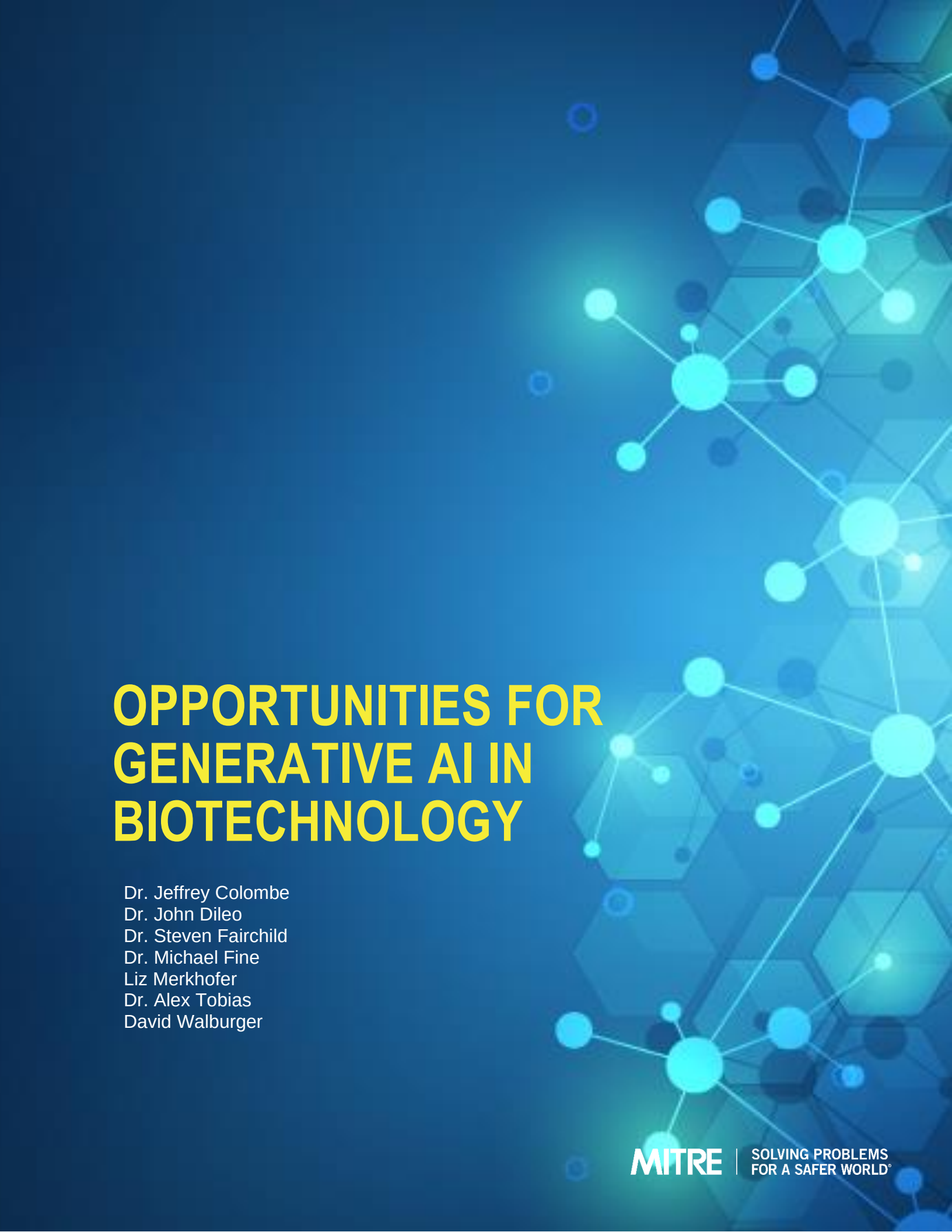

OPPORTUNITIES FOR GENERATIVE AI IN BIOTECHNOLOGY

Dr. Jeffrey Colombe
Dr. John Dileo
Dr. Steven Fairchild
Dr. Michael Fine
Liz Merkhofer
Dr. Alex Tobias
David Walburger

MITRE | SOLVING PROBLEMS FOR A SAFER WORLD®

EXECUTIVE SUMMARY

ChatGPT and related emerging machine learning based artificial intelligence (AI) tools have provoked a classic but unusually rapid hype cycle of public interest and doubt in the potential for AI to enable and/or disrupt a variety of economically significant labor roles. The unusual step of providing open public access for people to try these new tools has led to a great deal of speculation—and experimentation—into how such tools might create opportunities and risks.

MITRE hosted a Technical Exchange Meeting (TEM) titled, "Opportunities for generative AI in biotechnology" on 27 June 2023 in McLean, Virginia, with representatives from academia, industry, and government. The TEM focused on the potential for such emerging tools to impact progress in biotechnology based on their abilities to compose novel content, to engage in meaningful and informative dialogue with a human, and to explain ideas. While the impacts overall are unclear, recent progress suggests the potential for substantial disruption in biotechnology, both of positive value (major advances, for example in biomanufacturing and pharmaceutical development) and negative value (risks and dangers, for example in the development of chem-bioweapons capabilities, or sudden destabilization in economic competitiveness and self-sufficiency of the U.S. and its allies).

This paper summarizes key findings, and explores opportunities for emerging generative AI tools in biotechnology applications, including:

- Predictive design of useful biological systems that scale from molecules to organisms,
- Trustworthy and explainable results of AI outputs, including exposed chains of reasoning and references to scientific source literature or other forms of evidence,
- Exploratory mapping and tracking of trends in knowledge in the biotechnology field, and
- Development of conversational AI research assistant software.

Findings from the TEM include:

- *Current state of practice:* Generative AI is currently being developed to design biological products with desired properties, and to serve as research assistants with a conversational interface. Language-only models have inherent limitations for representation, prediction, and causal inference, even if trained on domain-specific data. Specialized generative models for specific biotechnology applications like the design of proteins and metabolic pathways may not support exposed reasoning and conversational query.
- *Performance metrics:* Performance measures are immature for generative AI models that are the first forms of AI to rival human cognitive performance in an array of tasks.
- *Hallucination:* Current large language models (LLMs) hallucinate plausible but nonexistent content such as statements about facts and authoritative citations, and have no means to discriminate truth value from creative synthesis of content derived from statistical patterns in training data.
- *Security risks:* Generative AI may be used to design chemical and biological weapons, and to conceal their function from conventional analytics used to screen out bioproduct synthesis with malicious intent. Generative AI used in biotechnology may also upset U.S. and allied economic and technical self-sufficiency.
- *Hardware and algorithms:* The coevolution of specialized high performance computing hardware and efficient algorithms for running neural network models drives both massive-scale disruption and small-scale democratized innovation.

- *Knowledge mapping:* Generative AI can be used to track topical trends in the biomedical literature, to map knowledge and gaps in knowledge, identify controversy and uncertainty, and generate hypotheses to test.

Recommendations from the TEM include:

1. *Multimodal training:* Continue to advance the research frontier of training multimodal models on biotechnology data, to include research literature (natural language, tables, and figures), gene sequences, protein 3-dimensional (3D) structure, reaction rates and binding affinities for enzymes and their products, biochemical pathway graphs, and physiologic and phenotypic variables at system and organism scales, among other modalities of data, with appropriate cross-references. Multimodal models are regarded as critical to realizing the full range of application needs in biotechnology, as language-only models have inherent limitations for representation, prediction, and causal inference.

2. *Performance measures:* Develop ways to meaningfully measure the performance of generative AI in task environments that have to date only been performed by humans. Cognitive science is still not mature in measuring various forms of human intelligence, and as such, ready-made metrics are imperfect and require advancement.

3. *Enhance trustworthiness and facticity of models by training on truth value:* Truth value can be trained using consensus or "gold standard" human curation, marking of authoritative sources prior to training and allowing verbatim citations of training data where appropriate, providing linguistic representations with sensory and/or motor meaning using multimodal models, training models to expose the chains of reasoning used to generate outputs, and development of intrinsic cognitive models that seek to optimize scale-model likeness of structure and function with the phenomena under study for accurate prediction and causal inference.

4. *Global security:* Track risks to global security that include disruption of employment, intellectual property issues, counter-WMD, and economic competitiveness of the U.S. and its allies.

5. *Promote compute capabilities to foster innovation:* Adapt both high performance computing (HPC) resources and algorithms in synergy to allow democratization and commoditization of research into novel capabilities.

6. *Develop knowledge-mapping models and conversational scientific assistants:* Adapt current AI tools to track trends in knowledge in biotechnology literature, and to identify controversies, uncertainties, and hypotheses worth investigating. Develop conversational assistant software to engage in dialogue with human users to answer queries and to collaborate on research activities. Develop standard formats for publication contents that facilitate multimodal machine learning of biotechnology and other scientific data, and cross-referencing of content in publications with scientific databases.

Technial Exchange Meeting

The Technical Exchange Meeting (TEM) titled, "Opportunities for generative AI in biotechnology" was hosted on 27 June 2023 at the MITRE Corporation in McLean, VA. The meeting involved experts and stakeholders across government, industry, and academia in a non-attributed forum designed to facilitate open conversation. The findings discussed in this paper are a summary of presentations at the TEM, as well as a literature review performed by MITRE subject matter experts. The recommendations are a synthesis driven by the conclusions presented in the remainder of this paper.

The following sections each treat a major component of the technical material deemed relevant to assessing prospects for generative AI to make an impact in biotechnology.

How do the emerging machine learning tools work?

ChatGPT and its competitors are large language models (LLMs). Language modeling dates to Claude Shannon, the father of information theory, and captures the simultaneous occurrence (a.k.a. *co-occurrence*) of words or characters based on observations from a large body of text. These models were initially useful for estimating the quality of language, for example, to select the better machine translation. In 2013, open source, high quality word *embeddings* were first developed by the natural language processing (NLP) community. Word embeddings, a mathmatical way to encode the meaning of a word, were learned from the co-occurrence of two or more words. These embeddings latently captured both regularities and meaning, and using them to encode text pushed forward the state of the art for tasks such as sentiment analysis (Mikolov et al., 2013). In 2017, the paradigm shifted from the static embedding of individual words to neural networks that tracked relationships between sequential data, like words in a sentence. These *transformer models*, most notably the Bidirectional Encoder Representations from Transformers (BERT), again learned about the world by "reading" large numbers of documents (Devlin et al., 2019). Transformer models are used by popular search engines such as Google and Bing.

More recently, language models have been trained with *instruction tuning* that allows them to recognize inputs as instructions to follow, allowing them to generate responses for many possible tasks. These models base their responses on the information trained into the model during the language modeling phase, which now includes tens of terabytes of text (Ouyang et al., 2022). Instruction-tuned LLMs are what the public often means by "LLMs," and many competitors to ChatGPT exist, such as FLAN-T5, Llama2 and Falcoln. Finally, some LLMs like ChatGPT undergo *alignment* using reinforcement learning based on human-provided scores for how good humans think the outputs are. This further refinement, after instruction tuning, leads to more appealing outputs (OpenAI, 2022).

Some generative machine learning (ML) tools also operate on non-language data such as continuous-valued quantitative variables (pixel intensities, temperatures, chemical concentrations, sound waves, and so on), and qualitative data such as Boolean (TRUE/FALSE) variables or pseudo-quantitative values that may be ordinal but not strictly numeric, such as Likert scale responses to survey questions (---, --, -, neutral, +, ++, +++, etc.). When these models learn on more than one kind of input data, they are called *multimodal*.

Prior to ChatGPT, multimodal models including DALL-E and Stable Diffusion (Rombach et al., 2022) were already producing impressive imagery outputs in response to queries, and Generative Adversarial Networks (GANs) had demonstrated

"deep-fake" abilities to render sensory imagery that rivaled natural photographic imagery.

Some major application areas for LLMs include interactive assistants, correcting human-generated content (e.g., autocorrect), chatbots, natural language generation, formal language generation (e.g., writing computer code), natural language processing (e.g., named entity recognition), and a variety of other functions via plug-in tools that can ground LLM responses in evidence outside the textual prompt.

The recent success of ChatGPT and related tools are enabled by the availability of very large samples of text and other data from the Internet, as well as at-scale high performance computing (HPC) resources to train and test ML models on these large samples of data. Scraping and storing large data from the Internet and deploying at-scale computing resources involves great expense, which makes the development of such tools competitive. However, many of these pre-trained models can be downloaded and fine-tuned for specific applications using less resources than were originally needed to produce them (e.g., BioGPT, Luo et al., 2023). Models may be *fine-tuned* by retraining in-house on custom data sets, or they may be *retrieval-augmented* by exposing them to custom prompt data.

Another key ingredient for the effective use of ML tools is knowledge of the architectures and algorithms that are used to train on source data, and where appropriate to operate on source data at runtime. There is no full scholarly description of training on GPT 3 and after. While these methods are available for commercial use via paid API access, there is not full transparency on the reinforcement learning phase of training, and actual training data is proprietary.

What are the apparent strengths and weaknesses, so far?

ChatGPT 3.5 can make novel inferences and explain a scientific topic reasonably well when asked by a human. When asked to give citations for its responses, ChatGPT may provide some genuine and relevant citations, but it also 'hallucinates' other citations that cannot be found. What's remarkable is that the hallucinated citations seem to have plausible author names, plausible dates, plausible paper titles, and plausible journal names, all in the correct format. This calls out a serious weakness of such tools: they are currently not truth-making in their answers to queries, they are statistics-making. They lack the ability to measure the truth value of their outputs.

Another potentially serious weakness is the readiness of human users to engage in "automation bias" (e.g., Goddard et al., 2012) about what they see as output from these decision aid tools that are still perhaps in their adolescence. In other words, it may require vigilance on the part of humans not to believe everything that ChatGPT "thinks." Social media and politically divided mass media have already facilitated the erosion of a vulnerable public's ability to discriminate fact from fancy (e.g., Kavanaugh and Rich, 2018). The emergence of deliberate or inadvertent misinformation, artfully assembled by sophisticated AI tools, will likely amplify already substantial truth decay and reality distortion in vulnerable human populations. It may be incumbent on civil society to develop a metaphorical 'immune system' for detecting and correcting untruths as they arise in both naturally and artificially intelligent systems.

A recent evaluation of an early version of GPT-4 was performed by Microsoft Research (Bubeck et al., 2023). The evaluators had concerns about whether GPT-4 was trained on arrays of standard benchmark test questions for domain-specific cognitive performance, and so developed an *ad*

hoc testing approach similar to laboratory tests in cognitive psychology. They found that GPT-4 performed, anecdotally, at or above average human performance in a variety of tasks, although below human performance in general. It is worth noting that there is not a uniform level of human performance in any cognitive task domain to compare to machine performance, as human performance varies widely. The authors of this study speculated that GPT-4 may represent a first, fledgling example of artificial general intelligence (AGI). Developing principled ways of testing LLM performance in context is a research frontier with substantial drive from application needs.

A relevant quote from the Microsoft evaluation of GPT-4 highlights the current conundrum in evaluations: "A question that might be lingering on many readers' mind is whether GPT-4 truly understands all these concepts, or whether it just became much better than previous models at improvising on the fly, without any real or deep understanding. We hope that after reading this paper the question should almost flip, and that one might be left wondering how much more there is to true understanding than on-the-fly improvisation. Can one reasonably say that a system that passes exams for software engineering candidates...is not really intelligent? Perhaps the only real test of understanding is whether one can produce new knowledge, such as proving new mathematical theorems, a feat that currently remains out of reach for LLMs."

These emerging tools have mined a repository of encoded human knowledge that is vastly larger than any one person could ever encounter first-hand. That is part of what makes their performance so impressive. In some cases, the generated outputs appear to reflect the human knowledge that is encoded in the training data, with novel compositions and inferences that are not simply a reconstruction of source material.

Machine learning is warranted when it isn't practical to encode human knowledge explicitly as a set of rules in a computer program, but when there is sufficient evidence in recorded data for flexible equations to develop knowledge through training. The most basic kind of knowledge is a simple record of events, but most useful knowledge involves generalizations about regularities in the world. Most ML approaches today are *supervised*, meaning that they seek to learn some input-output relationship from a large number of usually human-curated or human-provided examples of input-output pairs. These input-output relationships may take the form of questions and answers using natural language, or sequences of natural language in which the first part of the sequence is treated as input and the last item is treated as output, but input-output relationships may take other forms.

The results of ML may perform with very low error, if the structure in the training data has low uncertainty, low ambiguity, and low noise. However, in many cases there is no way to reduce errors in performance because there isn't sufficient information in training data for any machine to perform without error. It is useful to study the performance failures of any product of ML, to get an intuition about what went wrong, and to determine whether failures resulted from properties of the data or properties of the machine.

***Recommendation**: Performance measures: Develop ways to meaningfully measure the performance of generative AI in task environments that have to date only been performed by humans. Cognitive science is still not mature in measuring various forms of human intelligence, and as such, ready-made metrics are imperfect and require advancement.*

What is biotechnology?

Biotechnology is the use of biological systems to produce a desired engineering outcome. This may involve deploying existing biological systems, modification of biological systems, or *de novo* invention of biological systems, to perform a task.

Early examples of biotechnology include the making of bread, cheese, yogurt, beer, and wine via natural fermentation of foodstuffs by microorganisms like yeast, or the use of herbal medicines like willow bark, which contains a chemical similar to aspirin. Later examples include the early use of vaccination, and the use of bread mold to develop antibiotics against harmful bacteria.

Today, biotechnology includes the deliberate engineering of drug-like compounds and vaccines; the use of microorganisms to produce useful compounds for food, fuel, and industrial feedstocks for manufacturing; as well as engineering microorganisms to digest oil spills at sea and other harmful substances like environmental toxins or bioweapons. Current aspirational efforts in biotechnology include engineering replacement tissues or organs for humans, growing meat in laboratory environments that is cruelty-free and healthy to consume, using bioproduction for manufacturing and construction, and a wide range of other applications.

Proteins may be engineered to act as enzymes to catalyze chemical reactions, or to provide structural or functional materials. Metabolic pathways may be designed in a process known as retro-biosynthesis to produce a desired chemical product in a sequence of stages from inexpensive feedstock chemicals, all performed biologically in engineered microorganisms rather than using conventional chemical synthesis. Cells and tissues may be engineered with desired properties, as well as large-scale phenotypes (traits that require many underlying cellular and biochemical actions to be realized), up to the level of the design of whole organisms.

Applications of emerging machine learning tools that are specific to biotechnology

Generative AI has recently had major impacts on the ability of researchers to understand and characterize biological systems. This impact primarily arose through the open-source release of the AlphaFold software package, which can accurately predict three-dimensional protein structures (Jumper et al., 2021) and, as of time of writing this document, has nearly 7,000 citations. Researchers had been attempting to solve the problem of protein structure prediction for decades with limited success and no universal solution. AlphaFold solves the problem by using a generative AI model trained across large databases of over 170,000 known protein sequences and 3D structures elicited through X-ray crystallography of proteins. AlphaFold's ability to accurately predict protein structures is now allowing researchers to leverage large amounts of genomic data, which can be obtained through high-throughput DNA sequencing, to determine the structures of proteins that are encoded in genomes.

While AlphaFold is widely believed to involve LLM components, the algorithm and architecture are proprietary. One especially useful feature of AlphaFold is its reporting of confidence levels from 0-100 for the position of each amino acid residue in its predictions. The confidence measure, called predicted local-distance difference test (pLDDT), is based on a mathematical energy model of local protein structure. This confidence measure allows calibrated trust in model outputs on a granular scale.

As an example of its utility, AlphaFold enabled researchers to identify a novel therapeutic for liver cancer in 30 days and after synthesizing only 7 compounds to test (Ren et al., 2023). Going beyond protein structure prediction, emerging

research is demonstrating that AlphaFold can be leveraged further to accurately predict protein-protein interactions (Wallner, 2023). This provides even greater possibilities for understanding and designing biological systems and components, including identification of potential zoonotic diseases (via characterization of host-pathogen protein-protein interactions) and protein therapeutic design. Along with AlphaFold, other researchers have started using generative AI for biological molecule design, including the design of protein scaffolds around a known protein active site (Wicky et al., 2022).

Other ML based tools are emerging for enzyme and metabolic pathway design (reviewed in Jang et al., 2022), including: the prediction of chemical precursors in metabolic pathways, using methods originally developed for machine translation of language, using generative transformers operating on string-based representations of molecules, and use of generative LLMs such as ProGen in the predictive design of novel proteins for 3-dimensional (3D) structure, enzymatic function, and gain-of-function mutations (e.g., Shroff et al., 2020; Jang et al., 2022; Madani et al., 2023; Buehler et al., 2023). A key question for any predictive design activity is experimental validation, with costs that scale with the inaccuracy or uncertainty of model predictions.

An example potential application applied to biotechnology is designing a metabolic pathway to synthesize the statin drug Lipitor (Furberg, 1999) from acetate (Manzoni and Rollini, 2002). Source material for training ML to perform such tasks may include scientific literature on drug design, biochemistry, and related fields, as well as databases on high-throughput screening of candidate drug molecules, known enzyme reactions (substrates, products, and enzyme macromolecules), chemical reactions described in patent disclosures, and metabolic pathways (e.g., the KEGG database). Queries to a conversational interface might consist of natural language descriptions as above: "Design a metabolic pathway to convert acetate to the statin drug compound Lipitor."

Global security concerns

Tools like AlphaFold and ProGen have some conceivable dual use. They could potentially be used to facilitate the hiding of known harmful epitopes within new protein structures or sequences to a degree sufficient to evade current sequence similarity-based DNA screening protocols. Accomplishing this hiding would likely require experimental evaluation of dozens to thousands of predicted sequences at present, but this is not known for certain and could change rapidly with further developments or refinements to these AI tools. The International Gene Synthesis Consortium is a group of commercial synthetic DNA providers who have agreed to best practices in biosecurity, including automatic identification of orders with signatures of malicious intent. However, evasion of current screening methods is theoretically possible. "One example of misuse would be to use a bio-design tool to identify protein-based toxins that are predicted to be functionally similar to known toxins but are otherwise different enough from those found in nature that traditional safeguarding measures would be ineffective" (Walsh, 2023).

Governance of chatbots has largely been focused on preventing users from employing AI to discover existing information that may be misused, for example a list of companies that will synthesize DNA but who have not agreed to scan orders for malicious intent, which does not involve the synthesis of any new information. Governance of biotechnology tools used to generate biological systems with desired function will also need to consider preventing users from producing harmful new information. The Tianjin Biosecurity Guidelines for Codes of Conduct for Scientists provides guidance on ethical practices in biomedical research, and governance of AI systems used in biotechnology applications may

seek to align with this guidance. Two pieces of legislation were introduced in the U.S. Senate in 2023, the Artificial Intelligence and Biosecurity Risk Assessment Act and the Strategy for Public Health Preparedness and Response to Artificial Intelligence Threats Act, based on alarms about the potential for generative AI to be used to design chemical and biological weapons. The problem-solving organization Helena convened a group of leaders in AI-enabled biology to develop recommendations to promote biosecurity in the near term, summarized in the report Biosecurity in the Age of AI (2023).

At the TEM, a discussion took place on the great power competition between the U.S. and China, considering biotechnology as a focus for strategic advantage. It was proposed that new generative AI tools created in the U.S. might give China the ability to monopolize entire sectors of the emerging bioeconomy, involving pharmaceutical manufacturing facilities and entire supply chains. The technical aspects of this risk are augmented by concerns about cybersecurity, economics, destabilization of work roles and purchasing power for individuals, trade, individual rights, intellectual property, and domestic and foreign policy. It was proposed that export controls aren't meaningful for technologies like LLMs, which once released are impossible to recall.

***Recommendation**: Global security: Track risks to global security that include disruption of employment, intellectual property issues, counter-WMD, and economic competitiveness of the U.S. and its allies.*

Applications of emerging machine learning tools to science in general, relevant to biotechnology

We have identified two major areas of desirable function in LLMs used in broader scientific research and engineering, with indispensable uses in biotechnology. One of these is in trustworthy and explainable outputs, in which systems can explain and justify what they generate by revealing aspects of the reasoning process used to generate their results, as well as reaching back accurately to supply authoritative literature citations or references to other source data. Another is in mapping knowledge in a domain (such as biotechnology), showing topics, their interrelationships, their emergence over time, and measuring knowledge gaps, controversy, uncertainty, or other lack of well-informed consensus and convergence of thought around specific topics.

Trustworthy and explainable results

A major topic in AI research today involves efforts to make AI responsible, trustworthy, interpretable, and explainable (e.g., NIST Trustworthy and Explainable AI). "Black box" systems, whether proprietary systems or multilayered neural networks that lack indicators of how they transform queries into responses, are in some ways as difficult to trust as a human being who "thinks from their gut" without any exposure of reasoning processes or evidence used to justify conclusions.

It was noted above that LLMs can hallucinate their output, including generating bogus scientific citations to support their answers. There have been recent efforts to provide vetted source evidence for the models to cite, and to expose chains of thought used to generate answers to scientific questions. Non-parametric memory is when information relevant to the user's question is included alongside text of the prompt; retrieval augmented generation (RAG) is when the relevant information is retrieved automatically, e.g. from a database of relevant papers. These techniques lead to more grounded and accurate outputs than using only the bare question (Lewis et al., 2020). Paper QA (Andrew White) is an example of a transformer-based LLM that advertises generating answers to questions, "…with no hallucinations, by grounding responses with in-text citations."

ScienceQA is a dataset that has been used to train LLMs to answer scientific questions via chains of reasoning that are exposed to explain how answers to scientific questions were derived (Lu et al., 2022). Chern et al. (2023) have introduced a methodology for assessing the factuality of LLM outputs in a range of domains, including scientific literature review. There are strong drivers to assess and improve this dimension of LLM performance.

It was noted above that current LLMs appear unable to measure the truth value of their own generated outputs, although efforts have been underway to evaluate and improve this status (e.g., Bohnet *et al.*, 2022; Gao *et al.*, 2023). This limitation results in part from a lack of grounding of models in sensorimotor experience with the real world, which is where the non-arbitrary truth value of any linguistic proposition originates (Dretske, 1983). Mental representations, whether natural or artificial, that have measurable truth value are those whose analytic representations approximate the structure and function of phenomena in the real world (like size and shape in scale models, such as paper models of floor plans and cut-out furniture used to plan room arrangements, Waskan, 2006). Language acts as a set of pointers to such "intrinsic" cognitive representations, which are made up of substantially non-linguistic content (e.g., Pearson and Kosslyn, 2015). Most words bear an arbitrary relationship to their environmental referents (exceptions are onomatopoeic words like "bang" or "pop", whose pronunciation approximates a natural sound, and thus have intrinsic representational content). This means that words and sentences do not contain truth, but rather provide pointer-like access to truth that lies in the approximate structural and functional isometry between nonlinguistic mental representations and phenomena in the real world.

A substantial hurdle in reasoning about real-world phenomena using language-like or otherwise symbolic representations is the "frame problem" (Janlert, 1996), which may be understood as an attempt to tile a continuous world with a patchwork of discrete and disconnected beliefs about how things work, or what to do to achieve a desired outcome. To accurately model a continuous world with discrete buckets of reasoning, one may be required to make the buckets smaller and greater in size in a manner that makes learning and reasoning scale intractably.

For generative AI models to reason correctly about the real world and evaluate non-arbitrary truth, they will need more than just linguistic data to train on, and more than just linguistic representations to reason with (consider Bender and Koller, 2020). Emerging multimodal models are a frontier in this pursuit.

***Recommendation**: Multimodal training: Continue to advance the research frontier of training multimodal models on biotechnology data, to include research literature (natural language, tables, and figures), gene sequences, protein 3D structure, reaction rates and binding affinities for enzymes and their products, biochemical pathway graphs, and physiologic and phenotypic variables at system and organism scales, among other modalities of data, with appropriate cross-references. Multimodal models are regarded as critical to realizing the full range of application needs in biotechnology, as language-only models have inherent limitations for representation, prediction, and causal inference.*

In scientific inquiry, much of the time we are interested in causal processes and evaluating the truth value of competing hypotheses about causes. Recent DARPA programs have sought to ingest causal models from scientific literature and scientific databases, including the now-completed Big Mechanism program and the resulting INDRA (Integrated Network and Dynamical Reasoning Assembler) system for drug development.

Judea Pearl and colleagues (e.g., Pearl, 2009) have described a "causal ladder" for being able to make successively stronger claims about causal processes from available evidence. The lowest

level with the weakest claims is *prediction*, which only involves a joint distribution of measurements of variables. If you know something about one or more variables, you can make educated guesses about other jointly distributed variables. This situation is typical of observational scientific studies (e.g., "Building damage often occurs during earthquakes."). The intermediate level allowing stronger claims is *intervention*, in which an agent can force some variables into desired states, and then measure what other variables do in response. This is typical of experimental scientific studies and makes it possible to develop models of the mechanisms underlying observed constraints among measured variables (e.g., "Forces typical of earthquakes can cause damage to buildings."). The level affording the strongest claims about causality is *counterfactuals*, in which we imagine what might have happened if certain conditions were met (e.g., "If there had been no earthquake, this building would not have been damaged."). This level requires models that capture mechanistic dependencies among variables to reason accurately. AI tools that are intended to offer synthetic conclusions about causality in biotechnology and other fields may benefit from these theoretical insights.

***Recommendation**: Enhance trustworthiness and facticity of models by training on truth value: Truth value can be trained using consensus or "gold standard" human curation, marking of authoritative sources prior to training and allowing verbatim citations of training data where appropriate, providing linguistic representations with sensory and/or motor meaning using multimodal models, training models to expose the chains of reasoning used to generate outputs, and development of intrinsic cognitive models that seek to optimize scale-model likeness of structure and function with the phenomena under study for accurate prediction and causal inference.*

Mapping knowledge in a domain

A critical part of science and engineering labor includes understanding existing knowledge, and targeting investigations to discover the most useful new knowledge. Scholarship involves reading and understanding published literature, which involves learning about topics and concepts within a domain and their interrelationships.

Some formal instruments are used in scholarship. *Bibliometrics* is the measurement and analysis of scholarly literature, and *scientometrics* is its application within the domain of science. *Ontologies* are formal characterizations of topics and their relationships within a domain. *Taxonomies*, widely used in biology to represent categorical membership, for example among biological species or of evolved proteins, are a kind of ontology with strict "is-a-member-of" hierarchical structure. Taxonomies and ontologies more broadly evolve over time. For example, enzymes, or macromolecules with biochemical catalytic properties, were once thought to be exclusively proteins. The discovery of RNA molecules with catalytic properties, called ribozymes, upended this assumption, and ontologies needed to be revised.

Generative AI may be useful in mapping and tracking knowledge trends in a domain, including developing dynamic domain-specific ontologies and topic-specific maps. The Medical Subject Headings (MeSH) is a National Institutes of Health (NIH) maintained and professionally curated domain ontology for biomedicine and related topics. While updating MeSH with new concepts and extending old concepts with new variations happens regularly, the field of biotechnology evolves more rapidly than these updates. If generative AI can accurately generate or extend current ontologies through accurate identification of emergence of new concepts, or the branching and expansion of previous areas of study, then the mapping of knowledge in a particular domain

could be greatly simplified. Mapping of laboratory or observational data to logical and hierarchical knowledge graphs can advance the activities of discovering trends in concepts and topics over time. This may result in identifying existing competing research hypotheses and generating novel hypotheses to address, and measuring uncertainty, controversy, and knowledge gaps worthy of targeted new research.

Mapping knowledge in biotechnology requires adequate training data, and effective data preparation. More and better training data results in better AI predictions from tools like BioGPT (Luo et al., 2023). Extracting and combining datasets from the literature is expected to be a challenge due to non-standard formats and incomplete datasets in publications, especially for negative results. Better AI predictions for protein and metabolic pathway designs would be expected with AI tools that could read data in multiple figure formats, expanded opportunities for authors to release larger data sets, and standards for dataset formatting and content-addressable access.

Recent advances in generative AI have required data and computational time at scale to generate connections and relationships in a neural net sufficiently large to respond to questions of a general nature. These requirements for the development of models on par with ChatGPT are outside the data gathering and computational capabilities of many organizations. For generative AI deployments that are intended to be stand-alone *de novo* in organizations with limited staffing and compute resources, the relationships between concepts could either be generated or referenced from prebuilt concept models such as ontologies or taxonomies to avoid large costs.

It is possible that ontologies, perhaps curated by subject matter experts (SMEs) who can define relationships between focused concepts, could be leveraged to develop domain or task-specific generative AI, without needing to have access to the world's data and huge amounts of computation. The various formats of knowledge graphs, such as MeSH, that are available today could potentially reduce the volume of the unstructured training data necessary to determine concept relationships and potentially to reduce errors. For example, OpenAlex ("an open and comprehensive catalog of scholarly papers, authors, institutions, and more"), is a scientific knowledge graph of millions of books and papers and relationships between authors, institutions, and about 65,000 Wikidata concepts, produced via an automated hierarchical multi-tag classifier (Priem et al., 2022).

Ontologies and other hierarchical knowledge representations, implemented for example using the Web Ontology Language (OWL), are broadly used to provide context to non-ML algorithms for mapping knowledge on the Internet and the Internet of Things. These logical representations are easily parsed and used by non-ML algorithms, which have a much lower computational burden than at-scale ML approaches. Representations of an LLM's logic could perhaps be evaluated through the generation of hierarchical knowledge representations, and simplify the process of evaluating an LLM for trustworthiness and explainability. Once an LLM creates a hierarchical representation of data or processes, the evaluation of these logic structures could be performed by non-ML processes.

The above discussions of cost and efficiency in developing domain-specific LLMs may already be out of date. A recently published document advertised as an internal Google memo (Google, 2023) suggests that the massive investments of big tech in gathering data and training using high performance computing at scale may be unnecessary to achieve beneficial results. Further, open-source AI/ML methods may outpace at-scale efforts by big tech, which may level the playing field substantially. "Many of the new ideas are from ordinary people. The barrier to entry for training and experimentation has dropped from the

total output of a major research organization to one person, an evening, and a beefy laptop" (ibid.).

The TEM attendees also discussed prospects for conversational scientific assistant software that can interpret multimodal data and engage in exposed reasoning and question answering in collaboration with human researchers.

***Recommendation**: Develop knowledge-mapping models and conversational scientific assistants: Adapt current AI tools to track trends in knowledge in biotechnology literature, and to identify controversies, uncertainties, and hypotheses worth investigating. Develop conversational assistant software to engage in dialogue with human users to answer queries and to collaborate on research activities. Develop standard formats for publication contents that facilitate multimodal machine learning of biotechnology and other scientific data, and cross-referencing of content in publications with scientific databases.*

High performance computing and related algorithm development

A discussion at the TEM highlighted the coevolution between dedicated high performance computing hardware and tailored ML algorithms designed to make best use of that hardware, with specific use cases in biotechnology.

***Recommendation**: Promote compute capabilities to foster innovation: Adapt both high performance computing (HPC) resources and algorithms in synergy to allow democratization and commoditization of research into novel capabilities.*

Conclusions

ChatGPT and related emerging ML technologies offer substantial value in their ability to compose novel content, to engage in meaningful and informative dialogue with a human, and to explain ideas. Currently, such systems have substantial failure modes, which in the domain of scientific inquiry may emphasize hallucinating content that is not correct, but that is deceptively plausible in its persuasive style and format. Current LLM-based tools require substantial human supervision and systematic doubt for them to be of correct use in science and technology applications.

Specific application areas discussed include:

- Predictive design of useful biological systems that scale from molecules to organisms,
- Trustworthy and explainable results of AI outputs, including exposed chains of reasoning and references to scientific source literature or other forms of evidence,
- Exploratory mapping and tracking of trends in knowledge in the biotechnology field, and
- Development of conversational AI research assistant software.

Concerns exist about expensive resource allocation issues in developing these tools. Open-source tools are rapidly achieving performance levels that compete with at-scale big tech efforts, which may have a leveling effect on the ability of smaller players to achieve desired results.

Finally, linguistic representations have inherent limitations, in that language does not embody intrinsic cognitive modeling of the structure and function of phenomena in the world, which is the basis of truthful conclusions about those phenomena. Multimodal generative AI systems are thus regarded as a significant area of

investment in research and testing for applications in a wide range of domains, including biotechnology. Principled ways of testing the most meaningful dimensions of AI performance are a major focus of current efforts, and depend to a degree on human understanding of which dimensions are most meaningful.

About the author

Jeff Colombe is Principal Scientist at the MITRE Corporation. He has a bachelor of science in biomedical engineering and a doctorate in neurobiology. He has been involved in AI/ML, biotechnology, and cognitive science research and development for over 20 years, supporting federal agencies including NGA, DARPA, IARPA, ARPA-H, ONR, Army MRDC, Army USAMMDA, DIA, NSA, DOJ, NIH, NSF, and most recently has supported upskilling of staff at the CDC in data science inclusive of AI/ML. Outside of work, he has mentored over 60 high school and college students in STEM subjects.

References

Bender EM and Koller A (2020) Climbing towards NLU [Natural Language Understanding]: On meaning, form, and understanding in the age of data. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5185–5198. July 5 - 10, 2020.

Bohnet B et al. (2022) Attributed question answering: Evaluation and modeling for attributed large language models. arXiv:2212.08037v2 [cs.CL] Last revision 10 Feb 2023.

Bubeck S et al. (2023) Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv:2303.12712v3 [cs.CL] 27 Mar 2023.

Buehler MJ (2023) Generative pretrained autoregressive transformer graph neural network applied to the analysis and discovery of novel proteins. Journal of Applied Physics 134: 084902

Chern IC et al. (2023) FacTool: Factuality detection in generative AI, a tool augmented framework for multi-task and multi-domain scenarios. arXiv:2307.13528v2.

Devlin J et al. (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186.

Dretske F (1983) Knowledge and the Flow of Information. MIT Press.

Furberg CD (1999) Natural statins and stroke risk. Circulation 99:185-188.

Gao L et al. (2023) RARR: Researching and revising what language models say, using language models. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers, pages 16477–16508. July 9-14, 2023.

Goddard K et al. (2012) Automation bias: a systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association 19(1): 121–127.

Google (2023) "We have no moat, and neither does OpenAI" [Google internal memo]

Jang WD, Kim BG, Kim Y, and Lee SY (2022) Applications of artificial intelligence to enzyme and pathway design for metabolic engineering. Current Opinion in Biotechnology 73:101-107.

Janlert L (1996) The frame problem: Freedom or stability? With pictures we can have both. In: The Robot's Dilemma Revisited: The Frame Problem in Artificial Intelligence (Theoretical Issues in Cognitive Science). Ford K and Pylyshyn Z, Eds. Ablex.

Jumper J et al. (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596:583-589.

Kavanaugh J and Rich MD (2018) Truth decay: An initial exploration of the diminishing role of facts and analysis in American public life. Rand Corporation Research Report 2314.

Lewis P et al. (2020) Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv:2005.11401v4 [cs.CL] Last revised 12 Apr 2021.

Lu P et al. (2002) Learn to explain: Multimodal reasoning via thought chains for science question answering. The 36th Conference on Neural Information Processing Systems (NeurIPS), 2022.

Luo et al., (2023) BioGPT: Generative pre-trained Transformer for biomedical text generation and mining. arXiv:2210.10341v3 [cs.CL] 3 Apr 2023.

Madani et al. (2023) Large language models generate functional protein sequences across diverse families. Nature Biotechnology.

Manzoni M and Rollini M (2002) Biosynthesis and biotechnological production of statins by filamentous fungi and application of these cholesterol-lowering drugs. Applied Microbiology and Biotechnology 58(5):555-64. doi: 10.1007/s00253-002-0932-9. Epub 2002 Feb 14.

Mikolov T et al. (2013) Efficient estimation of word representations in vector space. arXiv:1301.3781v3 [cs.CL]. Last revision 7 Sep 2013.

Ouyang L et al. (2022) Training language models to follow instructions with human feedback. arXiv:2203.02155v1 [cs.CL]. 4 Mar 2022.

Pearl J (2009) Causality. Cambridge University Press.

Pearson J and Kosslyn SM (2015) The heterogeneity of mental representation: Ending the imagery debate. Proceedings of the National Academy of Sciences 112(33):10089-92.

Priem J et al. (2022) OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv:2205.01833v2 [cs.DL] Last revision 17 Jun 2022.

Ren F et al. (2023) AlphaFold accelerates artificial intelligence powered drug discovery: Efficient discovery of a novel CDK20 small molecule inhibitor. Chemical Science 14:1443–1452.

Rombach et al. (2022) High-resolution image synthesis with latent diffusion models (a.k.a. LDM & Stable Diffusion). Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2022, pp. 10684-10695.

Shroff R et al. (2020) Discovery of novel gain-of-function mutations guided by structure-based deep learning. ACS Synthetic Biology 9(11):2927-2935.

Wallner B (2023) AFsample: Improving multimer prediction with AlphaFold using aggressive sampling. bioRxiv preprint.

Walsh ME (2023) Why AI for biological design should be regulated differently than chatbots. Bulletin of the Atomic Scientists. 1 Sep 2023.

Waskan J (2006) Models and Cognition. MIT Press.

Wicky BIM et al. (2022) Hallucinating symmetric protein assemblies. Science 378(6615):56-61.