Fleming: An AI Agent for Antibiotic Discovery in Mycobacterium tuberculosis

Ziming Wei[1,*], Yasha Ektefaie[1,*,‡], Andrew Zhou[1], Dereje Negatu [5,6], Bree Aldridge[7,8,9,10], Thomas Dick[5,12,13], Michael Skarlinski[14], Andrew D. White[14], Samuel Rodriques[14], Sepideh Hosseiniporgham[11], Inna Krieger[11], James Sacchettini[11], Marinka Zitnik [1,3,4,δ] and Maha Farhat[1,2,‡,δ]

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[2]Division of Pulmonary and Critical Care, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA
[3]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[4]Harvard Data Science Initiative, Cambridge, MA 02138, USA
[5]Center for Discovery and Innovation, Hackensack Meridian Health, Nutley, New Jersey, USA
[6]Center for Innovative Drug Development and Therapeutic Trials for Africa (CDT-Africa), Addis Ababa University, Addis Ababa, Ethiopia
[7]Department of Molecular Biology and Microbiology, Tufts University School of Medicine, Boston, MA, United States
[8]Graduate School of Biomedical Sciences, Tufts University School of Medicine, Boston, MA, United States
[9]Stuart B. Levy Center for Integrated Management of Antimicrobial Resistance, Boston, MA, United States
[10]Department of Biomedical Engineering, Tufts University School of Engineering, Medford, MA, United States
[11]Departments of Biochemistry and Biophysics, Texas A&M University, College Station, Texas 77843, United States
[12]Department of Medical Sciences, Hackensack Meridian School of Medicine, Nutley, New Jersey, USA
[13]Department of Microbiology and Immunology, Georgetown University, Washington, DC, USA
[14]FutureHouse Inc., San Francisco, CA

*Contributed equally ‡Corresponding authors: yasha_ektefaie@hms.harvard.edu, maha_farhat@hms.harvard.edu δCo-senior authors

Tuberculosis (TB) remains a critical public health challenge, causing over 1.5 million deaths annually. Rising resistance to existing TB antibiotics threatens to reverse decades of progress in combating the disease. Developing new antibiotics is challenging due to high development costs and failure rates. Artificial intelligence offers a promising solution to streamline early drug discovery by predicting inhibitory properties of novel compounds, generating molecules with desired properties, and automating workflows through large language model agents. In this work, we introduce Fleming, an AI agent designed for TB antibiotic discovery. We curated and generated the largest dataset of TB inhibitors to date (n=114,933), and used these data to train both discriminative and generative models for TB inhibitor identification. Fleming orchestrates four specialized agents—a bacterial inhibition prediction agent, a molecular

generation agent, a molecular optimization agent, and an ADMET agent—to perform key tasks in early TB antibiotic discovery. These agents incorporate literature search capabilities and chemical optimization tools, creating an integrated platform for TB antibiotic discovery. Using experimental data and human medicinal chemist review, we demonstrate how Fleming enhances the antibiotic discovery process, and mirrors the decision-making of medicinal chemists through an intuitive natural language interface.

Background

Tuberculosis (TB) is a major global public health threat. In 2022, there were more than 10 million TB cases globally and 1.5 million deaths [8]. Multi-drug resistant (MDR) strains of *Mycobacterium tuberculosis* (MTB), affect an estimated 550,000 new cases annually. These resistant forms of TB contribute about 13% of all deaths worldwide associated with antibiotic resistance [53, 5]. New antibiotics to treat TB are needed, but their discovery and development face several challenges. Though high-throughput antibacterial screens (HTS) can screen hundreds to thousands of small molecule candidate compounds for their ability to inhibit growth of whole bacterial cells in culture, they ultimately study a limited and over-explored space of chemical candidates. Desirable candidates are structurally and functionally novel minimizing the likelihood of pre-existing MTB resistance but must meet a long list of additional criteria that include high chemical stability, ease of synthesis, favorable absorption, distribution, metabolism, excretion properties, and minimal to no toxicity (ADMET). These characteristics and others contribute to a small molecule's potential as a plausible antibiotic lead worthy of additional investment in the next stages of *in vivo* development.

Artificial intelligence (AI) has been proposed as a solution to support rapid and more efficient pre-clinical selection of antibiotic candidates [42, 52, 34]. Several machine learning models have successfully predicted molecular properties from a molecule's basic chemical properties and its 2D or 3D structural information. Previous models have used a variety of architectures, from regression and random forests [30, 58, 6, 38, 47] to graph neural networks (GNNs) [57, 39, 7, 55, 20] and transformers [59, 22, 60, 23, 15, 28, 24]. Despite the success of these models[42, 34], there are concerns about the novelty of candidates they identify compared with the chemicals used for training. More recently, generative models have been proposed as a solution to create structurally distinct small molecule candidates using diffusion models [12] but to date have not been applied for antibiotic development. Monte Carlo search algorithms with an inhibitor model oracle [44] have been used, but are expected to be more limited in generating structural novelty than diffusion models.

While machine learning (ML) models have precedent in predicting whole cell bacterial inhibitory effects of small molecules, finding an inhibitor is just one aspect in the drug discovery process. For example, molecules with detergent-like properties are reliably predicted as inhibitors using *in vitro* screens and ML models trained on screen data, but are not drug leads. ML pipelines need to consider a broad number of chemical and biological properties to streamline and automate the drug discovery process. These properties include favorable ADMET. ML methods to predict

ADMET are currently limited by the limited number of proxy ADMET metrics and relatively small low chemical diversity datasets with experimentally measured ADMET metrics. Insights from the chemical and medical literature codified as textual data can aid in characterizing drug-like properties and molecules that may be out-of-distribution or structurally distinct from ADMET ML model training data. Large language models (LLMs) are increasingly being used to evaluate molecules and incorporate insights beyond those captured by current models [46, 54, 27]. A particularly attractive feature of LLMs is their use as agents with access to both textual data and molecular ML models allowing for automation of a complex drug discovery pipeline. Another attractive feature of these agents is their interactivity and the ability to provide rationale and sources to support the recommendation of a particular drug lead [25].

Integrating individually validated AI components for small-molecule drug discovery can significantly increase the novelty and success rate of compounds identified during pre-clinical screening. We introduce Fleming, an AI agent designed for TB antibiotic discovery. Fleming utilizes molecular ML tools to perform tasks central to antibiotic discovery, including predicting small molecule whole-cell inhibition, providing rationales by identifying chemical substructures responsible for inhibition, generating novel inhibitory candidates, and predicting and optimizing ADMET properties of candidate compounds. Specifically, Fleming integrates three main state-of-the-art molecular ML models: (1) a small molecule MTB inhibition graph neural network (GNN) that we built for Fleming and experimentally validated to have an 83% enrichment rate, (2) a diffusion model for generating novel inhibitory compounds, (3) a GNN for ADMET property prediction, with an LLM agent for literature-text search, based on GPT4 and PaperQA2 (PQA2), to incorporate relevant scientific literature into decision-making processes. Fleming synthesizes the outputs of these models to make decisions at different steps in the drug-development pipeline. In ADMET property prediction, Fleming combines the ADMET GNN with literature-based insights retrieved by PQA2 to improve accuracy on ADMET benchmarks. We also find that Fleming increases the plausibility of the generated and selected candidates to 82.7%, compared to 58% achieved by a pipeline that runs the molecular ML models sequentially. This improvement was validated through a benchmarking study in which MTB medicinal chemists manually assessed 100 candidate molecules for plausibility as TB antibiotics. Fleming is an AI agent with the potential to accelerate the pre-clinical selection of candidates for developing antibiotics against tuberculosis.

Results

Overview of Fleming

Fleming is an agent that generates novel and plausible TB antibiotics (Figure 1a). It consists of a central manager agent that orchestrates the actions of four specialized agents: the inhibition agent, ADMET agent, molecular generation agent, and molecular optimization agent. Together, these agents collaborate to fulfill user-defined tasks in antibiotic discovery. The inhibition agent takes a small molecule and predicts its inhibition of MTB using a model trained on data from 114,933 distinct molecules and their *in vitro* effect on TB growth. This inhibition model also identifies the molecular substructures that contribute to its predictions. The agent then leverages an LLM model, GPT-4o, to assess the plausibility of the molecule based on its structure.

The ADMET agent operates similarly, leveraging ADMET.AI molecular ML models [43] to predict a small molecule's ADMET properties. After this step, the agent uses PaperQA2 to assess the same properties by reviewing the relevant literature and, lastly, making a decision by combining the predictions of the molecular ML model and the literature. The molecular generation agent utilizes a conditional small molecule diffusion model and a Monte-Carlo Tree Search model to design novel compounds filtered for synthesizability and novelty. The molecular optimization agent receives an input molecule and directs a secondary agent to identify structurally similar molecules in chemical space that exhibit enhanced ADMET properties. The optimal molecule is then sent back to the manager, which determines whether to proceed with further optimization or to conclude the process. When a user queries Fleming, the manager automatically selects the appropriate agents to address the specific request. For instance, to assess whether a specific compound could be plausibly a TB antibiotic, Fleming engages the inhibition and ADMET agents. To optimize the properties of ADMET of a molecule, the molecular optimization agent is used (Figure 1 b-e).

Inhibitory agent identifies small molecule MTB inhibitors

To train the inhibition model, we collected data on the measured inhibitory response of MTB on 113,919 distinct molecules using luminescence and/or resazurin readouts under different growth conditions (Methods). The experimental data was generated in the same laboratory using mc2-7000 strain of MTB over three different screening campaigns. To these data, we added MTB inhibition data of 1000 molecular fragments generated from a second laboratory [31] and 14 current TB antibiotics. The final training data contains 6,572 (5.7%) MTB growth inhibitory compounds/fragments and 108,361 (94.3%) non-inhibitory compounds/fragments (Figure 2b).

We compare five state-of-the-art molecular property prediction model architectures, specifically Directed Message Passing Neural Networks (DMPNN) [57], Equivariant Graph Neural Networks (EGNN) [39], Transformer-M [26], Feed Forward Neural Networks (FFNN), and Random Forest on two accuracy metrics the AUROC, AUPRC, and the area under the spectral performance curve (AUSPC) generalizability metric [3] (Methods, Figure 2c). In 5-fold cross-validation, the DMPNN has a significantly higher AUROC (0.76, paired t-test p-value < 0.001), and AUPRC (0.35, paired t-test p-value < 0.001) compared to the four other model architectures (Figure 2d). We measured AUSPC using the spectral framework for model evaluation (SPECTRA) [3] that generates dataset splits based on a spectral parameter ranging from 0 to 1. Higher spectral parameters (closer to 1) produce splits with lower train-test similarity, while lower spectral parameters (closer to 0) result in splits with higher train-test similarity. The area under the performance-versus-spectral-parameter curve (AUSPC) quantifies the model's generalizability across varying spectral parameter values. We trained and evaluated each model on splits generated with spectral parameters of 0, 0.5, 0.65, and 0.75. The Directed Message Passing Neural Network (DMPNN) consistently outperformed other models on splits with higher spectral parameters, achieving an AUSPC of 0.35. (Figure 2e).

High throughput screening of bacterial cell inhibition *in vitro* is recognized to have high error rates, to address this we trained DMPNN with an evidential Dirichlet loss [40], which provides a confidence level output for each prediction. Thresholds for confidence filtering were set at the measured confidence value for current TB antibiotics. Confidence filtering increased the AUROC of the inhibition model by 4.2% (0.76 to 0.79, paired t-test p-value = 0.02), and the AUPRC by 11.9% (0.18 to 0.20, p-value = 0.01), and the AUSPC by 8.6% (0.35 to 0.38). (Figure 2d,e).

Prospective *in vitro* validation of inhibition models

To validate the generalizability of the inhibition agent, we prospectively predicted and then experimentally tested 439 additional molecules for inhibitory effects against MTB. These molecules were distinct from those in the training data, molecules had a median maximum pairwise Tanimoto similarity against the training set of 0.45 (IQR: [0.39, 0.51]) (Figure 3b), confirming high diversity of the validated molecules. Compared with the EGNN, Transformer-M, FFNN and RF models, the DMPNN model with evidential classification demonstrated the highest true positive and lowest false negative rate in the prospective evaluation. The DMPNN based agent identified 6 of 439 molecules as inhibitory, of which 5 (83.3%) were confirmed as inhibitory *in vitro* (Meth-

ods). Among the 433 predicted non-inhibitory molecules, 393 were confirmed non-inhibitors, but 40 inhibitors were missed by the DMPNN based agent (Figure 3a). We hypothesized that the high false negative rate is due to the structural diversity of the bench-validated molecules and the limitations of the training data. To assess this, we compared the structural diversity of the tested molecules by calculating their Tanimoto similarity against molecules in the training data. For molecules correctly identified by the model as inhibitory of MTB, the median maximum similarity to the training set was 0.55 (IQR: [0.48, 0.60]). This value was significantly higher than the median maximum similarity of inhibitory molecules missed by the model, which was 0.44 (IQR: [0.38, 0.49]) (Wilcoxon Rank Sum Test, $p = 0.03$). These results confirmed our choice of using DMPNN with evidential classification as the model for the inhibition agent (Figure 3c).

The ADMET agent characterizes and classifies small molecule ADMET properties

The ADMET agent is a GPT4-based system that combines molecular models predicting ADMET properties (ADMET.AI, [43]) with additional insights extracted from the chemical literature. ADMET.AI consists of two DMPNN models, one for regression tasks and one for classification tasks, which are trained on Therapeutic Data Commons (TDC) datasets [13] for 41 ADMET metrics. The ADMET agent augments the ADMET.AI predictions using PaperQA2 (PQA2) [41], a general purpose literature summary agent which we prompted to extract relevant ADMET information from the chemical literature (Figure 4a, Methods section 'ADMET Agent'). We also evaluated a version of the ADMET agent that exclusively utilizes GPT4 (AGPT, Methods). In this approach, a textual description of the molecule is provided to GPT4, and feedback is requested on a specific property of interest.

We benchmarked the added value of PQA2 (Methods). We selected three toxicity datasets from the TDC: carcinogenicity (Carcinogen) [21, 1], drug-induced liver injury (DILI) [56], and hERG channel blockage (hERG) [48]. The agent's zero-shot performance was assessed on the test sets of these datasets. Across all three datasets, the inclusion of PQA2 consistently improved the agent's area under the curve (AUC) compared to using AGPT. Specifically, the PQA2+AGPT achieved an AUC of 0.72 on the DILI dataset, outperforming AGPT with AUC of 0.63. On the hERG dataset, PQA2+AGPT reached an AUC of 0.76 compared to 0.64 for AGPT. Similarly, on the Carcinogen dataset, PQA2+AGPT achieved an AUC of 0.73, exceeding AGPT's AUC of 0.69 (Figure 4b).

We designed and benchmarked an approach to combine the strengths of PQA2+AGPT and ADMET.AI. Specifically, when ADMET.AI outputs an "uncertain" prediction, defined as a probability between 0.5 and 0.6, we substitute its prediction with that of PQA2+AGPT, allowing us to ascertain if the agent can fill gaps in ADMET.AI's performance by providing more confident and accurate predictions in these uncertain cases. This hybrid approach improved the overall AUC across all datasets. For the Carcinogen dataset, the AUC increased from 0.83 to 0.85 and the combined model provided likelihoods closer to the true label for 4 out of 5 uncertain predictions; for DILI, the AUC increased from 0.96 to 0.97 with 3 of 4 uncertain predictions corrected; while for hERG, AUC increased from 0.89 to 0.90 with 4 of 6 uncertain predictions corrected (Figure 4c). While ADMET.AI already performs strongly on these datasets, these results highlight the agent's ability to correct uncertain predictions and enhance performance.

The molecular generation agent generates novel small molecules with MTB inhibitory potential

Chemical libraries used in high-throughput screening will naturally cover only a small fraction of the possible chemical diversity. The molecular generation agent uses conditional generative models to generate small molecules from unexplored parts of chemical space (Figure 5a, Methods section 'Molecular generation agent'). The molecular optimization agent is an agent that iteratively modifies molecules to improve ADMET properties while maintaining predicted inhibitory effect (Figure 5b, Methods 'Molecular optimization agent').

We implemented a conditional equivariant diffusion model [12], training it on the data used to develop the inhibition model; conditioning was on the level of inhibition (Methods, 'Diffusion Generative Model'). This approach contrasts with existing methods that typically rely on rule-based combinations of predefined molecular subcomponents that constrain the molecule form. As the conditional diffusion model is mainly constrained on inhibition, it has the potential to generate structurally distinct molecules. The conditional aspect of the model enables targeted molecule generation by explicitly associating inhibition levels with specific regions of chemical space. This guidance improves the model's ability to explore underrepresented areas, enhancing both the structural diversity and the relevance of the generated molecules. We compared this conditional diffusion model to SyntheMol, a state-of-the-art in vitro generative small molecule model based on Monte Carlo Tree Search (MCTS) [44] (Methods, 'SyntheMol'). We evaluated and compared the synthesizability, novelty, and chemical space coverage of molecules generated

by these methods.

After comparing 10,000 molecules generated by each method, we found that while diffusion-generated molecules had lower synthesizability scores than those from the MCTS method, over 99% of diffusion-generated molecules still achieved acceptable synthesizability scores (SAScore <6) [4] (Figure 5c). The molecules generated from the diffusion generative model have significantly higher novelty (Median Tanimoto similarity = 0.26 vs. 0.40, Wilcoxon Rank Sum Test p-value < 0.0001, Figure 5d). The molecules generated from two different methods had limited overlap in chemical space (Figure 5e, Methods section 'Chemical Space Visualization'). Based on these results, we proceeded to collect generated molecules from both methods, filtered for the molecules that have high novelty and synthesizability as the output of the generative models (Methods section 'Synthesizability and novelty filter').

Since ADMET properties are influenced by functional groups or side chains that are not always needed for inhibitory effects, modifications to these regions can enhance ADMET properties without diminishing efficacy. We developed an agent that, given a seed molecule coordinates a chemical modifier algorithm, SynSpace [51], the inhibitory model and the ADMET agent to optimize a molecule for more favorable ADMET while retaining MTB inhibitory activity [9] (Figure 5b). To demonstrate this, we ran the molecular optimization agent on 100 generated molecules that passed initial inhibition and ADMET filters (Methods, 'Molecular optimization agent'). After optimization, the median number of failed ADMET tasks decreased significantly from 9 (IQR: [7, 11]) to 6 (IQR: [4, 7]) (Wilcoxon Rank Sum Test p-value<0.0001) while maintaining stable inhibition probability (median probability 0.92 to 0.88, Figure 5b).

Fleming antibiotic plausibility assessment aligns with medicinal chemist review

We evaluated Fleming's ability to identify plausible antibiotic candidates by benchmarking it against a manual workflow involving its individual components. The manual process began with the conditional diffusion model, which was used to generate potential TB antibiotic candidates. These candidates were then filtered using the inhibition model, followed by the ADMET.AI models. Molecules with good inhibition properties and nearly acceptable ADMET properties were passed to a molecular optimization step, resulting in a final set of 100 TB antibiotic candidates. Medicinal chemists rated 58% of these manually generated molecules as plausible antibiotics worth further development.

We then provided these 100 molecules to Fleming, which selected a subset that it deemed potentially plausible. Medicinal chemists rated 82.6% of Fleming's selected candidates as plausible antibiotics worth further development. Furthermore, Fleming's reasoning for antibiotic plausibility closely aligned with the assessments of medicinal chemists (Figure 6). For example, Fleming rejected an acetyl-carbamothioate, reasoning that "molecules with thioester linkages may not be very stable due to susceptibility to hydrolysis," a conclusion that precisely matched the medicinal chemists' evaluation.

Discussion

As drug-resistant TB continues to rise [53], the need for new antibiotics has become increasingly urgent. Machine learning offers a promising approach to accelerate antibiotic discovery by enhancing throughput and exploring new regions of chemical space. Fleming integrates recent advances in machine learning, combining discriminative models that predict the properties of small molecules [10, 43], generative models that create novel molecules with specific characteristics [44, 12], and large language models that embed textual knowledge into machine learning workflows. Fleming coordinates these tools to effectively address user queries in natural language. We validate Fleming's sub-agents through *in silico* and *in vitro* testing, demonstrating that Fleming's approach better reflects the reasoning of medicinal chemists compared to traditional, non-agentic pipelines.

Fleming can be adapted for antibiotic discovery beyond MTB. With a dataset of small molecules affecting another target organism, users can retrain the inhibition and generative models and integrate them into Fleming to design antibiotics specific to that organism. Fleming's modular structure also allows for easy integration of new agents as relevant to drug discovery, enabling it to incorporate other critical factors such as mechanism of action prediction or chemical stability.

Using a conditional diffusion model enabled us to generate a higher number of novel molecules than traditional rule-based approaches using molecular substructures. Diffusion models produced molecules within acceptable synthesizability score ranges, demonstrating their ability to create novel compounds without compromising synthesizability. However, many generated molecules were invalid (e.g., incorrect bond numbers). Although the ability to sample thousands of molecules mitigates this issue, incorporating multi-objective optimization into the diffusion model could enhance its efficiency. This would allow the model to better explore the chemical space, generating molecules that are not only inhibitory but also valid and optimized for other relevant properties, thereby improving overall throughput.

There has been limited prospective evaluation of state-of-the-art machine learning models to assess enrichment, as most recent work has focused on final viable candidates that require significant human input and review to evaluate machine learning predictions. In this study, we performed a thorough prospective benchmarking of state-of-the-art machine learning models for inhibition prediction, including in vitro validation. Using the spectral framework for model evaluation, we observed that model performance rapidly deteriorates as train-test similarity decreases. This trend was confirmed during in-vitro validation, where models predominantly identified hits that closely resembled training data while struggling with structurally distinct compounds. These findings

highlight the critical importance of training data size and diversity in achieving robust model performance.

We showed that Fleming's reasoning closely parallels the decision-making of medicinal chemists. Providing such expertise in an accessible format will substantially accelerate the antibiotic discovery process. Furthermore, Fleming's use of natural language interfaces improves access to advanced machine learning models. Given that TB disproportionately impacts low-resource regions, Fleming's accessibility offers a crucial advantage, enabling TB researchers without extensive resources to engage with powerful computational tools.

Our work has several limitations. We used GPT4 [32] for the agents and did not evaluate alternatives such as Claude Sonnet [45] or Llama 3 [2], which might enhance Fleming's accessibility. Additionally, we employed prompting to initialize agents and guide their use of bioinformatic tools. However, other methods, such as instruction tuning, could improve Fleming's ability to understand and apply these tools effectively [50]. We also trained our inhibition model without incorporating drug concentration data. As a result, Fleming does not account for the possibility that certain hits may require concentrations that are too high to be practically viable.

As Fleming continues to evolve with advances in machine learning and bioinformatics, it is poised to redefine the pace and accessibility of antibiotic discovery, bridging critical gaps in global health. The future of drug discovery will be shaped by systems like Fleming, empowering scientists worldwide to address the most urgent public health threats of our time with unprecedented speed and precision.

Data Availability The data used to develop the inhibition prediction agent and molecule generation agent were experimentally tested in-house and are available upon request. The data for training the ADMET agent are publicly accessible through the Therapeutic Data Commons (https://tdcommons.ai/) [13].

Code Availability The code to run Fleming agent, together with the code to reproduce the results are available on GitHub at https://github.com/farhat-lab/Fleming .

Figures

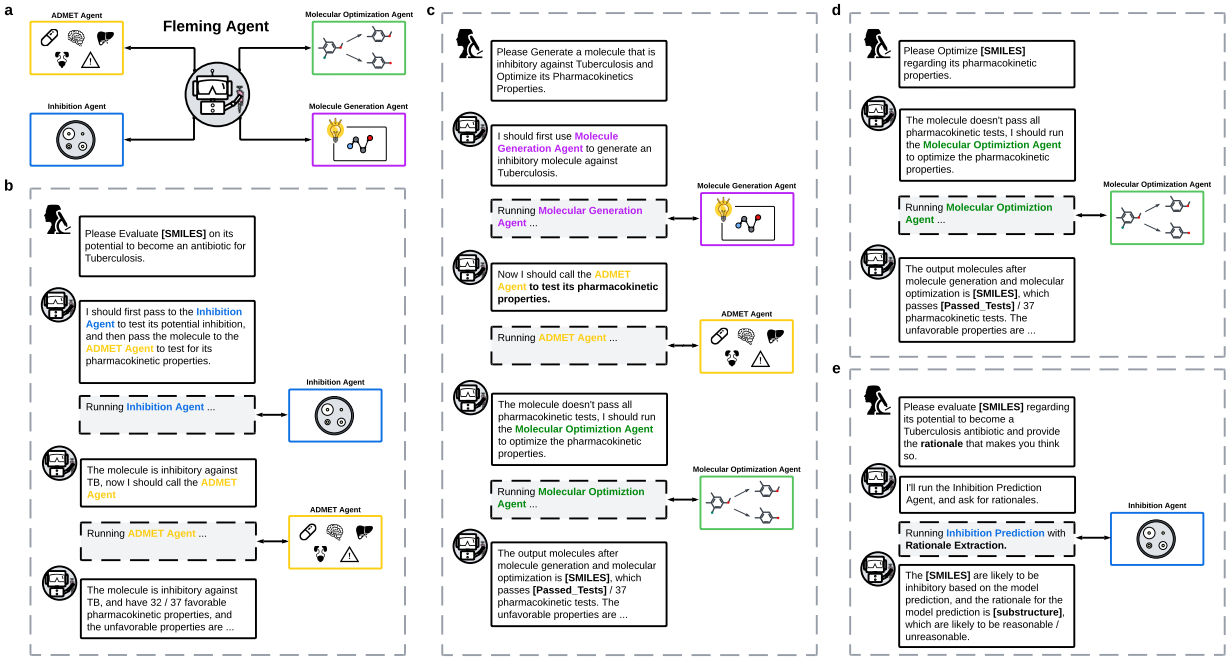

Figure 1: Fleming is an agent for antibiotic discovery in Tuberculosis (TB). (a) Fleming integrates four specialized agents—the ADMET agent, inhibition prediction agent, molecule generation agent, and molecular optimization agent—to perform tasks critical to TB antibiotic discovery. **(b)** In one application, a user queries Fleming with a SMILES structure to assess its potential as a TB antibiotic. Fleming deploys the inhibition prediction and ADMET agents to address the query. **(c)** In another application, a user prompts Fleming to generate a novel molecule with potential TB antibiotic properties. Fleming activates the molecule generation, ADMET, and molecular optimization agents to fulfill this request. **(d)** Fleming can also optimize user-provided molecules by engaging the molecular optimization agent. **(e)** Additionally, Fleming can identify and explain substructures within input molecules that contribute to the inhibition prediction.

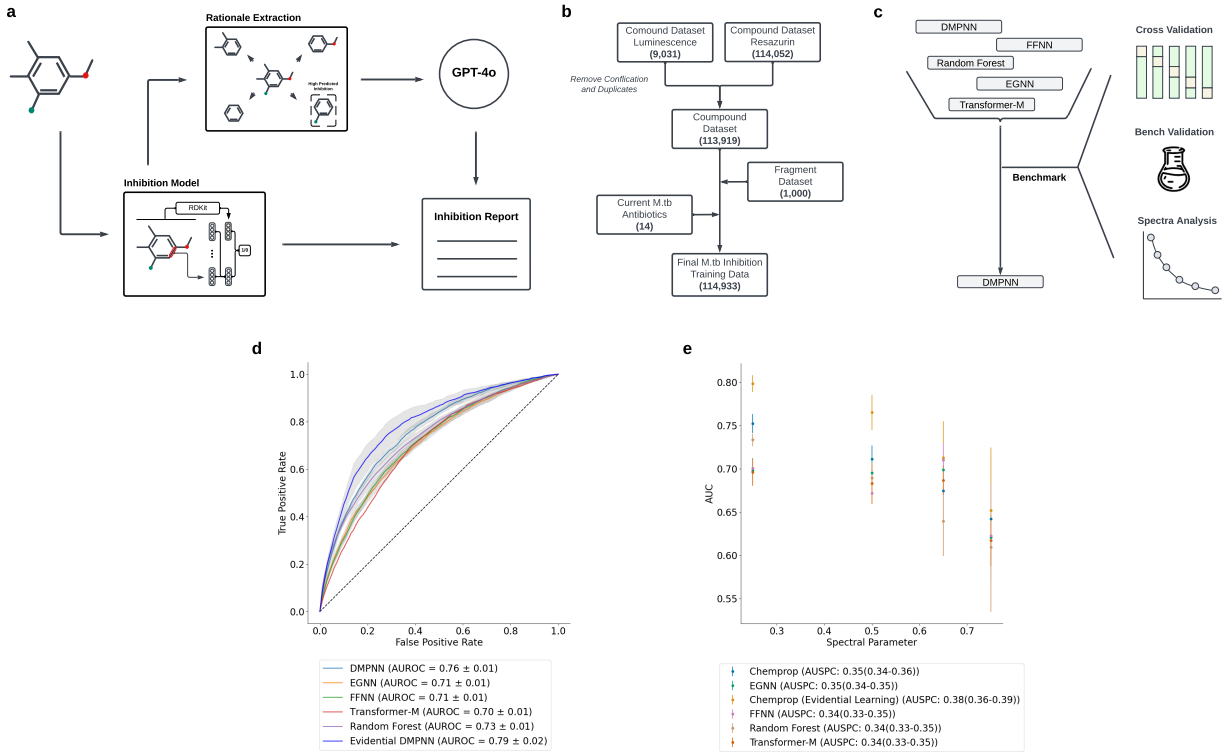

Figure 2: Inhibition agent identifies small molecules with inhibitory effects on TB. (a) Structure of the inhibition prediction agent, where input molecules are fed to an inhibition prediction model followed by optional rationale extraction which identifies the substructures relevant to inhibition prediction and passes them to PaperQA2 before returning an inhibition report that synthesizes information from literature from PaperQA2 and the inhibition model. **(b)** Training dataset for the inhibition prediction model, which includes two compound libraries with distinct assay readouts, a fragment dataset, and 14 antibiotics currently used to treat TB. **(c)** The inhibition prediction model was chosen among directed message-passing neural network (DMPNN), feed-forward neural network (FFNN), random forest, equivariant graph neural network (EGNN), and TransformerM. DMPNN trained with an evidential classification loss showed the best performance in cross-validation, bench validation, and SPECTRA analysis. **(d)** Area under receiver operating characteristic (AUROC) curve of benchmarked models on TB inhibition prediction. **(e)** Area under the spectral performance curve (AUSPC) of benchmarked models on TB inhibition prediction. AUSPC generated with spectral parameters of 0.25, 0.50, 0.65, and 0.75.

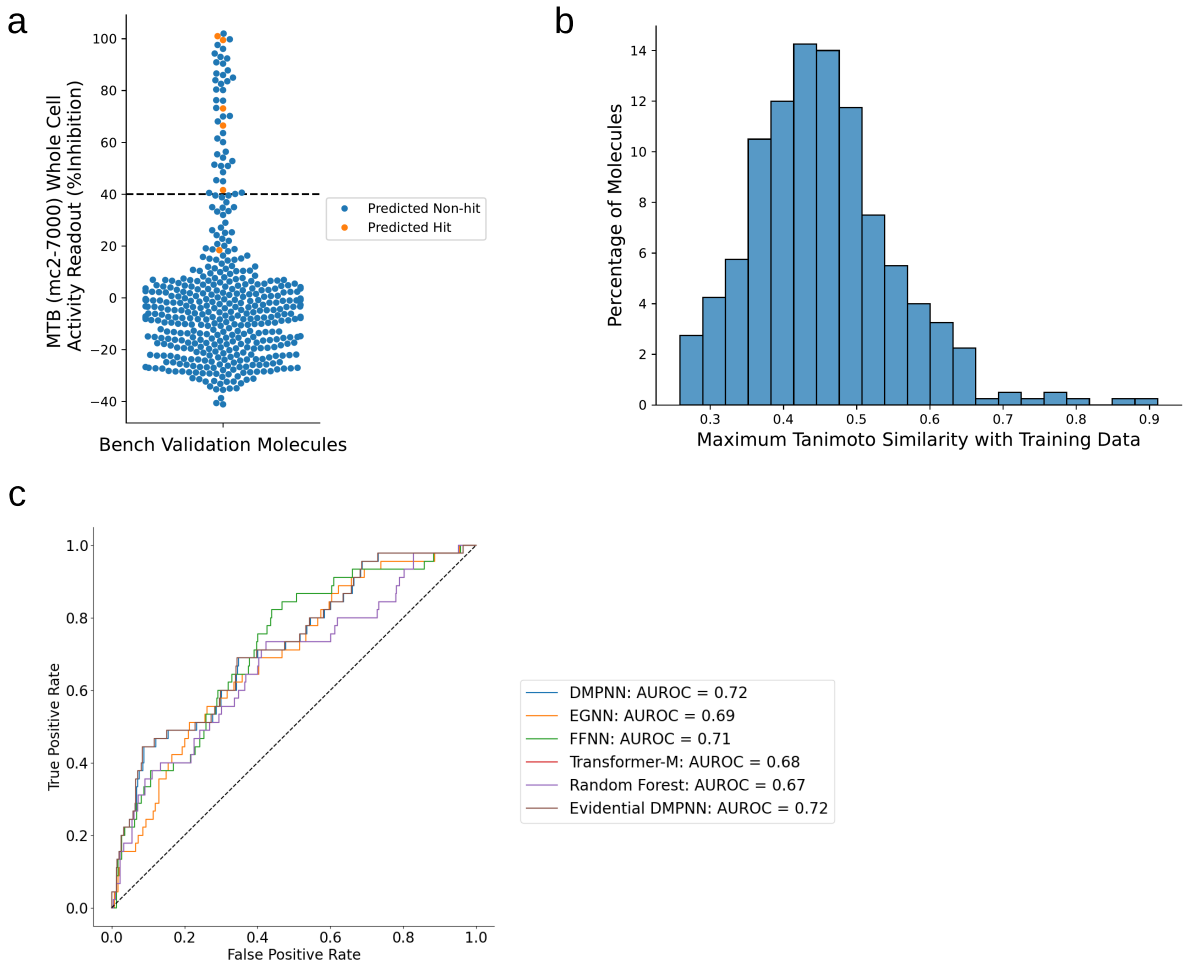

Figure 3: Bench validation results for the inhibition agent (a) %inhibition readouts of screened molecules on whole-cell activity against Mtb(mc2-7000). Blue dots are predicted non-hits, and orange dots are predicted hits. 40% is used as the threshold of inhibition binarization. **(b)** Maximum Tanimoto similarity distribution of the bench-validated molecules against the training data. **(c)** Area under receiver operating characteristic (AUROC) curve of the selected models on the bench validation set of 439 molecules.

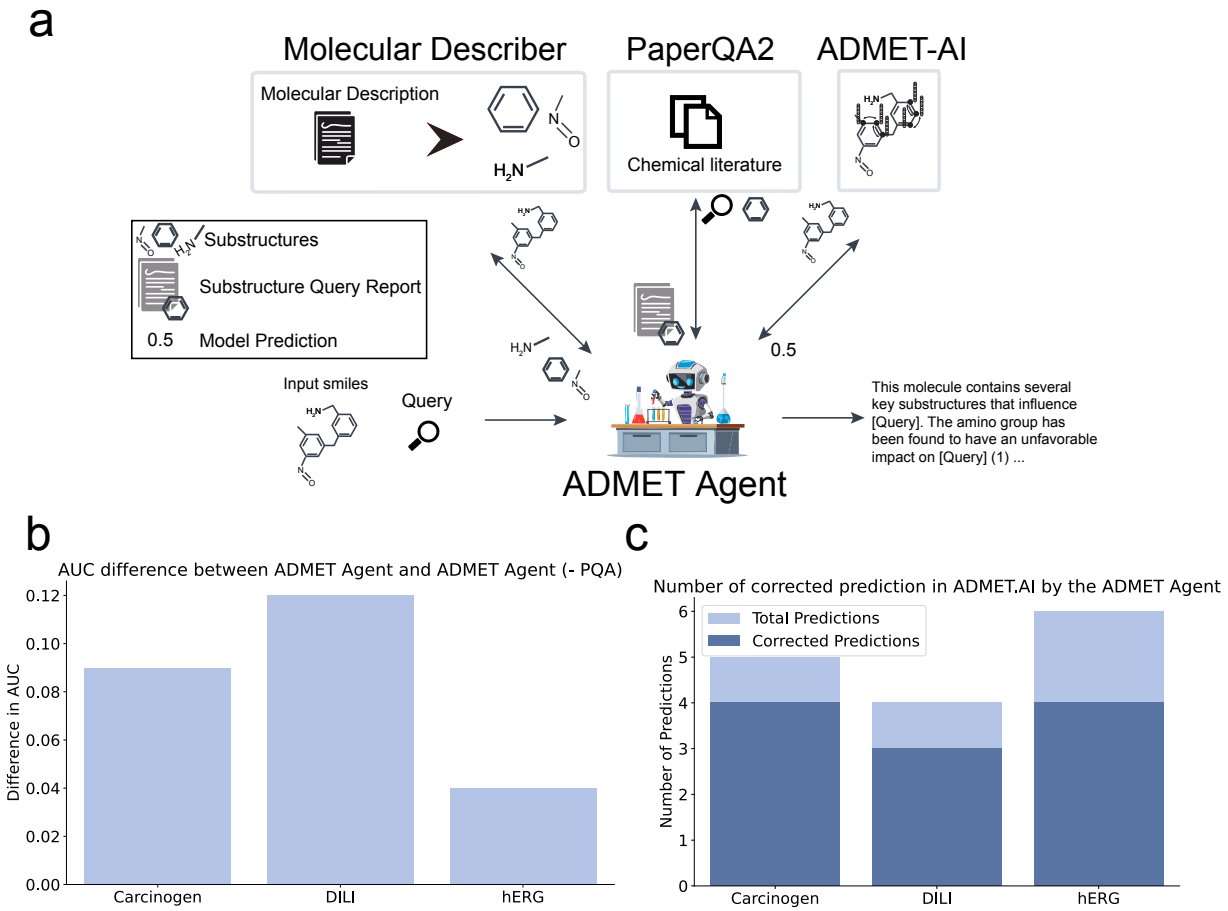

Figure 4: ADMET agent can identify molecules with favorable ADMET properties. (a) Overview of the ADMET agent's components and workflow. Given an input SMILES structure and a query, the ADMET agent integrates three tools: (1) a Molecular Describer that extracts molecular substructures and generates textual descriptions, (2) PaperQA2, which queries scientific literature based on the input query and identified substructures, and (3) ADMET-AI, a pretrained Chemprop-based model that predicts ADMET properties for the input molecule. Together, these tools provide an informed response to the query. **(b)** Improvement in AUC performance when using the ADMET agent with PaperQA2 (PQA) compared to without it across three tasks from the Therapeutic Data Commons. **(c)** Number of uncertain predictions (probability scores between 0.5 and 0.6) made by ADMET-AI that were corrected by the ADMET agent.

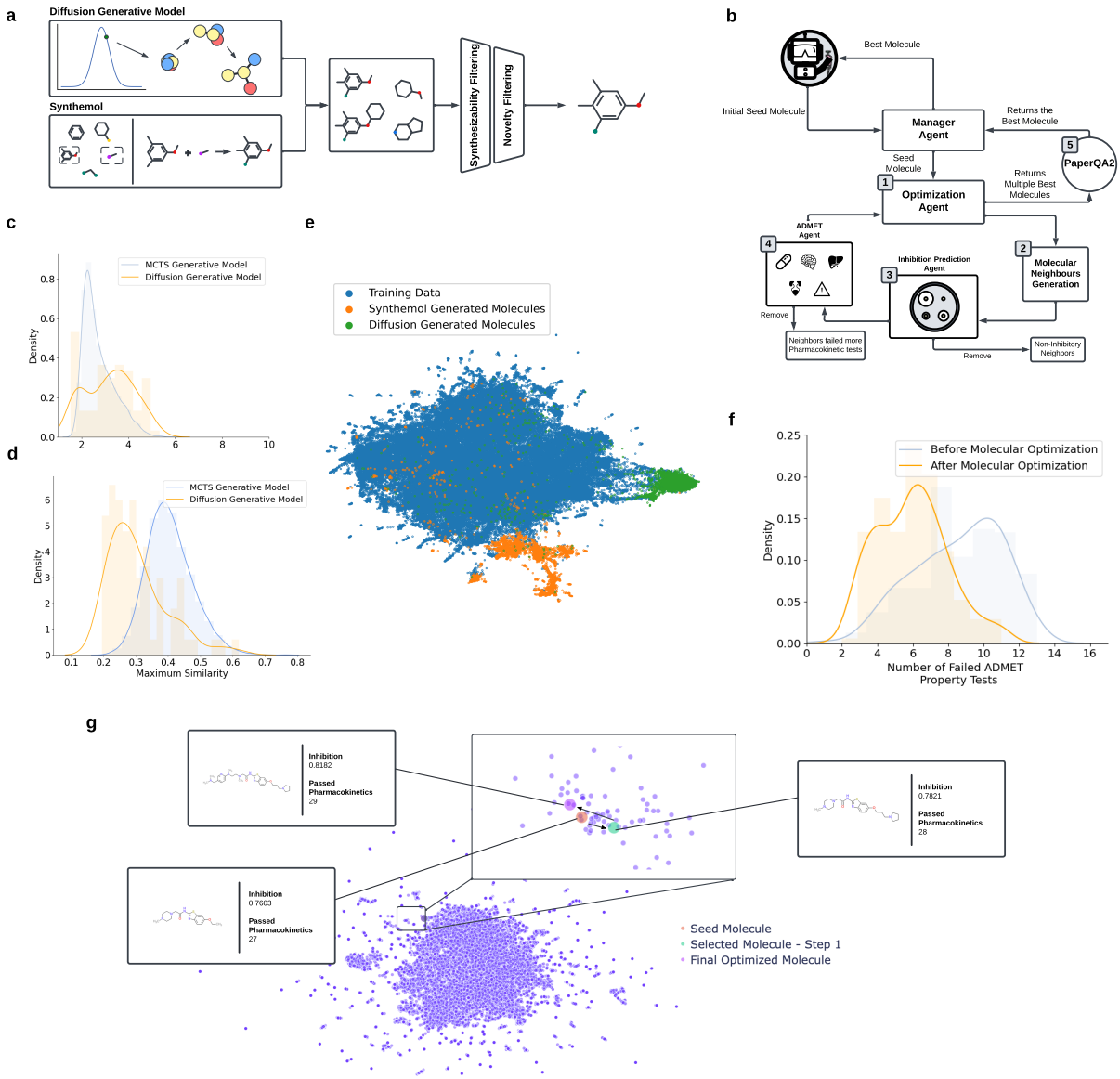

Figure 5: The molecular generation and optimization agents design novel TB inhibitors and enhance them for optimal ADMET properties. (a) The molecule generation agent generates novel TB inhibitors using a conditional equivariant diffusion model [12] or Synthemol [44]. Generated molecules undergo synthesizability and novelty filtering before output. **(b)** The molecular optimization agent inputs a molecule into a manager agent, which coordinates with an optimization agent to iteratively identify close chemical neighbors with improved ADMET properties while preserving similar inhibition. **(c)** Distribution of molecular synthesizability, measured by SAscore [4], for Synthemol and the diffusion model. A lower SAscore indicates easier synthesis, with molecules scoring below 6 considered synthesizable [4]. **(d)** Distribution of novelty, measured by tanimoto similarity to the training set, of generated molecules from Synthemol and diffusion model. The lower the similarity, the higher the novelty. **(e)** UMAP projection of morgan fingerprints (2048 bits) showing training data (blue), Synthemol-generated molecules (orange), and diffusion model-generated molecules (green). **(f)** Distribution of failed ADMET tasks for molecules before and after optimization by the molecular optimization agent. The less failed ADMET tasks, the better. **(g)** Illustration of molecular optimization represented in umap projection of morgan fingerprints. The seed molecule (orange) is optimized in two steps (light blue to light purple dot).

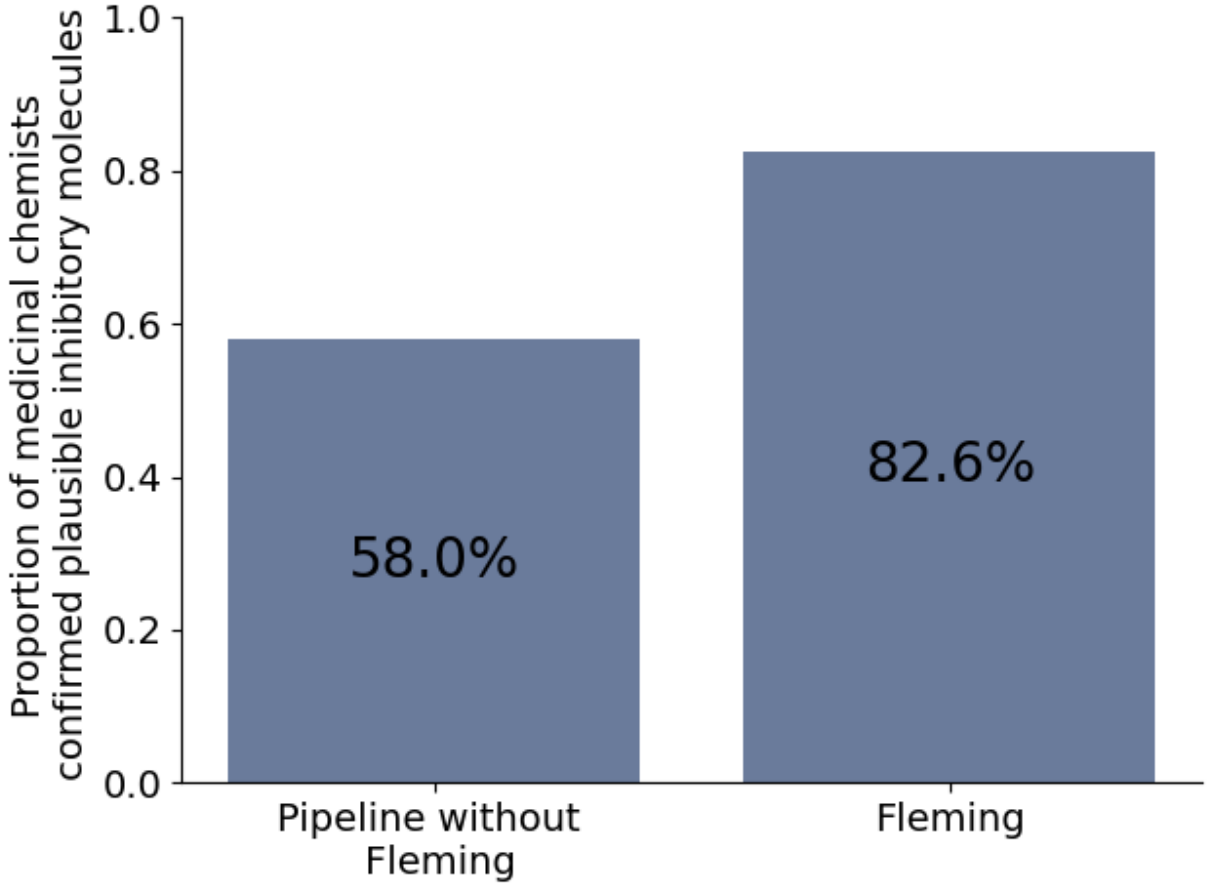

Figure 6: Fleming is able to capture medicinal chemist knowledge. Bar chart showing the proportion of 100 molecules correctly identified as potentially inhibitory, based on labels from medicinal chemists.

References

[1] Feixiong Cheng et al. "admetSAR: A Comprehensive Source and Free Tool for Assessment of Chemical ADMET Properties". In: *Journal of Chemical Information and Modeling* 52.11 (2012). _eprint: https://doi.org/10.1021/ci300367a, pp. 3099–3105. DOI: 10.1021/ci300367a. URL: https://doi.org/10.1021/ci300367a.

[2] Abhimanyu Dubey et al. *The Llama 3 Herd of Models*. Aug. 15, 2024. arXiv: 2407.21783[cs]. URL: http://arxiv.org/abs/2407.21783 (visited on 11/12/2024).

[3] Yasha Ektefaie et al. *Evaluating generalizability of artificial intelligence models for molecular datasets*. Feb. 28, 2024. DOI: 10.1101/2024.02.25.581982. URL: http://biorxiv.org/lookup/doi/10.1101/2024.02.25.581982 (visited on 07/24/2024).

[4] Peter Ertl and Ansgar Schuffenhauer. "Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions". In: *Journal of Cheminformatics* 1.1 (Dec. 2009), p. 8. ISSN: 1758-2946. DOI: 10.1186/1758-2946-1-8. URL: https://jcheminf.biomedcentral.com/articles/10.1186/1758-2946-1-8 (visited on 07/17/2024).

[5] Maha Farhat et al. "Drug-resistant tuberculosis: a persistent global health concern". In: *Nature Reviews Microbiology* 22.10 (Oct. 2024), pp. 617–635. ISSN: 1740-1526, 1740-1534. DOI: 10.1038/s41579-024-01025-1. URL: https://www.nature.com/articles/s41579-024-01025-1 (visited on 11/26/2024).

[6] Peter Gedeck, Bernhard Rohde, and Christian Bartels. "QSAR How Good Is It in Practice? Comparison of Descriptor Sets on an Unbiased Cross Section of Corporate Data Sets". In: *Journal of Chemical Information and Modeling* 46.5 (Sept. 1, 2006), pp. 1924–1936. ISSN: 1549-9596, 1549-960X. DOI: 10.1021/ci050413p. URL: https://pubs.acs.org/doi/10.1021/ci050413p (visited on 07/02/2024).

[7] Justin Gilmer et al. *Neural Message Passing for Quantum Chemistry*. June 12, 2017. arXiv: 1704.01212[cs]. URL: http://arxiv.org/abs/1704.01212 (visited on 07/02/2024).

[8] *Global Tuberculosis Report 2022*. URL: https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2022 (visited on 07/03/2024).

[9] Markus Hartenfeller et al. "A Collection of Robust Organic Synthesis Reactions for *In Silico* Molecule Design". In: *Journal of Chemical Information and Modeling* 51.12 (Dec. 27, 2011), pp. 3093–3098. ISSN: 1549-9596, 1549-960X. DOI: 10.1021/ci200379p. URL: https://pubs.acs.org/doi/10.1021/ci200379p (visited on 06/11/2024).

[10] Esther Heid et al. "Chemprop: A Machine Learning Package for Chemical Property Prediction". In: *Journal of Chemical Information and Modeling* 64.1 (2024). PMID: 38147829, pp. 9–17. DOI: 10.1021/acs.jcim.3c01250. eprint: https://doi.org/10.1021/acs.jcim.3c01250. URL: https://doi.org/10.1021/acs.jcim.3c01250.

[11] Tin Kam Ho. "Random decision forests". In: *Proceedings of 3rd international conference on document analysis and recognition*. Vol. 1. IEEE. 1995, pp. 278–282.

[12] Emiel Hoogeboom et al. *Equivariant Diffusion for Molecule Generation in 3D*. June 16, 2022. arXiv: 2203.17003[cs, q-bio, stat]. URL: http://arxiv.org/abs/2203.17003 (visited on 07/04/2024).

[13] Kexin Huang et al. *Therapeutics Data Commons: Machine Learning Datasets and Tasks for Drug Discovery and Development*. Aug. 28, 2021. arXiv: 2102.09548[cs, q-bio]. URL: http://arxiv.org/abs/2102.09548 (visited on 06/10/2024).

[14] J. D. Hunter. "Matplotlib: A 2D graphics environment". In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55.

[15] Md Shamim Hussain, Mohammed J. Zaki, and Dharmashankar Subramanian. *Triplet Interaction Improves Graph Transformers: Accurate Molecular Graph Learning with Triplet Graph Transformers*. June 9, 2024. arXiv: 2402.04538[cs]. URL: http://arxiv.org/abs/2402.04538 (visited on 07/02/2024).

[16] Plotly Technologies Inc. *Collaborative data science*. 2015. URL: https://plot.ly.

[17] National Cancer Institute. *NCI Chemical Identifier Resolver*. https://cactus.nci.nih.gov/chemical/structure. Accessed: 2024-11-05.

[18] Wengong Jin, Dr.Regina Barzilay, and Tommi Jaakkola. "Multi-Objective Molecule Generation using Interpretable Substructures". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 13–18 Jul 2020, pp. 4849–4859. URL: https://proceedings.mlr.press/v119/jin20b.html.

[19] Sunghwan Kim et al. "PubChem 2023 update". In: *Nucleic Acids Research* 51.D1 (Oct. 2022), pp. D1373–D1380. ISSN: 0305-1048. DOI: 10.1093/nar/gkac956. eprint: https://academic.oup.com/nar/article-pdf/51/D1/D1373/48441598/gkac956.pdf. URL: https://doi.org/10.1093/nar/gkac956.

[20] Thomas N. Kipf and Max Welling. *Semi-Supervised Classification with Graph Convolutional Networks*. Feb. 22, 2017. arXiv: 1609.02907[cs, stat]. URL: http://arxiv.org/abs/1609.02907 (visited on 07/02/2024).

[21] Alexey Lagunin et al. "Computer-Aided Prediction of Rodent Carcinogenicity by PASS and CISOC-PSCT". In: *QSAR & Combinatorial Science* 28.8 (Aug. 2009), pp. 806–810. ISSN: 1611-020X, 1611-0218. DOI: 10.1002/qsar.200860192. URL: https://onlinelibrary.wiley.com/doi/10.1002/qsar.200860192 (visited on 01/03/2025).

[22] Yi-Lun Liao and Tess Smidt. *Equiformer: Equivariant Graph Attention Transformer for 3D Atomistic Graphs*. Feb. 27, 2023. arXiv: 2206.11990[physics]. URL: http://arxiv.org/abs/2206.11990 (visited on 07/02/2024).

[23] Yi-Lun Liao et al. *EquiformerV2: Improved Equivariant Transformer for Scaling to Higher-Degree Representations*. Mar. 6, 2024. arXiv: 2306.12059[physics]. URL: http://arxiv.org/abs/2306.12059 (visited on 07/02/2024).

[24] Chengyou Liu et al. "ABT-MPNN: an atom-bond transformer-based message-passing neural network for molecular property prediction". In: *Journal of Cheminformatics* 15.1 (Feb. 26, 2023), p. 29. ISSN: 1758-2946. DOI: 10.1186/s13321-023-00698-9. URL: https://jcheminf.biomedcentral.com/articles/10.1186/s13321-023-00698-9 (visited on 06/11/2024).

[25] Zhiwei Liu et al. *AgentLite: A Lightweight Library for Building and Advancing Task-Oriented LLM Agent System*. Feb. 23, 2024. arXiv: 2402.15538[cs]. URL: http://arxiv.org/abs/2402.15538 (visited on 11/11/2024).

[26] Shengjie Luo et al. *One Transformer Can Understand Both 2D & 3D Molecular Data*. Mar. 27, 2023. arXiv: 2210.01765[cs, q-bio, stat]. URL: http://arxiv.org/abs/2210.01765 (visited on 04/24/2024).

[27] Andres M. Bran et al. "Augmenting large language models with chemistry tools". In: *Nature Machine Intelligence* 6.5 (May 8, 2024), pp. 525–535. ISSN: 2522-5839. DOI: 10.1038/s42256-024-00832-8. URL: https://www.nature.com/articles/s42256-024-00832-8 (visited on 06/19/2024).

[28] Dominic Masters et al. *GPS++: An Optimised Hybrid MPNN/Transformer for Molecular Property Prediction*. Dec. 6, 2022. arXiv: 2212.02229[cs, q-bio]. URL: http://arxiv.org/abs/2212.02229 (visited on 07/02/2024).

[29] Leland McInnes, John Healy, and James Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. Sept. 17, 2020. arXiv: 1802.03426[cs, stat]. URL: http://arxiv.org/abs/1802.03426 (visited on 07/24/2024).

[30] H. L. Morgan. "The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service." In: *Journal of Chemical Documentation* 5.2 (May 1, 1965), pp. 107–113. ISSN: 0021-9576, 1541-5732. DOI: 10.1021/c160017a018. URL: https://pubs.acs.org/doi/abs/10.1021/c160017a018 (visited on 04/24/2024).

[31] Dereje A. Negatu et al. "Whole-Cell Screen of Fragment Library Identifies Gut Microbiota Metabolite Indole Propionic Acid as Antitubercular". In: *Antimicrobial Agents and Chemotherapy* 62.3 (Mar. 2018), e01571–17. ISSN: 0066-4804, 1098-6596. DOI: 10.1128/AAC.01571-17. URL: https://journals.asm.org/doi/10.1128/AAC.01571-17 (visited on 06/11/2024).

[32] OpenAI et al. *GPT-4 Technical Report*. Mar. 4, 2024. arXiv: 2303.08774[cs]. URL: http://arxiv.org/abs/2303.08774 (visited on 08/23/2024).

[33] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[34] A. S. M. Zisanur Rahman et al. "A machine learning model trained on a high-throughput antibacterial screen increases the hit rate of drug discovery". In: *PLOS Computational Biology* 18.10 (Oct. 13, 2022). Ed. by Tobias Bollenbach, e1010613. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1010613. URL: https://dx.plos.org/10.1371/journal.pcbi.1010613 (visited on 07/02/2024).

[35] K. Rajan, A. Zielesny, and C. Steinbeck. "STOUT V2.0: SMILES to IUPAC name conversion using transformer models". In: *ChemRxiv* (2024). This content is a preprint and has not been peer-reviewed. DOI: 10.26434/chemrxiv-2024-089vs.

[36] *RDKit: Open-source cheminformatics*. http://www.rdkit.org. [Online; accessed 11-April-2013].

[37] Allison W Roberts et al. "Cas9\textsuperscript+ conditionally-immortalized macrophages as a tool for bacterial pathogenesis and beyond". In: *eLife* 8 (June 2019). Ed. by Wendy S Garrett and Bavesh D Kana. Publisher: eLife Sciences Publications, Ltd, e45957. ISSN: 2050-084X. DOI: 10.7554/eLife.45957. URL: https://doi.org/10.7554/eLife.45957.

[38] David Rogers and Mathew Hahn. "Extended-Connectivity Fingerprints". In: *Journal of Chemical Information and Modeling* 50.5 (May 24, 2010), pp. 742–754. ISSN: 1549-9596, 1549-960X. DOI: 10.1021/ci100050t. URL: https://pubs.acs.org/doi/10.1021/ci100050t (visited on 07/02/2024).

[39] Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. *E(n) Equivariant Graph Neural Networks*. Feb. 16, 2022. arXiv: 2102.09844[cs, stat]. URL: http://arxiv.org/abs/2102.09844 (visited on 04/24/2024).

[40] Murat Sensoy, Lance Kaplan, and Melih Kandemir. "Evidential Deep Learning to Quantify Classification Uncertainty". In: ().

[41] Michael D. Skarlinski et al. *Language agents achieve superhuman synthesis of scientific knowledge*. 2024. arXiv: 2409.13740 [cs.CL]. URL: https://arxiv.org/abs/2409.13740.

[42] Jonathan M. Stokes et al. "A Deep Learning Approach to Antibiotic Discovery". In: *Cell* 180.4 (Feb. 2020), 688–702.e13. ISSN: 00928674. DOI: 10.1016/j.cell.2020.01.021. URL: https://linkinghub.elsevier.com/retrieve/pii/S0092867420301021.

[43] Kyle Swanson et al. *ADMET-AI: A machine learning ADMET platform for evaluation of large-scale chemical libraries*. Dec. 28, 2023. DOI: 10.1101/2023.12.28.573531. URL: http://biorxiv.org/lookup/doi/10.1101/2023.12.28.573531 (visited on 06/10/2024).

[44] Kyle Swanson et al. "Generative AI for designing and validating easily synthesizable and structurally novel antibiotics". In: *Nature Machine Intelligence* 6.3 (Mar. 2024). Publisher: Nature Publishing Group, pp. 338–353. ISSN: 2522-5839. DOI: 10.1038/s42256-024-00809-7. URL: https://www.nature.com/articles/s42256-024-00809-7 (visited on 03/29/2024).

[45] "The Claude 3 Model Family: Opus, Sonnet, Haiku". In: URL: https://api.semanticscholar.org/CorpusID:268232499.

[46] Robert Tinn et al. "Fine-tuning large neural language models for biomedical natural language processing". In: *Patterns* 4.4 (Apr. 2023), p. 100729. ISSN: 26663899. DOI: 10.1016/j.patter.2023.100729. URL: https://linkinghub.elsevier.com/retrieve/pii/S2666389923000697 (visited on 07/03/2024).

[47] Gemma Turon et al. "First fully-automated AI/ML virtual screening cascade implemented at a drug discovery centre in Africa". In: ().

[48] Shuangquan Wang et al. "ADMET Evaluation in Drug Discovery. 16. Predicting hERG Blockers by Combining Multiple Pharmacophores and Machine Learning Approaches". In: *Molecular Pharmaceutics* 13.8 (2016). _eprint: https://doi.org/10.1021/acs.molpharmaceut.6b00471, pp. 2855–2866. DOI: 10.1021/acs.molpharmaceut.6b00471. URL: https://doi.org/10.1021/acs.molpharmaceut.6b00471.

[49] Jason Wei et al. "Chain of Thought Prompting Elicits Reasoning in Large Language Models". In: *CoRR* abs/2201.11903 (2022). arXiv: 2201.11903. URL: https://arxiv.org/abs/2201.11903.

[50] Jason Wei et al. *Finetuned Language Models Are Zero-Shot Learners*. Feb. 8, 2022. arXiv: 2109.01652[cs]. URL: http://arxiv.org/abs/2109.01652 (visited on 11/12/2024).

[51] Andrew White. *whitead/synspace*. original-date: 2022-12-28T00:59:14Z. Apr. 10, 2024. URL: https://github.com/whitead/synspace (visited on 06/11/2024).

[52] Felix Wong et al. "Discovery of a structural class of antibiotics with explainable deep learning". In: *Nature* 626.7997 (Feb. 1, 2024), pp. 177–185. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-023-06887-8. URL: https://www.nature.com/articles/s41586-023-06887-8 (visited on 04/24/2024).

[53] World Health Organization. *Global Tuberculosis Report 2024*. Licence: CC BY-NC-SA 3.0 IGO. Geneva, 2024. URL: https://www.who.int/publications/i/item/9789240101531.

[54] Tong Xie et al. *DARWIN Series: Domain Specific Large Language Models for Natural Science*. Aug. 24, 2023. arXiv: 2308.13565[cond-mat,physics:physics]. URL: http://arxiv.org/abs/2308.13565 (visited on 07/03/2024).

[55] Keyulu Xu et al. *How Powerful are Graph Neural Networks?* Feb. 22, 2019. arXiv: 1810.00826[cs,stat]. URL: http://arxiv.org/abs/1810.00826 (visited on 07/02/2024).

[56] Youjun Xu et al. "Deep Learning for Drug-Induced Liver Injury". In: *Journal of Chemical Information and Modeling* 55.10 (Oct. 26, 2015), pp. 2085–2093. ISSN: 1549-9596, 1549-960X. DOI: 10.1021/acs.jcim.5b00238. URL: https://pubs.acs.org/doi/10.1021/acs.jcim.5b00238 (visited on 01/03/2025).

[57] Kevin Yang et al. "Analyzing Learned Molecular Representations for Property Prediction". In: *Journal of Chemical Information and Modeling* 59.8 (Aug. 26, 2019), pp. 3370–3388. ISSN: 1549-9596, 1549-960X. DOI: 10.1021/acs.jcim.9b00237. URL: https://pubs.acs.org/doi/10.1021/acs.jcim.9b00237 (visited on 09/13/2022).

[58] Kevin Yang et al. "Are Learned Molecular Representations Ready For Prime Time?" In: ().

[59] Chengxuan Ying et al. *Do Transformers Really Perform Bad for Graph Representation?* Nov. 23, 2021. arXiv: 2106.05234[cs]. URL: http://arxiv.org/abs/2106.05234 (visited on 07/02/2024).

[60] Sheheryar Zaidi et al. *Pre-training via Denoising for Molecular Property Prediction*. Oct. 24, 2022. arXiv: 2206.00133[cs,q-bio,stat]. URL: http://arxiv.org/abs/2206.00133 (visited on 06/26/2024).

Methods

This section describes the details of how we constructed the (1) inhibition agent, (2) ADMET agent, (3) molecular optimization agent, and (4) molecular generation agent. We also detail how we benchmark these agents.

Inhibition Agent

The inhibition agent consists of an inhibition prediction model and a rationale extraction module. To determine whether a small molecule is inhibitory, the molecule is first processed by the inhibition prediction model, which provides an initial inhibitory prediction. The rationale extraction module then identifies specific substructures utilized by the model in making this prediction. These substructures are subsequently analyzed using GPT4 [32] to evaluate their potential involvement in inhibition. The insights from both the model's prediction and the GPT4's analysis are then integrated to produce a final inhibitory assessment.

MTB Growth Inhibition Data

The data we used to train the MTB inhibition prediction model and the conditional diffusion generative model consists of 113,919 chemically diverse compounds (Tanimoto similarity ¡ 0.7), 1,000 fragments [31], and 14 current in-use TB antibiotics. The compound readouts are from three high-throughput screen runs. In all cases, mc2-7000 strain of MTB (PMID: 12219086) starter culture was grown in 7H9 media supplemented with OADC, 0.04% Tyloxapol and 25 μg/ml of pantothenate to the OD600 of 1. For the screening, starter culture cells were diluted into testing media to the OD600 of 0.01 and dispensed into 384 well plate. The first set of 50,000 molecules was tested at a final concentration of 10 uM (compounds were added from 100% DMSO stock, to final 1% DMSO) using M9 minimal media supplemented with 0.5% dextrose, 0.2% sodium acetate (stock solution was pH adjusted to neutral prior to use), 0.01% NaCl, 0.04% Tyloxapol and 25 μg/ml of pantothenic acid. After 5 days of incubation at 37C, resazurin was added to a final concentration of 0.005 mg/ml to all wells to provide viability readout. Plates were incubated for additional 1-2 days at 37C, until negative control wells, containing 1% DMSO, were uniformly converted to resorufin pink. Resorufin fluorescence was recorded by OMEGA Polar Star plate reader with excitation wavelength of 540 nm, and emission at 590 nm. Percent inhibition for each compound was calculated using positive (1μM rifampicin) and negative (1% DMSO) controls on each corresponding plate. The second set was tested at a final concentration of 20 μM (2% DMSO final) in the 7H9 media supplemented with 0.5% dextrose, 0.1% sodium acetate pH 7.0, 0.05% tyloxapol,

25 μg/mL pantothenic acid, and 25 μg/mL kanamycin, using both resazurin and luminescence readouts. For the later, MTB mc2-7000 strain was transformed with luxCDABE operon containing plasmid enabling constituent expression of both light producing enzyme and the substrate for it (courteously shared with us by Cox lab) [37]. Resazruin conversion to resofurin was recorded as described above, luminescence was measured by OMEGA Clariostar plate reader at 37C after 5 days of incubation with the compound. Percent inhibition was calculated by normalizing between positive and negative controls included on each plate, as described above. Finally, a set of 100,000 molecules, partially overlapping with the first set, was tested under the same growth conditions as set two, using luciferase expressing mc2-7000 strain, and recording luminescence as a viability readout.

The 1000 compound fragment collection (Maybridge Ro3 library, MW ¡ 300 g/mol) was tested against the MTB H37Rv strain in complete Middlebrook 7H9 broth, with optical density (OD) at 600 nm as readout as described [31]. Due to the anticipated weak inhibitory effects of fragments, 1 mM was used in a single-point screen [31].

We binarized inhibition based on an internally calibrated threshold based on the distribution of inhibition across the assayed molecules. Molecules with normalized inhibition based on the corresponding plate controls at or more than 40% are labeled as inhibitors, while other molecules are labeled as non-inhibitory. For replicates across different batches, molecules with any replicate that exceed 40% inhibition are labeled as inhibitory. In the fragment dataset we used an inhibition threshold of 70%. All current TB antibiotics are considered as inhibitory when adding into the training data.

MTB Inhibition Model

We compared the performance of machine learning models that have been recently applied for in-silico antibiotic discovery [42, 52] or other state-of-the-art molecular property prediction[39, 26]. Specifically we tested, Directed Message Passing Neural Network (DMPNN) [57], Equivariant Graph Neural Network (EGNN) [39], Transformer-M [26], Feed Forward Neural Network, and Random Forests [11].

We evaluated model accuracy using cross-validation, and SPECTRA [3]. We used 5-fold cross-validation with an 80%/10%/10% scaffold-based split for training, validation, and test sets, grouping molecules by scaffold created by the chemprop package [10] to prevent data leakage and ensure robust evaluation. For SPECTRA, Tanimoto similarity was used as the metric to evaluate similarity between the train and test set. Spectral parameters of 0.25, 0.5, 0.65, and 0.75 were used

to evaluate the generalizability of the models on splits with decreasing levels of train-test similarity.

In-vitro Bench Validation

To further validate our inhibition prediction model, we screened an out-of-sample chemical library containing potential binders of the PKS13-TE protein in MTB and tested these molecules in-vitro. For consistency, we used the same experimental conditions as in the second batch of training data. The mc2-7000 strain of MTB (PMID: 12219086) starter culture was grown in 7H9 media supplemented with OADC, 0.04% Tyloxapol and 25 μg/ml of pantothenate to the OD600 of 1. Starter culture cells were diluted into testing media to the OD600 of 0.01 and dispensed into 384 well plates. Candidates are tested at a final concentration of 100 μM in the 7H9 media supplemented with 0.5% dextrose, 0.1% sodium acetate pH 7.0, 0.05% tyloxapol, 25 μg/mL pantothenic acid, and 25 μg/mL kanamycin, using resazurin readouts.

Directed Message Passing Neural Network

For Directed Message Passing Neural Network (DMPNN), the model constructs graph objects based on molecules' Simplified Molecular Input Line Entry System (SMILES) structure, and generates embedding of each molecule using message passing. The molecular embeddings are then concatenated with additional molecular features generated by the RDKit package [36] to predict the MTB inhibition of the molecules using a multi-layer perceptron. We trained both our sample classification model and evidential classification model using the chemprop package [10] with a batch size of 64, depth of 3, and dropout of 0.2. An exponential decay learning rate scheduler is used with two warm-up epochs, the initial learning rate of 0.0001, the max learning rate of 0.001, and a final learning rate of 0.0001.

To identify molecules with the highest likelihood of MTB inhibition while accounting for noisy training data, we trained DMPNN with evidential classification [40]. This approach enables the model to incorporate uncertainty in its predictions rather than categorizing each molecule strictly as inhibitory or non-inhibitory. Using prediction confidence from this model, we filtered out molecules with low confidence (confidence < 0.6) to enhance the inhibition model's overall performance.

Equivariant Graph Neural Network

The Equivariant Graph Neural Network (EGNN) [39] is a model architecture equivariant to translation, rotation, reflection, and permutation, leveraging the 3D positional data of each atom in the molecule as input. We used a batch size of 64, hidden layer dimension of 64, 3 layers, gradient

clipping at 0.2, and no dropout. Training employed an exponential decay learning rate scheduler with 4000 warm-up steps. We used grid-search to determine optimal hyperparameters.

Transformer-M

Transformer-M [26] is a state-of-the-art model that uses both 2D and 3D information of a molecule as model input. It incorporates a self-supervised 3D Positional Denoising objective [60] in the 3D channel as a co-objective to enhance model performance. We used a batch size of 256, noise scale of 0.2, channel weight of [1,0.01]. We also used an exponential decay learning rate scheduler with initial learning rate of 1e-5. We used grid-search to determine optimal hyperparameters.

Feed Forward Neural Network

We used a Feedforward Neural Network (FFNN) with RDKit Fingerprints to test the effectiveness of RDKit fingerprints alone in predicting the MTB inhibition. We used a 10-layer FFNN with a batch size of 64, a hidden layer dimension of 256, a gradient clip of 0.2, and no dropout. An exponential decay learning rate scheduler is used with 4000 warm-up steps. We used grid-search to determine optimal hyperparameters.

Random Forest

We used a random forest model from sklearn v1.5 [33] with 100 decision trees. The input to the random forest model is 200 RDKit features, which is the same as the input of the FFNN.

Rationale Module

For every molecule, the Rationale Module finds the molecular substructure that drives the model's prediction, and assesses whether this rationale is likely to be the core inhibitory element of an antibiotic. To do this, we employ a Monte Carlo Tree Search to identify subgraphs of the molecules that are between S_{min} and S_{max} atoms with predicted inhibition above δ [18]. This method has been used previously to identify rationales of antibiotic inhibition[52] and is implemented in Chemprop [10]. We used parameters: $\delta = 0.5, S_{min} = 8, S_{max} = 22, C_{PUCT} = 10$, rollouts=20. We use the rationale with the highest inhibition scores as the final rationale for the molecule. After obtaining natural language descriptions of the rationale and the molecule from the molecular describer tool, we used GPT-4o to assess whether it is mechanistically plausible that the molecule with this rationale is the inhibitory core of an antibiotic. Plausibility refers to the likelihood that the rationale presented is responsible for selective inhibition of an MTB biochemical process according to the existing chemical literature.

ADMET Agent

The ADMET agent comprises a set of ADMET models and an instance of PaperQA2 to analyze literature on molecule ADMET properties. First, the ADMET models evaluate the input molecule, followed by an analysis through PaperQA2, which assesses ADMET properties based on existing literature. The agent then integrates these findings to determine whether the molecule has favorable or unfavorable ADMET characteristics.

ADMET Models

We utilized 37 ADMET models that were trained on datasets from TDC Data Commons [13]. An ensemble of 5 DMPNN models was trained on each ADMET datasets with checkpoints from ADMET.AI [43]. The predicted properties are then binarized as pass/fail by comparing them to the predictions for currently used TB antibiotics generated by the same model. Screening molecules with predicted property values exceeding the top 90% threshold of values observed for all antibiotics are classified as pass; otherwise, they are classified as fail. Detailed pharmacokinetic properties and corresponding thresholds are provided in the supplement. Molecules are ranked based on the number of ADMET tasks they pass in each category.

PaperQA2 ADMET Analysis

PaperQA2 [41] was used to extract additional information from the existing literature on the properties of ADMET of potential tuberculosis (TB) inhibitors. Each molecule was processed using a molecular descriptor tool to identify and describe its substructures and functional groups. These descriptors were then distilled into core substructures. Each core substructure was individually input into PaperQA2, which retrieved relevant literature describing the effect of the substructure on ADMET properties. All substructure literature results, along with the molecule's overall descriptor, were then provided to GPT-4o, which was prompted to assess the molecule's ADMET properties as favorable or unfavorable based on literature evidence. This evaluation was subsequently integrated with the outputs from ADMET models to generate a comprehensive profile of each molecule's ADMET characteristics.

Molecular generation agent

The molecular generation agent employs two generative models: Synthemol [44] and an equivariant conditional diffusion model [12]. Molecules generated by either model are subsequently screened for novelty and synthesizability.

Conditional diffusion model

We trained a conditional diffusion model [12] on the MTB inhibition dataset using 4 H100 GPUs over a span of 3 days with recommended parameters. The model was conditioned on the level of inhibition and trained to generate small molecules present in the dataset. To generate a potentially inhibitory molecule, we first sample a molecule from the diffusion model and then evaluate it using the inhibition model to predict its inhibitory activity. If the sampled molecule is not predicted to be inhibitory, we resample iteratively until a molecule with predicted inhibitory activity is obtained.

SyntheMol

We used SyntheMol [44], which is a Monte-Carlo Tree Search algorithm (MCTS) based package to generate synthesizable molecules with MTB inhibition. Within Synthemol, scores related to the selected property are predicted by our inhibition model and then given to 132,000 building blocks. Molecules are then generated from the building blocks through 13 valid chemical reactions based on inhibition scores. The model then iteratively builds molecules that are likely MTB inhibitors.

Synthesizability and novelty filter

Molecules generated from the generative models were filtered by a synthesizability filter and a novelty filter to ensure their synthesizability and novelty. For the synthesizability filter, we use the SA score calculator from RDKit [36] and filter out molecules with an SA score larger than 4. The novelty filter calculates the Tanimoto Similarities of the Morgan Fingerprints between the output molecules and the inhibition model training data. Molecules with similarity larger than 0.3 are removed. A two-tailed t-test is used to determine the statistical significance of novelty and synthesizability for molecules generated from Synthemol and Diffusion Generative Model.

Molecular optimization agent

The molecular optimization agent was used to optimize the ADMET properties of the molecules output from the previous agents by searching its synthesizable molecular neighbors with better ADMET properties while maintaining potential inhibition. The agent consisted of two separate agents: a manager agent and an optimization agent. The optimization agent was responsible for running one iteration of molecular optimization, and the manager agent was used to determine how many iterations the molecule needed to be optimized.

Molecular optimization was performed using the SynSpace v0.3.0 package [51] to generate synthesizable molecular neighbors of the input seed molecule through 50 organic synthesis reactions [9]. We used Synspace to generate 1,000 molecular neighbors (including the seed molecule

itself) for the seed molecule. Molecular neighbors are passed through the inhibition model and ADMET models to get the predicted MTB inhibition and ADMET properties. The molecular neighbors that maintain MTB inhibition were ranked based on the total number of passed ADMET properties. The molecular neighbors with the most passed ADMET properties of the seed molecule were passed to PaperQA2. PaperQA2 was used to select the molecular neighbor with the most favorable ADMET properties among those that passed the same number of ADMET criteria from the ADMET models. The selected molecular neighbor was sent to the manager agent. If the optimized molecule differed from the seed molecule, the optimization was considered successful in finding a molecule with improved ADMET properties. The optimized molecule was then used as the new seed for another optimization iteration.

The final optimized molecule was output once no molecular neighbors outperformed the seed molecule or the predefined maximum number of iterations was reached. We used five as the maximum number of iterations.

Chemical Space Visualization

To visualize the chemical space, we used the UMAP [29] to reduce the Morgan Fingerprints nBits=2048 [30] with Jaccard distance into two dimensions. Matplotlib v2.9.0 [14] and Plotly v.5.23.0 [16] are used in visualizing the chemical space.

Molecular Describer

The biochemical literature typically does not describe molecules by their SMILES string. To enable effective literature queries, we convert the machine-readable SMILES strings to human-readable, natural language molecular descriptions. We first searched the two databases for the IUPAC name corresponding to the input SMILES string [19, 17]. If this does not exist in the database, we use STOUT V2.0 to generate a predicted IUPAC name [35]. In this case, we supplement the IUPAC information with various features extracted from RDKit [36]. We use GPT-4 [32] and one chain-of-thought example[49] to generate a description in the following format:

> Input SMILES: `COc1c(N2C[C@@H]3[C@H](C2)CCCN3)c(F)cc2c1n(cc(c2=O)C(=O)O)C1CC1`
>
> Molecular description: "This is a quinolone with a cyclopropyl substitution and a fluoro substitution. It also has a diazabicyclononane substituent and a methoxy substituent."