Roundtable Discussion on Understanding Artificial Intelligence Applications in Bioscience and Biotechnology:

How can the U.S. government support innovation and leverage advancements in biology and artificial intelligence while serving the public good?

Background - This paper reports on a November 15, 2018 Roundtable that explored opportunities and obstacles associated with applications of machine learning, deep learning, neural networks and other forms of "artificial intelligence" to bioscience, biotechnology and their application to biomedicine. The Roundtable was convened by IQT Labs, the research venture of In-Q-Tel (IQT), in collaboration with Lawrence Livermore National Laboratory (LLNL). Roundtable participants included multidisciplinary experts from industry, academia, finance and several U.S. government agencies. The discussion took place over a single day, included invited presentations from three participants, and was held on a not-for-attribution basis.

Introduction

Artificial Intelligence (AI) – a term we will use to reference advanced analytical techniques performed on "big data" (very large datasets) – encompasses a number of specific statistical techniques. These include machine learning, deep learning, neural networks, computer vision, and other types of mathematical approaches. AI has been successfully applied to a large range of fields and problems, but applications to biological research and biomedicine are relatively recent and remain largely at the research stage. Nonetheless, there is strong enthusiasm for AI's potential to greatly improve, or even transform, biological research, drug discovery, and disease management.

LLNL, a global leader in applied computational science, has partnered with universities and large pharmaceutical firms in efforts to probe biological mechanisms and advance drug discovery using AI in conjunction with mechanistic simulation. IQT has a long history of working with companies applying analytical techniques to a range of commercial and national security problems.
This roundtable was motivated by three core beliefs:

(1) AI will be a game-changing force in advancing the life sciences and the exceptionally complex field of biomedicine;
(2) excellence in AI techniques and diverse applications is important to U.S. national security and economic competitiveness; and
(3) achieving such excellence in the U.S. requires a deliberate, strategic, national approach.

This roundtable was a collaborative effort to identify actions that might be taken by government, the private sector, and academia to support and accelerate biological applications of AI techniques for the public good. We were interested in framing the opportunities, risks, barriers and technical challenges associated with AI applications in bioscience, biotechnology and biomedicine. Future Roundtables on additional AI+bio topics, including inquiries into technical problems and needed hardware, are anticipated.

Summary of Discussion

Discussion was organized around three presentations:
[1] *Strengths and weakness of AI*: presented by Dr. Casey Greene, Assistant Professor of Systems Pharmacology and Translational Therapeutics, U. Penn
[2] *Discovering the Drugs of Tomorrow*: presented by Dr. John Baldoni, Senior Vice President, DPU Head, ln Silico Discovery, GSK *(retired)*
[3] *Characterization of engineered DNA*: presented by Dr. Alec Nielsen, Founder and CEO, Asimov

For each discussion session, participants were asked to consider the following questions:

- How do we make best use of AI to advance capabilities in bioscience, biotech, biomed?
- What types of biological problems or applications are best suited to AI techniques?
- What are the "low-hanging fruit" versus "hard problems" in AI/bio applications?
- What are the key enabling factors and impediments associated with AI/bio applications?
- What actions or developments in the next 5 years to unlock the power of AI in biology?
- What ethical or public perception issues loom large?
- What would be the ideal roles of the US government, industry and research communities in developing AI/bio for the public good?

While these questions covered a range of topics in AI, most of our discussion focused on the need for robust biological datasets to fuel various biomedical applications of AI. Securing access to such high-quality datasets is an important early step to enabling many potential applications of AI to the biological sciences and was the predominant discussion topic at our roundtable.

Key Findings and Conclusions

We list here our key findings and discuss them in the sections below.

1) Because of the necessity of large, well-labeled datasets for the effective use of AI, biological data should be considered a "strategic national asset".
2) The distributed nature of biological research and healthcare in the U.S. make it challenging to amass, collate and share large biological datasets; but solutions may be on the horizon.
3) A national effort could encompass a range of opportunities, such as the creation or dedication of a national laboratory to harness and use large biomedical datasets; a new business model for healthcare laboratories; or requirements to immediately release data during a public health emergency
4) Finding, standardizing and harnessing the large amount of unused / underused biological and healthcare data ("biological data dust") in U.S. institutions, and the utilization of data simulations, may augment the ability to more fully explore AI applications in biomedicine.
5) China's pursuit of global leadership in AI presents a challenge to US economic and scientific competitiveness, particularly in the life sciences.

. . .

Biological data are a "strategic national asset": Very large, well-labeled datasets are essential to applications of AI techniques. The potential power of AI to transform our understanding of biology and biomedicine is such that biological data should be considered a strategic national asset. The availability of and accessibility to such "big data" will have profound implications for US economic competitiveness and national security, as well as for the future of health care, drug discovery and the management of epidemics.

Although the discussions throughout the day ranged widely and were rich with specific observations, insights and suggestions, the challenges surrounding the imperative for large volumes of labeled, well-curated biological data (e.g. clinical and genomic data) was a persistent topic.

The compelling need to treat data as a national asset has only recently been recognized by some in government, the private sector and academia. Progress in basic bioscience and in biomedical development would be significantly aided and accelerated by access to well curated "big data" sets.

The heterogeneity of the US biomed research enterprise and the fragmented and variable nature of the country's health care institutions and electronic health records were recognized as challenges to efforts to organize and treat "data as a strategic asset". One participant thought it too late to establish a national health database, given the finance-driven nature of most American electronic health records. But the diversity and biological heterogeneity of the American population were recognized as extremely valuable assets.

Sharing data remains critical, but solutions may be on the horizon: Successful development and implementation of AI tools for biomedical and bioscience applications requires a re-think of how to approach sharing of biological data; and new tools are on the horizon that can make this possible. Additionally, successful implementation of new ways of aggregating and sharing data – both biological data and personal and population health data – will require re-thinking who "owns" such data and what rights, privileges and obligations "ownership" entails.

Concerns about patient privacy must be addressed when assembling biomedical data collections, particularly when the data includes personal clinical information. The Health Insurance Portability and Accountability Act of 1996 (HIPAA), mandates stringent privacy and security rules that govern the way hospitals and health care providers handle – and share -patient medical records. Violations carry stiff penalties. The Privacy Rule has been controversial since its inception, largely because of paperwork requirements and implementations costs, but studies have shown that HIPAA resulted in less medical research of some types being done and higher research costs. The question is how to protect privacy while promoting medical research.

There is clear evidence that individuals are willing to share their personal medical data for some purposes. "Opt-in" models of data sharing, such as those in which individuals volunteer their own personal and clinical data to assist research into a specific disease, often one from which they are suffering, are becoming increasingly common. Microsoft, for example, has helped to support an open dataset that contains medical and personal data provided by people diagnosed with

Amyotrophic Lateral Sclerosis (AML), in efforts to use big data to better understand and treat the disease.

On the other hand, an initiative by the Health and Human Service's Agency for Public Health Preparedness and Response, which sought to assemble a national database of available hospital beds for use during public health emergencies failed, because individual hospitals were unwilling to share what they regarded as proprietary information. In general, health care systems regard their own data as a competitive advantage and are not inclined to share it. Companies such as Nebula Genomics and LunaDNA[1] may signal an early trend towards attempts to encourage data sharing by empowering individuals to share their own medical data only under conditions of strict privacy controls and for purposes of their choosing.

A National effort could encompass a range of opportunities: Many in the group thought that U.S. government could do more to address the need for scientific access to "big data" – specifically, enable collection, curation, and access to large biomedical datasets to serve the public good. Several suggestions for how to accomplish this were raised. The group favored approaches that "aligned incentives" rather than the imposition of regulations, although regulations were not ruled out, especially if public health were imperiled. Regulations compelling the sharing of genomic and clinical data during epidemics might be considered, for example[2]. Regulatory standards that made data sources more useful might also require legislation.

The U.S. Food and Drug Administration (FDA) could consider awarding priority review vouchers to drug companies who made their data (including negative data) available. The FDA's existing Priority Review voucher programs allow companies to receive expedited regulatory review of a drug which would meet certain specified criteria. Once FDA awards a voucher, the company can use it to expedite its next drug or can sell it to another company. Since time is literally money in the drug development business, vouchers are quite valuable and have successfully spurred drug development in areas where market forces alone did not. Similarly, the National Institutes of Health (NIH) might find ways to advantage grant applicants who agree to label and share their experimental data and who developed a track record for doing so.

The use of AI-powered image-analysis of healthy and diseased cells, and imaging of multi-cell environments, is beginning to prove itself a powerful tool in understanding molecular pathology. Some biotech companies have been formed around this technology and large pharmaceutical firms are using such approaches to investigate the nature of diseased versus healthy cells and tissues, and to spur drug development. Participants argued that these cell visualization methods – and many other AI-driven research efforts – should be "industrialized". That is, methods and outcomes should be held to agreed upon standards so that results could be widely compared and understood, thereby producing more rapid and useful insights and progress.

[1] Nebula Genomics (https://nebula.org) and LunaDNA (https://lunadna.com) are exploring the use of blockchain, homomorphic encryption, and other technologies to enable and incentivize a personal genomics economy.
[2] https://www.who.int/blueprint/what/norms-standards/gsdsharing/en/

The group wanted an entity with "authority" to organize the current confederation of individual companies, universities and researchers who generate and use such data. Some opined that the federal government should use its convening power for this purpose. It was also suggested that appealing to state governments and governors to establish pilot projects and demonstrations might be a more efficient and faster route to success.

Some additional ideas were discussed:

National Laboratory for AI driven bioscience -- A participant from the pharmaceutical industry suggested that the country dedicate or create a national laboratory to the mission of creating, curating, and maintaining high quality data repositories of biomedical data, and to developing high-fidelity data platforms and efficient means of making data publicly available.

Repurpose clinical laboratories -- It was noted that large commercial clinical laboratories (of which there are only a handful), that currently "report out" clinical lab data to hospitals and physicians, might consider an alternate business model. They could take advantage of the huge datasets generated by their lab services by offering "consultation services" to health care providers and researchers.

Public Health -- The data available during epidemics – epidemiological data and, in the case of an emergent or rare disease, biological data also – would be "streaming". Such data would become available and subject to analysis in batches as it was collected in real time. This poses particular analytical challenges, especially when operational decisions must rest on limited datasets and prior knowledge. Although AI applications of epidemiological data are conceivable, no epidemic has been subjected to such techniques in real-time due to the scant amount of data available. Establishing and invoking a regulation that requires immediate release of data during epidemics may facilitate further development and implementation of analytic methods that would be useful for managing an epidemic.

Different types and scales of data may prove important in AI analyses: Artificial intelligence methods, especially when investigating complex fields such as biology or human health, not only require large volumes of data, but *different types* of data. Any very large dataset will contain biases. A health-related dataset may include mostly males, or exclude people of color or children, for example. Genomic datasets are well known to be skewed towards Caucasians living in the Western hemisphere. When applying AI techniques to large data collections, the biases inherent in the data collection always outweigh the statistical uncertainty. Thus, the analytical results may be correct for that dataset but may not be widely generalizable. This problem can be mitigated to some extent by including different kinds of datasets. *The more datasets you have, the less important are the biases in a specific dataset.* But this brings us back to the imperative for large data collections – and collections of different types of data.

Simulations of data may be a useful approach when large and diverse datasets are not available:

One example of this is ATOM: Accelerated Therapies and Opportunities in Medicine; a collaborative enterprise among the pharmaceutical company Glaxo-Smith Kline (GSK), the Lawrence Livermore National Laboratory, and researchers from the National Institutes of Health and the University of California at San Francisco Medical Center. Established in March 2018, ATOM is a pre-competitive, open source, integrated framework of AI tools and capabilities to explore the use of AI analytics in the context of drug discovery and development efforts. ATOM is not trying to fit AI to the existing drug discovery pipeline paradigm but rather build a new value chain that is inspired by the unique needs and ways AI can learn. Using a Moonshot challenge, ATOM is building milestones, cultures, and organizations to pilot the future AI biopharmaceutical concept.

ATOM's database includes information GSK developed from its historical investigations into over two million compounds, in addition to other available biological data. Simulations are employed as "surrogate data" to decrease the noise in the biological data and to help close the gap between available data and reality. Machine learning is used to interpolate between available data and the simulation. The collaboration is in early stages, but already results have been used to drive new experiments and generate new data which are integrated into the models. If this proves to be reliably accomplished, it could greatly increase the pace and reduce costs of drug development, and increase the probability that clinical trials of a candidate drug would be successful. ATOM is demonstrating the importance of thinking anew, which will be important to fully appreciate AI as a new tool for the next generation of medicines.

Getting "digital dust" to work for biological problems:

IQT Labs Reflections -- Elements of AI Innovation, beyond data:

You can't do analytics without data. However, it is important to consider all of the many other components important to developing, testing and implementing AI, which include, but are not limited to:

Algorithms. Substantial ongoing research is focused on the development of novel algorithms. Of particular interest are methods that could be used on sparse data, such as drug development for rare diseases. A machine learning method called model transfers may offer a useful approach, like the algorithm called MultiPLIER. The model was trained from a large, public data compendia before being applied to the small, target dataset of interest. This approach resulted in models that were more effective at aligning biological processes to disease outcome than training on smaller, target specific datasets.

Explainability. Machine learning is mostly empirical today, where only benchmarks are used evaluate the performance and merit of models especially in computer vision. However, for applications that require rigor around decision-making, a theoretical approach (a.k.a. the how and why behind the model) may prove critical. A theory around AI would inform new research directions better than the throw everything against the wall approach that is common today.

Infrastructure. Artificial intelligence is a broad field that can take advantage of many kinds of computer hardware. But neural networks (a common form of deep learning) are almost exclusively trained on specialized hardware; most commonly a Graphics Processor Unit (GPU). The biggest vendor of GPUs is Nvidia Corporation, who holds an effective monopoly of which they have exploited to great financial effect. Other companies have tried to take Nvidia's market from them. However, Nvidia is vigorously competitive and nobody seems able to keep up.

AI Talent Pipeline. A single person with expertise in both domain expertise and AI expertise is incredibly rare, yet essential for developing accurate and effective analytic models. Therefore, teams are almost always an essential part to AI research and development, which can present challenges due to domain specific language for both AI and bio. It is critical that efforts be taken now to educate up-and-coming data scientists in biology, or other domain-specific terminology, to ensure cross-discipline talents are developed for the future.

One participant noted that the digital ecosystem of big tech companies, most notably the "FAAMG" companies (Facebook, Apple, Amazon, Microsoft, and Google), is largely powered by the "digital dust" created by smartphone users and online transactions, such as location info and web browsing data. This "digital dust" is a byproduct that users and consumers have, at least until recently, been mostly willing to allow companies to access. AI applications to biology might make more effective use of analogous "biological dust," such as patterns of signals across multiple experiments, the 'noise' in 'omic' experiments, and data found in lab notebooks (especially if digitized). Often, this data is discarded or discounted as unimportant. However, these data might reveal signals when taken collectively or in concert with other data; and might be easier to share or access than direct measures of markers or genomic information. Accessing medical health records may represent the richest trove of "biological dust," but privacy concerns, regulations and a lack of a unified electronic health records system in the U.S. present major obstacles.

China's pursuit of leadership role in AI: The group's discussions about the importance of access to large datasets frequently referred to China's activities in promoting AI techniques in biomedicine and in other spheres. China is clearly pursuing an aggressive strategy to develop its AI prowess. China enjoys a significant advantage in amassing big data, because of its huge population, the opportunity to mine available digital data including surveillance data based on ubiquitous CCTV cameras, bicycle share accounts, cell phone records, etc., and because its authoritarian approach obviates the need for delicate negotiations about privacy, proprietary rights or individual interests.

The Chinese government has made clear that it intends to make Chinese pre-eminence in Artificial Intelligence a top national priority. Considerable effort is being exerted to acquire talent, leverage the capabilities of China's large internet companies (Tencent, Baidu, etc.), and publish original research on AI applications and theory. In spite of all this, some participants with experience investing in Chinese companies viewed China's ability to innovate as less developed than US capabilities. It was also noted that China faces tremendous challenges in meeting the health care needs and expectations of its population. China's population is aging, and suffers the highest cancer incidence in the world, yet has access to only four of the latest biological cancer drugs. Hospitals and physicians are in short supply; waits in emergency rooms can last hours or even days. Using AI methods to drive medical innovation and to modernize China's health care capabilities makes sense, whatever other motives are at work.

These efforts notwithstanding, the structural advantages that China possesses in access to population-wide health data, including genomics, will present an enduring challenge to U.S. competitiveness in applying AI to biomedical problems.

. . .

What’s Next?

The discussants clearly articulated the imperative for well-labelled, large, and diverse datasets of biologically-relevant data. Significant challenges, but also opportunities for solutions, were discussed. While the US will never mimic the authoritarian approach taken by China, the US government should consider how it can use its convening power and incentive structures to establish coordinated efforts across academia, industry, and government. Opportunities to enable data sharing, such as those discussed here, should be further explored with representatives from relevant stakeholders. This may be a topic for a future IQT Labs roundtable. Additionally, more focused efforts need to evaluate the challenges and opportunities pertaining to algorithms, explainability, infrastructure, and talent.