Artificial Intelligence for Public Good Drug Discovery

Recommendations for Policy Development

November 2021

GPAI / THE GLOBAL PARTNERSHIP ON ARTIFICIAL INTELLIGENCE

Contributors

Co-Chairs - GPAI Drug Discovery Committee

Yoshua Bengio, Founder and Scientific Director, Mila – Quebec Artificial Intelligence Institute & Professor of Computer Science, University of Montreal
Alice Oh, Associate Professor of Computer Science, Korea Advanced Institute of Science and Technology (KAIST)

Contributing Authors

Allison Cohen, Applied AI Projects Lead, AI for Humanity – Mila
Elliot Layne, PhD Candidate - McGill University, School of Computer Science

Invited Specialists

Anurag Agrawal, Director, Council of Scientific and Industrial Research
Alan Aspuru-Guzik, Professor of Chemistry and Computer Science, University of Toronto
Regina Barzilay, Distinguished Professor for AI and Health, MIT
Joanna Bryson, Professor of Ethics and Technology, Hertie School
Carlo Casonato, Professor of Comparative Constitutional Law, University of Trento
Raja Chatila, Professor of Robotics and Ethics, Pierre and Marie Curie University
Enrico Coiera, Director of the Centre for Health Informatics, Australian Institute of Health Innovation
Payel Das, Research Staff Member and Manager, AI Science, IBM
Marc-Antoine de La Vega, Postdoctoral Fellow, Microbiology and Immunology, Galveston National Laboratory
Aled Edwards, Founder and Chief Executive of the Structural Genomics Consortium
Marc-André Gagnon, Professor of Political Economy, Carleton University
Yeong Zee Kin, Deputy Commissioner, Personal Data Protection Commission
Hiroaki Kitano, Director, The Systems Biology Institute. President & CEO, Sony Computer Science Laboratories
Gary Kobinger, Director, Galveston National Laboratory
Pierre Larouche, Professor of Innovation and Law, University of Montreal
Tze Yun Leong, Professor of Computer Science, National University of Singapore
Kim McGrail, Professor, School of Population and Public Health, UBC. Director of Research, UBC Health
Dewey Murdick, Director, Center for Security and Emerging Technology, Georgetown
Richard Naiberg, Litigator, Intellectual Property Law, Goodmans
Ziad Obermeyer, Distinguished Associate Professor of Health Policy and Management, Berkeley
Alan Paic, Senior Policy Analyst, OECD
Alejandro Pisanty, Director General for Academic Computing Services, National University of Mexico
Daniele Pucci, Head, DIC-lab, IIT. Visiting Lecturer, University of Manchester
Margarita Sordo, Instructor of General Internal Medicine and Primary Care, Department of medicine, Bringham and Women's Hospital, Harvard Medical School
Mirjana Stankovich, Senior Digital Policy Specialist, Center for Digital Acceleration, DAI
Gaël Varoquaux, Research Director, INRIA

About the Document

This recommendations document was created in consultation with over 25 global experts spanning industry, public sector and academia. Due to the multidisciplinary nature of this field, this document results from the contributions of individuals with many areas of expertise, which have been grouped below for the sake of simplicity. These areas include:

- Machine Learning
- AI Ethics
- AI for Medicine and Antibiotics
- Computational Systems in Biology
- Data Innovation and Protection
- Intellectual Property Law
- Medical Biophysics
- Medical Informatics in AI
- Pharmaceutical, Social and Health Policy
- Robotics Engineering
- Internet Governance
- Science and Technology Policy
- Cybersecurity
- High-Performance Computing

Furthermore, these committee members were geographically diverse, representing countries including Australia, Canada, France, Switzerland, India, Italy, Japan, Mexico, Singapore, South Africa, South Korea, the United Kingdom, and the United States.

From May to September 2021, this expert group met approximately once a month to discuss high priority questions associated with the development of this document. Between committee meetings, experts were also asked to provide asynchronous feedback on specific sections of the document itself.

In addition, this document has benefited from the input of relevant stakeholders and institutions working to promote research and innovation in drug development, largely in the context of neglected diseases and antimicrobial resistance. These individuals were consulted throughout, providing input on how to create, promote and incentivize an effective data sharing scheme for AI-uptake in the drug development process.

We are deeply grateful for the expertise, wisdom and generosity of all those consulted in the development of this document. We hope to continue our work with this committee and others to realize a more healthy, equitable and prosperous future.

Executive summary

The current drug discovery market is not responding sufficiently to health care needs where it is not adequately lucrative to do so. Unfortunately, there are a number of important yet non-lucrative fields of research in domains including pandemic prevention and antimicrobial resistance, with major current and future costs for society. In these domains, where high-risk public health needs are being met with low R&D investment, government intervention is critical. To maximize the efficiency of the government's involvement, it is recommended that the government couple its work catalyzing R&D with the creation of a drug development ecosystem that is more conducive to the use of high-impact artificial intelligence (AI) technologies.

The scientific and political communities have been ringing alarm-bells over the threat of bacterial resistance to our current antibiotics arsenal and, more generally, the evolving resistance of microbes to existing drugs. Yet, a combination of technical capacity issues and economic barriers has led to an almost complete halt of R&D into treatments that would otherwise address this threat. When a gap arises between what the market is incentivized to produce and the healthcare needs of society, governments must step in. The COVID-19 pandemic illustrates the importance of bridging that gap to ensure we are protected from future threats that would result in similarly devastating consequences.

Artificial intelligence (AI) capabilities have contributed to watershed moments across a variety of industries already. The transformative power of AI is showing early signs of success in the drug discovery industry as well. Should AI for drug discovery reach its full potential, it offers the ability to discover new categories of effective drugs, enable intelligent, targeted design of novel therapies, vastly improve the speed and cost of running clinical trials, and further our understanding about the basic science underlying drug and disease mechanics.

However, the current drug discovery ecosystem is suboptimal for AI research, and this threatens to limit the positive impact of AI. The field requires a shift towards open data and open science in order to feed the most powerful, data-hungry AI algorithms. This shift will catalyze research in areas of high social impact, such as addressing neglected diseases and developing new antibiotic solutions to incoming drug-resistant threats. Yet, while open science and AI promise successes on producing new compounds, they cannot address the challenges associated with market-failure for certain drug categories. Government interventions to stimulate AI-driven pharmaceutical innovation for these drug categories must therefore target the entire drug development and deployment lifecycle to ensure that the benefits of AI technology, as applied to the pharmaceutical industry, result in strong value added to improve healthcare outcomes for the public.

Summary of Recommendations

This document puts forward a set of recommendations that, taken together, task governments with the responsibility to promote:

1. Research and development in fields of drug discovery that are valuable to society and necessary to public health, but for which investments are currently insufficient because of market considerations.

2. Uptake of AI throughout the entire drug discovery and development pipeline.

3. A shift in culture and capabilities towards more open-data among stakeholders in academia and industry when undertaking research on drug discovery and development.

The Role of GPAI

The drug discovery and development process is a multidisciplinary, multi-industry behemoth, involving stakeholders ranging from pharmaceutical giants, biotech and tech start-ups, academia and healthcare systems. Thus, improvements to the R&D process must consider each of these stakeholders and thereby involve answering scientific and technical questions, as well as taking into account political and economic realities.

For infectious disease drug research, the international nature of the work adds a further challenge since pathogens, especially those that are resistant to drugs, can quickly become a global threat. Lower and middle-income countries (LMICs) often face unique challenges with regards to the fight against emerging and neglected diseases, and can become reliant on drugs whose development is controlled by a foreign organization or government that controls access to valuable intellectual property (IP) rights. New technologies and resulting therapies will often become available in developed nations first, but equitably and effectively disseminating them to LMICs is in the best interest of all parties, preventing localized public health threats from spiralling out of control and becoming a global menace. In particular, AI carries the promise of bringing about transformative changes to the drug discovery and development process; however, effectively capturing this game-changing technology will require navigating challenges in the scientific, political and economic arenas.

The Global Partnership on Artificial Intelligence ("GPAI") is a multi-stakeholder initiative with 19 member nations, guided by the principles that informed the OECD's recommendations on Artificial Intelligence. GPAI's roster of experts includes contributors from academia, industry and the public sector, providing the broad perspective that is needed to address the multi-faceted issue at hand. Thus, GPAI is well positioned to bring unique insight to the question of: "How can we create and effectively leverage AI for the purpose of re-orienting the drug discovery ecosystem towards more equitable and inclusive healthcare outcomes?"

This work has benefited from the full array of expertise at GPAI. This expertise includes that of the Organisation for Economic Co-operation and Development (OECD) and the Working Group on Data Governance. Furthermore, the large and diverse group of GPAI member nations have encouraged the promotion of international cooperation, which is critical given the global nature of many public health threats. The international membership of the organization has also reinforced the need for this work to be done equitably, in terms of the geographic distribution of research, affordability to and accessibility of lifesaving medicines, and investment in areas of global importance (including neglected diseases and antimicrobial resistance).

On the sensitivity of medical data

While this document makes the case for open science and data sharing more broadly, it is critically important to remember that not all data is created equal, which has implications for *how* we ought to be sharing data. These sensitivities vary depending on whether the data was collected in an academic or clinical setting, regards lab assays involving samples from people or not (e.g. purely chemical or on non-human cells), and/or is derived from a particular demographic and/or geographic group. As such, later sections acknowledge that there should be a difference in how data ought to be treated depending on its sensitivity to ensure that fundamental human rights are unequivocally respected within the context of this initiative. More detailed discussions of these issues can be found in the Risk and Risk Mitigation section.

Introduction

The danger that infectious agents pose to our health, lives, livelihoods, and even political stability seems, until recently, to have been underestimated by the public. The COVID-19 pandemic has starkly demonstrated our vulnerability, and also raised awareness about future threats. Indeed, analyses of global trends indicate that the number of emerging infectious diseases is expected to rise over time, highlighting the importance of better preparing ourselves against future pandemics[1].

COVID-19 vaccines can also serve to demonstrate the importance of rapidly developed countermeasures to infectious diseases. However, the speed with which COVID-19 vaccines were developed is unfortunately not the norm. For non-vaccine drugs, some estimates say that it typically takes about 10-12[2] years for a new medicine to complete the journey from discovery to the marketplace and the cost is on the order of 2.6 billion dollars[3]. Progress on development in some categories of drugs, such as antimicrobials targeting gram-negative bacteria, has almost completely ground to a halt, with very few chemically novel compounds being put into clinical trials over the last 50 years[4]. The trials themselves constitute an important time-intensive barrier; even in the case of COVID-19, it only took a couple of months to find the first drug candidate but it took 8 months to complete the trials. As such, the combination of technical challenges and the current state of the drug development ecosystem makes both research and health and safety a time- and resource-intensive endeavor, and, as a result, it is nearly impossible to quickly research and deploy vital medicine.

These issues are compounded by the concerning development of antimicrobial[5] resistant pathogens, as is well studied in the O'Neill report[6]. Indeed, the evolution of drug resistant pathogens radically increases the threat of infectious diseases by undermining the efficacy of existing antimicrobials. Since the entire healthcare system depends on access to effective antimicrobials (e.g., for **any** surgical procedure), it has been estimated that unless we change course, the emergence and spread of new antimicrobial resistant pathogens could increase deaths by more than 10 times, to 10 million per year by 2050 (which is much higher than the number of deaths due to COVID-19). The accumulated healthcare and economic cost of this scenario is estimated to be on the order of 100 trillion dollars. Already 2.8 million Americans get infected by antimicrobial resistant pathogens and 35,000 of them die each year. According to the CDC, the annual cost of antimicrobial resistance in the United States is $55 billion[7] ($20 billion in excess direct healthcare costs and $35 billion from productivity loss) [8].

As we have witnessed with COVID-19, infectious pathogens can reach a global scale and thereby necessitate coordinated international action in response. Although the whole of humanity is at risk (a potentially existential risk) of infectious agents, the poorer countries are often those which suffer the most (as illustrated by the case of malaria and tuberculosis). However, rich countries are also vulnerable[4] (as illustrated by the case of COVID-19). Thus, scientific methods and infrastructure that radically reduce the time and cost of drug development will be necessary to ensure that society is prepared to handle future infectious disease outbreaks.

Introduction to AI/ML

Machine Learning (ML) is the field of developing computer programs that can learn from data how to best perform a task of interest, without needing explicit instructions. ML techniques make up the largest and most significant

1 Jones et al. Global trends in emerging infectious diseases. Nature (2008)

2 Lexchin J. (2021) Time to market for drugs approved in Canada between 2014 and 2018: an observational study, BMJ Open

3 Biopharmaceutical Research & Development: The Process Behind New Medicines (2015)

4 "Antibiotic Resistance Threats In The United States, Centers for Disease Control and Prevention (2019)

5 Pathogenic microbes include bacteria, viruses and fungi, and all three types are subject to the evolution of drug-resistant variants.

6 O'Neill Review into Antibiotic Resistance, British House of Commons. (2017)

7 These numbers are calibrated according to the U.S. dollar price in 2008

8 https://www.cdc.gov/drugresistance/pdf/ar-threats-2013-508.pdf (2013)

portion of the R&D done in the broader field of Artificial Intelligence (AI), and in popular contemporary diction the two terms are often used interchangeably.

The success of AI methods depends on both the development of strong learning algorithms as well as the presence of useful datasets on which to train them, and computational resources to perform the training. A machine learning algorithm will extract useful patterns from its training dataset, and use these patterns to make accurate predictions about future, novel examples. For example, after training on a dataset consisting of many examples of drug-like molecules, along with their toxicity levels, an AI predictive system might learn about patterns of chemical sub-components that make a molecule likely to be toxic. Given an example of a new molecule, it could be able to estimate its most probable toxicity level. Another type of AI system could generate new molecular structures likely to have these desirable properties[9] as estimated by the AI predictive systems, thus being particularly useful as part of the exploration in the high-dimensional space of candidate drugs.

AI methods excel at tasks where there are large amounts of relevant data available to learn from. Once this prerequisite is met, AI's capabilities can exhibit impressive strength and flexibility, which has given rise to the use of AI methods across many industries today. This can include the ability to find relevant patterns from data that would be far too complex for a human to wrap their mind around, such as mapping the relationship between the sequence of amino-acids forming a protein and their 3-D shape[10] (which determines their function), or extracting useful information from the millions of components in a cell's DNA sequence and the nature of the organism that the cell will develop into. Beyond making predictions, AI methods have also excelled at training decision-making agents to work inside a dynamic system, as is seen with AI bots that can play complex games at a superhuman level, or even with the advent of self-driving cars.

Given AI's dependence on training data, insufficient training data (whether in size, quality, relevance or other forms of bias) will prevent AI from being accurate or effective. AI models perform most strongly when an answer can be derived by interpolating between various training examples. When a prediction requires extrapolating beyond the distribution of training examples, performance will start to drop off, in what is referred to as Out of Distribution (OOD) errors. For example, an AI model that is trained on a dataset exclusively consisting of a specific class of drug may struggle to make accurate predictions concerning molecules belonging to a separate class of drug with different molecular characteristics. Furthermore, modern machine learning methods are not capable of performing many tasks currently associated with human intelligence, such as deep conceptual understanding and creativity. Additionally, while the research areas of causal machine learning and AI explainability are enhancing our capacity to derive conceptual knowledge from the decisions of AI algorithms, many machine learning methods are still considered "black boxes". They may be capable of making accurate predictions, but for now often cannot be used to effectively explain the patterns in the training data that were leveraged during the prediction-making process in a way that is understandable to humans.

Despite the current limitations described above, AI has already become a paradigm-shifting force across a variety of industries and areas of research. Its influence is being felt across a number of fields, such as: search engines, smartphones, commerce, advertising, language translation, healthcare, robotics, finance and manufacturing. The pharmaceutical and biotech industry is beginning to realize that AI can similarly offer significant value to the arena of drug discovery, as witnessed by numerous recent investments. With carefully directed guidance, the GPAI committee believes that AI can become a powerful tool not just for greatly increasing the success of drug discovery in disease areas with clear commercial drivers, but also, with interventions from governments, for furthering public good with regards to the current and upcoming public health challenges facing the global community.

9 Bengio et al, Flow Network based Generative Models for Non-Iterative Diverse Candidate Generation, NeurIPS'2021.

10 Jumper et al. Highly accurate protein structure prediction with AlphaFold. Nature 596.7873 (2021): 583-589.

Given the revolutionary potential of AI in this area, one may wonder why the international community, governments or philanthropic organisations need to be involved at all. To answer that question, we must think about why and how the current academic and industrial research ecosystem in individual countries must and can be steered to catalyze AI's potential faster, at a lower cost and in ways that better serve global needs.

Subjects of Concern in the Drug Discovery Ecosystem

Introduction

The drug discovery and development ecosystem is exceptionally large, intricate and complex, belying attempts for sufficiently concise summarization. The drug discovery and development process involves players from a number of arenas, including academia, large pharmaceutical companies, smaller biotechs and other start-up ventures, healthcare services providers and insurers, governments and regulatory bodies, and other NGO entities such as the World Health Organization (WHO) or the Drugs for Neglected Diseases Initiative (DNDi).

The drug discovery process begins with the identification of some druggable target relevant to the disease of interest. Typically this is either a molecule present in and on cells of the human body or, for infectious diseases, a molecule in a bacterial cell or a viral particle. Identifying this target can involve scientific research into the causal biological mechanisms underlying the target disease.

This work is typically followed by the pursuit of a relevant drug-like molecule capable of interacting with the druggable target and producing some desirable effect. Concerns such as issues with binding affinity, toxicity, or off-target effects will screen out the vast majority of candidate drugs at this early stage.

Once candidate molecules are identified, the candidates are tested, beginning with laboratory tests and proceeding, for those candidates showing promise, in pre-clinical animal models and eventually multi-stage clinical trials in human patients. If a candidate drug makes it through the gauntlet of the clinical trials process, applications may be made to relevant regulatory bodies seeking approval to bring the drug to market as a treatment. The total cost of bringing a drug through the clinical trial and regulatory review process has been estimated to be between 1 and 3 billion dollars, and the process can take up to 10 years or more, in large part because of a high failure rate, i.e., because of candidate drugs that are selected and evaluated but end up not being appropriate. Being able to better predict and select which candidates are more likely to survive the drug development process would thus be a game changer.

Obtaining marketing approval is not the end of the development process. During the development process and beyond, drug developers invest in creating industrial processes to fabricate the drug at scale, in optimizing the formulation to be used to deliver the drug, running additional clinical trials that continue to test approved drugs in new indications and patient populations, as well as in promoting the drug with prescribers, marketing the drug and facilitating its delivery to prescribers and dispensers. Drug developers must also conduct post-approval surveillance and monitoring of the drug's performance in patients around the world.

There are a variety of factors that influence how far along the discovery/development pipeline a particular treatment will progress. Significant academic and industry interest in a disease may result in an increased understanding of the underlying biological processes, leading to a larger number of druggable targets to pursue. Government incentivization, such as pre-purchase agreements (or favorable terms), co-development, and even running joint clinical trials can accelerate the process or enhance the investment, as was seen with COVID-19 vaccines. Ultimately, as the system is currently structured, the single most significant factor is the commitment of large industrial players, as well as venture funds and the startups they support, to invest financially in the

development of a drug, which is typically dependent on the ability to foresee a sufficient profit-to-risk ratio for that investment.

Current Challenges

A form of market failure[11] has been identified within the drug discovery market, particularly for antibiotics[6] but also in the treatment for many other infectious diseases[12]. This market failure arises because of a mismatch between the societal value of innovations (like AI techniques to help discover new antimicrobials, or the novel antimicrobials themselves) and their commercial value[13]. One issue is that novel antibiotics are reserved for patients infected by bacteria resistant to older front-line drugs, reducing the volume of sales. This contributes to the market demand for new antibiotics being currently low, and thus the vast majority of pharmaceutical companies have essentially stopped research on such drugs. This is in spite of the trillions of dollars in damages to the global economy (between 60 and 100 trillions till 2050 according to the O'Neill report) and hundreds of millions of lives at stake in the coming decades. There also exists a large number of neglected diseases in LMIC experiencing the same lack of funding towards critically important treatments. Between 2000 and 2011, 11% of the global disease burden came from diseases that received little or no attention from the pharmaceutical industry, primarily due to lack of perceived economic opportunity[14].

Due to the mismatch between profitability and global public benefit, it is necessary for governments to step in[15]. They could provide financial incentives[16] (for example via subsidies, grants, tax credits, advance market commitments, procurement, or other reward mechanisms) for private or public organizations to fill the gap, although we have also seen substantial investments from philanthropy as well. Note that governments and philanthropy already pay for a large part of the research costs of drug discovery, as has been documented for the AstraZeneca-Oxford Covid-19 vaccine[17], as well as the Moderna vaccine[18], which is one example of multiple successful projects funded by the BARDA program in the United States[19].

Patents have been used to provide an indirect but powerful financial incentive for the discovery of many drugs. However, a patent-centric incentive system on its own is not conducive to the sharing of data and intermediate scientific results that would be beneficial to the progression of AI methodologies for drug discovery and development. A large number of costly innovation efforts are currently kept confidential for commercial purposes and experimental results are not shared, because until and unless a patented drug results from these efforts, firms have limited means of protecting the potential value of their data other than keeping it secret. This becomes a barrier to the progress and deployment of AI-based solutions for drug discovery, which requires pooling the available sources of data and scientific knowledge for maximum efficiency, reproducibility, and chances of

11 Levy M, Rizansky, (2014) A Market failure in the pharmaceutical industry and how it can be overcome: the CureShare mechanism, Eur J Health Econ.

12 Trouiller P et al. (2001 Drugs for neglected diseases: a failure of the market and a public health failure? Trop Med Int Health.).

13 Note that because research and clinical trials are conducted on a wide range of potential drugs, the successful drugs need to generate sufficient revenue to cover not only their own research and trial costs, but also the costs incurred in unsuccessful research ventures and clinical trials. Accordingly, the mere presence of some potential demand for a drug, even if significant, may not be sufficient to justify research efforts.

14 Yamey G et al.(2018) Funding innovation in neglected diseases. BMJ.

15 The WHO Council on the Economics of Health for All, (2021) Governing health innovation for the common good, Council Brief No. 1

16 Dutescu, Ilinca A., and Sean A. Hillier. (2021) Encouraging the Development of New Antibiotics: Are Financial Incentives the Right Way Forward? A Systematic Review and Case Study., Infection and Drug Resistance and Morel, Chantal M., et al (2020) Industry incentives and antibiotic resistance: an introduction to the antibiotic susceptibility bonus, The Journal of antibiotics

17 98% of the costs of the research leading to the AZ vaccine came from public funding, according to this study: Samuel Cross et al. (2021, forthcoming), Who funded the research behind the Oxford-AstraZeneca COVID-19 vaccine?

18 Sharing The NIH Moderna Vaccine Recipe, Public Citizen report (2021).

19 See the description of the program.

success. Legal and regulatory mechanisms such as data exclusivity rights can also further obstruct the dissemination of scientific data[20].

Furthermore, separate to the issue of IP concerns, in drug discovery as in many scientific fields, often data is not published if it concerns an unsuccessful experiment such as a failed lead. Yet, this data would be very useful from the point of view of training AI systems, which requires quality negative examples to learn from, in addition to the positive examples of successful drugs. This issue of the current R&D climate as it relates to data and knowledge sharing is further explored in the Open Science and Open Data section of this report.

Opportunities

Just as market agents tend to optimize for profitability (which can result in suboptimal conditions from a social standpoint), governments are expected to optimize for social good and influence markets accordingly. Therefore, it is governments and not the market agents which should play a leadership role in orienting the drug development industry in favor of socially beneficial outcomes. Although this shift in perspective concedes that governments will have to shoulder some of the financial risk, the governments' diversified portfolio of investments will allow the successful initiatives to compensate[21] for the loss of the unsuccessful ones, even more so if these investments are pooled across many countries.

As such, we recommend that governments increase their activities in funding or incentivizing drug discovery and development projects that are of high social value, but would not otherwise happen at a sufficient scale in the profit-driven market. This will result in significant benefits to the ability of the drug discovery and development ecosystem to deliver better value to society. As described below, these benefits include:

1. *Directing work towards proactively preparing for future threats.* A significant goal of the recommendations in this report is to ensure that global society is better prepared for future pandemics. Government-directed incentives can ensure that vital basic research is performed on vaccine technologies, novel antimicrobial drugs and the application of AI to drug discovery before there is a pandemic-created market demand and in anticipation of the evolution of future pathogens, greatly reducing the human and economic costs of future public health crises.

2. Unblocking the development of treatments for unprofitable but high social value targets such as neglected diseases and antimicrobial resistance. As described in the Current Challenges section, there are a number of disease targets for which work receives greatly under-proportional funding due to lack of foreseeable profitability. A significant benefit of introducing government-funded R&D and procurement-based incentives will be to ensure progress is made on these target areas for which there is currently low market value but high social value, as seen in the CureShare mechanism[22].

3. *Encouraging multidisciplinary open-science research practices.* The application of AI technologies to drug discovery will be an inherently multidisciplinary area of research resting on the access to appropriate data by researchers. By giving preference to funding projects with experts representing all relevant disciplines and committed to appropriate sharing of information, the pace and quality of research can be improved over what would be achieved by allowing progress to be siloed within each field, each lab or each company.

20 Data exclusivity in international trade agreements: What consequences for access to medicines?, Médecins Sans Frontières technical brief (2004)

21 Mazzucato, Mariana (2016) From market fixing to market-creating: a new framework for innovation policy. Industry and Innovation

22 Levy M, Rizansky, (2014) A Market failure in the pharmaceutical industry and how it can be overcome: the CureShare mechanism, Eur J Health Econ.

AI for Drug Discovery

Introduction

AI has brought about and/or accelerated massive developments across a variety of disciplines, including significant achievements inside the biomedical and pharmaceutical fields. AI and deep learning (DL) have already been successfully leveraged for drug discovery, being instrumental in the discovery of Halicin, one of the first novel antibiotics capable of inhibiting the class of gram-negative bacteria in the last 50 years[23]. In another notable example, AI was used to drastically improve performance on predicting the shape of folded proteins; a problem that had been considered a central question to the field for decades[24]. These examples should illustrate the potential for well-applied AI to enable domain experts to solve what had been previously intractable problems inside their field. AI should also be able to dramatically reduce the burden of expensive and time consuming wet-lab experiments during the early stages of drug development, allowing the efforts of a broader collection of academics and early-stage industrial entities to contribute meaningfully to progressing the field, as well as during late stages of drug development to reduce the costs of clinical trials, and with post-marketing surveillance, to identify potential, previously unidentified, adverse effects.

There is a large, diverse collection of sub-fields in the AI domain. Collectively, the field has the potential to be leveraged for the improvement of each stage in the drug discovery and development process, including post marketing and surveillance. AI can be useful not only for screening candidate molecules for useful biological properties (as in the case of Halicin[25]), but also in designing novel molecules to be synthesized to solve problems for which no currently existing compounds are suitable; for improving the design and cost of clinical trials, for evaluating the impact of a potential drug for an individual based on their genetic or their molecular make-up or that of their pathogens, and for improving the forecast, surveillance and detection of emerging infectious diseases before they break out into pandemics.

Current Challenges

Although we are seeing early evidence of AI's potential in drug discovery, many fundamental as well as applied research questions require significant efforts to bring this potential to fruition. Many of these questions will require highly multidisciplinary efforts as well as a scale of funding which pose challenges in areas where there is a mismatch between commercial value and societal value, for which appropriate organizations such as GPAI can play a useful coordinating role.

One of the challenges common to all interdisciplinary endeavours is that a high level of communication and coordination is required between players from different domains of expertise. Without sufficient synchronization, it is far too common to see work that is redundant or futile being undertaken by one side alone. AI experts in academia or start-ups could commit significant effort to drug discovery without understanding issues that seasoned drug discovery experts know to be most important. Similarly, without consulting with sufficient AI expertise, pharmaceutical researchers may miss the best opportunities to leverage AI inside of their existing R&D pipelines.

As described previously in the Introduction to AI/ML section, AI methods can only succeed when there are sufficiently large amounts of quality data available with which to train the learning systems. With some exceptions, there is currently a dearth of high-quality open datasets with which to train AI systems for drug discovery tasks, in particular for questions related to later stages in the drug discovery process such as clinical trials. This issue will be discussed in greater depth in the Open Science and Open Data section.

23 Stokes et al. (2020) A Deep Learning Approach to Antibiotic Discovery, Cell.

24 Jumper et al. (2021) Highly accurate protein structure prediction with AlphaFold, Nature.

25 Stokes et al. (2020) A Deep Learning Approach to Antibiotic Discovery, Cell.

In addition to requiring large amounts of data, some modern AI methods require access to large amounts of powerful and costly computer equipment. The scale of compute needed for state-of-the-art algorithms is becoming prohibitive for many smaller players, such as most academic labs or early-stage startups, or even larger companies. Without distributed access to significant computing architecture, there is the risk that progress will be increasingly made almost exclusively by large industrial players who have the financial means to make significant investments in computational resources.

Finally, there remain many tasks in drug discovery for which current AI methods will not be fully sufficient. Further research will be necessary in order to push the boundaries of what AI is capable of achieving, especially when we move from the traditional static dataset supervised learning framework, where most ML successes have been, to dealing with iterative active learning setups, where ML retraining and experimental assays are integrated in an interactive and non-stationary setting to maximize the efficiency of the search for information and therapeutic solutions. In order to address the gaps in capabilities relevant to the drug discovery and ecosystem pipeline, AI research will have to be specifically targeted towards relevant goals, opportunities and constraints in consultation and collaboration with medical and pharmaceutical subject matter experts.

Opportunities

The development and implementation of AI methods specific to drug discovery has the potential to vastly accelerate progress on areas of significant importance to public health. As such, we recommend that governments seek to increase the effective use of AI inside the drug discovery and development ecosystem.

As was previously described, there is a large number of open challenges in drug discovery where new AI methods can be fruitful. We outline below leading examples of how AI can accelerate the drug discovery process, with mention of what additional research needs to be done.

1. *Estimating the properties of a candidate drug from data.* This is an important element in determining whether a candidate drug would be appropriate as a therapeutic in general and to treat a particular disease. A supervised learning system can be trained with the appropriate data to associate candidate molecules (represented by their molecular structure or sequence) with properties of interest, such as drug likeness (is the molecule small enough and with the right properties to get into the right places in the body), toxicity, low production price, easy storing, long shelf life, easy transportation, easy manufacturing and of course, affinity to one or more targets of interest. Interesting challenges remain, however, in terms of how to process molecular descriptors (e.g., graph neural networks are currently doing best but research on their limits and properties is still in its infancy) and how to take advantage of multiple sources of related data (e.g., the affinity of a large set of drugs with a large set of proteins, protein-protein interactions, etc) as a form of multi-task learning (where, again, graph neural networks may be useful, but the datasets are much larger, posing other challenges).

2. *Improving relevant biological, chemical and physical models so they can be simulated. With* limited amounts of costly experimental data, or simply to obtain better generalization, one can take advantage of appropriate causal and domain knowledge, either in the design of the machine learning predictors introduced above or to generate cheaper approximate data regarding the expected associations between drugs and their effects. Such knowledge can be crucial when very little experimental data is available but these models can also benefit from different kinds of data (i.e. patient data) in order to tune some of their parameters, or machine learning can be used to approximate and accelerate simulation of these models. A particularly interesting research question is how to exploit results from experiments where biological interventions (intervening on genes or with drugs) are followed by measurements of high-dimensional biological data of cell state or dynamics (such as gene expression or images of cells undergoing these perturbations), in order to elucidate causal paths of influence characterizing the mechanisms of a cell of a particular type. Recent advances in causal machine learning could help combine existing biological knowledge with these observations and experiments, as well as suggest new large-scale experiments

using active learning, leading to progress in predicting the effect of drugs and in target identification (potentially multiple targets at once). Such models could also potentially help predicting side-effects or preventing resistance[26] or dependence.

3. *Active learning iterating between expensive experiments and machine learning.* This is necessary to go from the approximate knowledge in computational models to decisions based on real experiments and iterate back using the experimental results. Knowing that the machine learning predictors are going to be used to generate candidates not just once but multiple times means that a form of exploration and diversity of candidates can be beneficial in the generation or selection of these candidates (in contrast to picking only the best initial guess). The algorithms can be set up and optimized to take advantage of this interactive context in which the learning system can ask questions, propose experiments, in order to reduce epistemic uncertainty, acquire relevant knowledge, capture causal structure and explore efficiently the vast space of therapeutic interventions, molecules or of drug combinations. Existing research in active learning, sequential model optimization, Bayesian optimization and reinforcement learning is relevant, but this particular area at the intersection of drug discovery is still nascent and asks fundamental questions in AI research about purposeful knowledge acquisition.

4. *Learning to search in the space of molecules.* Combining all of the above advances, the question of learning a search policy for controlling the sequence of experimental data acquisitions can be framed in the context of reinforcement learning, and raises additional questions like figuring out the right abstract action space (how to move in the space of candidate molecules, e.g., adding atoms or larger blocks to an existing scaffold candidate, or consider appropriate mutations), and discovering the appropriate abstract space for representing molecules and planning in such a search.

5. *High-throughput screening technologies.* The speed at which experimental data can be acquired is going to be the key to success and rely mostly on research done outside of machine learning, for example in biological and chemical technologies enabling the simultaneous screening of hundreds of thousands or more candidates at once (e.g., with DNA Encoded Libraries[27], robotic chemistry platforms[28], synthetic biology[29] screening and similar ideas). Progress in robotics could also be important, with the longer-term objective of automating the entire experimental process and integrating the machine learning and experimental bench completely. Funding multidisciplinary research combining synthetic biology or robotics with machine learning to explore the space of candidate solutions is necessary to steer the development of more relevant biotechnology tools for high-throughput synthesis and screening.

6. *Incorporating patient-specific information, real-time monitoring.* The choice of treatment and the models in 1 and 2 above can be personalized by also feeding the machine learning models with patient data, such as genomic data, gene expression and other molecular and histopathological data, ethnicity, vital signs (possibly from new devices, like wearable sensors) and patient history (in particular immunity history, pathology imaging, etc). This is important because the appropriate treatment may depend on the stage of the disease and the particulars of the patient. This requires a very different kind of data, involving health records, medical and biomarker measurements from patients. Patient-specific combinations of drugs could also be considered, using machine learning to predict which combination is more appropriate given patient data. Antiviral drugs that aim to inhibit virus replication are most effective when the virus is actively replicating during the incubation period and early symptomatic period, whereas inhibitors of inflammation are best to introduce later, in order to properly control inflammation. This requires continuous and real-

26 A Bold New Strategy for Stopping the Rise of Superbugs, The Atlantic (2018)

27 Goodnow et al. (2017), DNA-encoded chemistry: enabling the deeper sampling of chemical spac. Nat Rev Drug Discov

28 Schwaller et al. (2019), Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. Chemical Science.

29 See the following reviews on synthetic biology:

1. Thaker MN, and Wright GD (2015), Opportunities for synthetic biology in antibiotics: expanding glycopeptide chemical diversity, ACS synthetic biology.
2. Kim et al. (2015), Reinvigorating natural product combinatorial biosynthesis with synthetic biology. Nature chemical biology.
3. Smanski et al. (2016), Synthetic biology to access and expand nature's chemical diversity. Nat Rev Microbiol.
4. Davis AM, Plowright AT, and Valeur E, (2017) Directing evolution: the next revolution in drug discovery? Nat Rev Drug Discov.

time monitoring to detect the disease stage and anomalies at the earliest possible time, as was found for COVID-19. One pioneering example of such attempts is seen in the Warrior Watch Study[30] at Mount Sinai.

7. *Improving and scaling up trial designs.* Machine learning can also take advantage of new trial designs aimed at reducing the cost and duration of clinical trials, while making them more representative and better targeted. This is achieved by targeting volunteers from the larger population of infected patients using digital technology to report the effects of treatments, potentially with randomized cross-trials and models trained on the background population as control (synthetic arm). Such trials could at least eliminate the need for unnecessary but costly classical clinical trials, for which the treatment is not working in the cheaper general population. Machine learning could also be used as part of adaptive clinical trials[31], to extract more useful information faster from them. Finally, machine learning can be used to generalize the results of a trial to a broader population than that included in the trial.

8. *Translating results across species and biological contexts.* There is a well-documented challenge in pharmacology research associated with using the results of pre-clinical animal studies in a laboratory setting to predict the effects of a compound during later stages of clinical trials[32]. Computational and AI approaches should be developed that can aid in the prediction of in-vivo human performance from an accurate in-silico computer human model.

9. *Development of privacy preserving methods.* Individual medical information is subject to significant privacy risks and corresponding regulatory protections, especially in the case of data modalities such as electronic health records (EHR). These protected data sources have the potential to provide invaluable information to researchers if they can be accessed safely and responsibly. Areas of ML research such as federated learning[33] are increasing the capacity to learn from data without ever needing to directly access individual records, in a privacy-preserving fashion that can unlock the use of more sensitive data modalities.

10. *Automating bespoke drug development for personalized medicine*. In the current context, drugs are only approved by regulatory authorities once, which can make the process of adapting drugs to suit the unique genome of individuals or of a mutating pathogen prohibitively expensive. However, with slight adaptations, these drugs can be tailor-made to suit the unique needs of patients (based on their genome and that of the pathogen). This context-specific drug development process can even be applied to adapting the drugs based on a rapidly mutating pathogen or cancer in the body of the patient. This can be used to ensure that vaccines retain their efficacy in the context of a mutation. However, in order for this to work, the regulatory approval process must be adapted (in the spirit of cell therapies), wherein the relevant regulatory bodies approve the general procedure that produces a therapy, but the drug developers would still have the flexibility to adapt the drug based on changes in the patient population and pathogen composition.

30 See the description on the project website.

31 Pallmann et al. (2018) Adaptive designs in clinical trials: why use them, and how to run and report them. BMC Med 16

32 Van Norman et. al. (2020) Limitations of Animal Studies for Predicting Toxicity in Clinical Trials: Part 2: Potential Alternatives to the Use of Animals in Preclinical Trials. JACC Basic Transl Sci.

33 Kairouz et. al. (2021), "Advances and Open Problems in Federated Learning", Foundations and Trends® in Machine Learning

Open Science and Open Data

Introduction

The goal of open science is to promote unhindered access to the outcomes of scientific research, including articles, methods, theories, insights, experimental results, metadata, experimental reagents, tools and datasets. The benefits of pursuing a culture of **open science** are well documented. As was described in the Canadian Roadmap for Open Science[34], the benefits of open science practices include increased reproducibility, accelerated knowledge transfer, enhanced diversity and inclusion, and greater opportunities for impact. It is for this reason that the OECD's Committee for Scientific and Technological Policy has been arguing for access to data to become a major policy priority within the OECD[35].

Science is, fundamentally, an incremental process, with each great discovery building upon previous work. AI is no exception to this rule. Thus, creating a culture of open science is one of the most effective ways of accelerating AI's capacity to aid in the drug discovery and development process.

Of particular interest in the context of AI for drug discovery is the principle of open data, which is the practice of making datasets used in and generated by R&D, or other industrial activities, open and freely accessible whenever possible, at least to researchers. This concept has been endorsed by 85 governments who adopted the "Open Data Charter"[36]. The Open Data Charter commits adopter countries to the development of open data policies that make data accessible and freely available while still ensuring the rights of peoples and communities are protected[37].

Open data is critically important because, as previously mentioned, AI methods are "data hungry". The largest family of AI methods are what is known as supervised learners, requiring high-quality labelled data in order to learn about the task that the AI is being trained to perform. The difference between a successful and unsuccessful AI-centric project is often simply the volume, diversity and fidelity of data that has been collected.

Many of AI's success stories to date, including advances in image classification and protein structure inference, involved large, high-quality benchmark datasets that had been made publicly available (including those used in the ImageNet and CASP challenges).

In order to best utilize the plethora of powerful AI techniques that are being developed in the context of drug discovery for infectious disease, it is imperative that larger and higher quality datasets are made available to all relevant practitioners. This is of particular concern to academia and smaller players within industry, for whom the financial barriers to obtaining sufficient data can be significant.

The application of AI to drug discovery and development is also making an increasing number of data modalities relevant. While research such as the discovery of Halicin[38] may have only required access to a dataset of drug-like molecules, there are a much larger number of opportunities for AI to contribute, as discussed in the AI for Drug Discovery section. Many of these applications will require datasets that are useful when analyzing a drug's performance in later stages of the clinical trials, a drug's efficacy in combination with other drugs, or how a drug's performance is affected by an individual's unique genetic makeup. The challenge of finding relevant accessible datasets for these sorts of tasks is often much more difficult than finding open chemical data relevant to early-stage drug discovery.

34 Office of the Chief Science Advisor of Canada (2020), Roadmap For Open Science

35 OECD (2021) "Recommendation of the Council Concerning Access to Research Data from Public Funding" page 3

36 Charter description here

37 Open Data Charter principles (2016)

38 Stokes et al. (2020) A Deep Learning Approach to Antibiotic Discovery, Cell.

Current Challenges

There are several factors contributing to the lack of high quality open data. For example, when data has been produced by a for-profit entity, there is often a disincentive to share in the absence of any mechanism to obtain compensation for third-party use of data that was costly to produce. Proprietary datasets can be rightly considered assets of significant value, especially as the capacity for AI and data-science to mine valuable patterns increases.

Furthermore, there are some datasets that are unethical or illegal to share openly. For example, data concerning the private medical information of identifiable humans is heavily regulated in many jurisdictions. This is a topic of significant relevance, as it is imperative that the pursuit of AI not result in data sharing practices that infringe on people's basic human rights. A more in-depth discussion of this issue is contained in the Risk and Risk Mitigation section of this report.

Finally, there are also types of data that are not currently collected properly to provide the size and quality that is required for AI techniques to excel. However, advancements in the fields of biotechnology, synthetic biology and robotics carry great promise to provide novel datasets of much larger size and lower cost in the near future.

The current lack of data can severely limit the ability for smaller players to perform high-end AI centric research. In order to learn to understand a concept, such as what makes a molecule a suitable drug, AI models require access to both positive and negative examples in great numbers. This factor further demonstrates the importance of data sharing. While the results of successful clinical trials must eventually be made public, unsuccessful research is not always published in peer-reviewed journals, resulting in redundant efforts as well as a lack of the types of negative examples that are needed for well-trained AI models to accelerate drug discovery across the whole ecosystem.

This barrier is especially concerning because it severely limits the capacity of stakeholders who are most willing to work on public health threats that are not profitable enough for large pharmaceuticals to solve (such as antimicrobial-resistant bacteria and neglected diseases). Through mechanisms that will be explored later in this report, government intervention in the domain of drug discovery and development will be much more impactful if it involves strongly encouraging data-sharing, to allow as many stakeholders as possible to leverage high-impact AI methodologies.

Opportunities

We recommend that governments promote a culture of open science and open data as much as possible throughout the drug discovery and development process, especially as it pertains to the use of AI technologies. Whenever possible, governments should make open data a requirement in order for access to bidding on government-funded drug discovery and development projects or procurement contracts.

The government should pay particular attention to opportunities that increase the capacity for biotechnology to produce massive and high-quality open datasets. When data is highly valuable to research but of a sensitive modality, we suggest that the recommendations discussed in the Risk and Risk Mitigation section be followed in order to determine the best opportunity for safely leveraging it for AI-driven drug discovery. In the case that some valuable and particularly hard to obtain datasets are being kept proprietary by for-profit entities, we suggest that the discussion in the Novel Data Sharing Frameworks section of this report be considered to help policy makers in their approach to alternative data-sharing incentivisation schemes.

Embracing a culture of open science and open data will greatly benefit the drug discovery and development ecosystem, especially as it pertains to exploiting the newfound capacities of AI technology. Some of the benefits associated with open data are summarized below:

1. Accelerating the speed of research. Policies that push the drug discovery ecosystem towards an open-science and open-data model will increase the pace of research for several reasons. First, greater access to data has been found to accelerate research. Second, sharing negative results in experiments will significantly reduce the likelihood of duplicated work across organizations and allow organizations to allocate their time to more efficient pursuits. Third, it will allow more groups to explore new AI techniques on the available data, greatly speeding up the rate of progress in AI. Finally, more datasets, which are necessary to train AI algorithms, give rise to more robust systems when fed with a diversity of rich data sources, yielding higher quality predictions.

2. Reducing research costs. In addition to the cost reductions directly resulting from the increase in research speed, it can be expected that an open-science and open-data environment will further decrease the cost of novel developments, as discussed in the Canadian Roadmap for Open Science[39]. In particular, when data-sharing is enforced across successive projects, the costs that would otherwise be associated with repetitive data-gathering procedures (which often require significant wet-lab investments) would be reduced by avoiding unnecessary duplication.

3. Equitably including R&D talent from LMIC. By facilitating the development of high quality and open-access datasets, a larger number of academic and industrial groups from around the world will be able to contribute more significantly to early-stage drug discovery research, even without access to large amounts of cost-prohibitive wet-lab infrastructure. In particular, AI research groups are lacking such wet-lab equipment and expertise. In addition to amplifying the contributions and rise of early-stage startups, this shift in the research model will make better use of the research talents within academic institutions, which have not yet been able to invest in laboratory equipment (including many high-ability researchers based out of LMIC). Data sharing thus has the potential of opening up the innovation process and creating a more global value chain[40].

Recommendations

1. Governments need to invest in multi-disciplinary academic research in the field of AI-driven drug discovery. Governments should especially fund research into applications of AI in public health concerns where there is currently insufficient commercial interest and investment. To benefit the R&D pipeline and society more broadly, government-funded academic research should be in an open-science, open-source and open-data legal framework. Additionally, some grants should be specifically targeted towards funding the construction of high quality open datasets, as well as cross-discipline collaborations that enable the development and testing of AI algorithms for tasks that are of particular interest to public health, such as that of novel antimicrobials.

2. In order to best facilitate multidisciplinary academic research in the field of AI-driven drug discovery, governments should incentivize AI-capacity building inside of the drug discovery and development ecosystem. This should include developing AI literacy across all aspects of the ecosystem, an emphasis on access to quality training programs, and the development of software tools and resources to facilitate increased AI uptake.

3. Governments should set up novel innovation procurement programs for stimulating and incentivizing the efforts of biotech, pharma, healthcare or public research organizations, downstream from the academic research, to go from academic prototypes (software, biotech methodologies, candidate drugs) to industrial-strength development pipelines and optimized drugs.

    a. One of the goals of these programs should be to substantially increase knowledge-sharing and data-sharing across organizations and disciplines compared to the current practice in industry

39 Office of the Chief Science Advisor of Canada (2020), Roadmap For Open Science

40 OECD (2021), Recommendation of the Council Concerning Access to Research Data from Public Funding

(ideally as open as with the academic work in (1)) in order to reduce costs, accelerate the pace of innovation, and enable more successful application of AI.

b. Another goal of these programs (for example through particular forms of licensing) should be to avoid abusive drug prices resulting from patent-supported monopolies, as well as favour low-cost access of the resulting drugs and fabrication methodology in LMIC.

Different contractual and licensing options for (a) and (b) are discussed below and should be expanded as per the incentives created. An important characteristic of the proposed procurement approaches is that they combine partial funding of the R&D costs (like grants) along with outcome-driven rewards, while allowing multiple private entities to compete in exploring different innovation approaches in parallel, with outcome-driven guarantees from governments that good results will be rewarded appropriately.

4. Governments should set up financial incentives (not necessarily regular patents, see below) to make sure that clinical trials are performed following successful outputs from the government's innovation procurement contracts, when the drugs being developed are sufficiently promising to address significant public health issues and when usual commercial incentives are not sufficient to motivate the pharmaceutical industry to fund the clinical trials themselves.

5. Governments should internationally coordinate the efforts recommended above to favour

    a. research collaborations, knowledge sharing and transfer of know-how across countries and in particular from richer countries to LMICs;

    b. more uniform innovation policies (across countries) in their procurement and incentive mechanisms to make it easier for companies that are involved to comply with similar legal and operational frameworks across countries;

    c. access to the resulting technologies and drugs at low prices in LMICs;

    d. joint funding on efforts with international scope;

    e. international mobility for students, researchers, and training activities, including industry and other stakeholder sectors;

    f. effective use of remote collaboration for capacity-building, access to computing resources and data, joint research, as well as regulation and ethics vigilance.

6. Follow-up this report with a deeper review and assessment of different procurement and incentive policies appropriate to reach the goals in (3) and (4), in particular regarding the objective of maximizing data sharing in a context where current regulatory instruments, such as intellectual property rights, are insufficient (patents do not apply, copyrights are insufficient, and trade secrets prevent sharing and lead to inefficiencies). This analysis should be specifically motivated by the consideration of how to best make available the data related to the results of pre-clinical assays and of clinical trials. The latter warrant additional attention as they will be far more likely to be obtained through industry partnerships rather than academic grants. Additionally, they will have considerable privacy concerns that may require the use of a Data Access Committee, or similar mechanism, to administer release. This body will ensure that there will be sufficient infrastructure to manage the regulatory considerations that come with incentivizing industrial organizations to share their data.

7. Either create a permanent international non-profit organization or leverage an existing one which will be responsible over the longer term to coordinate across countries, as well as manage internationally funded projects (from discovery to fabrication to deployment), and make sure to fund this organization to allow it to reach the desired goals. A discussion of the necessary criteria this organization should fulfill is included below.

Evaluating Candidate International Non-Profit Organizations

This section will lay-out the criteria that should be considered when designing or selecting an organization to implement recommendation (7) in this document. Examples of existing organizations that fulfill some or all of these criteria will follow.

1. *Country membership*.

    1.1. The organization should have diverse international membership in order to facilitate coordination as described in recommendation (5). There should be inclusion of the perspective from both developed nations and LMIC.

2. *Researchers and experts*.

    2.1. There should be sufficiently strong scientific and academic expertise within the organization or among its advisors to effectively guide research and development efforts. This will include the medical and epidemiological expertise in order to identify target areas of high importance to public health and social good. Broader scientific knowledge will be required in order to evaluate cutting edge research directions, such as novel AI driven methodologies, possibly via a committee of external evaluators. This will include the ability to balance both preventing the wastage of resources in ill-conceived directions, while still having the foresight to aggressively leverage upcoming opportunities in order to meet growing public health challenges head-on.

3. Mandate and outcome.

    3.1. The organization's mandate should focus on developing solutions for public health challenges where current market dynamics either fail or are not sufficient.

    3.2. An ideal organization would have the ability to push candidate drugs through the clinical trial process to market themselves, or will have exhibited the ability to negotiate favourable partnerships with industrial partners in order to do so.

    3.3. The organization's mandate should include a commitment to delivering low-cost therapies, especially for LMIC. When pursuing industrial partnerships, there must be an ability to avoid enabling predatory pricing schemes when drugs are ultimately delivered to market.

    3.4. The organization should contribute to the development of an open-data and open-science culture within the drug discovery ecosystem. There should be a directive that projects undertaken or supported by the organization will make an effort to ensure that resulting data are made publicly available (or at least available to credible researchers) whenever possible. When pursuing industrial partnerships, there should be a strong preference for partnerships with resulting IP structures that are favourable for public interests.

    - **Drugs for Neglected Diseases initiative (DNDi)** *[relevant criteria: 1.1, 2.1, 3.1, 3.2, 3.3]:* The DNDi is an example of an international NGO focused on developing drugs to treat neglected disease, while ensuring that resulting products are made accessible in regions where the target diseases are endemic. Through various industrial partnerships, the DNDi has developed nine treatments, with a major success story being the development of Fexinidazole, the first globally-approved treatment for human African trypanosomiasis ("sleeping sickness")[41].

41 Deeks ED (2019) Fexinidazole: First Global Approval, Drugs.

- **The Global Antibiotic Research and Development Partnership (GARDP)** *[relevant criteria: 1.1, 2.1, 3.1, 3.2, 3.3, 3.4]:* GARDP is a non-profit initially created in 2016 by the World Health Organization (WHO) and the DNDi, which became an independent organization in 2019. GARDP's mission is to develop treatments against drug-resistant infections for priority populations, whilst ensuring affordable and equitable access. GARDP has a history of developing partnerships with different industrial and academic partners for individual components of the drug-development life cycle from Discovery to Phase 3 and beyond, and has development programs that will help meet its strategic aims. Additionally, GARDP's experts have forward-thinking plans to evaluate the integration of AI into their drug discovery activities.

Policies

There are a number of policies that can be referenced as best practices in enabling open science and data sharing. These policies have been made at various levels of jurisdiction, including at the institutional, national, and international levels. Generally, these policies emphasize the importance of innovation, capacity building and participatory approaches to scientific discovery to meaningfully catalyze the move towards open science.

i) OECD's Recommendation on the Council on Health Data Governance[42]

For stakeholders to reap all of the benefits associated with open science, it is critical that there be an ecosystem conducive to data sharing. As such, governments are recommended to develop coordinated public policy frameworks that guide the ways in which healthcare data is shared. This recommendation document puts forward a strategy for implementing a national health data governance framework to encourage the availability and use of personal health data while maintaining the data's privacy and security.

ii) World Health Organization's Ethics and Governance of Artificial Intelligence for Health[43]

In order for AI's impact in the healthcare industry to be positive and sustainable, the technology must be applied in accordance with ethical norms and standards. The World Health Organization's guidance document identifies the core principles involved in promoting the ethical use of AI in the context of healthcare. These principles must be observed by stakeholders at each stage of the design, development and deployment process to ensure the technology achieves its mission of benefiting public health and medicine.

iii) European Union's Guidance on Innovation Procurement[44]

Procurement policies that are innovative in their approach and that seek to promote socially beneficial research and innovation are necessary in the context of developing AI for drug development. The European Unions' Guidance on Innovation Procurement offers a best practice solution for how governments can structure their bids to enhance the potential for innovative applications, wherein cross-disciplinary and experimental research is needed. According to the European Union, this approach ensures that the procured solution provides the most added value in terms of quality, cost-efficiency as well as environmental and social impact.

iv) National Institute of Health's Bridge2AI[45]

Open science is also enabled by an ecosystem of high quality researchers. The collaboration of these researchers is crucial for open science projects to benefit from multi-disciplinary expertise. To facilitate collaboration, the National Institute of Health (NIH) (medical research agency in the United States) developed a National Center of Excellence, which it hopes will catalyze the uptake of AI in the drug development process. In its announcement of the Bridge2AI program, NIH describes that the Center will be responsible for disseminating

42 OECD's Recommendation on the Council on Health Data Governance

43 *World Health Organization's Ethics and Governance of Artificial Intelligence for Health*

44 European Commision (2018), Guidance on Innovation Procurement

45 Discussed here.

products, best practices and training materials as well as being able to evaluate the merits of grant applications with its repository of in-house experts. The promise of centers of excellence is that they facilitate innovative, collaborative and participatory approaches to scientific discovery. Furthermore, they bring together a multidisciplinary group of experts that can generate novel yet socially relevant solutions. Fostering these types of clusters creates new opportunities to establish a competitive advantage in a new market. Furthermore, the public-private partnerships that are generated as a result give rise to synergies that are critical for economic growth and development.

Novel Data Sharing Frameworks

When the government funds a large fraction of the drug discovery and development effort, it can contractually impose data sharing clauses. Alternatively, or in addition, governments can adapt the regulatory environment to facilitate high-quality data sharing. In recent decades, the drug discovery process has been kept siloed with strong IP and data exclusivity protections and tightly-controlled trade secrets. Our recommendation (6) suggests studying and elaborating on legal protections to facilitate data sharing between profit-motivated stakeholders. In order to provide stakeholders with additional benefits (on top of government financing for having won the Request for Proposal) by virtue of becoming involved in a data sharing scheme, novel legal protections must be considered.

It is not expected that such changes could happen immediately but it is recommended that governments implement a long term strategy for incorporating data sharing into the innovation process. However, in the short term, we recommend that governments simply impose data sharing requirements (with appropriate protections for privacy) in their grants or innovation procurement contracts.

Options for Compensating Data-Sharing Firms

Option A: Financial Compensation

In order to financially compensate the data-sharing firm, one may consider the development of a new sui generis right. Such a right could confer the data-owning firm, that is engaging in the data-sharing scheme, with exclusive protections and benefits when the data they've shared is used and incorporated into a profit-making product by other stakeholder groups. This right may contain the following conditions:

- All data-sharing firms must register the data they've shared into a regulated system that records data about the creator of the dataset to ensure exclusive benefits are conferred onto that firm once the data has been released.
- The data-sharing firm cannot prevent other stakeholders from using the data that they have shared to the system.
- However, once a stakeholder profitably uses the data that has been shared, the data-owning firm is entitled to compensation in the form of royalties calculated according to FRAND[46] terms (Fair, Reasonable, and Non-Discriminatory) and linked to profits obtained from a product derived thanks to that data.
- If a firm leverages shared data and uses it to develop a successful product without compensating the data-owning firm, the data-owning firm retains the right to sue for a FRAND-based royalty, irrespective of any contractual dealings or negotiations between the two firms.

46 Layne-Farrar et al. (2007). "Pricing Patents for Licensing in Standard-Setting Organizations: Making Sense of FRAND Commitments". Antitrust Law Journal; and Contreras, Jorge (2015) "A Brief History of FRAND" Antitrust Law Journal

This option will likely involve the creation of a data trust, which is discussed further in the Risk and Risk Mitigation section below.

Option B: Recognition-Based Compensation

Alternatively, we can look to academia for the compensation model that works best in the context of government-procured work. Although academics are often required, according to the grant contract, to make their findings and their data publicly available, they retain the right to be compensated in the form of public recognition when another researcher leverages their work. In this context, it would be useful to develop a taxonomy for how collaborators are to be recognized for their contributions, including those responsible for the design, data acquisition, curation, analysis, vadiation, and final documentation[47].

Although public recognition does have the potential to generate commercial value, it is possible that the government will have to increase the dollar value of the contract with such a compensation structure. This is due to the fact that other stakeholders will be able to create products using the dataset, which can generate significant earnings that do not need to be shared with the data-generating company. As such, there are future earnings of untold amounts that the company is waiving their rights to. Therefore, without the possibility of securing financial benefit from the use of their data, companies may decide that the contract should cover not only the cost of the discovery but also the cost of foregone earnings from losing exclusive rights to their dataset.

As long as the social value of having the data be publicly available outweighs the increased cost to the government, this compensation scheme could work.

Option C: Government-funded compensation

Another option is to allow governments, or their representatives, (e.g. the organization managing procurement of drug discovery R&D) to assign a post-hoc value to the data that has been shared and compensate the data-generating firm accordingly. For example, those firms can mount a case to the government as to the value brought to society by their data (e.g., the data was used by another organization to develop a drug which turned out to have great public health value) and the governments or their representative can then decide on a fair compensation. This is similar to the notion of payments associated with project milestones and outcome-driven rewards in a procurement contract, except that those payments may come at an arbitrary time after the end of the original contract, and only if the data generated was instrumental in providing significant public health benefits.

Intermediate solutions are also possible, e.g., if the shared data enabled important advances on the path to obtaining a clear public health benefit (e.g., a new antibiotic has been discovered but clinical trials and deployment has not happened yet), the government could award an ad-hoc prize to the organization which generated and released the data. The objective is to spread the expected value of such contributions across earlier and less sparse rewards, reducing the risk for the data-generating firm obtaining no downstream reward for its work. This is also why we recommend that governments fund at least in part the initial data generating effort (when the data is publicly released and is deemed of good quality), reducing the time it takes for the data-generating firm to be compensated for its work.

Options for Compensating Governments

Option A: Financial Compensation

While the government would like to ensure that data is easily shared amongst relevant stakeholder groups, it is also possible that the government charge those who are not considered priority stakeholders for the data that is used. However, it is critical that this also be done in accordance with FRAND-based principles and does not contravene the purpose of data sharing for drug discovery.

47 OECD (2021), Recommendation of the Council Concerning Access to Research Data from Public Funding page 10

Option B: Recognition-Based Compensation

The government may also be compensated with formal recognition for the contribution that the dataset they funded made to the development of a product. Formal recognition of governments that were involved in facilitating data sharing and open science can be useful from the perspective of developing norms around data sharing. This is because recognition can signal collaboration and build trust among relevant stakeholder groups in the government's willingness to promote open science. Furthermore, this recognition-based compensation model improves the likelihood of good-faith-based interactions with international stakeholder groups and can pave the way for more formalized international agreements in the future[48].

Risk and Risk Mitigation

In the healthcare context, where data is personal and sensitive and the stakeholder groups are heterogeneous, there are a number of risks that data sharing can present. These risks include the possibility of contravening: i) privacy rights; ii) autonomy over one's own data; iii) equitable access to data sets; as well as, iv) the right to be free from bias and discrimination.

In order to balance the need for open science with the legitimate concerns for human rights, autonomy, sharing of benefits and freedom from bias and discrimination, we must emphasize the importance of well-governed and articulated data sharing schemes that limit access to sensitive information across stakeholder groups and ensure AI models perform equitably and as intended.

i) Privacy Rights

Since patient data is highly sensitive, privacy concerns are front and center when entering into data sharing agreements involving such data. In this context, privacy-preserving techniques, which encrypt sensitive patient data in ways that still allow machine learning models to use that data to train, is critically important. These techniques make it possible for organizations with large health datasets to share their data while maintaining the privacy rights of their patients.

Sharing sensitive patient data in rights-respecting ways has been a primary focus area for the UK Biobank (UKBB)[49]. The UKBB is a large-scale database with in-depth genetic and health information on half a million UK participants. This database has been made globally accessible to approved researchers who are studying life-threatening diseases. In order to maintain the privacy of each individual within their database, the UKBB allows selective access to the data and leverages encryption techniques that protect their patients' privacy.

More generally, there are a number of privacy-preserving techniques that are worth outlining. These privacy-preserving methods allow stakeholders to train their machine learning models on sensitive data without exposing any data on individual patients. Research in this field is ongoing and could have a significant impact on the ease of data sharing in the future.

1. Federated Learning & Differential Privacy

Federated learning[50] is a machine learning technique that allows many data owners (i.e. hospitals) to leverage a distributed collection of data when training their machine learning models. During this process, the data used to train AI models remains decentralized which, when combined with techniques like differential privacy, retains a high degree of confidentiality. The limitation of this method is that it requires the explicit coordination of a set number of stakeholders and does not enable data sharing more broadly.

2. Lightweight Encoding Techniques

48 Lewis (2011) Confidence-building and international agreement in cybersecurity, CyberNorms

49 Project description here.

50 Kairouz et. al. (2021), "Advances and Open Problems in Federated Learning", Foundations and Trends® in Machine Learning

Lightweight encoding techniques involve distorting images to the extent that humans are unable to identify their contents. However, these images can still be used to train machine learning models. The benefit of this approach is that it allows multiple data owners to work collaboratively on one dataset without risking patient privacy. However, this technique is not yet sufficiently robust against adversarial attacks[51].

3. Cryptographic Methods

Cryptographic methods[52] (including secure multi-party computation, fully homomorphic encryption and functional encryption) involve one or more data owners (i.e. hospitals) encrypting their data before sharing it with a third party. The benefit of this approach is that it generates a highly secure dataset against adversarial attack. However, this method can entail high computational overhead cost. Therefore, this method is recommended in the context of sufficient resourcing as well as a need to maintain confidentiality over every facet of the dataset.

ii) Autonomy Over One's Own Data

Whenever patient data is being collected, patients must be made aware of, and consent to, the ways in which their data is being/ will be used[53]. Specifically, patients should understand who owns or is responsible for that data whenever it is being leveraged by a stakeholder group. However, these data sharing policies should adhere to privacy protections guaranteed by local laws and regulations as well as being respectful of local customs and traditions.

A promising solution to this problem is that of data trusts[54]. A data trust is a legal agreement that involves entrusting one's data, and the rights thereof, with a designated data owner. The data owner, or data steward, has a fiduciary responsibility to represent the interests of those within the trust to those outside the trust. The data trusts provide individuals with greater transparency into how their data is being used and allows them to accrue potential benefits from that use as well, through the collective strength of a large number of individuals whose data are pooled by the data trust.

iii) Equitable Access to Datasets

In order for data sharing schemes to be just and sustainable, it is critical that all parties involved share in the benefits of partaking. Although the value that each party derives does not need to be identical substantively, they should all receive fair benefits from engagement. When there is an unequal power distribution among the stakeholders involved, it is important for long term collaboration that all parties act in good faith.

iv) Freedom from Bias and Discrimination

Data sets that contain a disproportionate amount of data on one community to the detriment of another can train machine learning models whose performance may be biased against a particular demographic group. Not to mention, these discriminatory models can perpetuate an uneven[55] quality of care to patients across the population.

In order to ensure the datasets being leveraged are robust, with a diverse group of demographics represented, machine learning developers must understand the strengths and limitations of the datasets they're using. Current dataset labelling techniques include Model Cards[56], Factsheets for AI Services[57], and Datasheets for Datasets[58].

51 Yala et al, (2021 forthcoming), NeuraCrypt: Hiding Private Health Data via Random Neural Networks for Public Training

52 See this review paper of various methods: Sharma et al (2021), A review on various cryptographic techniques & algorithms, Materials Today: Proceedings

53 WHO guidance on Ethics & Governance of Artificial Intelligence for Health (2021)

54 McDonald (2019), Reclaiming Data Trusts Centre for International Governance Innovation

55 Discussed in the article by Sobia Raza (2021), Minding the genomic data gap: COVID-19, genomics and health inequalities, Ada Lovelace Institute

56 Margaret et al (2019), Model Cards for Model Reporting, FACCT Proceedings

57 A further discussion of factsheets is: Mojsilovic (2018) Factsheets for AI Services, IBM Research Blog

58 Microsoft Research has a discussion on the topic.

The benefit of data labelling is that it can expose biases and other limitations that the users will, at least consider and at most mitigate, when building their models.

Additionally, evaluation processes for AI tools should be standardized so as to ensure that datasets are being effectively compared to one another, particularly along the dimension of demographic representativeness. Ensuring appropriate demographic representation in the dataset is necessary in order to prevent AI models from being biased in their treatment of different groups.

v) Preventing the dissemination of dangerous biological molecules

The ability to synthesize novel biological molecules, such as genetic sequences, is rapidly becoming more powerful and less expensive[59]. As the ability to generate custom molecules on demand rises, so will the risk in the dissemination of the make-up of dangerous molecules such as infectious microbes and deadly viruses. Of particular note is the practise of so-called “Gain of Function” research, where-in the particular mutations required to cause a virus to gain some adverse ability are considered. These lines of investigation are well-intentioned, and have the potential to increase our ability to combat threats of the future, the dangers to these sorts of sequence becoming public knowledge should be clear[60]. While the extent to which these controversial research directions should be pursued remains the subject of debate, it should be clear that the outputs of such work do not fall under the scope of this committee’s call to open science and open data practises. Careful consideration should be made to the growing present and future abilities of synthetic biology before making public biological knowledge that could be a threat to public safety in the hands of malicious actors.

Successful Usage Scenario

The goal of this document is to influence the drug development process to become cheaper, more efficient and include more stakeholders in order to generate drugs that are more responsive to healthcare needs. The impact of this process would be most evident in cases where public health is not well served by the current system. Thus, this recommendation would influence: i) the types of diseases that drug developers are attempting to cure and the accessibility of those treatments (e.g. neglected diseases); and ii) the success, speed and feasibility of drug development (e.g. antimicrobials). Both scenarios will be explored in greater detail below.

Investment in Once Neglected Diseases

A disease is circulating locally in the Tropics, transmitted by a small insect that is unable to survive in cooler climates further north. As a result of this disease, several millions of people's lives are put at risk. However, the overall market remains too small to attract the interest of global pharmaceutical giants.

In an effort to create a vaccine, a healthcare startup in Guinea, called Rose, conducts preliminary research to determine whether they can afford to undertake this work. Luckily, they found a group of researchers in England that had been provided funding to make their research on drug candidates publicly available. Rose decides to run their AI models on top of this dataset to screen for candidate drugs that may be effective at treating the disease. Due to the high social value of this endeavour, Rose is able to secure a government grant to put candidate drugs through the clinical trial process, which leads them to successfully developing a treatment.

Development of Drugs to Address Antibiotic Resistant Microbes

A new strain of an infectious microbe is discovered in Cambodia, making the bacteria resistant to treatment from all commonly available antibiotics. An international organization, which is responsible for monitoring potential outbreaks, flagged and notified the international community of this threat. Thankfully, AI algorithms had been deployed in India to enable the rapid development of modified antibiotics that were unaffected by the novel strain’s resistance capabilities. Furthermore, AI algorithms were used to predict the likelihood of success of each

59 Reports and Data (2021), Synthetic Biology Market By Products

60 Marc Lipsitch (2018), Why Do Exceptionally Dangerous Gain-of-Function Experiments in Influenza?, Methods Mol Biol

candidate drug among people, given their specific health profile, which allowed the clinical trials process to be sped up considerably. This AI-powered drug development approach shed years off the drug development process and resulted in an effective, safe, novel vaccine being sent to Cambodia with sufficient speed to prevent any international spread.