Supporting information

Bayesian symbolic learning to build analytical correlations from rigorous process simulations: Application to CO₂ capture technologies

Valentina Negriᵃ, Daniel Vázquezᵃ, Marta Sales-Pardoᵇ, Roger Guimeràᵇ,c,*, Gonzalo Guillén-Gosálbezᵃ,*

ᵃInstitute for Chemical and Bioengineering, Department of Chemistry and Applied Biosciences, ETH Zürich, Vladimir-Prelog-Weg 1, 8093 Zürich, Switzerland.

ᵇDepartment of Chemical Engineering, Universitat Rovira i Virgili, Tarragona 43007, Catalonia, Spain.

ᶜICREA, Barcelona 08010, Catalonia, Spain.

*Corresponding authors: Roger Guimerà: roger.guimera@urv.cat; Gonzalo Guillén-Gosálbez: gonzalo.guillen.gosalbez@chem.ethz.ch.

This document includes additional material to the content in the main article. Here we list the parameters of the analytical expressions and the MCMC steps, process models assumptions and limitations, and other secondary results. The latter comprise the residual plots, the models training, and an additional analysis using artificial neural networks.

Number of pages: 16
Number of figures: 7
Number of tables: 5
1. **Model parameters and MCMC steps**

The number of Markov chain Monte Carlo (MCMC) steps chosen to obtain the mathematical expression of each output variable, and the respective parameters are reported in Table S1 for the two case studies analyzed. The goodness of fit for the natural gas process variables results in an R-squared ($R^2$) above 99% for three out of four output variables, at low (1000-10000) MCMC steps. On the contrary, in the case of the flue gas, we require more steps (up to 50000) to achieve similar performance.

**Table S1:** MCMC steps and parameters of the dependent variables mathematical expressions for each case study analyzed.

| Case study | Variable       | MCMC steps | Parameters     |
|------------|----------------|-------------|----------------|
| Natural gas| Min CU         | 1000        | a0 4648399064  |
|            |                |             | a1 0.515205668|
|            |                |             | a2 0.08127544 |
|            |                |             | a3 -5.251595861|
|            |                |             | a4 22554.25141|
|            |                |             | a5 1.203644142|
| Min HU     |                | 5000        | a0 2.92526E+18 |
|            |                |             | a1 3.73775E-08|
|            |                |             | a2 351315.5    |
|            |                |             | a3 17571.65    |
|            |                |             | a4 9.752716    |
|            |                |             | a5 1752.488    |
|            |                |             | a6 -8.4617E+12 |
| Net power  |                | 1000        | a0 6.692334    |
|            |                |             | a1 -13459      |
|            |                |             | a2 -0.85238    |
|            |                |             | a3 -0.37024    |
|            |                |             | a4 0.507276    |
|            |                |             | a5 -7115.09    |
|            |                |             | a6 20657.74    |
| Amount of MEA|              | 10000       | a0 14112.78686 |
|            |                |             | a1 -2935.415061|
|            |                |             | a2 0.998725114 |
|            |                |             | a3 0.063845319 |
|            |                |             | a4 8.823491621 |
|            |                |             | a5 3.99600556  |
|            |                |             | a6 1467.217871 |
|            |                |             | a7 26.01699426 |
|                        | Min CU  | 50000 | a0     |                   | a1     |                   | a2     |                   | a3     |                   | a4     |                   | a5     |                   | a6     |                   | a7     |                   |
|------------------------|---------|-------|--------|-------------------|--------|-------------------|--------|-------------------|--------|-------------------|--------|-------------------|--------|-------------------|--------|-------------------|--------|-------------------|
| **Flue gas**           |         |       |        |                   |        |                   |        |                   |        |                   |        |                   |        |                   |        |                   |        |                   |
| **Min HU**             |         |       |        |                   |        |                   |        |                   |        |                   |        |                   |        |                   |        |                   |        |                   |
| **Net power**          |         |       |        |                   |        |                   |        |                   |        |                   |        |                   |        |                   |        |                   |        |                   |
| **Amount of MEA**      |         |       |        |                   |        |                   |        |                   |        |                   |        |                   |        |                   |        |                   |        |                   |

2. **Models assumptions and limitations**

2.1 Process design

The case studies in the main article are designed based on the following assumptions:

- The natural gas feed is considered a binary mixture, similarly to other literature studies about emerging technology such as membranes $^{1-3}$.
- With the underlying idea of including a wider range of product purity in the natural gas case study, we also consider scenarios where further processing of the sweet gas product
is required before pipeline distribution (about 50% of the results meet the specification for immediate distribution).

- The cooling and heating utilities requirements are calculated based on the composite curves for each point without determining the optimal heat exchanger network.
- Further optimization of the process design could be achieved in both case studies to decrease the amount of solvent, energy consumption, and increase the product purity. Both processes are simulated based on literature studies as a starting point. Then, sensitivity analyses are carried out on the characteristic parameter of each unit operation. For example, coolers and heaters outlet temperatures, compressor pressure, feed stage and number of trays in the absorber and stripper columns were varied in ranges of interest. The effect of the design on the product stream (sweet gas or clean gas for the first and second case study, respectively) and the CO\(_2\) stream was monitored. We note that no optimization tool was used to define the final design of either process.
- Our designs differ from previous literature studies as we include a purge in the process. We account for pressure drop in the unit operations to represent a real system more closely. The pressure drop is fixed to 0.2 bar in the heat exchangers and 0.04 bar/stage in the columns. No pressure drop is considered in the flashes.
- The energy requirement for CO\(_2\) injection and storage in a selected site is not considered in the models, and only the energy demand to obtain supercritical CO\(_2\) is accounted for.
- Specifically for the flue gas treatment process, the design of the base case has not been improved further to allow for greater flexibility to explore a range of six input variables, despite the low CO\(_2\) removal compared to the majority of literature studies. However, some literature about low CO\(_2\) capture rate is also available\(^4,5\). We here aim at exploring a wider range of input conditions while proving a lower bound for the proxies of the operating costs. Nonetheless, we report that for the same design in Figure 3 and an overall capture rate of 90%, the reboiler duty is 3.3 MJ/kg of CO\(_2\) captured. We obtain 6804 kW and 5280 kW from the Aspen energy analyzer as the target of cooling and heating utilities.
- The training and validation dataset depend on the ranges of the input variables chosen and are limited by the flowsheet convergence.
- To improve the goodness of fit, a higher number of MCMC steps can be explored for each output variable. At infinite MCMC steps, the error of the prediction should be zero.

Additionally, we report in Table S2 the upper and lower bounds of the two product streams in both case studies analyzed.

**Table S2:** Maximum and minimum flow rate values of the clean gas and CO\(_2\) stream obtained during the sampling of the two case studies.

| Case study                  | Clean gas [kg/s] | CO\(_2\) stream [kg/s] | CO\(_2\) stream [%mol. fraction CO\(_2\)] |
|-----------------------------|------------------|------------------------|------------------------------------------|
| Natural gas sweetening      | Minimum 16.08    | 11.47                  | 98.79                                    |
|                             | Maximum 22.12    | 17.32                  | 99.17                                    |
| Flue gas treatment          | Minimum 7.17     | 7.93E-3                | 99.57                                    |
|                             | Maximum 8.21     | 1.84E-2                | 99.59                                    |
The sampling and training steps are carried out on an Intel Core i9-9900 machine at 3.10 GHz and 32 GB RAM running Windows 10. The training of each output variable required 3 to 24 hours, depending on the number of MCMC steps. The validation is done using an Intel Core i7-10510U machine at 1.80 GHz and 16 GB RAM running Windows 10.

2.2 Variables selection

Table 1 reports the selected independent and dependent variables for the two case studies analyzed. Here, we are interested in two processes that provide a CO₂-free product, natural gas or flue gas, for a wide range of inlet conditions of the untreated gas. Therefore, the independent variables are the properties of the input flow, i.e., temperature, pressure and composition. On the contrary, the dependent variables are selected based on the goal of the analysis. In our work, we want to analyze the process performance with indicators related to the economics and environmental impacts, such as heating and cooling energy requirement. We note that different performance indicators can be chosen, e.g., CO₂ absorption efficiency or recovery, CO₂ storage conditions or alternative process design parameters. These parameters can be easily calculated when working with process simulators. However, it might be more challenging when the data are obtained from an experimental setup. In the case of experimental systems, it might not always be possible to measure the variables of interest, which can limit the choice of dependent variables substantially. The configuration of the setup might be very similar to the flowsheets presented in Figure 2 and 3, but on a smaller scale (see, for example, the pilot scale setup presented by Morgan and co-workers6).

2.3 Interpretability of the models

The most critical aspect of data-driven machine learning algorithms is the poor interpretability due to the lack of components based on chemical and physical laws in the algorithm's structure. We note that our model is trained with data based on the solution of a complex system of nonlinear equations comprising mass and energy balances and thermodynamic equations, which might be impossible to express in the form of a single simple equation based on first-principles. Noticeably, the original first-principles model is more accurate than the simplified model, but the latter comes with the great advantage of substantially simplifying the modeling7. As already mentioned, building mechanistic models can be time-consuming, especially when convergence issues arise. Therefore, the streamlined equations presented here can facilitate access to the modeling tools by experimental groups, with applications in a wide range of process systems engineering problems, as discussed in Section 5.3, and eventually in industry.

Although machine learning methods can be applied to a great variety of problems, including chemical engineering-specific problems, the trained models are valid within the range of the training set, often with very limited extrapolation capabilities. Consequently, the models presented in this work cannot be applied to a different solvent, either reactive or physical, or a different solvent composition. However different surrogate models could be conveniently introduced for all the cases. On the other hand, hybrid models are developed by integrating mass and energy balances and approximating other relationships with data-driven models. The advantage of the so-called black box models, such as those presented in our work, is having a single compact expression for each dependent variable representing the entire flowsheet. Regarding the drawbacks mentioned above, we argue that the lower accuracy shown by the...
simplified model, relative to the fully mechanistic analog, would also be found in gray box (hybrid) models.

2.4 Application of the Bayesian machine scientist to time-dependent analyses

In our work, we show how the BMS is able to retrieve simplified equations based on synthetic data generated from simulations at steady-state. However, the BMS is also suitable to deal with time-dependent analyses, provided that time is given as an input variable. Alternatively, the BMS can be used to model differential equations instead of the time-variable quantities. In the original publication, the BMS is shown to accurately recover a system of coupled nonlinear differential equations. More established machine learning methods, such as artificial neural networks (ANN), can also be applied to deal with time-dependent data. In this case, they can include internal "recycle" connections to deal with dynamic problems and time-series data.

3. Residual plots

Once the closed-form expression of a dependent variable has been obtained using the BMS, various metrics are calculated to determine the goodness of the regression. The $R^2$, mean relative error (MRE) and mean squared error (MSE) of the validation set are reported in the main article, while the residuals are shown in Figure S1 for the two case studies. The residuals are calculated as the difference between given and predicted values of the output variables in the training set.
**Figure S1**: Scatter plot of the residual vs. predicted values in the training set. The plots on the left column refer to the analytical expressions of the natural gas sweetening process (1), while on the right one to the flue gas treatment (2) for the dependent variables a) cooling, b) heating utilities, c) net power required and d) amount of MEA.
4. Model training performance

We generated 1200 and 2500 scenarios in the natural gas and flue gas cases, respectively, out of which 1174 and 1245 converged, which were employed to determine the closed-form mathematical expressions for each output variable of the two case studies reported in the main manuscript. The scatter plot of the data around the regression line is shown in Figure S2 and Figure S3 for the natural gas and flue gas training dataset, respectively. The corresponding \( R^2 \), MRE and MSE are reported in Table S3 and Table S4.

![Figure S2](image-url): Given vs. predicted values correlation for the four output variables in the training dataset: a) cooling and b) heating utilities, c) net power and d) amount of MEA for the natural gas sweetening process. The regression line is shown in black.
Table S3: $R^2$, MRE and MSE statistics for each output variable of the natural gas sweetening process in the training dataset.

| Case study | Variable     | $R^2$  | MRE   | MSE       |
|------------|--------------|--------|-------|-----------|
| Natural gas| Min CU       | 0.9842 | 0.0102| 1.27E+06  |
|            | Min HU       | 0.9922 | 0.0054| 5.04E+05  |
|            | Net power    | 0.9986 | 0.0071| 2.97E+02  |
|            | Amount of MEA| 0.9930 | 0.0051| 1.70E+07  |

Figure S3: Given vs. predicted values correlation for the four output variables in the training dataset: a) cooling and b) heating utilities, c) net power and d) amount of MEA for the flue gas treatment process. The regression line is shown in black.
Table S4: R², MRE and MSE statistics for each output variable of the flue gas treatment process in the training dataset.

| Case study | Variable    | R²   | MRE   | MSE       |
|------------|-------------|------|-------|-----------|
| Flue gas   | Mín CU      | 0.4909 | 0.0707 | 1.16E+05  |
|            | Mín HU      | 0.9434 | 0.0359 | 1.64E+04  |
|            | Net power   | 0.9993 | 0.0099 | 7.71E+01  |
|            | Amount of MEA | 0.9962 | 0.0046 | 7.77E+03  |

5. Comparison with neural networks

Lastly, we want to compare the performance of the BMS to a standard machine learning approach. Numerical methods are well established in the process engineering community, and successful examples of input-output relationships applied to industrial processes have been presented\textsuperscript{10–12}. However, we argue that symbolic regression offers more advantages in terms of interpretability, complexity and flexibility of the mathematical structure than neural networks, making it particularly suited to understanding complex systems\textsuperscript{13}. ANNs, even the simplest feedforward fully connected networks, present a specific predefined structure where the number of neurons, layers, and hyperparameters must be determined beforehand. Usually, these hyperparameters are chosen through different tuning methodologies. Similarly, Gaussian processes require the selection of a kernel function, together with its hyperparameters.

In these numerical methodologies, the model has a pre-defined structure that is trained, i.e., the weights are fitted. On the contrary, the BMS does not present a singular structure, but it develops one while fitting the parameters. Additionally, the BMS can successfully identify rigorous models with as little as 100 points using less than ten parameters\textsuperscript{8}.

We use the ANN built-in function in MATLAB\textsuperscript{14} and we activate the Bayesian regularization. The results of the ANN are given in Figure S4 and Figure S5 for the training and testing of natural gas, while Figure S6 and Figure S7 for training and testing of the flue gas, respectively. The R² value of the regression is reported on top of each scatter plot.

Notably, we find that a data fit with higher R² can be obtained using the ANN and the computational time is lower. However, the ANN is a black box model rather difficult to interpret and use, especially for the type of analyses proposed in the main article in Section 5. We also point out that those variables that are regressed with the lowest accuracy using the BMS show the same behavior when the analysis is carried out with ANN (Figure S6 a) .
Figure S4: ANN results of natural gas process training dataset using the Bayesian regularization. The blue line represents the fit, the circles the data and the dashed line is the regression output = target. The four output variables in the training dataset are a) cooling and b) heating utilities, c) net power and d) amount of MEA.
Figure S5: ANN results of natural gas process validation set using the Bayesian regularization. The green line represents the fit, the circles the data and the dashed line is the regression output = target. The four output variables in the testing dataset are a) cooling and b) heating utilities, c) net power and d) amount of MEA.
Figure S6: ANN results of flue gas treatment process training set using the Bayesian regularization. The blue line represents the fit, the circles the data and the dashed line is the regression output = target. The four output variables in the training dataset are a) cooling and b) heating utilities, c) net power and d) amount of MEA.
Figure S7: ANN results of flue gas treatment process validation set using the Bayesian regularization. The green line represents the fit, the circles the data and the dashed line is the regression output = target. The four output variables in the testing dataset are a) cooling and b) heating utilities, c) net power and d) amount of MEA.

We repeat the analysis carried out in Section 5.2 using ANN instead of the BMS, whose results are reported in Table S5. We find that the error on the extrapolation of the cooling utility ratio is 12%, while the one on the power is 3.5% compared to the result obtained within the training bounds. ANN proves once again that the new technology using Stirling coolers performs better in terms of energy consumption.

Table S5: Comparison of cooling and electricity requirements for the process by Song et al.\textsuperscript{15} (cry) and our BAU (BAU) using ANN. The values are calculated as the processes' energy requirement ratio: BAU/cryogenic normalized by the absorber top product mass flow rate.

| Case I: without heat integration | MinCU\textsuperscript{BAU}/MinCU\textsuperscript{cry} | Net power\textsuperscript{BAU}/Net power\textsuperscript{cry} |
|---------------------------------|---------------------------------|---------------------------------|
|                                 | Prod. low | Prod. high | Prod. low | Prod. high |
| CO\textsubscript{2} mol. in\textsuperscript{15} (0.5%) | 2.58      | 2.25      | 1.08      | 0.94      |
(1) Safari, M.; Ghanizadeh, A.; Montazer-Rahmati, M. M. Optimization of Membrane-Based CO\textsubscript{2}-Removal from Natural Gas Using Simple Models Considering Both Pressure and Temperature Effects. *Int. J. Greenh. Gas Control* **2009**, *3*, 3–10. https://doi.org/10.1016/j.ijggc.2008.05.001.

(2) Qi, R.; Henson, M. A. Optimization-Based Design of Spiral-Wound Membrane Systems for CO\textsubscript{2}-CH\textsubscript{4} Separations. *Sep. Purif. Technol.* **1998**, *13* (3), 209–225. https://doi.org/10.1016/S1383-5866(98)00044-6.

(3) Hoorfar, M.; Alcheikhhamdon, Y.; Chen, B. A Novel Tool for the Modeling, Simulation and Costing of Membrane Based Gas Separation Processes Using Aspen HYSYS: Optimization of the CO\textsubscript{2}/CH\textsubscript{4} Separation Process. *Comput. Chem. Eng.* **2018**, *117*, 11–24. https://doi.org/10.1016/j.compchemeng.2018.05.013.

(4) Zhang, Y.; Chen, C.-C. Modeling CO\textsubscript{2} Absorption and Desorption by Aqueous Monoethanolamine Solution with Aspen Rate-Based Model. In *Energy Procedia*; Elsevier, 2013; Vol. 37, pp 1584–1596. https://doi.org/10.1016/J.EGYPRO.2013.06.034.

(5) Danaci, D.; Bui, M.; Petit, C.; Mac Dowell, N. En Route to Zero Emissions for Power and Industry with Amine-Based Post-Combustion Capture. *Environ. Sci. Technol.* **2021**. https://doi.org/10.1021/acs.est.0c07261.

(6) Morgan, J. C.; Chinen, A. S.; Anderson-Cook, C.; Tong, C.; Carroll, J.; Saha, C.; Omell, B.; Bhattacharyya, D.; Matuszewski, M.; Bhat, K. S.; Miller, D. C. Development of a Framework for Sequential Bayesian Design of Experiments: Application to a Pilot-Scale Solvent-Based CO\textsubscript{2} Capture Process. *Appl. Energy* **2020**, *262*, 114533. https://doi.org/10.1016/J.APENERGY.2020.114533.

(7) Ding, C.; Ierapetritou, M. A Novel Framework of Surrogate-Based Feasibility Analysis for Establishing Design Space of Twin-Column Continuous Chromatography. *Int. J. Pharm.* **2021**, *609*, 121161. https://doi.org/10.1016/J.IJPHARM.2021.121161.

(8) Guimerà, R.; Reichardt, I.; Aguilar-Mogas, A.; Massucci, F. A.; Miranda, M.; Pallarès, J.; Sales-Pardo, M. A Bayesian Machine Scientist to Aid in the Solution of Challenging Scientific Problems. *Sci. Adv.* **2020**, *6*. https://doi.org/10.1126/sciadv.aav9971.

(9) Trinh, C.; Meimaroglou, D.; Hoppe, S. Machine Learning in Chemical Product Engineering: The State of the Art and a Guide for Newcomers. *Processes* **2021**, *9* (8), 1456. https://doi.org/10.3390/PR9081456.

(10) Feng, L.; Zhao, C.; Sun, Y. Dual Attention-Based Encoder-Decoder: A Customized Sequence-to-Sequence Learning for Soft Sensor Development. *IEEE Trans. Neural Networks Learn. Syst.* **2021**, *32* (8), 3306–3317. https://doi.org/10.1109/TNNLS.2020.3015929.

(11) Chai, Z.; Zhao, C.; Huang, B. Variational Progressive-Transfer Network for Soft Sensing of Multirate Industrial Processes. *IEEE Trans. Cybern.* **2021**. https://doi.org/10.1109/TCYB.2021.3090996.

(12) Qin, Y.; Zhao, C.; Huang, B. A New Soft-Sensor Algorithm with Concurrent Consideration of Slowness and Quality Interpretation for Dynamic Chemical Process. *Chem. Eng. Sci.* **2019**, *199*, 28–39. https://doi.org/10.1016/J.CES.2019.01.011.

(13) Bongard, J.; Lipson, H. Automated Reverse Engineering of Nonlinear Dynamical Systems. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104* (24), 9943–9948. https://doi.org/10.1073/PNAS.0609476104/SUPPL_FILE/09476FIG4.PDF.
(14) MATLAB - MathWorks https://ch.mathworks.com/products/matlab.html.

(15) Song, C.; Liu, Q.; Ji, N.; Deng, S.; Zhao, J.; Kitamura, Y. Advanced Cryogenic CO\textsubscript{2} Capture Process Based on Stirling Coolers by Heat Integration. Appl. Therm. Eng. 2017, 114, 887–895. https://doi.org/10.1016/j.applthermaleng.2016.12.049.