Accelerating Drug Discovery Throughput with Streaming Inference Pipelines: Real-Time Computational Analytics in Pharmaceutical R&D

Dr. Ekaterina Ovchinnikova, Associate Professor of Applied Mathematics and Computer Science, Saint Petersburg State University, Russia

1. Introduction

The generation of high volume and diverse data in pharmaceutical research has led to the application of sophisticated analytics in order to enhance the productivity of the pharmaceutical research organization. The major focus of these applications has been limited to the development of algorithms to solve and devise new alternative solutions to provide greater productivity. Additionally, the application of such analytics is generally not real-time, and this would limit their impact on individuals who participate and incorporate them into their work as per their convenience. There is a need to develop comprehensive real-time AI-powered analytics algorithms that can be seamlessly and quickly deployed in a pharmaceutical research environment to speed up the interpretation of results, pinpoint areas to focus on, identify reasons for variations, and predict clinical implications. In the changing environment of drug discovery and drug development, the driving forces are the need to innovate and bring new drugs faster and more economically. In such a scenario, it is important to have high-quality results and data interpretation at a very early stage in the discovery to development workflow.

Research professionals generate huge amounts of experimental, structure-based, and more recently patient-oriented data during their course of work, which, when handled, analyzed, and interpreted quickly, reliably, and usefully, reduces the time and cost of the drug discovery and development process. Pharmaceutical research today, more than ever, consumes massive amounts of resources and generates an inordinate volume of diverse data to feed the drug development pipelines. Time is the enemy of the pharmaceutical industry, and harnessing data in shorter times would provide greater

research advances. In this contemporary research environment, where many projects are conducted with fewer resources but high expectations, the 'early' element in the phrase 'early phase drug discovery' and 'early phase clinical research' is becoming even earlier. In addition, the application of recent methods of data acquisition tends to generate data in larger quantities and increased complexity. Only specialized statistical techniques and their robust computational ability can mine out the value of such data and identify trends and causation.

1.1. Background and Rationale

Pharmaceutical research, with associated drug discovery and development, is a challenging area to work in. Historically, development work has also been stifled at all levels of the history timeline by the cost and time taken to move a candidate through to a marketed approved drug, with time periods measured in double-digit years rather than months or single years. At the same time, the amount of data generated during research sets a rapidly expanding whirlwind of at least a multitude increase per decade. Concomitantly with the high throughput and all-omic generation techniques, more data than anticipated is increasingly becoming accessible at a vast rate, being frequently converted into a digital format. However, even from the outset, one of the well-recognized pushbacks has been the various and divergent technologies, together with widespread or isolative analytical methods, each with their deepening limitations as the data churn continues. Historically, just the filtering of the petabytes of available data has served to demonstrate how very limited the traditional data analytics systems have been in the drug discovery and development pipeline arena. Therefore, as a consequence of the mountain ranges of pharmacological, biological, chemical, or engineering datasets in place, informed holistic research to development decisions could become somewhat stultified or practically impossible within any one individual scientist or team. Partially as a result of the overarching reasons for considering possible generation iteration with an AI platform, it potentially meets all the strategic aims and elements of the research hypotheses that we present. From our utilitarian perspective, it especially covers the ability to allow automated analytical machine augmentation, possibly thereby providing resultant faster internal velocity of data range spectrum, in terms of potential acceleration in post-analysis inferences or effectively writing potential neutrally networked synthetic grey matter or alternative ways in knowledge capture, to improve, enrich, or suggest beneficial changes. More opportunistically, and perhaps strategically,

such fruits could, in turn, result in upgrading our real-time data spectrum time span, concurrency (depth), and an ultra-scale data-hosting capability concept of the new AI engine, thus achieving a two-way loop internal enhancement with future, particularly inextricable and congruent sentience, to present advantage. The elements discussed in the first capability proposes section header above regarding deep learning capacity in internal self-adaptation could be adding either individual or combined fidelity to the throughput data spectrum quantification enrichment advancements. In turn, ongoing further development on these interim modulations of such a future AI real-time box could enhance the bigger term possibility spectrum of involving the creation of new superior research AI agents. These transformative additive components could effectively outperform AI potential above or proposed here to elicit the three core enhancement hypotheses elements.

2. AI-Powered Analytics in Pharmaceutical Research

In the face of disruptions anticipated for the pharma industry, the adoption of data-driven decision-making processes will be imperative for research functions to boost productivity. As it stands, the magnitude of datasets is no longer keeping pace with the accelerating rate of knowledge accumulation. Workflow inefficiencies result in high failure rates and long product development timelines, leading to difficulties in decision-making, planning, and prioritization. Effective adoption of AI techniques such as deep learning, transfer learning, and adversarial learning, among others, could potentially overturn dominant pharmaceutical research methodologies for the better. For example, quantitation of actions at the macro scale, combined with mesoscopic views on cellular topology, can be used to dissect the contributions of gene expression regulation at different levels on the spatial profiles of the inflammation mediators in the context of penicillin allergy. This would enable the study of settings in which the complex multi-omics data 'cancel out' and indicate a contraindication.

Analytics at the early phases of drug discovery can help overcome hurdles such as lack of quality and quantity of data and uncertain clinical outcomes. By integrating into the current workflow, AI analytics can help in better interpretation of data at various decision nodes in the pipeline. Even though companies disagree on the specific dollar amount of mature drugs that they have identified through diagnostics, they nonetheless confirm that investments in precision medicine are particularly cost-effective. In clinical

trials, recruitment of research participants for precision medicine sub-studies is three times as expensive when conducted randomly as opposed to testing biomarkers prior to recruitment. In these cases, AI-based programs find participants suitable for therapy twice as effectively and at reduced cost, demonstrating the ROI of technologies that cater to diagnostics and therapeutics. In this way, the immuno-oncology therapeutic landscape is being shaped by AI, driving it closer to innovation trends. AI analytics will reduce the time spent collecting data and facilitate rapid hypothesis generation. Development and interpretation of AI study results will become more of a resource allocation issue. Whether at the academic, R&D, or patient levels, one will need to assess if the results of the AI study will create additional value.

2.1. Current Challenges in Pharmaceutical Research

Drug discovery and development are inherently risky and costly, with average out-of-pocket costs per successful drug estimated at over 1.5 billion USD. From stand-alone to mega pharma, the entire pharmaceutical industry is struggling with various research and development challenges. Today, many of these challenges are related to the complexities and unknowns of the underlying biological systems of many diseases, despite considerable research prospects promised by genomics and promising new medications identified by AI. More broadly, however, fundamental limitations are many. The challenges that most prompt and are felt by our collaborative partners in R&D are concerns over data fragmentation, regulatory compliance, and control over collaboration by seamless integration of components within and across the discovery to post-marketing R&D pipeline—with the intended outcome being improved research productivity.

Large volumes of high-dimensional data are commonly generated in discovery from a variety of technologies: genomics, proteomics, image analysis, clinical tests, patient histories, and high-throughput screens, to name but a few. Cutting-edge technologies such as advanced microscopy or mass spectrometry produce even larger amounts of data that require innovative mining methods. The data must be 'cleaned' and relationships identified. Unfortunately, the number of variables and the complexity of relationships and the right order of variables that are important in predicting the outcome of interest in a discovery data set is uncertain in any given experiment. So-called $p < 0.05$ methods, which have underpinned much of the most exciting biological

discovery of the past 20 years, creak under the weight of the complexities of modern biology. In addition, the very nature of discovery research is that researchers do not necessarily know what to measure nor indeed where to focus their collective efforts in a domain of interest. This can lead to wasted resources and effort; a specific concern given the challenges of escalating requirements for drug safety. But waiting for the results to come through from hypothesis-directed pathways is not fruitful in the faster soft-money making culture of today. Unified systematic collaborations are hard to manage against traditional team/function paradigms or a common one-size-fits-all solution due to lack of time from back-to-back meetings and heavy teaching loads. Taken together, many of these factors add to rising R&D and drug development costs.

2.2. Benefits of AI in Pharmaceutical Research

While the development of AI technologies in pharmaceutical research faces significant challenges, there are substantial potential advantages stemming from the deployment of successful solutions. Perhaps most importantly, machine learning models can process and analyze data far more quickly than human researchers, making it feasible to handle and interrogate large and valuable datasets in ways that previously would have been largely infeasible. Many modern methodologies are of this data-driven persuasion, with advantages manifesting in terms of increased efficiency, improved accuracy, and in some cases drawing novel insights that were previously entirely overlooked. In addition to making it possible to gain value from increasingly large and complex datasets, breaking down the silos between different teams and organizations led by people working on possibly complementary treatments, AI can help to drive collaboration among researchers working in these different areas by increasing the agility with which it is possible to access and interrogate large stores of historical data.

At a more granular level, AI technologies have the potential to enhance the accuracy of traditional predictive analytics. In the same way that machine learning models can identify so-called 'nonlinearities' – patterns and correlations that simple arithmetic measurements may miss – we propose that advanced AI-capable techniques can be used to quickly and more accurately interrogate clinical databases, allowing companies to resolve dose-related questions in early development, avoiding much of the wasted effort and time that characterizes some early phase R&D activities. Consolidating all these benefits, the ability to evaluate positive and negative signals in real time, combined with

a capacity for improved predictive methods and faster decision making, has the potential to enhance pharmaceutical research productivity for decades to come. Related to this, machine learning can identify deeply latent relationships that are much harder or even impossible to identify through orthogonal measurement techniques, delineating new mechanistic insights that were previously overlooked or misunderstood. In addition, AI can help on different levels to gain insights that are not easily interpretable via conventional measurements, establishing the applied relevance of the analysis.

Moreover, the companies that take the plunge to get into AI research discover that AI can potentially save costs in the long run. The potential benefits for R&D are substantial, with estimates of cost savings of 25–50% per trial. Competitive advantage in the pharmaceutical industry aside, the power of data capture, analytics, and use by manufacturers is driving the current wave of industry consolidation. Taken as a whole, the greater value, cost-effectiveness, and capacity for bulk data analysis of AI technologies hold a great deal of promise to transform the productivity of pharmaceutical research carried out in the AI analytics research ecosystem.

3. Machine Learning in Data Management

Machine learning (ML) is immensely powerful when it comes to deriving insights from data and is steadily being adopted throughout the research process within companies. Despite this, there is a famous saying within the data science community that goes, "garbage in, garbage out," which concisely highlights the necessity of ensuring that data management is robust and well-established prior to implementing AI techniques. This is because no matter how advanced the chosen methodology may be, the analysis will inherently be limited by the quality of the input data. Presently, high-quality data infrastructures are not developed in the pharmaceutical industry, meaning that the "frustration around ML today is really all a symptom of a lack of productive things to actually analyze."

Currently, widespread throughout the industry are manual data infrastructures, operating on a case-by-case basis, which slows data handling. Within our organization, data are frequently outsourced, and so the initial data handling that we perform is a system built for speed. Rather than hand-engineer multiple features for a specific task, we concentrate more on low-level issues—such as annotating missing values and handling structured vs. unstructured data—since dealing with those issues would

require additional time further downstream. Feature engineering consists of three widely grouped areas of data manipulation, the initial one being data collection. Various approaches can be adopted for this process, such as manual entry; electronic spreadsheets; or graphical user interfaces (GUIs). Nevertheless, the collection of high-quality data is essential, and using electronic means for both data entry and storage can help to mitigate collection errors. Post-collection comes data preprocessing and then feature selection, before actually employing any modern machine learning methods. Data preprocessing is focused on making the data appropriate for downstream analyses. This involves dealing with any missing or corrupted data, noisy data, and irrelevant data. Feature selection aims to reduce the number of predictor variables in data used for model training. Removing irrelevant or redundant predictors improves interpretability and yields more effective predictions, and thus taking a subset of the most relevant simple features and training the model with these results in an increase in the prediction accuracy compared to using all features. Several approaches can be used for feature selection, including multiple R-squared testing, LASSO, Ridge regression, and decision trees.

3.1. Data Collection and Preprocessing

Gathering and preprocessing suitable training and validation data is one of the most important aspects for an effective application of machine learning and artificial intelligence capability of transforming negatively biased data interpretation into evidence-based scientific knowledge. In the pharmaceutical industry, various data from in vitro assays, in vivo animal studies, and outcomes of clinical trials are possible sources for the application of such analytical and predictive models. Reflecting on cutting-edge genetic and parametric information as predictive tools in pharmaceutical research, the application of effective patient stratification for clinical trials or subgroup analyses becomes conceivable.

Several advanced machine learning algorithms have been evaluated and benchmarked, but they generally rely on multiple data from diverse origins and scales. Data inconsistency, missing data, and already observed clinical trial phase-driven typical data pattern changes need to be addressed in complex global research environments. It is also common to come upon unsuitable time zones for appropriate data exchange to obviate latencies, biases, and to satisfy strict data protection agreements. There is a considerable

number of different options one might consider in order to ensure a consistent and synchronized machine learning predictive analysis framework and setup. Necessary incoming data are, thus, transformed in such a way as to be used in robust model analyses. Data cleaning and data preprocessing are therefore essential preconditions for effective artificial intelligence and machine learning analyses. Medical data collection and preprocessing of such big data is a multi-step process that includes normalization and scaling, features and outcomes definitions, data imputation, and trend analyses.

3.2. Feature Engineering and Selection

Feature engineering and feature selection are crucial aspects of machine learning, especially in the context of pharmaceutical research. Feature engineering involves the creation of attributes, or features, that specifically define each instance, or input, into the predictive model. Features describe the data generated from the real world that are present in the application domain, which contains a set of important variables. The essence of feature engineering is to create domain-specific features that enable the predictive model to generalize better for future predictions, which can explain the model. In other words, the predictive models are based on real-world pharmaceutical research inputs to explain the outputs akin to a potential descriptive model. Feature selection refers to the process of selecting the relevant variables that significantly contribute to the performance of the prediction model.

In practical applications, different types of methods may be employed for both feature engineering and feature selection. For improved predictive model performance, domain knowledge can be utilized to combine and refine features, creating a stronger, more general relationship between the dataset and its potential domain outcomes. Overfitting can potentially occur when some of this domain knowledge is only marginally important to the model, in which case the feature engineering method must be used strategically to balance this risk. Selecting the top domain-baked features can potentially alleviate this problem. In addition to using domain-based feature engineering to improve model performance, other methods to incorporate domain knowledge in the modeling process have been issued concerning the feature selection approach. These methods employ smart ways of creating new features or removing existing attributes by adopting feature selection derived either manually or automatically. Automatic feature

selection allows for selecting those informative features from a broader field, which can efficiently streamline the model learning and prediction process.

4. Machine Learning in Experimentation

We now turn to the applications of machine learning in the experimental phases of pharmaceutical research. Within this space, machine learning has the potential to be transformative, enabling researchers to more systematically engage in the design of experiments to identify approaches with more significant impacts on outcomes or to operate experimental equipment more interactively.

Predictive modeling is an essential contribution of machine learning to experimental design. The ability to relate parameters to outcomes to create a model that predicts outcomes once the parameters of an experiment are specified is central to the approaches becoming useful in designing experiments. The predictive models are central to making rational decisions about the values of parameters chosen for experiments to be used in the search for better experiments. Optimization is another valuable area enriched through machine learning. The combination of customizable modeling and optimization techniques enables the production of powerful tools for the design of experiments in practice by predicting outcomes under candidate experiments.

At a more fundamental level, the tools for machine learning guide the operational decisions researchers make when choosing experiments to perform. When researchers have an opportunity to make choices regarding experiments, as they often do, the application of optimization methods can enhance decision-making. This is representative in cases where the decision-making involves the allocation of resources to different experimental options. Choosing the most promising experimental conditions to be tested with enthusiasm requires thoughtful consideration, posing an optimization problem to be handled with care. There are good reasons for refining experimentation that do not involve the realities of practice. Machine learning contributes theoretical considerations without explicitly adhering to decision-making and resource allocation as an operative guide. By concurrently satisfying the needs of theory and practice, the application of machine learning designs an impactful route to successful experimentation tailored to the resources available—two paradigms that can inform one another synergistically.

4.1. Predictive Modeling

Predictive modeling is a practice of deriving predictive models using machine learning and data mining techniques on historical data to uncover hidden patterns in predicting future outcomes. In drug development, predictive modeling plays a significant role in predicting compound properties, bioavailability, metabolism, pharmacokinetics, ADMET, and toxicology responses, which help in identifying potentially successful candidate drugs. Some common modeling techniques include linear and polynomial regression, decision trees, ensemble methods such as random forests and boosting, support vector machines, and modern deep learning techniques. In predictive modeling, the aim is to build a model that generalizes well to future predictions and responds well to cases beyond the training data. Results are often validated, but there are also some potential pitfalls with overfitting and not generalizing to new data. The data used to develop the model also needs to be validated, and model results replicated in additional independent datasets.

Predictive models are created using information from historical data, and while not perfect at predicting the future, they are extremely helpful in providing a guide on what could happen and what to try next. Other limitations of predictive modeling are that they are closed systems in which new knowledge and confirmation are difficult, and they are dependent on high-quality input data. In a research environment, predictive modeling can be used to identify what a potential output data profile might look like and then use the models in forward engineering to generate higher insight and novel objectives for experimentation. The stronger the predictive model in a research environment, the lower the amount of experimentation required to answer the hypothesis. Many datasets that pharmaceutical companies generate are suitable for predictive modeling, and their use is well-documented. A good example is their use to understand the interactions of compounds with nucleic acids. In predictive modeling, the aim is to build a model that generalizes well to future predictions and responds well to cases beyond the training data. Results are often validated, but there are also some potential pitfalls with overfitting and not generalizing to new data.

4.2. Optimization and Decision Making

As highlighted in Sub-Section 4.1, machine learning techniques can be applied on the back of high-quality data to predict outcomes and to identify cause and effect

associations. In pharmaceutical research, optimization is also important, which can be used alongside machine learning to identify the best options or combinations of components of an experiment that will yield the desired cut points. Genetic algorithms are perfect optimization choices, as being genetic algorithms, the solutions can be evolved and stored to help navigate around a complex decision matrix. Once a solution evolves that suggests the best combinations of decisions, data, or analysis in which to proceed, a more mathematical approach can be taken—for example, by using gradient methods that search around this point to evaluate approaches that have been previously identified as a good bet. In this way, we can begin to positively integrate prediction and optimization in an automated and real-time manner with increased confidence.

There is a very close relationship between prediction and optimization. The difference between the two is that in prediction, it is desired to identify the characteristics of a future set of outcomes whereas, in optimization, it is desired to determine more effective outcomes. Our goal with the combination of machine learning alongside optimization should be to approach the combined problem in a manner that can solve the problem with only a minimal number of experiments being incurred. Attempting to integrate simultaneously predictive models that can identify characteristics or investigations and optimization techniques in which the characteristics that are being sought are unknown can lead to future strategies that may lead quickly to better decisions and therefore be utilized. The interplay between using prediction, optimization, and experimentation to refine aims seems therefore to be the most effective and fastest way to reach a decision.

5. Case Studies and Applications

Selected Case Studies and Their Methodology In this topic, we presented case studies to illustrate AI/machine learning use in pharmaceutical research. This should demonstrate in a concrete way how to leverage AI-driven analytics to improve efficiency and productivity. We expect that new users who will set up work will benefit from several sessions on best practices in neuroscience research and related topics concerning AI/ML in precompetitive pharmaceutical research.

Analysis of Research Methodologies Each case study describes methodologies that include the application of machine learning and provides an indication of which part of the research process benefited from machine learning and also which part led to a research breakthrough. As the conference showed, using machine learning tools in

pharmaceutical research can lead to many benefits in terms of efficiency, effectiveness, and productivity. The results of this workshop will be used to inform best practices in pharma and AI in the future.

A detailed analysis of a five-year project on the use of big data and artificial intelligence in support of pharmaceutical research was conducted. The evidence indicates that the entry cost for such research is lower than the first envisaged study. Throughout the study, difficulties, but also successes, evolved around the key features of successful transformations, workflows, implementation, or positive objections. The study found that regardless of the objectives and techniques used, the eagerness to measure changes that had a direct or indirect impact on the results was essential.

5.1. Real-Life Examples in Pharmaceutical Research

Real-Life Examples in Pharmaceutical Research

Since the development of AI technologies, real-life examples of their successful application in different industries have only become more common. In this section, we will focus on the case studies where AI-powered analytics have already been applied successfully in pharmaceutical research.

Drug discovery was one of the first areas where AI technologies found their application. An example is the DeepStatus approach, which uses AI to explore treatment strategies for specific diseases and streamline the identification of potential anti-inflammatory drugs. The Quest to Learn project presented machine learning models aimed at predicting novel targets for treating colorectal cancer and Alzheimer's disease. The Quiet data-driven discovery methodology introduced novel drugs for clearer and safer MRI scanning using a combination of AI technologies. CAMP is an example of systems biology modeling that successfully found and supported the efficacy of a new target in acute decompensated heart failure. Another successful AI-powered application was in the area of optimizing a subpopulation for a clinical study. Finally, another example of AI-wide application in the pharmaceutical sector is SPORE2, which is responsible for system-based support and navigation of personalized cancer treatments for oncologists.

Retrospective research showed that some areas require sophisticated AI approaches; therefore, changing the type of research is hard but, at the same time, opens new opportunities in the pharmaceutical industry. This includes drug target discovery, drug

discovery beyond ADME, patient stratification, and drug repositioning. Some AI technologies could be utilized by the pharmaceutical industry. Most AI approaches can be used for known target discovery and non-targeted drug repositioning. These strategies, however, are still considered novel in pharmaceutical research, and not many companies are looking for a treatment of a disease outside of their core activities. It will be very exciting to see such AI approaches widely accepted in the pharmaceutical sector.

6. Future Direction

We believe that AI has the potential to be tightly integrated into all steps of pharmaceutical R&D, including the design and execution of experiments and the interpretation of the resulting data, holding great promise for future advancements. It is predicted that the capability of AI will grow significantly in terms of speed, accuracy, and broad diversity across data types, and will continue to drive innovations in efficiencies through the ongoing convergence of technologies. This may impact the methodologies used in pharmacological research, which in turn could potentially be prevented or become less frequently initiated in humans where success is less likely. Access for narrower interest groups, such as precision medicines, may also improve. Finally, it is likely that more opportunities for new combinations of drug products for repurposing to new uses will be explored. In terms of the constitution of R&D departments, there are likely to be more informaticians and digital scientists to curate and integrate data and extract knowledge.

Finally, a responsible approach to AI is of paramount importance; care should be taken in the use of all new technologies. The vast expansion of AI-generated models to self-parameterize and self-interpret opens new ethical areas with, as yet, unexplored risks and regulatory implications. The false development of a tool that harms human health could undermine decades of evidence-based medicine. In conclusion, the prognosis is optimistic; AI has great potential to improve productivity in a challenging environment. There are scientific, ethical, and collaborative opportunities that we encourage and are worth exploring, understanding good AI technologies in pharmacological research at the same time as stimulating innovation and new tools. Despite this, there are course requirement changes needed, particularly by focusing on conscientious adoption, to expedite and expand our current improvements. Potential barriers to this adoption and

how we could channel efforts to further progress in handling experiment outputs from AI-driven work pathways have yet to be fully explored.

7. Conclusion

In this essay, we considered the potential of the edge in accelerating pharmaceutical research processes to enhance industry productivity. During discussions, an underlying argument developed, many of the challenges facing the pharmaceutical industry were driven by the need for speed, as the ability to generate data outstripped the capabilities of the workforce. Primarily, the division highlighted two principal areas in the design–make–test–learn research cycle in which AI could exert a transformative effect: data optimization and experimental design. Employing AI tools, it is possible to reduce researcher workload and increase the range of hypotheses tested. Clearly, the use of machine learning and mathematical optimization techniques could prove to be invaluable for both managing incredibly large data sets and rapidly suggesting experimental parameters that are most likely to lead to the desired outcome. Throughout these discussions, however, it has been strategy that could facilitate the necessary change. The use of case studies should therefore provide actionable insights to guide the implementation of AI tools into broader research processes. In conjunction with the shift from the vertical to horizontal scaling, transforming decision making from the traditional top-down method to a more integrated approach could yield further benefits. With these considerations in mind, the scientific community can enhance the alignment of their own research agendas with the needs of the industry as we work together to establish AI as a transformative tool in pharmaceutical research. Evidence of this shift could encourage investment from the pharmaceutical sector, leading to a self-perpetuating cycle of further evolution within this space. Indeed, the integration of human factors with the AI-tool development strategy will be key moving forward. In conclusion, the impacts of the AI-revolution are incredibly diverse – as they span the intersection of technology and healthcare – and require real evidence to garner stakeholder buy in at all stages of AI development and use. Given that change is occurring, quite possibly at breakneck speed within research, the conversation around this subject matter is only just beginning.