香港中文大學
The Chinese University of Hong Kong

A Deep Learning Approach to Antibiotic Discovery

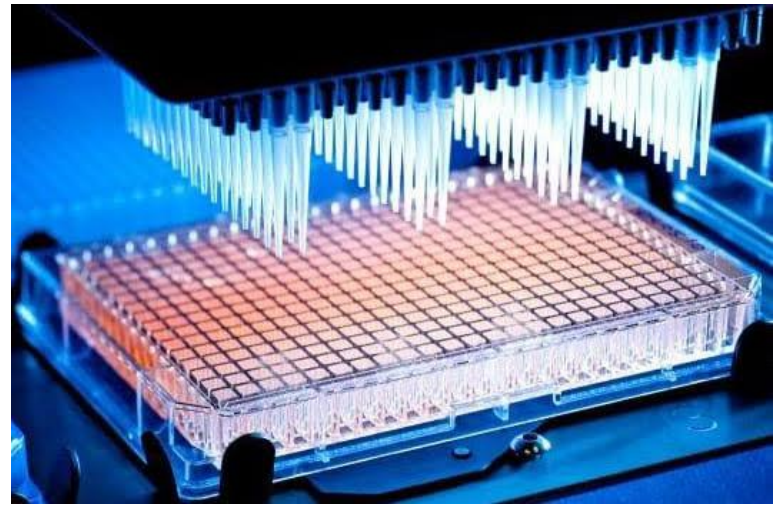
Andrei Macrineanu
24 July 2026

Background

What prompted this research to be conducted.

Main issues

"Since the implementation of high-throughput screening in the 1980s, no new clinical antibiotics have been discovered using this method"[1]



- HTS was a novel screening technique that increased efficiency greatly, but didn't produce any results
- Bacteria was becoming resistant to current drugs.
- Antibiotic discovery was unprofitable
- Screening would constantly come back to the same compounds.

1. Stokes, J. M., et al. (2020). "A Deep Learning Approach to Antibiotic Discovery." Cell **180(4): 688-702.e613.**

Their approach

How they tried to solve the problem

Methods

Using Machine Learning Models to predict activity and then screen those hits.

- Train a model once, use it forever.
- Training data was only 2,335 molecules
- Millions of molecules screened.

Model design

Design of the ML model

- D-MPNN – implemented through Chemprop
- Model would output whether a molecule has antibacterial properties, not trying to predict its activity
- Filtering structurally different compounds using Tanimoto similarity

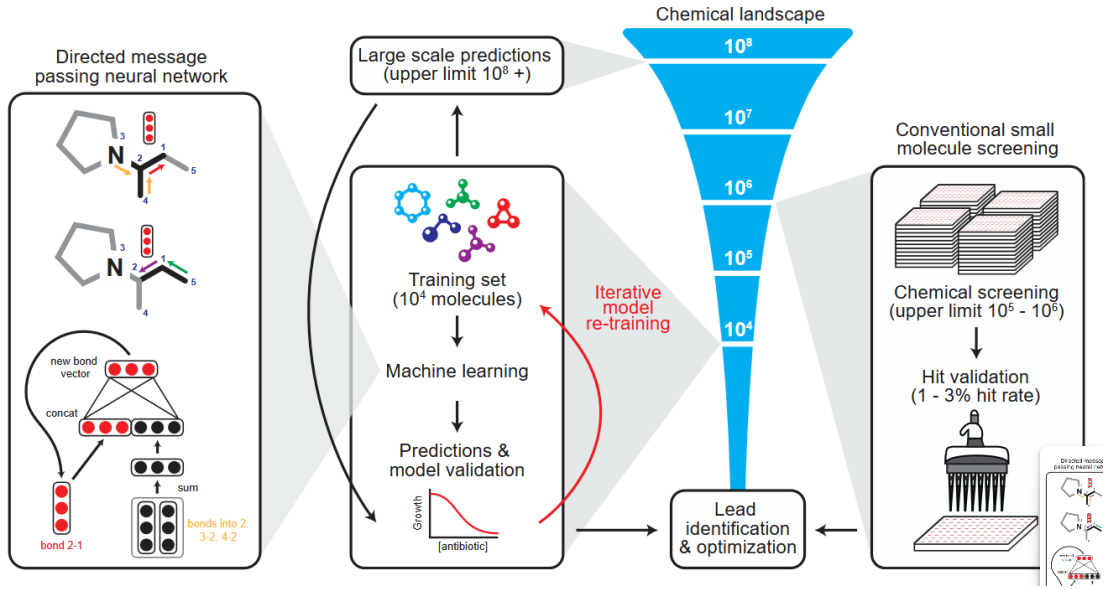


Figure 1. Stokes' ML model[1]

1. Stokes, J. M., et al. (2020). "A Deep Learning Approach to Antibiotic Discovery." Cell **180(4): 688-702.e613.**

D-MPNN

- D-MPNN – unlike MPNN, it avoids tottering by passing messages from bond to bond rather than atom to atom
- Model would output whether a molecule has antibacterial properties, not trying to predict its activity

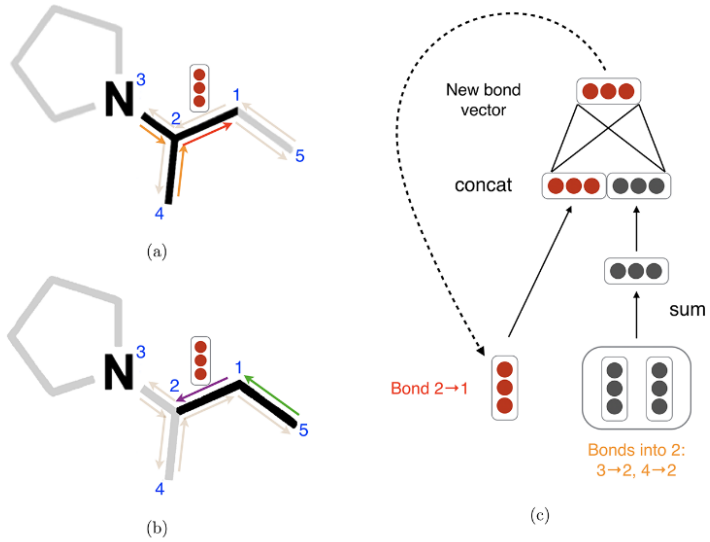


Figure 1. Illustration of bond-level message passing in our proposed D-MPNN. (a) Messages from the orange directed bonds are used to inform the update to the hidden state of the red directed bond. By contrast, in a traditional MPNN, messages are passed from atoms to atoms (for example, atoms 1, 3, and 4 to atom 2) rather than from bonds to bonds. (b) Similarly, a message from the green bond informs the update to the hidden state of the purple directed bond. (c) Illustration of the update function to the hidden representation of the red directed bond from diagram (a).

Figure 2. Graphical representation of how D-MPNN differs from regular MPNNs.[2]

2. Yang, K., et al. (2019). "Analyzing Learned Molecular Representations for Property Prediction." Journal of Chemical Information and Modeling **59(8): 3370-3388.**

Results

Did the architecture matter?

- Five other models were checked retrospectively to check validate the results of their model

- Other models ranked halicin anywhere from 270-2500, while Stokes' model ranked it 89

Why was halicin a big discovery?

- Halicin was previously only looked at as a diabetes drug candidate, never tested as an antibiotic.
- Very structurally different from other antibiotics, Tanimoto similarity of 0.21
- Different mechanism compared to other drugs.

Mechanism

- Halicin disrupts the bacterial membrane's pH gradient, rather than inhibiting it.
- Bacteria struggled to evolve resistance

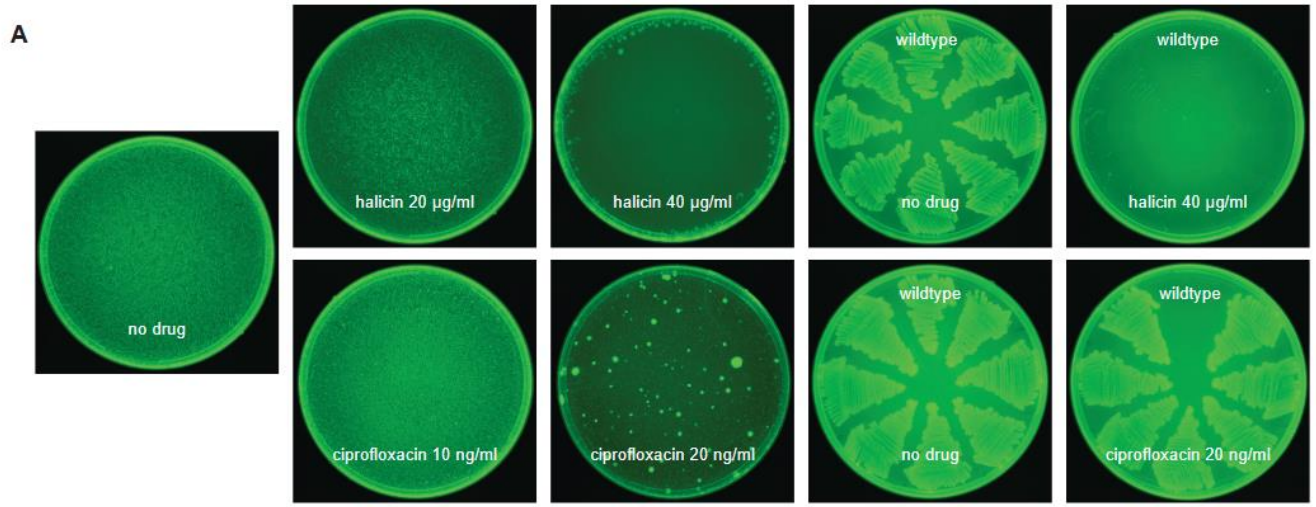


Figure 3. Plates showing that bacteria struggled to evolve resistance towards halicin.[1]

1. Stokes, J. M., et al. (2020). "A Deep Learning Approach to Antibiotic Discovery." Cell **180(4): 688-702.e613.**

Takeaways

- The choice in architecture was deliberate – it solved a known inefficiency in generic MPNNs
- Combining learned and hand-engineered features outperformed either alone – this allowed the use of a small molecular dataset.
- Screening this was is cheap, only after the training has been done.
- Proves that a fairly modest training data size can train a model that generalises well enough to find novel chemistry.

Thank You