Partial Knowledge Predictive Models for Hydrocarbon Storage

Daniele Giampaoli[1,2], Guido Parodi[1,2], Francesca Cipollini[1], Shaji Vattakunnel[1], Alberto Maria Gambelli[3], and Luca Oneto[2]

[1] aizoOn consulting, Genoa, Italy
[2] University of Genoa, Genoa, Italy
[3] University of Perugia, Perugia, Italy

Abstract. Clathrate hydrates enable the storage of hydrocarbons in solid form with considerably lower energy requirements than conventional methods. In this study, we use Machine Learning (ML) techniques to develop and validate a model that describes clathrate formation, drawing on laboratory data from propane, ethane, methane, and carbon dioxide. The ML-based framework augments traditional analytical and theoretical approaches by predicting equilibrium points of gas compounds and guiding the process parameters to meet scientific and engineering needs. In addition, we address critical challenges in framing the problem for ML modeling, including selecting pertinent input and output variables and incorporating ML outputs into the physical-chemical equations, thereby providing a physics-informed estimate of the stored gas volume.

Keywords: Hydrocarbon Storage · Clathrate · Partial Knowledge · Machine Learning

1 Introduction

The increasing economic efforts and investments in sustainable energy production make it urgent to define new competitive and environmentally friendly energy storage methods. With increasing consumption of methane, the storage and transportation of energy gases is gaining growing attention from researchers and industrialists [23]. Gas hydrates, or clathrates, have been drawing attention in relation to waste and energy gas storage [21]. These ice-like solid crystalline compounds are formed when water molecules create hydrogen bonds with surrounding molecules, building a solid structure, while gaseous molecules fill the cavities in the lattice. The interaction between water and gases is exclusively physical, so gas recovery exclusively needs the dissociation of the water cages [8]. For this reason, these compounds have the promising property of making hydrocarbons more efficiently storable; for instance, methane can be liquefied at $-83°C$ and at least 5MPa [10], while the production of methane hydrates would be feasible at milder thermodynamic conditions, resulting in lower costs and yielding comparable results in terms of stored quantities. It could be possible to achieve temperatures above the freezing point of water ($2-3°C$) and at pressures lower

than 6MPa, where one cubic meter of hydrates can contain up to $172m^3$ of methane [10]. Solid-form storage also allows us to easily reach the standards in terms of safety, avoiding pressure loss and, in case of an accident, the release of gas would be drastically smaller and more manageable. Regarding the possibility of storing hydrocarbons, it should be noted that in addition to the identification of new and potentially carbon neutral energy sources, the capture and storage of carbon dioxide is currently the most effective strategy to preserve economic development without neglecting the environmental issue [15]. In fact, since 1997, the disposal of CO_2 in deep ocean sediments has been considered one of the most effective solutions [13], and since then numerous researchers have focused their attention on the production of CO_2 hydrates in porous media, with particular attention to the formation rate and the storage efficiency of CO_2 under different operating conditions. The amount of gas per volume unit stored via gas hydrates is affected by the type of crystalline structure that can be obtained. Moreover, the size of gas molecules, the presence of one or more different gas species, the diffusion of gas within the hydrate lattice, the specific growth of crystals - and so on - can significantly affect the amount of gas that can be stored [14].

This article deals with the study of clathrate formation when composed of ethane, propane, carbon dioxide and/or methane, adopting Machine Learning (ML) to enhance the prediction of its temperature-pressure equilibrium. In this context, modeling the stochastic process of clathrate hydrates formation using ML can be beneficial; ML algorithms are inherently well-suited for modeling complex systems, such as clathrate hydrate production in suitable refrigerated and sensorized storage reactors (SR-H2). Existing methods can be found in the literature about predicting the phase equilibrium of clathrate hydrates as a function of a set of parameters and starting conditions. Some more traditional approaches are based upon using chemical-physical constants and parameters, interpolations from tables of known values, thermodynamic equations and statistical-thermodynamic modeling under specific conditions, including working with small masses, low temperatures and confined environments [1]. A few other approaches employ ML algorithms, either relying upon more classical algorithms, such as linear regression and Support Vector Machines (SVM) [5, 7], Random Forests (RF) and boosting algorithms, or more advanced algorithms, like Artificial Neural Networks (ANNs) [22]. Many variables operate in the whole process and influence its outcomes, for example, but not limited to, pressure (typically high), temperature (typically low), some porous or permeable material acting as a medium, water purity, impurities and electrolyte activity. We may as well include the presence of inhibitory (or promoter) chemical compounds, which, like sodium chloride, can be naturally found in gas hydrate reservoirs. Studies in the literature tackle the analysis of both the formation and equilibrium phase of pure hydrates in the presence of a single gas compound, and mixtures of gases and salts in different concentrations, which may include different gases, such as methane, ethane, propane, carbon dioxide, nitrogen, xenon, krypton and argon [19]. In addition to this, laboratory test results have also shown that the geometry and mechanics of the reactor used for the experiments can strongly in-

fluence the results in terms of hydrate formation and stability conditions. It was experimentally observed that the dissociation, or equilibrium, phase, produces simple temperature-pressure curves from a qualitative point of view, showing a pseudo-exponential trend for several observed gases. Conversely, the formation phase experimentally produces more complex curves, characterized by greater variance and irregularity, making it harder to model with respect to the former.

Since the adoption of ML has gained widespread popularity across multiple disciplines in recent years, mainly thanks to the capacity of efficiently and quickly deriving laws and nonlinear relationships from complex data [20], we propose a combination of two data-driven models to accurately describe and govern the formation process, to which we add a traditional thermodynamic equation in a Partial Knowledge fashion. Partial Knowledge allows ML to leverage incomplete or fragmented domain knowledge for solving complex problems [12]. Instead of relying solely on data-driven methods, these approaches incorporate insights from theoretical principles, empirical rules, or expert intuition to constrain or enhance the learning process. This hybrid approach is particularly useful in those scientific domains where acquiring complete data is impractical, but partial understanding can provide valuable guidance [17]. In this study, different types of ML models were trained, such as ANN, RF, and eXtreme Gradient Boosting (XGBoost), in order to learn different physical quantities of the system studied and combine them in a governing physical-chemical equation. By embedding a thermodynamic equation into the learning process, the difficulty of translating a problem into closed-form mathematical representations typical of the physical-chemical sector can be overcome. The data used for model training are the result of an industrial-scale accumulation system based on gas hydrate technologies that stages preliminary laboratory experiments and their subsequent application and validation in practical use cases.

In the following, Section 2 contains a state-of-the-art review of relevant ML methods and a description of the clathrate experimental setting for data acquisition. In Section 3, we present the validation tests of our ML models and a full discussion and analysis of the results. Finally, Section 4 concludes the paper.

2 Methodologies

2.1 Data acquisition

Experimental data for model creation was acquired from the system shown in Figure 1 and composed of:

- a reactor capable of operating, if required, at high pressures, whose details are reported in Figure 2;
- an internal temperature control system;
- devices to monitor pressure and its variation during hydrate formation;
- devices to directly observe the process evolution.

The shape and surface-volume ratio of the reactor were designed in order to favor an accurate monitoring of the internal temperature and reduce the formation of

Fig. 1: The reactor is placed on a cart and directly connected to the gas cylinders. The whole apparatus is placed inside a cooling room, except for the acquisition device placed outside. The two pictures on the left show the upper flange, opened above and closed below, with sensors and connections clearly visible.

gradients. Downstream of the design of an industrial-scale reactor (easily scalable for higher storage volume usage) and having an adequate flow and pressure control system for gases, together with the required internal control systems for temperature and pressure variations, a 1000 liter basic module will be realized in the future and installed in a mobile handling system which guarantees ease of use for tests and end users. Currently used data were acquired for several tests exploring different combinations of carbon dioxide (CO_2), methane (CH_4), ethane (C_2H_6) and propane (C_3H_8). The number of tests performed with their relative gas combination is reported in Table 1. For each test, the following features were acquired:

- Time [s]: counting from the experiment start;
- Pressure [bar]: measured with a digital manometer model MAN-SD with an accuracy of $\pm 0.5\%$ of full scale, supplied by Kobold;
- Temperature [$^{\circ}C$]: measured with type K thermocouples at different depths inside the reactor (5, 10, and 15cm from the flange), with accuracy class 1, supplied by TC Direct.

For each series of repeated experiments, the test number (ordinal) was also registered since the memory effect of water can influence and shift the equilibrium position [18]. An example of the relationship between Temperature and Pressure during the clathrate's formation and dissociation phases can be found in Figure 3.

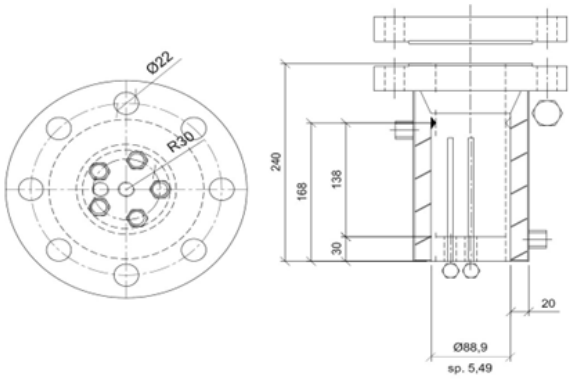

Fig. 2: Technical details of the reactor.

Table 1: Number of tests performed in each experimental setting and resulting number of samples (data points).

Test ID	Gas Combination	Number of Tests	Number of Samples
1	CH_4 (100%)	2	27354
2	CO_2 (100%)	7	40670
3	CO_2 (75%)+C_3H_8 (25%)	5	34662
4	C_2H_6 (100%)	14	72957
5	CH_4 (85%)+C_3H_8 (15%)	2	16567
6	C_2H_6 (75%)+ CH_4 (25%)	3	20082
7	C_2H_6 (50%)+ CH_4 (50%)	3	21170
8	C_2H_6 (25%)+ CH_4 (75%)	3	12582

2.2 Model Structure

ML models and algorithms define ways for artificial systems to extract patterns from raw data and gain knowledge on their own. For this reason, the performance of ML algorithms largely depends on the availability of data and how data is represented [24]. The main objective of all ML algorithms is learning a mapping function $y = f(x)$ to connect an input feature vector x, i.e., a sequence of numbers or categorical attributes, to an output prediction of some form y with some degree of uncertainty, governed by a probability distribution $p(y|x)$ [20]. How the mapping function is created is what distinguishes the different families of ML methods. One particular category of ML models is ANNs, computational models inspired by the structure of the human brain, consisting of interconnected layers of nodes (neurons) that process and transform data to recognize patterns or make predictions [11]. A different type of ML ensemble method is RF, which operates by constructing a collection of decision trees during training and outputting the majority vote for classification or the average prediction for regression [6]. Each tree is trained on a random subset of the data and features,

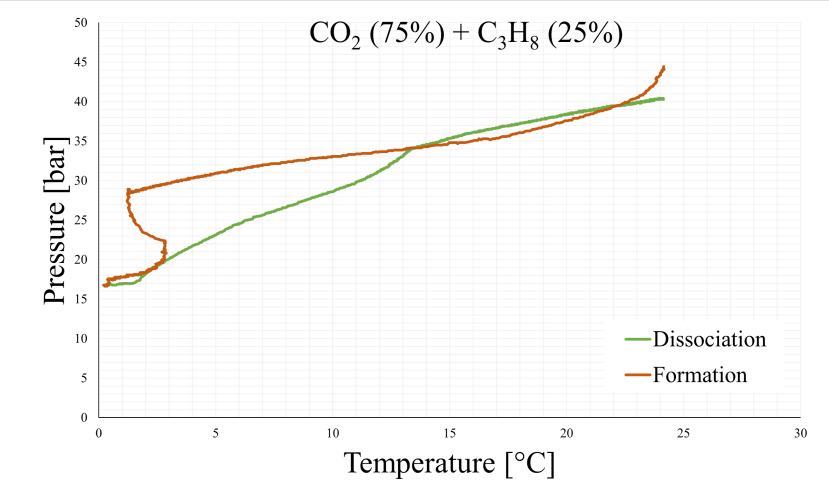

Fig. 3: The relationship between pressure and temperature in a Carbon Dioxide (75%) and Propane (25%) experiment.

introducing diversity and reducing overfitting compared to individual decision trees. Another tested ensemble method is XGBoost, which is based on gradient boosting and therefore, unlike RF, creates trees sequentially, where each new tree focuses on correcting the mistakes of the previous ones. XGBoost typically outperforms RF in structured data tasks due to its ability to minimize bias and variance more effectively through gradient-based optimization and regularization techniques.

Modeling the formation phase is relevant for storage purposes while the equilibrium/dissociation phase can determine how the hosted gas can be extracted and separated from additives included in the clathrate structure. These three categories of ML models were all considered as time-efficient and state-of-the-art models [22] in the Partial Knowledge architecture in order to predict the formation and phase equilibrium conditions of multiple hydrates in the presence of salts and organic substances while also considering technical features of the tools used in the experiments, wherever possible. Other state-of-the-art methods such as SVM and other ML Kernel methods were preliminarily evaluated for the proposed task but due to the available number of training samples, these methods resulted inefficient in terms of time and computational resources.

Two ML models with different input and output variables were created adopting the techniques proposed for this study. The global architecture showing how these two models are adopted together with the physical knowledge of hydrate formation is reported in Figure 4, based on the availability and selection of a subset of features described in Section 2.1. The first model (Model I) takes the initial fractional gas composition of the compound as input, together with the sample number (representing the timestamp), the initial temperature and pres-

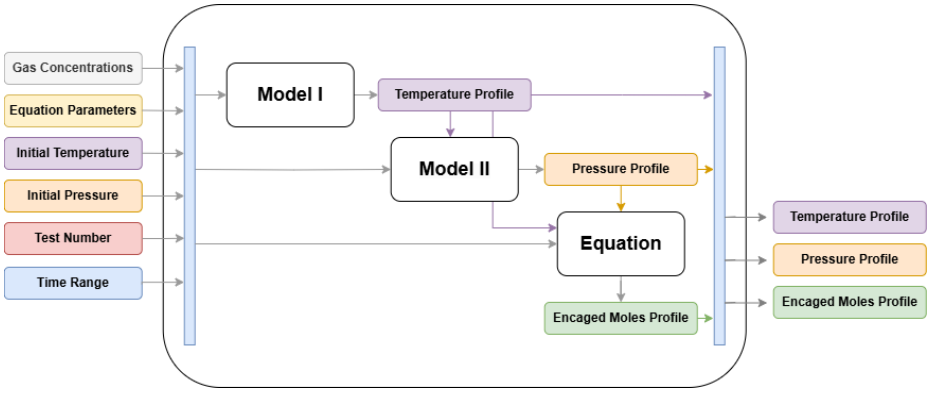

Fig. 4: Model structure.

sure values (known or set before the experiment starts) and the test number within a series of subsequent experiments (to account for the memory effect of water [18]). The data flows through a multi-layer perceptron neural network that outputs the experiment's predicted temperature at the given timestamp. By concatenating predicted samples in a given time range, we can reconstruct the entire time-temperature curve of the experiment as a result (Figure 5). The two-dimensional output curve is represented by pairs of time-temperature values, simulating the temporal evolution of the clathrate formation or dissociation process. For this model to learn how to satisfactorily associate time with temperature, several experiments with specific initial conditions are required. The

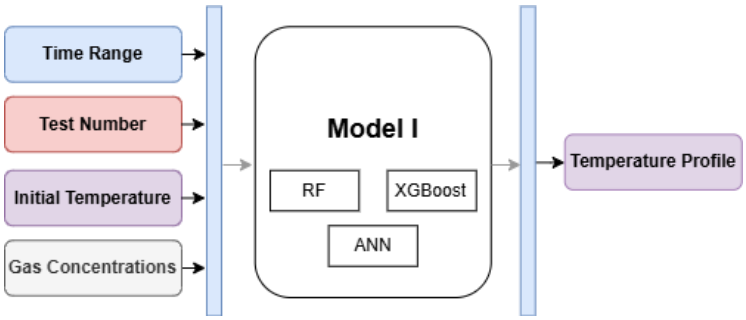

Fig. 5: Model I's schema.

second model (Model II) incorporates the experimental temperature profile into the set of input variables, one sample at a time, whose measurements are available at training time. Additionally, it includes the compound's fractional gas composition, the initial temperature and pressure values, the test number, and the sample number (acting again as a timestamp) (Figure 6). Similarly to Model

I, Model II's combined output is the entire pressure curve of the experiment throughout the selected time range, predicted one sample at a time. Adding the whole temperature profile to the input of the model is especially useful to predict the pressure profile. Firstly, the initial temperature can be set in advance, during laboratory setup, and then manually adjusted while the experiment is running. Secondly, it makes the learning process easier since, as observed experimentally and in the literature, temperature strongly correlates with pressure. The two

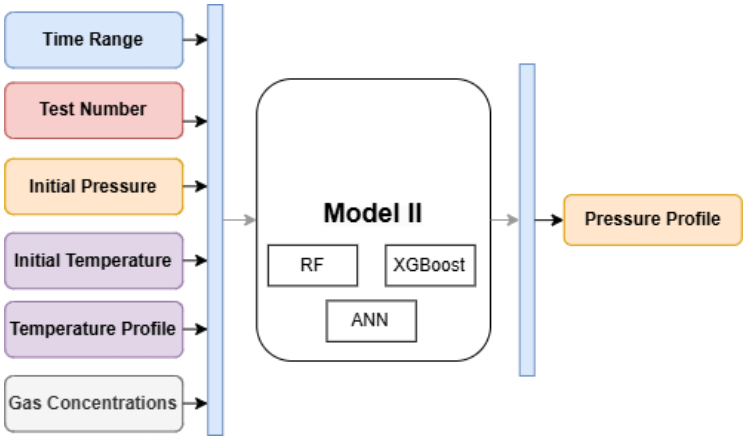

Fig. 6: Model II's schema.

models outlined above only partially address our ultimate task in which they cannot predict the amount of gas which is stored within the clathrate during the process. In a Partial Knowledge fashion, we include a third step in which the target stored gas moles are computed with Equation 1, as shown in [9], as follows:

$$mol_{HYD} = \frac{V_{PORE}(P_i Z_f - P_f Z_i)}{Z_f(RT - \frac{P_f}{\rho_{HYD}})} \tag{1}$$

In this equation, V_{PORE} represents the volume of sand pores available for gas hydrate formation (set to 253 cm^3 for all experiments); R and Z respectively describe the gas constant and the compressibility factor, calculated with the Peng-Robinson equation (8.314472 $J/(Kmol)$ and 0.87 (dimensionless) for all experiments, respectively); P (MPa, converted from bar) and T (K, converted from $°C$) describe pressure and temperature, while subscripts "i" and "f" were used to describe values measured at the beginning and the end of hydrate formation. Finally, ρ_{HYD} is the ideal molar density of hydrates and was defined according to [3, 25], and assumed to be 0.007339 mol/cm^3 in our experimental setting.

Estimating the amount of gas stored by solving an equation, as opposed to developing a proper third ML model, was a more viable option due to the limited

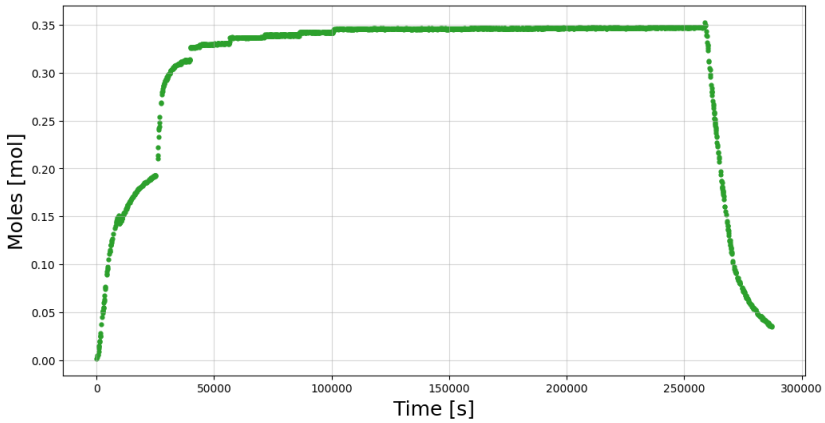

Fig. 7: An example of hydrate mole formation using Equation 1 for the same experiment shown in Figure 3 (Carbon Dioxide 75% and Propane 25%). It can be seen that the maximum value occurs at the end of the formation phase, before the dissociation phase starts.

availability of training experiments. Contrary to Model I and II scenarios, where each single time-temperature-pressure triplet point can be used for training, acquiring gas mole formation data would require pausing the experiment, opening the reactor and removing the clathrate for inspection, at small and possibly regular intervals. This is in contrast with what we can reasonably expect to have, i.e., one single measurement per test at the end of the formation phase - when the laboratory operator terminates the experiment - which would result in too few data points to train a ML model.

By plotting the mole formation profiles, as computed with Equation 1, it can be seen that the maximum yield of hydrate moles is obtained near the switching point between the formation and dissociation phase, i.e., near the end of the clathrate formation phase, as shown in Figure 7. Therefore, that neighborhood of points is the one we are most interested in correctly estimating.

3 Results and discussion

To validate the effectiveness of model training and assess which model type among ANN, RF, and XGBoost has the best performance, we employed two methods:

- the K-Fold Cross-Validation method is adopted in order to assess the learning capabilities of the different models [2];
- the Leave-One-Out (L-O-O) approach, to assess the models' extrapolation capabilities [4].

K-Fold Cross-Validation involves repeating the training and testing phases on different partitions of the original dataset. In this process, the dataset is split

into 10 equally-sized subsets ($K = 10$). Among these, 9 are combined to form the training set, while the remaining one is used as the test set. This procedure is repeated 10 times, each time removing a different subset from the training set, to be used as the test set. At the end of each iteration, the model's performance is computed on the test set. Upon completing the process, the overall performance of the model is calculated by averaging the results obtained over 10 iterations, while the standard deviation multiplied by two is used to compute the 95% confidence interval of the error.

Differently from K-Fold, the L-O-O approach not only evaluates the robustness of the model but also highlights its ability to generalize to unseen scenarios, which is the specific purpose of our application. Extrapolation ensures that the model can reliably predict outcomes for data points outside the range of the training set. By testing this capability, we can verify a model's potential to maintain performance in live settings, adapting effectively to new and unforeseen conditions while avoiding overfitting to the training data. This ensures that the ML solution is not only accurate but also versatile and operationally viable. For this purpose, since the global dataset includes N types of experiment data, each representing a different composition and/or concentration of gases used in experiments conducted in the reactor, N L-O-O scenarios can be simulated in order to compute model performances. In each scenario, a training set is created using data from all categories but one ($N - 1$), while data from the remaining category (the N-th) are moved to the test set. Model training and testing are performed for each scenario by leaving a different category out each time.

It is important to note that for both evaluation methods, the test set does not directly participate in the iterative estimation process of model parameters and is set aside during the training phase. Since test data are never used by the models' optimization algorithm, they effectively simulate a set of new and unseen data. Their purpose is to validate the models' ability to generalize, as much as possible, the salient characteristics of the entire process, thereby avoiding model overfitting on the training set.

We compared methods under scrutiny (ANN, RF, and XGBoost) with respect to state-of-the-art metrics, namely MAE, MSE, MAPE, sMAPE, and the coefficient of determination (R^2) [16]. All models were developed in Python, a programming language that hosts a wide range of specialized libraries for Artificial Intelligence and Machine Learning. Specifically, this includes NumPy and Pandas for data reading and processing, Matplotlib and Seaborn for data and result visualization and, finally, TensorFlow, XGBoost and Scikit-Learn for ML model development. The best hyperparameters for ANN and XGBoost were selected by adopting a grid-search procedure on a separate validation set extracted from the training set where the metrics were computed. In particular, for ANN models the ranges of values for each hyperparameter tested for both Model I and Model II were:

– number of layers in the neural network: 1 separate layer for each input variable (as seen in Figure 5 and Figure 6), 1 concatenation layer, 1 or 2 fully-connected layers and 1 output layer;

- number of neurons per layer in the neural network: $8, 16, 32$;
- number of training epochs, i.e., the number of times the neural network analyzes the entire training set: 400 with early stopping at around 30 (10 epochs of patience);
- batch size, i.e., the number of samples which the model analyzes before updating its internal parameters, during the training loop: $8, 16, 32, 64$;
- learning rate, i.e., used by the Adam algorithm during the optimization process: $1E$-$2, 1E$-3;
- λ term used to weigh the loss function's regularization term: $1E$-$8, 1E$-$6, 1E$-4.

For XGBoost, a grid-search procedure was performed on the following set of hyperparameters:

- number of trees: $50, 100, 500$;
- λ term, i.e., L2 regularization term on weights: $0, 0.1, 1$;
- α term, i.e., L1 regularization term on weights: $1, 5, 10$;

For RF models, 100 trees were used without making any hyperparameter tuning.

Table 2 reports the described error metrics computed on ANN, RF, and XGBoost methods for K-Fold Validation (with K=10) while Table 3 and Table 4 report results of the L-O-O Validation, respectively for Model I and II.

Table 2: Results of K-fold Cross Validation, K=10

Model	**Method**	**MAE**	**sMAPE**	**MSE**	R^2
I	ANN	$6.6E$-$1\pm1.5E$-1	$6.7E$-$2\pm1.4E$-2	$3.2E$-$2\pm1.8E$-2	$9.7E$-$1\pm1.8E$-2
	RF	$1.0E$-$2\pm2.2E$-4	$2.9E$-$3\pm2.2E$-4	$4.0E$-$6\pm1.9E$-6	$1.0E0\pm1.9E$-6
	XGBoost	$3.6E$-$1\pm5.8E$-1	$4.9E$-$2\pm7.2E$-2	$5.0E$-$3\pm1.0E$-2	$1.0E0\pm1.0E$-2
II	ANN	$3.9E$-$1\pm2.1E$-1	$8.0E$-$3\pm4.3E$-3	$5.4E$-$3\pm7.8E$-3	$9.9E$-$1\pm7.8E$-3
	RF	$3.3E$-$2\pm8.6E$-4	$5.8E$-$4\pm1.5E$-5	$3.5E$-$5\pm2.6E$-5	$1.0E0\pm2.6E$-5
	XGBoost	$3.5E$-$1\pm3.2E$-1	$7.0E$-$3\pm6.0E$-3	$3.0E$-$3\pm4.0E$-3	$1.0E0\pm4.0E$-3

In Figure 8, Figure 9, and Figure 10 sample profile predictions are reported to show the results obtained by Model I, Model II and Equation 1 for the different methods, respectively. Although our models can globally achieve high scores for all metrics in K-Fold Validation (Table 2), some gas combinations are predicted worse than others as shown in the L-O-O results in Table 3 and Table 4. In Table 3 and Table 4, some compounds achieve low R^2 scores – some below zero. This means that if we remove a certain gas combination from the training set, the model may have a hard time predicting it at test time. Among the tested algorithms, while RF proved to be the best model during K-Fold Validation, the different models have similar behaviours in extrapolation scenarios for both Model I and Model II. These performances may benefit from having a more accurate experimental design, where one carefully selects all the gas combinations that are worth exploring in the lab to describe the temperature-pressure relationship in the most generalized way. This would allow us to train a model on

Table 3: Results of L-O-O Validation for Model I

ML method	L-O-O Gas	MAE	sMAPE	MSE	R^2
ANN	CH_4(100%)	4.8E0	3.7E-1	2.8E-1	5.0E-2
	CO_2(100%)	4.4E0	4.3E-1	3.9E-1	5.0E-1
	CO_2(75%)+C_3H_8 (25%)	3.6E0	3.6E-1	1.9E-1	7.7E-1
	C_2H_6(100%)	5.1E0	2.0E-1	1.1E0	2.4E-1
	CH_4(85%)+C_3H_8(15%)	1.3E1	6.8E-1	2.6E0	-1.3E0
	C_2H_6(75%)+ CH_4(25%)	3.5E0	1.4E-1	2.2E-1	8.4E-1
	C_2H_6(50%)+ CH_4(50%)	7.4E0	3.4E-1	1.2E0	-6.9E-1
	C_2H_6(25%)+ CH_4(75%)	2.8E0	1.8E-1	2.3E-1	6.2E-1
RF	CH_4(100%)	4.4E0	4.1E-1	3.0E-1	-3.3E-4
	CO_2(100%)	4.7E0	3.5E-1	3.4E-1	5.6E-1
	CO_2(75%)+C_3H_8 (25%)	5.3E0	2.9E-1	4.2E-1	4.9E-1
	C_2H_6(100%)	2.9E0	1.6E-1	2.4E-1	8.3E-1
	CH_4(85%)+C_3H_8(15%)	8.5E0	5.4E-1	1.2E0	-9.5E-2
	C_2H_6(75%)+ CH_4(25%)	4.6E0	2.2E-1	3.2E-1	7.7E-1
	C_2H_6(50%)+ CH_4(50%)	4.0E0	3.1E-1	2.9E-1	5.8E-1
	C_2H_6(25%)+ CH_4(75%)	2.7E0	2.1E-1	1.9E-1	6.9E-1
XGboost	CH_4(100%)	4.1E0	2.4E-1	2.6E-1	1.2E-1
	CO_2(100%)	6.4E0	3.8E-1	6.6E-1	1.4E-1
	CO_2(75%)+C_3H_8 (25%)	7.6E0	5.1E-1	9.5E-1	-1.5E-1
	C_2H_6(100%)	2.4E0	1.1E-1	1.9E-1	8.7E-1
	CH_4(85%)+C_3H_8(15%)	8.4E0	4.3E-1	1.1E0	-7.5E-3
	C_2H_6(75%)+ CH_4(25%)	4.4E0	1.4E-1	2.8E-1	8.0E-1
	C_2H_6(50%)+ CH_4(50%)	7.5E0	3.3E-1	1.2E-1	-7.5E-1
	C_2H_6(25%)+ CH_4(75%)	2.4E0	1.6E-1	1.2E-1	8.0E-1

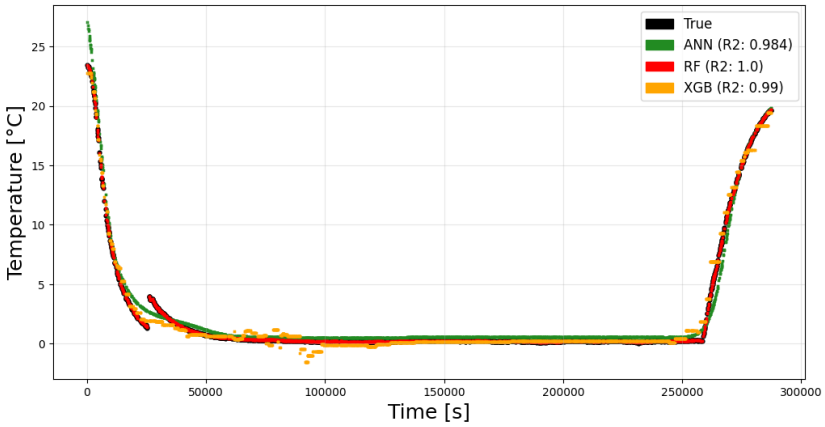

Fig. 8: An example of temperature profile predictions using Model I showing a good approximation of the true profile for the experiment reported in Figure 3 with Carbon Dioxide (75%) and Propane (25%).

Table 4: Results of L-O-O Validation for Model II

ML method	L-O-O Gas	MAE	sMAPE	MSE	R^2
ANN	$CH_4(100\%)$	$6.7E0$	$9.0E$-2	$5.2E$-1	-$2.1E$-1
	$CO_2(100\%)$	$5.2E0$	$1.1E$-1	$3.6E$-1	$3.0E$-1
	$CO_2(75\%)$+$C_3H_8(25\%)$	$4.8E0$	$9.0E$-2	$3.3E$-1	$6.1E$-1
	$C_2H_6(100\%)$	$1.1E1$	$2.2E$-1	$1.4E0$	-$1.0E0$
	$CH_4(85\%)$+$C_3H_8(15\%)$	$3.3E0$	$4.0E$-2	$1.4E$-1	$8.0E$-1
	$C_2H_6(75\%)$+ $CH_4(25\%)$	$3.6E0$	$5.0E$-2	$1.4E$-1	$7.0E$-1
	$C_2H_6(50\%)$+ $CH_4(50\%)$	$4.1E0$	$6.0E$-2	$2.0E$-1	$2.3E$-1
	$C_2H_6(25\%)$+ $CH_4(75\%)$	$3.0E0$	$4.0E$-2	$1.0E$-1	$7.9E$-1
RF	$CH_4(100\%)$	$5.3E0$	$4.4E$-1	$2.8E$-1	$3.5E$-1
	$CO_2(100\%)$	$2.7E0$	$3.6E$-1	$8.0E$-2	$8.4E$-1
	$CO_2(75\%)$+C_3H_8 (25%)	$3.2E0$	$2.2E$-1	$1.2E$-1	$8.5E$-1
	$C_2H_6(100\%)$	$1.2E1$	$6.1E$-1	$1.6E0$	-$1.3E0$
	$CH_4(85\%)$+$C_3H_8(15\%)$	$4.4E0$	$3.7E$-1	$1.6E$-1	$7.6E$-1
	$C_2H_6(75\%)$+ $CH_4(25\%)$	$3.3E0$	$3.4E$-1	$1.0E$-1	$7.8E$-1
	$C_2H_6(50\%)$+ $CH_4(50\%)$	$2.9E0$	$4.0E$-1	$8.2E$-2	$6.8E$-1
	$C_2H_6(25\%)$+ $CH_4(75\%)$	$2.9E0$	$4.3E$-1	$8.5E$-2	$8.2E$-1
XGboost	$CH_4(100\%)$	$5.6E0$	$7.2E$-2	$3.2E$-1	$2.7E$-1
	$CO_2(100\%)$	$3.7E0$	$6.5E$-2	$1.9E$-1	$6.3E$-1
	$CO_2(75\%)$+C_3H_8 (25%)	$3.4E0$	$6.9E$-2	$1.2E$-1	$8.5E$-1
	$C_2H_6(100\%)$	$9.9E0$	$2.0E$-1	$1.2E$-0	-$6.8E$-1
	$CH_4(85\%)$+$C_3H_8(15\%)$	$3.6E0$	$4.4E$-2	$1.1E$-1	$8.4E$-3
	$C_2H_6(75\%)$+ $CH_4(25\%)$	$2.3E0$	$2.9E$-2	$5.6E$-2	$8.8E$-1
	$C_2H_6(50\%)$+ $CH_4(50\%)$	$6.2E0$	$1.1E$-1	$4.7E$-1	-$8.3E$-1
	$C_2H_6(25\%)$+ $CH_4(75\%)$	$2.6E0$	$3.5E$-2	$7.5E$-2	$8.4E$-1

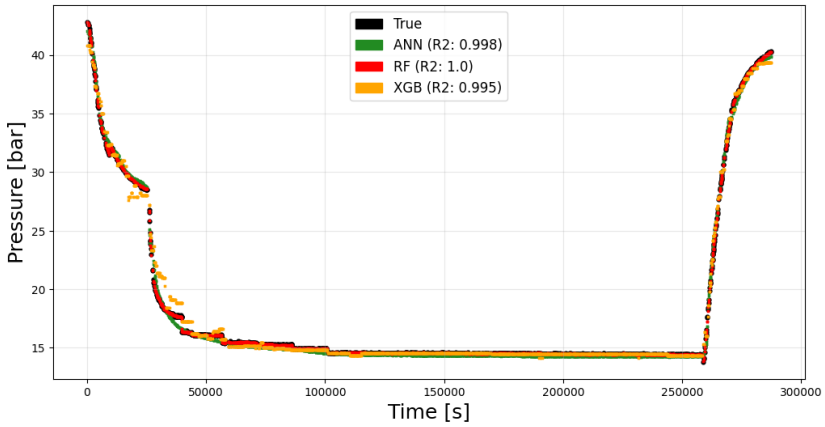

Fig. 9: An example of pressure profile predictions using Model II showing a good approximation of the true profile for the experiment reported in Figure 3 with Carbon Dioxide (75%) and Propane (25%).

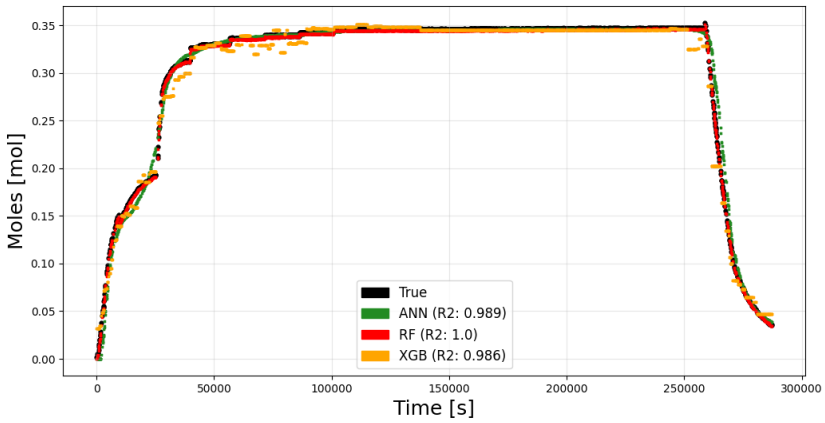

Fig. 10: An example of encapsulated mole profile predictions using Equation 1 showing a good approximation of the true profile for the experiment reported in Figure 3 with Carbon Dioxide (75%) and Propane (25%).

a wider-purpose training dataset and use it to accurately predict any combination of gases, either seen or unseen by the model. However, for the purpose of our study, we have seen that both models can predict well the temperature and pressure values around the switching point between formation and dissociation (when the curve abruptly starts growing, as seen in Figure 8 and Figure 9), which typically yields the maximum hydrate moles across the whole process. This is extremely important from the viewpoint of the researcher conducting the simulation as it is the core outcome of the whole experiment. More precisely, this result lets the researcher know whether a specific experimental setting is worth exploring in a hydrate storage production scenario and is a quantitative comparison metrics among different experimental settings.

4 Conclusions

Using clathrate hydrates to store hydrocarbon gases in solid form offers a new promising alternative to current storage technologies by significantly reducing the energy required for storage. This study addressed the potential of data-driven machine learning approaches to describe and predict the clathrate formation process, providing valuable support to laboratory experiments and highlighting the need for an accurate experimental design phase to identify the best gas combination for model training. Moreover, we addressed the key challenges in translating the problem into a Partial Knowledge framework, carefully selecting input and output variables and selecting the model to best meet the problem's demands. In this way, data-driven modeling can be a powerful complement to traditional analytical-theoretical descriptions, enabling accurate predictions and control of process characteristics in response to scientific and engineering requirements.

The best performing model, namely Random Forest, will be adopted in future studies to investigate the use of clathrate hydrates of hydrogen as an efficient storage system. In fact, hydrogen storage via clathrate hydrates offers a promising, energy-efficient alternative to traditional high-pressure and liquid hydrogen storage methods, using water as the storage medium and requiring relatively mild thermodynamic conditions. While requiring lower energy density per volume and mass unit than other methods, it achieves competitive results at comparable pressures, making it a viable option for safe, low-impact, and low-pressure storage. However, challenges include scaling the technology to ensure consistent hydrogen supply, achieving high hydrogen purity, and maintaining environmental and economic sustainability. We aim to address these issues by validating the scalability of clathrate-based hydrogen storage in real-world industrial applications, particularly in sectors like mobility and continuous industrial consumption, where hydrogen could replace natural gas and reduce CO_2 emissions.

Acknowledgments. The authors gratefully acknowledge the financial support derived by the PNRR project entitled: "High Efficiency Hydrogen Storage (HEHS)", ID: RSH2B_000052. The authors also acknowledge the technical and material contribution of the Civil and Environmental Engineering Department at the University of Perugia and Nippon Gases Industrial SrL.

References

1. Acharya, P.V., Bahadur, V.: Thermodynamic features-driven machine learning-based predictions of clathrate hydrate equilibria in the presence of electrolytes. Fluid Phase Equilibria **530**, 112894 (2021)
2. Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L., Ridella, S., et al.: The'k'in k-fold cross validation. In: ESANN (2012)
3. Aregbe, A.G.: Gas hydrate—properties, formation and benefits. Open Journal of Yangtze Oil and Gas **2**(1), 27–44 (2017)
4. Bachmann, G., Hofmann, T., Lucchi, A.: Generalization through the lens of leave-one-out error. arXiv preprint arXiv:2203.03443 (2022)
5. Baghban, A., Ahmadi, M.A., Pouladi, B., Amanna, B.: Phase equilibrium modeling of semi-clathrate hydrates of seven commonly gases in the presence of tbab ionic liquid promoter based on a low parameter connectionist technique. The Journal of supercritical fluids **101**, 184–192 (2015)
6. Breiman, L.: Random forests. Machine learning **45**, 5–32 (2001)
7. Eslamimanesh, A., Gharagheizi, F., Illbeigi, M., Mohammadi, A.H., Fazlali, A., Richon, D.: Phase equilibrium modeling of clathrate hydrates of methane, carbon dioxide, nitrogen, and hydrogen+ water soluble organic promoters using support vector machine algorithm. Fluid Phase Equilibria **316**, 34–45 (2012)
8. Gambelli, A.M., Presciutti, A., Rossi, F.: Review on the characteristics and advantages related to the use of flue-gas as co2/n2 mixture for gas hydrate production. Fluid Phase Equilibria **541**, 113077 (2021)
9. Gambelli, A.M., Rossi, F.: Thermodynamic and kinetic characterization of methane hydrate 'nucleation, growth and dissociation processes, according to the labile cluster theory. Chemical Engineering Journal **425**, 130706 (2021)

10. Gambelli, A.M., Rossi, F., Cotana, F.: Gas hydrates as high-efficiency storage system: Perspectives and potentialities. Energies **15**(22), 8728 (2022)
11. Goodfellow, I.: Deep learning (2016)
12. Guha, R.V.: Partial knowledge in embeddings. arXiv preprint arXiv:1710.10538 (2017)
13. Haugan, P.M.: Direct ocean storage of co2: Options and obstacles. In: ISOPE International Ocean and Polar Engineering Conference. pp. ISOPE–I. ISOPE (1997)
14. Makogon, Y.F.: Natural gas hydrates–a promising source of energy. Journal of natural gas science and engineering **2**(1), 49–59 (2010)
15. Mesbah, M., Soroush, E., Roham, M., Shahsavari, S.: Phase equilibrium modeling of semi-clathrate hydrates of the co2+ h2/ch4/n2+ tbab aqueous solution system. Petroleum Science and Technology **35**(15), 1588–1594 (2017)
16. Naser, M.Z., Alavi, A.H.: Error Metrics and Performance Fitness Indicators for Artificial Intelligence and Machine Learning in Engineering and Sciences. Architecture, Structures and Construction **3** (2021)
17. Oneto, L., Chicco, D.: Eight quick tips for biologically and medically informed machine learning. PLOS Computational Biology **21**(1), e1012711 (2025)
18. Ripmeester, J.A., Alavi, S.: Some current challenges in clathrate hydrate science: Nucleation, decomposition and the memory effect. Current Opinion in Solid State and Materials Science **20**(6), 344–351 (2016)
19. Rossi, F., Gambelli, A.M.: Thermodynamic phase equilibrium of single-guest hydrate and formation data of hydrate in presence of chemical additives: A review. Fluid Phase Equilibria **536**, 112958 (2021)
20. Shalev-Shwartz, S., Ben-David, S.: Understanding machine learning: From theory to algorithms. Cambridge university press (2014)
21. Sloan Jr, E.D., Koh, C.A.: Clathrate hydrates of natural gases. CRC press (2007)
22. Song, Y., Zhou, H., Wang, P., Yang, M.: Prediction of clathrate hydrate phase equilibria using gradient boosted regression trees and deep neural networks. The Journal of Chemical Thermodynamics **135**, 86–96 (2019)
23. Strobel, T.A., Koh, C.A., Sloan, E.D.: Thermodynamic predictions of various tetrahydrofuran and hydrogen clathrate hydrates. Fluid Phase Equilibria **280**(1-2), 61–67 (2009)
24. Sun, C., Shrivastava, A., Singh, S., Gupta, A.: Revisiting unreasonable effectiveness of data in deep learning era. In: IEEE international conference on computer vision (2017)
25. Takeya, S., Muromachi, S., Yoneyama, A.: X-ray imaging of clathrate hydrates as gas storage materials: absorption contrast of low-density and low-absorption materials using energy-dependent x-ray computed tomography. Energy & Fuels **36**(18), 10659–10666 (2022)