Federated Learning and Knowledge Graph Integration: AI-Driven Platforms for Cross-Institutional Pharmaceutical Research Collaboration

Dr. Javad Salehi, Professor of Electrical Engineering, University of Tehran, Iran

1. Introduction

Transformative changes in the ways by which scientists can collate data and formulate hypotheses for the R&D pipeline are timely and relevant to pharmaceutical research collaboration. The challenges and complexities of enhancing global health through sporadic research conducted by disparate entities without sharing data, and the difficulty in conducting intra-consortium studies unacceptable to multiple consortium members have been discussed. The idea of a just-in-time computing model has led to increased application of informatics infrastructures into a range of research arenas. Here, we seek to describe how established Artificial Intelligence-driven platforms can be integrated to enhance collaboration processes within pharmaceutical research.

The increasing complexity of phenome-genome data, for example, among certain pathogens, echoes the growing complexities in early phase biological systems-based drug discovery. Furthermore, pharmaceutical companies contain valuable patient data. A good collaboration process can be important since it may be hampered by the governance of the specific datasets involved. In light of these challenges, it is essential for sufficiently advanced analytical tools to integrate both sequence and clinical data to facilitate pharmaceutical research collaboration. In this capacity, driven by data analytics and machine learning, open access models can forge the new model of not only efficient well-characterized research but also access to alternative potential biomarkers and drug targets. Therefore, in research collaborative scenarios described here, it is necessary to consider how to achieve sophisticated AI-facilitated data analytics.

1.1. Background and Significance

The landscape of pharmaceutical and medical research collaboration has rarely been different for a long time. The increasingly complex questions faced by researchers in understanding molecular and disease processes require a variety of interdisciplinary contributions from diverse expertise – epidemiologists, pathologists, computational biologists, clinicians, pharmaceutical experts, and many others. Yet, while the grand challenges facing humankind become increasingly co-dependent on each other's insights, how we deliver on this shared scientific ambition remains rooted in the past, seeking value in a convoluted system based on independent, modular advancements that are increasingly frustrating to invest in due to sustainability limitations.

Two decades of work and investment in digital initiatives and, in turn, bringing computational research methods into the heart of research thinking – touched by numerous and painful failures – have borne little fruit when it comes to exploiting the potential benefits of a heavily networked global research community. Questions that continue to tax the industry now are how to practically enable research communities to collaborate closely and, in an open spirit, to share data fully. Digital technologies were supposed to create a new operating environment for every pharmaceutical R&D, making hypothesis-led research into a collaborative enterprise that could be supported by data.

Pharmaceutical companies are also increasingly keen to boost research in order to face growing competition in every pharma space. As more companies have their research programs farmed out to external academic teams and/or maintain would-be rival companies, it is increasingly important for the company to invest in digital technology that boosts currently slow, cumbersome drug discovery processes. The traditional field of cheminformatics and its legacy portfolios – essentially, drug-like molecules with optimized properties – will drive the pharmaceutical companies' success, and collaboration also underscores a deepening precompetitive spectrum, as the industry gets to grips with just how far it has gone in the wrong direction in the post-rational world. Moreover, the value of an appropriate computational, rapid-response element of interdisciplinary research – particularly in the boundary between clinical and systems research – is now firmly established as a means to accelerate time to activation with maximum impact. Across these driving principles, there is an increasing body of

research that shares one common component – the ability to exploit high-quality data on massive scales as a means of developing patient treatment plans. There is a complex route from the evidence-supporting concepts of limited clinical trials and individual case studies to large structured datasets characterizing larger-scale condition processes. A properly designed framework for interdisciplinary collaboration involves sound ethical practice and respect for privacy, seamless integration of clinical and experimental data, and obtaining actionable information. It also involves computational challenges that high-quality, timely cross-industry computation can help solve. At once, it is also ripe for challenges that would provoke an algorithmic rethink in drug discovery.

1.2. Research Objectives

This research aims to evaluate how AI-driven platforms can help enhance collaboration in pharmaceutical research. The following is a list of research objectives that were established to guide the evaluation:

Research Objective 1: Assess the impact of using AI-driven platforms on 'time-to-collaborate' from pre-clinical to phase 2 research phase collaborations. As an initial phase, this research first involves defining a cohort for research, specifically those involved in the pre-clinical to phase 2 research phase collaborations. This will be used to perform a quantitative analysis of change before implementation and post-implementation of the AI-driven knowledge platform. Based on data collected, research objective 1 will be answered. The collected data will be used to evaluate whether the time taken to search for and identify experts and one-sided contract completions are shortened by implementing the AI-driven knowledge platform.

Research Objective 2: Examine challenges and barriers encountered when implementing the AI-driven platform based on the least successful cases. From the pool of participants that have been identified as least successful, in terms of reducing 'time-to-collaborate' using AI-driven technology, this research will be able to identify what barriers are faced. By asking these less successful participants questions surrounding the barriers encountered, research objective 2 will be addressed.

The final research aims to address will require a more qualitative and creative approach to executing this project and involve the use of recommendations from lessons learned and success factors literature. By identifying the most successful participants in this

study, this research will be able to identify, through explicit discussion and analysis, the capability and design features that can: (a) carry co-creative value and (b) create better environments that support collaboration in knowledge sharing for research.

2. Machine Learning in Pharmaceutical Research

Machine learning refers to a type of artificial intelligence that enables computer systems to learn from observed data sets. This branch of AI mainly operates through algorithms and mathematical models to facilitate task optimization and decision-making processes based on the computational analysis of data collections. Machine learning provides systems the ability to independently recognize patterns in observed data and to subsequently make informed decisions about new data according to the identified patterns. As a result, machine learning has a variety of applications involving advanced data analysis and pattern recognition, with notable techniques including linear regression, decision trees, and k-means clustering.

Until recently, drug discovery and pharmaceutical research collaboration have frequently been hampered by long periods needed for complex data analysis. Nonetheless, the applications of machine learning techniques may help streamline the research processes involved. Each application brings its own novelty and benefit toward advancing fast-paced and complex data analysis. Machine learning has been frequently used in the process of drug repurposing to identify new use cases and applications for existing drugs. Usually, algorithms help to predict new biologically applicable uses by inferring biological target functions related to their drug response. The integration of biological network and gene expression data with machine learning may help to advance computational drug repurposing. Drug repurposing accurately predicted the known therapeutic properties of a variety of pharmacological agents based on their local genomic effects, which were largely unknown when designing and executing analyses and experiments. With the help of autoencoder machine learning to computationally drug target, a significant scientific team of researchers was able to identify a series of unsafe FDA-approved drugs. Integration of transcriptomic and metabolomic profiles with machine learning may help to predict whether anti-diabetic agents have multi-organ impacts.

2.1. Overview of Machine Learning

Machine learning (ML) is a field of artificial intelligence (AI) that uses statistical techniques to give computer systems the ability to learn from data without being explicitly programmed. A primary task of ML is to model patterns and recognize insightful structures in a given dataset so that these structures can be leveraged for insight, prediction, and decision-making. Various algorithms have been developed to model data for a given task in these paradigms. For example, the k-nearest neighbors algorithm models data based on its similarity to other cases, while a decision tree finds successively linear decision boundaries on the input data. Recent advances in computer hardware have enabled more complex algorithms—such as those modeled as neural networks—that amalgamate simpler units in a hierarchically structured manner to learn complex relationships. Neural networks have shown exceptional promise for modeling large volumes of data. These advances reflect the complexity of information currently being produced in pharmaceutical research, where researchers can be faced with making decisions on thousands to millions of complex molecular descriptors that pertain to biochemical assays, physical properties, and pharmacokinetics. Recent theoretical improvements and innovations in algorithms can process vast unstructured datasets and have achieved state-of-the-art success in many applications. Thus, ML models have the potential to support drug discovery through mining and analyzing this data. However, current methods are predicated on clean, well-curated, small, and mechanistically appropriate datasets several orders of magnitude smaller than larger 'real-world' datasets. Challenging aspects of 'real-world' application include the need for complex pipelines of feature selection and model validation.

2.2. Applications in Pharmaceutical Research

These advantages of machine learning can now be seen in several applications in the pharmaceutical research domain. One of the most clearly noticeable changes machine learning has brought about is in drug discovery, where the use of such algorithms has helped in transforming the traditional approach to finding effective therapeutics. Machine learning has been widely applied in pharmaceutical research. Key areas where machine learning enhances existing practices include aiding in the discovery of new drugs, regulatory approval processes—particularly during preclinical and clinical trials —and marketing new drugs. Moreover, machine learning can streamline patient condition prediction and has applications in epidemiology research. Machine learning

algorithms improve the reliability of analytic conclusions. They generate highly accurate predictions for future events. The accuracy of these predictions is such that scholars are now confidently using data mining directly to draw final inferences and perform further analysis instead of deriving conclusions from data points through basic statistical analysis. Making additional assumptions, a model might fail to make adequate adjustments even for informative covariates, such as comorbid conditions or socioeconomic factors. Also, machine learning can create accurate predictive models from these sources, often with far less complex assumptions than traditional statistical methods. The increased point prediction precision is of greater importance for research applications.

Such progress in our predictive precision can significantly improve the use of the predictive model in practice, make their extensions significantly less complex, speed up drug development, and bring about lower testing costs. The subsequent use of machine learning platforms for more predictive modeling in big data research is the key benefit of advanced methodologies. At the raw status, machine learning improves the accuracy of our predictions. Furthermore, machine learning allows for faster design of optimal predictive models. Several case examples of successfully applying machine learning models and combining them with other widely used techniques have surfaced in the literature. These studies have generally fallen under the research subfields of data mining and pharmacogenomics. All in all, several studies implicitly or explicitly mention the importance of using machine learning tools in combination with other methods to improve analytic precision, predictor design, and improve prediction model calibration. In recent years, such technique applications have drastically reduced the preparatory data analyses' required study time and have, in most cases, improved prediction accuracy and reduced prediction error. We hint that vast opportunities lie in pharmaceutical research for this research community to advance beyond internal validation capable content.

3. Collaborative Data Sharing Platforms

Advancements in science and technology have amplified the volume of research data produced. The efficiency of transferring data and translating it into valuable information is key to fostering innovative collaboration, creating new products, and improving the drug development process. The value of integrating diverse preclinical data to augment

understanding of disease mechanisms and therapeutic responses is paramount for unlocking scientific discoveries. Collaborative data-sharing platforms are one way of achieving a common sharing space among multiple researchers from different areas. They are commonly platforms that enable a secure, efficient, and transparent method of data sharing, with tools for querying capabilities also available. The function of each of these offerings can vary and include data deposition from a data provider, access to data from the landlord, tools to allow simple query options of the data, and more complex data interrogation tools such as statistical packages or data exploration capabilities.

The use of these collaborative platforms to either access or store data has shown multiple benefits for both the provider and consumer. In general, it can lead to increased transparency, improved project and study tracking, standardization of data collection, collaboration initiation and advancement, lowering costs, and speeding up R&D through data exchange. Despite the benefits, the successful implementation of such a platform comes with certain risks and challenges that include the balance between security and usability, particularly the challenge with sensitive patient data requiring proper permissions and data access governance. Legal issues relating to public and sensitive patient data must be considered, and ethical issues regarding data sharing in sensitive health areas are to be addressed as well. Additionally, all platforms must ensure they provide data-sharing protocols that are consistent with international standards and local regulations. The technical challenge of data platforms is to ensure they have common harmonized ways to link to each other. Ensuring data can be integrated from multiple sources can be most effectively achieved through adopting global standard data formats.

3.1. Benefits and Challenges

A structured collaborative data-sharing platform, especially when driven by AI technologies, offers several advantages. One of the main benefits of a system like this is addressing the main challenge of such collaborations, which is communication. These platforms can allow timely and appropriate communication among all interested parties; setting up audits and notification systems ensures a seamless experience when sharing and utilizing the data and associated results. This, in turn, can reduce the number of failed projects due to either redundant efforts, insufficient access to certain types of data, lack of expertise, or availability of the scientific community. Indeed, better access to data

can also allow faster progress on projects, accelerating project timelines and reducing the time to discover and develop new drugs. By tipping the balance towards collaboration from the currently siloed and highly competitive landscape, such collaborative platforms could lessen such constraints.

Another advantage of data-driven collaborations is the access they lend to smaller institutions and companies, effectively democratizing their participation in larger projects. Given that biotech and pharmaceutical outputs are highly uneven, speculation is that as success rates for drug discovery increase, there will be a growing number of disease areas and drug targets researched, keeping these organizations busy with target development opportunities. Such technologies offer an easier route to collaboration with larger organizations for the development of drugs, giving smaller companies access to technologies, therapeutic areas, and markets that might not be possible due to the relatively small size of the internal team. On the flip side, several challenges exist around such a platform in terms of data-sharing. Channels of data provision need to be laid out, and these systems must include clear demarcations of ownership and possible commercial linkage. Although data gathered or generated as part of larger consortia may be the driving force behind a collaborative platform, there is a pertinent danger of platforms being a means for data generosity for larger companies as a tool for outsourcing with little equity to be shared. In research consortia, this can pose a problem for the level of scientific relationships, which could also have been inherited into such platforms. The benefit of such a platform for a data-generating company is that their data could go no further than just all supporting molecules where there is no further interest in the target. Alternatively, there is limited publication value and priority application readout. Such concerns must be dealt with at a legal level, which requires the setting up of early project governance. Indeed, the spotty, incomplete, and mistake-ridden nature of external data can also be a worry. A future user community must maintain the quality of their submissions to uphold the integrity of the community's work. Therefore, given these concerns about data-sharing, cautionary statements arise when imagining the broad application of these platforms to consortia. Parameters governing data-sharing should be made explicit, given that consortia involve competing actors. Because a level of competition would make any one user more likely to withhold the best data, research groups are generally afraid of sharing. Structures already exist that allow donors to direct regions or control what kind of experiments/distribution

occurs with their samples on a platform. All this said, any such platform should ensure that data provision is enforced either by sharing data requirements from consortia agreements or negotiation of such agreements. Experiencing minimal use of statins caused people to experience muscle damage; these are some examples of non-public, industry-provided information that would need to find another mechanism for access. From a user perspective, a clear plan that balances management direction and membership demands about data contributions must be developed.

3.2. Key Features

Collaborative data-sharing platforms must be user-friendly because their success is built on the active engagement of diverse stakeholder communities. The importance of ease of use is seen in the increasing number of platforms that are available for collaborative research. Basic to any modern data platform is the inclusion of advanced analytics tools that are easily used through a user-friendly interface. Also, collaboration platforms must support many of the features described in the platforms for data analysis. Aspects such as secure data access, version control, and audit trails are particularly important as they serve the functions of validating data sharing, ensuring the integrity of the data, and securing a level of compliance with regulatory and ethical directives. An essential technological feature for any modern data analysis solution is to base itself on the use of the latest cloud computing technologies.

The normal operation of a data sharing and analysis platform will generate substantial data. Research and development in the pharmaceutical sector are pushing the limits of the infrastructure required for data storage and analysis. A scalable solution to data storage and analysis is therefore essential. Not all potential users of a collaborative platform will want or be able to engage with it in the same way. Users are likely to have very different requirements for their engagement with a collaborative data analysis platform. The time spent optimizing the method will give a much better idea of the platform's performance and value to a user. Ultra-secure access to data and analysis solutions makes it possible to engage legal users with proprietary and confidential computational tools. The security of the operations is ensured through a connection to every cloud solution. The ideal data analysis platform should interface with existing, commonly used data analysis tools. Such tools are somewhat dependent on the datasets that are being explored. A publicly available database is compatible with multiple

platforms and analysis tools that are commonly found in use by biologists, medicinal chemists, and computational scientists.

Another key feature is the ease of collaboration. To allow collaboration between a variety of users, the platform needs to support multiple user accounts with differing privileges. Finally, legal and ethical obligations concerning the protection of potentially sensitive research project data need to be accommodated. Web security protocols, restricted access, and a need for all public data to be free of personally identifiable information are all factors that the platform developers need to consider. Regular software updates are essential to allow ongoing correction of vulnerabilities. In a platform setting where data sharing and analytical pipelines are linked, version control is essential both for the toolkit and the uploaded data. This is particularly essential as data analysis advances at a fast pace, and as a result, the data formats typically change very quickly.

4. Case Studies in AI-Driven Pharmaceutical Research Collaboration

Case Study 1: Collaborating to Improve the Discovery Process. This case study details a collaboration that uses a treatment-naive approach to understand pandemic-driven, life-threatening infections. Challenges faced: proprietary data sources and multicentric recruitment; quantification of a complex multisystem disorder with spillovers to environmental analysis. AI solution: unsupervised clustering to reveal biologically relevant subtypes prior to biomarker testing, sparse canonical correlation to reproduce randomized trial results, and causal regression models to bring biological significance to the data interpretation. Case Study 2: Collaborating to Improve Preclinical Translation. This case study exemplifies a collaboration between universities seeking to push the boundary of LR through strategic partnerships. RM will graduate from an integrated MSc/PhD. In 2022, this sophisticated drug discovery programme plans to partner with scientific consultancies as potential clients. Collaborating to tackle unmet needs in immune-mediated rare diseases, orphan drug discovery, and development. Challenges faced: identifying biomarkers of efficacy and safety when traditional endpoints often do not apply, small patient numbers, and a novel regulatory pathway with evolving acceptability criteria. AI solution: unsupervised learning to bring meaning to transcriptomic analysis in an otherwise small data set and Bayesian networks to produce expert knowledge-sourced, safety-reduced signatures. Case Study 3: Collaborating to

Aid the Regulatory Approval of a New Drug. This case study discusses a new drug for treatment indications in neurology and immunology; we have limited enrollment in clinical trials where individuals with Parkinson's disease will receive the drug over 52 weeks. Parkinson's disease biomarkers. AI solution: data quality analysis to maintain the fidelity of data generated through high-field neuroimaging analysis. Lessons learned: team collaboration across a medtech company, a sponsor company, and academics to benchmark and identify robust candidates for modeling before regulatory approval. Case Study 4: Collaborating to Predict Genome Editing Outcomes. This case study demonstrates a unique form of collaboration, where researchers from three different academic departments are studying a rare genetic disease.

5. Future Directions and Challenges

AI is a constantly evolving area of research, with advances already being seen in the field and further breakthroughs expected in the future. Among these developments, a greater understanding of the algorithms themselves, such as counterfactual methods in preparation for evaluating clinical data, is highly anticipated. Additionally, emerging visualization techniques are expected to make AI and predictions more accessible to researchers. Advanced networks should include inputs from more complex data such as linked datasets and social media for an even better understanding of the research environment. Research that demonstrates impact on outcomes is expected to be facilitated with the development of techniques to examine the correlation between research knowledge and decisions, quantitative research methods, etc. The use of AI for making all phases of the research process more effective raises various questions relating to scientific methodologies, the need for high-quality input data, ethics and transparency of AI models, and the appropriate regulatory environment for this. It needs careful consideration of how to ensure that improvements are made to the system while also guaranteeing fairness and equity. Additionally, predictions can be improved by revisiting old data and models as part of the retraining process. However, biases may be inadvertently introduced through revisiting models and expunging bad data, as well as attempting to make the AI system adapted to a continually evolving technological environment. As early career researchers are generally the best equipped to keep abreast of new technologies, the constant refresh, retrain, and maintenance leaves open the question of how to best maximize the potential of a research area that is constantly adapting. We see the development of AI-driven platforms for pharmaceutical research

collaboration as manageable; however, the translation of these outputs into practice will be complex to implement across all stakeholders of the research process—from pharmaceutical companies to academia, to regulators and charities. It will require skilled staff and continuing development and improvement and be subject to rigid data protection. The merging of different AI tools and techniques is also likely to require further research, development, and training from within and outside of the pharmaceutical industry. Ensuring there remains a level of trust and stakeholder involvement in the design of such platforms will also be critical to successful implementation. Research will need to address the culture within the pharmaceutical industry and beyond to encourage cross-company data collaboration, in a similar remit to work performed in commercial Digital Innovation Partnerships. In addition, there will be a requirement for technology appraisal of AI systems to demonstrate the clinical and cost-effectiveness of new disease pathways and treatments. Technologies may need to adapt across multiple healthcare systems, countries, and technological backgrounds, ensuring that all political perspectives are considered and upheld. Future AI technologies will also need to ensure the upliftment of less well-known and technologically assisted perspectives to the same level for full coverage. This reflects the view that AI sciences are inherently multidisciplinary, and therefore future planning of AI shifts to incorporate and enhance some of the oldest perspectives in new innovation while also ensuring that those perspectives can be accessed by newer disciplines. This will require significant cross-training and education and breaking down of traditional subject study barriers.

6. Conclusion

In this essay, we have provided an interdisciplinary review of several AI-driven platforms for improving the collaboration and efficiency of research into new pharmaceuticals. The four case studies present applications of AI to improve (1) our understanding of the fundamental mechanisms of disease biology and therapeutics, (2) the prediction of which of the many drug targets are most relevant to pursue, (3) the development of new low-toxicity drugs starting from single targets, and (4) the aggregation of old and emerging data around a network of drug-target interactions. We discuss how these AI methods may revolutionize our possibilities for conducting science by mirroring the acceleration of information gathering and knowledge representation of the last 50 years, even though developing such tools also faces a myriad of challenges.

The experiments are therefore interesting but only present initial steps into changing the current model of the limited possibilities of drug development. They are the first in a hopefully expanding collection of papers looking into improving AI platforms for drug discovery and personal healthcare. A major challenge to be overcome will be bringing together different interdisciplinary paradigms involved in this project. We suggest that the best approach is to establish special, more informally organized research projects rather than attempting to bring together individuals working from different philosophical, methodological, and scientific agendas. In this regard, the present essay is an example of how to bring together different lines of inquiry in achieving industry-relevant research, and we hope that we shall see more such attempts reported in the literature.