Machine Learning Applied to Single-Molecule Activity Prediction

Kendric Hood
khood5@kent.edu
Department of Computer Science.
Kent State University
Kent, Ohio, USA

Li Zuo
lzuo4@kent.edu
Department of Chemistry and
Biochemistry. Kent State University
Kent, Ohio, USA

Hao Shen
hshen7@kent.edu
Department of Chemistry and
Biochemistry. Kent State University
Kent, Ohio, USA

Hanbin Mao
hmao@kent.edu
Department of Chemistry and
Biochemistry. Kent State University
Kent, Ohio, USA

Qiang Guan
qguan@kent.edu
Department of Computer Science.
Kent State University
Kent, Ohio, USA

ABSTRACT

Catalytic processes are used in about 1/3 of US manufacturing, from the field of chemical engineering to renewable energy. Assessing the activity of single-molecules, or individual molecules, is necessary to the development of efficient catalysts. Their heterogeneity structure leads to particle-specific catalytic activity. Experimentation with single-molecules can be time consuming and difficult. We purpose a Machine learning (ML) model that allows chemical researchers to run shorter single-molecule experiments to obtain the same level of results. We use common and widely understood ML methods to reduce complexity and enable accessibility to the chemical engineering community. We reduce the experiment time by up to 83%. Our evaluation shows that a small data set is sufficient to train an acceptable model. 300 experiments are needed, including the validation set. We use a well understood multilayer perceptron (MLP) model. We show that more complex models are not necessary and simpler methods are not sufficient.

CCS CONCEPTS

• **Applied computing**; • **Computing methodologies** → **Neural networks**; **Machine learning**; **Artificial intelligence**;

KEYWORDS

data sets, neural networks, chemistry, single-molecule, machine learning

ACM Reference Format:
Kendric Hood, Li Zuo, Hao Shen, Hanbin Mao, and Qiang Guan. 2023. Machine Learning Applied to Single-Molecule Activity Prediction. In *Workshops of The International Conference on High Performance Computing, Network, Storage, and Analysis (SC-W 2023), November 12–17, 2023, Denver, CO, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3624062.3626083

SC-W 2023, November 12–17, 2023, Denver, CO, USA

ACM ISBN 979-8-4007-0785-8/23/11…$15.00
https://doi.org/10.1145/3624062.3626083

1 INTRODUCTION

Catalysts are used to speed up chemical reactions by providing a lower energy pathway for the reaction. Catalytic processes are used in a wide range of chemical engineering fields including renewable energy, solar energy conversion, and batteries to name a few. Catalytic processes are used in over 1/3 of US manufacturing overall [9, 13]. The activity of individual molecules have a direct impact on how effective a catalysts is. Much work in the chemical engineering community has been spent on developing new methods to assess the reactivity of single-molecules [3, 12, 13]. These methods range from practical, in the laboratory, experiments to computational methods. Computational methods include simulation and theoretical calculations.

Each practical and computational method has major benefits and draw backs. The practical experimentation can give the most accurate results, as it allows real world measurements. However, they suffer from time and labour intensive costs to preform. Computational Methods are, comparatively, cheap in terms of time and labour. But they suffer limitations in scale.

Machine learning (ML), specifically artificial neural networks (ANN), remain underutilized [8]. ANN's offer a unique opportunity, when compared to the theoretical calculations, to extract features based on real experiments. Bridging the gap between the practical experimentation and theoretical calculation. However, they require the researcher to have a good understanding of current ML technology. Including how to adjust (add, remove, modify) layers of an ML model and tune hyperparameters. Learning how best to build, train, and evaluate a ML model serves as a barrier to applying ML methods in the field of chemical engineering and research. In addition, most are evaluated with *simulated experimentation data*, as freely available data sets are very limited or non-exists [8].

We evaluate how effective a simple multilayer perceptron (MLP) model is when applied to inferring single-molecule activity. MLP was chosen for its well understood properties and widely available documentation. We evaluated our MLP model with 6 single-molecule experiment data sets. Each set is a collection of *real* experiments under different experimental conditions. We used 200 experiments from each set to train 6 models, one for each data set, and 100 for validation. Each model had the same number of neurons per layer, number of layers, learning rate, loss function, and optimizer. This demonstrates that chemical engineers can use

our model without having to evaluate different hyperparameters. Thus, allowing chemical engineers with no background in ML to use the model. We summarize our contribution as follows;

- We offer a simple ANN model based on well understood principles, using a common and easy to use framework PyTorch [11]
- We evaluate our model with 6 different single-molecule experiment data sets, each with unique experiment conditions.
- We compare our model's accuracy with a naive approach that requires the same level of expertise on the chemical engineers part. We also compare our results with a more complex Recurrent Neural Network (RNN) model.
- We show that a naive approach is insufficient, and the RNN model retains similar or worse accuracy but with higher complexity.

The remainder of this paper is organized as follows. Section 2 introduces the background in the field of chemical engineering for both practical and computational methods. In Section 3 we review literature related to current methods for measuring single-molecule activity with ML. Section 4 describes our data set and problem definition. In Section 5 we offer a overview of our model, and explain model and training parameters. Section 6 we present our evaluation. Section 7 we overview our plans for future work and possible research directions moving forward. Section 8 we conclude the paper with a summary of our work.

2 BACKGROUND

Within this section, we present a concise overview of single-molecule assessment through both laboratory experimentation and computational methodologies. Practical experimentation is time consuming, while computational have hard limitations and are normally supplemented with some level of laboratory experimentation. Thus there is a need to accelerate the practical experimentation via machine learning (ML).

2.1 Practical Single-Molecule Experimentation

Here we overview some of the common practical methods used to assess the quality of a products particle-specific activity. The structural heterogeneity of molecules leads to distinctive catalytic activity associated with individual particles. Thus it is necessary to evaluate particle-specific catalytic activity due to this inherent heterogeneity.

Several methods have been employed to measure particle-specific activity including X-ray spectroscopy, Raman spectroscopy, and fluorescence microscopy [13]. Any form of spectroscopy uses the absorption and emission of light and other radiation to obtain measurements. X-ray absorption spectroscopy (XAS) has a spatial resolution to 10 nm. However it is limited in distance, posing challenges to nanoreactor design. It can also cause significant damage to the sample [13]. surface-enhanced Raman scattering (SERS) allows real-time monitoring of chemical reactions and can yield complementary chemical and structural information. A notable constraint is the necessity for a specific material within the experimental setup to amplify the Raman signal [13]. Fluorescence microscopy is a sensitive and informative technique in life and materials sciences that uses fluorescent indicators to allow researchers to watch cell physiology[12, 14]. The resolution of the acquired images remains a limiting factor [14]. Dual-trap laser-tweezers instrument is another method. Using focused laser beams, this method combines two optical traps, also known as optical tweezers, allowing researchers to hold, move, and manipulate nanoparticles with extreme precision [2]. This method can becomes problematic when hour-long observations are desired and results from large numbers of nanoparticles are required.

Each of these approaches demands a substantial number of experiments, ranging from hundreds to tens of thousands [12]. While each individual experiment consume minutes to hours, when scaled up to identify trends within the data, this process can extend to days, weeks, or even months, contingent on the total number of experiments required. In addition more than one method may be used, for example in [17] a combination of methods where used including fluorescence microscopy, observation with a dual-trap laser-tweezers instrument and fluorogenic probing.

2.2 Computational Methods

Computational methodologies have been extensively employed, encompassing density functional theory (DFT), high-throughput theoretical calculations, reactive force field (ReaxFF)-based molecular dynamics simulations, and kinetic modeling [13, 17]. These approaches predominantly revolve around simulations and theoretical calculations, encountering inherent limitations in terms of scale, size of catalytic systems under scrutiny, and available data resources [8]. Randall H. Goldsmith et al use a microfluidic trapping device to obtain prolonged solution-phase measurement of single enzymes in solution [7]. The result of these measurements where time-tagged photon detection data. For data analysis, they utilize a Matlab implementation of a change-point-finding algorithm, specifically tailored for their context. Gillespie et al purpose a stochastic simulation algorithm in [6]. A unique strength of their approach lies in its inherent capacity to avoid approximations for infinitesimal time increments. All computational method suffer from very small data sets if any are avalable at all [5].

3 RELATED WORK

In this section we briefly summarize the current application of machine learning (ML) in the field of chemical engineering and research, specifically single-molecule experimentation.

Masateru Taniguchi el al use ML to help identify single molecules in DNA and RNA [15]. Identification is done with histograms of single molecules in DNA, when these histograms overlap identification is problematic. The authors use a combination of fast Fourier transform (FFT) to create features and a support vector machine (SVM) to classify the molecules. Thomsen et al apply ML to single molecule Forster resonance energy transfer (FRET) with DeepFRET in [16]. DeepFRET is an open-source software package that implements a deep convolutional neural network (CNN) with a graphical user interface (GUI). The authors add a long short-term memory (LSTM) layer after the convolutional layers to assist in their classification. They provide a trained deep neural network (DNN). Their training data is 150,000 *simulated* FRET traces. Celik et al purpose a Deep-Channel model architecture in [4] to help

Table 1: AuNP concentrations, Data set size, and [minimum, maximum] value

AuNP concentration	Number of experiments (size)	Range (total events)
50	640	[35, 327]
100	499	[50, 382]
400	477	[50, 289]
800	564	[55, 321]
1200	794	[40, 335]
1800	888	[40, 314]

classify single molecule data of loss, or dysregulation, of ion channels. They combine 1D convolution layers with several pooling and LSTM layers. They use a *simulated* data set using the computational methods presented by Gillespie et al in [6]. Parker et al purpose a multi-stage machine learning workflow to identify the correct structure/property relationships of Pt nanoparticles [10]. Their workflow consists of unsupervised clustering via iterative label spreading (ILS), supervised classification via Extra Trees Classifier (ETC) and finally supervised regression with Extra Trees Regressor (ETR). They use a data set presenting 1300 optimized platinum (Pt) nanoparticle configurations for data-driven studies.

To the best of our knowledge these methods rely on a chemical engineers ability to chose hyperparameters and modify the model as needed. Choosing hyperparameters may be difficult but doable for simple models that have limited, or well documented, parameters to set. However, modifying the model, especially when dealing with more complex LSTM layers, can be overly difficult. Acting as a barrier for ML methods in chemical engineering.

4 SINGLE-MOLECULE EXPERIMENT DATA SET

4.1 Experiment Setup and Description

Our data set is a collection of experiments done by the Kent State University Chemistry and Biochemistry department, published in [17]. Their goal is to create and evaluate a artificial enzymes. They combine gold nanoparticles (AuNPs), a nanozymes, and DNAzymes. They used a combination of practical experimentation and computational methods to evaluate AuNP@DNA, the combination of gold and DNA. Multiple practical experimentation methods where preformed. We only consider the experiments that used a spin casted coronazyme under fluorescence microscope. Table 1 summarizes the different levels of AuNP concentration, the number of experiments run, and the range in total number of events in a single experiment, [min,max]. Each Concentration is considered a different data set. All other variables where the same.

4.2 Output and Goal

While each spin casted coronazyme experiment is preformed a recording is made, called a *movie*. Each movie is taken at a specified *frame rate* and is *60* minuets long. This frame rate is how many times a second the single-molecule (AuNP@DNA) is checked for a reaction, or *event*. All experiments where recorded with a frame rate of 20 frames per second (FPS). Each frame is numbered according to it's order in the recording, the first frame's 1, the second is 2 and so on. The finale output is a list of frame numbers in a movie where an event was recorded.[1] The total number of frames in a movie, ones with a recorded event, are then counted for the *total number of events* in the experiment. The total number of events is the desired result of the experment, this is what chemical engineers use to gauge how well a product will preform as a catalysts.

In summary, Our data set is a collection of experiments. Each element of our data set is an experiment's accosted movie. In total we have 6 data sets, each under a different concentration of AuNP. Our goal is to shorten the experiment time needed, by createing a ML model that reads in the first 10mins of a movie and then can predict, within a margin of error, the total number of events that movie would have at 60mins. Thus reducing the experiment time by 83%.

5 MODEL OVERVIEW

We purpose a multilayer perceptron (MLP) model. We demonstrate that MLP can be applied to different *real experimentation data sets* with no modification to the majority of its hyperparameters. Thus, demonstrating that chemical engineers with no background can train and use our model easily and out of the box. We will first present an overview of our model and model parameters, we will then discuss implementation details.

5.1 Multilayer Perceptron Model

Our MLP model it made up of 1 input layer, 2 hidden layers and 1 output layer. The input layer can vary in size depending on the size of the movie input to the model. That is to say, it is determined by the practical single-molecule experimentation used. More details on the input from our datasets is available in the following section 5.2. Our hidden layers have 700 neurons each. Our output layer is a single neuron. We use dropout between the input, first hidden, and second hidden layers. Neurons have a probability to be zeroed of 20%, there is no dropout between the second hidden layer and the output layer. We use ReLU activation functions for both hidden layers and the output layer. See figure 1 for illustration.

5.2 Implementation Details and Prepossessing

We implement our model with PyTorch [11]. We use their basic frame work, with no extraneous modification or alteration. This allows easy installation and setup of the necessary dependencies for running our model. We use mean squared error (MSE) as our loss function with sum reduction. We use L2 regularization for our weights. We use the Adam optimization algorithm for gradient decent with a learning rate of 0.000001.

Our prepossessing is kept as simple as possible in order to promote generalization and easy of use. Movies frames with no events are omitted from the original data. However, we are able to infer which movie frames are missing as we know the total run time of the movie, the frame rate and all the frames that did have an event. We add in missing frames in to the movies. Frames with events are recorded as having a value of 1 while frames without events are

[1]A copy of the data can be downloaded at https://github.com/khood5/SNN-DNA-project/blob/fe24f113c1fa389e4d640d202c44da934ebc5c71/DNA-SC23_workshop_paper/concentration_experments.7z

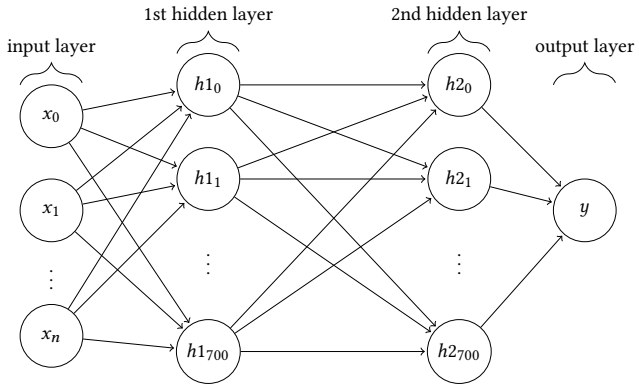


Figure 1: MLP model. x_0 to x_n is movie frame $0 - n$. y is the final predicted activity for the experiment if run for 60min

recorded as 0. That is each movie is converted to a bit string. Our model is thus trying to predict the total number of 1s in the bit string given a beginning portion of the bit string. That is, the total number of 1s is the target output. The original experiments where run for a total of 60min creating 72, 000 long bit strings. We use as input the first 10min of an experiment, or 12, 000 bits.

We make our implementation, preliminary work, and other research materials freely available at [1]

6 EVALUATION

We evaluate our model on all 6 data sets. For each set we train a new model, however all hyperparameters are the same for each model. The only variable is the data set. We compare our results to a naive approach and an RNN model. For each model we use 300 of the total available data for training and validation. Of the 300 training/validation set 200 is used for training and 100 is used for validation. The remaining data is used for testing and evaluation. A prediction is considered accurate if it is within a marrgin of error. For each approach we use a margin of error while training, validating, and testing of ±20 events, unless specified otherwise. A margin of error of ±20 events was selected as it is the smallest average difference between targets, total number of events per movie, among all the data sets.

6.1 Multilayer Perceptron Model Approach

Our MLP model is trained with 10000 epoch and check pointing whenever the accuracy on the validation set increases. Figure 2 shows how MLP preformed on each data set. The left bar is the average target value, average number of event in an experiment's movie, of the test set with error bars set as the standard deviation of that set. The right bar is the average predicted number of events across the test set, with error bars set as the standard deviation of the predicted number of events. The accuracies shown are thous of the test set.

For each of the data sets we see that the standard deviation for the predicted number of events has a slightly varied overlap with the actual standard deviation. This is due to difference in the distributions of values in each data set, that is the AuNP concentration of 1800 data set has a more varied distribution than the AuNP concentration of 50 data set. We also see that our model biases lower activity, i.e. lower predicted value. This is because our data sets have outliers that exaggerate the average. For a deeper look we also compare the mode and medium for each set of predicted vs actual in tables 2 and 3. Table 2 shows the mode of predicted number of events per test set vs the actual mode. Included is the count of how many times this value occurred. In table 3 we show the medium of predicted number of events per test set vs the actual medium. What is observable with the mode is that MLP can closely approximate the most common value. The predicted and actual mode are all within 14 events. The predicted medium values also correspond with the actual medium. Both the mode and medium shows that MLP bias towards the lower reactivity for each data set. This is because the higher reactivity levels are outliers not always captured in the training set. As we assume that the chemical engineer or researcher will not know ahead of time what would be considered an outlier for their data prior to training the model we do not filter these out in our data sets.

The high overlap in standard deviation for all data sets, the accuracy of the mode and medium all suggest that the MLP model was able to capture the underlying dynamic of each data set with the limited number of samples given. This means that a chemical engineer or researcher can run less experiments at full length to gain a fair easement of the reactivity of a given product, AuNP concentration. For an AuNP concentration of 1800 this would have reduced to time needed from 888 hours, one hour for each experiment to 398 hours, reducing the overall time by more than half.

6.2 Recurrent Neural Network Approach

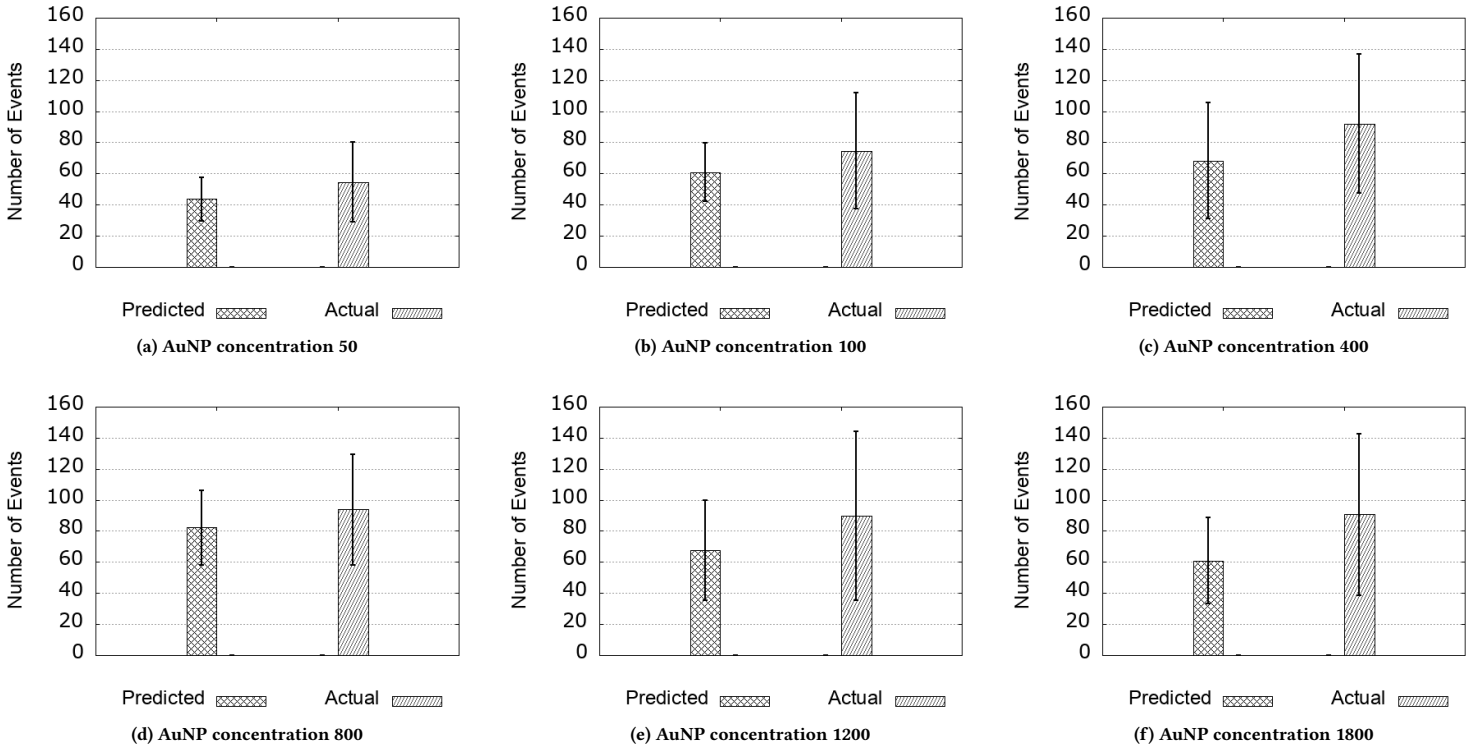
The Recurrent Neural Network (RNN) model is constructed using the PyTorch framework [11] and shares a structural resemblance to the MLP model. Comprising several layers, the RNN's architecture consists of an input layer, two hidden layers, and an output layer. Each movie is presented to the RNN in sequential segments, each spanning a minute in duration, (1,200 bits). This contrasts with the holistic approach of the MLP model, which takes the entire first 10mins of the movie (12,000 bits) as an input at once. As the RNN receives each minute of the movie individually, the output of the hidden layers in the previous time step (reflecting the preceding minute of the movie) serves as supplementary input for the current time step, capturing temporal aspect of the movie.

To elaborate, while each movie remains condensed to a fixed duration of 10 minutes (equivalent to 12,000 bits), the RNN processes these minutes incrementally. This means that at each time step, the network digests a new portion of the movie, allowing for the incorporation of temporal dependencies within the data.

Each hidden layer is still 700 neurons, with ReLU activation functions, the output layer is still a single neuron. All other hyper parameters (optimizes, wight regularization, etc) are the same as the MLP model.

6.3 Naive Approach

Our naive approach is to count the number of reactions in the first 10min for each experiment setup. Then multiply this number by 6, for the total movie length of 60min. This approach is intended to be



(a) AuNP concentration 50 (b) AuNP concentration 100 (c) AuNP concentration 400

(d) AuNP concentration 800 (e) AuNP concentration 1200 (f) AuNP concentration 1800

Figure 2: MLP performance on test data sets

Table 2: MLP predicted vs actual mode by data set (AuNP concentrations)

Data set	predicted mode, count	actual mode, count
50	38, 135	36, 30
100	52, 36	50, 13
400	47, 56	50, 7
800	70, 79	56, 13
1200	53, 160	40, 21
1800	47, 161	40, 17

Table 3: MLP predicted vs actual medium by data set (AuNP concentrations)

Data set	predicted medium	actual medium
50	38	47
100	55	64
400	55	78
800	71	81
1200	57	72
1800	52	72

the simplest method that could be used with the data set available and no additional information. It does not require any statistical analysis of the data set, as we believe a MLP model could be used without any such analysis. In this way it is a fair comparison.

Table 4: data set (AuNP concentration) vs accuracy of MLP, RNN models and naive approach all with error margin of ±20

AuNP concentration	Naive	*MLP*	RNN
50	51%	82%	82%
100	39%	80%	79%
400	9%	64%	38%
800	11%	75%	34%
1200	24%	65%	33%
1800	27%	61%	33%

6.4 Comparison

We summarize the accuracy for each data set in table 4. Here we can see that the performance of the MLP model is impacted by the distribution of the original data set. However, even in the worse case, AuNP concentration of 1800, the MLP model more than doubles the accuracy of the naive approach. We also see that MLP is competive with the more complex RNN model. Notice that the RNN model is not only more diffcult for chemical researchers to understand, but is more likely to need additional modification (adding/removing layers) to work well under different experimental data sets.

Table 5: Comparison between MLP, RNN models and naive approach under different error margins (average accuracy across data sets)

error margins	Naive	*MLP*	RNN
10	14%	38%	24%
11	15%	41%	26%
12	16%	44%	28%
13	18%	46%	31%
14	19%	48%	34%
15	20%	50%	36%
16	22%	52%	39%
17	24%	54%	41%
18	25%	55%	44%
19	26%	56%	47%
20	27%	57%	48%

Table 5 shows the comparison between accuracies of the MLP, RNN models with the naive approach. Accuracy is the average accuracy across all 6 data sets. The error tolerance is the number of events a prediction must be within, plus or minus, to be considered accurate. For example, an error tolerance of 10 means if the predicted number of events is 140 and the actual is 150 it is considered accurate, if the actual is 151 it is considered inaccurate.

As shown in the table the MLP model performs better than the more complex RNN model. Both ANN models perform much better than the naive approach.

6.5 Discussion

Based on our evaluation we have shown that MLP is a versatile model that can be applied easily by chemical engineers and researchers with little or no background in ML. We have shown that MLP performs better than a naive approach that requires the same amount of background knowledge of statistics as our MLP model. We have demonstrated that the more complex RNN model, that would require more background in ML to tune, has similar or worse accuracy to MLP. We show this in table 4 where MLP performs at worst 1.6 times better than the naive approach. While it matches the more complex RNN model accuracy on the least varied 50 AuNP data set, the easiest to predict, and perfoms better on all other data sets. In figure 2 we show that the standard deviation of the predictions has high overlap with the test set. This indicates that it accurately represents the experimental data at large, and suggests that it should be accurate for new experimental data. In short all you need is MLP.

The RNN model does however have one major advantage over the MLP model. As the input size grows the RNN model will scale better at the time of inference as it can load the data in fixed sequence, while the MLP model will need to load all of the input data at once.

It is our belief that this should demonstrate that simple ML methods can be applied to fields outside of computer science with easy. We offer a basic MLP model with predetermined hyperparameters that should be easy for chemical engineers and researchers to apply to their own work.

We also highlight that MLP performs well under different experiment settings, and can be used to accelerate pratical experimentation. Allowing chemical researchers to save time and labour.

7 FUTURE WORK

We hope to continue our work by examining other areas of chemical engineering can befit from ML. We also seek to further investigate models that take advantage of the time domain. We believe that such models could leverage patterns in time to offer new insights to the chemical engineers. Specifically, we aim to investigate the application of spiking neural networks to chemical engineering as we believe that can both take advantage of the time domain and also require fewer resources than comparable classic ML models, such as RNN or long-short term memory (LSTM) models. We of course, want to maintain the befits of simplicity. We seek to find a model that can generally fit a wide range of chemical research experiment data sets and take advantage of the time domain.

8 CONCLUSION

Catalytic processes play an important role in a large portion of US manufacturing. Chemical engineers have taken an interest in evaluating how single-molecules perform as catalysts. Evaluating single-molecules is a lengthy process that is costly in terms of time and labour. We have demonstrated that an MLP model can significantly speed up the evaluation of single-molecules by bridging the gap between practical and computational evaluation methods. We demonstrates that an MLP model performs significantly better than a naive approach, and at the same level as more complex ML models, namely an RNN model.

ACKNOWLEDGMENTS

The authors want to thank the support from NFS, 2212465, 2217104, and 2238734.

REFERENCES

[1] [n. d.]. SNN-DNA-project. https://github.com/khood5/SNN-DNA-project.
[2] Arthur Ashkin, James M Dziedzic, John E Bjorkholm, and Steven Chu. 1986. Observation of a single-beam gradient force optical trap for dielectric particles. *Optics letters* 11, 5 (1986), 288–290.
[3] Mark Bates, Bo Huang, Graham T Dempsey, and Xiaowei Zhuang. 2007. Multicolor super-resolution imaging with photo-switchable fluorescent probes. *Science* 317, 5845 (2007), 1749–1753.
[4] Numan Celik, Fiona O'Brien, Sean Brennan, Richard D Rainbow, Caroline Dart, Yalin Zheng, Frans Coenen, and Richard Barrett-Jolley. 2020. Deep-Channel uses deep neural networks to detect single-molecule events from patch-clamp data. *Communications biology* 3, 1 (2020), 3.
[5] Tao Chen, Fengxia Tong, Jorg Enderlein, and Zhaoke Zheng. 2020. Plasmon-driven modulation of reaction pathways of individual Pt-modified Au nanorods. *Nano Letters* 20, 5 (2020), 3326–3330.
[6] Daniel T Gillespie. 1977. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry* 81, 25 (1977), 2340–2361.
[7] Randall H Goldsmith, Leandro C Tabares, Dorota Kostrz, Christopher Dennison, Thijs J Aartsma, GW Canters, and WE Moerner. 2011. Redox cycling and kinetic analysis of single molecules of solution-phase nitrite reductase. *Proceedings of the National Academy of Sciences* 108, 42 (2011), 17269–17274.
[8] Yani Guan, Donovan Chaffart, Guihua Liu, Zhaoyang Tan, Dongsheng Zhang, Yanji Wang, Jingde Li, and Luis Ricardez-Sandoval. 2022. Machine learning in solid heterogeneous catalysis: Recent developments, challenges and perspectives. *Chemical Engineering Science* 248 (2022), 117224.
[9] Mahdi Hesari, Justin B Sambur, Xianwen Mao, Won Jung, and Peng Chen. 2019. Quantifying photocurrent loss of a single particle–particle interface in nanostructured photoelectrodes. *Nano Letters* 19, 2 (2019), 958–962.

[10] Amanda J Parker, George Opletal, and Amanda S Barnard. 2020. Classification of platinum nanoparticle catalysts using machine learning. *Journal of Applied Physics* 128, 1 (2020).

[11] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[12] Zoran Ristanović, Marleen M Kerssens, Alexey V Kubarev, Frank C Hendriks, Peter Dedecker, Johan Hofkens, Maarten BJ Roeffaers, and Bert M Weckhuysen. 2015. High-resolution single-molecule fluorescence imaging of zeolite aggregates within real-life fluid catalytic cracking particles. *Angewandte Chemie* 127, 6 (2015), 1856–1860.

[13] Justin B Sambur and Peng Chen. 2014. Approaches to single-nanoparticle catalysis. *Annual Review of Physical Chemistry* 65 (2014), 395–422.

[14] Michael J Sanderson, Ian Smith, Ian Parker, and Martin D Bootman. 2014. Fluorescence microscopy. *Cold Spring Harbor Protocols* 2014, 10 (2014), pdb–top071795.

[15] Masateru Taniguchi. 2020. Combination of single-molecule electrical measurements and machine learning for the identification of single biomolecules. *ACS omega* 5, 2 (2020), 959–964.

[16] Johannes Thomsen, Magnus Berg Sletfjerding, Simon Bo Jensen, Stefano Stella, Bijoya Paul, Mette Galsgaard Malle, Guillermo Montoya, Troels Christian Petersen, and Nikos S Hatzakis. 2020. DeepFRET, a software for rapid and automated single-molecule FRET data classification using deep learning. *Elife* 9 (2020), e60404.

[17] Li Zuo, Kehao Ren, Xianming Guo, Pravin Pokhrel, Bishal Pokhrel, Mohammad Akter Hossain, Zhao-Xu Chen, Hanbin Mao, and Hao Shen. 2023. Amalgamation of DNAzymes and Nanozymes in a Coronazyme. *Journal of the American Chemical Society* 145, 10 (2023), 5750–5758.