ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

Machine learning in research and development of new vaccines products: opportunities and challenges

Paul Smyth, Gaël de Lannoy, Moritz Von Stosch, Alexander Pysik, Amin Khan

GSK, Technical Research and Development
Rue de l'institut 83, 1332 Rixensart, Belgium

Abstract. Modern high-throughput technologies deployed in research and development of new vaccine products have opened the door to machine learning applications that allow the automation of tasks and support for data-driven risk-based decision making. In this paper, the opportunities and the challenges faced for the deployment of machine learning algorithms in the field of vaccines development are discussed.

1 Introduction

Aside from clean water, no other health intervention saves more human lives than vaccines. The development of vaccines is a complex process due to the natural variability in biological material and the many steps required in the manufacturing of biological medicines that do not allow an exact replication of the molecular micro-heterogeneity. Furthermore, the development of modern vaccines, for diseases such as HIV or respiratory syndromes, require more advanced technologies during the exploration of a vaccine candidates than those used in previous generations. These new technologies typically generate a much larger amount of data and therefore need much more advanced data processing techniques to distinguish the signal from the noise. In this paper we will review the opportunities and the challenges for the deployment of machine learning techniques during the research and development of new vaccines products.

2 Opportunities

In this section, a non-exhaustive review of machine learning applications in the process of vaccines development is discussed. This section is separated in three parts: firstly, the drug substance part (i.e. generation of the active substance, the antigen), secondly the drug product part (i.e. the formulation of the drug substance up to its final container form) and finally the analytical methods used to characterize the product.

2.1 Drug substance - Cell culture and bacterial fermentation

Images of the cell culture in case of viral diseases, or of the fermentation production for bacterial diseases, allow to keep track of the yield and of the quality of the harvest. Machine learning techniques such as image segmentation and recognition are used to automatically detect the amount and the aggregation of cells or of bacteria colony forming units, but also their nature.

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

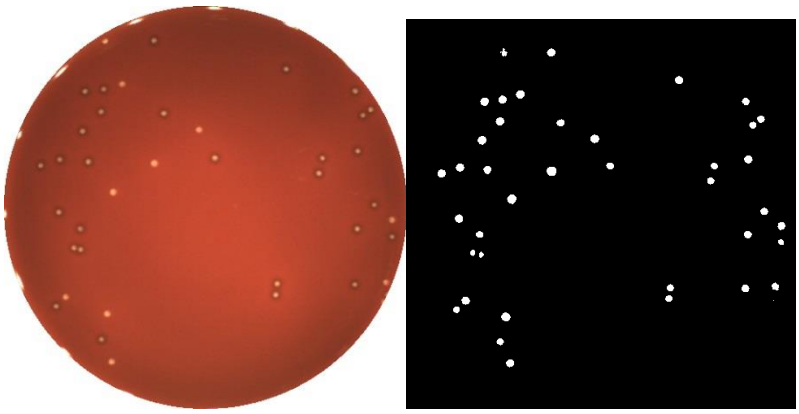

Figure 1: A raw image of a petri dish displaying colony forming units and the related mask used for training a UNet architecture.

Typical algorithms are the well-known convolutional neural network architectures, such U-Net [1], applied to images of colony forming units on petri dishes (see Figure 1). This case provides multiple interesting directions for future work, including the application of generative adversarial networks for data augmentation and the role of non-translation symmetries e.g. spherical and group-equivariant convolutions (see e.g. Cohen et al [2]).

2.2 Drug product – Formulation and lyophilization

A large proportion of vaccines are lyophilized, i.e. freeze-dried, in order to improve their thermostability. While the physics of freeze-drying is well understood in idealized situations, real-world lyophilization is complex and the subject of many interesting studies (see also hybrid models, below) [3]. From the point of view of machine learning, the main opportunities here are two-fold: the analysis of the multivariate time-series data being generated during the lyophilization process, and analysis of the images of the ‘cakes’ produced. The goal of our work here is to develop a deeper knowledge of the freeze-drying process across multiple vaccines and to assess the feasibility of online anomaly detection algorithms to assess the product quality and performance.

2.3 Analytical methods

Vaccines research and development, and the subsequent manufacturing process, dedicates the large proportion of time to ensure the safety and efficacy of its products through analytics methods. Indeed 70% of the manufacturing throughput time for any given vaccines is used just for product quality control. The majority of the analytic techniques employed are highly refined biotechnology methods that are already rigorously controlled and validated in accordance with international regulatory guidance documents. However some more qualitative methods remain and play a valuable role in vaccine development, with a classic example being sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-Page) western blot gel. In this method,

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

molecules of different mass are separated and the resulting bands are compared to 'eye-ball' the molecules that appear. The opportunity here is again two-fold. One could imagine using image recognition techniques directly on images on the classic, qualitative gel. Arguably a more promising direction would be the to develop a data analysis pipeline for more modern methods that measure the intensity of the multiple bands after the addition of fluorescent stains.

Additional, interesting use cases in this area include image segmentation for electron microscope imagery, the analysis of quantification of images with cluttered or agglomerated cells, the analysis of holographic images of live and dead cells, and automated data preprocessing for flow cytometry data.

3 Challenges

Despite the numerous appealing opportunities previously listed, many practical challenges make it difficult to apply machine learning models in the medical field and particularly for biological products like vaccines. In this section, the major encountered challenges are discussed.

3.1 **Variability of biological products**

Biologicals products have a large intrinsic variability compared to traditional drugs. The data is therefore largely polluted by multiple additive variability components such as batch to batch variability due to raw materials and to the manufacturing process itself, as well as the analytical variability of the assays used to characterize the product. For instance, in-vivo immunological assays used to characterize the potency of the vaccine by ELISA can typically have a variability as high as 50%.

Because of this high intrinsic variability, the models require high sample sizes to reach reliable accuracy and precision. However, each run of a process involves many raw materials and lots of steps to yield only one batch in the end. Each single observation can be thus very expensive to generate both in terms of resources. For example, one cell culture run usually takes up to 20 days for the simpler ones.

3.2 **Data collection**

The vaccines development process is so complex that is not easy to link all the systems, and to have annotated results for supervised classification algorithms. For example, in the lyophilization example above one has a mix of structured and unstructured metadata for the particular lyophilization cycle, the process time-series data, images of the resulting 'cakes' and their analytic characterizations. This leads to data management challenges, along with the more technical aspects of data augmentation for supervised machine learning problems.

3.3 **Regulatory considerations**

Like traditional drugs intended at human use, vaccine Companies face a strict external and internal regulation. External regulations are governmental bodies which control the

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

registration of new products and changes made to commercial products, and edit guidelines that Companies must comply with such as ICH (International Committee on Harmonization) framework. In the USA, the authority for vaccines is the CBER (Center for Biologics Evaluation and Research) and in Europe the EMA (European Medicines Agency). Internal regulation within companies ensure that the requirements edited by authorities are met in their facilities, through policies and standard operating procedures.

Given this context, one can understand that generating training data in a commercial facility, for example where the process parameters are varied on purpose, is a highly difficult task since each variation might be subject to regulation aspects. For this reason, the training data is typically generated at small scale in a R&D facility, but then comes the question on the scalability and the generalization of the classifier to real-life data.

Furthermore, machine learning classifiers are statistical models facing uncertainty in their results and in the estimation of their parameters. There is a high probability that the parameters of the model are not going to be identical when estimated on two distinct training datasets generated by the same process, so does their classification performance on an independent dataset. For this reason, it might be difficult to convince regulators on the reliability of the classifier if it is intended for decision making purposes.

3.4 **Mechanistic modeling versus statistical modeling**

In the industrial context, modeling, whatever its nature, must add value rather than being a mental exercise. Hence, the most appropriate modeling approach should be selected case-by-case considering the likely benefits and risks, e.g. likelihood that the model will fail to describe the system.

It is beneficial to develop mechanistic models when the impact of the factors on the system response is understood and can be mathematically described [4]. In order to exploit the model for optimization/control, the model parameters should not change from experiment to experiment. Consider, for instance, modeling the release of a drug from a carrier-particle into the bloodstream [5]. The cumulative drug release can relatively easy be described by material balances, considering different transport phenomena. The model parameters can be fitted from data of a very limited number of experiments (one or two) as the impact of particle size on the release rate is explicitly described, i.e. the structure of the model is known but the parameters are not - parametric modeling. The model will remain valid as long as the assumptions hold and it can then be used to optimize the size of the particle such that a desired release profile is obtained. However, if the drug or composition of the particle is changed, then the model parameters would have to be updated as otherwise the model cannot describe the behavior of the system.

The development of statistical (data-driven) models is beneficial when the systems is poorly understood including the impact of the factors on the system response, but experiments can be performed or data (with variations of the factors in these data) are

available [4]. The quality of the data is important, since typically both, the structure and parameters of the model have to be inferred from the data - nonparametric modeling. In the drug-release example, one could imagine changing the composition of the particle and experimentally study its impact on the drug release rate. Subsequently, the data can be used to build a model that links the changes in composition to those in the release rate, allowing to optimize the composition as long as the optimal composition is similar to the already studied ones. However, changes in the particle size or drug would render the model useless.

Hybrid models, which combine statistical and mechanistic models, should be developed when some parts of the system are mechanistically understood, but others are not [4,6]. Data are required to build the statistical model, representing the unknown part, and they need to contain variations in those factors whose impact is to be described by the statistical model. However, as the impact of some of the factors is described by the mechanistic model, overall the data requirements are reduced. Consider the changes in the particle composition for the development of the statistical model as well as the impact of the particle size described by the mechanistic model, both described before. The parameter(s) in the mechanistic model that are specific for one particle composition could become functions by describing the effect of changes in the composition on these parameters using the statistical model. Without performing additional experiments, a model that can describe the impact of changes in both, particle composition and size, on the drug release rate has become available [5].

Prior to starting the modeling exercise, as might have become apparent, one should assess what knowledge is available and how many experiments, and in which setting, one is willing to perform. However, most of all one should consider what the added value of the developed model would be and adopt the modeling objective accordingly [7].

4 Acknowledgements

The authors would like to acknowledge our collaboration with Prof. John Lee and Thomas Beznik of UCL (Université catholique de Louvain) on image segmentation and recognition.

5 Conflict of Interest Declaration

All authors have declared the following interests: All authors are employees of the GSK group of companies. Gaël de Lannoy, Moritz Von Stosch, Alexander Pysik and Amin Khan report ownership of GSK shares and/or restricted GSK shares.

All authors were involved in drafting the manuscript or critically revising it for important intellectual content.

References

[1] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015.

[2] T. Cohen, M. Geiger, J. Köhler and M. Welling, Convolutional Networks for Spherical Signals. https://arxiv.org/abs/1709.04893.

[3] E. Bourlès, G. de Lannoy, B. Scutellà, F. Fonseca, I. C. Trelea, S. Passot, Scale-Up of Freeze-Drying Cycles, the Use of Process Analytical Technology (PAT), and Statistical Analysis, in Lyophilization of Pharmaceuticals and Biologicals: New Technologies and Approaches, Springer 2019.

[4] J. Glassey and M. von Stosch, Hybrid Modeling in Process Industries, CRC Press, 2018.

[5] C.R. de Azevedo, M. von Stosch, M.S. Costa, A.M. Ramos, M.M. Cardoso, F. Danhier, V. Préat and R. Oliveira, Modeling of the burst release from PLGA micro-and nanoparticles as function of physicochemical parameters and formulation characteristics, International journal of pharmaceutics, 2018, 532 (1), 229-240.

[6] M. von Stosch, R. Oliveira, J. Peres, S.F. de Azevedo, Hybrid semi-parametric modeling in process systems engineering: Past, present and future, Computers & Chemical Engineering, 2014 60, 86-101.

[7] D. Bonvin, C. Georgakis, C. C. Pantelides, M. Barolo, M. A. Grover, D. Rodrigues, R. Schneider, and D. Dochain; Linking Models and Experiments, Industrial & Engineering Chemistry Research, 2016, 55, 25, 6891-6903.