GenVS-TBDB: An Open AI-Generated and Virtual-Screened Small-Molecule Database for Tuberculosis Drug Discovery

Xiaoying Lv[1+], Han Guo[1+], Xi Lu[1], Zhuo Tang[1], Kai Yu[1], Gengwei Xu[1], Shawn Chen[1], Rumin Zhang[1], Jinjiang Guo[1*]

[1.] Global Health Drug Discovery Institute, Beijing, China, 100192

* Corresponding author Jinjiang.guo@ghddi.org

+ These authors contributed equally to this work

Abstract

Tuberculosis (TB) remains a leading global health threat, with over 10 million new cases and 1.25 million deaths reported in 2023. Current TB therapies rely on a limited drug repertoire and prolonged treatment courses that have changed little in four decades. Here, we present GenVS-TBDB, an open AI-generated and virtual-screened small-molecule database, which expands the chemical space against Mycobacterium tuberculosis (Mtb) essential proteins. We firstly identified 458 probable small-molecule binding pockets across 376 essential Mtb proteins by integrating crystallographic ligand binding sites, homology to ligand-bound templates, literature review and computational pocket predictions. Then, by leveraging the target-aware generative model TamGen, over 1.2 million novel small molecules were produced tailored to these pockets. All compounds were evaluated using AutoDock Vina, yielding binding affinity distribution for each target. An anti-TB specific graph neural network, Ligandformer, further enriched the compounds for whole-cell activity. We also computed their key physicochemical properties (i.e., logP, molecular weight, topological polar surface area) and flagged any Pan-Assay Interference Compounds (PAINS) to ensure medicinal chemistry tractability, enriching a set of promising, drug-like molecules targeting diverse Mtb proteins. GenVS-TBDB, accessible at http://datascience.ghddi.org/tuberculosis/view, provides downloadable docking scores, cellular activity probabilities and physicochemical annotations, thereby offering an open resource for rapid hit generation and lead optimization. Our work demonstrates the power of generative AI and

virtual screening in expanding the chemical search space for tuberculosis therapeutics and provide a valuable resource for accelerating early drug discovery.

Keywords: Mycobacterium tuberculosis, Essential Targets, Generative AI, Virtual Screening, De Novo Drug Design, Hit Generation

Introduction

Tuberculosis, caused by Mycobacterium tuberculosis (Mtb), is one of the top infectious causes of death worldwide. Despite decades of control efforts, TB caused an estimated 10.8 million new cases and 1.25 million deaths in 2023 (1). The rise of multi-drug-resistant TB (MDR-TB) further exacerbates the crisis, with ~450,000 new rifampicin-resistant cases in 2021. Standard treatment for drug-sensitive TB requires a 6-month regimen of multiple antibiotics that has remained essentially unchanged for 40 years (2). This lengthy therapy and its side effects lead to adherence challenges and treatment failure. Only a few new TB drugs have emerged in recent years – for example, bedaquiline (approved in 2012) was the first new TB drug in over four decades (3) – and these are mostly reserved for resistant cases. Clearly, there is an urgent need to discover novel TB drug candidates that can shorten therapy and overcome resistance.

One promising strategy is structure-based drug design (SBDD), which leverages the 3D structures of essential microbial proteins to guide inhibitor discovery. Mtb has approximately 600 essential genes (roughly 15% of its genome) required for in vitro growth (4), representing a wealth of potential drug targets across various pathways. With advances in structural biology – including X-ray crystallography and AlphaFold predictive model (5) – we now have structural data or high-quality models for many of these essential proteins. Structure-based virtual screening can identify small molecules that fit into a target's binding pocket (6), enabling rational targeting of enzymes and proteins vital to Mtb survival. However, traditional virtual screening of large chemical libraries is limited by library diversity and size. Here, generative artificial intelligence (AI) can expand the chemical space: generative models can design novel compounds optimized for a given protein pocket, and predictive models can forecast biological activity (7).

Recent breakthroughs in AI-driven drug discovery have demonstrated the ability to rapidly generate and optimize novel compounds. For example, deep generative models have produced potent kinase inhibitors in a matter of weeks (7). TamGen (Target-aware molecule generation) is a state-of-the-art target-aware molecule generator that employs a GPT-like chemical language model conditioned on protein pocket context (8)(9). By integrating protein structural information into the generation process, TamGen can propose compounds specifically tailored to a target binding site, improving the likelihood of binding and physicochemical properties. In parallel, AI predictive models like Ligandformer (10) (a graph neural network originally developed for compound property prediction) enable in silico phenotypic screening, i.e. predicting which

compounds are likely to exhibit anti-mycobacterial activity. Such models can learn from datasets of active vs. inactive compounds to prioritize molecules with desirable bioactivity profiles.

In this work, we present GenVS-TBDB, an open AI-generated and virtual-screened small-molecule database, to accelerate TB drug discovery. We focus on essential Mtb proteins as targets, identifying binding pockets and generating novel small molecules that could potentially bind to these targets. Then, we dual-rank these molecules based on their docking scores and AI-predicted anti-TB cellular activity scores. The compounds are assessed for bioactivity potential, drug-likeness and synthetic feasibility. Importantly, we share the GenVS-TBDB, including compounds' virtual screening analysis data, openly with the scientific community, providing an open resource to jump-start experimental testing and lead optimization. This approach exemplifies how generative AI and computational tools can greatly expand the hit discovery funnel against a high-priority infectious disease.

In this work, we present GenVS-TBDB, an open AI-generated and virtual-screened small-molecule database, to accelerate TB drug discovery. We focus on essential Mtb proteins as targets, identifying binding pockets and generating novel small molecules that could potentially bind to these targets. Then, we dual-rank these molecules based on their docking scores and AI-predicted anti-TB cellular activity scores. The compounds are assessed for bioactivity potential, drug-likeness and synthetic feasibility (Figure 1). Importantly, we share the GenVS-TBDB, including compounds' virtual screening analysis data, openly with the scientific community, providing an open resource to jump-start experimental testing and lead optimization. This approach exemplifies how generative AI and computational tools can greatly expand the hit discovery funnel against a high-priority infectious disease.

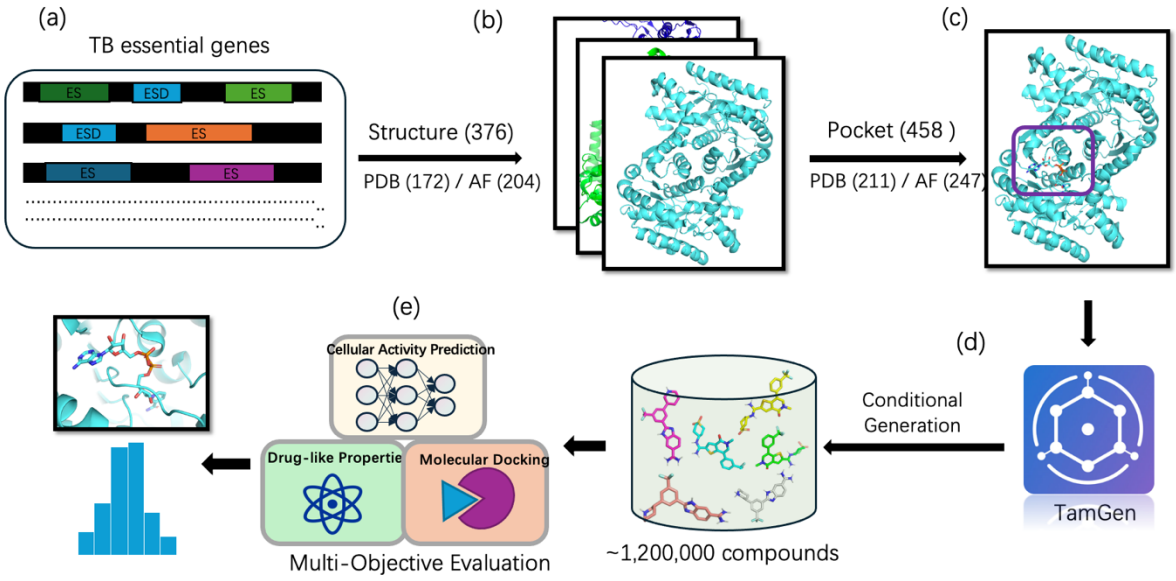

Figure 1: Workflow for constructing GenVS-TBDB. Key steps: (a) selection of an essential Mtb protein target; (b) retrieval or modeling of its 3D structure; (c) identification of binding pockets via co-crystal ligands, homology alignment, and pocket prediction algorithms (Fpocket); (d) compilation of all likely pockets to be used for compound generation; (e) Multi-objective evaluation with molecular docking, AI-predicted cellular activity and drug-like property for all compounds in GenVS-TBDB.

Methods

Target Selection

We selected a panel of Mtb protein targets based on genetic essentiality and druggability. Essential genes were defined according to genome-wide transposon mutagenesis studies (4), which have identified hundreds of loci indispensable for Mtb growth in vitro. From a comprehensive essential-gene list (4), we downloaded 1,157 protein crystal structures of these targets from the Protein Data Bank (PDB) database. In cases where a target had multiple crystal structures, we selected 172 crystal structures considering their resolution, protein length and presence of bound ligands to ensure reliable pocket identification. For essential targets lacking crystal structures, we harnessed 204 predicted models from the AlphaFold Protein Structure Database, which provides high-confidence structures for most Mtb proteins and allows us to include many otherwise intractable targets (5). This target selection process yielded a diverse set of proteins, each critical to Mtb survival, forming the basis for downstream pocket identification and ligand generation.

Binding Pocket Identification

For each chosen target protein, we identified 458 candidate small-molecule binding pockets using a combination of structure-based analyses from highest to lowest confidence level as shown in Table S1:

Known ligand sites: If a target's crystal structure contained a co-crystallized ligand (such as inhibitor, substrate or analog), we assumed that vicinity as a binding site. The bound ligand's coordinates defined the pocket location and shape. For this category of pocket, a total of 159 binding sites were detected.

Homology-based pockets: For targets without known ligands, we searched for homologous proteins (e.g. in other bacteria H37Ra, M.smeg, E.coli) with ligand-bound structures. By aligning the target's structure with its homolog, we inferred 110 corresponding ligand-binding pockets. Corresponding crystal target proteins were re-constructed using homologous proteins as template to ensure the suitable side chain position in binding pocket.

Binding Pockets Deduced from Literature Review: Our investigation extended to a review of pertinent literature focusing on related crystal structures, with the goal of identifying potential binding sites in the absence of co-crystallized proteins with homology. As a result of this comprehensive literary analysis, we pinpointed 20 binding pockets specific to this category.

Computational predictions: For targets devoid of the aforementioned information, we utilized the Fpocket, an open-source pocket detection tool, which is adept at discovering surface cavities that possess the requisite dimensions and shape conducive to ligand binding (6). Typically, we select one or two pockets per target based on their druggability scores. Employing this technique, we successfully identified 169 pockets.

These targets without obvious cavity in surface were excluded in our protein library. All these sources of information were integrated to build a comprehensive pocket map for each protein (Table S1). In many cases, multiple pockets were identified per protein (e.g. orthosteric active site and secondary/allosteric sites). We catalogued 458 distinct pockets across all targets. Figure 1 illustrates this workflow of binding site identification. Each pocket was annotated with its originating evidence (ligand-bound, homologous, literatures or predicted). This pocket compendium guided our molecular generation, with each pocket considered a separate design scenario for creating tailored compounds.

Molecular Generation with TamGen

We applied the TamGen algorithm (8) to design novel small molecules for each identified pocket. TamGen is a deep learning-based generative model that extends a transformer-type chemical language model with protein context encoding. In brief, TamGen's architecture includes a protein encoder that processes the target pocket information (e.g. sequence of pocket residues or 3D environment features) and a compound decoder that outputs novel molecular structures in string format (such as SMILES notation). By conditioning the generation on pocket features via a cross-attention mechanism, TamGen produces compounds predicted to fit and interact favorably with the given binding site. This approach effectively biases the chemical output toward each specific target site, unlike generic de novo design.

For each pocket (all 458), we initiated TamGen to generate an extensive set of candidate compounds. Generation was performed in iterative batches to maximize diversity: initial compound proposals were obtained by random sampling from the model, then a subset of promising structures (based on heuristic desirability or a preliminary docking score) could be used as "seeds" for further refinement generations (TamGen supports iterative compound refinement). This procedure yielded over 1.2 million unique small-molecule structures overall, with roughly thousands of compounds per pocket on average. Notably, TamGen is designed to optimize multiple objectives simultaneously; beyond just pocket shape complementarity, it considers drug-likeness metrics during generation. The model inherently tends to avoid overly large or undruggable structures, aiming for compounds in a medicinal chemistry-friendly range of size and complexity. The chemical space spanned by these generated molecules is enormous, covering a wide variety of scaffolds, functional group combinations, and three-dimensional geometries not found in existing libraries. By generating this AI-driven library, we circumvented the limitations of predefined screening collections and created bespoke molecules for Mtb targets.

Virtual Screening on Bioactivity Potential

To evaluate the binding potential of the generated molecules, we performed molecular docking for each compound into its designated target pocket. We used AutoDock Vina 1.2.5, a widely used open-source docking program known for its speed and accuracy (11). Vina 1.2.5 offers enhancements including a reoptimized scoring function and a convenient batch mode for high-throughput docking – critical features for handling our large library. Prior to docking, protein structures were prepared by adding missing hydrogen atoms, assigning protonation states appropriate for pH 7, and defining a docking box around the pocket region. Ligand structures were

generated by TamGen in SMILES, which we converted to 3D conformations and minimized. Each ligand was docked into the pocket using multiple runs to sample different binding poses. Vina outputs a docking score (in kcal/mol) for the top poses, which is an estimate of the free energy of binding (more negative indicates stronger predicted binding). The positive docking scores of molecules were set to empty, as they are unlikely to bind to their target pockets.

We recorded the best docking score for each compound–target pair. These scores allowed us to rank compounds and also compare the relative druggability of different pockets. As an internal check, for some targets with known inhibitors, we docked the known ligand as a reference. In general, compounds with scores better (more negative) than that of a known moderate inhibitor (~–7 kcal/mol) were empirically considered as interesting hits. After docking the entire set, we compiled a distribution of docking scores for each target pocket (illustrated in Figure 2a). We also summarized results across all targets to see overall trends. Many targets had at least one compound with a very strong score (e.g. < –10 kcal/mol), indicative of high-affinity interactions in those pockets.

In parallel with docking-based scoring, we employed an AI phenotypic activity predictor to further sort the compounds. We used a deep learning model, namely Ligandformer, which is a graph neural network framework tailored to predict compound bioactivity (10). Specifically, we trained/adapted Ligandformer to predict anti-TB cellular activity, using publicly available data on known TB-active (Minimum Inhibitory Concentration, i.e., MIC ≤10uM) and inactive molecules (12). The model takes a compound's structure as input and outputs a probability score (ranging from 0 to 1) which estimates the likelihood that the compound's MIC against Mtb is less than 10 uM.In 5-fold cross-validation, the model achieved the average Area Under the Receiver Operating Characteristic Curve (AUROC) = 0.82 and Area Under the Precision-Recall Curve (AUPR) = 0.77 at an active probability threshold of 0.5 (Figure S1 (a)). We also curated an independent validation set containing 394 GHDDI internal pipeline compounds, on which our model achieved a 50 % hit-rate in the top-20 ranked molecules (2.1-fold enrichment over the 23.6 % random-screening baseline) and maintained ≥40 % hit-rate through the top-100, confirming its ability to rapidly surface true actives for downstream experimentation (Figure S1 (b)). Such a phenotypic predictor also complements the structure-based docking: docking focuses on binding to a specific target, whereas the Ligandformer model can learn aspects of cell permeability, metabolic stability or other features that influence whether a compound actually works in a whole-cell context. Therefore,

after docking, we sorted compounds according to their predicted probabilities of cellular activity by the AI model. The compounds with favorable computational profiles thus satisfied two criteria: good binding score and good predicted cellular activity. This dual-ranking strategy balances the pocket-specific design with a data-driven likelihood of translating into actual antimicrobial effect.

Molecular Drug-like Properties

All generated compounds were characterized by key physicochemical descriptors relevant to drug-likeness and synthesizability. We calculated properties including partition coefficient (log P), molecular weight (MW), topological polar surface area (TPSA), and number of rotatable bonds. These descriptors inform Lipinski's Rule of Five – a guideline for oral drug-likeness (MW < 500, logP < 5, TPSA ≤150) (13). We evaluated these compounds conformed to established another drug-like parameters, as defined by the Quantitative Estimate of Drug-likeness (QED) criteria (14). Additionally, we computed the synthetic accessibility (SA) score for each compound (using RDKit) to gauge how easily the molecule might be synthesized (15). Where necessary, we applied an SA cutoff to eliminate highly complex structures unlikely to be synthetically tractable. To facilitate visual interpretation of the synthesized compound library, a comprehensive comparative analysis was performed between the generated structures and the TamGen-produced protein conformations from the CrossDocked2020 dataset (16), employing the previously described multidimensional evaluation metrics.

Another crucial descriptor was the alert of PAINS (Pan-Assay Interference Compounds) and other frequent hitters. We screened all structures against known PAINS substructures (17)– chemically reactive or promiscuous motifs that often give false positives in assays. Any compound containing a PAINS alert (such as catechols, aggregating dyes, etc.) was flagged in the database to avoid pursuing problematic chemotypes. By applying these property-based descriptors, we ensured that the final set of proposed molecules not only have high predicted efficacy but also possess favorable drug-like profiles and fewer liabilities. All cheminformatics analyses were performed with standard libraries (e.g. RDKit for property calculation).

Results

Docking Score Distribution across Targets

This study evaluated the Vina docking score distributions across 1.2 million small molecules screened against 458 essential protein pockets. The analysis identified distinct patterns in ligand-target interactions, with significant variations in binding affinities observed across different

pockets. Figure 2a shows the distribution of docking scores (AutoDock Vina scores, in kcal/mol; more negative is better) aggregated over all targets. The distribution is roughly normal: a majority of compounds achieved moderate scores around –6 to –10 kcal/mol (comparable to micromolar to sub-micromolar affinities). The overall median docking score was approximately –7.5 kcal/mol. More than half of the targets had at least one compound with a docking score better than –10, suggesting that TamGen successfully found high-affinity chemotypes for many essential proteins. This indicates that for most target pockets, TamGen was able to generate some ligands with reasonable predicted affinity.

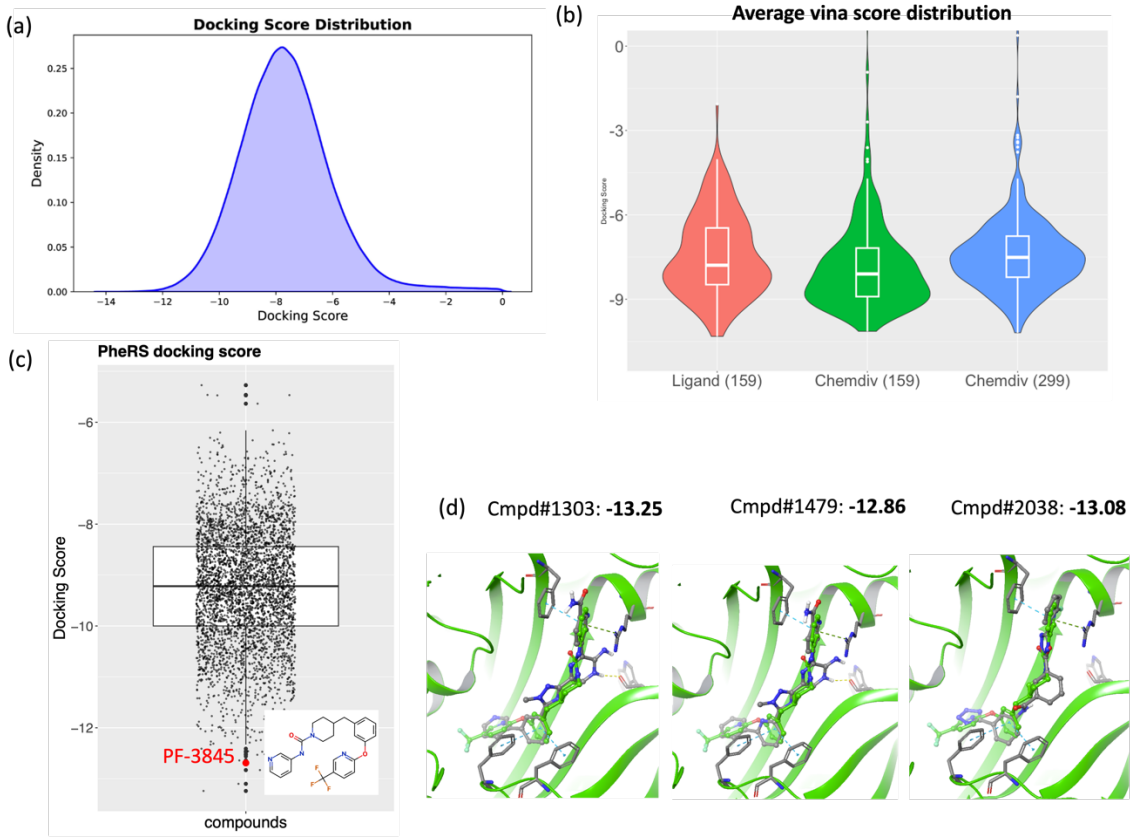

Figure 2: Distribution of docking scores for AI-generated compounds against Mtb targets and related compounds against PheRS target. (a) The histogram shows the frequency of AutoDock Vina scores across all ~1.2 million compound–target docking simulations. A lower (more negative) score indicates a better predicted binding affinity. (b) An inset compares score distributions for different target classes (enzymes vs. other proteins). Vina average docking scores across three distinct experimental groups. The 159 and 299 in bracket represents these co-crystallized and apo essential proteins. "Ligand" and "Chemdiv" represents the seed of TamGem from co-crystal ligands and chemdiv library, respectively. (c) Distribution of docking scores for related compounds against PheRS target. The red dot means the docking score of co-crystal PF-3845 compound. (d) The binding poses for three compounds with better docking score compared with PF-3845.

This study also evaluated the Vina average docking scores across three distinct experimental groups: (1) 159 proteins using native co-crystal ligands as seed molecules for TamGen, (2) the same 159 proteins seeded with ten ChemDiv-derived small molecules, and (3) 299 apo proteins similarly seeded with ten ChemDiv compounds. Notably, the second group exhibited the lowest average docking scores (Figure 2b), suggesting enhanced binding affinities compared to the co-crystal ligand group and the apo protein cohort. This disparity may arise from the ChemDiv compounds in the second group adhering more closely to the Rule of Five, compared with the native co-crystal ligands (Group 1), while the apo proteins (Group 3) may accommodate diverse ChemDiv molecules due to their unoccupied, flexible binding sites. These findings underscore the interplay between ligand physicochemical properties and protein conformational states in docking performance, highlighting the need for context-specific compound libraries in structure-based drug discovery. We also note that the distribution of top scores correlates with known target druggability; pockets known to bind potent inhibitors (like the proteasomal enzyme active site) showed tighter score ranges, whereas more challenging or novel pockets had broader, lower-affinity distributions.

At the per-target level, we observed that certain targets yielded consistently better docking hits. The top five representative targets are RmlB (-10.19), Glf (-10.14), EmbC (-10.10), LigA (-9.94), and MurA (-9.84) with best average docking score. The broad molecular binding observed in Mtb essential targets—RmlB, Glf, EmbC, CtaD, LigA, and MurA—stems from their critical roles in bacterial survival and structural adaptations to diverse substrates. Enzymes such as RmlB (cell wall biosynthesis) and MurA (peptidoglycan synthesis) exhibit flexible substrate-binding pockets to accommodate nucleotide sugar derivatives, while glycosyltransferases (Glf, EmbC) require tolerance for branched polysaccharide precursors, enabling interactions with varied ligands. LigA's dynamic NAD+-dependent DNA repair mechanism permits promiscuous binding. Evolutionary pressures under host stress conditions may have further selected for structural plasticity in these targets, combining conserved catalytic cores with peripheral flexibility. This functional and structural adaptability highlights their vulnerability to inhibition but underscores the need for precision in drug design to balance potency and selectivity. These findings position these targets as high-priority nodes for anti-tubercular therapeutic development.

Conversely, targets Rv1110 (11.38), Rv1990c (31.15), and CarA (32.17) displayed the highest (least favorable) scores, indicating potential challenges in ligand binding or suboptimal pocket geometries. For these three targets with the poorest performance, the limited space within their

binding pockets presents a consistent obstacle for accommodating small molecule docking, as illustrated in Figure S2.

An essential Phenylalanyl-tRNA synthetase (18) had ~3,000 compounds docked in ATP and amino acid pocket, a median score of –9.2, and at least 3 compounds with scores stronger than –12.69 (the score of co-crystallized PF-3845 with Ki = 0.73 ± 0.06 μM), with the best at –13.25 kcal/mol (Figure 2c). Molecular docking revealed three compounds exhibiting superior binding affinities compared to PF-3845, as indicated by lower docking scores: Cmpd#1303 (-13.25), Cmpd#1479 (-12.86), and Cmpd#2038 (-13.08). These compounds demonstrated favorable physicochemical properties, with logP/MW/TPSA values of 2.646/448.5/139.2 (Cmpd#1303 and #1479) and 3.243/427.4/122.5 (Cmpd#2038), suggesting drug-like potential. The only difference between molecules 1303 and 1479 lies in the position of an N atom at benzene ring (Figure 2d). Structural analysis illustrated that all three compounds adopt binding modes analogous to PF_3845, occupying a similar position within the catalytic pocket and forming conserved interactions with key residues. The enhanced docking scores, coupled with comparable binding geometries, highlight these molecules as promising PheRS inhibitors for further optimization. This work provides a foundation for developing novel therapeutics targeting aminoacyl-tRNA synthetases.

These docking results validate the target-aware generation strategy: by designing compounds specifically for each pocket, we achieved good docking fits for many targets. This provides insight into which target classes might be most effectively addressed by our compounds. Importantly, docking alone is an imperfect predictor of true binding affinity; nonetheless, the scoring step was invaluable for cutting down the enormous library to a manageable list of high-scoring candidate molecules for each target.

AI-Predicted Anti-TB Cellular Activity

Integrating the Ligandformer AI predictions with the docking results greatly enhanced our hit prioritization. After obtaining docking scores, we input each compound into the Ligandformer model to predict its anti-tubercular activity probability. The model output is a score from 0 to 1, where higher values indicate greater likelihood that the compound will inhibit Mtb growth (based on learned patterns from known actives). We set a conservative threshold (tuned on a validation set of compounds with known activities) to label compounds as "predicted active." Roughly speaking, about 45% of the generated compounds was predicted to be active against Mtb by this model (with a probability >0.5 for a cellular MIC≤ 10uM for each compound) (Figure 3). When

we dual-ranked with docking, an encouraging trend emerged: compounds with better docking scores tended to also receive higher activity predictions. This enrichment suggests our generative design succeeded in producing molecules that are not only good binders but also carry features associated with cellular activity (possibly due to TamGen's built-in bias toward drug-like chemistry).

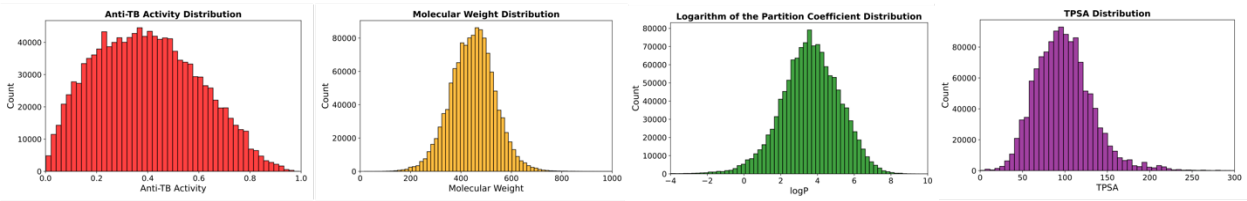

Figure 3: AI-predicted anti-TB cellular activity and Physicochemical property profiles of generated molecules and example structures. (a) Histograms of molecular weight, cLogP, and TPSA for the compounds, indicating adherence to drug-like ranges.

We dual-ranked the compounds according to their docking scores and predicted cellular activity probabilities as prioritization guide. This prioritizes the top of target-specific compounds that more likely exhibit bioactivity, which is a tractable number for further analysis or experimental testing. The AI model cannot yet pinpoint the mechanism of action, but it flags compounds with profiles similar to known TB-active molecules. This serendipitous identification of known motifs is reassuring, and more importantly, the model also identified novel scaffolds with high predicted activity that have not been previously reported.

In summary, the AI-predicted activity and docking prioritized a refined set of high-potential hits: molecules that not only bind well to their intended Mtb target but also carry a computational imprimatur of likely whole-cell efficacy. These prioritized compounds form the basis of our top candidate selection in the subsequent analysis.

Library Drug-Likeness and Molecular Properties

We next measured the physicochemical profiles of the library to evaluate whether they meet desirable drug-like criteria. Figure 3 depicts the distribution of key molecular properties (molecular weight, cLogP, and topological polar surface area) for all compounds after measurement. The compounds span a range of sizes: molecular weight ranges from about 250 up to 500+ Da, with a median around 360 Da. This is within a typical small-molecule drug range and suitable for potential oral bioavailability. The clogP (calculated octanol-water partition coefficient) distribution centers around 2–3, with the majority of compounds between 1 and 4 log units, indicating moderate lipophilicity. Only a small fraction has clogP > 5, which could signal solubility

issues. The topological polar surface area (TPSA) values cluster in the 50–100 Å^2 range, consistent with good permeability (generally TPSA <140 Å^2 is favorable for oral drugs).

We also compared the drug-like properties of molecules of the GenVS-TBDB with those from other datasets, using metrics such as QED, molecular diversity, synthetic accessibility (SA), and Lipinski rule compliance. For instance, the GenVS-TBDB molecules achieved a molecular diversity score 0.866, outperforming the other datasets (Table 1). The synthetic accessibility (SA) score, reflecting favorable synthetic feasibility, was calculated as 0.782 for our dataset, which is comparable to or exceeds values reported for other reference libraries. In contrast, the quantitative estimate of drug-likeness (QED) and Lipinski's rule compliance scores were lower in our dataset (0.498 and 0.608, respectively) compared to the reference dataset (QED = 0.596; Lipinski's rule compliance = 0.964). These results highlight opportunities for further optimization of drug-like properties while retaining robust synthetic tractability in our library.

	QED	**Diversity**	**SA**	**Lipinski**
CrossDock2020	**0.596**±0.187	0.863±0.022	0.764±0.087	**0.964**
GenVS-TBDB	0.498±0.183	**0.866±0.007**	**0.782±0.062**	0.608

Table 1. Comparison of molecular properties (QED, diversity, SA, Lipinski compliance) between the GenVS-TBDB and other datasets. The highest score in each column was highlighted by bold.

Notably, after applying PAINS (Pan-Assay Interference Compounds) and reactivity descriptor filters, only 0.25% of the compounds in our library exhibited structural motifs associated with frequent assay interference (Table 2). In comparison to established libraries containing FDA-approved drugs or clinical-stage candidates, the molecules by TamGen demonstrated a PAINS rate of 0.25% (3298 compounds), which is significantly lower than those of the FDA Library (2.6%) and ReFrame Library (2.1%). This stark contrast underscores the reduced risk of promiscuous binding in GenVS-TBDB, thereby enhancing their reliability and prioritization potential for drug discovery pipelines.

	Total Count	**PAINS Count**	**PAINS%**
FDA	2,664	68	2.6%
Prc&Cli	1,129	21	1.9%
ReFrame	11,599	244	2.1%
GenVS-TBDB	1,252,274	3298	0.26%

Table 2. PAINS counts and percentages in GenVS-TBDB versus known libraries.

Open-access of the GenVS-TBDB

The GenVS-TBDB (accessible at http://datascience.ghddi.org/tuberculosis/view) provides a comprehensive repository of Mtb essential proteins, their structural data, and associated small-molecule interactions. Users can search for targets using three primary identifiers: (1) locus name (e.g., Rv3280), (2) gene name (e.g., accD5), or (3) structural ID (e.g., PDB ID 2BZR). Upon querying, the platform dynamically displays interactive visualizations, including 3D protein structures and center position coordinates, and predicted cellular activity profiles of associated compounds. For example, searching for the structural ID "2BZR" directs users to a dedicated page (http://datascience.ghddi.org/tuberculosis/detail?pid=2BZR) showcasing the crystallographic structure of AccD5 (Rv3280), annotated pockets with spatial coordinates, and a list of docked molecules ranked by Vina scores.

The database supports bulk downloads of all target metadata (e.g., locus tags, gene names, structural IDs, pocket counts) in CSV formats. Additionally, individual target pages enable users to download detailed molecular datasets, including compound smiles physicochemical properties (e.g., molecular weight, LogP, HBA, HBD), and docking scores. This feature facilitates downstream analyses such as structure-activity relationship (SAR) studies or virtual screening workflows.

The integration of structural biology data with cheminformatics tools positions this resource as a critical platform for TB drug discovery, enabling researchers to prioritize targets, optimize lead compounds, and explore novel binding mechanisms.

Discussion

Our study demonstrates a novel convergence of structure-based methods and generative AI design to address the urgent need for new tuberculosis therapeutics. By targeting a wide array of essential Mtb proteins, we aim to expand the therapeutic arsenal beyond the current limited set of TB drugs. The results yielded multiple potential drug candidates worthy of experimental follow-up.

The pipeline succeeded in generating high-affinity ligands for many Mtb targets, as evidenced by likely sub-micromolar range docking scores and supportive AI activity predictions. The ability to generate targeted inhibitors for different pathways – cell wall synthesis, protein degradation, fatty acid metabolism, etc. – raises the possibility of combination therapies drawn entirely from the novel compounds, which could attack Mtb on several fronts and potentially shorten treatment.

Moreover, current first-line TB drugs (isoniazid, rifampicin, ethambutol, pyrazinamide) were mostly discovered decades ago through phenotypic screens, and they hit a small set of targets. Newer drugs like bedaquiline (targets ATP synthase) (3) and pretomanid (multiple effects via nitric oxide release) emerged after long gaps and are used for specific cases. Our GenVS-TBDB provides fresh chemical matter against both the classic targets and entirely new ones. Notably, many of our top compounds have novel scaffolds unrelated to existing TB medications.

Despite the powerful integration of methods, several limitations of our study must be acknowledged. First, only one structure for an essential target was selected for molecular generation despite a protein potentially possessing multiple pockets. It's important to note that variations in side-chain positions within the same pocket, as well as different protein conformations, were not taken into account. Second the accuracy of docking and scoring is finite. AutoDock Vina's scoring function, while useful for relative ranking, can produce false positives or negatives due to protein flexibility not being fully accounted for and the approximate nature of the scoring. We attempted to mitigate this by focusing on top-ranked compounds and using an orthogonal AI prediction, but some highly scoring compounds may not bind as well experimentally (e.g., if the protein undergoes an induced fit that was not captured). Third our Ligandformer AI model, being a statistical learner, is only as good as the data it was trained on. If certain chemotypes were absent or underrepresented in training data, the model might mispredict those. We treated the AI output as a guide rather than a strict rule – some compounds we set aside as "predicted inactive" might still be worth testing if they have exceptional docking and properties, because the model could be wrong.

It is important to contextualize that while our computational approach generated many ideas, actual drug development will require experimental validation at each step. This project sets the stage for future endeavors, including experimental validation of compounds, optimization using AI with human input, enhancing AI models, and potentially expanding the approach to other diseases.

Conclusion

Targeting 458 druggable binding pockets on 376 essential TB proteins, we curated a vast library of novel small molecules for TB drug discovery, namely GenVS-TBDB, enabled by an innovative generative AI and virtual screening pipeline. The molecular generation model, TamGen, produced over 1.2 million small molecules, which were systematically assessed for binding affinity using

docking evaluations and further enriched for anti-TB cellular activity with the AI predictor Ligandformer, yielding top candidates with high potency and favorable drug-like properties. In an unprecedented move, we have made this entire library publicly available in a public website to foster global collaborative research and expedite TB drug R&D, believing that shared data can spur discoveries by multiple groups simultaneously.

Our work demonstrates the power of generative AI and multi-objective virtual screening in uncovering efficacious antitubercular agents and offers a framework that can extend to other infectious diseases. The next phase involves experimental validations, while the larger vision is to leverage this AI-driven method to address other formidable pathogens, thereby progressing toward ending the TB epidemic through open science and computational advances.

References

1. Organization. Fact sheets of Tuberculosis from WHO [Internet]. 2025. Available from: https://www.who.int/news-room/fact-sheets/detail/tuberculosis

2. Dartois VA, Rubin EJ. Anti-tuberculosis treatment strategies and drug development: challenges and priorities. Nat Rev Microbiol. 2022 Nov;20(11):685–701.

3. Mirsaeidi M. After 40 years, new medicine for combating TB. Int J Mycobacteriology. 2013 Mar;2(1):1–2.

4. DeJesus MA, Gerrick ER, Xu W, Park SW, Long JE, Boutte CC, et al. Comprehensive Essentiality Analysis of the Mycobacterium tuberculosis Genome via Saturating Transposon Mutagenesis. mBio. 2017 Jan 17;8(1):e02133-16.

5. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 2022 Jan 7;50(D1):D439–44.

6. Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: an open source platform for ligand pocket detection. BMC Bioinformatics. 2009 Jun 2;10:168.

7. Zhavoronkov A, Ivanenkov YA, Aliper A, Veselov MS, Aladinskiy VA, Aladinskaya AV, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat Biotechnol. 2019 Sep;37(9):1038–40.

8. Wu K, Xia Y, Deng P, Liu R, Zhang Y, Guo H, et al. TamGen: drug design with target-aware molecule generation through a chemical language model. Nat Commun. 2024 Oct 29;15(1):9360.

9. Schuh MG, Hesse J, Sieber SA. AI-guided Antibiotic Discovery Pipeline from Target Selection to Compound Identification [Internet]. arXiv; 2025 [cited 2025 Apr 23]. Available from: https://arxiv.org/abs/2504.11091

10. Guo J, Liu Q, Guo H, Lu X. Ligandformer: A Graph Neural Network for Predicting Compound Property with Robust Interpretation [Internet]. arXiv; 2022 [cited 2025 Apr 23]. Available from: http://arxiv.org/abs/2202.10873

11. Eberhardt J, Santos-Martins D, Tillack AF, Forli S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. J Chem Inf Model. 2021 Aug 23;61(8):3891–8.

12. Lane T, Russo DP, Zorn KM, Clark AM, Korotcov A, Tkachenko V, et al. Comparing and Validating Machine Learning Models for Mycobacterium tuberculosis Drug Discovery. Mol Pharm. 2018 Oct 1;15(10):4346–60.

13. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev. 2001 Mar 1;46(1–3):3–26.

14. Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL. Quantifying the chemical beauty of drugs. Nat Chem. 2012 Jan 24;4(2):90–8.

15. Ertl P, Schuffenhauer A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. J Cheminformatics. 2009 Jun 10;1(1):8.

16. Francoeur PG, Masuda T, Sunseri J, Jia A, Iovanisci RB, Snyder I, et al. Three-Dimensional Convolutional Neural Networks and a Cross-Docked Data Set for Structure-Based Drug Design. J Chem Inf Model. 2020 Sep 28;60(9):4200–15.

17. Baell JB, Holloway GA. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. J Med Chem. 2010 Apr 8;53(7):2719–40.

18. Wang H, Xu M, Engelhart CA, Zhang X, Yan B, Pan M, et al. Rediscovery of PF-3845 as a new chemical scaffold inhibiting phenylalanyl-tRNA synthetase in Mycobacterium tuberculosis. J Biol Chem. 2021;296:100257.