IFP

A Million-Peptide Database to Defeat Antibiotic Resistance

How to build a large peptides database to train the AlphaFold for new antibiotics | **Maxwell Tabarrok**

A Million-Peptide Database to Defeat Antibiotic Resistance

How to build a large peptides database to train the AlphaFold for new antibiotics | Maxwell Tabarrok

This essay is part of The Launch Sequence, a collection of concrete, ambitious ideas to accelerate AI for science and security.

Summary

Antibiotic-resistant infections already kill over 1.2 million people annually worldwide. This rate is rising, threatening to return humanity to pre-antibiotic mortality rates. Antimicrobial peptides offer a promising solution because they resist resistance, are programmable through machine learning (ML), and are easy to manufacture. However, current peptide databases contain only thousands of sequences, far below the scale needed for effective ML models. We propose creating a million-peptide database through targeted data infrastructure investment by the NIH, following successful precedents like PubChem, the Human Genome Project, and the Protein Structure Initiative. Using high-throughput synthesis methods like SPOT synthesis, this database could be completed within five years for less than $350 million — a fraction of the $4.6 billion annual treatment cost of just six drug-resistant infections in the US. This moonshot would provide the data foundation for an explosion of ML-driven breakthroughs in antimicrobial peptide research, potentially saving millions of lives.

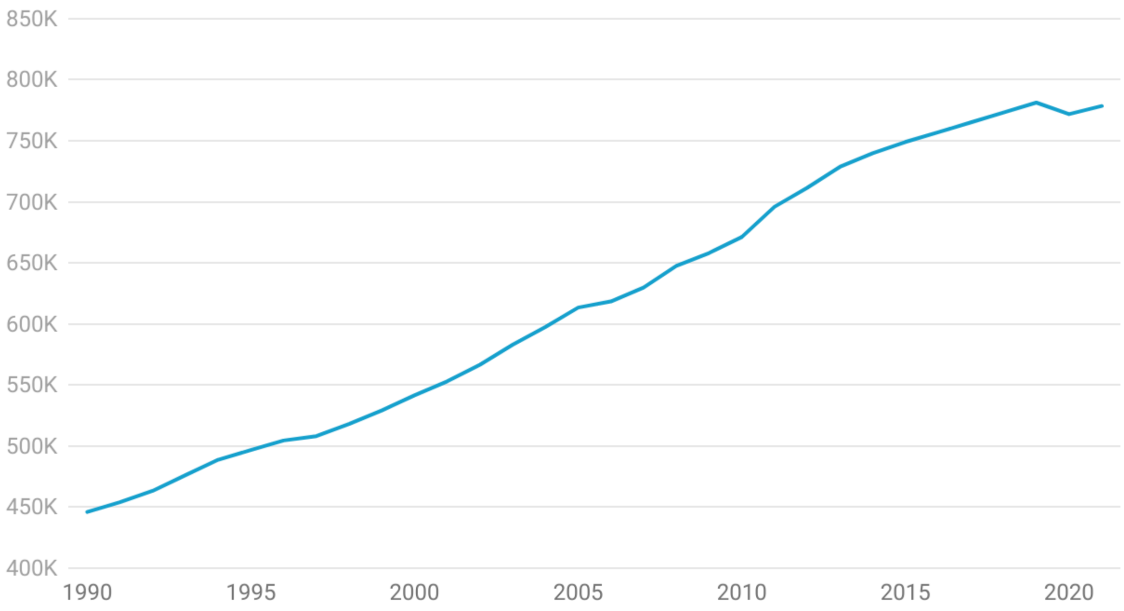
Motivation

For nearly all of human history, infectious disease has been our deadliest foe. In the first decades of the 20th century, nearly one in a hundred Americans would die of an infectious disease every year. To put that into context, the American infectious disease death rate was 10 times lower during the height of the COVID-19 pandemic in 2021.

But if nothing is done, antibiotic resistance promises a return to the historical norm of frequent death from infectious diseases. As humans use more antibiotics, we inadvertently run the world's largest selective breeding program for bacteria which can survive our onslaught of drugs. We have discovered more powerful antibiotic drugs, but our usage of them is accelerating, while our discovery rate is, at best, stagnating.

As a result, antibiotic resistance is spreading. Today, certain forms of Staphylococcus aureus, like MRSA, are resistant to even our most powerful antibiotics, and the disease results in 20 thousand deaths every year in the US, and more globally.

Global deaths associated with antibiotic resistant S. Aureus



Maxwell Tabarrok | Source: IHME, Oxford | Created with Datawrapper

The way out: AI-accelerated antimicrobial peptide research

Antimicrobial peptides can defeat this crisis. Peptides have in vitro effects on the toughest antibiotic-resistant infections including MRSA, HIV, fungal infections, and even cancer. Peptides are especially promising as a scalable solution to antibiotic-resistant pathogens for three reasons. First, they resist resistance — it is very hard for pathogens to develop a resistance to peptides, as they are able to do for antibiotics.[1] Second, they are programmable. Peptides are direct outcomes of linear amino acid chains, so it's relatively easy to work with them computationally and apply machine learning (ML) to program desired properties. Third, once the right peptide is identified, they are significantly, quicker, easier and cheaper to manufacture than antibiotics.[2]

But peptide antibiotics still aren't common on pharmacy shelves or in hospital treatments. Other peptide treatments, like insulin or GLP-1s, are widespread, and some current clinical trials might bring more peptide antibiotics to market, but many of these are just clinical applications of peptides that are already produced by the human body, like LL37. The main barrier to defeating antibiotic resistance with peptides is still in the fundamental research problem of finding sequences that are both effective against a particular disease and non-toxic.

The good news is that this research problem is perfectly poised to benefit from rapid advances in computation. Scientists are searching for relationships between linear sequences of text, i.e., the amino acid chain, and quantitative outcomes like

[1] While antibiotics often target narrow pathways into a bacteria's metabolism or particular proteins found in their cytoplasm; peptides target general properties of a bacteria's entire membrane, like charge or lipid composition. This means that antibiotics are more specific, but also that they are easy to resist. Changing one residue in a target protein is much easier than changing the electric charge over the entire bacterial surface. This general targeting has allowed antimicrobial peptides to be effective first defenses against pathogens for millions of years without changing much.

[2] There are dozens of companies that will synthesize custom proteins on demand for reasonable prices. By rapidly synthesizing and testing hundreds of different peptides, you can screen for effective and non-toxic treatments and scale them up in six or seven days. This is a stark contrast to small molecule antibiotic manufacturing, where figuring out how to synthesize a particular chemical can take years of trial and error, and making that synthesis efficient can take even longer.

toxicity and antimicrobial activity. Fitting the relationship between quantitative inputs and outputs like this is a textbook machine learning problem that gets easier as your stock of compute power and data grows.

Indeed, scientists already use small machine learning models to find or generate effective peptide treatments. Researchers at the frontier of this field use machine learning to predict new sequences with antimicrobial properties or to search through known sequences for promising candidates. They use similar techniques to Google's AlphaFold, which made significant progress in predicting the 3D structure of proteins. However, ML models of peptides more directly target the medical properties of the peptides, like human toxicity or activity against bacteria, rather than predicting their 3D structure. ML prediction on peptides may also be even more tractable than AlphaFold because peptides are so much shorter than most proteins. But current progress is too slow to meet the challenge of antibiotic resistance.

The problem: Data

ML needs big data. Google's AlphaGo trained on 30 million moves from human games and orders of magnitude more from games it played against itself. The largest language models are trained on at least 60 terabytes of text. AlphaFold was trained on just over 100,000 3D protein structures from the Protein Data Bank.

The data available for antimicrobial peptides is nowhere near these benchmarks. Some databases contain a few thousand peptides each, but they are scattered, unstandardized, incomplete, and often duplicative. For example, the APD3 database is small, with just under 4,000 sequences, but it is among the most tightly curated and detailed. However, most of the sequences available are from frogs or amphibians due to path-dependent discovery of peptides in that taxon. Another database, CAMPR4, has on the order of 20,000 sequences, but around half are "predicted" or synthetic peptides that may not have experimental validation, and contain less information about source and activity. The formatting of each of these sources is different, so it's not easy to put all the sequences into one model. More inconsistencies and idiosyncrasies stack up for the dozens of other datasets available.

The need for public data production

Market failure

Expanding the dataset of peptides and including negative observations is feasible and desirable, but no one in science has an incentive to do it. Anyone can costlessly copy-paste a dataset, making it difficult to put it behind a paywall. Therefore, we can't rely on private pharmaceutical companies to invest sufficiently in manufacturing and testing peptides to support this kind of open data infrastructure. Even if they did, they would fight hard to keep this data a trade secret. This would help firms recoup their investment, but it would prevent other firms and scientists from using the data, undercutting the reason it was so valuable in the first place.

Academia will not step in to fill the data gap either. Non-monetary rewards in academia, like publications and prestige, point toward splashy results in big journals, not toward foundational infrastructure like open datasets. Scientists are, however, enthusiastic for these resources to exist; in the field of antimicrobial peptides, researchers host small-scale open peptide databases and prediction tools free for anyone to use. They are motivated by a genuine desire to see progress in this field, but genuine desire doesn't pay for all of the equipment and labor required to scale up these databases to ML-efficient size. To solve this data problem, we need public investment in data production.

Successful precedents

This strategy of targeted data infrastructure investments has three successful precedents: PubChem, the Human Genome Project, and ProteinDB.

1. The NIH's **PubChem** is a database of 118 million small-molecule chemical compounds that contains nearly 300 million biological tests of their activity, e.g., their toxicity or activity against bacteria. With an annual budget of $3 million, PubChem exceeded the size of the leading private molecule database from Advanced Chemistry Development by around 10,000x in 2011 and made the data free. PubChem is credited with supporting a renaissance in ML for chemistry.

2. Another success is the **Human Genome Project**. Unlike PubChem, the Human Genome Project couldn't rely on collating existing data, and had to industrialize DNA sequencing to record all 3 billion base pairs of human DNA. This 13-year effort began in the early 1990s and cost about $3.8 billion. Over the course of the project, the per-base cost of DNA sequencing plummeted by ~100,000-fold. Before the HGP, gene therapies were less than 1% of clinical trials; today they comprise more than 16%, all building off the data infrastructure foundation laid by the project.

3. Perhaps the closest analog to the million-peptide database proposal is **ProteinDB**, a database of around 150,000 complex proteins and their 3D structure. Like PubChem, ProteinDB has become the primary depository for protein structure discoveries — and like the Human Genome Project, ProteinDB was paired with a large data generation program, the Protein Structure Initiative (PSI).

   > The PSI was a $764 million project funded by the US National Institute of General Medical Sciences between 2000 and 2015. The hundreds of thousands of detailed 3D protein structures in the PSI databank became the essential training data behind the success of AlphaFold.

Solution

A dataset of one million peptide sequences and their antimicrobial properties (or lack thereof) would accelerate progress toward new drugs that can kill antibiotic-resistant pathogens. This would put us on track to defeat drug-resistant diseases before they roll back the clock on the medical progress of the past century.

There are no significant scientific barriers to generating a 1,000x or 10,000x larger peptide dataset. Several high-throughput testing methods have been successfully demonstrated, with some screening as many as 800,000 peptide sequences at once. These methods will need to be scaled up, not only by testing more peptides, but also by testing them against different bacteria, checking for human toxicity, and testing other chemical properties, but scaling is an infrastructure problem, not a scientific one.

Creating the million-peptide database

Agency

The NIH is well-placed to create a million-peptide database, having been responsible for the successful precedents for this project: ProteinDB, the Protein Structure Initiative, PubChem, and the Human Genome Project. In particular, the National Institute of General Medical Sciences (NIGMS) can apply the organizational and funding models learned during the Protein Structure Initiative to the challenge of creating a million-peptide database.

Timeline

Phase I: Data merging and standardization (<1 year)

- NIGMS should replicate the success of PubChem by merging and standardizing existing peptide datasets, and open them to all. A researcher today who wants to use all available peptide data in their model has to collect dozens of files, interpret poorly documented variables, and filter everything into a standardized format. Hundreds of researchers are currently duplicating all of this work. Organizing this data once and for all and establishing a central repository for all future peptide sequence discoveries would save thousands of hours of researcher time.

- The initial Request For Applications for the Molecular Libraries Screening Centers Network, of which PubChem was a part, went out in April of 2004. PubChem was on-line by September of the same year and by August of 2005 PubChem already held over 10 million records.

- However, collecting existing data will not be nearly enough to get to a million-peptide database. The next step is to industrialize peptide testing.

Phase II: Scale up testing to expand the dataset to a million peptides (<5 years)

- Phase II of the million-peptide database project would follow the playbook of the Protein Structure Initiative's production phase. This would begin by soliciting cooperative agreements with both academic research centers and national laboratories to research and implement high-throughput testing methods.

- Mass-produced protein synthesis and testing are already well-established techniques in the field, so this project won't need any 100,000x advances in technology to succeed like the Human Genome Project did. The NIGMS only needs to support scaling up existing techniques.
- A million-peptide database could be completed in less than five years. A single researcher can synthesize 400 peptides on a 20×20 cm cellulose sheet in 6 days using SPOT synthesis and can probably perform tests for antimicrobial activity, human toxicity, and other traits in another week. With an automated pipetting machine the yield increases to 6,000–8,000 peptides in the same six days. A rate of 8,000 peptides synthesized and tested every two weeks would get to a million peptides in 1,800 days, just under five years. Most importantly, almost all of these processes are highly parallelizable, so scaling up the number of peptides you want to test doesn't necessarily increase the amount of time it takes if you can set up another researcher or pipetting machine working in parallel.

Funding

The cost of Phase I would be small, perhaps less than $5 million.

- PubChem, a similar database of chemical compounds, had a budget of only $3 million in its first year and the annual costs for upkeep remained around that figure in later years.

The cost of Phase II would be less than $350 million.

- Retail custom proteins cost $5-$10 per amino acid. At an average peptide length of 20 amino acids, that is around $200 per peptide. That cost is just for the synthesis, not all of the time and labor required for testing, so a reasonable upper bound on the cost of a million-peptide database is $350 million.
- Even this large upper bound cost is likely justified by the potential impact of antimicrobial peptides. The direct treatment costs for just six drug-resistant infections is around $4.6 billion annually in the US, with a far greater cost coming from the excess mortality and damaged health.

- The actual cost is likely far less than this $350 million upper bound. Performing protein synthesis in-house and in bulk, rather than buying retail, can greatly reduce costs. Additionally, high-throughput methods like SPOT synthesis can be less than 1% of the cost per peptide and allow researchers to synthesize thousands of peptides at once. Clinical use of the tested peptides would probably require retesting with more expensive, higher purity methods, but only the few most promising candidates would require retesting. For the purpose of supplying millions of data points to an ML model, the purity of this high-throughput method is more than sufficient.

Implementation

The NIGMS Director, Jon Lorsch, Ph.D., is the critical decision maker for this proposal.

Director Lorsch has the authority to:

- Approve the project concept
- Submit proposal to the NIGMS Advisory Council for approval
- Allocate funding from extramural grants budget
- Instruct NIGMS program leadership to draft new RFAs
- Coordinate peer review with the Office of External Research

A single, concentrated effort over several years would lay a foundation for an explosion in ML-driven antimicrobial peptide research, making possible effective new treatments for some of the world's deadliest and intransigent diseases.

Further resources

- Antimicrobial Peptide Database (APD3), "Antimicrobial Peptide Database (APD3)," n.d.

  Current leading peptide database with ~4,000 sequences, demonstrating both the quality of existing curation efforts and the scale limitations that this moonshot would address.

- Waghu et al., "Machine Learning Approaches for Antimicrobial Peptide Discovery," Antibiotics, 2022.

  Comprehensive review of current ML methods in peptide research, demonstrating both the computational advances already achieved and the critical data limitations that constrain further progress in the field.