Technical Sciences
Academy of Romania
www.jesi.astr.ro

Journal of Engineering Sciences and Innovation
Volume 6, Issue 4 / 2021, pp. 473-480

H. Inter – and Transdisciplinarity in Science and Technology

Received ***30 August 2021*** *Accepted* ***19 November 2021***
Received in revised form ***25 October 2021***

Interdisciplinarity: artificial intelligence and chemical engineering

ALEXANDRU WOINAROSCHY[*]

Politehnica University Bucharest, Romania
Technical Sciences Academy of Romania, Bucharest, Romania

Abstract. There are exposed implications of artificial intelligence in addressing important problems from a technical and economic point of view whose traditional solution would not be possible or would require a huge computational time: selection of the most efficient catalyst, and determination of the optimal sequence of processes for the separation into individual components of the multicomponent mixtures in the liquid state.

Keywords: artificial intelligence, chemical process optimization, catalyst selection, separation sequences, multicomponent mixtures.

1. Introduction

Interdisciplinarity or interdisciplinary studies involve combining two or more academic disciplines in a single activity (for example, a research project), gathering information from several fields by thinking across disciplinary boundaries. Large engineering teams are usually interdisciplinary, although the term "interdisciplinary" is sometimes limited to academia. Interdisciplinarity can also be applied to complex topics that can only be understood by combining the perspectives of two or more areas.

Artificial intelligence (AI) was founded as an academic discipline in 1955, and in the years since has experienced several waves of optimism, followed by disappointment and again renew by new successfully approaches. The field was founded on the assumption that human intelligence "can be so precisely described that a machine can be made to simulate it". AI has proven its usefulness in a wide range of scientific and technical applications in various fields and is particularly effective in solving problems that include combinations or a huge number of possible solutions. The computer time for solving them traditionally, even on high-

[*] Correspondence address: a_woinaroschy@chim.upb.ro

performance supercomputers, is overwhelming. Chemistry, biochemistry, genetics, chemical engineering are mainly fields involving AI. The number of applications is huge. In the following, we will limit from thousand applications in the field of Chemical Engineering to only two: catalysts design via machine learning and synthesis of separation sequences of multicomponent mixtures by minimization of the informational entropy of learning heuristic rules sets.

2. Catalysts design via machine learning

The chemical space of all possible compounds is almost inconceivably vast, with just the number of small organic molecules estimated to be $>10^{60}$. Clearly, it is impossible to synthesize and characterize all possible compounds. So, there is great interest in the development of efficient methods to search for molecular compounds and materials with desirable properties without having to test all possible structures. In particular, the search for efficient catalysts is central to a wide range of chemical processes, with 80% of all manufacturing requiring catalysis at one or more steps of the production mechanisms. The importance of catalysis selection and design by computer is in chemistry as the Holy Grail [1].

Traditionally, the search for molecular compounds and materials with desired properties has been based on the so called direct method where a library of promising candidates is generated and then experimentally screened to identify compounds with suitable properties. However, the number of possible molecules in the library grows exponentially with the number of sites that could be modified, so the cost and time needed for synthesis and testing can be massive. Computational methods can reduce the experimental burden by narrowing the range of possibilities. It is appreciate [2] that it need 5000 years of experimentation time of a chemist who averages 50 compounds a year, 100 days of calculation time using 200 CPUs, and 16 days calculation time using cloud computing.

The inverse design strategy is the reverse of the direct method since it starts with a desired target property and tailors a structure with that property. Often, an initial reference structure is gradually changed by following the gradients of the expectation value of the property with respect to the parameters that define the chemical identity.

Several algorithms that have been developed for inverse design of molecules and materials especially methods relevant to catalysts and/or catalytic properties. Three major techniques are: 1. gradient-based methods such as the linear combination of atomic potentials, 2. alchemical transformations, and 3. machine learning techniques. We present here machine learning (ML).

ML is the study of computer algorithms that improve automatically through experience and by the use of data [3]. It is seen as a part of artificial intelligence. ML algorithms build a model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so. ML algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks[3]. ML involves computers discovering how they can perform tasks without being explicitly programmed to do so. It involves computers learning from data provided so that they carry out certain tasks. For simple tasks assigned to computers, it is possible to program algorithms telling the machine how to execute all steps required to solve the problem at hand; on the computer's part, no learning is needed. For more advanced tasks, it can be challenging for a human to manually create the needed algorithms. In practice, it can turn out to be more effective to help the machine develop its own algorithm, rather than having human programmers specify every needed step [4]. The discipline of ML employs various approaches to teach computers to accomplish tasks where no fully satisfactory algorithm is available. In cases where vast numbers of potential answers exist, one approach is to label some of the correct answers as valid. This can then be used as training data for the computer to improve the algorithm(s) it uses to determine correct answers. Largely, ML is capable of solving classification and regression problems. This means that if one wants to use ML for chemistry they must first phrase their problem either as separating data into classes based on differences in the values of descriptors used to describe that data (classification problem) or looking for a relationship between an input set of features and an output (regression problem). As a trivial example, a classification problem may be inputting free energies and separating reactions into spontaneous and nonspontaneous. A regression problem may be relating ligand-withdrawing capability to reaction rate. Ensuring that the problem fits the capabilities of machine learning is the first hurdle to applying these methods. One of the most common limiting factors in applications of ML to chemistry is the availability of enough data for reliable parameterization of classification models. Databases such as ChemDB, ChemSpider, The PubChem Project, and The National Chemical Database Service hosted by the Royal Society of Chemistry provide valuable repositories of structures and data. Expanding past these databases often requires a great deal of computer science skill, though some have attempted to make this process easier. One such example was the development of Algorithm for Chemical Space Exploration with Stochastic Search (ACSESS). ACSESS allows for the systematic identification of missing components of already explored chemical space and the expansion into unknown regions to generate new libraries. Extensions to the algorithm have added preference toward the exploration of diverse molecules with desired properties. Though this method has only yet been used to explore small organic molecules, it presents a promising start for those wishing to explore the frontiers of chemical space. Few studies have as of yet been reported as using ML for inverse design of catalysts. Most of these efforts have been limited to prediction of catalytic activity and catalytic reaction pathways. As the range of possible pathways based on available sites, possible structures, and experimental conditions involved in a typical reaction can be enormous, it makes sense to invert design techniques to explore them. For example, the reaction of carbon monoxide with hydrogen gas has more than 2000 possible pathways. Nevertheless, a significant reduction in the range of possibilities was achieved by using ML conjunction with principal component analysis and with the assistance of group additively. Through the use of ML, a screening of pathways at lower computational cost was achieved, while viable pathways for catalytic reaction were identified for additional study, with some aspects of selectivity already confirmed in experimental literature. Even when a reaction pathway is known, the catalyst which will assist the most in lowering the reaction activation barrier remains to be determined. This problem was addressed for the Suzuki cross-coupling reaction using machine learning to predict the reaction energies of catalysts and plotting them on a reference volcano plot. The goal of volcano plots is to identify catalysts that bind substrates strongly, but not too strongly, thus setting them at the activity peak of the “volcano” [1]. This ML method varied both the ligands and metals and discovered 557 catalysts that fit in the Goldilocks region of the volcano plot. Starting from a database of 25116 possible realistic structure candidates, this study revealed the ability of ML to generalize patterns across varied metal and ligand types, even when some of the test ligands were not present in the training sets. Further studies using this technique could work on eliminating the reliance on reference volcano plots.

Looking practically from the experimental side, reaction yield can be of great importance not only for turnover numbers but also especially for multistep reactions where product may be lost at every step. Ahneman et al. [4] examine the use of high throughput experimentation for generating output labels in the form of reaction yield for the Pd-catalyzed Buchwald−Hartwig C−N cross-coupling reaction the presence of isoxazoles. Using publicly available reagent geometries and scripts, catalyst descriptors were generated and used as input for multiple ML methods. Performing additional statistics to verify generalization uncovered the importance of understanding the underlying chemistry in train/test set splitting and found that active splitting performed better in generalization tests than random splitting. This additional testing is the result of communication with data scientists and highlights the crucial nature of applying null hypotheses and other statistical methods when utilizing ML.

One of the important challenges of computational modelling of catalytic systems is the description of the relative stability of configurations. ML could significantly improve the computational efficiency of configurational sampling. In fact, ML has already been applied to a wide range of studies where the relative energies of different configurations are critical, including the description of reactive gas-surface dynamics of N_2 on Ru (0001), CO_2 adsorption on Au/Cu alloy surfaces, and formation energies of elpasolites made from all main-group elements up to Bi.

Transfer learning is a ML technique to repurpose a model trained on one task for a second related objective. In the simplest form, transfer learning ads on layers to a network retrained for a similar task to fit a new problem while retaining the lessons from the previous purpose. Transfer learning is an excellent way to save computational time by generating structures from potentials rather than using the forward methodology. At the same time, this avoids repetition of previous calculations by tailoring previous work to new research objectives.

Transfer learning is one method to bridge the bottlenecks of machine learning in chemistry. Explicitly, these bottlenecks are access to data, covering enough of a property subspace to describe a desired relationship, complexity of chemical interactions increasing size of model needed, generalizability to vast chemical space, and representation of system. Data are the key unlocking relationships with ML. The generation of large data sets of chemical data, similar to those generated in the biochemistry field but particularly for catalysis, is needed for continued development of catalysts with machine learning. Such data sets must reach widely across chemical space but also deeply into the pockets of space that are examined so that complex chemical relationships can be unwrapped. In particular, relating catalytic performance to molecular properties is an open problem that nonlinear regression methods may be able to successfully tackle given enough well-selected data. Additionally, more complex relationships often require deeper models, which increase memory and processing requirements. With the improved data sets, models could be trained to be generalizable to wider spaces of data, increasing the utility of individual models.

Finally, representation of systems in ML remains an open question. The creation of a compact data structure containing all pertinent chemical interactions offers them promise of model training and pre-processing speedup, allowing individual researchers to compete more evenly with industrial giants. Each of these bottlenecks presents opportunities for research in chemistry as well as math and computer science. In the interim though, it is clear that advances in ML for the inverse design of catalysts have laid a foundation on which further research may be built.

3. Synthesis of separation sequences of multicomponent mixtures by minimization of the informational entropy of learning heuristic rules sets

Synthesis of separation sequences of multicomponent mixtures is concerned with the optimal selection and sequencing of separation operations. Several such problems may exist in a chemical process which include, for example, feed separation, reactants recovery and products separation, products finishing, and waste treatment section.

The choice of appropriate separation operation and their sequencing is a huge combinatorial problem [5]. However, in many processes, ordinary distillation using only energy-separating agent is feasible and frequently the most economically attractive means of separating more-or–less pure components or groups of components from a multicomponent mixture. Although the number of possible sequences for a unique separation is greatly reduced from the general separation process system, the problem may still be lowering when a large number of components must be separated.

If the mixture contains only 3 components A, B, C, two distillation operation are necessary, and two different sequence of the two operations are possible. If the order of decreasing volatilities is A, B, C the two possible sequences are:

a) The so-called “direct” sequence consists of separating A (more volatile component), from B and C in the first operation, and then separating B from C in the second distillation;

b) The so-called “indirect” sequence proceeds by first separating C (less volatile component), from A and B, followed by the separation of A from B.

As the number of components to be separated increases, the number of possible sequences increases quickly. If n_C components are separated into n_c products then $n_C – 1$ ordinary distillation operations are required, assuming that each receives a single feed and produce a single one top and a single one bottom products.

The number of such possible sequences n_S for n_C components separated into n_c products is [5]:

$$n_S = [\, 2\, (\, n_C - 1)\,]\,! \ \ / \, [\, n_c\, !\, (n_c - 1)\, !\,]$$

n_c	n_s
2	1
3	2
4	5
5	14
6	42
...	...
11	16796 .

Fortunately, there are very few industrial separation cases which contain six or more different products. When there are considered a number of T different types of separation operation, the number of sequences calculated by previous relation must be multiplied by $T^{(nc-1)}$ (for T = 2 and n_c = 11, then 2^{10} = 1024 and n_s is approx. 17 millions).

In an attempt to overcome these combinatorial difficulties several researchers have proposed the use of a learning algorithm based on sets of heuristic rules. Examples of such heuristic rules are:

1. Direct sequence is generally favoured when the remaining heuristic rules do not apply.
2. Sequences which give a more nearly equimolar division of the feed between the distillate and bottom product should be favored.
3. Separation where the relative volatility of the two adjaacent components is close to unity should be performed in the absence of other components. That is, such separation should be reserved until last in the sequence.
4. Separation involving high specified recovery fractions should be reserved until last in the sequence.
5. Remove any toxic or corrosive components early in the sequence.

In fact, the number of heuristic rules is more higher. It can be observed that the heuristic rules can be in agreement or in opposition. For synthesis of a separation process with n_c components and T=1 is necessary to select n_c-1 separation stages, therefore there are n_c-1 decisions. If for each decision stage there are a number of r_{si} heuristic rules, the learning algorithm based of these rules is:

1. Initial, for each decision stage and each heuristic rules for weighting coefficients wij for i = 1.. n_c - 1 and j = 1...r_{sj} are assigned the same initial value (e.g. w_{ij} = 1 for all i and j) and also a high arbitrar value for the performance index (e.g. the total investment and operation cost) of the process.
2. The process is synthesized using the heuristic rules with the highest values of the weighting coefficients.
3. If the current value of performance index is better like the best previous value, (lower in case of minimization) the values of the used weighting coefficients are increased with Δw. If the current value of performance index is weaker like the best previous value of performance index, than the values of the used weighting coefficients are decreased with Δw. In both cases the values of unused weighting coefficients remain unchanged.
4. The procedure is repeated from the step 2 up to an imposed maximum number of learning stages L_{max}.

This learning algorithm can be evaluated and improved by the use of informational entropy [7]. If in a system are present N decision rules, the informational entropy of the system is:

$$S = -\sum_{i=1}^{N} p_i \ln p_i$$

The probability p_{ij} of the heuristic rule at i decision stage is:

$$p_{ij} = \frac{w_{ij}}{\sum_{j=1}^{r} w_{ij}}$$

The total informational entropy is the sum of informational entropy of each decision stage i = 1.. n_c - 1 :

$$S = \sum_{i}\sum_{j} - p_{ij} \ \ln \ p_{ij}$$

After a learning step the total entropy can be decreased if the one or more used rules are promising, can be increased if one or more used rules are improper, or can be remain constant if some used rules are good and others unfit. If after several learning steps, for one or more stages the entropy are not increasing, then the set of corresponding rules are improper. In these ways we can assist and increased the performance of the learning process. Of course it cannot be assumed that the final best solution is the optimal solution. But this happen always when are used heuristic elements.

A good design of a separation system has a high economic importance.

5. Conclusion

In chemical engineering there are thousands problems very difficult or impossible to be solved by classical means. In these cases the interdisciplinarity between chemical engineering and AI is till now the unique method of solving. In many cases the solution of these problems has a great theoretical and/or economic importance.

References

[1] Freeze, J. G., Kelly H. R., Batista V. S., *Search for Catalysts by Inverse Design: Artificial Intelligence, Mountain Climbers and Alchemists,* Chem. Rev., 119, 2019, p. 6595−6612.

[2] Weiser J., *Digital Transformation in Chemistry and Material Research,* web event CHEManager, 22 june 2021

[3] Mitchell T., *Machine Learning*, McGraw Hill, New York, 1997.

[4] Alpaydin E., *Introduction to Machine Learning* (Fourth ed.). MIT, Boston, 2020.

[5] Ahneman D. T., Estrada J. G., Lin S., Dreher S. D., Doyle, A. G., *Predicting Reaction Performance in C−N Cross-Coupling Using Machine Learning*., Science, 360, 2018, p. 186−190.

[6] Woinaroschy A., *Unit Operations in Chemical Engineering*, UPB Printing House, Bucharest, 1994

[7] Guiașu S., Theodorescu R., *Teoria matematică a informației*, Editura Academiei RSR, Bucuresti, 1966.