Computer-Aided Design of New Physical Solvents for Hydrogen Sulfide Absorption

Alexey Orlov, Gilles Marcou, Dragos Horvath, Alvaro Echeverria Cabodevilla, Alexandre Varnek, Frédérick de Meyer

Computer-aided design of new physical solvents for hydrogen sulfide absorption

Alexey A. Orlov.[§], Gilles Marcou[§], Dragos Horvath[§], Alvaro Echeverria Cabodevilla†, Alexandre Varnek[§], Frédérick de Meyer[†‡,3*].*

[§] Laboratory of Chemoinformatics, Faculty of Chemistry, University of Strasbourg, Strasbourg, 67081 France

† TOTAL SA, Total Exploration Production, Development and Support to Operations, Liquefied Natural Gas – Acid Gas Entity, Paris, 92078 France

‡ MINES ParisTech, PSL University, Centre de thermodynamique des procédés (CTP), 35 rue St Honoré 77300 Fontainebleau, France

*corresponding authors

ABSTRACT

Treatment of hydrogen sulfide (H_2S) is important in many industrial processes including oil refineries, natural and biogas processing, coal gasification. The most mature technology for the selective H_2S capturing is based on its absorption by chemical or physical solvents. However, only several compounds are currently used as physical solvents in industry, and the search for the new ones is an important task. The experimental screening of physical solvents requires a lot of time and resources, while solubility modeling might enable one to reduce the number of solvents for the experimental evaluation. In this study, a workflow for the *in silico* discovery of new physical solvents for H_2S absorption was suggested and experimentally validated. A dataset composed of 99 H_2S physical solvents was collected and predictive quantitative structure-property relationships for H_2S solubility were built using random forest algorithm and two types of molecular descriptors: ISIDA fragments and quantum-chemical descriptors. Virtual screening of industrially produced chemicals and their structural analogs enabled to identify the ones with high predicted solubility values. They can be suggested as starting points for further exploration of H_2S physical solvents chemical space. The predicted solubility value for one of the compounds found in virtual screening, 1,3-Dimethyl-2-imidazolidinone, was confirmed experimentally.

INTRODUCTION

Hydrogen sulfide (H_2S) is a colorless highly toxic gas. It exists naturally in crude petroleum, natural gas, and biogas. The main anthropogenic sources of H_2S are oil refineries and coal gasification, wherein a conversion of the sulfur content of these resources into H_2S occurs[1]. For any gas stream, including natural gas, biogas, or syngas in order to be useful for energy or chemical industrial application, unacceptable amounts of H_2S should be removed. Numerous approaches for H_2S capturing were suggested: absorption with chemical and physical solvents, ionic liquids, adsorption with metals, metal oxides and metal-organic frameworks, zeolites, membrane separation, cryogenic distillation.[1] However, the most common industrial approaches for H_2S removal are based on absorption with either chemical solvents (alkanolamines solutions), which rely on the reversible chemical reaction between a solvent and a gas, or physical solvents (methanol, N-methyl-2-pyrrolidone, polyethylene glycols ethers) in which no chemical reaction occurs.[1] Purely physical solvents are usually used for bulk H_2S removal in cases when the amount of H_2S is sufficiently large.[1] Although chemical and physical gas absorption is a mature process, existing for nearly 100 years, only a limited number of physical solvents is being currently used in industry and the search for new perspective physical solvents for H_2S capture is an important task.

Experimental screening of physical solvents is a time and cost consuming process. In order to reduce the number of solvents being subjected to experimental measurements, predictive modeling enabling to suggest the most promising solvents for the experimental evaluation can be utilized. In the pioneering work of Bryk et al.[2] the authors used a multiple linear regression approach with six experimentally determined properties (refractive index, dielectric constant, Palm basicity, Reichardt electrophilicity, Hildebrand solubility parameter, molar volume) as descriptors, in order to model H_2S solubility in 49 solvents at 293.15 K, and in 11 solvents at 298.15 K. The calculations for all 49

solvents resulted in equations with low multiple correlation coefficient and sequential rejection of the outliers (dioxane, ethyl cellosolve, cyclohexane, and dimethylformamide) was performed. In this way, a six-parameter equation with a sufficient fitting accuracy for the H_2S solubilities in 45 solvents was obtained, although there was no validation on the external set performed. An analysis of the signs and significances of the terms in the equations obtained by the authors showed that the major parameters that control H_2S solubility are the cohesive energy density and the solvent basicity. The former decreases the solubility, since the more associated the solvent is, the more energy is consumed to incorporate a foreign molecule into the structure of the liquid.

The main disadvantage of the modeling approaches based on experimentally determined physico-chemical parameters is the limited number of compounds for which the parameters are available, and thus the narrow set of solvents for which the prediction is possible. Although these physico-chemical parameters can, in turn, be predicted,[3,4] the accuracy of such predictions was shown to be low. An alternative approach is based on the chemoinformatics-driven methodology, wherein only some parameters directly derived from chemical structures are required. Technically, structures are encoded by real-value vectors of molecular descriptors used as variables in quantitative structure-property relationships (QSPR) [5] built with the help of machine learning algorithms. To our knowledge, only one conference paper related to QSPR modeling of H_2S solubility has been published so far [6]. In this paper, multiple linear regression method with genetic algorithm-based features selection and descriptors calculated with the Dragon software[7] were used to model H_2S solubility at 293.15 K on a set of 44 solvents.[6]

In this *proof-of-concept* paper, we investigated the possibility to use machine learning for the *in silico* design of new H_2S physical solvents. A dataset consisting of 99 mole fraction solubility values (χ) at 298.15 K and 1 atm extracted from the IUPAC report, recent scientific papers and patents and two data points experimentally measured by Total S.A. was compiled. The models for $-\log(\chi)$ were built using random forest algorithm and two different types of descriptors: ISIDA

fragments[8] and some parameters issued from quantum chemical calculations[9]. Models with reasonable predictive performance were used in virtual screening of the industrially produced chemicals. It enabled to suggest several potent H_2S absorbents one of which, 1,3-Dimethyl-2-imidazolidinone, was then studied experimentally. Experiment solubility value was found close to the predicted one, thus showing good predictive performance of the obtained model.

MATERIALS AND METHODS

Data collection and preprocessing

Data points (χ values) for H_2S solubility were collected from several sources: IUPAC solubility report[10], patents[11,12], and papers[13-32]. In cases, where there were no data available at 298.15 K, but several data points at other temperatures were available, the values were extrapolated or interpolated. The data points have been extrapolated only if there were at least more than two data points at different temperatures available and the deviation of the closest temperature among these points was not larger than 10 K. In cases where the pressure needed to be adjusted to 1 atm, solvent vapor pressure was estimated by the Antoine equation. Extrapolation of the data to 298.15 K was performed using either equations suggested in the original publications or eq. 1:

$$\ln\chi = A + B \times \ln(T), \qquad (1)$$

where χ – mole fraction solubility value, T – temperature (K), A, B – constants.

Mole fraction solubilities were converted to the Kuenen coefficients *S* using eq. 2:

$$S = \frac{R \times T \times P}{Mw} \times \frac{\chi}{1-\chi}, \qquad (2)$$

S – Kuenen coefficient ($Nm^3 \times kg^{-1}$), R – ideal gas constant (8.314 $m^3 \times Pa \times K^{-1} \times mol^{-1}$), T and P – standard temperature and pressure (273.15 K and 101.325 kPa), Mw – molecular weight of compound (kg/mol), χ – mole fraction solubility value.

According to the IUPAC's definition[10], Kuenen coefficient is the volume of saturating gas at 273.15 K and 1 atm pressure, which is dissolved by unit mass of pure solvent at the temperature of measurement and partial pressure of 1 atm. This parameter is widely used in industry applications, as it enables one to directly estimate the efficiency of the particular solvent related to its cost and dimensions of the required industrial unit (design-capital expenses cost CAPEX). Here, Kuenen coefficients were used for the data analysis and models interpretation.

For the modeling, all χ values were transformed to logarithmic scale, i.e. negative value of decimal logarithm was taken.

Chemical structures standardization

All compounds structures were standardized using in-house standardization procedures based on the RDKit tool,[33] which included aromatization, stereochemistry depletion, etc.

Molecular descriptors

Two different types of descriptors were used in the modeling: ISIDA fragments[8] and some parameters issued from quantum chemical calculations[9]. 193 different types of ISIDA fragment descriptors were generated using Fragmentor17 software.[8,34] These fragments represent either sequences (the shortest topological paths with explicit presentation of all atoms and bonds), atom pairs or triplets (all the possible combinations of 3 atoms in a graph with the topological distance between each pair indicated).

Quantum chemical descriptors resulted from DFT calculations in gas phase, with model wB97X-D 6-31G* performed with the Spartan 18.0 program.[9] Default QSAR descriptors available in Spartan including energy, dipole moment, E_{HOMO}, E_{LUMO} were calculated.

Machine-learning method.

Random forest (RF) algorithm implemented in sci-kit learn library (v. 0.22.1) was used. The following hyperparameters were tuned during optimization (grid search): number of trees (100, 300, 1000), number of features (all features, one third of all features).

Model training and validation workflow

Modeling workflow was implemented using sci-kit learn library (v. 0.22.1) in python 3.7 scripting language. At the first stage, a random forest (RF) algorithm (sci-kit learn implementation) was applied to the entire dataset (parent set) with various hyperparameters (Figure 1a).

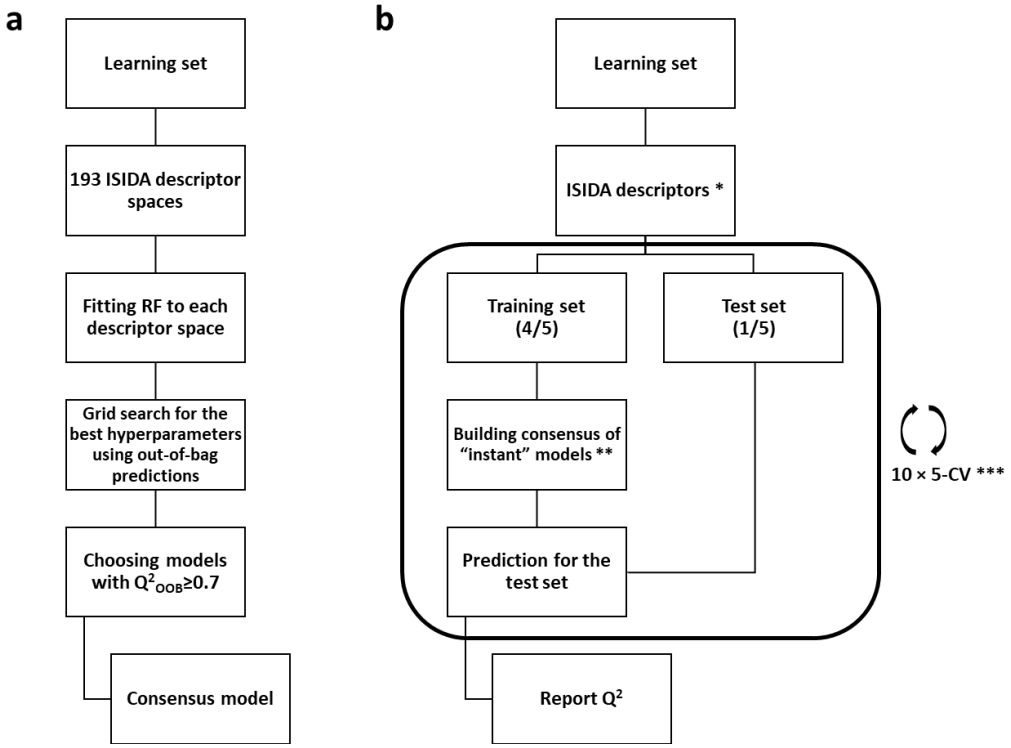

Figure 1. Workflows for **(a)** building ISIDA consensus model and **(b)** nested cross-validation procedure. * Only ISIDA descriptors selected for the consensus model in workflow (a)

are used. ** Each "instant" model corresponds to particular type of descriptors. Its hyperparameters are optimized on related OOB in the grid search *** 5-fold cross-validation repeated 10 times with reshuffling.

We used bootstrap aggregation in the process of building the decision trees. For each training sample -logχ values were calculated as the mean of the values predicted by the trees that did not have this training sample in their bootstrap sample. These out-of-bag predictions were used for the estimation of model's performance. The best model for each descriptor space was chosen according to the largest coefficient of determination for out-of-bag predictions (Q^2_{OOB}). Only the models for which Q^2_{OOB} was no less than 0.7 were saved as the "local" models forming the consensus predictor trained on the given data set. Two other measures ($RMSE_{OOB}$, MAE_{OOB}) were also reported. The following equations were used to calculate the measures:

$$Q^2_{OOB} = 1 - \frac{\sum_{i=1}^{n}(y_{i,exp} - y_{i,pred})^2}{\sum_{i=1}^{n}(y_{i,exp} - \bar{y})^2} \qquad (3)$$

$$RMSE_{OOB} = \sqrt{\sum_{i=1}^{n}\frac{(y_{i,exp} - y_{i,pred})^2}{n}} \qquad (4)$$

$$MAE_{OOB} = \sum_{i=1}^{n}\frac{|y_{i,exp} - y_{i,pred}|}{n} \qquad (5)$$

Above, **n** is the number of compounds in the parent set, $\mathbf{y_{i,exp}}$, $\mathbf{y_{i,pred}}$ experimental and out-of-bag predicted values for compound **i** from the parent set.

Each of the selected models was then associated to an Applicability Domain (AD), defined as boundary box for Spartan descriptors and by the presence of all fragments from the test set in the training set (fragment control) for ISIDA descriptors. The pool of selected models extracted from the given data set 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 (a) over the predictions returned by each model in the pool or, alternatively, (b) over the predictions returned by only those models having the candidate within their AD.

In the above workflow, the models' parameters have been selected in such a way to optimize the predictive performance on OOB data. In order to assess a propensity to predict data never seen during the models training, a nested cross-validation procedure[35] has been implemented. Here the method hyperparameters were found by optimizing the model performance on OOB for each training set in the 5-fold cross-validation loop (see Figure 1b). The fragment control and boundary box applicability domains were applied upon the model application on a test set compounds. In order to avoid a bias with the compounds numbering in the parent set, this procedure was repeated 10 times after compounds reshuffling. In such a way, the overall performance of the model ($Q^2_{NCV\text{-}AD}$, $RMSE_{NCV\text{-}AD}$, $MAE_{NCV\text{-}AD}$) were estimated as an average of related statistical parameters obtained for each (out of 10) individual cross-validation loop.

Outlier analysis

Outlying data points were defined as the data points, for which absolute errors ($|\chi_{exp}-\chi_{pred}|$) for out-of-bag predictions were larger than $2\times RMSE_{NCV\text{-}AD}$ threshold.

Y-randomization test

The absence of chance correlation was checked through the Y-randomization procedure. Y-randomization test was performed in the following way: negative $\log_{10}\chi$ values (y values) were shuffled, random forest models were built using shuffled values and the out-of-bag values were calculated. This procedure was repeated 200 times and the maximum values of out-of-bag coefficient of determination were reported.

Virtual screening

In-house dataset, comprising several hundred industrially produced compounds and their structural analogs, was screened in the following way. Only structures containing the same atoms (C, H, N, O, S, P, halogens) as in the parent set were kept. All structures were standardized and ISIDA descriptors were calculated for them as described above. Then, predictions were made using

the ISIDA consensus model. Compounds, that were inside AD for at least three ISIDA fragment types, were considered as being inside AD of the ISIDA consensus model.

Software implementation

Developed model was implemented into the ISIDA-Predictor software.[8]

Experimental measurement of H_2S solubility

For the synthetic gas solubility measurements (isothermal *P-x* data) a static apparatus was used. In this synthetic method, the system pressure is measured at constant temperature for different overall compositions. The apparatus can be operated at temperatures between 200 K and 500 K and pressures up to 20 MPa.

To determine the global compositions, the quantities of pure substances charged into the stirred equilibrium cell, which is evacuated and placed in a thermostatic liquid bath, need to be known precisely. The purified and degassed solvents are charged into the cell as compressed liquids using thermostatted piston injectors. Then, the gas is added stepwise as a liquefied gas using the same injection pumps or as a gaseous component using a thermo-regulated gas bomb. Knowing the pressure, temperature, and volume of the gas bomb, the amount of gas inside the bomb can be calculated using correlated *PvT* data of the gas. Thus, the injected amount of gas can be obtained from the pressure difference in the bomb before and after each injection.

Since only temperature, pressure, total loadings and total volumes are measured, the compositions of the coexisting phases need to be determined by evaluation of the raw data. From the known amount of solvent, the liquid phase volume is determined using precise information about the density of the liquid solution inside the equilibrium chamber. From the total volume of the cell, the remaining gas phase volume can be calculated precisely. At given equilibrium conditions (temperature, gas phase volume and gas pressure) the amounts of gas in the gas phase and thus, also in the liquid phase are obtained. In this approach, several effects have influence on the resulting

liquid phase compositions. These effects are the small amounts of solvents in the gas phase, the compressibility of the solvent under the gas pressure, the partial molar volume of the dissolved gas and the solvent activity coefficient. All effects are considered in an iterative isothermal and isochoric algorithm by solving the mass and volume balances.

The partial pressure is obtained during the iterative procedure:

$P_{gas} = P_{sys} - P_{solvent}$ (6)

where P_{gas} – partial pressure of the acid gas in the system, P_{sys} – total pressure in the system, $P_{solvent}$ – partial pressure of a solvent vapour.

RESULTS AND DISCUSSION

Data collection, preprocessing and analysis

The core of our dataset was composed from the H_2S solubility data available in the IUPAC report (Volume 32 of the IUPAC solubility data series) representing the most complete and carefully curated source of solubility data.[10] Most of the data points for various solvents were measured at 298.15 K, and therefore, these data were chosen for modeling. As noted by IUPAC's experts, reliability of solubility measurements varies as a function of experimental technique used. Thus, solubility values measured by chromatographic methods are assumed to be less reliable as they are prone to errors due to surface effects. In general, common experimental error estimated by IUPAC's expert varies in the range of ±2% − ±10% of measured solubilities. We relied on the IUPAC's expert opinion on the data reliability in all cases, where it was possible, and retained only those data points, that were considered as reliable. Data from recent publications either at 298.15 K or obtained by extrapolation of the data measured at close temperatures were also added to the dataset.[11-32]

Besides the data collected from IUPAC report and literature, data points for two compounds − hexametapol (HMPA) and thiodiglycol (TDG), for which there were no data available at 298.15 K and 1 atm, were measured by Total and incorporated in our dataset (Figure 2a; Table S1 in Supporting Information). TDG is employed in a commercial mixed chemical/physical solvent formulation for mercaptan rich sour gas treating (HySWEET technology) developed by Total S.A.[36] In principle, the solubility value for TDG can be estimated by extrapolation of Vahidi et al.[26] measurements to 298.15 K. However, taking into account a limited availability of data related to H_2S solubility in sulfur-containing solvents, we decided to independently measure solubility in TDG at 298.15 K.

HMPA is being used as a solvent for polymers, gases and in organic synthesis.[37] The only data available for HMPA is its Henry coefficient (1.61 atm) from Lenoir's paper[24]. As noted by IUPAC expert[10], even if the value for HMPA is reliable it is incorrect to assume a linear variation of mole fraction solubility with partial pressure to 1.013 bar, as it would give a very high mole fraction solubility value (0.62).

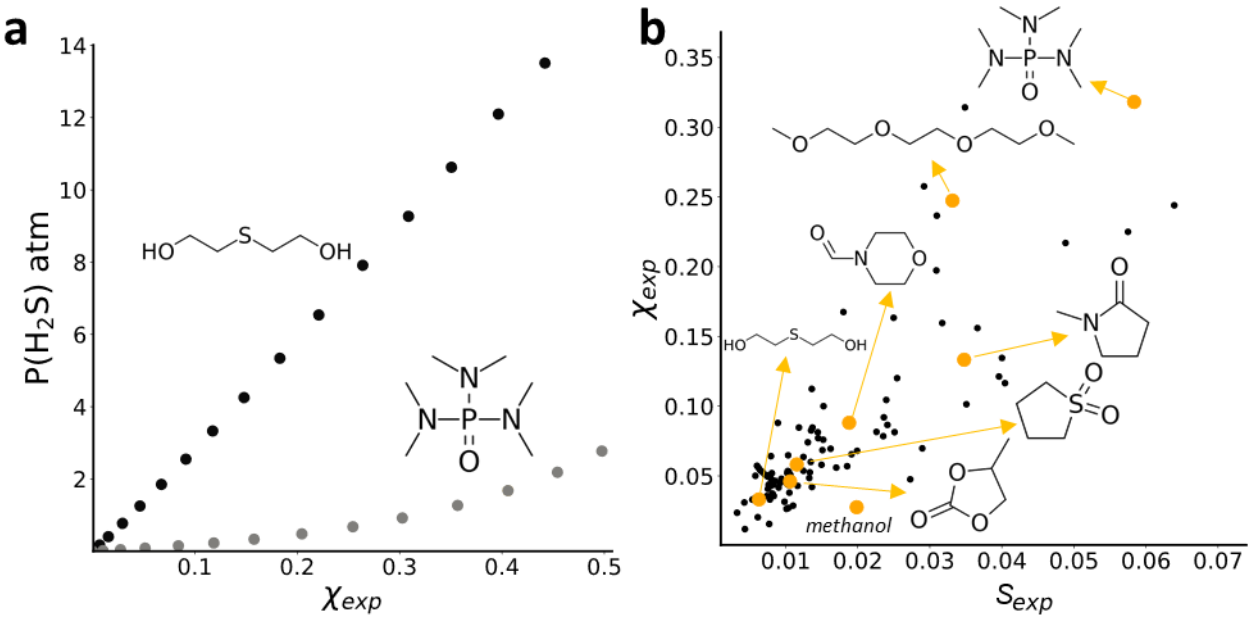

Figure 2. (a) Variation of mole fraction with partial pressure for H_2S in TDG (black) and HMPA (grey) at 298.15 K experimentally measured for this paper; **(b)** Plot of experimental molecular fraction values (χ_{exp}) vs Kuenen coefficients (S_{exp}) at 298.15 K and 1 atm. Solvents, that are used for gas absorption in industry, are shown in orange.

Variation of mole fraction with partial pressure for H_2S in TDG (green) and HMPA (dark yellow) at 298.15 K was experimentally measured using static apparatus method. There exists a linear relationship of mole fraction with partial pressure for TDG (Figure 1a). Estimated mole fraction solubility of TDG at 1 atm is 0.0332, which is close to the value extrapolated from Vahidi et al. (0.0296)[26]. On the contrary, for HMPA the relationship is highly non-linear. One can assume that non-linearity could be caused by strong ability of HMPA to form hydrogen bonds[38].

The final dataset consisted of 99 compounds from diverse classes including non-aromatic hydrocarbons (alkanes, cycloalkanes), aromatic hydrocarbons, alcohols, ethers and esters, halogenated compounds, nitrogen-, phosphorus-, sulfur- containing compounds. We analyzed the data through the prism of interrelationships between mole fractions solubility values and Kuenen coefficients of the solvents. Figure 2b shows that, although the mole fractions solubilities in some solvents can be large, these solvents are not necessarily the most interesting for the industrial application due to relatively low *S*. The largest mole fraction solubility value was found for HMPA (Figure 2b), which is slightly larger than that for esters of phosphoric acid, long-chain ethers, lactams and the only representative of amines in the dataset − 1-methylpiperidin-4-one. On the other hand, short-chain alcohols (methanol, ethanol, ethylene glycol, etc.), acetic acid and hydrocarbons (bicyclohexyl, diphenyl methane) were the worst solvents. The maximal value of Kuenen coefficient was found for 1-methylpiperidin-4-one followed by HMPA and 1-methylpiperidin-2-one (six-membered ring analog of industrially used 1-methylpyrrolidin-2-one (NMP) solvent). Notably, the only representative of sulfoxides in the dataset, dimethyl sulfoxide

(DMSO), was also among the best solvents according to the Kuenen coefficient. To our knowledge, the collected dataset is the largest among publicly available ones. Although, we could not directly compare its content with commercial Dortmund Data Bank,[39,40] we checked that the largest part of data for H_2S-solvent binary mixtures at 298.15 K available in Dortmund Data Bank is present in our dataset.

Quantitative structure - solubility relationships

The QSPR modeling workflow (see Materials and Methods section) was applied to the collected dataset. Distribution of the experimental $-\log\chi$ values, which were used as end-points for modeling, is shown in Figure 3a. 16 ISIDA individual models corresponding to different types of ISIDA descriptors with $Q^2_{OOB} \geq 0.7$ were selected for the consensus modeling. The propensity of the model to return accurate predictions with respect to novel solvent candidates was checked using a nested cross-validation procedure (Figure 2b). The validation statistics obtained from the outer cycle of nested cross-validation and out-of-bag values for the parent set are present in Table 1. The usage of ISIDA consensus approach indeed allows one to achieve reasonably good predictive performance in the nested cross-validation. The value of $MAE_{NCV\text{-}AD}$ (0.094) is close to the variance in experimental data. For example, the standard deviation of experimentally measured mole fractions for 1-methylpyrrolidin-2-one (NMP) can be estimated from Shokouhi et al. [18] and equals 0.7 log units. The absence of chance correlations was checked using y-randomization procedure (Table 1).

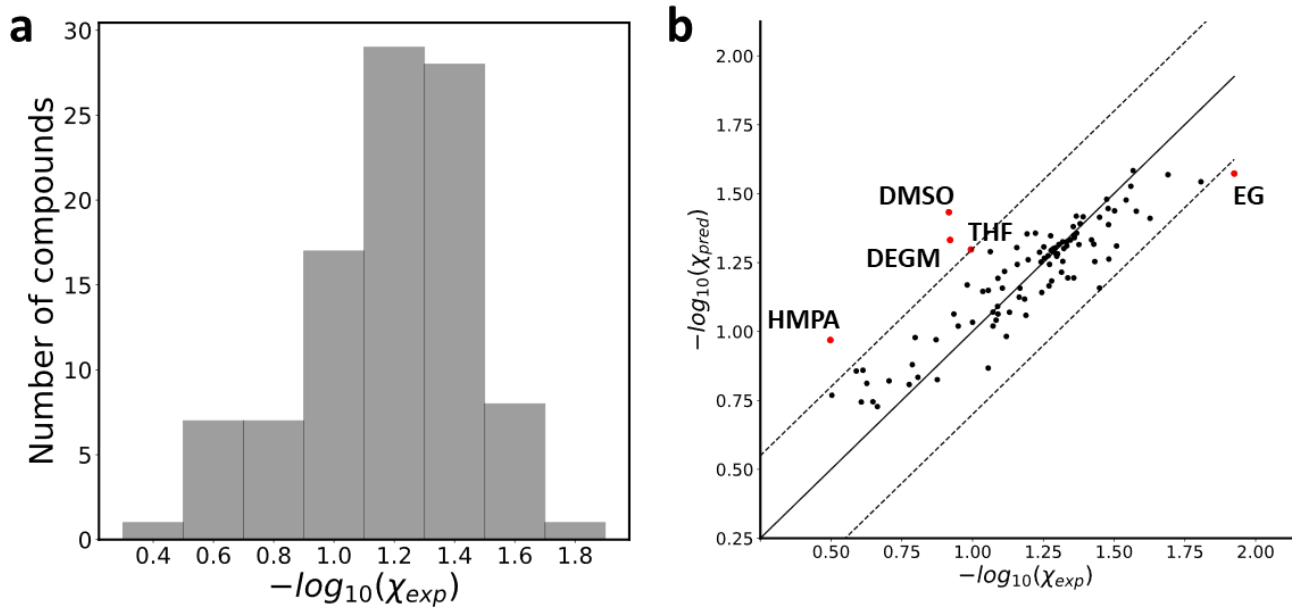

Figure 3. (a) Negative logarithm of mole fraction solubility values distribution; **(b)** Plot of predicted ($-\log_{10}\chi_{pred}$) vs experimental ($-\log_{10}\chi_{exp}$) values for ISIDA consensus model in nested-cross validation procedure. Compounds for which absolute errors were larger than $2\times RMSE_{NCV}$ are shown in red. Dash lines indicate $\pm 2\times RMSE_{NCV}$ threshold.

Table 1. Model validation statistics.

Model	Nested cross-validation			Final model			
	Q^2_{NCV}	$RMSE_{NCV}$	MAE_{NCV}	Q^2_{OOB} *	$RMSE_{OOB}$	MAE_{OOB}	Y_{rand} **
ISIDA consensus (16 models)	0.73±0.01	0.15±0.01	0.10±0.01	0.75±0.03	0.14±0.01	0.10±0.01	0.11
Spartan	0.72 ±0.01	0.15±0.01	0.10±0.01	0.74	0.14	0.10	0.03

* For ISIDA consensus model Q^2_{OOB}, $RMSE_{OOB}$, MAE_{OOB} statistics were calculated as an average of the ones of 11 individual ISIDA models participating in consensus.

** Maximum value among all individual ISIDA models is shown.

Unsurprisingly, the model's largest absolute errors were for HMPA and DMSO containing rare fragments. These compounds are the only representatives of phosphoramides and sulfoxides in the dataset. The absolute error values for diethylene glycol monomethyl ether (DEGM), ethylene glycol (EG) and oxolane (THF) were also found larger than $2\times RMSE_{NCV}$ threshold (Figure 3b). Solubility

of H_2S in THF was measured only by Short et al.[25] and according to the IUPAC's expert can be used on a tentative basis[10]. Solubility of its closest structural analog – 1,4-dioxane – is 20% lower. Further experimental investigation of similar cyclic ethers systems is required to clarify structure-solubility relationships in this class of compounds.

In contrast to the above examples, solubility of H_2S in alcohols, glycols and their ethers was systematically studied and, hence, the parent dataset contained many structural analogs of EG and DEGM. However, the structure-solubility relationships in these compounds' classes is rather complex: small changes in structure (e.g. replacement of hydrogen atom to methyl group in DEG/DEGM) may lead to large changes in solubility (see Figure 4, Table S1 in Supporting Information). Related pairs of compounds are called "activity cliffs" [41].

In order to check if some other descriptor types may improve predictive performance, the models were built on quantum chemical descriptors calculated with the Spartan software. The performance of Spartan descriptors-based model was comparable to the one involving ISIDA fragments obtained on the parent set, and lower to the latter in the nested cross-validation (Table 1, Figure S1 in Supporting Information). Averaging of the results of the ISIDA consensus model and the Spartan model predictions does not improve predictive performance (Q^2_{NCV} 0.76±0.02). It should be noted that computation of ISIDA fragment descriptors is very fast comparing to time-consuming quantum chemical calculations and, therefore, it enables one to apply ISIDA-based models for the virtual screening of large compound libraries[42].

Solubility of gases in liquids is governed by gas-solvent and solvent-solvent interactions.[43,44] Generally, strong gas-solvent and weak solvent-solvent interactions lead to greater solubility. This consideration may help to interpret variation of solubility in small congeneric series of solvents: heptane-octane-nonane (Series 1, Figure 4) and TDG-DEG-DEGM-diglyme (Series 2). One may see that solubility in alkanes marginally increases with the number of carbon atoms, while terminal oxygens of glycols significantly decrease solubility. The replacement of the oxygen atom (strong

hydrogen bond acceptor) in DEG with sulfur (TDG) does not significantly affect solubility. Note, that the compounds inside DEG-DEGM-diglyme and heptane-octane-nonane series are different from each other only by one CH_2 group. Nevertheless, the H_2S solubility in these compound series changes very differently: while the solubility increases significantly from DEG to diglyme, it practically does not change from heptane to nonane. This fact can be explained by the strength of solvent-solvent interactions, which can be estimated by the values of cohesive energy density. The cohesive energy density is the amount of energy needed to completely remove unit volume of molecules from their neighbors to infinite separation. While the cohesive energy density is only slightly increasing from heptane to nonane (231 – 243 MPa)[45], it is steeply decreasing from DEG (615 MPa)[46] to diglyme (296 MPa)[45]. This significant drop in the cohesive energy density values can be explained by very strong hydrogen bonding between DEG molecules. In general, solvent-solvent hydrogen bonding seem to play a key role in H_2S solubility in polar solvents (compare e.g. solubility in aniline/dimethyl aniline, NMA/DMA, etc., Table S1 in Supporting Information): intermolecular H_2S – solvent interactions are weak and cannot compensate unfavorable solvent-solvent bonds breaking. Nevertheless, one cannot completely neglect the role of H_2S – solvent interactions. For example, significantly higher solubility in diglyme can be explained by stronger gas-solvent interactions, that apart from London dispersion forces are driven by dipole-dipole interactions and hydrogen bonding.[47] In line with it, extremely high solubility of H_2S in HMPA can be due to exceptional hydrogen bonding capacity of HMPA.[38]

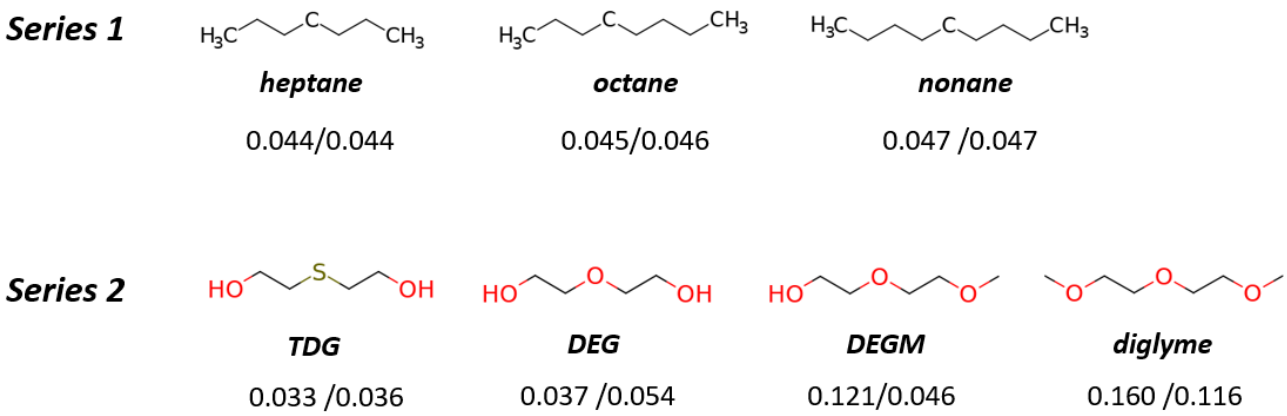

Figure 4. Variation of H_2S solubility in two congeneric series of solvents. The numbers correspond to experimental / predicted χ values.

Virtual screening

In order to find new solvents with high H_2S solubility, the developed models were used to screen the in-house library of industrially produced chemicals and their close structural analogs. Only the compounds containing the same atoms (C, H, N, O, S, halogens) as in the parent set were considered. In total, the screening library comprised more than 8,417 chemicals. Compared to the dataset used for the model building, the screening library contained heavier compounds, with a larger number of H-bond acceptors and smaller number of H-bond donors, see Figure S2 in Supporting Information. However, more than one third of the compounds from the screening set (35%) appeared to be inside of AD.

Although, predicted solubilities for the screening set compounds do not exceed their maximal values observed for the experimentally studied solvents (see Figure 5a), the screening results can still be useful for the design of industrially applicable solvents. The design of new solvents is, by its nature, a multi-objective optimization task, as far as apart from high H_2S solubility, a perspective solvent should possess high selectivity with respect to other gases, and appropriate physico-chemical parameters, including density, viscosity, boiling point, flammability, etc. Hence, even the

compounds with H_2S solubility, that is lower, than the one of the best solvent − HMPA, can still be useful as they can be more selective or possess more preferable physico-chemical properties.

Our calculations show that long-chain ethers (e.g. for pentaglyme $\chi_{pred} = 0.23$) display the highest H_2S solubility, while structural analogs of the industrially used NMP ($\chi_{exp} = 0.133$), for instance, 1-methylazocan-2-one ($\chi_{pred} = 0.21$), have the highest predicted Kuenen coefficients. For the experimental validation of screening results a compound with high predicted Kuenen coefficient − 1,3-Dimethylimidazolidin-2-one (DMI) ($\chi_{pred} = 0.146$; $S_{pred} = 0.33$) has been chosen. DMI is a well-known industrially produced solvent with high thermal and chemical stability,[48] which was suggested as a possible replacement for chemical solvent – aqueous *N*-methyl diethanolamine[49]. However, data on H_2S solubility in pure DMI is still lacking. . Thus, the solubility in DMI was measured using the same static pressure method as for HMPA and TDG (see Method section). The mole fraction solubility at 1 atm was estimated assuming its linear dependence on H_2S partial pressure (P) in the range P = 0.5 – 2 atm (Figure 5b). Experimental solubility (0.149) well matched the predicted one (0.146).

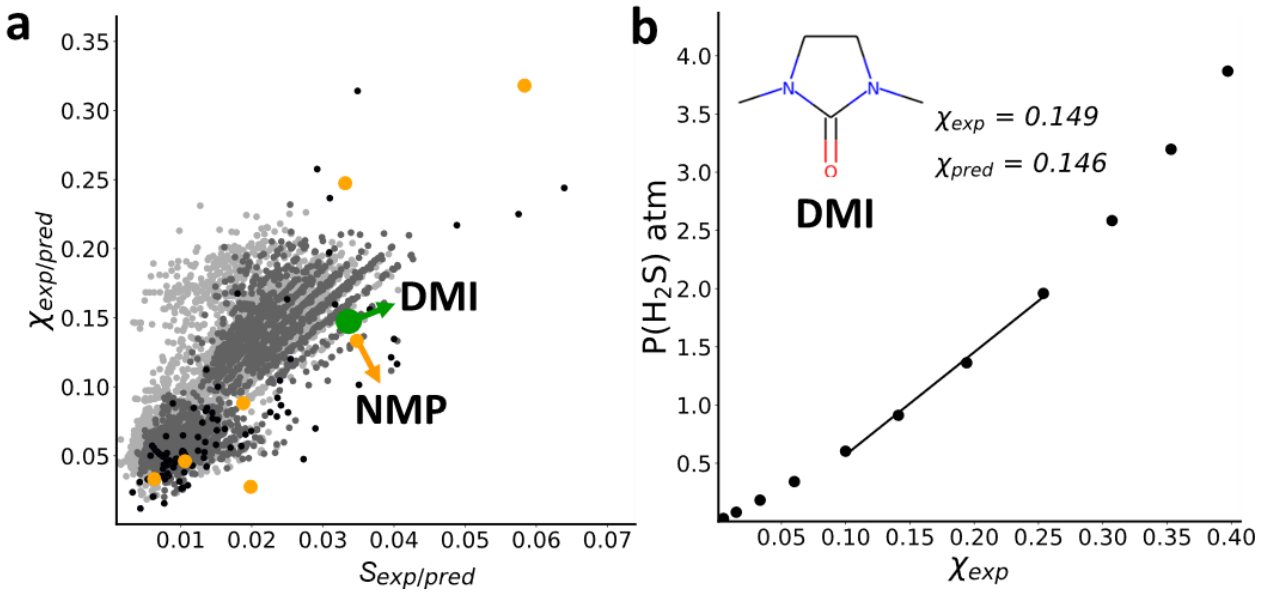

Figure 5. (a) Plot of molecular fraction values vs Kuenen coefficients: predicted values for compounds inside AD – grey, outside AD – light grey, for DMI – green. Experimental values –

black and orange (solvents used in the industry); **(b)** Variation of mole fraction with partial pressure for H_2S in DMI at 298.15 K experimentally measured in this paper; χ_{exp} – experimental mole fraction value at 1 atm and 298.15 K, χ_{pred} – predicted value. Fitted linear curve used for the interpolation – green solid line.

CONCLUSIONS

In this *proof-of-concept* study we showed that the design of new physical solvents for H_2S capturing can be rationalized via statistical models able to predict solubility as a function of molecular structure. A dataset containing 99 mole fraction H_2S solubility values at 298.15 K and 1 atm has been collected. Machine learning algorithm (random forest) and two types of molecular descriptors were then used for the modeling. The models displayed reasonably predictive performance: the mean absolute error in solubility was about 0.10 log units. Virtual screening of a library comprising >8400 industrially produced chemicals and their structural analogs resulted in several hits with reasonably high values of solubility and Kuenen coefficients. They can be considered as hot spots for further experimental exploration of H_2S physical solvents chemical space. The predicted solubility value for one of the retrieved hits − 1,3-Dimethylimidazolidin-2-one − was confirmed experimentally. Further accumulation of data will allow building more robust models which facilitate the progress in the rational design of solvents.

ASSOCIATED CONTENT

Supporting Information. The following file are available free of charge. Supporting Information – a pdf file containing Table S1, Figure S1, Figure S2.

AUTHOR INFORMATION

Corresponding Authors

*Professor Alexandre Varnek. Laboratory of Chemoinformatics, Faculty of Chemistry, University of Strasbourg, 4, Blaise Pascal Str., 67081, Strasbourg, France. email: varnek@unistra.

*Doctor Frédérick de Meyer. TOTAL SA, Total Exploration Production, Development and Support to Operations, Liquefied Natural Gas – Acid Gas Entity, Paris, 92078 France. email: frederick.de-meyer@total.com.

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding Sources

This work was supported by the Gas Solutions and Sustainable Development R&D program from Total S.A. Exploration & Production.

ACKNOWLEDGMENT

The authors are grateful to Dr. Fanny Bonachera for her help with the models implementation. The authors also thank Dr. Igor Baskin for fruitful discussion.

ABBREVIATIONS

DEG – 2-(2-hydroxyethoxy)ethanol (diethylene glycol)

DEGM – 2-(2-Methoxyethoxy)ethan-1-ol (diethylene glycol monomethyl ether)

diglyme – 1-methoxy-2-(2-methoxyethoxy)ethane

DMF – *N*,*N*-dimethylformamide

DMI – 1,3-Dimethylimidazolidin-2-one

DMSO – methylsulfinylmethane (dimethyl sulfoxide)

EG – ethane-1,2-diol (ethylene glycol)

HMPA – *N*-[bis(dimethylamino)phosphoryl]-*N*-methylmethanamine (hexametapol)

M2CA – methyl 2-cyanoacetate

NFM – morpholine-4-carbaldehyde (*N*-formylmorpholine)

NMP – 1-methylpyrrolidin-2-one

PC – 4-methyl-1,3-dioxolan-2-one (propylene carbonate)

TDG – 2-(2-hydroxyethylsulfanyl)ethanol (thiodiglycol)

pentaglyme – 1-methoxy-2-[2-[2-[2-(2-methoxyethoxy)ethoxy]ethoxy]ethoxy]ethane

THF – oxolane (tetrahydrofuran)

TPrP – tripropyl phosphate

χ – mole fraction solubility

S – Kuenen coefficient

SI – Kuenen coefficients selectivity index

REFERENCES

(1) Shah, M. S.; Tsapatsis, M.; Siepmann, J. I. Hydrogen Sulfide Capture: From Absorption in Polar Liquids to Oxide, Zeolite, and Metal–Organic Framework Adsorbents and Membranes. *Chem. Rev.* 2017, *117* (14), 9755–9803.

(2) Bryk, S. D.; Makitra, R. G.; Pal'chikova, E. Ya. Solubility of Hydrogen Sulfide in Organic Solvents. *Russ. J. Inorg. Chem.* 2006, *51* (3), 506–511.

(3) Sanchez-Lengeling, B.; Roch, L. M.; Perea, J. D.; Langner, S.; Brabec, C. J.; Aspuru-Guzik, A. A Bayesian Approach to Predict Solubility Parameters. *Adv. Theory Simul.* 2019, *2* (1), 1800069.

(4) Bradley, J.-C.; Abraham, M. H.; Acree, W. E.; Lang, A. S. Predicting Abraham Model Solvent Coefficients. *Chem. Cent. J.* 2015, *9* (1), 12.

(5) Muratov, E. N.; Bajorath, J.; Sheridan, R. P.; Tetko, I. V.; Filimonov, D.; Poroikov, V.; Oprea, T. I.; Baskin, I. I.; Varnek, A.; Roitberg, A.; Isayev, O.; Curtalolo, S.; Fourches, D.; Cohen, Y.; Aspuru-Guzik, A.; Winkler, D. A.; Agrafiotis, D.; Cherkasov, A.; Tropsha, A. QSAR without Borders. *Chem. Soc. Rev.* 2020, *49* (11), 3525–3564.

(6) H. Rostami; Riahi, S. Quantitative Structure–Property Relationship Study on Solubility of Hydrogen Sulfide in Organic Solvent; Kish, Iran, 2014.

(7) Kode srl, Dragon (software for molecular descriptor calculation) version 7.0.8, 2017, https://chm.kode-solutions.net

(8) Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solov'ev, V.; Hoonakker, F.; Tetko, I.; Marcou, G. ISIDA - Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors. *Curr. Comput. Aided-Drug Des.* 2008, *4* (3), 191–198.

(9) Spartan 18.0 www.wavefun.com.

(10) Fogg, P. G. T., Young, C. L. 2 - Hydrogen Sulfide in Non-Aqueous Solvents. In Hydrogen Sulfide, Deuterium Sulfide & Hydrogen Selenide; Fogg, P. G. T., Young, C. L., Eds.; Pergamon: Amsterdam, 1988; pp 166–326.

(11) Martial Atlani; Roben Loutaty; Claude Wakselman; Charles Yacono. Method of Purifying a Gas Mixture Containing Undesirable Gas Compounds. 4504287, March 12, 1985.

(12) Robert Frederick James Barber; Terence John Ritter; Christopher William Sweeney. Removing Sulfur Compounds from Gases. 2245889A, January 15, 1992.

(13) Xu, Y.; Schutte, R. P.; Hepler, L. G. Solubilities of Carbon Dioxide, Hydrogen Sulfide and Sulfur Dioxide in Physical Solvents. *Can. J. Chem. Eng.* 1992, *70* (3), 569–573.

(14) Gerrard, W. Solubility of Hydrogen Sulphide, Dimethyl Ether, Methyl Chloride and Sulphur Dioxide in Liquids. The Prediction of Solubility of All Gases. *J. Appl. Chem. Biotechnol.* 2007, *22* (5), 623–650.

(15) Shokouhi, M.; Farahani, H.; Hosseini-Jenab, M.; Jalili, A. H. Solubility of Hydrogen Sulfide in *N* -Methylacetamide and *N,N* -Dimethylacetamide: Experimental Measurement and Modeling. *J. Chem. Eng. Data* 2015, *60* (3), 499–508.

(16) Jou, F.-Y.; Deshmukh, R. D.; Otto, F. D.; Mather, A. E. Solubility of H2S, CO2 and CH4 in N-Formyl Morpholine. *J. Chem. Soc. Faraday Trans. 1 Phys. Chem. Condens. Phases* 1989, *85* (9), 2675.

(17) F.-Y. Jou; F.D. Otto; A.E. Mather. Solubility of H2S and CO2 in diethylene glycol at elevated pressures. *Fluid. Phase. Equilib.* 2000, 175 (1), 53–61.

(18) Shokouhi, M.; Salooki, M. K.; Ahari, J. S.; Esfandyari, M. Thermodynamical and Artificial Intelligence Approaches of H2S Solubility in N-Methylpyrrolidone. *Chem. Phys. Lett.* 2018, *707*, 22–30.

(19) Murrieta-Guevara, F.; Romero-Martinez, A.; Trejo, A. Solubilities of Carbon Dioxide and Hydrogen Sulfide in Propylene Carbonate, N-Methylpyrrolidone and Sulfolane. *Fluid. Phase. Equilib.* 1988, *44* (1), 105–115.

(20) Hayduk, W.; Pahlevanzadeh, H. The Solubility of Sulfur Dioxide and Hydrogen Sulfide in Associating Solvents. *Can. J. Chem. Eng.* 1987, *65* (2), 299–307.

(21) Tremper, K. K.; Prausnitz, J. M. Solubility of Inorganic Gases in High-Boiling Hydrocarbon Solvents. *J. Chem. Eng. Data* 1976, *21* (3), 295–299.

(22) Shokouhi, M.; Farahani, H.; Hosseini-Jenab, M. Experimental Solubility of Hydrogen Sulfide and Carbon Dioxide in Dimethylformamide and Dimethylsulfoxide. *Fluid Phase Equilibria* 2014, *367*, 29–37.

(23) Haertel, G. H. Low-Volatility Polar Organic Solvents for Sulfur Dioxide Hydrogen Sulfide and Carbonyl Sulfide. *J. Chem. Eng. Data* 1985, *30* (1), 57–61.

(24) Renon, Henri.; Lenoir, J. Y.; Renault, Philippe. Gas Chromatographic Determination of Henry's Constants of 12 Gases in 19 Solvents. *J. Chem. Eng. Data* 1971, *16* (3), 340–342.

(25) Short, I.; Sahgal, A.; Hayduk, W. Solubility of Ammonia and Hydrogen Sulfide in Several Polar Solvents. *J. Chem. Eng. Data* 1983, *28* (1), 63–66.

(26) Vahidi, M.; Shokouhi, M. Experimental Solubility of Carbon Dioxide and Hydrogen Sulfide in 2,2′-Thiodiglycol. *J. Chem. Thermodyn.* 2019, *133*, 202–207.

(27) Rivas, O. R.; Prausnitz, J. M. Sweetening of Sour Natural Gases by Mixed-Solvent Absorption: Solubilities of Ethane, Carbon Dioxide, and Hydrogen Sulfide in Mixtures of Physical and Chemical Solvents. *AIChE J.* 1979, *25* (6), 975–984.

(28) Koolivand Salooki, M.; Shokouhi, M.; Farahani, H.; Keshavarz, M.; Esfandyari, M.; Sadeghzadeh Ahari, J. Experimental and Modelling Investigation of H2S Solubility in N-Methylimidazole and Gamma-Butyrolactone. *J. Chem. Thermodyn.* 2019, *135*, 133–142.

(29) Shokouhi, M.; Rezaierad, A. R.; Zekordi, S.-M.; Abbasghorbani, M.; Vahidi, M. Solubility of Hydrogen Sulfide in Ethanediol, 1,2-Propanediol, 1-Propanol, and 2-Propanol: Experimental Measurement and Modeling. *J. Chem. Eng. Data* 2016, *61* (1), 512–524.

(30) D.S. Tsiklis; G.M. Svetlova. The Solubility of Gases in Cyclohexane. *Zh. fiz. Khim.* 1958, *32*, 1476–1480.

(31) Sciamanna, S.; Lynn, S. PhD thesis. Development Of A Process For Simultaneous Desulfurication, Drying, And Recovery Of Natural Gas Liquids From Natural Gas Streams, University of California, Berkeley, Lawrence Berkeley National Laboratory, 1986.

(32) Sweeney, Christopher W; Ritter, Terence J; McGinley, Eric B. A Strategy For Screening Physical Solvents. *Chem. Eng.* 1988, *95* (9), 119–125.

(33) *RDKit: Open-Source Cheminformatics; http://www.Rdkit.Org*.

(34) Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev, V. P. Substructural Fragments: An Universal Language to Encode Reactions, Molecular and Supramolecular Structures. *J. Comput. Aided Mol. Des.* 2005, *19* (9–10), 693–703.

(35) Baumann, D.; Baumann, K. Reliable Estimation of Prediction Errors for QSAR Models under Model Uncertainty Using Double Cross-Validation. *J. Cheminformatics* 2014, *6* (1), 47.

(36) Total S.A. HySWEET process https://www.ep.total.com/en/areas/liquefied-natural-gas/hysweetr-delivers-affordable-performance# (accessed Oct 05, 2020)

(37) Reichardt, C.; Welton, T. *Solvents and Solvent Effects in Organic Chemistry: Reichardt:Solv.Eff. 4ED O-BK*; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2010.

(38) Ivanov, E. V.; Kustov, A. V. Volumetric Properties of (Water+hexamethylphosphoric Triamide) from (288.15 to 308.15) K. *J. Chem. Thermodyn.* 2010, *42* (9), 1087–1093.

(39) Onken, U.; Rarey-Nies, J.; Gmehling, J. The Dortmund Data Bank: A Computerized System for Retrieval, Correlation, and Prediction of Thermodynamic Properties of Mixtures. *Int. J. Thermophys.* 1989, *10* (3), 739–747.

(40) Dortmund Data Bank web-site http://www.ddbst.com/ (accessed Oct 05, 2020).

(41) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inform.* 2010, *29* (6–7), 476–488.

(42) Lin, A.; Horvath, D.; Marcou, G.; Beck, B.; Varnek, A. Multi-task generative topographic mapping in virtual screening. *J. Comput. Aided Mol. Des.* 2019, *33*, 331-343.

(43) Battino, R.; Clever, H. L. The Solubility of Gases in Liquids. *Chem. Rev.* 1966, *66* (4), 395–463.

(44) Pierotti, R. A. A Scaled Particle Theory of Aqueous and Nonaqueous Solutions. *Chem. Rev.* 1976, *76* (6), 717–726.

(45) Abboud, J.-L. M.; Notari, R. Critical Compilation of Scales of Solvent Parameters. Part I. Pure, Non-Hydrogen Bond Donor Solvents. *Pure Appl. Chem.* 1999, *71* (4), 645–718.

(46) Zeng, W.; Du, Y.; Xue, Y.; Frisch, H. L. Solubility Parameters. In *Physical Properties of Polymers Handbook*; Mark, J. E., Ed.; Springer New York: New York, NY, 2007; pp 289–303.

(47) Biswal, H. S.; Bhattacharyya, S.; Bhattacherjee, A.; Wategaonkar, S. Nature and Strength of Sulfur-Centred Hydrogen Bonds: Laser Spectroscopic Investigations in the Gas Phase and Quantum-Chemical Calculations. *Int. Rev. Phys. Chem.* 2015, *34* (1), 99–160.

(48) Leahy, E. M. 1,3-Dimethyl-2-Imidazolidinone. In *Encyclopedia of Reagents for Organic Synthesis*; John Wiley & Sons, Ltd, Ed.; John Wiley & Sons, Ltd: Chichester, UK, 2001; p rd342.

(49) Ganizheva, L. L.; Ponomarenko, D. B.; Borisova T. Yu. Accelerated Method Of Choosingselective Absorbents For Extractionof Hydrogen Sulfide From Hydrocarbon Gases. *Chem. Pet. Eng.*, 2011, *46*, 11–12.