FOPAM

Foundations of Process/product Analytics and Machine learning, UC Davis, July 30-Aug 3, 2023

Poster Abstracts

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

THE UNIFIED ACTIVE LEARNING FRAMEWORK FOR MOLECULAR MATERIAL DESIGN

Jiali Li[1], Shomik Verma[3], Kevin Greenman[3], Yizhe Chen[2], Haoyu Yin[2], Zhihao Wang[2], Rafael Gomez- Bombarelli[3], Aron Walsh[4], Xiaonan Wang[2,*]

1. National University of Singapore, Singapore, Singapore.
2. Tsinghua University, Beijing, China.
3. Massachusetts Institute of Technology, Cambridge, MA, United States.
4. Imperial College London, London, United Kingdom.

Keywords: Active Learning, Material Design, Data Efficient

Solar photovoltaic (PV) technology has received immense, global interest in recent years. Many novel devices have been created beyond conventional silicon solar cells, such as thin film PV, organic PV, and perovskites. Unfortunately, all such single-junction solar cells suffer from an efficiency cap known as the detailed-balance limit, which limits solar cell efficiency to 33.7%. This limit is primarily due to solar cells unable to absorb light below their bandgap, and inefficiently absorbing light above their bandgap.

One strategy to improve PV efficiency is to use certain organic molecules that up- or down-convert photon energies using interplays between their excited states. Two common types of photon conversion processes are triplet-triplet annihilation (TTA) up-conversion and singlet fission (SF) down-conversion. However, designing efficient TTA and SF molecules comes with several challenges. Namely, the design space of organic photon conversion molecules is massive; for example, 166 billion organic molecules exist with less than 17 atoms. Further, experiments or first-principles simulations to evaluate the excited state energies of these molecules are time- and resource-consuming.

Fortunately, large scale virtual screening and inverse design with machine learning are promising solutions, as they allow accelerated evaluation of properties and efficient exploration of chemical space. Supervised machine learning comes with its own challenges, however. As a primarily data-driven method, it is limited by slow acquisition of labeled data for model training. Further, the chemical design space is diverse, so a large amount of labeled data is required for training to ensure sufficient coverage. Finally, tasks may be very different or require non-overlapping data. For these reasons, a model with high accuracy for desired properties is difficult to obtain.

Therefore, it is useful to develop an efficient, unified strategy for generating training sets and suggesting candidate molecules under different conditions. Active learning is one promising strategy for achieving this, as it efficiently explores chemical space. This study presents an active learning framework for designing energy- relevant molecules, where molecules suitable for TTA up-conversion and SF down-conversion are taken as a case study.

First, an ultra-fast chemical simulation method based on machine-learned calibrations to tight binding is developed for accelerating the labeling process. The calibration training set is carefully curated to ensure both breadth of chemical space and depth in space of molecules of interest, namely large aromatic molecules with pi-conjugated bonds.

Next, we use this accelerated labeling to generate a large molecular database and benchmark various active learning strategies with different priorities over this database. Namely, we come out with a unified active learning framework as shown in (Figure 1) and implement different settings of key components for comparsion.

Finally, a highly accurate generative model based material design workflow is developed based on the informative database derived from the unified active learning framework. The workflow considers both suitability and synthetic accessibility to reduce experimental effort. A pool of molecules suitable for TTA up-conversion and SF down-conversion are proposed, verified with higher-fidelity computational chemistry methods, and finally demonstrated with experiments.

Overall, a unified active learning framework is developed, and molecules suitable for TTA and SF are proposed. The suitability of these materials as photon conversion materials to improve PV efficiency is a promising demonstration of the utility of this approach for designing energy-relevant molecules.

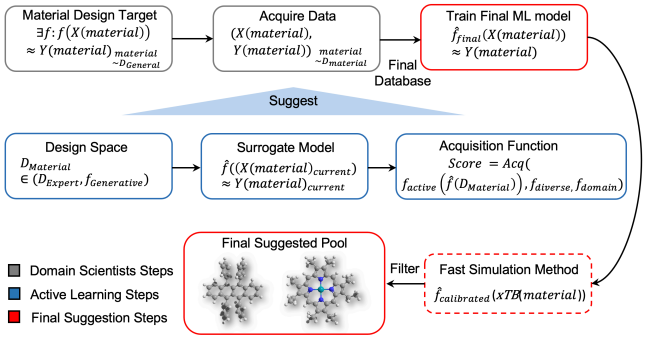

Figure 1 A unified active learning framework for molecular materials

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

DESCRIPTIVE PROCESS ATTRIBUTE SEQUENCE TO SEQUENCE ROLLING FOR ZERO-SHOT FAULT DIAGNOSIS

Wen-An Lee[a], Yuan Yao[a], and Jia-Lin Kang[b,*]
[a]Department of Chemical Engineering,
National Tsing Hua University, Hsinchu 300044, Taiwan
[b]Department of Chemical and Material Engineering,
National Yunlin University of Science and Technology, Yunlin 64002, Taiwan

Keywords: Fault diagnosis, Seq2Seq, Generators, Supervised learning, Process monitoring

Introduction

Artificial intelligence has emerged as an invaluable tool for fault diagnosis in diverse industries, facilitating the efficient and precise detection of faults and malfunctions within intricate systems. While conventional supervised classifiers primarily classify known faults, the reality often entails a combination of known and unknown faults. Consequently, our objective is to develop the capability to identify unknown faults without compromising the classification accuracy of known faults. To this end, we present a novel approach called the Descriptive Process Attribute Sequence to Sequence Neural Network (DPAS).

Methodology

Based on Kang's 2020 research, RNNs excel in classifying the Tennessee Eastman Process (TEP). Thus, our DPAS model incorporates a two-layer RNN (Figure 1). We enhanced the process description attributes from Feng and Zhao's study (2020), categorizing them into three groups: dynamic behavior (e.g., step changes, random variations), fault location (e.g., reactor, condenser), and fault cause (e.g., flow rate, temperature changes). Each TEP fault was defined with a 16-length attribute array.

For predicting the manipulate variable (MV), we used a three-layer fully connected neural network. It takes historical process data (MV and attributes) as input and outputs predicted MV. The encoder takes the historical process data as input, generates a hidden state, and passes it to the decoder. The decoder uses the predicted MV and attribute input to produce predicted process data without the MV. Combining the predicted MV with this output yields the predicted process data with the MV.

By employing a rolling approach, we can generate process data corresponding to the fault attributes without relying on any external data. This approach involves merging the predicted results with the past data input at each step. By iteratively updating and incorporating the predicted results into the input data, we can generate a sequence of process data that aligns with the fault attributes. This allows us to simulate the behavior of the system and generate valuable insights without the need for additional data sources.

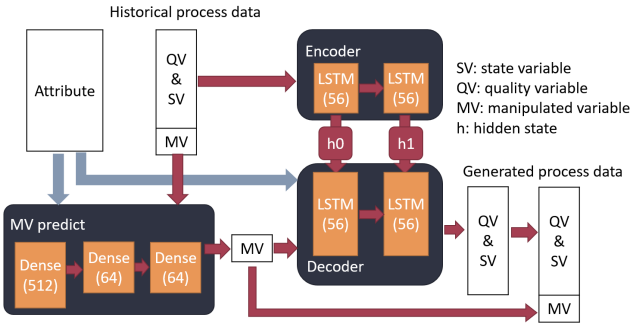

Figure 1. DPAS model structure

Results

In our tests, we selected one unknown fault from IDV1 to IDV14 in the TEP. Known faults were a mix of generated and original data (1:1 ratio), while the unknown fault consisted entirely of generated data. The random forest classifier accuracy improved from 72.06% to 75.05%, and the RNN classifier increased from 80.49% to 81.92%. Training the RNN classifier solely with generated data achieved an average accuracy of 71.99%. These findings demonstrate the effectiveness of our approach in fault classification, even with unknown faults, and the value of generated data.

References

J.-L. Kang (2020). Visualization analysis for fault diagnosis in chemical processes using recurrent neural networks. Journal of the Taiwan Institute of Chemical Engineers. 112, 137-151.

L. Feng and C. Zhao (2021). Fault description based attribute transfer for zero-sample industrial fault diagnosis. IEEE Transactions on Industrial Informatics. 17(3), 1852-1862.

* All correspondence should be addressed to jlkang@yuntech.edu.tw.

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

DISCRIMINATIVE EDGE-GROUP SPARSE PRINCIPAL COMPONENT ANALYSIS FOR PROCESS FAULT DIAGNOSIS

PO-CHUN MAO and YUAN YAO *
Department of Chemical engineering, National Tsing Hua University,
Hsinchu 30013, Taiwan, ROC

Keywords: Discriminative edge-group sparse principal component analysis, Fault diagnosis, Process monitoring

Introduction

In modern industrial processes, data is often high-dimensional and noisy, with complex and nonlinear interactions between variables, making fault diagnosis a significant challenge. Principal component analysis (PCA) has been a popular process monitoring tool for decades, known for its ability to reduce dimensionality and extract features. As a result, it is a preferred method for screening process status. However, its limitations include a lack of sparsity and the inability to account for the topological mechanisms of the process. This work introduces a discriminative edge-group PCA (DESPCA) method to address these issues and improve fault diagnosis.

Methodology

Fig/ 1 depicts the flowchart of DESPCA, which takes process data and process topology represented as a graph as input information. The process data are segmented into several groups based on changes in process status, such as two groups for fault occurrence or three groups for the initial steady state, transition period, and final steady state. Next, a correlation weight factor is calculated for each variable, reflecting its discriminatory power across different groups. Subsequently, scores and loadings are calculated using an iterative algorithm that incorporates sparsity into the loading vectors. This is achieved by utilizing the graph and weights on the edges connecting each pair of nodes, i.e., variables. The use of sparsity and topology information in the DESPCA method ensures that each principal component corresponds to one or multiple physically connected process units. This often reflects the propagation path of the fault. Furthermore, the weights contain information on both data variability and discriminability, which can significantly influence the placement of variables affected by the fault into the first principal component (PC), thus facilitating fault diagnosis.

Results

This study demonstrates the effectiveness of the proposed model using the IDV1 fault (A/C feed ratio, B composition constant) of the Tennessee Eastman process as an example. Based on the knowledge of process experts, V23, which represents the composition of component A in stream 6, is the variable closest to the root cause of the fault. Along with V1, V44, and V6, it contributes to the first PC of DESPCA. Causality analysis among these four variables can therefore diagnose the root cause of the fault easily.

References

Miao, R., Dong, X., et al. (2022). Dynamic Meta-Data Network Sparse PCA for Cancer Subtype Biomarker Screening. *Frontiers in Genetics*, **13**: 869906.

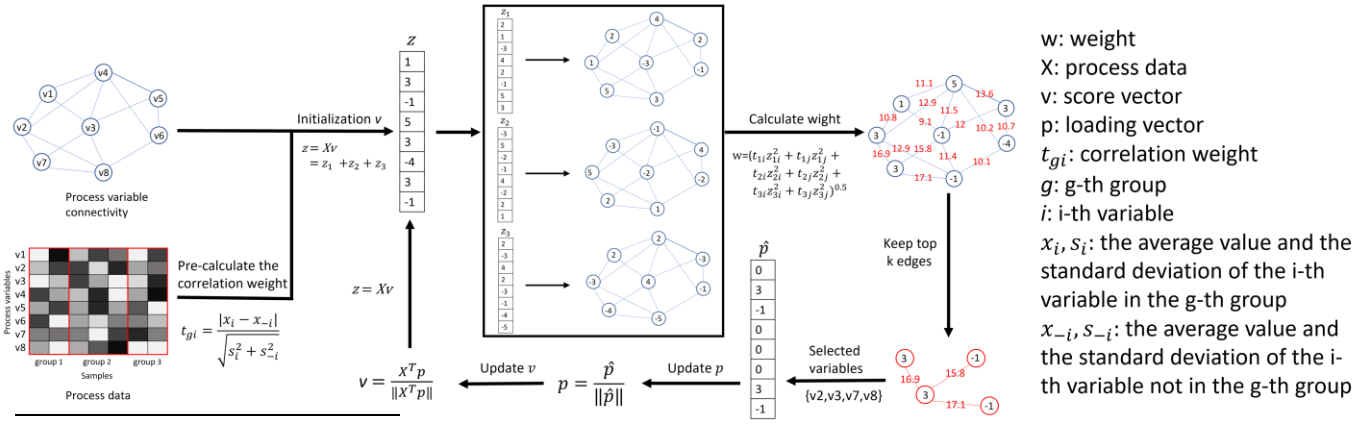

Figure 1. Algorithm of DESPCA model

* To whom all correspondence should be addressed; yyao@mx.nthu.edu.tw.

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

OVERCOMING MODELING CHALLENGES IN THERMOSET SHAPE MEMORY POLYMERS: A MACHINE LEARNING APPROACH

Ama Darkwah, Cheng Yan*, Patrick Mensah
Department of Mechanical Engineering, Southern University and A&M College

Baton Rouge, LA, 70807

Keywords: *Shape Memory Polymer, Thermomechanical property, Machine Learning, glassy modulus*

Predicting the thermomechanical properties of thermoset shape memory polymers (TSMPs) is crucial for their effective design and application. However, the complex topological nature of TSMPs presents a significant challenge in modeling these properties, often resulting in time-consuming processes. To address this challenge and expedite the prediction of the glassy modulus of shape memory polymers (SMPs), we propose the application of machine learning (ML) techniques, aiming to overcome the limitations inherent in traditional modeling methods.

Methodology

In this study, we utilized a small dataset of about 214 datapoints obtained from various references to train our ML model. Ninety percent of the data was used for training and the remainder for validation. Our approach involved employing three distinct ML models: Gaussian Process Regression (GPR), Support Vector Machine (SVM), and Artificial Neural Network (ANN). Each model was trained using the collected dataset, enabling us to investigate their performance in predicting the glassy modulus of SMPs. We carried out an exhaustive analysis comparing the effectiveness of three distinct models.

Results and Discussions

The results clearly indicate that the ANN model stands out, demonstrating superior predictive accuracy compared to the other two. In particular, the ANN model recorded a mean absolute percent error (MAPE) of just 13%, which is notably lower than the 18% and 17% achieved by the GPR and SVM models, respectively.

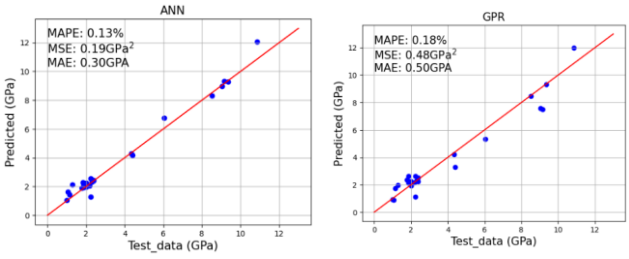

Figure 1 Visualization comparisons for glass modulus based on two different ML methods.

Moreover, the ANN model excelled in reducing errors, as evidenced by its mean squared error (MSE) of 0.19 GPa² and mean absolute error (MAE) of 0.30 GPa.

Table 1 Comparison of errors for the models

MODELS	MAPE (%)	MSE (GPA^2)	MAE (GPA)
ANN	**13**	**0.19**	**0.30**
SVR	17	1.96	0.66
GPR	18	0.48	0.50

Furthermore, we carefully dissected the underlying factors contributing to the observed disparities. The distinctions among various models can be attributed to three primary factors. First and foremost, the efficacy of these models can be influenced by the quantity of parameters they encompass. ANN typically contains a significantly higher number of parameters compared to SVM, which enables ANN to characterize a greater diversity of features. Secondly, ANNs employ piecewise functions that confer a higher degree of adaptability in making predictions, as opposed to the rigidity of the exponential functions utilized by SVMs. The third factor concerns the inherent characteristics of the data points, which can result in the diminished effectiveness of GPR. Specifically, GPR is susceptible to the 'curse of dimensionality', a phenomenon whereby its performance deteriorates with the escalation in the number of input dimensions.

Conclusion

In conclusion, our study provides a feasible approach for predicting the glassy modulus of SMPs. By utilizing these three algorithms, we have significantly accelerated the prediction process, enabling efficient material design and optimization. Moreover, our analysis for the three ML techniques explores the adaptability of material science, which will deepen the understanding for the investigators in new material design.

Acknowledgement

This work is supported by the U.S. National Science Foundation under grant number OIA-1946231 and the Louisiana Board of Regents for the Louisiana Materials Design Alliance (LAMDA).

* Corresponding author; email address: cheng.yan@sus.edu

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

Application of Experimental Design via Bayesian Optimization to Pharmaceutical Process Characterization

J. E. Tabora*, Jun Li, J. Stevens, M. Williams, J. Fu, D. Reyes and D. Skliar

Bristol Myers Squibb
Chemical Process Development
New Brunswick, NJ 08903

Keywords: Gaussian Process, Pharmaceutical Development, Garrido Projection

Previous studies have demonstrated the applicability of Experimental Design via Bayesian Optimization (EDBO) to chemical synthesis and chemical processes. In addition, multiple research groups have published software that facilitates the programmatic incorporation of EDBO into a research organization (Torres, 2022; Wang, 2021).

In pharmaceutical development, the late-stage process characterization workflow, is enabled by the application of Design of Experiments (DOE) which allow the identification of a robust control strategy that takes into account the multivariate nature of the underlying unit operation. Typically, an experimental design of a targeted optimality criteria is performed from which models are built to estimate the level of the responses (critical quality attributes, cQA's). The quantification of the responses allows the construction of a design space which provides assurance of control.

In this work we demonstrate that applying EDBO to the characterization workflow results in an efficient and streamlined workflow for control strategy design. We apply the procedure in two instances of reaction characterizations with multiple factors and responses. We propose a methodology to seed the algorithm with an appropriate subset of the DOE set to improve the efficiency of process characterization. Finally, we suggest a visualization strategy (the Garrido Projection) to enable interpretability of the surrogate process model to facilitate the design of a multivariate control strategy.

Process Characterization

The characterization of pharmaceutical processes involves the quantification of the impact of process parameters on the quality attributes of the unit operations that are critical to the final quality of the pharmaceutical drug product (DP). Typically, this task is accomplished by developing a surrogate mathematical model that provides adequate estimates of the desired observables (the quality attributes) given the set points of the controlled process parameters that are part of the process description. Figure 1 shows the resulting implementation of a Gaussian Process and the corresponding visualization using a Garrido Projection of ~5E4 process conditions from which a reasonable control strategy can be devised.

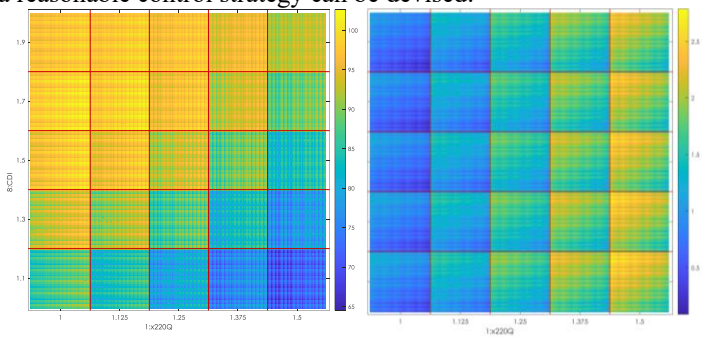

Figure 1. Garrido Projections of the estimated values of a key impurity (left) and the yield (right) estimated by a Gaussian Process trained on experimental data from a chemical transformation.

Systematic visualization of the two-dimensional projection of the high-dimensional model and the corresponding quality attributes (two in this example) allows scientists to efficiently determine a suitable set of conditions under which the appropriate quality conditions are met.

Finally, we show the efficiency of the Gaussian Process to estimate process outcomes with limited data, allowing for a graded completion of the experimental design resulting in significant reduction of experimental effort.

References

Torres JA, Lau SH, Anchuri P, Stevens JM, Tabora JE, Li J, Borovika A, Adams RP, Doyle AG. A Multi-Objective Active Learning Platform and Web App for Reaction Optimization. Journal of the American Chemical Society. 2022 Oct 19;144(43):19999-20007.

Wang Y, Chen TY, Vlachos DG. NEXTorch: a design and Bayesian optimization toolkit for chemical sciences and engineering. Journal of Chemical Information and Modeling. 2021 Oct 25;61(11):5312-9.

Foundations of Process/Product Analytics and Machine learning

FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

PREDICTING THERMAL CONDUCTIVITY OF ADDITIVELY MANUFACTURED ALLOYS USING MACHINE LEARNING BASED MODELS

Evelyn Dede Quansah, Congyuan Zeng, Patrick Mensah *,
Department of Mechanical Engineering, Southern University and A&M College,
Baton Rouge, LA 70807

Keywords: *Machine Learning, Thermal Conductivity, Laser Powder Bed Fusion, Additive Manufacturing*

Thermal conductivity is a crucial property for the performance of additively manufactured alloys in service applications. However, predicting the thermal conductivity of novel alloys is a challenging task due to the impact of various factors, such as processing temperature, microstructure, thermal cycling history, and alloy compositions. Computationally aided simulation together with experimental approaches for predicting thermal conductivity are expensive and time-consuming, particularly for parts produced using additive manufacturing (AM). To address this issue, a machine learning (ML) model was developed for predicting the thermal conductivity of AM alloys. Herein, a large dataset of experimentally determined thermal conductivity (TC) values was compiled for various temperature ranges (25-1000 °C) and several laser powder bed fusion (L-PBF) AM alloy families, including Nickel, Copper, Iron, and Cobalt-based alloys. These datasets were used to train and validate classical regression ML models (linear regression, stochastic gradient descent, support vector machine) and an Artificial Neural Network (ANN) ML model. Among the models, the Artificial Neural Network showed exceptional performance, with a mean absolute percentage error (MAPE) of 0.07.

Table 1. ML Error Matrix Comparison

ML models	Root mean square error (W/mK)	Mean absolute Error (W/mK)	Mean absolute percentage error %
Support Vector Regressor	47.62	14.42	13
Stochastic gradient descent	11.97	5.51	17
Linear Regression	11.15	4.5	11
Artificial Neural Network	5.2	2.23	7

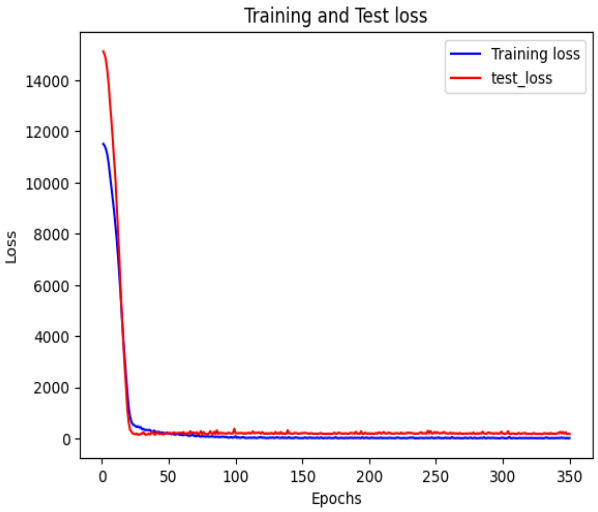

Figure 1. Representation of Artificial Neural Network Model generalization.

The ANN model was then tested on predicting the TC of Inconel 718 alloys manufactured using the L-PBF AM process. This approach has not been documented in the literature to the author's knowledge. The results showed excellent agreement between the ML predictions and the experimental outcome, indicating a strong correlation. Our study demonstrates that machine learning models can efficiently predict the thermal conductivity of new L-PBF AM alloys, enabling rapid identification of appropriate alloys for specific industrial applications and predicting TC for new alloy development. Additional TC measurement data can be added to continue to expand the ML method and further validate its accuracy.

*Corresponding Author: patrick_mensah@subr.edu

Foundations of Process/Product Analytics and Machine learning FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

Automated Outlier Detection and Estimation of Missing Data

J. Rhyu[1], D. Bozinovski[2], A. B. Dubs[1], N. Mohan[2], E. M. Cummings Bende[2], A. J. Maloney[1], M. Nieves[2], J. Sangerman[2], A. E. Lu[1], M. S. Hong[1], A. Artamonova[2], R. W. Ou[3], P. W. Barone[2], J. C. Leung[2], J. M. Wolfrum[2], A. J. Sinskey[2,3], S. L. Springs[2], R. D. Braatz[1,2*]

[1]Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139
[2]Center for Biomedical Innovation, Massachusetts Institute of Technology, Cambridge, MA 02139
[3]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139

Keywords: Multivariate statistics, Principal component analysis, Outlier detection, Biomanufacturing

Industrial process datasets commonly have missing values. The distribution of missing data points can be characterized as being within four classes: (1) random missing data, which exhibit no explicit pattern, (2) sensor drop-out, in which the missing values are correlated in time, (3) multi-rate, in which the missing data occur periodically, and (4) censoring, in which there exist thresholds for censoring so that the measurements outside the range are not recorded (Imtiaz and Shah, 2008; Severson et al., 2017).

The presence of missing values inhibits the use of the data in process modeling, analysis, and control. Even basic data analytics methods such as principal component analysis (PCA) and partial least squares (PLS) require a full data matrix, that is, without any missing values. The simplest way to deal with missing values is to only consider the observations with full measurements. Removing every observation with missing values, however, can cause significant data loss in the specific time period when the missing values are agglomerated, which makes capturing the process dynamics challenging. In order to fully utilize the given dataset, many general-purpose methods such as mean imputation, alternating algorithm, PCA data augmentation (Imtiaz and Shah, 2008), Bayesian PCA (Oba et al., 2003), singular value thresholding (Cai et al., 2010), and augmented Lagrange multiplier (Lin et al., 2010) have been developed to filling in missing values in a structured way, which are briefly reviewed in this poster.

The aforementioned matrix recovery algorithms for filling in missing datapoints, however, are vulnerable to outliers. Not removing the outliers before filling in missing values emphasizes the effect of outliers and degrades the accuracy and reliability of results obtained by subsequent data analytics. These outliers can be detected by using the contribution map for T^2 and Q statistics.

To the best of our knowledge, there has been no software that simultaneously detects outliers and fills in missing values in an automated way. In this poster, we introduce an open-source software that automatically detects outliers, fills in missing values, and evaluates each algorithm used for matrix recovery. This poster describes the framework of outlier detection and missing value estimation used in the software, and the demonstration of the software to data collected from a continuous biomanufacturing pilot facility at the Massachusetts Institute of Technology. The data are from the production of a monoclonal antibody produced by Chinese Hamster Ovary cells in a perfusion bioreactor.

To provide a thorough validation of the methods and software, the software is applied to a variety of datasets are varying types and extent of missing data constructed from an initial dataset in which all of the measurements are initially available. The performance of the various methods for filling in missing values are compared using several metrics including the normalized root-mean-squared error, the number of imputed values outside the boundaries, the number of imputed values considered outliers, and the computational time. Fifty simulations on each missing data scenario were conducted to obtain the distribution of these metrics. The matrix completion methods were the most effective except for the censoring case where probabilistic PCA methods were the most effective.

References

Cai, J. F., Candès, E. J., Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982.

Imtiaz, S. A., Shah, S. L. (2008). Treatment of missing values in process data analysis. *Canadian Journal of Chemical Engineering*, 86(5):838–858.

Lin, Z., Chen, M., Ma, Y. (2010). The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices. arXiv:1009.5055.

Oba, S., Sato, M.-A., Takemasa, I., Monden, M., Matsubara, K.-I., Ishii, S. (2003). A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19(16):2088–2096.

Severson, K. A., Molaro, M. C., Braatz, R. D. (2017). Principal component analysis of process datasets with missing values. *Processes*, 5(3):38.

* To whom all correspondence should be addressed; email address of corresponding author: braatz@mit.edu

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

A MIXED-INTEGER REFORMULATION FOR GLOBAL TRAINING AND REGULARIZATON OF DEEP LEARNING UNDER DATA SCARCITY

H. Sildir[*,a] and O. Deliismail[b]

[a]Department of Chemical Engineering, Gebze Technical University, Kocaeli, 41400, Türkiye
[b]SOCAR Turkey R&D and Innovation Co., Izmir, 35800, Türkiye

Keywords: McCormick envelope, Piecewise linearization, Mixed-integer optimization, Deep learning

A mixed-integer approximation is proposed for the nonlinear and nonconvex terms, based on McCormick envelopes and piecewise approximations to ensure simultaneous global training and regularization of deep learning architectures.

Introduction

Deep learning architectures nonlinearly map the input vector to outputs through succeeding layers containing high number of parameters. A typical training problem is given by:

$$\begin{aligned} Min_{w_k,b_k} \quad & \sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ij}^{*}-y_{ij}^{exp}\right)^2+\gamma\sum_{k=1}^{K}\|w_k\| \\ s.t. \quad & \\ u_{k+1} &= f_k(w_k u_k+b_k), \quad k=1,\dots,K \end{aligned} \tag{1}$$

where u_k is the input vector to the kth layer; u_{k+1} is the output vector from the kth layer; w_k and b_k are parameters in the kth layer; y^* is the ultimate output obtained from the last layer; y^{exp} is the experimental data; f_k is the activation function at the kth layer; γ is tuning parameter for the tradeoff between fitting and regularization.

Deep learning architectures suffer from identifiability issues due to over-parameterization or data scarcity. In turn, their test performance might become poor unless traditional training and regularization problem is tailored to deal with nonlinear and nonconvex terms in Eq. 1.

Methodology

Modified mixed-integer training problem is given by:

$$\begin{aligned} Min_{w_k,b_k} \quad & \sum_{i=1}^{N}\sum_{j=1}^{M}F_1\left\{y_{ij}^{*}-y_{ij}^{exp}\right\} \\ s.t. \quad & \\ u_{k+1} &= F_3\{F_2\{w_k u_k\}+b_k\}, \quad k=1,\dots,K \\ \sum_{k=1}^{K}\sum|w_k| &\le w_d, \quad k=1,\dots,K \\ |w_k| &\ge w_{kb}w_{min}, \quad k=1,\dots,K \\ \sum_{k=1}^{K}\sum w_{kb} &\le w_{kbd}, \quad k=1,\dots,K \end{aligned} \tag{2}$$

where F_1, F_2, F_3 are mixed-integer approximations based on McCormick envelopes, piecewise linearization () and reformulations to account for matrix multiplication of continuous variables, nonlinear activation functions and absolute value function. Proposed formulations provide a mixed integer linear formulation and deliver significant theoretical advancements. The regularization term appears in the constraints, in contrast to Eq. 1 where a multi-objective optimization is performed, for an explicit formulation. Those additional constraints prune the architecture through binary variables which account for the existence of a particular connection by linking constraints, serving similar purposes to dropout methods. A further weight regularization is performed through defining a minimum weight magnitude, w_{min}, to ensure a significant weight value once the connection exists in the network and constraint summation of weights.

Results

Proposed approach is implemented on actual dataset collected from a refinery, in which data are scarce. Table 1 provides a summary of some statistics on the performance of the proposed method (PM) and compared to traditional regularization method (TM):

Table 1. Performance comparison

Criteria	PM	TM
Connection	8	18
RMSE	0.026	0.47
$\sum_{k=1}^{K}\sum\|w_k\|$	0.05	0.15

References

Sildir, H., Aydin, E., (2022). A Mixed-Integer linear programming based training and feature selection method for artificial neural networks using piece-wise linear approximations, *Chemical Engineering Science*, **249**:117273.

* hasansildir@gtu.edu.tr

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

PHYSICS GUIDED MACHINE LEARNING MODEL PREDICTIVE CONTROL OF A HIGH DENSITY POLY-ETHYLENE SLURRY REACTOR

Zhen-Feng Jiang[a], David Shan-Hill Wong[a]*, Yuan Yao[a]*, Jia-Lin Kang[b]*,
Yao-Chen Chuang[c], Shi-Shang Jang[a], John D.Y. Ou[c]
[a]Department of Chemical Engineering, National Tsing Hua University, Hsinchu 30043, Taiwan
[b]Department of Chemical and Materials Engineering, National Yunlin University of Science and Technology, Yunlin 64002, Taiwan
[c]Center for Energy and Environmental Research, National Tsing Hua University, Hsinchu 30043, Taiwan
Keywords: Physics Guided Neural Network , Model Predictive Control, HDPE reactor

INTRODUCTION

As the process being controlled by model predictive control (MPC) becomes more complex, use of deep learning (DL) models has received a lot of interests (Ren et al 2022). It is well known that physics guidance (PG) can improve the generalization ability of a DL model (Karpartne et al 2017). In this work, a physics guided sequence-tosequence model with memory layer (PG-StS+ML) for high-density polyethylene (HDPE) reactor (Jiang et al 2022) is presented. Consistency of process gains of output sensor variables with respect to manipulated variables in key control loops are guranteed. Reults of MPC using this model are presented.

METHOD

An ASPEN Dynamics™ simulator based on the kinetics of Khare et al 2002 was developed. Grade transitions and steady state productions with varying catalyst activities of a local plant were simulated. There are many sensors variables (SV) and mauipulated variables (MVs) in the process. MPC of two key controlled variables, hydrogen to ethylene ratio and pressure, by catalyst and hydrogen feeds as manipulative variables are consided.

The StSML model is illustrated in Figure 1. The encoder uses operation and sensor data in a past window of length W to estimate hidden states. A decoder predicts sensor output in a future horizon, of length H, given a future plan of MVs. Gated recurrent units were used in the encoder and decoder. Daily operations of a local plant in a two-year period were simulated as training,validation, and testing data. Gain consistencies between the four pairs of key MVs and SVs (Table 1) were incorprated as part of the loss function in training. The future operation plan can be optimized in a nonlinear MPC(NMPC) procedure using differential evolution algorithm. The key distinction between our NMPC approach and more traditional MPC is that there is no need of plant tests for model development.

RESULTS

Figure 2 illustrates results of directing a smooth grade transition using our PG-StSML NMPC. Results of ASPEN DMCPlus, which is not expected to work because of its linear nature, was incorporated as comparison.

CONCLUSIONS

A novel hybrid model (PG-StS+ML) was proposed for NMPC of a HDPE reactor. This success of the approach offers an alternate approach to nonlinear MPC of complex chemical processes.

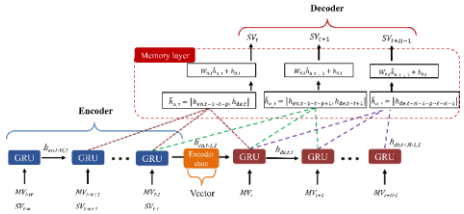

Figure 1: Sequence-to-Sequence Model with Memory Layer

Table 1: Gain Consistency Table

	HER	Pressure
Catalyst flow	+	–
Hydrogen flow	+	+

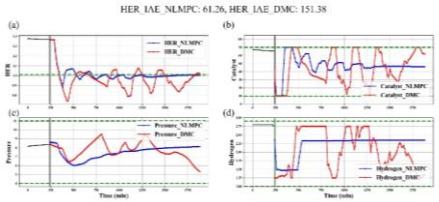

Figure 2 Grade Transition Control Performances

References

Jiang, Z. F., Wei, X. Z., Wong, D. S. H., Yao, Y., Kang, J. L., Chuang, Y. C., Jamg S.S. & Ou, J. D. Y. (2022). Model Predictive Control of Grade Transition with Attention Base Sequence-to-Sequence Model. In Computer Aided Chemical Engineering (Vol. 49, pp. 367-372). Elsevier.

Karpatne, A., Watkins, W., Read, J., & Kumar, V. (2017). Physics-guided neural networks (pgnn): An application in lake temperature modeling. arXiv preprint arXiv:1710.11431, 2.

Khare, N. P., Seavey, K. C., Liu, Y. A., Ramanathan, S., Lingard, S., & Chen, C. C. (2002). Steady-state and dynamic modeling of commercial slurry high-density polyethylene (HDPE) processes. Industrial & Engineering chemistry research, 41(23), 5601-5618.

Ren, Y. M., Alhajeri, M. S., Luo, J., Chen, S., Abdullah, F., Wu, Z., & Christofides, P. D. (2022). A tutorial review of neural network modeling approaches for model predictive control. Computers & Chemical Engineering, 107956.

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

FARM: A FAST, ACCURATE, ROBUST FAULT DETECTION AND DIAGNOSIS FRAMEWORK FOR INDUSTRIAL PROCESS MONITORING

Alireza Miraliakbar, Zheyu Jiang *
School of Chemical Engineering, Oklahoma State University, Stillwater, OK 74078

Keywords: Fault detection and diagnosis, Statistical Process Control, Support Vector Machine (SVM), Tangent space

Problem Statement

Modern manufacturing plants and industrial processes are continuously monitored by a large number of sensors. Despite having access to large volumes of historical and online sensor data, industrial practitioners still face challenges in the era of Industry 4.0 in effectively utilizing them to perform online process monitoring and fast fault detection and diagnosis. This is mainly because 1) fault scenarios in most industrial processes are complex and cannot be exhaustively enumerated or predicted, 2) sensor measurements continuously produce massive arrays of high-dimensional big data streams that are often nonparametric and heterogeneous, and 3) there is an intrinsic trade-off between fault detection speed and accuracy of any process monitoring tool.

The FARM Framework

To address these challenges, in this talk, we present a fast, accurate, and robust fault detection and diagnosis framework, which we call it FARM, for any general industrial process monitoring task (see Figure 1). FARM is a holistic framework that consists of a nonparametric statistical process control (n-SPC) module for fast fault detection and a modified SVM-based classification module for accurate and reliable fault diagnosis. The n-SPC module can detect any process mean shift or anomaly from heterogeneous high-dimensional sensor data streams as early as possible while maintaining a pre-specified in-control average run length (Jiang, 2023). Inspired by the work of Smith et al. (2022), our modified SVM-based classification module takes the covariance matrices of historical dataset containing different fault scenarios. Each covariance matrix, which corresponds to one of the fault scenarios, is then mapped to its tangent space of the Riemannian manifold. Following this preprocessing step, we use standard SVM with radial basis function (RBF) kernel to learn the mapping between the covariance matrix in the tangent space to its corresponding fault scenario.

Once offline training of FARM is complete, online sensor measurements will continuously be sent to FARM for simultaneous fault detection and diagnosis. First, they are monitored by the n-SPC module to detect any process anomaly in real time. Only if a process anomaly is detected will the online data be sent to the fault diagnosis module for accurate fault classification. Unlike general-purpose process monitoring frameworks, FARM's hierarchical architecture decomposes process monitoring tasks into two subtasks (fault detection and diagnosis), each of which is accomplished by specialized techniques. This allows fast, accurate, and robust fault detection and diagnosis to be simultaneously accomplished by FARM.

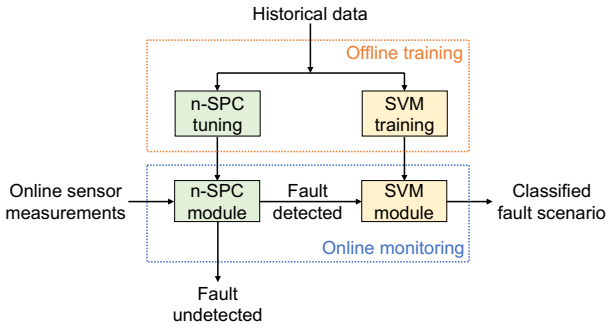

Figure 1. FARM's hierarchical architecture.

Results

We test our FARM framework on the benchmark case study of the Tennessee Eastman Process (Rieth et al., 2017). Our FARM framework achieves outstanding performance in fault detection time and accuracy.

References

Jiang Z. (2023). Online Monitoring and Robust, Reliable Fault Detection of Chemical Process Systems, *Comp. Aided Chem. Eng.*, **52**, Accepted.

Smith A., Laubach B., Castillo I., Zavala V.M. (2022). Data analysis using Riemannian geometry and applications to chemical engineering, *Comp. Chem. Eng.*, **168**, 108023.

Rieth C.A., Amsel B.D., Tran R, Cook M.B. (2017). Additional Tennessee Eastman Process Simulation Data for Anomaly Detection Evaluation, https://doi.org/10.7910/DVN/6C3JR1, Harvard Dataverse, V1

* To whom all correspondence should be addressed; zheyu.jiang@okstate.edu

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

FAULT DETECTION AND IDENTIFICATION FOR CHEMICAL PROCESS BASED ON 3D-CNN WITH CONTINUOUS WAVELET TRANSFORM

C. Ukawa*, and Y. Yamashita
Tokyo University of Agriculture and Technology, 2-24-16 Naka-cho, Koganei, Tokyo 184-8588, JAPAN

Keywords: Fault Detection, Fault Identification, Neural Network, Wavelet Transform

Introduction

A Variety of data-driven fault detection and identification (FDI) methods, especially machine learning (ML) methods, have been investigated and proposed for chemical plant in recent years. Data preprocessing is one of the most important steps in developing accurate machine learning methods. Applying appropriate preprocessing helps to improve the model performance. Signal analysis methods are widely used for data preprocessing in various fields, such as bearing diagnosis, electrocurdiogram, and also in chemical engineering fields, for instance, oscillations detection (Bounoua, 2023), and fault detection and diagnosis in batch process (Liu, 2021). Foulier Transform (FT) and Wavelet Transform (WT) are the most used methods in signal analysis. WT has gained attention because of its performance in nonsteady signal analysis. It transforms the original data into scalogram based on time-shift parameter and scaling parameter of the mother wavelet. This feature causes its capability for non-steady and non-linear data. It suggests that WT can be useful for extracting the feature from chemical process data. It is expected that the new perspectives from time-frequency domain contribute for FDI for chemical process.

This work proposed a novel FDI methods using continuous wavelet transform (CWT) and three-dimensional convolutional neural network (3D-CNN) and evaluated by applying the proposed method to Tennessee Eastman process datasets (TEP).

Proposed Method

Chemical process data was divided by a time-shifting window. Each divided data was transformed into a scalogram by using CWT to highlight a specific feature in the time-frequency domain. This preprocessing step was applied to each process variable, resulting in a 3D input data array consisting of the frequency axis, time axis, and number of process variables. The generated input data was then fed into a 3D-CNN model to obtain outputs representing the classification results of chemical process conditions.

* s193040z@st.go.tuat.ac.jp

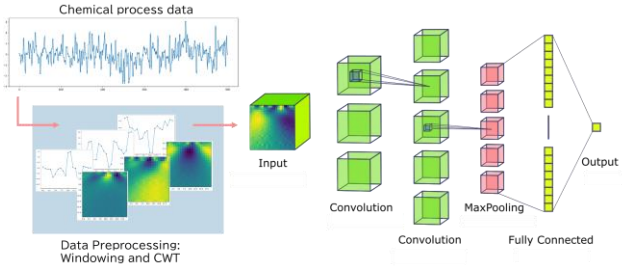

Figure 1. Overall view of proposed method.

Application and Result

The proposed method is applied to the TEP dataset introduced by Rieth in 2017. 22 measuring variables and 11 manipulated variables were selected as explanatory variables for the proposed methods. It consists of 1 normal operating condition and 20 fault conditions.

The result of proposed method were compared to the models with Fast Fourier Transform and Short Term Fourier Transform. The proposed method showed better performance in detecting 15 faults. For example, Fault 5 is one of the TEP faults that previous methods such as Principal Component Analysis could not identify with a high degree of accuracy. However, the proposed method achieved a higher identification rate and the input scalograms showed clear differences between normal operating conditions and Fault 5 conditions. These facts indicate that the proposed method effectively captures time-frequency domain features that are useful for detecting and identifing faults in chemical processes.

References

Bounoua, W., Aftab, M. F., Omlinm, C. W. P. (2023). Online detrended fluctuation analysis and improved empirical wavelet transform for real-time oscillations detection in industrial control loops, *Computers & Chemical Engineering*, Volume 172, 108173, ISSN 0098-1354

Liu, J., Wang, D., Chen, J. (2021). Global-local based wavelet functional principal component analysis for fault detection and diagnosis in batch processes, *Chemometrics and Intelligent Laboratory Systems*, Volume 212, 104279, ISSN 0169-7439

Rieth, C. A., Amsel, B. D., Tran, R., Cook, M. B., 2017, "Additional Tennessee Eastman Process Simulation Data for Anomaly Detection Evaluation", https://doi.org/10.7910/DVN/6C3JR1, Harvard Dataverse, V1

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

MACHINE LEARNING TO DETERMINE SOLID PROPERTIES OF SLURRIES THROUH RAMAN ATTENUATION OF SOLUTION PHASE

R. Prasad, S. H. Crouse, R. W. Rousseau, M. A. Grover *,
School of Chemical and Biomolecular Engineering, Georgia Institute of Technology,
Atlanta, GA 30332, USA

Keywords: Slurries, Raman spectroscopy, slurry density, particle size, machine learning

Background

Slurries are solid-liquid mixtures whose behavior during processing depends on properties such as size and concentration of solid particles (slurry density). Monitoring these slurry properties (size and concentration) is relevant in chemical, pharmaceutical, and nuclear waste industries. Slurry concentration and particle size of solids influence the settling behavior and the amount of water required to transfer slurries across tanks. These parameters are of particular interest at Hanford, where a waste treatment plant is being constructed to remediate nuclear waste slurries into glass for long-term disposal.

Multiple sensors could be deployed to obtain slurry properties for process control. However, employing a single process analytical technology (PAT) tool to obtain the required information is a beneficial and cost-effective strategy for any process. Therefore, this work is aimed to demonstrate the feasibility of optical spectroscopy, specifically *in-situ* Raman, to obtain information about the slurry density and particle size of slurry systems. A knowledge of solid-state properties of nuclear waste slurries in real-time may assist plant operators at Hanford to maintain waste remediation rates.

Results

The Raman intensity of a species dissolved in solution (nitrate anion) was monitored in the presence of solids. The Raman intensity of nitrate decreased as the concentration of solids increased, and the attenuation was observed to be dependent on the particle size of the solids (Fig. 1). This demonstrates that the solid properties of the slurries may be predicted from the solution phase Raman attenuation behavior.

Auto-machine learning algorithms such as Tree-based Pipeline Optimization (TPOT)[1] were employed to model the attenuation behavior and could predict slurry density with an R^2 of 0.86 and particle size with an R^2 of 0.75. These values are of reasonable accuracy considering the thickness of the slurries (containing up to 35 wt% solids). Although the ability of Raman to quantify solids in such dense slurries has been demonstrated previously[2], the current study demonstrates that Raman spectroscopy may also be used to determine additional solid properties of slurries using attenuation as an additional feature of the input-space.

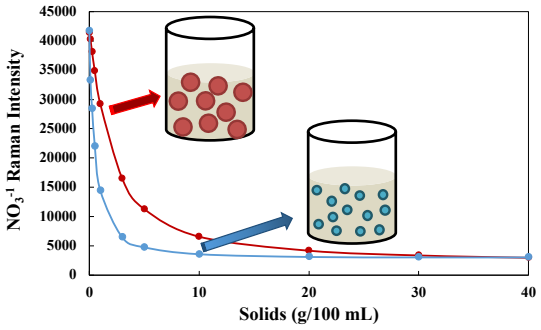

Figure 1. Attenuation of Raman intensity of nitrate anion (in solution) as a function of added solids of different particle sizes.

Acknowledgement

This work was produced by Battelle Savannah River Alliance, LLC under Contract No. 89303321CEM000080 with the U.S. Department of Energy. Support by the U.S. Department of Energy under Cooperative Agreement DE-FC01-06EW07053, entitled "The Consortium for Risk Evaluation with Stakeholder Participation III," is gratefully acknowledged.

References

Olson, R. S., Moore, J. H. (2016). TPOT: A tree-based pipeline optimization tool for automating machine learning. In: Proceedings of the Workshop on Automatic Machine Learning. 66–74.

Prasad, R., Crouse, S., Grover, M., Rousseau, R. (2022). Applicability of in-situ ATR-FTIR and Raman Spectroscopy to Measure Dense Slurries: A Hanford Case Study. AIChE Annual Meeting, Phoenix, AZ.

* To whom all correspondence should be addressed: martha.grover@chbe.gatech.edu.

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

DEVELOPMENT OF ALGORITHMS FOR MASS CONSTRAINED DYNAMIC NEURAL NETWORKS

Angan Mukherjee, Debangsu Bhattacharyya*
Department of Chemical and Biomedical Engineering, West Virginia University
Morgantown, WV 26506

Keywords: Mass conservation, Neural networks, Inverse problem, Forward problem, Constrained Optimization

Developing accurate first-principles models for nonlinear dynamic systems can be computationally expensive. Neural network (NN) models are relatively easier to develop but have their disadvantages in accurately representing nonlinear dynamic systems. Moreover, the measurement data available for training the NNs for any chemical engineering process may not necessarily satisfy mass conservation and other physics of the system. If these constraints are not satisfied during machine learning and during simulation (i.e., inverse and forward problems), model predictions can violate the conservation laws and therefore may not be meaningful. This work develops algorithms where mass balance constraints are exactly satisfied during inverse and forward problems, even though the corresponding training data violate the same.

Recent years have seen the development of physics-informed neural networks (PINNs) which aim to impose certain physics constraints by penalizing the objective function of typical NN training algorithms, thus only 'approximately' satisfying such constraints (Raissi et al., 2019). But in most chemical engineering applications, it is expected that certain mathematical relationships are exactly satisfied, which cannot be guaranteed by typical PINNs. Moreover, most hybrid approaches focused on exactly conserving mass of a system require rigorous understanding of the process for formulating physics-based differential/algebraic constraints and hence become system-specific. In this work, a novel class of network models is proposed, namely Mass Constrained Neural Networks (MCNNs), that 'exactly' satisfies mass balance constraints using only a subset of input and output boundary conditions. The mass conservation laws, expressed as species molar/atom balance equations, are posed as equality constraints in the nonlinear parameter estimation problem, thus providing flexibility to apply this algorithm to model any generic nonlinear chemical system. Efficient training algorithms are developed for solving both inverse and forward problems in presence of noise injected to simulated data for generating steady-state as well as dynamic training data for the MCNNs.

Unlike steady-state, developing a fully data-driven dynamic modeling approach by exactly satisfying mass balance equations can be significantly challenging, since conservation of mass during transience is difficult to check in general due to insufficient information about the holdup of a system. Therefore, in addition to steady-state modeling, efficient parameter estimation algorithms have also been developed for dynamic MCNNs represented by hybrid series/parallel all-nonlinear static-dynamic neural network models (Mukherjee and Bhattacharyya, 2023), which have been shown to perform significantly superior to many state-of-the-art approaches in terms of both computational expense and predictive capability. The proposed structures and algorithms are applied to model various nonlinear dynamic chemical processes. It is observed that the outputs from the MCNN exactly satisfy mass conservation, even though the data used for training the network violates the same. The optimal MCNNs developed in this work have also been shown to accurately capture the system truth, provided the data for model training is sufficiently rich (see Figure 1).

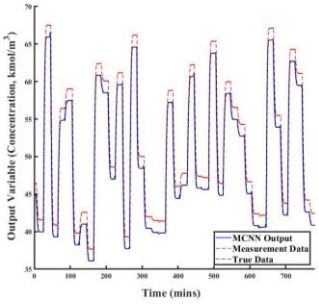

Figure 1. Simulation results from MCNN vs Measurement (training) and True data

Acknowledgement

The U.S. DOE financial support (Grant #: DE – FE0031768) for the project is gratefully acknowledged.

References

Raissi, M., Perdikaris, P. & Karniadakis, G. E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **378**, 686–707 (2019).

Mukherjee, A. & Bhattacharyya, D. Hybrid Series/Parallel All-Nonlinear Dynamic-Static Neural Networks: Development, Training, and Application to Chemical Processes. *Ind. Eng. Chem. Res.* **62**, 3221–3237 (2023).

* Corresponding author. Email address: Debangsu.Bhattacharyya@mail.wvu.edu

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

A REINFORCEMENT LEARNING-FUZZY OPTIMIZED ARTIFICIAL NEURAL NETWORKS FRAMEWORK FOR CONTROL OF NONLINEAR PROCESSES

Jiaxin Zhang, Gade Pandu Rangaiah and Lakshminarayanan Samavedham[1]
Department of Chemical and Biomolecular Engineering, National University of Singapore, Singapore. 117585

Keywords: Fuzzy Optimized Artificial Neural Network; Reinforcement Learning; Process Control; Machine Learning

Introduction

Many chemical processes exhibit nonlinear behavior over their range of operation. Linear control strategies, perform poorly when large changes occur in the operating conditions. We propose the coupling of reinforcement learning (RL) strategies with fuzzy optimized artificial neural networks (FOANNs) to effectively control nonlinear systems. FOANNs have achieved significant improvement in model generalization and function approximation abilities. RL based on the DDPG (Deep Deterministic Policy Gradient) algorithm demonstrates excellent decision-making capabilities. The proposed RL-FOANN method was tested for its control capability on an interacting two-tank liquid level process (Figure 1).

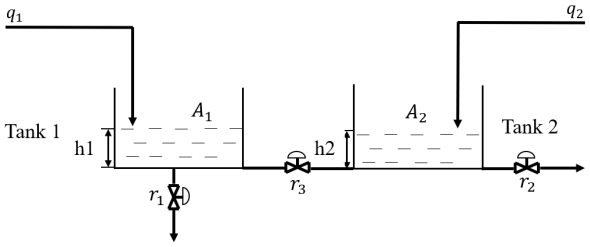

Figure 1: The Two-tank System

A discretized version of the process model (equations 1 and 2) is used to find the next state according to the current action and the current state.

$$\frac{dh_1}{dt} = \frac{(q_1 - r_1\sqrt{h_1} - r_3\sqrt{h_1 - h_2})}{A_1} \quad (1)$$

$$\frac{dh_2}{dt} = \frac{(q_2 - r_2\sqrt{h_2} + r_3\sqrt{h_1 - h_2})}{A_2} \quad (2)$$

Results and Discussion

Parameters A_1, A_2, r_1, r_2 and r_3 are set equal to 1. Flow q_2 is kept at 0 initially. The set point for h_1 is 10. Figure 2 shows the liquid level trajectory starting from initial value of $h_1 = 1$ and $h_2 = 0$. Starting from the previous steady state, a disturbance is introduced into the second tank by changing q_2 from 0 to 2. The results shown in Figure 3 illustrates the ability of the control strategy to effectively reject the disturbance.

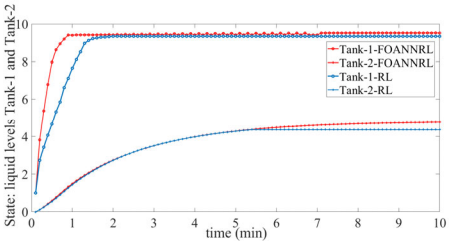

Figure 2. Set Point Tracking Comparison

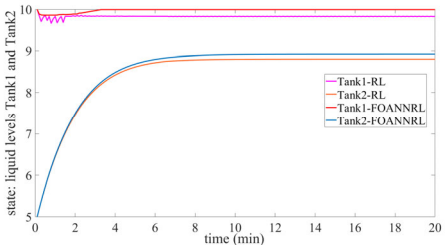

Figure 3. Disturbance Rejection Comparison

Mean Square Error in the controlled variable under Deep Q-network (DQN)-RL control, PID control and RL-FOANN compared in Figure 4 shows that RL-FOANN performs better than DQN-RL & PID.

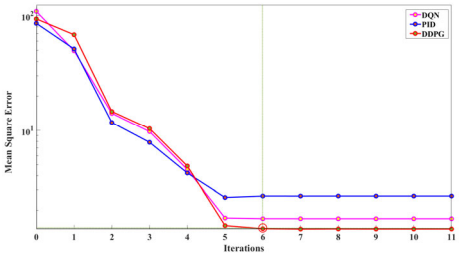

Figure 4. Performance Comparison

References

Hamoodi, S. A., Sheet, I. I., & Mohammed, R. A. (2019). A Comparison between PID controller and ANN controller for speed control of DC Motor. Paper presented at the 2019 2nd International conference on electrical, communication, computer, power and control engineering (ICECCPCE).

[1] To whom all correspondence should be addressed; chels@nus.edu.sg

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

A NOVEL FRAMEWORK TO DETERMINE COMPLEX PROCESS FEASIBILITY

Margherita Geremia[1], Fabrizio Bezzo[1], Marianthi G. Ierapetritou[2]*
[1]CAPE_Lab – Computer-Aided Process Engineering Laboratory. Department of Industrial Engineering, University of Padova, via Marzolo 9, 35131 Padova (PD), Italy
[2]Department of Chemical and Biomolecular Engineering, University of Delaware, Newark, DE, USA

Keywords: Feasibility analysis, Design Space, topological data analysis, adaptive sampling, surrogate models.

Feasibility analysis can be exploited to identify the subset of combinations of uncertain input parameters that satisfy all the process and quality constraints, i.e., the design space (DS) of the process. In the presence of disjoint feasibility regions and for computationally expensive nonconvex problems, the use of surrogate-based approaches has been successfully adopted to properly predict the boundaries of the process feasibility space (Wang and Ierapetritou, 2017). It is worth noticing that the choice and prediction accuracy of a suitable surrogate model strongly depends on the specific process of interest, and on the dataset that is available for training. In this context, we aim at investigating how to correctly identify the process feasibility space relying on the available dataset, and determining the minimum number of sampling points that are necessary to uncover the complexity of the original feasibility function. The aim is to compare the performance of different candidate surrogates, while uncovering the complexity of the process feasibility space based on the inclusion of additional sample points up to the attainment of a pre-set level of prediction accuracy.

Methodology and Results

We propose a novel framework to acquire information on the complexity of the feasible space and accurately predict the feasibility boundaries with the minimum number of training data. First, we couple Topological Data Analysis (TDA) (Smith and Zavala, 2021) and data interrogation (Sun and Braatz, 2021) to reconstruct the dataset complexity and restrict the number of candidate surrogate models to the most promising ones that can be further trained. Then, we compute the Bayesian information criterion (BIC) to evaluate quality of fitting and predictive performance of the different models. If none of the trained surrogates guarantees the preset level of accuracy (i.e., stop criterion), new sampling points are needed. The implementation of an adaptive algorithm locates additional points along the boundaries of the feasible region, which are included in the training dataset. The procedure is repeated until the stopping condition is achieved based on the preset level of accuracy. The methodology is tested on pharmaceutical case studies and numerical problems with complex feasible regions, such as the Gomez function (Sasena et al., 2002). Although the initial number of points and sampling technique affect the reconstruction of the problem complexity, the inclusion of adaptive points promotes a fast identification of all disjointed feasibility regions and significant accuracy improvement. Figure 1 shows the surrogate-based feasibility approximation of the Gomez function with 98% accuracy, that is attained after the addition of 154 adaptive points to the initial training dataset (25 Sobol's samples). The workflow can be efficiently utilized for higher-dimensional case studies; thus, it can be implemented for real manufacturing processes, where the accurate description of the DS is central for process development.

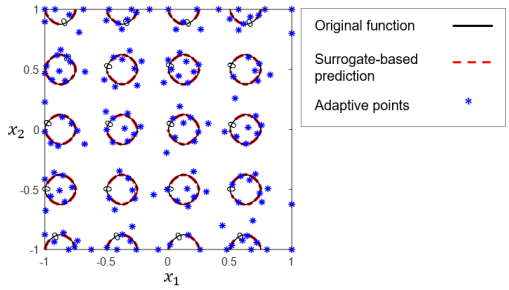

Figure 1. Gomez case study: surrogate-based feasibility approximation after 154 adaptive samples.

References

Sasena, M.J., Papalambros, P., Goovaerts, P. (2002). Exploration of metamodeling sampling criteria for constrained global optimization. *Eng. Optim*., **34**:3, 263–278.

Smith, A., Zavala, V. M. (2021). The Euler characteristic: A general topological descriptor for complex data. *Comput. Chem. Eng.*, **154**, 107463.

Sun, W., Braatz, R. D. (2021). Smart process analytics for predictive modeling. *Comput. Chem. Eng.*, **144**, 107134.

Wang, Z., Ierapetritou, M. (2017). A novel feasibility analysis method for black-box processes using a radial basis function adaptive sampling approach. *AIChE J.*, **63**:2, 532–550.

* To whom all correspondence should be addressed: mgi@udel.edu

A Novel Machine Learning-based Method for Tracking Renewable Carbon during Biomass Co-processing

Liang Cao, Yankai Cao, Siang Lim, R. Bhushan Gopaluni *
Department of Chemical and Biological Engineering, The University of British Columbia,
Vancouver, British Columbia, V6T 1Z3, Canada

Keywords: Renewable carbon, Soft sensors, Interpretable neural networks, Biomass Co-processing

Decarbonization of the oil refining industry is crucial for reducing carbon emissions and mitigating climate change. Co-processing biomass at existing oil refineries offers a promising strategy for achieving this goal. (Su, Jianping, et al, 2022) However, accurately quantifying the renewable carbon content of co-processed fuels remains challenging due to the complex processes involved, with current methods relying on expensive offline ^{14}C measurements at specialized labs. This study proposes a novel data-driven approach utilizing machine learning and high-quality, large-scale commercial data to develop soft sensors for estimating real-time renewable carbon. The method employs interpretable deep neural networks for input selection, robust linear regression, and bootstrapping techniques for renewable content estimation. Four previous ^{14}C measurements during co-processing at the fluid catalytic cracker demonstrate the feasibility of the proposed method. This method allows the industry and policymakers to quantify renewable carbon content, generate credits, and accelerate the transition to a more sustainable energy system.

Methods

Co-processing is a promising method that involves the simultaneous treatment of biomass and petroleum-based feedstocks in a refinery, as shown in Figure 1. This process enables the production of biofuels and other value-added products by subjecting biomass and petroleum-based feedstocks to refining processes like hydrotreating and catalytic cracking. Co-processing offers numerous advantages such as reducing dependence on oil, lowering carbon emissions, promoting sustainable development in agriculture and forestry, and diversifying energy sources.

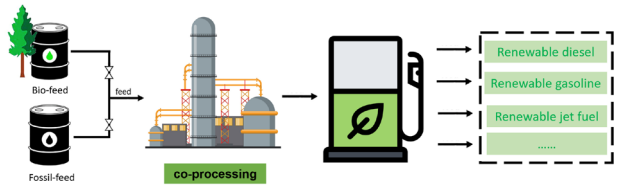

Figure 1. A diagram of co-processing

The integration of Artificial Intelligence (AI) into the field of biomass co-processing has the potential to provide new and innovative solutions for tracking renewable carbon. In this work, we use soft sensors, deep neural networks, and SHAP for feature selection (as shown in Figure 2), as well as bootstrapping and robust linear regression for building predictive models.

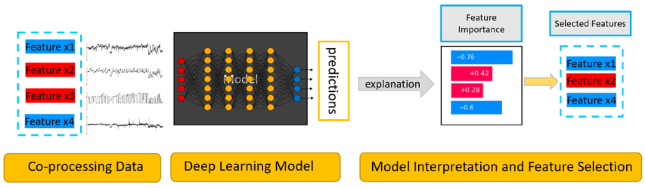

Figure 2. Feature selection with SHAP and DNNS

Results

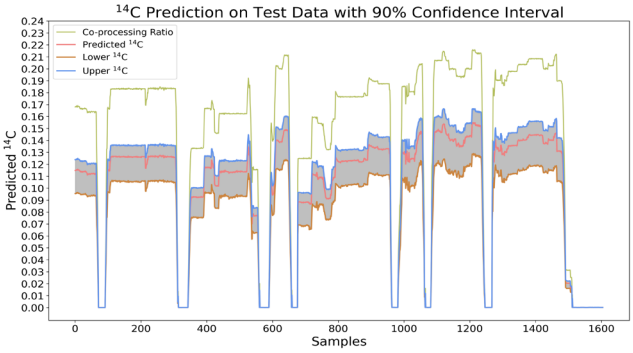

Figure 3. Predicted renewable carbon (^{14}C)

Figure 3 displays the predicted ***^{14}C*** values and their corresponding 90% confidence interval with bootstrapping and robust linear regression. The industrial partner conducted four ***^{14}C*** experiments with co-processing ratios and ***^{14}C*** of 10.75% and 6.30%, 12.24% and 7.30%, 12.32% and 7.80%, and 12.22% and 7.40%, respectively. We compared these experimental results with our AI model's renewable content predictions. and the results are presented in Table 1.

Table 1. Experimental and AI Results of renewable carbon (14C) during Co-processing

Experiment			AI model			
	ratio	^{14}C	ratio(mean)	^{14}C(mean)	^{14}C(lower)	^{14}C(upper)
Sample 1 01:00	10.75%	6.30%	10.75%	6.71%	5.57%	7.85%
Sample 2 14:00	12.24%	7.30%	12.24%	7.78%	6.62%	8.82%
Sample 3 01:00	12.32%	7.80%	12.32%	7.79%	6.44%	9.15%
Sample 4 14:00	12.22%	7.40%	12.22%	7.72%	6.63%	8.92%

References

Su, Jianping, et al. (2022). Tracking the green coke production when co-processing lipids at a commercial fluid catalytic cracker (FCC): combining isotope 14 C and causal discovery analysis. *Sustainable Energy & Fuels*, *6*(24), 5600-5607.

* Corresponding author: R. Bhushan Gopaluni; Email address: bhushan.gopaluni@ubc.ca.

Foundations of Process/Product Analytics and Machine learning FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

Optimization of Modern Manufacturing for Thermoelectric Material using Machine Learning and Data Science

K. Wang, A. W. Dowling*
Department of Chemical and Biomolecular Engineering, University of Notre Dame
Notre Dame, IN 46556

Keywords: Bayesian optimization, Model-based design of experiments, Machine learning, Data science

Objectives & Relevance

Thermoelectric (TE) materials convert thermal energy into electricity and show promise for niche applications. A transition from traditional Edisonian optimization methods to data-driven frameworks is necessary for the efficient discovery of high-performance TE materials. Failure to utilize computational frameworks for systematically optimizing the manufacturing process can lead to inefficient resource utilization and cannot guarantee success. This work demonstrates the use of existing computational frameworks for the effective optimization of modern manufacturing processes.

Background & Relevant Prior Contributions

TE materials improve energy efficiency and reduce emissions by converting waste heat into electricity. The performance of TE materials depends on the dimensionless figure of merit (zT) (Jeffrey, and Toberer, 2008). Sintering is a crucial step in the TE manufacturing process, transforming TE particles into dense materials. Two such methods are flash sintering, which utilizes intense pulsed light, and plasma sintering, which relies on electrically generated plasma. These methods enable the sintering of TE materials under controlled conditions, optimizing their properties without damaging the substrate. When optimizing the sintering process of TE materials, experimentalists often employ a trial-and-error approach, known as Edisonian search, to enumerate various possible experiments. However, this methodology becomes impractical when dealing with complex, high-dimensional problems due to the vast number of potential experiments, e.g., $O(10^5)$, and the associated economic and labor costs. (Wang and Dowling, 2022) To address this challenge, adaptive sampling strategies, such as Bayesian optimization (BO) and model-based design of experiments (MBDoE), have been proposed for optimizing manufacturing processes in the field of thermoelectric materials.

Review of Main Results: Bayesian optimization

Bayesian optimization (BO) operates under the assumption that the relationships between manufacturing process parameters and desired material properties are black-box functions. BO employs a stochastic surrogate model, such as Gaussian process regression (GPR), to approximate these black-box functions, enabling the construction of an inexpensive acquisition function that recommends the next experiments to be conducted. Within the BO process, the acquisition function balances the tradeoff between exploration (sampling where uncertainty is high) and exploitation (sampling where the objective mean is high) to optimize the objective function. In this study, we demonstrate the use of GPR to model the black-box relationship between flash sintering manufacturing parameters and power factor. Subsequently, we implement BO for adaptive experimental design in flash sintering, achieving state-of-the-art thermoelectric material performance in just 40 experiments. (Mortaza et al., 2022)

Review of Main Results: Model-based design of experiments

Model-based design of experiments (MBDoE) operates under the assumption that relationships between manufacturing process parameters and desired material properties are partially known or can be derived from first principles. MBDoE incorporates scientific models with physically meaningful parameters, offering a systematic framework for recommending new experiments based on prior observations. By considering metrics such as D-optimality, MBDoE recommends the most informative experiments, efficiently identifying the optimal model and its corresponding parameters among candidate models. In this study, we demonstrate the use of a first-principle model to emulate the behavior between the manufacturing process and film thickness. We then apply MBDoE to recommend a single experiment to further postulate and calibrate the model for quantitative control of thickness. (Wang et al., 2023)

References

Snyder, G. Jeffrey, and Eric S. Toberer. Nature materials 7.2 (2008): 105-114.
Saeidi-Javash, Mortaza, et al. Energy & Environmental Science 15.12 (2022): 5093-5104.
Wang, Ke, and Alexander W. Dowling. Current Opinion in Chemical Engineering 36 (2022): 100728.
Wang, Ke, et al. Digital Chemical Engineering 6 (2023): 100076.

Foundations of Process/Product Analytics and Machine learning FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

EFFICACY OF COMPUTER AIDED MOLECULAR DESIGN (CAMD) BY LARGE LANGUAGE MODEL GPT-3.5

Yuan-Cyún Liao[a], Jia-Lin Kang[b]*, David Shan-Hill Wong[a]*
[a]Department of Chemical Engineering, National Tsing Hua University, Hsinchu 30043, Taiwan
[b]Department of Chemical and Materials Engineering, National Yunlin University of Science and Technology, Yunlin 64002, Taiwan

Keywords: Computer aided molecular design, ChatGPT

Introduction

Recently large scale language model (LLM), Chat Generative Predictive Transformer (ChatGPT) has received a lot of interests because of its artificial generative intelligence (AGI) properties (Bubeck et al 2023). In this work we explored potential of transforming GPT-3.5 into a computer aided molecular design (CAMD) tool with a toy problem of finding a molecule with specific value of solubility parameter (desirable property) with boiling points provided asauxiliary molecular information.

Method

We ask ChatGPT how to do molecule design with the following prompt and answer (Figure 1).

Prompt :
How do I use GPT 3 to implement Molecular Design for multitasks with thermodynamic properties as inputs and SMILES format output?

ChatGPT :
To implement molecular design with GPT-3, you would need to:
1. Train GPT-3 on a large dataset of molecular data, such as SMILES format and thermodynamics properties.
2. Fine-tune GPT-3 on a smaller, task-specific dataset that includes the desired input-output mapping (thermodynamics properties as inputs and SMILES format as output).
3. Use the fine-tuned GPT-3 model to generate molecular designs for your specific task by inputting the desired thermodynamics properties and obtaining the output in SMILES format.
4. Validate the generated molecular designs using molecular simulation or experimental methods to ensure accuracy and feasibility.
5. Repeat the process as necessary to improve the performance of your molecular design model.

Figure 1 Asking ChatGPT to do molecular design

Let NA be the number of SMILES structure with auxiliary properties we provide, NT be the number SMILES structure with target property we provide. Valid SMILES (NV) are the numbers of valid molecular structures, and good SMILES (NG) are the number of SMILES generated with desired properties within 5% of target. Random search from a database of more than 10000 compounds with known boiling point and solubility parameters curated from ASPEN Plus was used as a benchmark. In case that a SMILE structure that is not found in the original database was generated, a deep learning surrogate model developed using the database was used to predict the solubility parameter (Huang et al 2022). Each experimental setting was repeated 10 times and 2-sample t-tests were used to determine the superiority of the model in the row to that in the column of Table 1.

Results

Figure 2 shows the change in valid SMILES as NA increases. Obviously providing enough quantitative structure information of an easily available molecular properties increase the efficacy of the generative model.

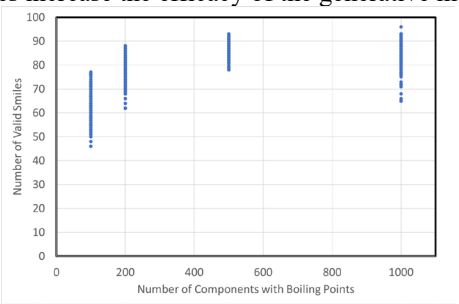

Figure 2 Increase of NV with NA

Table 1 shows that changes in good SMILES when target values are 23100 and 13100 $\sqrt{J/m^3}$. The results show that even with no prior information provided on solubility parameters, ChatGPT3-5 procedure is superior to random search. Provision of scant amount of target information is of little help, but enough quantitative structure information is important.

Table 1 Good SMILES with various NT (N=1000)

	NG	Random	NT=0, NA=1000	NT=10 NA=1000	NT=0, NA=500
Solubility Parameter=23100					
Random	7.2 ± 2.4	--			
NT=0,NA=1000	9.8 ± 3.1	0.036	--		
NT=10,NA=1000	9 ± 2.5	0.052	0.359	--	
NT=0,NA=500	7.9 ± 3.5	0.319	0.220	0.426	--
Solubility Parameter=13100					
Random	1.5 ± 1.5	--			
NT=0,NA=1000	3.4 ± 1.4	0.015	--		
NT=10,NA=1000	2.8 ± 1.7	0.093	0.313	--	
NT=0,NA=500	1.9 ± 0.9	0.291	0.015	0.068	--

Conclusions

The results showed that AGI LLM like CHATGPT can be fine-tuned into an adequate CAMD expert system with minimal coding. However, the true potential of this approach needs to be explored in depth.

References

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Huang J.S. Wong D.S.H., Chiu, C.T.,Kang J.L., Tai S.H., Tang P.H. Graph based deep learning quantitative structure properties relation for solvents, (2022) PSE Asia Chennai,India

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

Machine-learning-powered molecular design: optimal solvents for hybrid extraction-distillation

Johanna Lindfeld, Luca Bosetti, Benedikt Alexander Winter, Johannes Schilling, and André Bardow*
Energy and Process System Engineering, ETH Zurich, 8092 Zurich, Switzerland

Keywords: molecule design, process design, thermodynamic properties

Computer-aided molecular and process design (CAMPD) simultaneously optimizes molecules and processes, relying on accurate molecular property predictions. Group contribution (GC) and quantum mechanics-based (QM) methods have traditionally been used (Papadopoulos 2018), but machine learning (ML)-based frameworks have recently demonstrated higher accuracy and efficiency due to their ability to capture complex relationships. A natural language model has been shown to outperform GC and QM methods in calculating activity coefficients for various molecules (Winter 2022).

The full potential of incorporating ML-based property prediction into the CAMPD framework remains unexploited. The accuracy and speed of property predictions make this approach promising for exploring a larger design space in an efficient manner. In this work, we combine ML-based property prediction and genetic algorithm-based (GA) molecule design with pinch-based process models to optimize solvent molecules for extraction-distillation.

The present investigation focuses on the hybrid extraction–distillation process for the separation of γ-valerolactone (GVL) from aqueous solutions. GVL is a promising bio-based platform chemical. However, its separation is challenging. Our aim is therefore to reduce the energy demand for separation. While the conventional process involves only distillation, the inclusion of an extraction unit is expected to lower the overall energy consumption and production costs (Scheffczyk 2018). The liquid-liquid equilibrium between water and a suitable solvent is exploited in the extraction unit, which necessitates careful consideration of the solvent design.

Method and results

To include an extraction step in a process, a suitable solvent must be identified, which requires precise methods to calculate its thermodynamic properties. The SMILES-to-Properties Transformer (SPT) machine learning model, is a natural language processing model and takes SMILES code as input to predict molecular properties. SPT is originally trained to estimate activity coefficients (Winter 2022). Here, we further trained SPT to estimate properties relevant to extraction processes, such as melting and boiling temperatures. The thermodynamic properties can either be used directly as heuristic objectives for solvent design by the GA or serve as input for the process model to calculate a process-level objective function, such as the energy demand for the entire hybrid extraction-distillation process. The GA iteratively ranks and refines the solvent design according to the objective function.

In this abstract, we only report the solvent design based on the minimum energy demand of the entire process, which is driven by the heat required in the reboiler of the distillation column. By generating solvents composed of only carbon, oxygen, and hydrogen atoms, the best separation performances, according to the minimum energy demand, have been achieved with isobutyric acid (Q_{min} = 3.814 MJ/kmol$_{\text{Feed}}$) at a GA runtime of 8h. When comparing isobutyric acid to the best-performing molecule of Scheffczyk et al. (2018), divinyl ether, it was found that using isobutyric acid resulted in a 14.8% reduction in the minimum energy required for the hybrid extraction-distillation process.

Conclusion

In summary, we introduce a machine-learning-powered framework for the solvent design of the separation of a GVL from water. This design framework allows us to screen rapidly and accurately a vast number of solvents for which the thermodynamic properties are not experimentally available. This framework can be expanded to include conceptual design, enabling the integrated computer-aided molecular and process design (CAMPD).

Acknowledgments

The project leading to this application has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 101036854 and from NCCR Catalysis (grant number 180544) a National Centre of Competence in Research funded by the Swiss National Science Foundation.

References

Papadopoulos, A. I., Tsivintzelis, I., Linke, P., & Seferlis, P. (2018). *Chem., Mol. Sci. and Chem. Eng*.

Winter, B., Winter, C., Schilling, J., & Bardow, A. (2022). *Digital Discovery*, 1(6), 859-869.

Scheffczyk, J.; Schäfer, P.; Fleitmann, L.; Thien, J.; Redepenning, C.; Leonhard, K.; Marquardt, W.; Bardow, A. *Mol. Syst. Des. Eng.* (2018), 3 (4), 645–657.

* To whom all correspondence should be addressed: abardow@ethz.ch

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

LEARNING THE BEST CONFIGURATION FOR CONTROLLING MODULAR DYNAMIC SYSTEMS

Yi Dai and Andrew Allman *
University of Michigan
Ann Arbor, MI 48109

Keywords: Reconfiguration, Machine Learning, Modular Facilities, Model Predictive Control

Mixed-integer model predictive control (MPC) allows for discrete variables to be optimized with continuous-valued control decisions in a feedback control law. The presence of discrete decisions, which are a common feature in industrial processes, renders the optimization problem challenging to solve in a time scale relevant to control. A potential solution to this issue is the use of surrogate models which have been studied for problems in modeling, feasibility analysis, and optimization due to their ability to exploit the black-box nature and the computational simplicity (Bhosekar and Ierapertritou, 2018). Two widely used approaches for surrogate modeling in optimization exist: model order reduction and data-driven modeling (Biegler et al., 2014). The former aims to acquire a lower dimensional system that has similar response characteristics to the original system. The latter relies on data without the requirement of explicit formations of an existing model. Recent advances in machine learning have made these approaches more popular, particularly for reducing system complexity and creating mathematical models based on data generated from simulation optimization or "black box" models (Cozad et al., 2014).

In this work, we present an approach to determine integer control decisions *a priori* to solving the MPC problem using data-driven machine learning classifier algorithms. Our previous work (Dai et al., 2023) considered control of numbered-up modular systems with fixed configurations, and which demonstrated that operational conditions could affect selection of the optimal configuration of modules. Using the same benchmark modular three reactor system, here we collect data on which modular configuration achieves the best control performance for various initial conditions of states, values of disturbances and set points of process outputs. Data are obtained by solving the optimal control problem offline for all possible configurations under the following workflow: first, values of set points, disturbances, and initial conditions are chosen from a space-filling sampling of the parameter space. Next, additional data points are generated using a moving horizon simulation to update initial conditions, as would be done in practical implementation of the MPC. Once the data are collected, various classifier models are trained which fall within three major types: support vector machines (SVM), decision trees, and k-nearest neighbors (KNN). These provide mathematical guidelines for reconfiguration to help the online MPC quickly make decisions on dynamically selecting the optimal configuration. Principal component analysis (PCA) is also implemented to identify the parameters that have the largest effect on the configuration decision making. By fine-tuning machine learning classifiers with linear PCA in order to determine the best module configuration offline, we develop an online configuration switching approach determined by the dynamic system condition in the MPC of numbered-up modular facilities.

Results demonstrate that KNN classifier has the greatest accuracy on predicting the optimal configurations compared to the other two methods. Encouragingly, even when the classifier predicts the incorrect configuration, many of the wrong predictions match the second-best configuration of the corresponding initial conditions, particularly when the best and the second-best configurations have extremely similar performance index. Based on that, we built a metric to demonstrate how similar the performance of the predicted configuration is to that of the target configuration. KKN classifier still has the highest score, while the other two methods perform better than only counting prediction accuracy.

This work was supported in part by the U.S. Department of Energy, Office of Science, USA, Biological and Environmental Research (BER), USA under Award # AWD018868, which is gratefully acknowledged.

References

A. Bhosekar and M. Ierapetritou, (2018), "Advances in surrogate based modeling, feasibility analysis, and optimization: A review," *Comp. Chem. Eng.*, vol. 108, pp. 250-267.

L.T. Biegler, Y. Lang, and W. Lin, (2014). "Multi-scale optimization for process systems engineering," *Comp. Chem. Eng.*, vol. 60, pp. 17-30.

A. Cozad, N. V. Sahinidis, and D. C. Miller, (2014). "Learning surrogate models for simulation-based optimization," *AIChE J.*, vol. 60, pp. 2211-2227.

Y. Dai, S. Fay, and A. Allman, (2023) "Analysis of model predictive control in numbered-up modular facilities," *Dig. Chem. Eng.*, vol. 7, p. 100088.

* To whom all correspondence should be addressed; allmanaa@umich.edu.

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

INDUSTRIAL PROCESS MONITORING USING DEEP LEARNING BASED PROCESS ANALYTICS AND FEATURE EXTRACTION

Cheng Ji[1], Fangyuan Ma[1,2], Jingde Wang[1] and Wei Sun[1,*]
[1]College of Chemical Engineering, Beijing University of Chemical Technology, 100029, North Third Ring Road 15, Chaoyang District, Beijing, China

[2]Center of process monitoring and data analysis, Wuxi Research Institute of Applied Technologies, Tsinghua University, 214072 Wuxi, China

Keywords: Process system engineering, Process modeling and monitoring, Deep learning, Latent space regularization

Practical process operation is highly affected by environmental factors and equipment conditions, making it not exactly consistent with the mechanism model. Process modeling and monitoring can be employed to extract process information from historical operation data to compensate for the limitations of the mechanism model, and at the meantime identify the process operation status. In recent years, deep learning models have been widely applied in state-of-the-art process monitoring methods because of their powerful information processing capability aided by connections among neurons. Theoretically, the feature extraction performance of deep learning models can be improved with increasing numbers of neurons in the hidden layers, but it also brings to higher requirements on the amount of training data, which may not be satisfied, coupled with poor interpretability of the black-box models, making them not trustable enough for industrial community. In terms of the above issues, extensive researches have been conducted to improve the feature extraction performance and the generalization ability of the deep learning models. Generally, it can be achieved through modifying the structure of the models or applying regularization to the latent space. For example, a series of recurrent neural networks (Cheng, et al., 2019) and convolutional neural networks (Ma, et al., 2022) have been proposed to extract temporal characteristics and spatial characteristics of data to improve process monitoring performance. On the other hand, massive process monitoring methods based on graph neural networks (Xiao, et al., 2023) and variational autoencoder (Lee, et al., 2019) are also proposed to improve the generalization ability of the model by regularizing the latent feature space. Generally speaking, the deep learning models are modified according to process information and data characteristics of industrial processes. Actually, there are still various complex features in industrial operation data that have not been fully considered, which provides a great potential for research on deep learning based process monitoring. This work presents several newly developed deep learning models, which have been applied to monitor different chemical industrial processes. A three-dimensional convolutional neural network is proposed to extract the spatial characteristics resulted from the spatial arrangement of chemical instrument. Then a difference-embedded recurrent neural network is proposed to extract the short-term nonstationary characteristics of data by embedding difference layers to the latent space. Moreover, a contrastive loss is proposed in the Siamese network to perform the latent space regularization, by which both the local correlation among training samples at different time stamp and the process information contained in historical fault samples can be effectively extracted. Above methods have been verified through several industrial case studies.

Acknowledgments

Authors gratefully Acknowledge the support from National Natural Science Foundation of China (Grant Number: 22278018).

References

Cheng, F., He, Q. P., Zhao, J., (2019). A novel process monitoring approach based on variational recurrent autoencoder, Comput. Chem. Eng., **129**, 106515.

Lee, S., Kwak, M., Tsui, K.-L., Kim, S. B., (2019). Process monitoring using variational autoencoder for high-dimensional nonlinear processes, Eng. Appl. Artif. Intell., **83**, 13-27.

Ma, F., Ji, C., Wang, J., Sun, W., (2022). Early identification of process deviation based on convolutional neural network, Chin. J. Chem. Eng.

Xiao, L., Yang, X., Yang, X., (2023). A graph neural network-based bearing fault detection method, Sci. Rep., **13**: (1).

* Wei Sun; sunwei@mail.buct.edu.cn.

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

BAYESIAN OPTIMIZATION FOR NONLINEAR MODEL CALIBRATION

Montana N. Carlozo[1], Bridgette J. Befort[1], Edward J. Maginn[1], and Alexander W. Dowling[1*]
[1]The University of Notre Dame

Keywords: Bayesian optimization, Gaussian process, molecular simulation, nonlinear regression

Nonlinear model calibration is essential in many facets of engineering including reaction kinetics, additive manufacturing, and pharmaceutics. However, standard derivative-based optimization methods are intractable for computationally expensive models, such as molecular simulations, because they require too many function evaluations. In this case, Gaussian process regression can be used with Bayesian optimization (GPBO) to estimate model parameters rapidly.

Prior work suggests that a GP which directly emulates the results of a simulation will yield higher accuracy results than a GP which trains on an error metric between the experimental results and model predictions (Befort et al., 2022; Madin & Shirts, 2022).

Method and Demonstration

In this work, we demonstrate and compare the use of five formulations of GPBO for parameter estimation and compare their accuracy to a traditional least-squares approach. As an example, we regress parameters $\boldsymbol{\theta}$ from data generated via Equation 1:

$$y_i = \theta_1 x_i + \theta_2 {x_i}^2 + {x_i}^3 + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, 0.01) \tag{1}$$

Given Equation 1, we then seek to solve Equation 2

$$\min_{\underline{\boldsymbol{\theta}} \le \boldsymbol{\theta} \le \bar{\boldsymbol{\theta}}} \sum_{i=1}^{i=N} (f(\boldsymbol{\theta}, x_i) - y_i)^2 \tag{2}$$

where $y_i \sim f(\boldsymbol{\theta}, x_i)$ for sample $i \in \{1, \dots N\}$.

The first two methods use a GP to model either the error (1A) or the logarithmic error (1B) between the simulation and experimental results. The last three methods use a GP to emulate the model using an independence approximation (2A), an independence approximation and logarithmic scaling (2B), or a sparse grid (2C) approximation for the expected improvement (EI) acquisition function used by BO.

Results

Our results, indicated in Table 1, show that emulator-based (type 2) GPBO methods can potentially converge 10x faster than standard (type 1) GPBO methods. Additionally, type 2 methods provide greater insight into function behavior without sacrificing accuracy or reproducibility.

Table 1. Abbreviated Results of This Work

Method	Evaluations (Best - Left) (Median -Right)		Error (L2 Norm)	Median Error (L2 Norm)
1A	45	5	6.77×10^{-3}	1.03
1B	100	68	6.74×10^{-3}	8.51×10^{-3}
2A	7	32	5.48×10^{-3}	9.85×10^{-3}
2B	6	9	5.52×10^{-3}	7.44×10^{-3}
2C	6	2	5.00×10^{-3}	1.55×10^{-1}

Conclusions and Outlook

High global warming potential hydrofluorocarbons (HFCs) must be phased out of use and replaced, but a lack of data on HFCs limits separation process design (Raabe, 2019). Molecular modeling force fields (FF) can generate missing HFC data, however, these models require accurately tuned parameters to generate useful predictions. Therefore, future work will examine applying method 2C to this challenging problem.

Acknowledgments

We thank the Shiflett group from The University of Kansas for their collaboration. We also acknowledge support from the Graduate Assistance in Areas of National Need fellowship from the Department of Education via grant number P200A210048, the National Science Foundation via Award numbers CBET-1917474 and EFRI-2029354, and the University of Notre Dame College of Engineering and Graduate School.

References

Befort, B. J., Defever, R. S., Maginn, E. J., & Dowling, A. W. (2022). Machine Learning-Enabled Optimization of Force Fields for Hydrofluorocarbons. Comput. Aided Chem. Eng., 49, 1249–1254.

Madin, O. C., & Shirts, M. R. (2022). Using Physical Property Surrogate Models to Perform Multi-Fidelity Global Optimization of Force Field Parameters. ChemRxiv. https://doi.org/10.26434/chemrxiv-2022-7bmzv-v2

Raabe, G. (2019). Molecular Simulation Studies on Refrigerants Past – Present – Future. Fluid Phase Equilib., 485, 190–198. https://doi.org/10.1016/J.FLUID.2018.12.022

* To whom all correspondence should be addressed: adowling@nd.edu

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

A MULTI-LANGUAGE PROCESS MONITORING FRAMEWORK FOR CHEMICAL INDUSTRY

Chengyu HAN, Jingzhi RAO, Jingde WANG, Wei SUN*
College of Chemical Engineering, Beijing University of Chemical Technology
Beijing, 100029, China

Keywords: Process monitoring, System architecture, Micro services

Scripting languages are often used by researchers to develop new monitoring algorithms. When applying these algorithms to industry, the algorithms need to be wrapped or rewritten using a compiled language. Deploying the new monitoring algorithm as a microservice and then communicating with a cluster of microservices written in other languages allows the new monitoring algorithm to be quickly applied to the industry.

DCS systems are commonly used in the chemical industry, and the large amount of historical data as well as the real-time data from DCS provides a good data base for data-driven process monitoring algorithms (Ji and Sun, 2022). Researchers often use scripting languages for algorithm development. Scripting languages do not need to be compiled, which allows algorithms to be iterated quickly. Newly developed algorithms need to be packaged into a complete system before they can be made available to users. Compiled languages such as Java, C#, and C++ are usually used for monitoring system development. The inconsistency of the language used in algorithm development and system development adds barriers to the process of applying new algorithms to industry.

When deployed using a microservice architecture, as opposed to a monolithic architecture, the microservices are independent of each other and can also be written in different languages (De Lauretis, 2019). In this work, a multi-language microservice monitoring framework is proposed, where algorithms written in Python language are wrapped as a microservice running independently. The input and storage of data are provided by other microservices.

System framework

The proposed framework is shown in Figure 1, and it could be divided into three layers. The View layer provides an interface for user interaction, including front-end interfaces such as Web pages and mobile apps. The microservices layer consists of business logic and core monitoring algorithms. Each individual functional module runs as a standalone microservice. Microservices can communicate with each other. The data source layer is an abstraction of the data sources in the factory. Data from different data sources are pre-processed and then sent to the Data Service microservice for processing in a unified way.

Based on the proposed monitoring framework, an intelligent coal chemical monitoring system is developed. All the algorithm parts are written in Python and run as independent microservices. Other parts of the system such as Gateways, Data Services and other microservices are still written in Java. The proposed framework avoids the problem of rewriting algorithms using compiled languages and accelerates the application of new monitoring algorithms in industry.

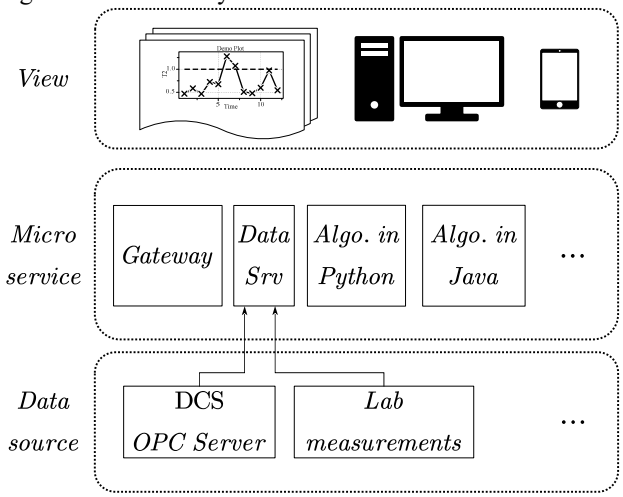

Figure 1. Monitoring Framework

Acknowledgments

Authors highly appreciate China Coal Energy Group Co., Ltd. for its financial support and all assistance on testing.

References

Ji, C., & Sun, W. (2022). A review on data-driven process monitoring methods: Characterization and mining of industrial data. Processes, 10(2), 335.

De Lauretis, L. (2019). From monolithic architecture to microservices architecture. In 2019 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW) (pp. 93-96). IEEE.

* Wei Sun; sunwei@mail.buct.edu.cn

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

ON THE IDENTIFIABILITY OF HYBRID MODELS

K.D. Jones and A.W. Dowling
Department of Chemical and Biomolecular Engineering, University of Notre Dame
Notre Dame, IN, 46556

Keywords: hybrid modeling, uncertainty quantification, optimization under uncertainty

In process systems engineering, it is becoming commonplace to utilize Kennedy & O'Hagan (KOH) hybrid models (Kennedy and O'Hagan, 2001) to minimize data collection and expedite decision-making. However, recovering the actual values of the model parameters is not guaranteed as the KOH model is generally unidentifiable. In this work, we present best practices for deploying KOH models in an engineering setting. Utilizing an illustrative isotherm model, we demonstrate identifiability with the traditional Bayesian hierarchical approach, frequentist maximum likelihood estimation (MLE), and frequentist sequential inference. We posit that well-defined Bayesian and frequentist sequential approaches are best suited for parameter estimation and reserve uninformative simultaneous approaches for prediction near the test data.

Background

In their seminal work, statisticians Kennedy and O'Hagan developed a framework for Bayesian calibration of computer models (Kennedy and O'Hagan, 2001). In the KOH model, an output y from the physical system $\zeta(\mathbf{x})$ is observed with measurement error $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ for N independent experimental controls $\mathbf{x}_\mathrm{n} \in \mathbb{R}^J, n = 1, \ldots, N$. The physical reality $\zeta(\mathbf{x})$ is approximated by a first-principles model $\eta(\mathbf{x}, \boldsymbol{\theta})$ where $\boldsymbol{\theta} \in \mathbb{R}^P$ is a vector of parameters. When $\eta(\cdot,\cdot)$ is misspecified in functional form, a model discrepancy function $\delta(\cdot)$ quantifies bias between the model and the data. That is,

$$y = \zeta(\mathbf{x}) + \epsilon = \eta(\mathbf{x}, \boldsymbol{\theta}) + \delta(\mathbf{x}) + \epsilon. \quad (1)$$

The modeler assumes $\delta(\cdot)$ follows a Gaussian process (GP) denoted $\delta(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), \sigma^2 K(\mathbf{x}, \mathbf{x}', \boldsymbol{\phi}))$. Briefly, a GP is characterized by a mean function $m(\mathbf{x})$ and covariance function $\sigma^2 K(\mathbf{x}, \mathbf{x}'; \boldsymbol{\phi})$ where $\boldsymbol{\phi}$ are hyperparameters of the correlation function $K(\cdot,\cdot)$ and σ^2 is the process variance. The goal of calibration is to find a set of model parameters $\boldsymbol{\omega} = [\boldsymbol{\theta}, \sigma^2, \boldsymbol{\varphi}, \sigma_\epsilon^2]$ that best reproduce the reality of the experimental data y.

Although model calibration is typically employed to improve the predictive performance of models, estimating the physically meaningful parameters may also have an intrinsic scientific value. In statistics, a model is said to be identifiable if realizing the actual values of its parameters is theoretically possible after obtaining infinitely many observations. Although technically well-defined with informative priors (Higdon et al., 2005), Eq. (1) is generally unidentifiable. Consider two different values $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ for $\boldsymbol{\theta}$ and write $\delta_1(\mathbf{x}) = \zeta(\mathbf{x}) - \eta(\mathbf{x}, \boldsymbol{\theta}_1)$ and $\delta_2(\mathbf{x}) = \zeta(\mathbf{x}) - \eta(\mathbf{x}, \boldsymbol{\theta}_2)$. Both sets $(\boldsymbol{\theta}_1, \delta_1)$ and $(\boldsymbol{\theta}_2, \delta_2)$ yield the same distribution for $\zeta(\mathbf{x})$ and are thus generally impossible to infer uniquely (Wong, Storlie, and Lee, 2017).

Results

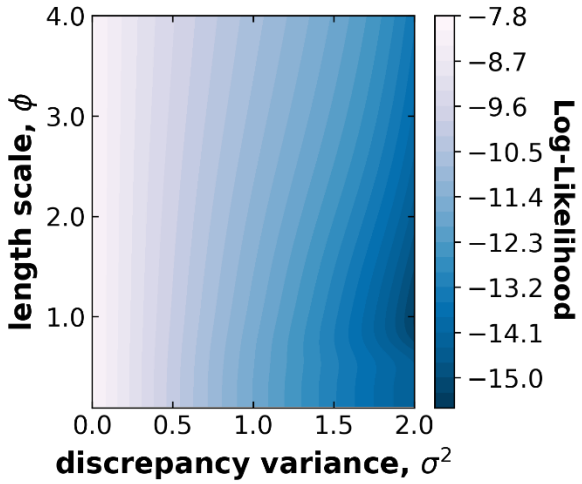

Figure 1. Multiple maxima of log-likelihood as a function of the discrepancy length scale and variance show identifiability challenge for frequentist MLE.

References

Kennedy, M.C., and O'Hagan, A. (2001). Bayesian calibration of computer models, *J R Stat Soc Series B Stat Methodol*, **63**(3): 425-464.

Higdon, D., Kennedy, M.C., Cavendish, J.C., Cafeo, J.A. and Ryne, R.D. (2005). Combining field data and computer simulations for calibration and prediction. *SIAM J Sci Comput*, **26**(2):448–466.

Wong, R.K.W., Storlie, C.B., and Lee, T.C.M. (2017). A Frequentist Approach to Computer Model Calibration. *J R Stat Soc Series B Stat Methodol.* **79**(2):635-648.

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

APPLICATION OF BIBLIOMETRIC DATA ANALYSIS SOFTWARE IN CORONARY HEART DISEASE RESEARCH

S. F. Y. Li *, S. Y. Lim

Department of Chemistry, National University of Singapore, 3 Science Drive 3, Singapore 117543

Keywords: Glycomics; Metabolomics; Lipidomics; Metallomics; Bibliometrics; Coronary Heart Disease

Coronary heart disease (CHD) continues to be the top cause of death worldwide despite notable advancements in disease diagnosis and treatment. The pursuit of omics studies to develop alternative/orthogonal biomarkers as well as to uncover unique insights into disease mechanisms is necessary due to the numerous practical hurdles currently present in clinical settings. The omics beyond the core dogma (OBCD; for example, metabolomics, lipidomics, glycomics, and metallomics) have obvious contributions and potential in CHD research, while being relatively young in comparison to the omics frontrunners (genomics, transcriptomics, and proteomics). With an emphasis on the more active domains of metabolomics and lipidomics, we characterized the global trends in publication/citation outputs, collaborations, and research hotspots regarding OBCD-CHD in this bibliometric analysis. There were insufficient publication records on the applications of metallomics and glycomics. Thus, we examined and supported their potential in CHD research, reviewed their applicability in general health/disease research, and recommended significant/ promising study directions.

Results and discussion

Here, a bibliometrics technique has been used to analyze the trends in OBCD applications in CHD research, and a qualitative literature analysis has added additional viewpoints. Characterizing their global trends is extremely relevant for growing and fast increasing omics domains including metabolomics, lipidomics, glycomics, and metallomics. Thus, this work serves as an exhaustive review for omics and CHD researchers, both established and new, to have a deeper grasp of the current situation and determine appropriate future research approaches.

The results suggest that interest in metabolomics and lipidomics is growing, but also that there are many research gaps that merit investigation and the formation of new study topics. Nutrition is a heavily impacted issue with a high frequency of citations in CHD-OBCD publications, with a current emphasis on the Mediterranean diet for the prevention of CVD and on gut microbiota. The transition from a reductionist (single-omics) to a global (multi-omics) approach is further highlighted by analysis of the research hotspots. Narrow discipline competence may no longer be sufficient to address such research difficulties as a result of this shift. Multidisciplinary partnerships and interdisciplinary education may become even more important. This is relevant not only to the generation of high-quality research data for the research team's exclusive use, but it will also be crucial for various groups to promote proactive and open data exchange between omics domains. While rarely used in CHD research, glycomics and metallomics have a lot of potential as shown by their applicability to other diseases and fields of study. The application of these omics methods should no longer be constrained by existing analytical platforms and procedures, and they should be incorporated into the broader multi-omics agenda.

Conclusion

This work demonstrates the use of a software tool to give a first and complete image of OBCD applications in CHD by summarizing information gathered both numerically and qualitatively, helping the establishment of future research directions.

* To whom all correspondence should be addressed; chmlifys@nus.edu.sg.

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

EXPLOITING HIGH-THROUGHPUT EXPERIMENTS IN BAYESIAN OPTIMIZATION

L. D. González and V. M. Zavala*
University of Wisconsin-Madison
Madison, WI 53706

Keywords: Bayesian optimization, high-throughput experiments, parallelization

High-throughput experimentation (HTE) has become a powerful tool for driving discovery in research domains that span large parameter spaces. These platforms permit experiments to be automated and run in parallel, decreasing costs and experiment wall-clock time. However, due to the large number of variables involved, determining the optimal conditions at which to perform experiments is often an infeasible task for a human. As a result, many HTE labs rely on screening, a process where experiments are performed at points along a grid of the parameter space to identify potential candidate solutions. While this approach has proven useful, coupling HTE with an effective design of experiments (DoE) strategy can greatly augment the capabilities of these platforms by reducing the time and materials required for discovery.

Bayesian optimization (BO) has proven to be an effective and sample-efficient machine learning (ML) algorithm (Snoek et al., 2015). Unfortunately, the sequential nature of BO makes it incompatible with DoE on HTE platforms. Ad-hoc modifications to the BO algorithm that would give it parallelization capabilities have been developed and demonstrated to provide better performance than sequential BO (Young et al., 2020). However, these approaches are limited in the degree of parallelization they can achieve and can increase the complexity of the algorithm.

In this work, we develop a set of methods to parallelize the BO algorithm and exploit HTE platforms. These strategies are centered around modifications to the optimization of the acquisition function (AF), the decision-making mechanism for BO. Our proposed paradigms, shown in Figure 1, focus on new and effective ways for partitioning the parameter space, allowing the AF to select multiple sampling points in tandem. We then assign a module to explore and optimize a specified objective in each partition. By processing the data collected into a central model that is shared with each module, we capture global rather than local trends which further improves the optimization routine. The methods we propose are scalable to any desired number of experiments, fully parallel, and designed to prevent redundant sampling.

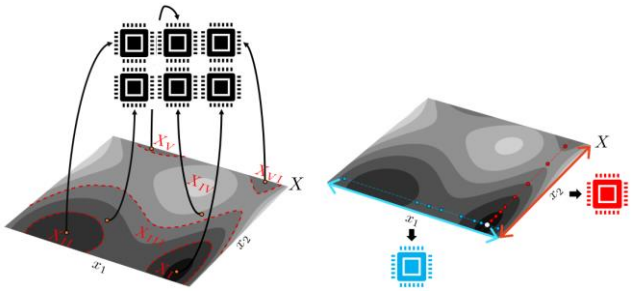

Figure 1. BO parallelization via level-set (left) and variable (right) partitioning.

We apply our approaches to the case study of a chemical reactor network where the aim is to select the temperature of each reactor that will minimize the yearly operating cost of the system. In addition to sequential BO, we also compare the performance of our parallel BO algorithm with existing parallelization techniques found in the literature such as Hyperspace (Young et al., 2018), NxMC (Snoek et al., 2012), and AF optimization over a set of exploratory parameters (Hutter et al., 2012).

References

Hutter, F., Hoos, H. H., and Leyton-Brown, K. (2012). Parallel algorithm configuration. In *International Conference on Learning and Intelligent Optimization*, pages 55–70. Springer.

Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems, volume 25*, pages 2951–2959. Curran Associates, Inc.

Snoek, J., Rippel, O., Swersky, K., Kiros, R., Satish, N., Sundaram, N., Patwary, M., Prabhat, M., and Adams, R. (2015). Scalable Bayesian optimization using deep neural networks. In *International Conference on Machine Learning*, pages 2171–2180, PMLR.

Young, M. T., Hinkle, J., Ramanathan, A., and Kannan, R. (2018). Hyperspace: Distributed Bayesian hyperparameter optimization. In *2018 30th International Symposium on Computer Architecture and High Performance Computing*, pages 339–347, IEEE.

Young, M. T., Hinkle, J. D., Kannan, R., and Ramanathan, A. (2020). Distributed Bayesian optimization of reinforcement learning algorithms. *Journal of the Parallel and Distributed Computing,* **139**:1, 43–52.

* To whom all correspondence should be addressed; victor.zavala@wisc.edu.

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

BLIND SOURCE SEPARATION FOR APPLICATIONS IN REAL-TIME IR SPECTROSCOPY MONITORING

S. H. Crouse,† S. Kocevska,† S. Noble,‡ R. Prasad,† A. M. Howe,‡ D. P. Lambert,‡ R. W. Rousseau,† M. A. Grover†*
†School of Chemical and Biomolecular Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA
‡Environmental, Materials, and Energy Sciences Directorate, Aiken, SC 29803, USA

Keywords: Blind Source Separation, Infrared Monitoring, Nuclear Waste

Overview

On-line infrared absorbance spectrometers enable rapid measurement of solution-phase molecular species. Many spectra-to-concentration models exist for spectral data, with some models able to handle overlapping spectral bands and nonlinearities. However, every model's accuracy is limited by the quality of training data used in model fitting. The process spectra (spectra taken in-line) of nuclear waste simulants at the Savannah River Site display incongruity from training spectra: the glycolate spectral signature in the training data does not match the glycolate signature in Savannah River National Laboratory process data. A novel blind source separation (BSS) algorithm is proposed that preprocesses spectral data so that process spectra more closely resemble training spectra, thereby improving model quantification accuracy when unexpected sources of variation appear in process spectra. The novel blind source separation preprocessing algorithm is shown to improve nitrate quantification from an R^2 of 0.934 to 0.988 and from 0.267 to 0.978 in two instances analyzing nuclear waste simulants from the Slurry Receipt Adjustment Tank and Slurry Mix Evaporator cycle at the Savannah River Site.

Results

Based on prior work done by Kocevska et al., spectra are preprocessed using BSS and then quantified using partial least squares regression.[1] In Figure 1a, the present source removal technique is shown to remove the contributions of glycolate (spectral signature at 1080 cm^{-1} and 1440 cm^{-1}). Figure 1b highlights model improvement, where prediction accuracy is improved with BSS preprocessing.

While the present BSS algorithm improves quantification, its utility as a process analysis tool extends beyond quantification accuracy improvements. In the present application, the detected change in glycolate peak signature prompted an investigation into solution chemistry, where pH was determined to be an unaccounted source of variation. Real time implementation of the BSS algorithm is additionally studied in this work.

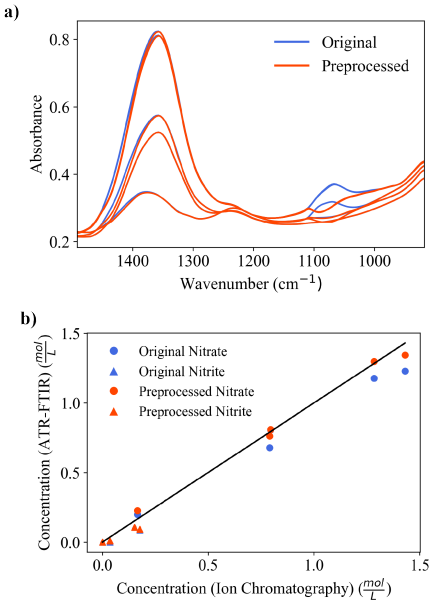

Figure 1. (a) Overlay of spectra before and after BSS-preprocessing; (b) parity plot comparing the concentration predictions for spectra with no preprocessing (blue) and BSS-preprocessed spectra (red).

Acknowledgement

This work was produced by Battelle Savannah River Alliance, LLC under Contract No. 89303321CEM000080 with the U.S. Department of Energy. This material is based upon work supported under an Integrated University Program Graduate Fellowship. Support by the U.S. Department of Energy under Cooperative Agreement DE-FC01-06EW07053, entitled "The Consortium for Risk Evaluation with Stakeholder Participation III," is also gratefully acknowledged.

References

Kocevska, S. *et al.* (2021) Spectroscopic Quantification of Target Species in a Complex Mixture Using Blind Source Separation and Partial Least-Squares Regression: A Case Study on Hanford Waste, *Industrial and Engineering Chemistry Research*, 60(27), pp. 9885–9896. doi: 10.1021/acs.iecr.1c01387.

*Corresponding Author Email: martha.grover@chbe.gatech.edu

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

INTEGRATED PLANNING AND SCHEDULING FOR CRUDE OIL OPERATIONS USING DATA-DRIVEN BILEVEL MULTI-FOLLOWER OPTIMIZATION

H. Nikkhah[1], V.M. Charitopoulos[2], S. Avraamidou[3] and B. Beykal[1*]

[1]Department of Chemical & Biomolecular Engineering, Center for Clean Energy Engineering, University of Connecticut, Storrs CT, 06269, USA

[2]Department of Chemical Engineering, Sargent Centre for Process Systems Engineering, University College London, Torrington Place, London WC1E 7JE, UK

[3]Department of Chemical & Biological Engineering, University of Wisconsin-Madison, Madison WI, 53706, USA

Keywords: Data-driven Optimization, Mixed-Integer Nonlinear Bilevel Programming, Planning, Scheduling.

Planning and scheduling are two distinct, yet interconnected elements in supply chain management and enterprise-wide optimization. The former evaluates market factors (i.e., product demands) to establish production goals, while the latter defines how tasks will be sequenced and which units will be responsible for completing them to achieve the production targets. The interlink between these different layers of the supply chain can be mathematically realized with bilevel multi-follower optimization (Eq.1), where one optimization problem (leader-planning) is bounded by other optimization problems (followers-scheduling). This formulation, however, introduces significant algorithmic challenges, including NP-hardness, nonconvexity, and discontinuity. While certain challenges can be overcome for general bi-level formulations (i.e., reducing the original formulation to a single-level problem using KKT conditions), this is not possible for integrated planning and scheduling because mixed-integer scheduling formulations are restrictive. Hence, an algorithmic strategy that can handle these challenges while guaranteeing feasibility is necessary.

$$\begin{aligned} &\min\ Total\ cost\ of\ planning \\ &\quad S.t.\ \ Inventory\ and\ Balalnce\ equations \\ &\qquad \min\ \ Production\ Cost\ for\ Each\ Planning\ Period \\ &\qquad\ S.t.\ \ \ Scheduling\ Constraints \end{aligned} \tag{1}$$

We tackle the mathematical difficulties of planning problems constrained with mixed-integer nonlinear (MINLP) scheduling levels using data-driven optimization. We specifically use the DOMINO framework that performs a Design of Experiments-based sampling procedure to determine production targets at the planning level and then solves the scheduling level for global optimality at those targets across the full planning period (Beykal et al., 2020). The optimization is then carried out by feeding the input samples into preexisting subroutines and using the output optimality information to identify promising regions.

In this work, the DOMINO framework is applied for the solution of an MINLP crude oil operation case study (Jia et al., 2003), which covers unloading, transfers, and charging of oil mixes to crude-oil distillation units (CDUs) and is subject to nonlinear component balances and constraints. We use a 15-day planning period for the demand of 2 oil mixes from 2 charging tanks. The results for the second oil mix are provided in Figure 1, where product demands are satisfied for the entire planning period with scheduling levels solved to global optimality. This work represents how DOMINO can be effective in solving these types of problems. This research is supported by the NIEHS Superfund Research Program P42 ES027704.

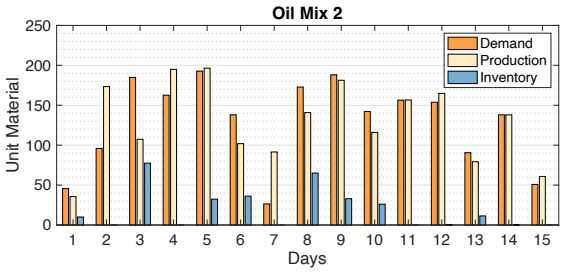

Figure 1. Demand, production, and inventory profiles of oil mix 2 for a 15-day planning period.

References

Beykal, B., Avraamidou, S. Pistikopoulos, I.P.E, Onel, M., Pistikopoulos, E.N (2020). DOMINO: Data-driven Optimization of bi-level Mixed-Integer NOnlinear Problems, *J. Glob. Optim.* **78**, 1-36.

Jia, Z., Ierapetritou, M., Kelly, J.D. (2003). Refinery Short-Term Scheduling Using Continuous Time Formulation: Crude-Oil Operations, *Ind. Eng. Chem. Res.* **42**:13, 3085-3097.

* To whom all correspondence should be addressed; beykal@uconn.edu.

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

MODEL IDENTIFICATION FROM HISTORICAL DATA USING DATA MINING TECHNIQUES: INDUSTRIAL APPLICATION

Ammar S. Bakhurji *, Rohit Patwardhan, and Manaf A. Alahmadi
Saudi Aramco,
Dhahran, KSA 31331

Keywords: Model identification, data mining, historical data

System identification from process data is a critical task in process control, where accurate models of the underlying process are necessary for effective control and optimization. While conventional step testing based identification of system models is widely used, it is time-consuming and requires significant effort, making it impractical for control applications requiring periodic model updates. Therefore, automation of this task using data mining techniques can provide significant benefits in terms of maintaining the application performance.

In this work, we discuss the application of automated data mining techniques for system identification in process modeling and control. The work involves the development of algorithms that can pre-process industrial real-time data using machine learning and statistical methods to automatically identify information-rich regions where sufficient excitation is available to identify the underlying process dynamics for SISO, MISO, MIMO systems, or steady-state models.

Operating data can often be rich in information and this can be exploited for modeling purposes. The automation of system identification and information-rich region identification can significantly reduce the time required to develop or maintain process models used in Model Predictive Controllers (MPC), re-tune PID controllers, and inferential modeling. Additionally, the work has the potential to improve the accuracy of process models leading to enhanced process efficiency, reduced energy consumption, and improved product quality.

Motivation

Model-based control applications are widely deployed in industry. These appear in the form of model predictive controllers (MPC) to handle complex multivariable processes or in simpler forms such as PID control where a model is used to determine the ideal tuning parameters. Even inferentials or soft sensors are essentially a type of model used to estimate infrequently measured process variables of importance.

For example, at Saudi Aramco alone there are 100s of MPCs and inferentials and well over 10,000 PID controllers (Patwardhan et al. 2019). Maintaining such a large number of applications requires automated tools to monitor, diagnose and improve the closed loop performance periodically (Forbes et al. 2015). The most time-consuming step in any modeling effort is the data preprocessing step which can easily consuming more than 50% of the total effort. Data preprocessing involves bad data removal and replacement, outlier treatment, identification of appropriate time ranges and variables. Most MPC and PID tuning software will depend on experienced engineers to carry out this task manually.

Results & Discussion

The proposed techniques are applied on both SISO and multivariable use cases. The SISO application for identifying information rich windows in historical data and then estimating new models and tuning parameters from this data.

The multivariable applications focus on identifying rich data sets for estimating MISO or MIMO models. Examples of re-identifying MPC models and inferentials are shared.

References

Patwardhan R., Hamadah H. A., Patel, K. M., Hafiz, R. H., Al-Gwaiz, M. M., Applications of Advanced Analytics at Saudi Aramco: A Practitioners' Perspective, *Industrial & Engineering Chemistry Research*, 2019.

Forbes, M. G., Patwardhan, R. S., Hamadah, H., ·Bhushan Gopaluni, R., "Model Predictive Control in Industry: Challenges and Opportunities", *In Proceedings of Advanced Control of Chemical Processes (ADCHEM)*, Whistler, Canada, 2015

* ammar.bakhurji@aramco.com

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

A NOVEL TRAINING METHODOLOGY FOR HYBRID MODELS WITH TIME-VARYING DISTURBANCES: AN APPLICATION TO BUILDINGS

P. Krishna, L. dela Rosa, and M.J. Ellis*
University of California, Davis
Davis, CA 95616

Keywords: Model predictive control, Building systems, Hybrid models, Three-step training methodology

The combined energy consumption of the residential and commercial building sectors in the United States has been increasing at about 1.3% per year over the past decade (EIA, 2023), making efficient building operations more crucial than ever (Alvarez and Molnar, 2021). Studies have shown that model predictive control (MPC) can improve building energy efficiency (Afram et al., 2017). However, MPC is still not widely adopted in buildings due to the challenges in developing and training a control-oriented model (Ellis, 2021). This task is challenging because it results in a coupled state/disturbance parameter estimation problem due to the presence of unmeasured time-varying heat disturbances brought on by building occupancy, solar radiation, and electrical devices.

Hybrid models, combining physics-based and data-driven models, have emerged as a promising solution for control-oriented building thermal modeling. In our previous study (Krishna and Ellis, 2023), our approach integrated a parameterized low-order physics-based thermal resistance-capacitance (RC) network model, derived from first principles, with a feedforward neural network used for forecasting the unmeasured heat disturbances. This work proposed a method for simultaneously training both model parameters. In an approach by (Kumar et al., 2023), a physics-based model was developed to capture the underlying thermal dynamics of the building space and a neural network to forecast the unmeasured disturbances. However, in their approach, the two models were trained separately. Simultaneously training the hybrid model parameters is beneficial for minimizing overall prediction errors and can potentially avoid compounding prediction errors resulting from training models separately. However, due to the non-convex nature of training problems, the estimation of model parameters is subject to the influence of initial parameter estimates provided to the solver.

In the case of building thermal models, limited data is available to estimate the model parameters owing to practical considerations (i.e., minimize potential comfort violations to building occupants). In this context, second-order or quasi-Newton methods like the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm can be considered as an alternative to first order methods. To this end, a novel heuristic three-step training methodology is proposed to improve the prediction of the building thermal dynamics with a lower computational time compared to using a single solver. The proposed approach combines two steps of gradient descent and one step of BFGS, taking advantage of BFGS while addressing the necessity to find suitable initial parameter estimates to provide to the solver. A model validation approach is also proposed as part of the training methodology. The simulation results highlight the effectiveness of the proposed three-step training methodology for the hybrid building model in predicting the building thermal dynamics compared to using a single solver. Specifically, the three-step training methodology achieves a higher accuracy with fewer iterations.

References

U.S. Energy Information Administration (2023). “March 2023 monthly energy review”, Technical report.

Afram, A., Janabi-Sharifi, F., Fung, A. S., and Raahemifar, K. (2017). Artificial neural network (ANN) based model predictive control (MPC) and optimization of HVAC systems: A state of the art review and case study of a residential HVAC system, *Energy and Buildings* **141**, 96–113.

Alvarez, C. F., Molnar, G. (2021). “What is behind soaring energy prices and what happens next?”, Technical Report, International Energy Agency.

Ellis, M. J. (2021). “Machine Learning Enhanced Grey-Box Modeling for Building Thermal Modeling,” In Proceedings of the American Control Conference, New Orleans LA USA, pp. 3927-3932, 25-28 May 2021.

Krishna, P., and Ellis, M. J. (2023). "Control-Oriented Hybrid Modeling Framework for Building Thermal Modeling," in Energy Systems and Processes: Recent Advances in Design and Control, edited by Mingheng Li (AIP Publishing, Melville, New York), Chapter 9, pp. 9-1–9-28.

Kumar, P., Rawlings, J. B., Wenzel, M. J., Risbeck, M. J. (2023). Grey-box model and neural network disturbance predictor identification for economic MPC in building energy systems, *Energy and Buildings* **286**, 112936.

* To whom all correspondence should be addressed: mjellis@ucdavis.edu

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

MULTIPERIOD OPTIMIZATION OF INTEGRATED ENERGY SYSTEMS WITH MACHINE LEARNING SURROGATES TO PREDICT MARKET IMPACTS

X. Chen[1*], R. T. Gooty[2], D. Guittet[3], A. W. Dowling[1]
[1] Department of Chemical and Biomolecular Engineering, University of Notre Dame
Notre Dame, IN 46556
[2] National Energy Technology Laboratory, Pittsburgh, PA 15236
[3] National Renewable Energy Laboratory, Golden, CO 80401

Keywords: Integrated Energy System, Surrogate Model, Conceptual Design

Integrated energy systems (IES) explore synergies between different energy resources and technologies to improve the efficiency of energy systems. One of the widely used assumptions in the IES design and operation optimization is the price-taker. The price-taker assumes that the electricity grid is an "infinite" bus and can take and give any amount of electricity without affecting the market price. However, this assumption ignores the interaction between IES and the energy market, which results in misleading analyses and conclusions of IES design and operation.

ML Surrogate-assisted Optimization

We present a new multiperiod design and operations co-optimization formulation that explicitly considers IES/market interactions using machine learning surrogate models (e.g., time-series clustering, neural networks). Our framework is organized into four steps.

Step 1: Perform Production Cost Model (PCM) simulations to assemble initial training data.

First, we sampled different combinations of design variables from the domain of the IES design space. For each sample, we fix the IES design decisions and simulate the annual electricity market using the production cost model (PCM), Prescient. PCM solves unit commitment and economic dispatch problems in a rolling horizon fashion and yields detailed day-ahead and real-time dispatch profiles of all generators and the day-ahead and real-time locational marginal prices (LMPs) at all nodes.

Step 2: Train surrogate models to predict market outcomes as a function of IES characteristics.

Next, we identify a few representative dispatch profiles for the IES using time-series clustering of daily real-time dispatch profiles obtained from the library of annual PCM simulations. We then train a ML surrogate model to predict the weight/frequency of each representative dispatch profile as a function of the IES design variables. We also train ML surrogate models to predict the total electricity revenue (determined using the nodal LMPs) as a function of the IES design. Together, these surrogate models capture the effect that the IES has on the overall electric grid/market.

Step 3: Solve conceptual design optimization problems with embedded market surrogates.

We then formulate and solve IES conceptual design optimization problem using the surrogate models for both revenue and dispatch and solve it to obtain the optimal design of the IES. The optimization model is mixed with ML models and OMLT is employed to represent neural networks within the Pyomo optimization environment.

Step 4: Verify results via multiscale PCM simulation.

Finally, we use the multiscale simulation framework to simulate the sub-hourly operation of the optimal IES design over an entire year.

Results

Using this framework, we assess the economic benefits of retrofitting an existing baseload nuclear generator with a low-temperature electrolysis unit. The electrolyzer enables the nuclear generator to ramp down the power output to the electricity market and participate in the hydrogen market during periods of low electricity demand. We demonstrate that the price-taker approach yields a sub-optimal size of the electrolyzer since it substantially underestimates the revenue generated from the electricity market, and thereby illustrates the need for the proposed workflow. Compared with the historical LMP signals, we observed that after operating the nuclear generator with electrolysis unit reduces the near zero LMP frequency by 5%.

References

Gao, X.; Knueven, B.; Siirola, J. D.; Miller, D. C.; Dowling, A. W. Applied Energy 2022, 316, 119017.

Watson, J. P.; Knueven, B.; Concepcion, R.; Melander, D.; Short, A.; Zhang, P.; Woodruff, D.; USDOE Prescient, Version 1.0, 2020.

Ceccon, F., Jalving, J., Haddad, J., Thebelt, A., Tsay, C., Laird, C. D., & Misener, R. (2022). The Journal of Machine Learning Research, 23(1), 15829-15836.

Frew, B.; Levie, D.; Richards, J.; Desai, J.; Ruth, M. Applied Energy 2023, 329, 120184.

Martinek, J.; Jorgenson, J.; Mehos, M.; Denholm, P. Applied energy 2018, 231, 854–865.

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

IMPROVING SYSTEM IDENTIFICATION OF KINETIC NETWORKS USING NEURAL STOCHASTIC DIFFERENTIAL EQUATIONS

Krystian Ganko[*, 1], Nathan M. Stover[*, 1], Utkarsh[2], Richard D. Braatz[1], and Christopher Rackauckas[**, 2, 3]
[1] MIT Department of Chemical Engineering, Cambridge, MA 02139
[2] MIT CSAIL, Cambridge, MA 02139
[3] JuliaHub, Inc., Cambridge, MA 02139 and Pumas-AI, Inc., Centreville, VA 20120

Keywords: Scientific Machine Learning, Structural Identifiability, MINLP, Kinetic Networks

Many chemical and biological systems are experimentally observed to have both deterministic drift and noisy character. State-space stochastic differential equations (SDEs) model this noise, to more realistically describe the stochastic dynamics compared to ordinary differential equation (ODE) descriptions that employ the continuum approximation (Turner et al., 2004). Moreover, for physics that are poorly understood, neural networks may be embedded to increase model expressivity (Raissi et al., 2019). However, modelling noise terms requires additional parametric structures, which frequently introduces structural identifiability issues—i.e., the non-uniqueness of parameter sets that produce similarly optimal fits to experimental data—relative to ODE models. As such, the capacity of parameter-identified SDE models to predict system dynamics, as well as for any neural networks embedded in such models to learn the dynamics, is inherently limited.

This work demonstrates that—under assumptions allowing the normal distribution approximation to the Poisson processes governing a larger ensemble of reacting species (Gillespie, 2000)—structural identifiability and generalizability of embedded neural networks are substantially improved for compartment models of reacting chemical networks. This improvement simplifies the Chemical Master Equation description of a kinetic network (e.g., see Jang et al., 2016 and citations therein) to an SDE in the form of the Chemical Langevin Equation (CLE) (Browning et al., 2020). Figure 1 depicts this simplification, and positions the CLE with respect to the Chemical Master Equation and often-used mass-action ODE system description for such systems.

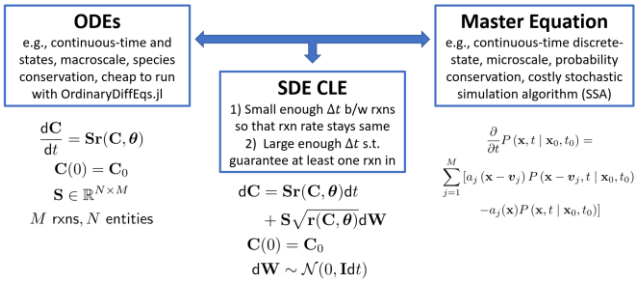

Figure 1. Formulation of the SDE CLE for kinetic networks.

In full, this approach enables the exploration of more expensive experimental spaces involving both fully and partially observable reaction networks.

References

Turner, T. E., Schnell, S. & Burrage, K. (2004). Stochastic approaches for modelling in vivo reactions. *Computational Biology and Chemistry*, 28, 165–178.

Raissi, M., Perdikaris, P. & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378, 686–707.

Gillespie, D. T. (2000). The chemical Langevin equation. *The Journal of Chemical Physics*, 113, 297–306.

Jang, H., Kim, K.-K.K., Braatz, R. D., Gopaluni, R. B. & Lee, J. H. (2016). Regularized maximum likelihood estimation of sparse stochastic monomolecular biochemical reaction networks. *Computers & Chemical Engineering*, 90, 111–120.

Browning, A. P., Warne, D. J., Burrage, K., Baker, R. E. & Simpson, M. J. (2020). Identifiability analysis for stochastic differential equation models in systems biology. *Journal of The Royal Society Interface*, 17(173), 1–23.

[*] Authors contributed equally to this work.

[**] To whom all correspondence should be addressed; crackauc@mit.edu.

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

MACHINE LEARNING FOR SENSOR DATA INTEGRATION AND QUANTITATIVE EARLY DETECTION OF PLANT DISEASES

Sina Jamalzadegan, Giwon Lee, and Qingshan Wei (qwei3@ncsu.edu)
North Carolina State University
Raleigh, NC 27606

Keywords: Wearable Sensor, Plant Health, Machine Learning, Smart Agriculture, Biosensor

Wearable plant sensors hold tremendous potential for smart agriculture. We report a lower leaf surface-attached multimodal wearable sensor for continuous monitoring of plant physiology by tracking both biochemical and biophysical signals of the plant and its microenvironment. Sensors for detecting volatile organic compounds (VOCs), temperature, and humidity are integrated into a single platform. A machine learning model was developed to analyze multichannel sensor data for quantitative detection of tomato spotted wilt virus as early as 4 days after inoculation. The model also evaluates different sensor combinations for early disease detection and predicts that minimally three sensors are required including the VOC sensors.

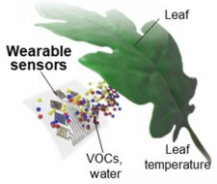

Figure 1. Schematic representation of multimodal wearable sensor

Machine Learning Model and Results

To quantitatively assess our multimodal sensors for the early detection of pathogens, an unsupervised machine learning approach based on principal components analysis (PCA) was used to analyze the real-time sensor data. For the demonstration, we used the tomato spotted wilt virus (TSWV) inoculation data as an example. The multichannel wearable sensor data from the same plant was first divided into different days (e.g., days 0, 1, 2, 3, etc.). Day 0 data were used as the healthy control and compared to other days. Data from different days were clustered by PCA with reduced data dimensions. Then, the centroid and Euclidean distance between two centroids of clusters (two different days) were calculated. The separation of the clusters was quantitatively assessed by a parameter called "discriminability" (D), as defined by equation (1) where D, E, and R denote discriminability, Euclidean distance, and radius (or SD) of the cluster, respectively. Figure 2 represents the best sensor composition for each number of sensors. VOC_C1, C2, F1, F2, H, and T denote four different types of VOC sensors, leaf surface relative humidity sensor, and leaf temperature sensor, respectively.

$$D = E - (R_{STD,1} + R_{STD,2}) \quad (1)$$

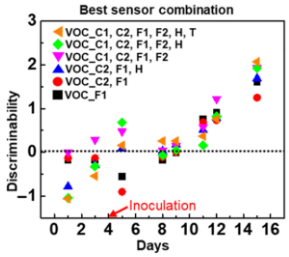

Figure 2. Discriminability with the best sensor composition for each number of sensors.

According to the discriminability values, a minimum of three sensors (VOC_C2, VOC_F1, and H) is needed for the early detection of TSWV. The results suggest that for effective disease detection, the biochemical VOC sensor is probably the most important sensor that is needed in each sensor combination; in addition, the leaf surface humidity sensor works slightly more effectively than the leaf temperature sensor in disease detection. Such a machine learning analysis pipeline can help find the most impactful sensor (and sensor combination) for a particular application and potentially reduce the total number of redundant sensors, which would be particularly useful to reduce sensor cost while maintaining sensor performance.

Acknowledgments

We gratefully acknowledge the funding support from the NCSU Game-Changing Research Incentive Program for the Plant Science Initiative (GRIP4PSI), USDA (no. 2019-67030-29311), USDA APHIS Farm Bill grant (no. 3.0096), and NSF (nos. 1728370 and 2134664).

References

Lee, G. et al. (2023). Abaxial leaf surface-mounted multimodal wearable sensor for continuous plant physiology monitoring, *Sci. Adv.***9**:15, eade2232.

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

APPLICATIONS OF MACHINE LEARNING IN PETROLEUM SYSTEMS: OPTIMIZATION UNDER UNCERTAINTY

D.J Robbins *
Chevron Technology Center
San Ramon, CA 94583

Keywords: machine learning, deep learning, physics informed learning, Bayesian methods, uncertainty

Optimization of the productivity of petroleum systems has been an ongoing endeavor for the oil and gas industry for many decades. Historically, methods have progressed from empirical and semi-empirical extrapolations through to simplified theory of the partial differential equations describing diffusion and mass transport and full numerical solvers for the continuity equations. These tools have been developed and modified to account for some of the uncertainty faced by petroleum engineers and earth scientists for both the input data describing the petroleum system and the production rates themselves.

The provisioning of data for petroleum systems alongside the exponential growth in computational power and complexity has greatly improved within this time frame, which has led to more data-rich estimates of well and reservoir productivity. It is now possible for engineers and geologists to make multiple, rapid estimates of potential reservoir productivity with several tools in parallel in order to assess uncertainty of the resource in addition to the impact of pseudo-measurable uncertainties of the inputs. This facilitates more informed data collection, surveillance and intervention operations.

In this talk, application of contemporary machine learning methods to production data will be demonstrated, from simple decision-tree based algorithms through to deep learning and physics-informed methods. We discuss how there is no ‘one sized fits all’ solution, but rather the necessity for the data scientist and petroleum engineer to understand the dominant physics of the system to select the most appropriate method. Finally, we will discuss the application of Bayesian and physics-informed neural operator methods to optimize under uncertainty to generate estimates of well productivity and remaining resource.

* To whom all correspondence should be addressed; email address of corresponding author.

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

BAYESIAN NEURAL NETWORK FOR PREDICTION OF PROTEIN–LIGAND DISSOCIATION KINETICS

Y. Zhao, Q. Liu *, L. Zhang, J. Du, and Q. Meng
State Key Laboratory of Fine Chemical, Frontiers Science Center for Smart Materials Oriented Chemical Engineering, Institute of Chemical Process Systems Engineering, School of Chemical Engineering, Dalian University of Technology, Dalian 116024, China

Keywords: Small molecule drug; Brønsted–Evans–Polanyi relation; Bayesian neural network; Dissociation kinetic constant

Small molecule drugs are indispensable to human health and social development. Binding kinetic properties, particularly the dissociation rate constant (k_{off}), are crucial factors affecting drug potency and have become critical indexes in discovering better- or first-in-class drugs. In kinetic measurements, the application of conventional methods is challenging due to the fact that k_{off} is dependent on the Gibbs free energies of intermediate transition states. Moreover, it generally takes several minutes or even days to observe the experimental response of k_{off} using conventional analysis methods such as the X-ray crystallography or the nuclear magnetic resonance. For these reasons, computational methods are paid more and more attention to provide insights into the protein–ligand binding/unbinding processes. Currently, the physics-driven molecular dynamics simulations and data-driven surrogate modeling methods have been used to develop single(multi)-protein models to predict k_{off}. Even with these exploratory studies, it remains challenging to develop binding kinetic models. For example, the high computational cost of the molecular dynamics simulations hinder the high-throughput screening of drug candidates. For the surrogate modeling methods, on the one hand, although the multi-protein models have a wide application range on different proteins, the generalization ability of the current models is not acceptable. On the other hand, the prediction accuracy of single-protein models is unstable due to their unclear application scenarios.

This work focuses on addressing the prediction stability issue of single-protein models by developing a Bayesian Neural Network (BNN) -based binding kinetic surrogate model for interval estimations of k_{off} using the descriptors of the van der Waals and electrostatic interaction energies of protein–ligand complexes. A case study involving HSP90 inhibitors are presented to highlight the feasibility and effectiveness of the proposed binding kinetic model. The HSP90 binding kinetic data is collected from Ganotra and Wade (2018). The samples in the dataset are ranked from low to high -logk_{off} values and every fifth sample in the list is selected for the test dataset. The BNN model is composed of an input layer, a hidden layer, and an output layer, where the optimizer, loss function, and activation function are Adam, the negative evidence lower bound, and Sigmoid, respectively. Its optimal hyper-parameters are determined via the knowledge-based trial and leave-one-out cross-validation methods. The commonly used Partial Least Squares Regression (PLSR) model is also developed and compared with our BNN model. The regression results of the HSP90 dataset are given in Table 1.

Table 1 The regression results of the BNN and PLSR models for the HSP90 dataset

Method	R^2_{test}	MAE_{test}
PLSR	0.908	0.232
BNN	0.947	0.184

As shown in Table 1, the BNN model developed in this work exhibits a great power to predict the -logk_{off} of HSP90 inhibitors with R^2_{test}=0.947 and MAE_{test}=0.184, and outperforms the commonly used PLSR model with R^2_{test}=0.908 and MAE_{test}=0.232. In addition to the interval prediction ability of the BNN method, another possible reason leading to the outstanding regression results maybe that the relationship between -logk_{off} and the van der Waals/electrostatic interactions is complex and more likely to be nonlinear. In our future work, the binding kinetic model will be integrated with the *de novo* drug design methods to fast design novel drug candidates with well-performed dissociation kinetics.

References

Ganotra, G. K., Wade, R. C. (2018) Prediction of Drug–Target Binding Kinetics by Comparative Binding Energy Analysis, *Acs Med. Chem. Lett.* **9**:11, 1134-1139.

* To whom all correspondence should be addressed; liuqilei@dlut.edu.cn.

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

INTEGRATING SMART MANUFACTURING TECHNIQUES INTO UNDERGRADUATE EDUCATION

Mrunal Sontakke[1]*, Lucky E. Yerimah[1], Andreas Rebmann[1], Sambit Ghosh[1], Craig Dory[2], Ronald Hedden[1], B Wayne Bequette[1]
[1]Dept. of Chemical and Biological Engineering, Rensselaer Polytechnic Institute, Troy, NY - 12180
[2]CESMII Smart Manufacturing Innovation Center, Rensselaer Polytechnic Institute, Troy, NY – 12180

Keywords: Smart Education, Machine Learning, Human-Machine Interface, Heat Exchanger

The process systems domain is undergoing the fourth industrial revolution, implementing Smart Manufacturing (SM) techniques with technological advancements like the Industrial Internet of Things, data storage and computing, Artificial Intelligence, and cloud computing (Yerimah et al., 2022). In parallel, in the field of machine learning and data analysis, data-based modeling has been expanding. These advancements have helped industries leap into digitization and optimized production techniques. However, SM methodologies are yet to be incorporated into process control applications in undergraduate chemical engineering education (Bequette, 2019a).

The work presented here takes a crucial step toward bridging the industry-academia gap by enabling students to conduct better experiments using an experiment assistant and access to larger datasets to catalyze their interest in developing predictive models for fault detection. The demonstration is implemented on a shell and tube heat exchanger using the Smart Manufacturing Innovation Platform (SMIP) (*SM Innovation Platform - CESMII*, 2023).

Introduction

Experiment Assistant

For students to collect accurate data and perform better experiments, a human-machine interface (HMI) is designed to provide them with all the valuable information and guidance while conducting experiments. This aims to acclimatize and train them for a control room they will encounter in the future. The display caters to the needs of the students, with prompts for incorrect steps taken during the experiment, faults detected by a fault detection algorithm, and suggestions for necessary actions.

The developed HMI can be used to study and conduct lab-scale studies to understand human-machine interaction and the Human-in-the-loop aspect of decision-making (Bequette, 2019b; Ghosh & Bequette, 2020).

Fault detection

A fully connected, convolutional and recurrent neural network was trained and developed using data collected from the heat exchanger. The models are used to develop tutorials to demonstrate the use of the framework. Students can train and test their models with smooth implementation with real-world data.

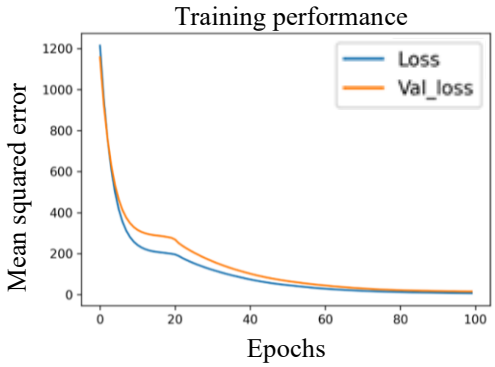

Figure 1: Fault detection model training

Acknowledgements

This material is based upon work supported by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under the CESMII Award Number DE-EE0007613.

References

Bequette, B. W. (2019a). Process control practice and education: Past, present and future. Computers & Chemical Engineering, 128, 538–556. https://doi.org/10.1016/J.COMPCHEMENG.2019.06.011

Bequette, B. W. (2019b). 110th Anniversary: Commentary: The Smart Human in Smart Manufacturing. Ind. Eng. Chem. Res., 58(42), 19317–19321. https://doi.org/10.1021/ACS.IECR.9B03544

Ghosh, S., & Bequette, B. W. (2020). Process Systems Engineering and the Human-in-The-Loop: The Smart Control Room. Ind. Eng. Chem. Res. 59(6), 2422–2429. https://doi.org/10.1021/ACS.IECR.9B04739

SM Innovation Platform - CESMII. (2023). https://www.cesmii.org/platform/smip/ (Date accessed: 04/09/2023)

Yerimah, L. E., Ghosh, S., Wang, Y., Cao, Y., Flores-Cerrillo, J., & Bequette, B. W. (2022). Process prediction and detection of faults using probabilistic bidirectional recurrent neural networks on real plant data. J. Adv. Manuf. Process., 4(4), e10124. https://doi.org/10.1002/AMP2.10124

*mrunalsontakke@gmail.com

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

OPTIMIZATION IN ENGINEERING WITH EMBEDDED LINEAR MODEL DECISION TREES

Bashar L. Ammari [a,*], Emma S. Johnson [b], Georgia Stinchfield [a], Taehun Kim [c], Michael Bynum [b], William E. Hart [b], Joshua Pulsipher [a], Carl D. Laird [a]

[a] Carnegie Mellon University, Pittsburgh, PA 15213
[b] Sandia National Laboratories, Albuquerque, NM 87123
[c] Georgia Institute of Technology, Atlanta, GA 30332

Keywords: Mixed-integer Programming, Decision Trees, Flexibility Analysis, Heat Exchanger Design

Machine learning (ML) models have become increasingly popular due to their ability to accurately represent complex functional relationships. While many resources have been developed to streamline training and evaluation of these models, new research efforts have focused on embedding trained ML models into optimization problems. This is of key interest in several research communities as the ability to optimize over ML surrogates allows for solution of the inverse problem (solving for inputs given objectives or constraints on outputs). Mixed-integer programming formulations for embedding ML models such as neural networks and tree ensembles into optimization problems exist in the literature. Piecewise-linear ML surrogates that can be represented within mixed-integer linear programs (MILPs) have been especially favored due to advances in MILP solvers. In this work we present several mixed-integer programming formulations of linear model decision trees and demonstrate their utility as surrogates in engineering applications.

Linear model decision trees differ from standard decision trees in that they return linear functions at the leaf nodes. They are advantageous in their ability to represent discontinuous functions and can produce smaller tree representations with reduced error. This work expands on the work of Mistry et al. (2021) and utilizes Generalized Disjunctive Programming (GDP) to formulate linear model decision trees as mixed-integer quadratic constrained programs (MIQCPs) and MILPs. We compare the performance of these formulations using a process family design problem (Stinchfield et al., 2022). In addition, we exhibit their use case on a flexibility index problem and an optimal heat exchanger design problem.

Flexibility Analysis

In the flexibility index problem, we convert the traditional multi-level optimization problem (Swaney and Grossmann, 1985) into a single level optimization by approximating the inner control problem with an ML surrogate. Results show that linear model trees have the potential to provide a better approximation than neural network surrogates due to their ability to represent discontinuous functions. Furthermore, by approximating the optimal control policy with an ML surrogate, we obtain a conservative estimate of the flexibility index, regardless of nonconvex constraints. This may be advantageous over current solution methods such as the active set strategy or vertex enumeration.

Heat Exchanger Design

In this case study, we consider the optimal design of a molten salt and steam heat exchanger. We take traditional black-box thermodynamic property calculations (such as those utilized in the IDAES framework) and replace them with linear model decision tree surrogates. This allows for use of deterministic global optimization methods through solvers such as BARON. We show that linear model decision trees can achieve comparable solutions using BARON with the solvers within IDAES (Lee et al., 2021). Lastly, we comment on the interpretability of linear model decision trees with respect to the physics of the problem.

References

Lee, A., Ghouse, J. H., Eslick, J. C., Laird, C. D., Siirola, J. D., Zamarripa, M. A., Gunter, D.,Shinn, J. H., Dowling, A. W., Bhattacharyya, D., Biegler, L. T., Burgard, A. P., and Miller, D. C. (2021). "The IDAES process modeling framework and model library—flexibility for process simulation and optimization." *Journal of Advanced Manufacturing and Processing*, 3.

Mistry, M., Letsios, D., Krennrich, G., Lee, R. M., and Misener, R. (2021). "Mixed-integer convex nonlinear optimization with gradient-boosted trees embedded." *INFORMS Journal on Computing*, 33:1103–1119.

Stinchfield, G., Biegler, L. T., Eslick, J. C., Jacobson, C., Miller, D. C., Siirola, J. D., Zamarripa, M. A., Zhang, C., Zhang, Q., and Laird, C. D. (2022). "Optimization-based approaches for design of chemical process families using ReLU surrogates." *in proceedings of Foundations of Computer Aided Process Operations*, Jan. 2023.

Swaney, R. E. and Grossmann, I. E. (1985). "An index for operational flexibility in chemical process design." *AIChE Journal*, 31:621–630.

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

ENABLING ENQUIRY BASED LEARNING THROUGH ONLINE VISUALIZATION-BASED DATA SCIENCE MODULES

Joseph Menicucci[a], Raghuram Thiagarajan[b], Srinivas Rangarajan[a]
[a]Chemical & Biomolecular Engineering, Lehigh University, PA; [b]Pratt&Miller Engineering, MI

Keywords: Interactive modules, data science, python-based visualization, enquiry-based learning

This work demonstrates an ongoing multi-year project to develop interactive online visualization-based modules to explain STEM concepts broadly, and chemical engineering principles, in particular. Such modules are envisioned to serve as pedagogies for personalized learning as they are easy to use and enable enquiry-based learning of concepts facilitated by an instructor. This year, with support from CAChE and Lehigh U, we are developing modules for teaching data science to students of chemical engineering and allied disciplines.

Background and Results

Data science, considered the fourth pillar of science and technology is rapidly becoming a sought-after skillset in the chemical industry and academia alike. Most chemical engineering curricula do not include components of statistical learning or data-driven modeling to study chemical data although the last half a decade has seen tremendous advances in both research and education in artificial intelligence.

Our goal, in this context, is to develop interactive, online, visualization-based tools to teach general science and engineering concepts and, specifically, data science concepts. Computational modules are valuable pedagogical tools to teach complex engineering concepts. Excellent examples of such tools include the Etomica [1] project and the LearnChemE modules [2]. Our thesis therefore is that providing such online modules to demonstrate various aspects of building data-driven models can facilitate instructors to employ them in relevant core chemical engineering courses (e.g., introductory engineering courses, numerical methods course, as well as undergraduate lab courses).

In the previous years, we have developed and hosted a number of online modules [3], ranging from simple ones such as a system of sequential reactions to complicated ones such as an infectious disease model as well as a basic catalysis data science module. Currently, we are developing two additional data science modules. First one focuses on comparing building physics-based, purely-data driven, and hybrid models using biological data. The scope of this module is to enable students to build different kinds of models using noisy experimental datasets of biomolecular systems and thereby understand the advantages and disadvantages of these methods. The second module will focus on techniques and concepts within data science to prevent overfitting of a model to a specific data set and to build models of optimal complexity.

This poster will present the general idea of interactive online visualization-based pedagogical modules, our experience with their adoption in classroom so far, the new data science modules, and possible future directions in the development of these modules.

References

[1] Etomica, https://www.etomica.org/ - Accessed May 12 2023
[2] Interactive simulations, LearnChemE https://learncheme.com/simulations/ - Accessed May 12 2023
[3] STEM visualizations - https://srrweb.cc.lehigh.edu/app/

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

THREE-BODY PROBLEM: SMALL DATA, INTERPRETATION AND EXTRAPOLATION IN ML

L. A. Briceno-Mena[1], C. G. Arges[2], and J. A. Romagnoli[1] *
[1]Cain Dept. of Chemical Eng., Louisiana State University, Baton Rouge, LA, US.
[2]Dept. of Chemical Eng., The Pennsylvania State University, University Park, PA, US

Keywords: Hybrid modeling, Machine Learning, Optimization

Introduction

Development and exploitation of Machine Learning (ML) models in engineering applications comes with challenges beyond the computer and data science aspect of it. Scarce availability of data, model interpretation, and physics awareness are three key issues that need to be tackled for ML models to become more ubiquitous in the industry. In this contribution, we propose a framework for the development of hybrid models oriented to address these three elements. We then show its applicability for the optimization of electrochemical systems with newly synthesized materials.

Methods

Compositional modeling has been shown to be effective in obtaining useful models in situations where extensive data and/or physical knowledge are not available. The modeling task is formulated as problem such that given the physical structure of the system, a set of statements about its behavior, a library of model fragments and a set of rules constraining their use, we seek the most useful, coherent scenario model for answering a query. This formulation enables the introduction of physics-based model fragments where available, and data driven model fragments where needed.

From a compositional model one can obtain an ML-based surrogate model with some desired characteristics. In doing so, Transfer Learning (TL) becomes useful. TL[2] is a ML technique in which a data-driven model previously trained (general training) for a given task is used as the base to build a model for a new task, with less data being required for the new training stage (task-specific training). In this work, the general training is performed on simulations generated using a compositional model. The target-specific training is done on experimental data.

While the lack of physics awareness in ML models is an ongoing area of research, its effects could be mitigated by tracking the model's extrapolation behavior. Here, extrapolation is addressed by detecting if the new sample to be predicted lays out of the training data distribution. Intuitively, if a point is located close enough to a set of training examples, it is not a point of extrapolation. The out-of-distribution detection was implemented using the k-nearest neighbors' distance [3].

Results

Results reveal a trade-off between extrapolation and predicted performance (as peak power density) (Fig. 1A). From the analysis of the solutions in the pareto front (Fig.1B), the ranges allowed for pressure and ion exchange capacity of the membrane are not fully exploited by the optimizer, although direct proportionality is identified. This is consistent with the physical intuition that additional hindering mechanisms exist, such as mass transfer of the reactant to the electrocatalyst surface through the ionomer binder.

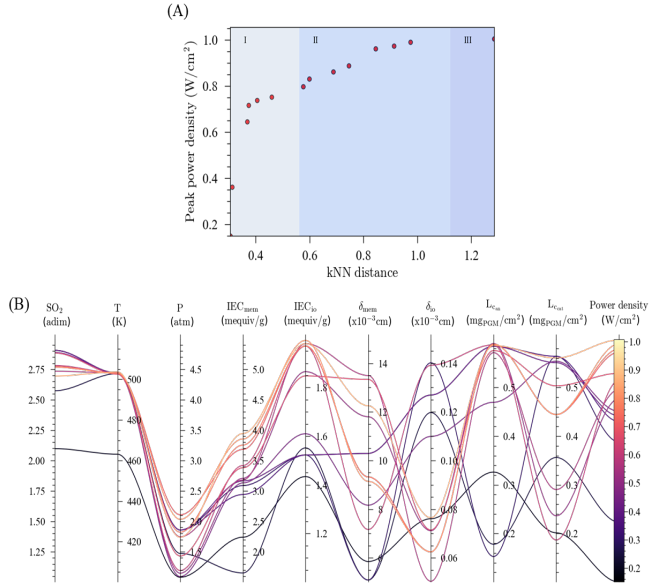

Figure 1. (A) optimization with extrapolation monitoring and (B) analysis of the solutions in the pareto front.

References

Falkenhainer, B.; Forbus, K. D., Compositional modeling: finding the right model for the job. *Artificial Intelligence* **1991,** *51* (1), 95-143.

Pan, S. J.; Yang, Q., A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering* **2010,** *22* (10), 1345-1359.

Sun, Y. a. M., Yifei and Zhu, Xiaojin and Li, Yixuan, Out-of-Distribution Detection with Deep Nearest Neighbors. *arXiv* **2022**.

* To whom all correspondence should be addressed; jose@lsu.edu.

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

LEARNING-BASED CONTROL IN ATOMIC-SCALE PROCESSING OF SEMICONDUCTOR MATERIALS

P.S Ramesh, S. Nath, K. Cho, and A. Mesbah*
University of California, Berkeley
Berkeley, CA 94720

Keywords: Atomic Layer Etch, Semiconductor Manufacturing, Learning-based Control, Surface Kinetic Model

The adoption of atomic layer processing for sub-nanometer device fabrication has been slow due to long processing times and challenges in effective control of small features. This work presents a learning-based control strategy for optimal control of atomic layer etch (ALE), with the objective of minimizing the time to obtain a specified target etch-depth, to maximize the productivity of the etch process. We propose a cycle-to-cycle (C2C) model predictive control (MPC) scheme to minimize the ALE cycle time in a single run, wrapped by an exterior learning-based control strategy to reduce the number of cycles in run-to-run (R2R) control. The outer learning-based control loop explores the design space of MPC parameters in a data-efficient manner to enable control policy adaptation based on offline measurements of etch depth for R2R control.

Introduction

ALE, which can cyclically remove single monolayers from a substrate due to its self-limiting nature, is gaining traction for its use in etching sub-nanometer features to provide enhanced uniformity and smoothness in next generation semiconductor devices (Kanarik *et al.*, 2018). A batch of semiconductor wafers is fed into an ALE process run, which operates over a pre-specified number of cycles. The long processing times of ALE is a current challenge that limits its widespread adoption in semiconductor manufacturing. Increased etch rates generally lead to increased surface roughness and, hence, balance between the two has been hard to achieve due to the lack of adequate process models and control schemes for the complex surface processes that occur during ALE.

Model Predictive Control

A first-order surface kinetic model for ALE of Si with Cl/Ar, which has been validated against experiments (Park *et al.*, 2004), is utilized for implementing a C2C MPC scheme. The kinetic model considers adsorbed Cl concentration, surface active site concentration, and etch per cycle as state variables, with flow rates and cycle times as input variables. Recent studies have shown that optical emission spectra can be used to monitor ALE processes, thereby facilitating state feedback at the end of each cycle. With a fixed number of cycles, the MPC controller seeks to minimize the total process time for each run.

Learning-Based Control for MPC Policy Adaptation

As the ALE process undergoes drifts and changes with time, the controller must be adapted in a R2R fashion to ensure performance in view of the time-varying nature of the process. This necessitates the implementation of a learning-based R2R control. Closed-loop performance data acquired cumulatively from different runs can be effectively utilized for setting up a Bayesian optimization framework for automated adaptation of the MPC policy for R2R control based on etch-depth. Bayesian optimization involves the construction of a probabilistic surrogate model to relate the MPC parameters with the closed-loop performance metrics. The uncertainty description provided by the probabilistic surrogate model is then used to successively identify better candidates for the parameters with the use of an acquisition function, which balances the exploitation-exploration tradeoff in efficiently exploring the potentially large design space of MPC parameters. Bayesian optimization is capable of searching the policy parameter space in a very data-efficient manner to identify parameter candidates using data accumulated from a limited number of process runs. In this sense, Bayesian optimization is superior to policy search methods that rely on gradient information, which require a much larger dataset to converge to the optimal candidates. We demonstrate that the proposed learning-based control approach is able to realize the desired etch rate in a limited number of process runs. Such control approaches would be essential for enabling nanofabrication of next generation quantum materials.

References

Keren J. Kanarik et al., (2018). Atomic Layer Etch: Rethinking the Art of Etch, J. Phys. Chem. Lett. 9, 16, 4814–4821.

Park *et al.*, (2004). Precise depth control of silicon etching using chlorine atomic layer etching, *Jap. J. Appl. Phys*. 44(1A): 389-393.

* To whom all correspondence should be addressed; mesbah@berkeley.edu.

A MACHINE LEARNING APPROACH FOR OPTIMIZING FLUSHING OPERATIONS IN LUBE OIL MANUFACTURING AND PACKAGING PLANTS

S. S. Jerpoth, B. Gao, E. A. Aboagye, R. P. Hesketh, C. S. Slater, and K. M. Yenkie*
Department of Chemical Engineering, Rowan University, Glassboro, NJ 08028

Keywords: Lube Oil Blending, Multi-Product Plants, Optimization, Flushing, Random Forest Classification

Introduction

The commercial lubricant industry manufactures over a thousand unique lube oil products by blending different base oils with numerous groups of additives. To meet the growing market demand, the number of product formulations and operation complexities are increasing as well. Therefore, the production system comprising blending vessels, ancillary equipment, and a complex pipeline network must be reused multiple times for numerous batch productions. The lube oil industry poses stringent product quality requirements therefore, it is a necessity to avoid cross-contamination during changeovers and ensure efficient cleaning of the pipeline network. The use of external agents is strictly prohibited therefore, an upcoming product batch is used for cleaning/flushing the residues of the previous product. This results in the formation of commingled/mixed oil that is regarded as a low-value product. The existing operations lack a standardized procedure that results in large volumes of commingled oil leading to economic losses exceeding millions of dollars annually. To achieve desired product specifications, attain quality targets, and maximize the productivity of assets, it is of paramount importance to enhance process control, minimize human errors and improve the resource management footprint of these industries. To this end, we present a machine learning approach to learn from the data of existing operations and strategically optimize the flushing operations.

Methodology

We collaborated with a commercial lube oil blending plant (LOBP) to get insights into their plant configuration and manufacturing processes.

Based on their end-use, the finished products are classified into families such as Engine, Gear, Industrial, Turbine, and Synthetic oils. Each of these oil families has distinguishing features and consists of over 100 different products. To ensure product integrity, the respective family must remain separate from each other. Hence, the pipeline configuration at the partnered facility is separated into five lines for each product family.

Our present study focuses on using machine learning to predict distinguishing features of the lube oil blends, allow better in-line controllability, and minimize the oil downgrade during product changeovers. We performed a cross-validation and based on the score we chose a random forest classification algorithm to solve the classification problem for success/failure of the flushing operation [1], [2]. A total of 1432 data points were used, with 80% being used for training and 20% for testing. The feature set comprised documented flush, kinematic viscosities at 40°C and 100°C, average ambient temperature, product family, and the system type (2 oil/ 3 oil mixture). After tuning the hyperparameters for the model, the performance was evaluated on the test set using the confusion matrix. The goal was to predict which product batch failed and which passed depending on the compatibility of the consecutive products.

Results

Figure 1 illustrates the results from the model where out of 116 test records, it predicted the classification of 110 records correctly with an accuracy of 96% and a recall value of 0.33.

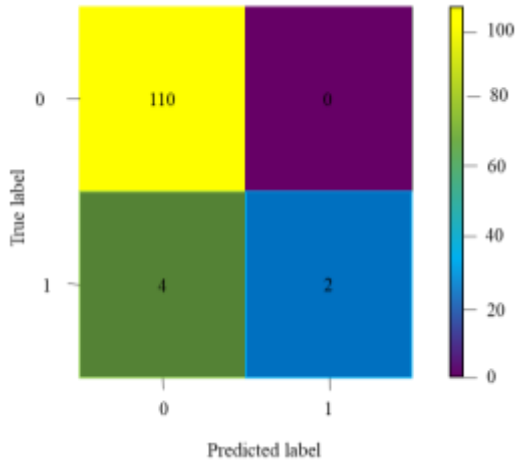

Figure 1: Confusion Matrix of the test dataset prediction by Balanced Random Forest Classifier

Summary

Our work will help the lube oil industries to make more informed decisions and predict the pass and fail scenarios depending on the compatibility of the consecutive products and mixture type.

References

[1] H. Sun and W. Zhou, "Classification of failure modes of pipelines containing longitudinal surface cracks using mechanics-based and machine learning models," *J. Infrastruct. Preserv. Resil.*, vol. 4, no. 1, p. 5, Jan. 2023, doi: 10.1186/s43065-022-00062-5.

[2] M. A. Marins *et al.*, "Fault detection and classification in oil wells and production/service lines using random forest," *J. Pet. Sci. Eng.*, vol. 197, p. 107879, Feb. 2021, doi: 10.1016/j.petrol.2020.107879.

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

A SURROGATE-BASED FRAMEWORK FOR FEASIBILITY-DRIVEN OPTIMIZATION

Huayu Tian, Marianthi Ierapetritou*

Department of Chemical and Biomolecular Engineering, University of Delaware, Newark, DE, USA

Keywords: Surrogate-Based Optimization, Feasibility, Simulation

Introduction

Surrogate-based optimization approaches have been widely adopted in industrial problems due to its potential to reduce the number of simulation runs required in the optimization process, among which, Kriging-based optimization, also known as Bayesian optimization (Mockus, 2012) has gained increasing popularity. In this work, we developed a framework based on (Wang, Escotet-Espinoza, Singh, & Ierapetritou, 2017), with feasibility problem considered as a classification problem, and additional stages introduced to improve local exploitation and global exploration.

Methodology

First, support vector machine and Gaussian process models are constructed for feasibility (Badejo & Ierapetritou, 2022) and objective functions, respectively, based on initial sampling. Next, new samples are iteratively added to improve model accuracy at regions of interest. Different stage including different infill criteria focusing on feasibility characterization, objective minimization and global exploration are included in the framework. Switches between stages are determined considering model changing and data characteristics as the iteration proceeds. Finally, the algorithm terminates and the returns the best solution.

Results

The framework is applied to several test problems and a continuous pharmaceutical manufacturing case study. Due to space limitations, results for some test problems are summarized in Table 1. Results show that the algorithm is able to converge to the optimum with a small number of samples for test problems. Also, the distribution of sampling cost across different stages is problem-dependent, indicating that the stage index determination criteria are related to problem characteristics.

Conclusions

Classification models are applicable for feasibility characterization and are easier to train and be used for high dimensional/categorical data than regression models. Flexible allocation of sampling budget for different purposes according to the characteristics of the problem can reduce the total number of simulation runs.

Table 1. Results for test problems (Wang & Ierapetritou, 2017)

Problem	Average number of iterations across 20 runs (percentage of total sample points)				
	Feasibility	Constrained optimization	Local refinement	Global optimization	Total
Ex3	14.3 (93.46%)	1 (6.54%)	0 (0%)	0 (0%)	15.3
New Branin	34.2 (64.04%)	10 (18.73%)	1 (1.87%)	8.2 (15.36%)	53.4
Gomez	52.1 (56.20%)	26.6 (28.69%)	4.3 (4.64%)	9.7 (10.46%)	92.7

Acknowledgement

The authors acknowledge financial support from US Food and Drug Administration (FDA) through grants FDABAA-20-00123 and DHHS-FDA-1U01FD006487.

References

Badejo, O., & Ierapetritou, M. (2022). Integrating tactical planning, operational planning and scheduling using data-driven feasibility analysis. *Computers & Chemical Engineering, 161*. doi:10.1016/j.compchemeng.2022.107759

Mockus, J. (2012). *Bayesian approach to global optimization: theory and applications* (Vol. 37): Springer Science & Business Media.

Wang, Z., Escotet-Espinoza, S., Singh, R., & Ierapetritou, M. (2017). Surrogate-based Optimization for Pharmaceutical Manufacturing Processes. In *27th European Symposium on Computer Aided Process Engineering* (pp. 2797-2802).

Wang, Z., & Ierapetritou, M. (2017). A novel feasibility analysis method for black-box processes using a radial basis function adaptive sampling approach. *AIChE Journal, 63*(2), 532-550. doi:10.1002/aic.15362

* Corresponding author: mgi@udel.edu

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

HYBRID MODELING FOR SOLID TRANSPORT CRITICAL VELOCITY

Su Meyra Tatar[a], Yushi Deng[a], Haijing Gao[b], Shuxing Cheng[b], Simona Duplat[b] and Selen Cremaschi[a]*
[a]Department of Chemical Engineering, Auburn University, Auburn, AL 36849, USA
[b]Chevron Technical Center, Houston, TX 77002, USA

Keywords: solid transport, machine learning, hybrid modeling, semi-mechanistic models

Critical velocity is the carrier fluid velocity that transports solid particles without depositing them in wellbores/pipelines. Accurate critical velocity estimates and their uncertainty are essential for preventing solids deposition in multiphase flow applications (Dabirian et al., 2016). This contribution adopts a parallel structure hybrid modeling approach (Deng et al., 2022) (Fig. 1) to calculate critical velocity and its uncertainty. In Fig. 1, X represents the inputs, y^m is the predicted velocity by the semi-mechanistic model, and $\hat{y}^e$ is the hybrid model prediction with its confidence interval (CI). Additionally, δ represents the model discrepancy, the difference between the semi-mechanistic model's prediction (y^m) and experimentally observed critical velocity (y^e), and $\hat{\delta}$ is the model discrepancy estimate with its mean (μ) and variance (σ^2) calculated using the data-driven model. We selected Oroskar and Turian (Oroskar and Turian, 1980) and Mantz (Mantz, 1977) as the semi-mechanistic models because they are widely used (Fig. 1). Gaussian Process Modeling (GPM) is employed to build the data-driven model because of its accuracy in regression with limited data sets. The data set includes 910 and 979 critical velocity measurements for Oroskar and Turian and Mantz models, including the combinations of solid concentrations, liquid densities and viscosities, solid densities, and particle and hydraulic diameters. While Oroskar and Turian model uses all inputs, Mantz Model doesn't include solid concentration. A GPM model is trained to estimate discrepancy for each semi-mechanistic model using the data set with GPy.

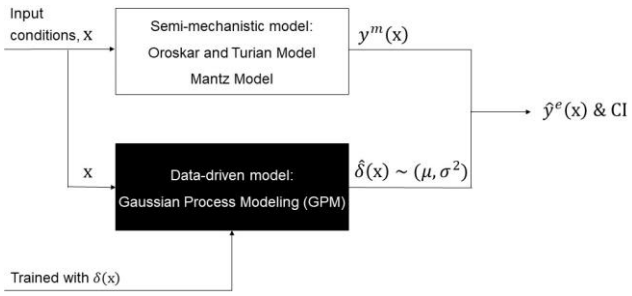

Figure 1. Hybrid modeling approach (Deng et al., 2022)

The critical velocity predictions of Mantz (Fig. 2(a)) and Oroskar and Turian (Fig. 2(b)) models, corresponding hybrid models including 95% CIs (Figs. 2(c) and 2(d)) are given in Fig. 2. Hybrid model predictions are obtained using k-fold cross-validation with four folds. The root mean square errors (RMSEs) are 0.876 and 0.491 m/s for Mantz and Oroskar and Turian models. The corresponding hybrid models reduce RMSEs to 0.229 and 0.164 m/s. Comparison of Figs. 2(c) and 2(d) reveal that the hybrid model with the Oroskar and Turian model yields lower RMSE and narrower CIs compared to the Mantz model.

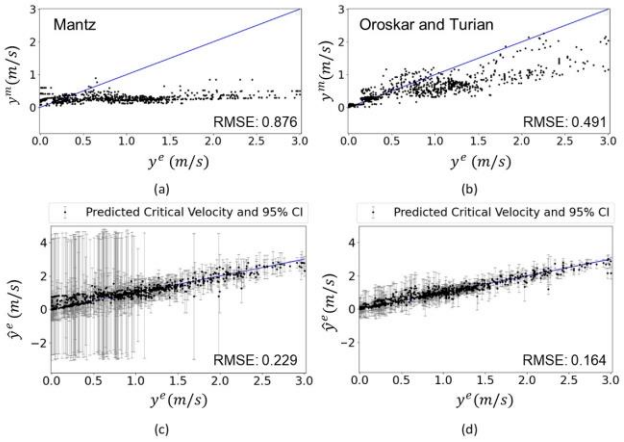

Figure 2. Critical velocity estimates of semi-mechanistic and hybrid models using Mantz (a) & (c) and Oroskar and Turian (b) & (d) models.

References

Dabirian, R., Mohan, R., Shoham, O., Kouba, G., 2016. Critical sand deposition velocity for gas-liquid stratified flow in horizontal pipes. *J Nat Gas Sci Eng* **33**, 527–537.

Deng, Y., Avila, C., Gao, H., Mantilla, I., Eden, M.R., Cremaschi, S., 2022. A hybrid modeling approach to estimate liquid entrainment fraction and its uncertainty. *Comput Chem Eng* **162**.

Mantz, P.A., 1977. Incipient Transport of Fine Grains and Flakes by Fluids — Extended Shields Diagram. *Journal of the Hydraulics Division* **103**, 601–615.

Oroskar, A.R., Turian, R.M., 1980. The critical velocity in pipeline flow of slurries. *AIChE Journal* **26**, 550–558.

* To whom all correspondence should be addressed; selen-cremaschi@auburn.edu.

Deep Reinforcement Learning-based Adaptive Tuning of Control for Complex Dynamical Systems

Myisha Chowdhury and Qiugang (Jay) Lu*
Texas Tech University
Lubbock, TX 79409

Keywords: Deep reinforcement learning, PID control, complex dynamical systems

Most controllers for dynamical systems are designed based on a fixed model under a given operating condition. Thus, they often require re-tuning once the operating condition changes. Traditional controller tuning strategies are either time-consuming, e.g., heuristic and trial-and-error methods, or require an accurate plant model, e.g., rule-based methods (Borase et al., 2021). How to enable model-free automatic and adaptive tuning of controllers under varying operating conditions is critical. To this end, deep reinforcement learning (DRL) has attracted attention for controller tuning due to its model-free nature and online learning capability. However, DRL algorithms suffer from respective issues in their applications to the controller auto-tuning. Specifically, stochastic DRL methods, e.g., proximal policy gradient, often have poor sample efficiency. On the contrary, deterministic actor-critic methods, e.g., deterministic deep policy gradient (DDPG), suffer from under-exploration to find global solutions (Silver et al., 2014). This motivates our work to develop a new DRL method that overcomes the issues of low sample efficiency and under-exploration, to assist the auto-tuning of controllers for dynamical systems.

Entropy-maximization TD3 (EMTD3) method

In this work, we present a novel DRL method to enable sample efficiency and sufficient exploration in training the agent and discovering global solutions. Specifically, the proposed method integrates the advantages of (stochastic) soft actor-critic (SAC) and (deterministic) twin-delayed DDPG (TD3). For our method, termed as EMTD3, a stochastic actor is deployed at the beginning where an extropy maximization (EM) term, similar to SAC, is added to ensure sufficient explorations. Such an entropy maximization term adds uncertainty to the policy to facilitate the exploration of the action space, leading to improved solution than traditional deterministic methods. Our method runs such a stochastic actor for a duration to gain initial experience. After that, a deterministic actor is deployed to focus on local exploitation. This deterministic actor employs the TD3 method due to its sample efficiency, high stability, and fast convergence (Chowdhury and Lu, 2022). By integrating stochastic and deterministic DRL algorithms sequentially, the EMTD3 method can outperform typically DRL approaches in terms of sample efficiency and convergence to solutions.

EMTD3-based Controller Tuning

The proposed EMTD3 method is implemented to the PID controller tuning for a nonlinear continuous stirred tank reactor (CSTR). Specifically, for the formulated DRL problem, the action of the agent is defined as the PID parameters, whereas the environment state is defined as the trajectories of CV, MV, and setpoints of a closed-loop simulation under a given set of PID parameters. The reward is defined as the closed-loop tracking error. To assess the performance of our method, we compare it with other methods including SAC and TD3, as well as fixed PID control, under varying setpoint values that correspond to different operating conditions, shown below:

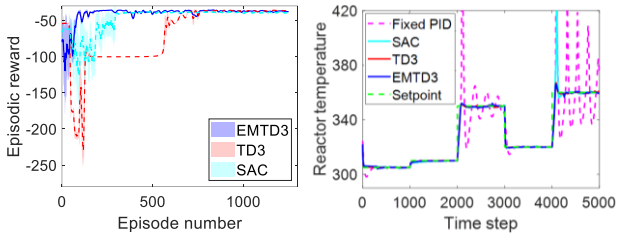

Fig. 1. Left: Learning curves for DRL methods; Right: The CV and MV under the last-episode PID parameters.

As shown in the left plot of Fig. 1, among three candidate DRL methods, the proposed EMTD3 method can achieve the fastest training speed and thus the highest sample efficiency. The right plot shows the adaptivity of the trained DRL agents under various setpoint changes for this nonlinear CSTR system. It shows that our method can outperform the others in enabling adaptive PID tuning. Further study shows that our method can give solutions that are much closer to the global optimum than the others.

Acknowledgments. The authors acknowledge the support from the Texas Tech University.

References

Borase, R.P., Maghade, D.K., Sondkar, S.Y. and Pawar, S.N., (2021). A review of PID control, tuning methods and applications. *Int. J. Dyn. Control.*, **9**, pp. 818-827.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., & Riedmiller, M. (2014). Deterministic policy gradient algorithms. In *ICML*, pp. 387-395.

Chowdhury, M. A., & Lu, Q. (2022). A novel entropy-maximizing TD3-based reinforcement learning for automatic PID tuning. *arXiv preprint arXiv:2210.02381*.

* Email: jay.lu@ttu.edu

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

REAL TIME GUIDANCE FOR MODEL PREDICTIVE CONTROL SYSTEM THROUGH AUTOMATED ANALYSIS

K. C. Ridley *
Aspen Technology, Inc.
Houston, TX 77042

Keywords: APC, simulation, cognitive analytics

Maximizing the utilization and effectiveness of Advanced Process Control (APC) remains a challenge despite its long-standing benefits in optimizing continuous manufacturing process units. Interacting with APC controllers can sometimes be complex for operators because changing an input constraint in an APC application does not generally produce the same effect on the process as adjusting the same input directly in a single-input single-output (SISO) controller. To address this challenge, we present an interactive tool called Aspen Virtual Advisor (AVA) that simulates the controller using cognitive analytics to help operators and engineers understand why the identified solution occurs and how to modify the problem to achieve a more desired outcome. The tool is capable of automatically running different scenarios and provides recommendations for improving control strategies.

Balancing Optimization and Variable Requirements

APC controllers solve real-time constrained optimization problems. The optimization problem changes when there are changes in the limits, tuning, or process disturbances. If a constraint is modified, variables that have no direct relationship may move due to the interplay between the process model, objective function, and constraints. If the APC controller was turned off, changing the input would only directly affect that input and any other changes would be determined by the process response. Though the APC controller can find the optimal solution, there is a trade-off between achieving that optimal solution and meeting specific requirements for certain variables. The problem is that finding a solution to this trade-off can be difficult, especially for large controllers. Running manual simulations to identify a solution is impractical because it requires educated guesses changing constraints based on "What-If" analysis of the online controller. By the time the best course of action is identified, the controller will have likely already transitioned to a new state.

Using Cognitive Analytics for Desired Trade-Off

One potential solution is to leverage the predictive capabilities of APC controllers. By applying the same predictive capability of the controller at the current time, specific scenarios can be quickly simulated to determine which constraint changes lead to the desired behavior. This empowers operators to identify their best course of action. AVA for DMC3, AspenTech's APC technology, enables users to ask why variables are not behaving as expected or how to get variables to behave in desired ways. By asking AVA questions, operators can discover novel solutions.

When an operator asks a question, AVA creates a copy of the online controller, runs specific simulations, and returns a set of recommendations, to change constraints, that are ordered based on how well they improve the DMC3 economic objective function. This ranking provides alignment between the optimal solution and the trade-off in getting specific variables to behave as desired. By combining the operator's knowledge about the plant and the information AVA provides from the simulations, this technology augments the operator's intelligence enabling them to make better use of DMC3.

To gain operator acceptance, AVA scans their controller history data to see which variables the operators actually change and by what magnitude. It develops a profile that captures how comfortable operators are adjusting limits for variables based on the current operating point of the plant. As operators open limits closer to their true constraints, they tend to get more conservative. So as AVA pushes a limit further open approaching a true constraint it knows the operators comfort level and recommends smaller moves.

By leveraging cognitive analytics, operators, and the AVA can work together, using DMC3 to optimize the plant's performance. To improve the APC controller, ask AVA a question and identify the best course of action.

* To whom all correspondence should be addressed; k.c.ridley@aspentech.com.

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

INTEGRATING PROCESS ANALYTICS AND MACHINE LEARNING INTO THE CHEMICAL ENGINEERING CURRICULUM

Aycan Hacioglu, Ph.D. *
MathWorks
Natick, MA 01760

Keywords: Machine Learning, Process Analytics, Chemical Engineering, Curriculum, MATLAB

Many chemical engineering departments are looking for ways to integrate machine learning and process analytics into chemical engineering curricula to meet the increased demand from the industry.

Traditional chemical engineering curriculum helps students excel at first-principles modeling to solve chemical engineering problems in fluid mechanics, heat transfer, mass transfer, thermodynamics, or other core chemical engineering courses. To introduce process analytics or machine learning to chemical engineering students, some departments offer dedicated courses or minor degree programs, some departments add curriculum modules to existing courses to cover relevant topics.

In addition to theoretical foundations, process analytics, and machine learning require programming knowledge. Instructors consider students' prior programming experience, the dominant language in the job market, availability of programming tools and teaching resources, and ease of use when they select the programming tool for their courses. MATLAB is a popular computational tool among engineers for scientific computing. It also offers capabilities to develop data-driven or hybrid models in addition to first-principles models.

This work focuses on how process analytics and machine learning can be integrated into chemical engineering using MATLAB either as a standalone course or as a module in the existing courses. Fluid mechanics, chemical engineering laboratory courses, process control, and thermodynamics are some examples that could benefit from the addition of data-driven techniques. Examples of adding machine learning to first-year chemical engineering courses using MATLAB can be found in the literature. (Joss and Müller, 2019). Stand-alone, elective courses on predictive maintenance, machine learning, and process analytics offer students great exposure to data-driven techniques as well. (Hedengren, 2023).

There are various educational resources available for chemical engineering instructors to introduce machine learning. This work summarizes MATLAB-based curriculum resources for machine learning: Free, interactive, modular course modules (Live Scripts), self-paced online training courses, domain-specific documentation examples for applications such as predictive maintenance and visual inspection of products in pharmaceutical or semiconductor applications, course projects derived from industry-relevant problems, books, educational videos, and automated grading tool, MATLAB Grader.

This work highlights user stories derived from chemical processing industries which could be used to motivate students to learn machine learning. Additionally, making data science competitions part of the curriculum offers a way to challenge students to apply their knowledge to novel problems by being part of an interdisciplinary team. Examples of such competitions will be shared in this presentation.

References

Joss, L., & Müller, E. A. (2019). Machine learning for fluid property correlations: classroom examples with MATLAB. *Journal of Chemical Education*, *96*(4), 697-703.

Hedengren, J. (2023). Learn Data-Driven Engineering with Interactive Modules. 2023 AIChE Spring Meeting and 19th Global Congress on Process Safety, March 15, Houston, Texas

* To whom all correspondence should be addressed; ahaciogl@mathworks.com

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

IOT SENSOR NETWORKS AND MACHINE LEARNING FOR ANALYZING PLANT RESPONSE TO SALT AND ACETIC ACID TREATMENTS

Chinwe Aghadinuno, A'Ishah Trahan, Eman El Dakkak, Yadong Qi, Wesley Gray,
Jiecai Luo, Yasser Ismail, Fred Lacy*, Patrick Mensah.

Southern University and A&M College.
Baton Rouge, LA 70807

Keywords: plant health, Internet of Things (IoT), sensor networks, Machine Learning,

This research investigates the application of Internet of Things (IoT) sensor networks and machine learning techniques to analyze the response of Buxus Microphyll (Boxwood) and Rhododendron Indicum (Azalea) plants to various salt and acetic acid treatments. It presents an approach to analyze plant responses to varying concentrations of salt and acetic acid treatments. The objective of the research was to assess the changes in plants subjected to different stressors and to predict their impact on plant health and growth.

For the experiment, plants were treated in a controlled environment, with diverse percentages of acetic acid and salt concentrations. Subsequent to the treatments, IoT sensors and surveillance cameras were utilized to collect a comprehensive set of data and images, documenting the plant health and growth over time.

The IoT sensors monitored parameters such as temperature, moisture, light intensity, and soil fertility, which are essential for assessing plant health. Light intensity was measured as Photosynthetically Active Radiation (PAR), while dielectric permittivity was utilized to measure soil moisture. Soil fertility was determined based on soil electrical conductivity (EC), and atmospheric temperature was captured in Celsius by built-in thermostats.

Images of the treated plants were collected using surveillance cameras. The backgrounds were removed to focus on the plants' physical attributes and a machine learning model was implemented to analyze pixel values in the images. A percentage error was calculated to evaluate the accuracy and reliability of the analysis.

Initial findings suggest that IoT sensor networks and machine learning can effectively monitor and analyze plant responses to different stressors. By monitoring vital parameters and analyzing images, this approach can provide valuable insights into plant physiology and help develop strategies to improve plant resilience to environmental stressors. Further research is needed to refine the model and validate its efficacy across a wider range of plant species.

References

[1] G. G. and A. P. J., "Identification of plant leaf diseases using a nine-layer deep convolutional neural network," Computers & Electrical Engineering, vol. 76, pp. 323–338, Jun. 2019, doi: https://doi.org/10.1016/j.compeleceng.2019.04.011.

[2] D. L. Corwin, "Climate change impacts on soil salinity in agricultural areas," European Journal of Soil Science, vol. 72, no. 2, pp. 842–862, Aug. 2020, doi: https://doi.org/10.1111/ejss.13010.

[3] K. Golhani, S. K. Balasundram, G. Vadamalai, and B. Pradhan, "A review of neural networks in plant disease detection using hyperspectral data," Information Processing in Agriculture, vol. 5, no. 3, pp. 354–371, Sep. 2018, doi: https://doi.org/10.1016/j.inpa.2018.05.002.

[4] N. Suma, S. Samson, S. Saranya, G. Shanmugapriya, and R. Subhashri, "International Journal on Recent and Innovation Trends in Computing and Communication IOT Based Smart Agriculture Monitoring System," 2017.

[5] S. L. Ullo and G. R. Sinha, "Advances in IoT and Smart Sensors for Remote Sensing and Agriculture Applications," Remote Sensing, vol. 13, no. 13, p. 2585, Jul. 2021, doi: https://doi.org/10.3390/rs13132585.

[6] L. García, L. Parra, J. M. Jimenez, J. Lloret, and P. Lorenz, "IoT-Based Smart Irrigation Systems: An Overview on the Recent Trends on Sensors and IoT Systems for Irrigation in Precision Agriculture," Sensors, vol. 20, no. 4, p. 1042, Feb. 2020, doi: https://doi.org/10.3390/s20041042.

[7] V. Singh et al., "Detection of unhealthy region of plant leaves using Image Processing and Genetic Algorithm

Corresponding author's name: fred_lacy@subr.edu

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

MODELING COMPLEX COMBUSTION WITH MODIFIED TRANSFORMER NEURAL NETWORKS

Ethan Gallup, Jacob Tuttle, Blake Billings, Jacob Immonen, and Kody M. Powell*
University of Utah Department of Chemical Engineering
50 Central Campus Dr, Salt Lake City UT 84112

Keywords: Transformer neural networks, Dynamic optimization, Model Predictive Control

The growing popularity of renewable energy sources is adding to the variability of the already fluctuating load on the grid. This makes load following in coal-fired power plants increasingly difficult and opens opportunities to improve dynamic control systems.

Recent research has shown that transformer neural networks used as multi-step ahead time-series prediction models significantly enhance both control performance and computation efficiency of model predictive control (MPC) algorithms when compared with long-short-term memory (LSTM) neural networks (Park 2022). Popular transformer models lack the ability to efficiently correlate variables individually. This project designs a transformer with modified positional embedding layers from vision models to increase understanding of the relationships between variables (Ramachandran et al. 2019).

Transformer Neural Networks

The transformer network used requires the data to be structured as shown in Figure 1.

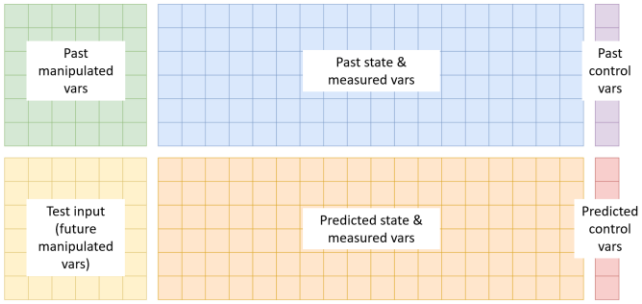

Figure 1. Data structure used by the transformer model.

The encoder of the transformer takes the past variables as input and passes a hidden state to the decoder which takes the test inputs and predicts control variables.

Modifications for Combustion Modeling

The generic transformer takes its own outputs as an input to the decoder. The new model takes manipulated variables as input to the decoder allowing it to draw stronger correlations between them and the control variable. The structure and inputs of the model are in Figure 2. The attention mechanism has a global positional embedding layer based on visual attention models.

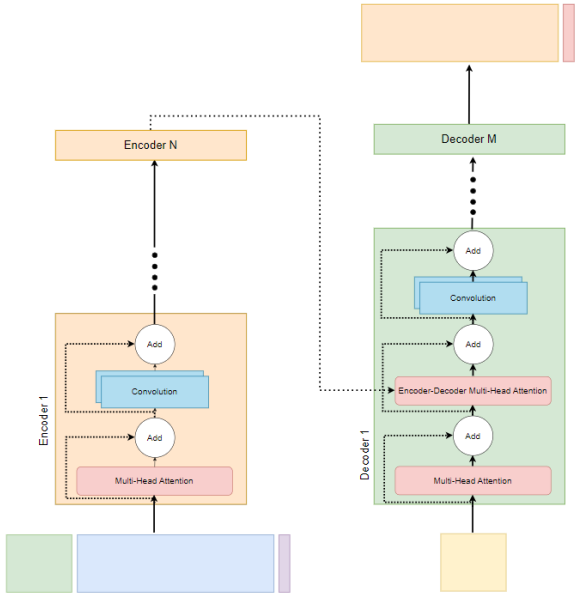

Figure 2. general transformer model with manipulated variables as the decoder inputs.

Results

The new model was tested against GRU, LSTM, bidirectional LSTM, and previous transformer models. The newly developed model showed at least a 15% increase in computational efficiency over all recurrent models and a 5% increase in response to manipulated variables over the previous transformer model.

References

Park, J. (2022). Hybrid Machine Learning and Physics-Based Modeling Approaches for Process Control and Optimization (dissertation).

Ramachandran P., Parmar N., Vaswani A., Bello I., Levskaya A, Shlens J. (2019). Stand-Alone Self-Attention in Vision Models, *arXiv*.

* kody.powell@ utah.edu.

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

UTILIZING MACHINE LEARNING TO BRIDGE ION TRANSPORT MODELS IN IONIC LIQUIDS

J. E. Umaña, R. K. Cashen, M. A. Gebbie, and V. M. Zavala*
University of Wisconsin-Madison
Madison, WI 53706

Keywords: Ionic Liquids, Conductivity, Physics-Informed Modeling, Machine Learning

Ionic liquids (ILs) have emerged as a promising alternative to flammable organic electrolytes. ILs uniquely possess low melting points and high chemical stability due to their ionic composition (Gebbie, et al., 2015). As such, high conductivity ILs have become increasingly investigated materials in electrochemical energy storage research (Baskin, et al., 2022). Unfortunately, the coulombic interactions which stabilize these molecules produce emergent interactions that are not well understood, limiting the systematic exploration of the IL molecular design.

The transport models often used to model the electrical conductivity of liquids often assume hydrodynamic mechanisms. In particular, the Nernst-Einstein (NE) conductivity model is largely used to benchmark IL electrolyte behavior using a material's viscosity. This has led to a framework which maximizes conductivity by minimizing viscosity. Recent studies, however, show many ILs exhibiting faster ion transport than their viscosity should allow and suggest the presence of structural transport mechanisms in ILs (Cashen, et al., 2022). Similarly, solid electrolytes demonstrate that non-hydrodynamic transport mechanisms can define electrolyte performance. The Arrhenius equation is used in literature to model ion transport in solid and polymeric materials and can be applied to ILs. However, this equation is empirical, limiting its ability to provide physical insight to ion motion (Caradant, et al., 2021).

This work builds on existing models using machine learning to improve model accuracy and interpretability. We used RDKit cheminformatic software to simulate IL molecular descriptors and combined this with experimentally reported conductivity to create an IL database. We then used these molecular descriptors to correct the NE equation and discover the physical origin of deviations from NE, shown in equation 1, where N_a is Avogadro's number, z is ion charge, e is an electron charge, μ is viscosity, R is the radius of each ion, and $f(X)$ is our learned correction. Further, IL descriptors were used to estimate Arrhenius equation parameters, $A(X)$ and $E_a(X)$, and interpret the significance of molecular IL descriptors in the kinetic modeling of ion transport as illustrated in equation 2 where R is the gas constant and T is temperature.

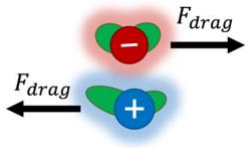

Figure 1. Hydrodynamic model of IL transport.

$$\Lambda_{NE} = \frac{N_a z^2 e^2}{6\pi\mu} \cdot \left(\frac{1}{R_+} + \frac{1}{R_-}\right) + f(X) \quad (1)$$

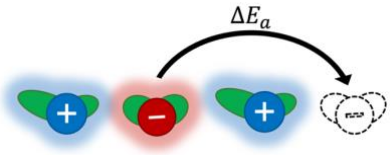

Figure 2. Kinetic model of IL transport.

$$\Lambda_{Arrhenius} = A(X) \cdot e^{-\frac{E_a(X)}{RT}} \quad (2)$$

The utilization of theoretical frameworks in machine learning allows us to gain real intuition for IL. We suggest a new theoretical framework for ILs by introducing interaction parameters. This combines hydrodynamic and kinetic models describe emergent IL transport phenomena. This new framework provides a new avenue for identifying IL candidates with desirable transport properties while providing inherent electrolyte stability.

References

Cashen, R. K., Donoghue, M. M., Schmeiser, A. J., and Gebbie, M. A. (2022). Bridging Database and Experimental Analysis to Reveal Super-hydrodynamic Conductivity Scaling Regimes in Ionic Liquids. *The Journal of Physical Chemistry B,* **126**:32, 6039-6051.

Gebbie, M. A., Dobbs, H. A., Valtiner, M., and Israelachvili, J. N. (2015) Long-range Electrostatic Screening in Ionic Liquids. *Proceedings of the National Academy of Sciences,* **112**:24, 7432-7437.

Caradant, L., Verdier, N., Foran, G., Lepage, D., Prébé, A., Aymé-Perrot, D., and Dollé, M. (2021) Extrusion of Polymer Blend Electrolytes for Solid-State Lithium Batteries: A Study of Polar Functional Groups. *ACS Applied Polymer Materials,* **3**, 6694-6704.

Baskin, I., Epshtein, A., Ein-Eli, Y. (2022). Benchmarking Machine Learning Methods for Modeling Physical Properties of Ionic Liquids. *Journal of Molecular Liquids,* **351**.

Foundations of Process/Product Analytics and Machine learning
FOPAM 2023, University of California, Davis CA, July 30-Aug 3, 2023

MODEL-BASED DESIGN OF EXPERIMENTS AND PYOMO.DOE

H.G. Lynch, J. Wang, and A.W. Dowling *
University of Notre Dame, Department of Chemical and Biomolecular Engineering
Notre Dame, IN 46556

Keywords: Nonlinear regression, model-based design of experiments, parameter estimation

Chemical engineers have been at the forefront of developing data science methods for nonlinear physics-informed models and computational optimization algorithms. At the confluence of these topics, chemical engineers often employ the iterative nonlinear model frame work shown in Figure 1 that include model selection, parameter estimation, sensitivity analysis, and model-based design of experiments (MBDoE). In fact, chemical engineers are some of the pioneers for MBDoE methods, which maximize one or more information metrics to design experimental campaigns that improve parameter precision or discriminate between models with the least amount of data (Franceschini and Macchietto, 2008).

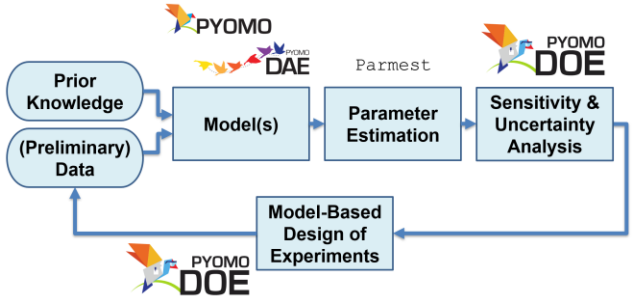

Figure 1. Nonlinear Modeling Workflow.

This poster has three goals: (1) present a self-contained mathematical primer on nonlinear regression and MBDoE; (2) survey recent advances and applications in MBDoE and sequential design of experiments; (3) and a tutorial of Pyomo.DoE (Wang and Dowling, 2022), an open-source package for MBDoE in Python.

Mathematical Primer

Inferring uncertain parameters from a model requires solving a (nonlinear) regression (or maximum likelihood estimation) optimization problem. The goal is to minimize the sum of square residuals between the model predictions and observed data (or maximize the log likelihood) constrained by the model, which is often a set of (partial) differential-algebraic equations (PDAEs). Parameter estimation is performed using observed experimental conditions.

In contrast, MBDoE determines the experimental conditions that will provide the maximum information gain, measured by Fisher Information Matrix (FIM). Often, MBDoE is applied sequentially, as shown in Figure 1, and the model(s) are updated/recalibrated before determining the next best experiments.

Advances and Applications for Design of Experiments

Recent advances include "self-driving laboratories" that intelligently combine MBDoE (or similar adaptive optimization strategies) with robotics to elect the next experimental conditions to maximize a material property or discover the dominant kinetics of a complex reaction environment (Langner et al., 2020). There are many barriers to fully leveraging physics-informed data science methods for molecular-to-systems applications. Existing methods may not scale to the nonlinear and large-scale (many equations) nature of multiscale models. As such, MBDoE is most often used by optimization experts.

Pyomo.DoE Tutorial

To reduce this barrier, Pyomo.DoE seeks to make FIM-based MBDoE easy-to-use for average Pyomo users. In this poster, we use an illustrative tutorial problem to demonstrate the improved Pyomo.DoE user interface (Wang and Dowling, 2022). Execution in the Pyomo ecosystem includes implementing the mathematical model into Pyomo, defining inputs for the model, and generating an experiment. Methodology aims at detailing the methods for computing the FIM, optimization, and exploratory analysis through enumeration. Results are examined through heatmaps and sensitivity curves.

References

Franceschini, Gaia, and Sandro Macchietto. Chemical Engineering Science 63, no. 19 (2008): 4846-4872.

Langner, Stefan, et al. Advanced Materials 32, no. 14 (2020): 1907801.

Wang, Jialu, and Alexander W. Dowling. AIChE Journal (2022): e17813.

* To whom all correspondence should be addressed; adowling@nd.edu