# *GRICU 2022, Ischia, (Italy), July 3-6, 2022*



**Automated Machine Learning Framework for Surrogate Model Generation: Application to the Aspen HYSYS Process Simulator**

Andrea Galeazzi1,2, Kristiano Prifti1,2, Flavio Manenti1,2\*

1. *Politecnico di Milano, Dipartimento di Chimica, Materiali ed Ingegneria Chimica “Giulio Natta”, Piazza Leonardo da Vinci 32, 20133 Milano, Italy;*
2. *Consorzio Interuniversitario Nazionale per la Scienza e Tecnologia dei Materiali, Via Giusti 9, 50121 Firenze, Italy;*

*\*Corresponding author E-Mail:* [*flavio.manenti@polimi.it*](mailto:flavio.manenti@polimi.it)

## Introduction

A typical task of a chemical engineer is the modelling of complex systems with great accuracy. Usually, the main obstacle to overcome is the non-linearity of chemical processes which enhances exponentially the computational load required to obtain a significant solution. In the field of process systems engineering, one well-established tool created for lowering the complexity of these mathematical systems is the surrogate model [1]. Their purpose is to recreate the behaviour of a complex system through the usage of simpler mathematical tools that are easier to manipulate. For example, a simple surrogate model could be created using a polynomial equation for which it is possible to calculate its derivatives fast. This makes it possible to start using derivative- based optimization methods on systems where the derivative is either computationally costly or analytically impossible to evaluate [2].

The use case described in this work is for the complexity reduction of chemical process flowsheet simulations. In this specific case, the advantage is that by surrogating a steady-state simulation it could be possible to export the model (e.g. in C++, Python, etc.) and use it later for optimization purposes, ideally in real-time and online, without the need of using directly a licensed process simulator (e.g. Aspen HYSYS, PRO/II, gPROMS, etc.).

## Methods

A process simulation is simplified using a surrogate model composed of many selected algorithms (i.e., in order of complexity, linear regression, second-order polynomial regression, support vector regression, decision tree, random forest, AdaBoost), chosen based on the accuracy with respect to the predicted variable and its complexity. Thus, the regressed models are evaluated with a cross-validation methodology against all the streams' predicted variables of a flowsheet simulation, and the algorithm with the smallest residual error is chosen. The variables identified for this study are the ones required to fully saturate the thermodynamic degrees of freedom of units, and in particular, streams (i.e. temperature, pressure, flow rate, and composition).

Step zero in generating a surrogate model is to choose the specific process or unit operation, or a mix of both to be approximated. It is possible to surrogate either an entire plant or a single specific piece of equipment. The drawbacks in approximating an entire plant with a lot of machinery with a single surrogate model are mainly the data generation and, consequently, the accuracy. On the other hand, surrogating only a single piece of equipment may not solve at all the problem of computational load since the units would be numerically connected as they were before the surrogate creation, thus recycle loops and other computationally intensive operations are still present.

The data generation, which is the first step, is conducted using a design of experiment (DoE) approach [4]. The analyzed domain depends on the number of existing variables and it grows exponentially larger as the number of variables increases. For this reason, surrogating a process with a high number of streams and equipment may result in a need for generating an enormous amount of process data to evaluate every possible interaction between the variables. A reasonable compromise must be found, for example by splitting a large process into several subprocesses in the case where some independence exists between one another.

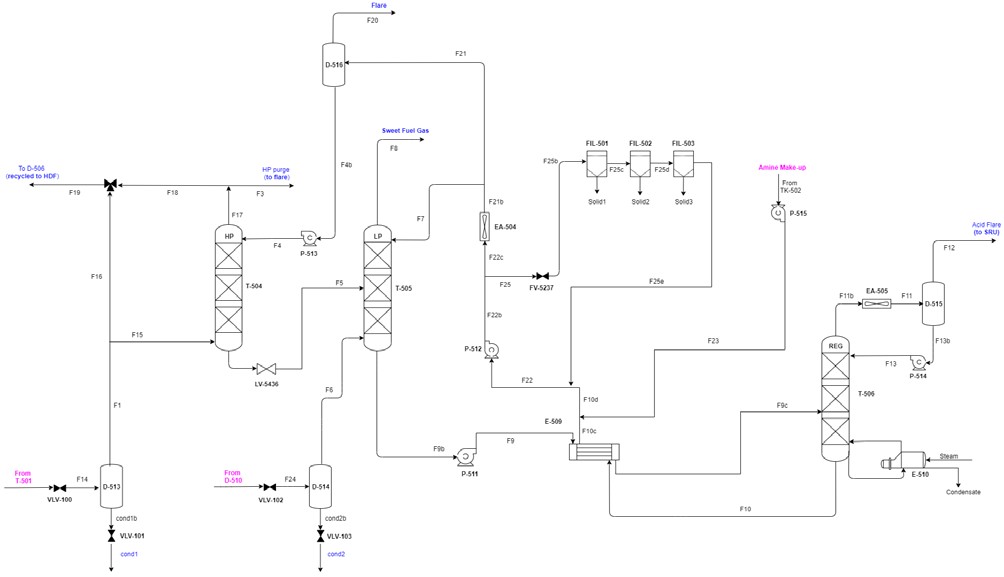
## Results and discussion

The framework described above has been applied on a steady-state digital twin (DT) of a sweetening process using diethanolamine in the exhausted oil circular refinery of Itelyum Regeneration, in Pieve Fissiraga (LO), Italy. A process flow diagram (PFD) is shown in Figure 1. The process main units are the three absorbers, one for the regeneration (REG, T-506) of the amine while the other two for high-pressure (HP, T-504) and low- pressure (LP, T-505) separation. The streams entering the process are rich in H2S, H2, CO2, and hydrocarbons. The DT has been developed using Aspen HYSYS and the accuracy has been validated against process data coming from the distributed control system.

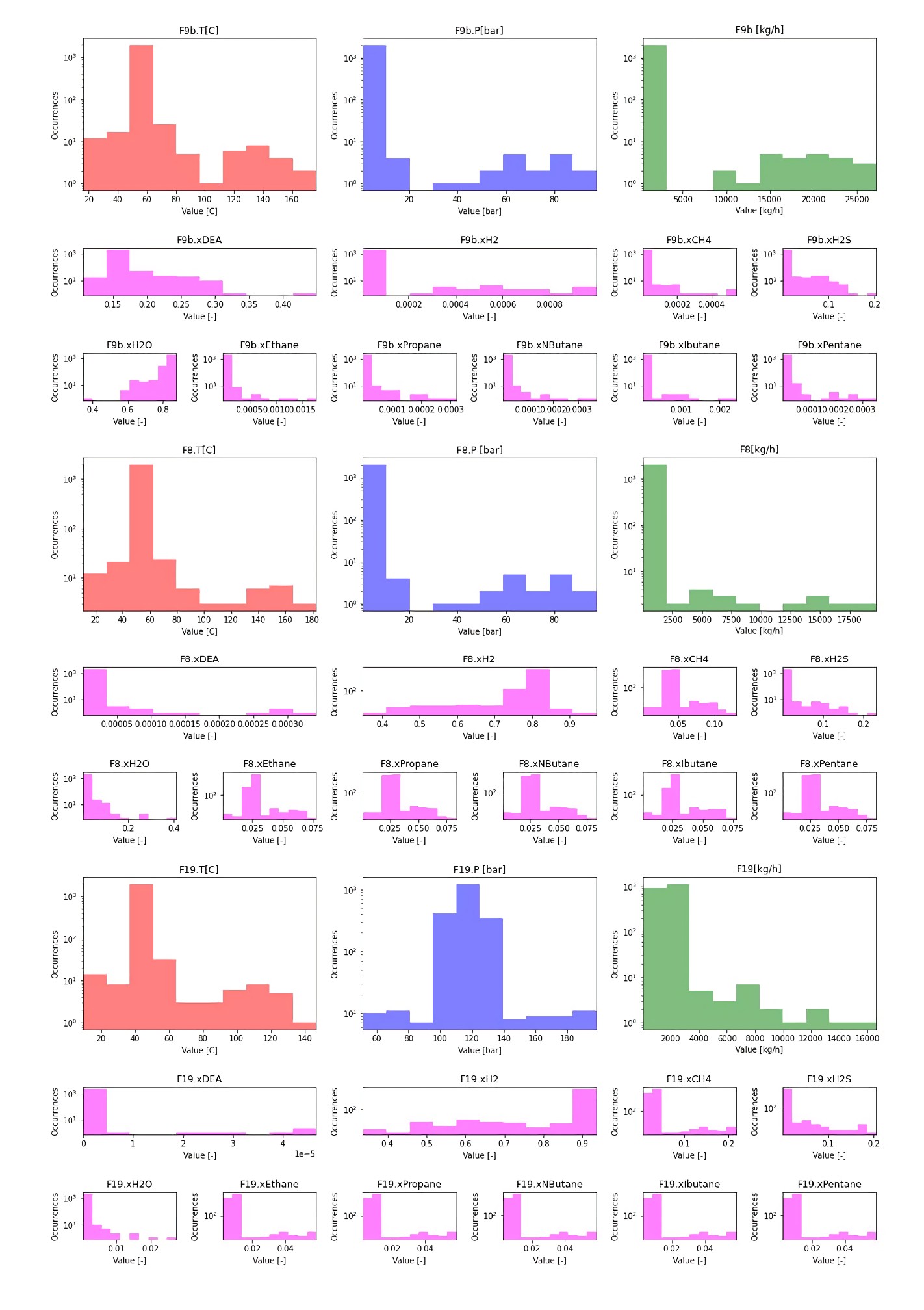
The process has been split into two parts, the regeneration section, and LP/HP section, thus creating two sets of surrogate models. The data generation step has been conducted using a Latin-Hypercube DoE [3] in two splits, with a total of 4000 data points for each section and a cut of 50% between each split. The first split is a domain closer to the steady-state (±10%) while the second one is far from the steady-state (±40%). The latter is also subject to convergence problems due to the possibility of encountering operating conditions that are thermodynamically unfeasible.

After the training of both surrogate models, the residual error is 2% and 5% for the smaller domains (±10%) of the HP/LP and the regeneration section, respectively. The error on the wider domain (±40%) is higher by orders of magnitude and this is caused by the unfeasibility of a high number of operating conditions which reduces the available converged solutions to a total of, e.g., 64 out of 2000 simulations, for the HP/LP section. Thus, the algorithms have a big disadvantage in that domain since there is approximately 2000 data point in the closest range so the wider range is neglected.

The resulting algorithms trained and chosen for each thermodynamic variable of the simulation are shown in Table 1. It is worth noting that simple linear regression is sometimes the better option with respect to other more advanced methods like random forests. This is especially true when there is a linear pressure drop imposed on a unit in Aspen HYSYS. Artificial neural networks have been tested but their efficacy has proven unworthy of their additional complexity and training time. For this reason and the lower accuracy, they have never been selected as the better option, as seen in Table 1.



**Figure 1.** PFD of the amine washing section of Itelyum Regeneration exhausted oil refinery.



**Figure 1.** Histograms of the occurrences of data points in the range of process variables for several selected streams.

**Table 1.** Number of occurrences for each best-performing algorithm for the regenerator and HP/LP absorbers sections

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Regenerator section |  | Absorbers section |  |
|  | Occurrences (#) | Occurrences (%) | Occurrences (#) | Occurrences (%) |
| Linear regression | 2 | 7.70% | 3 | 7.70% |
| Polynomial order 2 | 1 | 3.80% | - | - |
| SVR | 12 | 46.20% | - | - |
| Decision tree | 2 | 7.70% | - | - |
| Random forest | 6 | 23.10% | 28 | 71.80% |
| AdaBoost | 3 | 11.50% | 8 | 20.50% |

## Conclusions

In conclusion, the surrogate modelling approach has been applied for a case study of a real industrial process. The approach described using a careful design of experiment together with domain split near and far from the steady-state operating conditions makes it possible to reduce the computational load of a flowsheet simulation, in this case, made in Aspen HYSYS, using simpler or less computationally expansive machine learning algorithms. This method could be suitable for creating surrogate models of steady-state digital twins of industrial plants, given the existence of a reliable digital twin. Thanks to the development of such a surrogate model it is now possible to subsequently perform optimization or other costly computational procedures, like, e.g., sensitivity analysis. This methodology should be further enhanced to make it able to treat dynamic digital twins. Furthermore, the final goal of this method is to substitute the need for a process simulator that generates the data and takes this data directly from the DCS of the industrial plant, automating the process of surrogate models’ creation.

## References

[1] K. McBride, K. Sundmacher. Chemie Ingenieur Technik 91 (2019), 228–239.

[2] A. Shokry, P. Baraldi, E. Zio, A. Espuña. Industrial & Engineering Chemistry Research 59 (2020), 15634–15655.

[3] M. D. McKay, R. J. Beckman, W. J. Conover. Technometrics 21 (1979), 239.

[4] J. Eason, S. Cremaschi. Computers & Chemical Engineering 68 (2014), 220–232.