2020-08-03

Active learning for efficient microfluidic design automation

Version	Published version
Citation (published version):	David McIntyre, Ali Lashkaripour, Douglas Densmore. 2020. "Active Learning for Efficient Microfluidic Design Automation." https://www.iwbdaconf.org/2020/docs/IWBDA2020Proceedings.pdf. 12th International Workshop on Bio-Design Automation (IWBDA-2020). Online, 2020-08-03 - 2020-08-05.

https://hdl.handle.net/2144/41356

Boston University

Active Learning for Efficient Microfluidic Design Automation

David McIntyre
dpmc@bu.edu
Department of Biomedical Engineering
Boston University
Boston, MA

Ali Lashkaripour
lashkari@bu.edu
Department of Biomedical Engineering
Boston University
Boston, MA

Douglas Densmore
dougd@bu.edu
Department of Computer and Electrical Engineering
Boston University
Boston, MA

1 INTRODUCTION

Droplet microfluidics has the potential to eliminate the testing bottleneck in synthetic biology by screening biological samples encapsulated in water-in-oil emulsions at unprecedented throughput [2]. Sophisticated screens require functional and complex devices that perform exactly as designed. Effective performance characterization and predictive design of droplet microfluidic components has been hampered due to low-throughput and expensive fabrication with standard soft lithography techniques. This has limited droplet microfluidics to proof-of-concept devices. Even when some of these barriers are removed through rapid prototyping, developing a robust dataset to effectively represent all parameters as a "lookup table" is near impossible.

One solution to explore how design parameters affect performance in microfluidics is through machine learning. Although machine learning can make accurate microfluidic design automation tools, standard development pipelines require a large, naively-generated training set **(Figure 1, left)**. These approaches become intractable in cases where generating labeled data is particularly time or money-intensive.

Training data-restricted models can benefit from active learning algorithms, in which the model queries an "oracle" (the user) during the training process to only generate or label the data it predicts would best improve model performance **(Figure 1, right)** [6]. Through structured data generation, the amount of training data needed for an accurate model can be significantly reduced, speeding up the time to predictive design and eliminating unproductive user efforts.

Here, we present a novel experimental paradigm to rapidly generate microfluidic design automation tools. Efficacy of this method was tested against a previously generated dataset for a droplet generator design tool (DAFD) [3, 4]. This method can be extended to additional microfluidic components or fabrication methods, provided a method for data generation is high-throughput enough.

2 RESULTS

Efficient data generation for active learning algorithms necessitates evaluation of the quality of unlabeled data (informativeness and/or diversity) used in each round of model training [6]. Informativeness is the predicted amount that a specific datapoint can improve the model, whereas diversity is the spread of the data used across the design space. Here, previously generated data is pooled as "chips" (i.e., all datapoints generated using the same microfluidic device), which includes $\tilde{1}000$ datapoints pooled as 43 chips that were fabricated with previously developed rapid prototyping workflows [5]. Data was pooled in this way to minimized future microfluidic devices that need to be made, the most resource-intensive step in the data generation process.

To initially explore the advantages of active learning, three data quality metrics were implemented: (1) random choice; (2) greedy sampling (GS), which chooses the most different chip to the training set [7]; and (3) query by committee (QBC), which chooses the most informative chip [1]. In all cases, the model is seeded with one chip randomly picked from the training set.

In greedy sampling, optimal candidates are chosen by the maximum average distance of the geometric features of the chip from the existing labeled training set **(Equation 1)**. All features of each datapoint is normalized to avoid bias.

$$d_{dp} = \frac{1}{N_{train}} \sum_{i=1}^{N_{train}} ||\mathbf{x_{dp}} - \mathbf{x_i}|| \tag{1}$$

Alternatively, the potential "information" gained through adding a specific datapoint can be evaluated with QBC **(Equation 2)**.

$$I(x) = \frac{1}{P} \sum_{p=1}^{P} \frac{(\hat{y}_p(x) - \bar{y}(x))^2}{\bar{y}(x)} \tag{2}$$

In QBC, the quality of each unlabeled point is evaluated by the variance of each prediction across P regressors. Each regressor is trained on a bootstrapped collection of the training set. Points with high information are estimate to be those with a large variance in predicted value. In this study, results were normalized by the mean prediction to avoid bias for larger values. Each iteration, the chip with the max average variance was chosen as the next datapoint.

These methods were implemented into the DAFD framework, consisting of 4 neural networks (NN) predicting the droplet size and generation rate in the dripping and jetting

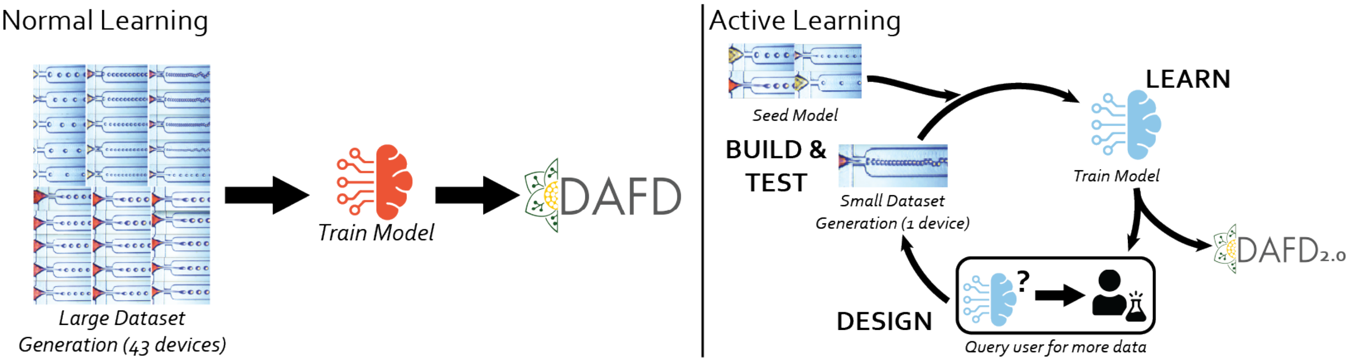

Figure 1: Comparison between a normal machine learning pipeline (left), in which a large training set it fed into the model and active learning (right), in which the model queries a user for more data after "seeding" with a small initial dataset.

regimes **(Figure 2)**. Regressor accuracy was tested on a randomly partitioned 20% of the total dataset and evaluated using root-mean-square error (RMSE). Across all NNs, GS performed better than or equivalent to random choice. This was distinct in regime 2: an RMSE of 0.9 was achieved with 100 and 150 fewer datapoints for size and generation rate, respectively. This indicates that diversity of data is the most important characteristic of the training set. QBC had improved performance than random choice in some cases, however, performed worse when predicting droplet size in regime 2. Poor performance by QBC could be from poor initialization or balancing data selection over the 4 regressors.

3 CONCLUSION & FUTURE WORK

Here, we have shown that active learning can provide a design framework to streamline the experimental workflow for developing design automation tools. While model improvement was variable while simultaneously training four regressors, this framework can be improved through development of a more sophisticated algorithm accounting for both diversity and informativeness. Model seeding could also be improved through formal Design of Experiments (DoE), giving a high-quality base model for further data generation and model evaluation cycles.

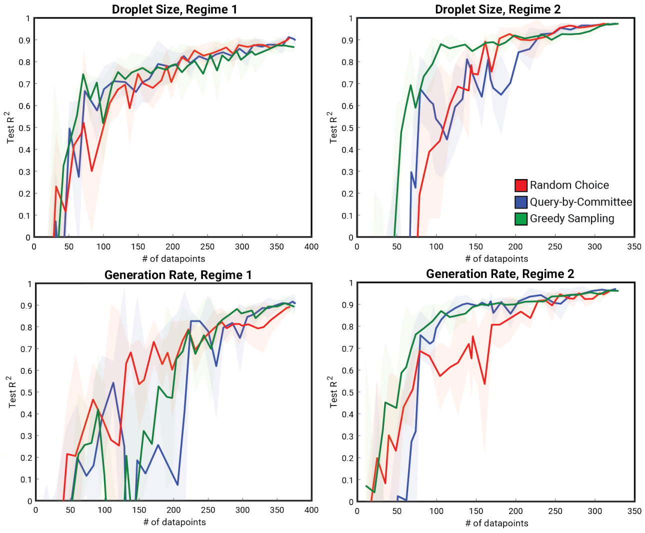

Figure 2: RMSE error across all four regressors with different active learning algorithms. Curves and shaded regions are the mean and standard deviation, respectively (N=3)

While this first study has used an existing dataset exploring how microfluidic device parameters affect droplet generation, we can extend this approach to *de-novo* models of different components (droplet sorter, merger, etc.). This method can also be used to rapidly perform transfer learning for using a device with custom fluid classes or different fabrication methods. Development of a streamlined pipeline for design automation is a necessary step for the standardization of microfluidics, and further spread its adoption by non-experts.

REFERENCES

[1] Burbidge, R., Rowland, J. J., and King, R. D. Active Learning for Regression based on Query by Committee. Tech. rep.

[2] Guo, M. T., Rotem, A., Heyman, J. A., and Weitz, D. A. Droplet microfluidics for high-throughput biological assays. *Lab on a Chip 12*, 12 (may 2012), 2146.

[3] Lashkaripour, A., Rodriguez, C., Mehdipour, N., McIntyre, D., and Densmore, D. Modular microfluidic design automation using machine learning. 11th International Workshop on Bio-Design Automation (IWBDA-19).

[4] Lashkaripour, A., Rodriguez, C., Ortiz, L., and Densmore, D. Performance tuning of microfluidic flow-focusing droplet generators. *Lab on a Chip 19*, 6 (2019), 1041–1053.

[5] Lashkaripour, A., Silva, R., and Densmore, D. Desktop micromilled microfluidics. *Microfluidics and Nanofluidics 22*, 3 (mar 2018), 31.

[6] Wu, D. Pool-Based Sequential Active Learning for Regression. Tech. rep., 2018.

[7] Yu, H., and Kim, S. Passive sampling for regression. In *2010 IEEE International Conference on Data Mining* (2010), IEEE, pp. 1151–1156.