Journal of Molecular Graphics and Modelling 118 (2023) 108361

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/jmgm

De novo design of anti-tuberculosis agents using a structure-based deep learning method

Sowmya Ramaswamy Krishnan [a], Navneet Bung [a], Siladitya Padhi [a], Gopalakrishnan Bulusu [a,b], Parimal Misra [b], Manojit Pal [b], Srinivas Oruganti [b], Rajgopal Srinivasan [a], Arijit Roy [a,*]

[a] TCS Research (Life Sciences Division), Tata Consultancy Services Limited, Hyderabad, 500081, India
[b] Dr. Reddy's Institute of Life Sciences, University of Hyderabad Campus, Gachibowli, Hyderabad, 500046, India

ARTICLE INFO

Keywords:
Drug design
Computational chemistry
Deep learning
Tuberculosis
Chorismate mutase

ABSTRACT

Mycobacterium tuberculosis (*Mtb*) is a pathogen of major concern due to its ability to withstand both first- and second-line antibiotics, leading to drug resistance. Thus, there is a critical need for identification of novel anti-tuberculosis agents targeting *Mtb*-specific proteins. The ceaseless search for novel antimicrobial agents to combat drug-resistant bacteria can be accelerated by the development of advanced deep learning methods, to explore both existing and uncharted regions of the chemical space. The adaptation of deep learning methods to under-explored pathogens such as *Mtb* is a challenging aspect, as most of the existing methods rely on the availability of sufficient target-specific ligand data to design novel small molecules with optimized bioactivity. In this work, we report the design of novel anti-tuberculosis agents targeting the *Mtb* chorismate mutase protein using a structure-based drug design algorithm. The structure-based deep learning method relies on the knowledge of the target protein's binding site structure alone for conditional generation of novel small molecules. The method eliminates the need for curation of a high-quality target-specific small molecule dataset, which remains a challenge even for many druggable targets, including *Mtb* chorismate mutase. Novel molecules are proposed, that show high complementarity to the target binding site. The graph attention model could identify the probable key binding site residues, which influenced the conditional molecule generator to design new molecules with pharmacophoric features similar to the known inhibitors.

1. Introduction

Tuberculosis is a respiratory infection caused by the bacterium, *Mycobacterium tuberculosis* [1]. Despite being a respiratory pathogen, *Mtb* can also affect multiple organ systems of the human body, including spine, kidney and brain, leading to extra-pulmonary tuberculosis (EPTB) [2]. Clinically, two types of tuberculosis have been identified based on the advent of the infection after exposure to the pathogen: latent tuberculosis infection (LTBI) and tuberculosis disease (TB). In case of LTBI, the infection is asymptomatic despite exposure to the pathogen, making diagnosis and treatment equally challenging [3]. Further, human immunodeficiency virus (HIV) infection has been known to increase susceptibility to TB, due to the weakening of the immune system by the former [2]. As of 2020, an estimated 5.8 million people have been infected with TB globally, with India, Indonesia and Phillipines being the worst affected countries [4]. With increasing incidence of drug resistance in TB patients and the emergence of extensively drug-resistant TB strains (XDR-TB), it is essential to rapidly identify novel anti-tuberculosis agents to tackle the infection [5].

Multiple virulence factors and potential targets of hypervirulent *Mtb* strains have been identified for targeted therapy [6–8]. These targets are predominantly involved in the following essential pathways of the bacterium: lipids and fatty acid metabolism, mycolic acid biosynthesis, complex lipid biosynthesis, cholesterol catabolism, transport, secretion, apoptosis and protein degradation. However, protein biosynthesis pathways in *Mtb* remain the least explored, although they encompass several pathogen-specific pathways with limited chances to induce side effects [9]. One such pathogen-specific pathway of interest is the shikimate pathway, which catalyzes the biosynthesis of aromatic compounds in bacteria, fungi, algae, and plants, including the essential amino acids phenylalanine and tyrosine. The first crucial step of this pathway is the conversion of chorismate to prephenate catalyzed by the enzyme chorismate mutase (EC 5.4.99.5). This is a unique reaction of interest, due to the chair-like endo oxabicyclic transition state formed

* Corresponding author.
E-mail address: roy.arijit3@tcs.com (A. Roy).

https://doi.org/10.1016/j.jmgm.2022.108361
Received 19 August 2022; Received in revised form 10 September 2022; Accepted 7 October 2022
Available online 13 October 2022

during the reaction [9]. Consequently, a few studies have attempted to design potential transition state analogues (TSA) inhibiting the enzyme [10–12], and preventing the biosynthesis of essential amino acids necessary for survival of the bacterium.

In this study, we used a structure-based *de novo* drug design method [13], to design potential small molecules that can inhibit the *Mtb* chorismate mutase enzyme. The structure-based method involves a semi-supervised multimodal deep learning model utilizing a graph representation of the protein binding site structures and SMILES representation of the ligand. This model learns from experimentally determined protein-ligand complexes to design novel small molecules for any target protein with known structure. A multimodal drug-target affinity (DTA) prediction model is used to formulate a reward function for target-specific bioactivity maximization, which is utilized as the objective to optimize the molecule generation process in a reinforcement learning framework.

With just the knowledge of the conformation of binding site residues extracted from the available crystal structure of a TSA-CM complex [14], the binding site graph was constructed and used to design novel anti-tuberculosis agents. The designed molecules were compared with the existing inhibitors reported for chorismate mutase [9]. The method could produce molecules with similarity to existing inhibitors and new molecules with high complementarity to the *Mtb* chorismate mutase binding site. The generated molecules also preserved features of the existing inhibitors although the model had information about only the binding site of the target protein. Finally, based on the graph attention model, a set of key binding site residues were identified which could be responsible for favorable interaction of the generated small molecules with *Mtb* chorismate mutase.

2. Materials and methods

2.1. Learning the binding site features of the target protein using the GAT-VAE model

The binding site is the region where the small molecule binds and modulates the function of a target protein. Consequently, the aim of the model is to learn the structure and the type of interactions between the key amino acid residues forming the binding site, which can potentially influence the molecule generation process. To facilitate this, the binding site was represented in a ligand-agnostic manner using a residue interaction graph where, nodes represent residues and edges represent interactions between residues within a 4 Å distance cutoff [15,16]. Each node was featurized using a 7-class classification derived from literature [17] along with two binary bits representing the hydrogen bond accepting and donating potential of the residue. A dataset of 5981 such non-redundant binding site graphs were collated from the PDBbind [18] and scPDB [19] databases. Binding sites were also filtered such that they contained only the 20 standard amino acids. A variational autoencoder (VAE) model composed of graph attention (GAT) layers was trained with the one-hot encoded node feature vector and the unweighted adjacency matrix of the binding site graphs as input (Fig. 1a). The GAT-VAE encoder included five parallel attention heads of 128 dimensions each and a single head GAT layer for output aggregation from the heads of the parallel layer. The encoder embeds the input binding site graph into a 256-dimensional latent vector (z), which is input to the GAT-VAE decoder. The dimension of latent vector for the GAT-VAE model was determined based on the smallest latent vector dimension for the SMILES-VAE model that could produce the best benchmarking results in terms of the GuacaMol benchmark metrics (see supplementary information 1 - Table S1). The decoder is trained to reconstruct the adjacency matrix from the latent vector. The model was trained using the Adam optimizer with an initial learning rate of 0.001 and a batch size of 256 for 100 epochs in a Tesla V100 GPU. All implementations were done using PyTorch.

2.2. Learning the grammar of small molecules using the SMILES-VAE model

The dataset of ~1.6 million drug-like small molecules was obtained from the ChEMBL database in SMILES format. The SMILES dataset was pre-processed by following the procedure from our previous study [13] using the RDKit library. The SMILES-VAE model is composed of an encoder and a decoder (Fig. 1b) made of stack-augmented bidirectional gated recurrent units (GRU). An embedding layer was used to pass the input to the encoder and a dense layer with log softmax activation was used to convert model outputs to probabilities at the decoder. The model was trained using the AMSGrad optimizer with an initial learning rate of 0.0005 and a batch size of 256 for 100 epochs on a Tesla V100 GPU. Learning rate decay and gradient clipping were used to prevent vanishing and exploding gradients, respectively. All implementations were done using PyTorch.

2.3. Combining the pre-trained VAE models to form the conditional molecule generator

The pre-trained GAT-VAE and SMILES-VAE models were combined to obtain the conditional VAE model (Fig. 1c). A joint latent vector was

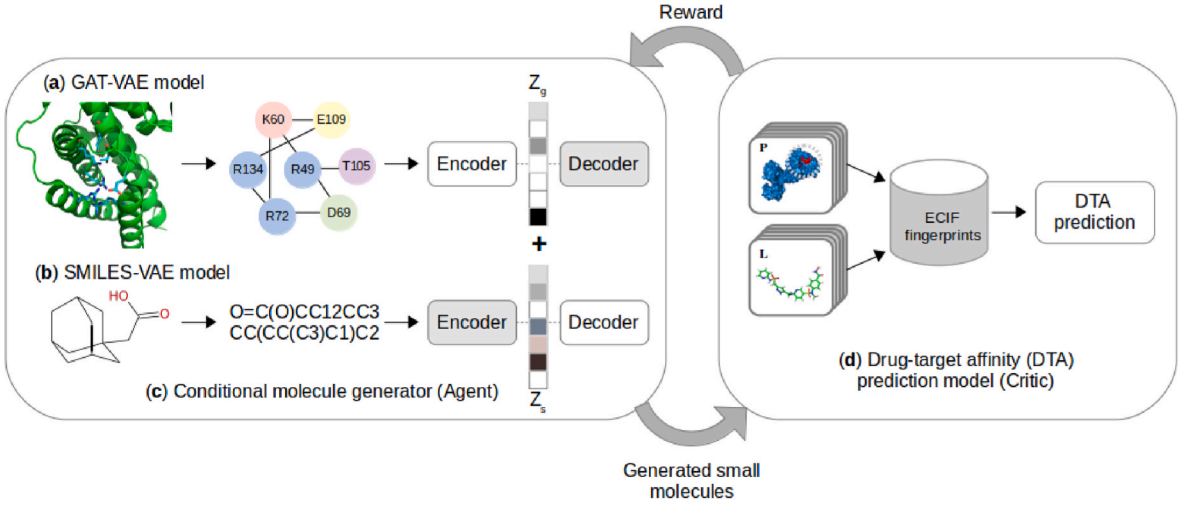

Fig. 1. Structure-based drug design pipeline. The components of the pipeline are: (**a**) Learning the binding site graph using the GAT-VAE model; (**b**) Learning the molecular grammar using the SMILES-VAE model; (**c**) Conditional molecule generator; (**d**) Drug-target affinity (DTA) prediction model.

constructed by combining the binding site graph latent vector (z_g) from the GAT-VAE model, and the primer string latent vector (z_s) from the SMILES-VAE model. This joint latent vector was used to condition the decoder of the SMILES-VAE model to generate novel target-specific small molecules. It is a pre-requisite for the latent vectors z_s and z_g to be of same dimensions to enable addition [20,21]. The conditional molecule generator was subject to a short pre-training phase with a dataset of binding site graph – small molecule pairs curated from the PDBbind dataset [18]. This enabled the SMILES-VAE decoder to learn how to decode the joint latent vector into chemically valid small molecules, and retain model stability during further optimization with reinforcement learning.

2.4. Optimizing the conditional molecule generator using a drug-target affinity (DTA) prediction model

With the pre-trained conditional molecule generator as the agent and a generic drug-target affinity (DTA) prediction model as the critic, the binding affinity of the generated small molecules was optimized using reinforcement learning (RL). The generic DTA model was trained to be target-agnostic, and can predict the binding affinity towards any given target protein (Fig. 1d) using extended connectivity interaction fingerprints (ECIF) as the input representation [22]. The model was trained, tuned and validated based on the model architecture and hyperparameters indicated in the previous study [22]. PDBbind core set and Astex diversity set [23] were used for model validation and testing, respectively.

ECIF fingerprints for a protein-ligand complex were obtained through on-the-fly docking of molecules to the target binding site using gnina [24]. The DTA model predicts the pIC_{50} value of the protein-ligand interaction, which is used to optimize the model based on the following reward formulation (1).

$$r(x) = \exp\left(\frac{x}{3.0}\right) \quad (1)$$

where, x refers to the predicted pIC_{50} value of the generated molecule. The reinforcement learning process is terminated when the bioactivity distribution for the generated small molecules is well optimized. Termination of the RL training process is target protein-dependent, and multiple criteria are considered as discussed in our previous study [21].

2.5. Validation of the generated small molecules after RL training

An *in silico* validation was performed to understand the quality of the generated molecules. Since *Mtb* chorismate mutase is an under-explored target protein, only 37 inhibitors could be identified in literature which were tested against the chorismate-binding site [9]. These 37 inhibitors, for which bioactivity data is available, were considered as the validation set for comparison with the generated molecules after RL. A set of 10000 small molecules with predicted bioactivity values was obtained after RL training. As a first step of validation, the Tanimoto similarity of generated small molecules with the 37 known inhibitors was calculated [25], along with similarity of various physicochemical property distributions. Due to the unavailability of a large enough validation dataset, a pharmacophore-based screening was also performed to understand if the diverse designed small molecules have spatial features similar to the known chorismate mutase inhibitors. The PharmaGist program [26] was used for ligand-based pharmacophore analysis. A random set of 32 molecules from the set of known inhibitors was used as input to the PharmaGist program. From the PharmaGist output, coverage of binding site was used to choose the top 2 composite ligand-based pharmacophores for further screening of generated small molecules.

The generated set of small molecules were filtered further based on their solvent accessible surface area (SASA) when bound to the chorismate mutase binding site based on the complex structure obtained from docking using gnina [24]. This criterion filtered out the small molecules which can have regions partially residing outside the binding site. The free and bound SASA values for the small molecule were calculated using the FreeSASA program [27] with the Shrake-Rupley algorithm [28] and a default probe radius of 1.4 Å. Generated small molecules with less than 10% of exposed surface area when bound to chorismate mutase were chosen for further analyses. These molecules were clustered using Butina clustering [29] with a Tanimoto distance cut-off of 0.5 and 1024-bit ECFP4 fingerprints for molecular representation. The interaction of each cluster with the binding site residues were analyzed to elucidate the major binding site residues governing the interaction of generated small molecules with chorismate mutase.

3. Results and discussion

3.1. Performance of the pre-trained models

The accuracy, uniqueness and novelty of the SMILES-VAE model were found to be 93.22%, 99% and 96%, respectively, as per the GuacaMol distribution learning benchmark (v0.5.3) [30]. The other benchmarking results of the SMILES-VAE model are discussed in detail in the supplementary information 1 (Section S1). The ROC score for the GAT-VAE model (4 Å edge cutoff) was 0.89. Five-fold cross validation of the DTA model yielded Pearson correlation coefficients (R_p) of 0.851 (RMSE = 1.21) and 0.565 (RMSE = 1.52) for the PDBbind core set and Astex diversity set, respectively (Table S2). Further, analysis of SMILES-VAE and GAT-VAE latent vectors using PCA showed that atleast 200 dimensions are required to capture 90% of variance in the input data (see Supplementary information 1, Section S2).

3.2. Generating novel small molecules targeting chorismate mutase

The structure-based drug design method was used to generate novel small molecules specific to chorismate mutase (PDB ID: 2FP2) of *Mycobacterium tuberculosis*. The limited number of inhibitors reported in literature for inhibition of the conversion of chorismate to prephenate, highlights the gap in designing a large and diverse set of small molecules targeting this essential step. The binding site graph was extracted from the experimental rotamer conformations observed when chorimsate mutase is bound to a transition state analogue (PDB ID: 2FP2). The final bioactivity distribution obtained after optimization is given below (Fig. 2a). The resultant target-specific conditional VAE model was used to sample 10000 small molecules, which were processed to omit chemically invalid molecules and canonicalized. The set of physicochemical property distributions after reinforcement learning in comparison to that of the known chorismate mutase inhibitors, is provided in the supplementary information 1 (Supplementary Fig. S1). A random set of diverse molecules from the model, along with their key physicochemical properties are also provided in the supplementary information 1 (Table S3).

3.3. Analysis of the generated small molecules

3.3.1. (A) Similarity of generated molecules based on Tanimoto coefficient

With ECFP4 fingerprints [31] as the molecular representation and Tanimoto coefficient (TC) as the distance metric, similarity was quantified to known chorismate mutase inhibitors. Generated molecules with high similarity to existing inhibitors were distinguished based on a TC value higher than 0.60. The comparison identified 28 generated small molecules of high similarity (Fig. 2b). However, TC-based scoring cannot identify pharmacophore-level similarity to existing inhibitors based on the spatial arrangement of functional groups of similar characteristics, which are crucial for biological response [32].

3.3.2. (B) Similarity of the generated molecules based on ligand-based pharmacophores

PharmaGist program uses ligand-based pharmacophores to screen

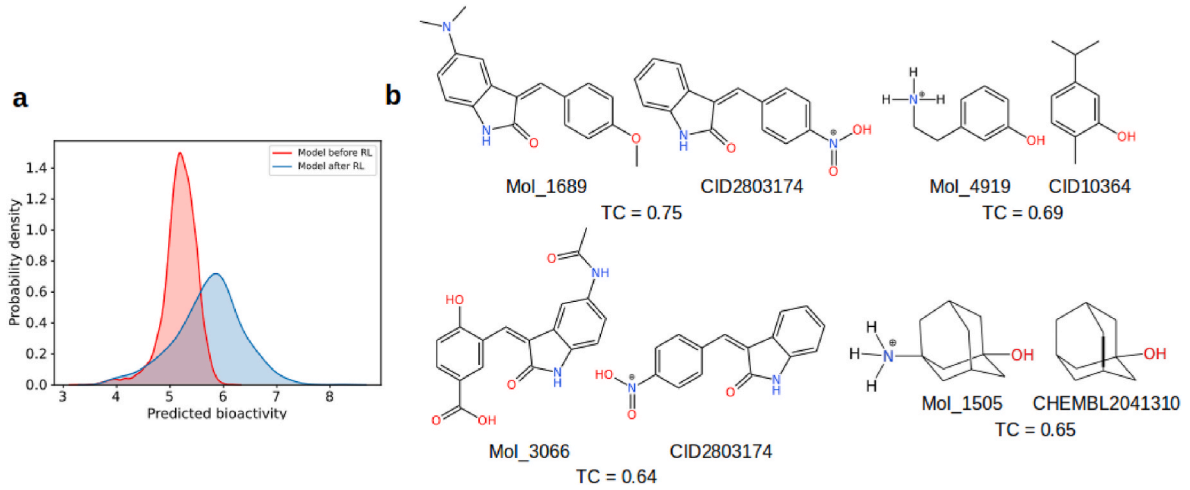

Fig. 2. (**a**) Predicted bioactivity distribution before and after optimization of the generated small molecules specific to *Mtb* chorismate mutase (**b**) examples of generated small molecules with high similarity to existing inhibitors of chorismate mutase. The similarity between the generated molecule and the known inhibitor is provided in terms of the Tanimoto coefficient (TC).

datasets of small molecules and provides a feature overlap score summarizing the functional group-level similarity to the pharmacophores. Molecules with high score can therefore be considered as efficient inhibitors of the target protein, irrespective of their Tanimoto similarity to existing inhibitors. A small molecule was considered as a hit, if the feature overlap score of the molecule with the target pharmacophore was at least half of the maximum feature overlap score [21]. Based on the results (Table 1), the two selected composite pharmacophores (Supplementary Figs. S2–a) could cover 100% of the chorismate mutase-specific generated molecules.

From the pharmacophore-based screening results, the generated small molecules were found to capture the key pharmacophore features of the target binding site. To further confirm the pharmacophore-level similarity of the generated small molecules to the existing inhibitors, two pharmacophore fingerprints (ErGFP and PharmacoPFP) were calculated following a recent study [33] and compared using cosine similarity. The distribution of the cosine similarity values from all pairwise comparisons (Supplementary Figs. S2–b) shows that above 75% of the generated small molecules have high pharmacophore-level similarity (cosine similarity above 0.8) to existing inhibitors. It is important to note that only limited number of small molecules (molecules from the validation dataset) have been tested against chorismate mutase, which is still an ongoing research area. The current comparison of generated small molecules with the validation set through Tanimoto coefficient and the ligand-based pharmacophore analysis, were performed for *in silico* validation of the proposed approach.

3.4. GAT-VAE model identified important binding site residues

The attention coefficients from the GAT-VAE model were analyzed for each residue (node) and its neighborhood in the binding site graph. The biological significance of the latent representation from the GAT-VAE model can be understood by elucidating residue pairs with high attention coefficient (Fig. 3).

Residue pairs with attention coefficient above 0.5 were considered important. For the binding site of *Mtb* chorismate mutase (PDB ID: 2FP2), only 10 of the 244 interactions had attention coefficient (α_{ij}) above 0.50 from the GAT-VAE model. These 10 interactions involved the following 16 binding site residues: Leu50, Ala53, Lys60, Pro66, Ile67, Asp69, Arg72, Val73, Phe98, Ile102, Ala107, Lys117, Arg103, Glu109,

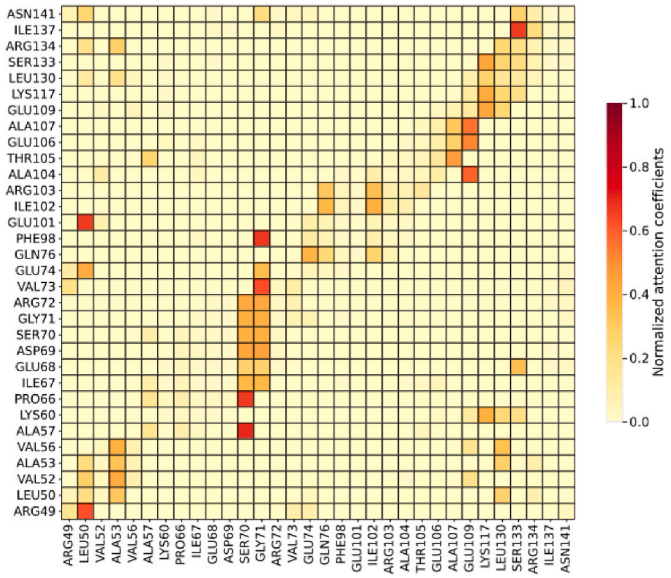

Fig. 3. Attention coefficient heatmap for the binding site residues of *Mtb* chorismate mutase (PDB ID: 2FP2). Darker boxes indicate more importance to the residue pair in the binding site graph.

Table 1
Results from the pharmacophore-based screening of generated small molecules for chorismate mutase (CM).

Protein	Pharmacophore	Hits[a] (%)	Screened count[b]	Not screened count	Screened by the other pharmacophore	Not screened by both pharmacophores
CM validation set	Pharmacophore 1	71.42	32	4	2	2
	Pharmacophore 2	68.57	28	8	6	
CM generated set	Pharmacophore 1	88.49	7483	973	973	0
	Pharmacophore 2	67.41	5701	2755	2755	

[a] Molecules with at least half the maximum overlap score.
[b] Any molecule with a positive overlap score.

Ile137, and Asn141. Based on the interactions of the transition state analogue with these binding site residues in the crystal structure [14], the importance of the interactions in stabilizing the protein-ligand complex was verified. According to the previous literature, Lys60 interacts with Glu109, and both residues coordinate the ether oxygen of the transition state analogue through hydrogen bonding. Both these interactions with the ligand are thought to be crucial for the enzyme's catalytic mechanism [14]. Arg72, being a highly conserved residue in the chorismate mutase subfamily, was found to coordinate the two carboxylate groups of the transition state analogue (Supplementary Fig. S3). Apart from these polar residues, the hydrophobic residues Ile67, Val73 and Ile102 were also observed to form strong van der Waal's interactions with the ligand in the crystal structure. Overall, the residue pairs with higher attention coefficients were found to provide stability to the generated molecules and their role in enzyme activity is also known from previous literature [14]. These residues can influence the molecule generation process, which can be deduced from the complementarity of interactions at the chorismate mutase binding site (Fig. 4).

3.5. Clustering and interaction analysis identified key interactions between binding site residues and the generated small molecules

The set of 8941 generated small molecules specific to chorismate mutase were filtered based on their exposed surface area in the docked complex structure, resulting in a set of 4041 molecules satisfying the chosen criterion (see Supplementary information 2). These molecules were clustered using the Butina clustering method [29]. As most of the clusters were singletons, only clusters with at least 10 molecules were chosen for further analysis, amounting to 13 distinct clusters of generated molecules. Each cluster was found to contain a representative molecule with a minimal scaffold, and other molecules in the cluster with various substituents attached to one or more positions of the minimal scaffold. The topmost cluster (cluster 1) included 78 generated molecules with similar scaffold. The molecules belonging to the top 10 clusters were represented using a TMAP [34] to understand the variation between clusters (Fig. 5). Based on an analysis of the best docking pose of each molecule at the chorismate binding site, the interactions predominant among the molecules within each cluster were identified (Tables S4 and S5).

Among the polar binding site residues, Arg72, Arg134, Gln76, Lys60 and Thr105 contributed to a major percentage of interactions in all 11 clusters. Similarly, among the non-polar binding site residues, Ile137, Leu130, Val56 and Ile102 were found to interact with at least 50% of molecules in each cluster. Histograms of interaction percentages for the polar and non-polar binding site residues for the top 5 clusters are shown in Fig. 6. Interestingly, Arg72, Lys60, Ile137 and Ile102 were also involved in interactions with high attention coefficients identified from the binding site graph using the GAT-VAE model (Fig. 3). This indicates the capability of conditional molecule generation with the GAT-VAE model to focus on the key binding site residues and design molecules with favorable interactions at the binding site.

4. Conclusions

Most of the current generative models are ligand-based (Jin et al., 2018; Balakrishnan et al., 2021; [13,35,36], which limits the applicability of ligand-based generative models against novel target proteins with limited amount of target-specific ligand dataset. In this study, a new structure-based deep learning method was used for designing novel anti-tuberculosis agents. The method utilized the binding site information of *Mtb* chorismate mutase enzyme to condition the molecule generation process. The graph attention model was able to distinguish the key binding site residues and inter-residue interactions through the attention coefficients, which was visualized using the attention coefficient heatmap. An *in silico* validation was performed, where the conditional generative model was found to generate small molecules having high similarity with respect to the existing inhibitors of chorismate mutase. The generated small molecules were also found to preserve the key pharmacophoric features required to efficiently bind to the binding site of the target protein. Analysis of the attention coefficients followed by clustering and interaction analysis identified key binding site residues responsible for substrate binding. The approach presented in this work explains the suitability of conditional molecule generation using the GAT-VAE model, which provides opportunity for designing drug-like molecules against novel target proteins whose structure is known.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data used to train the models is opensource and appropriate reference has been provided. The data generated in the manuscript has been provided in Supplementary material

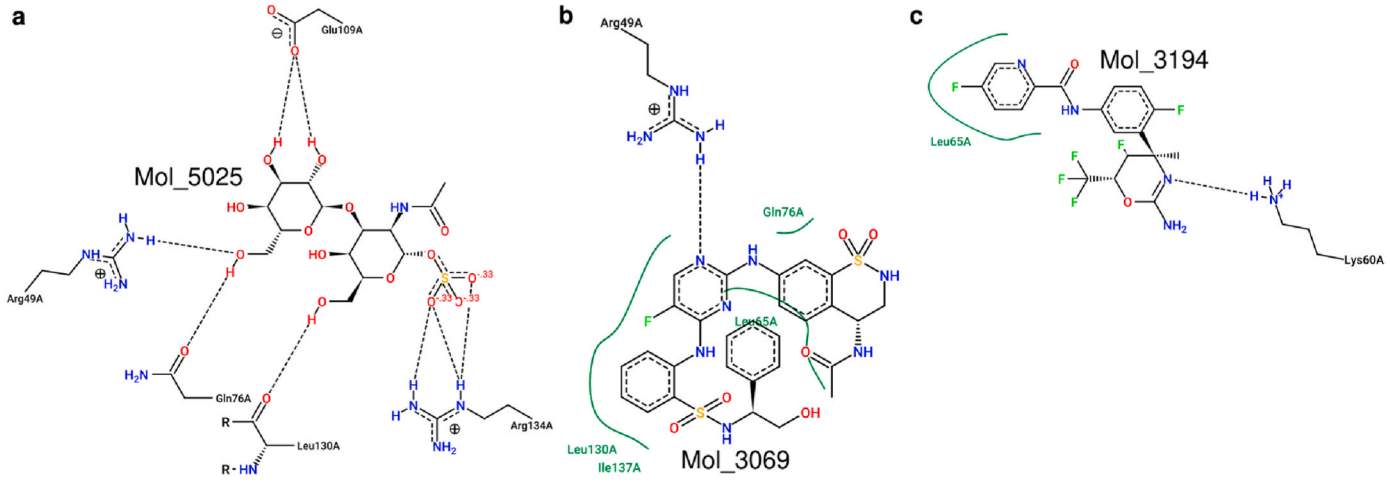

Fig. 4. 2D interaction diagrams of the representative molecules from the top three clusters with binding site residues. Hydrogen bonds are shown as dotted lines. Non-polar hydrogens have been omitted for clarity. The interaction diagrams were prepared using the PoseView tool.

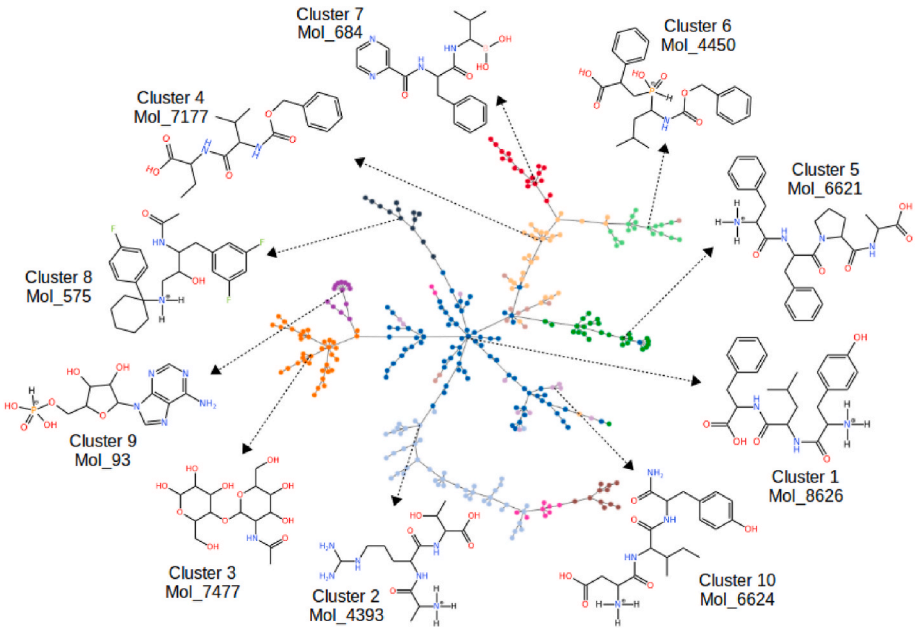

Fig. 5. TMAP of the generated small molecules with the top 10 clusters highlighted in different colors. The representative molecule of each cluster is also shown. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

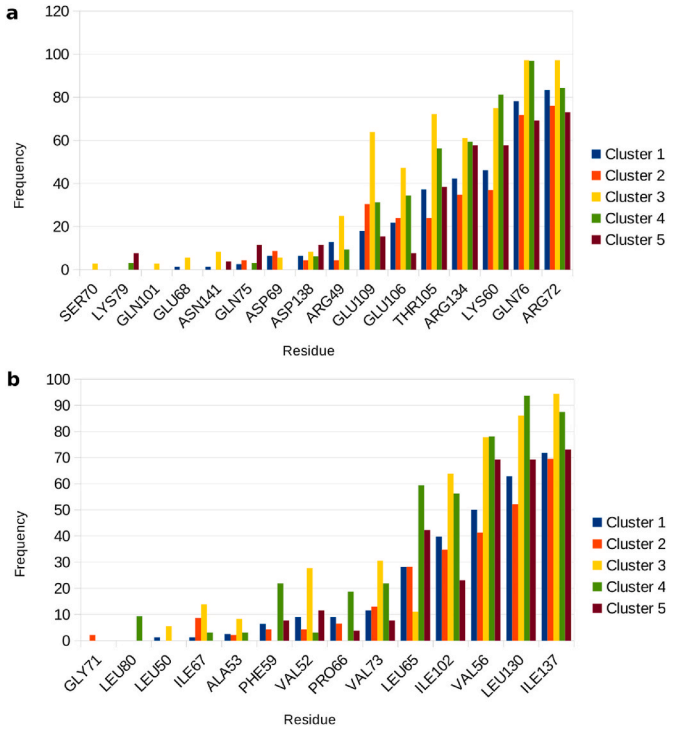

Fig. 6. Histogram of the percentage of interactions observed with (**a**) polar binding site residues and (**b**) non-polar binding site residues for the top 5 clusters of generated small molecules.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmgm.2022.108361.

References

[1] M. Pai, M.A. Behr, D. Dowdy, K. Dheda, M. Divangahi, C.C. Boehme, A. Ginsberg, S. Swaminathan, M. Spigelman, G. Getahun, D. Menzies, M. Raviglione, Tuberculosis. *Nat. Rev. Dis. Primers.* 2 (2016), 16076, https://doi.org/10.1038/nrdp.2016.76.

[2] A. Natarajan, P.M. Beena, A.V. Devnikar, S. Mali, A systemic review on tuberculosis, Indian J. Tubercul. 67 (3) (2020) 295–311, https://doi.org/10.1016/j.ijtb.2020.02.005.

[3] N.I. Paton, L. Borand, J. Benedicto, M.M. Kyi, A.M. Mahmud, M.N. Norazmi, N. Sharma, C. Chuchottaworn, Y. Huang, N. Kaswandani, H.L. Van, G.C.Y. Lui, T. E. Mao, Diagnosis and management of latent tuberculosis infection in Asia: review of current status and challenges, Int. J. Infect. Dis. 87 (2019) 21–29, https://doi.org/10.1016/j.ijid.2019.07.004.

[4] World Health Organization (WHO), Global Tuberculosis Report 2021, World Health Organization, Geneva, 2021.

[5] K.J. Seung, S. Keshavjee, M.L. Rich, Multidrug-resistant tuberculosis and extensively drug-resistant tuberculosis, Cold Spring Harb. Perspect. Med. 5 (9) (2015), a017863, https://doi.org/10.1101/cshperspect.a017863.

[6] M.A. Forrellad, L.I. Klepp, A. Gioffre, J.S. García, H.R. Morbidoni, M.P. Santangelo, A.A. Cataldi, F. Bigi, Virulence factors of the *Mycobacterium tuberculosis* complex, Virulence 4 (1) (2013) 3–66, https://doi.org/10.4161/viru.22329.

[7] J. Sun, P.A. Champion, F. Bigi, Editorial: cellular and molecular mechanisms of *Mycobacterium tuberculosis* virulence, Front. Cell. Infect. Microbiol. 9 (2019) 331, https://doi.org/10.3389/fcimb.2019.00331.

[8] A. Ly, J. Liu, Mycobacterial virulence factors: surface-exposed lipids and secreted proteins, Int. J. Mol. Sci. 21 (11) (2020) 3985, https://doi.org/10.3390/ijms21113985.

[9] M. Khanapur, M. Alvala, M. Prabhakar, K.S. Kumar, R.K. Edwin, P.S.V.K. S. Saranya, R.K. Patel, G. Bulusu, P. Misra, M. Pal, Mycobacterium tuberculosis chorismate mutase: a potential target for TB, Bioorg. Med. Chem. 25 (6) (2017) 1725–1736, https://doi.org/10.1016/j.bmc.2017.02.001.

[10] H.S.I. Chao, G.A. Berchtold, Inhibition of chorismate mutase activity of chorismate mutase-prephenate dehydrogenase from Aerobacter aerogenes, Biochemistry 21 (11) (1982) 2778–2781, https://doi.org/10.1021/bi00540a031.

[11] A.P. Campbell, T.M. Tarasow, W. Massefski, P.E. Wright, D. Hilvert, Binding of a high-energy substrate conformer in antibody catalysis, Proc. Natl. Acad. Sci. U.S.A. 90 (18) (1993) 8663–8667, https://doi.org/10.1073/pnas.90.18.8663.

[12] A. Mandal, D. Hilvert, Charge optimization increases the potency and selectivity of a chorismate mutase inhibitor, J. Am. Chem. Soc. 125 (19) (2003) 5598–5599, https://doi.org/10.1021/ja029447t.

[13] S.R. Krishnan, N. Bung, G. Bulusu, A. Roy, Accelerating *de novo* drug design against novel proteins using deep learning, J. Chem. Inf. Model. 61 (2) (2021) 621–630, https://doi.org/10.1021/acs.jcim.0c01060.

[14] M. Ökvist, R. Dey, S. Sasso, E. Grahn, P. Kast, U. Krengel, 1.6 Å crystal structure of the secreted chorismate mutase from Mycobacterium tuberculosis: novel fold topology revealed, J. Mol. Biol. 357 (5) (2006) 1483–1499, https://doi.org/10.1016/j.jmb.2006.01.069.

[15] R. Zamora-Resendiz, S. Crivelli, Structural Learning of Proteins Using Graph Convolutional Neural Networks, BioRxiv, 2019.

[16] W. Torng, R.B. Altman, Graph convolutional neural networks for predicting drug-target interactions, J. Chem. Inf. Model. 59 (10) (2019) 4131–4149, https://doi.org/10.1021/acs.jcim.9b00628.
[17] J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, H. Jiang, Predicting protein–protein interactions based only on sequences information, Proc. Natl. Acad. Sci. USA 104 (11) (2007) 4337–4341, https://doi.org/10.1073/pnas.0607879104.
[18] Z. Liu, Y. Li, L. Han, J. Li, J. Liu, Z. Zhao, W. Nie, Y. Liu, R. Wang, PDB-wide collection of binding data: current status of the PDBbind database, Bioinformatics 31 (3) (2015) 405–412, https://doi.org/10.1093/bioinformatics/btu626.
[19] J. Desaphy, G. Bret, D. Rognan, E. Kellenberger, sc-PDB: a 3D-database of ligandable binding sites - 10 years on, Nucleic Acids Res. 43 (2015) D399–D404, https://doi.org/10.1093/nar/gku928.
[20] J. Born, M. Manica, A. Oskooei, J. Cadow, M.R. Martínez, Paccmann^{RL}: designing anticancer drugs from transcriptomic data via reinforcement learning, *Proceedings of the International Conference on Research in Computational Molecular Biology,* June 22–25 (2020) 231–233.
[21] S.R. Krishnan, N. Bung, S.R. Vangala, R. Srinivasan, G. Bulusu, A. Roy, *De novo* structure-based drug design using deep learning, J. Chem. Inf. Model. (2021), https://doi.org/10.1021/acs.jcim.1c01319.
[22] N. Sánchez-Cruz, J.L. Medina-Franco, J. Mestres, X. Barril, Extended connectivity interaction features: improving binding affinity prediction through chemical description, Bioinformatics 37 (10) (2020) 1376–1382, https://doi.org/10.1093/bioinformatics/btaa982.
[23] M.J. Hartshorn, M.L. Verdonk, G. Chessari, S.C. Brewerton, W.T.M. Mooij, P. N. Mortenson, Diverse, high-quality test set for the validation of Protein−Ligand docking performance, J. Med. Chem. 50 (4) (2007) 726–741, https://doi.org/10.1021/jm061277y.
[24] A.T. McNutt, P. Francoeur, R. Aggarwal, T. Masuda, R. Meli, M. Ragoza, J. Sunseri, D.R. Koes, Gnina 1.0: molecular docking with deep learning, J. Cheminf. 13 (1) (2021) 43, https://doi.org/10.1186/s13321-021-00522-2.
[25] A.H. Lipkus, A proof of the triangle inequality for the Tanimoto distance, J. Math. Chem. 26 (1999) 263–265, https://doi.org/10.1023/A:1019154432472.
[26] D. Schneidman-Duhovny, O. Dror, Y. Inbar, R. Nussinov, H.J. Wolfson, PharmaGist: a webserver for ligand-based pharmacophore detection, Nucleic Acids Res. 36 (2008) W223–W228, https://doi.org/10.1093/nar/gkn187.
[27] S. Mitternacht, FreeSASA: an open source C library for solvent accessible surface area calculations, F1000Res 5 (2016) 189, https://doi.org/10.12688/f1000research.7931.1.
[28] A. Shrake, A.J. Rupley, Environment and exposure to solvent of protein atoms. Lysozyme and insulin, J. Mol. Biol. 79 (2) (1973) 351–371, https://doi.org/10.1016/0022-2836(73)90011-9.
[29] D. Butina, Unsupervised data base clustering based on daylight's fingerprint and Tanimoto similarity: a fast and automated way to cluster small and large data sets, J. Chem. Inf. Comput. Sci. 39 (4) (1999) 747–750, https://doi.org/10.1021/ci9803381.
[30] N. Brown, M. Fiscato, M.H.S. Segler, A.C. Vaucher, GuacaMol: benchmarking models for de Novo molecular design, J. Chem. Inf. Model. 59 (3) (2019) 1096–1108, https://doi.org/10.1021/acs.jcim.8b00839.
[31] D. Rogers, M. Hahn, Extended-connectivity fingerprints, J. Chem. Inf. Model. 50 (5) (2010) 742–754, https://doi.org/10.1021/ci100050t.
[32] N. Stiefl, I.A. Watson, K. Baumann, A. Zaliani, ErG: 2D pharmacophore descriptions for scaffold hopping, J. Chem. Inf. Model. 46 (1) (2006) 208–220, https://doi.org/10.1021/ci050457y.
[33] W.X. Shen, X. Zeng, F. Zhu, Y. Wang, C. Qin, Y. Tan, Y.Y. Jiang, Y.Z. Chen, Out-of-the-box deep learning prediction of pharmaceutical properties by broadly learned knowledge-based molecular representations, Nat. Mach. Intell. 3 (2021) 334–343, https://doi.org/10.1038/S42256-021-00301-6.
[34] D. Probst, J. Reymond, Visualization of very large high-dimensional data sets as minimum spanning trees, J. Cheminf. 12 (1) (2020) 12, https://doi.org/10.1186/s13321-020-0416-x.
[35] N. Bung, S.R. Krishnan, G. Bulusu, A. Roy, *De novo* design of new chemical entities for SARS-CoV-2 using artificial intelligence, Future Med. Chem. 13 (6) (2021) 575–585, https://doi.org/10.4155/fmc-2020-0262.
[36] N. Bung, S.R. Krishnan, A. Roy, An *in silico* explainable multiparameter optimization approach for *de novo* drug design against proteins from the central nervous system, J. Chem. Inf. Model. 62 (11) (2022) 2685–2695, https://doi.org/10.1021/acs.jcim.2c00462.