Artificial Intelligence for Synthetic Biology: Opportunities and Challenges

DOI:10.1145/3500922

The opportunities and challenges of adapting and applying AI principles to synbio.

BY MOHAMMED ESLAMI, AARON ADLER, RAJMONDA S. CACERES, JOSHUA G. DUNN, NANCY KELLEY-LOUGHNANE, VANESSA A. VARALJAY, AND HECTOR GARCIA MARTIN

Artificial Intelligence for Synthetic Biology

BIOLOGY HAS DRAMATICALLY changed in the last two decades, enabling the effective engineering of biological systems. The genomic revolution,[17] which provided the ability to sequence a cell's genetic code (DNA), is the primary driver of this dramatic change. One of the most recent discoveries and tools enabled by this genomic revolution is the ability to precisely edit DNA in vivo using CRISPR-based tools.[11] Higher-level manifestations of the genetic code, such as the production of proteins, are known as phenotype (as shown in Figure 1 and the accompanying table). The combination of high-throughput phenotypic data with precision DNA editing provides a unique opportunity to link changes in the underlying code to phenotype.

Synthetic biology (synbio) aims to design biological systems to a specification[3] (for example, cells that produce a desired amount of biofuel, or that react in a specific manner to an external stimulus). To this end, synthetic biologists leverage engineering design principles to use the predictability of engineering to control complex biological systems. These engineering principles include standardized genetic parts, and the Design-Build-Test-Learn (DBTL) cycle, iteratively used to achieve a desired outcome. The synbio DBTL cycle adapts the expected four stages to this discipline as follows:

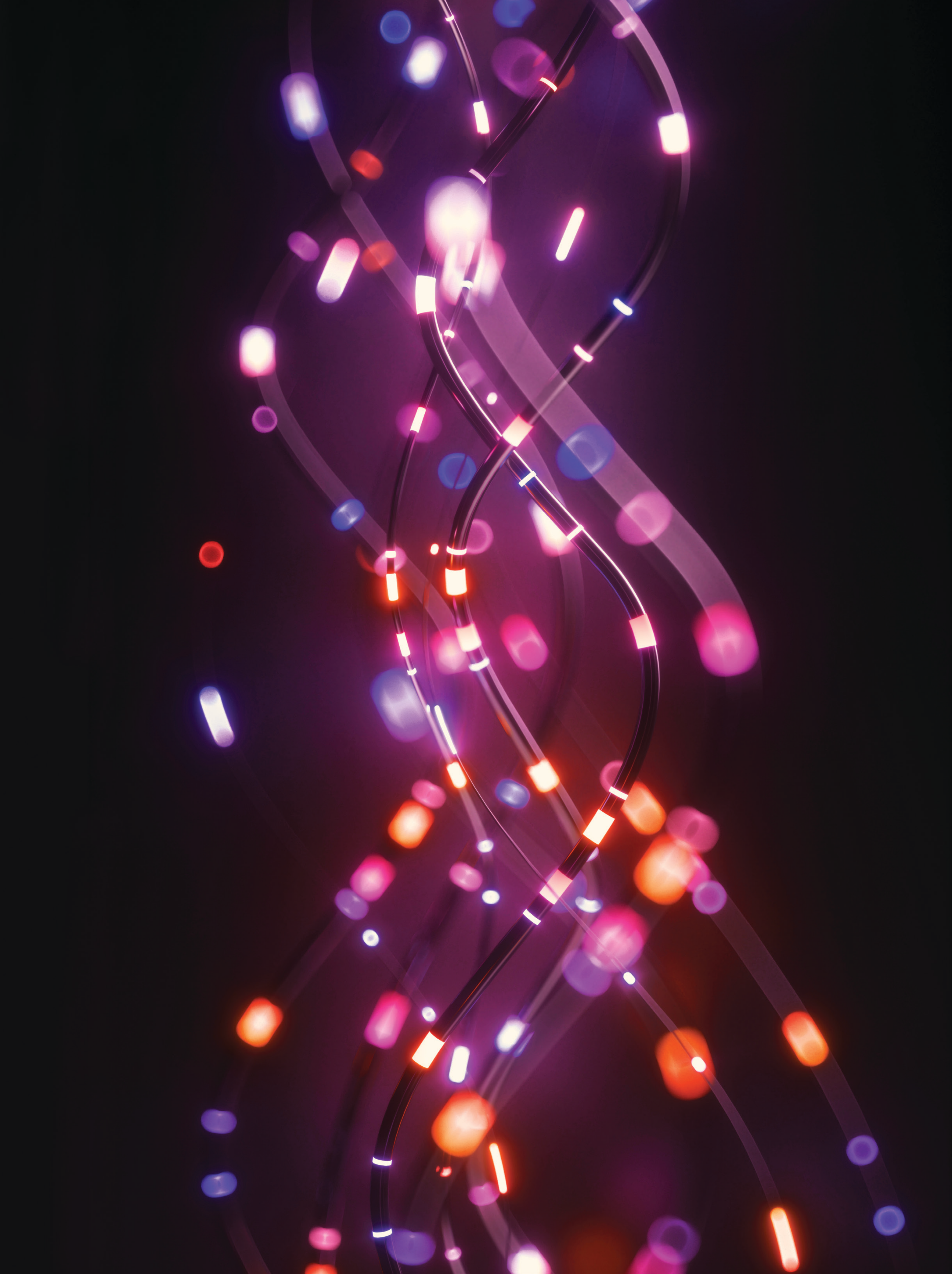
1. *Design:* Hypothesize a DNA sequence or set of cellular manipulations that can achieve a desired design goal.
2. *Build:* Implement the design steps on the biological system. This primarily involves the synthesis of the DNA fragment and its successful transformation into a cell.
3. *Test:* Generate data to check how closely the measured phenotype achieves the desired goal and evaluate the impact of any off-target or unforeseen side effects.
4. *Learn:* Leverage the test data to learn principles that drive the cycle to the desired goals more efficiently than might be accomplished by a random search. This often includes the diagnosis of failures that arise from unforeseen off-target effects. Artificial intelligence (AI) can be used here to inform the next set of designs, thereby reducing the number of DBTL iterations needed to achieve the desired outcome.

More specifically, synbio typically involves manipulations at the genomic

» key insights

- **AI and synbio naturally complement each other and have world-changing applications for the environment, agriculture, medicine, energy, and materials.**
- **AI has begun to make its way into various synbio applications, but major technological (data, models, metrics) and sociological (different cultures) hurdles continue to separate the fields.**
- **The budding interdisciplinary field needs more researchers to fully flourish: we recommend several community-wide, strategic efforts to support interdisciplinary research.**

IMAGE BY WACOMKA



level to push a cell to create specific products or behave in a certain way

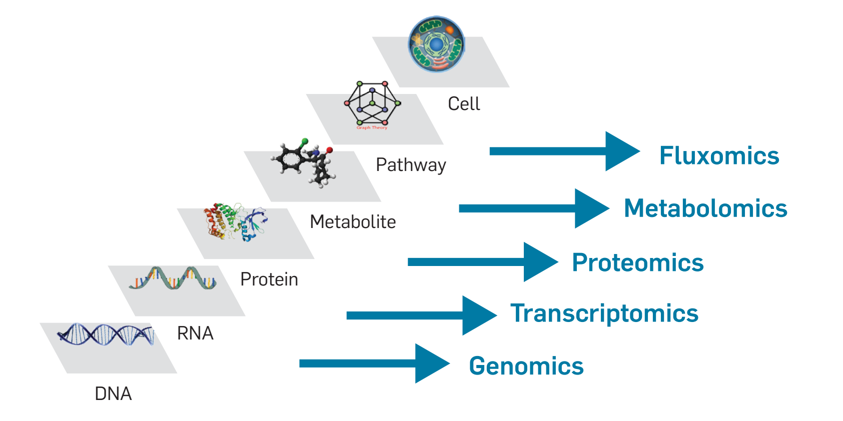
We are a group of AI practitioners looking to adapt and apply principles of AI to synbio in a variety of applications. In this article, we seek to provide other AI practitioners with an overview of the potential of this domain, some initial successes, and the main challenges faced when applying AI technologies to the synbio domain. Our goal is to motivate AI practitioners to address these challenges and promote involvement in a discipline that will significantly impact society in the future. There have been major breakthroughs in AI when large datasets and technology enthusiasts have met. Image and natural language processing are perfect examples of this. We believe biology, and specifically synbio, provides an unparalleled opportunity for breakthroughs in both domains.

The Potential of Synbio

Synbio is primed to have a transformative impact on every activity sector in the world: food, energy, climate, medicine, and materials[29] (see Figure 2). Synbio has already produced insulin without the need to sacrifice pigs for their pancreases (in a previous stage, as genetic engineering), synthetic leather, parkas made of spider silk that have never seen a spider, antimalarial and anticancer drugs, meatless hamburgers that taste like meat, renewable biofuels, hoppy flavored beer produced without hops, the smell of extinct flowers, synthetic human collagen for cosmetic applications, and gene drives to eliminate dengue-bearing mosquitos. Many believe this is just the tip of the iceberg because the ability to engineer living beings provides seemingly unlimited possibilities, and there is a growing level of investment, both public and private, in this field[8] (see Figure 3).

Figure 1. Omics data embody the high-level manifestations of the cell's genetic code (DNA).

This genetic code is transcribed into RNA, which is then translated into proteins (central dogma of biology). Proteins enable a variety of reactions, among them the transformation of metabolites (chemical species) into other metabolites. Several reactions are combined into pathways, which carry on vital metabolic processes for the existence and survival of the cell. Transcriptomics, proteomics, metabolomics, and fluxomics are all examples of omics data.



Datasets/data types frequently used in biology (not a comprehensive list).

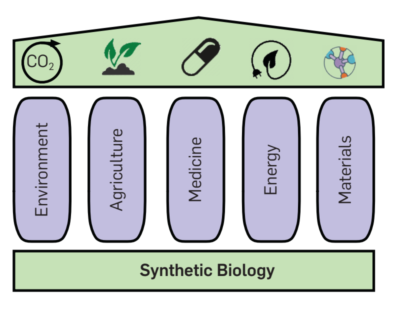
Data Type	Description	Format
Genomics	Underlying code (DNA) that drives cellular processes.	DNA sequences and quality scores associated with each token in the sequence. Often aligned with a reference sequence to highlight any potential changes measured.
Transcriptomics	The amount of transcript (RNA) created from each piece of code.	Transcript sequences and a number that indicates their abundance within the sample.
Proteomics	The amount of decoded product (proteins), the main functional units of life.	Protein sequences and a number that indicates their abundance in the sample.
Metabolomics	Set of chemical species involved in all the reactions in the cell.	Metabolites (small molecules) and a number that indicates their abundance in the sample.
Fluxomics	Set of all metabolic reactions in a cell.	A number that indicates the rate for each metabolic reaction in a cell.

Furthermore, as AI enters a third wave, focusing on incorporating context into models, its potential to impact synbio increases. It is well known that an organism's genotype is not so much a blueprint for a phenotype, but an initial condition in a complex, interconnected, dynamic system. Biologists have spent decades building and curating a large set of properties such as regulation, association, rate of change, and functions, to characterize this complex, dynamical system. Additional resources such as gene networks, known functional associations, protein-protein interactions, protein-metabolite interactions, and knowledge-driven dynamical models for transcription, translation, and interactions provide a rich set of resources to enrich AI models with context. Model explainability is also critical to uncover novel design principles. These models provide biologists an opportunity to answer significantly more complex questions about the biological system and build integrative, explainable models to expedite discovery. The increase in knowledge and resources is clear in the number of synbio publications along with the commercial opportunities in synbio (Figure 3).

AI and Its Current Impact in Synbio

AI has had a limited impact in synbio compared with its potential to influence the synbio field. We have seen successful applications of AI, but they are still limited to a particular dataset and research question. The challenge remains to see how generalizable these approaches are to broader applications, and other datasets. Data mining, statistics, and mechanistic modeling are currently the primary drivers of computational biology and bioinformatics in the field, and the line between them and AI/machine learning (ML) is often blurred. For example, clustering is a data mining technique that identifies patterns and structure in gene expression data, and these patterns can indicate if the engineered modifications lead to a toxic outcome for the cell. These clustering techniques can also serve as unsupervised learning models that find structure in unlabeled datasets. These classical techniques and novel AI/ML approaches in development will have a much-expanded role and impact in the future of synbio as larger datasets become customarily available. Transcriptomics data volume doubles every seven months, and high-throughput workflows for proteomics and metabolomics are becoming increasingly available.[5] Furthermore, the gradual automation,[35] and miniaturization through microfluidics chips[14,24] of laboratory work hints at a future where data processing and analysis are the main productivity multipliers in synbio. DARPA's Synergistic Discovery and Design (SD2, 2018–2021) program was focused on building AI models to address this gap. This is also evident in some companies operating at the state-of-the-art of the field (for example, Amyris, Zymergen, or Ginkgo Bioworks). AI and synbio intersect in a few ways: applying existing AI/ML to existing datasets; generating new datasets (for example, the upcoming NIH Bridge2AI); and creating new AI/ML techniques to apply to new or existing data. Although SD2 did some work in the last category, much work and potential remains.

Figure 2. Synbio can potentially impact every activity sector in the world.



A fundamental challenge in synbio, which AI can help surmount, involves predicting the impact of bioengineering approaches on the host and the environment.[15] Without the ability to predict the bioengineering outcome, synbio's goal of engineering cells to a specification[4] (that is, inverse design) can only be achieved through arduous trial-and-error. AI offers an opportunity to use both publicly available and experimental data to predict the impacts on the host and environment.

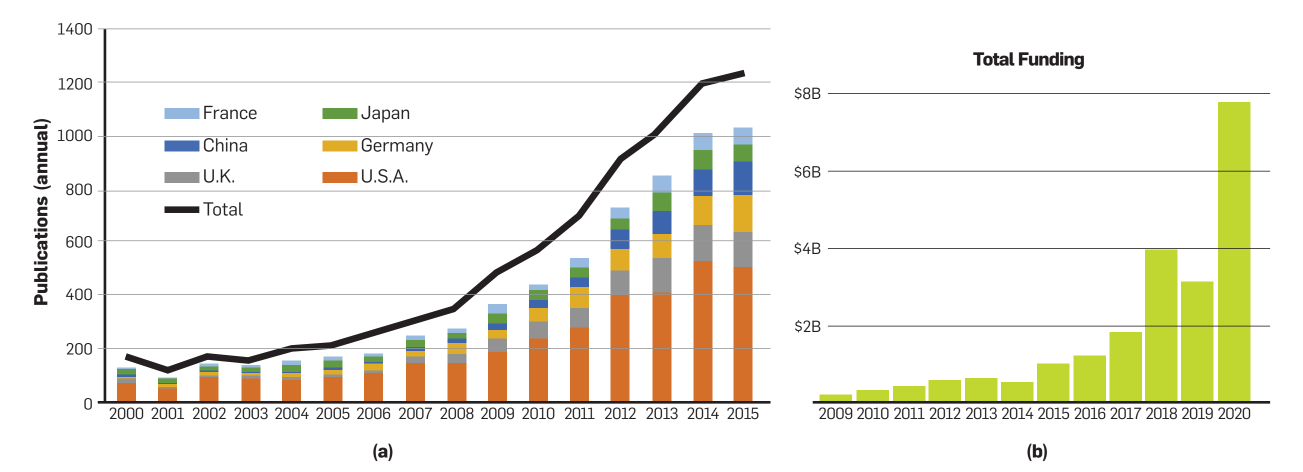
Design of genetic constructs for programming cells. Many synbio efforts have focused on engineering genetic constructs/circuits,[3] which present very different challenges from designing electronic circuits. The genetic constructs are designed to elicit a specific reaction from the cell, much like electronic circuits are designed to provide control of an electronic system. Whereas we can synthesize DNA and transfer it into cells, the global impact of this transfer on the cellular machinery of the dynamic, living organism is not entirely known or currently predictable. Electrical engineers, in contrast, have the tools to "static" design electronic circuit boards to perform a variety of function(s), and not impact the board in a detrimental way. The rules behind the physics and biology of living cells are complex, intertwined, and require significant effort for discovery. In summary:

- **Design on Circuit Boards**
  - Known set of parts to achieve desired circuit output.
  - Impact of printed circuit board on gates/circuit and vice versa are negligible.
  - Qualitative and quantitative models of parts and circuit board exist to predict circuit performance robustly.
- **Design on Living Cells**
  - Genetic constructs are designed to achieve certain response from a cell.
  - Impact of living cell on construct and vice versa cannot be ignored.
  - Models to predict performance must account for both host and construct dynamics.

AI techniques have been leveraged that combine known biophysical, machine learning, and reinforcement learning models to effectively predict the constructs' impact on the host and vice versa, but there is much room for improvement. For example, for machine-assisted gene circuit design, a variety of

Figure 3. Significant growth in both academic (a) and commercial (b) domains provide rich sources of information, data, and context for the application of AI in the synbio field.[8]

Figure 3(a) from Shapira et al.[33] under Creative Commons License.
Figure 3(b) used with permission.

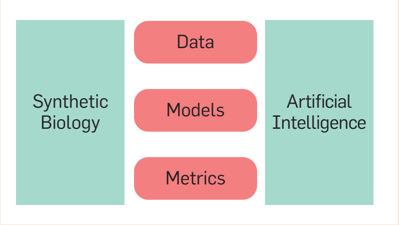


AI techniques have been applied. They include expert systems, multi-agent systems, constraint-based reasoning, heuristic search, optimization, and machine learning.[2,30,38] Sequence-based models and graph convolutional networks have also gained traction in the domain of engineering biological systems. Factor-graph neural networks[27] have been used to incorporate biological knowledge into deep learning models. Graph convolutional networks have been used to predict functions of proteins[1,39] from protein-protein interaction networks. Sequence-based convolutional and recurrent neural network models have been used to identify potential binding sites of proteins,[1] the expression of genes,[36] and the design of new biological constructs.[12] Some of the most useful applications of AI will be in the development of comprehensive models that will reduce the number of experiments (or designs) that need to be conducted (or tested).

Metabolic engineering. In metabolic engineering, AI has been applied to almost all stages of the bioengineering process.[22,23,31] For example, artificial neural networks have been used to predict translation initiation sites, annotate protein function, predict synthetic pathways, optimize the expression level of multiple heterologous genes, predict strength of regulatory elements, predict plasmid expression, optimize nutrient concentration and fermentation conditions, predict enzyme kinetic parameters, understand genotype-phenotype associations, and predict CRISPR guide efficacy. Clustering has been used to find secondary metabolite biosynthetic gene clusters and identify enzymes that catalyze a specific reaction. Ensemble approaches have been used to predict pathway dynamics, optimal growth temperatures, and find proteins that confer higher fitness in directed evolution approaches. Support vector machines have been used to optimize ribosome binding site sequences and predict the activity of CRISPR guide RNAs. The most promising metabolic engineering stages for the application of AI are: process scale-up, a significant bottleneck in the field,[6,37] and downstream processing (for example, systematic extraction of the produced molecule from the fermentation broth).

Figure 4. Challenges of integrating AI techniques with synbio applications.

Data is often multimodal, difficult to integrate, and lacks metadata. Models have been developed for leveraging large amounts of data and lack explainability and uncertainty quantification. Metrics need to be rethought to truly rank the available models in a larger context.



Experiment automation. AI impact has reached well beyond the "Learn" phase of the DBTL cycle, in helping to automate lab work and recommending experimental designs. Automation is slowly becoming a key practice as the most reliable way to obtain the high-quality, high-volume, low-bias data needed to train AI algorithms and enable predictable bioengineering.[4] Automation offers the opportunity to rapidly transfer and scale complex protocols to other labs. As an example, liquid-handling robotic stations[35] form the backbone of biological foundries and cloud labs.[20] These foundries have seen their capabilities revolutionized by robotics and planning algorithms, enabling fast iterations through the DBTL cycles. Semantic networks, ontologies, and schemas have revolutionized the representation, communication, and exchange of designs and protocols. These tools have enabled rapid experimentation and the generation of significantly more data in a structured, queryable format. In a domain where most context was either lost or captured manually in lab notebooks, the promise of AI has forced a significant change in the domain to reduce the barrier to generate data.

Microfluidics[14,24] represent an alternative to macroscopic liquid handlers that provide higher throughput, less reagent consumption, and cheaper scaling. Indeed, microfluidics might be the key technology that enables self-driving labs,[19] which promise to substantially accelerate the discovery process by augmenting automated experimentation platforms with AI. Self-driving labs involve fully automated DBTL cycles in which AI algorithms actively search for promising experimental procedures by hypothesizing about their results based on previous experiments. As such, they may represent the largest opportunity for AI researchers in the synbio field. While automated DBTL loops have been demonstrated in liquid-handling robotic stations, the scalability, high-throughput capabilities, and fabrication flexibility provided by microfluidic chips may provide the final technological leap that makes scientist AIs a reality.

Challenges

AI has begun to make its way into various synbio applications, but major technological and sociological hurdles continue to separate both fields.

Technological challenges. The technical challenges of applying AI to synbio (see Figure 4) are that data is scattered in different modalities, difficult to combine, unstructured, and often lack the context in which they were collected; models require significantly more data than is often collected in a single experiment and lack explainability and uncertainty quantification; and there are no metrics or standards to effectively evaluate model performance in the larger design task at hand. Furthermore, experiments are often designed to explore only positive outcomes, complicating or biasing the evaluation of the model.

Data challenges. The lack of appropriate datasets remains the first major hurdle to merging AI with synthetic biology. Applying AI to synthetic biology requires large volumes of labeled, curated, high-quality, contextually rich data from individual experiments. Although the community has made progress in setting up databases[28] containing various biological sequences (even whole genomes) and phenotypes, there is a scarcity of labeled data. By "labeled data" we mean phenotypic data mapped to measurements that capture their biological function or cellular responses. It is the presence of such measurements and labels that will drive the maturity of AI/ML and synbio solutions to rival human competency, as it has done in other fields.

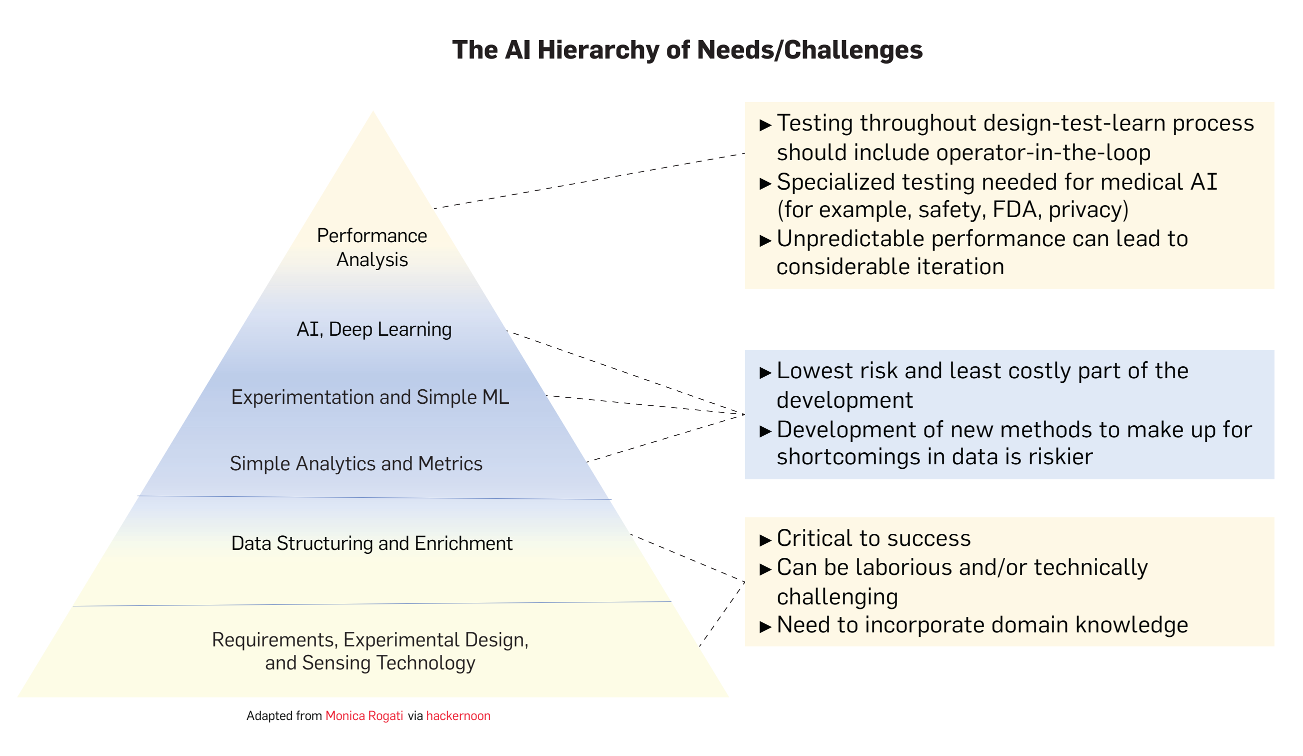
A lack of investment on data engineering is partially responsible for the lack of appropriate datasets. Advancements in AI techniques often overshadow the computing infrastructure requirements that support and ensure its success. The AI community refers to this canonical infrastructure as the *pyramid of needs*[32] (see Figure 5), of which data engineering is an important component. Data engineering encapsulates the experimental planning, data collection, structuring, accessing, and exploration steps. Successful AI application stories involve a data engineering step that is standardized, consistent, and reproducible. While we now can collect biological data at unprecedented scale and detail,[5] this data is often not immediately suitable for machine learning. There are still many hurdles in the adoption of community-wide standards to store and share measurements, experimental conditions, and other metadata that would make them more amenable to AI techniques.[13,28] Rigorous formalization work and consensus is required to make such standards rapidly adoptable and to promote common metrics of data quality evaluation. In short, AI models require consistent and comparable measurements across all experiments, which prolongs experimental timelines. This requirement adds significant overhead for the experimentalists who are already following complex protocols to make scientific discoveries. Thus, the long-term needs of the data collection are often sacrificed to meet the tight deadlines often imposed on such projects.

This situation often results in sparse data collections that represent only a small part of the multiple layers that form the omics data stack (shown in Figure 1). In these cases, data representation has a significant impact on the ability to integrate these siloed datasets for comprehensive modeling. Today, significant effort is spent across a variety of industry verticals performing data cleansing, schema alignment, and extract, transform, and load operations (ETL) to collect and prepare unruly digital data into a form suitable for analysis. These tasks account for nearly 50% to 80% of a data scientist's time, limiting their ability to extract insights.[26] Dealing with a large variety of data types (data multimodality) is a challenge for synthetic biology researchers, and the complexity of preprocessing activities increases dramatically as a function of data variety compared to data volume.

Modeling/algorithmic challenges. Many of the popular algorithms fueling current AI advances (for example, in the computer vision and natural language processing fields) are not robust when it comes to analyzing omics data. Traditional application of these models often suffers from the "curse of dimensionality" when applied to data collected in a specific experiment (see Figure 6). For example, a single experimentalist can produce genomics, transcriptomics, and proteomics data for an organism under a particular condition that will provide more than 12,000 measurements (dimensions). The number of labeled instances (for example, success or failure) for such an experiment is often in the tens to hundreds, at most. The dynamics of the system (time resolution) are seldom captured for these high-dimensional data types. These measurements gaps make driving inferences about complex, dynamical systems a significant challenge.

Omics data share similarities and differences with other data modalities such as sequential data, text data, and network-based data, but classical approaches are not always applicable. The shared data characteristics include positional encoding and dependencies, as well as complex interaction patterns. Yet there are some fundamental differences such as: their underlying representation, context required for meaningful analyses, and associated normalizations across modalities to make biological

Figure 5. A canonical AI/ML infrastructure can support synbio research. The middle stages are often a focus of attention, but the base is crucial and needs significant resource investment.


The AI Hierarchy of Needs/Challenges
Performance Analysis
AI, Deep Learning
Experimentation and Simple ML
Simple Analytics and Metrics
Data Structuring and Enrichment
Requirements, Experimental Design, and Sensing Technology

- ▸ Testing throughout design-test-learn process should include operator-in-the-loop
- ▸ Specialized testing needed for medical AI (for example, safety, FDA, privacy)
- ▸ Unpredictable performance can lead to considerable iteration

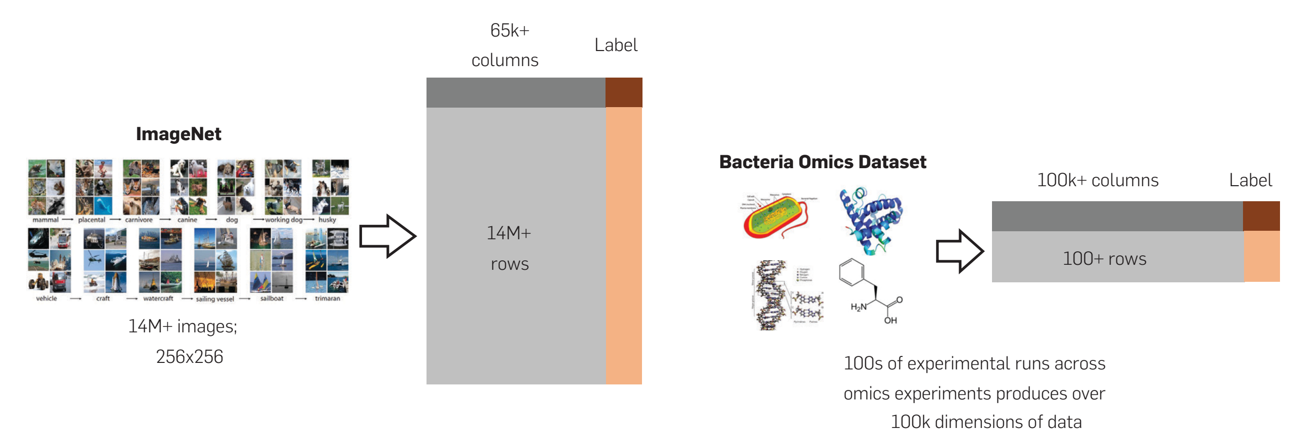
- ▸ Lowest risk and least costly part of the development
- ▸ Development of new methods to make up for shortcomings in data is riskier

- ▸ Critical to success
- ▸ Can be laborious and/or technically challenging
- ▸ Need to incorporate domain knowledge

Adapted from Monica Rogati via hackernoon

Figure 6. The curse of dimensionality.

Traditional datasets used in deep learning applications consist of millions of instances in a high-dimensional space. ImageNet, for example, has more than 14M images at a resolution of 256x256, which leads to a 65,536-dimensional representation of images.[9] Omics datasets, on the other hand, typically have 100s of instances (rows) across an even higher omic-dimensional space that can grow beyond 100k dimensions.



meaningful comparisons. Consequently, it's rare to find robust classes of generative models (akin to Gaussian models or stochastic block models[18]) that can accurately characterize omics data. Furthermore, biological sequences and systems represent sophisticated encodings of biological functions, but there are few systematic approaches to interpret these encodings in a similar way that we interpret semantics or context from written text. These different characteristics make it challenging to extract insights via data exploration and generate and verify hypotheses. Engineering biology involves the challenge of learning about a black box system, where we can observe input and output, but we have limited knowledge about the inner workings of the system. Considering the combinatorial, large parameter space that these biological systems operate in, AI solutions that strategically and efficiently design experiments to probe and interrogate biological systems for hypothesis generation and verification present a tremendous need and opportunity in this space.[19]

Lastly, many of the popular AI algorithmic solutions do not explicitly account for uncertainty and do not display robust mechanisms for controlling errors under input perturbations. This fundamental gap is particularly critical in the synbio space, considering the inherent stochasticity and noise in the biological systems we are trying to engineer.

Metrics/evaluation challenges. Standard AI evaluation metrics based on prediction and accuracy are insufficient for synbio applications. Metrics such as $\mathbb{R}^2$ for regression models or accuracy for classification-based models do not account for the complexity of the underlying biological system that we are trying to model. Additional metrics that quantify the degree to which a model can elucidate the inner workings of the biological system and capture existing domain knowledge are equally important in this field. To this end, AI solutions that incorporate principles of interpretability and transparency are key in supporting iterative and interdisciplinary research. Also, the ability to properly quantify uncertainty requires the creative development of novel metrics to gauge the effectiveness of these approaches.

Metrics for proper experimental design are also needed. Evaluation and validation of models in synbio will at times call for additional experiments, requiring extra resources. A handful of misclassifications or small errors can have a drastic impact on the research goal. These costs should be integrated into objective functions or evaluations of the AI models to reflect the real-world impact of a misclassification.

Sociological challenges. Sociological hurdles can be more challenging than technical ones in leveraging AI to benefit synbio (and vice versa). It is our impression that many impediments stem from a lack of coordination and understanding between the very different cultures involved. While there are certain initiatives that have begun to overcome these challenges, it is interesting to note that persistent themes remain problematic in both academia and industry.

The root of sociological challenges. These challenges spring from the need to blend expertise from two very different groups: computational scientists and bench scientists.

Computational and bench scientists are trained very differently (see Figure 7). Computational scientists, by training, tend to focus on abstractions, be enthused about automation and computational efficiency and disruptive approaches. They naturally lean toward task specialization and look for ways to hand off repeated tasks to an automated computer system. Bench scientists are practical, trained in working with concrete observations, and prefer explainable analyses to accurately describe the specific outcome of an experiment.

These two worlds profess different cultures, reflected not only in how they solve problems, but also which problems they consider worth solving. For example, there is a continuous tension regarding the amount of effort devoted to building infrastructure that supports general research versus aiming to study a specific research question. The computational scientist favors providing reliable infrastructure that can be counted on for a variety of projects (for example, an automated pipeline for strain construction or a centralized database collecting all relevant data); whereas bench scientists tend to focus on the final goals (for example, producing a desired molecule in commercially meaningful amounts), even if that means relying on bespoke approaches that are valid only for that specific case. In this regard, computational scientists like to develop mathematical models that explain and predict the behavior of the biological systems, whereas bench scientists prefer producing qualitative hypotheses and testing them experimentally as soon as possible (at least when working with microorganisms, since those experiments can be completed quickly: 3–5 days). Furthermore, the computational scientists can often only get excited and energized about lofty, blue-sky goals, such as bioengineering organisms to terraform Mars, writing a compiler of life able to create DNA to fulfill a desired specification, reengineering trees to adopt a desired shape, bioengineering dragons in real life, or substituting scientists by AIs. The bench scientists see these lofty goals as "hype," are rather burnt because of previous examples of overpromising and underdelivering by computational types, and would rather only consider goals that can be attained using the current state of technology.

Addressing sociological challenges. The solution to these sociological hurdles is to appreciate interdisciplinary teams and requirements. Admittedly, achieving this inclusive environment may be easier in a company (where the team sinks or succeeds together) than in an academic environment (where a graduate student or postdoc pursue publication of a few first-author papers to claim success, without the need for integration with other disciplines).

A possible route for this integration is creating cross-training courses, where bench scientists are trained in programming and machine learning, and computational scientists are trained in experimental work. In the end, both communities are bringing something valuable, unique, and necessary to the table. The sooner this is readily apparent to everyone involved, the faster synbio can advance. In the long term, we need university curriculums that combine the teaching of biology and bioengineering with automation and math. While several initiatives are currently underway, they are just a drop in the bucket of the needed workforce.

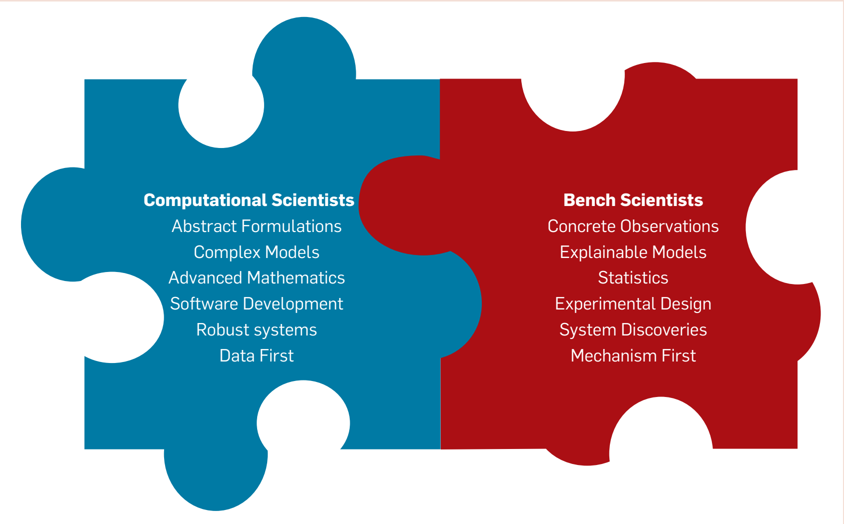
Perspectives and Opportunities

AI can radically enhance synbio and enable its full impact by opening a third axis in the engineering phase space: physical, chemical, and biological. Most obviously, AI can produce accurate predictions in bioengineering outcomes, enabling effective inverse design. Furthermore, AI can support the scientist in designing experiments and choosing when and where to sample, a problem that currently requires a highly trained expert. AI can also support automated search, high throughput analysis and hypothesis generation from large data sources including historical experimental data, online databases, ontologies, and other technical material. AI can augment the knowledge of the synbio domain expert by allowing faster exploration of large design spaces and by recommending interesting, "outside the box" hypotheses. Synbio presents some unique challenges for the current AI solutions that, if resolved, will lead to fundamental advances in both the synbio and AI fields. Engineering a biological system is intrinsically reliant on the ability to control the system; this is the ultimate test for understanding the fundamental laws that govern the system. Therefore, an AI solution that can enable synbio research must be able to describe the mechanism that led to the best prediction.

While recent AI techniques based on deep learning architectures have changed our perspective on how we approach feature engineering and pattern finding, they are still at their infancy in terms of their ability to reason and interpret their learning mechanisms. To this effect, AI solutions that incorporate causal reasoning, interpretability, robustness, and uncertainty estimation requirements have immense potential impact in this interdisciplinary area. The complexity of biological systems is such that AI solutions based purely on brute-force correlation finding will fail to efficiently characterize the system's intrinsic features. A new class of algorithms that smoothly incorporates physics and mechanistic models with data-driven models is an exciting new research direction. We see some initial positive results in climate science and computational chemistry and hopefully similar advancements will follow in the study of biological systems.[16,25]

Synbio can also inspire new AI approaches, since it provides the tools to modify biological systems. Let us not forget that biology inspired such staples of AI as neural networks, genetic algorithms, reinforcement learning, computer vision, and swarm robotics. It would be surprising if biology could not provide further inspiration. Indeed,

Figure 7. Computational and bench scientists come from different research cultures that must learn to work together to fully benefit from combining AI and synbio.



more online

For more information, and a robust list of references, see supplemental material at http://dl.acm.org/doi/10.1145/3500922

there are many biological phenomena that would be desirable to emulate digitally. Gene regulation, for example, involves an exquisitely crafted network of interactions that allows cells to not only sense and react to the environment but also to keep the cell alive and stable. Keeping homeostasis (the state of steady internal, physical, and chemical conditions maintained by living systems) involves producing the right components of the cell at the right moment, and at the right amount, sensing internal gradients, and carefully regulating the cell's exchange with its environment. Can we understand and leverage this capability to produce truly self-regulating AIs or robots? Another example involves emergent properties (that is, properties exhibited by the system but not by its constituent parts). For instance, ant colonies behave and react as a single organism that is much more sophisticated that the sum of its parts (the ants). In a similar fashion, consciousness (that is, sentience or awareness of internal or external existence) is a qualitative trait that arises from a physical substrate (for example, neurons). Swarm robots that self-organize and collectively build structures already exist. Could we use a general theory of emergence to create hybrids of robots and biological systems? Could we create consciousness from a very different physical substrate (for example, transistors instead of neurons)? A final possible example involves self-healing and replication: even the least sophisticated example of life exhibits the ability to self-repair and reproduce. Could we understand the quandaries of this phenomenon to produce self-repairing and replicating AIs?

While this kind of biological mimicry has been considered before, the beauty of synbio lies in providing us with the capability to "tinker" with biological systems to test the models and underlying principles of biomimicry. For example, we can now tinker with cell gene regulation at a genomic-scale to modify it and test what we believe to be the underlying reasons for its remarkable resilience and adaptability. Or we can bioengineer ants and test what kind of ant colony behavior ensues, and how it affects its survival rate. Or we can alter cell self-repair and self-replication mechanisms and test the long-term evolutionary effects on its ability to compete.

Furthermore, in cell modeling we are very close to a good understanding of the involved biological mechanisms. While there is little hope that understanding how a neural network detects the shape of an eye would reveal how the brain does the same, that is not the case in synbio. Mechanistic models are not perfect in their predictions,[21] but produce qualitatively acceptable results. Combining these mechanistic models with the predictive power of ML can help bridge the gap between both and provide biological insight into why some ML models are more effective at predicting biological behavior than others. This insight can lead into new ML architectures and approaches.

AI can help synbio, and synbio can help AI; but it is ultimately the interaction of these two disciplines in a continuous feedback loop that will create possibilities we cannot even fathom right now. In the same way Benjamin Franklin could not imagine his discovery of electricity would someday enable the Internet.

Getting Involved

The interface between AI and synbio is a budding interdisciplinary field that needs more AI researchers to fully flourish. How can you get involved? We recommend several community-wide, strategic efforts to support interdisciplinary research in AI-enabled synbio:

- Attendance at and formation of conferences that support standardization of data collection and storage and facilitate sharing of synbio-related benchmark data for comparing and evaluating AI solutions.
- Democratization and ease of access to AI and synbio tools.
- Supporting and requesting tracks in conferences in both domains, such as SEED's track on Computational Biology and Artificial Intelligence, and the AAAI symposia on AI and Synthetic Biology.
- Identification of canonical synbio challenge problems similar to protein structure and CASP challenge.[7]

Furthermore, there is significant public funding for research in these fields. Public investment from the U.S. Department of Defense (DoD) and the Department of Energy (DoE) have funded research in the domain for years from applications of identifying new materials to the production of biofuels. The DoE has been the leading organization in its investment in applications from biofuel production, agriculture, and energy conversion.[34] The DoD has invested significantly in synthetic biology. Among the many DoD programs in the field, the DARPA Living Foundries program further focused efforts through automation. They successfully managed to reduce the time and cost to engineer organisms 10x. The access, analysis, and understanding of the data generated has led to an explosion in our understanding of biology, making it more accessible and predictable. Another effort involves the Applied Research for the Advancement of Science and Technology Priorities (ARAP) on developing capabilities for Synthetic Biology for Military Environments (SBME), funded by the Office of the Secretary of Defense (OSD) in 2017–2019. This $45 million tri-service effort leveraged DoD laboratory expertise to use biological systems for defense and resulted in long-term infrastructure and community resources for synthetic biology. SBME initiated annual Synthetic Biology for Defense workshops which resulted in even further collaboration with academia and industry on synthetic biology efforts. Through SBME, each of the Air Force, Army, and Navy Research Laboratories also mentored a team for the International Genetically Engineered Machine (iGEM) Foundations iGEM Competition, which enables students to address challenge problems using synthetic biology. Moving forward with commitment to synbio in December 2019, the DoD announced the establishment of the Biotechnology Community of Interest which will enhance coordination, collaboration, and communication across the DoD biotechnology research and development (R&D) components and wider biotechnology community, including public-private partnerships with academia and industry.[10] The DoD, as part of the biotechnology modernization priority, is also currently investing in a Bioindustrial Manufacturing Innovation Institute to scale biomanufacturing processes and biotechnologies with industry and academia. The services continue to collaborate through these new initiatives to drive the bioeconomy to meet military needs.

The National Science Foundation (NSF) and National Institutes of Health (NIH) have also begun to define their initiatives on synbio. NIH has stood up a synbio consortium to have researchers identify roadmaps for the application of synbio in vaccine development, immunotherapy, and other areas of research applicable to healthcare applications. The NSF, on the other hand, is taking a broader approach to understand the mechanistic modeling of a variety of biological networks, computational methods, and molecular to systemwide rules in natural or synthetic microbial communities that could lead to a fundamental understanding of these biological systems.

Another approach would be a large "moon-shot" project, executed over 10–20 years, bringing together experimentalists, theorists, and computationalists, with a strong educational component to educate the next generation of practitioners. For example, plants are notoriously hard to engineer for a variety of reasons, including their long growth cycles. A possible project could address engineering plants to reduce climate change by making plants more resilient and increasing the amount of carbon the plants sequester.

Finally, private industries in domains ranging from drug discovery to materials science, to food and beverages are all turning to synbio for their next wave of products. Companies such as Amyris, Conagen, Ginkgo Bioworks, and Zymergen have embraced the Living Foundries vision to engineer and automate the design of DNA to rapidly loop through the design-build-test-learn cycle and have cells produce or detect an item of interest.

References

1. Alipanahi, B., Delong, A., Weirauch, M., and Frey, B. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology 33*, 8 (Aug. 2015), 831–838; https://doi. org/10.1038/nbt.3300
2. Bilitchenko, L., Liu, A., and Densmore, D. The Eugene language for synthetic biology. *Methods in Enzymology 498* (2011), 153–172; https: //doi.org/10.1016/B978-0-12-385120-8.00007-3
3. Cameron, D., Bashor, C., and Collins, J. A brief history of synthetic biology. *Nature Reviews Microbiology 12*, 5 (Apr. 2014), 381–390; https://doi.org/10.1038/nrmicro3239
4. Carbonell, P., Radivojevic, T., and Martín, H. Opportunities at the intersection of synthetic biology, machine learning, and automation. *ACS Synthetic Biology 8*, 7 (Jul. 2019), 1474–1477; https://doi. org/10.1021/acssynbio.8b00540
5. Chen, Y., et al. Automated "cells-to-peptides" sample preparation workflow for high-throughput, quantitative proteomic assays of microbes. *J. Proteome Research 18*, 10 (Oct. 2019), 3752–3761; https://doi.org/10. 1021/acs.jproteome.9b00455
6. Chubukov, V., Mukhopadhyay, A., Petzold, C., Keasling, J., and Martín, H. Synthetic and systems biology for microbial production of commodity chemicals. *NPJ Systems Biology and Applications 2* (Apr. 2016), 16009; https://doi.org/10.1038/npjsba.2016.9
7. Croll, T., Sammito, M., Kryshtafovych, A., and Read, R. Evaluation of template-based modeling in CASP13. *Proteins 87*, 12 (Aug. 2019), 1113–1127; https://doi. org/10.1002/prot.25800
8. Cumbers, J. *Meet Eight Tech Titans Investing in Synthetic Biology*; https://bit.ly/3ItFelL
9. Devopedia. ImageNet, 2019; https://devopedia.org/imagenet
10. DiEuliis, D., Terrell, P., and Emanuel, P. Breaching the department of defense's biotech bottleneck. *Health Security 18*, 2 (2020), 139–144; https://doi.org/10.1089/hs.2019.0150
11. Doudna, J. and Charpentier, E. Genome editing: The new frontier of genome engineering with CRISPR-Cas9. *Science 346*, 6213 (Nov. 2014), 1258096; https://doi. org/10.1126/science.1258096
12. Eastman, P., Shi, J., Ramsundar, B., and Pande, V. Solving the RNA design problem with reinforcement learning. *PLoS Computational Biology 14*, 6 (Jun. 2018), e1006176; https://doi.org/10.1371/journal.pcbi.1006176
13. El Karoui, M., Hoyos-Flight, M., and Fletcher, L. Future trends in synthetic biology—a report. *Frontiers in bioengineering and biotechnology 7* (Aug. 2019), 175; https://doi.org/10.3389/fbioe.2019.00175
14. Gach, P., Iwai, K., Kim, P., Hillson, N., and Singh, A. Droplet microfluidics for synthetic biology. *Lab on A Chip 17*, 20 (Oct. 2017), 3388–3400; https://doi. org/10.1039/c7lc00576h
15. Gardner, T. Synthetic biology: From hype to impact. *Trends in Biotechnology 31*, 3 (Mar. 2013), 123–125; https://doi.org/10.1016/j.tibtech.2013.01.018
16. Gaw, N., et al. Integration of machine learning and mechanistic models accurately predicts variation in cell density of glioblastoma using multiparametric MRI. *Scientific Reports 9*, 1 (Jul. 2019), 10063; https://doi. org/10.1038/s41598-019-46296-4
17. Gersbach, C. Genome engineering: The next genomic revolution. *Nature Methods 11*, 10 (Oct 2014), 1009–1011; https://doi.org/10.1038/nmeth.3113
18. Gupta, S., Dukkipati, A., and Castro, R. Restricted boltzmann stochastic block model: A generative model for networks with attributes. *arXiv* (Nov. 2019).
19. Häse, F., Roch, L., and Aspuru-Guzik, A. next-generation experimentation with self-driving laboratories. *Trends in Chemistry 1*, 3 (Mar 2019), 282-291. https://doi. org/10.1016/j.trechm.2019.02.007
20. Jessop-Fabre, M. and Sonnenschein, N. Improving reproducibility in synthetic biology. *Frontiers in bioengineering and biotechnology 7* (Feb. 2019), 18; https://doi.org/10.3389/fbioe.2019.00018
21. Karr, J., et al. A whole-cell computational model predicts phenotype from genotype. *Cell 150*, 2 (Jul. 2012), 389–401; https://doi.org/10.1016/j.cell. 2012.05.044
22. Kim, G., Kim, W., Kim, H., and Lee, S. Machine learning applications in systems metabolic engineering. *Current Opinion in Biotechnology 64* (Sep. 2019), 1-9; https://doi. org/10.1016/j.copbio.2019.08.010
23. Lawson, C., et al. Machine learning for metabolic engineering: A review. *Metabolic Engineering 63* (2021), 34–60; https://doi.org/10.1016/j.ymben.2020.10.005.
24. Le, K., et al. A novel mammalian cell line development platform utilizing nanofluidics and optoelectro positioning technology. *Biotechnology Progress 34*, 6 (Sept. 2018), 1438–1446; https://doi.org/ 10.1002/btpr.2690
25. Lessler, J., Azman, A., Grabowski, M., Salje, H., and Rodriguez-Barraquer, I. Trends in the mechanistic and dynamic modeling of infectious diseases. *Current Epidemiology Reports 3*, 3 (Jul. 2016), 212–222; https:// doi.org/10.1007/s40471-016-0078-4
26. Lohr, S. *For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights.*; https://www.nytimes. com/2014/08/18/ technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html
27. Ma, T. and Zhang, A. Incorporating biological knowledge with factor graph neural network for interpretable deep learning. *arXiv* (Jun 2019).
28. Morrell, W., et al. The experiment data depot: A web-based software tool for biological experimental data storage, sharing, and visualization. *ACS synthetic biology [electronic resource] 6*, 12 (Dec 2017), 2248-2259. https://doi.org/10.1021/acssynbio.7b00204
29. National Academies Press. Committee on industrialization of biology: A roadmap to accelerate the advanced manufacturing of chemicals, board on chemical sciences, technology, board on life sciences, division on earth, life studies, and national research council. *Industrialization of biology: A roadmap to accelerate the advanced manufacturing of chemicals* (2015); https://doi.org/10.17226/19001
30. Pedersen, M. and Phillips, A. Towards programming languages for genetic engineering of living cells. *Journal of the Royal Society, Interface* 6 Suppl 4 (Aug. 2009), S437–50; https://doi.org/10.1098/rsif.2008.0516.focus
31. Presnell, K. and Alper, H. Systems metabolic engineering meets machine learning: A new era for data-driven metabolic engineering. *Biotechnology J. 14*, 9 (Sep. 2019), e1800416; https://doi.org/10.1002/biot.201800416
32. Rogati, M. *The AI Hierarchy of Needs* (2017); https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007
33. Shapira, P., Kwon, S., and Youtie, J. Tracking the emergence of synthetic biology. *Scientometrics 112*, 3 (Jul. 2017), 1439–1469; https://doi.org/ 10.1007/s11192-017-2452-5
34. Si, T. and Zhao, H. A brief overview of synthetic biology research programs and roadmap studies in the United States. *Synthetic and Systems Biotechnology 1*, 4 (Dec. 2016), 258–264; https://doi.org/10.1016/j. synbio.2016.08. 003
35. Unthan, S., Radek, A., Wiechert, W., Oldiges, M., and Noack, S. Bioprocess automation on a Mini Pilot Plant enables fast quantitative microbial phenotyping. *Microbial Cell Factories* 14, 1 (Dec. 2015), 216; https: //doi.org/10.1186/s12934-015-0216-6
36. Wang, M., Tai, C., Weinan, E., and Wei, L. DeFine: deep convolutional neural networks accurately quantify intensities of transcription factorDNA binding and facilitate evaluation of functional non-coding variants. *Nucleic Acids Research 46*, 11 (Jun. 2018), e69; https://doi.org/10.1093/nar/gky215
37. Wehrs, M., Tanjore, D., Eng, T., Lievense, J., Pray, T.R., and Mukhopadhyay, A. Engineering robust production microbes for large-scale cultivation. *Trends Microbiology 27*, 6 (Jun 2019), 524–537; doi:10.1016/j.tim.2019.01.006
38. Yaman, F., Bhatia, S., Adler, A., Densmore, D., and Beal, J. Automated selection of synthetic biology parts for genetic regulatory networks. *ACS Synthetic Biology 1*, 8 (Aug. 2012), 332–344; https://doi.org/10.1021/sb300032y
39. Zitnik, M. and Leskovec, J. Predicting multicellular function through multi-layer tissue networks. *Bioinformatics 33*, 14 (Jul. 2017), i190–i198; https://doi. org/10.1093/bioinformatics/btx252

Mohammed Eslami is Chief Data Scientist and co-founder at Netrias, LLC, Arlington, VA, USA.

Aaron Adler is Senior Scientist at Raytheon BBN, Columbia, MD, USA.

Rajmonda S. Caceres is Senior Technical Staff at MIT Lincoln Laboratory, Lexington, MA, USA.

Joshua G. Dunn is Head of Design at Ginkgo Bioworks, Boston, MA, USA.

Nancy Kelley-Loughnane is Biological Materials and Processing Research Team Lead at Air Force Research Laboratory, Wright Patterson Air Force Base, OH, USA.

Vanessa A. Varaljay is Bioinformatics Lead at Materials and Manufacturing Directorate (recently moved to be Genomics Lead at the 711 Human Performance Wing), Air Force Research Laboratory, Wright-Patterson, OH, USA.

Hector Garcia Martin is Staff Scientist at Lawrence Berkeley National Laboratory, Learn co-lead at Agile BioFoundry, Group Lead at Joint BioEnergy Institute, and External Scientific Member at Basque Center for Applied Mathematics at Emeryville, CA, USA.

Watch the authors discuss this work in the exclusive *Communications* video. https://cacm.acm.org/videos/ai-for-synthetic-biology