OXFORD

Briefings in Bioinformatics, 2025, **26(5)**, bbaf498
https://doi.org/10.1093/bib/bbaf498
Problem Solving Protocol

TCM-navigator, a deep learning-based workflow for generation and evaluation of traditional Chinese medicine-like compounds for drug development

Feiying Chen [1,2,‡], Victor Jun Yu Lim [2,‡], Mingyu Li [1], Hao Fan [2,3,*]

[1]Medicinal Chemistry and Bioinformatics Center, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China
[2]Bioinformatics Institute (BII), Agency for Science, Technology and Research (A*STAR), 30 Biopolis Street, #07-01 Matrix Building, Singapore 138671, Republic of Singapore
[3]Duke-NUS Medical School, 8 College Rd, Singapore 169857, Republic of Singapore

*Corresponding author. Bioinformatics Institute (BII), Agency for Science, Technology and Research (A*STAR), 30 Biopolis Street, #07-01 Matrix Building, Singapore 138671, Republic of Singapore. E-mail: hfan2006@gmail.com
‡Feiying Chen and Victor Jun Yu Lim contributed equally to this work.

Abstract

Traditional Chinese Medicine (TCM) has long been regarded as a valuable resource for modern drug discovery. However, the limited availability of recorded entities and information, the complexity and sparsity of the herb–ingredient–target–disease network, and inconsistencies in data representation hinder the effectiveness of high-throughput screening approaches. While some therapeutically valuable compounds from TCM have been discovered through manual experimental screening, such methods are time-consuming and require substantial human resources. To address these challenges, we developed a data-driven and deep learning–based workflow, TCM-navigator, which enables the *in-silico* generation, quality control, and physics-based evaluation of TCM-like molecules. The generation is done by TCM-Generator, a transfer learning- and Long Short-Term Memory (LSTM)-based chemical language model that generates standardized, hierarchically structured, and high-throughput–friendly datasets of TCM-like molecules. In this study, we generated a target-nonspecific dataset comprising 3.7 million TCM-like molecules, expanding the number of entities in existing TCM datasets by more than 100-fold. The workflow also enables flexible, goal-driven molecule generation customized for specific targets, yielding three target-specific datasets and multiple high-potential target-ligand pairs. The quality control is done by TCM-Identifier, the first quantitative model specifically designed to capture unique characteristics of TCM, using an AttentiveFP framework with message passing neural networks. TCM-Identifier is expected to serve as an essential evaluation and guidance tool for TCM-related drug development. Our workflow bridges cutting-edge data science—including deep learning—with biomedical research to tackle longstanding challenges in target identification and molecular design. Its adaptable framework is also transferable to interdisciplinary innovation beyond drug development.

Keywords: traditional Chinese medicine; TCM-navigator; TCM-generator; TCM-identifier; deep learning; chemical language model

Introduction

Traditional Chinese Medicine (TCM) has been explored, discovered, and validated through extensive practice for thousands of years, encompassing processes such as prevention, treatment, and health regulation [1]. Statistics indicate that approximately one-third of modern medicines are derived from natural herbs [2]. Coupled with the generally low toxicity of natural compounds, TCM holds immense potential and has become an important resource for modern drug discovery. Notably, certain TCM compounds have undergone further exploration and validation at the ingredient and molecular levels, leading to their successful application in clinical treatments. The most well-known example is artemisinin, discovered by Nobel Laureate Tu Youyou, which has been widely adopted as an effective treatment for malaria [3]. In recent years, the modernization of TCM has gathered significant attention. To standardize and unify existing TCM knowledge, researchers have established high-quality TCM-related databases, such as TCMBank [2], ETCM [4], and TCM-ID [5]. The emergence of TCM databases has enabled data-driven approaches that accelerate TCM-related drug discovery, bridging the gap between ancient practices and modern pharmacology [6]. Although these knowledge bases have deepened researchers' understanding and inspired further exploration, TCM remains an underexploited resource in contemporary drug development. This underuse stems from several intrinsic challenges. TCM operates across multiple scales—from macroscopic herbs to microscopic molecules—and its natural origins give rise to a particularly complex herb–ingredient–target–disease network. This complexity results in limited ingredient availability, intricate molecular compositions, and largely unresolved mechanisms of action. These characteristics hinder the systematic and high-throughput utilization of TCM knowledge bases. Moreover,

Received: May 30, 2025. **Revised:** August 17, 2025. **Accepted:** September 3, 2025
© The Author(s) 2025. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

Downloaded from https://academic.oup.com/bib/article/26/5/bbaf498/8266408 by guest on 14 May 2026

manual experimental validation is time-consuming and labor-intensive, further slowing the progress of TCM-related drug discovery and limiting its clinical translation potential.

To address the above issues, we hypothesized that chemical language models, pre-trained with large chemical libraries and fine-tuned with TCM compound databases, can generate novel chemicals that exhibit biological, chemical, or pharmaceutical properties similar to those of original TCM molecules (TCM-like). We then developed TCM-navigator, the first deep learning-based, end-to-end workflow bridging between traditional TCM knowledge and modern drug discovery pipelines. TCM-navigator enables the generation, quality control, and physics-based evaluation of TCM-like compounds. It comprises three core modules: (i) TCM-generator, leverages the existing knowledge in the TCM knowledge base through transfer learning and an LSTM architecture, to expand and refine the original chemical space of the TCM database, allowing for flexible adaptation to target-specific TCM-like compound generation; (ii) TCM-identifier, the first quantitative TCM scoring model, leverages an Attentive FP framework with message passing neural networks (MPNNs), for compound quality assessment; and (iii) a binding affinity prediction module, integrating virtual screening and molecular dynamics (MD) simulations for structure-based evaluation. Beyond the enlarged and refined chemical space, the molecules in our TCM-like datasets exhibit improved synthesizability, drug-likeness, and favorable ADMET profiles compared to original TCM compounds, offering enhanced drug development potential.

To demonstrate the application of our TCM-like compounds, we prepared case study that comprises three important drug targets, including epidermal growth factor receptor (EGFR) for cancer, SARS-CoV-2 Main protease (Mpro) for viral infection, and Sirtuin 1 (SIRT1) for diabetes and metabolic diseases. These three targets are of important clinical applications for the treatment of various conditions, and have been targeted by TCM molecules [7–9]. In our case study, molecules from the TCM-like datasets show better predicted binding potential compared to TCM molecules known to interact with these targets, suggesting promising ligand candidates.

Beyond accelerating TCM-related drug development, more importantly, our workflow addresses a critical challenge in data-driven biology: standardizing and expanding limited datasets. Despite the proliferation of high-throughput technologies and deep learning-driven pipelines, small-to-medium-sized datasets in domains such as medicinal chemistry and synthetic biology frequently suffer from inadequate standardization, limited scalability, and incomplete data, rendering manual processing inefficient and error-prone. Our workflow tackles this by enabling *in silico* molecular generation, quality control, and physics-based evaluation, transforming sparse data into enriched, standardized resources for experimental exploration. This integrative approach unifies diverse biomedical domains, enabling scalable, efficient, and streamlined research.

Results and Discussions

We proposed an end-to-end workflow — TCM-navigator — for TCM-like compound generation and evaluation, which integrates multiple data-driven and deep learning-based components. Our workflow is organized into three core design modules: two innovative components, including TCM-generator (Fig. 1A) and TCM-iidentifier (Fig. 1B), and physics-based evaluation (Fig. 1C). TCM-generator is a large-scale molecular generation module that employs a stagewise learning strategy consisting of three phases: pre-training, fine-tuning, and state-tuning. This phased approach empowers TCM-generator to progressively encode the chemical signatures of different compound libraries, achieving efficient feature extraction across stages [10]. TCM-Identifier is an AttentiveFP-based molecular distinguisher developed to evaluate the TCM-likeness of generated compounds [11]. As the first deep-learning model for TCM-like property assessment, TCM-identifier systematically captures global molecular information while preserving high interpretability. In the physics-based evaluation module, we employ two key criteria: (i) static binding affinity—molecular docking is employed to predict the binding conformations and evaluate the structural and chemical complementarity of drug-like molecules with target proteins; (ii) dynamic binding stability—MD simulations are conducted to assess the stability of shortlisted TCM-like compounds within the binding pocket. Our workflow establishes an innovative end-to-end paradigm for TCM-related drug development (Fig. 1).

TCM-Generator: Chemical Language Model for generating TCM-like molecules

Generation of the TCM-like molecules. To enable the chemical language model to comprehensively learn the syntax and feature distribution of SMILES sequences, we pre-trained it on the ZINC15-filtered dataset (see Supplementary Methods 'Datasets'). We selected LSTM as our generative model for its generalizability, flexibility, and ability to capture TCM-specific local features. While transformer-based models like GPTs show promise in ligand discovery, they require extensive training data and prioritize global molecular topology during the generation process [12], de-emphasizing pharmacophoric features central to our approach. All parameters in the embedding layer and LSTM cell were optimized via gradient descent optimization during the current pre-training phase (Supplementary Fig. S1B) [13]. The final pre-trained model demonstrated stable performance on the test set split from the ZINC15-filtered dataset, achieving a loss <0.75, accuracy >0.75, and perplexity <2.5, indicating effective learning of chemical representations. Subsequently, we utilized the pre-trained model as the point of initialization and fine-tuned it using the TCM dataset. The fine-tuning process selectively updated the fully connected layer parameters within the LSTM architecture. Model performance converged on the test set, achieving a loss <0.85, accuracy >0.68, and perplexity <0.25. Using the above fine-tuned model (TCM-Generator), we generated a large-scale TCM-like dataset containing 6 million molecular entries, expanding the number of entities in the TCM dataset by over 100-fold. Thereafter, we systematically characterized the dataset's chemical space and TCM-specific features to assess its utility in drug development.

Evaluation of chemical space expansion capability. To investigate whether TCM-Generator can expand the chemical space of TCM rather than merely 'memorizing' existing molecules or interpolating within the known chemical space, we conducted the extrapolative generation test [14]. We first quantified the chemical space of TCM by computing the Morgan Fingerprint representations for each molecule (see Supplementary Methods 'Evaluation of Chemical Space'). Principal component analysis (PCA) was then applied to the fingerprint data to extract the top three principal components (PC1, PC2, and PC3), and each molecule was projected onto a three-dimensional coordinate system formed by these principal axes (Supplementary Fig. S2A- C). To visualize the chemical space distribution of the TCM dataset, we applied K-means clustering, dividing the dataset into five distinct clusters (Fig. 2A, Supplementary Fig. S2D). Next, we excluded molecules belonging to cluster 4, which comprises 22.23% of the molecules and is

Downloaded from https://academic.oup.com/bib/article/26/5/bbaf498/8266408 by guest on 14 May 2026

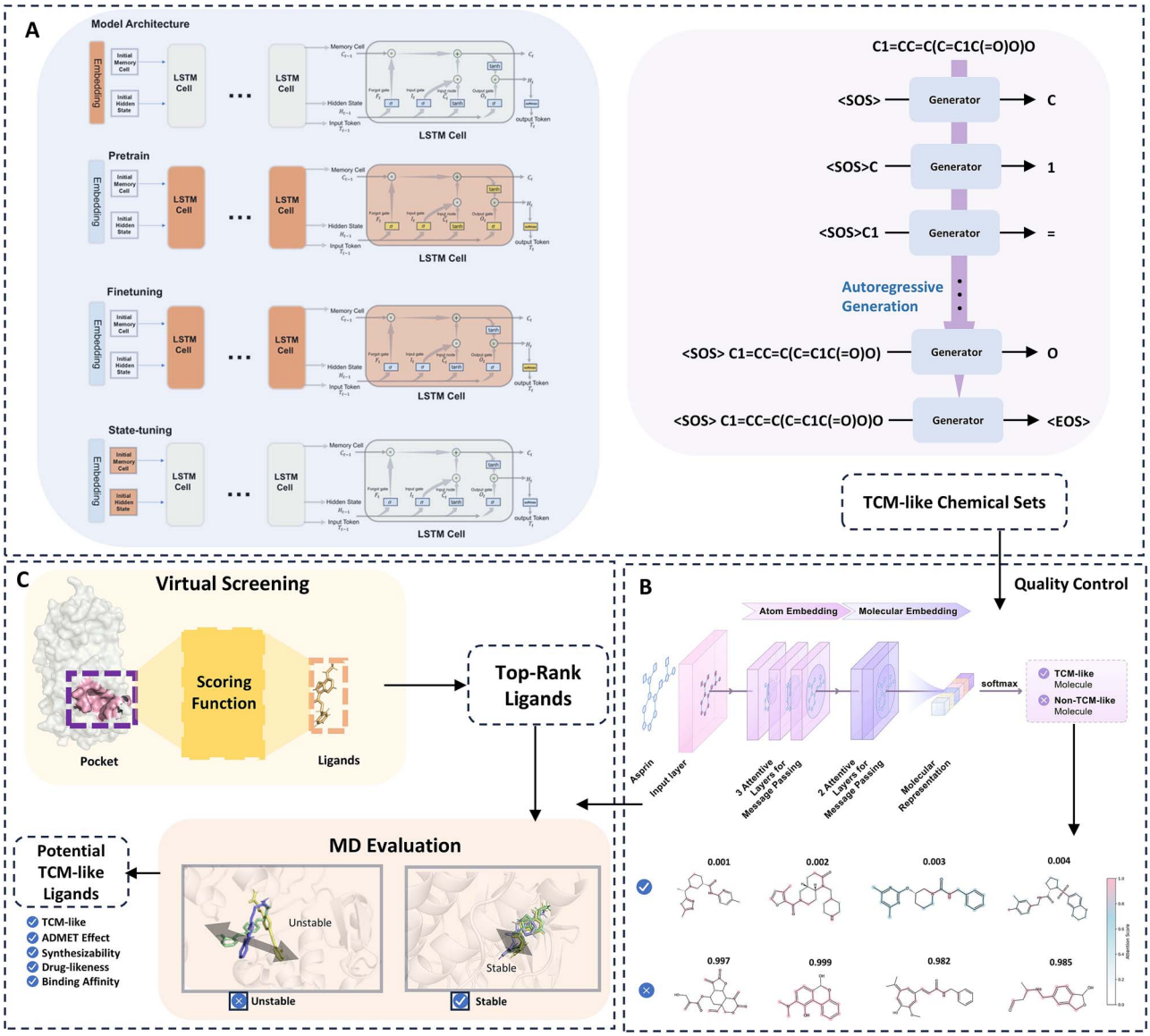

Figure 1. Overview of the TCM navigator workflow. TCM navigator facilitates TCM-related drug development through a multi-component approach, providing a comprehensive workflow from *de novo* generation to quality control and physics-based evaluation. This workflow consists of (A) TCM-generator, (B) TCM-identifier, and (C) the physics-based evaluator. We developed a transfer learning method based on the LSTM architecture for molecular generation (A), designed a model incorporating message passing and attention mechanisms to evaluate TCM-like properties (B), and integrated molecular docking with MD simulations to validate ligand binding (C).

located at the edge of the chemical space defined by the TCM dataset (Fig. 2A), generating a reduced dataset named as TCM Subset (Fig. 2B). Using the pre-trained model as the initialization point, we fine-tuned the model for 200 epochs using TCM Subset instead of the full TCM dataset. The parameters from epoch 200 were selected as the final generative model parameters, and 100 000 molecules were generated using the subset-trained model (Fig. 2C).

Interestingly, when we projected the generated molecules onto the PC1, PC2, and PC3 axes, we observed a denser molecular distribution within the chemical space, with significantly increased number of molecules occupying the same space. Notably, the region originally occupied by cluster 4 in the TCM dataset, which had been intentionally removed in the TCM Subset, was now repopulated with a high-density molecular distribution (Fig. 2C). These results confirm that the model can effectively amplify molecular density within known chemical domains while simultaneously expanding into unexplored regions of chemical space.

The capacity to systematically explore uncharted chemical space represents a cornerstone of *de novo* design, transcending the limitations of existing knowledge bases to unlock novel therapeutic candidates with high clinical potential. We demonstrated that TCM-generator does not rely on memorization but instead learns underlying patterns through the extrapolation test. The capability to generate novel structures beyond the given TCM dataset is likely attributed to the extensive pretraining on the ZINC15-filtered dataset, which exposed the model to a wide range of molecular motifs and representations, enhancing its generative diversity.

Examination of the generated TCM-like dataset. For each molecule in the generated TCM-like dataset, we measured its five highest chemical similarity scores to compounds in the TCM dataset, allowing for a smoother and more representative distribution of top similarity scores in the subsequent analysis. Such

Downloaded from https://academic.oup.com/bib/article/26/5/bbaf498/8266408 by guest on 14 May 2026

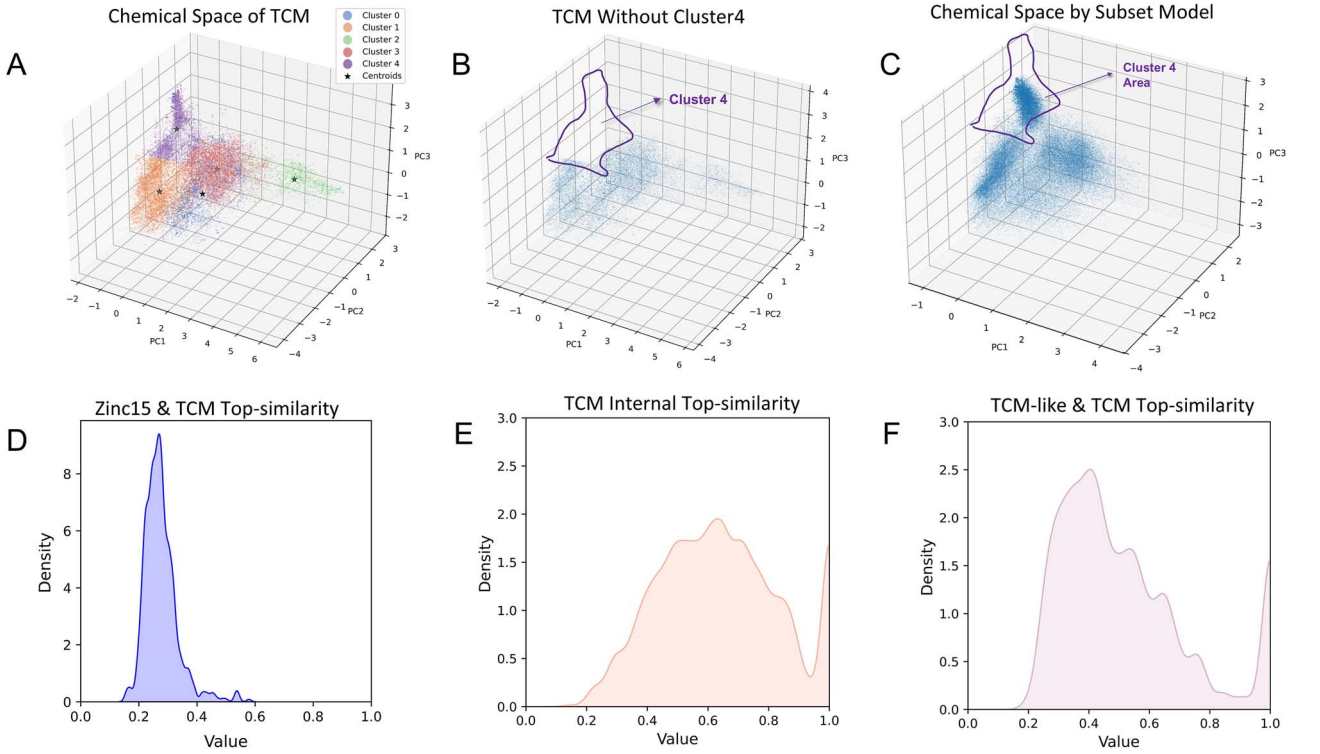

Figure 2. Evaluation of the fine-tuned Model's chemical space expansion and performance. Visualization of the chemical space of TCM (A) based on Morgan fingerprints and a reduced subset (B) by removing cluster 4. The chemical space generated by the subset model fills the missing region of cluster 4 (C). (D) Calculated Top5-similarity distributions between ZINC15 and TCM. (E) TCM internal Top5-similarity, a subset sampled from the TCM dataset calculated against the remaining TCM compounds. (F) Top5-similarity between TCM-like and TCM.

top5-similarity distributions were also calculated between the ZINC15-filtered and TCM datasets (Fig. 2D), as well as within the TCM dataset (Fig. 2E), as controls. Our analysis revealed that the top5-similarity distribution between the TCM-like and TCM datasets (Fig. 2F) is significantly skewed toward higher values than that between the ZINC15-filtered and TCM datasets, and resembles the intra-set similarity of the TCM dataset. These results suggest that fine-tuning had a strong influence on the model's generation behavior, and the generated dataset (TCM-like) likely captures domain-specific features present in the fine-tuning set (TCM). Notably, some TCM compounds are large or complex molecules that exhibit limited potential for small-molecule drug development. Such molecules are also present in the TCM-like dataset. To optimize the TCM-like dataset's drug potential and synthetic feasibility, we applied selection criteria (a) (Supplementary Table S1), resulting in the TCM-like-filtered dataset of 3.7 million entries.

The TCM-like-filtered dataset exhibits improved ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) profiles and enhanced drug-like physicochemical properties (Fig. 3A) [15, 16] (see Supplementary Methods 'ADMET and Chemical Properties'). Notably, the TCM-like-filtered dataset outperforms the TCM dataset in several toxicity-related prediction metrics, including ClinTox, DILI, hERG, and LD50_Zhu, suggesting that TCM-like compounds may show reduced toxicity. In terms of drug-like properties, the compounds in the TCM-like-filtered dataset exhibit a higher distribution of Lipinski scores and a more favorable QED profile (Fig. 3B). Additionally, they display lower SA (synthetic accessibility) scores, indicating improved synthetic feasibility. To further explore the global chemical space shifts across the TCM, TCM-like, and TCM-like filtered datasets from a unified perspective, we performed clustering on all molecules from these datasets, and identified 80 clusters with each cluster representing a group of molecules sharing similar ADMET-related features. The centroids of the 80 clusters were projected into a 2D space using PCA, based on the high-dimensional feature space (Fig. 3C). Additionally, we calculated the correlation between each chemical property and the first two principal components (PC1 and PC2) (Supplementary Fig. S3). The results showed that PC1 and PC2 were most strongly correlated with Lipinski score, molecular weight, logP, TPSA, and solubility, followed by QED, synthetic accessibility (SA), and toxicity-related metrics (hERG and LD50_zhu). These findings indicate that such properties are key variables in distinguishing molecular distributions and cluster structures. Comparing both the TCM-like dataset and the TCM-like filtered dataset to the TCM dataset, the proportional distribution of some clusters changed notably (Fig. 3C), while the number of molecules in most clusters increased significantly. These observations indicate that while the TCM-Generator substantially expands the number of molecules based on the TCM dataset, it also induces a noticeable shift in the chemical space.

Furthermore, we investigated the clusters whose proportions increased or decreased significantly, as such disproportionate changes suggest a shift in the chemical space. We extracted clusters that exhibited a disproportionately high increase (Fig. 3D) or decrease (Fig. 3E) in the TCM-like filtered dataset, along with representative molecules from each, for further analysis. We found that molecules in the significantly expanded clusters displayed distinct structural features—for example, Cluster 43 was enriched with conjugated alkynes, and Cluster 45 with epoxy groups. These expanded clusters were also enriched with molecules of lower

Downloaded from https://academic.oup.com/bib/article/26/5/bbaf498/8266408 by guest on 14 May 2026

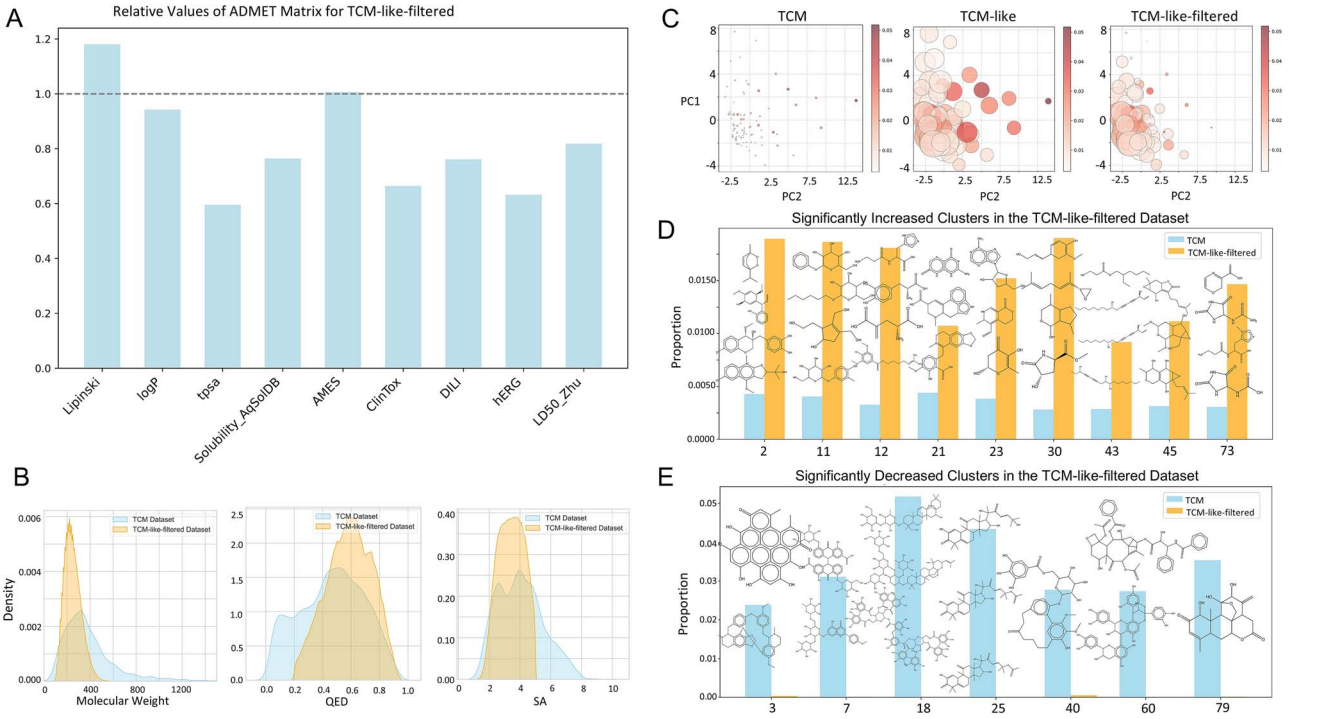

Figure 3. Systematic evaluation of the TCM-like-filtered dataset. (A) Relative ADMET values of the TCM-like-filtered dataset. (B) Distribution of molecular weight, QED, and SA between the TCM and the TCM-like-filtered. (C) Chemical space analysis using multi-Center clustering and PCA for TCM, TCM-like, and TCM-like-filtered datasets based on molecular properties. Molecules were clustered into 80 groups, with each circle representing a cluster. The color intensity (from light to deep red) represents the proportion of TCM molecules in each cluster, while the circle size represents the relative proportions of molecules from three datasets in each cluster. (D) Significantly increased clusters and (E) significantly decreased clusters in the TCM-like-filtered dataset compared to the TCM dataset, along with representative molecular examples from each cluster.

molecular weight and higher drug-likeness (Fig. 3D). In contrast, the reduced clusters primarily consisted of molecules with higher molecular weights and more complex ring systems (Fig. 3E). These findings provide cluster-level evidence that the TCM-Generator leads to both an expansion and refinement of the chemical space, which also in turn explains the improved synthetic accessibility, reduced toxicity, and enhanced drug-likeness observed in the TCM-like filtered dataset.

TCM-identifier: Measurement of TCM-likeness

Retaining TCM properties is essential for TCM-like compound generation. However, these properties are complex and abstract, encompassing multiple physicochemical and biological features [17]. To efficiently quantify TCM-likeness, a standardized end-to-end model is required. To enhance the generalization ability of our TCM-Identifier and effectively capture global and geometric molecular features, we adopted AttentiveFP architecture [11], which integrates attention mechanisms with message-passing neural networks (MPNN) and iterative atom- and molecular-level representations (Fig. 1B). We randomly sampled molecules from the ZINC15-filtered dataset as negative samples and used molecules from the TCM dataset as positive samples to construct the full dataset. We then split 10% of the data as the test set. Due to the class imbalance—negative samples being approximately five times more than positive samples—we applied a weighted learning rate strategy during training. After sufficient training, the TCM-identifier achieved an AUPR (area under the precision-recall curve) of 0.978 on the test set for the task of identifying TCM-like molecules (Supplementary Fig. S4). Additionally, attention score visualization provides interpretability into the model's decision-making process.

We first used the TCM-identifier to calculate the TCM-like score for each molecule in the ZINC15-filtered dataset (sampled), the TCM dataset, the natural product dataset obtained from COCONUT [18], and the TCM-like dataset. We then fitted and compared the TCM-like score distributions of these datasets (Fig. 4A–D). The ZINC15-filtered dataset had over 95% of molecules with TCM-like scores in the 0–0.05 range, whereas more than 90% of molecules in the TCM dataset scored between 0.8 and 1. The natural product dataset displayed a bimodal distribution, with peaks at 0–0.05 (>16%) and 0.95–1 (>48%), reflecting its mixed composition of both TCM/TCM-like and TCM-dislike compounds. The TCM-like dataset closely resembled the TCM dataset, with over 90% of molecules scoring between 0.8 and 1, confirming its strong TCM-like characteristics. These distributions underscore the capability of TCM-Identifier to capture structural features that are unique in TCM but not universally present in all natural products. We hypothesize that, since TCM compounds are primarily used for therapeutic purposes, TCM and TCM-like molecules inherently possess safety- and efficacy-related characteristics that are often lacking in non-medicinal natural products.

Recently, transformer-based models like ChemBERTa [19, 20] and MolGPT [21] have demonstrated to be able to successfully generate novel molecules. For comparison with our TCM-Generator, a single-layer transformer-decoder model was trained using similar parameters to those in TCM-Generator with the same training data. An in-house transformer model was employed in preference to available pre-trained options, as it allows us to compare models with comparable parameters and

Downloaded from https://academic.oup.com/bib/article/26/5/bbaf498/8266408 by guest on 14 May 2026

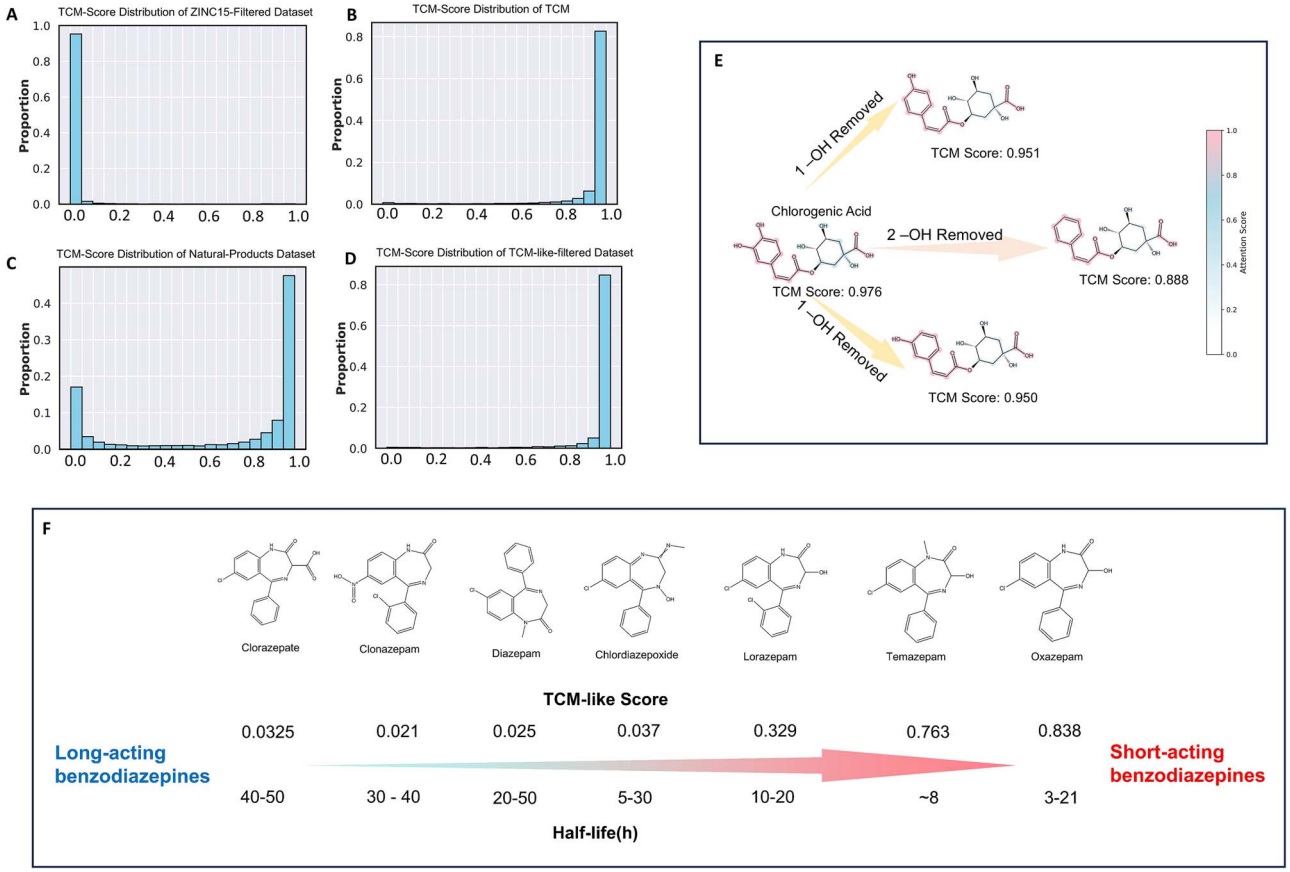

Figure 4. TCM-like score measurement with TCM identifier. Distribution of TCM-like scores in the (A) ZINC15-filtered dataset, (B) TCM dataset, (C) natural products dataset, and (D) TCM-like-filtered dataset. (E) TCM Identifier's attention primarily focuses on the phenol, ester bonds, and carboxyl functional group features. Removal of the phenolic hydroxyl group results in a decrease in the molecule's acidity, functional changes, and a decrease in the TCM-like score. (F) TCM-like scores of common long-acting, medium-acting, and short-acting benzodiazepine drugs and their half-life.

training. After applying selection criteria (a) (Supplementary Table S1), 389,574 unique compounds were generated with the transformer-based model for comparison. The ADMET properties and TCM-like scores were computed for compounds generated by the transformer-based model and compared to the TCM-like filtered dataset from TCM-Generator (Supplementary Fig. S5). The dataset generated by the transformer-based model exhibits ADMET properties similar to those of the compounds generated by TCM-Generator (Supplementary Fig. S5A). The distribution of TCM-like scores was also comparable to that of the TCM-Generator compounds, with more than 90% of the compounds having TCM-like scores above 0.8 (Supplementary Fig. S5B). Interestingly, TCM-like compounds generated by TCM-Generator have a slightly higher proportion of scores above 0.95 than those generated by the transformer-based model. This suggests that TCM-Generator has a more nuanced understanding of TCM-like properties, but this may also be due to differences in the number of compounds generated. Overall, a single-layer transformer-based model does not offer any significant advantage over the TCM-Generator for the compounds generated, although a more complex model may offer better performance.

We further investigated the interpretability of the TCM-Identifier using a representative TCM molecule—chlorogenic acid (TCM-like score: 0.976), a phenolic acid-type TCM ingredient [22]. Phenolic acids, which are common in TCM, derive their acidity from phenol hydroxyl and carboxyl groups. Our TCM-Identifier assigned higher attention scores to the catechol and carboxyl groups—key acidic fragments—compared to the cyclohexane core (Fig. 4E), demonstrating its ability to identify critical chemical moieties for classification. Moreover, the model showed sensitivity to subtle structural variations, as evidenced by the decrease in TCM-like scores when the phenol hydroxyl groups were removed.

Interestingly, we observed a potential link between TCM-like scores and drug half-lives in another case study, supporting our initial hypothesis that TCM-related properties may correlate with various other physicochemical and biological characteristics. Specifically, we illustrate this connection using benzodiazepines—a widely used class of synthetic drugs [23]. Notably, despite their structural similarities, benzodiazepines exhibit distinct TCM-like scores that positively correlate with their half-lives, a key indicator of drug excretion. For example, Temazepam and Oxazepam, which have higher TCM-like scores (0.763 and 0.838), exhibit significantly shorter half-lives (~8 and 3–21 hours, respectively), compared to Clorazepate, which has a lower TCM-like score (0.033) and a much longer half-life (40–50 hours) (Fig. 4F). Interestingly, although Oxazepam and Clorazepate differ only slightly in structure—by a hydroxyl and a carboxyl group adjacent to the amide—this minor variation results in notable differences in both their TCM-like scores and pharmacokinetics. This case highlights the TCM-like score's sensitivity to subtle structural features and its potential predictive power for key pharmacokinetic properties. We propose that this correlation arises from the fact that TCM-like scores are grounded in natural compound characteristics, which are

Downloaded from https://academic.oup.com/bib/article/26/5/bbaf498/8266408 by guest on 14 May 2026

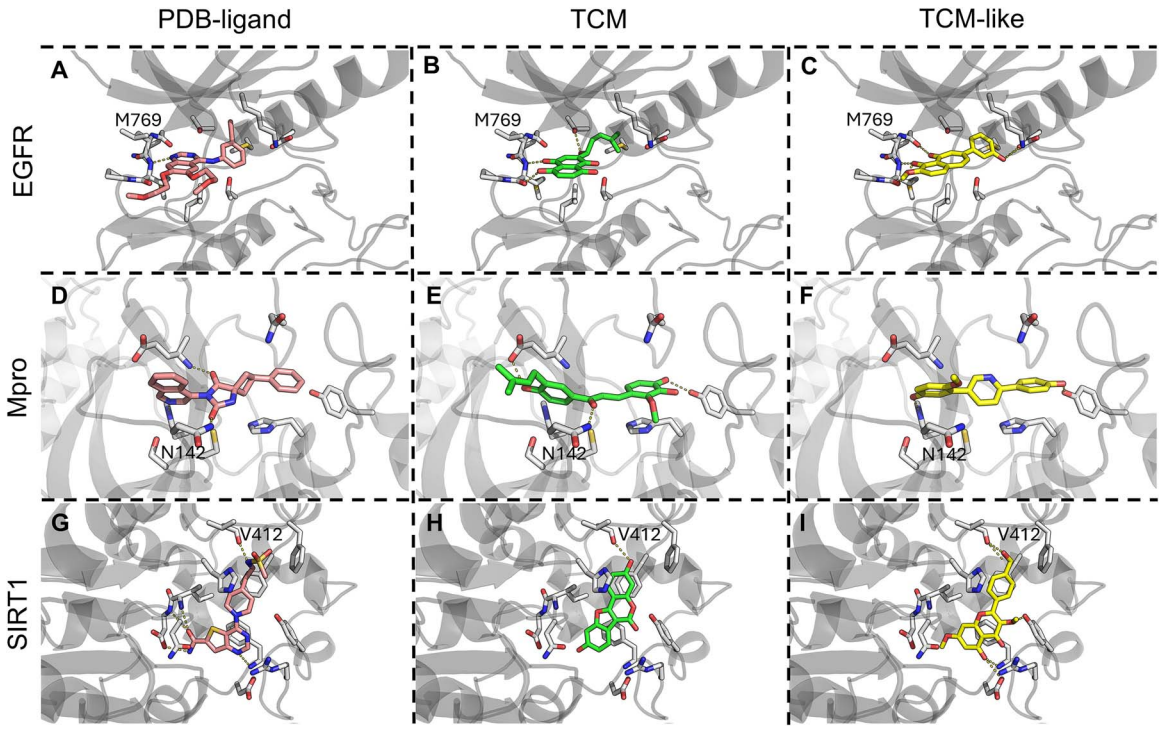

Figure 5. Selected docking poses of PDB-ligand, TCM, and TCM-like compounds for EGFR, Mpro and SIRT1. (A) PDB ligand of EGFR, erlotinib, (B) docking pose of EGFR TCM, Shikonin, and (C) EGFR-TCML1. (D) PDB ligand of Mpro, (E) docking pose of Mpro TCM, and (F) Mpro-TCML1. (G) PDB ligand of SIRT1, (H) docking pose of SIRT1 TCM, and (I) SIRT1-TCML1.

generally associated with faster metabolism and excretion, making favorable TCM-like scores indicative of shorter half-lives.

Generation of target-specific TCM-like datasets

Three targets, EGFR, Mpro, and SIRT1 were selected to demonstrate our model's ability to generate target-specific TCM-like datasets. They were chosen based on their clinical relevance, the availability of crystal structures with reported binding pocket for docking and MD simulations, and the presence of verified TCM molecules for state-turning. We collected 20, 22, and 24 experimentally validated TCM ligands for EGFR, Mpro, and SIRT1, respectively. The relatively small number of TCM ligands for these targets made them ideal candidates for expanding their target-specific TCM-like molecules chemical space. To address the issue of limited ligand data, we preserved TCM-Generator's core framework and employed a state-tuning strategy, by only updating the initial hidden states and memory cells during training (see Supplementary Methods 'TCM-Generator', Supplementary Fig. S1). To ensure the generated compounds possess favorable drug-like potential and binding affinity to the target, we applied selection criteria (b) (Supplementary Table S1), yielding 6179, 4528, and 6179 molecules for the EGFR, Mpro, and SIRT1, respectively. All datasets exhibited favorable ADMET/physicochemical profiles, with higher drug potential and lower synthetic complexity compared to the TCM dataset (Supplementary Fig. S6).

The TCM-like scores of these three datasets displayed a unimodal distribution, with over 75% of molecules scoring between 0.8 and 1, demonstrating TCM-like characteristics (Supplementary Fig. S4B–E., Supplementary Table S1). These distributions exhibited slight deviations from the TCM dataset, due to the application of selection criteria (b). These results demonstrate that state-tuning effectively preserves the knowledge embedded in the TCM-like compound distribution, yielding negligible differences in ADMET/physicochemical profiles (Supplementary Fig. S6). While preserving the ADMET and physicochemical profiles characteristic of TCM-like compounds, the state-tuning process effectively aligned the generation distribution of the TCM-Generator with the SMILES sequences of target-specific TCM datasets. Specifically, after state-tuning, the next-token prediction accuracy of the TCM-generator improved significantly compared to the model trained solely on the general TCM dataset, increasing from 0.742 to 0.817, 0.748 to 0.800, and 0.690 to 0.735 across different targets. TCM-Generator's multi-stage training strategy enables efficient state- tuning with low computational cost and can be universally adapted to any target protein with reported TCM ligands, while consistently preserving ADMET and physicochemical property distributions that remain largely unaffected by variations in the training set composition.

Physics-based binding affinity evaluation of TCM-like compounds

The target-specific TCM and TCM-like datasets for EGFR, Mpro, and SIRT were docked to their respective PDB structures, and the docking scores for the top 500 TCM-like compounds for each protein were evaluated. For all three proteins, the top 500 TCM-like compounds showed similar or improved docking scores compared to their TCM (Supplementary Fig. S7). This hints at the capability of the TCM-Generator to generate compounds that can potentially bind comparable or better than TCM. For each protein, one TCM and one TCM-like compound, alongside the crystal ligands (Fig. 5), were selected for further analysis and MD simulations.

Analysis of docking results for selected compounds. The EGFR reference ligand, Erlotinib (Fig. 5A), binds to the ATP pocket of EGFR kinase domain, forming a hydrogen bond with the backbone of

Downloaded from https://academic.oup.com/bib/article/26/5/bbaf498/8266408 by guest on 14 May 2026

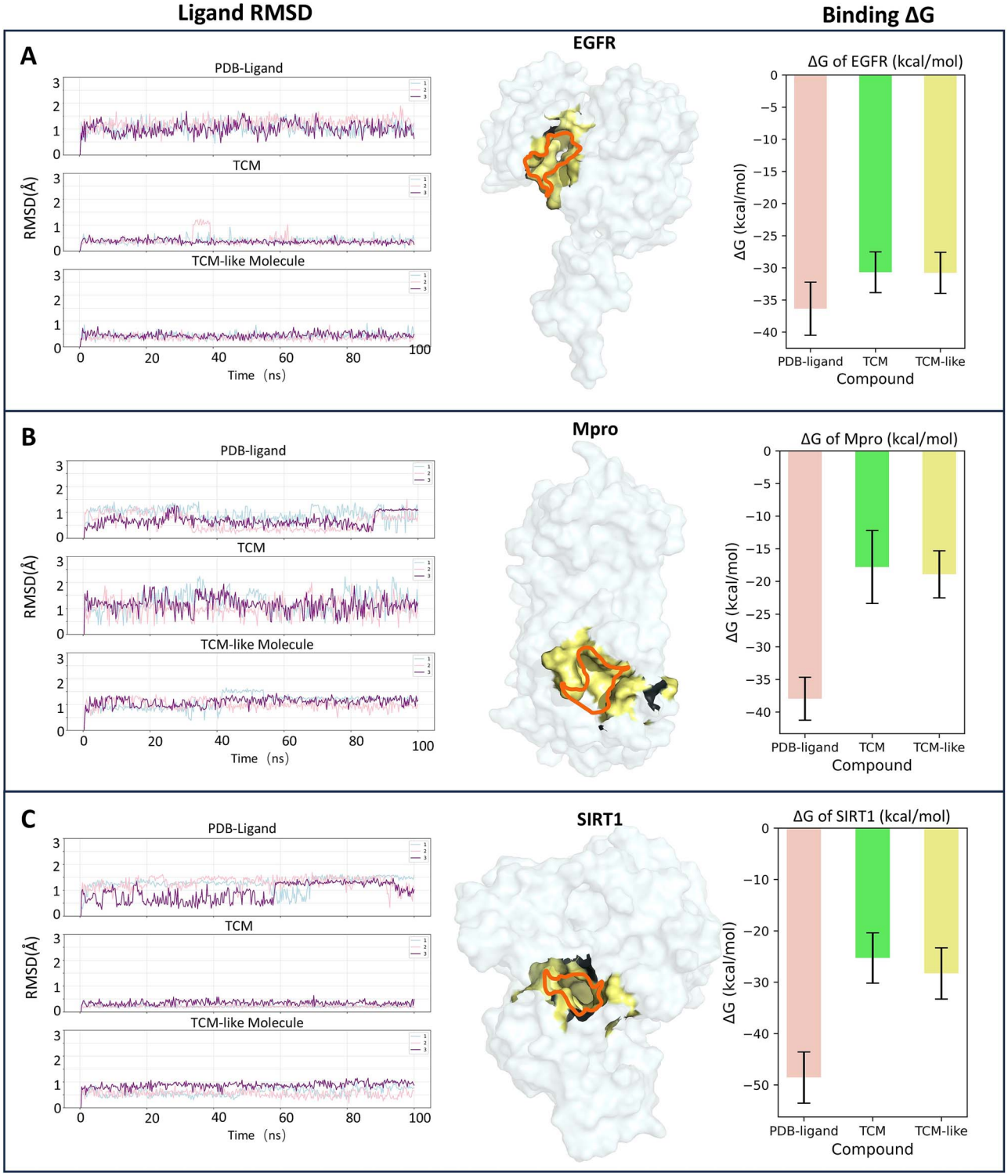

Figure 6. MD-based evaluation of ligand binding affinity. For EGFR (A), Mpro (B), and SIRT1 (C), the RMSD values of the PDB ligand, TCM molecule, and TCM-like molecule are shown on the left, with the lines representing three independent replicates. The middle panel shows the surface representation of each protein, with the targeted binding pocket highlighted. On the right are the MMGBSA binding free energies of the selected compounds for each target.

M769. The selected TCM (Fig. 5B) docks to EGFR with a docking score of −7.68 kcal/mol, while the TCM-like compound (Fig. 5C) has a docking score of −9.21 kcal/mol. The selected TCM and TCM-like compounds for EGFR are both heterocyclic compounds that form hydrogen bonds with the backbone of M769, similar to the PDB ligand. For Mpro, the selected TCM and TCM-like compounds (Fig. 5E, F) show docking scores of −6.37 kcal/mol and −7.84 kcal/mol, respectively. Although the selected TCM-like compound does not form a hydrogen bond with N142 as the PDB ligand and the selected TCM compound do, its docking score is not compromised, probably due to its lower number of rotatable bonds of three compared to the selected TCM with six rotatable bonds. For SIRT1, the selected TCM and TCM-like compounds (Fig. 5H, I) show docking scores of −8.33 kcal/mol

Downloaded from https://academic.oup.com/bib/article/26/5/bbaf498/8266408 by guest on 14 May 2026

and − 8.67 kcal/mol, respectively. All three selected compounds of SIRT1 form a hydrogen bond with the backbone of V412. These examples of TCM-like compounds illustrate the comparable docking conformation of the generated compounds compared to the corresponding PDB ligands and TCM compounds.

Analysis of MD simulations. The MD simulations of EGFR PDB ligand, Erlotinib, show higher RMSD values (Fig. 6A, average RMSD: 1.09 Å) compared to the selected TCM and TCM-like compounds (average RMSD: 0.37 Å and 0.40 Å, respectively). However, Erlotinib has the best average MMGBSA ΔG of −36.35 kcal/mol, as expected from the PDB ligand. The TCM-like compound exhibits a similar average binding ΔG of −30.78 kcal/mol to that of the TCM compound (−30.68 kcal/mol). Furthermore, analysis of hydrogen bonding with the backbone of M769 revealed that our TCM-like compound forms a stable interaction in 67% of the frames, compared to 46% for the TCM molecule and 17% for Erlotinib. For Mpro (Fig. 6B), the average RMSD of the PDB ligands, TCM, and TCM-like compounds are 0.74 Å, 1.12 Å, and 1.07 Å, respectively, and the average MMGBSA binding ΔG are −37.95 kcal/mol, −17.80 kcal/mol, and − 18.91 kcal/mol, respectively. Similar to EGFR, the SIRT1 PDB ligand also shows a higher average RMSD of 1.16 Å compared to TCM (0.26 Å) and TCM-like compound (0.68 Å) (Fig. 6C). Nevertheless, the average MMGBSA binding ΔG of the PDB ligand is the most favourable (−48.57 kcal/mol), outperforming the TCM (−25.27 kcal/mol) and the TCM-like compound (−28.29 kcal/mol). In all three case studies, the TCM-like compounds demonstrate performance comparable to their corresponding TCM compounds though not surpassing the PDB ligands, which are typically highly optimized for binding affinity. This highlights the ability of our methods to generate TCM-like compounds with *in-silico* performance on par with TCM compounds.

Conclusions

In this study, we established TCM-Navigator, a data-driven and deep learning-based workflow for TCM-related drug development consisting of TCM-Generator, TCM-Identifier, and a physics-based evaluator. To the best of our knowledge, this is the first work to generate TCM-like datasets and quantitatively define TCM-likeness. The TCM-Generator used a stagewise training strategy consisting of pre-training, fine-tuning, and state-tuning, enabling the efficient utilization of the limited TCM knowledge. The generated TCM-like dataset meets modern drug design requirements with improved ADMET properties over TCM. The TCM-Identifier leveraged MPNN and self-attention mechanisms to accurately capture molecular features from atomic-level details to global characteristics, providing a reliable metric and valuable guidance for studies involving TCM-like molecules. Lastly, our physics-based approach employs docking and MD simulations to evaluate the *in-silico* binding of generated compounds.

Our approach bridges TCM with contemporary drug discovery, combining historical bioactivity with modern physicochemical standards.

Methods

The detailed description of the methods can be found in Supplementary Methods where the following were described in greater details: 'Datasets', 'TCM-Generator', 'Compounds generation with TCM-Generator', 'Evaluation of Chemical Space', 'ADMET and Chemical Properties', 'TCM-Identifier', 'Transformer-based Model', 'Molecular Docking' and 'MD simulations'.

Key Points

- We introduced TCM-navigator, a deep learning-based, end-to-end workflow for the generation and evaluation of TCM-like compounds for drug development.
- TCM-navigator consists of TCM-generator, TCM-identifier, and a physics-based evaluator.
- TCM-Generator is an LSTM model for generating TCM-like molecules.
- TCM-Identifier is the first deep-learning-based model for TCM-like molecule identification.
- Using the workflow, we generated protein-specific TCM-like compounds and *in*-silico evaluation of selected TCM-like compounds yields better or comparable affinity with respect to TCM compounds.

Acknowledgements

The authors acknowledge funding support from Bioinformatics Institute, A*STAR.

Author contributions

Feiying Chen and Hao Fan conceptualised and designed the project. Feiying Chen and Victor Jun Yu Lim generated and analyzed results, and all authors (Feiying Chen, Victor Jun Yu Lim, Mingyu Li, and Hao Fan) wrote the manuscript.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest: None declared.

Funding

None declared.

Data availability

The TCM-Generator and TCM-Identifier executables are publicly available for download at GitHub, https://github.com/cfy2yue/TCM-Navigator. The generated TCM-like compounds SMILES are available at https://zenodo.org/uploads/15518561.

References

1. Cheung FTCM. Made in China. *Nature* 2011;**480**:S82–3. https://doi.org/10.1038/480S82a
2. Lv Q, Chen G, He H. *et al.* TCMBank: Bridges between the largest herbal medicines, chemical ingredients, target proteins, and associated diseases with intelligence text mining. *Chem Sci* 2023;**14**:10684–701. https://doi.org/10.1039/D3SC02139D
3. Tu Y. The discovery of artemisinin (qinghaosu) and gifts from Chinese medicine. *Nat Med* 2011;**17**:1217–20. https://doi.org/10.1038/nm.2471
4. Xu H-Y, Zhang Y-Q, Liu Z-M. *et al.* ETCM: An encyclopaedia of traditional Chinese medicine. *Nucleic Acids Res* 2019;**47**:D976–82. https://doi.org/10.1093/nar/gky987

Downloaded from https://academic.oup.com/bib/article/26/5/bbaf498/8266408 by guest on 14 May 2026

5. Chen X, Zhou H, Liu YB. *et al.* Database of traditional Chinese medicine and its application to studies of mechanism and to prescription validation. *Br J Pharmacol* 2006;**149**:1092–103. https://doi.org/10.1038/sj.bjp.0706945
6. Li M, Zhang J. A focus on harnessing big data and artificial intelligence: Revolutionizing drug discovery from traditional Chinese medicine sources. *Chem Sci* 2023;**14**:10628–30. https://doi.org/10.1039/D3SC90185H
7. Han L, Wang B, Sun K. *et al.* A SARS-CoV-2 M^{pro} fluorescent sensor for exploring pharmacodynamic substances from traditional Chinese medicine. *Analyst* 2024;**149**:3585–95. https://doi.org/10.1039/D4AN00372A
8. Wu Z, Zhu Q, Zhang Y. *et al. EGFR-Associated Pathways Involved in Traditional Chinese Medicine (TCM)-1-Induced Cell Growth Inhibition, Autophagy and Apoptosis in Prostate Cancer*. Rep: Mol. Med (Spandidos Publications), 2018. https://doi.org/10.3892/mmr.2018.8818.
9. Luttens A, Gullberg H, Abdurakhmanov E. *et al.* Ultralarge virtual screening identifies SARS-CoV-2 main protease inhibitors with broad-Spectrum activity against coronaviruses. *J Am Chem Soc* 2022;**144**:2905–20. https://doi.org/10.1021/jacs.1c08402
10. Amabilino S, Pogány P, Pickett SD. *et al.* Guidelines for recurrent neural network transfer learning-based molecular generation of focused libraries. *J Chem Inf Model* 2020;**60**:5699–713. https://doi.org/10.1021/acs.jcim.0c00343
11. Xiong Z, Wang D, Liu X. *et al.* Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *J Med Chem* 2020;**63**:8749–60. https://doi.org/10.1021/acs.jmedchem.9b00959
12. Özçelik R, De Ruiter S, Criscuolo E. *et al.* Chemical language modeling with structured state space sequence models. *Nat Commun* 2024;**15**:6176. https://doi.org/10.1038/s41467-024-50469-9
13. Tay DWP, Yeo NZX, Adaikkappan K. *et al.* 67 million natural product-like compound database generated via molecular language processing. *Sci Data* 2023;**10**:296.
14. Hatakeyama-Sato K, Oyaizu K. Generative models for extrapolation prediction in materials informatics. *ACS Omega* 2021;**6**:14566–74. https://doi.org/10.1021/acsomega.1c01716
15. Tetko IV, Bruneau P, Mewes H-W. *et al.* Can we estimate the accuracy of ADME–tox predictions? *Drug Discov Today* 2006;**11**:700–7. https://doi.org/10.1016/j.drudis.2006.06.013
16. Swanson K, Walther P, Leitz J. *et al.* ADMET-AI: A machine learning ADMET platform for evaluation of large-scale chemical libraries. *Bioinformatics* 2024;**40**:btae416. https://doi.org/10.1093/bioinformatics/btae416
17. Schoretsanitis G, Deligiannidis KM, Paulzen M. *et al.* Drug-drug interactions between psychotropic medications and oral contraceptives. *Expert Opin Drug Metab Toxicol* 2022;**18**:395–411. https://doi.org/10.1080/17425255.2022.2106214
18. Sorokina M, Merseburger P, Rajan K. *et al.* COCONUT online: Collection of open natural products database. *J Chem* 2021;**13**:2. https://doi.org/10.1186/s13321-020-00478-9
19. Chithrananda S, Grand G, Ramsundar B. ChemBERTa: *Large-Scale Self-Supervised Pretraining for Molecular Property Prediction*. ArXiv Preprint 2020.
20. Ahmad W, Simon E, Chithrananda S. *et al. ChemBERTa-2: Towards Chemical Foundation* Models, ArXiv Preprint 2022.
21. Bagal V, Aggarwal R, Vinod PK. *et al.* MolGPT: Molecular generation using a transformer-decoder model. *J Chem Inf Model* 2022;**62**:2064–76. https://doi.org/10.1021/acs.jcim.1c00600
22. Naveed M, Hejazi V, Abbas M. *et al.* Chlorogenic acid (CGA): A pharmacological review and call for further research. *Biomed Pharmacother Biomedecine Pharmacother* 2018;**97**:67–74. https://doi.org/10.1016/j.biopha.2017.10.064
23. Greenblatt DJ, Shader RI, Divoll M. *et al.* Benzodiazepines: A summary of pharmacokinetic properties. *Br J Clin Pharmacol* 1981;**11 Suppl 1**:11S–6. https://doi.org/10.1111/j.1365-2125.1981.tb01833.x

Downloaded from https://academic.oup.com/bib/article/26/5/bbaf498/8266408 by guest on 14 May 2026

© The Author(s) 2025. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.
Briefings in Bioinformatics, 2025, **26(5)**, bbaf498
https://doi.org/10.1093/bib/bbaf498
Problem Solving Protocol