Bioprocess Parameter Optimisation and Cell Culture Intelligence: Machine Learning Approaches to Enhanced Biopharmaceutical Development and Manufacturing

Dr. Byung-Woo Kim, Professor of Automotive Engineering, Korea University, South Korea

1. Introduction to Biopharmaceutical Research and Machine Learning

The pharmaceutical industry is key to the healthcare ecosystem; within this, biopharmaceuticals or biological drugs are a significant domain. They are valued for their importance in treating patients with malignant diseases, genetic disorders, or producing vaccines against diseases. This field, compared to conventional pharmaceutical products, deals with drugs derived from living organisms that are far more complicated and more difficult for quality assurance. In this light, we explore the potential of operationalizing machine learning techniques for enhancing biopharmaceutical research. Such a transformation has never been more relevant as computational technologies and analytical methods meld with core biological research.

Biopharmaceutical research has relied on traditional methods, thereby restricting pathways to the development of biological drugs. However, this large-scale unstructured data can prove vital to building enhanced biopharmaceutical research methodologies designed for artificial intelligence-embedded tools. This is specifically relevant for the knowledge discovery domain that researchers and practitioners straddle across discussion sections of biology and machine learning, thus forging a bridge between them. Coupling disease treatment catered via biopharmaceutical products to better outcomes requires the successful integration of traditional sub-disciplines with those that cater to the domain. Mapping optimal and novel avenues bolstered by machine learning contributes directly to this process. This addition is necessary, educating practitioners in the biotechnology industry to understand and delve into simplistic machine learning methodologies, especially in text and signal-based biopharmaceutical research. Regulatory efficacy and efficiency are warranted by replacing the traditional system with in-silico modeling methodologies. In conclusion,

tools and techniques unique to the biopharmaceutical research domain can be carved with the use of machine learning linked algorithms and strategies. This can be used as a complementary methodology to traditional measurements for recruitment, clinical performance, and outcomes for removing population-level disparities. In the subsections of the subsequent section, we discuss multiple applications of machine learning in the biopharmaceutical domain to reduce complexity and engage researchers better. We could then present an ensemble of methodologies designed within various levels of expertise attempting to solve the issues in the problem statement.

1.1. Overview of Biopharmaceuticals

1.1.1. Conceptual Consideration Biopharmaceuticals, also known as biologics, are a special class of drugs, differing in their physicochemical properties and production approach compared to traditional pharmaceuticals. Rather than small organic molecules, biopharmaceuticals are typically large, complex molecules. Biotechnology, utilizing biological systems and living organisms, is mainly employed in their production. Some popular examples include antibodies, hormones, enzymes, and nucleic acids. Monoclonal antibodies are indeed one of the most rapidly expanding market segments of biopharmaceuticals. Other examples range from vaccines to cell and gene therapies. These classes of drugs have remarkable potential for personalized medicine, as exemplified by cancers treated with chimeric antigen receptor T cells or congenital diseases treated with gene therapies. These novel therapies are revolutionizing the treatment of diseases for which there were previously limited management options.

1.1.2. Rationale for Special Session: Challenges Faced by Biopharmaceuticals Biological systems are multifaceted and intricate, and the translation of therapies into these biological systems involves several challenges. Medicines should comply with standards, and their structures, compositions, and functions have to be characterized by several analytical methods. Because biopharmaceuticals are subject to stringent regulatory review and authorized by regulatory agencies before their entry into the market, they increase public health through the regulatory assessment of optimal pharmaceutical development, approval, and post-marketing safety. Moreover, their structure can be highly heterogeneous, and the relationship between higher-order structure and biological activity is not well known due to denaturation and pathological unfolding of the protein; therefore, completely characterizing their structures and the

pattern of their sequences becomes extremely complex. Additionally, small alterations can have a profound effect on biopharmaceutical efficacy and can compromise safety. Machine learning tools or analytical methods can assess an overall global tertiary or denominated structure. To date, some predictive models for the accurate determination of biopharmaceuticals with artificial intelligence tools have been developed.

2. Importance of Predicting Biologic Drug Performance

Biologics account for an increasing number of new drugs developed today. These complex molecules may come in the form of monoclonal antibodies, fusion proteins, hormones, and nucleic acid-based products. Understanding their behavior, before and during development, is critical for predicting their potential success during clinical trials and after marketing. Predictive models would provide significant advantages for timely decision-making, resource allocation, and risk reduction in the pharmaceutical research setting. Differentiating between biologics that will show desirable or unfavorable behavior in the intended clinical settings is complicated by the molecules' large size, unique variable distribution of amino acids in their structure, and a variety of post-translational modifications. Classical approaches for the prediction of pharmaceutical successes, which involve statistical or mechanistic methods, have had limited success in this domain. Data-driven machine learning approaches that capitalize on the unique features of biopharmaceuticals for prediction are expected to outperform other prediction methods given their ability to model complex, non-linear relationships.

The stakes of misidentifying biologics with inadequate performance are high. Failure of these agents in the clinic during the clinical trial phase or upon reaching the market comes with significant public health and financial implications. Development of drugs can cost hundreds of millions of dollars and takes more than a decade. Establishing paradigms and tools that predict performance prior to drug development can significantly reduce investment costs and time to bring successful biopharmaceuticals to the market. ML-based prediction models can, in principle, support companies in rapidly making the best choices about a biologic to pursue before advancing it through costly development programs. Data-driven models offer the advantage of learning iteratively from information, automatically adjusting their predictions with new accumulated clinical and preclinical data, ultimately providing more accurate predictions than classical methods. Can we make a promise that an ML-based model will outperform any

other available options in predictive equations? Unfortunately not, especially for systems as complex as the human body or the components of a biologic, where multiple combined interactions of organs and tissues in the case of the body or regions in the case of the biologic could be at play in determining overall performance. Highly accurate prediction is hard to guarantee at a prespecified acceptable uncertainty level in testing, even if the underlying signals may be predictive using systems that are generating the data.

Our objective should not be in defining thresholds for highly accurate predictions. It is in assembling a predictive framework that, like a building, is fortified with the right ingredients to stand the test of time. In the absence of this level of accuracy for predictions, predictive platforms can help us better understand the much broader questions of the associated data cohorts. For biologic drug products, their predictability involves the extensive data concentrations around the effect of the biological surroundings and the design and class discrimination of the product itself. We believe that one stark conclusion of the capabilities of this prediction in biopharmaceuticals is that data and model-agnostic features for processing and prediction will be unable to produce the globally most predictive models. In response, the capabilities of models in biopharmaceutical research can best be harnessed to the unique and high-dimensional data properties of the biological objects in question, biologics themselves. In biopharmaceutical research, approaches that adapt to the new possibilities of biological data are most relevant.

2.1. Challenges in Traditional Drug Development

Biologics are used to analyze, treat, and prevent diseases that were previously hard to cure using small molecule drugs. A biologic drug is characterized by its high molecular mass, diversity, and in some cases, life systems. The drug approval rate for all biologic candidates undergoing clinical development is only 17 out of 100. The cost of successfully creating a drug and bringing it to the market was estimated to be $2.6 billion. Developing new biologics is a long and expensive process. As a result, there is an increasing risk that poor investment decisions in specialized facilities and talent will harm developers.

Nevertheless, there are a few drawbacks to traditional methods of drug production. The lack of an in-depth understanding and validation of the interactions between a drug and

its target often contributes to the failure of complex molecule drugs in the development phase. Along with the growth of pharmaceutical businesses, large quantities of resources are expected to be used to ensure that a drug is effective, suitable, and prepared for customer consumption. Similarly, holding a biological entity via biophysical and structural analyses at a large scale in the development cycle is both expensive and traditionally time-consuming. As a result, large batches of protein production, which is a time-consuming part of the creation process and can be demanding in terms of resources, are one explanation for the high failure rates. Scaling up protein synthesis to match the anticipated customer demand must be addressed. Whether embracing the growing public demand for growth, innovative ways are continually sought to reduce the number of times a candidate fails and minimize the high expenditure of the drug production process. In conclusion, increasing competition in the biopharmaceutical sector has become complex. Indeed, rising consumer demand pressures the industry to devise new strategies to satisfy the rate of development of the pharmacological sector.

3. Fundamentals of Machine Learning in Biopharmaceutical Research

Today's science and drug development are rapidly adapting to the era of big data, and machine learning approaches are increasingly involved in the related areas of research and applications. Machine learning has many facets ranging from supervised to unsupervised and reinforcement learning. A widely applied category of machine learning in pharmaceutical and biotechnology research is supervised learning. "Supervised" refers to the learning process being guided by labels associated with input vectors to guide an algorithm in learning a mapping from input features to the corresponding output value. This type of machine learning is commonly applied to address classification or regression problems in pharmaceutical research. Unsupervised learning, on the other hand, is used to draw inferences from datasets consisting of input vectors without any labeled output. The methodology is often used for clustering a dataset into meaningful partitions or for discovering patterns, regularities, or statistical associations from the dataset. Reinforcement learning is a different paradigm, and it is loosely applicable to pharmaceutical and biopharmaceutical research in adaptive strategies.

Machine learning approaches have been increasingly applied in various fields of pharmaceutical and biopharmaceutical research to support decision-making related to both research and development. Machine learning models in pharmaceutical and biopharmaceutical applications have been used as tools for drug design and lead optimization, target selection, safety assessment, and more. Data preprocessing, data splitting, feature selection, and model validation are important steps in the development and deployment of machine learning methods. All models built with machine learning approaches bear the risk of some kind of bias and assume some level of certainty regarding the training dataset. Some applications of machine learning, including deep learning strategies, do not support explaining predictions made on new, previously non-assessed cases. It is important to be aware of potential biases and limitations when introducing machine learning methods in any study and to validate a model on independent datasets preferably split from an extermination cohort.

3.1. Supervised, Unsupervised, and Reinforcement Learning

Machine learning, a branch of artificial intelligence, has become a powerful tool in enhancing the workflow in virtually all biopharmaceutical research areas. This section further segments machine learning methodologies into three primary categories: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning employs labeled data to establish the relationship between input and output, potentially leading to predictive models or classification insights into the outcome. Unsupervised learning, on the other hand, is utilized when exploring the clustering structure of a dataset is an end in itself. By allowing the algorithm to analyze unlabeled data in a manner where there is no precise notion of "correct" results, data points can be parsed in useful ways. As the majority of research questions in biopharmaceutical laboratories fall outside the construct set by supervised classification, unsupervised approaches are expectedly highly popular. Reinforcement learning, which typically yields dynamic outcomes, is an emerging approach for which there can be no definitive labeled data. This type of learning is also more related to systems optimization that can be useful in drug development processes since biopharmaceutical research areas are strongly converging towards process-oriented drug target identification and screening of drug candidates.

In the supervised learning approach, and particularly in the area of drug discovery research, input data are structured descriptors or "features" relating to critical drug-like properties of small molecules used for creating predictive algorithms to predict the likelihood of a desired output, such as an in vitro or in vivo activity or tasked classification of the biological target site. The supervised classification algorithm is the most common type employed in drug discovery research, as this methodology allows for a yes-or-no situation where an algorithm can make a "best guess" value judgment to the input features. Many types of regression and classification algorithms, including random forest, decision tree, artificial neural network, Bayesian methods, and support vector machine, are useful in biopharmaceutical research. For example, the random forest and XGBoost algorithms have been employed to classify active versus inactive anti-cancer cell lines. In the area of unsupervised learning, "bags of representations" for molecules, which are probabilistic models for the high-dimensional "molecular fingerprints" data obtained from assays in large-scale bioactivity databases, have been employed through high-speed screening of molecular structure–activity datasets. The unsupervised learning approach of Gaussian mixture modeling over latent space has been employed to find the "clustering moments" related to the properties of drugs with respect to their respective target classes. The unsupervised clustering moments help in exploring whether the interaction score has some correlation with the scoring system adopted within that target class. Some cases of reinforcement learning in healthcare are worth mentioning. The most well-known is the reinforcement-driven nonmyopic clinical trial design model to reduce uncertainty in assessing treatment effects. The reinforcement learning framework has been proposed through reinforcement-driven ablation to ablate a combination of pharmaceuticals to facilitate the complete ablation of large pancreatic tumor volume. Furthermore, deep reinforcement learning has been employed for personalized dosing in trials.

4. Applications of Machine Learning in Predicting Biologic Drug Performance

Advancing the field of biopharmaceutical research strongly depends on choosing the best and most promising candidate drugs for further clinical development. Learning from historical data helps to build models that forecast how new drugs are going to work in terms of efficacy and safety, and can also project the financial and clinical outcomes. These broad predictive applications usually use quantitative models, in the same way a conventional statistical model in some field can make inferences or

decisions. At the center of developing quantitative models for predicting biologic drug performance is the identification of so-called predictive biomarkers of response, markers mechanistically associated with drug performance. As such, predictive biomarkers of response could improve the selection of leads and candidates by providing decision support. In this section, we discuss the machine learning methodologies, as opposed to traditional analytically explainable methods previously discussed, to produce predictive biomarker scores.

One advantage of machine learning approaches is that they reveal any hidden causes and multidimensional interactions considered as contributors to the efficacy or safety outcomes. Diseases are generally driven by a complex set of multidimensional factors, beyond the genetic and physiological hyper/hypo-regulation, such as personal lifestyles, dietary habits, geographical locations, economic level, immunizations, and urban or rural locality, among others. These multitudinous factors detract to a large extent from the success of personalized medicine by using quantitative models for the prediction of drug efficacy and safety outcomes individually for every patient. Machine learning-based predictive biomarkers of response might also be able to use the feature slopes to suggest the presence of derisking factors operating at different signs of effect of the predictive biomarker on efficacy, or derisking factors that only appear in patients with a high predictive biomarker value. Historically, multivariate quantitative models defining predictive factors leading to efficacy and/or safety in patient populations are mainly defined through univariate and low-dimensional studies. More recently, the application of omics technologies has dramatically increased the dimensionality of predictive biomarkers of response used in predictive whole-omics studies; however, it is also possible to use high-dimensional predictive biomarkers of response in low-tens-of-conventional-variables studies. Accurate evaluations of the in silico prediction accuracy or robustness of quantitative model methods may be hampered by scarce independent unseen data to test such methods and also by extensive overfitting methodologies available to compare hypotheses sharpened to fit the training data.

4.1. Drug Target Identification and Validation

4.1. Drug Target Identification and Validation. Target identification is typically the first step in drug discovery systems. It is defined as a reasoning pipeline where new targets are based on collected information from various scientific sources. The limitation of

these approaches, such as molecular profiling assays and automated text content generators, is the requirement to validate them using experimental evidence, which is labor-intensive and high-cost. Additionally, the incomplete biological view is supported by computational techniques that demand a large number of biological targets for supporting evidence. In this context, machine learning approaches can be used to minimize these problems by studying large-scale data distributed across the various levels of cellular products, including genetics, genomics, proteomics, metabolomics, and others. With such levels of complexity, the discovery of new candidates or the already developed drugs that can be redirected to a new therapeutic target based on useful criteria remains a major step in drug discovery.

The application of any machine learning algorithm to further optimize the drug discovery process and the target validation phase is also useful, rather than just target identification. This additional phase can also be performed before the implementation of in vivo or in vitro screening tests. The availability of large datasets, efficient algorithms, multi-omic databases, and common chambers of omics data can easily overlap this scenario. According to this information, many studies have already tested and demonstrated the efficiency of various machine learning algorithms and have also suggested advances through the integration of omics data and machine learning approaches. In addition, since the drug discovery process involves a multi-faceted team mainly composed of technological computing scientists, bioinformaticians, specialists in the field of biochemistry, chemists, and doctors, it would be more effective to develop strong international collaboration in this field.

5. Optimization Techniques in Biologic Drug Development

The biopharmaceutical industry resorts to various techniques to increase the chances of success and to streamline time and costs in the optimization of therapeutic proteins during early discovery and late-stage preclinical development. Due to their unique properties and increasing prominence among therapies, biopharmaceuticals present their own optimization challenges. Manufacturing variability and molecular structural complexity range from primary to quaternary structures, disordered regions, additional and diverse post-translational modifications, and cross-species differences. By making use of the optimization techniques, it is possible to facilitate a systematic adjustment of previously unexplored or less well-described but often equally relevant processes and

parameters. Not seldom, the positive outcomes encompass an increased potency and stability of the protein at each step of the process, as well as its improved manufacturability.

There are several methods to achieve optimization, including design of experiments, response surface methodology, and various optimization algorithms. Design of experiments establishes a multivariable space to perform combinations of factor settings and their levels, thereby providing insight into what combinations of factors tend to generate the best results relevant to drug performance or manufacturability, thereby facilitating rational decision-making towards further preclinical pathway proceeding. Optimization algorithms can be used in several experimental designs and data sets resulting, for example, in the identification of more variables and the prediction of the best settings of parameters for execution. Machine learning tools, based on predictive analytics algorithms, are increasingly used to attract maximum information from the data. A machine learning-assisted design of experiments approach, integrating deep neural network or evolutionary algorithm-based predictions, can guide experimental designs to explore, confirm, or optimize the design space across a wide range of bioprocess goals. Furthermore, an ongoing, iterative, machine learning-assisted exploration of design space can offer the promise of real-time optimization towards clinical trials and scale-up through the creation and use of quality-by-design pathways. In addition, discussions can also be created by making proper use of machine learning focusing on adding known bioprocess designs or experimental historical data into the making available such scientific discussions. This brief will present the concept of using various optimization approaches as decision support for achieving more effective biological proteins with desired activity in manufacturing. Aggregation, activity, and half-life of high molecular weight therapeutic proteins do play a role in the concept of exploring and improving them based on known scientific information carried out by a few case studies with noticeable occurring immunogenic events.

5.1. Genetic Algorithms

5.1. Genetic Algorithm Genetic algorithms are the most prominent representatives of algorithmic optimization. They are used to find good or even the optimal solution for a given complex problem. The optimization of a population is conducted by performing a selection of the best individuals according to a specific cost function. The best

individuals in a population will survive and multiply. Other less performant individuals in a population will not have guaranteed survival. From time to time, an algorithm applies crossover, the unification of genes from two better solutions, on a selection of individuals. Crossover creates new individuals that have the possibility of being better than the old ones. Furthermore, mutation, which adds a small characteristic to an individual, can also be performed to prevent stagnation of the algorithm in a local minimum, which might not necessarily be the optimal point in the whole solution space. By applying these mechanisms, genetic algorithms are able to find a desirable solution from a wide solution space.

Therefore, this technique works more efficiently in numerical optimization or optimization problems involving biological systems or other complex systems. A few examples already exist where genetic algorithms assist with an optimization process in the pharmaceutical industry—both small molecules and biologics. The primary aim is to improve product quality while aiming to reduce operating costs. Additionally, it appears that there is limited use of genetic algorithms with a dataset that combines both drug product and process parameters for biological pharmaceutical drugs. There are numerous use cases for machine learning when evaluating the influence of components or input factors on a desired output, which is likely due to the human-designed procedures. As mentioned above, the biopharmaceutical industry is transitioning towards AI and machine learning for innovation. A genetic algorithm fits very well into this new trend as it effectively combines machine learning with natural selection. This technique is most beneficial in speeding up the formulation process, which in turn improves cost-effectiveness and efficacy of the biopharmaceutical product. Therefore, there is value in exploring other computational techniques when producing proteins.

6. Future Direction

6. Future Directions

Advances in computational power and data availability are on the horizon to offer much-needed impetus for innovative applications of machine learning in various fields of biopharmaceutical research, including drug discovery. In the present data-driven science, new research directions are on the way, such as the integration of artificial intelligence and machine learning to enable predictive capabilities, optimize workflows, and generate valuable insights into big data. Importantly, a wealth of data alongside

ethical considerations and attentive regulatory requirements will be on the agenda, accommodating more recent, disruptive approaches of AI and machine learning, as they stand ready to disrupt global economies, including healthcare and multimodal solutions to global challenges. These new frontiers are being innovated and set in motion in parallel for responsible, sound ways of using and benefiting from these technologies, leading to a departure from the current risk-averse research environment.

Therefore, it is a prime time to transform the conventional, uni-disciplinary research approach into an interdisciplinary framework, where biostatisticians and life/drug discovery researchers and practitioners work hand in hand for the prompt identification of intricate biological challenges present in big and complex data related to drug discovery and exploring new drug targets. Future machine learning research would keep pace not by huge reanalyses of enormous datasets, but by updating the algorithms using both additional data and new primary studies to incorporate into the models inevitably produced by better drugs and therapies that are subsequently developed. To take truly individualized or personalized medicine to the next level will require new generations of biomedical and clinical advances driven by machine learning methodologies. The fine-scale molecular, metabolic, and perturbation data patterns observed will enable new targets and interactions to be revealed, and predictive trajectories of future response and new therapy optimization paradigms to be proposed. Thus, as the algorithms learn, this potentially enables the next stage of therapy generations often called hot-spotting or new transactional medicine therapies throughout the disease and therapeutic pathways through carefully derived combination therapies.

7. Conclusion

This review demonstrated the large impact of machine learning on developing biologic drug performance prediction tools that can improve drug development efficiency. The data used in biopharmaceutical research is usually large and complex, which poses a high level of difficulty for traditional research methods. Thus, it is of great importance to integrate advanced data analytics tools and use them to help applications in diverse stages of drug discovery processes. Supervised and unsupervised learning techniques have been used to make breakthroughs for a wide range of research applications and reduce concern areas in some cases, which has improved the operation of the

biopharmaceutical industry. From a technology standpoint, machine learning algorithms are highly accessible and widely used in various fields, solving practical bottlenecks of the past. This review included recent successful case studies and demonstrated the feasibility to solve application problems in early and late stage drug development, and also illustrated the realistic benefits to biopharmaceutical companies. Although there are a large number of potential innovations associated with machine learning techniques that are likely to revolutionize biopharmaceutical development in the near future, many challenges and areas of research remain. Becoming aware of the limitations as well as the future potential of machine learning tools can energize and enable biopharmaceutical companies to bring new approaches and results to drug development. Given the complexity of biopharmaceutical research, the successful use of machine learning technology in drug development in the future will depend on how the industry collaborates with the relevant scientific community.