Retrobiosynthesis, Biosecurity, and Large Language Models

Prasanna Muthukumar, Nicholas Roehner, Kemper Talley, Sean Colbath, Jacob Beal

Raytheon BBN
10 Moulton Street, Cambridge, Massachusetts 02138 USA
prasannakumar.muthukumar@rtx.com

Challenge Overview

Large Language Models (LLM) are cracking open a biosecurity Pandora's box. One of the major challenges of synthetic biology has been the engineering of enzymatic pathways for production of desired molecules from metabolic precursors, also known as *retrobiosynthesis*. Currently, retrobiosynthesis is an extremely challenging problem, typically requiring a high degree of cross-disciplinary expertise coupled with large-scale investments such as in the DARPA 1000 Molecules program or high-throughput laboratory operations at large leading-edge companies like Ginkgo Bioworks or Amyris.

Large language models like chatGPT (openAI 2022) have become extremely proficient at summarizing information and transforming that information from one written style to another. Extrapolating from recent results in the application of LLMs to design of chemical reactions (Mahjour, Hoffstadt, and Cernak 2023), we predict that we are not far from the day when LLMs can lower the barriers for retrobiosynthesis by providing easy-to-use "recipes" that a layperson can use to synthesize any compounds they desire while only using commercial nucleic acid synthesis and standard or DIY biochemistry apparatus. Models like the one proposed in (Boiko, MacKnight, and Gomes 2023) reduce barriers further by enabling LLMs to interact directly with cloud labs without requiring an expert in the loop. While such LLM-enabled systems are likely to usher in a new era of scientific exploration, democratization of retrobiosynthesis also vastly increases the potential for misuse. A malicious actor will now be able to plug instructions from a future LLM into a cloud lab and get a highly toxic substance shipped out to anywhere they want.

Safeguards do exist in nucleic acid synthesis companies and cloud labs today, but they have primarily focused on screening for dangerous chemical reagents and/or DNA sequences directly associated with pathogenicity or toxicity. These approaches are effective for non-biological chemical synthesis and for protein toxins such as enterotoxins, ricin, or conotoxins. Small molecule toxins such as saxitoxins or amatoxins, however, are produced indirectly via enzymatic reactions that modify an otherwise harmless metabolite or short peptide into the toxin molecule. Current screening tools are unable to warn of such indirect relationships, and thus malicious retrobiosynthesis is likely to evade all current screening methods. It is therefore imperative that we develop better screening processes that can evaluate the potential of an enzyme to be used for producing dangerous small molecule toxins.

Technical Discussion of Solution

A currently popular line of research is to investigate techniques that prevent LLMs from divulging sensitive information. Such research runs the gamut from imbuing LLMs with human morality (Jiang et al. 2021) to using external safeguards that wrap around the LLM. While we fully support such investigations, we strongly believe that such approaches are a purely stopgap solution, at least in the context of retrobiosynthesis.

The ability to build a foundational model is currently restricted to a small handful of well-resourced institutions, all of whom have pledged to build models that restrict the release of harmful information. Unfortunately, humans have proven to be adept at tricking LLMs into divulging such material (Bond 2023). Moreover, even if we assume that LLM safety techniques massively improve in the near future, and are perfect at protecting every bit of sensitive information, these safeguards will only apply to those LLMs that are created with good intentions. The cost of hardware to train these models is only expected to decrease, and future software tools will also lower the software engineering barriers for model training at scale. Adding to these issues are more conventional problems like models getting leaked before protections can be put in place. As experts with several years of experience in both AI and biosecurity, our opinion is that effectively securing sensitive information in an AI model is impossible. We need to instead focus on *biological chokepoints*.

To explain our suggested approach, let us consider the analogous problem of nuclear safety. Information on how to build a nuclear device is not difficult to find. (McPhee 1974), for example, gives an extremely detailed description of how nuclear fission bombs are constructed. The proliferation of nuclear weapons is instead controlled by restricting access to fissile material. No amount of nuclear weapon design knowledge can overcome the handicap of being unable

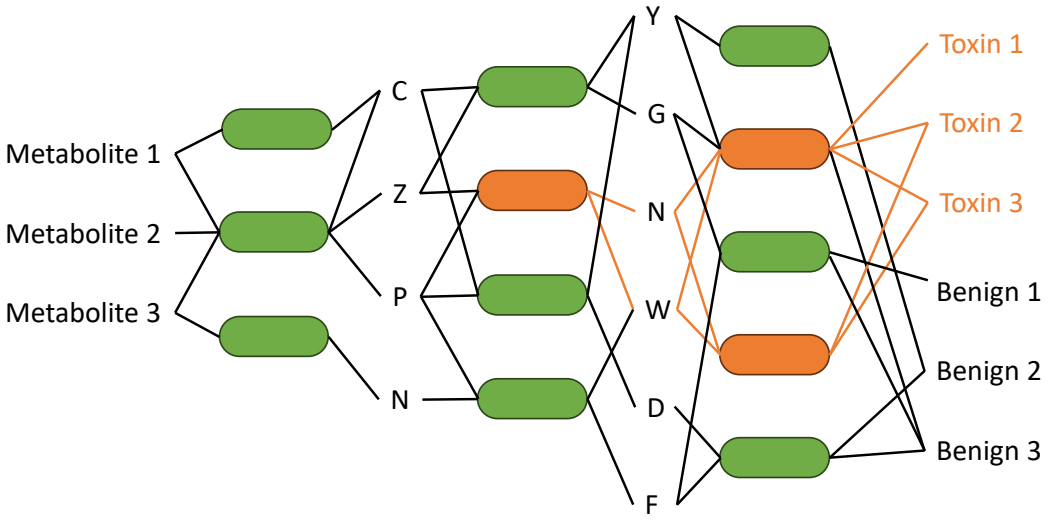

Figure 1: Notional network of retrobiosynthesis pathways generated from an LLM. The rounded rectangles depict enzymes which act to transform each reactant linked on their left into a corresponding product on their right. The pathways and enzymes highlighted in orange lead primarily or exclusively to toxins. Controlling nucleic acid synthesis of the enzymes highlighted in orange therefore can act as effective chokepoints (the network depicted here is completely hypothetical and is drawn as above purely for illustrative purposes).

to get one's hands on uranium and plutonium. These materials are therefore a critical chokepoint that governments the world over have restricted.

Retrobiosynthesis of small molecule toxins involves a series of enzymatic chemical reactions with intermediate compounds culminating in the toxin itself. While the intermediate compounds are only produced during the actual chemical reactions, the enzymes necessary for the reactions are large proteins that are produced by nucleic acid sequences long enough for conventional screening to be applicable. Most enzymes, however, cannot serve as a chokepoint because they are also useful for making many benign molecules. At the same time, many classes of toxins have distinctive features in their chemical structures, which in turn are likely to be linked with specific classes of enzymes that can produce those chemical features. The primary difficulty of our task therefore lies in identifying those specific enzymes that lead primarily or exclusively to small-molecule toxins.

Our proposed approach involves starting with a list of toxic substances from T3DB, a list of completely benign compounds, and enzyme information from databases like KEGG and UniProt. We will then leverage current LLM techniques to generate retrobiosynthesis pathways starting from a standard set of precursor metabolites. Our plan is to then fuse the various pathways into a network. Analyzing this network then helps us identify the enzymes that can be used as chokepoints, since they are primarily or exclusively used in pathways for toxin synthesis.

Conclusion

The likely LLM-driven increase in accessibility of retrobiosynthesis is a critical emerging biosecurity concern. While we assess that it is likely to be impossible to prevent access to AI-supported retrobiosynthetic design of toxin synthesis pathways, we propose that it is likely to be possible to "blacklist" enzymes that are key chokepoints for the biosynthesis of dangerous toxins. More investigation is necessary to determine whether the hypothesized chokepoints actually exist, but if they do this approach would enable a first line of defense against malicious retrobiosynthesis. As future LLMs continue to improve, however, they may also improve in the ability to evade such chokepoints, necessitating further iterations of the same process of chokepoint identification on the evolving landscape of capabilties.

References

Boiko, D. A.; MacKnight, R.; and Gomes, G. 2023. Emergent autonomous scientific research capabilities of large language models. *arXiv preprint arXiv:2304.05332*.

Bond, S. 2023. https://www.npr.org/2023/08/15/1193773829/what-happens-when-thousands-of-hackers-try-to-break-ai-chatbots. Accessed: 2023-08-29.

Jiang, L.; Hwang, J. D.; Bhagavatula, C.; Bras, R. L.; Liang, J.; Dodge, J.; Sakaguchi, K.; Forbes, M.; Borchardt, J.; Gabriel, S.; et al. 2021. Can machines learn morality? the delphi experiment. *arXiv preprint arXiv:2110.07574*.

Mahjour, B.; Hoffstadt, J.; and Cernak, T. 2023. Designing Chemical Reaction Arrays using phactor and ChatGPT.

McPhee, J. 1974. *The Curve of Binding Energy: A Journey Into the Awesome and Alarming World of Theodore B. Taylor*. Macmillan.

openAI. 2022. https://openai.com/blog/chatgpt. Accessed: 2023-08-29.