Generative AI for designing and validating easily synthesizable and structurally novel antibiotics

Kyle Swanson*
Stanford University
swansonk@stanford.edu

Gary Liu*
McMaster University
liug34@mcmaster.ca

Denise B. Catacutan
McMaster University
catacud@mcmaster.ca

James Zou
Stanford University
jamesz@stanford.edu

Jonathan M. Stokes
McMaster University
stokesjm@mcmaster.ca

Abstract

The rise of pan-resistant bacteria is creating an urgent need for structurally novel antibiotics. AI methods can discover new antibiotics, but existing methods have significant limitations. Property prediction models, which evaluate molecules one-by-one for a given property, scale poorly to large chemical spaces. Generative models, which directly design molecules, rapidly explore vast chemical spaces but generate molecules that are challenging to synthesize. Here, we introduce SyntheMol, a generative model that designs easily synthesizable compounds from a chemical space of 30 billion molecules. We apply SyntheMol to design molecules that inhibit the growth of *Acinetobacter baumannii*, a burdensome bacterial pathogen. We synthesize 58 generated molecules and experimentally validate them, with six structurally novel molecules demonstrating potent activity against *A. baumannii* and several other phylogenetically diverse bacterial pathogens.

1 Introduction

The global dissemination of antibiotic resistance determinants is one of the most significant challenges of modern medicine. In 2019, an estimated 4.95 million deaths were associated with drug-resistant infections [15], and this number is projected to grow to 10 million per year by 2050. Six bacterial species known as the ESKAPE pathogens are especially virulent and drug-resistant [18]. One of them, the Gram-negative bacterium *Acinetobacter baumannii*, is particularly burdensome in clinical settings and is a critical priority according to the World Health Organization [22]. Structurally and functionally novel antibiotics are urgently needed to address this problematic pathogen [11].

Artificial intelligence (AI) methods have shown that they can rapidly and accurately identify promising drug candidates, including antibiotics [21, 5]. For example, property prediction AI models can evaluate chemical libraries to identify compounds with desirable properties [8]. An important limitation is that these models evaluate molecules one-by-one, which is time consuming for large chemical spaces. In contrast, generative models build molecules from scratch, assembling molecular fragments into larger molecules with the desired property [2]. Generative models thus directly design promising molecules without a slow evaluation of many compounds.

However, a major limitation of generative models is that the generated compounds are often synthetically intractable [6], thereby preventing experimental validation. Although some generative methods that incorporate synthesizability have been proposed with promising *in silico* results [3, 4, 9, 7, 17],

*Denotes co-first author.

NeurIPS 2023 Generative AI and Biology Workshop.

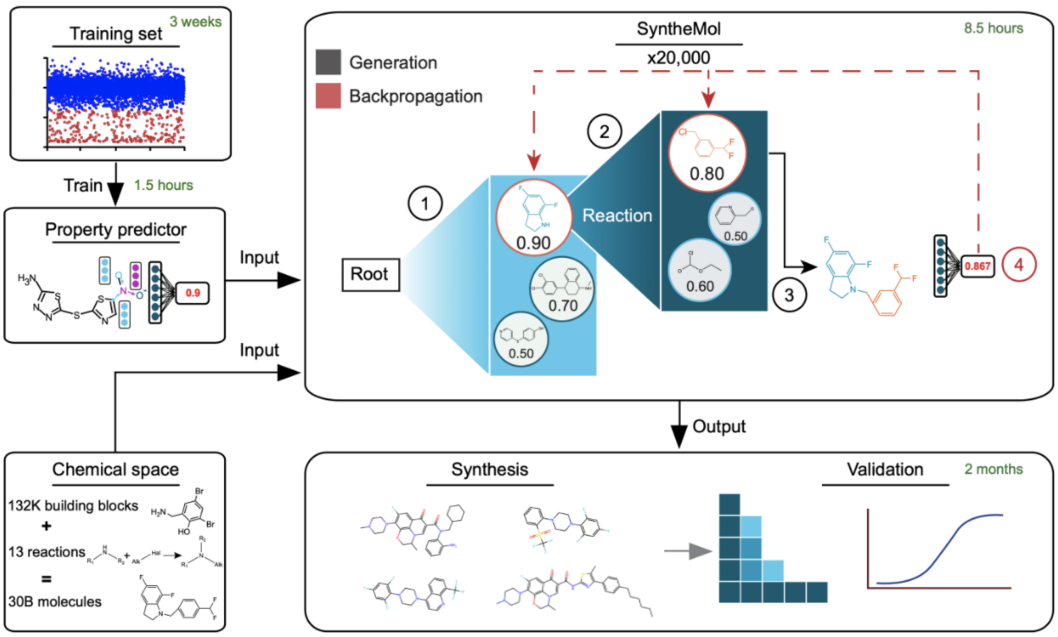

Figure 1: An overview of SyntheMol, our generative AI method for designing novel antibiotics.

very few studies synthesized and experimentally tested any generated molecules [2], even though such experiments are necessary to validate the generative method. Moreover, none of these methods have been applied to antibiotic development, which urgently requires state-of-the-art approaches.

In this study, we developed SyntheMol, a generative AI model that uses a Monte Carlo tree search to assemble novel compounds using ∼132,000 molecular building blocks with known reactivities and 13 well-validated chemical synthesis reactions (Figure 1). The resulting chemical space contains nearly 30 billion molecules that are easy to synthesize, with synthesis success rates of over 80% within 3 to 4 weeks. We trained SyntheMol to design molecules with antibiotic activity against *A. baumannii*, and we synthesized and experimentally validated 58 generated molecules, with six showing potent activity against *A. baumannii* and several other phylogenetically diverse bacterial pathogens.

2 Property Prediction Models

To establish a training dataset, we physically screened 13,524 molecules and measured growth inhibition of *A. baumannii* ATCC 17978 when treated with each chemical, resulting in 470 active compounds and 13,054 inactive compounds. Using this dataset, we built three models for predicting antibacterial activity against *A. baumannii*. **Chemprop** is a directed message passing neural network (MPNN) for molecular property prediction [25]. **Chemprop-RDKit** is a variant of Chemprop that concatenates the MPNN embedding with a set of 200 molecular features computed by RDKit prior to the feed-forward layers. **Random Forest** is a random forest classifier that uses the 200 RDKit features as input to 100 decision trees. We trained each model type on our training dataset using 10-fold cross-validation with splits containing 80% train, 10% validation, and 10% test data. All three model types trained in less than 90 minutes on 16 CPU cores and performed similarly, with ROC-AUCs in the range 0.80–0.84 and PRC-AUCs in the range 0.35–0.40. During molecule generation, we treated the models from the 10 cross-validation folds as an ensemble of 10 models.

3 SyntheMol

We designed SyntheMol, a generative model that builds easily synthesizable molecules from a combinatorial chemical space, which consists of readily purchasable molecular building blocks along with well-validated chemical reactions that combine two or three building blocks.

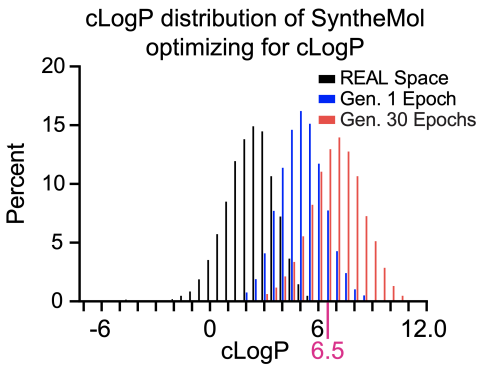

Figure 2: cLogP for random REAL molecules and molecules generated by SyntheMol with a Chemprop predictor for cLogP trained for 1 or 30 epochs. The target is cLogP > 6.5.

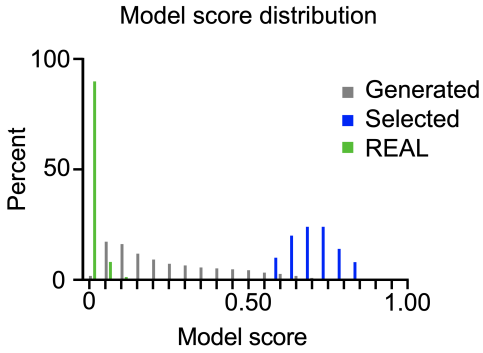

Figure 3: Chemprop antibacterial prediction scores on the generated or selected (filtered) compounds and on random REAL molecules.

Combinatorial Chemical Space. The combinatorial chemical space we use is the Enamine REadily AccessibLe (REAL) Space [10]. The REAL Space consists of 31 billion single-reaction molecules produced by applying 169 chemical reactions to 138,085 molecular building blocks (132,479 unique molecules). For simplicity, we restricted SyntheMol to use 13 of the most common REAL reactions, which account for 29.6 billion molecules (93.9% of REAL Space).

SyntheMol MCTS Algorithm. SyntheMol uses a Monte Carlo tree search (MCTS) [20] guided by a property prediction model to search through a vast combinatorial chemical space for molecules with promising molecular properties (Figure S1). SyntheMol works by repeatedly performing MCTS rollouts, during which it constructs a molecule using the 132,479 building blocks and 13 chemical reactions and then evaluates the constructed molecule using a molecular property prediction model. Nodes in the search tree represent one or more building blocks or full molecules. Each step of MCTS involves scoring each potential next node N according to $S(N) = \frac{Q(N)+P(N)\cdot U(N)}{D(N)}$ where $Q(N)$ is the exploit score that prioritizes nodes leading to high scoring molecules, $P(N)$ is the average property prediction score of molecules in the node, $U(N)$ is the explore score that prioritizes unexplored nodes, and $D(N)$ is the diversity penalty that encourages use of different building blocks (see Appendix B). During the rollouts, SyntheMol updates these scores as it learns which building blocks and chemical reactions produce molecules with high property prediction scores.

4 Generation Results

Prior to running SyntheMol for antibiotic discovery, we evaluated it *in silico* using cLogP, the computed octanol-water partition coefficient [24].

Generating cLogP Molecules. To create a cLogP training set, we computed cLogP values for the 13,524 molecules in our antibiotics training set and binarized them with cLogP > 6.5 as "active," resulting in 495 (3.7%) active molecules. We trained a cLogP Chemprop model (ROC-AUC = 0.97, PRC-AUC = 0.74) and ran SyntheMol with this model for 20,000 rollouts ($\sim$9 hours). Among the 25,550 generated molecules, 61.42% were active, representing a 1,396x increase in hit rate compared to 0.044% active molecules in a random sample of 25,000 molecules from REAL Space (Figure 2). Even when using a weaker cLogP model trained for one epoch instead of 30 to better match the antibiotic model's performance (PRC-AUC = 0.20 cLogP vs 0.35 antibiotic), 11.78% of the generated molecules were active, representing a 268x increase in hit rate. These results demonstrate that SyntheMol can rapidly and effectively search a huge combinatorial space for active molecules.

Generating Antibiotics. We next applied SyntheMol to discover potential antibiotic candidates against *A. baumannii* by using our three antibiotic property predictors. Since the results were similar for all three models, we present the results for SyntheMol with Chemprop. Over the course of 20,000 rollouts ($\sim$8.5 hours), SyntheMol with Chemprop evaluated 452 million intermediate nodes containing

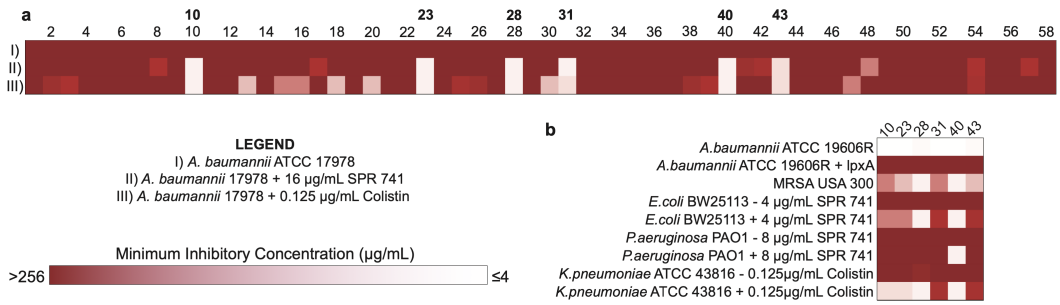

Figure 4: **(a)** A heat map summarizing the minimum inhibitory concentrations (MIC) of the 58 synthesized molecules generated by SyntheMol against *A. baumannii* ATCC 17978 with or without a permeabilzation agent. Bold numbers indicate molecules with high activity (MIC $\leq 8\mu$g/mL). **(b)** Six molecules with high activity were additionally tested against a panel of ESKAPE species.

diverse combinations of molecular building blocks and generated 24,335 complete molecules. The generated molecules had significantly higher Chemprop antibacterial property prediction scores than a random sample of REAL molecules, with 2,868 (12%) of the 24,335 generated molecules scoring ≥ 0.5 compared to 1 (0.004%) of 25,000 random REAL molecules (Figure 3). Notably, the generated molecules included a diverse set of 10,846 building blocks and all 13 reactions.

Filtering Antibiotics. Since we aimed to select a diverse set of structurally novel molecules with high property prediction scores for experimental validation, we developed a set of three filters. First, to ensure structural novelty of the generated molecules, we computed the Tversky similarity [23] between Morgan fingerprints [19] of each generated molecule and all of the 470 active training set molecules as well as 1,005 antibacterial molecules in the ChEMBL database [13] and removed molecules with a Tversky similarity > 0.5. Second, to obtain effective molecules, we kept only the molecules with the top 20% of prediction scores. Third, to select structurally diverse molecules, we applied k-means clustering [1] using Tanimoto distance [12] between the Morgan fingerprints of the remaining molecules to obtain 50 clusters of molecules. We selected the highest scoring molecule in each cluster for a total of 50 molecules. Applying this filtering to the generated set from each of the three property prediction models resulted in 150 molecules for experimental validation.

5 *In Vitro* Validation of Generated Molecules

We next aimed to synthesize the generated compounds and validate their bioactivity. Among the 150 compounds, 70 were available from Enamine. Of those, 58 (83%) were synthesized in 4 weeks with 26 molecules from Chemprop, 22 from Chemprop-RDKit, and 10 from random forest.

***A. Baumannii* Validation.** We validate those 58 synthesized molecules by performing growth inhibition assays against *A. baumannii* ATCC 17978, the same strain used for training set curation. Because *A. baumannii* is a Gram-negative species with challenging permeability characteristics due to its highly impermeable outer membrane [16], we added low concentrations of the permeabilization agents SPR 741 [26] or colistin [14]. Results reveal exceptional antibacterial activity of six molecules, as defined by a minimum inhibitory concentration (MIC) ≤ 8 μg/mL, when combined with $\frac{1}{4}$ MIC SPR 741 or $\frac{1}{4}$ MIC colistin (Figure 4a). This represents a remarkable 10% hit rate. As a control, we tested 58 randomly selected molecules from the Enamine REAL Space. None of these compounds displayed antibacterial activity against *A. baumannii* ATCC 17978, either alone or when combined with SPR 741, as defined by our threshold of an MIC ≤ 8 μg/mL.

Broad-Spectrum Validation. To assess broad-spectrum activity, we tested these six compounds against the Gram-negative species *Escherichia coli* BW25113, *Pseudomonas aeruginosa* PAO1, *Klebsiella pneumoniae* ATCC 43816, *A. baumannii* ATCC 19606R (a lipopolysaccharide-deficient polymyxin-resistant mutant), and the Gram-positive species *Staphylococcus aureus* USA 300. Remarkably, when used with an appropriate permeabilization agent (if needed), all six compounds

displayed broad-spectrum antibacterial activity against all species except *P. aeruginosa*, which is likely due to the impermeability commonly displayed by the cell envelope of this species (Figure 4b).

6 Conclusion

We developed SyntheMol, a novel generative AI model for small molecule drug design that uses molecular property prediction models in conjunction with MCTS to explore a vast combinatorial chemical space for promising molecules. We applied SyntheMol to design antibacterial compounds against *A. baumannii*, and we synthesized and experimentally tested 58 structurally novel and diverse generated compounds. Remarkably, six molecules had activity against *A. baumannii* and other phylogenetically diverse ESKAPE species. This work demonstrates the utility of generative AI to design structurally novel, synthetically tractable, and effective small molecule antibiotic candidates.

References

[1] David Arthur and Sergei Vassilvitskii. K-Means++: The Advantages of Careful Seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pages 1027–1035, USA, 2007. Society for Industrial and Applied Mathematics. event-place: New Orleans, Louisiana.

[2] Camille Bilodeau, Wengong Jin, Tommi Jaakkola, Regina Barzilay, and Klavs F. Jensen. Generative models for molecular discovery: Recent advances and challenges. *WIREs Computational Molecular Science*, 12(5), September 2022.

[3] John Bradshaw, Brooks Paige, Matt J. Kusner, Marwin H. S. Segler, and José Miguel Hernández-Lobato. A Model to Search for Synthesizable Molecules. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[4] John Bradshaw, Brooks Paige, Matt J. Kusner, Marwin H. S. Segler, and José Miguel Hernández-Lobato. Barking up the Right Tree: An Approach to Search over Molecule Synthesis DAGs. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. event-place: Vancouver, BC, Canada.

[5] Paula Carracedo-Reboredo, Jose Liñares-Blanco, Nereida Rodríguez-Fernández, Francisco Cedrón, Francisco J. Novoa, Adrian Carballal, Victor Maojo, Alejandro Pazos, and Carlos Fernandez-Lozano. A review on machine learning approaches and trends in drug discovery. *Computational and Structural Biotechnology Journal*, 19:4538–4558, 2021.

[6] Wenhao Gao and Connor W. Coley. The Synthesizability of Molecules Proposed by Generative Models. *Journal of Chemical Information and Modeling*, 60(12):5714–5723, December 2020.

[7] Wenhao Gao, Rocío Mercado, and Connor W. Coley. Amortized Tree Generation for Bottom-up Synthesis Planning and Synthesizable Molecular Design. In *International Conference on Learning Representations*, 2022.

[8] Thomas Gaudelet, Ben Day, Arian R Jamasb, Jyothish Soman, Cristian Regep, Gertrude Liu, Jeremy B R Hayter, Richard Vickers, Charles Roberts, Jian Tang, David Roblin, Tom L Blundell, Michael M Bronstein, and Jake P Taylor-King. Utilizing graph machine learning within drug discovery and development. *Briefings in Bioinformatics*, 22(6):bbab159, November 2021.

[9] Sai Krishna Gottipati, Boris Sattarov, Sufeng Niu, Yashaswi Pathak, Haoran Wei, Shengchao Liu, Karam J. Thomas, Simon Blackburn, Connor W. Coley, Jian Tang, Sarath Chandar, and Yoshua Bengio. Learning to Navigate the Synthetically Accessible Chemical Space Using Reinforcement Learning. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[10] Oleksandr O. Grygorenko, Dmytro S. Radchenko, Igor Dziuba, Alexander Chuprina, Kateryna E. Gubina, and Yurii S. Moroz. Generating Multibillion Chemical Space of Readily Accessible Screening Compounds. *iScience*, 23(11):101681, November 2020.

[11] Chang-Ro Lee, Jung Hun Lee, Moonhee Park, Kwang Seung Park, Il Kwon Bae, Young Bae Kim, Chang-Jun Cha, Byeong Chul Jeong, and Sang Hee Lee. Biology of Acinetobacter baumannii: Pathogenesis, Antibiotic Resistance Mechanisms, and Prospective Treatment Options. *Frontiers in Cellular and Infection Microbiology*, 7, March 2017.

[12] Gerald Maggiora, Martin Vogt, Dagmar Stumpfe, and Jürgen Bajorath. Molecular Similarity in Medicinal Chemistry: Miniperspective. *Journal of Medicinal Chemistry*, 57(8):3186–3204, April 2014.

[13] David Mendez, Anna Gaulton, A Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F Mosquera, Prudence Mutowo, Michał Nowotka, María Gordillo-Marañón, Fiona Hunter, Laura Junco, Grace Mugumbate, Milagros Rodriguez-Lopez, Francis Atkinson, Nicolas Bosc, Chris J Radoux, Aldo Segura-Cabrera, Anne Hersey, and Andrew R Leach. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research*, 47(D1):D930–D940, January 2019.

[14] Jennifer H. Moffatt, Marina Harper, Paul Harrison, John D. F. Hale, Evgeny Vinogradov, Torsten Seemann, Rebekah Henry, Bethany Crane, Frank St. Michael, Andrew D. Cox, Ben Adler, Roger L. Nation, Jian Li, and John D. Boyce. Colistin Resistance in *Acinetobacter baumannii* Is Mediated by Complete Loss of Lipopolysaccharide Production. *Antimicrobial Agents and Chemotherapy*, 54(12):4971–4977, December 2010.

[15] Christopher JL Murray, Kevin Shunji Ikuta, Fablina Sharara, Lucien Swetschinski, Gisela Robles Aguilar, Authia Gray, Chieh Han, Catherine Bisignano, Puja Rao, Eve Wool, Sarah C Johnson, Annie J Browne, Michael Give Chipeta, Frederick Fell, Sean Hackett, Georgina Haines-Woodhouse, Bahar H Kashef Hamadani, Emmanuelle A P Kumaran, Barney McManigal, Ramesh Agarwal, Samuel Akech, Samuel Albertson, John Amuasi, Jason Andrews, Aleskandr Aravkin, Elizabeth Ashley, Freddie Bailey, Stephen Baker, Buddha Basnyat, Adrie Bekker, Rose Bender, Adhisivam Bethou, Julia Bielicki, Suppawat Boonkasidecha, James Bukosia, Cristina Carvalheiro, Carlos Castañeda-Orjuela, Vilada Chansamouth, Suman Chaurasia, Sara Chiurchiù, Fazle Chowdhury, Aislinn J Cook, Ben Cooper, Tim R Cressey, Elia Criollo-Mora, Matthew Cunningham, Saffiatou Darboe, Nicholas P J Day, Maia De Luca, Klara Dokova, Angela Dramowski, Susanna J Dunachie, Tim Eckmanns, Daniel Eibach, Amir Emami, Nicholas Feasey, Natasha Fisher-Pearson, Karen Forrest, Denise Garrett, Petra Gastmeier, Ababi Zergaw Giref, Rachel Claire Greer, Vikas Gupta, Sebastian Haller, Andrea Haselbeck, Simon I Hay, Marianne Holm, Susan Hopkins, Kenneth C Iregbu, Jan Jacobs, Daniel Jarovsky, Fatemeh Javanmardi, Meera Khorana, Niranjan Kissoon, Elsa Kobeissi, Tomislav Kostyanev, Fiorella Krapp, Ralf Krumkamp, Ajay Kumar, Hmwe Hmwe Kyu, Cherry Lim, Direk Limmathurotsakul, Michael James Loftus, Miles Lunn, Jianing Ma, Neema Mturi, Tatiana Munera-Huertas, Patrick Musicha, Marisa Marcia Mussi-Pinhata, Tomoka Nakamura, Ruchi Nanavati, Sushma Nangia, Paul Newton, Chanpheaktra Ngoun, Amanda Novotney, Davis Nwakanma, Christina W Obiero, Antonio Olivas-Martinez, Piero Olliaro, Ednah Ooko, Edgar Ortiz-Brizuela, Anton Yariv Peleg, Carlo Perrone, Nishad Plakkal, Alfredo Ponce-de Leon, Mathieu Raad, Tanusha Ramdin, Amy Riddell, Tamalee Roberts, Julie Victoria Robotham, Anna Roca, Kristina E Rudd, Neal Russell, Jesse Schnall, John Anthony Gerard Scott, Madhusudhan Shivamallappa, Jose Sifuentes-Osornio, Nicolas Steenkeste, Andrew James Stewardson, Temenuga Stoeva, Nidanuch Tasak, Areerat Thaiprakong, Guy Thwaites, Claudia Turner, Paul Turner, H Rogier van Doorn, Sithembiso Velaphi, Avina Vongpradith, Huong Vu, Timothy Walsh, Seymour Waner, Tri Wangrangsimakul, Teresa Wozniak, Peng Zheng, Benn Sartorius, Alan D Lopez, Andy Stergachis, Catrin Moore, Christiane Dolecek, and Mohsen Naghavi. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet*, 399(10325):629–655, February 2022.

[16] Hiroshi Nikaido. Molecular Basis of Bacterial Outer Membrane Permeability Revisited. *Microbiology and Molecular Biology Reviews*, 67(4):593–656, December 2003.

[17] Aryan Pedawi, Pawel Gniewek, Chaoyi Chang, Brandon M. Anderson, and Henry van den Bedem. An efficient graph generative model for navigating ultra-large combinatorial synthesis libraries. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[18] Louis B. Rice. Federal Funding for the Study of Antimicrobial Resistance in Nosocomial Pathogens: No ESKAPE. *The Journal of Infectious Diseases*, 197(8):1079–1081, April 2008.

[19] David Rogers and Mathew Hahn. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, May 2010.

[20] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, January 2016.

[21] Jonathan M. Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M. Donghia, Craig R. MacNair, Shawn French, Lindsey A. Carfrae, Zohar Bloom-Ackermann, Victoria M. Tran, Anush Chiappino-Pepe, Ahmed H. Badran, Ian W. Andrews, Emma J. Chory, George M. Church, Eric D. Brown, Tommi S. Jaakkola, Regina Barzilay, and James J. Collins. A Deep Learning Approach to Antibiotic Discovery. *Cell*, 180(4):688–702.e13, February 2020.

[22] Evelina Tacconelli, Elena Carrara, Alessia Savoldi, Stephan Harbarth, Marc Mendelson, Dominique L Monnet, Céline Pulcini, Gunnar Kahlmeter, Jan Kluytmans, Yehuda Carmeli, Marc Ouellette, Kevin Outterson, Jean Patel, Marco Cavaleri, Edward M Cox, Chris R Houchens, M Lindsay Grayson, Paul Hansen, Nalini Singh, Ursula Theuretzbacher, Nicola Magrini, Aaron Oladipo Aboderin, Seif Salem Al-Abri, Nordiah Awang Jalil, Nur Benzonana, Sanjay Bhattacharya, Adrian John Brink, Francesco Robert Burkert, Otto Cars, Giuseppe Cornaglia, Oliver James Dyar, Alex W Friedrich, Ana C Gales, Sumanth Gandra, Christian Georg Giske, Debra A Goff, Herman Goossens, Thomas Gottlieb, Manuel Guzman Blanco, Waleria Hryniewicz, Deepthi Kattula, Timothy Jinks, Souha S Kanj, Lawrence Kerr, Marie-Paule Kieny, Yang Soo Kim, Roman S Kozlov, Jaime Labarca, Ramanan Laxminarayan, Karin Leder, Leonard Leibovici, Gabriel Levy-Hara, Jasper Littman, Surbhi Malhotra-Kumar, Vikas Manchanda, Lorenzo Moja, Babacar Ndoye, Angelo Pan, David L Paterson, Mical Paul, Haibo Qiu, Pilar Ramon-Pardo, Jesús Rodríguez-Baño, Maurizio Sanguinetti, Sharmila Sengupta, Mike Sharland, Massinissa Si-Mehand, Lynn L Silver, Wonkeung Song, Martin Steinbakk, Jens Thomsen, Guy E Thwaites, Jos WM van der Meer, Nguyen Van Kinh, Silvio Vega, Maria Virginia Villegas, Agnes Wechsler-Fördös, Heiman Frank Louis Wertheim, Evelyn Wesangula, Neil Woodford, Fidan O Yilmaz, and Anna Zorzet. Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. *The Lancet Infectious Diseases*, 18(3):318–327, March 2018.

[23] Amos Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, July 1977.

[24] Scott A. Wildman and Gordon M. Crippen. Prediction of Physicochemical Parameters by Atomic Contributions. *Journal of Chemical Information and Computer Sciences*, 39(5):868–873, September 1999.

[25] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Analyzing Learned Molecular Representations for Property Prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, August 2019.

[26] Daniel V. Zurawski, Alexandria A. Reinhart, Yonas A. Alamneh, Michael J. Pucci, Yuanzheng Si, Rania Abu-Taleb, Jonathan P. Shearer, Samandra T. Demons, Stuart D. Tyner, and Troy Lister. SPR741, an Antibiotic Adjuvant, Potentiates the *In Vitro* and *In Vivo* Activity of Rifampin against Clinically Relevant Extensively Drug-Resistant Acinetobacter baumannii. *Antimicrobial Agents and Chemotherapy*, 61(12):e01239–17, December 2017.

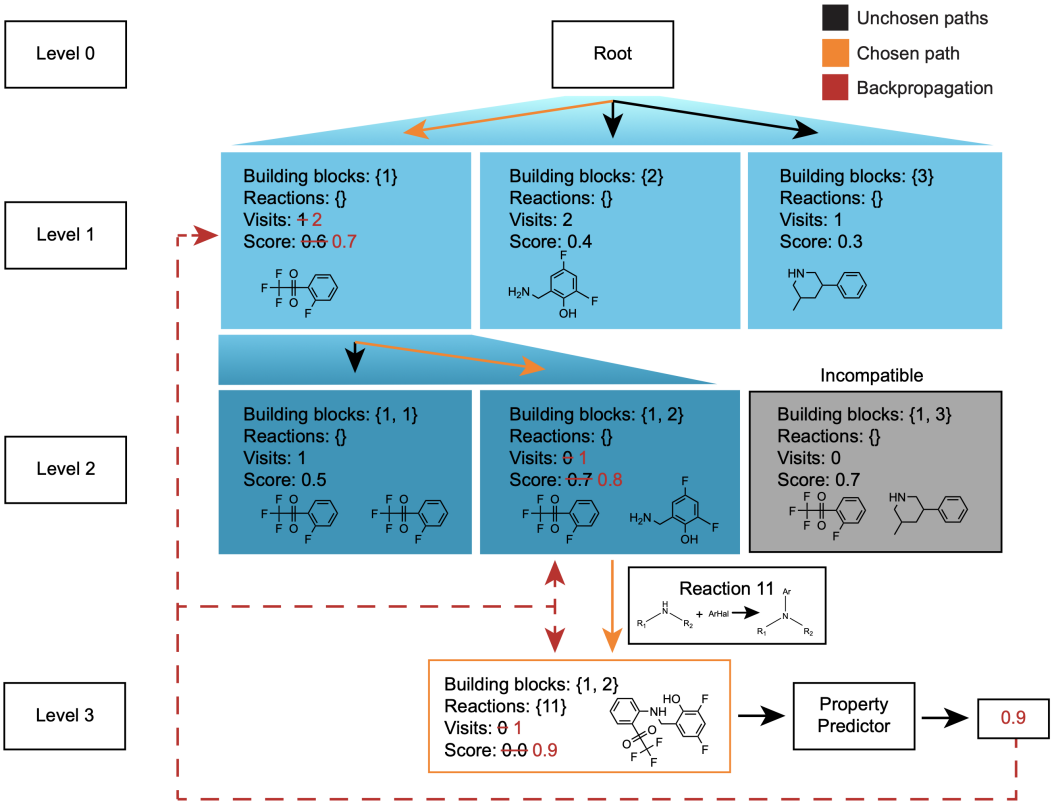

Figure S1: SyntheMol uses a Monte Carlo tree search to generate molecules with desired molecular properties by combining building blocks with chemical reactions.

Appendix

A Code and Data

Code and data are available in this Google Drive folder: `https://drive.google.com/drive/folders/15UBuo9O6UjPuN9e1Qi7oX9vzG6cXA_cv`

B SyntheMol Monte Carlo Tree Search

Below, we describe the SyntheMol Monte Carlo tree search (MCTS) algorithm in detail. Table 1 summarizes the notation used below and Figure S1 illustrates the algorithm.

Let C be the set of all molecules. We assume that we have a set of molecular building blocks $B \subset C$, which are molecules that are small and easy to purchase from commercial vendors. We also have a set of chemical reactions R where each chemical reaction $r \in R$ combines two or more molecules into a single molecule (ignoring byproducts and catalysts). We then build a property predictor $M : C \rightarrow \mathbb{R}$, which is a function, such as a neural network, that predicts a property of a molecule. In our case, B is a set of 132,479 REAL molecular building blocks, R is a set of 13 REAL chemical reactions, and M is either a Chemprop, Chemprop-RDKit, or random forest model that is trained to predict A. baumannii growth inhibition, with prediction values in the range $[0, 1]$.

We define a synthesis tree T that represents every possible synthetic route that creates a molecule using molecular building blocks from B and chemical reactions from R. The tree consists of a set of nodes $N \in T$, each of which represents a discrete step in the synthetic route. Specifically, each node N contains a set of one or more molecules $N_{mols} \subset C$, which are either building blocks from B or molecules that can be produced from building blocks in B and reactions in R. All molecules

Name	Notation	Type	Description
Chemical space	C	Set of molecules	The set of all molecules.
Building blocks	B	$B \subset C$	A set of building block molecules, which are molecules that are small and easy to purchase.
Chemical reactions	R	Set of chemical reactions	A set of chemical reactions that combine two or more molecules in C into a single molecule in C (ignoring byproducts and catalysts).
Property predictor	M	$M : C \to \mathbb{R}$	A function, such as a neural network, that predicts a property of a molecule.
Synthesis tree	T	Set of nodes	A synthesis tree that represents every possible synthetic route that creates a molecule using molecular building blocks from B and chemical reactions from R.
Node	N	$N \in T$	A node in the synthesis tree.
Node molecules	N_{mols}	$N_{mols} \subset C$	The molecules represented by node N, which are either building block molecules from B or molecules produced by combining building blocks from B with chemical reactions from R.
Node children	$N_{children}$	$N_{children} \subset T$	The child nodes of node N, which consist of all nodes that contain N_{mols} along with one more building block molecule from B or contain the product of applying a reaction $r \in R$ to N_{mols}.
Node siblings	$N_{siblings}$	$N_{siblings} \subset T$	The sibling nodes of node N, which are all nodes created at the same time as N by the parent node of N.
Node visits	N_{visit}	$N_{visit} \in \mathbb{N}$	The number of times node N has been visited (i.e., selected during a rollout).
Node value	N_{value}	$N_{value} \in \mathbb{R}$	The value of the node, which is the sum of the property prediction scores of all final molecules produced by a synthetic route that passes through node N.
Node diversity	$N_{diversity}$	$N_{diversity} \in \mathbb{N}$	The building block diversity of the node, which is the maximum number of times that any of the building blocks used in any of the molecules in N_{mols} has been used in non-building block molecules generated so far.
# rollouts	$n_{rollout}$	$n_{rollout} \in \mathbb{N}$	The number of rollouts to run the SyntheMol MCTS algorithm.
# reactions	$n_{reaction}$	$n_{reaction} \in \mathbb{N}$	The maximum number of reactions allowed during a rollout.

Table 1: Notation used to describe the SyntheMol algorithm.

N_{mols} in a given node N must be able to participate together in at least one reaction in R (although additional reactants may be needed).

Additionally, each node in the tree has a set of child nodes, $N_{children} \subset T$, which can come from two sources. First, for each reaction $r \in R$ where the node's molecules N_{mols} match all the reactants in r, we apply r to N_{mols} and add a child node to $N_{children}$ for each unique product molecule of the reaction. Note that there may be multiple ways to run a reaction for a given set of reactant molecules resulting in multiple possible products and, thus, multiple child nodes. The second source of child nodes comes from creating nodes that contain all of the molecules in N_{mols} along with one molecular building block from B that is compatible with all of the molecules in N_{mols} in at least one reaction in R. Note that for the root node, which has no molecules, the child set is all nodes that have exactly one molecular building block in B.

In order to generate molecules, SyntheMol employs an MCTS algorithm that searches through the chemical tree T to find nodes N that contain molecules that are predicted to have high molecular property scores according to the property predictor M. Specifically, SyntheMol runs $n_{rollout}$ rollouts through the chemical tree T, where each rollout begins at the root node, which is an empty node, and proceeds to search through the tree as outlined below.

At each node, SyntheMol selects a child node by scoring all of the child nodes of the current node using a scoring function $S(N)$ (defined below), and it then selects the node with the highest score. This scoring and selection is then repeated for this child node, and the process continues until a node is found that contains a single molecule $m \in C$ produced with $n_{reaction}$ chemical reactions. Every

node N that is selected ("visited") during this rollout increments its visit count N_{visit} by one and increments its value N_{value} by $M(m)$, which is the model score of the final molecule of the rollout.

The node score is $S(N) = \frac{Q(N)+P(N)\cdot U(N)}{D(N)}$, which balances exploitation with $Q(N)$, molecular property prediction with $P(N)$, exploration with $U(N)$, and building block diversity with $D(N)$.

The exploitation factor is $Q(N) = \frac{N_{value}}{N_{visit}}$ where N_{value} is the sum of property prediction scores of all final molecules discovered on rollouts that visit node N, and N_{visit} is the number of times node N has been visited. This factor encourages SyntheMol to follow routes through the chemical tree T that lead to high scoring final molecules.

The property prediction factor is $P(N) = \frac{1}{|N_{mols}|}\sum_{i=1}^{|N_{mols}|} M(N_{mols}^i)$ where M is the property prediction model and N_{mols}^i is the ith molecule in node N. This factor represents the average property prediction score of the molecules in the node and encourages selection of nodes with high-scoring molecules that could potentially form a single high-scoring molecule when combined by a chemical reaction.

The exploration factor is $U(N) = c \cdot \frac{\sqrt{1+N_{visit}+\sum_{N' \in N_{siblings}} N'_{visit}}}{1+N_{visit}}$ where $c = 10$ is a hyperparameter controlling the exploration-exploitation tradeoff, $N_{siblings}$ is the set of sibling nodes of N (i.e., all nodes created at the same time as N by the same parent node), and N_{visit} is the visit count of the node. This factor encourages SyntheMol to select child nodes that have not been visited frequently compared to their sibling nodes.

The building block diversity factor is $D(N) = e^{\frac{N_{diversity}-1}{100}}$ where $N_{diversity}$ is the maximum number of times that any of the building blocks used in any of the molecules in N_{mols} has been used in molecules across all of the nodes searched so far. This factor penalizes SyntheMol for selecting nodes with molecules containing building blocks that have already been used many times in previously visited nodes.

After $n_{rollout}$ rollouts (we use $n_{rollout} = 20,000$), SyntheMol stops and returns a list of all the nodes it encountered during the search. This list is then filtered to only keep nodes that contain a single molecule that was produced using at least one chemical reaction (i.e., excluding the building blocks themselves). In order to ensure rapid, inexpensive, and easy synthesis, we use $n_{reaction} = 1$ to generate single-reaction molecules, which is equivalent to searching the REAL Space since it only contains single-reaction molecules. However, SyntheMol can be directly applied to generate molecules that require multiple chemical reactions. Even allowing just 2–3 chemical reactions per molecule would result in a chemical space of 10^{20} to 10^{30} molecules, illustrating the potential of SyntheMol to explore truly huge combinatorial chemical spaces.