Adversarial AI-Assisted Protein Design for Evading Hazard Detection: Methods, Results, and Mitigations

Bruce Wittmann and Eric Horvitz

Office of the Chief Scientific Officer
Microsoft
One Microsoft Way, Redmond, WA 98052
bwittmann@microsoft.com; horvitz@microsoft.com

October 27, 2023

1 Summary

Over the last five years, artificial intelligence (AI) methods applied to protein design have greatly expanded our ability to engineer proteins. The evolving approaches and trajectory of advancement are providing tools for custom-tailoring proteins to take on numerous challenges in medicine, bioscience, chemistry, and material science. We anticipate that progress in AI-driven protein design will be foundational to breakthroughs in many scientific and engineering fields, with great promise for elevating our quality of life. However, as much as these new AI-assisted protein engineering tools have the power to do good, they can just as easily be used by bad actors to do harm. For instance, using AI methods, it is now possible to dramatically alter a protein's sequence while preserving its function. We are concerned that this hitherto impossible diversification capability has potential to allow the nefarious generation of variants of dangerous proteins that are undetectable by the hazard detection software used today by synthetic DNA providers to screen customer orders. The new capabilities enhance the ability of would-be bioterrorists to obtain hazardous protein material. We explore a set of methods for performing adversarial AI-assisted protein design (AAIPD) aimed at evading traditional methods for detecting the intention to produce hazardous proteins.

We report on a preliminary investigation into the capabilities of existing hazard detection software used by synthetic DNA providers to identify potentially dangerous proteins after well-known default sequences of amino acids are modified. As an initial AAIPD study, we present several sample methods for producing variants of base proteins and demonstrate in a case study how these methods can be employed to create dissimilar sequences with a high likelihood of retaining the function of dangerous base proteins.

After generating several tens of thousands of alternate sequences of wild-type ricin, we engaged with Twist Bioscience and Raytheon BBN—two companies that employ nucleotide screening—to understand the value of current methods to identify the modified proteins. We found that existing hazard detection software was incapable of identifying AI-generated variants of dangerous proteins, with up to 100% of AI-generated variants of the potent toxin ricin passing through the software undetected. While it is unclear how many of our designed sequences would retain (or, indeed, exceed) native-like function, the methods we use have proven success in other applications, so it is not unreasonable to expect that at least some of the variants that passed the filter would retain toxicity. Regardless, at a minimum, this study indicates the potential of AAIPD to allow potential bad-actors to order genes encoding potentially dangerous variants.

Working with Twist and Raytheon, hazard detection methods were improved to detect hazardous variants produced by AAIPD strategies that employ current AI-assisted protein engineering tools. However, as AI-assisted protein engineering methods continue to improve, hazard detection systems will need to be updated as well. We thus propose an approach to DNA hazard detection analogous to cybersecurity, where white-hat individuals continue to attempt to "hack" their way around existing security measures, then work to address identified vulnerabilities in advance of public disclosure.

2 Background

Proteins are macromolecules critical to almost all of life's processes. The instructions for generating every protein in every organism are encoded in that organism's DNA. Specifically, in a protein-coding gene, three-letter DNA codons encode one-letter protein amino acids, which are the monomeric units that, when chemically chained together, form a protein. The full sequence of codons in a gene thus defines the full sequence of amino acids in a protein.

To synthesize a protein, biologists and others working with proteins order the DNA sequence that encodes it from one of numerous synthetic DNA providers, then express that synthesized DNA in a host organism. That is, the scientists seeking the protein rely on the biological machinery present in a host organism to build the protein for them using the instructions present in the DNA. To prevent their customers from either accidentally or intentionally ordering DNA that encodes dangerous proteins, synthetic DNA providers have developed software that identifies potentially hazardous sequences and prevents them from being ordered. While the exact strategies used by existing hazard detection software vary from provider to provider, in general, the detection methodology works by looking for similarities between known hazardous DNA sequences (and the proteins they encode) and the ordered DNA.

Recent advances in AI-assisted protein engineering facilitate the diversification of protein sequences associated with known proteins while maintaining function, an engineering capability that was previously extremely difficult. Because existing hazard detection software rely on identifying similarities between known hazardous sequences and ordered sequences, the newfound simplicity with which protein variants can now be designed raises the question as to whether these software are sufficient for threat mitigation. By extension, it also asks how current methods can be extended to be more robust to adversarial manipulations of sequence performed in a manner that does not significantly degrade function.

3 Methods

3.1 Protein Variant Production

For our case study, we selected ricin as a model toxic protein. Naturally derived from castor beans, ricin is a potent ~~toxic~~ [toxin] that can enter the human body by inhalation, injection, or ingestion. Due to its relatively low LD_{50} (~30 $\mathrm{mg/kg}$ body weight in mice when ingested, ~4 $\mu\mathrm{g/kg}$ body weight in mice when inhaled) and the ease with which it can be distributed as a powder, ricin is a substance of concern for use in bioterrorism.

To build ricin sequence analogs, we downloaded its structure (PDB: 2AAI) from the Protein Data Bank (PDB) and fed it into the open-source software, ProteinMPNN. ProteinMPNN was developed to solve the inverse folding problem, where a model takes the backbone coordinates of a protein as input and returns a proposed protein sequence that should fold to that structure as output. Thus, by feeding the ricin PDB structure into it, we could generate thousands of ricin analogs.

Importantly, not all positions in a protein structure are created equal, and some will be more important to a protein's function than others. Indeed, it is generally observed that by fixing the identities of important residues to those of the wild-type amino acid sequence, while still allowing other residues to be varied, the success rate of ProteinMPNN can be improved.

It is also important to note that, because existing DNA hazard detection software rely on calculating similarities between known and potential dangerous sequences, it is reasonable to expect that ricin analogs with less sequence identity to the wild type would be more difficult to detect. Thus, there is a direct correlation between the probability that our ricin analogs will maintain function and the probability that they will be detected by hazard detection software: the more residues we fix to the wild-type identity, the greater the sequence identity will be between the variants produced by ProteinMPNN and the wild-type sequence, so although the probability of maintaining wild-type-like function will be improved, the probability of detection by hazard detection software will increase as well.

The expected correlation between detection rates and expected likelihoods of preservation of function is somewhat comforting. We nonetheless have sought a deeper understanding of the degree of diversification of sequences for ricin needed before sequences of can evade identification by existing hazard detection software. We thus created six sets of ~10k protein variants each, all with differing levels of wild-type residues conserved:

1. Set A: All residues in the protein were allowed to be changed by ProteinMPNN.
2. Set B: Residues known to be important for ricin toxicity that were previously identified in the literature were held constant.
3. Set C: Every residue in Set ~~A~~ [B] as well as all residues with a heavy atom within 1.5 Å of any heavy atom of the residues in Set ~~A~~ [B] were held constant.
4. Set D: All residues with any heavy atom within 5 Å of a bound RNA substrate (ricin's toxicity is derived from catalytic activity on rRNA) in PDB 3RTJ were held constant.
5. Set E: The same as Set C, only using a 3.5 Å cutoff.
6. Set F: The same as Set D, only using a 10 Å cutoff.

From Set A to Set F, we roughly expect the difficulty of detection to decrease and the probability of function to increase—Set F is thus the one most likely to contain dangerous proteins while Set A is the least likely.

3.2 DNA Variant Production

Synthetic DNA providers receive specifications of DNA sequences from their customers, not protein sequences. Thus, the sequences built in Section 3.1 must be reverse translated into DNA before being passed into hazard detection software. This provides another opportunity to test the robustness of these software, as explained in the three methods we used for the reverse translation process:

1. The first strategy produced a baseline: As multiple codons equivalently encode individual amino acids, we replaced the codon for each amino acid in each protein sequence by a randomly selected codon encoding that amino acid.
2. Because codons are 3 letters long, there are 3 possible reading frames along a single-stranded DNA molecule. For instance, the DNA strand ATGGCAAA can be read as ATG|GCA, TGG|CAA, or GGC|AAA, corresponding to the three different amino acid sequences MA, WQ, and GK, respectively. Thus, to translate DNA into protein, the reading frame must be known; otherwise, the incorrect protein sequence will be read.
   Recognizing the above, in the second strategy we prepended a randomly generated DNA sequence for sperm whale myoglobin and a segment of junk DNA to ricin DNA sequences generated using the first approach. Junk DNA was also appended to the ricin DNA sequence. The idea here was that the junk DNA added before and after the ricin DNA sequence would place the ricin DNA in a different reading frame relative to myoglobin. Because the myoglobin DNA sequence perfectly matched a known sequence while the ricin analogs would not, the hazard detection system would ~~possible~~ [possibly] choose the myoglobin reading frame as the most probable and thus fail to translate the ricin analog properly, effectively obscuring it.

3. Of course, our second reverse translation strategy could be bypassed easily if the hazard detection software simply tested all DNA reading frames rather than choosing the most probable (a strategy that is already employed in hazard detection software). To make detection of full ricin analogs impossible in any reading frames, in our last reverse translation strategy, each ricin DNA sequence was split into five fragments, the order of these fragments was shuffled, and random-length pieces of sperm whale myoglobin were inserted between them all. Additionally, each ricin DNA fragment was replaced with its reverse complement with 50% probability. The idea behind this approach was that, by breaking up the ricin sequence and changing the reading frames of the resulting fragments relative to one another, it would become impossible to translate any segment of the DNA directly to a full ricin protein sequence. Any software used to detect ricin would need to perform a homology search on all possible fragments of the submitted DNA in all possible reading frames. Of course, a molecular biologist could still quite easily reconstruct the full, in-frame ricin analog sequence from DNA ordered in this fashion.

DNA sequences were prepared for each set using the first method. Additionally, DNA was produced for Set A using methods 2 and 3. The DNA for Set A produced using reverse translation approach 1 was labeled as Set A_0; the DNA produced using reverse translation approach 2 was labeled as Set A_1, and so on.

3.3 Submitting DNA to Twist Biosciences and Raytheon BBN

Set A_0 was submitted to Twist Biosciences; all sets of DNA were submitted to Raytheon BBN (the organization that provides DNA hazard detection software to IDT). These two companies use different hazard detection software, though the specific details of how these software function are unknown to us.

4 Results

A summary of results can be found in Table 1.

4.1 Twist Biosciences

Twist Bioscience submitted DNA Set A_0 to their ThreatSeq software. Of the ~10k sequences in this set, ThreatSeq detected only one of them as a threat—the wild-type ricin sequence that was included as a positive control. The remaining sequences were not detected as a threat, indicating that, absent any other intervention, they could currently be ordered through Twist Bioscience.

As a follow-up, in the context of our submitted challenge, Twist Bioscience passed all protein sequences of Set A through ESMFold to predict their three-dimensional structures, then aligned those predicted structures against the AlphaFold-predicted structure for ricin to calculate a TM score for each, which is a measure of structural similarity with values on the range (0, 1], where 1 indicates a perfect match between structures. Generally, TM scores below 0.17 correspond to unrelated, randomly chosen proteins while scores above 0.5 correspond to proteins that have the same overall fold. The team found that all of the variant sequences had a TM score above 0.17 and ~9k had a TM score above 0.5, indicating that ESMFold-predicted structures could identify analogs reliably where the ThreatSeq software could not.

4.2 Raytheon BBN

Raytheon BBN submitted all DNA sets to their internal threat detection software. As with ThreatSeq, only the positive control sequence was identified as a threat—all ricin variants went undetected and so could currently be ordered through IDT absent any other intervention.

Following the failure to detect variants, the Raytheon BBN team had an insight regarding another approach that could be more resilient to AI-diversified harmful proteins. The details of the new

methods are unknown to us; however, the updated approach was far more effective at identifying the submitted DNA as encoding ricin analogs, as summarized in Table 1. By report of the Raytheon BBN team, the revised threat detection methodology had a false-positive rate (defined as how frequently the system incorrectly identified a random sequence from a set of ~8k as a ricin analog) of 0.000%.

Set Name	Sequence Identity (%, $\mu \pm \epsilon$)	N Variants	Detection Rate Twist (%)	Detection Rate Raytheon Old (%)	Detection Rate Raytheon New (%)
A_0	41.46 ± 1.23	9985	0.00	0.00	100.00
A_1	41.46 ± 1.23	9985	N/A	0.00	100.00
A_2	41.46 ± 1.23	9985	N/A	0.00	99.94
B	43.03 ± 1.20	10000	N/A	0.00	100.00
C	44.24 ± 1.19	10000	N/A	0.01	100.00
D	43.73 ± 1.19	10000	N/A	0.00	100.00
E	47.06 ± 1.13	10000	N/A	0.18	100.00
F	48.79 ± 1.12	10000	N/A	4.36	100.00

Table 1: Reported detection rate of ricin analogs using the original methods employed by Twist and Raytheon BBN as well as using the new method employed by Raytheon BBN.

5 Conclusions and Future Directions

The existing DNA hazard detection systems of both Twist Biosciences and IDT—two of the leading synthetic DNA providers—were unable to identify any AI-diversified ricin analogs. This result clearly demonstrates that, given newfound AI-assisted protein engineering strategies, DNA hazard detection systems that had been in place are insufficient for mitigating bioterrorism threats. We believe the results make salient the need to make changes in our existing threat detection systems to address this shortfall.

Future threat detection methods must satisfy three criteria: (1) the false negative rate must be low (i.e., threats must be detected), (2) the false positive rate must be low (i.e., proteins that are not threats must not be detected), and (3) the method must be computationally inexpensive to operate.

These criteria make the structure-based method proposed by the Twist team a challenge to implement in practice, as predicting protein structure is an expensive computational operation to perform. Not to mention, the third reverse translation system that we describe above could easily bypass a structure-based search. As for Raytheon BBN's alternative strategy, we do not know the details of it and cannot comment at this time on its broader feasibility. However, we assume that it still relies on a sequence-based search.

If the current rapid pace of advancement in AI-assisted protein engineering continues as is, the nature of the threats with which DNA hazard detection software will have to contend will likely evolve quickly and become ever-more sophisticated. We should not expect that sequence-based strategies alone will be sufficient in the long-term, as we envision a future where AI-assisted generation of proteins produces sequences unlike any seen in nature. In addition to improving sequence-based hazard detection software, then, we should pursue the development of new approaches as well. For instance, there is likely value in investigating homology detection via the learned embedding space of large protein language models, research into which is already thoroughly underway. Similarly, the detection of hazardous proteins among a sea of nonhazardous proteins is at its heart a form of anomaly detection, a problem that has already received extensive attention among the broader machine learning community.

Regardless of the threat mitigation strategies developed to date and in response to the challenges presented in this paper, it is important to recognize that the future of biosecurity will be ever changing. In the same way that cybersecurity teams must constantly develop new strategies to mitigate new

threats as new AI approaches are developed, those dedicated to preventing acts of bioterrorism will have to continue to adapt as well.

6 Acknowledgments

We thank James Diggans and Emily Leproust at Twist Bioscience, Mark Behlke, Adam Clore, Gavin Kurgan, and Jeff Manthey at IDT, and Jake Beal and Tom Mitchell at Raytheon BBN for participating in this study. In particular, we thank James Diggans at Twist Bioscience and Jake Beal and Tom Mitchell at Raytheon BBN for running the ricin variant sets that we sent them through their respective DNA hazard detection software and analyzing the outputs. We thank David Baker at the Institute for Protein Design at the University of Washington for feedback on design of variant sets.