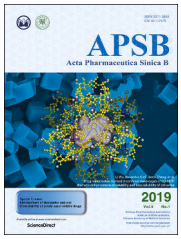
Acta Pharmaceutica Sinica B xxxx;xxx(xxx):xxx

Chinese Pharmaceutical Association
Institute of Materia Medica, Chinese Academy of Medical Sciences

Acta Pharmaceutica Sinica B

www.elsevier.com/locate/apsb
www.sciencedirect.com



ORIGINAL ARTICLE

Unveiling the bioactive landscape of drug inactive ingredients (DIGs) using deep transfer learning

Minjie Mou[a,b,†], Jinsong Zhang[a,†], Xingang Liu[c], Hao Yang[c], Tingting Fu[a], Hengbin Zhang[a], Yimiao Zhu[a], Tianle Niu[c], Xuedong Li[c], Yichao Ge[a], Ziqi Pan[a], Xinyu Liu[c], Huaicheng Sun[a], Tianyuan Zhang[a], Yang Zhang[c,*], Feng Zhu[a,b,*], Jianqing Gao[a,b,*]

[a]*College of Pharmaceutical Sciences, State Key Laboratory of Advanced Drug Delivery and Release Systems, Zhejiang University, Hangzhou 310058, China*
[b]*Department of Pharmacy, The Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou 310009, China*
[c]*School of Pharmacy, Hebei Medical University, Shijiazhuang 050017, China*

Received 29 May 2025; received in revised form 3 September 2025; accepted 30 December 2025

KEY WORDS

Drug inactive ingredients;
Excipients;
Excipient–protein interactions;
Bioactive landscape;
Deep transfer learning;
Transformer;
Drug formulation;
Drug safety

Abstract In a drug product, the major components by mass are the drug inactive ingredients (DIGs), which raises great concerns about their unwanted effects and clinical toxicities. It is demanded to unveil their proteome-wide bioactive landscape using computational methods. However, existing methods are impeded by either incapability to scan human proteome or inaccuracy in DIGs' bioactivity prediction. Here, a cross-attention transformer model, titled *TransDIG*, leveraging cross-module deep transfer learning was therefore developed to map the bioactive landscape of DIGs using minimal experimental data. First, the generalizability and interpretability of this model was verified by the prediction of zero-shot proteins and identification of key atoms/residues, respectively. Then, the bioactive landscape of hundreds of DIGs was unveiled using *TransDIG*, and thousands of potential bioactivities were found for the DIGs currently employed in pharmaceutical industry. Finally, the bioactivities of four popular DIGs were identified based on the landscape and experimentally validated by activity assay. As a result, the colorant β-carotene was validated to inhibit a critical drug transporter, and our study presented the

This article is part of special issue entitled: Machine Learning in Drug Discovery.

*Corresponding authors.
E-mail addresses: zhangyang@hebmu.edu.cn (Yang Zhang), zhufeng@zju.edu.cn (Feng Zhu), gaojianqing@zju.edu.cn (Jianqing Gao).
[†]These authors made equal contributions to this work.

Peer review under the responsibility of Chinese Pharmaceutical Association and Institute of Materia Medica, Chinese Academy of Medical Sciences.

https://doi.org/10.1016/j.apsb.2026.01.042

first *in vitro* evidence of the bioactivity of the antioxidant dodecyl gallate that has not previously been reported to regulate any human protein. This study might offer insights for the design of drug formulation and its clinical utilization.

1. Introduction

Drug inactive ingredients (DIGs, also known as excipients) have been generally regarded as biologically inert[1,2], but recent studies identify several DIGs exhibiting bioactivities against human proteins[3-5]. Because the amount of DIGs in a drug product outweighs the active ingredient by 10 to 100 folds[6], it is critical to uncover their activities for avoiding unwanted effect on drug efficacy and mitigating consequential toxicity[7-9]. Recently, a landmark study has been conducted to assess the likelihood of FDA-approved DIGs binding to human protein, drawing extensive attention on the communities of pharmaceutical and food sciences[10]. It experimentally tested the activities of 38 DIGs against 44 proteins, which revealed that many DIGs bound to multiple proteins[10]. Given the pervasiveness and diversity of DIGs, existing studies on assessing DIGs' bioactivity are considered as 'only scratched the surface'[11], and an elucidation of the comprehensive bioactive landscape for each DIG across all proteins in human proteome is therefore highly demanded[12].

Due to the huge number of proteins (~22,000[13]) in human proteome, it is impossible for experimental methods (characterized by time-consuming and resource-intensive) to depict the comprehensive bioactive landscape of studied DIGs[14], which asks for the development of computational approach[15]. So far, a variety of computational strategies (that may facilitate the elucidation of the bioactive landscape of DIGs) have been proposed[16-19]. Particularly, a series of machine learning-driven regression models have been constructed based on the similarities among ligands[16], which helped to predict the bioactivity of DIGs through scanning 1188 models of corresponding human proteins[16]. However, such type of models relies heavily on the availability of bioactive molecules (BAMs) for each studied protein[20], and this is the reason why the evaluation on DIGs' bioactivity in existing study[16] is limited to only 5% of human proteome (1188 proteins). Moreover, in-depth analysis reveals that 46.5% and 28.6% of the 1188 proteins have <200 and < 100 BAMs, respectively, which makes many of the models questionable due to their limited training data[21,22].

To address the above issue, other computational methods may also be considered to elucidate the bioactive landscape of DIG[17-19], which work by predicting compound-protein interactions (CPIs). Compared with the above ligand-based methods, these CPI-based ones can substantially broaden the scope of assessable human proteins[23,24]. However, as shown in Fig. 1A, the distribution of DIGs' bioactivities differs markedly from that of BAMs (the peak activity of DIGs is around two orders of magnitude higher than that of BAMs, indicating much lower bioactivities of DIGs than BAMs). Such difference in distributions can result in inaccurate prediction of DIGs' bioactivities, if the CPI-based model is trained solely based on BAMs[17-19]. Moreover, if the CPI-based strategy is directly applied to train model using DIG data, the generalizability of the resulting models will be impeded by the extremely limited amounts of DIG activities and interacting proteins available in published DIG database[25]. All in all, it is urgently needed to have a model tailored for unveiling the proteome-wide bioactive landscape of DIGs, but no such model has been available yet.

In this study, a model, named *TransDIG*, for unveiling the bioactive landscape of DIG was therefore developed. First, the strategy of cross-module deep transfer learning was introduced to train a cross-attention transformer model based on a minimal set of experimentally-validated bioactive DIGs. Second, the generalizability and interpretability of the developed model were validated by the accurate prediction of DIG activities in zero-shot protein tasks and the precise capture of both protein residues and DIG atoms essential for bioactivities. Third, the bioactive landscape of hundreds of DIGs was unveiled using *TransDIG*, and thousands of potential bioactivities were found for the DIGs employed in modern pharmaceutical industry. Finally, the bioactivities of four popular DIGs were identified based on the landscape and experimentally validated by activity assays. Particularly, the colorant β-carotene was validated to regulate a famous drug transporter, and our study provided the first *in vitro* evidence of the bioactivity of the antioxidant dodecyl gallate that has not previously been reported to modulate any human protein. All in all, this study might offer useful insights for the design of drug formulations and their subsequent clinical utilization.

2. Results and discussion

2.1. Constructing TransDIG for the prediction of DIG bioactivities

To uncover the bioactive landscape of DIG against various clinically important proteins, a deep learning framework, *TransDIG*, was designed based on the cross-attention transformer algorithm. *TransDIG* employed a unique cross-module deep transfer learning for its training process.

2.1.1. The deep learning framework of TransDIG

The overall framework of *TransDIG* is illustrated in Fig. 1B. *TransDIG* used the protein sequence and the Simplified Molecular Input Line Entry Specification (SMILES) of compound as inputs, which were then processed through four main modules: a protein encoder, a compound encoder, an interaction decoder and a binary classifier. It ultimately generated activity labels of compound-protein pairs. First, the protein sequence was encoded using a pre-trained word2vec model to generate residue encodings. These encodings were then processed through a multi-layer gated convolutional networks (GatedCNN) to capture both local and global contextual features of residues, thereby producing the protein embedding. For compound encoding, atom encodings were initially obtained by calculating its physicochemical features using RDKit[26]. The resulting encodings were then used to

Please cite this article as: Mou Minjie et al., Unveiling the bioactive landscape of drug inactive ingredients (DIGs) using deep transfer learning, Acta Pharmaceutica Sinica B, https://doi.org/10.1016/j.apsb.2026.01.042

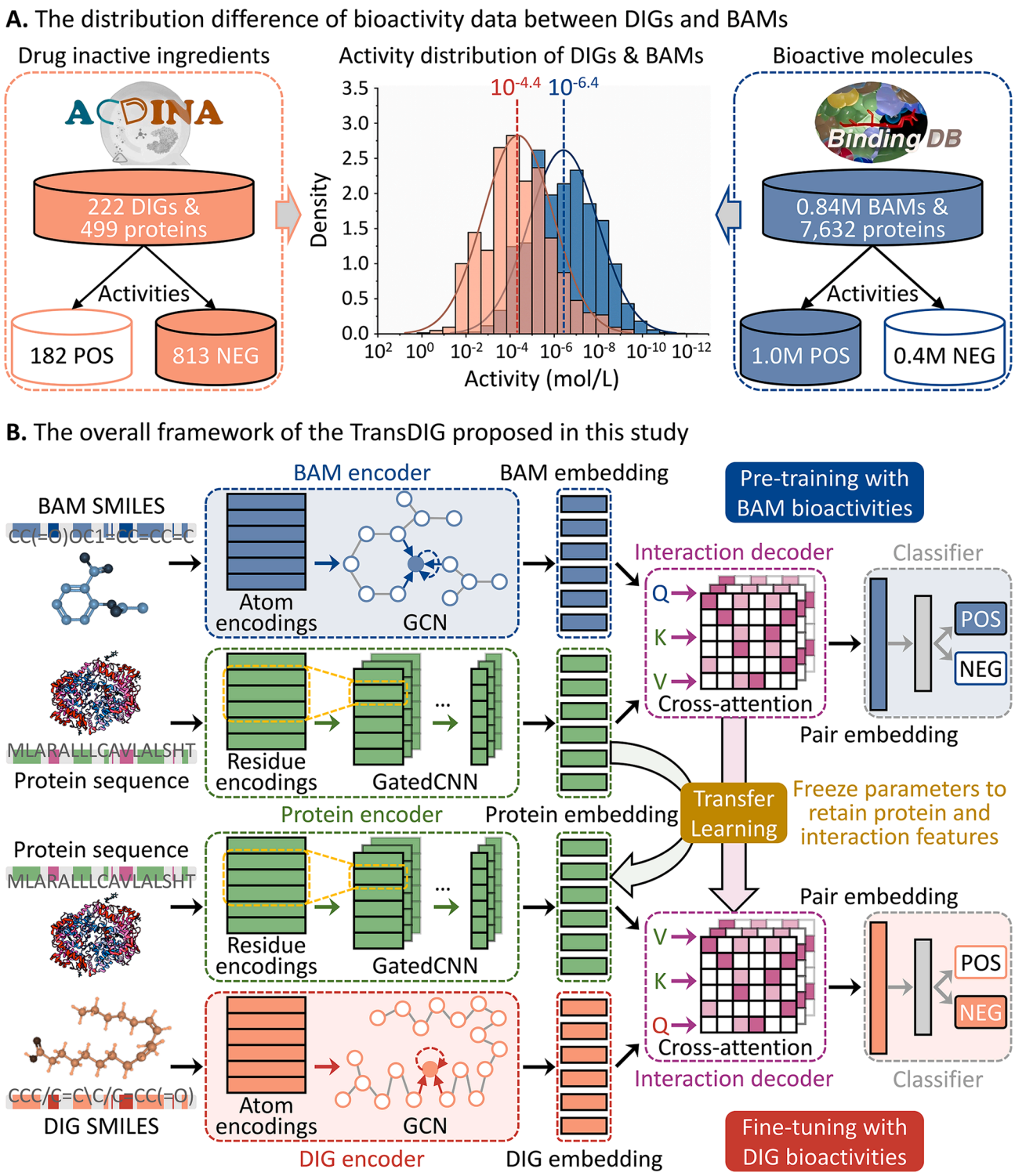


Figure 1 The activity data of drug inactive ingredients (DIGs) and the framework of *TransDIG*. (A) The distribution difference of bioactivity data between DIGs and bioactive molecules (BAMs). The activities of DIGs collected from the ACDINA database were relatively limited, comprising 182 positive and 813 negative data. In contrast, the activities of BAMs retrieved from the BindingDB database included nearly 1 million positive and 0.4 million negative data. The peak activity value of DIGs was approximately two orders of magnitude higher than that of BAMs, highlighting a substantially reduced activity level in DIGs relative to BAMs. (B) The overall framework of the *TransDIG* proposed in this study. *TransDIG* was designed based on the cross-attention transformer algorithm. It used the protein sequence and the SMILES of compound as inputs, which were then processed through a protein encoder, a compound encoder, an interaction decoder and a classifier. The training process of *TransDIG* comprised two phases: pre-training with BAM activities and fine-tuning with DIG activities. First, the model was pre-trained using large-scale BAM activity data. Then, the parameters of the pre-trained model were used to initialize *TransDIG*, followed by fine-tuning on DIG bioactivities. The parameters of protein encoder and interaction decoder were frozen to retain the learned protein and interaction features. M, million.

Please cite this article as: Mou Minjie et al., Unveiling the bioactive landscape of drug inactive ingredients (DIGs) using deep transfer learning, Acta Pharmaceutica Sinica B, https://doi.org/10.1016/j.apsb.2026.01.042

generate the compound embedding through a graph convolutional network (GCN), which learned the structural relationships between atoms. The dual encoding architecture ensured the comprehensive feature extraction of proteins and compounds. Then, the protein embedding and compound embedding were input into the interaction decoder, which was a modified transformer decoder incorporating the cross-attention algorithm. By leveraging the multi-head cross-attention mechanism, the interaction decoder captured the physical interactions between protein residues and compound atoms. This process led to a pair embedding that encapsulated the interaction features. In prediction, the pair embedding was processed by a classifier composed of multiple fully connected layers (FCs) for activity prediction.

2.1.2. *Training TransDIG using cross-module transfer learning*

To develop an accurate and well-generalized prediction model using the limited DIG bioactivity data, a cross-module transfer learning strategy was applied for training *TransDIG*. Transfer learning referred to a machine learning strategy that used generalizable insights gained from other relevant tasks to expedite the learning process for a distinct task, especially when only a small dataset was available[27]. In the field of drug discovery, deep transfer learning was the most commonly used type of transfer learning[28,29], which froze pre-trained modules to retain domain-invariant information acquired from the source domain. In contrast, fine-tuning the pre-trained modules allowed it to capture subject-specific information from the target domain[30].

Before training *TransDIG*, in addition to analyzing the differences in activity distribution between BAMs and DIGs as shown in Fig. 1A, we also examined the distributions of physicochemical properties and spatial conformations of DIGs and BAMs using the druglikeFilter tool[31], as well as the structural similarity between them. According to the results presented in Supporting Information Fig. S1A and S1B, DIGs differed significantly from the BAMs in terms of several physicochemical properties and spatial conformational features. For instance, DIGs exhibited lower numbers of atoms and lower molecular weights compared to BAMs, indicating reduced structural complexity. DIGs also displayed lower log*P* values, suggesting enhanced hydrophilicity, which could improve the solubility of poorly soluble drugs and thereby promote drug absorption. Moreover, DIGs had fewer aromatic rings and fewer rotatable bonds, indicating distinct conformational features compared to BAMs. We further analyzed the structural similarity between DIGs and BAMs. As shown in Fig. S1C, over half of the DIGs exhibited a maximum Tanimoto similarity of less than 0.7 to all BAMs, indicating low (<0.5, dark blue) or moderate (0.5−0.7, light blue) structural similarity between these DIGs and BAMs[32].

Due to the physicochemical and structural differences between DIGs and BAMs, the training process of *TransDIG* was designed in two phases: pre-training with BAM activities and fine-tuning with DIG activities. Specifically, the model underwent pre-training using large-scale BAM activities collected from BindingDB, enabling the protein encoder and interaction decoder to learn comprehensive representations of numerous proteins and the interaction patterns between small molecules and proteins, respectively. The parameters of the pre-trained model, PreDIG, were then used to initialize *TransDIG*, followed by fine-tuning DIG encoder and classifier using DIG bioactivity data. In other words, during the fine-tuning phase, the parameters of protein encoder module and interaction decoder module were frozen, and only the DIG encoder and classifier parameters were optimized. This unique cross-module transfer learning strategy retained learned large-scale protein and interaction features by freezing pre-trained protein encoder and interaction decoder, and extended the pre-trained compound encoder and classifier towards the DIG input domain to enhance specificity for learning the distinctive chemical features and activity distribution of DIGs. In summary, our deep transfer learning effectively mitigated data scarcity issue of DIG activities by transferring generalized interaction knowledge from BAMs, thereby enhancing model robustness against overfitting.

2.2. *Evaluating the performance of the TransDIG*

2.2.1. *Performance evaluation using five-fold cross validation*

Due to the scarcity of known DIG bioactivities, 5-fold cross-validation (CV) was employed to evaluate the performance of *TransDIG*. Three tasks were established using different data splitting strategies, namely 5-fold CV with random pair split, 5-fold CV with zero-shot protein setting and 5-fold CV with zero-shot DIG setting. *TransDIG* was compared with three deep learning methods for predicting CPIs, namely MolTrans[24], HyperAttentionDTI[33] and DeepDTAGen[34], using the same training data and transfer learning strategy as *TransDIG*. Three popular machine learning models were also evaluated, including support vector machine (SVM), random forest (RF) and extreme gradient boosting (XGBoost). As illustrated in Fig. 2A, in the 5-fold CV with random pair split task, *TransDIG* outperformed other models across all evaluation metrics, including ACC, AUROC, AUPRC and MCC. Specifically, *TransDIG* achieved a mean MCC of 0.534, surpassing the second-best method HyperAttentionDTI by 0.196 and the worst model DeepDTAGen by 0.470. The results demonstrated the high accuracy and robustness of *TransDIG*.

Due to the limited number of known proteins that interact with DIGs, it is necessary to evaluate the model's generalization ability for unseen proteins. The 5-fold CV with zero-shot protein setting task was designed to evaluate the model's capability in predicting DIG bioactivity toward previously unseen proteins, where proteins in the validation set entirely differed from those in the training set. The sequence identity between validation and training proteins in this task was analyzed using DIAMOND[35]. As the statistics presented in Supporting Information Fig. S2A, the majority of proteins in the validation set of each fold exhibited low (<30%, purple) or moderate (30%−70%, pink) maximum sequence identity with the corresponding proteins in the training set[36,37]. As illustrated in Fig. 2B, *TransDIG* still showed superior performance in the zero-shot protein setting, with its MCC metric significantly outperforming other models, achieving a mean value of 0.454. By contrast, the second-best model, MolTrans, only achieved a mean MCC of 0.352. Similarly, in the 5-fold CV with the zero-shot DIG setting, Fig. S2B showed that over half of the DIGs in the validation set of each fold shared a maximum Tanimoto similarity of less than 0.7 to all training DIGs, indicating low or medium structural similarity between the DIGs in the validation and training sets. As illustrated in Supporting Information Fig. S3A, *TransDIG* maintained good performance, achieving the highest scores in both ACC and MCC metrics, with mean values of 0.832 and 0.367, respectively. These results demonstrated the generalization ability of *TransDIG*. The source data for the evaluation results of all baseline models across the three 5-fold CV tasks were provided in Supporting Information Tables S1−S3.

Please cite this article as: Mou Minjie et al., Unveiling the bioactive landscape of drug inactive ingredients (DIGs) using deep transfer learning, Acta Pharmaceutica Sinica B, https://doi.org/10.1016/j.apsb.2026.01.042

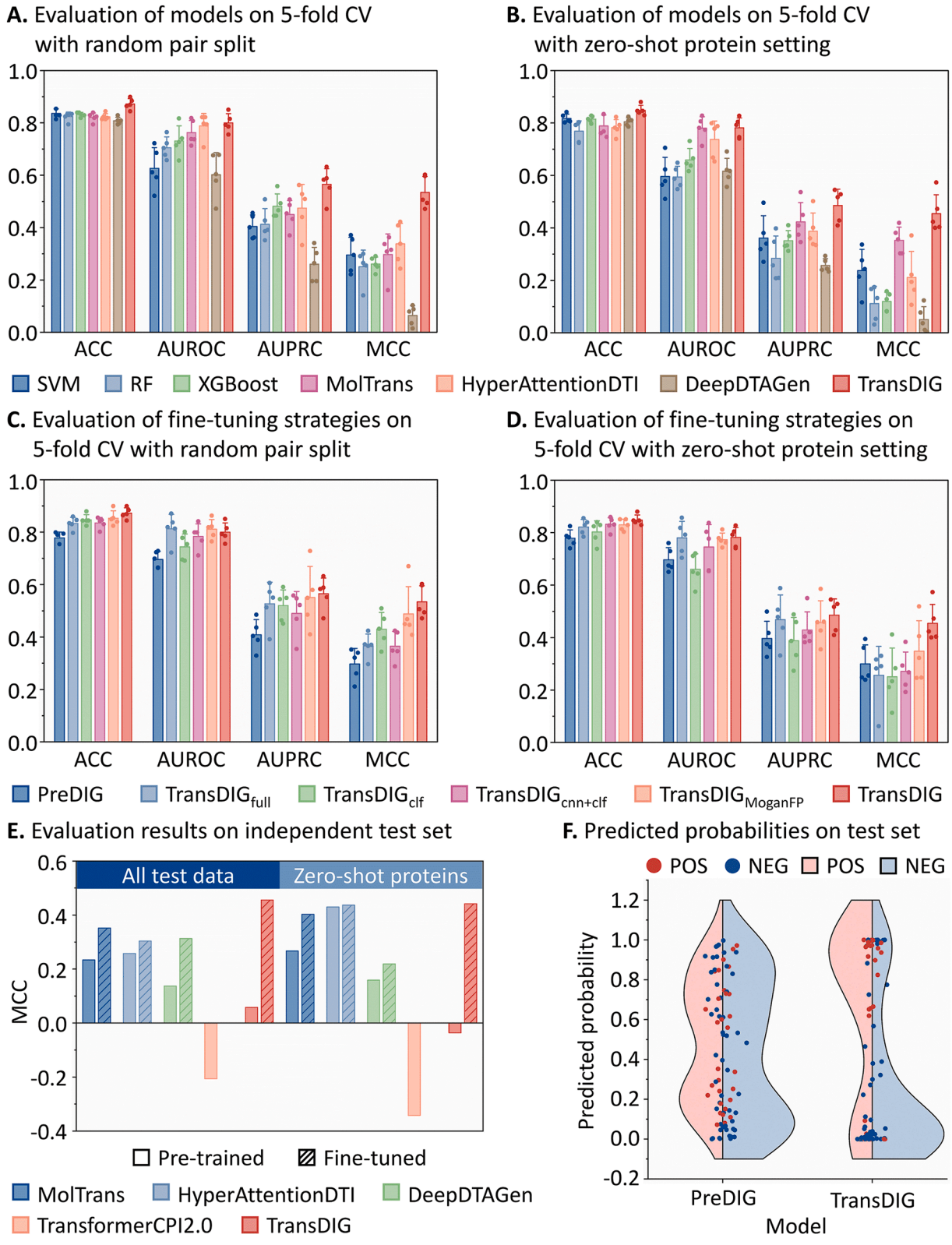


Figure 2 Performance evaluation of *TransDIG*. (A) Performance evaluation of *TransDIG* and baseline models on 5-fold CV with random pair split. (B) Performance evaluation of *TransDIG* and baseline models on 5-fold CV with zero-shot protein setting. (C) Performance evaluation of various fine-tuning strategies on 5-fold CV with random pair split. (D) Performance evaluation of various fine-tuning strategies on 5-fold CV with zero-shot protein setting. All error bars represented mean ± SD. (E) Performance evaluation of various pre-trained and fine-tuned models on independent test dataset. All models showed some degree of performance improvement on the independent test set after applying the transfer learning. (F) The predicted probabilities of *PreDIG* and *TransDIG* on test data. *PreDIG* exhibited nearly identical prediction probability distributions for both positive and negative data. In contrast, *TransDIG* successfully adjusted the prediction probability distributions, achieving a clear and effective separation between positive and negative data. Source data of all benchmark evaluations were provided in Tables S1−S10.

Please cite this article as: Mou Minjie et al., Unveiling the bioactive landscape of drug inactive ingredients (DIGs) using deep transfer learning, Acta Pharmaceutica Sinica B, https://doi.org/10.1016/j.apsb.2026.01.042

Owing to *TransDIG*'s retention of learned protein and interaction features during the fine-tuning process, it is likely that *TransDIG* has already captured the interaction patterns between zero-shot proteins in the validation set and BAMs during the pre-training phase. Given the same nature of compound-protein interactions, this pre-training strategy provided highly relevant knowledge for enhancing the activity prediction between zero-shot proteins and DIGs. *TransDIG*'s robust generalization to zero-shot proteins ensured its practicality in large-scale investigations of bioactive landscape between DIGs and various clinically important proteins.

2.2.2. Performance evaluation of various transfer learning strategies

Moreover, we also investigated the performance of several *TransDIG* variants, including models employing different fine-tuning strategies and a model incorporating molecular fingerprints for enhanced compound representation. Specifically, we evaluated three additional fine-tuning strategies: full-parameter fine-tuning ($TransDIG_{full}$), fine-tuning only the classifier ($TransDIG_{clf}$), and jointly fine-tuning the protein encoder and the classifier ($TransDIG_{cnn\ +\ clf}$). The results showed that *TransDIG* consistently outperformed other fine-tuning strategies across all 5-fold CV tasks. As illustrated in Fig. 2C and Fig. S3B, *TransDIG* achieved superior performance over models with alternative fine-tuning strategies in terms of ACC, AUPRC, and MCC in both the 5-fold CV with random pair split and the 5-fold CV with zero-shot DIG setting. In contrast, the pre-trained model *PreDIG* only exhibited performance comparable to that of various machine learning models. In the 5-fold CV with zero-shot protein setting, *TransDIG* outperformed all other fine-tuning strategy variants across all evaluation metrics, as shown in Fig. 2D. In the 5-fold CV with random pair split, zero-shot protein and zero-shot DIG settings, *TransDIG* achieved MCC scores that were 0.104, 0.183, and 0.049 higher than the second-best fine-tuned model, respectively. Notably, compared to the pre-trained model *PreDIG*, the majority of fine-tuning strategies improved predictive performance on the 5-fold CV with random pair split and the 5-fold CV with zero-shot protein setting tasks, highlighting the effectiveness of deep transfer learning strategy. However, in the 5-fold CV with zero-shot DIG setting, none of the other fine-tuned models surpassed *PreDIG* in terms of AUROC, AUPRC, and MCC. In contrast, *TransDIG* still achieved a 0.049 higher MCC than *PreDIG*. The source data for evaluation results of various transfer learning strategies were provided in Supporting Information Tables S4−S6.

We also applied the transfer learning strategy of fine-tuning the compound encoder and classifier to other deep learning models, which yielded similar results. The performance of the pre-trained baseline models on the three 5-fold CV tasks were presented in Supporting Information Tables S7−S9. Compared with performance of fine-tuned models presented in Tables S1−S3, the results showed that this fine-tuning strategy could improve the performance of various models on the majority of tasks, fully demonstrating the effectiveness of the strategy that involved fine-tuning both the compound encoder and the classifier.

Furthermore, given the structural and physicochemical differences between DIGs and BAMs, we constructed an additional model, $TransDIG_{MorganFP}$, to investigate whether the inclusion of molecular fingerprints could enhance model performance. $TransDIG_{MorganFP}$ incorporated Morgan fingerprints (MorganFP) alongside GCN-derived molecular representations, and was trained using the same protocol as *TransDIG*. The MorganFP of each molecule was processed through a FC layer and then concatenated with the GCN-extracted molecular representation. As shown by the orange bars in Fig. 2C and D, and Fig. S3B, $TransDIG_{MorganFP}$ exhibited only a marginally higher AUROC than *TransDIG* in the 5-fold CV with random pair split task, but showed no notable performance improvement across the other evaluation tasks. Thus, the inclusion of MorganFP could not further improve the model's predictive performance.

2.2.3. Performance evaluation using independent test data

To develop an application-oriented version, all available DIG bioactivity data were utilized to retrain the *TransDIG* model. A highly reliable independent test dataset containing 85 DIG bioactivities was curated from a previous study[10] and employed to evaluate the performance of *TransDIG*. As shown in Fig. S2C, the test set comprised 31 DIGs and 39 human proteins, including 15 zero-shot proteins and 15 zero-shot DIGs that were absent from the training set. These 15 zero-shot proteins were involved in 40 bioactivities, and 4 of them shared low sequence identity (<30%) with proteins in the training set. Among the 15 zero-shot DIGs, 11 exhibited a maximum Tanimoto similarity lower than 0.5 to all DIGs in the training set.

We trained all competing deep learning models using the same training scheme and evaluated both their pre-trained and final fine-tuned versions on the independent test set. The MCC values of the different models were illustrated in Fig. 2E, with source data provided in Supporting Information Table S10. The results revealed that all models showed some degree of performance improvement on the independent test set after applying the transfer learning. Specifically, the fine-tuned *TransDIG* achieved the best performance on both the complete test data and the zero-shot protein subset, with MCC values of 0.456 and 0.442, respectively. Although the pre-trained *PreDIG* model exhibited poor performance on both test subsets, *TransDIG* demonstrated the most substantial performance improvements compared to other models using the same transfer learning strategy, with MCC increases of 0.398 and 0.478 on the respective datasets. Among all pre-trained models, HyperAttentionDTI showed the best initial performance; however, it showed only a slight improvement following fine-tuning. Moreover, we also evaluated a pre-trained CPI prediction model, TransformerCPI2.0[38]. It should be noted that the poor performance of TransformerCPI2.0 was primarily due to the unavailability of its training code, which prevented retraining on the data used for *TransDIG* construction. It was originally trained on fewer pre-training samples and used a different label definition for the training data compared to the test set.

For a more rigorous and unbiased assessment, we stratified the entire test set into subsets according to difficulty levels, following the established data splitting protocols from prior studies on evaluation schemes for pair-input computational prediction tasks[39,40]. The test set was stratified into two subsets based on similarity to the training data: subset C1, pairs involving DIGs and proteins that were either previously seen or sequentially similar to those in the fine-tuning dataset; subset C2, pairs in which either the DIG or the protein exhibited low similarity to any instance in the fine-tuning dataset. Notably, the test set did not include any samples where both the DIG and the protein were dissimilar to those used for fine-tuning. As a result, the subset C1 contained 59 pairs, while the subset C2 included 15 pairs involving 11 dissimilar DIGs and 11 pairs involving 4 dissimilar proteins.

Please cite this article as: Mou Minjie et al., Unveiling the bioactive landscape of drug inactive ingredients (DIGs) using deep transfer learning, Acta Pharmaceutica Sinica B, https://doi.org/10.1016/j.apsb.2026.01.042

All models were re-evaluated on subsets C1 and C2. As shown in Supporting Information Fig. S4A, *TransDIG* achieved the best performance on both subsets. It significantly outperformed all other models on the lower-difficulty subset C1, and performed comparably to the second-ranked model, DeepDTAGen, on the more challenging subset C2 (with source data provided in Supporting Information Table S11). Compared to the pre-trained model *PreDIG*, *TransDIG* demonstrated consistent and notable performance improvements across both subsets. Furthermore, we conducted a detailed analysis of model performance on subset C2. As illustrated in Fig. S4B and Supporting Information Table S12, for the 15 pairs in C2 involving 11 dissimilar DIGs, *TransDIG* achieved the highest predictive performance, with the MCC value 0.051 higher than that of the second-best model, TransformerCPI2.0. The ability to predict activities of dissimilar DIGs was crucial, as it reflected the model's reliability when applied to novel DIGs without previously reported activities. In contrast, HyperAttentionDTI and DeepDTAGen showed the weakest performance on these samples. For the 11 pairs involving 4 dissimilar proteins, DeepDTAGen achieved the best performance, with *TransDIG* ranking second. This slight gap was attributed to DeepDTAGen correctly predicting one additional true positive instance. These results collectively indicated that, compared to existing methods, *TransDIG* exhibited good generalization capability in DIG activity prediction.

The detailed information of test set was provided in Supporting Information Table S13. The predicted probabilities of the test data were further analyzed, with the results illustrated in Fig. 2F. *PreDIG* exhibited nearly identical prediction probability distributions for both positive and negative data, making it ineffective at distinguishing between the them. Notably, *TransDIG* successfully adjusted the prediction probability distributions, achieving a clear and effective separation between positive and negative data. These results fully exhibited the robust capability and practical utility of *TransDIG* in predicting new DIG bioactivities.

2.3. Analyzing the interpretability of TransDIG

Owing to the cross-attention mechanism equipped in the interaction decoder module, *TransDIG* exhibited good interpretability both in terms of DIGs and proteins. To analyze its interpretability, two DIG−protein pairs were selected from the independent test set: propyl gallate and human polyunsaturated fatty acid 5-lipoxygenase (ALOX5), and methylene blue and human muscarinic acetylcholine receptor M2 (CHRM2). First, *TransDIG* successfully predicted the activity labels of both pairs as positive and subsequently calculated the attention scores for each residue in the protein and each atom in the DIG. Next, the residues and atoms with high attention scores were highlighted. Then, molecular docking and molecular dynamics (MD) simulations were employed to analyze and visualize the binding conformations between DIGs and proteins, aiding in understanding the specific interactions.

As presented in Fig. 3A, computational simulation trajectory analysis revealed that propyl gallate formed stable hydrogen bonds with residues ASN555 and GLN558, with occupancies of 64.14% and 19.99%, respectively. Moreover, the molecule also exhibited a certain probability of forming hydrogen bonds with Tyr559 and Gln364. It was worth noting that the protein residues and propyl gallate atoms involved in these hydrogen bonds were successfully assigned high attention scores by *TransDIG*. The protein residues and ligand atoms with high attention score were highlighted in red and yellow circles, respectively.

Similarly, as shown in Fig. 3B, four of the top ten CHRM2 residues ranked by *TransDIG*'s attention scores were located within 5.0 Å of the methylene blue in the docking pose, including ILE72, SER76, PHE181, and PHE195. To further investigate the contribution of these four residues to bioactivity, we performed MD simulations and free energy decomposition analysis. MD simulations were conducted to achieve sufficient conformational sampling. Three independent parallel production runs (3×100 ns) were performed to ensure the robustness and reproducibility of the simulation results. Using the docking-derived initial conformation as a reference, the average root-mean-square deviation (RMSD) values of the protein, the ligand, and the residues within 5.0 Å of methylene blue (defined as the binding site) were calculated. These metrics were employed to quantitatively assess the conformational stability of the receptor-ligand complex throughout the simulations. As shown in Supporting Information Fig. S5, during the 100 ns MD simulations, the average RMSD values of the ligand (black curve) and the key residues within the binding site (red curve) across the three trajectories remained below 2.0 Å. This indicated that both the ligand conformation and the key binding-site residues exhibited only minor fluctuations and rapidly converged throughout the simulations. In contrast, the RMSD values of the protein backbone (blue curve) displayed relatively large fluctuations in the early stage of the simulation (0−20 ns), but gradually stabilized after 20 ns and eventually converged to approximately 3.0 Å, further confirming that the system reached equilibrium in the later stage of simulations.

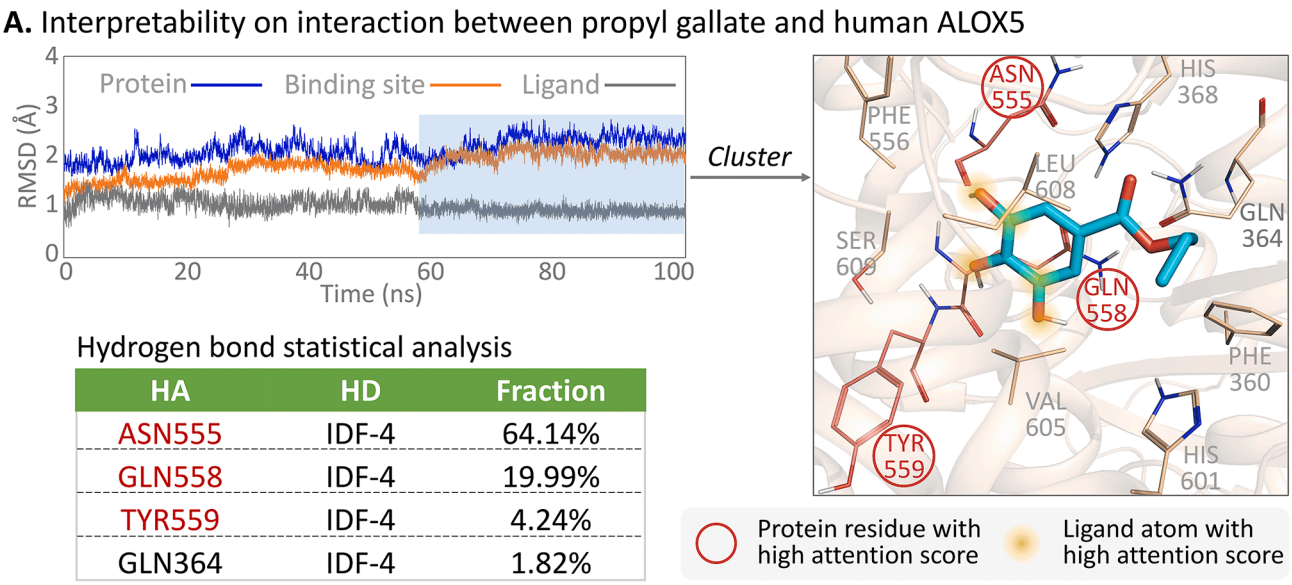
Based on free energy decomposition and representative conformational analysis, the binding mode between CHRM2 and methylene blue was systematically examined. As shown in the bottom-left corner of Fig. 3B, free energy decomposition analysis identified key residues with favorable contributions to methylene blue-CHRM2 binding affinity, including ILE72 (−0.65 kcal/mol), SER76 (−0.32 kcal/mol), PHE181 (−0.71 kcal/mol) and PHE195 (−0.43 kcal/mol). These interpretability analyses illustrated that *TransDIG* has effectively captured the interaction patterns between DIGs and proteins using only sequence information.

2.4. Unveiling the bioactive landscape of DIGs

We then used *TransDIG* to conduct the proteome-wide mapping of bioactive landscape for DIGs. To facilitate the discovery of DIG bioactivities, we have included partial scanning results. Specifically, we provided the bioactivities between 152 commonly used DIGs (listed in Supporting Information Table S14) and four classes of clinically important proteins, including 111 drug-metabolizing enzymes (DMEs), 42 drug transporters (DTPs), 460 drug therapeutic targets (DTTs) and 54 toxicity-related proteins (TRPs). This landscape, which included 5153 potential bioactivities, was offered in Supporting Information Tables S15−S18. Detailed statistical analysis of the bioactive landscape was presented in Fig. 4.

As shown in Fig. 4A, there were widespread bioactivities between DIGs and human proteins, with up to 93 DIGs (>61%) exhibiting potential bioactivity with these clinically important human proteins. Potential bioactivity was observed across all 10 functional classes of DIGs, and only a small fraction of the four critical human protein classes were predicted to exhibit no bioactivity with any DIG. Fig. 4B illustrates that the number of identified bioactivities varied significantly between each DIG

Please cite this article as: Mou Minjie et al., Unveiling the bioactive landscape of drug inactive ingredients (DIGs) using deep transfer learning, Acta Pharmaceutica Sinica B, https://doi.org/10.1016/j.apsb.2026.01.042



Hydrogen bond statistical analysis

HA	HD	Fraction
ASN555	IDF-4	64.14%
GLN558	IDF-4	19.99%
TYR559	IDF-4	4.24%
GLN364	IDF-4	1.82%

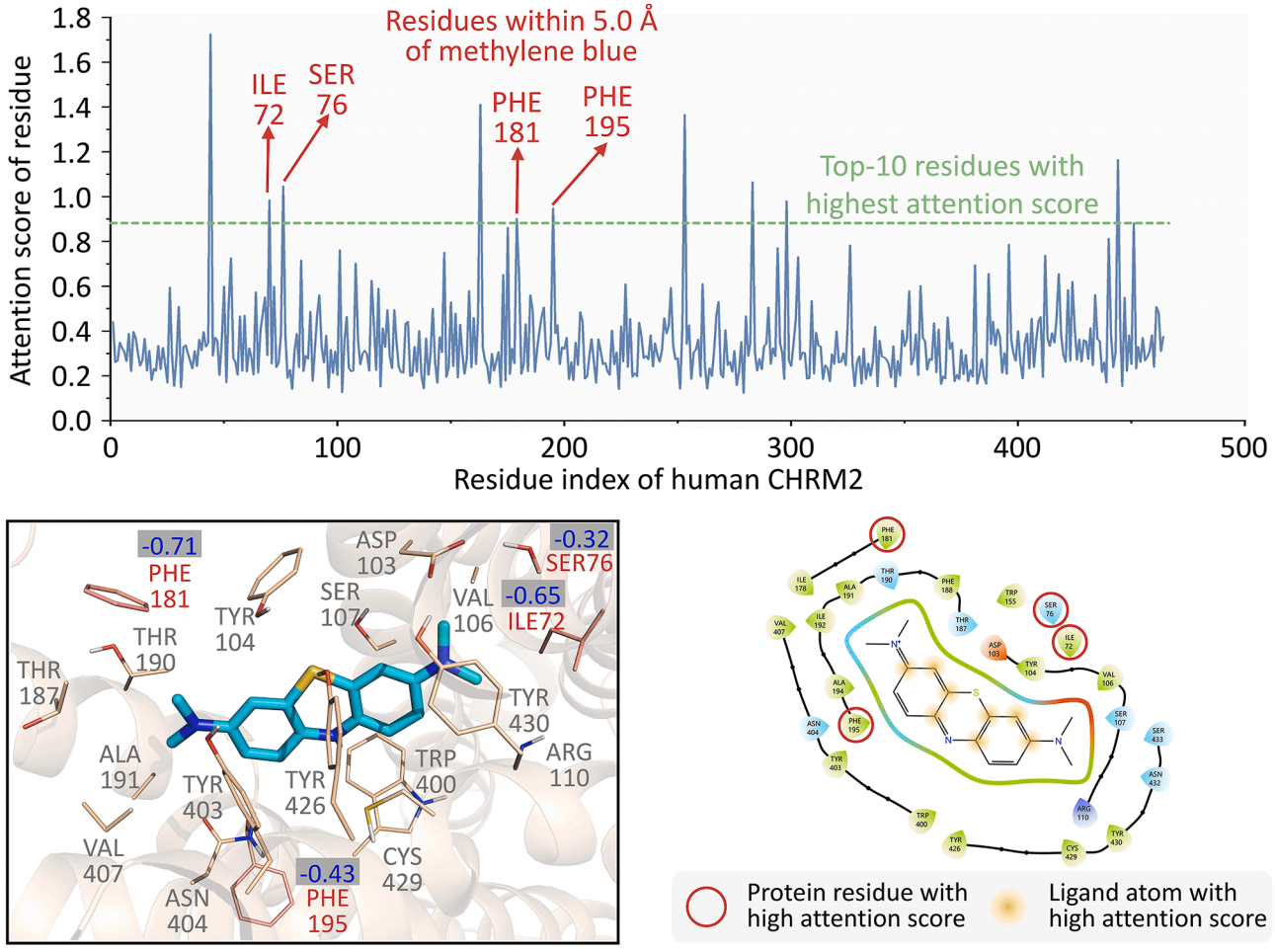


Figure 3 The interpretability analysis of *TransDIG* using two samples from independent test set. (A) Interpretability on interaction between propyl gallate and human ALOX5. The crystal structure of ALOX5 used for molecular simulation was extracted from PDB (PDB ID: 3V99). (B) Interpretability on interaction between methylene blue and human CHRM2. The crystal structure of CHRM2 used for molecular simulation was extracted from PDB (PDB ID: 5ZKC). The top panel showed the attention scores of residues in predicting the interaction between methylene blue and CHRM2. The bottom-left panel showed the result of free energy decomposition analysis for methylene blue−CHRM2 complex. The values highlighted in blue indicated the energy contribution of the residues to ligand's binding. The binding conformations were visualized in PyMOL.

Please cite this article as: Mou Minjie et al., Unveiling the bioactive landscape of drug inactive ingredients (DIGs) using deep transfer learning, Acta Pharmaceutica Sinica B, https://doi.org/10.1016/j.apsb.2026.01.042

class and each protein class. Coloring agents were predicted to have the highest number of potential bioactivities (1336), followed by flavoring agents (934). In contrast, buffering agents had the fewest predicted bioactivities, with only 127 interactions detected. Given the varying numbers of members within each DIG or protein class, additional analysis was conducted to determine the proportion of predicted bioactivities within each DIG−protein class pair. As the results shown in Fig. 4C, coloring agents and DMEs formed the highest percentage of bioactive DIG−protein pairs at 24.2%, while buffering agents and DTPs had the lowest percentage at just 0.5%.

The top 30 proteins from each class with the highest number of potential bioactivities were selected to illustrate the bioactive landscape of all 152 DIGs, as shown in Fig. 4D. The large-scale DIG bioactive landscape established in this study was expected to accelerate systematic investigations into DIG activities, highlighting their complex roles in drug therapy. DIGs could impact drug efficacy through interactions with DMEs and DTPs, potentially resulting in side effects *via* DIG−drug interactions. Additionally, they might induce toxicity by interacting with TRPs or achieve therapeutic effects by modulating DTTs. The prevalent bioactivity between DIGs and proteins underscored the need for attention from pharmacologists. These findings provided novel insights for drug formulation design and clinical usage of drugs.

2.5. *Validating the identified bioactivities for DIGs*

To valid the practical utility of *TransDIG* in real-world scenarios, we then performed protein activity inhibition assays based on the bioactive landscape. We selected representative DIGs from four distinct functional classes, namely surfactants, colorants, antioxidants, and antimicrobial preservatives. Specifically, for antioxidants and antimicrobial preservatives, we ranked the DIGs within each category based on the number of predicted interactions. The dodecyl gallate (antioxidant) and chlorhexidine (antimicrobial preservative) with the highest predicted activity counts were selected for experimental validation. For surfactants and colorants, in addition to the number of predicted activities, we further considered the availability as well as the frequency of use in drug formulation and daily diet. Accordingly, we selected linoleic acid from 15 surfactants and β-carotene from 17 colorants for experimental validation, both of which are widely present in pharmaceutical formulations and daily diet. Finally, for each selected DIG, all predicted interactions were ranked in descending order based on their predicted probabilities, and the protein with the highest probability was chosen for subsequent experimental validation. As a result, four bioactivities from the identified landscape were selected for validation, involving four distinct types of DIGs and three different classes of proteins (as provided in Table 1).

2.5.1. *Validating the bioactivity of linoleic acid against human SPHK1*

First, we validated the bioactivity between the surfactant linoleic acid and the human DME sphingosine kinase 1 (SPHK1), as demonstrated in Fig. 5A. Linoleic acid was widely utilized in pharmaceutical formulations such as chlordiazepoxide hydrochloride capsules, various liposomal systems, cosmetics, and dietary supplements[41,42]. SPHK1 was a key cytoplasmic lipid-metabolizing enzyme, which catalyzed the phosphorylation of sphingosine to sphingosine-1-phosphate and was implicated in the metabolism of immunomodulatory drugs such as the fingolimod[43,44]. Additionally, aberrant activation of SPHK1 were reported to be associated with the pathogenesis of various diseases, including multiple cancers and inflammatory responses[45,46].

The results revealed that linoleic acid exerted a pronounced concentration-dependent inhibition on SPHK1 enzymatic activity, with an IC_{50} value of 80.7 $\pm$ 14.3 μmol/L. Representative conformation analysis and detailed binding mode characterization (Fig. 5A, middle panel) further elucidated the interaction between linoleic acid and SPHK1. These findings suggested that linoleic acid-mediated inhibition of SPHK1 during clinical use might interfere with the enzyme's metabolism of SPHK1-dependent drugs, potentially influencing drug efficacy or safety.

2.5.2. *Validating the bioactivity of β-carotene against human OCT1*

We next investigated the inhibitory effect of the colorant β-carotene on the human DTP organic cation transporter 1 (OCT1), as illustrated in Fig. 5B. β-Carotene was a compound extensively incorporated into pharmaceutical formulations such as oral emulsions, topical creams, and granules (*e.g.*, colestipol hydrochloride granules), and was also ubiquitously present in daily dietary sources, including kale, spinach, carrots, and cantaloupe[47,48]. OCT1 was primarily expressed in the liver and was a polyspecific transporter with a broad substrate spectrum[49]. It mediated the cellular uptake of various important drugs, such as metformin (a first-line drug for type 2 diabetes), oxaliplatin and other cationic chemotherapeutic agents (*e.g.*, cisplatin and carboplatin)[50].

The experimental result demonstrated that under low-dose β-carotene stimulation, the uptake of the cationic substrate ASP^+ by OCT1 increased, suggesting that low doses of β-carotene might enhance OCT1 activity. Conversely, when the concentration of β-carotene was increased further, the accumulation of the cationic substrate ASP^+ significantly decreased. As the concentration of β-carotene progressively increased, the accumulation of ASP^+ showed a dose-dependent decline. These findings indicated that at a high concentration, β-carotene demonstrated an inhibitory effect on the transport activity of OCT1. Analysis of the experimental data revealed that the IC_{50} of β-carotene on OCT1 was 138.4 $\pm$ 31.9 μmol/L. The molecular simulation results and the detailed binding mode between β-carotene and OCT1 were presented in the middle panel of Fig. 5B. Considering the widespread presence of β-carotene in daily life[51], its regulatory impact on OCT1 could influence various real-life scenarios, especially the potential pharmacokinetic interactions during co-administration with OCT1-dependent medications.

2.5.3. *Validating the bioactivities of dodecyl gallate and chlorhexidine against human EGFR*

Finally, the bioactivities of the antioxidant dodecyl gallate and the antimicrobial preservative chlorhexidine on the well-known DTT human EGFR were experimentally validated, as shown in Fig. 5C and D, respectively. Dodecyl gallate was commonly used in

Protein residues within 5.0 Å of DIGs were displayed in lines, and those with high attention scores were highlighted in red. DIGs structures were shown in cyan. The interactions between DIGs and proteins were visualized using Schrödinger. The protein residues and DIG atoms with high attention scores were marked in red and yellow circles, respectively.

Please cite this article as: Mou Minjie et al., Unveiling the bioactive landscape of drug inactive ingredients (DIGs) using deep transfer learning, Acta Pharmaceutica Sinica B, https://doi.org/10.1016/j.apsb.2026.01.042

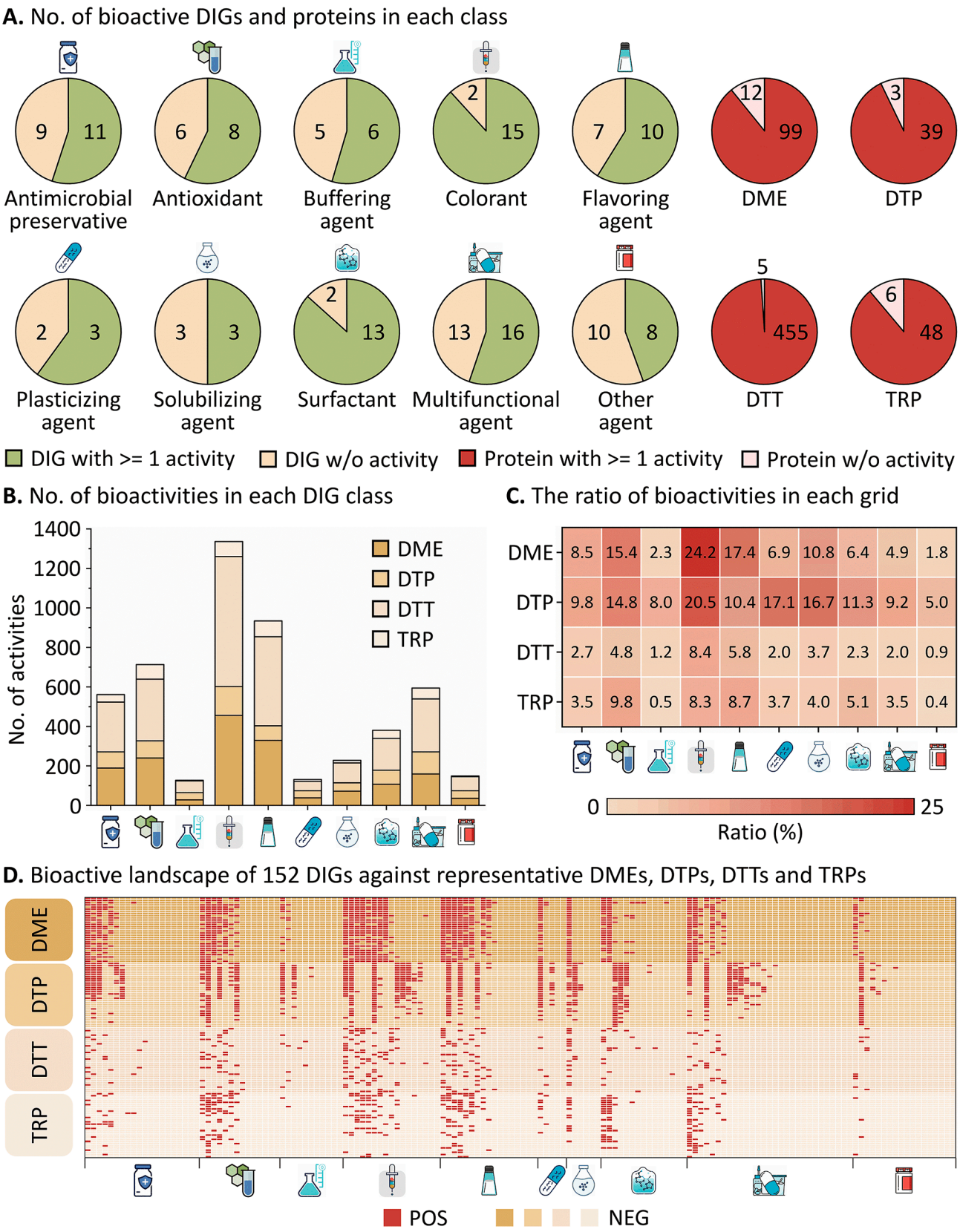


Figure 4 Statistics on the comprehensive bioactive landscape of DIGs unveiled by *TransDIG*. (A) The number of bioactive DIGs and bioactive proteins within ten classes of DIGs and four classes of proteins. DIGs with at least one activity and those without activity were displayed in green and yellow, respectively. Proteins with at least one activity and those without activity were shown in red and pink, respectively. (B) The number of identified bioactivities in each DIG class. The number of identified bioactivities varied across each DIG—protein class pair, with the four classes of

Please cite this article as: Mou Minjie et al., Unveiling the bioactive landscape of drug inactive ingredients (DIGs) using deep transfer learning, Acta Pharmaceutica Sinica B, https://doi.org/10.1016/j.apsb.2026.01.042

Table 1 The summary of identified bioactivities of four DIGs from different functional classes.

DIG name	Functional class	Uses	Family of training protein (gene of training protein)	Family of identified protein (gene of identified protein)	Bioactivity (IC_{50}, μmol/L)
Linoleic acid	Surfactant	Used in drug formulations such as chlordiazepoxide hydrochloride capsule, various liposomes, cosmetics and dietary supplements.	Fatty-acid binding protein family (FABP3, FABP4); G-protein coupled receptor 1 family (FFAR1, FFAR4); Nuclear hormone receptor family (PPARA)	DME: Sphingosine kinase family (SPHK1)	80.7 ± 14.3
β-Carotene	Colorant	Used in drug formulations such as oral emulsion, topical creams, and granules such as colestipol hydrochloride granules.	ABCB family (ABCB11); Organo anion transporter family (SLCO1B1)	DTP: Organic cation transporter family (SLC22A1)	138.4 ± 31.9
Dodecyl gallate	Antioxidant	Used in various pharmaceutical creams and emulsions, cosmetics, and various foods.	Not available	DTT: Tyr protein kinase family (EGFR)	27.0 ± 0.8
Chlorhexidine	Antimicrobial preservative	Used in oral pellicles such as compound chlorhexidine and dexamethasone pellicles, various pharmaceutical creams and sprays, mouthwashes, and cosmetics.	STE Ser/Thr protein kinase family (MAP4K2); Organic cation transporter family (SLC22A1, SLC22A2, SLC22A3); Multi antimicrobial extrusion family (SLC47A1, SLC47A2); Flavin monoamine oxidase family (SMOX)	DTT: Tyr protein kinase family (EGFR)	59.2 ± 4.2

DME, drug-metabolizing enzyme; DTP, drug transporter; DTT, drug therapeutic target.

various pharmaceutical creams and emulsions, cosmetics, and foods[52,53]. Chlorhexidine was utilized in oral pellicles such as compound chlorhexidine and dexamethasone pellicles, various pharmaceutical creams and sprays, mouthwashes, and cosmetics[54,55]. EGFR was a well-studied tyrosine kinase receptor which was widely overexpressed in multiple types of tumor cells, and it played a crucial role in the proliferation, differentiation, survival, and migration of tumor cells[56,57].

In EGFR inhibition assays, dodecyl gallate demonstrated a notable concentration-dependent inhibitory effect on EGFR activity. As the concentration of dodecyl gallate increased, the inhibition rate of EGFR activity progressively rose. Experimental analysis determined that the IC_{50} of dodecyl gallate on EGFR was 27.0 ± 0.8 μmol/L. Similarly, chlorhexidine also exhibited inhibitory activity against EGFR. Experimental data indicated that the IC_{50} of chlorhexidine on EGFR was 59.2 ± 4.2 μmol/L. Notably, due to the poor aqueous solubility of chlorhexidine, compound precipitation occurred at concentrations above 71 μmol/L during EGFR inhibition assays. This solubility limitation prevented further increases in chlorhexidine concentration, thereby precluding the attainment of a clear plateau in its maximal EGFR inhibition, as observed in Fig. 5D. Although this solubility limitation may compromise the accuracy of IC_{50} curve fitting, the data still provide clear evidence that chlorhexidine exhibits inhibitory activity against EGFR at high concentrations.

The MD simulation results of dodecyl gallate and chlorhexidine with EGFR were shown in the middle panels of Fig. 5C and D, respectively. These results indicated that both dodecyl gallate and chlorhexidine could inhibit tumor growth by suppressing the EGFR activity. These findings indicated that in certain non-EGFR-targeted anticancer drug formulations, the inclusion of these DIGs could enhance therapeutic efficacy through synergistic effects.

In summary, a key observation from our study is that many DIGs exhibit direct biological activities against pharmacologically relevant proteins. The summary of the experimental results is shown in Table 1. These results indicate that low bioactivity in interactions between DIGs and proteins is a prevalent and physiologically plausible phenomenon, as evidenced by the independent test set in which eight out of 85 DIG activities exhibit IC_{50} or K_i values exceeding 50 μmol/L, and five surpass 200 μmol/L. Notably, among the discovered DIG activities, dodecyl gallate has not been previously reported to exhibit any biological activity with human proteins. Moreover, the families of identified bioactive protein for the other three DIGs were entirely different from their

human proteins represented by different colors. (C) The ratio of identified bioactivities in each DIG−protein class pair. Coloring agents and DMEs formed the highest percentage of bioactivities at 24.2%, while buffering agents and DTPs had the lowest percentage at just 0.5%. (D) Bioactive landscape of 152 DIGs against 30 representative DMEs, DTPs, DTTs and TRPs. Identified bioactive DIG−protein pairs were classified as positives and highlighted with red squares, while those predicted to be inactive were classified as negatives and highlighted with yellow squares.

Please cite this article as: Mou Minjie et al., Unveiling the bioactive landscape of drug inactive ingredients (DIGs) using deep transfer learning, Acta Pharmaceutica Sinica B, https://doi.org/10.1016/j.apsb.2026.01.042

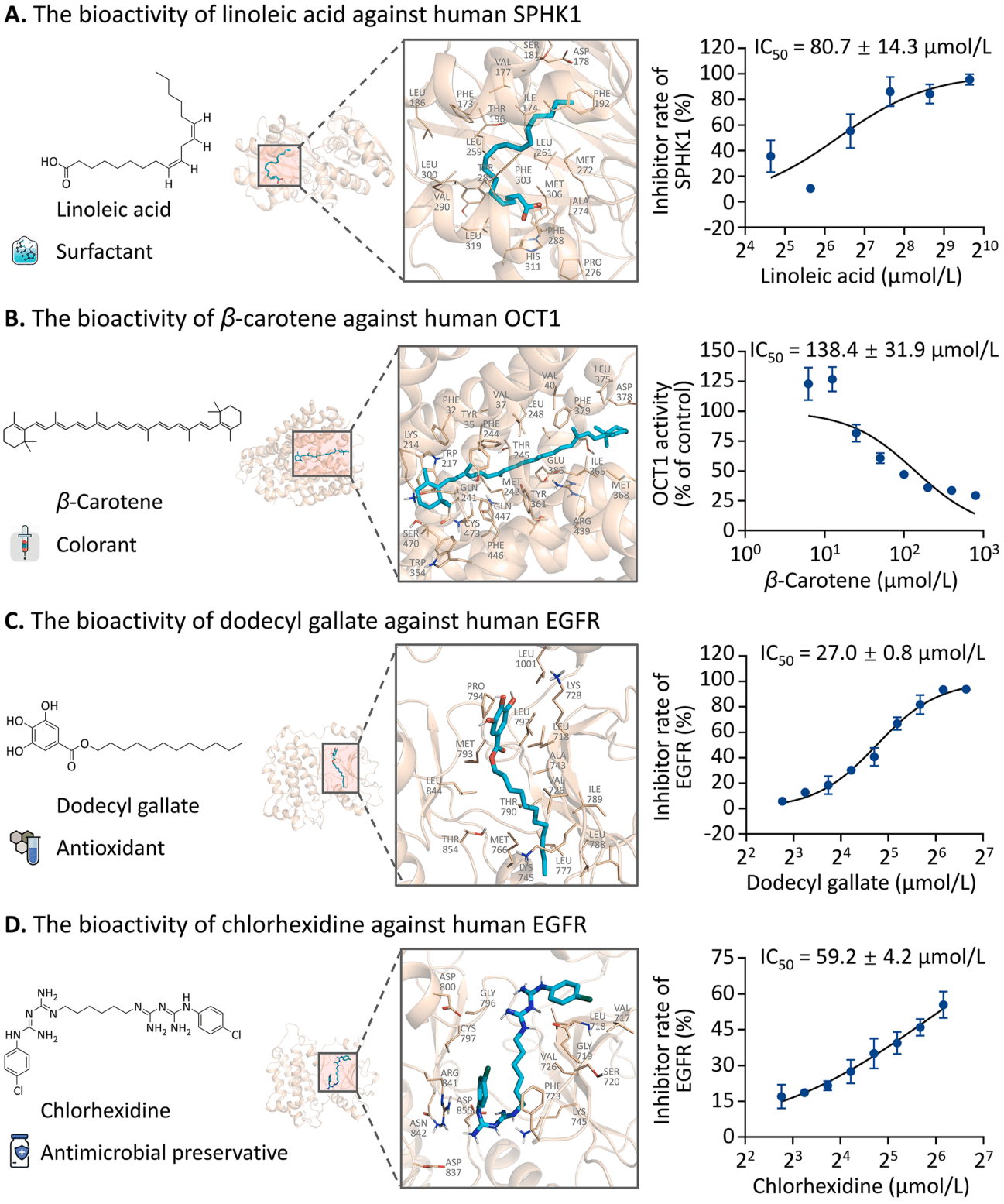


Figure 5 The experimentally validated bioactivities of DIGs against human proteins. (A) The bioactivity of linoleic acid against human SPHK1. The crystal structure of SPHK1 used for molecular simulation was extracted from the PDB (PDB ID: 4V24). (B) The bioactivity of β-carotene against human OCT1. The crystal structure of OCT1 used for molecular simulation was extracted from the PDB (PDB ID: 8ET8). (C) The bioactivity of dodecyl gallate against human EGFR. The crystal structure of EGFR used for molecular simulation was extracted from the PDB (PDB ID: 8SC7). (D) The bioactivity of chlorhexidine against human EGFR. The crystal structure of EGFR used for molecular simulation was extracted from the PDB (PDB ID: 8SC7). All error bars represented mean ± SD. The detailed binding modes were visualized in PyMOL (middle panel). DIGs structures were shown in cyan. The residues within 5.0 Å of DIGs were shown.

Please cite this article as: Mou Minjie et al., Unveiling the bioactive landscape of drug inactive ingredients (DIGs) using deep transfer learning, Acta Pharmaceutica Sinica B, https://doi.org/10.1016/j.apsb.2026.01.042

respective known protein families in the training data. These findings further demonstrated the robust generalization capability of *TransDIG*.

Generally, a 10 μmol/L threshold is widely adopted as the activity cutoff for identifying lead compounds, a value established by considering multiple factors such as therapeutic dosing, and the achievable drug concentrations *in vivo* or within cells[58]. In contrast, many of DIGs are administered at substantially higher doses than the active pharmaceutical ingredient (API). Although the systemic exposure to most DIGs is generally limited at the regulated maximum levels in individual medications, the widespread and chronic use of multiple drugs, often containing overlapping DIGs, can lead to cumulative DIG intake exceeding the permitted amount in any single product, thereby increasing overall systemic exposure. This situation may result in DIG-drug interactions between DIGs and APIs from different formulations, potentially leading to additive, synergistic, or antagonistic effects. For instance, certain solubilizing agents or surfactants might alter the bioavailability or metabolic profile of co-administered drugs, potentially resulting in unexpected pharmacokinetic outcomes.

Furthermore, the pervasive presence of many DIGs in foods, drinks, and cosmetics may significantly exacerbate total exposure. For instance, β-carotene is also widely present in daily diets, with particularly high concentrations found in foods such as mangoes, apples, and persimmons[59,60]. Therefore, although the bioactivities of DIGs discovered were relatively low compared to those of drug-like molecules, their ubiquitous presence in drug formulations and high-dose usage in daily life should raise significant concerns among drug developers and users. Thus, these validated DIG bioactivities might hold implications for optimizing drug formulation.

To fully assess DIG safety, future work should extend the current framework of *TransDIG* to predict DIG-drug interactions. A promising approach is the prediction of protein-mediated indirect interaction, which integrates DIG-protein and drug-protein interaction predictions to identify potential interference, such as competitive effects at shared targets like cytochrome P450 enzymes or transporters. Another approach is to build a joint model for the direct prediction of DIG-drug interactions. Both of these strategies rely on the availability of known DIG-drug interaction data. Crucially, incorporating quantitative information on DIG dosage and formulation concentration is essential, as the biological effects of a DIG are often dose-dependent. Therefore, future efforts to collect more comprehensive quantitative data on DIG and drug usage in clinical formulations, along with experimentally validated DIG-drug interactions, are critically important.

3. Experimental

3.1. Datasets collection and data processing

3.1.1. Datasets collection for dual-phase model training

The training data for the *TransDIG* model consisted of two main components, namely bioactivity data of BAMs from BindingDB database[61] and bioactivity data of DIGs from the ACDINA database[25]. The former BindingDB dataset was used for pre-training to develop the *PreDIG* model, while the latter ACDINA dataset was used for fine-tuning to obtain the *TransDIG* model.

To build a generalized deep learning model applicable to all types of proteins, we selected the BindingDB database (version 2024) to construct a large-scale BAM bioactivity dataset for pre-training the *PreDIG* model. The data preprocessing procedure mainly included the following steps: *a*) Data lacking explicit ligand names and SMILES or protein names and sequences were removed; *b*) Pairs with more than one protein chain in the target were excluded to ensure that the remaining data represented single-chain protein activity; *c*) Data with binding activity metrics of IC_{50}, EC_{50}, K_i, and K_d were selected and quantitatively standardized to the micromolar (μmol/L) unit; *d*) Data without any literature support were excluded; *e*) Pairs with activity values $<$ 1 μmol/L were labeled as high-affinity positive data[38], and those with activity values $>$ 10 μmol/L were labeled as low-affinity negative data[62], establishing a clear boundary to enhance the model's discrimination between active and inactive pairs; *f*) pairs with activity values between 1 and 10 μmol/L were excluded to prevent potential confusion caused by ambiguous data; *g*) Pairs with opposite labels were excluded from model construction, and duplicate pairs were de-duplicated.

The bioactivity data of DIGs were primarily collected from the ACDINA database, which was constructed by our group, and processed using the same procedure described above. Additionally, to ensure the comprehensiveness of the DIG activities, a complete list of DIGs was first retrieved from ACDINA. Then, all activity data relevant to DIGs were extracted from the cleaned BindingDB dataset and integrated into the ACDINA dataset. Ultimately, as shown in Fig. 1A, the BindingDB dataset used for pre-training contained 996,265 positive samples and 398,059 negative samples, involving over 7600 unique proteins and 0.84 million BAMs. Meanwhile, the ACDINA benchmark dataset used for fine-tuning included 182 positive and 813 negative samples, covering 499 proteins and 222 DIGs. Notably, there was no overlap in compounds between the BindingDB and ACDINA datasets to prevent information leakage during model evaluation.

In addition, we collected 134 experimentally validated bioactivities from a previous reliable study, involving 38 DIGs and 44 proteins[10]. Following the same label annotation rule described above, an independent test set was generated from these data with 32 positive and 53 negative samples (as provided in Table S13), which was used to assess the generalization capability of *TransDIG*. This independent test set included 31 DIGs and 39 human proteins, with 15 proteins absent from the ACDINA dataset forming a zero-shot protein subset comprising 40 activities. To enable a more objective evaluation, we partitioned the entire test set into subsets based on difficulty levels, following the data splitting strategies proposed in two previous studies on evaluation schemes for pair-input computational predictions[39,40]. Specifically, we first assessed the similarity of all DIGs and proteins in the test set relative to the entire fine-tuning dataset. Protein sequence identity was computed using DIAMOND[35], and molecular structural similarity was measured using the Tanimoto coefficient[32]. As shown in Fig. S2C, the analysis revealed that 4 proteins in the test set were dissimilar to the 499 proteins in the fine-tuning dataset (maximum sequence identity $<$30%), and 11 DIGs were dissimilar to the 222 DIGs in the fine-tuning dataset (maximum Tanimoto similarity $<$0.5). According to the proposed test split strategy based on the defined difficulty levels, the entire test set was divided into two subsets: C1, 59 test pairs involving DIGs and proteins that are either seen or similar to those in the fine-tuning set; C2, 26 test pairs where either the DIG or the protein is dissimilar to those in the training set. Notably, there were no test samples in which both the DIG and the protein were dissimilar to the training set. Furthermore, the subset C2 included 15 pairs with dissimilar DIGs (5 positives, 10 negatives) and 11 pairs with dissimilar proteins (6 positives, 5 negatives).

Please cite this article as: Mou Minjie et al., Unveiling the bioactive landscape of drug inactive ingredients (DIGs) using deep transfer learning, Acta Pharmaceutica Sinica B, https://doi.org/10.1016/j.apsb.2026.01.042

3.1.2. Data preparation for mapping the bioactive landscape of DIGs

To uncover the bioactive landscape of DIGs, we collected information on commonly used DIGs and clinically significant human proteins from public databases and literatures. Specifically, by analyzing the usage of DIGs in drug formulations as provided by the ACDINA database[25], we gathered information on 152 commonly used DIGs spanning ten main functional classes, including their names, functional classifications and SMILES (as detailed in Table S14). Additionally, through an extensive literature review, four categories of clinically important proteins were identified, namely DMEs, DTPs, DTTs, and TRPs. The relevant data were collected from the INTEDE[63], VARIDT[64], and TTD[65], which were previously constructed by our group, yielding 111 DMEs, 42 DTPs and 460 approved DTTs. Moreover, 54 TRPs were obtained from a prior study focusing on assessing drug safety[66]. The sequences of these proteins was downloaded from UniProt[67] for subsequent analysis of DIGs' bioactive landscape.

3.2. Model architecture of TransDIG

The architectures of *TransDIG* and *PreDIG* are identical, though their parameters differ. Both models comprise a protein encoder, a compound encoder, an interaction decoder, and a classifier, as illustrated in Fig. 1B.

3.2.1. Protein encoder

To transform protein sequences into sequential representations, the natural language processing technology Word2Vec was employed to generate residue encodings[68]. The Word2Vec model was pre-trained on all human proteins from UniProt[67] and all proteins extracted from the above BindingDB dataset. Following 30 epochs of training, each protein sequence was converted into overlapping 3-g amino acid subsequences, with each 3-g subsequence being represented as a 100-dimensional vector through the Word2Vec embedding approach.

The residue encoding matrix of a protein was then passed through the protein encoder. In contrast to the original transformer architecture, the protein encoder of *TransDIG* employed a GatedCNN as a replacement for the self-attention layers[69]. The residue encoding matrix of a protein was first transformed into an $L \times 64$ matrix using a FC layer and subsequently fed into the GatedCNN. Each gated convolution layer consisted of a convolution unit and a gated linear unit. The convolution unit was essentially a 1D convolutional layer (Conv1D) that captured the contextual representation of residues with local biases and learned global protein features by assembling local features. The gated linear unit improved the model's ability to handle nonlinear information and generate more informative sequence representations. As reported, the GatedCNN performed exceptionally well in feature extraction, speeding up training, significantly shortening training time, and preventing gradient vanishing issues[69]. The GatedCNN of *TransDIG* comprised a total of l gated convolution layers, with the computation formula for layer i shown in Eq. (1):

$$h_i\ (\boldsymbol{X}) = (\boldsymbol{X} * \boldsymbol{W}_1 + \boldsymbol{b}_1) \bigotimes \sigma(\boldsymbol{X} * \boldsymbol{W}_2 + \boldsymbol{b}_2) \tag{1}$$

where $\boldsymbol{X} \in \mathbb{R}^{n \times m_1}$ is the input of layer i, $\boldsymbol{W}_1 \in \mathbb{R}^{k \times m_1 \times m_2}$, $\boldsymbol{W}_2 \in \mathbb{R}^{k \times m_1 \times m_2}$, $\boldsymbol{b}_1 \in \mathbb{R}^{m_2}$ and $\boldsymbol{b}_2 \in \mathbb{R}^{m_2}$ are trainable parameters, l is the number of gated convolution layers, n is the length of protein sequence, m_1 and m_2 are the dimension of input and hidden features of each gated convolution layer, respectively, k is the kernel size of Conv1D, σ refers to the sigmoid function and $\bigotimes$ represents the element-wise product between matrices. In this study, l is 3, m_1 is 64, m_2 is 64 and k is 7. Additionally, the protein encoder incorporated the residual connection and layer normalization to address the over-smoothing issue[70]. The resulting $L \times 64$ matrix from the protein encoder constituted the final protein embedding.

3.2.2. Compound encoder

The topological graph structure of each compound was constructed based on the compound SMILES. Each compound atom was initially encoded as a 34-dimensional vector using the RDKit package in Python, resulting in the feature matrix for all atoms. The list of encoded features for each atom was summarized in Supporting Information Table S19. Next, an adjacency matrix was constructed based on the covalent bonds between atoms. A single layer of GCN was employed to aggregate neighboring features and update atom features, thereby extracting the compound embedding. After processing by the GCN, each compound (DIG or BAM) was represented as an $a \times 64$ matrix, where a denoted the number of atoms.

3.2.3. Interaction decoder

The interaction decoder of our *TransDIG* was designed to capture the physical interaction features between compounds and proteins. It was essentially a modified transformer decoder that incorporated the multi-head cross-attention algorithm. The cross-attention transformer has been widely used in various interaction prediction tasks, such as predicting antigen binding specificity to both HLA and TCR molecules[71]. The interaction decoder module consisted of three decoder layers, each of which contained a self-attention layer, a cross-attention layer, and a feedforward layer. Each attention layer required three inputs: query ($\boldsymbol{Q}$), key ($\boldsymbol{K}$), and value ($\boldsymbol{V}$), where the attention weights were calculated using $\boldsymbol{Q}$ and $\boldsymbol{K}$. The computation formula for attention was presented in Eq. (2):

$$\text{attention}\ (\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{2}$$

where d_k is a scaling factor depending on the dimension of hidden layer and the number of heads. For the first decoder layer, the self-attention mechanism was first applied to extract updated compound features. $\boldsymbol{Q}$, $\boldsymbol{K}$ and $\boldsymbol{V}$ for this step were all computed from the compound embedding. Since the order of atom vectors did not affect modeling, positional embeddings were removed in *TransDIG*. Then, the cross-attention layer was adopted to capture the interactions between compounds and proteins. Here, $\boldsymbol{Q}$ was calculated from the output of self-attention layer, while $\boldsymbol{K}$ and $\boldsymbol{V}$ were derived from protein embedding. Notably, the mask operation in the original transformer framework was adapted in the interaction decoder to ensure access to the entire sequence information. Next, a feedforward layer was used to further refine the features, enhancing their expressiveness through nonlinear transformation[72]. After each attention layer and feedforward layer, the residual connection and layer normalization were further applied. In the remaining two decoder layers, the $\boldsymbol{Q}$, $\boldsymbol{K}$ and $\boldsymbol{V}$ for the self-attention layer were all computed from the output of previous decoder layer, while the rest of the process remained unchanged. The attention weights for all residues in a protein sequence were calculated using $\boldsymbol{Q}$ and $\boldsymbol{K}$ from the last cross-attention layer.

After processing by the last decoder layer, an $a \times 64$ interaction matrix was obtained, where each interaction vector $\boldsymbol{x}_i$ represented the interaction between atom i and the entire protein. The weights for all interaction vectors were then computed by

Please cite this article as: Mou Minjie et al., Unveiling the bioactive landscape of drug inactive ingredients (DIGs) using deep transfer learning, Acta Pharmaceutica Sinica B, https://doi.org/10.1016/j.apsb.2026.01.042

applying the softmax function to their moduli, and all interaction vectors were then aggregated using a weighted sum based on these weights to derive the final pair embedding $\boldsymbol{z}$, as described in Eq. (3). Inspired by a previous study[73], the attention score for atom i was determined by the cosine similarity between $\boldsymbol{x}_i$ and $\boldsymbol{z}$, with the calculation formula provided in Eq. (4). The higher the similarity between $\boldsymbol{x}_i$ and $\boldsymbol{z}$, the more crucial the contribution of atom i to the interaction, indicating its greater importance in the interaction.

$$\boldsymbol{z} = \sum_{i=1}^{a} \alpha_i \boldsymbol{x}_i \tag{3}$$

$$\text{cosine}\ (\boldsymbol{x}_i, \boldsymbol{z}) = \frac{\boldsymbol{x}_i \cdot \boldsymbol{z}}{\|\boldsymbol{x}_i\|_2 \times \|\boldsymbol{z}\|_2} \tag{4}$$

where a is the number of atoms in a compound, α_i is the weight of interaction vector $\boldsymbol{x}_i$, $\boldsymbol{z}$ is the final pair embedding.

3.2.4. Binary classifier

The pair embedding $\boldsymbol{z}$ was further passed through a binary classifier for interaction prediction. The binary classifier was consisted of two FCs with 256 and 2 neurons, respectively. The outputs of the first FC were non-linearly converted based on the Rectified Linear Unit (ReLU) activation function[74]. Finally, the softmax function normalized the final layer's output into two-class probabilities, and the binary cross-entropy loss (Eq. (5)) was calculated to optimize the model parameters under supervised training.

$$\text{Loss} = -\sum_{i=1}^{N} y_i \log \widehat{y_i} + (1 - y_i)\ \log(1 - \widehat{y_i}) \tag{5}$$

where y_i is the true label of pair i, $\widehat{y_i}$ is the predicted probability of pair i being positive, and N is the total number of all predicted pairs.

3.3. Training, evaluation and implementation

The training of *TransDIG* consisted of two phases: pre-training and fine-tuning. First, the stratified sampling strategy was used to divide the BAM activity data from BindingDB dataset into training and validation sets at a 4:1 ratio, which were used for pre-training and validating the *PreDIG* model, respectively. When the model achieved the best performance on validation dataset, the *PreDIG* model was obtained and its model parameters were saved. Then, the parameters of *PreDIG* were used to initialize the *TransDIG* model, and the parameters of DIG encoder and binary classifier of *TransDIG* were fine-tuned using DIG activity data.

Due to the limited DIG activity data, we first employed 5-fold CV to evaluate the performance of *TransDIG*. To comprehensively assess the generalization ability of *TransDIG*, three distinct tasks were designed using the ACDINA benchmark dataset, namely 5-fold CV with random pair split, 5-fold CV with zero-shot protein setting, and 5-fold CV with zero-shot DIG setting. Specifically, in the 5-fold CV with random pair split task, all DIG activity data from the ACDINA dataset were randomly divided into five folds. Among these five folds, four folds were used to train *TransDIG*, with the remaining fold serving as the validation dataset for evaluation. This procedure was repeated five times, ensuring that each fold served as the validation dataset once. For the 5-fold CV with zero-shot protein setting and 5-fold CV with zero-shot DIG setting, all proteins and DIGs in the ACDINA dataset were randomly divided into five folds, respectively. The activity data involving four folds of proteins or DIGs were used as the training set, while the activity data associated with the remaining fold of proteins or DIGs served as the validation set. The average of the 5-fold CV results was calculated to obtain the overall evaluation result. Based on these results, the best hyperparameters were selected, and the best average result was reported. Using the optimal hyperparameters, the final *TransDIG* model was re-trained on the entire ACDINA dataset and evaluated on an independent test dataset. An early stopping strategy based on the AUROC metric with a patience of ten epochs was used to reduce overfitting and training costs[75]. To facilitate training convergence and enhance generalization ability, regularization methods including dropout and weight decay were used during the training of *TransDIG*[76].

During the pre-training phase, the learning rate, dropout rate, and batch size were set to 0.001, 0.2, and 256, respectively. During fine-tuning, these three key hyperparameters were optimized based on the model's predictive performance on the validation dataset. The optimization results for the 5-fold CV with random pair split were displayed in Supporting Information Fig. S6. The optimal hyperparameter combination was determined as a learning rate of 0.0008, a dropout rate of 0.1, and a batch size of 32. A complete summary of all hyperparameter settings for *TransDIG* was provided in Supporting Information Table S20.

The LookAhead and RAdam optimizers were applied to train *TransDIG*, incorporating a weight decay of 1E-4. Four metrics were used to evaluate the performance: accuracy (ACC), Matthews correlation coefficient (MCC), area under the receiver operating characteristic curve (AUROC), and area under the precision-recall curve (AUPRC). MCC is very comprehensive because it addresses the imbalance in interacting data[77]. All four metrics were calculated using Scikit-learn[78]. *TransDIG* was implemented in Python 3.7.12 and Pytorch 1.2.0 (http://pytorch.org/). All models were developed on the platform with Intel(R) Xeon(R) Gold 6132 CPU @ 2.60 GHz, NVIDIA(R) Tesla(R) V100 32 GB GPU and 263 GB RAM on CentOS Linux release 7.9.2009 (Core).

3.4. Baseline models

The prediction of DIG-protein interactions is essentially a CPI prediction task. Three advanced CPI prediction methods, namely MolTrans[24], HyperAttentionDTI[33] and DeepDTAGen[34], were thus employed for performance comparison with *TransDIG*. *TransDIG* differs from these models in its overall architecture. Specifically, MolTrans decomposes compounds and proteins into substructures, uses a transformer to extract their respective features, and finally employs a convolutional neural network (CNN) to learn the substructure-level interactions between the compound and the protein. The key difference between HyperAttentionDTI and *TransDIG* lies in the compound encoder. HyperAttentionDTI uses a CNN instead of a GCN for compound learning, which limits its ability to capture molecular topological structure. In comparison with *TransDIG*, DeepDTAGen also uses a GCN and a GatedCNN to extract features of small molecules and proteins, respectively, but it does not employ the attention mechanism to model interactions between compound atoms and protein residues. Moreover, since DeepDTAGen was originally designed to simultaneously predict binding affinity and generate novel molecules, we modified its affinity prediction module for fair comparison. Specifically, the final dense layer of its affinity prediction module was adjusted to output two values, which were then passed through a sigmoid function to yield interaction probability. All

Please cite this article as: Mou Minjie et al., Unveiling the bioactive landscape of drug inactive ingredients (DIGs) using deep transfer learning, Acta Pharmaceutica Sinica B, https://doi.org/10.1016/j.apsb.2026.01.042

three CPI prediction models were trained using the same dataset and training scheme as *TransDIG*. They were pre-trained on BAM activity data, followed by fine-tuning compound encoder and classifier on DIG activities. Moreover, the performance of TransformerCPI2.0 was also evaluated[38]. TransformerCPI 2.0 was a model pre-trained by its developers, differing from *TransDIG* primarily in the use of TAPE-BERT for protein representation and its adoption of a transformer the protein encoder. However, since its training code was not provided, the model could not be retrained, and thus only a single prediction result was obtained by loading the pre-trained model parameters.

Additionally, we constructed three DIG-protein interaction prediction models based on commonly used machine learning algorithms (SVM, RF, and XGBoost) to compare and evaluate the performance of *TransDIG*. All DIG-protein pairs from the ACDINA dataset were represented by concatenating MorganFPs (radius 2, 1024 bits) of DIGs with CTD encoding features of proteins[79,80], and min-max normalization was applied to scale the encoding vectors to the range [0, 1]. These baseline models adopted the same training and evaluation scheme as *TransDIG*. For SVM, the radial basis function kernel was used, and its hyperparameters C and gamma were systematically optimized using grid search. Key hyperparameters of RF (including n_estimators, max_depth and max_features) and XGBoost (including n_estimators, learning_rate, max_depth, and gamma) were optimized.

3.5. Methods for the experimental validation

3.5.1. Materials and reagents

ADP-Glo kinase assay kit (Promega biotech, USA); ATP (Promega biotech, USA); β-carotene (HY-N0411, 7235-40-7, MedChemExpress, China); bicinchoninic acid assay (BCA) (Beyotime, China); chlorhexidine (T1000, 55-56-1, Targetmol, USA); dodecyl gallate (T20648, 1166-52-5, Targetmol, USA); Dulbecco's modified Eagle's medium (DMEM), fetal bovine serum (FBS) (Gibco, USA); EGFR kinase assay kit (BPS bioscience, USA); Hanks' balanced salt solution (HBSS) (Biosharp, China); kinase-Glo luminescent kinase assay (Promega biotech, USA); 4-(4-(dimethylamino)styryl)-*N*-methylpyridinium iodide (ASP^+) (Aldrich, China); linoleic acid (HY-N0729, 60-33-3, MedChemExpress, China); penicillin/streptomycin (New cell & molecular biotech, China); sodium dodecyl sulfate (SDS) (Sinopharm chemical reagent, China); SPHK1 (Carna biosciences, Japan); D-sphingosine (Aldrich, China).

3.5.2. SPHK1 activity assay

The activity of SPHK1 was measured according to the protocol established by Wuxi AppTec. The production of ADP in the SPHK1 reaction was quantified using the ADP-Glo kinase assay kit to assess SPHK1 activity. In a 384-well plate, 5 μL reaction mixture, containing 2 μL SPHK1 (0.2 ng), 2 μL substrate/ATP mix (10 μmol/L ATP, 100 μmol/L D-sphingosine) and 1 μL linoleic acid solution, was added to each well, followed by incubation at 25 °C for 60 min. Then, 5 μL of ADP-Glo reagent was added to each well, and the plate was incubated at 25 °C for 40 min to terminate the kinase reaction and deplete residual ATP. Next, 10 μL of kinase detection reagent was added to each well, and the plate was incubated at 25 °C for 30 min to quantify the ADP generated in the kinase reaction. The luminescence values of each well were measured using a multifunctional microplate reader to analyze the inhibitory effect of the linoleic acid.

3.5.3. Construction of MDCK-hOCT1 cells

MDCK cells (parental Madin-Darby canine kidney II) were obtained from Peking Union Medical College Hospital. Following previously established techniques[81,82], MDCK-hOCT1 cells stably expressing human OCT1 (hOCT1) were generated using the plasmid pcDNA3.1(+) containing the full-length hOCT1 cDNA sequence (GenBank accession number NM_003057). The cells were cultured in DMEM supplemented with 10% FBS and 1% penicillin/streptomycin, and maintained in a humidified incubator at 37 °C with 5% CO_2.

3.5.4. OCT1 activity assay

Based on the previously established method[81,83], the activity of OCT1 in stably transfected cells was validated using the model substrate ASP^+. ASP^+ is a compound known to be transported into cells *via* OCT1 and can be used to assess the inhibitory effect of β-carotene on OCT1 activity by measuring its accumulation in MDCK-hOCT1 cells[82]. MDCK-hOCT1 cells were seeded in 24-well plates at a density of 2×10^5 cells per well and cultured for 3 days at 37 °C in a 5% CO_2 atmosphere. The culture medium was removed and the cells were washed twice with pre-warmed HBSS (pH = 7.4) at 37 °C. Then, 200 μL of incubation buffer containing 10 μmol/L ASP^+ was added to each well. For the inhibition group, the incubation buffer additionally contained β-carotene at specified concentrations, while the control group contained no β-carotene. After incubation at 37 °C for 5 min, the incubation buffer was rapidly removed, and cells were washed with ice-cold PBS to terminate the accumulation process. Cells were then lysed with 100 μL of 0.1% SDS and thoroughly homogenized by repeated pipetting. The fluorescence intensity of ASP^+ in the lysates was measured to determine its cellular uptake. Additionally, the total protein content in each well was quantified using a BCA assay kit. The results of ASP^+ uptake were normalized to the total protein content in each well.

3.5.5. EGFR activity assay

According to the manufacturer's instructions, the effects of two DIGs (namely dodecyl gallate and chlorhexidine) on the biological activity of EGFR were evaluated using an EGFR activity assay kit. In a 96-well plate, 25 μL of mixture, which containing 6 μL of Kinase Buffer (5 ×), 1 μL of ATP (500 μmol/L), 1 μL of polyglutamic acid-tyrosine substrate (Glu:Tyr = 4:1, 10 mg/mL) and 17 μL of water, was added. Subsequently, 5 μL of DIG solution at different concentrations was added to each well, followed by the initiation of the reaction using 20 μL of diluted EGFR enzyme. The plate was then incubated at 30 °C for 40 min. After the reaction, Kinase-Glo MAX reagent was added to each well, and the plate was incubated at room temperature for 15 min. Finally, the luminescence values of each well were measured using a multifunctional microplate reader to evaluate the inhibitory effects of two DIGs.

3.6. Molecular docking and molecular dynamics simulation

Molecular docking was performed using the Induced Fit Docking embedded in Schrödinger software suite[84], employing the default parameters to predict the ligand−receptor binding poses. Receptor structures were prepared with the Protein Preparation Wizard, and grid box dimensions were defined based on the original ligands in the binding site. The 3D structures of the studied ligands, retrieved from the PubChem database[85], were processed using LigPrep[86], employing the OPLS-2005 force field for energy minimization.

Please cite this article as: Mou Minjie et al., Unveiling the bioactive landscape of drug inactive ingredients (DIGs) using deep transfer learning, Acta Pharmaceutica Sinica B, https://doi.org/10.1016/j.apsb.2026.01.042

The ionization states were assigned using Epik[87] at a physiological pH of 7.0 ± 2.0. The DIG-protein binding poses were visualized using PyMOL (https://pymol.org/). The DIG-protein interactions were visualized using Schrödinger.

MD simulations were performed using the Amber 2020 software package[88]. Prior to simulation, the system underwent energy minimization and equilibration procedures. The energy optimization was carried out in two consecutive steps. First, a positional restraint of 10.0 kcal/(mol·Å^2) was applied to both the protein and the ligand, followed by a second step in which all restraints were removed to allow full relaxation of all atoms. The energy minimization consisted of a total of 10,000 steps, including 5000 steps of the steepest descent and 5000 steps of the conjugate gradient. Subsequently, the system was subjected to a stepwise equilibration protocol. Under the NVT ensemble, with the protein and ligand restrained by a force constant of 10.0 kcal/(mol·Å^2), the system was gradually heated to 100 K and then further raised to 310 K. Following this step, a 5-ns MD simulation was performed under the NPT ensemble (T = 310 K, P = 1 atm), during which the positional restraints were gradually released, allowing the system to fully equilibrate under periodic boundary conditions. Finally, a 100-ns MD simulation was conducted using the GPU-accelerated PMEMD, with three independent replicates performed. In this section, a hierarchical agglomerative approach was applied to perform conformational clustering on the simulation trajectories from the equilibration phase, and the representative conformation was extracted from the largest cluster. Structural analysis was performed using the cpptraj module in AmberTools, including RMSD calculation and hydrogen bond analysis.

3.7. Statistical analysis

The average results from 5-fold CV and the result from independent test set were used for benchmark evaluation. For all protein activity assays, each group included three independent experiments. All error bars in this study represented mean ± standard deviation (SD). All fitted curves and IC_{50} values were generated in the GraphPad Prism 8.0 using the 'inhibitor *vs*. normalized response' function. Mann−Whitney U test was used to determine the statistical significance.

Acknowledgments

This work was supported by Natural Science Foundation of Zhejiang (RG25H300001, China); National Key R&D Programs of China (2024YFA1307503, China); National Natural Science Foundation of China (22220102001 and 82373790); Priority-Funded Postdoctoral Research Project, Zhejiang Province, China (Grant No.: ZJ2024046); The Key Laboratory of Neural and Vascular Biology, Ministry of Education, Hebei Medical University, Shijiazhuang, China (Grant No.: NV20250011). The authors thank Information Technology Center and State Key Lab of CAD&CG, Zhejiang University, for providing computational resources.

Author contributions

Feng Zhu, Jianqing Gao, Yang Zhang, and Minjie Mou conceived the idea, designed the research and wrote the manuscript. Minjie Mou constructed the models. Jinsong Zhang, Xingang Liu, Hengbin Zhang, Xinyu Liu, and Tianyuan Zhang conducted the protein activity assays. Tingting Fu, Yimiao Zhu, and Tianle Niu collected the data. Hao Yang, Ziqi Pan, Xuedong Li, Yichao Ge, and Huaicheng Sun performed data analysis. All authors have approved the latest version of manuscript.

Conflicts of interest

The authors declare no competing interests.

Data availability

All benchmark datasets are available on GitHub (https://github.com/TransDIG-AI/TransDIG/). Source data of benchmark evaluations are fully provided in the Supporting Information of this manuscript. The information of all DIGs and the identified bioactive landscape are available as Supporting Tables for this paper.

Code availability

All codes of *TransDIG* are available on GitHub (https://github.com/TransDIG-AI/TransDIG/).

Appendix A. Supporting information

Supporting information to this article can be found online at https://doi.org/10.1016/j.apsb.2026.01.042.

References

1. Burdock GA, Carabin IG. Generally recognized as safe (GRAS): history and description. *Toxicol Lett* 2004;**150**:3−18.
2. Ruiz-Picazo A, Lozoya-Agullo I, González-Álvarez I, Bermejo M, González-Álvarez M. Effect of excipients on oral absorption process according to the different gastrointestinal segments. *Expet Opin Drug Deliv* 2021;**18**:1005−24.
3. Yuan H, Wu X, Wu Q, Chatoff A, Megill E, Gao J, et al. Lysine catabolism reprograms tumour immunity through histone crotonylation. *Nature* 2023;**617**:818−26.
4. Zani F, Blagih J, Gruber T, Buck MD, Jones N, Hennequart M, et al. The dietary sweetener sucralose is a negative modulator of T cell-mediated responses. *Nature* 2023;**615**:705−11.
5. Brown KA, Gould TD. Disassociating drug active ingredients from inactive: ketamine-like synaptic effects of a ketamine excipient. *Neuropsychopharmacology* 2024;**49**:301−2.
6. Zou L, Spanogiannopoulos P, Pieper LM, Chien HC, Cai W, Khuri N, et al. Bacterial metabolism rescues the inhibition of intestinal drug absorption by food and drug additives. *Proc Natl Acad Sci U S A* 2020;**117**:16009−18.
7. Reker D, Blum SM, Steiger C, Anger KE, Sommer JM, Fanikos J, et al. "Inactive" ingredients in oral medications. *Sci Transl Med* 2019;**11**:eaau6753.
8. Mai Y, Ashiru-Oredope DAI, Yao Z, Dou L, Madla CM, Taherali F, et al. Boosting drug bioavailability in men but not women through the action of an excipient. *Int J Pharm* 2020;**587**:119678.
9. Martinez-Mayorga K, Peppard TL, López-Vallejo F, Yongye AB, Medina-Franco JL. Systematic mining of generally recognized as safe (GRAS) flavor chemicals for bioactive compounds. *J Agric Food Chem* 2013;**61**:7507−14.
10. Pottel J, Armstrong D, Zou L, Fekete A, Huang XP, Torosyan H, et al. The activities of drug inactive ingredients on biological targets. *Science* 2020;**369**:403−13.
11. Kingwell K. The secret life of excipients. *Nat Rev Drug Discov* 2020;**19**:585.

12. Turner MA, Duncan JC, Shah U, Metsvaht T, Varendi H, Nellis G, et al. Risk assessment of neonatal excipient exposure: lessons from food safety and other areas. *Adv Drug Deliv Rev* 2014;**73**:89−101.
13. Pertea M, Salzberg SL. Between a chicken and a grape: estimating the number of human genes. *Genome Biol* 2010;**11**:206.
14. Tang J, Fu Q, Wang Y, Racette K, Wang D, Liu F. Vitamin E reverses multidrug resistance *in vitro* and *in vivo*. *Cancer Lett* 2013;**336**: 149−57.
15. Falah K, Zhang P, Nigam AK, Maity K, Chang G, Granados JC, et al. *In vivo* regulation of small molecule natural products, antioxidants, and nutrients by OAT1 and OAT3. *Nutrients* 2024;**16**:2242.
16. Reker D, Shi Y, Kirtane AR, Hess K, Zhong GJ, Crane E, et al. Machine learning uncovers food- and excipient-drug interactions. *Cell Rep* 2020;**30**:3710−3716.e3714.
17. Bai P, Miljković F, John B, Lu H. Interpretable bilinear attention network with domain adaptation improves drug−target prediction. *Nat Mach Intell* 2023;**5**:126−36.
18. Lu Z, Song G, Zhu H, Lei C, Sun X, Wang K, et al. DTIAM: a unified framework for predicting drug-target interactions, binding affinities and drug mechanisms. *Nat Commun* 2025;**16**:2548.
19. Singh R, Sledzieski S, Bryson B, Cowen L, Berger B. Contrastive learning in protein language space predicts interactions between drugs and protein targets. *Proc Natl Acad Sci U S A* 2023;**120**:e2220778120.
20. Yang SQ, Zhang LX, Ge YJ, Zhang JW, Hu JX, Shen CY, et al. *In-silico* target prediction by ensemble chemogenomic model based on multi-scale information of chemical structures and protein sequences. *J Cheminf* 2023;**15**:48.
21. Zdrazil B, Felix E, Hunter F, Manners EJ, Blackshaw J, Corbett S, et al. The ChEMBL database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res* 2024;**52**:D1180−92.
22. Ackloo S, Antolin AA, Bartolome JM, Beck H, Bullock A, Betz UAK, et al. Target 2035—an update on private sector contributions. *RSC Med Chem* 2023;**14**:1002−11.
23. Hadipour H, Li YY, Sun Y, Deng C, Lac L, Davis R, et al. GraphBAN: an inductive graph-based approach for enhanced prediction of compound-protein interactions. *Nat Commun* 2025;**16**:2541.
24. Huang K, Xiao C, Glass LM, Sun J. MolTrans: molecular interaction transformer for drug−target interaction prediction. *Bioinformatics* 2021;**37**:830−6.
25. Zhang C, Mou M, Zhou Y, Zhang W, Lian X, Shi S, et al. Biological activities of drug inactive ingredients. *Briefings Bioinf* 2022;**23**:bbac160.
26. Bento AP, Hersey A, Félix E, Landrum G, Gaulton A, Atkinson F, et al. An open source chemical structure curation pipeline using RDKit. *J Cheminf* 2020;**12**:51.
27. Yang X, Wang Y, Byrne R, Schneider G, Yang S. Concepts of artificial intelligence for computer-assisted drug discovery. *Chem Rev* 2019; **119**:10520−94.
28. Cai C, Wang S, Xu Y, Zhang W, Tang K, Ouyang Q, et al. Transfer learning for drug discovery. *J Med Chem* 2020;**63**:8683−94.
29. Liu H, Song Z, Zhang Y, Wu B, Chen D, Zhou Z, et al. *De novo* design of self-assembling peptides with antimicrobial activity guided by deep learning. *Nat Mater* 2025;**24**:1295−306.
30. Cui W, Xiang Y, Wang Y, Yu T, Liao XF, Hu B, et al. Deep multiview module adaption transfer network for subject-specific EEG recognition. *IEEE Transact Neural Networks Learn Syst* 2025;**36**:2917−30.
31. Mou M, Zhang Y, Qian Y, Zhou Z, Liao Y, Niu T, et al. druglikeFilter 1.0: an AI powered filter for collectively measuring the drug-likeness of compounds. *J Pharm Anal* 2025;**15**:101298.
32. He J, Nittinger E, Tyrchan C, Czechtizky W, Patronov A, Bjerrum EJ, et al. Transformer-based molecular optimization beyond matched molecular pairs. *J Cheminf* 2022;**14**:18.
33. Zhao Q, Zhao H, Zheng K, Wang J. HyperAttentionDTI: improving drug−protein interaction prediction by sequence-based deep learning with attention mechanism. *Bioinformatics* 2022;**38**:655−62.
34. Shah PM, Zhu H, Lu Z, Wang K, Tang J, Li M. DeepDTAGen: a multitask deep learning framework for drug−target affinity prediction and target-aware drugs generation. *Nat Commun* 2025;**16**:5021.
35. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015;**12**:59−60.
36. Lihan M, Lupyan D, Oehme D. Target−template relationships in protein structure prediction and their effect on the accuracy of thermostability calculations. *Protein Sci* 2023;**32**:e4557.
37. Somervuo P, Holm L. SANSparallel: interactive homology search against Uniprot. *Nucleic Acids Res* 2015;**43**:W24−9.
38. Chen L, Fan Z, Chang J, Yang R, Hou H, Guo H, et al. Sequence-based drug design as a concept in computational drug design. *Nat Commun* 2023;**14**:4217.
39. Hamp T, Rost B. More challenges for machine-learning protein interactions. *Bioinformatics* 2015;**31**:1521−5.
40. Park Y, Marcotte EM. Flaws in evaluation schemes for pair-input computational predictions. *Nat Methods* 2012;**9**:1134−6.
41. Chen Y, Xiao J, Zhu X, Fan X, Peng M, Mu Y, et al. Exploiting conjugated linoleic acid for health: a recent update. *Food Funct* 2025; **16**:147−67.
42. Sun L, Wang H, Du J, Wang T, Yu D. Ultrasonic-assisted extraction of grape seed procyanidins, preparation of liposomes, and evaluation of their antioxidant capacity. *Ultrason Sonochem* 2024;**105**:106856.
43. Billich A, Bornancin F, Dévay P, Mechtcheriakova D, Urtz N, Baumruker T. Phosphorylation of the immunomodulatory drug FTY720 by sphingosine kinases. *J Biol Chem* 2003;**278**:47408−15.
44. David OJ, Kovarik JM, Schmouder RL. Clinical pharmacokinetics of fingolimod. *Clin Pharmacokinet* 2012;**51**:15−28.
45. Lau P, Zhang G, Zhao S, Liang L, Zhang H, Zhou G, et al. Sphingosine kinase 1 promotes tumor immune evasion by regulating the MTA3−PD-L1 axis. *Cell Mol Immunol* 2022;**19**:1153−67.
46. Jin L, Zhu J, Yao L, Shen G, Xue BX, Tao W. Targeting SphK1/2 by SKI-178 inhibits prostate cancer cell growth. *Cell Death Dis* 2023;**14**:537.
47. Putthong C, Panmanee T, Charoensit P, Ross S, Tongpoolsomjit K, Viyoch J. Efficacy of natural β-carotene chewable tablets derived from banana (Musa AA) pulp in reducing UV-induced skin erythema. *Nutrients* 2024;**17**:65.
48. Teng YN, Sheu MJ, Hsieh YW, Wang RY, Chiang YC, Hung CC. β-Carotene reverses multidrug resistant cancer cells by selectively modulating human P-glycoprotein function. *Phytomedicine* 2016;**23**: 316−23.
49. Chen EC, Khuri N, Liang X, Stecula A, Chien HC, Yee SW, et al. Discovery of competitive and noncompetitive ligands of the organic cation transporter 1 (OCT1; SLC22A1). *J Med Chem* 2017;**60**: 2685−96.
50. Granados JC, Nigam SK. Organic anion transporters in remote sensing and organ crosstalk. *Pharmacol Ther* 2024;**263**:108723.
51. Ettefaghdoost M, Navirian H, Haghighi H. Effects of dietary β-carotene supplementation on growth performance, biochemical indices, hemato-immunological parameters, and physio-metabolic responses of the oriental river prawn (*Macrobrachium nipponense*). *Aquac Nutr* 2025;**2025**:5184405.
52. Cai D, Wang X, Wang Q, Tong P, Niu W, Guo X, et al. Controlled release characteristics of alkyl gallates and gallic acid from β-cyclodextrin inclusion complexes of alkyl gallates. *Food Chem* 2024;**460**: 140726.
53. Morin CB, Sasseville D. Expanding patch testing beyond the baseline series: usefulness of customized antimicrobials, vehicles, and cosmetics series. *Dermatitis* 2020;**31**:367−72.
54. Takeuchi M, Obara H, Furube T, Kawakubo H, Kitago M, Okabayashi K, et al. Efficacy of aqueous olanexidine compared with alcohol-based chlorhexidine for surgical skin antisepsis regarding the incidence of surgical-site infections in clean-contaminated surgery: a randomized superiority trial. *Br J Surg* 2025;**112**:znaf065.
55. Çoğulu D, Aşık A, Yılmaz Süslüer S, Yücel Er C, Topaloğlu A, Uzel A, et al. *In vitro* analysis of various mouthwashes: cytotoxic, apoptotic, genotoxic and antibacterial effects. *Clin Oral Invest* 2025; **29**:183.
56. Kim J, Park S, Ku BM, Ahn MJ. Updates on the treatment of epidermal growth factor receptor-mutant non-small cell lung cancer. *Cancer* 2025;**131**:e35778.

Please cite this article as: Mou Minjie et al., Unveiling the bioactive landscape of drug inactive ingredients (DIGs) using deep transfer learning, Acta Pharmaceutica Sinica B, https://doi.org/10.1016/j.apsb.2026.01.042

57. Hayes TK, Aquilanti E, Persky NS, Yang X, Kim EE, Brenan L, et al. Comprehensive mutational scanning of EGFR reveals TKI sensitivities of extracellular domain mutants. *Nat Commun* 2024;**15**:2742.
58. Wang Z, Liang L, Yin Z, Lin J. Improving chemical similarity ensemble approach in target prediction. *J Cheminf* 2016;**8**:20.
59. Karami S, Ali L, Ahsin M, Walsh SJ, Van Vliet S, Wisler A, et al. Effect of UV-A light dehydration on micronutrients of mango and apple: a comprehensive metabolomic study. *Food Chem* 2025;**493**:145858.
60. Lee CD, Lee HD, Ma GB, Lee B, Lee S. HPLC quantification of carotenoids in astringent and non-astringent persimmon peel and flesh. *Food Chem* 2025;**493**:145880.
61. Liu T, Hwang L, Burley SK, Nitsche CI, Southan C, Walters WP, et al. BindingDB in 2024: a FAIR knowledgebase of protein-small molecule binding data. *Nucleic Acids Res* 2025;**53**:D1633−44.
62. Chan WK, Zhang H, Yang J, Brender JR, Hur J, Özgür A, et al. GLASS: a comprehensive database for experimentally validated GPCR-ligand associations. *Bioinformatics* 2015;**31**:3035−42.
63. Zhang Y, Liu X, Li F, Yin J, Yang H, Li X, et al. INTEDE 2.0: the metabolic roadmap of drugs. *Nucleic Acids Res* 2024;**52**:D1355−64.
64. Yin J, Chen Z, You N, Li F, Zhang H, Xue J, et al. VARIDT 3.0: the phenotypic and regulatory variability of drug transporter. *Nucleic Acids Res* 2024;**52**:D1490−502.
65. Zhou Y, Zhang Y, Zhao D, Yu X, Shen X, Zhou Y, et al. TTD: therapeutic target database describing target druggability information. *Nucleic Acids Res* 2024;**52**:D1465−77.
66. Bowes J, Brown AJ, Hamon J, Jarolimek W, Sridhar A, Waldron G, et al. Reducing safety-related drug attrition: the use of *in vitro* pharmacological profiling. *Nat Rev Drug Discov* 2012;**11**:909−22.
67. Consortium U. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Res* 2023;**51**:D523−31.
68. Yue X, Wang Z, Huang J, Parthasarathy S, Moosavinasab S, Huang Y, et al. Graph embedding on biomedical networks: methods, applications and evaluations. *Bioinformatics* 2020;**36**:1241−51.
69. Li T, Zhao XM, Li L. Co-VAE: drug−target binding affinity prediction by co-regularized variational autoencoders. *IEEE Trans Pattern Anal Mach Intell* 2022;**44**:8861−73.
70. Rassil A, Chougrad H, Zouaki H. Augmented graph neural network with hierarchical global-based residual connections. *Neural Netw* 2022;**150**:149−66.
71. Yu C, Fang X, Tian S, Liu H. A unified cross-attention model for predicting antigen binding specificity to both HLA and TCR molecules. *Nat Mach Intell* 2025;**7**:278−92.
72. Luo G, Zhou Y, Sun X, Wang Y, Cao L, Wu Y, et al. Towards lightweight transformer *via* group-wise transformation for vision-and-language tasks. *IEEE Trans Image Process* 2022;**31**:3386−98.
73. Chen L, Tan X, Wang D, Zhong F, Liu X, Yang T, et al. TransformerCPI: improving compound−protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics* 2020;**36**:4406−14.
74. Eckle K, Schmidt-Hieber J. A comparison of deep networks with ReLU activation function and linear spline-type methods. *Neural Netw* 2019;**110**:232−42.
75. Wenzel J, Matter H, Schmidt F. Predictive multitask deep neural network models for ADME-Tox properties: learning from large data sets. *J Chem Inf Model* 2019;**59**:1253−68.
76. Mummadi SR, Al-Zubaidi A, Hahn PY. Overfitting and use of mismatched cohorts in deep learning models: preventable design limitations. *Am J Respir Crit Care Med* 2018;**198**:544−5.
77. Xu Z, Shen D, Kou Y, Nie T. A synthetic minority oversampling technique based on Gaussian mixture model filtering for imbalanced data classification. *IEEE Transact Neural Networks Learn Syst* 2024;**35**:3740−53.
78. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011;**12**:2825−30.
79. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model* 2010;**50**:742−54.
80. Rao HB, Zhu F, Yang GB, Li ZR, Chen YZ. Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res* 2011;**39**:W385−90.
81. Tu M, Sun S, Wang K, Peng X, Wang R, Li L, et al. Organic cation transporter 1 mediates the uptake of monocrotaline and plays an important role in its hepatotoxicity. *Toxicology* 2013;**311**:225−30.
82. Tu M, Li L, Lei H, Ma Z, Chen Z, Sun S, et al. Involvement of organic cation transporter 1 and CYP3A4 in retrorsine-induced toxicity. *Toxicology* 2014;**322**:34−42.
83. Li L, Tu M, Yang X, Sun S, Wu X, Zhou H, et al. The contribution of human OCT1, OCT3, and CYP3A4 to nitidine chloride-induced hepatocellular toxicity. *Drug Metab Dispos* 2014;**42**:1227−34.
84. *Schrödinger release 2025-1*. New York, NY: Maestro, Schrödinger, LLC; 2025.
85. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2025 update. *Nucleic Acids Res* 2025;**53**:D1516−25.
86. *Schrödinger release 2025-1*. New York, NY: LigPrep, Schrödinger, LLC; 2025.
87. *Schrödinger release 2025-1*. New York, NY: Epik, Schrödinger, LLC; 2025.
88. Lee TS, Allen BK, Giese TJ, Guo Z, Li P, Lin C, et al. Alchemical binding free energy calculations in AMBER20: advances and best practices for drug discovery. *J Chem Inf Model* 2020;**60**:5595−623.

Please cite this article as: Mou Minjie et al., Unveiling the bioactive landscape of drug inactive ingredients (DIGs) using deep transfer learning, Acta Pharmaceutica Sinica B, https://doi.org/10.1016/j.apsb.2026.01.042