eq. (14).

$$n_i^{\text{L}} = z_i n^{\text{T}} - y_i n^{\text{V}} \quad (14)$$

The solubility is determined with Eq. (15).

$$x_i = \frac{n_i}{\sum n_j} \quad (15)$$

Data availability

All the experimental data are available in Supplementary Materials and at https://github.com/AxelRolov/CO2_chemical_solvents. Structures of compounds, descriptors and predicted values are also available at https://github.com/AxelRolov/CO2_chemical_solvents. The data are also deposited into a DOI-minting repository ZENODO: <https://doi.org/10.5281/zenodo.6010667>.

Code availability

Jupyter notebooks containing the Python code used for model building, evaluation and virtual screening are available at https://github.com/AxelRolov/CO2_chemical_solvents. The code is also deposited into a DOI-minting repository ZENODO: <https://doi.org/10.5281/zenodo.6010667>. Python libraries used for machine-learning and OPERA software are freely available. ISIDA-Fragmentor is available upon reasonable request to Prof. Alexandre Varnek.

Received: 21 October 2021; Accepted: 23 February 2022;

Published online: 18 March 2022

References

1. Birol, F., Cozzi, L., & Gül, T. Net Zero by 2050—Analysis. *IEA* <https://www.iea.org/reports/net-zero-by-2050> (2021).
2. Hepburn, C. et al. The technological and economic prospects for CO₂ utilization and removal. *Nature* **575**, 87–97 (2019).
3. Bui, M. et al. Carbon capture and storage (CCS): the way forward. *Energy Environ. Sci.* **11**, 1062–1176 (2018).
4. Rochelle, G. T. Amine scrubbing for CO₂ capture. *Science* **325**, 1652–1654 (2009).
5. Brickett, L. *Carbon Dioxide Capture Handbook*. (US Department of Energy (DOE)/NETL, 2015). <https://www.netl.doe.gov/sites/default/files/netl-file/Carbon-Dioxide-Capture-Handbook-2015.pdf>.
6. Smit, B. Carbon Capture and Storage: introductory lecture. *Faraday Discuss* **192**, 9–25 (2016).
7. Borhani, T. N. & Wang, M. Role of solvents in CO₂ capture processes: the review of selection and design methods. *Renew. Sustain. Energy Rev.* **114**, 109299 (2019).
8. de Meyer, F. & Bignaud, C. The use of catalysis for faster CO₂ absorption and energy-efficient solvent regeneration: an industry-focused critical review. *Chem. Eng. J.* **428**, 131264 (2022).
9. Li, L. et al. Amine blends using concentrated piperazine. *Energy Procedia* **37**, 353–369 (2013).
10. Lin, L.-C. et al. In silico screening of carbon-capture materials. *Nat. Mater.* **11**, 633–641 (2012).
11. Boyd, P. G. et al. Data-driven design of metal–organic frameworks for wet flue gas CO₂ capture. *Nature* **576**, 253–256 (2019).
12. Conway, W. et al. Designer amines for post combustion CO₂ capture processes. *Energy Procedia* **63**, 1827–1834 (2014).
13. Kuenemann, M. A. & Fourches, D. Cheminformatics modeling of amine solutions for assessing their CO₂ absorption properties. *Mol. Inf.* **36**, 1600143 (2017).

14. Khareshi, S., Riahi, S., Mohammadi-Khanaposhtani, M. & Shokrollahzadeh, H. Prediction of amines capacity for carbon dioxide absorption based on structural characteristics. *Ind. Eng. Chem. Res.* **58**, 8763–8771 (2019).
15. Rezaei, B., Riahi, S. & Gorji, A. E. Molecular investigation of amine performance in the carbon capture process: least squares support vector machine approach. *Korean J. Chem. Eng.* **37**, 72–79 (2020).
16. Cheng, J. et al. Quantitative relationship between CO₂ absorption capacity and amine water system: DFT, statistical, and experimental study. *Ind. Eng. Chem. Res.* **58**, 13848–13857 (2019).
17. Gonfa, G., Bustam, M. A. & Shariff, A. M. Quantum-chemical-based quantitative structure-activity relationships for estimation of CO₂ absorption/desorption capacities of amine-based absorbents. *Int. J. Greenh. Gas. Control* **49**, 372–378 (2016).
18. Porcheron, F. et al. Graph machine based-QSAR approach for modeling thermodynamic properties of amines: application to CO₂ capture in postcombustion. *Oil Gas. Sci. Technol. – Rev. D'IFP Energ. Nouv.* **68**, 469–486 (2013).
19. Rozanska, X., Wimmer, E. & de Meyer, F. Quantitative kinetic model of CO₂ absorption in aqueous tertiary amine solvents. *J. Chem. Inf. Model.* **61**, 1814–1824 (2021).
20. Chowdhury, F. A., Yamada, H., Higashii, T., Goto, K. & Onoda, M. CO₂ capture by tertiary amine absorbents: a performance comparison study. *Ind. Eng. Chem. Res.* **52**, 8323–8331 (2013).
21. Kim, S. et al. PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.* **49**, D1388–D1395 (2021).
22. Kim, S. et al. PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.* **49**, D1388–D1395 (2021).
23. Sterling, T. & Irwin, J. J. ZINC 15—ligand discovery for everyone. *J. Chem. Inf. Model.* **55**, 2324–2337 (2015).
24. John J. Irwin & Brian K. Shoichet. ZINC – A Free Database of Commercially Available Compounds for Virtual Screening *J. Chem. Inf. Model.* **45**, 177–182 (2005).
25. Mansouri, K., Grulke, C. M., Judson, R. S. & Williams, A. J. OPERA models for predicting physicochemical properties and environmental fate endpoints. *J. Cheminformatics* **10**, 10 (2018).
26. Varnek, A. et al. ISIDA—platform for virtual screening based on fragment and pharmacophoric descriptors. *Curr. Comput. Aided-Drug Des.* **4**, 191–198 (2008).
27. Ruggiu, F., Marcou, G., Varnek, A. & Horvath, D. ISIDA property-labelled fragment descriptors. *Mol. Inf.* **29**, 855–868 (2010).
28. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
29. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, 2016).
30. Tropsha, A., Gramatica, P. & Gombar, V. K. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* **22**, 69–77 (2003).
31. Landrum, G. *RDKit: Open-source cheminformatics*; <http://www.rdkit.org> (2021).
32. Berthold, M. R. et al. KNIME - the Konstanz Information Miner: Version 2.0 and Beyond *SIGKDD Explor. Newsl.* **11**, 26–31 (ACM, New York, NY, USA, 2009).
33. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
34. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
35. Baumann, D. & Baumann, K. Reliable estimation of prediction errors for QSAR models under model uncertainty using double cross-validation. *J. Cheminformatics* **6**, 47 (2014).
36. Coquelet, C., Valtz, A. & Thèveveau, P. *Experimental Determination of Thermophysical Properties of Working Fluids for ORC Applications. Organic Rankine Cycles for Waste Heat Recovery - Analysis and Applications* (IntechOpen, 2019).
37. Kunz, O. & Wagner, W. The GERG-2008 wide-range equation of state for natural gases and other mixtures: an expansion of GERG-2004. *J. Chem. Eng. Data* **57**, 3032–3091 (2012).
38. Lemmon, E.W., Bell, I.H., Huber, M.L. & McLinden, M.O. NIST standard reference database 23: reference fluid thermodynamic and transport properties-refprop, version 10.0, national institute of standards and technology, standard reference data program, Gaithersburg, <https://doi.org/10.18434/T4/1502528> (2018).

Acknowledgements

This work was supported by the Carbon Capture Utilization and Storage (CCUS) transverse R&D program from TotalEnergies S.E.

Author contributions

A.A.O. performed machine learning, analyzed, interpreted the data, and contributed to the writing of the manuscript. X.R. and E.W. performed the molecular simulations. A.Valtz and C.C. performed the experimental part of the work. G.M. and D.H. contributed to the machine-learning models. B.P. contributed to the planning of the research. A. Varnek conceived, planned, and guided the part of the research related to building machine-learning models. F.D.M. conceived, planned, guided the research, analyzed, and interpreted the data, and wrote the manuscript. All authors critically analyzed data, edited, and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42004-022-00654-y>.

Correspondence and requests for materials should be addressed to Alexandre Varnek or Frédéric de Meyer.

Peer review information *Communications Chemistry* thanks Agilio Padua and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022