Multi-objective reinforcement learning for self-optimization of flow chemistry

Ashish Yewalea, Yihui Yangb, Neda Nazemifardb, Charles D. Papageorgioub, Chris D. Riellya, Brahim Benyahiaa\*

a Department of Chemical Engineering, Loughborough University, LE11 3TU Leicestershire, UK

b Process Chemistry and Development, Takeda Pharmaceuticals International Company, 40 Landsdowne Street, Cambridge, Massachusetts 02139, United States

b.benyahia@lboro.ac.uk

Abstract

The pharmaceutical industry is undergoing a paradigm shift with the increased adoption of digitalization leading to more effective process design and operating strategies. The identification of more selective, robust, and cost-effective synthetic pathways for active pharmaceutical ingredients is crucial during early development stages. In addition, more sustainable process designs and production strategies are increasingly being adopted, which require multi-objective optimization strategies to find the best decision compromises and feasible operating windows. Hence, this work proposes a Multi-Objective Deep Deterministic Policy Gradient (MODDPG) method to handle conflicting objectives and find Pareto optimal solutions for complex continuous-flow reaction schemes. The method is validated using Claisen-Schmidt reactions in a tubular reactor and compared against benchmark methods such as Genetic Algorithms, and Bayesian Optimization.

**Keywords**: Reaction Optimization, Multi-objective Reinforcement Learning, Flow Chemistry, Multi-Objective Deep Deterministic Policy Gradient (MODDPG).

* 1. Introduction and background

The synthesis of the Active Pharmaceutical Ingredient (API) is a crucial step in the production of all pharmaceuticals. The growing demand for safer and more effective medications has led to significant changes in the manufacturing technologies, requiring accelerated development and integration strategies. This has led to an increased demand for more flexible, cost-effective, and advanced technologies to deliver high quality medicines, while achieving stringent regulatory and environmental considerations. Achieving some of these critical objectives requires more effective process design and optimization of the synthetic pathways to meet urgent healthcare requirements by significantly reducing production time, while upholding the greatest standards of quality and safety.

The complex chemical reactions involved in API manufacturing require specialists to evaluate a wide range of reaction parameters, including discrete and continuous decision variables. This problem is commonly addressed based on single objective experiment intensive methodologies or even on trial and error, which can be time-consuming and may lead to poor or suboptimal solutions. To address these limitations and speed up the development process, computer-aided optimization algorithms combined with continuous flow chemistry and advanced process analysis techniques (PAT) have become the focus of scientists and engineers in recent years. These methods minimize human intervention and inherent bias, ensuring enhanced key performance indicators such as high productivity, reduced reaction time, and improved risk management. Integrating these technologies allows real-time self-optimization of reaction processes, enhancing efficiency and adaptability. It also prevents suboptimal solutions commonly associated with off-line model-based strategies due to model uncertainties or/and operational bias.

Single-objective algorithms such as Nelder-Mead Simplex (Fath et al., 2020), and SNOBFit (Clayton et al., 2020) have been extensively used to optimize single objectives of chemical reactions. However, for API and high value chemical manufacturing, it is highly critical to consider multiple performance criteria. Often, multi-objective optimization problems are converted into single objective optimization problems using a weighted sum. However, this approach requires prior knowledge of the relative importance of each objective, which is commonly highly uncertain in most real-world problems.

In this research, an artificial intelligence-based multi-objective reinforcement learning (MORL) technique is used to self-optimize the operating conditions of a multistep continuous reaction process. The proposed RL, which is a sequential decision-making process, is combined with the first principle knowledge of the chemistry and the reaction process to quickly identify the optimal reaction conditions under multiple objective optimization settings (Yewale et al 2023, Benyahia et al., 2021). In recent years, RL methods have increasingly been implemented to identify optimal reaction conditions or to improve process control (Zhou et al., 2017, and Neumann and Palkovits 2020). However, most of these studies were designed for single objective optimization. Hence, this will be the first attempt to apply a multi-objective reinforcement learning approach to address self-optimization of flow chemistry, where conflicting objectives can be considered simultaneously.

* 1. Problem statement or Methodology

Multi-objective optimization problems (MOPs) which are characterized by several competing or conflicting objectives are encountered in most real-world problems (Benyahia, et al., 2010; Liu et al., 2023). Instead of a single optimal solution, these problems have a set of Pareto-optimal solutions (PS), which represent the best trade-offs between the objectives. In this work, a MORL methodology was developed to address a flow chemistry problem in a dynamic environment where the agents can interact, learn and adapt in real-time. The optimization process involves finding a policy that achieves the best compromises amongst multiple objectives. The MORL's adaptable structure and transfer learning capabilities allow the agents to quickly adapt to new process conditions, unlike the traditional multi-objective optimization strategies which are set at fixed conditions leading to poor flexibility and very limited transferability.

* + 1. Introduction of the MODDPG

RL uses a Markov decision process to guide successive agent-environment interactions to maximize the long-term rewards. In this procedure, the agent begins with a set of specific environment conditions (i.e., reaction conditions represented in the state space). The agent then implements the chosen actions to the environment, and as a result it receives rewards or penalties which help evaluate the overall quality of the executed actions. Based on this recurring interaction between the environment and the agent through the states, actions, and rewards, the agent learns to identify the best possible actions that maximize the cumulative rewards.

The proposed MORL approach is a generalized version of the standard RL, which involves the extension of scalar reward to a multiple reward vector. In essence, the MORL evaluates a vector of rewards associated with the selected objectives based on the proposed actions. An algorithm based a combination of Deep learning and RL (DRP), allows the agent to learn and gain knowledge about the complex reaction environment based on repetitive interactions and adjustments of the actions. A Deep Deterministic Policy Gradient (DDPG) method, which extends the Deterministic Policy Gradient (DRL) based framework, is proposed to solve the current MOP. DRL uses a neural network (NN) to map out the set of inputs onto a set of outputs. This complete framework will be defined as a Multi-Objective Deep Deterministic Policy Gradient (MODDPG) as described by the architecture shown in Figure 1 (a). To extract temporal information from sequential acts and Q-values, NNs are employed as the actor and the critic approximator networks.



Fig.1 a) Framework of the MODDPG RL agent b) Reaction set up for the proposed Claisen-Schmidt condensation reaction.

* + 1. Process Model/ Environment

The MODDPG is implemented and tested using a multistep continuous Claisen-Schmidt condensation reaction between acetone and benzaldehyde where the objective is to maximize benzalacetone yield over dibenzalacetone (Fig 1(b)). It is worth noting that it is difficult to produce benzalacetone because the two α-Hs in acetone reacts with benzaldehyde quickly, often resulting in dibenzalacetone which is the undesired product. As a result, identifying and continuously monitoring the optimal reaction conditions is necessary to maximize the yield and selectivity of benzalacetone.

Before deploying the RL agent in a real system, it is important to train, test and refine it using a mathematical model. This study uses a mechanistic model for the training/optimization framework. The model was constructed based on several assumptions including: (i) the reaction is homogeneous, (ii) the reactor is plug-flow (i.e., no axial diffusion), and (iii) the heat transfer coefficient is constant.

The mass balances of the reactants and products in the tubular reactor are presented by Eq(1) below.

$\frac{∂C\_{iz}}{∂t}=-v\_{z}\left(z\right)\frac{∂C\_{iz}}{∂z}\pm r\_{i}$ (1)

where $c\_{iz}$ represents the concentration of reactant or product $i$, *vz* is the velocity in the $z$ direction, and $r\_{i}$is the rate of reaction.

The energy equation accounts for the heat of reaction, diffusive flux, and heat exchange between the reaction side and the jacket. The energy balance equation is given by Eq (2).

$\sum\_{}^{}C\_{p}\frac{∂T\_{z}}{∂t}=-v\_{z}\left(z\right)C\_{p}\frac{∂T\_{z}}{∂z}\pm \sum\_{}^{}r\_{i}\*∆H\_{react}+UA\left(T\_{c}-T\_{z}\right)$ (2)

where $C\_{p}$ refers to the specific heat capacity, $T\_{z}$ and $T\_{c}$ represent the temperature of the tubular reactor and coolant respectively, UA is the overall heat transfer coefficient, and $∆H\_{react} $is the reaction enthalpy.

The reaction rates of the different species are given by

$r\_{A}=-k\_{1}C\_{A}C\_{B}$ (3)

$r\_{B}=-k\_{1}C\_{A}C\_{B}-k\_{2}C\_{B}C\_{D}$ (4)

$r\_{D}=k\_{1}C\_{A}C\_{B}-k\_{2}C\_{B}C\_{D}$ (5)

$r\_{U}=k\_{2}C\_{B}C\_{D}$ (6)

The reaction rate constants ($k\_{1}$ and $k\_{2}$) follow an Arrhenius law.

For the sake of computational effectiveness, the method of lines (MOL) was used to convert the partial differential equations above into a set of ordinary differential equations which can be solved by using ordinary solvers.

As mentioned in section 2.1, MODDPG uses a vector of reward functions to address the MOP. Here, two objective functions were considered namely: the reaction conversion and the selectivity. The objective functions given below were both maximized.

$Conversion \left(X\_{B}\right)=\frac{C\_{B,0}-C\_{B}}{C\_{B,0}}$ (7)

$Selectivity \left(S\_{D}\right)=\frac{C\_{D}}{C\_{B,0}-C\_{B}}$ (8)

* 1. Results and discussion

It is important to accurately simulate the reactor behavior and use the simulations as an environment to train the agent. The mechanistic model introduced in the previous section can be used as an environment to perform the intended MODDPG optimization. Firstly, a comprehensive sensitivity analysis was carried out to rank all possible inputs or manipulated variables according to their impact on the outputs. The results suggest that 4 inputs are enough to generate optimal impact on the objective functions and effectively deliver a good trade-off between exploration and exploitation.

The training performance of the MODDPG agent is shown in Fig. 2. Each of the colored stripes in Fig. 2 represents 10 training episodes, where each circle represents the reward obtained at the end of each episode. In each episode, at least 10 e-greedy actions must be undertaken in order to get rewards based on the performance associated with the selectivity and conversion. Fig. 2 also suggests that the range of variation or exploration associated with the conversion is narrower compared to selectivity. This may be attributed to the inherently fast kinetics which in this case result in fast and nearly complete conversion of the limiting reactants.



Fig.2. Training performance of the MODDPG agent for: a) conversion, and b) selectivity.

Figure 3 (a) shows the evolution of the Pareto front identified by the MODDPG agent over the training episodes. Clearly, the Pareto front undergoes significant improvement compared to the initial Pareto, which was identified at earlier training episodes. Based on Figure 3(b), the conversion, considered here as the first objective function, spans a range of variation of 9%, whereas the selectivity (i.e., second objective function) spans a range of variation of 29%. These observations are overall consistent with the discussions associated with figure 2 above and confirm that all Pareto solutions exhibit relatively high conversions. Nevertheless, the range of variation of the selectivity is still broad enough which highlights the inherent conflicting nature of the selected objectives and the need for a set of compromises to help achieve well-informed decision making.



Fig.3. a) Pareto solutions generated by the MODDPG agent over the training episodes, and b) Final Pareto front obtained by MODDPG agent vs. genetic algorithm.

The results obtained based on the MODDPG agent were compared against those obtained based on a popular non-dominated sorting genetic algorithm (NSGA-II). A hypervolume (HV) evaluation metric, also called the Lebesgue measure, is used to compare the quality of the Pareto solutions. The Lebesgue measure evaluates the quality of the approximated solutions and distribution of the nondominated solution along the Pareto front based on an HV indicator score. Higher HV values indicate better Pareto results and suggest that the obtained solutions are closer to the true Pareto front. The proposed method produces HV indicator values of 0.78 and 0.53 for MODDPG and NSGA-II, respectively. These scores clearly demonstrate that the at the proposed MODDPG method outperforms the standard NSGA-II. Figure 3(b) clearly confirms these results as the MODDPG method delivered solutions that fully dominate those obtained by NSGA-II, which suggests that more optimal and reliable solutions were identified by the proposed method.

* 1. Conclusion

In this study, a multi-step continuous-flow Claisen-Schmidt reaction was optimized based on a multi-objective DRL framework. More precisely, a MODDPG agent was used to maximize two objectives, namely the conversion and the selectivity. The proposed MODDPG is a novel and more efficient method to explore high-dimensional design space from mutually incompatible locations, compared to the traditional or standard multi-objective evolutionary algorithms. Compared to the single objective optimization, the high-quality Pareto solutions obtained by the proposed method show the advantages of maintaining maximum conversion, while satisfying the requirement of high selectivity in the synthesis of API. This research may prove very valuable and highlights the benefits of RL in real-time self-optimization of real-world problems which inevitably involve multiple objectives that must be optimized simultaneously.

**Acknowledgements:**

This research work is sponsored by Takeda pharmaceutical international company, USA.

References

1. Fath, V., Kockmann, N., Otto, J., & Röder, T. (2020). Self-optimising processes and real-time-optimisation of organic syntheses in a microreactor system using Nelder-Mead and design of experiments. Reaction Chemistry and Engineering, 5(7), 1281–1299.
2. Clayton, A. D., Schweidtmann, A. M., Clemens, G., Manson, J. A., Taylor, C. J., Niño, C. G., Chamberlain, T. W., Kapur, N., Blacker, A. J., Lapkin, A. A., & Bourne, R. A. (2020). Automated self-optimisation of multi-step reaction and separation processes using machine learning. Chemical Engineering Journal, 384.
3. Yewale, A., Yang, Y., Nazemifard, N., Papageorgiou, C., Rielly, C., & Benyahia, B. (2023). Multi-Agent Reinforcement Learning and RL-Based Adaptive PID Control of Crystallization Processes. Computer Aided Chemical Engineering, 52, 1667–1672.
4. Benyahia, B., Anandan, P., Rielly, C. (2021). Control of batch and continuous crystallization processes using reinforcement learning. Computer Aided Chemical Engineering, 67, 1371–1376. https://doi.org/10.1016/B978-0-323-88506-5.50211-4.
5. Benyahia, B., Latifi, M.A., Fonteix, C., Pla, F. (2010). Multicriteria dynamic optimization of an emulsion copolymerization reactor. Computer Aided Chemical Engineering 28, 457-462.
6. Zhou, Z., Li, X., & Zare, R. N. (2017). Optimizing Chemical Reactions with Deep Reinforcement Learning. ACS Central Science, 3(12), 1337–1344.
7. Neumann, M., & Palkovits, D. S. (2022). Reinforcement Learning Approaches for the Optimization of the Partial Oxidation Reaction of Methane. Industrial and Engineering Chemistry Research, 61(11), 3910–3916.
8. Liu, J., & Benyahia, B. (2023). Single and multi-objective superstructure optimization of an integrated continuous multistage reaction-crystallization-filtration process with recycles. Computer Aided Chemical Engineering, 52, 667–672.