

Generative AI for the optimal design of seawater desalination processes

Valentin ZARLENGA^{abc*}, Antonio ROCHA AZEVEDO^{ab}, Alvaro MARTINEZ-TRIANA^b and Thibaut NEVEUX^{ab}

^a Université de Lorraine, CNRS, LRGP, F-54000, Nancy, France

^b EDF R&D Chatou, 6 quai Watier, 78400 Chatou, France

^c Université de Strasbourg, ECPM, 25 rue Becquerel, 67200 Strasbourg, France

* Corresponding Author: valentin.zarlenga@univ-lorraine.fr

ABSTRACT

In recent years, research for systematic process design approaches has gained traction, especially with the rise in popularity of generative machine learning models and reinforcement learning. However, works from the literature will often focus on proof-of-concept studies, limited to a specific process synthesis problem. Despite showing promising results, it is not clear how easily these methodologies could be transposed to new applications, and whether they would be successful. In this context, this work evaluates the possibility of using a Natural Language Processing model, which has already proven itself for thermodynamic cycle generation, for another different case: seawater desalination. The processes generated by this model will initially be those using reverse osmosis processes aimed at desalinating a seawater solution containing 25000 ppm of NaCl. Results show that the model has been successful in designing structural reverse osmosis desalination processes without defining assembly rules or a superstructure. Given the large number of these generated processes, a clustering method based on their similarity has been developed. This made it possible to identify different known heuristics (like multi-stage) in process engineering. This adaptation was made possible by modifying aspects external to the original model: a dedicated vocabulary, design rules and objective function.

Keywords: Process synthesis, Seawater desalination, Artificial Intelligence, SFILES, Space visualization

1. INTRODUCTION

In the work by Nabil et al. [1] and Rocha Azevedo et al. [2], a Generative Artificial Intelligence (AI) model for process synthesis was successfully applied to the synthesis of supercritical CO₂ power cycles. It works by treating the process synthesis problem as a Natural Language Processing one: flowsheets are represented by sequences ("sentences") of tokens ("words"). Then, data augmentation and fine-tuning methods are used to achieve targeted generation – generation of relevant flowsheets for a given objective function (defined by the user for each case study). The model identified known design heuristics despite having no prior domain knowledge of the application, while also discovering a novel, counter-intuitive heuristic not previously reported in the literature. This capability proves how useful these methods may be for innovative applications.

This work proposes to extend the method used in [1] and [2] for the application of water desalination. Indeed, water resources have been under increasing stress leading to their depletion. The scarcity of this resource is due to the combined effects of global warming and over-exploitation in irrigation, industrial and urban sectors. Faced with this challenge, it is becoming necessary to adopt more sustainable water resource management strategies. The recovery of industrial effluents therefore appears to be an interesting solution.

However, its deployment is often hindered by the high energy and equipment costs associated with conventional water treatment processes. The specific case of sodium chloride / water solution treatment for drinking water production, generally named desalination with seawater, is one of the most investigated problems. Designing optimal process architectures that balance energy, environmental, and economic criteria becomes a major

challenge considering the emerging water management difficulties. Furthermore, the development of new technologies for water treatment, such as membrane distillation, requires further studies to determine their relevance. The use of tools for autonomously evaluating process alternatives could be useful for quickly identifying new and useful solutions.

The primary objective is to assess how this method behaves to this new case study, with a focus on a single-technology seawater desalination system using Reverse Osmosis (RO). The second objective will focus on the generation of the results and how can they be visualized and interpreted. This well-known case study will allow for a critical analysis of its results and comparison with traditional industrial practices. Then, a section will be dedicated to future perspectives.

This generative synthesis approach is expected to explore a wide design space without prior assumptions, potentially identifying novel and highly performant process configurations. This design space will vary depending on the objective function and the constraints defined by each scenario.

2. METHODS

2.1. Data representation

The work by Rocha Azevedo et al. [2] is based on the use of a symbolic language dedicated to the representation of processes: SFILES 2.0. This language converts a process graph, comprising unit operations and their connections, into a sequence of tokens (“words”) representing equipment, branches, recycles and tags in the process [3]. It only allows the process flowsheet to be represented without internal parameters.

This notation had previously been modified to facilitate the model’s learning process for cyclic processes. The main difference between this version and the original is that a unit operation hierarchy is defined for the SFILES generation (canonical order), improving the consistency of the representation and the model’s performance [4]. This version was retained, defining the seawater solution inlet as the starting point and prioritizing the equipment defined for transcribing the process into a sequence of tokens.

In the case of this study, the processes generated will include a raw seawater entry and two global exits represented by the permeate (solution with the lowest salt concentration after treatment) and retentate (solution with the highest salt concentration after treatment). The units available for the synthesis of these processes are: mixer, splitter, pump and reverse osmosis.

2.2. Dataset generation

Currently, there is no large-scale SFILES dataset dedicated to desalination processes. Developing this

database is a crucial task. It must contain unique, random (no bias in token assembly) and valid sequences.

The algorithm proposed by Nabil et al. [1] for the generation of supercritical CO₂ cycles is adapted to the case of seawater desalination. The objective is to generate directed graphs (representing the processes), where the nodes correspond to unit operations and the edges to connection flows. The algorithm begins by creating the most basic desalination process: a raw seawater solution, a reverse osmosis or a membrane distillation (equal occurrence) and two exits.

An edge is then randomly selected before applying a mutation. Each mutation is characterized by the random insertion of one of the units available between two nodes linked by an edge. This step is repeated according to a random number between 4 and 20. When the maximum number of mutations is reached, the additional exits added during the mutation step are connected with mixers to ensure that there are only two exits named Retentate and Permeate (randomly). Finally, for each separation operation, the exit edges are randomly labelled “{P}” and “{R}” to distinguish the permeate and retentate flows. This directed graph is then converted into a SFILES sequence and a pre-established set of rules (see Supplementary Material) is used to certify the validity of the processes. 300 000 unique, random and valid sequences were thus generated for both cases.

The definition of the base dataset used to train the model is of utmost importance, as it strongly conditions the research space on which the model will operate. For this reason, it is desired that the dataset is as varied as possible. While in this work we try to achieve this by using purely random mutations (avoiding bias in this step), there isn’t a clear metric for studying the variety of such a dataset. As the number of available and total units increases, this task becomes all the more complicated, but the use of exploration methods (such as thermal sampling, which will be mentioned in the next section) can also help the model explore the research space and even go beyond its original scopes.

2.3. Model architecture and implementation detail

In order to explore a large design space, the original article is based on the sequential model including a long short-term memory (LSTM) proposed by Nabil et al. [1]. The model is trained iteratively to produce valid SFILES sequences. There are two distinct phases in this model.

The first phase consists in training the model on the syntactic rules (what makes a SFILES string represent a valid process). In it, the model is trained with random but valid SFILES, stemming from the dataset described on section 2.2, plus other valid sequences generated by the model itself, in a data augmentation loop.

The second phase will finetune the model to focus

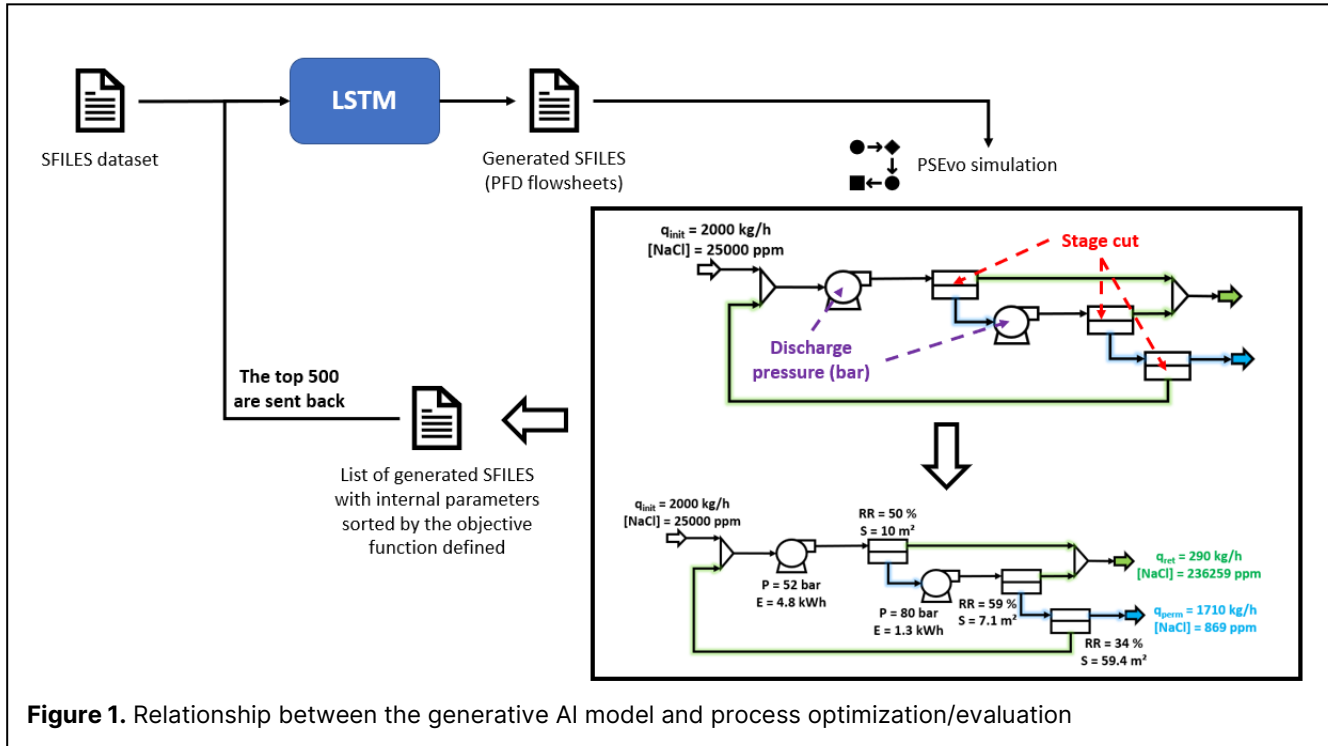


Figure 1. Relationship between the generative AI model and process optimization/evaluation

on the generation on relevant solutions according to a given objective function (which will be discussed in the next section). In this phase, the model is used for generating flowsheets in the form of SFILES. These SFILES processes are sent to the internal EDF simulator (PSEvo) where a simulation is launched. The aim is to maximize the objective function that has been defined by interacting with the design data (such as discharge pressure for the pumps or recovery rate for the reverse osmosis). After the optimization is complete, the process indicators (such as membrane surface, permeate NaCl concentration, permeate flow, ...) and value of the objective function will be extracted in order to evaluate the process efficiency. This internal simulator was crafted for efficiency and distributed calculations, which allowed the method to optimize over 20 000 processes in around 8 000 CPU hours (just a few real-time days thanks to the parallel computations performed by 10 Intel Xeon Platinum 8260 processors). The details of how this software works have been detailed in the article of Rocha Azevedo et al. [2].

Then, following the strategy proposed in [2], the best 500 flowsheets generated so far are used to fine-tune the model (see Figure 1). This process is repeated 10 times, iteratively refining the model and improving the quality of the flowsheets it generates. To improve the flowsheet exploration, a strategy of thermal sampling has been developed. This method allows the distribution of the tokens probability predictions to be influenced. A high temperature flattens the probability distribution, allowing more randomness and novelty in the generated SFILES sequences, but can also lead to more mistakes. Preliminary tests have shown that a temperature of 1.5

generates new design areas (see Supplementary Material), for that reason, it is used throughout the results presented in this work.

2.4. Objectives functions and weights

In the case of seawater desalination by reverse osmosis, we used various criteria to assess the suitability of the generated processes. These criteria are:

- The recovery rate of pure water produced (noted RR)
- The total membrane surface area used by the desalination units (noted S)
- The electrical power consumption required to generate a pressure greater than the necessary osmotic pressure (noted E)
- The NaCl concentration of the treated solution (noted x). This solution must be less than 1000 ppm (noted c_0) total dissolved salts for the process to be considered interesting.

Mathematically, all of these criteria were represented using the following formula:

$$f_{obj} = k_{Tr} \cdot RR - \left(\frac{1}{q_{perm}} \cdot (k_w \cdot E + k_S \cdot S) \right) - \begin{cases} \frac{k_c \cdot e^{m \cdot x}}{m \cdot e^{m \cdot c_0}} & \text{if } x < 1000 \\ k_c \cdot (x - c_0) + \frac{k_c}{m} & \text{if } x \geq 1000 \end{cases} \quad (1)$$

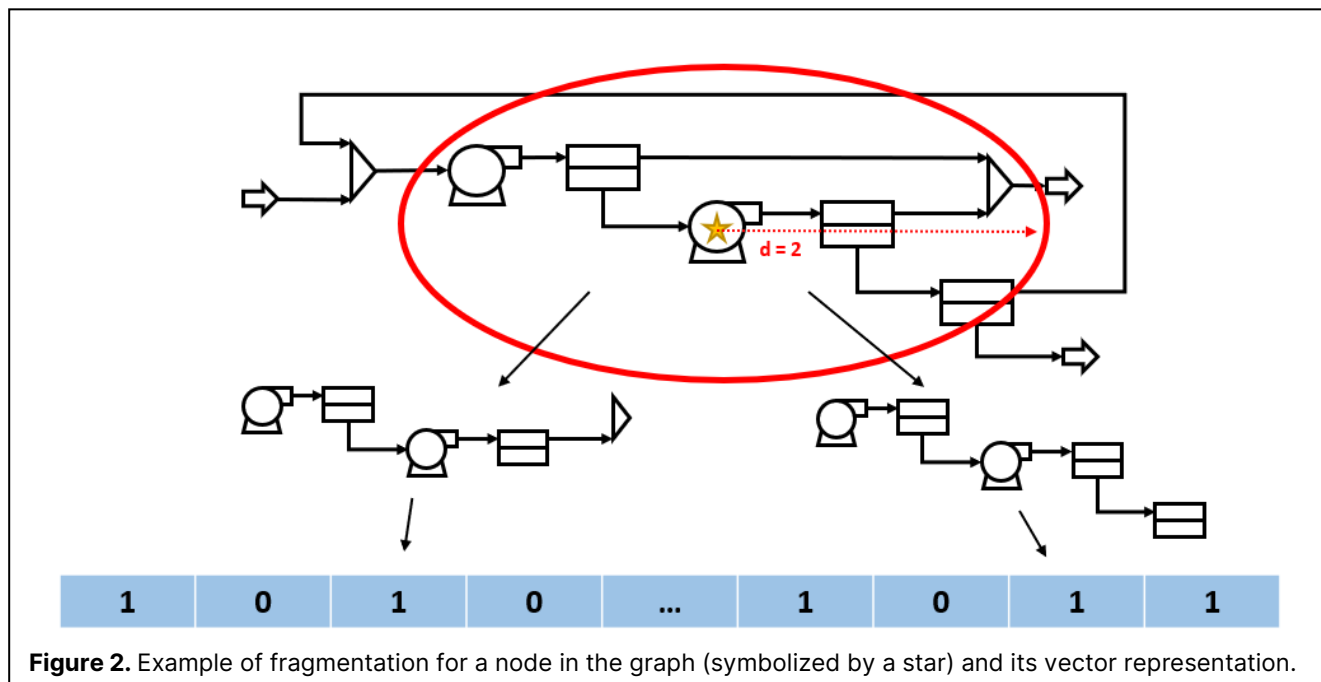


Figure 2. Example of fragmentation for a node in the graph (symbolized by a star) and its vector representation.

Here, weights (k) have been assigned to each of these parameters, in order to standardize their physical magnitude and/or prioritize certain parameters over others (see Supplementary Material). The NaCl concentration penalty will be strong if the permeate concentration is greater than 1000 ppm, and quickly (exponentially) otherwise. The term “ m ” is another weight that influences the slope of the exponential decrease.

2.5. Visualization of the generated design space

The AI method presented generates a large number of processes, making it impossible to analyze each process individually. It is therefore necessary to develop a global analysis strategy to extract a set of relevant information. The essential information is the common substructures of the processes and the search space explored.

Many articles in computational chemistry using AI models transform molecules into binary vectors (vectors whose elements are 0 or 1 called fingerprints) of fixed dimension. This representation is particularly suitable for visualization or similarity analysis. There are various vectorization processes (the most well known being Extended Connectivity Fingerprints [5]) based on rules specific to chemistry. The aim is not to replicate these algorithms but to redefine a digital representation strategy adapted to the processes field.

To do this, the SFILES sequences generated by the model are transformed into a directed graph. For each node in the graph, a subgraph containing all nodes at a distance less than or equal to d from the first node is isolated. This fragmentation is applied to all the processes

to be studied. When all the unique subgraphs have been extracted, they are associated with a component of a vector, the dimension of which is fixed by the number of unique subgraphs. For each process, this vector is initialized with zeros. If the process includes a particular subgraph, its component takes the value 1, otherwise it remains at 0. This mechanism allows the sequences generated by the model to be represented as a binary vector indicating the presence of the local patterns that constitute it. It is important to note that the distance d is a parameter set by the operator (see Figure 2). However, it should be not too low (low discriminating patterns) or too high (many long and unique patterns creating sparse vectors with little information). The value set in this work is 2.

To visualize the search space, the UMAP dimensionality reduction method is applied. UMAP (Uniform Manifold Approximation and Projection) allows complex non-linear data to be projected into a two-dimensional space while preserving the local and global topologies of the processes (suitable for revealing significant clustering patterns) [6]. The more structurally similar two processes are, the most similar the points representing them in the reduce space will be. Once the projection had been made on all the processes, a density-based clustering technique (HDBSCAN [7]) was applied to group similar processes together. Thanks to the binary vectors, we can quickly identify the most common or dominant substructures in these process families.

3. RESULTS AND DISCUSSION

This section focuses on the reverse osmosis desalination processes generated. The first subsection will

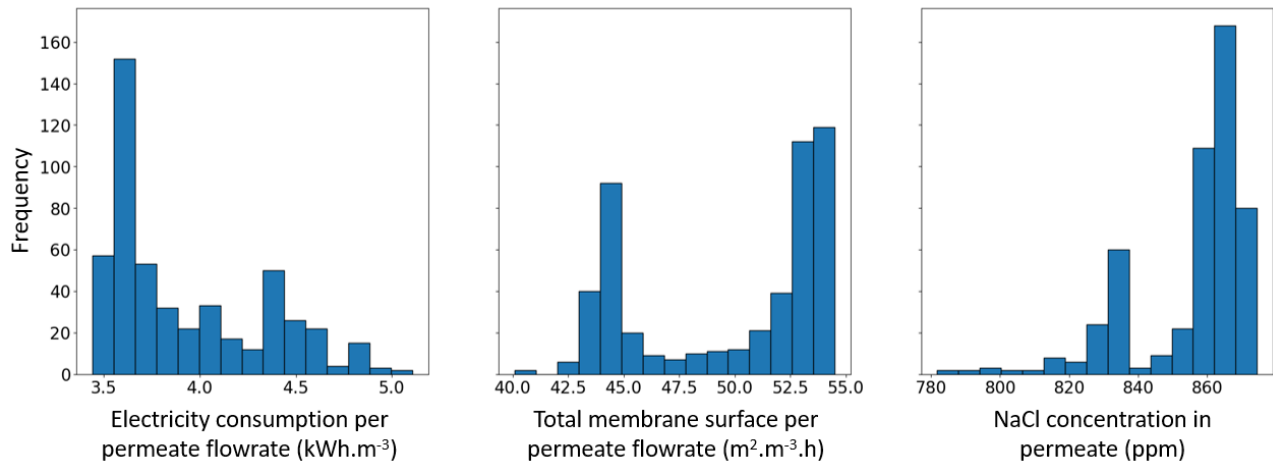


Figure 3. Histogram of internal parameter for the 500 best processes studied in section 3.2

investigate the steps necessary to adapt the initial model. Then, the next subsection focuses on the model exploration to determine if it converges towards a single zone (meaning that a single global architecture is found) or if several zones are explored. The 500 best solutions proposed by the model will be then investigated. Because this case study is well documented, it will be the basis for evaluating the model, as we can compare the flowsheets the model generates to known relevant processes. Finally, the perspectives of this work will be presented.

3.1. Model adaptation

This section lists the steps required to modify the Rocha Azevedo et al. model [2] in order to apply it to our case study.

The first step was to create the tokens used in the application. Each token has its own name, as well as an inlet number, an outlet number and a tag name for these currents (see Supplementary Material). A set of rules was created to verify the physical validity of the generated processes (see Supplementary Material). Otherwise, non-physical processes could have been simulated. When the valid token sequence needs to be optimized and then evaluated, it must be transformed into an incidence matrix where each column corresponds to a specific equipment. The algorithm used to do this had therefore to be adapted to assign the correct token to the correct simulator equipment. The same applies to the permeate/retentates flowrates. Another aspect that been modified is the post processing of the processes. Equipment that is not used for the process or unused flows has already been removed, but this post processing has been extended. The objective function has been also modified.

The LSTM model has not been modified except for the integration of thermal sampling strategy.

3.2. Evaluation of model exploration

In order to visualize how the model explores the research space, the search space has been modeled (see section 2.5) for the original random dataset (iteration “0”) plus the flowsheets generated by the method at each fine-tuning iteration (iterations 1 – 10). Only the flowsheets with a feasible value of objective function (equation 1) are considered. The objective is to highlight the structural evolution of the processes and the detection of areas that improve the objective function.

The results (see Figure 4) show that the generated processes differ significantly from those sent to the model in terms of structure and efficiency. Analysis of the objective function of the processes has highlighted regions that are not particularly relevant to the application. Future processes generated will tend to move away from these regions in order to exploit new designs in known regions or explore other regions.

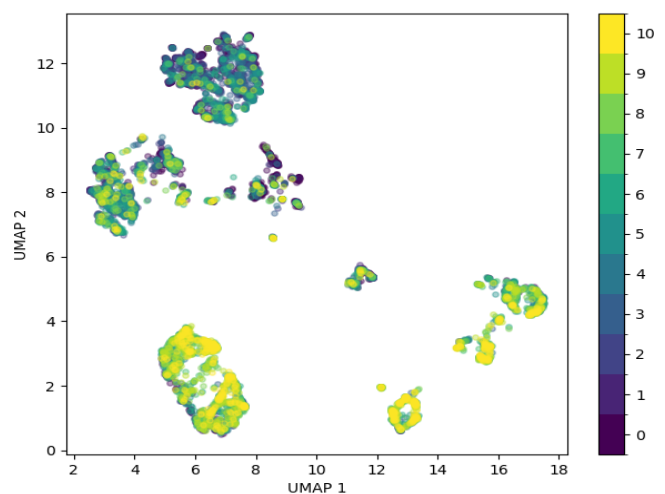


Figure 4. : Visualization of the processes generated by the model based on iteration (colorbar).

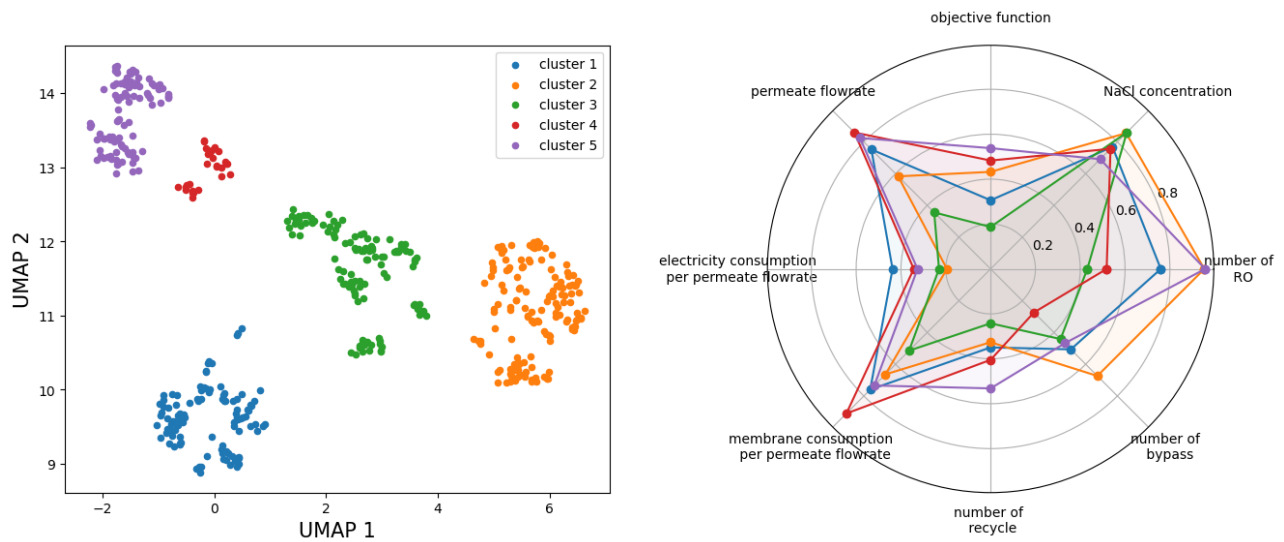


Figure 5. Process clustering (left) and Kiviati diagram of normalized data for each cluster (right) for the most relevant processes.

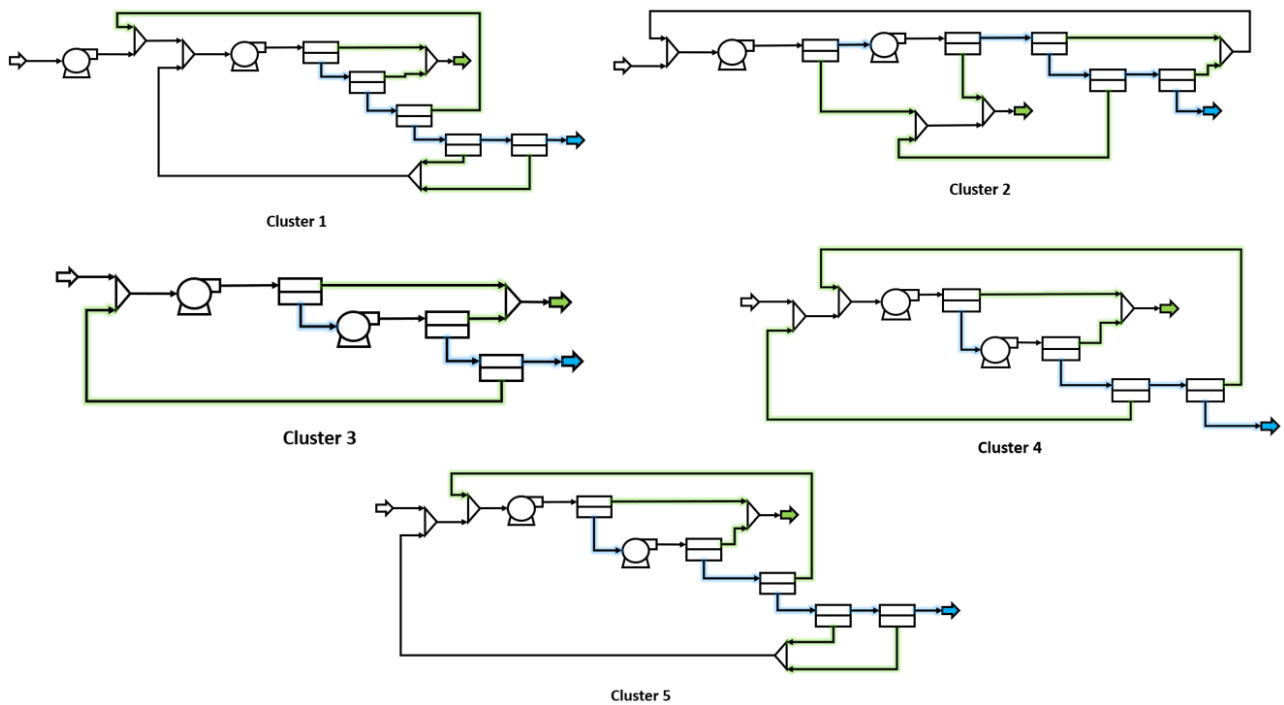


Figure 6. Randomly selected process of the each cluster. The green element correspond to retentate flows and the blue element correspond to permeate flows.

3.3. Exploitation of the most relevant processes

The 500 best processes have been selected for this study. Analysis of the results show that electricity consumption ranges from 3 and 5 kWh/m³ and total membrane surface used is between 40 and 55 m²/m³. All these processes produce water with less than 1000 ppm in salt (see Figure 3).

Using the method described in section 2.5, five clusters were detected, each with different design architectures. According to Kiviati diagram (see Figure 6), the elements of cluster 5 have the highest objective function. A random process was therefore selected for each cluster in order to analyze its architecture in detail (overview of these processes are given in Supplementary Material).

As illustrated in the Figure 6, a number of well-

known and widely documented heuristics were found in these processes, such as:

- Use of multi-stage reverse osmosis (between 3 and 5 stages) allowing separation requirements to be distributed across several levels.
- Excessively concentrated rejects are not reprocessed to avoid the use of very high hydraulic pressure or excessive membrane use by sending them directly to the retentate.
- Low concentration effluents are reprocessed to maximize water recovery.
- Use of partial recirculation

The processes of clusters 2, 3, 4 and 5 present a heuristic that those in cluster 1 do not share: a second pressurization stage after the first reverse osmosis unit. This allows us to reduce the level of electricity needed to obtain the same permeate flowrate (see Supplementary Material). The cluster 1 processes start directly at the highest possible pressure, which requires much more electricity to operate.

Unlike the other processes, those generated in cluster 3 have less reverse osmosis. Looking at the appearance of the processes, they are generated at the beginning (iteration 1) and then their number are falling in favor of processes incorporating more complex designs (see Figure 7). This proves that the model tends to move away from simple solutions to explore possibilities for designs that are more relevant to a process engineering expert.

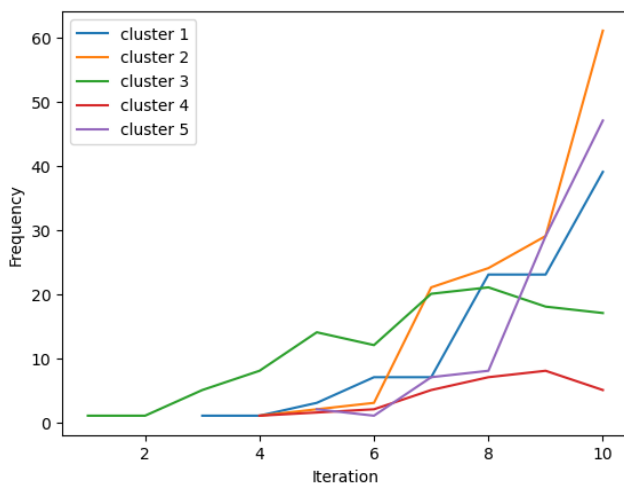


Figure 7. Appearance of processes in each cluster based on iterations.

3.4. Future perspectives

This work has demonstrated the ability of AI to generate processes incorporating reverse osmosis and certain known heuristics with strong potential for industrial application. However, several perspectives can be

implemented in order to refine the results. First, the objective function can be redefined to include more elements related to CAPEX or OPEX. This economic dimension could potentially lead to identifying simpler processes (with less separation stages) for achieving the same performance indicators. The penalty for NaCl concentration in the permeate could also be increased to obtain a weaker NaCl solutions in the permeate for the most interesting processes. It would also be interesting to have a multi-objective function and study the Pareto fronts of the generated solutions.

Nevertheless, scientific literature mentions numerous studies based exclusively on this separation technology. The integration of multi-technology processes appears to be a more relevant study with the use of membrane distillation. This type of hybrid process is generating interest despite of a lack of knowledge. A working perspective of readjusting the model and then retraining it to design this type of hybrid process (reverse osmosis and membrane distillation) should also be explored.

Finally, one could imagine coupling this synthesis approach with others methods. This could either be done: upstream, mixing known heuristics to the dataset during training, for nudging the method in the right direction; or downstream, automatically extracting useful patterns from the data generated for the development of superstructures which can better refine the research space.

4. CONCLUSION

The objective of this work was to evaluate the integration of AI to propose relevant processes for water recovery. Although the original model was proposed for thermodynamic cycle design, processes using reverse osmosis for seawater desalination were developed. These processes offer satisfactory energy and material efficiency and incorporate design architectures commonly used in industry (multi-stage reverse osmosis, ...). Generative AI has therefore made it possible to find known heuristics without prior knowledge. These heuristics were easily highlighted thanks to the development of a visualization method using UMAP and then clustering.

5. NOMENCLATURE

- RR: Recovery Rate
- S: membrane surface used
- E: electricity power consumption
- q_{perm} : permeate flowrate
- [NaCl]: NaCl concentration
- P: Pressure discharge
- q_{ret} : retentate flowrate

- q_{init} : initial flowrate

6. DIGITAL SUPPLEMENTARY MATERIAL

The supplementary material contains additional data and incorporates additional references [8]. See <https://psecommunity.org/LAPSE:2026.0030>.

7. ACKNOWLEDGEMENTS

The authors would like to express gratitude to the EDF R&D for supporting this research.

8. AUTHOR IDENTIFIERS

Author ORCIDs:

Valentin ZARLENGA: 0009-0006-6937-6884

Antonio ROCHA AZEVEDO: 0000-0002-0633-1252

Alvaro MARTINEZ-TRIANA: 0000-0002-0259-9833

Thibaut NEVEUX: 0000-0002-7460-702X

9. REFERENCES

1. Tahar Nabil, Mohamed Noaman et Tatiana Morosuk. "Data-driven structural synthesis of supercritical CO₂ power cycles". In: *Frontiers in Chemical Engineering* 5 (16 oct. 2023). Publisher: Frontiers. issn: 2673-2718. doi: <http://dx.doi.org/10.3389/fceng.2023.1144115>. url: <https://www.frontiersin.org/journals/chemical-engineering/articles/10.3389/fceng.2023.1144115/full>.
2. Rocha Azevedo A, Nabil T, Loubière V, Privat R, Neveux T, Commenge JM. Comparing generative process synthesis approaches with superstructure optimization for the conception of supercritical CO₂ brayton cycles. *Computers & Chemical Engineering* 201:109255 (2025). <https://doi.org/10.1016/j.compchemeng.2025.109255>
3. Vogel, G., Hirtreiter, E., Schulze Balhorn, L., Schweidtmann, A.M., 2023. SFILES 2.0: An extended text-based flowsheet representation. *Optim. Eng.* 24 (4), 2911–2933. <http://dx.doi.org/10.1007/s11081-023-09798-9>. URL: <https://link.springer.com/10.1007/s11081-023-09798-9>.
4. Antonio Rocha Azevedo et al. Representation Learning for flowsheets: generating structures for Process Synthesis. Publication Title: 34th European Symposium on Computer Aided Process Engineering / 15th International Symposium on Process Systems Engineering (ESCAPE34- PSE24). Mai 2024. 372-377. doi: <http://dx.doi.org/10.3303/BOA2401>. Url: <https://hal.science/hal-04733044>.
5. David Rogers et Mathew Hahn. "Extended-Connectivity Fingerprints". In: *Journal of Chemical Information and Modeling* 50.5 (24 mai 2010), p. 742-754. issn: 1549-9596, 1549 960X. doi: <http://dx.doi.org/10.1021/ci100050t>. Url: <https://pubs.acs.org/doi/10.1021/ci100050t>.
6. Leland McInnes, John Healy et James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 18 sept. 2020. doi: <https://doi.org/10.1038/s43586-024-00369-5> Url: <http://arxiv.org/abs/1802.03426>.
7. McInnes L, Healy J. Accelerated hierarchical density based clustering. 2017 IEEE International Conference on Data Mining Workshops (ICDMW) :33-42 (2017). <https://doi.org/10.1109/icdmw.2017.12>
8. Hinton, Geoffrey & Dean, Jeff & Vinyals, Oriol & Rachmad, Yoesoep. (2014). Distilling the Knowledge in a Neural Network. 1-9. url: <https://arxiv.org/pdf/1503.02531>.

© 2026 by the authors. Licensed to PSEcommunity.org and PSE Press. This is an open access article under the creative commons CC-BY-SA licensing terms. Credit must be given to creator and adaptations must be shared under the same terms. See <https://creativecommons.org/licenses/by-sa/4.0/>

