Technical Exchange Meeting: "MITRE Generative AI in Biotechnology Roundtable"

April 26, 2024

Introduction

The MITRE Corporation hosted a Technical Exchange Meeting titled, "MITRE Generative AI in Biotechnology Roundtable" on Friday, April 26, 2024, in McLean, VA. Approximately 80 in-person and 80 virtual attendees included personnel from the U.S. Government, academic research institutions, biotechnology industry laboratories, and federally funded research and development centers.

Topics of interest included biosecurity, the economic competitiveness of the U.S. and its allies, AI-enabled advances in medicine, biomanufacturing, and synthetic biology (small molecules, proteins, pathways, tissues, organisms), and using conversational LLM interfaces to augment human performance.

Synthesis and summary of findings

Artificial intelligence for biotechnology tools (AIBTs) currently exist in two relatively different forms, namely, large language model (LLM) based chatbots like ChatGPT, and mostly non-linguistic AI tools that apply principally to data from biological experiments. The latter tools include gene and protein sequences, 3D positions of atoms in proteins, enzyme-mediated biochemical reactions and associated biochemical networks and pathways, pathogen mutation and evolution, and human-curated knowledge representations like taxonomies and relational databases.

LLM-based chatbots

LLM-based chatbots are proving useful to scientists in accessing and digesting collections of published literature that are far too large for humans to access, discovering topics, tracking trends, summarizing document collections and topics at varying degrees of granularity and detail, identifying relationships among topics, identifying uncertainty and controversy, and generating useful hypotheses to test using laboratory experiments.

LLMs are rapidly evolving in the quality of their replies to human questions and requests due to advances that include fine tuning and retrieval augmented generation (RAG), and in training on multimodal datasets beyond text-only data. The latter is especially important, as it allows direct problem-solving, for example, design of proteins with desired properties, rather than being limited to conversations about problem-solving.

An important emerging use of LLMs in science and technology is as a natural language interface to the remaining non-linguistic tools. Chatbots can "raise the floor" for users of any level of expertise, explaining technical topics in plain language, and giving intuitive and user-friendly instructions for wielding technical resources such as biotechnology wet laboratory equipment or programming languages. Andrej Karpathy of Stanford University was recently quoted as saying, "The hottest new programming language is English," because tools like CodeGPT can take human instructions and write well-formed and fully functioning code to meet human user requests. The same kind of translation capability applies to complicated lab equipment. Chatbots are being trained to manage equipment interfaces that might otherwise require a human user to climb a steep learning curve.

Methods for evaluating performance in LLMs are steadily developing, with an expanding set of qualitative and quantitative measures reflecting performance in information retrieval (e.g., accuracy, relevance), risk analysis and security (privacy, malicious use), human values (fluency, demographic parity), and system usability (interpretability, explainability).

The potential for malicious use of LLMs is currently relatively low, as studied by formal red-team exercises, but risks for malicious use of LLMs may increase, and such increases should be monitored.

(Mostly) non-linguistic AI-Bio Tools:

Non-LLM AIBTs exist in wide variety, some created by commercial entities, and many developed by academic research labs. DeepMind's AlphaFold has had a significant impact on a long-standing hard problem in science, which is predicting the 3D structures of proteins from their 1D amino acid sequences. This opens the door to better understanding the functions of existing proteins, rational design of new proteins, prediction of how proteins will interact with each other, and optimizing the use of proteins as enzymes to catalyze chemical reactions for medical and industrial uses. Other tools are emerging for retrobiosynthesis of desired small drug candidate molecules (or, in a dual use scenario, nerve toxins) using biochemical pathways mediated by known, enhanced, or novel enzymes.

Non-LLM AIBTs are in clear need of datasets for training and testing machine learning models that are greatly improved in quality, scope, scale, and coverage of problem domains. One speaker named a "wish list" for training and testing biotechnology AI tools that included known proteins physically mutated at each amino acid position, for each of the 20 amino acids, then 3D imaged (for example using cryo electron microscopy rather than the more expensive and finicky X-ray crystallography technique).

The potential for malicious dual use of non-LLM AIBTs remains relatively low because of the low prediction accuracy of the tools (10-20%) and the resulting need for extensive skilled laboratory validation. Disgruntled skilled users such as graduate students, or well-funded nation state actors, pose a risk of being enabled by existing AIBTs. The accuracy and enabling capability of AIBTs is steadily increasing, and risks of these tools enabling malicious uses should be closely monitored and countermeasures discussed in the community of tool providers and government agencies with responsibility for protecting the public.

Self-driving laboratories

Toolsets including both LLMs and non-LLM AI tools are being assembled into automation workflows and pipelines, connecting in loops to provide laboratory-as-a-service and data-as-a-service models. A target for such automation is the "self-driving laboratory" (by analogy to self-driving vehicles) that can take natural language queries from human users and then specify and implement entire iterative loops of ideation (identifying hypotheses to test and theoretical models to inform), experimentation, problem reformulation, and refined hypotheses and experiment design. With a sufficient degree of autonomy, automated systems could become relatively independent of human intellectual capital and manual labor. Governance, quality assurance, and safety and security issues arise with increasing decoupling from the need for human skill and participation.

Conclusions and next steps

Participants engaged in directed discussion of topics of interest for investments in targeted research, collaboration, and future technical exchanges. Recommendations were as follows.

- Investments should be made in "self-driving laboratories" (by analogy to self-driving vehicles) and integration of components of emerging automation pipelines
- LLM-based natural language interfaces should be integrated with deep technical resources such as laboratory equipment and procedures, computer programming, and data analytics to "raise the floor" of accessibility for users of all levels of expertise, including novices
- LLMs should be fine-tuned with domain specific content and value-tuned for specific niche technical communities of interest, such as biotechnology, or science and technology audiences more broadly
- U.S. and allied governments should invest in data sources that are developed specifically to train and test AI systems in biotechnology, with a focus on data quality and coverage of design spaces
- Pooled laboratory experiments should be performed to go beyond simple high throughput screening
- Government and industry should reassess their complementary roles for maximum public benefit in public-private partnerships (e.g., government-provided incentives for vaccine development by private companies)
- The research and security communities should pursue deeper involvement in understanding potential dual and malicious uses of off-the-shelf tools
- Screening of DNA sequence orders for potential malicious use should be based on predicted 3D structure and biochemical function, to go beyond simple 1D base pair / amino acid sequence similarity to known harmful agents

About MITRE

MITRE is a not-for-profit corporation that manages several Federally Funded Research and Development Centers (FFRDCs), chartered by Congress to serve the public interest. MITRE has a long-term relationship with government sponsor organizations with deep institutional knowledge and a commitment to mission success. MITRE is not a vendor, and does not manufacture or sell products. Our job is to help you successfully navigate technical risk and organizational change.

To inquire about uses of artificial intelligence in biotechnology, please contact bio@mitre.org.

MITRE | SOLVING PROBLEMS FOR A SAFER WORLD®

Appendix A: Meeting agenda

- 8:00 – 8:30 Registration and light breakfast
- 8:30 – 8:45 Introductory remarks, Dr. Jeff Colombe, Life Sciences, MITRE, and Dr. Chris Fall, Vice President of Applied Sciences, MITRE
- 8:45 – 9:30 Large Language Models: Capabilities, Trends and Assurance, Dr. Ben Wellner, MITRE
- Threat Perspective
  - 9:30 – 10:30 REPORT: Informing Threat Awareness at the Nexus of Artificial Intelligence and Biotechnology, Dr. Alex Tobias, Life Sciences, MITRE
  - Industry
    - 10:30 – 11:00 Assessing the Impact of Artificial Intelligence on the Deliberate Biological Threat Landscape, Dr. Matt Walsh, Johns Hopkins University Center on Health Security
    - 11:00 – 11:30 The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study, Dr. Caleb Lucas and Dr. Chris Mouton, RAND Corporation
- 11:30 – 1:00 Lunch and networking
- Science
  - 1:00 – 1:30 AI in Protein Engineering: Progress, Realities, and Hype, Dr. Raghav Shroff, Houston Methodist Research Institute
  - 1:30 – 2:30 AI for Molecular and Chem/Bio Synthesis Design, Dr. Connor Coley, Massachusetts Institute of Technology
  - 2:30 – 3:00 Risk Estimation and AI-Enabled Protein Design, Dr. James Diggans, Twist Bioscience
  - 3:00—3:30 Lab Data as a Service, Dr. Susan Buckhout-White, Ginkgo Bioworks
- 3:30 – 4:00 Structured discussion session
- 4:00 – 5:00 Networking session

Appendix A: Meeting agenda

"Large Language Models: Capabilities, Trends and Assurance" —Dr. Ben Wellner, MITRE

This presentation sought to review the history and way forward for progress in the use of large language models (LLMs) in AI. The discussion was not specific to uses in biotechnology. Dr. Wellner gave an overview and history of LLMs, which are probability distributions over word sequences. LLMs are trained to estimate probabilities of specific word sequences, to generate word sequences according to their learned distributions, and interact with human users by exchanging word sequences.

Key points:

- Recent innovations in LLM training and performance were reviewed (generative pre-training, instruction tuning, value tuning, bidirectional encoding and autoregressive decoding, fine-tuning, LLMs trained for planning and tool use applications, and retrieval-augmented generation)
- LLMs are being used for sensemaking across document repositories that are too large for humans to read directly
- Performance is sensitive to learned representations (called embeddings) for text content and meaning
- Automatic creation of synthetic training data can be a solution to expensive training data curation by humans (an example is inversion of existing statements to form questions and their answers)
- Research seeks to improve performance in sensemaking across corpora of different genres with different ways of expressing similar topics (example is patent literature vs. scientific publications)
- Quality assurance for LLM performance is based on families of quality metrics (information-retrieval, usability, risk-analytic, and other measures of quality), and is an active area of research
- Existing public datasets for training and evaluation, and the pretrained models that result, may not capture your domain needs
- Model evaluation may benefit from model-based-evaluation, in which one LLM trained on human expert evaluations may be used to evaluate another LLM's performance
- Identification of the labor-intensiveness of steps in LLM evaluation can help to combine automation with human labor for maximum cost-benefit advantage

"Report: Informing Threat Awareness at the Nexus of Artificial Intelligence and Biotechnology" –Dr. Alex Tobias, Life Sciences, MITRE

This presentation reviewed the results of a study performed by MITRE to assess malicious dual uses of AIBTs. Dr. Tobias reported that the potential dual-use nature of AIBTs for malicious uses is evident. AIBTs are diverse and augment differently from more broadly used AI models like ChatGPT. The real-world limitations for malicious dual use of AIBTs include the need for tacit domain knowledge on the part of users, access to laboratory resources and reagents, the ability to perform technically difficult experiments or laboratory procedures, and the current lack of available training data to support highly effective AI models. Policy-makers are racing to catch up with rapid progress in AIBT development, with examples including the Executive Order of October 2023, and the S.2399 AI & Biosecurity Risk Assessment Act.

Key points:

- In the study detailed in the report, three types of biothreats were considered: "traditional" biothreats such as weaponized naturally occurring anthrax or plague, "enhanced" biothreats in which natural organisms are modified for increased effect, and "novel" biothreats involving de novo designed agents
- AI tools for biotechnology assessed for their malicious use included LLM chatbots and AI-based nonlinguistic tools like AlphaFold and Rosetta
- LLMs tend to be closed and accessed via API, and while current models can be induced to provide information on how to create bioweapons, they did not prove much more effective than internet search to find the same answers
- Non-LLM AIBTs often required advanced technical skills to install and use, and advanced laboratory skills to implement and validate
- As of February 2024, OpenAI's Risk Preparedness Framework for malicious use of GPT rated the risk of acquisition of CBRN (chem/bio/radiologic/nuclear weapons of mass destruction) information to be low
- Tools were evaluated for malicious applications of protein folding / design / interactions; enzyme pathway design; and directed viral mutation and evolution
- Profiles for malicious actors were assessed: the "do-it-yourself" science dabbler or terrorist or non-state actor with little technical depth; the "disgruntled graduate student" with high technical depth and limited supervision but limited resources at scale; and the "nation-state" with high technical depth and deep resources
- Assessments of the risks of AIBTs include measuring their accessibility, usability, data requirements, compute resource needs, interpretability of results, accuracy and validity, biosecurity concerns, and model development and provenance
- Safeguards, mitigation strategies, indicators, and warnings were described

"Assessing the Impact of Artificial Intelligence on the Deliberate Biological Threat Landscape" –Dr. Matt Walsh, Johns Hopkins University Center on Health Security

This presentation reviewed the results of a study performed by the Johns Hopkins University Center on Health Security to assess malicious dual uses of AI-Bio Tools. Dr. Walsh described an emerging AIBT chain that has an LLM-based chatbot-enabled conversational user interface (e.g., ChatGPT), that connects into non-linguistic AIBTs (e.g., AlphaFold), which in turn connects to laboratory robotics and automation systems. Human users are approaching the ability to query and program the entire chain using natural language, and leverage analytics for predictive design, and then self-driving laboratories for automated implementation, validation, and industrial production. The benefits of such tools include greatly increased utility of results, delivered faster and for less cost. Dr. Walsh called for increased clarity in the purpose of assessments that measure the impact of AI on the deliberate biological threat landscape to better design and employ assessment methods.

Key points:

- Considerations of AIBTs for misuse include directly causing harm to humans, animals, or plants (e.g., crops) and man-made objects
- Malicious actors may include lone wolves, non-state groups, and nation-states
- Indirect effects might include the ability to discriminate against individuals on a genetic basis, and to use biology in a manner inconsistent with established norms (e.g., editing of germ lines in human populations)
- A bioweapons kill chain was described that includes acquisition, production, weaponization, and deployment/delivery (not all of which are always required to produce damaging results)
- The combinatorial space of Misuse Scenarios x Components (types of actors, sophistication, kill chain steps, and design-build-test-learn) x AI may involve uncertainties and gaps, which may interfere with fully evaluating the impact of AI on the biological threat landscape
- The Weapons of Mass Destruction Proxy (WMDP) benchmark is a curated dataset with 3,668 multiple choice questions relating to hazardous knowledge in biosecurity, cybersecurity, and chemical security (https://wmdp.ai). "WMDP serves as both a proxy evaluation for hazardous knowledge in large language models (LLMs) and a benchmark for unlearning methods to remove such knowledge."
- OpenAI developed an early warning system for LLM-aided biological threat creation involving a test with human users divided into expert and student groups, half of each using LLM tools and half using unaided internet searches to answer biosecurity-relevant questions

"The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study" –Dr. Caleb Lucas and Dr. Chris Mouton, RAND Corporation

This presentation reviewed the results of a study performed by The RAND Corporation to assess malicious dual uses of AIBTs. Dr. Lucas described a red-team study of the potential for misuse of AIBTs to plan and execute a bioattack. A red team study involves game-theoretic assignment of nefarious motives to human volunteers when assessing the enabling capability of tools to achieve an outcome.

Key points:

- In this study, human actors were divided into groups based on level of expertise, and each group was divided into sub-groups ("cells") that had unaided internet access versus also having access to an LLM-based chatbot
- The findings were that bioattack planning assistance is beyond the capability of the LLM tested, that LLMs produce unfortunate outputs that are not substantially better than unaided internet access, and that current LLMs may be inefficient for attackers due to the time needed to verify the information that they provide
- Each cell was asked to produce an operational plan (OPLAN) for a bioattack using the expertise and resources available to them, which was then evaluated via expert adjudication for biological and operational feasibility
- LLM chat logs and internet searches were evaluated to expose the reasoning used to seek information, and human participants were interviewed to further elicit their reasoning in their use of internet and (if available) LLM
- A caveat of the study was that only "unimodal" or language-only LLM based chatbots were used, meaning that they did not have access to anything other than natural language training data
- Another caveat was that human participants did not work in wet biotechnology laboratories
- OPLANs developed during the exercise were not deemed sufficient to enact a bioattack

"AI in Protein Engineering: Progress, Realities, and Hype" –Dr. Raghav Shroff, Houston Methodist Research Institute

This presentation described ways that AI-Bio Tools are being used to engineer protein molecules with desired properties. Dr. Shroff described qualitative dimensions of protein design that included affinity and selectivity for binding targets, stability and solubility, viscosity, aggregation, and deimmunization.

Key points:

- Engineered proteins that are already commercially deployed were discussed, including the mRNA COVID-19 vaccine that required a solitary spike protein to be stabilized using two directed mutations, and a plastic-digesting enzyme
- Protein design usually occurs using iterative cycles of predictive design and laboratory validation, but the ultimate goal of research is de novo design of proteins with little or no need for laboratory validation and iterative experimentation
- Deep learning paradigms for protein engineering include, 1) predict the 3D structure given a 1D protein amino acid sequence, 2) predict the 1D amino acid sequence that best fills in a missing part of a 3D protein structure, 3) predict sequences given a specific condition
- AlphaFold has been pivotal in protein engineering, allowing achievements in about 1 hour for what used to require an entire Ph.D. thesis, and has accurately modeled protein-protein interactions and how protein complexes form, expanded the number of valid protein sequences by 10x, and been used to accelerate new therapeutics
- Applications in rapid vaccine development were discussed
- Cryo electron microscopy (CryoEM) may be used to validate 3D protein structure prediction models without a full X-ray crystallography process under certain conditions
- Applications in de novo antibody design were discussed
- Current results show promise, but are still very limited
- Training data is limited and needs investment, with a "wish list" example given of mutational scanning data in which every residue in the protein is mutated to each of the 20 amino acids

"AI for Molecular and Chem/Bio Synthesis Design" –Dr. Connor Coley, Massachusetts Institute of Technology

This presentation described the use of AIBTs in the design and biological synthesis of molecules with desired properties. Dr. Coley described a research program that involves AI for synthetic organic chemistry, medicinal chemistry, and analytic chemistry, with evolving foundational capabilities that include chemistry-tailored neural network architectures for molecular representation, data sharing to facilitate modeling for chemistry and drug discovery, and autonomous chemistry laboratories for molecular and reaction discovery.

Key points:

- Methods were described for planning the retrobiosynthesis of desired small, drug-like molecules
- Currently, AI and novice chemists need every step to be specified, whereas expert chemists are capable of filling in missing details
- Synthesis constrains the set of small molecules that can be considered at all, and has influence on what can be accessed easily, in the larger space of desirable chemical properties
- Computer-aided retrosynthesis for drug-like compounds works well via data-driven programs; small, highly functionalized compounds present a challenge because they look quite different from the training data (e.g., fentanyl = easy, nerve agents = harder)
- Synthesis planning beyond retrosynthesis is necessary when testing novel routes experimentally; models can't predict reaction conditions with sufficient precision (yet)
- De novo generative design still requires good property prediction models to be "steered" effectively; generative AI helps with sampling but not necessarily scoring
- Synthetically constrained enumeration is straightforward to do in a coarse manner, requires no synthetic creativity, and can easily produce billions of hypothetical structures
- AI/ML is lowering the expertise requirement by improving the accessibility of information, even if the information is public knowledge
- Important considerations in open access v. open source AIBTs: ability to monitor usage, and ability to remove basic safeguards, e.g., filters on allowable molecules for which to design syntheses

"Risk Estimation and AI-Enabled Protein Design" –Dr. James Diggans, Twist Bioscience

This presentation described uses of AIBTs to design proteins with desired properties, and efforts to screen DNA synthesis orders for potential malicious use. Dr. Diggans described innovations in massively parallel DNA synthesis at Twist Bioscience, and reviewed several advances in the broader field, including assessments of gene-length homology, pathway constructs, grouping functional domains, and AI assisted construct design.

Key points:

- Order and customer screening for synthesized DNA at Twist Biosciences was described
- Customer screening verifies legitimacy with respect to an institution and identity with respect to an individual:
- Is the customer or institution on any lists maintained by the Departments of Commerce, State, and Treasury?
- Is the customer licensed to carry out work on regulated pathogens?
- Ensure the shipping address is not a P.O. box or private residence
- Sequence screening assesses the potential risk of orders:
- Does the sequence pose a significant dual-use risk?
- Is the sequence a 'best match' to a regulated bacterial or viral pathogen?
- Is the sequence legal to manufacture and ship within the United States?
- Is the sequence from a gene that can endow or enhance pathogenicity?
- Does the sequence require an export license to ship overseas?
- The U.S. Government provided guidance in 2010 and 2023, and the International Gene Synthesis Consortium (IGSC) has 35+ members who voluntarily screen DNA orders
- "Only models will defend against models", as screening for biothreat agents must move toward functional analyses
- A challenge is developing tools to detect malicious designs, that are then the best tools to create malicious designs

"Lab Data as a Service: Enabling BioAI" –Dr. Susan Buckhout-White, Ginkgo Bioworks

This presentation reviewed uses of AIBTs to automate discovery and validation in biotechnology, and to provide usable data through automated high throughput laboratory experiments. Dr. Buckhout-White described the BioAI space as having specific challenges and concerns.

Key points:

- Unique challenges in the BioAI space include data dimensionality, data quality, and data availability
- Unique concerns in the BioAI space include ethics and privacy, bias and fairness, massive computation costs, and technology misuse
- Ginkgo Bioworks has pioneered an end-to-end synthetic biology pathway, that includes computational enzyme design, discovery, and improvement, strain prototyping and optimization, high-throughput cell engineering and screening, and downstream processing and manufacture scale-up.
- Ginkgo actively develops automation systems in-house that can be calibrated to the needs of a specific project for most flexibility (manual), most scale (integrated workcells), or somewhere in the middle ("walk-up" systems)
- Flexibility and scale at the same time are afforded by Reconfigurable Automation Carts (RACs), that are assembled into workflows and pipelines to satisfy experimentation and production needs
- Gingko provides Lab Data as a Service, and produces new data as needed, with a library of over 2 billion genes, including data on antibody developability and enzyme function
- Recommendations for new types of data that the community should be creating include:
- Broad DNA and protein sequence data that represents complete biodiversity
- Metabolomics and transcription data surrounding classes of biomolecules or types of microorganism function
- Data that intentionally addresses a distribution of experimental dimension
- Data with rigorous metadata that can be amended based on new learning around latent variable space
- Intentional supplementation for existing data sets based on recognized data quality concerns
- Benchmarking data to allow for quantitation of model performance
- Data that allows for tracing across models to understand data security