Deep Neural Network Enhanced Stochastic Optimization of Multistep Enzyme Catalyzed Industrial Codeine Bioproduction

Alexis Krywula Joannidis

School of Chemical and Biomedical Engineering, The University of Melbourne, Parkville, Melbourne, 3010, Australia

1. Abstract

The multistep enzyme catalyzed bioconversion of codeine from thebaine presents a more ecologically friendly route for codeine production, but industrial adoption hinges upon achieving economic viability, necessitating optimization of the process. To achieve this, maximum codeine production for the process was framed as an optimization problem, setting reaction parameters including enzyme introduction time, temperature and pH, as decision variables, and simulated with a python kinetics model.

Enzymes, as complex biological molecules, are prone to deviation from their expected performance, and hence the role of enzymes in this bioconversion embeds uncertainty into the optimization problem. This uncertainty was mitigated by development of a 2-step stochastic optimization method, whereby initial reaction parameters were optimized according to expected k_{cat} values - representing the expected enzyme efficiency of the system. After an initial, incomplete period of reaction, a deep neural network model was utilized to predict the true k_{cat} values from the incomplete reaction kinetics time-series data, allowing the incomplete reaction to be reoptimized according to this more accurate data. A 2.6% average increase in codeine production over the non-reoptimized process was observed for the stochastically reoptimized process.

Support vector machine models were compared to deep neural networks for several reaction kinetic data representations for detection of competitive contaminant inhibition of the steps of the reaction. Both methods performed similarly, with both performing best using PCA-reduced Euclidean distance representations of the kinetics time series data (~95% detection accuracy).

2. Introduction

The global contribution of industrial chemical production to mounting ecological decay[1], resource depletion[2] and public health crises[3], by processes that utilize harsh reagents and polluting effluents, has necessitated urgent efforts for the development of sustainable and clean alternative methods of chemical production for every implicated industry. Though the immense variety of chemical structures necessitates specialized approaches for different chemical syntheses, a field of applied research showing bright promise, owing to its sustainability and versatility of application, is the use of bioprocesses, and particularly, enzyme catalyzed synthetic methods[4].

These approaches mimic natural biological pathways, utilizing their enzymes to catalyze the production of bioproducts, without the toxic reagents or pollution. Non-naturally occurring chemicals are not excluded from this technique, as advances in genetic engineering unlock the customization potential of synthetic enzymes, or of modified microorganisms, engineered to express them.

Amongst natural compounds however, codeine exists as a prime example. A World Health Organization listed essential medicine[5], more than 300 tons of codeine is consumed annually, as the world's most abundantly used opiate[6]. Although codeine is naturally produced in the opium poppy plant *papaver somniferum* [7], 85-90% of its current global production occurs the synthetic methylation of morphine, involving harsh organic solvents, a toxic methylating agent and a carcinogenic byproduct[7]. With the cleaner, biological pathway on display for inspiration, using thebaine as a precursor (another naturally occurring opioid with less potential for misuse than morphine), presents an attractive alternative. Despite the approach's clear benefits however, the reality is that its industrial adoption will only be successful if it is economically attractive. The optimization of the process is therefore critical to its adoption, and such optimization forms the basis of this study.

The overall process, (only recently fully characterized)[8], occurs via a 3-step enzyme catalyzed reaction. The enzyme T6ODM catalyzes the conversion of thebaine to neopinone, which in turn is converted to codeinone, catalyzed by the enzyme NISO. COR, the final enzyme, ultimately catalyzes the conversion of codeinone to the desired product, codeine.

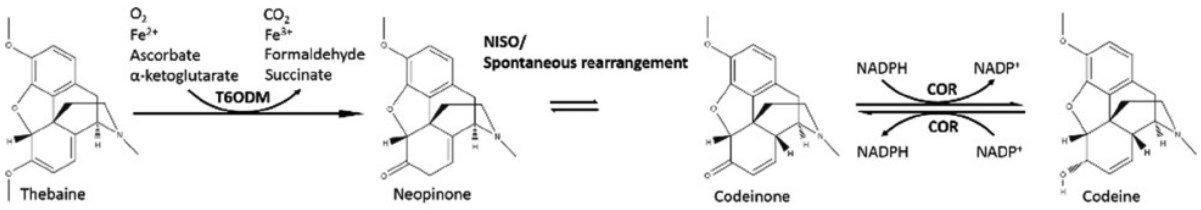

Fig. 1: In vitro pathway for enzyme-catalyzed conversion of thebaine to codeine, ignoring unwanted side-reaction conversion of neopinone to neopine[9]

Enzyme catalyzed reactions such as this may be modelled using reaction kinetic equations, such as the Michaelis-Menten equation:

$$\frac{d[S]}{dt} = \frac{k_{cat}[E]_0[S]}{\frac{k_{-1}+k_{+2}}{k_{+1}} + [S]} = -\frac{d[P]}{dt}$$

Where [S], [E], [P] represent the concentrations of the substrate (precursor), catalyzing enzyme and product respectively, and k_1, k_{-1}, k_2 represent the rate constant for the association and disassociation of the substrate-enzyme intermediate, and for the formation of the product from the intermediate, respectively. k_{cat} represents the turnover rate of enzyme – a measure of the enzyme's conversion effectiveness.

Hence, the first step, thebaine conversion, can be modelled as follows.

$$\frac{d[thebaine]}{dt} = \frac{k_{cat\,T6ODM}[T6ODM]_0[thebaine]}{\frac{k_{-1}+k_{+2}}{k_{+1}} + [thebaine]}$$

In many multi-step enzyme catalyzed reactions, including the thebaine to codeine conversion, one chemical species of the process may inhibit another species from its targeted reaction. In these cases, controlling the concentration of each species at each time step in the reaction becomes an important consideration to optimize conversion. An effective way to do so is to control the introduction of each enzyme progressively.

As biological compounds, enzymes are sensitive to their environmental conditions, and operate ideally within narrow temperature and pH ranges. In a multi-step reaction involving several enzymes with moderately different optimal conditions, these parameters become additional key optimization variables.

The thebaine to codeine conversion reaction represents one of numerous such multistep reactions where the above considerations are important optimization parameters. Using this reaction as a concrete example for a model that could represent other similar reactions, an optimization study may be framed around optimizing the enzyme presence, temperature and pH conditions of the process to maximize the production of codeine.

3. Batch Reactor Parameter Optimization

3.1 Objective Variable

In optimizing bioprocess output, several design choices must be made. Output goals may be specified in several ways; productivity, (rate of substrate conversion to product), yield, and time to reach a yield, are all common maximization objectives. For the purpose of this study, optimization is focused on reaction parameters – species presence, and physical properties, temperature and pH. Economic factors (cost of substrates, enzymes, reactor operation) are not considered here, hence the reaction timeframe, t_{final} is a predefined constant, and a suitable objective considers only thebaine conversion to codeine, that is, productivity. If the initial quantity of thebaine, $[thebaine]_0$ is also fixed, productivity is equivalent to final codeine conversion $[codeine]_{t\ final}$; hence the objective variable is labelled as

$$\max [codeine]_{t\ final}$$

By extending the single reaction Michaelis-Menten equation, the multi-step reaction can be modelled by defining the concentration of each subsequent species as the product of the previous step. Each species is then modelled as follows

$$\frac{d[thebaine]}{dt} = \frac{k_{cat\ T6ODM}[T6ODM]_0[thebaine]}{\frac{k_{-1}+k_{+2}}{k_{+1}} + [thebaine]}$$

$$\frac{d[neopinone]}{dt} = -\frac{d[thebaine]}{dt} + \frac{k_{cat\ COR}[COR]_0[neopinone]}{\frac{k_{-1}+k_{+2}}{k_{+1}} + [neopinone]}$$

$$\frac{d[codeinone]}{dt} = -\frac{d[neopinone]}{dt} + \frac{k_{cat\ NISO}[NISO]_0[codeinone]}{\frac{k_{-1}+k_{+2}}{k_{+1}} + [codeinone]}$$

And finally,

$$\frac{d[codeine]}{dt} = \frac{k_{cat\ NISO}[NISO]_0[codeine]}{\frac{k_{-1}+k_{+2}}{k_{+1}} + [codeine]}$$

Ultimately, [codeine], and hence the objective function, becomes a function of the concentration of each prior species

$$\max [codeine]_{t\ final}(x) = f\left(\frac{d[thebaine]}{dt}, \frac{d[neopinone]}{dt}, \frac{d[codeinone]}{dt}, \frac{d[T6ODM]}{dt}, \frac{d[NISO]}{dt}, \frac{d[COR]}{dt}\right)$$

3.2 Decision Variables

Bioprocess optimization can be subject to a plethora of parameters. Disregarding economic factors enables the narrowing of the scope of this study to focus on the fundamental reaction parameters, allowing for an intimate understanding of these factors before undertaking more comprehensive analyses. As such, the decision variables are as follows.

3.2.1 Temperature

Enzymes are complex biological proteins that developed to function at specific biological conditions. As such, they are sensitive to temperature, and their effectiveness dramatically decreases with deviation from ideal T. Each enzyme involved in a multi-step process may have a different optimal temperature, and hence temperature becomes a key decision variable.

Limitations in the python package used to model the reactions, Kinetics, allowed temperature to be modelled at only a constant value, instead of a custom profile, as a reactor would allow. As such, the constant temperature value, T, is the associated decision variable.

Although temperature effect is dramatic for high deviation, as enzymes denature and become completely inactive at high temperatures, most enzymes involved in multi-step reactions have similar T_{ideal} values within ~10 K of each other. As such, T effect on each reaction step is modelled in a simplified linear manner where deviation from T_{ideal} for the step reduces the enzyme effectiveness, K_{cat}:

$$\frac{d[S]}{dt} = \frac{\frac{|T_{ideal}-T|}{T_{ideal}}\ k_{cat}[E]_0[S]}{\frac{k_{-1}+k_{+2}}{k_{+1}} + [S]}$$

As T_{ideal} for each enzyme involved in codeine synthesis lies between 30-40° C, this decision variable may be constrained by 30<T<40 °C.

3.2.2 pH

Enzymes are similarly sensitive to environmental acidity, and have characteristic ideal pH values. As with temperature, the pH_{ideal} values for the relevant enzymes are all similar, within pH 7-8, hence for this small range, the pH effect can be simplified to be modelled by affecting K_{cat} in the same linear manner as temperature (taken together with temperature effect), with a constant value pH becoming the associated decision variable:

$$\frac{d[S]}{dt} = \frac{\frac{|pH_{ideal} - pH|}{pH_{ideal}} \frac{|T_{ideal} - T|}{T_{ideal}} k_{cat}[E]_0[S]}{\frac{k_{-1} + k_{+2}}{k_{+1}} + [S]}$$

As pH_{ideal} for each enzyme involved in codeine synthesis lies between7-8, this decision variable may be constrained by 7<pH<8.

3.2.3 Enzyme Introduction Time

Chemical inhibition occurs when one species of a reaction, though various mechanisms, reacts with another species (enzyme or substrate), limiting its ability to take part in its associated step of the reaction. In the codeine reaction, inhibition appears to occur in complex ways. As a simplification, for this study it is taken that the only inhibition occurring between the desired species is the competitive inhibition by codeine, on the conversion of thebaine to neopinone, where the inhibition is modelled as follows (including T and pH effects)

$$\frac{d[S]}{dt} = \frac{\frac{|pH_{ideal} - pH|}{pH_{ideal}} \frac{|T_{ideal} - T|}{T_{ideal}} k_{cat}[E]_0[S]}{\frac{k_{-1} + k_{+2}}{k_{+1}}(1 + \frac{[inhibitor]}{k_{inhib}}) + [S]}$$

As such, it is desirable to limit the production of codeine - the final product of the reaction - until the bulk of the thebaine has been converted into neopinone. As the production of codeine requires and is catalysed by COR, an effective method to control codeine production is to control the timing of COR introduction to the reactor. The Kinetics python package does not directly cater for such control; controlling the concentration of an enzyme is limited to defining d[E]/dt as a continuous ODE. As such, the best solution found was to govern enzyme concentration by defining d[E]/dt as a logistic differential equation

$$\frac{d[E]}{dt} = r\,[E](1 - \frac{[E]}{[E]_{max}})$$

So that [E] follows a logistic introduction profile from initial to maximum concentration

$$[E] = \frac{[E]_{max}}{1 + (\frac{[E]_{max} - [E]_{init}}{[E]_{init}})e^{-rt}}$$

where r is the parameter that influences the delay and slope of the logistic profile – hence it determines the delay and rate of introduction of enzyme. For this study, r becomes the decision variable associated with enzyme introduction time. As the only inhibition occurring involves a later species inhibiting a previous process, there is no need for negative sloped profiles, so r can be constrained to non-negative values. r was optimized for the final two enzymes – NISO and COR. For the study, r_{T6ODM} was ignored as none of the species involved in the first reaction step inhibit any other processes, hence T6ODM could be maximized from t=0.

3.3 Objective Function

With the decision variables defined, the full objective function may be formulated:

$$\max [codeine]_{t\ final}(x) = f\left(\frac{d[thebaine]}{dt}, \frac{d[neopinone]}{dt}, \frac{d[codeinone]}{dt}, \frac{d[T6ODM]}{dt}, \frac{d[NISO]}{dt}, \frac{d[COR]}{dt}\right)$$

$$for\ decision\ variables\ x_0 = T, x_1 = pH, x_2 = r_{NISO}, x_3 = r_{COR},$$

$$where\ generally, \frac{d[species]}{dt} = g(t, pH, T, [substrates], [products], [inhibitors], [E])$$

$$and\ \frac{d[E]}{dt} = h(t, pH, T, r)$$

$$subject\ to$$
$$30 \leq x_0 \leq 40$$
$$7.0 \leq x_1 \leq 7.8$$
$$0 \leq x_2$$
$$0 \leq x_3$$

3.4 Initial Optimization Implementation

The python Kinetics package was used to model the overall reaction, calculating the concentration of each species at each timestep, governed by the ODE rate equations specified in previous sections. Pyomo and scipy packages, particularly scipy.optimize, were then used for the parameter optimization, calling the Kinetics model to calculate $[codeine]_{t\ final}$, taking this value as the objective.

For t_{final} = 50 min, and the reaction parameters as given in Appendix A, the optimal values for the decision variables are presented in table 1, with the corresponding concentration curves presented in fig. 2.

Reaction variable	T (°C)	pH	r_{NISO}	r_{COR}
Optimized value	35.6	7.30	0.69	0.42

Table 1. Optimized reaction variables for thebaine to codeine conversion

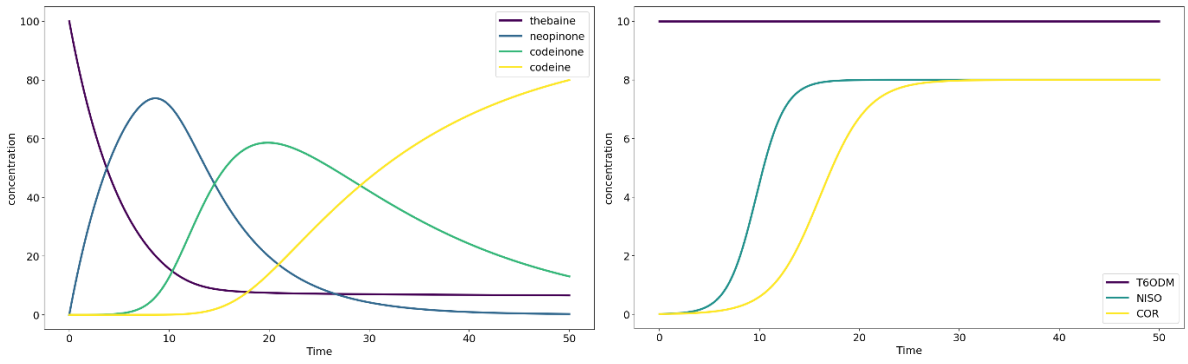

Figure 2. Reaction kinetics for optimized thebaine to codeine conversion:
a) reaction species concentrations **b)** Enzyme introduction profiles

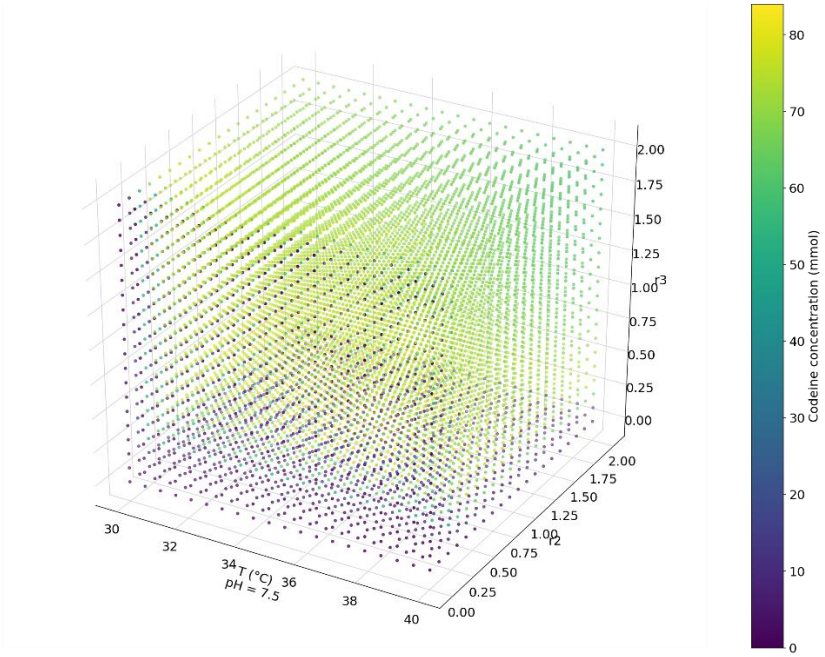

Figure 3. Codeine yield for varied decision variable values

The model can be extended to cater for an arbitrary quantity of enzyme catalyzed steps, optimizing the introduction profile of each enzyme according to inhibitions within the process. Below shows the optimization for an arbitrary 6 enzyme process with more complicated inhibition interactions.

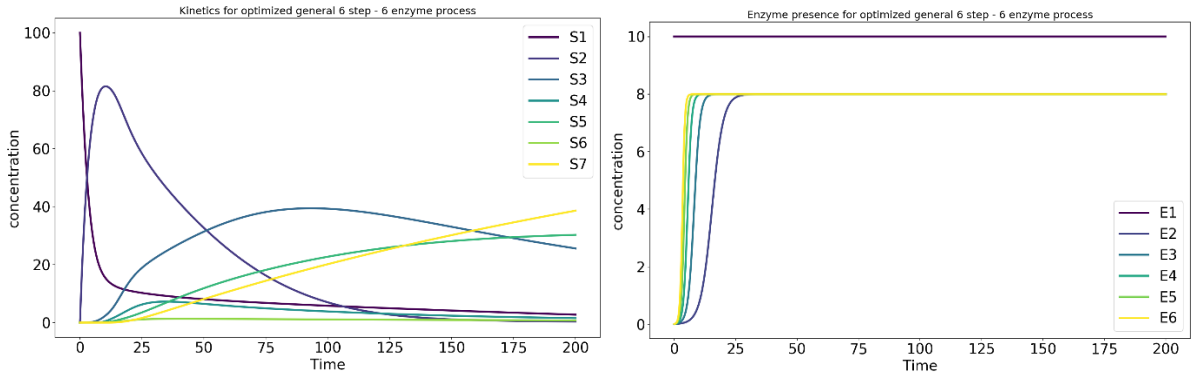

Figure 4. Reaction kinetics for optimized generic 6 step conversion reaction:
a) reaction species concentrations **b)** Enzyme introduction profiles

3.5 Stochastic Two-Step Optimization

Enzymes, as complex biological proteins, are prone to deviation from expected behaviour. As such, enzyme catalyzing performance may vary from theoretical performance. Enzyme kinetics models such as the one described here represent enzyme performance by the enzyme's $k_{cat,}$ or turnover number. In the optimization problem described above, parameters are optimized assuming a specific k_{cat} value for each enzyme. With the objective function a function of k_{cat}, if the actual values, $k_{cat\,exp}$ differ from the theoretical values, the 'optimized' decision variables x are likely sub-optimal. Hence, it is valuable to determine $k_{cat\,exp}$.

$K_{cat\ exp}$ can only be determined empirically by observation of the kinetics of a reaction. This challenge allows the problem to be framed as a stochastic optimization one with the following steps: X is optimized for assumed $k_{cat\ theor}$ values. Once the reaction has partially proceeded, at $t_{partial}$, evaluate $k_{cat\ exp}$ from the incomplete reaction kinetics. Then reoptimize X given $k_{cat\ theor}$, and finally, apply the reoptimized X values to complete the reaction.

3.5.1 Deep Neural Network K_{cat} Prediction

Deep neural networks (DNNs) are a class of neural network utilizing stacked hidden layers between input and output layers. They are effective for classification problems given a set of features.

A DNN model was applied to estimate $k_{cat\ exp}$ from the incomplete reaction kinetics (see appendix a for model architecture). The incomplete reaction data comprised of the concentration of each reaction species for each timestep from t_0 to $t_{partial} = 20$ min, as a separate time-series array for each species. To transform the data for the DNN, each time-series was simply concatenated to form one long array.

Training/test data of the same format was generated by simulating n=1000 reactions with the Kinetics model, with randomized k_{cat}, k_1 values for variance, using the reaction data as the feature vector y, and the predetermined k_{cat} value as the ground-truth label, X. The test/training data split followed a 60/40 ratio. The training results are as follows.

Following training, the model was utilized to evaluate $k_{cat\ exp}$ in the context of the two-step stochastic optimization problem. With reoptimization occurring at $t_{partial} = 20$ min, an average increase in codeine yield of 2% was observed compared with when x is optimized for $k_{cat\ theor}$ only. Fig. 5 illustrates the reaction kinetics for reoptimized vs non-reoptimized reactions with the same $k_{cat\ exp}$. Note the visible divergences at t=20, after which the reoptimized reactions typically convert more codeinone into codeine.

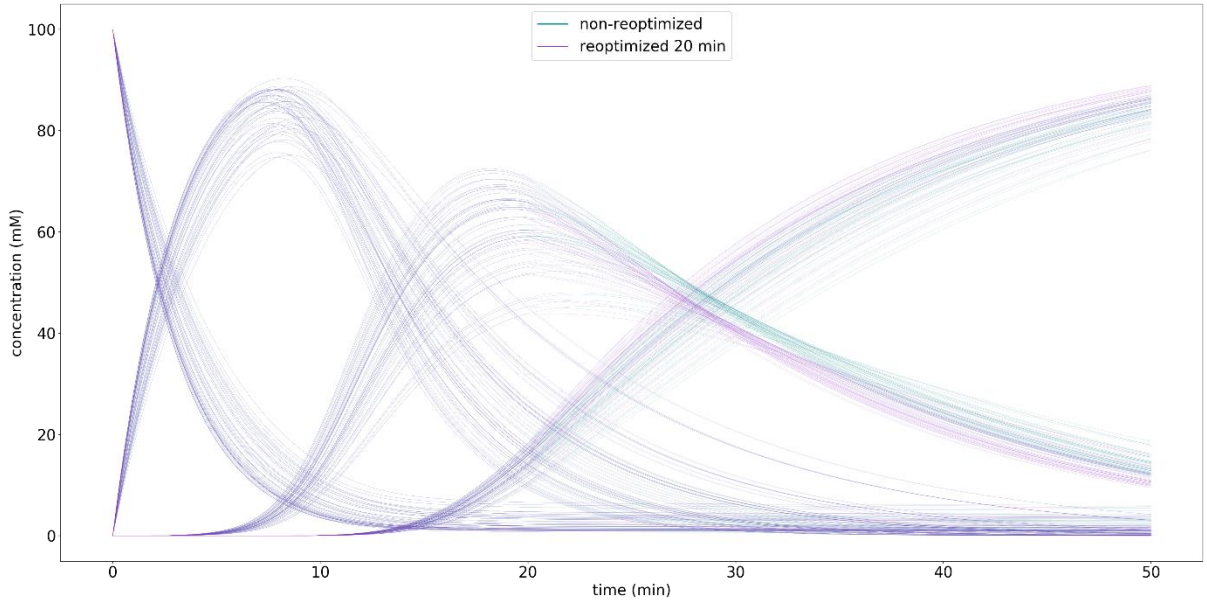

Figure 5. Comparison of kinetic concentration curves for 50 codeine conversion reactions, for reoptimization at $t_{partial} = 20$ min vs no reoptimization.

3.5.2 $t_{partial}$ optimization

The success of the two-step method poses an optimization problem of its own: For what value of $t_{partial}$ is codeine yield optimized? Lower $t_{partial}$ allows more of the reaction to occur with the superior reoptimized X values, but risks insufficient reaction data from which to accurately evaluate $k_{cat\ exp}$, while higher $t_{partial}$ reverses this tradeoff proposition.

To investigate this, the reoptimization process was repeated for t_partial = 10 min to compare against the initial 20 min result. Plotting the reaction kinetics for the reoptimized at t=10, 20 mins vs non-reoptimized reactions shows visible divergence after the reoptimization times (fig. 6). Although the plot shows several $t_{partial}$=10 min reactions generating the highest codeine yield, on average it performed worse than $t_{partial}$=20 (table 2), indicative that its k_{cat} predictions were more varied and less reliable, as corroborated by its higher average MSE from the true values.

Reoptimization time, $t_{partial}$ (min)	Average codeine yield (mM)	Average % increase over non-reoptimized reaction	Average DNN k_{cat} MSE
No reoptimization	83.5	-	-
10	85.4	2.32	1.06
20	85.7	2.63	0.887

Table 2. Optimized reaction variables for thebaine to codeine conversion

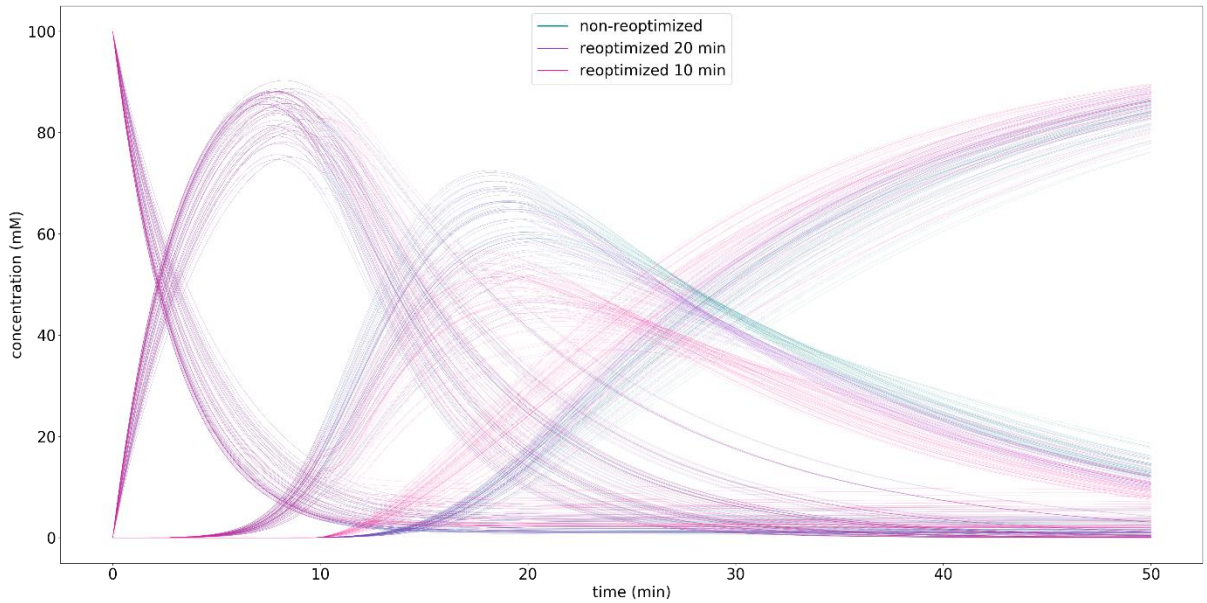

Figure 6. Comparison of kinetic concentration curves for 50 codeine conversion reactions, for reoptimization at $t_{partial}$ = 10 min, $t_{partial}$ = 20 min, and no reoptimization.

A more rigorous optimization problem was not performed here due to computation time constraints, but may be framed by setting the objective function as the maximum average $[codeine]_{t\ final}$ for the reoptimized process as a function of $t_{partial}$, evaluated by formulating this function by the procedure followed in 3.5.1.

$$\max[codeine]_{t\ final} = f(t_{partial})$$

4 Contaminant Detection

Enzymes utilized in bioprocesses are typically contained by microbes that are engineered to express them, and then grown via industrial fermentation processes using sustainably derived feedstock. The organic nature of the microbes, as well as the feedstock, provides sources of contamination in the final bioprocess.

Some contaminants have inhibitory effects on one or more steps of the reaction, such that a viable method of detecting these contaminants is to identify unexpected inhibition within the reaction. Reaction kinetics data of the kind used in this study is ideal for this purpose, and machine learning techniques may be applied to this data to detect inhibitory contamination. Such inhibition was simulated for the modelled codeine bioprocess, and two machine learning approaches – Support Vector Machines (a classical machine learning technique); and Deep Neural Networks (a modern approach) - were applied and compared in detecting contamination.

4.1 Contaminant Simulation

For the scope of this study, only competitive inhibition – whereby a contaminant species targets the active site of the enzyme, preventing the substrate from accessing it) – was examined. To simulate competitive inhibition, the Kinetics codeine model used throughout this study was modified via inbuilt functions to include competitive inhibition, represented by an inhibition rate coefficient, k_i, an inhibitor concentration, [i], and a specific target enzyme, e.

Inhibition properties for 1000 simulations were generated by randomizing [i] within 5-25 mM and the target enzyme (T60DM, NISO, COR, or no enzyme to represent no inhibition), holding k_i constant to focus on [i] effects for this initial study. Kinetics data for 1000 corresponding reaction simulations was generated by running the Kinetics model with these properties, by slightly randomizing k_{cat} for variance. The resultant kinetics data was modified with slight noise to more accurately resemble real data.

Setting [i], e, for each simulation as the target, y, and the reaction data as the feature, X, the data is prepared for machine learning analysis.

4.2 Classical Machine learning: PCA transformed SVM

For each simulation, X comprised of concentration time-series of thebaine, neopinone, codeinone and codeine, at each of 1000 timepoints. To make the data more tractable for further processing, for each element of X, the Euclidean distance between each species time series, and a time series of the same species for a reference reaction with no contaminant inhibition, was calculated; thereby reducing each X element to a vector of 4 Euclidean distance values.

This 4-dimensional data was then reduced to 3-dimensional and 2-dimensional representations using PCA transformation, allowing the data to be visualized to gain an intuition for its separation, and so that SVM classification could be compared between 4-dimensional and 3-dimensional data, as well as for a concatenated time-series representation as outlined in section 2.5.1.

Data Representation	4-d Euclidean distance	3-d Euclidean distance	4-d Euclidean distance to 3-d PCA	4-d Euclidean distance to 2-d PCA	Concatenated time-series
Binary Accuracy (%)	95.3	92.3	93.6	89.1	92.8
Multilabel Accuracy (%)	91.1	90.3	94.0	91.5	91.4

Table 3. SVM performance for binary and multilabel reaction step inhibition detection

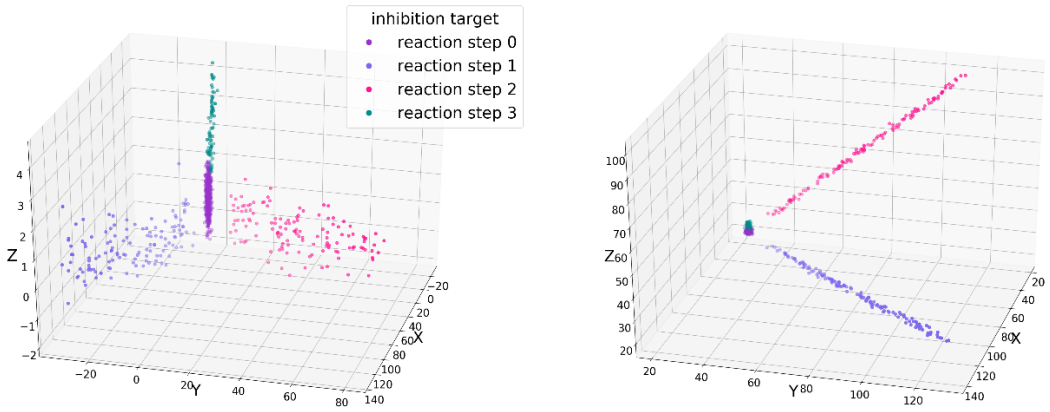

Figure 7: 3D plot of reaction instances and true inhibition labels for **a)** 3D PCA reduced 4D Euclidean-distance representation and **b)** 3D Euclidean distance representation

Applying a 60/40 training/test split, both binary and multilabel SVM models with linear separators were applied to each data representation. Although individual SVM models which utilize a single boundary vector can only separate spatial data into two groups, and hence can only perform binary-classification, multiple SVM models can be combined for multilabel classification. The 'One vs Rest' (OvR) method was the multilabel SVM appear used here; OvR operates by dedicating one binary SVM model to each class, separating data with the specified class label, from all other data. Combined, the model functions as a multilabel classifier.

Unsurprisingly, for the Euclidean distance and time series representations, binary classifiers performed better than the multilabel models. Notably, the PCA reduced representations both yielded better accuracy for multilabel classification however, and in fact were the best performing multilabel models. By plotting the 3-d spatial arrangement of the 3-d Euclidean distance and 3-d PCA reduced data, it is observed that the PCA data, designed to maximize variance between data, is better separated, and hence likely more amenable to linear separation, suggesting a cause for their improved performance (fig. 7). The multilabel performance improvement of the PCA representation over binary classification could in part be due to the capacity of multiple binary OvR classifiers to better capture dependencies between classes than a single binary classifier.

4.3 Modern Machine Learning: Deep Neural Networks

A deep neural network model was applied to each of the X representations of the previous section.

Data Representation	4-d Euclidean distance	3-d Euclidean distance	4-d Euclidean distance to 3-d PCA	4-d Euclidean distance to 2-d PCA	Concatenated time-series
Accuracy (%)	94.3	92.4	95.8	94.6	81.1

Table 4. DNN performance for multilabel reaction step inhibition detection

Although the concatenated time-series was the only data representation not involving a loss of information in conversion, it performed the worst. The retention of all the data itself may be an issue, as all present noise was retained. The scale of each concatenated time series (4000 time points) may also pose a difficulty for the model, with only a small portion of the data likely to contain information indicating the target label. The large data with small variance may also have made the model prone to overfitting, which was likely observed, with the model performing erratically, and worsening at epochs <15.

Conversely, 3-d PCA reduced Euclidean distance data, retaining less original information than all representations except its 2-d counterpart, performed best. This may indicate that the PCA reduction, designed to capture the most significant features of the data, effectively filtered out noise and irrelevant features from its input representation, retaining relatively more significant variations. The 2-d representation filtered out noise and irrelevant variations also, as indicated by its improved classification performance compared to its input 4-d data, though its worse performance than its 3-d counterpart indicates that the reduction to 2-dimensions necessarily diminished the retention of some significant features. That the pre-PCA processed Euclidean representations performed better than the time-series data may be indicative that this data-reduction mitigated the present noise, and reduced the overfitting of the large scale data.

5 Further Steps

The growing importance of industrial bioprocess application and the inherent variability and uncertainty of bioprocess characteristics, coupled with the positive results of this study, are encouragement for the benefit and feasibility of machine learning characterization and stochastic optimization of bioprocesses. The precise mathematical modelling of reaction kinetics lends itself well to optimization and characterization, and highly sophisticated optimization frameworks may reasonably be developed with this one as foundation. Some possible areas for development are as follows.

5.1 Objective Function

The objective function defined here considered only reaction parameters as decision variables. A more sophisticated function may be developed to model dependencies on other important industrial considerations, including economic factors such as expenditure of reaction equipment, operational expenditure of energy consumption and reagent costs, and time constraints, allowing for optimization of additional associated decision variables.

5.2 Kinetic Modelling

The Kinetics package provided an efficient foundation for modelling bioprocess reaction kinetics, however it was restrictive for more sophisticated reaction control, and more sophisticated modelling, such as modelling temperature and pH effects, required manual development. With some of this work already done, developing a kinetic modelling package

more tailored to parameter control and optimization would be feasible, and would allow for greater freedom in controlling the reaction.

Control of enzyme concentration, for instance, was severely restricted within the Kinetics model, and had to be done using non-ideal reverse engineered logistic function profiles. Development of a model allowing for greater freedom of enzyme concentration profiles would allow enzyme concentration to be more effectively optimized.

Setting profiles for reaction parameters such as temperature and pH was similarly restrictive, and more control over these profiles would similarly allow for greater optimization where different stages of reaction benefit from different reaction conditions. More sophisticated mathematical modelling of temperature, pH and other reaction parameter effects too, would allow for models of greater resemblance to their physical counterparts and hence more accurate modelling and optimization.

5.3 Hyperparameter and Model Optimization

The scope of this study emphasized the application of optimization and machine learning models to industrially important bioprocess procedures. The machine learning models used in this study are themselves subject to optimization, and while the models were extensively experimented with until satisfactory performance was achieved, their design and hyperparameters were not exhaustively optimized. Such optimization of the models themselves would lead to improved performance of their applications in this system.

Trying SVM separators other than linear separation, and adjusting the SVM c parameter, for instance, could boost performance. Similarly, the neural networks utilized could be further optimized by adjusting the architecture of the models (for instance, adding more layers, experimenting with dropout to reduce overfitting, and improving time-series learning with Recurrent Neural Networks), and by scrutinizing hyperparameters such as batch size and activation functions would lead to improved performance.

Such parameter optimization may be rigorously framed as optimization problems and explored using sampling techniques such as grid-search and monte-carlo sampling to find optimal hyperparameter values.

5.4 Stochastic Optimization

As demonstrated, the two-step stochastic optimization approach utilized here is subject to its own hyperparameter, $t_{partial}$, which may be subjected to a rigorous optimization process. Additionally, the stochastic approach used could itself be extended, with experimentation into incorporating multiple optimization stages beyond a two-step model, to the point of a dynamic optimization approach, to determine if this yields better bioprocess optimization.

6 Conclusion

Codeine production via the enzyme catalyzed multistep conversion of thebaine was maximized by modelling the kinetics of the reaction and formulating an optimization problem with reaction parameters T, pH, and enzyme introduction time coefficients as the decision variables.

This optimization problem formed the basis for a stochastic 2-step optimization method, designed to mitigate the challenge of optimizing the reaction given the uncertainty of the enzyme effectiveness – the enzyme k_{cat} values. The stochastic method utilized a deep neural

network model to predict $k_{cat\ actual}$ values based on incomplete reaction data, allowing for reoptimization and achieving a 2.6% average increase in codeine yield over non-reoptimized simulations.

SVM and DNN models were utilized to detect competitive inhibition by contaminants, from reaction concentration time-series, with both approaches achieving 95% accuracy in detecting which reaction step was inhibited. PCA transformed Euclidean distance representations of the original time-series data achieved the highest accuracy.

7 References

1. International Energy Agency. (2013). Technology Roadmap - Energy and GHG Reductions in the Chemical Industry via Catalytic Processes. IEA, Paris. Retrieved from https://www.iea.org/reports/technology-roadmap-energy-and-ghg-reductions-in-the-chemical-industry-via-catalytic-processes

2. Rissman, J., Bataille, C., Masanet, E., Aden, N., Morrow, W. R., Zhou, N., et al. (2020). Technologies and policies to decarbonize global industry: Review and assessment of mitigation drivers through 2070. Applied Energy, 266, Article 114848. https://doi.org/10.1016/j.apenergy.2020.114848

3. Organisation for Economic Co-operation and Development. (2012). OECD Environmental Outlook to 2050: The Consequences of Inaction. OECD Publishing, Paris. https://doi.org/10.1787/9789264122246-en

4. Wu, S., Snajdrova, R., Moore, J. C., Baldenius, K., & Bornscheuer, U. T. (2021). Biocatalysis: Enzymatic Synthesis for Industrial Applications. Angewandte Chemie (International ed. in English), 60(1), 88–119. https://doi.org/10.1002/anie.202006648

5. World Health Organization. (2019). WHO Model List of Essential Medicines, 21st List. Retrieved from https://www.who.int/medicines/publications/essentialmedicines/en/

6. International Narcotics Control Board. (n.d.). Estimated World Requirements for 2019 - Statistics for 2017. INCB Technical Report. Retrieved from https://www.incb.org/incb/en/

7. Symons, G. M., Hudson, C. J., & Livermore, M. L. (2009). Papaver somniferum with high concentration of codeine. WO 2009143574 A1.

8. Dastmalchi, M., Chen, X., Hagel, J. M., Chang, L., Chen, R., Ramasamy, S., Yeaman, S., & Facchini, P. J. (2019). Neopinone isomerase is involved in codeine and morphine biosynthesis in opium poppy. Nature Chemical Biology, 15(4), 384–390. https://doi.org/10.1038/s41589-019-0247-0

9. Li, X., Krysiak-Baltyn, K., Richards, L., Jarrold, A., Stevens, G. W., Bowser, T., Speight, R. E., & Gras, S. L. (2020). High-Efficiency Biocatalytic Conversion of Thebaine to Codeine. ACS Omega, 5(16), 9339–9347. https://doi.org/10.1021/acsomega.0c00282