Title: AI for Biology

Authors: Soha Hassoun, Xinghua Mindy Shi, Felicia Jefferson, Brian Stucky, Epaminondas Rosa, Jin Wang

Author emails:
soha@cs.tufts.edu
mindyshi@temple.edu
jeffersonf@fvsu.edu
stuckyb@flmnh.ufl.edu
erosa@ilstu.edu
Jin-Wang02@utc.edu

Summary

Despite best efforts to integrate research across different subdisciplines of biology, the scale of integration remains limited. We hypothesize that future generations of Artificial Intelligence (AI) technologies that are specifically adapted for biological sciences will enable the reintegration of biology. These technologies will allow us to collect, connect and analyze data at unprecedented scales. They will allow building comprehensive predictive models that span various subdisciplines and allow for both targeted (testing specific hypotheses) and untargeted discoveries. AI for biology will be the ***cross-cutting technology*** that will impact our ability to do biological research at every scale. We expect the impact of AI on biology to revolutionize biology in the same way that statistics transformed biology. There are many challenges, including data curation and assembly, development of new science in the form of theories that connect the subdisciplines, and new predictive and interpretable AI models that are more suited to biology than existing machine learning and AI techniques. Development efforts will require strong collaborations between biological and computational scientists. This white paper provides a vision for AI for Biology and highlights some challenges.

1. What's the big question? What's the exciting science?

Concern about the fragmentation of biology into specialized subdisciplines, and calls for its reintegration, have been appearing in the scientific literature for years. So far, though, grand reunification remains elusive. Human intellectual limits in collecting and integrating data and testing hypotheses spanning multiple subdisciplines are the primary and most important reasons biology became fragmented in the first place, and reintegration will be impossible without overcoming those limits. Stated differently, key biological systems and related information, at all levels of biological organization, are simply too complex for humans to understand with sufficient depth to permit generalized, human-driven reintegration. Here, we make the case that advances in AI methods and technologies will provide our best hope for

overcoming the human cognitive limitations that have splintered biology into ever-more-specialized subdisciplines.

Our vision for reintegrating biology recognizes the enormous potential of existing AI techniques to accelerate biological research, but it also extends **far beyond current AI techniques**. Current Machine Learning (ML) and AI methods are already having an impact on biology (discussed in more detail below), but their application is largely limited to relatively narrow data integration and processing tasks. We envision new suites of AI tools, developed for biological inquiry and perhaps even inspired by biological systems, powering biological investigation at unprecedented scales.

2. What's the potential impact?

The development of statistics and electronic computers transformed 20th-century biology, and we foresee AI having a similarly transformative impact on 21st-century biology. AI-driven reintegration of biological disciplines will establish a new kind of biology that will allow us to answer deep biological questions that span multiple time and biological scales in ways that are impossible today. Such questions will cut across biological subdisciplines and integrate across the scales of biological inquiry (spatial, temporal, and organizational). We offer some examples as illustrations, arranged in order of increasing difficulty of implementation.

Example 1: Biological knowledge discovery and assembly. Have you spent hours searching for papers relevant to a specific protein or organism on google scholar or pubmed? Have you spent a considerable amount of time reading and highlighting relevant information? Have you then spent a huge amount of time sifting through various databases to examine and perhaps gather appropriate relevant information to assemble a dataset or discover relationships to other proteins or organisms? Imagine what AI can do for you: you ask your biological question, and you get back all known information about your question, organized and visualized in a coherent and potentially *customizable* way, while also highlighting and imputing missing information.

Comprehensive data collection and analysis at this scale is nearly impossible today. AI-driven data generation and integration across the spectrum of data modalities and sources would make this possible. AI will utilize a variety of known and new techniques to collect and assemble these data: using text mining, semantic analysis, and missing link prediction in existing multilevel and hierarchical knowledge graphs. Put simply we need a search engine on steroids, capable of unearthing known and predicted biological knowledge.

Example 2: Behavioral ecology. Suppose that, for some species of bird, we would like to understand the relationship between individual fitness and environment, including the birds' social environment. Ideally, this task would draw upon data from a wide range of biological and spatial scales (e.g., vocalizations and communication, social networks, movement, morphometrics, parasite loads, genetics, biomarkers, etc.) and sources (e.g. imaging, videos, audio recordings, tracking tags, DNA sequencers, etc.). Currently, such analysis is usually done

using one or a few data modalities with relatively small numbers of individuals (e.g. using RFID tags to collect movements and social network analysis to understand social behaviors of birds). Simultaneous advances in AI and automated data collection will make it possible to answer these questions using a holistic approach that goes far beyond current capabilities, which will allow us to answer ever more complicated biological questions (e.g., How does genetics affect social behaviors that in turn affect collective behaviors like migration?).

Example 3: Genes to phenotypes. Predicting an organism's phenotype is extraordinarily difficult because it requires integrating processes and information across multiple scales of biological organization, from molecules to an organism's environment. The general solutions to this problem are beyond the grasp of today's AI technologies, but future advances in machine reasoning, learning, and causal inference, combined with continual growth in data, collection, and computational capacity, will help transform our understanding of how phenotypes emerge. Specifically, these technologies will allow us to use heterogeneous data (e.g., DNA sequence data, phylogenetic information, environmental data) and knowledge (e.g., gene function, results of prior experiments) to elucidate and test hypotheses about the inputs that shape phenotypes. For instance, we could investigate how data collected over diverse labs and fields (e.g. imaging of cells, genomics, epigenomics, proteomics, metabolomics, metagenomics in soils) can predict the cellular decision making or phenotypic changes that affect productivity of crops like corn.

This new, AI-augmented biology will generate tools, methods, and knowledge that will translate to a host of biology-adjacent disciplines, such as bioengineering, biophysics, biochemistry and medicine. In addition, we anticipate that new AI tools, in concert with open data, will help democratize participation in biology, allowing researchers at more resource-limited institutions to participate in cutting-edge biological research.

3. Why Now

The time for AI in biology has arrived. There are now sensors, Internet of Things (IoT), and environmental monitors that allow the collection of data at unprecedented scales. Large, heterogeneous datasets at the confluence of multiple information streams are rapidly growing in size. We now have multivaried data across time, space, and biological scales that needs to be analyzed in an integrated manner to discover system-wide, multiscale phenomena that can lead us to understand fundamental rules of life and their application to other systems.

The AI infrastructure to support these efforts is beginning to emerge. There are now outstanding computational capabilities in the form of storage, CPU/GPU computing, and large-scale distributed computing which, combined with the increasing availability of software tools for AI, is enabling the rapid exploration and development of novel techniques and applications. These resources continue to grow and will enable the next generation of AI for the most complex problems in biology.

4. State-of-the-art technologies and applications

Although machine learning (ML) has recently entered the popular lexicon and is often conflated with AI in general, AI is a broad field with a long history, and it provides a diverse set of tools and approaches that encompass much more than ML. A variety of these tools have already been used to help solve some biological problems. For example, methods from symbolic AI have been used to develop sophisticated software pipelines for integrating highly heterogeneous sources of information about plant development and to help elucidate possible links between gene function and phenotype. Statistical learning, and "deep learning" in particular, have recently found application in the automated analysis of biological imagery at various scales (e.g., field photographs of plants, radiological images, and satellite imagery), bioacoustic data, genomic analyses, and classifying protein function based on their amino acid sequences.

5. Barriers

Many critical barriers need to be addressed to enable the next generation of AI for biology.

Data: Data are critical in all aspects of data-driven hypothesis generation, validation, and unstructured biological discoveries toward understanding the Rules of Life. New technologies need to be developed for the automatic collection of biological data with varied data modalities (e.g. images, videos, molecular profiles) and comprehensive measurements of biological systems at various biological, spatial and temporal scales. Furthermore, data quality is a concern with large, noisy datasets, so data scientists must work with biologists to ensure the data we generate are as useful as possible. Key challenges include identifying outliers and biases, mitigating known biases, understanding variation, and improving signal-to-noise ratios. To enable the open sharing of data, tools should be developed to allow for transparent data sharing, with consideration of provenance, security, privacy, and fairness. Other researchers can use these shared data to form new hypotheses and build new theory.

Theory: Development of theory from multiple disciplines will enable the development of new AI technologies for biology. For example, theory in biology, chemistry, physics, and social sciences, can be utilized to develop more appropriate AI models for understanding biological systems. Mathematical and statistical theory should be developed to not only design new AI methods but also further our understanding of current and emerging AI technologies. Novel development and incorporation of evolving and updated theory will be conducted in a feedback loop, with AI data analysis and evaluation leading to the development of improved methods.

Model: Novel AI models need to be developed that are bio-meaningful, bio-inspired, and bio-integrated at scale. AI models should incorporate biological hierarchical structures and feedbacks/loops. Although blackbox models are appropriate for some type of modeling tasks, AI models that are interpretable, explainable, and visualizable should be encouraged. AI models must be able to learn data representations and provide meaningful integration. AI models

should be able to perform life-long and reinforced learning using small samples. AI models should be robust and resilient, allowing for redundancy and plasticity. AI models should enable unsupervised learning or semi-supervised learning when labelled data are missing, limited or insufficient.

Computing Infrastructure: Current computing storage and throughput will be challenged by the amount and scale of future biological data. Accordingly, storage and performance of computing systems must scale. Traditional computing models (Von Neumann architectures) may not be well suited for biological tasks. Emerging technologies such as quantum and neuromorphic computing might provide appropriate alternatives. Focusing AI on biology will open up novel opportunities for developing hardware, software, and new computing mediums that are more appropriate for biological applications. There are also exciting opportunities to explore novel computing-biological interfaces at the intersection of biology and computing.

Ethics: Central to this effort is making progress in an ethical fashion, ensuring fairness, access, equity, diversity, governance, privacy and security of data at all development stages. There are already well-known cases of biases in ML data and algorithms, which can then be exacerbated as data and models become more complicated. These ethical issues must be addressed head on as a first-class concern. Developers and users need to be trained to be aware of these issues. Further, our workforce must be sufficiently diversified to ensure no one is left behind. Further, we all should be aware of potential misuse of AI to harm humans or the environment and the utmost care must be taken to assess and address these issues.

Training: Training is an integral part of effort that needs to be addressed in a more systematic and cross-institutional/disciplinary manner. A new generation of diverse scientists must be trained at the cross of biology and computer science, starting with undergraduate studies and through graduate and postdoctoral opportunities. Users of AI systems must be trained to interpret the results and use the various tools judiciously. Vocational pathways need to reward cross-disciplinary work.

6. Broader impact

AI for Biology has broad, and potentially long-lasting, impacts on biological science and beyond. AI enables biologists with powerful tools to ask and solve ambitious questions, such as investigating and integrating complex mechanisms across a wide range of scales (from genes, to cells, to organisms, populations, and ecosystems), and developing theoretical machines to understand biological and ecological systems at extremely large scale, all of which would be impossible (or severely limited) without AI. Meanwhile, feedback from biology helps to re-define AI concepts and improve AI computing. In addition, we are witnessing unprecedented applications of AI to bio-related fields such as medicine, public health, and biomedical engineering, as well as many other scientific disciplines such as physics, chemistry, and engineering. Overall, AI is playing an increasingly important role in our world, in re-defining the

theories and methodologies in a variety of disciplines, and in advancing our science and technology.

7. What disciplines might be needed?

We envision that the advancement of AI and its application to biology reintegration will demand multidisciplinary collaboration. In addition to biology and computer science, which constitute the core of this bold effort, many other scientific disciplines will play an important role. For example, mathematical and statistical theory contribute to the underlying principles of AI and can lead to potential revolutions in the methodology; physics, chemistry, and other natural sciences not only strengthen the knowledge base of AI but also enrich its application domains; economics and social sciences can lend to important theories that expand the horizon of AI; ethics and philosophy provide essential guidance in the management of data and in the education and training of next-generation biologists and AI professionals. Additionally, electrical and computer engineering, in the form of sensors, environmental monitoring, IoT devices and their networks, robotics, and AI hardware will be needed to improve AI for biology.

8. Intended audience of the paper

Within the next decade, we expect AI for biology to impact biologists and other scientists at all levels. This document serves to educate graduate students, scientists, engineers, government agents, on the potential impact of AI and why we must act *NOW* to enable this revolution.