# DEEP LEARNING OPTIMIZATION TO IMPROVE REFINERY OPERATIONS

Roberto Linares, Ph.D. and Tim Spinner, Ph.D.
oPRO.ai, Inc.
Houston, Texas
roberto.linares@opro.ai

## I. INTRODUCTION

The refining industry has the important task of manufacturing transportation and other fuels, and it currently faces multiple challenges such as demand disturbances, excess capacity, and environmental regulations. Refining is a business of margins which requires an efficient and optimal performance of the refinery. To enable this, we must drive to obtain profit-maximizing control of individual units and the synergistic coordination of the refinery as a whole. In a highly competitive market where companies can see decreasing margins, optimum operations can be a differentiated capability to satisfy demand at the lowest cost while still satisfying any business constraints.

Historically, refining companies have invested in control systems, advanced process control and real time optimization applications to optimize their process and operations. To enhance the efficiency already achieved with the current automation platforms, artificial intelligence (AI) offers opportunities to operate at the higher yields and margins by operating closer to product specification and diminishing product giveaway. In addition, AI can enhance the current automation platforms and applications by providing real time predictions and faster identification of operational issues.

This paper describes the implementation of advanced process control and real time optimization through deep learning optimization to achieve the multiple goals of the refining processes while safeguarding the safety and environmental integrity of operations.

## II. METHODOLOGY

The implementation of AI-based advanced process control applications through deep learning optimization (DLO) starts with the identification of the optimization goals. These are the variables that need to be optimized (e.g., maximized, minimized, or stabilized to a target), the identification of constraints that must be respected, and the selection of the prescribed ("handles") variables to achieve the given optimization. To accelerate this identification process, oPRO.ai has developed the methodology to capture this information efficiently in collaboration with the process control engineers and process engineers responsible for the operation of the refinery.

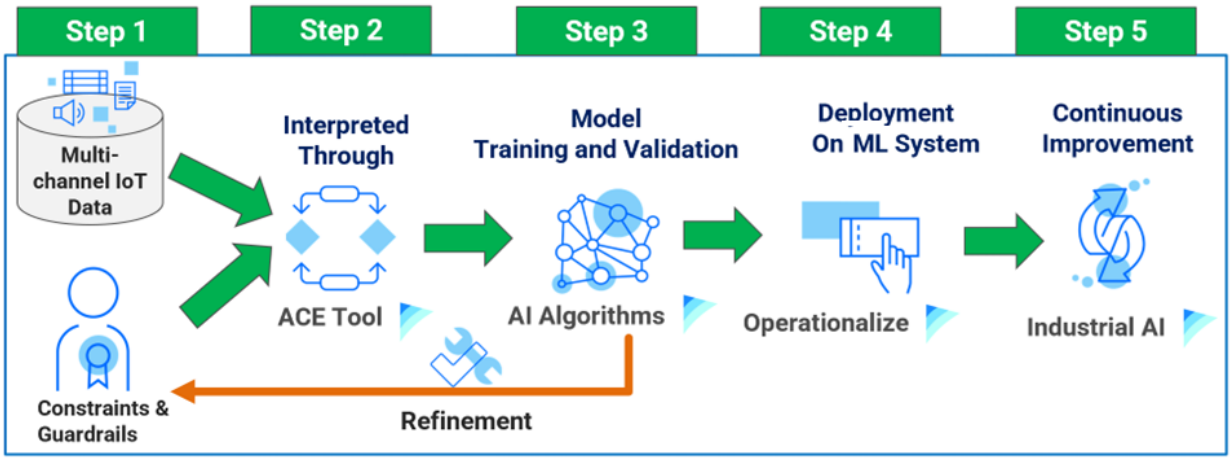


**Figure 1.** oPRO.ai methodology to implement reliable AI based applications.

Once the variables have been identified and a long-term process history has been provided, data issues will typically be discovered. Due to the vast size of the datasets, efficient algorithms are used to perform this validation and to identify issues early in the implementation. Multiple issues can exist, such as data out of range, stale tags, periods with no data available, and saturation of process instrument and actuator signals. To facilitate the reconciliation process, a detailed report is provided to the customer to discuss how these issues can be addressed prior to the development of the dynamic model.

The information stored in the distributed control systems (DCS) is commonly short term (e.g., 3 months) and is typically not enough to train a deep learning model. DLO frequently requires six months to a year of data, and can take advantage of more, at one-minute frequency from the refinery process. This is not a limitation since most refineries store real time data within the process data historian for each tag. The recommendation for customers in this instance that are planning an AI based control project is to standardize the level of compression and archiving frequency for all tags. The modeling process will take advantage of whatever data the customer has, but improving ongoing data quality will be useful for future model validation and retraining. We have also had some success on modeling with relatively small amounts of good quality data. In these cases, the identified model gain and dynamics may not generalize as well as a model trained on large periods of data, but could still be acceptable for variables that show more linear behavior or operate in a small range.

After addressing all data issues, a model is developed to learn the dynamic relationship of the inputs versus outputs of the process. The model is developed using an oPRO.ai proprietary AI Platform where the configuration of the AI pipeline is performed. This platform allows experimenting with different algorithms to capture the dynamics of the process accurately. The objective of the model is to capture both dynamic and long-term predictions of the process consistent with the length of the control horizon for the optimization process. Predictions are

assessed with different statistical metrics to ensure that results are sufficiently accuracy to support the optimization process.

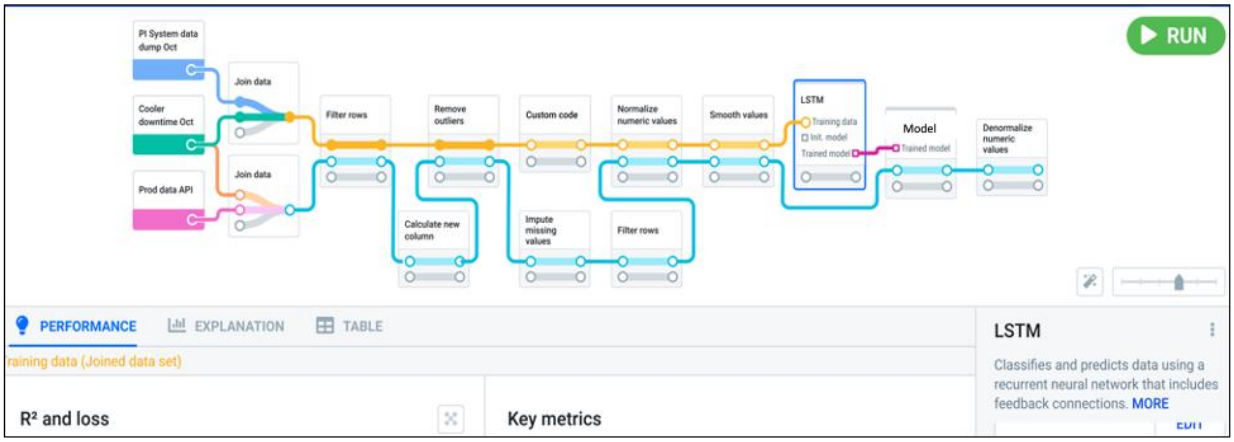


**Figure 2.** AI pipeline configuration within platform

The dynamic model is generated with a deep neural network that emulates a nonparametric model, which appears as a black box to the operator. Inherent in the definition of a black box, is that it is difficult to understand the behavior of the model and perform a validation against knowledge of the process that is either empirical, engineering based or known by conservation principles. To overcome this issue, oPRO.ai has developed a “window” into the deep learning optimization neural models using the concept known as explainable artificial intelligence.

The output of this software artifact is the generation of graphs that simulate the effect of each input (manipulated variables) to all its related outputs (controlled variables) for different conditions of the process operation. The step-response of an input-output pair generated by the oPRO.ai platform includes a family of curves, versus the single static curve generated by traditional linear APC modeling. From these graphs, the customer can validate the gains and the time expected to steady state under different process conditions.

These responses can be examined by experienced process and process control engineers. As an example, Figure 3 shows the reactor temperature response to an increase of cooling flow rate (1 engineering unit) for different operation conditions. We can observe the non-linearities in the response and how fast the reactor reaches the new steady state along with the expected final value of the reactor temperature. To generate these graphs, the system samples different conditions of the reactor found in the historical data. This information is then passed to the neural network which predicts the values for a fixed set of the manipulated variables to ensure that the neural networks predictions are not perturbed by current process dynamics. After the steady state is reached, a step change is introduced, and the responses are trended for analysis.

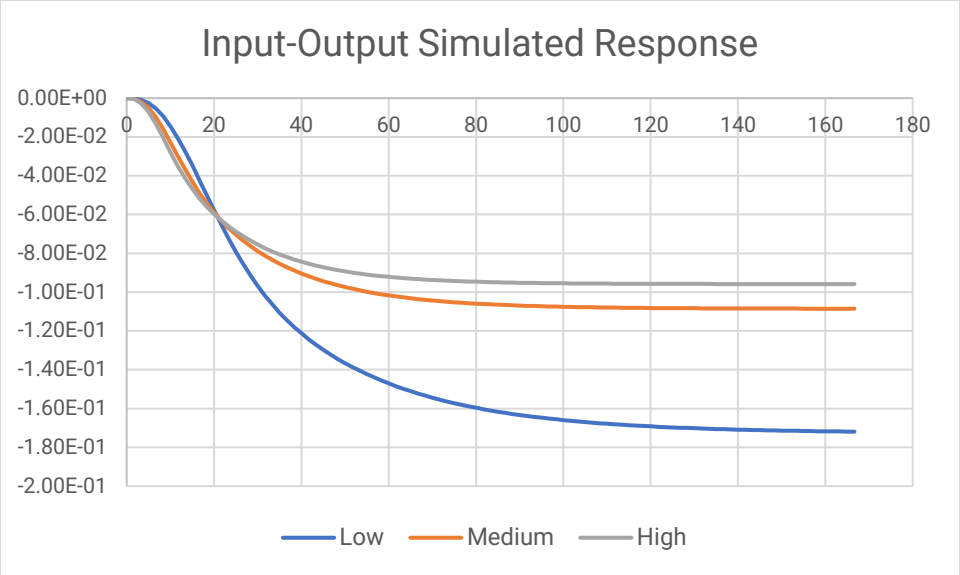


**Figure 3.** Input and output response from LSTM model.

In summary, in order to validate the robustness of the dynamic model, the following aspects have to be considered: 1) agreement of the prediction directionality which should guarantee the appropriate prescription of the manipulated variable to control certain process variables; 2) the distribution of the gains across the whole operation range – if a change of direction has to be justified by the physics of the process; and 3) the accuracy of its predictions. A model that has passed the threshold of this validation requirements can now be assembled into the optimization model. The oPRO.ai AI platform enables this process and the evaluation and automation automatically.

The model is later deployed and made available to operators and subject matter experts. To generate predictions and prescriptions, the model is connected to the DCS using OPC connectivity to consume the real time data of the input variables. oPRO.ai can deploy the model as a cloud or on prem solution; however, refining customers prefer the latter due to the strict data security requirements of this industry. Predictions and prescriptions can be viewed by using the provided dashboards and HMI of the application. An example of prediction versus real time value is shown in Figure 4.

For the generation of prescriptions, the optimizer finds the best candidate values of the inputs that optimizes the objective function which includes the user-defined objective as well as penalties for constraint violations. Other parameters, such as maximum step change and frequency of prescription, are configured after these values are validated with the subject matter experts and control engineers.

The model is thoroughly evaluated in closed control loop within the simulated environment to validate complex relationships and corner cases. Once the model is deployed, the operator and oPRO.ai execute open loop testing followed by closed loop control testing.

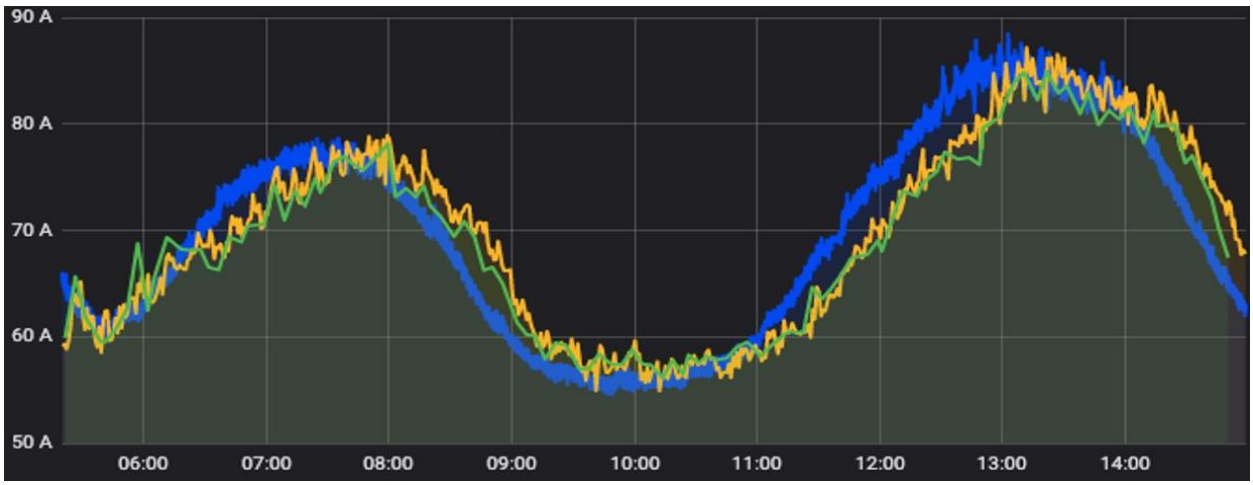


**Figure 4.** Figure showing actual value versus 5-minute ahead prediction for one CV.

## III. ALGORITHMS FOR APC AND RTO

Refinery processes and unit operations are designed to process large volumes of crude oil, intermediate products, and fuels. In general, large units exhibit prolonged delays which represent a process control challenge since the model dynamics need to capture the delay to adequately prescribe the manipulated variables values. For example, the composition in distillation column can react slowly to a change in a manipulated temperature of the column as the effect has to travel through all of the trays. To aggravate this issue, applications may involve process units in series, which increases even more the control horizon.

Depending on the requirements of dynamic model, oPRO.ai can use a variety of machine learning algorithms built into the platform. Specifically, for the processes with lengthy delays, oPRO.ai uses deep learning neural network using the LSTM (Long short-term memory) cell-type which is a type of recurrent neural network (RNN) architecture used in the generation. LSTM addresses the issues of the vanishing gradients found in plain RNN.

In addition, to the use of the LSTM models, the oPRO.ai platforms allow us to assemble multiple dynamic models that represent part of the process into a single overall model. This ensembled model increases the modularity and addresses the issue of the removal of non-casual relationships, since it is statistically possible to find correlation between any variable pair, but from the physics perspective that correlation is not considered relevant or even real. The following figure shows a diagram depicting an LSTM neural network cell.

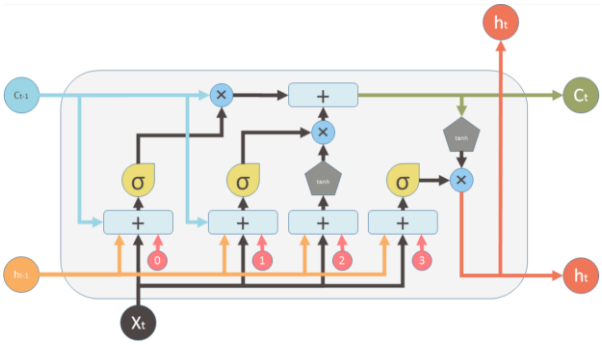

**Figure 5.** LSTM neural network cell.

Since the deep-learning model may exhibit very nonlinear or even discontinuous behavior reflecting the process, a stochastic search algorithm is used to find the optimum values of the manipulated variables. This method is based on the principle of natural selection, and it has become a tool of choice to solve non-linear optimization problems.

Industrial APC applications typically execute at a frequency of around 1-minute. Unfortunately, not all process information is available at high frequency. Such is the case with laboratory and analyzer values which can range in available frequencies from tens or minutes to days or longer. Also, this type of signal often does not contain sufficient datapoints to be used to directly train a dynamic deep-learning model. In these circumstances, an inferential or soft sensor is developed to estimate the values at the same frequency that is used for other signals.

oPRO.ai offers multiple methods to develop these inferentials. Among the linear methods available in the platform are multiple linear regression and lasso, which is a regularized linear regression. There is also a group of non-linear methods such as Random Forests (RF), Adaptive Boosting (AdaBoost), Gradient Boosting, and Multi-Layer Perceptron (a static neural network). These methods allow to develop regression models that can be used to generate a synthetic dataset to augment the real time dataset for the dynamic model development and online control.

## IV. EXAMPLE USE CASE

In chemical engineering, exothermic reactors play a crucial role in various industrial processes. These reactors release heat due to the chemical reactions, making them both fascinating and challenging to model.

We developed a simulation of a chemical reactor where we manipulate input variables to observe how the system responds. In this case, we are dealing with an exothermic reactor, which means that as reactions occur, heat is generated. Our goal? To simulate an entire year's worth of data, capturing the dynamic interplay of temperature, concentration, and other critical parameters.

To achieve this, we employ numerical methods and differential equations (A. Vasickaninová and M. Bakosova). By discretizing time, we create a computational grid that represents the reactor. The governing equations—mass balances, energy balances, and reaction kinetics—are solved iteratively. As the simulation progresses, we track temperature profiles, reactant and product concentrations, and heat release rates.

We collect our simulated data randomly, varying input conditions, and feed it into a deep neural network. The network learns from this rich dataset, uncovering hidden patterns and correlations. It becomes adept at predicting reactor behavior, even in scenarios it has not explicitly encountered. It grasps the intricate relationship between inputs and outputs, capturing non-linear relationships that classical linear models struggle to represent.

Figure 6 shows the steady-state gain for one MV-CV pair (feed to byproduct concentration). It can be observed the strong multivariate dependence of the gain and the magnitude of the disturbance variables. The increase of gain of more than 10-fold indicates that a linear model will not be adequate to perform predictions accurately for the whole operation range and that a deep learning model is required for the process identification of this reactor.

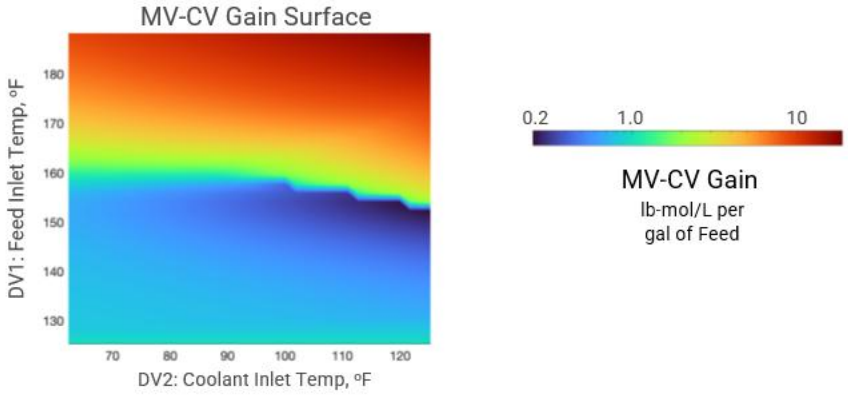


**Figure 6.** Gain analysis from process differential equations.

Once our neural network has undergone rigorous training, it is time to put it to the test. We feed it unseen data—our test set—and observe how well it predicts outcomes. This evaluation step is critical for assessing the network's generalization ability and ensuring it does not merely memorize the training data. We partition our dataset into two subsets: the training set (used for training) and the test set (reserved for evaluation).

We employ various metrics to quantify how well the neural network performs. One common metric is the Mean Squared Error (MSE). Suppose our neural network achieves low MSE on the test set. It means our model generalizes well beyond the training data.

Visualizing predicted versus actual values using plots helps us pinpoint areas of improvement. Figure 7 shows predicted versus measured values for three CVs of the reactor system.

Prepared with insights from evaluation, we fine-tune our neural network. Adjusting hyperparameters, adding regularization, or modifying the architecture can enhance performance. In summary, our neural network is not just a black box—it is a tool for understanding complex systems. By evaluating its predictions and quantifying accuracy, we gain confidence in its ability to manage the reactor scenarios.

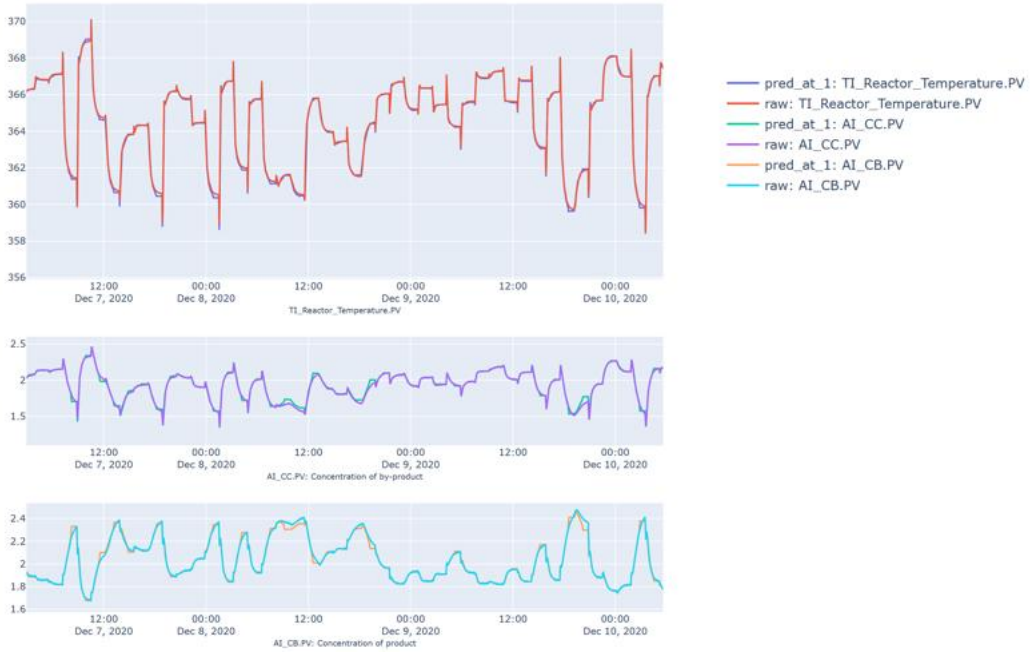


**Figure 7.** Prediction versus measurement of CVs.

Using the same neural network, we compute the input and output response as described in the previous section. The following plot (Figure 8) shows the reactor temperature dynamics caused by an increase of 1 unit change of the feed temperature.

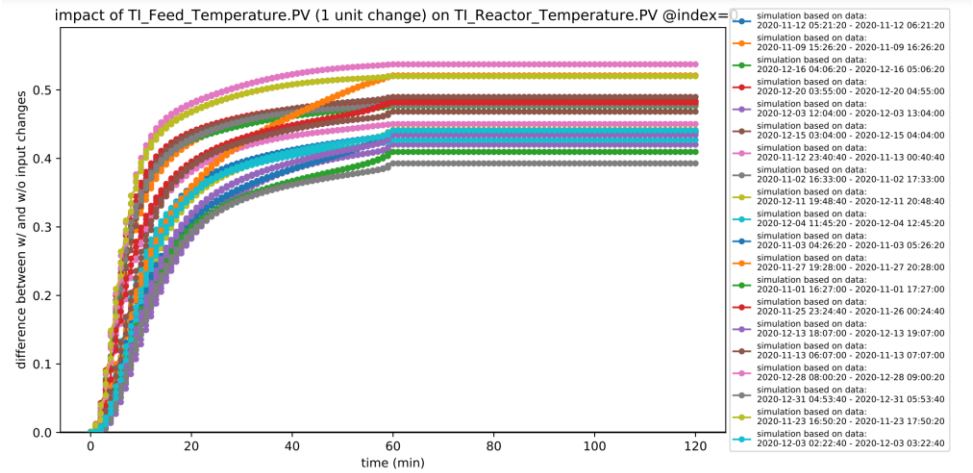


**Figure 8.** Step response simulated using a trained neural network.

Our trained neural network has become our tool that predicts reactor behavior. Armed with this foresight, we are ready to optimize. But first, let us define our objective. What are we optimizing? Efficiency? Safety? Profit? Perhaps all three. But for the purpose of our example, we will only consider the maximization of yield, which is the product of feed rate and concentration of product of interest (product B). For given values of DVs, we can plot the unconstrained values of the optimization function.

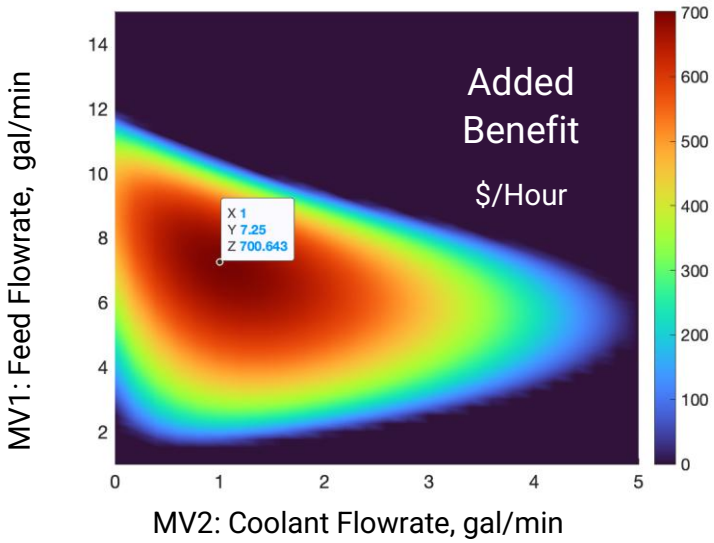


**Figure 9.** This plot shows the value of the optimization function.

In order to optimize the objective function, we use a genetic algorithm that allows us to find the optimal values of the MV moves so that we can find the global peak of the objective function especially in cases where constraints do not determine the optimal value.

The next figure shows the progression of MVs for a given update of the reactor temperature high limit.


**Figure 10.** MV values for the reactor optimization application.

## V. INDUSTRIAL DLO APPLICATION FOR NGL FRACTIONATION

An APC application was implemented on an NGL fractionation process using the oPRO.ai platform. The fractionation system comprises four units: deethanizer, depropanizer, debutanizer, and deisobutanizer columns (Figure 11). The primary objectives to enhance the unit's performance are: (1) maximizing the feed rate to its capacity limits to increase refinery revenue; (2) optimizing and controlling product impurity to reduce product giveaway; and (3) reducing the unit's energy intensity by operating at lower temperatures. The application meets these goals (based on their economic priorities) while adhering to constraints related to safety, quality, productivity, and energy as defined by operator limits.

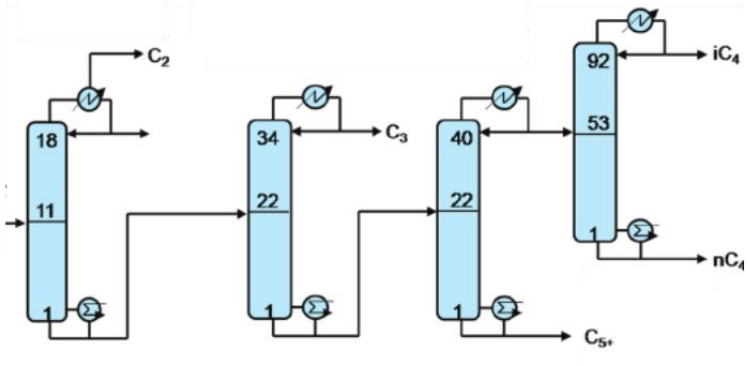


**Figure 11.** Process flow diagram of NGL fractionation system.

The APC problem, considered a typical-scale application for this industry, involves 10 MVs, 25 CVs, and 10 DVs, and is executed every 1 minute. The table below illustrates the incremental value achieved by a 4% increase in throughput. While additional benefits related to product

giveaway and energy savings are not quantified here, they are anticipated to enhance the overall value of the application.

| Variable | | Traditional APC | oPRO.ai |
|---|---|---|---|
| Capacity | Barrels per day | 100,000 | 104,000 |
| Margin | $ per Mscf | 1.38 | 1.38* |
| Margin increase | $ per year | -- | $4.8 Million |

**Table 1.** Marginal increase due to APC-DLO application.

Two of these fractionation optimization and dynamic control applications are now running on parallel units in the same refinery. One application has been running in closed-loop for over two years. The second was able to go from project-start to closed-loop commissioning in just over 3 months. This speed of execution serves as a demonstration of both oPRO.ai technology and project methodology.

## VI. INDUSTRIAL RTO APPLICATION FOR NAPHTHA DESULFURIZATION COMPLEX

A real-time optimization (RTO) system was successfully deployed on an industrial naphtha desulfurization complex using the oPRO.ai platform. The process (Figure 12) consists of three parallel desulfurization units treating light, middle, and heavy naphtha fractions, respectively, and an upstream fractionation column that splits the full-range cat naphtha sourced from two FCC units into these three streams by adjusting two cut-point temperatures. By manipulating these cut points, the optimizer indirectly controls the feed rate to each hydrotreater.

The primary objective of the RTO is to maximize the overall economic margin of the unit by simultaneously achieving two goals: (i) optimally distributing the total feed among the three hydrotreaters to exploit differences in octane (RON) loss, and (ii) meet other key product specifications—Reid vapor pressure (RVP) and sulfur content. These objectives, along with all operational and quality constraints, are encapsulated in a single, rigorously derived economic objective function.

Despite having only two manipulated variables (MVs), the optimization problem is highly nonlinear and involves twelve state and output variables used for operating constraints or in the economic objective function. A significant challenge arises from the infrequent laboratory analyses of key product properties, ranging from 2x daily to weekly. To enable closed-loop optimization on a minute-by-minute basis, accurate inferential models (soft sensors) based on deep neural networks and other statistical methods were developed and integrated into the RTO layer, providing real-time estimates of RVP, sulfur content, and octane number with high fidelity.

To aid in monitoring and control of the economic optimization, oPRO.ai introduced a new "Economics" HMI tab that allows detailed insights and regulation of optimizer behavior. For the current gasoline-pool application, this included breaking down economics by stream (LCN, ICN, HCN), and by product quality (RVP, sulfur, octane). Tabular and trend views of each economic indicator compare current, open-loop steady-state, and optimal steady-state operating points. As with other HMI tabs, the entries in the display can be configured for view/edit by the appropriate user roles (Operator, Engineer). Users with appropriate permissions can manipulate economic values (such as $/octane-barrel) or turn on/off elements of the objective function with granular controls. Values are also configured to stream from PI or DCS sources for automated switching between economic or operating scenarios. Overall, the new interface provides unprecedented clarity into the actions and economic value of the closed-loop optimization stack.

Although this multi-unit gasoline treating application performs only a steady-state optimization and supplies external targets to existing lower-level APC controllers, it is fully implemented using oPRO.ai's dynamic deep-learning models. The advantage here is that the process itself does not need to be at steady-state for optimization to occur; even under dynamic changes in inputs and disturbances, the models can predict where the process is headed and optimize with this knowledge. Thus, unlike traditional RTO technologies, optimization can occur every cycle at the configured execution frequency (5 minutes here); the DLO application never needs to wait for a process steady-state in order to provide economically optimal targets to the lower-level APC controls.

The implemented RTO system can deliver estimated economic benefits exceeding $4 million per year compared to the pre-optimization baseline. This application demonstrates the substantial value of multi-unit coordinated optimization even in systems with existing APC controllers. The key advances by oPRO.ai that enable these benefits include real-time nonlinear optimization powered by high-fidelity deep-learning models and inferentials paired with autonomous detection and seamless transition between operating modes to provide continuous economically-optimal operating targets.

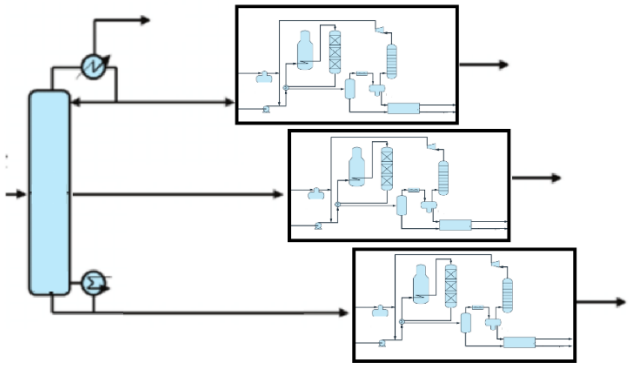

**Figure 12.** Process flow diagram of naphtha desulfurization complex.

## VII. CONCLUSIONS

Advanced Process Control powered by Deep-Learning Optimization (APC-DLO) is now a reality for the refining industry. Due to the capabilities of the APC-DLO, the adopters of this technology can improve the current efficiency of their process and operations. This efficiency can be translated into higher throughput, energy savings and reduction of product giveaway.

The adoption of AI can be challenging, especially within the context of an advanced process control application. But the challenges can be faced choosing a robust AI software platform and following a proven methodology for a successful implementation of a sustainable application.

This paper presented the methodology and the key features of the oPRO.ai platform to implement APC-DLO applications within the refinery processes and operations. The applications already implemented continue to perform optimally according to the design, using the platform to retrain models as needed by any process updates or changes to the data patterns needed to continue with an optimal control solution.

## REFERENCES

P. Acharyya, S. D' Rosario, R. Flor, R. Joshi, D. Li, R. Linares, H. Zhang - Autopilot of Cement Plants for Reduction of Fuel Consumption and Emissions, ICML 2019.

R. Linares, P. Acharyya – Implementation of AI to Optimize Clinker Coolers in International Cement Review, December 2020, 33-36.

A. Vasickaninová, M. Bakosova Neural Network Predictive Control of a Chemical Reactor, presented at17th International Conference on Process Control 2009. June 9–12, 2009, Strbske Pleso, Slovakia.