nature machine intelligence

Article

https://doi.org/10.1038/s42256-024-00809-7

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

Received: 10 March 2023

Accepted: 8 February 2024

Published online: 22 March 2024

Check for updates

Kyle Swanson[1,4], Gary Liu[2,4], Denise B. Catacutan[2], Autumn Arnold[2], James Zou[1,3] ✉ & Jonathan M. Stokes[2] ✉

The rise of pan-resistant bacteria is creating an urgent need for structurally novel antibiotics. Artificial intelligence methods can discover new antibiotics, but existing methods have notable 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 new compounds, which are easy to synthesize, from a chemical space of nearly 30 billion molecules. We apply SyntheMol to design molecules that inhibit the growth of *Acinetobacter baumannii*, a burdensome Gram-negative bacterial pathogen. We synthesize 58 generated molecules and experimentally validate them, with six structurally novel molecules demonstrating antibacterial activity against *A. baumannii* and several other phylogenetically diverse bacterial pathogens. This demonstrates the potential of generative artificial intelligence to design structurally novel, synthesizable and effective small-molecule antibiotic candidates from vast chemical spaces, with empirical validation.

The global dissemination of antibiotic resistance determinants is one of the most pressing challenges of modern medicine. In 2019, an estimated 4.95 million deaths were associated with drug-resistant infections. This number is projected to grow to 10 million per year by 2050 as the propagation of antimicrobial resistance determinants continues to outpace the discovery of novel antibiotics[1]. Six bacterial species known as the ESKAPE pathogens are especially virulent and drug-resistant, posing a critical threat to medicine globally[2,3]. One of those pathogens, the Gram-negative bacterium *Acinetobacter baumannii*, is particularly burdensome in clinical settings and is recognized as the highest priority for new antibiotic development, according to the World Health Organization[4]. Indeed, few treatment options remain for *A. baumannii* infection and novel antibiotics are urgently needed to address this pathogen[5].

Artificial intelligence (AI) methods have shown that they can identify promising drug candidates, including antibiotics[6]. One type of AI method that is commonly leveraged in the context of drug discovery is called a property prediction model, which is trained to predict drug properties of molecules[7]. As a concrete example, Stokes et al.[8] trained a property prediction model to predict growth inhibitory activity against the bacterium *Escherichia coli* and made predictions on a library of ~107 million molecules. They experimentally validated the

[1]Department of Computer Science, Stanford University, Stanford, CA, USA. [2]Department of Biochemistry and Biomedical Sciences, Michael G. DeGroote Institute for Infectious Disease Research, David Braley Centre for Antibiotic Discovery, McMaster University, Hamilton, Ontario, Canada. [3]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. [4]These authors contributed equally: Kyle Swanson, Gary Liu. ✉e-mail: jamesz@stanford.edu; stokesjm@mcmaster.ca

https://doi.org/10.1038/s42256-024-00809-7

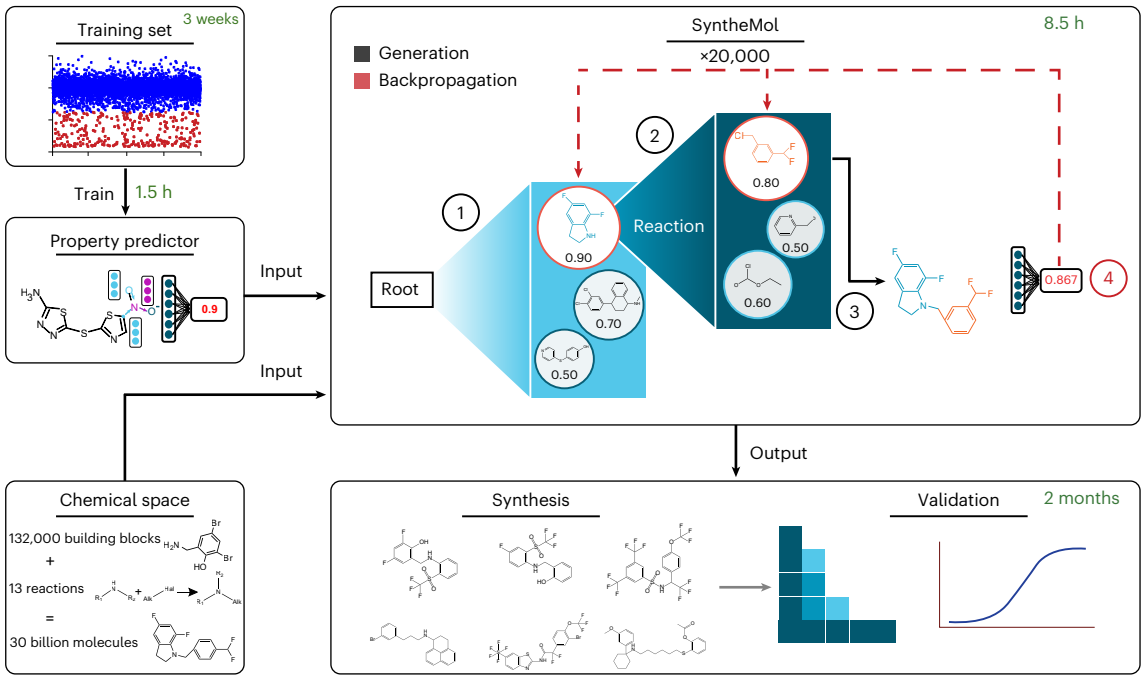


Fig. 1 | Generative AI for antibiotic discovery. An overview of our generative AI method, SyntheMol, for designing novel antibiotics. First, we curated a training set of ~13,000 molecules and performed growth inhibition assays to determine their bioactivity against *A. baumannii*. We subsequently used these chemical screening data to train a property prediction model to predict antibacterial activity. For molecule generation, we selected a chemical space consisting of nearly 30 billion molecules, each of which can be synthesized by applying one of 13 chemical reactions to combine two or three molecular building blocks from a set of ~132,000 building blocks. SyntheMol explores this chemical space for antibiotic candidates by using a MCTS guided by the property predictor. SyntheMol iteratively selects building blocks (steps 1 and 2) that are combined through a chemical reaction (step 3) to form molecules that are scored by the property predictor (step 4). These scores are backpropagated (red dashed lines) through the synthetic route to inform future selections by SyntheMol. After 20,000 iterations, we filtered the generated molecules to obtain a set of structurally novel and diverse high-scoring compounds, which were synthesized and experimentally tested against a phylogenetically diverse set of bacterial species in vitro.

top predictions and discovered several structurally novel molecules with strong antibiotic activity. Similarly, Rahman et al.[9] built a property prediction model for the antibiotic-resistant bacterium *Burkholderia cenocepacia* and applied it to a library containing ~224,000 molecules. Despite these successes, property prediction models have limitations. These molecular property prediction models must evaluate molecules one by one from enumerated chemical libraries, which prevents them from exploring truly vast chemical spaces in a reasonable time. Moreover, such models are limited to evaluating compounds from curated chemical libraries and are unable to generate truly novel chemical matter[10].

By contrast, generative AI models design new molecules from scratch, rather than evaluating given compounds[11]. Generative AI models can directly generate molecules with desired properties without the costly enumeration and evaluation of many compounds. Additionally, generative models design molecules from a vast chemical space, enabling the discovery of novel structural classes of molecules that might not be found within the in silico chemical libraries that can be processed by molecular property prediction models[12]. This is particularly important for antibiotics, where structurally and functionally novel molecules are desirable to overcome existing resistance determinants[13]. However, a major limitation of these generative models is that they often generate synthetically intractable compounds[14]. Without a practical method to chemically synthesize these in silico generated molecules, they are unable to be experimentally validated against bacteria. For this reason, there has been considerable interest in developing generative AI models that design synthesizable molecules. While methods have been proposed with promising in silico results[15–19], very few studies have synthesized and experimentally tested any of the generated molecules[11]. Moreover, none of these previous methods have been applied to small-molecule antibiotic development.

In this study, we developed SyntheMol, a generative AI model that uses a Monte Carlo tree search (MCTS)[20,21] to assemble novel compounds using ~132,000 molecular building blocks with known reactivities and 13 well-validated chemical synthesis reactions (Fig. 1). These building blocks allow for the exploration of a chemical space of nearly 30 billion molecules that are easy to synthesize, with synthesis success rates of over 80% within 3–4 weeks[22]. We trained SyntheMol to design molecules with antibiotic activity against *A. baumannii*, and we synthesized and experimentally validated 58 of our generated molecules. Six structurally diverse molecules displayed potent antibacterial activity against *A. baumannii* and several other phylogenetically diverse bacterial pathogens. Furthermore, these structurally novel compounds also retained antibacterial activity against clinical isolates harbouring an array of functionally diverse resistance determinants. These results demonstrate the use of generative AI models to design structurally novel, synthetically tractable and efficacious small-molecule antibiotic candidates from vast chemical spaces.

Results

Property prediction model development

Our generative approach relies on a molecular property prediction model to evaluate the potential of generated molecules to inhibit the growth of *A. baumannii*. Therefore, we began by physically screening three distinct chemical libraries to use as a training dataset. Library 1 and Library 2 are collections of bioactive compounds with 2,371 and 6,680 molecules, respectively. Library 3 is a synthetic commercially available small-molecule screening collection with 5,376 molecules. To acquire our training dataset, we grew *A. baumannii* ATCC 17978

https://doi.org/10.1038/s42256-024-00809-7

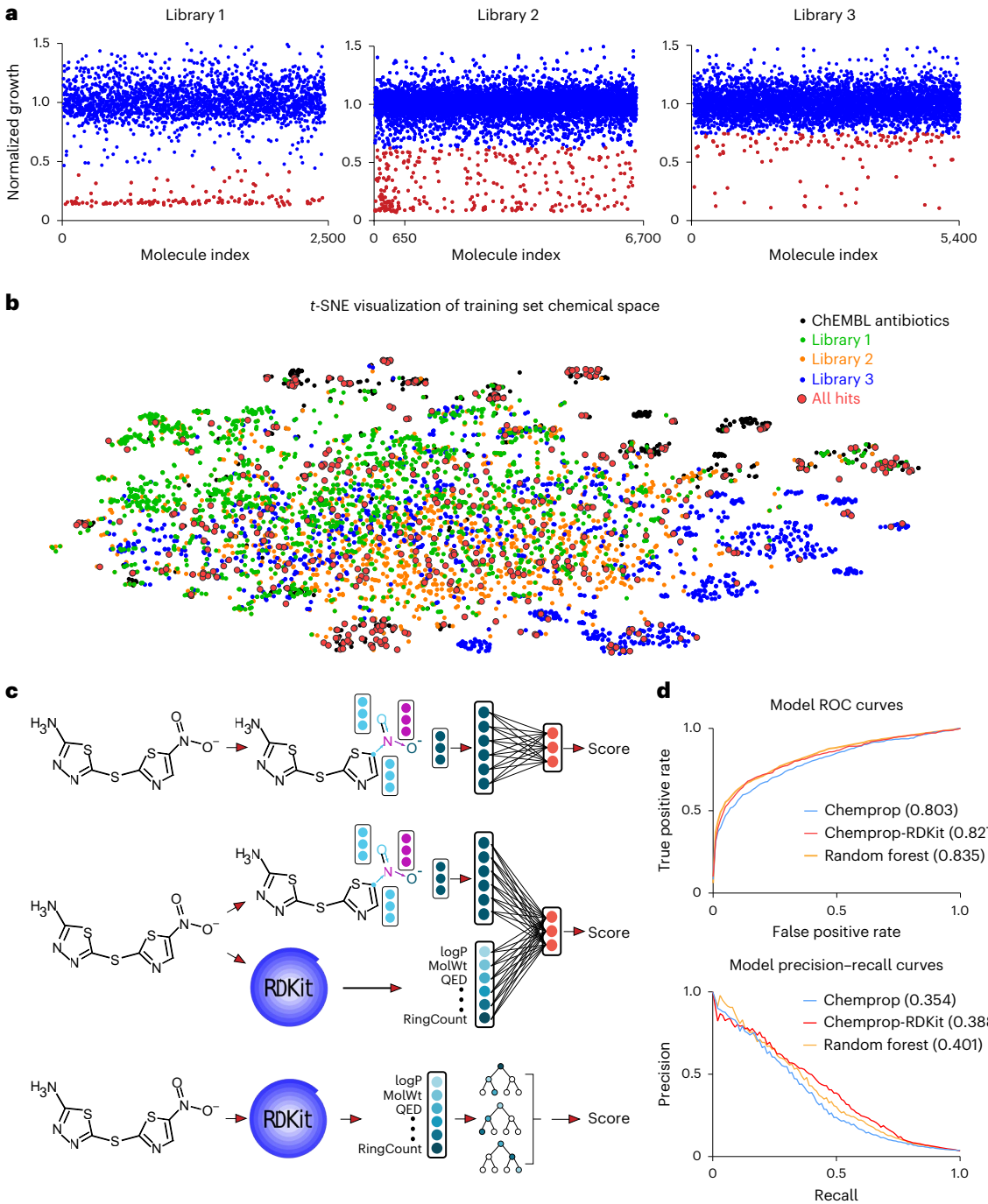


Fig. 2 | Property prediction model development. a, Three training datasets of diverse small molecules screened against *A. baumannii* ATCC 17978 for growth inhibition. Each plot shows the mean normalized growth across two biological replicates for all molecules in the training set. A threshold of the mean minus two standard deviations is used to binarize the growth values, with blue indicating non-active compounds (above the threshold) and red indicating active compounds (below the threshold). The first 650 compounds in Library 2 were previously tested to validate the property predictor for antibacterial activity against *A. baumannii*[70]. **b**, A *t*-SNE visualization showing 1,005 known antibacterial molecules curated from ChEMBL (black), each of the three training libraries (green, orange, blue), as well as the active molecules from these training libraries (red). **c**, Visual depictions of the Chemprop, Chemprop-RDKit and random forest property prediction models. **d**, ROC and PRC for each of the three property prediction models. Each curve is the average across the ten models in the ensemble. AUC is indicated next to each model.

in the presence of each chemical at 50 µM, in a volume of 100 µl, in biological duplicate. After 16 hours of incubation, we measured the endpoint optical density at 600 nm (OD_{600}). Next, for each library separately, we computed the mean μ and standard deviation σ OD_{600} value across the library and used $\mu - 2\sigma$ as a threshold for binarizing the optical density values into active and inactive molecules (Fig. 2a, Extended Data Fig. 1a and Supplementary Data 1)[23]. We then merged the three binarized libraries and removed duplicate compounds with conflicting activity labels (Methods). This resulted in a combined set of 13,524 unique molecules, with 470 active compounds and 13,054 inactive compounds. After binarization, we performed a *t*-stochastic neighbour embedding (*t*-SNE) visualization[24] of our training dataset

https://doi.org/10.1038/s42256-024-00809-7

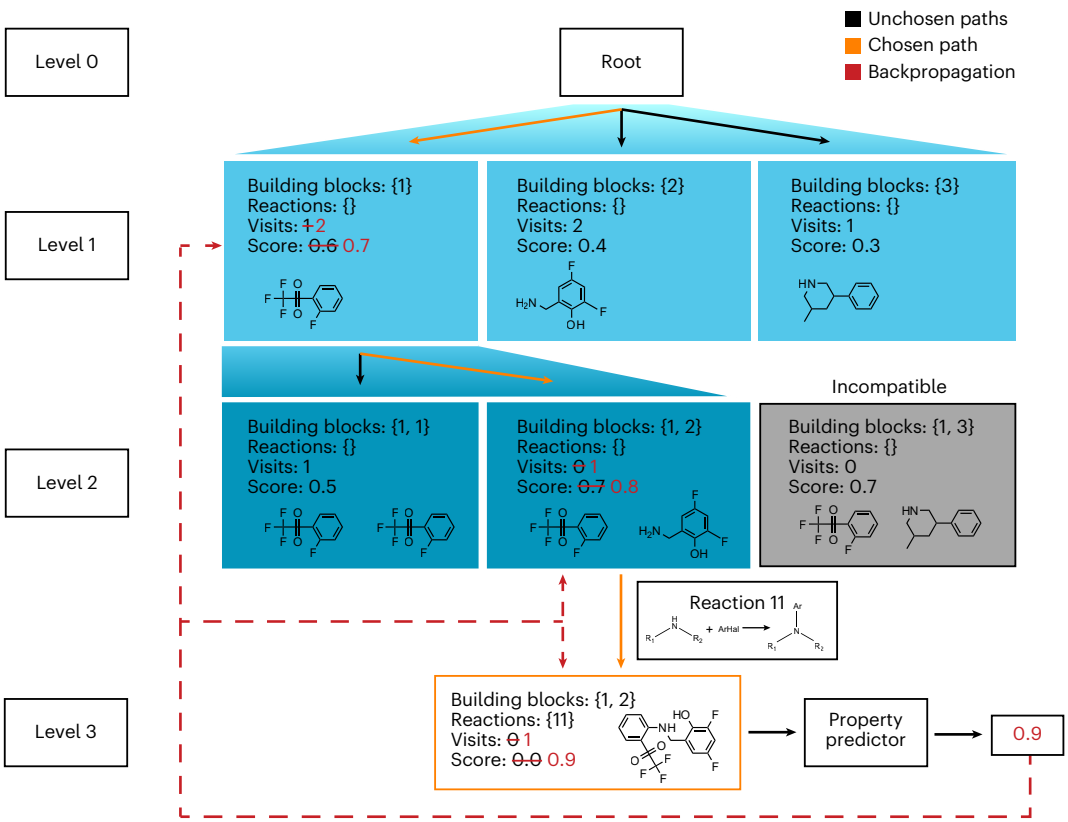


Fig. 3 | SyntheMol. SyntheMol generates molecules using a MCTS that explores combinatorial chemical space guided by a property prediction model. The SyntheMol MCTS algorithm begins at the root node (level 0), which is an empty node. This node is expanded by adding one child node for each building block molecule in the chemical space (level 1). For each child node, SyntheMol computes a score that balances both exploitation of nodes that are known to lead to high-scoring molecules as well as exploration of nodes that have rarely been visited during the search (Methods). SyntheMol then selects the node on level 1 with the highest score (orange arrow) and expands it by creating child nodes that contain the selected building block along with one other building block (level 2). However, if the two building blocks in a node are not synthetically compatible using at least one chemical reaction, the node is removed (grey node). As with level 1, the scores of the nodes on level 2 are computed and the node with the highest score is selected (orange arrow). If the building blocks in the selected node are synthetically compatible in a reaction that requires additional reactants, then the process repeats and child nodes are added that contain the two building blocks along with a third building block. If, as in this case, the building blocks fulfil all the reactants for a reaction, then a child node is created with the product of that reaction and the product molecule is evaluated by the antibacterial property predictor (level 3). The property prediction value is propagated back to all nodes on the path to this molecule (red dashed arrows), and each node's score and visit count are updated before the next rollout (red numbers). This entire process represents one rollout of the tree search. SyntheMol performs a set number of rollouts, with each rollout updating the node scores to guide future rollouts towards regions of chemical space that contain high-scoring molecules.

compounds in the context of 1,005 molecules from ChEMBL[25] with known antibacterial activity (Fig. 2b and Supplementary Data 2). This analysis indicates that our active compounds cover both known and new chemical space for antibiotics.

After we acquired our training dataset, we leveraged three different property prediction models for predicting antibacterial activity against *A. baumannii* (Fig. 2c): (1) Chemprop, a graph neural network[26]; (2) Chemprop-RDKit, a variant of Chemprop that incorporates a set of 200 molecular features computed by the cheminformatics package RDKit[27] and (3) a random forest model[28] that uses the 200 RDKit features as input to a set of 100 decision trees. We trained each of these three property prediction models on our *A. baumannii* training dataset using tenfold cross-validation with splits containing 80% train, 10% validation and 10% test data. All three model types trained in less than 90 minutes on a machine with 16 CPUs (Supplementary Data 3) and performed similarly, with receiver operating characteristic-area under the curve (ROC-AUC) in the range 0.80–0.84 and precision–recall curve-AUC (PRC-AUC) in the range 0.35–0.40 (Fig. 2d and Extended Data Fig. 1b–e). During molecule generation, for each property prediction model type, we used the average prediction score of the ensemble of ten models (from the ten cross-validation folds) trained on the full *A. baumannii* training dataset as the score function within our generative model.

Generative model development

Our generative model, SyntheMol, designs molecules within a large combinatorial chemical space, which is a chemical space in which every molecule can be produced by applying a series of chemical reactions to a defined set of molecular building blocks (Fig. 3). The building blocks are small molecules that are readily purchasable, and the chemical reactions are well-validated reactions that can be applied to a wide variety of these building blocks. To ensure rapid and cost-effective synthesis, we limited SyntheMol to design molecules that could be constructed using a single chemical reaction applied to two or three building blocks. SyntheMol can be applied to generate multi-reaction molecules if desired, which would expand the chemical space available to the model by several orders of magnitude.

The combinatorial chemical space we used here was the Enamine REadily AccessibLe (REAL) Space[22]. The REAL Space consists of 31 billion single-reaction molecules that can be produced by applying 169 chemical reactions to 138,085 molecular building blocks (Extended Data Fig. 2a,b). After processing and deduplicating the building blocks

(Methods), we were left with 132,479 unique building block molecules, with molecular weights ranging from 17 to 503 Da (1–34 heavy atoms) and cLogP ranging from −2.98 to 7.89 (Extended Data Fig. 2c,d). Although the REAL compounds occupy a different chemical space from our training set (Fig. 4a), the building blocks that are assembled into these larger molecules show notable overlap with our training set, indicating similar molecular features that our property prediction models can recognize and evaluate. For simplicity, we restricted SyntheMol to use 13 of the most common REAL reactions (Fig. 4b and Extended Data Fig. 2e). Applying these 13 reactions to our collection of 132,479 building blocks can produce 29.6 billion molecules, which is 93.9% of REAL Space (Supplementary Data 4).

SyntheMol uses a MCTS[29] guided by a property prediction model to search through a vast combinatorial chemical space for promising antibiotic candidates with activity against *A. baumannii* (Fig. 3). During each MCTS rollout, SyntheMol constructs a molecule by selecting building blocks and combining them with chemical reactions. The generated molecule is then evaluated by a property prediction model, which provides feedback to the MCTS algorithm. Over the course of the rollouts, SyntheMol learns which building blocks and chemical reactions tend to produce molecules with high property prediction scores. This is particularly important given that most building blocks have low scores (Fig. 4c and Extended Data Fig. 3a,b), and yet many molecules constructed from these building blocks have high scores that are poorly predicted by their average building block score (Fig. 4d and Extended Data Fig. 3c,d) but could be identified by MCTS. Notably, there are also many low-scoring molecules that contain high-scoring building blocks (Fig. 4d and Extended Data Fig. 3c,d), so MCTS could also learn to avoid such building blocks. Furthermore, SyntheMol balances exploration and exploitation by computing a score (Methods) that values both previously unselected building blocks (exploration) as well as building blocks that are known to lead to high-scoring molecules (exploitation). We additionally extend the standard MCTS score to specifically prioritize molecules that include a diverse set of building blocks (Extended Data Fig. 4a–c). After a set number of rollouts, SyntheMol outputs all the generated compounds along with the specific synthetic scheme—the building blocks and chemical reactions in order—required to synthesize each molecule.

Before running SyntheMol for antibiotic discovery, we evaluated this method in silico by applying it to a computed molecular property, thereby allowing us to evaluate the generated molecules rapidly and inexpensively. Specifically, we selected the property cLogP—the computed octanol-water partition coefficient—as determined by RDKit, with the goal of generating molecules with cLogP > 6.5. Using a binary classification Chemprop predictor for cLogP > 6.5, we ran SyntheMol for 20,000 rollouts. Among the 25,550 generated molecules, 61.42% were active (RDKit cLogP > 6.5), representing a 1,396 times increase in hit rate (Fig. 4e and Supplementary Data 5) compared to 0.044% active molecules in a random set of 25,000 REAL Space molecules (Extended Data Fig. 2d). Even when using a weaker cLogP Chemprop model trained for one epoch instead of 30 to better reflect the antibiotic Chemprop model's performance, 11.78% of the SyntheMol-generated molecules were active, representing a 268× increase in hit rate (Fig. 4e and Supplementary Data 5). These results gave us confidence that SyntheMol can rapidly search a huge combinatorial space for active molecules with a pronounced enrichment in hits.

On the basis of these compelling results, we next applied SyntheMol to discover potential antibiotic candidates against *A. baumannii* by using our antibiotic property predictors. We separately performed three sets of generations with SyntheMol, one each with our Chemprop, Chemprop-RDKit and random forest models. Since the analysis of the three generated sets is qualitatively similar, here we present results using Chemprop within SyntheMol and present the corresponding results for Chemprop-RDKit and random forest in the Extended Data figures and tables.

Over the course of 20,000 rollouts (less than 8.5 hours), SyntheMol with Chemprop evaluated 452 million intermediate nodes containing diverse combinations of molecular building blocks and generated 24,335 complete molecules, of which 2,868 had a Chemprop antibacterial prediction score of at least 0.5 (Supplementary Data 6–8). This outperforms an AI-based virtual screening approach in which Chemprop scored 10 million randomly sampled REAL molecules (8 hours) and only identified 374 molecules with a score of at least 0.5 (only 13% as many as SyntheMol).

SyntheMol generated high-scoring molecules throughout all rollouts, but these high-scoring molecules were particularly concentrated in early rollouts, with 1,035 (36%) of the 2,868 molecules with a Chemprop score of at least 0.5 generated in the first 2,000 (10%) of the 20,000 rollouts (Fig. 4f and Extended Data Fig. 5a,b). While 20,000 rollouts only explore a fraction of the nearly 30 billion-molecule chemical space, these results indicate that SyntheMol generated many of the highest scoring compounds rapidly, with diminishing returns from additional rollouts. The generated molecules included a diverse set of 10,846 unique building blocks, with each building block appearing in at most 137 different complete molecules (Extended Data Figs. 6a, 7a and 8a), and they used all 13 reactions with varying frequencies (Extended Data Figs. 6b, 7b and 8b).

Among the compounds generated by SyntheMol, we aimed to select a diverse set of structurally novel molecules with high property prediction scores for downstream validation; we developed a set of three filters to facilitate this process. First, to ensure structural novelty of the generated molecules, we computed the Tversky[30] similarity between Morgan fingerprints[31] of each generated molecule and all of the 470 active molecules from the training dataset (Fig. 4g and Extended Data Figs. 7c and 8c), as well the 1,005 antibacterial molecules in the ChEMBL database (Fig. 4h and Extended Data Figs. 7d and 8d). We then removed any molecules with a Tversky similarity greater than 0.5 to ensure structural novelty of the remaining generated molecules. Second, to obtain molecules that were predicted to have the greatest antibiotic activity, we ranked the remaining molecules according to their model prediction score and we kept only the top 20% of molecules (Fig. 4i and Extended Data Figs. 7e and 8e). Third, to select a structurally diverse set of molecules, we applied k-means clustering[32] with $k = 50$ using Tanimoto[33,34] distance between the Morgan fingerprints of the remaining molecules to obtain 50 clusters of molecules. We then selected the highest scoring molecule in each cluster for a total of 50 molecules (Fig. 4j and Extended Data Figs. 7f and 8f). These 50 selected molecules contain a diverse set of 73 different building blocks and use eight of the 13 reactions, and 33 (66%) have a Tanimoto similarity to all other selected molecules below 0.5.

We applied this filtering procedure to the generated set from each of the three property prediction models, resulting in a set of 150 diverse molecules for experimental validation (Fig. 4k, Extended Data Fig. 6c and Supplementary Data 6–8). Since many compounds that can theoretically be produced by REAL building blocks and reactions do not appear in the REAL Space, probably due to challenges with synthesis, we curated a set of 70 of the 150 compounds that were available in the REAL Space. Of those, 58 (83%) were successfully synthesized in about 4 weeks by Enamine, with 26 molecules from SyntheMol with Chemprop, 22 molecules from SyntheMol with Chemprop-RDKit and ten molecules from SyntheMol with random forest (Supplementary Data 9).

In vitro validation of AI generated molecules

We next set out to validate the bioactivity of the 58 synthesized molecules against *A. baumannii* in the laboratory. We began by performing growth inhibition assays using the synthesized molecules against *A. baumannii* ATCC 17978: the same strain used for training set curation. Because *A. baumannii* is a Gram-negative bacterium with challenging permeability characteristics due to its highly impermeable outer membrane[35], we added low concentrations of the cell envelope

https://doi.org/10.1038/s42256-024-00809-7

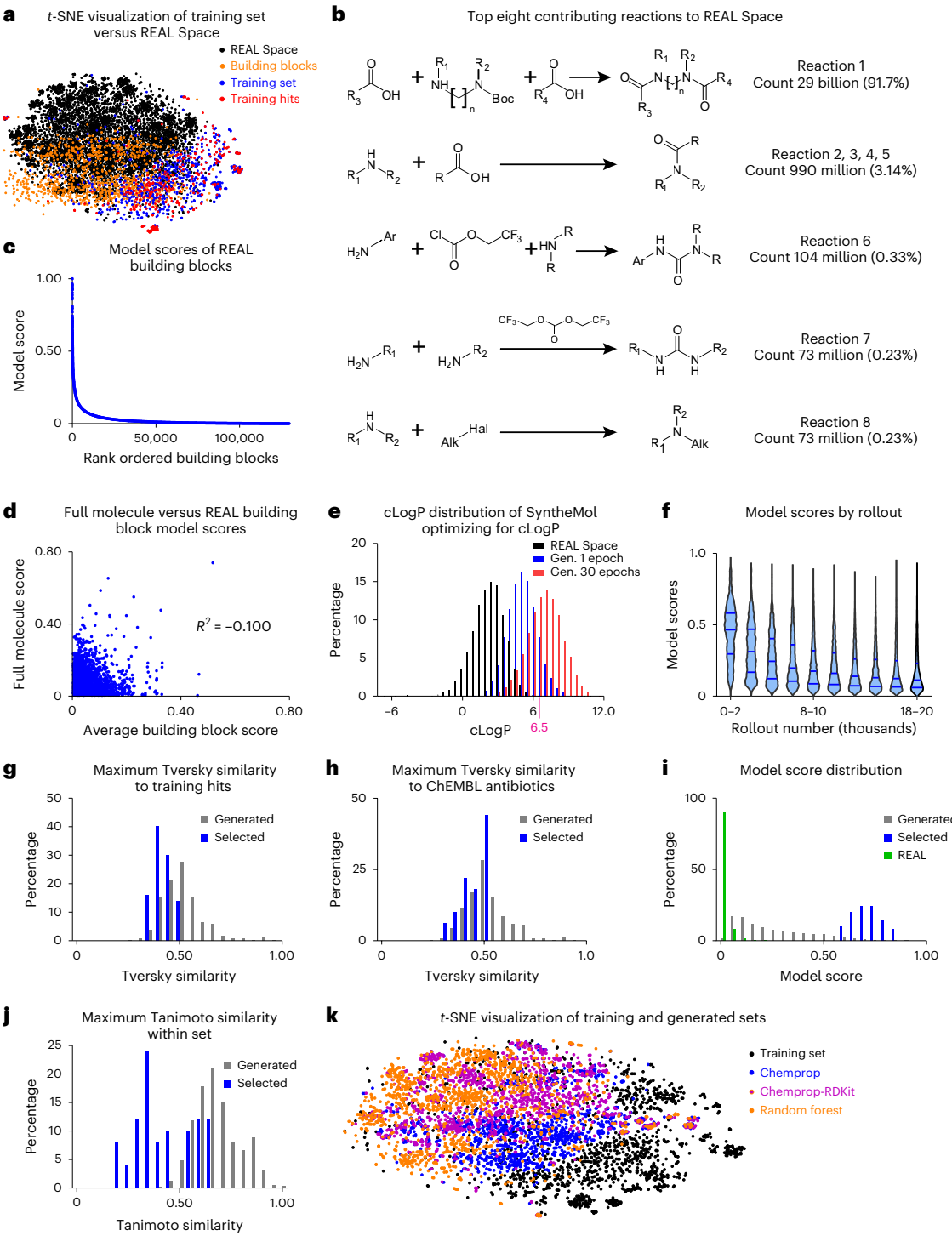


Fig. 4 | Generative model development. a, A *t*-SNE visualization of the Enamine REAL Space molecules (black), REAL building blocks (orange), the training set (blue) and active molecules within the training set (red). **b**, The top eight of the 13 REAL chemical reactions we applied in SyntheMol, along with the number and percentage of REAL molecules produced by each reaction. Note that four reactions (2, 3, 4 and 5) have common reactants and products, but have different catalysts, and are therefore grouped together. **c**, The distribution of Chemprop antibacterial model scores across the REAL building blocks. **d**, The correlation between the Chemprop antibacterial score of a REAL molecule and the average Chemprop score of its constituent building blocks. The R^2 value is the coefficient of determination. **e**, The distribution of cLogP values for a random sample of REAL molecules (black) and the molecules generated using SyntheMol with a Chemprop predictor for cLogP, either trained for one epoch (weak model, blue) or for 30 epochs (strong model, red). The threshold for binarization of the data is shown in pink (cLogP = 6.5). **f**, Violin plot of the distribution of Chemprop antibacterial model scores for every 2,000 rollouts of SyntheMol over 20,000 rollouts (n = 24,335 molecules). The lines in each violin indicate the first quartile, the median and the third quartile. **g**–**j**, A comparison of the properties of the 24,335 molecules generated by SyntheMol with the Chemprop antibacterial model and the 50 molecules selected from that set after applying post hoc filters. **g**, The distribution of nearest neighbour Tversky similarities between the generated or selected compounds and the active molecules in the training set. **h**, The distribution of nearest neighbour Tversky similarities between the generated or selected compounds and the ChEMBL antibacterial molecules. **i**, The distribution of Chemprop antibacterial model scores on the generated or selected compounds, as well as on a random set of 25,000 REAL molecules. **j**, The distribution of nearest neighbour Tanimoto similarities among the generated or selected compounds. **k**, A *t*-SNE visualization of the training set and the sets of molecules generated by SyntheMol with each of our three property predictor models.

permeabilization agents SPR 741 or colistin to enhance intracellular accumulation of the synthesized molecules. SPR 741 is a structural analogue of polymyxin B that disrupts the Gram-negative outer membrane without perturbing the inner membrane, rendering it less toxic than other polymyxins[36]. It has been investigated in clinical trials[37]. Colistin is a polymyxin antibiotic that is the last-line treatment for drug-resistant Gram-negative infections. Despite nephrotoxicity concerns, it is used in combination therapies in clinical settings[38]. Colistin rapidly disrupts both the Gram-negative outer membrane, as well as the inner membrane at slightly elevated concentrations. Results from antibiotic potency analyses shown in Fig. 5a reveal excellent antibacterial activity of six molecules (Enamine 10, 23, 28, 31, 40 and 43; Supplementary Data 9), as defined by a minimum inhibitory concentration (MIC) ≤ 8 μg ml^{-1}, when combined with a quarter MIC SPR 741 or a quarter MIC colistin (Extended Data Fig. 9a). This represents an exceptional 10% hit rate. This is nearly threefold larger than the 3.5% hit rate of the training set, which is already enriched for compounds with antibiotic activity.

We next tested our six bioactive molecules against *A. baumannii* ATCC 19606R, a lipopolysaccharide-deficient polymyxin-resistant mutant, to genetically validate our observed 741-mediated potentiation[38]. Lipopolysaccharide is the major component of the outer leaflet of the Gram-negative outer membrane that contributes to the impermeability of the Gram-negative cell envelope. Consistent with our potentiation results with SPR 741, all six compounds showed highly potent growth inhibitory activity against *A. baumannii* ATCC 19606R (MICs ≤ 4 μg ml^{-1}; Fig. 5b,c). Moreover, these six molecules displayed antibacterial efficacy against methicillin-resistant *Staphylococcus aureus* USA 300 (MICs ranging from 1 to 64 μg ml^{-1}; Fig. 5b,c). For reference, *S. aureus* is a Gram-positive bacterial pathogen that lacks an outer membrane. These results emphasize the inherent bioactivity of these six generated molecules, despite their limitations in outer membrane penetration.

To probe the frequencies with which resistance can spontaneously evolve against these six molecules in vitro, we quantified their frequencies of resistance (FOR) using *A. baumannii* ATCC 19606R growing on solid media as a model. Briefly, 100 μl of an overnight culture of *A. baumannii* 19606R was spread onto solid media supplemented with each generated molecule at concentrations ranging from 0× to 20× MIC. Plates were then incubated and monitored for the emergence of resistant colonies over 72 hours. We observed the concentration-dependent emergence of colonies, where lower concentrations of compound resulted in higher observed FOR (Supplementary Table 1). Indeed, at 20× MIC, we only observed the emergence of colonies for En-28; no other molecules permitted colony growth after 72 hours. For concentrations in which resistant colonies emerged during the 72 hour incubation period, the calculated FOR ranged from 10^{-7} (higher concentrations) to 10^{-5} (lower concentrations), which is consistent with known antibiotics. For reference, the FOR of rifampicin ranges from 10^{-8} to 10^{-5}, with single nucleotide polymorphisms occurring almost exclusively in the *rpoB* gene[39–41].

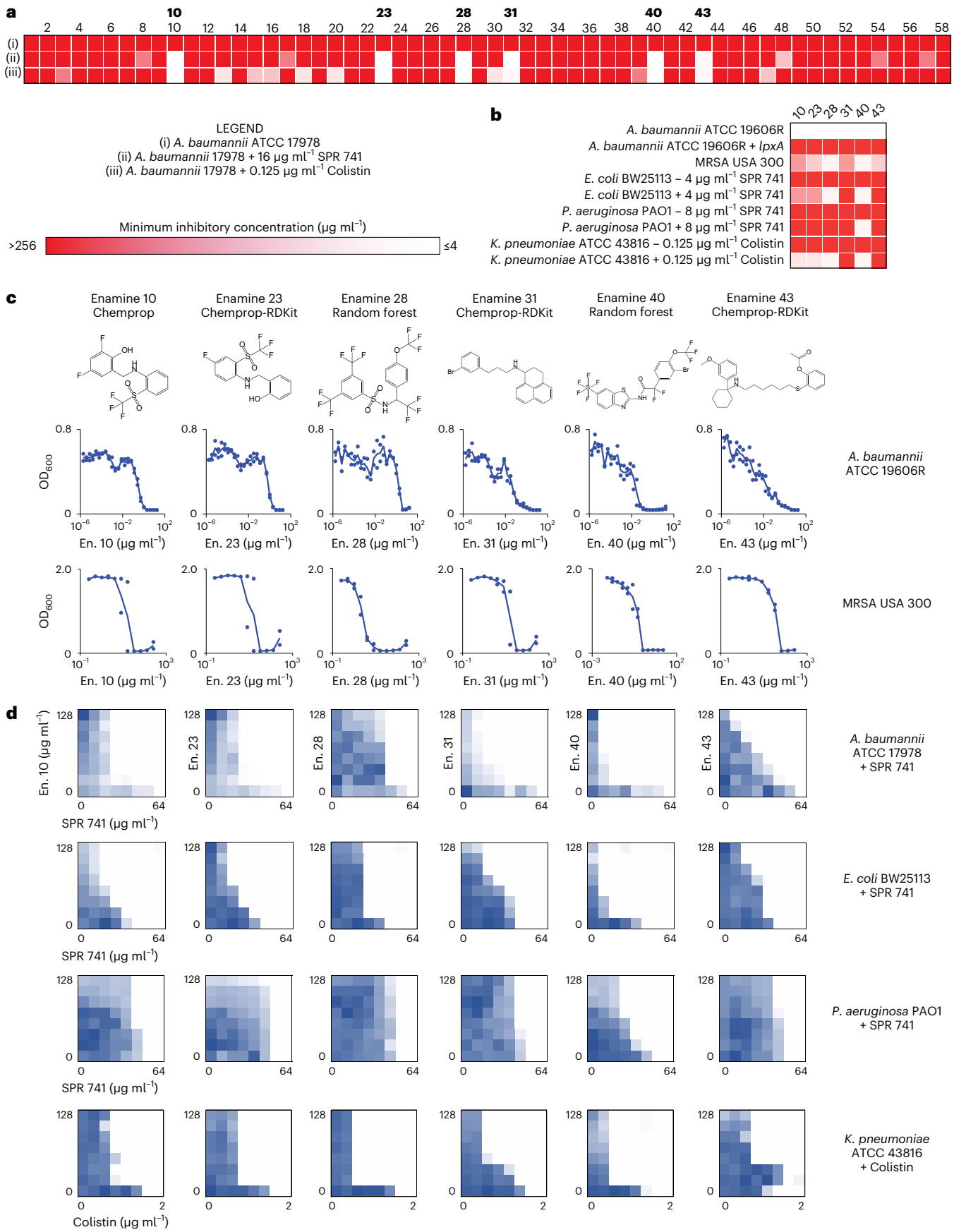
With data suggesting that the FOR of these efficacious generated molecules are within typical ranges for downstream development, we next ventured to understand the growth inhibitory activity of these molecules against multidrug-resistant clinical isolates of *A. baumannii* and *S. aureus*. We selected four strains of *S. aureus* and eight strains of *A. baumannii* from the Centers for Disease Control and Prevention Antibiotic Resistance Isolate Bank (ARIsolate Bank), which together capture all diverse resistance mechanisms present in each isolate panel (Supplementary Tables 2 and 3). The six generated compounds could overcome all resistance determinants encoded by these *S. aureus* and *A. baumannii* isolates (Supplementary Tables 2 and 3), consistent with their structural novelty relative to current clinical antibiotics. Together with their acceptable intrinsic FOR, these clinical isolate data further support potential downstream use of the generated molecules.

We also tested a set of 58 randomly selected molecules from the Enamine REAL Space for comparison (Extended Data Fig. 9b). From this random set of 58 compounds, none displayed antibacterial activity against *A. baumannii* ATCC 17978, alone or when combined with a quarter MIC SPR 741, as defined by our threshold of MIC ≤ 8 μg ml^{-1}. In combination with a quarter MIC colistin, three compounds displayed activity, which is just half of that observed from our generated set. Moreover, these random REAL molecules only displayed MICs from 4 to 8 μg ml^{-1}, while four of the six SyntheMol-generated compounds displayed an MIC ≤ 2 μg ml^{-1} (Supplementary Table 4). This potentiation by colistin, but not the less toxic outer membrane-specific compound SPR 741, suggests that these three random molecules may have nonspecific activity at the inner membrane, rather than a specific intracellular target. Taken together, although our model was not directly trained to predict Gram-negative outer membrane permeability, the prediction score was reflective of overall likelihood of bioactivity. Indeed, the success rate of the SyntheMol-generated molecules is impressive, particularly given that we emphasized selecting structurally novel compounds during post hoc filtering.

Given that we observed antibacterial activity of these six generated molecules against *A. baumannii*, as well as the phylogenetically distant bacterium *S. aureus*, we assessed whether these compounds would display broad-spectrum antibacterial activity against a wide range of bacterial pathogens. To this end, we tested these six compounds against the Gram-negative species *E. coli* BW25113, *Pseudomonas aeruginosa* PAO1 and *Klebsiella pneumoniae* ATCC 43816. The generated compounds were tested against *E. coli* and *P. aeruginosa* in combination with a quarter MIC SPR 741, and against *K. pneumoniae* in combination with a quarter MIC colistin (Fig. 5b,c and Extended Data Fig. 9a). All six compounds displayed broad-spectrum antibacterial activity against *E. coli* when tested in combination with a quarter MIC SPR 741, and against *K. pneumoniae* when tested in combination with a quarter MIC colistin (Fig. 5b,c). Only Enamine 40 retained its antibacterial activity in combination with a quarter MIC SPR 741 against *P. aeruginosa*. This is probably due to the high impermeability commonly displayed by the cell envelope of this species. Indeed, representative molecules from

Fig. 5 | In vitro validation of generated molecules. a, Heat map summarizing the MIC of the 58 synthesized molecules generated by SyntheMol against *A. baumannii* ATCC 17978 in (i) LB medium, (ii) LB medium + a quarter MIC SPR 741 and (iii) LB medium + a quarter MIC colistin. Compounds were tested at concentrations from 256 to 4 μg ml^{-1} in twofold serial dilutions. Lighter colours indicate lower MIC values for each generated molecule. Bold numbers indicate generated molecules with high activity (MIC ≤ 8 μg ml^{-1}) in conditions ii and iii. Experiments were performed in at least biological duplicate. **b**, Six molecules that showed high activity from **a** were tested against a panel of ESKAPE species, with and without an outer membrane disrupting agent (SPR 741 or colistin) at a quarter MIC. No potentiator was used for the lipopolysaccharide-deficient strain *A. baumannii* ATCC 19606R (mutation in *lpxA*) or the Gram-positive species *S. aureus*. Note that in trans expression of *lpxA* restores the native outer membrane of *A. baumannii* ATCC 19606R. Lighter colours indicate lower MIC values for each molecule. Experiments were performed in at least biological duplicate. **c**, Growth inhibition of lipopolysaccharide-deficient *A. baumannii* ATCC 19606R and *S. aureus* USA 300 by each of the six molecules in dose. Structures of compounds are shown. Experiments were performed using twofold serial dilution series. Experiments were performed in biological duplicate. Error bars represent absolute range of optical density measurements at 600 nm. **d**, Chequerboard analysis of the six compounds to quantify synergy, as defined by FICI, with SPR 741 or colistin against a panel of Gram-negative species. Chequerboard experiments were performed using twofold serial dilution series, with the maximum and minimum concentrations of the potentiator (*x* axis) and generated compound (*y* axis) shown in μg ml^{-1}. Darker blue represents higher bacterial growth. Experiments were performed in at least biological duplicate. The mean growth of each well is shown.

https://doi.org/10.1038/s42256-024-00809-7



many clinical antibiotic classes also failed to show potent antibacterial activity against *P. aeruginosa* PAO1 (Supplementary Table 5).

With data showing that our generated molecules displayed antibacterial efficacy against the diverse Gram-negative species *A. baumannii*, *E. coli*, *P. aeruginosa* and *K. pneumoniae* when combined with a quarter MIC SPR 741 or colistin, we ventured to quantify these chemical–chemical synergies. To this end, we performed dose-response chequerboard experiments to determine the fractional inhibitory concentration indices (FICI) (Supplementary Table 6) of all six generated molecules in combination with SPR 741 (*A. baumannii* ATCC 17978, *E. coli* and *P. aeruginosa*) or colistin (*A. baumannii* ATCC 17978 and *K. pneumoniae*) (Fig. 5d and Extended Data Fig. 9c,d). As expected, we observed synergy (FICI < 0.5) with all six compounds against *A. baumannii* ATCC 17978 (Fig. 5d and Extended Data Fig. 9c), *E. coli* (Fig. 5d) and *K. pneumoniae* (Fig. 5d). Dose-dependent synergy was only observed with Enamine 40 against *P. aeruginosa* (Fig. 5d), consistent with previous observations (Fig. 5b). Together, these in vitro experiments highlight the ability of SyntheMol to generate novel antibacterial molecules with potent laboratory-validated activity.

According to a Chemprop-RDKit molecular property prediction model trained on the ClinTox dataset[42], all six bioactive molecules are predicted to have a probability of human toxicity that is within the distribution of predictions for known non-toxic molecules in the ClinTox dataset, indicating a relatively low likelihood of toxicity (Extended Data Fig. 10). En-10 and En-23 did not induce any toxic phenotypes in mice—hunched posture, reduced exploratory behaviour, whole body piloerection and intermittent abdominal contractions—after an intraperitoneal injection of 50 mg kg^{-1}, which is an exceptional dose, in agreement with model predictions. Only En-10 and En-23 could be tested for toxicity in mouse models due to insufficient aqueous solubility of the other four generated antibacterial molecules. Indeed, SyntheMol was not trained to generate molecules with high aqueous solubility. Nevertheless, the favourable in vivo toxicity profiles of En-10 and En-23 may suggest that the additional four molecules, which received similarly low toxicity prediction scores, would also be non-toxic in mice. Therefore, we posit that these novel generated molecules represent exciting candidates for further antibiotic development in the context of translational investigations.

Discussion

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 antibiotic candidates. We applied this method to design antibacterial compounds against *A. baumannii*, for which new antibiotics are urgently needed. Among the compounds generated by SyntheMol, we synthesized and experimentally tested 58 structurally novel and diverse compounds. We discovered six structurally novel molecules, a 10% hit rate, with activity against *A. baumannii*, as well as other phylogenetically diverse ESKAPE species[2,3].

SyntheMol successfully designed molecules that are both synthetically tractable and effective against *A. baumannii* based on empirical wet laboratory validation, advancing beyond previous work applying generative methods for drug discovery tasks. Some studies have developed generative models for the design of antimicrobial peptides[43,44], which are potentially more straightforward to design and synthesize than small molecules, which can have complex synthetic routes. However, antimicrobial peptides can suffer from poor stability, low oral bioavailability and a short in vivo half-life due to proteolytic degradation, so small molecules may be preferred as systemically bioavailable medicines[45,46].

Generative models for small-molecule design can take a variety of forms[10–12]. A common method is the variational autoencoder, which learns to reproduce molecules from a given chemical space and can be conditioned to optimize for molecules with a desired property[47–53]. However, the molecules generated by these methods are often difficult to synthesize, limiting their practical use[14]. Therefore, recent methods have focused on improving synthesis tractability, either by generating molecules with high predicted synthesizability scores, which encourages but does not guarantee synthesizability[54], or by explicitly enforcing synthesizability in the generative model architecture[15–19]. SyntheMol belongs to the latter category; it treats the generative process as a search for effective molecules within a combinatorial chemical space composed of defined molecular building blocks and chemical reactions, which explicitly provides a scheme to synthesize every generated molecule.

Previous implementations of similar synthesis-aware approaches have either used a greedy search guided by molecular docking[55] that does not take advantage of the power of AI for property prediction and intelligent search, or they have used autoencoders[15,16,19], reinforcement learning[16,17] or genetic algorithms[18] that can be slow or challenging to optimize. By contrast, SyntheMol uses MCTS, which is fast and flexible, enabling its use with any combinatorial chemical space and property prediction model. MCTS has been used in previous studies for generating molecules[56,57], identifying functional groups[58] and planning retrosynthetic routes[59,60], but SyntheMol uses MCTS in combination with a property predictor to generate novel molecules from a multi-billion compound chemical space.

Crucially, we also introduce several post hoc filtering procedures that ensure the structural novelty, diversity and predicted antibacterial efficacy of the generated molecules. Structural novelty is particularly important to avoid generating analogues of known antibiotics, against which resistance is likely to rapidly emerge[13], and to ensure that the generative model does not simply replicate known active molecules or functional groups from the training set[61]. Notably, unlike most previous studies, which only perform in silico evaluations of generated molecules using imperfect predictors of synthesizability and efficacy[11], we chemically synthesized and experimentally tested 58 generated molecules, of which six were efficacious in the laboratory, representing an exceptional in vitro hit rate. Our entire pipeline—including training set curation, model training, molecule generation, chemical synthesis and experimental validation—could be performed in about 3 months (Fig. 1), demonstrating that generative AI is a powerful tool for rapidly exploring vast chemical spaces for new drug candidates that are easy to acquire in the laboratory.

The six highly potent molecules generated by SyntheMol may have potential clinical use on the basis of our in vitro experiments and in silico analysis of human toxicity. As single agents, these molecules display growth inhibitory activity against the Gram-positive bacterium methicillin-resistant *S. aureus*, as well as the lipopolysaccharide-deficient and colistin-resistant Gram-negative isolate *A. baumannii* ATCC 19606R. When administered in combination with an outer membrane perturbing agent, SPR 741 or colistin, all six molecules exhibit broad-spectrum activity across the diverse Gram-negative species *A. baumannii*, *E. coli* and *K. pneumoniae*, and one molecule, Enamine 40, also shows activity against *P. aeruginosa*. Due to these observed synergies in vitro, it may be possible to combine these generated compounds with potentiators at low dose in vivo, thereby reducing potential toxicities while maintaining antibacterial activity in dose-sparing combination therapies. In particular, the in vivo toxicity data presented here for En-10 and En-23 indicate that these compounds have minimal toxicity and are therefore promising as potentially safe and effective antibiotic candidates. Further discussion of future directions for these compounds and for SyntheMol is in the 'Extended Discussion' in the Supplementary Information.

While additional research is warranted to develop these generated molecules into reasonable antibiotic candidates, and to further improve generative AI methods for synthesizable molecule design, this work represents an important step towards the practical application of generative AI approaches for antibiotic discovery and drug discovery more broadly.

Methods

Training set

The training set consists of three libraries of compounds. Library 1 consists of 2,371 molecules from the Pharmakon-1760 library (1,360 Food and Drug Administration (FDA)-approved drugs and 400 internationally approved drugs) and from a set of 800 natural products isolated from plant, animal and microbial sources. Library 2 is the Broad Drug Repurposing Hub with 6,680 molecules, many of which are FDA-approved drugs or clinical trial candidates[62]. Library 3 is a synthetic small-molecule screening collection with 5,376 molecules, which were randomly sampled from a larger chemical library at the Broad Institute. All three libraries were screened in two biological replicates against *A. baumannii* ATCC 17978 for growth inhibitory activity. Cells were grown overnight at 37 °C in 2 ml of Luria-Bertani (LB) medium and then diluted 1:10,000 in fresh LB. Next, 49.5 µl (or 99 µl) of cells were added to every well of Corning 384-well (or 96-well) flat-bottom plates either manually or using an Agilent Bravo liquid handling system. Each compound was then added to a final screening concentration of 50 µM in a final volume of 50 µl (or 100 µl). Plates were then incubated at 37 °C without shaking for 16 h. Plates were then read at 600 nm using a SpectraMax M3 plate reader (Molecular Devices) and data were normalized by plate using interquartile mean before data compiling and hit identification (below).

Training set processing

For each library separately, we computed the average normalized OD_{600} for each compound from the two biological replicates. Next, we computed the mean μ and standard deviation σ of these average normalized OD_{600} values across the compounds of the library. We used the threshold $\mu - 2\sigma$ to binarize these values, with all values below the threshold labelled active and all values at or above this threshold labelled inactive. Next, we canonicalized the SMILES[63] for each dataset using RDKit[27] v.2022.03.4 and then combined the three libraries into a single dataset with 14,427 data points. In this combined dataset, for any data points with matching SMILES and binary activity labels, we kept one data point and removed the others, resulting in 13,594 data points. For any data points with matching SMILES and conflicting binary activity values (at least one with an active label and at least one with an inactive label with the same SMILES), we removed all samples to avoid noisy activity labels. This resulted in a final dataset of 13,524 unique molecules. Among these molecules, 470 (3.5%) are active and 13,054 (96.5%) are inactive.

ChEMBL antibiotics

To obtain a set of known antibiotics for comparison, we queried the ChEMBL database[25] using the search terms 'antibiotic' and 'antibacterial' on 8 August 2022. The term 'antibiotic' returned 636 molecules (https://www.ebi.ac.uk/chembl/g/#search_results/compounds/query=antibiotic), of which 587 had SMILES. The term 'antibacterial' returned 604 molecules (https://www.ebi.ac.uk/chembl/g/#search_results/compounds/query=antibacterial), of which 589 had SMILES. We merged the two sets of compounds, removed molecules with missing SMILES, converted the SMILES to canonical SMILES using RDKit and then deduplicated the compounds on the basis of canonical SMILES. This resulted in a set of 1,005 unique molecules.

t-SNE visualizations

t-SNE visualizations[24] were created using scikit-learn's *t*-SNE applied to the Morgan[31] fingerprints of molecules with Jaccard (Tanimoto) as the distance metric, squared distances and a principal components analysis initialization. Morgan fingerprints were computed with radius 2 and 2,048 bits using RDKit's GetMorganFingerprintAsBitVect function. For large datasets, molecules were randomly sampled to represent the dataset in the *t*-SNE visualization.

Property predictor architectures

Chemprop. Chemprop is a molecular property prediction model that uses a directed message passing neural network to process molecules and make predictions about their molecular properties[26]. Chemprop extracts simple atom and bond features from the molecular graph, such as the type of each atom and the type of each bond, to create a feature vector for each atom and bond. Chemprop then applies three message passing steps, which use a neural network layer to iteratively merge information from neighbouring atoms and bonds. After the message passing steps, Chemprop sums all the merged feature vectors to create a single feature vector that represents the whole molecule. This feature vector is passed through a feed-forward neural network with two layers to predict the molecular property, which in this case is the probability of *A. baumannii* growth inhibition. We used Chemprop v.1.5.2 with PyTorch v.1.12.0.post2 (ref. 64).

Chemprop-RDKit. The Chemprop-RDKit model is a variant of the Chemprop model described above. This model applies the same message passing procedure to obtain a single feature vector representing the molecule, but before applying the two feed-forward neural network layers, the molecular feature vector from message passing is concatenated with 200 molecular features computed by RDKit. We refer to these features as 'RDKit features'. The concatenated vector is then used as input to the feed-forward neural network to make the property prediction.

Random forest. A random forest model uses a set of decision trees with rules based on the input features to make predictions[28]. Our random forest model takes the molecular graph and computes the same 200 RDKit features as in the Chemprop-RDKit model. These 200 RDKit features are fed as input to a random forest classifier model with 100 decision trees, each of which makes a binary classification prediction. The prediction of the random forest is the average of the predictions of the decision trees. We used the RandomForestClassifier from scikit-learn v.1.1.1 with all default settings besides n_jobs=−1 and random_state=0 (ref. 65).

Model training

All three models were trained using the same data splits. The data was randomly split into 80% train, 10% validation and 10% test for each of ten cross-validation folds. The two Chemprop models used the validation data for early stopping, while the random forest model did not use the validation data. The Chemprop models were trained for 30 epochs using the Adam optimizer with a binary cross-entropy loss. The models were evaluated on the test data using both ROC-AUC and PRC-AUC. When using each model type to guide SyntheMol, we computed the average score of the ensemble of ten models (one model from each of the ten folds) as the model score.

To estimate the ability of the models to generalize to structurally novel molecules, we evaluated the ability of our model to transfer across our three chemical libraries. Specifically, we used the same procedure as above to train a set of ten models of each of the three model types on each of the three chemical libraries separately. For a given library, we measured the mean ROC-AUC and PRC-AUC values on the 10% test data from each of the ten cross-validations. Then, to measure generalization, we used the ensemble of ten models trained on one library to make predictions on the other two libraries, and we evaluated the ROC-AUC and PRC-AUC across those whole libraries.

Enamine REAL Space

The Enamine REAL Space consists of 31 billion make-on-demand molecules that can be synthesized using in-house validated one-pot synthetic procedures, which are applied to an in-house qualified set of chemical building blocks as reactants[22]. The synthesis time is typically 3–4 weeks with an average success rate of over 80%. We used the

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}	$N_{\text{mols}} \subset C$	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}	$N_{\text{children}} \subset T$	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}	$N_{\text{siblings}} \subset T$	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}	$N_{\text{visit}} \in \mathbb{N}$	The number of times node N has been visited (that is, selected during a rollout).
Node value	N_{value}	$N_{\text{value}} \in \mathbb{R}$	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	$N_{\text{diversity}}$	$N_{\text{diversity}} \in \mathbb{N}$	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.
No. of rollouts	n_{rollout}	$n_{\text{rollout}} \in \mathbb{N}$	The number of rollouts to run the SyntheMol MCTS algorithm.
No. of reactions	n_{reaction}	$n_{\text{reaction}} \in \mathbb{N}$	The maximum number of reactions allowed during a rollout.

November 2021 version of the REAL reactions consisting of 169 chemical reactions and the 2021 q3-4 version of the REAL building blocks consisting of 138,085 building block molecules. We downloaded the 2022 q1-2 version of the REAL Space consisting of 31,507,987,117 molecules that can be produced using the building blocks and chemical reactions.

To prepare the building blocks for use in our model, we first used RDKit to convert the building blocks SDF file to SMILES. All the building blocks were successfully converted. Then, we deduplicated the molecules by SMILES, which left 134,609 unique molecules due to a lack of stereochemistry in the converted SMILES. We applied the RDKit salt remover to remove salts from the building blocks to prevent incorrect reaction template matching during generation, and we deleted 25 molecules whose salts could not be correctly removed. This left us with 132,479 unique molecules as our building block set (with 138,060 unique building block IDs due to duplicate SMILES without stereochemistry and salts).

We chose 13 of the most common REAL Space that account for 96.6% of REAL Space. Applying these 13 reactions to our collection of 132,479 processed building blocks can produce 29,575,293,692 billion molecules, which is 93.9% of REAL Space. For each of these reactions, we manually converted the visual reaction template in the reactions PDF from Enamine to an atom-mapped SMARTS[66] reaction template. If a given set of building blocks matches the SMARTS reaction template, then it is likely, but not guaranteed, that those building blocks can successfully participate in that chemical reaction.

To improve the chance that a set of building blocks that matches a SMARTS template can successfully participate in a given chemical reaction to form a REAL Space molecule, we iterated through the REAL Space to collect a set of building block molecules that appear as each reactant in each reaction. For example, we collect a set of all the building blocks that appear at least once as the first reactant in REAL reaction 1. Then, to determine whether certain building blocks can participate in a reaction, we check both whether the building blocks match the SMARTS template and whether the building blocks appear in the set of building blocks that are used at least once as that reactant. If both conditions hold, then it is very likely that the building blocks do participate in the chemical reaction and create a product molecule in the REAL Space.

SyntheMol

SyntheMol is a generative model that explores combinatorial chemical spaces, which are composed of molecules formed by molecular building blocks and chemical reactions, to find molecules with a desired property. SyntheMol uses a MCTS algorithm similar to that used by AlphaGo[29] to efficiently search this chemical space for desirable molecules. SyntheMol not only quickly identifies promising molecules, but it also specifies the synthetic route (that is, a series of one or more chemical reactions combining molecular building blocks) to construct that molecule. Below, we summarize the mathematical notations used to describe the SyntheMol MCTS algorithm and we provide pseudocode.

SyntheMol MCTS algorithm

```
require synthesis tree T, property prediction model
M, maximum number of rollouts n_rollout, maximum
number of reactions n_reaction
function MCTS():
    for i = 1 to n_rollout do:
      rollout(T.root)
    end for
    return all visited nodes in T with 1 molecule
and ≥ 1 reaction
function rollout(N):
    if node N has undergone ≥ n_reaction reactions
then
      return property prediction score of M applied
to molecules in N
    end if
    E ← expand_node(N)
    S ← select child node in E with largest MCTS
score
    return rollout(S)
function expand_node(N):
    E ← empty set of nodes
    foreach reaction R do
      if R is compatible with molecules in N then
        Add new node to E with each product of R
```

```
applied to molecules in N
      end if
    end for
    foreach building block B do
      if any reaction is compatible with B and
molecules in N then
        Add new node to E with B and molecules in N
      end if
  end for
  return E
```

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 $M : C \to \mathbb{R}$, 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 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 $N_{\text{mols}} \subset C$, which are either building blocks from B or molecules that can be produced from building blocks in B and reactions in R. All molecules 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, $N_{\text{children}} \subset T$, 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.

To generate molecules, SyntheMol uses a 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 $S(N) = \frac{Q(N)+P(N)U(N)}{D(N)}$, 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 $Q(N) = \frac{N_{\text{value}}}{N_{\text{visit}}}$ 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 $P(N) = \frac{1}{|N_{\text{mols}}|}\sum_{i=1}^{|N_{\text{mols}}|} M\left(N^i_{\text{mols}}\right)$ where M is the property prediction model and N^i_{mols} 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 $U(N) = c\frac{\sqrt{1+N_{\text{visit}}+\sum_{N' \in N_{\text{siblings}}} N'_{\text{visit}}}}{1+N_{\text{visit}}}$ where $c = 10$ is a hyperparameter controlling the exploration–exploitation tradeoff, N_{siblings} is the set of sibling nodes of N (that is, 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 $D(N) = e^{\frac{N_{\text{diversity}}-1}{100}}$ where $N_{\text{diversity}}$ 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 $n_{\text{rollout}} = 20{,}000$), 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 (that is, excluding the building blocks themselves). To ensure rapid, inexpensive and easy synthesis, we use $n_{\text{reaction}} = 1$ 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.

Generating molecules with SyntheMol

To evaluate SyntheMol in silico before applying it to antibiotic discovery, we selected the property cLogP, the computed octanol-water partition coefficient, which can be computed for any generated molecule. We computed cLogP values for the 13,524 molecules in our antibiotic training set with RDKit's MolLogP function, which uses the Wildman–Crippen atom-based scheme[67]. We then binarized cLogP values by labelling molecules with cLogP > 6.5 as 'active' and molecules with cLogP ≤ 6.5 as 'inactive'. We selected this threshold because it resulted in 495 (3.7%) active molecules in the training set, which is comparable to the 470 (3.5%) antibiotic hits. We trained two Chemprop models on the binary cLogP data: one model with the standard 30 epochs of training (ROC-AUC = 0.97, PRC-AUC = 0.74) and one model with only one epoch of training to better match the performance of the antibiotic property prediction models (ROC-AUC = 0.86, PRC-AUC = 0.20). Both models were evaluated using tenfold cross-validation. Note that we did not train a Chemprop-RDKit or random forest model on cLogP since both models are given the cLogP value as part of the 200 RDKit input features. We then applied SyntheMol with the two Chemprop models trained to predict cLogP for 20,000 rollouts each (~9 hours). We next applied SyntheMol to discover potential antibiotic candidates against *A. baumannii* by using each of the three antibiotic property prediction models for 20,000 rollouts each.

Filtering generated molecules

We filtered the SyntheMol-generated compounds to achieve three goals: (1) low structural similarity to known antibiotics and antibiotic functional groups to ensure discovery of novel chemical structures; (2) high model score to maximize the probability of antibiotic activity

and (3) high diversity within the set of selected molecules to enable the discovery of a variety of structurally novel molecules.

First, to obtain structural novelty, we compare the generated molecules to two sets of molecules with known antibiotic effect: (1) the 470 active molecules in our training set, and (2) the 1,005 antibiotic or antibacterial molecules from ChEMBL. For each of these two reference sets, we compute the Tversky[30] similarity between the Morgan fingerprints of molecules in the SyntheMol-generated set and molecules in the reference set. If we let X be the Morgan fingerprint of a generated molecule and Y be the Morgan fingerprint of a reference molecule, then the Tversky similarity between the two molecules is defined as

$$T_{\alpha,\beta}(X,Y) = \frac{|X \cap Y|}{|X \cap Y| + \alpha |X \setminus Y| + \beta |Y \setminus X|}$$

for some $\alpha,\beta \geq 0$. We use $\alpha = 0$ and $\beta = 1$, which simplifies to

$$T_{0,1}(X,Y) = \frac{|X \cap Y|}{|Y|}.$$

Thus, $T_{0,1}(X,Y)$ is asymmetric and measures the proportion of the chemical substructures in Y (the reference molecule) that also appear in X (the generated molecule). This contrasts with the typically used Tanimoto[33,34] similarity, also known as the Jaccard index, which is defined as

$$J(X,Y) = T_{1,1}(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

and is a symmetric measure of the ratio of shared substructures to total substructures in the two molecules. The $T_{0,1}$ Tversky similarity is preferred over the Tanimoto similarity in our application because it assigns high similarity to generated molecules that contain most or all of the substructures of a reference molecule, even if the generated molecule also contains many other substructures and would thus have a low Tanimoto similarity. Therefore, high Tversky similarity allows us to identify and filter out generated molecules that simply repeat antibiotic functional groups from the reference set, potentially with additional irrelevant atoms and bonds, rather than contain novel functional groups.

For each generated molecule, we compute the Tversky similarity between that molecule and every molecule in the reference set, and we determine the most similar reference molecule. If the Tversky similarity between the generated molecule and the most similar reference molecule is less than or equal to 0.5, then the generated molecule is kept, otherwise the molecule is removed since it is not structurally novel. We apply this filtering to the generated molecules using both the active molecules from the training set and the antibiotic and antibacterial molecules from ChEMBL.

After filtering by Tversky similarity to the two reference sets, we filter by model prediction score to ensure that we are selecting high-scoring molecules that are likely to have antibiotic activity. We rank the remaining generated molecules according to their prediction score and keep the top 20% of molecules.

To ensure structural diversity among the structurally novel, high-scoring compounds, we apply clustering to the remaining molecules. Specifically, we apply k-means clustering[32] with $k = 50$ (with KMeans from scikit-learn) using Morgan fingerprints as the molecular features and Tanimoto similarity as the distance metric. Here, we use Tanimoto similarity instead of Tversky similarity since we are interested in a symmetric similarity comparison between molecules within the generated set. After obtaining the 50 clusters, we select the molecule with the highest score in each cluster, giving us 50 molecules. Performing this filtering process for each of the three sets of generated molecules, one for each property prediction model, gave us a set of 150 molecules generated by SyntheMol.

Compound synthesis

Although all the generated molecules were designed using REAL building blocks and chemical reactions, many molecules that could theoretically be produced by these building blocks and chemical reactions do not appear as part of the REAL Space, probably due to additional chemical filtering rules to remove molecules that might be difficult to synthesize. For this reason, several of our generated molecules were not part of REAL Space and were not available. Among our 150 compounds, we ordered 70 from Enamine that were available for synthesis, with 28 compounds from SyntheMol with Chemprop, 25 compounds from SyntheMol with Chemprop-RDKit and 17 compounds from SyntheMol with random forest. Among these molecules, 58 molecules (83%) were successfully synthesized with purity greater than 90% in about 4 weeks, with 26 compounds from SyntheMol with Chemprop, 22 compounds from SyntheMol with Chemprop-RDKit and ten compounds from SyntheMol with random forest. Purity was verified using liquid chromatography–mass spectrometry (LC–MS) except in cases of poor solubility, compound instability under LC–MS conditions, or non-informative LC–MS. In these cases, proton nuclear magnetic resonance (^{1}H-NMR) was used to assess chemical purity.

Antibacterial potency analyses

A. baumannii ATCC 17978, *A. baumannii* clinical isolates (ARIsolate Bank), *E. coli* BW25113, *P. aeruginosa* PAO1, *K. pneumoniae* ATCC 43816, *S. aureus* USA 300 and *S. aureus* clinical isolates (ARIsolate Bank) were grown overnight at 37 °C in 3 ml of LB medium. *A. baumannii* ATCC 19606R and *A. baumannii* ATCC 19606R + *lpxA* were grown overnight at 37 °C in 3 ml of LB medium supplemented with 10 μg ml^{-1} colistin or 200 μg ml^{-1} ampicillin, respectively. Overnight cultures were then diluted 1:10,000 into fresh LB (*A. baumannii* ATCC 17978, *A. baumannii* clinical isolates, *A. baumannii* ATCC 19606R (with 10 μg ml^{-1} colistin), *A. baumannii* ATCC 19606R + *lpxA* (with 200 μg ml^{-1} ampicillin), *E. coli* BW25113, *P. aeruginosa* PAO1, *K. pneumoniae* ATCC 43816, *S. aureus* USA 300 and *S. aureus* clinical isolates), LB with a quarter MIC SPR 741 (*A. baumannii* ATCC 17978, *A. baumannii* clinical isolates, *E. coli* BW25113 and *P. aeruginosa* PAO1) or LB with a quarter MIC colistin (*A. baumannii* ATCC 17978, *A. baumannii* clinical isolates and *K. pneumoniae* ATCC 43816). Cells were then introduced to twofold serial dilutions of each generated or control compound, in a final volume of 100 μl, in Costar 96-well flat-bottom plates. Plates were incubated at 37 °C without shaking (*A. baumannii* ATCC 17978, *A. baumannii* ATCC 19606R, *A. baumannii* ATCC 19606R + *lpxA*, *A. baumannii* clinical isolates, *E. coli* BW25113, *P. aeruginosa* PAO1 and *K. pneumoniae* ATCC 43816) or with shaking at 900 rpm (*S. aureus* USA 300 and *S. aureus* clinical isolates) until untreated control cultures reached stationary phase. Plates were then read at 600 nm using a BioTek Neo2 plate reader.

FOR analyses

A. baumannii ATCC 19606R was grown overnight at 37 °C in 3 ml of LB medium. For FOR quantification, 6.6×10^6 CFU in 100 μl of volume was deposited onto solid LB plates supplemented with each Enamine compound at the noted concentrations. After 24, 48 and 72 h of incubation at 37 °C, colonies were counted and these values were divided by 6.6×10^6 CFU to quantify the FOR in each condition.

Chequerboard analyses

A. baumannii ATCC 17978, *E. coli* BW25113, *P. aeruginosa* PAO1 and *K. pneumoniae* ATCC 43816 were grown overnight at 37 °C in 3 ml of LB medium. Cells were then diluted 1:10,000 into fresh LB. Enamine chemical potentiation by either SPR 741 (*A. baumannii* ATCC 17978, *E. coli* BW25113 and *P. aeruginosa* PAO1) or colistin (*A. baumannii* ATCC 17978 and *K. pneumoniae* ATCC 43816) was determined by conducting standard chequerboard broth microdilution assays with eight twofold serially diluted concentrations of each generated molecule, and SPR 741 or colistin, against cells in a final volume of 100 μl. Costar 96-well flat-bottom plates were used for bacterial growth. Plates were

incubated without shaking at 37 °C until untreated control cultures reached stationary phase. Plates were then read at 600 nm using a BioTek Neo2 plate reader.

FICI were calculated as follows:

$FIC_i = \frac{MIC_{ac}}{MIC_a} + \frac{MIC_{bc}}{MIC_b} = FIC_a + FIC_b$ where MIC_a is the MIC of compound A alone; MIC_{ac} is the MIC of compound A in combination with compound B; MIC_b is the MIC of compound B alone; MIC_{bc} is the MIC of compound B in combination with compound A; FIC_a is the FIC of compound A and FIC_b is the FIC of compound B. Synergy is defined as FICI ≤ 0.5. Antagonism is defined as FICI ≥ 4.0.

Toxicity prediction

To estimate the toxicity profile of the generated molecules, we developed a predictor of clinical toxicity. Specifically, we trained a Chemprop-RDKit model on the ClinTox toxicity dataset[42], which consists of 1,478 molecules with two binary labels for each molecule: one indicating whether the compound was FDA approved and one indicating whether the compound failed clinical trials due to toxicity reasons. Since the two labels are almost perfectly inversely correlated (lack of clinical toxicity almost always implies FDA approval and vice versa in this dataset), we only used the clinical toxicity label. According to this label, 112 (7.58%) of the 1,478 molecules are toxic. We trained an ensemble of ten Chemprop-RDKit models on this data using the same model settings as with the antibacterial prediction model. The model obtained an average test ROC-AUC of 0.881 ± 0.045 and an average test PRC-AUC of 0.514 ± 0.141 across tenfold cross-validation. We applied this model to our generated molecules to make toxicity predictions using the ensemble average prediction, with lower numbers indicating less toxicity. For comparison, we also made predictions on all molecules in the ClinTox dataset, where the prediction for each molecule in the dataset is made by the one model for which that molecule was in the test set.

Animal toxicity model

Mouse model experiments were conducted according to the guidelines set by the Canadian Council on Animal Care, using protocols approved by the Animal Review Ethics Board and McMaster University under Animal Use Protocol no. 22-04-10. No animals were excluded from the analysis, and blinding was considered unnecessary. Six- to eight-week-old female C57BL/6N mice were administered 50 mg kg^{-1} En-10 or En-23 via intraperitoneal injection ($n = 3$). Phenotypic observations were monitored at 3, 6 and 24 h postinjection. For chemical preparation, generated compounds were weighed and solubilized in 5% DMSO + 20% polyethylene glycol (PEG) 300 + 75% sterile distilled water. The solution was mixed thoroughly to ensure homogeneity. For mice in control groups ($n = 3$), the same weight-normalized volume of vehicle (DMSO + PEG 300 + sterile distilled water) was administered via intraperitoneal injection.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data used in this paper, including training data and generated molecules, are available in the Supplementary Data. The data, along with trained model checkpoints and LC–MS and ^{1}H-NMR spectra, are available at https://doi.org/10.5281/zenodo.10257839 (ref. 68). The ChEMBL database can be accessed from www.ebi.ac.uk/chembl.

Code availability

Code for data processing and analysis, property prediction model training and SyntheMol molecule generation is available at https://github.com/swansonk14/SyntheMol (ref. 69). This code repository makes use of general cheminformatics functions from https://github.com/swansonk14/chemfunc as well as Chemprop model code from https://github.com/chemprop/chemprop.

References

1. Murray, C. J. et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet* **399**, 629–655 (2022).
2. Rice, L. B. Federal funding for the study of antimicrobial resistance in nosocomial pathogens: No ESKAPE. *J. Infect. Dis.* **197**, 1079–1081 (2008).
3. Ma, Y. et al. Considerations and caveats in combating ESKAPE pathogens against nosocomial infections. *Adv. Sci.* **7**, 1901872 (2020).
4. Tacconelli, E. et al. Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. *Lancet Infect. Dis.* **18**, 318–327 (2018).
5. Lee, C. R. et al. Biology of *Acinetobacter baumannii*: pathogenesis, antibiotic resistance mechanisms, and prospective treatment options. *Front. Cell. Infect. Microbiol.* **7**, 55 (2017).
6. Carracedo-Reboredo, P. et al. A review on machine learning approaches and trends in drug discovery. *Comput. Struct. Biotechnol. J.* **19**, 4538–4558 (2021).
7. Gaudelet, T. et al. Utilizing graph machine learning within drug discovery and development. *Brief. Bioinform.* **22**, bbab159 (2021).
8. Stokes, J. M. et al. A deep learning approach to antibiotic discovery. *Cell* **180**, 688–702.e13 (2020).
9. Rahman, A. S. M. Z. et al. A machine learning model trained on a high-throughput antibacterial screen increases the hit rate of drug discovery. *PLoS Comput. Biol.* **18**, e1010613 (2022).
10. Zeng, X. et al. Deep generative molecular design reshapes drug discovery. *Cell Rep. Med.* **3**, 100794 (2022).
11. Bilodeau, C., Jin, W., Jaakkola, T., Barzilay, R. & Jensen, K. F. Generative models for molecular discovery: recent advances and challenges. *WIREs Comput. Mol. Sci.* **12**, e1608 (2022).
12. Bian, Y. & Xie, X. Q. Generative chemistry: drug discovery with deep learning generative models. *J. Mol. Model.* **27**, 71 (2021).
13. Liu, G. & Stokes, J. M. A brief guide to machine learning for antibiotic discovery. *Curr. Opin. Microbiol.* **69**, 102190 (2022).
14. Gao, W. & Coley, C. W. The synthesizability of molecules proposed by generative models. *J. Chem. Inf. Model.* **60**, 5714–5723 (2020).
15. Bradshaw, J., Paige, B., Kusner, M. J., Segler, M. H. S. & Hernández-Lobato, J. M. A model to search for synthesizable molecules. In *Proc. 33rd International Conference on Neural Information Processing Systems* (eds Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F. & Fox, E. B.) 7937–7949 (Curran Associates Inc., 2019).
16. Bradshaw, J., Paige, B., Kusner, M. J., Segler, M. H. S. & Hernández-Lobato, J. M. Barking up the right tree: an approach to search over molecule synthesis DAGs. In *Proc. 34th International Conference on Neural Information Processing Systems* (eds Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F. & Lin, H.) 6852–6866 (Curran Associates Inc., 2020).
17. Gottipati, S. K. et al. Learning to navigate the synthetically accessible chemical space using reinforcement learning. In *Proc. 37th International Conference on Machine Learning* (eds Daumé III, H. & Singh, A.) 3668–3679 (PMLR, 2020).
18. Gao, W., Mercado, R. & Coley, C. W. Amortized tree generation for bottom-up synthesis planning and synthesizable molecular design. In *Proc. 10th International Conference on Learning Representations* (2022); https://openreview.net/forum?id=FRxhHdnxt1
19. Pedawi, A., Gniewek, P., Chang, C., Anderson, B. M. & Bedem, H. van den. An efficient graph generative model for navigating ultra-large combinatorial synthesis libraries. In *Proc. 36th International Conference on Neural Information Processing Systems* (eds Oh, A. H., Agarwal. A., Belgrave, D. & Cho, K.) (2022); https://openreview.net/forum?id=VBbxHvbJd94

20. Kocsis, L. & Szepesvári, C. Bandit based Monte-Carlo planning. In *Proc. European Conference on Machine Learning, ECML 2006* Vol. 4212 (eds Furnkranz, J. et al.) 282–293 (Springer, 2006).
21. Coulom, R. Efficient selectivity and backup operators in Monte-Carlo tree search. In *Proc. International Conference on Computers and Games, CG 2006* Vol. 4630 (eds van den Herik, H. J. et al.) 72–83 (Springer, 2007).
22. Grygorenko, O. O. et al. Generating multibillion chemical space of readily accessible screening compounds. *iScience* **23**, 101681 (2020).
23. Stokes, J. M., Davis, J. H., Mangat, C. S., Williamson, J. R. & Brown, E. D. Discovery of a small molecule that inhibits bacterial ribosome biogenesis. *eLife* **3**, e03574 (2014).
24. van der Maaten, L. & Hinton, G. Visualizing data using *t*-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
25. Mendez, D. et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **47**, D930–D940 (2019).
26. Yang, K. et al. Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* **59**, 3370–3388 (2019).
27. RDKit: open-source cheminformatics. *RDKit* https://www.rdkit.org/. Accessed 28 Mar 2022.
28. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
29. Silver, D. et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
30. Tversky, A. Features of similarity. *Psychol. Rev.* **84**, 327–352 (1977).
31. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
32. Arthur, D. & Vassilvitskii, S. K-Means++: the advantages of careful seeding. In *Proc. Eighteenth Annu. ACM-SIAM Symp. Discrete Algorithms* 1027–1035 (SIAM, 2007).
33. Maggiora, G., Vogt, M., Stumpfe, D. & Bajorath, J. Molecular similarity in medicinal chemistry: miniperspective. *J. Med. Chem.* **57**, 3186–3204 (2014).
34. Tanimoto, T. T. *IBM Internal Report* (IBM, 1957).
35. Nikaido, H. Molecular basis of bacterial outer membrane permeability revisited. *Microbiol. Mol. Biol. Rev.* **67**, 593–656 (2003).
36. Zurawski, D. V. et al. SPR741, an antibiotic adjuvant, potentiates the in vitro and in vivo activity of rifampin against clinically relevant extensively drug-resistant *Acinetobacter baumannii*. *Antimicrob. Agents Chemother.* **61**, e01239-17 (2017).
37. Eckburg, P. B. et al. Safety, tolerability, pharmacokinetics, and drug interaction potential of SPR741, an intravenous potentiator, after single and multiple ascending doses and when combined with β-lactam antibiotics in healthy subjects. *Antimicrob. Agents Chemother.* **63**, e00892-19 (2019).
38. Moffatt, J. H. et al. Colistin resistance in *Acinetobacter baumannii* is mediated by complete loss of lipopolysaccharide production. *Antimicrob. Agents Chemother.* **54**, 4971–4977 (2010).
39. O'Neill, A. J., Cove, J. H. & Chopra, I. Mutation frequencies for resistance to fusidic acid and rifampicin in *Staphylococcus aureus*. *J. Antimicrob. Chemother.* **47**, 647–650 (2001).
40. Björkholm, B. et al. Mutation frequency and biological cost of antibiotic resistance in *Helicobacter pylori*. *Proc. Natl Acad. Sci. USA* **98**, 14607–14612 (2001).
41. Nicholson, W. L. & Maughan, H. The spectrum of spontaneous rifampin resistance mutations in the rpoB Gene of Bacillussubtilis 168 spores differs from that of vegetative cells and resembles that of *Mycobacterium tuberculosis*. *J. Bacteriol.* **184**, 4936–4940 (2002).
42. Wu, Z. et al. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **9**, 513–530 (2018).
43. Melo, M. C. R., Maasch, J. R. M. A. & de la Fuente-Nunez, C. Accelerating antibiotic discovery through artificial intelligence. *Commun. Biol.* **4**, 1050 (2021).
44. Yan, J. et al. Recent progress in the discovery and design of antimicrobial peptides using traditional machine learning and deep learning. *Antibiotics* **11**, 1451 (2022).
45. Mahlapuu, M., Håkansson, J., Ringstad, L. & Björn, C. Antimicrobial peptides: an emerging category of therapeutic agents. *Front. Cell. Infect. Microbiol.* **6**, 194 (2016).
46. Mahlapuu, M., Björn, C. & Ekblom, J. Antimicrobial peptides as therapeutic agents: opportunities and challenges. *Crit. Rev. Biotechnol.* **40**, 978–992 (2020).
47. Gómez-Bombarelli, R. et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **4**, 268–276 (2018).
48. Kang, S. & Cho, K. Conditional molecular design with deep generative models. *J. Chem. Inf. Model.* **59**, 43–52 (2019).
49. Krenn, M., Häse, F., Nigam, A., Friederich, P. & Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): a 100% robust molecular string representation. *Mach. Learn. Sci. Technol.* **1**, 045024 (2020).
50. Liu, Q., Allamanis, M., Brockschmidt, M. & Gaunt, A. L. Constrained graph variational autoencoders for molecule design. In *Proc. 32nd International Conference on Neural Information Processing Systems* (eds Wallach, H. M., Larochelle, H., Grauman, K. & Cesa-Bianchi, N.) 7806–7815 (Curran Associates Inc., 2018).
51. You, J., Liu, B., Ying, R., Pande, V. & Leskovec, J. Graph convolutional policy network for goal-directed molecular graph generation. In *Proc. 32nd International Conference on Neural Information Processing Systems* (eds Wallach, H. M., Larochelle, H., Grauman, K. & Cesa-Bianchi, N.) 6412–6422 (Curran Associates Inc., 2018).
52. Jin, W., Barzilay, R. & Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. *ICML* **80**, 2323–2332 (2018).
53. Jin, W., Barzilay, R. & Jaakkola, T. Hierarchical generation of molecular graphs using structural motifs. *ICML* **119**, 4839–4848 (2020).
54. Bilodeau, C. et al. Generating molecules with optimized aqueous solubility using iterative graph translation. *React. Chem. Eng.* **7**, 297–309 (2022).
55. Sadybekov, A. A. et al. Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. *Nature* **601**, 452–459 (2022).
56. Yang, X., Zhang, J., Yoshizoe, K., Terayama, K. & Tsuda, K. ChemTS: an efficient python library for *de novo* molecular generation. *Sci. Technol. Adv. Mater.* **18**, 972–976 (2017).
57. Qian, H., Lin, C., Zhao, D., Tu, S. & Xu, L. AlphaDrug: protein target specific de novo molecular generation. *PNAS Nexus.* **1**, pgac227 (2022).
58. Jin, W., Barzilay, R. & Jaakkola, T. Multi-objective molecule generation using interpretable substructures. *ICML* **119**, 4849–4859 (2020).
59. Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).
60. Coley, C. W. et al. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **365**, eaax1566 (2019).
61. Walters, W. P. & Murcko, M. Assessing the impact of generative AI on medicinal chemistry. *Nat. Biotechnol.* **38**, 143–145 (2020).
62. Corsello, S. M. et al. The Drug Repurposing Hub: a next-generation drug library and information resource. *Nat. Med.* **23**, 405–408 (2017).
63. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **28**, 31–36 (1988).

64. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. In *Proc. 33rd International Conference on Neural Information Processing Systems* (eds Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F. & Fox, E. B.) 8026–8037 (2019).
65. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
66. Daylight Theory. SMARTS - a language for describing molecular patterns. *Daylight Chemical Information Systems Inc.* www.daylight.com/dayhtml/doc/theory/theory.smarts.html (2022).
67. Wildman, S. A. & Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* **39**, 868–873 (1999).
68. Swanson, K. et al. Generative AI for designing and validating easily synthesizable and structurally novel antibiotics: data and models. *Zenodo* https://doi.org/10.5281/zenodo.10257839 (2023).
69. Swanson, K. & Liu, G. swansonk/SyntheMol: SyntheMol. *Zenodo* https://doi.org/10.5281/zenodo.10278151 (2023).
70. Liu, G. et al. Deep learning-guided discovery of an antibiotic targeting *Acinetobacter baumannii*. *Nat. Chem. Biol.* **19**, 1342–1350 (2023).

Acknowledgements

This research was kindly supported by the Weston Family Foundation (POP and Catalyst to J.M.S.); the David Braley Centre for Antibiotic Discovery (J.M.S.); the Canadian Institutes of Health Research (J.M.S.); a generous gift from M. and M. Heersink (J.M.S.) and the Chan-Zuckerberg Biohub (J.Z.). We thank Y. Moroz for his help accessing and answering our questions about the Enamine REAL Space building blocks, reactions and molecules. We thank G. Dubinina for help obtaining generated compounds. We thank M. Karelina, J. Miguel Hernández-Lobato, A. Tripp and M. Segler for insightful discussions about our property prediction and generative methods. We thank J. Boyce and A. Wright for providing mutant strains of *A. baumannii* used in this study. K.S. acknowledges support from the Knight-Hennessy scholarship.

Author contributions

Conceptualization was carried out by K.S., G.L., J.Z. and J.M.S. Model development was performed by K.S. and G.L. Biological validation was carried out by D.B.C., A.A. and J.M.S. K.S., G.L., D.B.C., J.Z. and J.M.S. wrote the paper. J.Z. and J.M.S. supervised the work.

Competing interests

These authors declare the following competing interests: K.S. is employed part-time by Greenstone Biosciences; J.Z. is on the scientific advisory board of Greenstone Biosciences and J.M.S. is cofounder and scientific director of Phare Bio. The other authors declare no competing interests.

Additional information

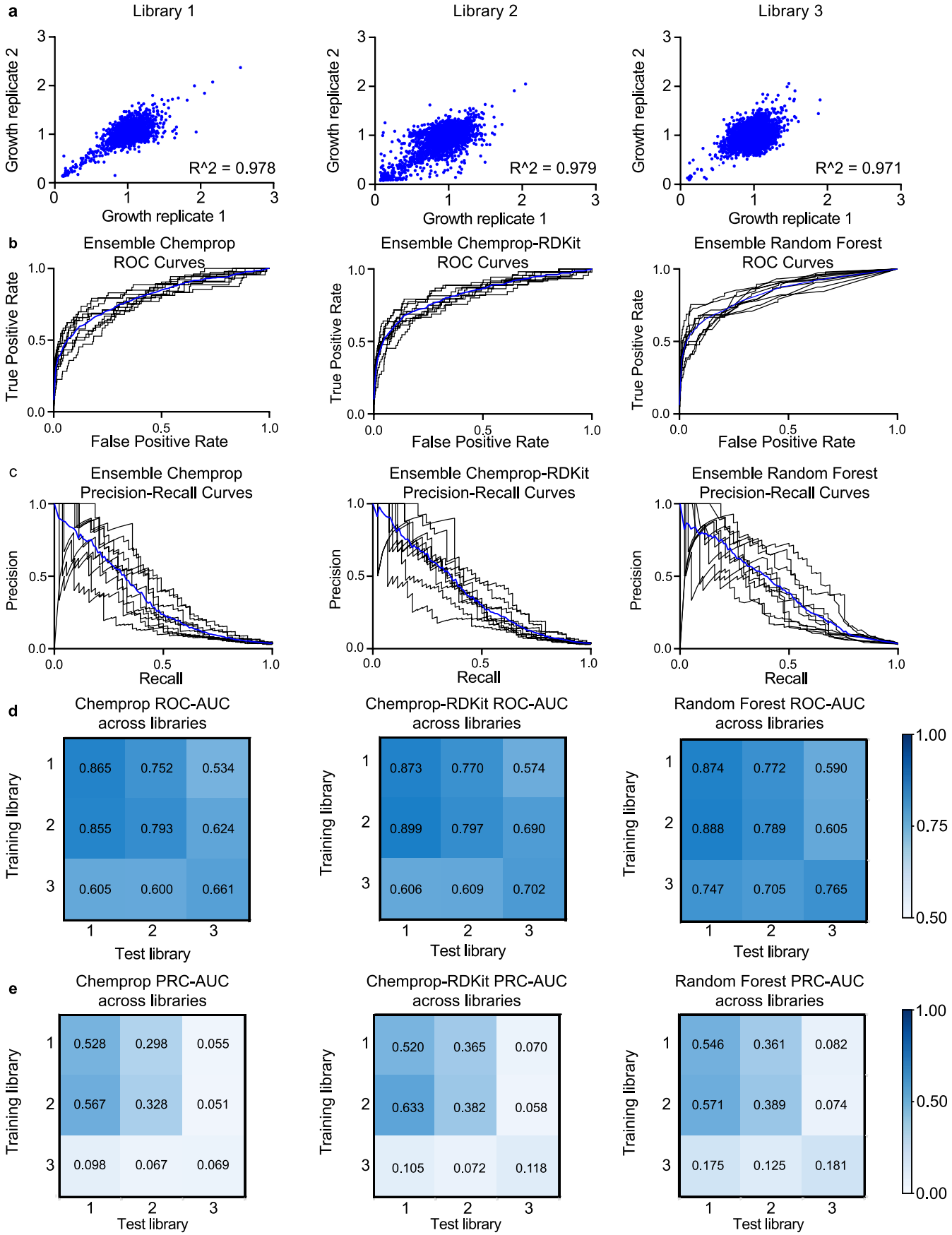
Extended data is available for this paper at https://doi.org/10.1038/s42256-024-00809-7.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-024-00809-7.

Correspondence and requests for materials should be addressed to James Zou or Jonathan M. Stokes.

Peer review information *Nature Machine Intelligence* thanks Feixiong Cheng and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s42256-024-00809-7

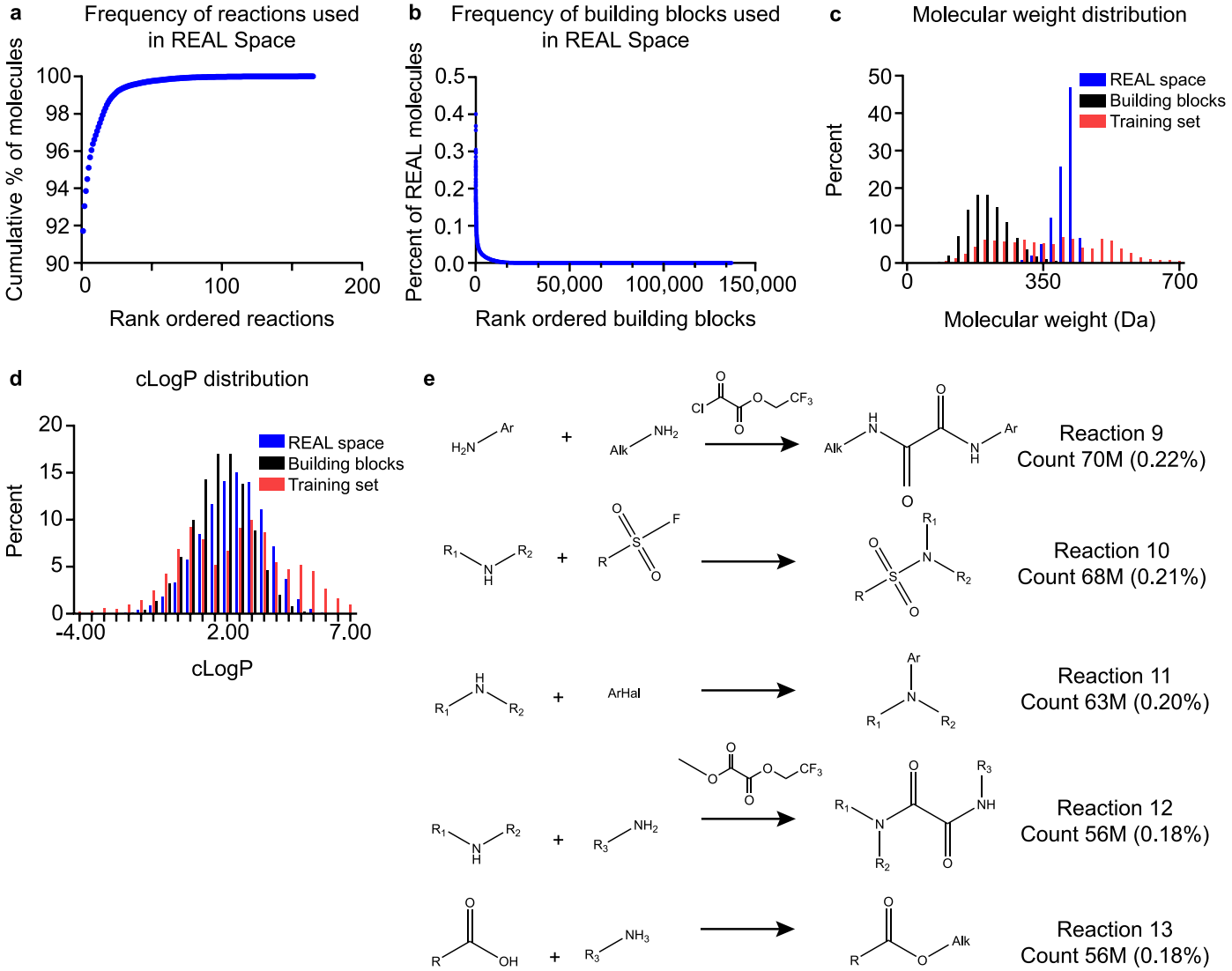


Extended Data Fig. 1 | See next page for caption.

https://doi.org/10.1038/s42256-024-00809-7

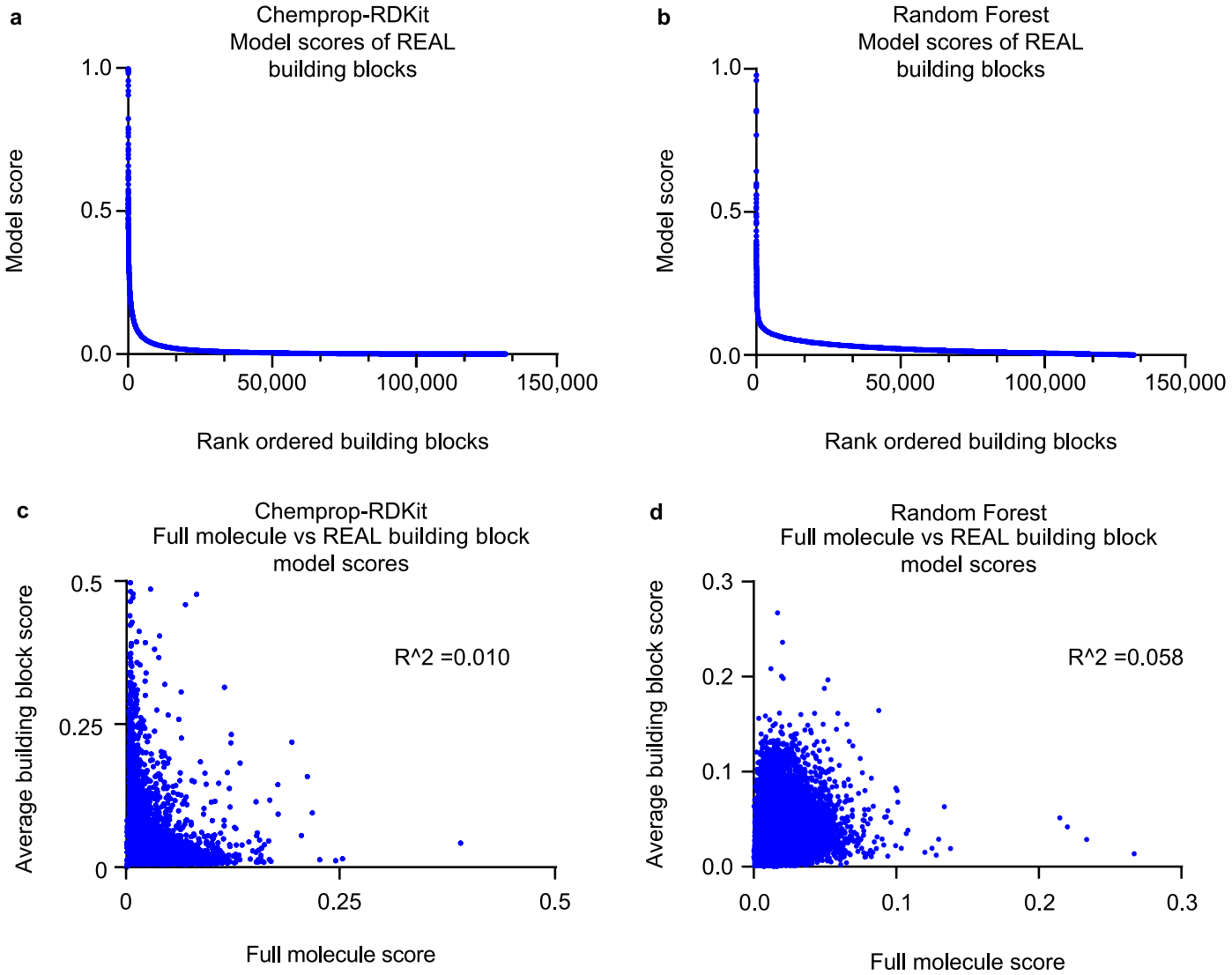
Extended Data Fig. 1 | Additional Property Prediction Model Development. (**a**) Normalized growth of *A. baumannii* ATCC 17978 in biological duplicate for each of the three training set chemical libraries. Note: compounds with normalized growth > 3 are removed for visual purposes. The R^2 values are the coefficient of determination. (**b**) Receiver operating characteristic (ROC) curves and (**c**) precision-recall (PRC) curves for the Chemprop, Chemprop-RDKit, and random forest models. For each model, the black lines show the performance of each of the ten models in the ensemble and the blue curve shows the average. (**d**, **e**) Model performance of each of our three property prediction models when generalizing between our three training set libraries. Values on the diagonal are the average test set performance of a model on a single library across tenfold cross-validation. Values on the off-diagonals are the result of applying an ensemble of ten models trained on one library to a different library and evaluating those predictions. (**d**) Performance measured by area under the receiver operating characteristic curve (ROC-AUC). (**e**) Performance measured by area under the precision-recall curve (PRC-AUC).

https://doi.org/10.1038/s42256-024-00809-7



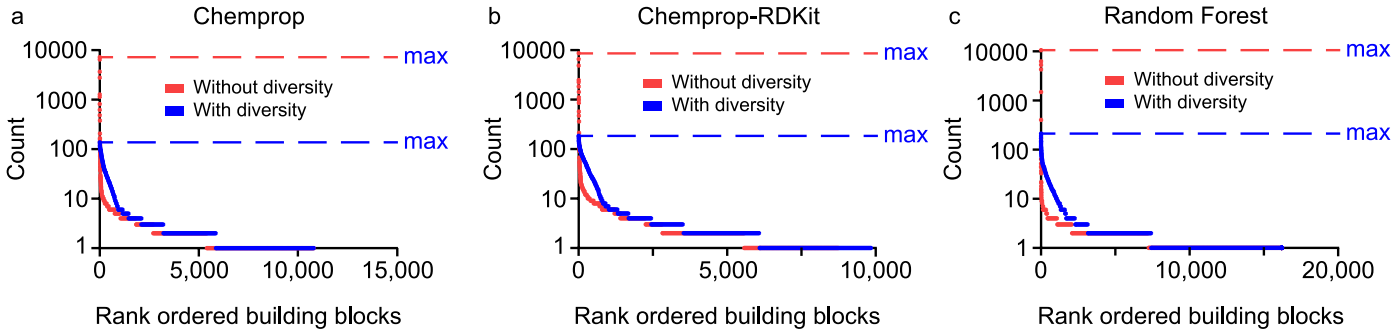
Extended Data Fig. 2 | REAL Space Analysis, Comparisons to Training Set, and Reactions. (**a**) The cumulative percent of molecules in REAL Space that can be produced by each of the 169 REAL chemical reactions. (**b**) The percent of molecules in REAL Space that include each of the REAL building blocks. (**c**) The molecular weight distribution of a random sample of 25,000 REAL molecules (blue), the REAL building blocks (black), and our training set molecules (red). (**d**) The cLogP distribution of a random sample of 25,000 REAL molecules (blue), the REAL building blocks (black), and our training set molecules (red). (**e**) The remaining 5 REAL chemical reactions we used (first 8 in Fig. 4b), along with the number and percent of REAL molecules produced by each reaction.

https://doi.org/10.1038/s42256-024-00809-7



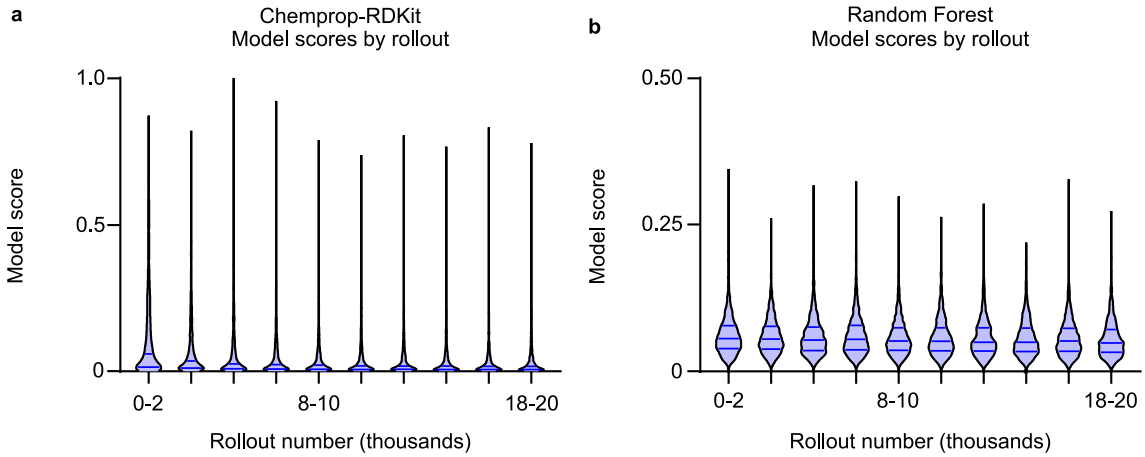
Extended Data Fig. 3 | REAL Building Block and Full Molecule Scores from Chemprop-RDKit and Random Forest. (**a**, **b**) The distribution of antibacterial model scores on the REAL building blocks using the (**a**) Chemprop-RDKit or (**b**) random forest models. (**c**, **d**) The correlation between the antibacterial model score of a REAL molecule and the average antibacterial model score of its constituent building blocks using the (**c**) Chemprop-RDKit or (**d**) random forest models. The R^2 values are the coefficient of determination.

https://doi.org/10.1038/s42256-024-00809-7



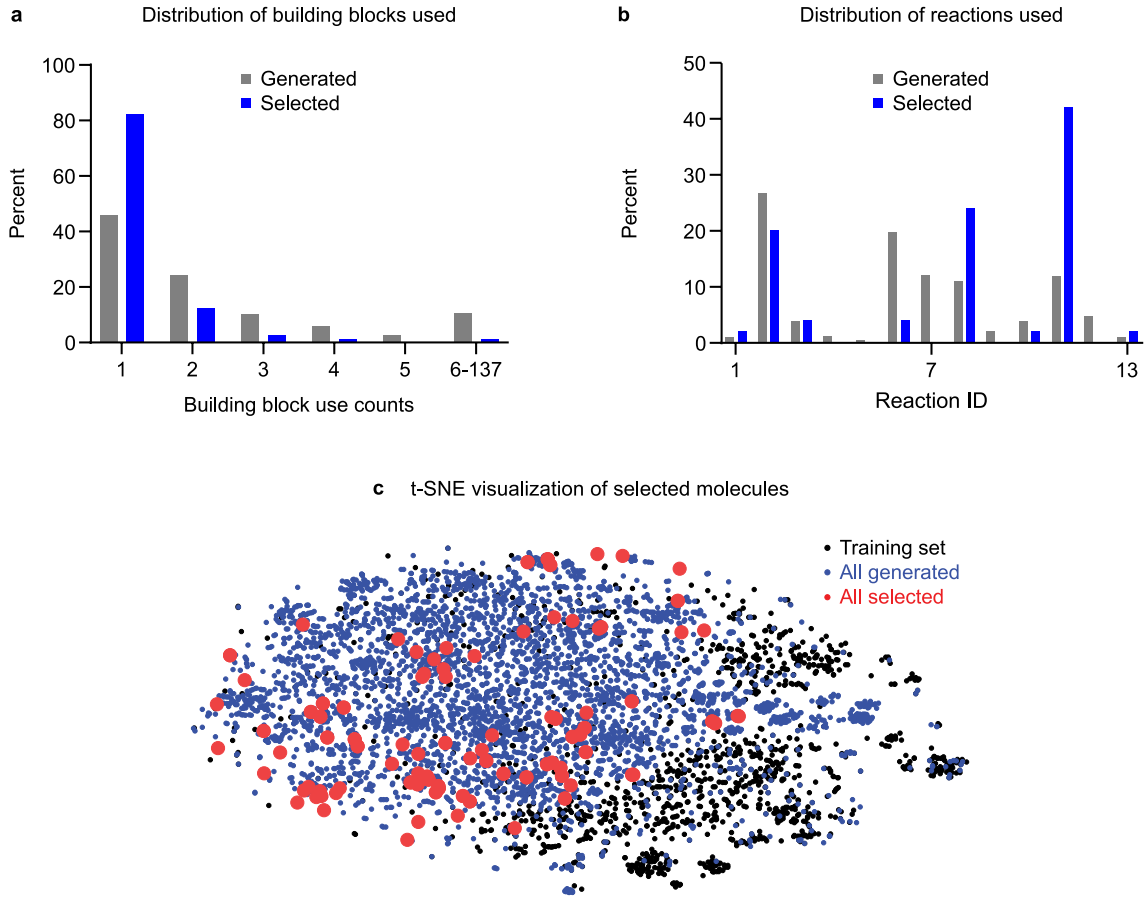
Extended Data Fig. 4 | Comparison of Generated Sets with and without Building Block Diversity. (**a**–**c**) The frequency with which building blocks were used in the generated molecules of SyntheMol, with and without the building block diversity score penalty for (**a**) Chemprop, (**b**) Chemprop-RDKit, and (**c**) random forest.

https://doi.org/10.1038/s42256-024-00809-7



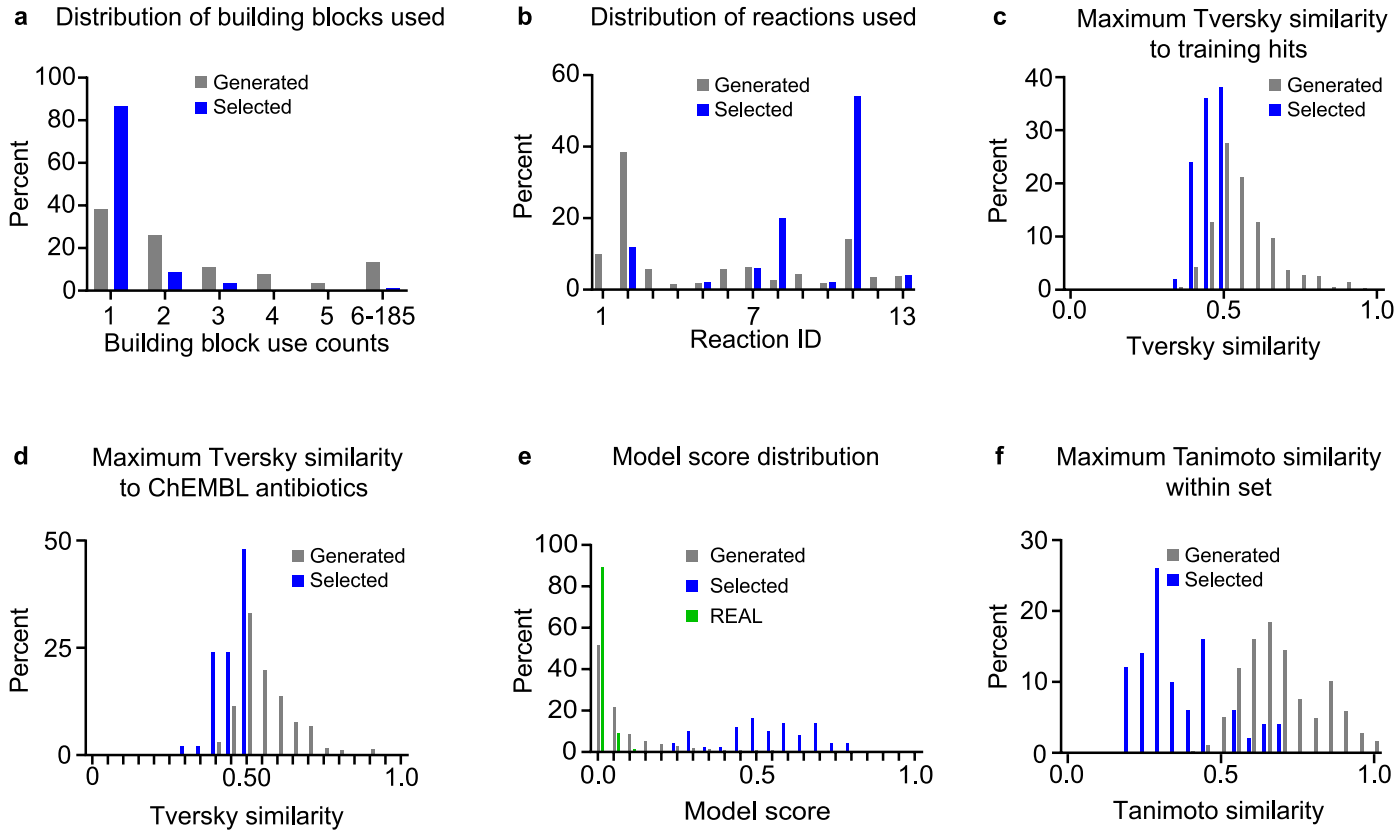
Extended Data Fig. 5 | Model Scores by Rollout from Chemprop-RDKit and Random Forest. (**a**, **b**) Violin plots of the distribution of antibacterial model scores for every 2,000 rollouts of the MCTS algorithm over 20,000 rollouts. SyntheMol uses the (**a**) Chemprop-RDKit or (**b**) random forest models for antibacterial prediction scores. The lines in each violin indicate the first quartile, the median, and the third quartile.

https://doi.org/10.1038/s42256-024-00809-7



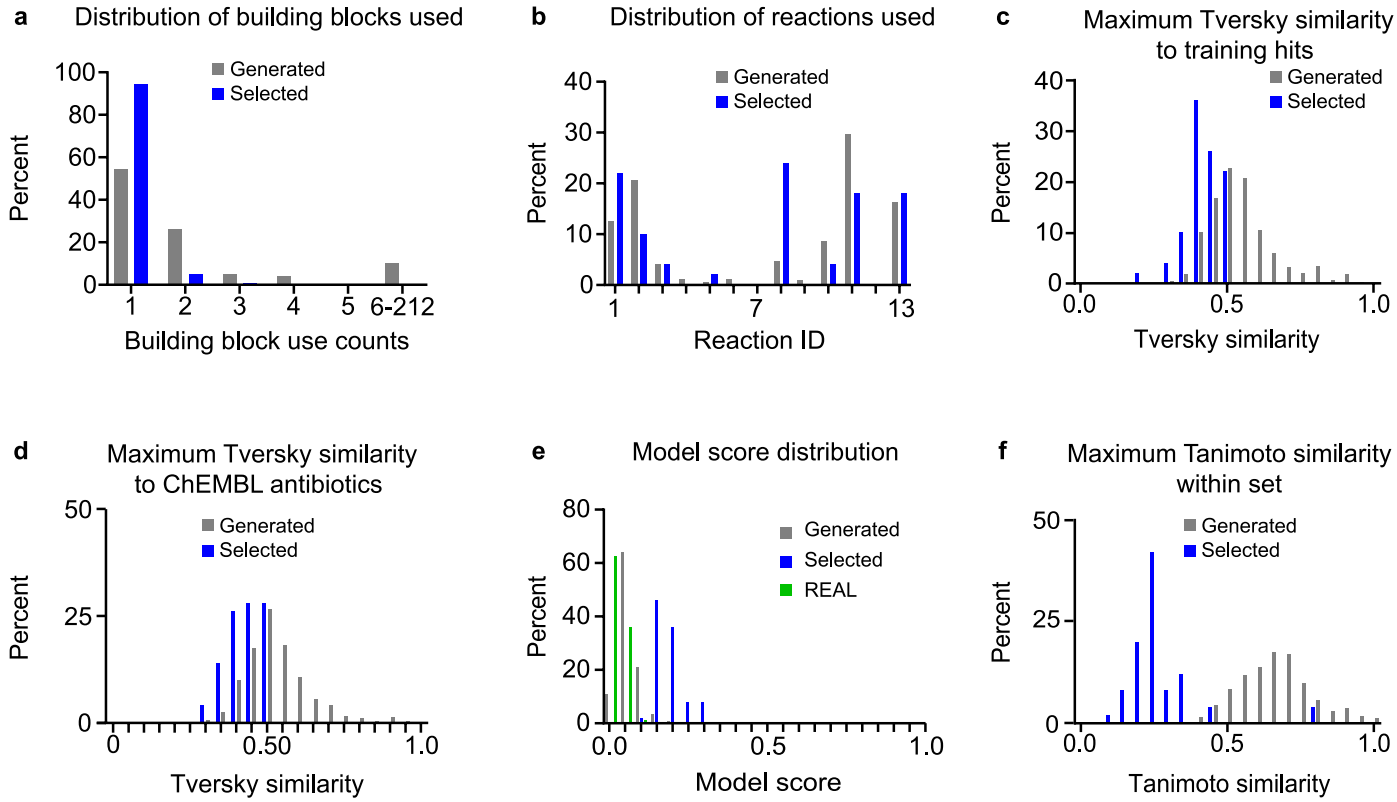
Extended Data Fig. 6 | Additional Analysis of Chemprop Generated and Selected Sets. (**a**) The percent of building blocks that appear at different frequencies among the generated or selected compounds by SyntheMol with Chemprop. Building blocks are assigned to bins on the x-axis based on the number of generated or selected compounds that contain that building block, with the final bin including building blocks that appear in at least six compounds (max 137). (**b**) The distribution of chemical reactions used by the generated or selected compounds by SyntheMol with Chemprop. (**c**) A t-SNE visualization of the training set along with all generated and selected molecules from each of the three property predictor models.

https://doi.org/10.1038/s42256-024-00809-7



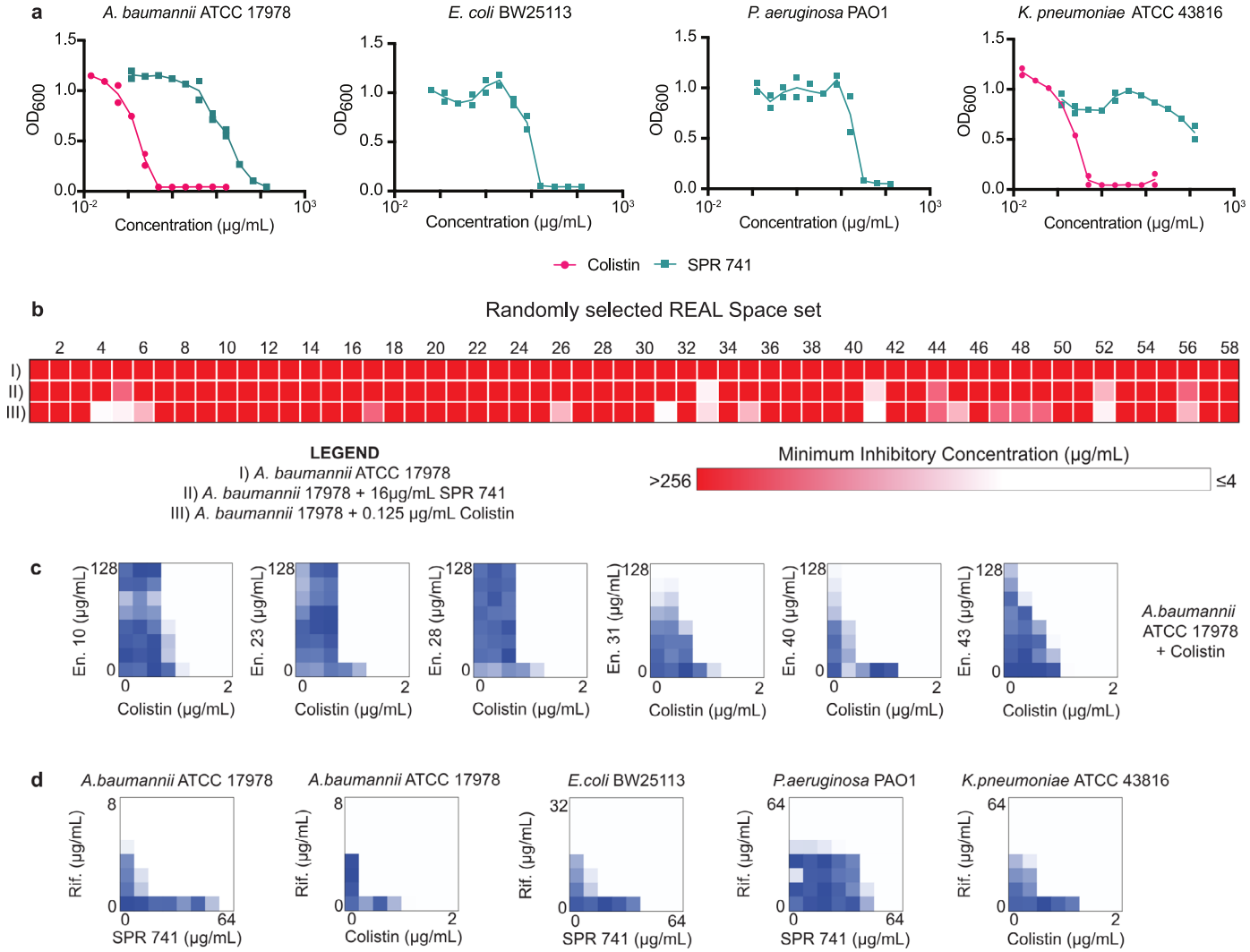
Extended Data Fig. 7 | Analysis of Chemprop-RDKit Generated and Selected Sets. (**a**) The percent of building blocks that appear at different frequencies among the generated or selected compounds by SyntheMol with Chemprop-RDKit. Building blocks are assigned to bins on the x-axis based on the number of generated or selected compounds that contain that building block, with the final bin including building blocks that appear in at least six compounds (max 185). (**b**) The distribution of chemical reactions used by the generated or selected compounds by SyntheMol with Chemprop-RDKit. (**c**–**f**) A comparison of the properties of the 25,828 molecules generated by SyntheMol with the Chemprop-RDKit antibacterial model and the 50 molecules selected from that set after applying post-hoc filters. (**c**) The distribution of nearest neighbour Tversky similarities between the generated or selected compounds and the active molecules in the training set. (**d**) The distribution of nearest neighbor Tversky similarities between the generated or selected compounds and the known antibacterial compounds from ChEMBL. (**e**) The distribution of Chemprop-RDKit antibacterial model scores on the generated or selected compounds, as well as on a random set of 25,000 REAL molecules. (**f**) The distribution of nearest neighbor Tanimoto similarities among the generated or selected compounds.

https://doi.org/10.1038/s42256-024-00809-7



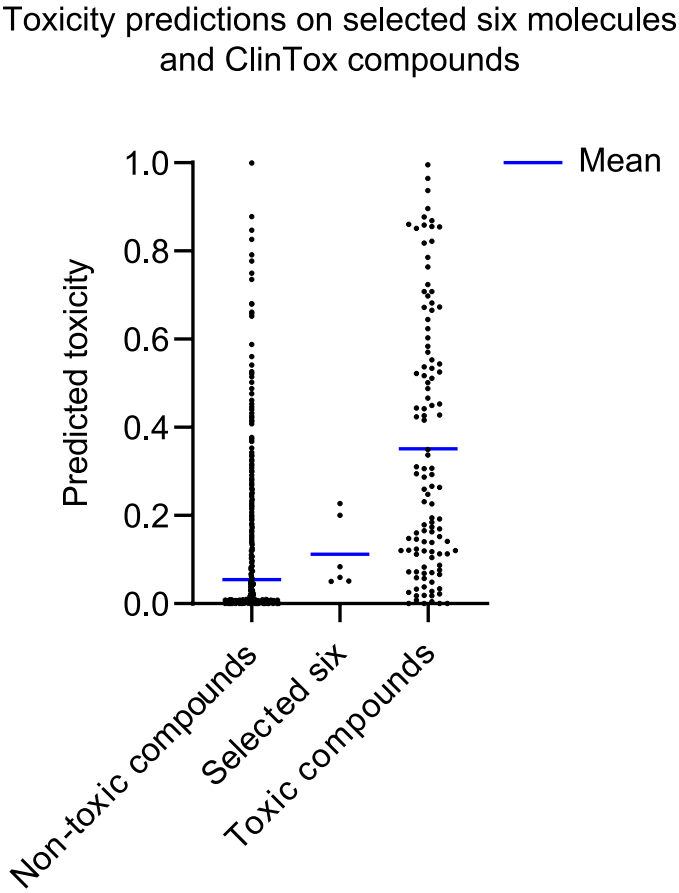
Extended Data Fig. 8 | Analysis of Random Forest Generated and Selected Sets. (**a**) The percent of building blocks that appear at different frequencies among the generated or selected compounds by SyntheMol with random forest. Building blocks are assigned to bins on the x-axis based on the number of generated or selected compounds that contain that building block, with the final bin including building blocks that appear in at least six compounds (max 212). (**b**) The distribution of chemical reactions used by the generated or selected compounds by SyntheMol with random forest. (**c**–**f**) A comparison of the properties of the 27,396 molecules generated by SyntheMol with the random forest antibacterial model and the 50 molecules selected from that set after applying post-hoc filters. (**c**) The distribution of nearest neighbor Tversky similarities between the generated or selected compounds and the active molecules in the training set. (**d**) The distribution of nearest neighbor Tversky similarities between the generated or selected compounds and the known antibacterial compounds from ChEMBL. (**e**) The distribution of random forest antibacterial model scores on the generated or selected compounds as well as on a random set of 25,000 REAL molecules. (**f**) The distribution of nearest neighbor Tanimoto similarities among the generated or selected compounds.

https://doi.org/10.1038/s42256-024-00809-7



Extended Data Fig. 9 | Additional In Vitro Validation. (**a**) Gram-negative bacterial isolates tested for growth inhibition against SPR 741 or colistin. Experiments were performed in biological duplicate. Error bars represent absolute range of optical density measurements at 600 nm. (**b**) Heat map summarizing MICs of 58 randomly selected compounds from the REAL Space against A. baumannii ATCC 17978 in I) LB medium, II) LB medium + a quarter MIC SPR 741, and III) LB medium + a quarter MIC colistin. Compounds were tested at concentrations from 256 μg/mL to 4 μg/mL in two-fold serial dilutions. Lighter colours indicate lower MIC values for each random REAL molecule. No compounds displayed potent antibacterial activity using the threshold of MIC ≤ 8 μg/mL. Experiments were performed in at least biological duplicate. (**c**, **d**) Chequerboard analysis to quantify synergy, as defined by FICI, with SPR 741 or colistin against Gram-negative isolates. Chequerboard experiments were performed using two-fold serial dilution series with the maximum and minimum concentrations of the potentiator (x-axis) and compound (y-axis) shown in μg/mL. Darker blue represents higher bacterial growth. Experiments were performed in biological duplicate. The mean growth of each well is shown. (**c**) Chequerboard assays using the six bioactive compounds, in combination with colistin, against A. baumannii ATCC 17978. (**d**) Chequerboard assays using rifampicin – a control antibiotic – in combination with SPR 741 or colistin against a panel of Gram-negative bacterial species.

https://doi.org/10.1038/s42256-024-00809-7



Extended Data Fig. 10 | Toxicity Predictions. Predictions of the probability of clinical toxicity using an ensemble of ten Chemprop-RDKit models trained on the ClinTox dataset. 'Non-toxic compounds' show the toxicity predictions of the model on the non-toxic molecules in the dataset (n = 1,372), where each molecule's prediction score comes from the one model in the ensemble for which that molecule was in the test set. 'Toxic compounds' shows the same toxicity predictions for the toxic molecules in the dataset (n = 112). 'Selected six' shows the average prediction of the ensemble of ten models on the six potent generated molecules. Blue horizontal lines represent the mean predictions for each set.

nature portfolio

Corresponding author(s): Jonathan M Stokes
James Y. Zou

Last updated by author(s): December 7, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	RDKit version 2022.03.4 Chemprop version 1.5.2 (w/ PyTorch version 1.12.0.post2.68) sckit-learn version 1.1.1 Additional custom software (e.g. SyntheMol) was developed. It is publicly available at https://github.com/swansonk14/SyntheMol (and archived at https://doi.org/10.5281/zenodo.10278151). This code repository makes use of general cheminformatics functions from https://github.com/swansonk14/chem_utils as well as Chemprop model code from https://github.com/chemprop/chemprop.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The data used in this paper, including training data and generated molecules, is available in the Supplementary Data. The data, along with trained model checkpoints and LC/MS and 1H-NMR spectra, are available here: https://doi.org/10.5281/zenodo.10257839.
ChEMBL Database: https://www.ebi.ac.uk/chembl/

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Not applicable
Reporting on race, ethnicity, or other socially relevant groupings	Not applicable
Population characteristics	Not applicable
Recruitment	Not applicable
Ethics oversight	Not applicable

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Mouse experiment sample size was determined using field convention group sizes, (for example, see Sarathy 2014). All in vitro experiments were conducted using biologically independent replicates, also based on field convention. A Lethal Murine Infection Model for Dengue Virus 3 in AG129 Mice Deficient in Type I and II Interferon Receptors Leads to Systemic Disease Vanessa V. Sarathy , Mellodee White , Li Li , Summer R. Gorder , Richard B. Pyles , Gerald A. Campbell , Gregg N. Milligan , Nigel Bourne , and Alan D. T. Barrett
Data exclusions	No data was excluded from any analyses.
Replication	All in vitro experiments were conducted using biologically independent replicates. We define biological independence as experimental replicates being conducted using identical protocols on different days.
Randomization	Randomization was only applicable to mouse studies . Before treatment, mice were relocated at random from housing cages to single-occupancy control or treatment cages. For in vitro antibiotic testing, randomization is not necessary since the same bacterial culture is aliquoted and each aliquot received the desired treatment with the appropriate controls running side-by-side.
Blinding	Investigators were not blinded for any in vitro or animal experiments. We were analyzing the antibacterial properties of new molecules against A. baumannii. Many of the analyses performed pertained to antibacterial efficacy, which is quantifiable using common laboratory approaches. Therefore, blinding was considered unnecessary.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	Six- to eight-week-old female C57BL/6N mice were purchased from Charles River Laboratories. Animals were housed in a specific pathogen free barrier unit under Level 2 conditions, temperature controlled (21℃) 30-50% humidity, 12h light and dark cycle environment (dark from 7pm to 7am) and were fed regular chow ad libtum.
Wild animals	No wild animals were used in the study.
Reporting on sex	All mice were female to allow for cohabitation of several mice per cage. Sex was also not deemed a relevant factor for mouse toxicity studies.
Field-collected samples	No field collected samples were used in the study.
Ethics oversight	Mouse model experiments were conducted according to the guidelines set by the Canadian Council on Animal Care, using protocols approved by the Animal Review Ethics Board and McMaster University under Animal Use Protocol #22-04-10.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Plants

Seed stocks	No plants were used in this study.
Novel plant genotypes	No plants were used in this study.
Authentication	No plants were used in this study.