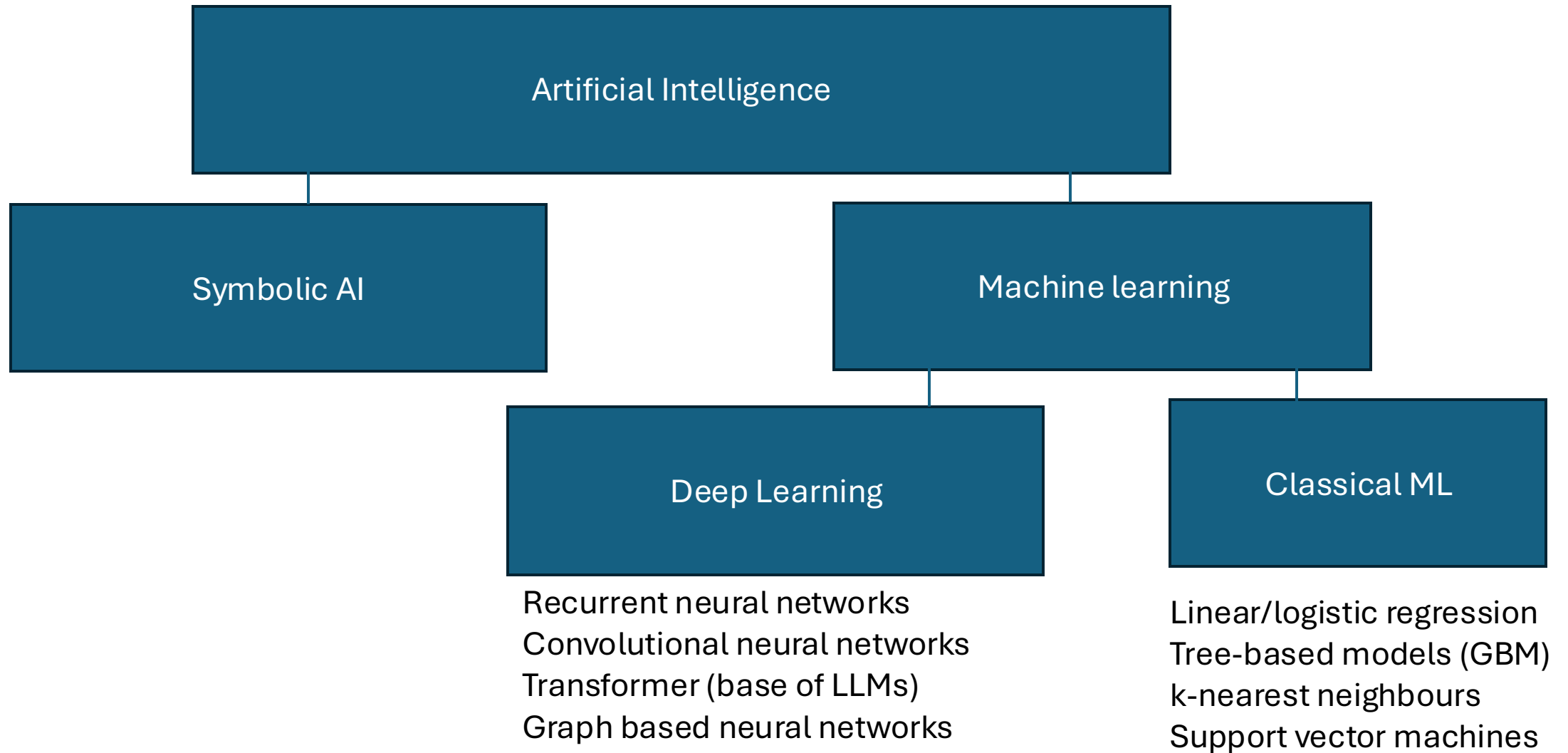


AI/ML for drug discovery

Lou Kohler Voinov

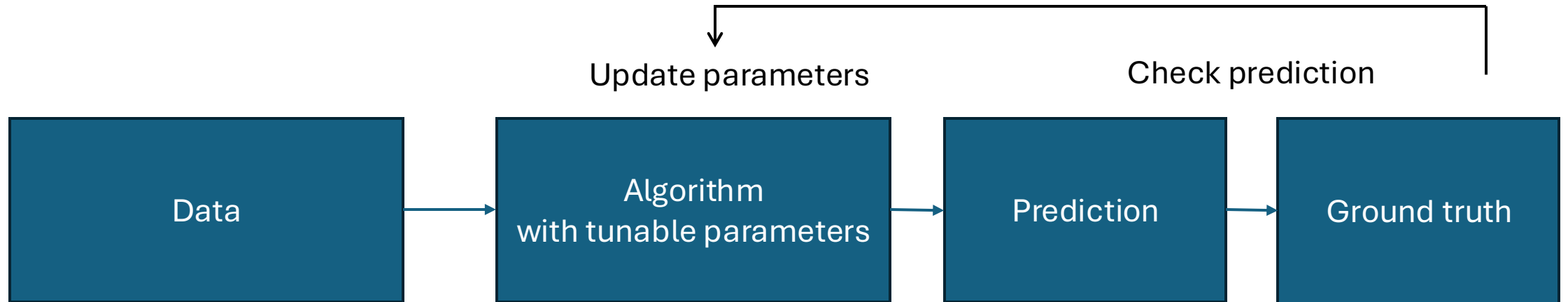
AI vs Machine learning



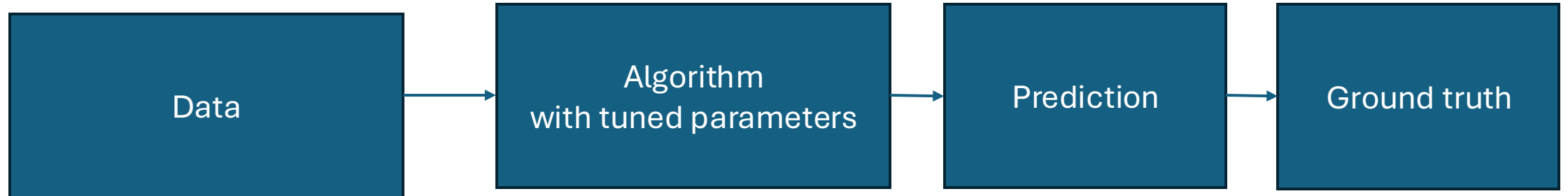
What is machine learning?

Mapping inputs to outputs

Phase 1: Training

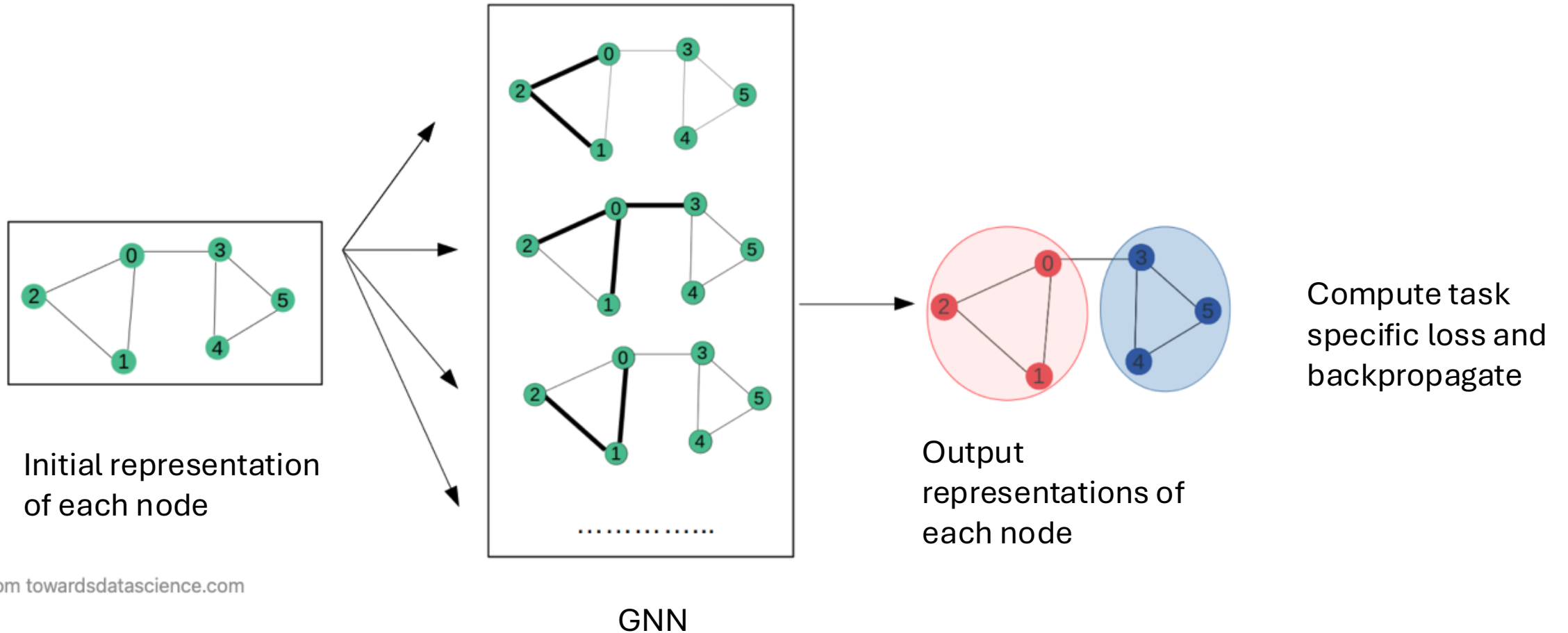


Phase 2: Testing



Graph neural networks

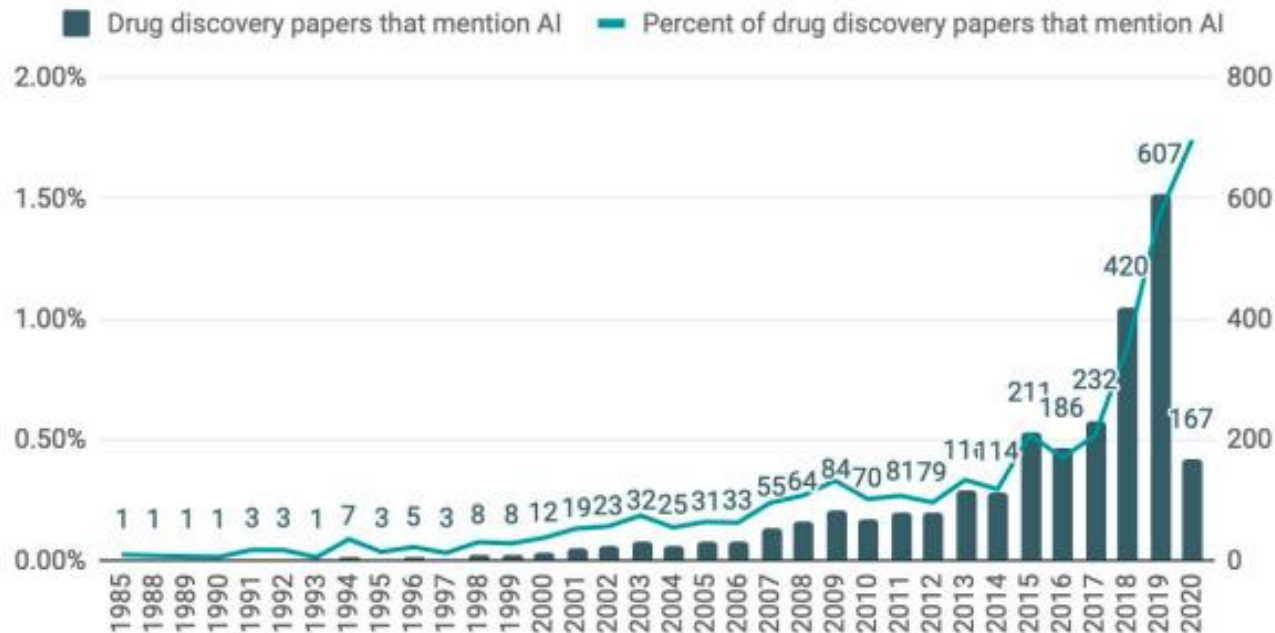
Graphs = (Nodes, Edges)



The use of AI in medicine/drug discovery

AI in drug discovery papers published per year

Source: blog.benchsci.com/artificial-intelligence-in-drug-discovery-trends-and-statistics



Article | [Open access](#) | Published: 15 July 2021

Highly accurate protein structure prediction with AlphaFold

[John Jumper](#), [Richard Evans](#), [Alexander Pritzel](#), [Tim Green](#), [Michael Figurnov](#), [Olaf Ronneberger](#), [Kathryn Tunyasuvunakool](#), [Russ Bates](#), [Augustin Židek](#), [Anna Potapenko](#), [Alex Bridgland](#), [Clemens Meyer](#), [Simon A. A. Kohl](#), [Andrew J. Ballard](#), [Andrew Cowie](#), [Bernardino Romera-Paredes](#), [Stanislav Nikolov](#), [Rishub Jain](#), [Jonas Adler](#), [Trevor Back](#), [Stig Petersen](#), [David Reiman](#), [Ellen Clancy](#), [Michal Zielinski](#), ... [Demis Hassabis](#)

[+ Show authors](#)

Nature **596**, 583–589 (2021) | [Cite this article](#)

ProteinBERT: a universal deep-learning model of protein sequence and function

[Nadav Brandes](#), [Dan Ofer](#), [Yam Peleg](#), [Nadav Rappoport](#), [Michal Linial](#) [Author Notes](#)

Bioinformatics, Volume 38, Issue 8, March 2022, Pages 2102–2110, <https://doi.org/10.1093/bioinformatics/btac020>

Article | [Open access](#) | Published: 11 July 2023

De novo design of protein structure and function with RFdiffusion

[Joseph L. Watson](#), [David Juergens](#), [Nathaniel R. Bennett](#), [Brian L. Trippe](#), [Jason Yim](#), [Helen E. Eisenach](#), [Woody Ahern](#), [Andrew J. Borst](#), [Robert J. Ragotte](#), [Lukas F. Milles](#), [Basile L. M. Wicky](#), [Nikita Hanikel](#), [Samuel J. Pellock](#), [Alexis Courbet](#), [William Sheffler](#), [Jue Wang](#), [Preetham Venkatesh](#), [Isaac Sappington](#), [Susana Vázquez Torres](#), [Anna Lauko](#), [Valentin De Bortoli](#), [Emile Mathieu](#), [Sergey Ovchinnikov](#), [Regina Barzilay](#), ... [David Baker](#)

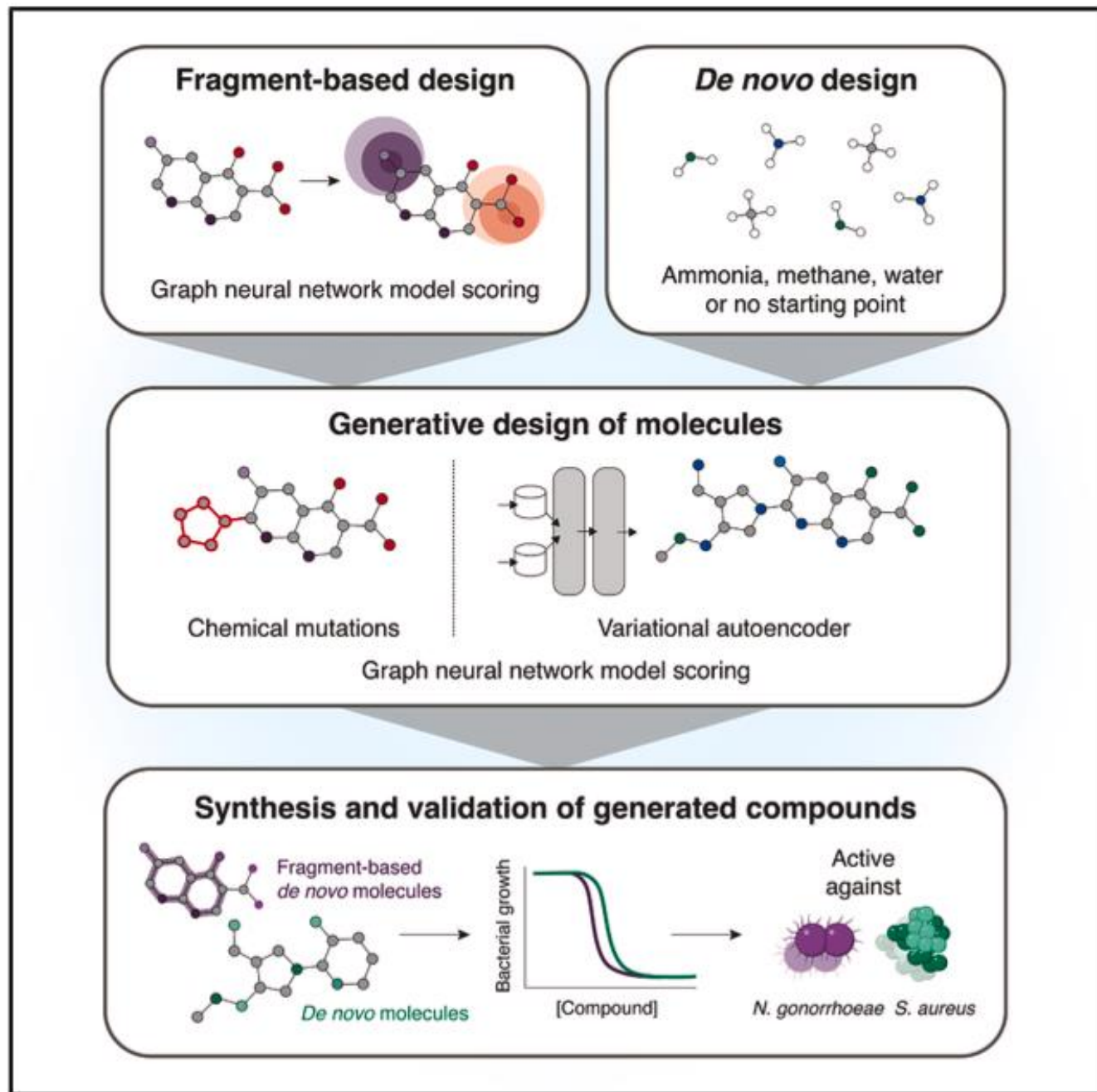
[+ Show authors](#)

Nature **620**, 1089–1100 (2023) | [Cite this article](#)

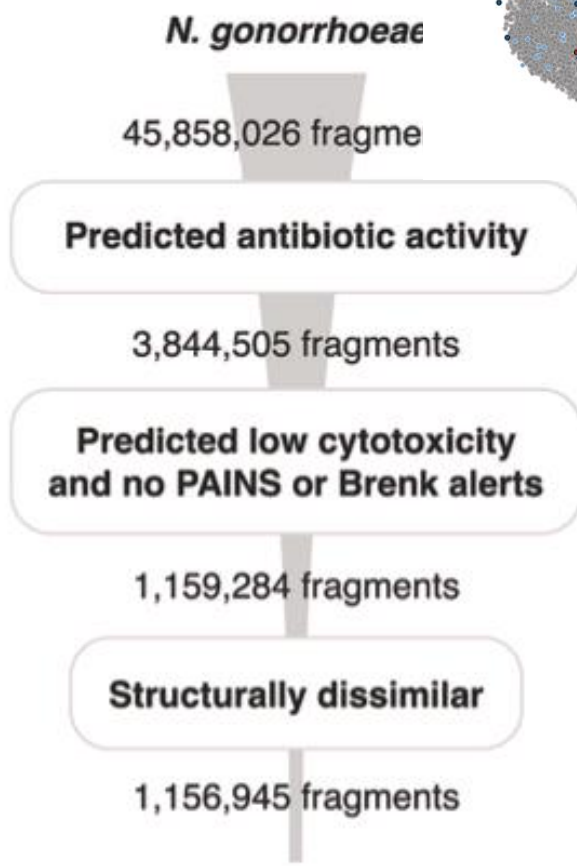
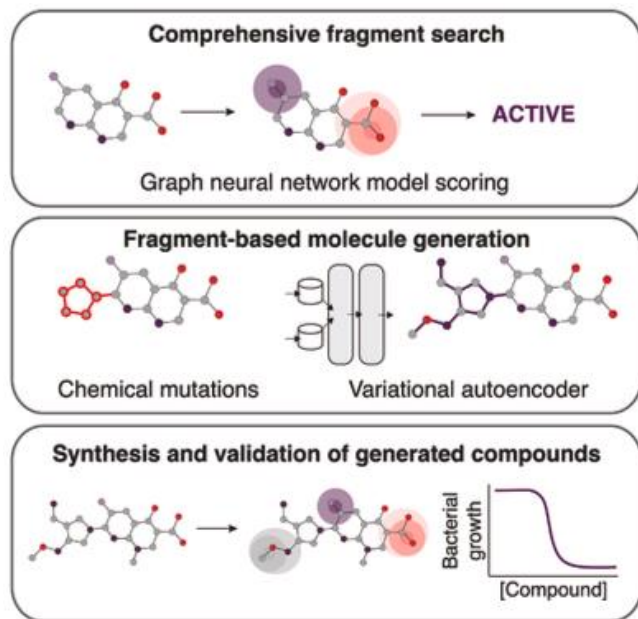
Article

A generative deep learning approach to *de novo* antibiotic design

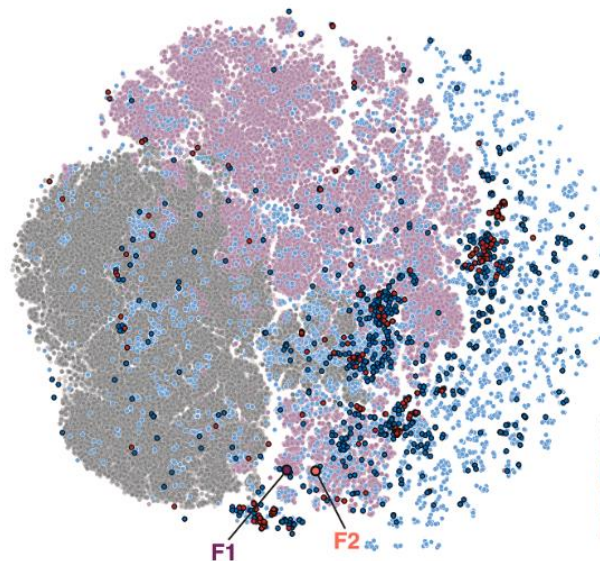
Aarti Krishnan,^{1,2,3,4,25} Melis N. Anahtar,^{1,2,4,5,25} Jacqueline A. Valeri,^{1,2,4,25} Wengong Jin,⁶ Leif Sieben,^{1,2,4,8} Andreas Lutten, ^{1,2} Yu Zhang,^{1,2,4} Seyed Majed Modaresi,^{1,2,4} Andrew H. Parijat Bandyopadhyay,^{1,2} Jonathan C. Chen,^{1,2} Danyal Rehman,¹⁰ Ronak Desai,^{1,11,12} Paige Marie-Stéphanie Aschtgen,¹⁴ Margaux Gaborieau,¹⁴ Massimiliano Gaetani,^{15,16} Samantha Lutete Khonde,¹⁸ Yuri S. Moroz,^{19,20,21} Bruce Blough,²² Chunyang Jin,²² Edmund Loh,^{14,23} Amir Ata Saei,¹⁴ Connor W. Coley,^{9,24} Felix Wong,^{1,2,13} and James J. Collins^{1,2,4,26,*}



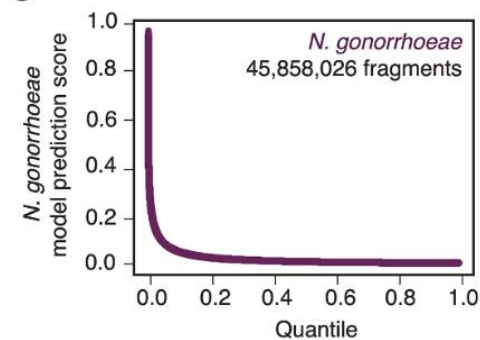
Training the GNN



B



C

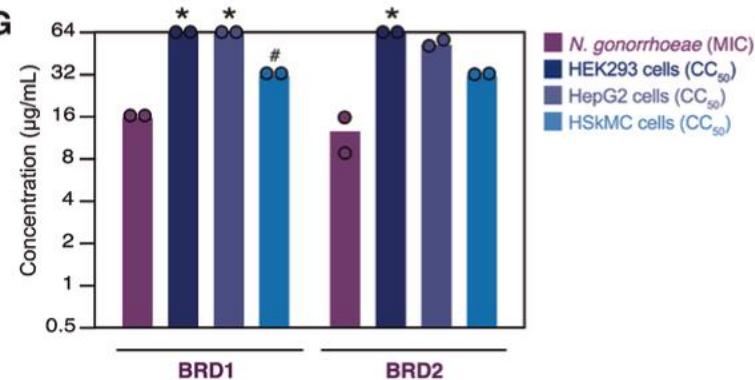


GDB-11: 100,000 fragment sample
 Enamine fragments (including F1 and F2)
 Internal 39K actives
 Internal 39K inactives
 Known antibiotics (559 compounds)

F

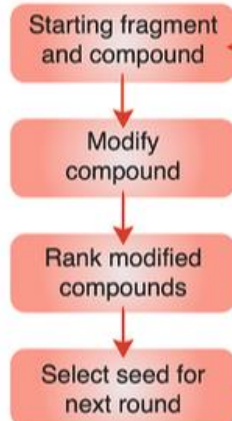


G

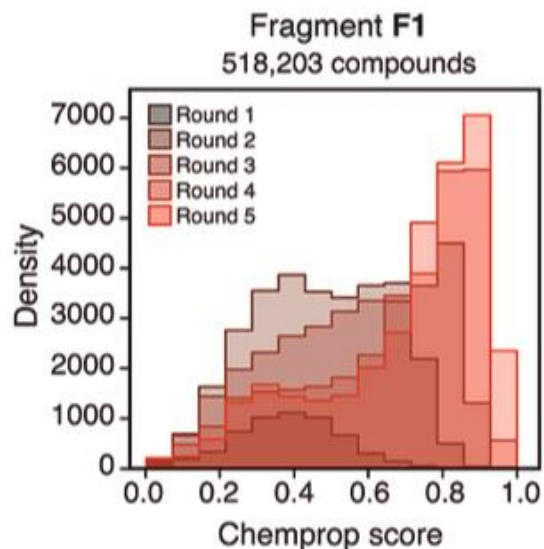
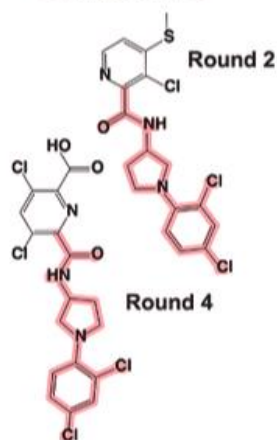


Generative ML algorithms

Genetic algorithm based on F-CReM

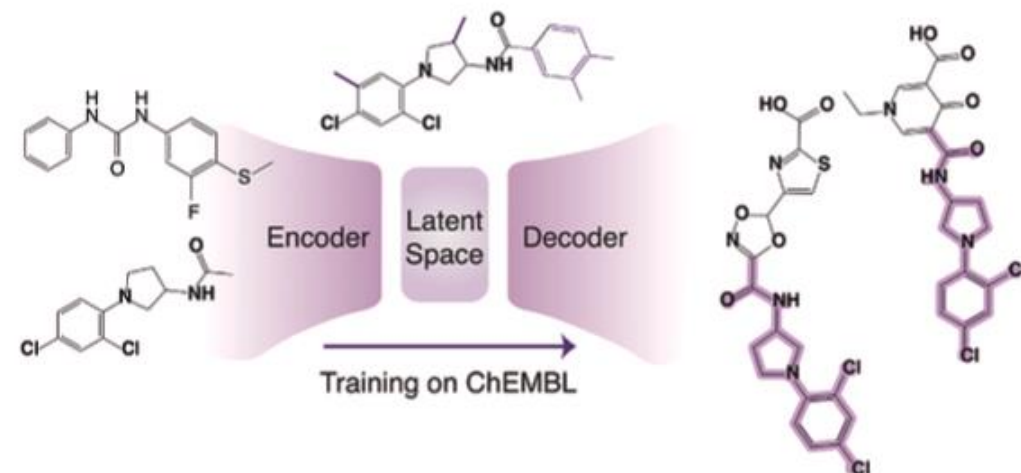


F-CReM generated compounds



Fragment-based variational autoencoder (F-VAE)

F-VAE generated compounds



F-CReM

518,203 molecules

F-VAE

6,937,677 molecules

Predicted antibiotic activity
Predicted low cytotoxicity
Predicted synthesizability
Structurally dissimilar

285 molecules

678 molecules

Highest scoring molecules PS > 0.7
Dissimilar between each other

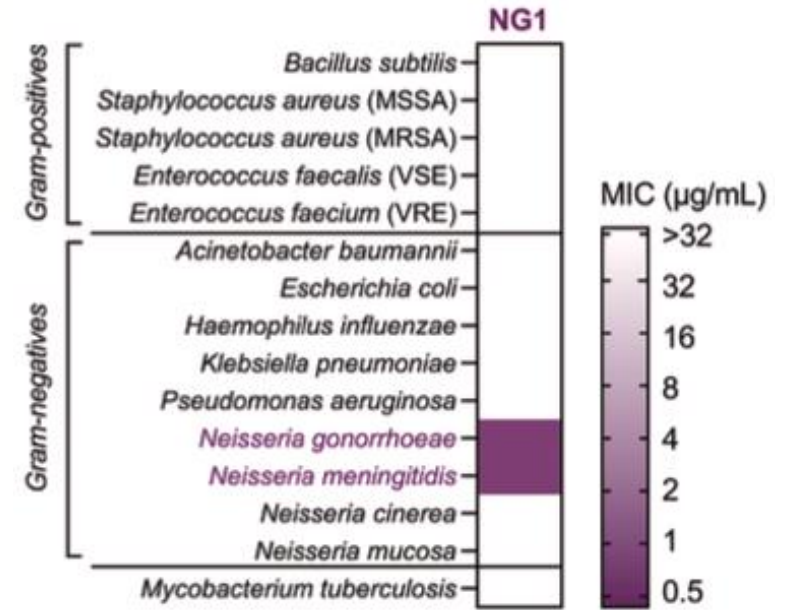
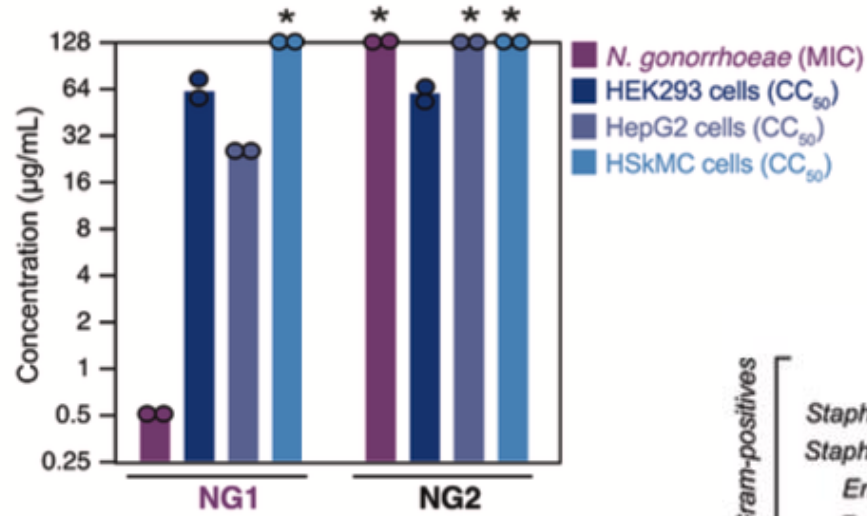
40 molecules

40 molecules

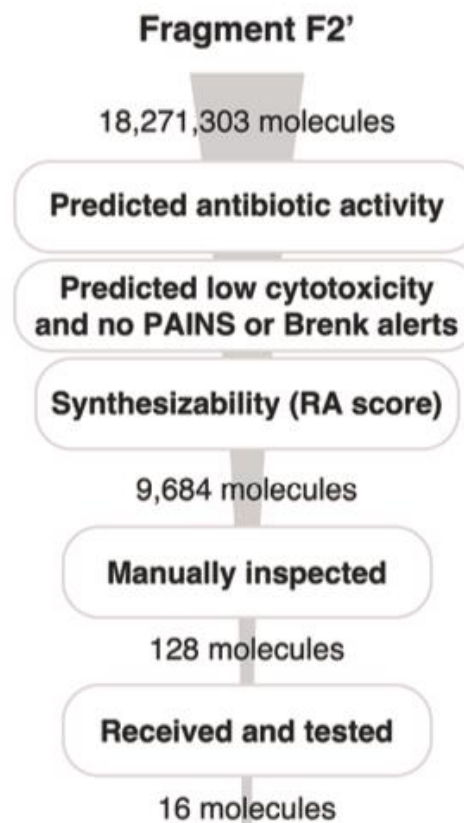
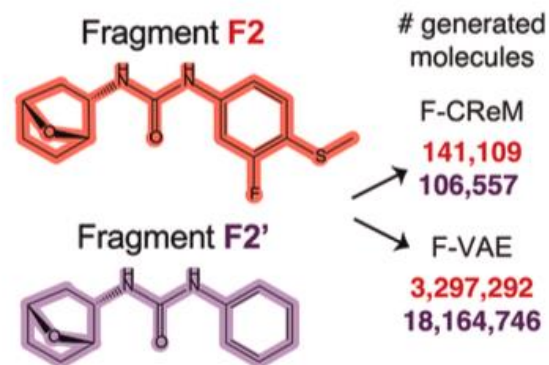
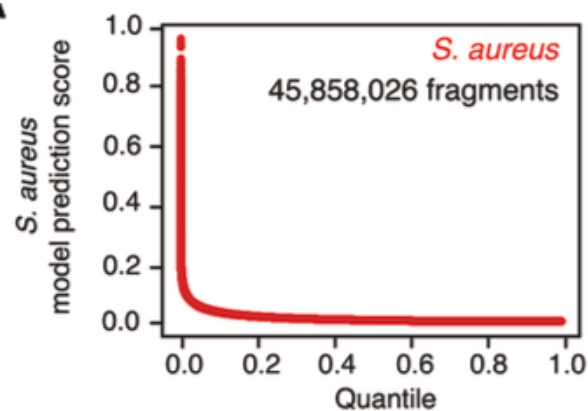
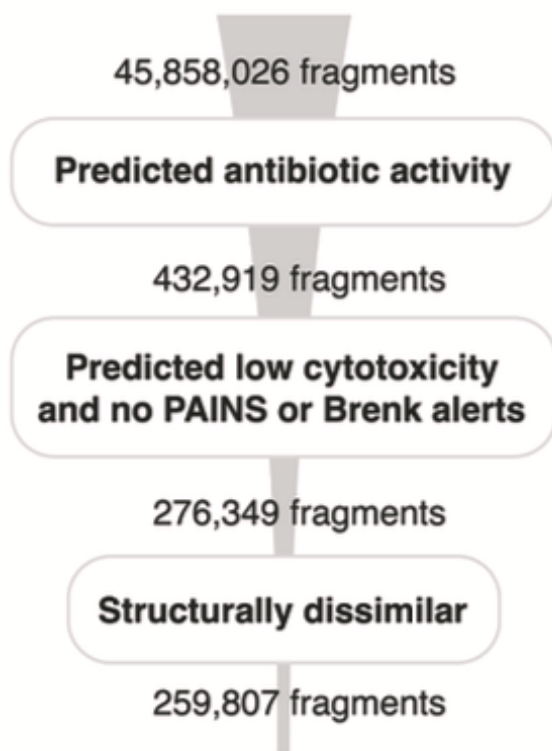
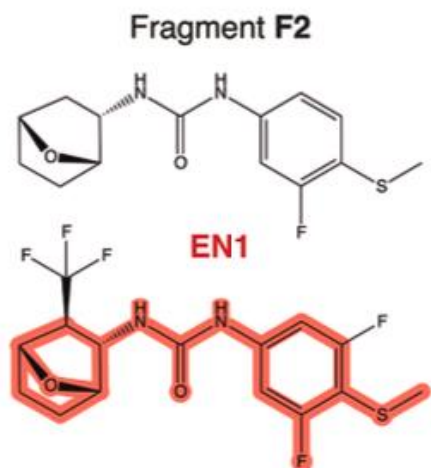
De novo molecules
synthesized in two reaction steps

2 molecules

Synthesis and experimental validation of fragment-based designed compounds



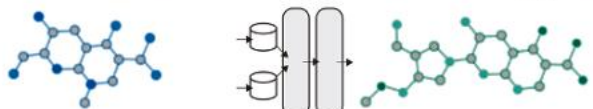
Design of compounds with activity against *S. aureus*



Generation, synthesis, and experimental validation of *de novo* designed compounds

De novo molecule generation

Without the need for a fragment as a starting point



Chemical mutations

Variational autoencoder

Synthesis and validation of generated compounds



Bacterial growth

[Compound]

Genetic algorithm based on CReM

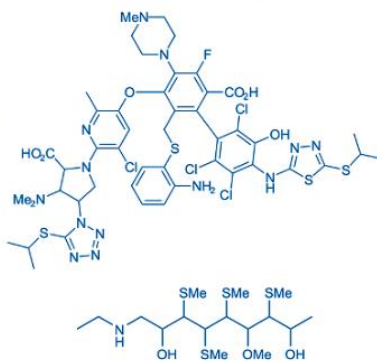
Starting with ammonia, methane or water

Modify compound

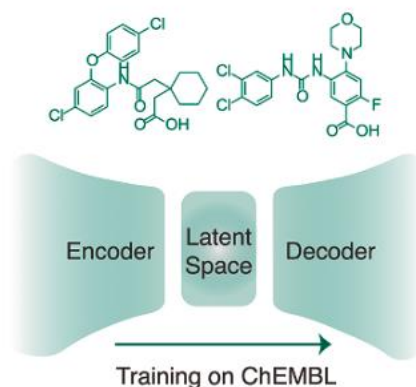
Rank modified compounds

Select seed for next round

CReM generated compounds

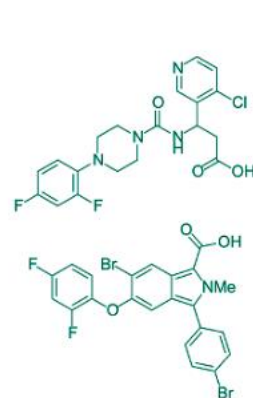


Junction-tree variational autoencoder (JT-VAE)



Training on ChEMBL

JT-VAE generated compounds



JT-VAE

28,534,490 molecules

Predicted antibiotic activity

Structural dissimilarity to known antibiotics

Synthesizability (RAscore)

4,831 molecules

Manually inspected

90 molecules

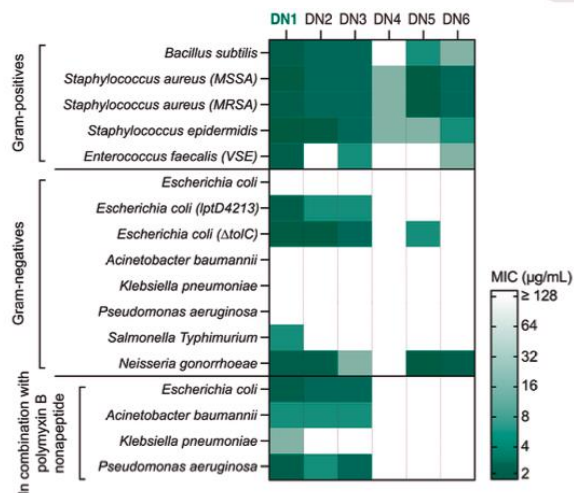
Received and tested

22 molecules

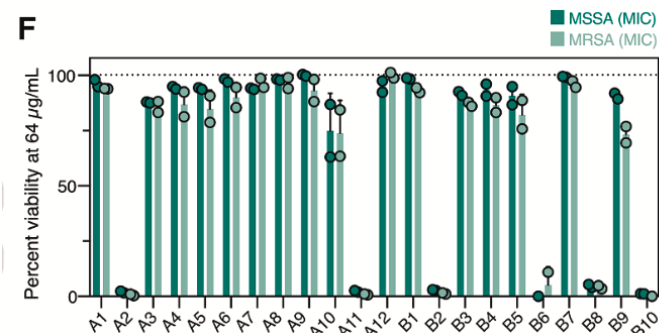
Active at $\leq 64 \mu\text{g/mL}$

6 molecules

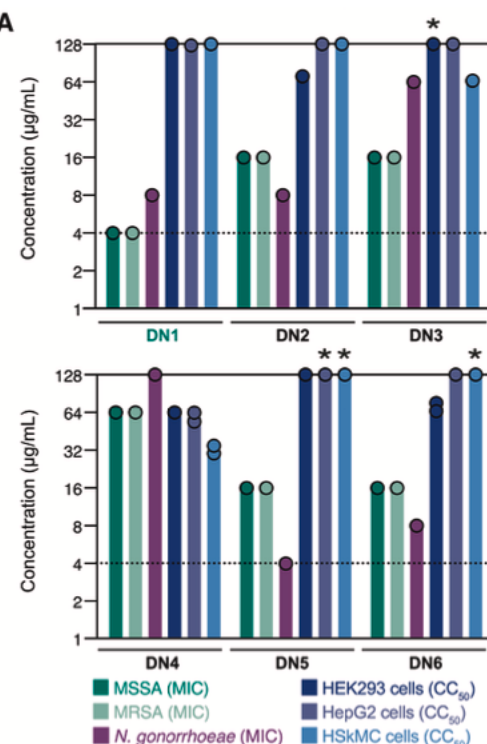
(27.27%)



F



A



Explainable AI

nature protocols

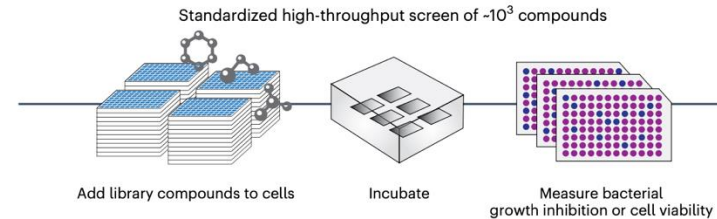
<https://doi.org/10.1038/s41596-024-01084-x>

Protocol

 Check for updates

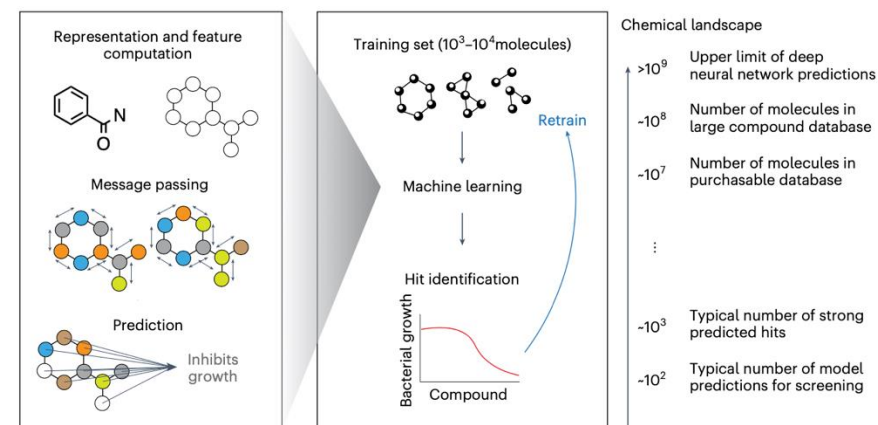
An explainable deep learning platform for molecular discovery

Felix Wong^{1,2,3}, Satotaka Omori^{1,3}, Alicia Li³, Aarti Krishnan^{1,2}, Ryan S. Lach³, Joseph Rufo^{4,5,6,7}, Maxwell Z. Wilson^{3,4,5,6,7} & James J. Collins^{1,2,8} ✉



Stage 1, steps 1–20:
data generation
1 d to 1 week

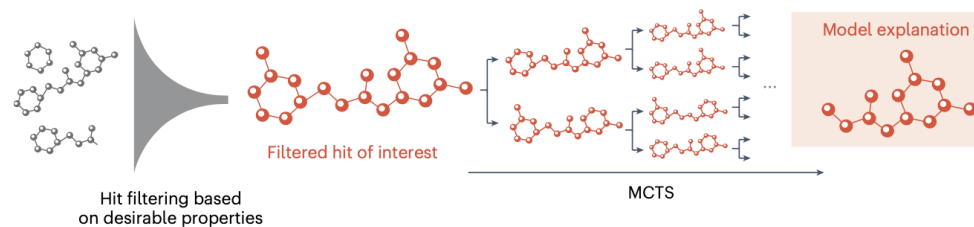
Key output: training data



Stage 2, steps 21–35:
model training and benchmarking
1 h to 1 week

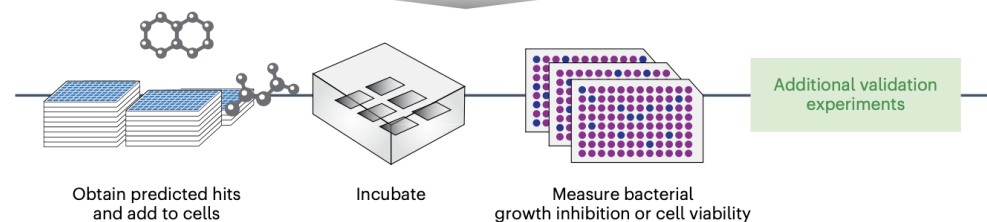
Key output: trained Chemprop models and predictions

Filtering, graph search and structural analysis with rationale explanations



Stage 3, steps 36–44:
rationale analysis and filtering
1 h to 2 d

Key output: chemical structure rationales and filtered predictions



Stage 4, steps 45 and 46:
prediction testing
1 d to 1 week

Not without concerns

Data driven Concerns
“Garbage in garbage out”

Technical & Model-Related Concerns
Generalization, overfitting, black box problem

Biological & Clinical Reality Concerns
Oversimplification of biology, AI is trained on what is known, making finding truly novel targets difficult

Practical & Operational Concerns
Experimental validation must follow, efforts must be made to integrate into traditional workflows

Ethical, Legal, and Regulatory Concerns
Regulatory scrutiny, reproducibility, biosafety & biosecurity