nature communications

Article

https://doi.org/10.1038/s41467-025-66311-9

Semi-inductive dataset construction and framework optimization for practical drug target interaction prediction with ScopeDTI

Received: 8 April 2025

Accepted: 29 October 2025

Published online: 13 December 2025

Check for updates

Yigang Chen[1,2,3,7], Xiang Ji[1,2,7], Ziyue Zhang[1,2,7], Zihao Zhu[1,2,7], Yuming Zhou[1], Chang Su[1], Yang-Chi-Dung Lin[1,2,3,4], Hsi-Yuan Huang[1,2,4], Kangping Wei[1], Yi Lai[1,2], Ke Chen[1,2], Xingqiao Lin[1], Yangyi Zhang[1], Jiehui Fu[1], Yixian Huang[1,2], Shidong Cui[1,2], Shih-Chung Yen[1,2], Tao Zhang[3], Arieh Warshel[5] & Hsien-Da Huang[1,2,3,4,6]

Deep learning-based drug-target interaction (DTI) prediction methods have demonstrated strong performance; however, real-world applicability remains constrained by limited data diversity and modeling complexity. To address these challenges, we propose SCOPE-DTI, a unified framework combining a large-scale, balanced semi-inductive human DTI dataset with advanced deep learning modeling. SCOPE-DTI is constructed from 13 public repositories and expands data volume by up to 100-fold compared to common benchmarks such as the Human dataset. The SCOPE model integrates three-dimensional protein and compound representations, graph neural networks, and bilinear attention mechanisms to effectively capture cross domain interaction patterns and outperform state-of-the-art methods across various DTI prediction tasks. Additionally, SCOPE-DTI provides a user-friendly interface and database. We further demonstrate its effectiveness by experimentally identifying anticancer targets of two bioactive natural compounds. By offering comprehensive data, advanced modeling, and accessible tools, SCOPE-DTI accelerates drug discovery research.

Identifying drug-target interactions (DTIs) is essential for drug discovery, drug repurposing, toxicity prediction, and improving the success rate of clinical trials[1–4]. For decades, experimental approaches have dominated DTI identification due to their high accuracy and reliability[5]. However, their high cost and extensive time requirements limit scalability, especially considering the immense chemical and biological search spaces[6].

The accumulation of extensive experimental datasets has enabled the development of data-driven computational methods[7,8]. Machine learning-based approaches leverage these large-scale datasets to learn and generalize interaction patterns, facilitating rapid and cost-effective prediction of previously uncharacterized drug-target interactions (DTIs). These computational models provide efficient predictions, guiding experimental validation toward the most promising

[1]School of Medicine, The Chinese University of Hong Kong, Shenzhen, Longgang District, Shenzhen, Guangdong, China. [2]Warshel Institute for Computational Biology, School of Medicine, The Chinese University of Hong Kong, Shenzhen, Longgang District, Shenzhen, Guangdong, China. [3]Better Way Group—Chinese University of Hong Kong (Shenzhen) Warshel Joint Laboratory for skin health and active molecule innovation, Longgang District, Shenzhen, Guangdong, China. [4]Guangdong Provincial Key Laboratory of Digital Biology and Drug Development, The Chinese University of Hong Kong, Shenzhen, Longgang District, Shenzhen, Guangdong, China. [5]Department of Chemistry, University of Southern California, Los Angeles, CA, USA. [6]Department of Endocrinology, Key Laboratory of Endocrinology of National Health Commission, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, People's Republic of China. [7]These authors contributed equally: Yigang Chen, Xiang Ji, Ziyue Zhang, Zihao Zhu.
e-mail: huanghsienda@cuhk.edu.cn

candidate interactions, thereby accelerating drug discovery and significantly reducing associated costs[9].

Recent advances in deep learning have significantly advanced computational DTI prediction. Proteochemometrics (PCM), a widely used computational framework, represents drugs and targets as feature vectors to predict DTIs through supervised classification[10]. Embedding methods for molecules and proteins have rapidly evolved along two parallel fronts. Sequence-based approaches have advanced from one-dimensional convolutional neural networks[11,12] to powerful pre-trained foundation models (FM) like BERT that capture deep semantic context[6,13,14]. Concurrently, structure-awared methods have progressed from 2D topological representations using graph neural networks (GNNs)[15,16] to embeddings that explicitly model 3D spatial arrangements[17,18]. Alongside these advances in representation, model architectures have evolved from simple multilayer perceptrons (MLPs) to interactive mechanisms like interaction maps[19] and bilinear attention networks (BANs)[15]. These innovations have enabled state-of-the-art models to achieve notable performance improvements on standard benchmarks, including BindingDB[20,21], human[22], and KIBA datasets[23].

However, substantial challenges remain when applying these models in practical settings. In practical drug discovery applications, accurately predicting interactions between previously uncharacterized compounds and druggable targets is essential. Therefore, models must demonstrate the ability to achieve reliable predictive performance even when operating beyond the limitations of their training data. Chatterjee et al.[24] categorize drug-target interaction (DTI) prediction into three distinct scenarios: transductive, semi-inductive, and inductive. The inductive scenario, where both drugs and proteins are absent from the training set, is crucial for previously uncharacterized DTI discovery but remains highly challenging. Current state-of-the-art models under inductive conditions achieve only moderate predictive accuracy, with AUROC typically below 0.7 and correlation coefficients below 0.5[15,17], limiting their practical utility. Conversely, the transductive scenario, where both drugs and proteins are present in the training set but their interactions are unknown, yields high performance but is often compromised by overfitting and dataset biases[21,25], thus restricting generalization to real-world tasks.

In contrast, the semi-inductive scenario-particularly the prediction of interactions involving previously uncharacterized drugs against known proteins—offers significant potential compared to purely inductive or transductive approaches[17,24]. Although current models perform well in such scenarios, their success remains limited by the size and diversity of available data. Expanding and integrating existing datasets to encompass more druggable targets could substantially enhance predictive accuracy and real-world applicability, enlightening the practical value of the semi-inductive approach.

In this work, we present SCOPE-DTI, a unified framework that enhances the practical utility of DTI prediction models. We construct a large, well-balanced semi-inductive human DTI dataset by integrating data from 13 public repositories, expanding available training data 20–100-fold compared to standard benchmarks[20,21]. Using this dataset, the SCOPE model incorporates three-dimensional structural embeddings, graph neural networks, and a bilinear attention mechanism to capture cross-domain interaction patterns and achieve improved predictive performance over state-of-the-art methods across diverse tasks. In addition, we provide a user-friendly web interface and searchable database to facilitate accessibility. We further demonstrate the effectiveness of SCOPE-DTI by experimentally validating previously uncharacterized anti-cancer targets of Ginsenoside Rh1 and Celastrol, underscoring its real-world applicability in drug discovery.

Results

SCOPE dataset development

Our study aims to predict DTIs for previously uncharacterized compounds within a semi-inductive framework, addressing the critical question: Given any previously uncharacterized active compound, how can we accurately identify its binding targets across a comprehensive set of druggable proteins? To tackle this challenge, we constructed a large-scale, high-quality human-focused DTI dataset.

As illustrated in Fig. 1A, we aggregated and curated data from 13 primary DTI repositories, integrating information from over 20 sources in total, including referenced datasets (Supplementary Table 1). Each dataset entry comprises a protein, a compound, and an interaction label. Proteins were annotated using UniProt identifiers[26], retaining only human proteins, and categorized into pharmacological families (e.g., GPCRs, kinases) according to the IUPHAR database[27]. To enable structural modeling, we generated 3D protein structures using AlphaFold2[28]. Compounds were annotated with identifiers from PubChem[29] and ChEMBL[30], and their 3D structures were constructed using RDKit and optimized with the Merck Molecular Force Field (MMFF)[31]. Interaction labels were systematically assigned using standardized measurement-specific cutoff values derived from PubChem Bioactivity data, classifying interactions as positive (1) or negative (0) (Supplementary Fig. 1; Supplementary Table 2). Additional details on labeling methods are described in Supplementary Note 1.

Recognizing that target-level class imbalance—where certain proteins predominantly display interactions from one class—could bias semi-inductive predictions, we implemented a filtering approach (Fig. 1B). For proteins whose interactions exceeded 75% from a single class, we randomly removed interactions using stratified sampling, balancing the classes within a 50–75% range. Proteins with insufficient interactions or skewed class distributions were entirely excluded to maintain dataset integrity. Figure 1C illustrates that this filtering approach significantly mitigated imbalance, ensuring a more balanced distribution of interaction labels across targets (Supplementary Note 2 and Supplementary Fig. 2 provide further details).

To further assess the dataset's relevance to the druggable proteome, we analyzed its coverage of major pharmacological families (Fig. 1D). For kinases and nuclear hormone receptors (NHRs), SCOPE dataset demonstrates exceptional coverage, encompassing the vast majority of known proteins in these families, a feature largely preserved post-filtering. This high retention suggests a wealth of rich and relatively balanced interaction data for these well-studied targets. In contrast, while the initial dataset for ion channels (ICs) was substantial, the number of proteins was moderately reduced after filtering, indicating that some members suffer from sparser or highly skewed interaction data. Notably, the coverage for G-protein-coupled receptors (GPCRs) is more limited, reflecting a comparative scarcity of public interaction data and underscoring GPCRs as a critical family warranting further investigation and data curation efforts.

Finally, we quantitatively compared the resulting SCOPE dataset with widely used benchmarks (Fig. 1E), revealing a substantial increase in scale—20- to 100-fold more interactions, accompanied by significantly more proteins and compounds. Detailed dataset statistics are summarized in Supplementary Table 3. Overall, to our knowledge, SCOPE represents the large, most balanced semi-inductive human DTI dataset available, providing a robust foundation for developing and validating predictive models for previously uncharacterized drug-target interactions.

SCOPE framework

Figure 2 provides a schematic illustration of the proposed SCOPE-DTI framework for predicting drug-target interactions. The framework begins with the transformation of drug molecules and target proteins from the SCOPE dataset into their respective three-dimensional (3D) conformations, generated using RDKit for compounds and AlphaFold2 for proteins. To effectively capture the complex structural features of both proteins and compounds, their 3D structures are encoded using heterogeneous graph neural networks (HGNNs)[32,33] and geometric vector perceptrons (GVPs)[17,34], respectively. The protein encoding

https://doi.org/10.1038/s41467-025-66311-9

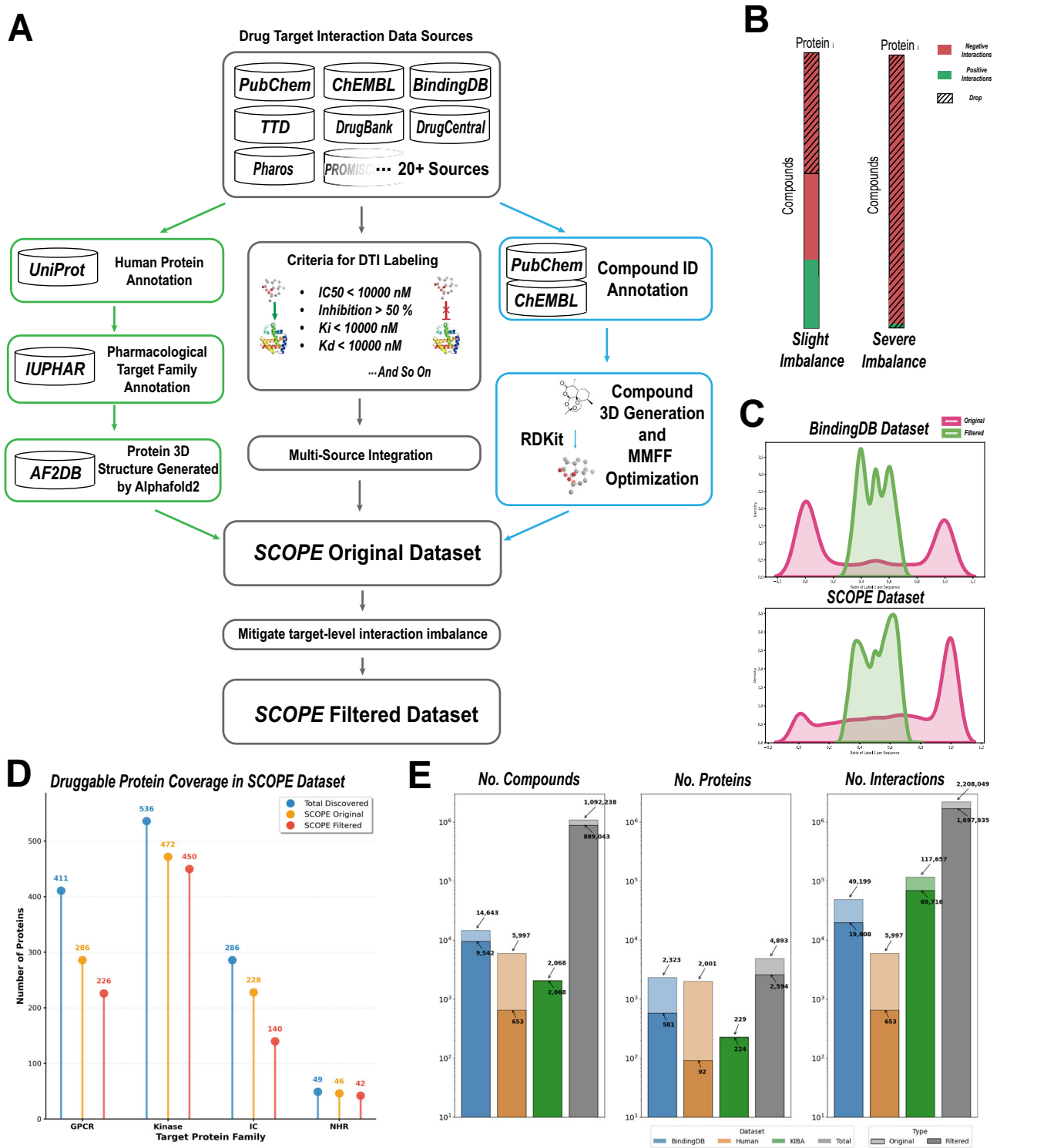

Fig. 1 | Overview of the SCOPE dataset construction pipeline and data characteristics. A Schematic representation of the data integration and preprocessing steps used to construct the SCOPE dataset. Molecular structures were visualized using ChemDraw and Chem3D. The protein structures were visualized using PyMOL. **B** Strategy to mitigate target-level imbalance: for proteins dominated by one class, we randomly remove interactions to achieve balance; in extreme cases, we discard all of that protein's interactions. **C** The effect of data filtering. The red and green distributions represent datasets before and after filtering, respectively. **D** The analysis of the coverage of different target protein families in the SCOPE dataset. **E** Comparison of data volume between SCOPE and previous datasets in terms of the number of compounds, targets, and interactions separately. Source data are provided as a Source Data file.

module simultaneously encodes the sequential relationships among residues and their spatial neighborhood information captured via radius graphs constructed from residue coordinates. The compound encoding incorporates atom-specific features (atom type, charge, hybridization state) defined by DGL-LifeSci's encoding scheme[35], along with spatial coordinates representing their 3D structural information. Further details of these encoding methods are provided in Supplementary Table 4.

To account for the scale disparity between small molecules and large proteins, atom-level features were aggregated via global pooling

https://doi.org/10.1038/s41467-025-66311-9

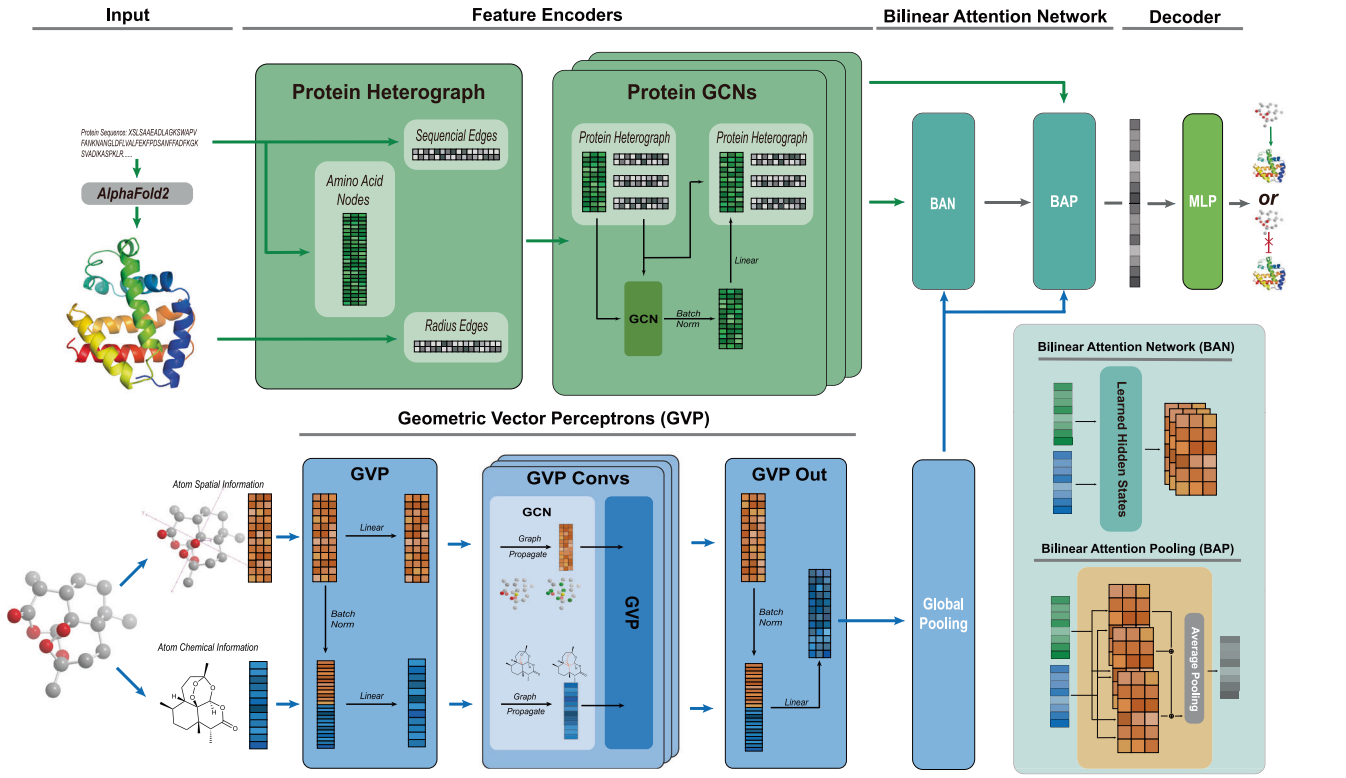

Fig. 2 | Schematic representation of the SCOPE model for drug-target interaction prediction. The SCOPE model integrates 3D structural information of proteins and compounds to predict drug-target interactions. To capture intricate structural features, the 3D protein structures and molecular graphs are encoded using heterogeneous graph neural networks (HGNNs) and geometric vector perceptrons (GVPs) with global pooling. These encoded representations are then processed through a bilinear attention network, which consists of a bilinear attention layer followed by bilinear pooling, to generate a joint representation that models local interactions between the drug and the target protein. The final predictive score is computed by a fully connected classification layer, representing the likelihood of an interaction. Molecular structures were visualized using ChemDraw and Chem3D. The protein structures were visualized using PyMOL. Source data are provided as a Source Data file.

to generate fixed-size molecular representations for downstream processing. These pooled molecular vectors are then combined with the encoded protein representations and fed into a BAN. This network comprises a bilinear attention layer, followed by a bilinear pooling layer, enabling the model to capture cross-domain interactions between the drug and the target protein residues[15]. Ultimately, the combined representations are passed through an MLP classifier, which computes the final predictive score, representing the probability of a drug-target interaction.

Evaluation strategies and metrics

We evaluated the classification performance of our model using three publicly available datasets—BindingDB[20,21], human[22], and KIBA[23]—as well as our proprietary SCOPE dataset. Our evaluation simulates a real-world scenario by adopting a strict semi-inductive framework, where models must predict interactions for previously uncharacterized compounds against a known set of target proteins (Fig. 3A). To achieve this, we partitioned datasets exclusively by compound, randomly assigning them to training, validation, and test sets (7:1:2 ratio). This approach guarantees that while the test set contains compounds unseen during training, the entire protein vocabulary remains consistent across all data splits. Detailed information about the dataset splits is provided in Supplementary Note 3.

We utilized the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC) as the primary metrics for assessing classification performance. Additionally, we reported accuracy, sensitivity, and specificity at the threshold corresponding to the optimal F1 score. To ensure robustness, we performed at least five independent runs with different random seeds for each dataset split. The model that achieved the highest AUROC on the validation set was selected for final evaluation on the test set, and the performance metrics were then recorded.

Performance comparison

A persistent challenge in DTI prediction is the risk of performance inflation from dataset biases. We first investigated the impact of target-level class imbalance, a common issue where a target's known interactions are overwhelmingly positive or negative. To isolate the effect of this bias, we conducted a controlled experiment on the BindingDB dataset with the SCOPE model (Fig. 3B). The original, imbalanced data yielded an optimistic AUROC, which was further inflated under a "label polarization" control that intentionally maximized this bias. In contrast, our proposed "label balancing" filter produced a more realistic AUROC baseline. Importantly, both the label-balanced and label-polarized datasets were constructed to have a similar overall class balance and sample size, confirming that the observed performance gap stems directly from the per-target bias distribution rather than dataset size or global imbalance. This result validates our filtering strategy as essential for a fair and rigorous evaluation, which we applied to all subsequent experiments.

Having established a robust evaluation framework, we next assessed whether our newly constructed SCOPE dataset provides a superior benchmark. We performed a head-to-head comparison focused on the kinase target family, pitting the widely used KIBA dataset against our SCOPE Kinase dataset (Fig. 3C). This ensures that performance differences are attributable to data quality and scale within a similar biological context. Across all tested models, performance on the SCOPE Kinase dataset was consistently and significantly

https://doi.org/10.1038/s41467-025-66311-9

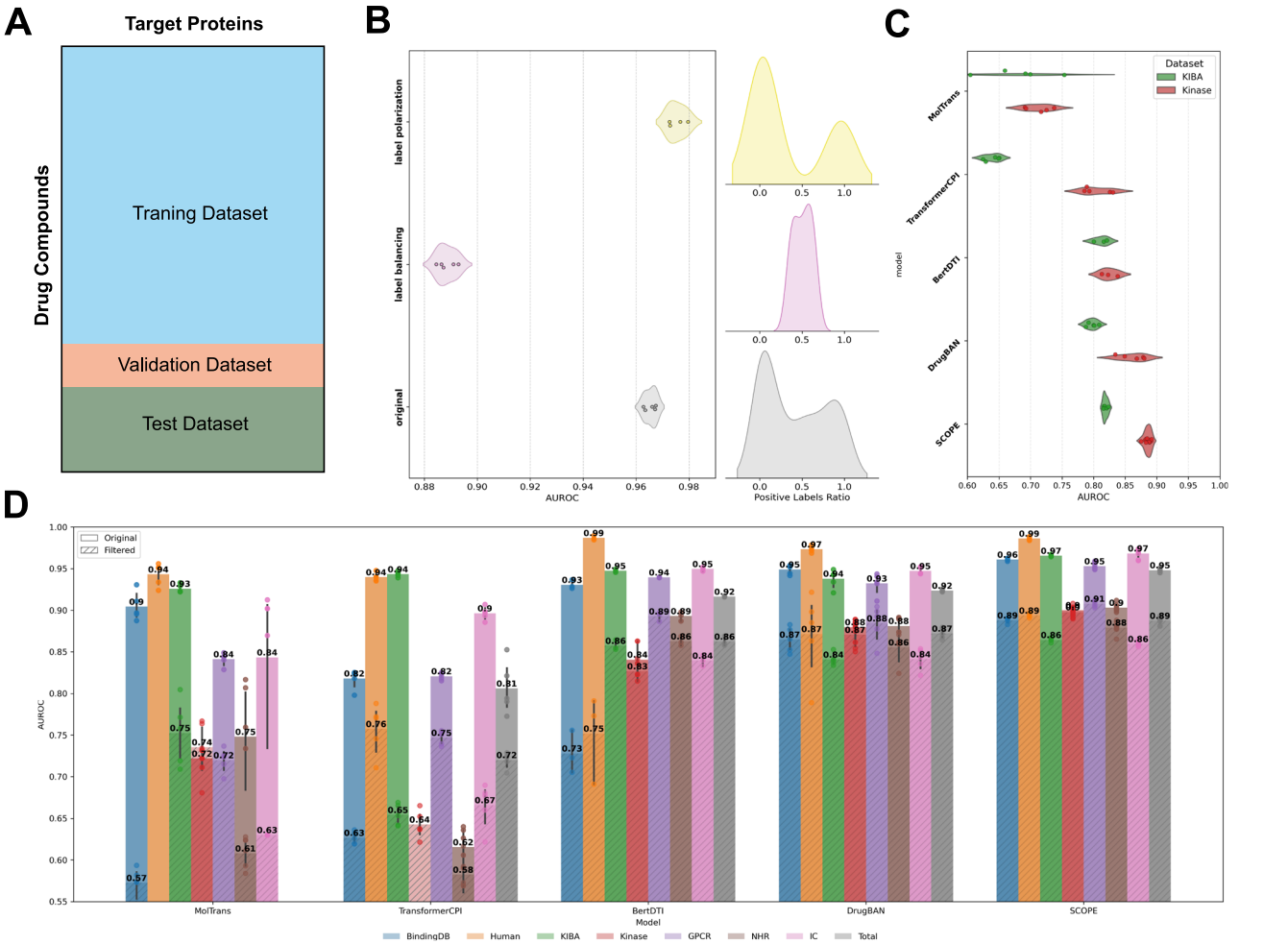

Fig. 3 | Evaluating DTI model performance changes on multiple datasets under a bias-aware, semi-inductive setting. A The semi-inductive compound evaluation framework. **B** Target-level class imbalance inflates model performance. Our Label Balancing (pink) corrects this bias, providing a realistic benchmark, while a Label Polarization control (yellow) confirms that high AUROC on unfiltered data is an artifact. **C** Models achieve superior AUROC on the larger, filtered SCOPE Kinase dataset compared to the standard KIBA benchmark, demonstrating the value of our curated data. **D** Overall performance before and after filtering. Filtering reveals the brittleness of simpler models, whereas advanced models, especially our 3D-aware SCOPE model, demonstrate robust and superior performance on the more realistic filtered datasets. Each bar represents the mean performance over $n = 5$ independent training runs performed with different random seeds and dataset splits. Error bars indicate ± standard deviation (SD). Central values correspond to the mean. Source data are provided as a Source Data file.

higher. This finding is further supported by a scaling analysis, which shows a clear positive correlation between data volume and model performance (Supplementary Fig. 3 and Supplementary Note 4). This demonstrates that the increased scale and richer data curated in SCOPE dataset translate directly into a higher-quality benchmark, enabling more powerful and generalizable models.

We then expanded our analysis to dissect the robustness of different model architectures under this rigorous framework. Our benchmark suite included classical models (SVM)[36], simple deep learning models (MLP)[37], and four state-of-the-art architectures: MolTrans[19] and TransformerCPI[25], which use simple transformers; BertDTI[14], which leverages pre-trained FMs; and DrugBAN[15], which introduced a Bilinear Attention Network (BAN). As shown in Fig. 3D, the balancing filter and semi-inductive split acted as powerful differentiators. The simpler transformer-based models (MolTrans, TransformerCPI) proved brittle, suffering substantial performance degradation after filtering, indicating their reliance on dataset artifacts. The FM-based BertDTI showed greater resilience on larger datasets but still experienced a notable performance drop on smaller ones like Human or BindingDB. In contrast, models employing the BAN architecture—DrugBAN and our SCOPE model—demonstrated exceptional stability across all datasets, regardless of size. This highlights the critical role of the BAN module in achieving robust DTI prediction.

To provide a definitive quantitative assessment, we summarized the average AUROC of all models across the eight filtered datasets (Table 1). The results chronicle a clear and consistent performance improvement driven by advances in both molecular encoding and feature fusion technologies. The technological progression reveals several key insights. While a basic transformer architecture like MolTrans did not offer a significant leap in peak performance over a simple Deep MLP, it provided a notable improvement in training stability and convergence consistency, albeit at the cost of substantially increased computational overhead. Performance and stability only improved with the introduction of richer, graph-based embeddings, as seen in TransformerCPI. A major leap in performance, however, came with the advent of more sophisticated methods. The pre-training and fine-tuning paradigm of BertDTI, the advanced fusion mechanism of DrugBAN, and our 3D encoding with a BAN module each propelled model performance to a level of practical applicability. Within this top tier, the benefits of 3D structural information became decisive, allowing SCOPE model to achieve state-of-the-art performance across

Table 1 | Average AUROC performance of SCOPE and six baseline models across eight filtered datasets

Method	SVM 1D+SVM	Deep MLP 1D+MLP	MolTrans 1D+transformer	TransformerCPI 2D+transformer	BertDTI FM+MLP	DrugBAN 2D+BAN	SCOPE 3D+BAN
BindingDB	0.600 ± 0.102	0.711 ± 0.104	0.560 ± 0.029	0.622 ± 0.008	0.748 ± 0.022	0.865 ± 0.018	**0.888 ± 0.013**
Human	0.581 ± 0.151	0.785 ± 0.026	NaN	0.676 ± 0.037	0.760 ± 0.023	0.854 ± 0.041	**0.877 ± 0.043**
KIBA	0.560 ± 0.051	0.766 ± 0.016	0.682 ± 0.055	0.640 ± 0.012	0.811 ± 0.010	0.798 ± 0.008	**0.819 ± 0.013**
Kinase	NaN	0.607 ± 0.123	0.713 ± 0.020	0.805 ± 0.022	0.824 ± 0.012	0.862 ± 0.020	**0.885 ± 0.025**
GPCR	NaN	0.707 ± 0.054	0.649 ± 0.011	0.649 ± 0.005	0.865 ± 0.003	0.856 ± 0.028	**0.882 ± 0.022**
NHR	NaN	0.688 ± 0.137	0.444 ± 0.051	0.614 ± 0.035	0.853 ± 0.007	0.848 ± 0.021	**0.872 ± 0.035**
IC	NaN	0.792 ± 0.019	0.492 ± 0.010	0.638 ± 0.020	0.838 ± 0.004	0.841 ± 0.013	**0.862 ± 0.021**
Total	NaN	0.533 ± 0.072	NaN	0.717 ± 0.006	0.855 ± 0.004	0.857 ± 0.005	**0.875 ± 0.014**

The comparison spans from sequence-based (1D) and graph-based (2D) methods to more advanced approaches using Foundation Models (FM) and 3D structures with different feature fusion methods, demonstrating the effectiveness of incorporating higher-dimensional structural information (**Best,** Second Best).

Table 2 | Ablation study on the filtered SCOPE dataset (averaged over five random runs)

Protein encoding	Compound encoding	Backbone	AUROC	AUPRC	F1
3D Graph HGNN	1D Fingerprint	BAN+MLP	0.829 ± 0.012	0.841 ± 0.015	0.771 ± 0.019
3D Graph HGNN	2D Graph	BAN+MLP	0.859 ± 0.015	0.873 ± 0.015	0.785 ± 0.016
3D Graph HGNN	3D Graph GVP no Pooling	BAN+MLP	0.860 ± 0.014	0.874 ± 0.014	0.786 ± 0.015
1D Onehot	3D Graph GVP	BAN+MLP	0.827 ± 0.016	0.846 ± 0.017	0.758 ± 0.012
1D CNN	3D Graph GVP	BAN+MLP	0.873 ± 0.008	0.871 ± 0.007	**0.806 ± 0.010**
3D Graph HGNN	3D Graph GVP	MLP	0.859 ± 0.024	0.857 ± 0.025	0.789 ± 0.026
3D Graph HGNN	3D Graph GVP	BAN+MLP	**0.875 ± 0.014**	**0.888 ± 0.013**	0.803 ± 0.016

The first three models evaluate the compound embedding design, followed by two models assessing the protein embedding design. The impact of the Bilinear Attention Network (BAN) layer is shown in the subsequent model. The final model integrates all components of the SCOPE design (**Best,** Second Best).

all eight datasets. A comprehensive analysis, including additional metrics, is provided in the Supplementary Notes and Supplementary Tables. Notably, this robust superiority is achieved with remarkable efficiency; our model's computational requirements are comparable to those of DrugBAN and significantly lower than resource-intensive FM-based methods (Supplementary Note 5). These findings confirm that by integrating 3D structural information with an advanced attention mechanism, SCOPE model establishes a state-of-the-art system that is not only highly accurate and robust but also computationally efficient.

Ablation study

We conducted an ablation study to assess the individual contributions of protein encoding, compound encoding, and backbone design to the overall performance of the SCOPE model. Table 2 summarizes the key findings, highlighting the critical role each module plays in the model's effectiveness, with detailed experimental results provided in Supplementary Table 7.

To evaluate the impact of compound encoding using the 3D Graph GVP, we constructed three variants of the SCOPE model, each employing different compound encoding methods: 1D fingerprint, 2D Graph[15], and 3D Graph GVP without global pooling, where the maximum atom size was set to 300. The results demonstrate that increasing the representational dimension of the compounds led to a significant improvement in model performance. Notably, the inclusion of pooled molecular encoding resulted in an over 1% performance gain, highlighting the critical role of pooling techniques in enhancing molecular representations. Notably, our analysis shows that the choice of compound encoder has a negligible impact on the overall computational cost (Supplementary Note 5).

For protein encoding, we conducted a comparative analysis of our proposed 3D Graph HGNN architecture against two alternative 1D methods: One-hot encoding and a CNN-based approach[15]. Our 3D architecture leverages protein structures derived from AlphaFold; we chose this path because we posit that incorporating 3D information is a crucial step towards creating more physically realistic and accurate models, a strategic choice we elaborate on in Supplementary Note 6. The ablation results revealed a nuanced performance landscape. The 1D CNN-based encoding performed surprisingly well, achieving results comparable to our 3D design and even outperforming it on the F1 score. However, our 3D Graph HGNN architecture maintained a distinct advantage in both AUROC and AUPRC, the primary metrics for this task. This suggests that while 1D methods can effectively capture sequence-level patterns, the 3D design unlocks access to critical structural information that ultimately leads to superior predictive power. This performance gain from 3D features, however, comes with a computational trade-off. As we dissect in detail in Supplementary Note 5, this is primarily an I/O overhead from loading the large structural embeddings, rather than a model complexity issue.

Finally, we examined the impact of the BAN layer within the backbone. Remarkably, removing the BAN layer caused a substantial drop in performance, bringing it to levels comparable to DrugBAN, which utilizes a 1D CNN and a 2D Graph for protein and compound encoding. This underscores the significance of both the encoding strategies and the backbone design in achieving optimal model performance. The efficiency of the BAN module itself ensures that the primary computational trade-off lies within the data encoding, not the fusion architecture.

In summary, the ablation study confirms the critical contributions of each module in the SCOPE model, validating the importance of its overall design and its role in improving performance.

Webserver and database

To enhance accessibility and usability, we have developed a webserver and database for the SCOPE framework, accessible at https://awi.cuhk.edu.cn/SCOPE/. The platform allows users to input a SMILES string of a compound to perform search or prediction tasks. In search mode, the server identifies and returns all compounds from the SCOPE dataset with a structural similarity greater than 0.9 to the input molecule,

https://doi.org/10.1038/s41467-025-66311-9

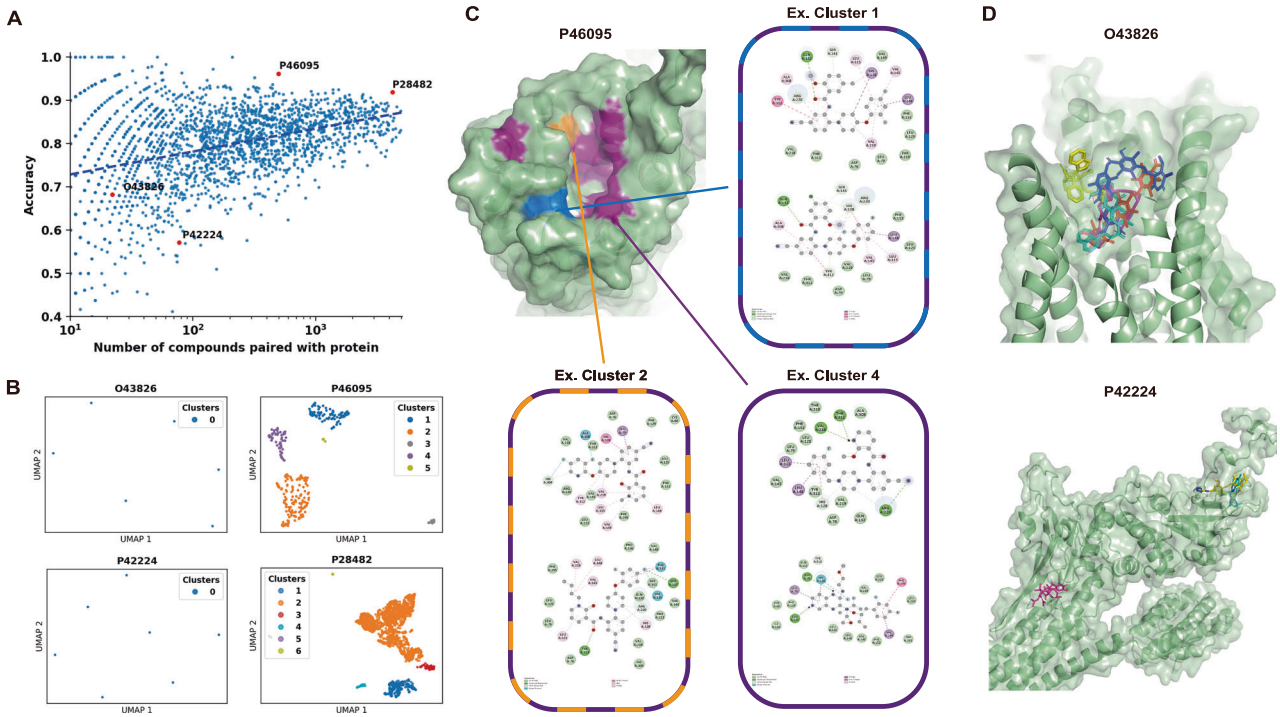

Fig. 4 | Interpretability analysis of the model. A Relationship between predictive performance and the number of known interactions per protein. As more known interactions are identified, accuracy stabilizes around 0.8-0.9, whereas fewer known interactions lead to greater variability. **B** UMAP-OPTICS clustering of BAN attention vectors for four representative proteins: two with robust predictive performance (P46095 and P28482) and two with lower accuracy (O43826 and P42224). Proteins enriched with interaction data (P46095 and P28482) exhibit clear clustering, suggesting that the model captures common binding features. In contrast, proteins with limited data (O43826 and P42224) do not show discernible clusters. **C** Docking analysis of ligands binding to P46095, highlighting shared and unique interaction residues across distinct ligand clusters. Unique critical residues are marked in blue and yellow, and universally important residues are shown in purple, indicating both global and cluster-specific binding determinants. Protein structure was visualized using PyMOL. **D** Analysis of proteins with relatively lower predictive accuracy. Although O43826 achieves an accuracy of about 0.7, its five known ligands—despite targeting the same pocket—adopt highly divergent poses, impeding a clear consensus. In the case of P42224 (accuracy <0.6), the small number of known ligands not only bind in differing orientations but also occupy distinct pockets, challenging the model's predictive capability. Source data are provided as a Source Data file.

calculated using RDKit. In prediction mode, the input molecule is paired with all proteins in the SCOPE library, providing interaction predictions with semi-inductive accuracy. Additionally, the complete SCOPE dataset is available for direct download from the website. Detailed information about the web development process and user instructions can be found in Supplementary Note 8 and Supplementary Fig. 4.

Model interpretability

As shown in Fig. 4A, prediction accuracy improves progressively with an increasing number of recorded protein interactions, highlighting the critical importance of sufficient DTI data for robust model training. Proteins with limited known interactions exhibit considerable variability in predictive performance, whereas accuracy stabilizes at approximately 0.8-0.9 as interaction data expands.

To further explore this relationship, we examined four representative proteins (Fig. 4B). Proteins O43826 and P42224, characterized by fewer known interactions, displayed lower predictive accuracy. In contrast, proteins P46095 and P28482, benefiting from richer datasets, achieved consistently higher accuracy. Interaction embeddings derived from the BAN attention module were analyzed using UMAP[38] for dimensionality reduction and OPTICS[39] for clustering. Proteins with extensive interaction data (P46095 and P28482) demonstrated clear, distinct clusters, indicating that the model effectively captures meaningful binding features. Conversely, sparse interaction data for O43826 and P42224 resulted in indistinct or no observable clusters, reflecting difficulties in recognizing consistent binding patterns.

A deeper analysis of docking results for the GPCR protein P46095 (Fig. 4C) revealed that clusters identified by the model correspond to chemically distinct binding modes within the canonical binding pocket. Specifically, interactions in clusters 1, 2, and 4 commonly involved residues 79, 128, 132, 145, 219, 220, 312, and 315, while unique residues 295 and 304 characterized cluster 1, and residue 141 was distinctive for cluster 2. These findings confirm that our model captures nuanced and chemically relevant features governing ligand specificity.

Further examination of proteins with relatively low predictive accuracy (Fig. 4D) revealed additional complexity. Protein O43826, despite an accuracy of around 0.7, had five known ligands adopting significantly different binding poses within the same pocket, complicating the identification of consistent patterns. Similarly, protein P42224 (accuracy <0.6) had limited ligand binding in various orientations and distinct pockets, presenting significant modeling challenges.

Collectively, these analyses emphasize the necessity of sufficient and diverse interaction data to enable accurate predictions of drug-target interactions. They also illustrate the interpretative capability of our model, highlighting its ability to identify both global and ligand-specific binding features critical for understanding molecular interactions. For further methodological details and additional analyses, see the Supplementary Note 7.

Efficient target discovery for bioactive natural compounds using SCOPE-DTI

Natural products have historically served as an important source of therapeutic agents and continue to garner significant attention in

https://doi.org/10.1038/s41467-025-66311-9

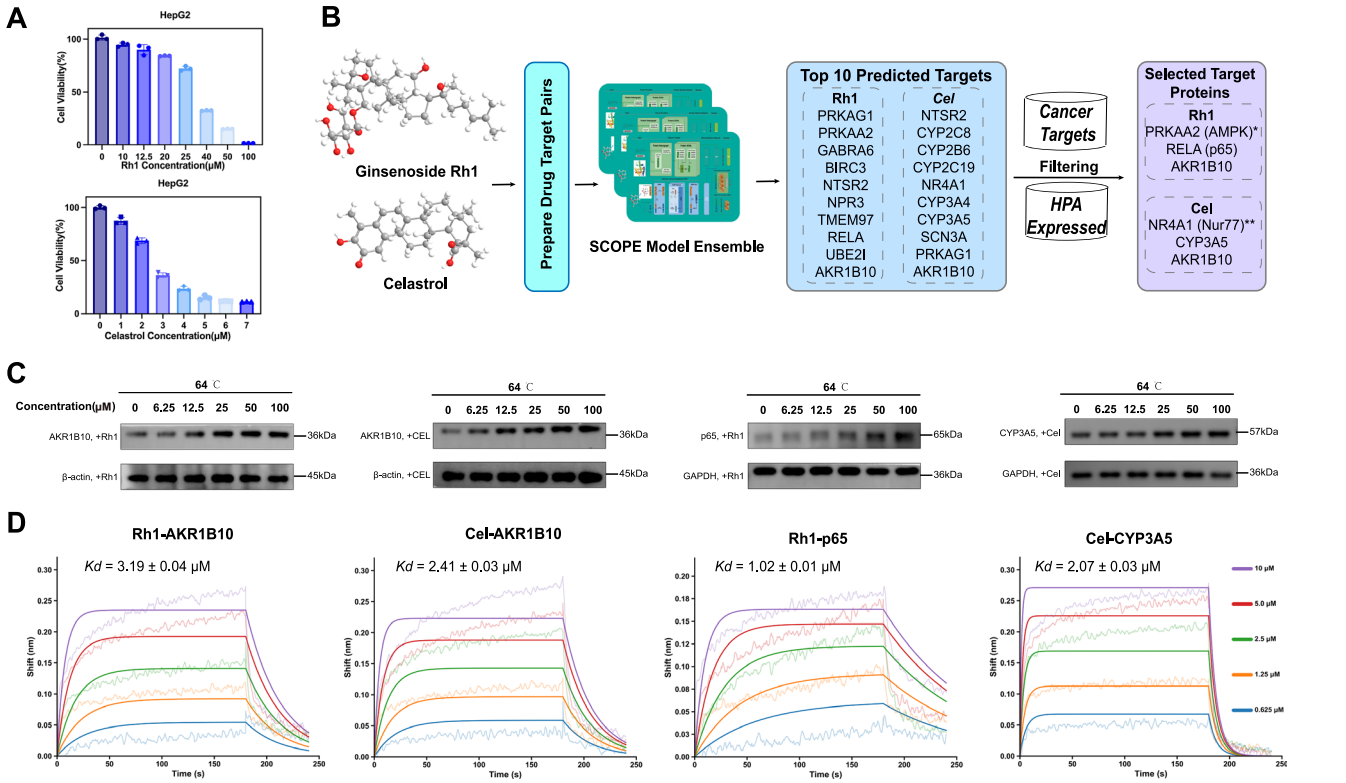

Fig. 5 | Experimental validation of SCOPE-DTI-predicted targets for natural products Ginsenoside Rh1 and Celastrol. A Viability of HepG2 cells after 24-h treatment with Ginsenoside Rh1 (top) and Celastrol (bottom), measured by CCK-8 assay. Data represent the mean ± SD from six biological replicates, normalized to controls. **B** The target identification workflow. SCOPE-DTI predicted Top 10 targets for each compound, which were then filtered by cancer relevance and expression in HepG2 cells to select final candidates. Molecules were visualized using Chem3D. * PRKAA2 (AMPK) was excluded after failing the initial CETSA screen. ** The Celastrol-NR4A1 (Nur77) interaction was previously validated in the literature. **C** Isothermal dose-response CETSA (ITDR-CETSA) showing increased thermal stability of target proteins in HepG2 cell lysates upon drug treatment at 64 °C, indicating direct binding. β-actin and GAPDH served as loading controls. Representative blots are shown. Each experiment was independently repeated three times with similar results. Uncropped blots are provided in the Source Data file. **D** Bio-Layer Interferometry (BLI) confirming direct binding and quantifying affinities. The calculated dissociation constants (K_d) for all four validated pairs are in the micromolar range, meeting our model's <10 μM success criterion. K_d values are mean ± standard error, calculated from kinetic parameter fitting. Source data are provided as a Source Data file. Additional fitting parameters, including Rmax and chi^2 values, are also available in the Source Data file.

modern drug discovery[40]. These compounds often possess complex chemical scaffolds and exert their biological effects not through a single, high-affinity target, but by modulating a network of targets with moderate affinities to produce a systems-level response. This polypharmacological profile, while therapeutically advantageous, makes their target identification a formidable challenge, rendering them ideal test cases for advanced predictive models like SCOPE-DTI. We therefore selected two representative anti-cancer natural products for our validation study: Ginsenoside Rh1, a major in vivo metabolite of the widely studied ginsenosides[41–43], and Celastrol, a potent bioactive compound from *Tripterygium wilfordii* known for its strong anti-cancer effects and multi-target mechanism[44]. As a first step, we evaluated the cytotoxic effects of Rh1 and Celastrol on HepG2 cells using the CCK-8 assay. As shown in Fig. 5A and Supplementary Table 8, the IC_{50} value of Rh1 was determined to be 32.04 μM. Celastrol exhibited substantially higher potency, with IC_{50} value of 2.53 μM, confirming that both compounds significantly inhibit HepG2 cell proliferation.

Having established their inhibitory effects, we next utilized the SCOPE framework to predict potential protein targets for both compounds. As depicted in the workflow in Fig. 5B, Rh1 and Celastrol were computationally screened against all candidate proteins in our database, and the SCOPE model ensemble ranked the potential interactions. From the top 10 ranked proteins for each compound, we applied two stringent filtering criteria to prioritize candidates for validation: (1) inclusion of targets with reported anti-cancer relevance[45], and (2) confirmation of protein expression in HepG2 cells via the Human Protein Atlas (HPA)[46]. This refinement process yielded three high-confidence candidates for Rh1 (PRKAA2/AMPK, RELA/p65, and AKR1B10) and three for Celastrol (NR4A1/Nur77, CYP3A5, and AKR1B10). Notably, the interaction between Celastrol and NR4A1 has been recently validated in the literature[47]; therefore, we excluded it from our subsequent biophysical assays, focusing our efforts on the remaining five previously uncharacterized pairs.

To experimentally confirm these computational predictions, we first employed the cellular thermal shift assay combined with Western blot (CETSA-WB) as a primary screen for direct physical engagement in a cellular context. HepG2 cells were treated with each compound or a vehicle control (DMSO) and subjected to a temperature gradient (40 to 74 °C). As shown in Supplementary Fig. 5, a clear thermal stabilization was observed for four of the five tested pairs: Rh1-p65, Rh1-AKR1B10, Celastrol-CYP3A5, and Celastrol-AKR1B10. In contrast, the Rh1-AMPK interaction did not exhibit a significant thermal shift and was thus excluded from further validation. To corroborate these findings, we then performed Isothermal Dose-Response CETSA (ITDR-CETSA) for the four positive hits. The results, shown in Fig. 5C, demonstrate clear, dose-dependent stabilization for all four pairs, providing strong qualitative evidence of direct physical binding in situ.

Finally, to obtain a definitive, quantitative measure of these interactions, we performed Bio-Layer Interferometry (BLI) assays. The results, presented in Fig. 5D, provide conclusive evidence of direct binding with clear sensorgrams and excellent fitting statistics. The measured dissociation constants (K_d) were 3.19 ± 0.04 μM for Rh1-

AKR1B10, 1.02 ± 0.01 μM for Rh1-p65, 2.41 ± 0.03 μM for Celastrol-AKR1B10, and 2.07 ± 0.03 μM for Celastrol-CYP3A5. All measured affinities are well within the <10 μM cutoff used by our model, confirming them as successful predictions. In total, including the previously validated Celastrol-NR4A1 interaction ($K_d \approx 292$ nM)[47], five out of the six prioritized candidate targets were confirmed, corresponding to a success rate of over 80%. This high success rate, which closely aligns with the accuracy metrics from our large-scale computational benchmarks, powerfully demonstrates the reliability and real-world applicability of the SCOPE-DTI framework for discovering previously uncharacterized targets of bioactive compounds. We also assessed the performance of our baseline models on these validated pairs. SCOPE-DTI was the only model to successfully rank all targets within the top 10, demonstrating its superior real-world applicability. A detailed breakdown of each model's ranking performance is provided in Supplementary Table 9.

Discussion

In this study, we introduced SCOPE-DTI, a previously uncharacterized framework developed to enhance the practical utility of deep learning-based drug-target interaction (DTI) prediction. By constructing the large semi-inductive human DTI dataset to date and incorporating robust target-level data balancing, we have established a more reliable foundation for evaluating model performance under real-world conditions. Our approach, which leverages three-dimensional structural representations of both proteins and compounds within an attention-based architecture, demonstrated superior predictive performance and structural interpretability across multiple benchmarks. We further showcased SCOPE-DTI's practical applicability by experimentally identifying and validating previously uncharacterized anti-cancer targets for two bioactive natural products. To foster broader adoption, we have made our model and curated dataset publicly accessible via a user-friendly web interface.

While the case studies in this manuscript focused on identifying anti-cancer targets, the potential applications of SCOPE-DTI extend far beyond traditional drug discovery. Bioactive natural compounds are ubiquitous, influencing human health through everyday consumables such as foods[48], dietary supplements[49], and skin care[50]. AI-driven target identification tools like SCOPE-DTI have the potential to elevate our understanding of these compounds from empirical observation to a new, mechanism-centric era. By systematically mapping the interactions between these molecules and the human proteome, we can begin to unravel the molecular basis of daily wellness and preventative health. However, moving from target identification to mechanistic insight requires acknowledging that physical binding is a critical starting point, not an endpoint. A truly comprehensive understanding of a compound's systemic effects necessitates a multifaceted approach. To translate physical binding into functional outcomes, downstream functional assays are essential. Furthermore, to capture the dynamic nature of cellular responses, future studies should integrate context-specific data, such as transcriptomics from treated cells, to identify targets whose expression is modulated by the compound itself.

Despite these accomplishments, several areas remain for future improvement. Our choice to incorporate 3D structural information was motivated by its clear benefits to performance and interpretability. However, our ablation studies revealed a notable discrepancy: 3D compound structures currently offer more significant performance gains than 3D protein structures. We hypothesize this is directly linked to the inherent limitations of current protein structure prediction methods like AlphaFold2. While these AI-based tools have revolutionized structural biology, their predictions, as others have noted, must be treated as exceptionally useful hypotheses rather than ground truth[51]. Their accuracies vary, and critically, they do not account for the influence of ligands, covalent modifications, or other cellular factors that can alter protein conformation. Consequently, even high-confidence predictions can exhibit local or global deviations from the biologically active state. Looking forward, we anticipate that the next generation of predictive tools, such as AlphaFold3, will provide more accurate, dynamic, and context-aware structural data. Our modular framework is well-positioned to seamlessly integrate these future advancements, which we expect will unlock significant further gains in DTI prediction accuracy. Furthermore, while the bilinear attention mechanism effectively captures interaction patterns, its interpretability remains most meaningful at the cluster level. Future work will therefore focus on incorporating more granular data, such as pocket-level binding information, and refining the model architecture to enhance both the residue-specific accuracy and the interpretability of our framework. Finally, on the data front, while SCOPE represents the large dataset of its kind, our analysis revealed that coverage for certain crucial protein families, such as GPCRs, remains limited due to the scarcity of public data. Our scaling analysis confirms that model performance is currently data-limited. Therefore, future efforts in large-scale, systematic data generation and curation for these underrepresented target classes will be paramount to building truly comprehensive and universally applicable DTI prediction models.

Methods

SCOPE dataset development

We developed the SCOPE dataset as a comprehensive, multi-source, and well-annotated resource tailored for drug-target interaction (DTI) prediction. This dataset integrates DTI data from a wide range of public sources, including ChEMBL activity[30], PubChem activity[29], DrugBank[52], BindingDB[20], DrugCentral[53], TTD[54], Pharos[55], PROMISCUOUS[56], GtoPdb[27], Human[22], BioSNAP[57], KIBA[23], and DAVIS[58]. Additionally, these sources encompass more than 20 reference data sources. However, the raw data presented several challenges: (1) lack of unified annotations, (2) varying scales and pharmacological evaluation metrics, and (3) inconsistent data quality with potential erroneous entries. To address these issues, we performed comprehensive data cleaning and standardization.

Protein and compound annotation. We unified protein annotations by the Uniprot database and filtered all human proteins, focusing on druggable targets for interaction prediction with small molecules. Proteins were classified into pharmacological target families (e.g., G-protein-coupled receptors (GPCRs), kinases) using the IUPHAR database. To provide structural insights, we generated three-dimensional (3D) structures for all proteins using AlphaFold2. Compounds were annotated using identifiers from the PubChem and ChEMBL databases. We generated 3D structures of the compounds using RDKit and optimized their conformations with the Merck Molecular Force Field (MMFF).

Interaction label unification. To harmonize different scales and pharmacological evaluation metrics across various data sources, we assigned interaction labels based on PubChem activity distributions, classifying them as either positive (1) or negative (0). Specific classification criteria are detailed in Supplementary Note 1 and Supplementary Table 2. This approach allowed us to standardize experimental information from different sources into a consistent binary classification.

Multi-source data integration. To further improve data quality, we consolidated multiple data sources by retaining only interactions consistently labeled as positive across all sources. If any source labeled an interaction as negative, we considered the interaction negative in our dataset. This conservative approach maximized the reliability of positive samples.

Mitigating target-level interaction imbalance. To address target-level interaction imbalance—where a protein p_i exhibits a disproportionate number of interactions from one class – we implemented a dataset filtering strategy. For each protein p_i, we consider its interaction set $\mathcal{I}(p_i) = \{(p_i, c_j, l_{ij})\}$, where c_j denotes a compound and $l_{ij} \in \{0, 1\}$ indicates the interaction label. The imbalance is quantified through two counters: N_0(negative interactions) and N_1 (positive interactions) within $\mathcal{I}(p_i)$.

If p_i had more than 75% of its interactions belonging to one class, we performed stratified random sampling to obtain a subset $\mathcal{I}'(p_i)$ that satisfies:

$$\frac{\min(N_0', N_1')}{N_0' + N_1'} \geq 0.25, \tag{1}$$

where N_0' and N_1' are the counts of negative and positive interactions in $\mathcal{I}'(p_i)$, respectively. This ensures that the minority class is represented at least 25% of the interactions for p_i, achieving a class ratio between 25% and 75%. Proteins with insufficient interactions after filtering were excluded to maintain data quality.

Visualization of data cleaning effects. To demonstrate the impact of our data cleaning and balancing procedures, we analyzed the distribution of interaction ratios per protein before and after filtering. For each protein p_i, we calculated the ratio of positive interactions:

$$R(p_i) = \frac{N_1}{N_0 + N_1}, \tag{2}$$

where N_0 and N_1 are the counts of negative and positive interactions, respectively, for protein p_i. We grouped proteins into intervals based on $R(p_i)$ with a specified interval length (e.g., 0.02) and plotted the number of proteins within each interval. This visualization highlighted the reduction of proteins with extreme interaction imbalances after filtering.

Dataset statistics and comparative analysis. We quantified the SCOPE dataset by calculating the number of unique compounds, targets, and total interactions. Our dataset encompasses a significantly larger number of compounds and interactions compared to existing datasets, both before and after the filtering procedures. To illustrate the scale and comprehensiveness of SCOPE, we generated plots that compare the data volume with that of previous datasets, highlighting our dataset's superiority in size and diversity. Detailed statistical information and comparative analyses are provided in Fig. 1 and Supplementary Table 3.

SCOPE framework architecture

HGNN for protein structure. A protein $P_i \in \mathcal{P}$, composed of M amino acid residues, can be represented by its primary sequence $S_i = (v_1, v_2, \ldots, v_M)$, where each residue v_m belongs to one of the 20 standard amino acid types. In their physicochemical environment, these sequences fold into stable three-dimensional (3D) structures. To capture both the sequence and structural information of protein P_i, we construct a residue-level heterogeneous protein graph $G_P^{(i)} = (V_P^{(i)}, E_P^{(i)}, R)$, where $V_P^{(i)}$ represents the residues as nodes, $E_P^{(i)}$ denotes the edges connecting them, and R specifies the edge types.

In this work, we use two types of edges to represent the relationships between residues. Sequential edges connect residues that are adjacent in the primary sequence, preserving the natural order of the amino acids. Radius edges connect residues whose spatial Euclidean distance, calculated based on the geometric centers of all their atoms, is below a predefined threshold d_r. This approach ensures that both local sequence context and spatial proximity within the folded protein are encoded.

To encode the heterogeneous protein graph $G_P = (V_P, E_P, R)$, we employ a Heterogeneous Graph Neural Network (HGNN)[32]. Each edge type $r \in R$ is associated with a weight matrix $W_r^{(l)}$ at layer l, and a shared weight matrix $W_h^{(l)}$ is used to combine the aggregated messages from different edge types.

The node embeddings are updated iteratively across L layers. At layer l (where $1 \leq l \leq L$), the embedding of each residue v_m is computed as:

$$h_m^{(l)} = \mathrm{BN}\left(\mathrm{ReLU}\left(W_h^{(l)} \cdot \sum_{r \in R} \sum_{v_n \in \mathcal{N}_r(m)} W_r^{(l)} h_n^{(l-1)}\right)\right), \tag{3}$$

where $h_m^{(0)} = x_m$ represents the initial input feature of residue v_m, and $h_m^{(l)}$ is the embedding at layer l. The neighbors $\mathcal{N}_r(m)$ are defined as residues connected to v_m by edge type r. The batch normalization function $\mathrm{BN}(\cdot)$ and rectified linear unit activation $\mathrm{ReLU}(\cdot)$ are applied for stable training, while $W_r^{(l)} \in \mathbb{R}^{d \times d}$ and $W_h^{(l)} \in \mathbb{R}^{d \times d}$ are learnable weight matrices. This formulation ensures that the heterogeneous nature of the graph is effectively captured, integrating sequence and structural information into a unified representation.

This formulation ensures that messages from different edge types are appropriately transformed and aggregated.

GVP for drug encoding. Given the three-dimensional (3D) coordinates of atoms in a molecule, we represent the molecular structure as a graph $G_d = (V_d, E_d)$, where the nodes V_d correspond to the atoms of the molecule, and the edges E_d are defined between pairs of atoms whose Euclidean distance is less than 4.5 Å[59]. This representation captures both the chemical and spatial relationships within the molecule.

Node Features: For each atom i, we construct a node feature $h_i^{(0)} = (h_i^{(0),v}, h_i^{(0),s})$, which consists of both a vector component $h_i^{(0),v}$ and a scalar component $h_i^{(0),s}$. The vector feature $h_i^{(0),v} = c_i \in \mathbb{R}^3$ represents the atom's 3D coordinates in space. The scalar feature $h_i^{(0),s} \in \mathbb{R}^{74}$ is a 74-dimensional integer vector that describes the atom with eight types of information: the atom type, the atom degree, the number of implicit hydrogens, the formal charge, the number of radical electrons, the atom hybridization, the number of total hydrogens, and whether the atom is aromatic[35].

Edge Features: For each edge $(i, j) \in E_d$, we define an edge feature $e_{ji} = (e_{ji}^v, e_{ji}^s)$. The vector feature $\mathbf{e}_{ji}^v = \frac{\mathbf{c}_j - \mathbf{c}_i}{\|\mathbf{c}_j - \mathbf{c}_i\|} \in \mathbb{R}^3$ is the unit vector pointing from atom i to atom j. The scalar feature $e_{ji}^s = \mathrm{RBF}(\| c_j - c_i \|) \in \mathbb{R}^{16}$ encodes the pairwise distance using 16 Gaussian radial basis functions (RBFs) with centers evenly spaced between 0 and 4.5 Å.

Molecular Graph Neural Network: To learn a representation for the input molecule, we utilize a Graph Neural Network based on Geometric Vector Perceptrons (GVPs)[34]. Each node embedding $h_i^{(l)} = (h_i^{(l),v}, h_i^{(l),s})$ consists of a vector and a scalar component.

At each layer l, the node embeddings are updated using the following equation:

$$h_i^{(l)} = h_i^{(l-1)} + \mathrm{GVP}\left(h_i^{(l-1)}, \sum_{j \in \mathcal{N}(i)} \mathrm{GVP}\left(h_j^{(l-1)}, \mathbf{e}_{ji}\right)\right) \tag{4}$$

where $\mathcal{N}(i)$ denotes the set of neighboring nodes of node i, $\mathrm{GVP}(\cdot, \cdot)$ represents a GVP layer that processes node and edge features, and $h_i^{(0)} = v_i$ is the initial feature of node i. This update rule incorporates both the features of neighboring atoms and the geometric information from edge features.

After L layers of message passing, we obtain the final node embeddings $h_i^{(L)}$ for all nodes in the graph. To derive a fixed-size representation of the entire molecule, we apply a global add pooling

operation over all node embeddings:

$$h_d = \sum_{i \in V_d} h_i^{(L)}, \tag{5}$$

where $h_d \in \mathbb{R}^d$ is the learned representation of the input molecule.

Pairwise interaction learning. To capture pairwise local interactions between drugs and proteins, we employ a bilinear attention network module comprising two key layers: a bilinear interaction map to compute pairwise attention weights, and a bilinear pooling layer over the interaction map to extract a joint drug–protein representation.

Given the hidden representations from the encoders $-H_d \in \mathbb{R}^{N \times D_d}$ for the drug and $H_p \in \mathbb{R}^{M \times D_p}$ for the protein—we compute the pairwise interaction matrix $I \in \mathbb{R}^{N \times M}$ as:

$$I = \left(\sigma(H_d U)\, q^\top\right) \odot \sigma\left(H_p V\right)^\top, \tag{6}$$

where $U \in \mathbb{R}^{D_d \times K}$ and $V \in \mathbb{R}^{D_p \times K}$ are learnable weight matrices, $q \in \mathbb{R}^K$ is a learnable weight vector, $\sigma(\cdot)$ is an activation function (e.g., sigmoid), $\odot$ denotes element-wise multiplication (Hadamard product), and the superscript $\top$ indicates matrix transpose. Each element $I_{i,j}$ represents the interaction score between the i-th atom of the drug and the j-th residue of the protein.

For intuitive understanding, an element $I_{i,j}$ can be expressed as:

$$I_{i,j} = q^\top \left(\sigma\left(U^\top h_d^i\right) \odot \sigma\left(V^\top h_p^j\right)\right), \tag{7}$$

where h_d^i and h_p^j are the embeddings of the i-th atom and the j-th residue, respectively.

The joint representation $f' \in \mathbb{R}^K$ is then computed via bilinear pooling:

$$f' = \left(\sigma(H_d U)^\top I\, \sigma\left(H_p V\right)\right) 1, \tag{8}$$

where $1 \in \mathbb{R}^M$ is a vector of ones used for summation over the protein residues. To reduce dimensionality, we apply sum pooling:

$$f = \mathrm{SumPool}(f', s), \tag{9}$$

resulting in the final feature vector $f \in \mathbb{R}^{K/s}$.

To compute the interaction probability, we feed the joint representation f into a decoder consisting of a fully connected layer followed by a sigmoid activation function:

$$p = \sigma(W_o f + b_o), \tag{10}$$

where W_o and b_o are learnable weight parameters.

Finally, we jointly optimize all learnable parameters using backpropagation. The training objective is to minimize the cross-entropy loss with L2 regularization:

$$\mathcal{L} = -\sum_i \left[y_i \log p_i + (1 - y_i) \log(1 - p_i)\right] + \frac{\lambda}{2} \| \theta \|_2^2, \tag{11}$$

where θ represents the set of all learnable parameters, y_i is the ground-truth label for the i-th drug–protein pair, p_i is the predicted probability, and λ is the regularization hyper-parameter. Overall, this pairwise interaction learning approach is highly inspired by and adapted from DrugBAN[15].

Experimental setting

Datasets. We evaluated the SCOPE model and five state-of-the-art baseline models on four public DTI datasets: BindingDB, KIBA, Human, and our constructed SCOPE dataset. The SCOPE dataset includes various protein families annotated by IUPHAR, such as G-protein-coupled receptors (GPCRs), kinases, ion channels, and nuclear hormone receptors, along with our total dataset. All datasets were tested in both their original and debiased versions. The BindingDB dataset[20] is a web-accessible database of experimentally validated binding affinities between small drug-like molecules and proteins; we utilized a low-bias version of this dataset[21]. The KIBA dataset[23] integrates various bioactivity measurements to provide a comprehensive set of drug-target interactions, focusing particularly on kinase inhibitors. The Human dataset, constructed by Liu et al.[22], includes highly credible negative samples generated via an in silico screening method; following previous studies[18,25], we used the balanced version containing an equal number of positive and negative samples.

Implementation. SCOPE model is implemented in Python 3.9 and PyTorch 2.2.0[60], along with functions from PyG 2.5.2[61], DGL 2.2.5[62], DGLlifeSci 0.3.2[35], Scikit-learn 1.0.2[63], Numpy 1.20.2[64], Pandas 1.5.2, and RDKit 2021.03.2. The batch size is set to be 64, and the Adam optimizer is used with a learning rate of 5e-5. We allow the model to run for at most 100 epochs for all datasets. The best performing model is selected at the epoch giving the best AUROC score on the validation set, which is then used to evaluate the final performance on the test set. The protein encoder HGNN utilizes an embedding dimension of 320 and processes protein sequences with a maximum allowed length of 2000 amino acids. To construct the protein 3D graph, we set an edge cutoff distance of 10 Å. The HGNN comprises 4 layers, and we include a fully connected layer with bias to enhance model capacity. For the drug feature encoder, the atom input dimensions are set to [74, 1], capturing both scalar and vector features. The atom's hidden dimensions are [320, 64], allowing the model to learn complex representations. Similarly, the edge input dimensions are [16, 1], and the edge hidden dimensions are [32, 1]. The GVP model consists of 3 layers and incorporates a dropout rate of 0.1 to prevent overfitting. An edge cutoff distance of 4.5 Å is used, and we compute 16 radial distribution functions to capture spatial relationships between atoms. Notably, there is no maximum length restriction for compounds in our model. In the bilinear attention module, we employ two attention heads to enhance interpretability while capturing intricate interactions between drugs and proteins. The latent embedding size is set to 768, and we use a sum pooling window size of 3 to aggregate features effectively. The decoder is a fully connected network with 512 hidden neurons, enabling the model to make accurate predictions based on the learned representations.

Baselines. We compare the SCOPE model with the following five methods for DTI prediction: (1) **SVM**, a shallow machine learning algorithm; (2) **MLP**, a simple deep neural network with hidden dimensions [2048, 512, 128, 32], applied to the concatenated ECFP4 and PSC fingerprint features; (3) **MolTrans**, a deep learning model that uses the transformer architecture to encode drug and protein features, with a CNN-based module to capture sub-structural interactions; (4) **TransformerCPI**, a Transformer-based model with an encoder for protein sequences and a decoder for molecular graphs, leveraging multi-head attention to extract interaction features; (5) **BertDTI**, a model that leverages large, pre-trained foundation models for representation learning. It utilizes ChemBERTa to encode drug SMILES and ProtBERT to encode protein sequences, with a simple MLP decoder for the final DTI prediction. (6) **DrugBAN**, a model that encodes drug molecules and protein sequences using graph convolutional networks and 1D-convolutional neural networks, followed by a bilinear attention network to capture pairwise interactions and a fully connected decoder for DTI prediction. For the above deep DTI models, we follow the recommended model hyper-parameter settings described in their original papers.

Interpretability. Representative AlphaFold structures (UniProt IDs P46095, O43826, and P42224) were retrieved from the AlphaFold Protein Structure Database (AlphaFold DB; DOI: 10.1093/nar/gkab1061), model version 4 (corresponding to AlphaFold DB release 4), in September 2024. These structures were used in the interpretability analysis to visualize model-derived interaction patterns and evaluate structure-based consistency of binding features. The associated pLDDT-colored models and Predicted Aligned Error (PAE) plots are provided in Supplementary Fig. 6.

Experimental validation of anti-cancer targets of Ginsenoside Rh1 and Celastrol

Cell viability assay. HepG2 cells (human hepatocellular carcinoma, male, 15 years old, Caucasian; CellCook, cat. no. CC0101) were used in this study. HepG2 cells were seeded in 96-well plates at a density of 1×10^4 cells per well and incubated overnight at 37 °C with 5% CO_2 to allow adhesion. Cells were treated with Ginsenoside Rh1 (Aladdin, cat. no. G107710-20mg) at final concentrations of 0, 10, 12.5, 20, 25, 40, 50, and 100 μM, or with Celastrol (Aladdin, cat. no. C107671-10mg) at final concentrations of 0, 1, 2, 3, 4, 5, 6, and 7 μM for 24 h. Untreated cells served as controls. Cell viability was evaluated using the Enhanced Cell Counting Kit-8 (CCK-8; Beyotime Biotechnology, cat. no. C0042). After treatment, 10 μL of CCK-8 solution was added to each well, followed by incubation at 37 °C for 2 h. Optical density (OD) at 450 nm was measured using a microplate reader to quantify viable cells. All conditions were tested in six independent replicates to ensure reproducibility.

Cellular thermal shift assay followed by Western blotting (CETSA-WB). HepG2 cells were lysed in RIPA buffer (Beyotime, cat. no. P0013C), and the resulting lysate was incubated with either the test compound (100 μM Ginsenoside Rh1 or 10 μM Celastrol) or a vehicle control (1% DMSO, v/v) at room temperature for 30 min. The lysate was divided into aliquots and incubated at defined temperatures (40, 46, 52, 58, 64, and 74 °C) for 3 min using an Applied Biosystems PCR analyzer (Thermo Scientific, USA). Following heat treatment, the samples were cooled at 4 °C for 10 min and transferred to low-adsorption centrifuge tubes. To each 100 μL aliquot, 40 μL of PBS (Gibco, cat. no. C14190500BT) was added and mixed thoroughly. Protein concentrations were adjusted to 2000 μg/mL and quantified using a BCA assay (Coolaber, cat. no. SK1070). The samples were centrifuged at 20,000 × *g* for 30 min at 4 °C to remove precipitated proteins. The resulting supernatants were mixed with 6× loading buffer (Beyotime, cat. no. P0015F) and denatured at 100 °C for 10 min. The denatured samples were then cooled on ice for 10 minutes, followed by centrifugation at 14,000 × *g* for 10 min at 4 °C. The final supernatants, containing soluble protein fractions, were collected for Western blot analysis to evaluate the thermal stability of target proteins.

Protein concentrations of the CETSA supernatants were quantified using a BCA assay to ensure equal loading. Samples were mixed with 6× loading buffer (Beyotime, cat. no. P0015F). Equal amounts of total protein were loaded onto a BeyoGel™ Plus Precast PAGE Gel (10% HEPES, 15 wells, Beyotime, cat. no. P0509M). Proteins were resolved by electrophoresis (60 V for 30 min, then 110 V for 40 min). Proteins were transferred onto PVDF membranes (Millipore, cat. no. ISEQ00010) using a rapid transfer apparatus (eBlot® L1, GenScript, cat. no. L00686C) after membrane activation with methanol and equilibration buffer (Genescript, cat. no. L00734C). Membranes were blocked with QuickBlock™ Western Block solution (Beyotime, cat. no. P0252) for 15 min at room temperature. Membranes were incubated overnight at 4 °C with the following primary antibodies (all at 1:1000 dilution): anti-AKR1B10 (Abcam, cat. no. ab96417), anti-AMPKα2 (Abcam, cat. no. ab32047), anti-NF-κB p65 (Abcam, cat. no. ab16502), and anti-CYP3A5 (Abcam, cat. no. ab108624). After washing three times with TBST, membranes were incubated with HRP-labeled Goat Anti-Rabbit IgG(H+L) (Beyotime, cat. no. A0208, 1:1000 dilution) or HRP-labeled Goat Anti-Mouse IgG(H+L) (Beyotime, cat. no. A0216, 1:1000 dilution) for 1 hour at room temperature. Protein bands were visualized using an ECL substrate (Beyotime, cat. no. P0018FM) and imaged using Touch Imager Pro (e-BLOT Life Science, Shanghai, China). Beta Actin Polyclonal antibody (Proteintech, cat. no. 20536-1-AP, 1:1000 dilution) or GAPDH Monoclonal antibody (Proteintech, cat. no. 60004-1-Ig, 1:1000 dilution) were probed as loading controls on the same membranes after stripping.

Isothermal dose-response CETSA followed by Western blotting (ITDR-CETSA-WB). For targets that exhibited a clear thermal shift in the initial CETSA screen, ITDR experiments were carried out. HepG2 lysates were incubated for 30 min with a six-point concentration series of each compound, prepared by serial 1:2 dilutions from a starting concentration of 100 μM (i.e., 100, 50, 25, 12.5, 6.25, and 0 μM), along with a vehicle control. All samples were then heated to a single, fixed temperature chosen from the thermal gradient screen to maximize the observed shift (e.g., 64 °C). Following heat treatment, samples were processed as described for CETSA-WB, and the soluble protein fractions were subjected to Western blotting to assess dose-dependent stabilization of the target proteins.

Bio-layer interferometry (BLI) assay. Binding kinetics and affinities were quantified using a Gatorprime Bio-Layer Interferometry system (Gator Bio, Palo Alto, CA, USA). All experiments were conducted at 30 °C in 96-well black microplates with a working volume of 200 μL per well. The kinetic buffer used throughout the assay consisted of phosphate-buffered saline (PBS) supplemented with 0.02% Tween-20. Recombinant human proteins, including CYP3A5 (Biorbyt, cat. no. orb1477061), AKR1B10 (Abcam, cat. no. ab85415), and p65 (SinoBiological, cat. no. 12054-H09E), were purified and biotinylated in-house using a standard NHS-biotinylation pipeline. SAS probe (Gator Bio, cat. no. 88-0002) was first hydrated in kinetic buffer and then used to immobilize the biotinylated proteins to a target density over 5 nm. The compounds, Ginsenoside Rh1 and Celastrol, were prepared in a five-point, 1:2 serial dilution series, starting from 10 μM, following a standard protocol. To ensure data accuracy, a double reference subtraction was employed during data analysis using the Gatorprime software suite. This involved subtracting the signal from parallel reference sensors (immobilized protein exposed to buffer only, to correct for signal drift) and from a blank reference subtraction (non-immobilized sensors exposed to the compound, to correct for non-specific binding). The resulting sensorgrams were globally fitted to a 1:1 binding model to determine the association rate constant (k_{on}), dissociation rate constant (k_{off}), and the equilibrium dissociation constant (K_d). The final K_d was calculated from the ratio of k_{off}/k_{on} with a standard error.

Statistical analysis. Data were analyzed using GraphPad Prism (v10.5.0) and are presented as mean ± SD. Statistical significance was defined as $p < 0.05$.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The SCOPE dataset generated in this study has been deposited in Zenodo under accession code https://doi.org/10.5281/zenodo.17217469 and is also accessible via our webserver: https://awi.cuhk.edu.cn/SCOPE/. All other datasets used in this study are publicly available from established resources, including DrugBank (https://go.

drugbank.com), ChEMBL (https://www.ebi.ac.uk/chembl), BindingDB (https://www.bindingdb.org), DrugCentral (http://drugcentral.org), the Therapeutic Target Database (http://db.idrblab.net/ttd/), PubChem (https://pubchem.ncbi.nlm.nih.gov), Pharos (https://pharos.nih.gov), PROMISCUOUS, and the IUPHAR/BPS Guide to PHARMACOLOGY. Additional resources include TransformerCPI (https://github.com/lifanchen-simm/transformerCPI), the SNAP library (http://snap.stanford.edu), and previously published kinase inhibitor datasets[23,58]. Protein structures were obtained from AlphaFold DB (https://www.alphafold.ebi.ac.uk/) and the RCSB PDB (https://www.rcsb.org/). Representative protein structures used for molecular visualization were obtained from AlphaFold DB and the Protein Data Bank, including https://alphafold.ebi.ac.uk/files/AF-P46095-F1-model_v4.pdb, https://alphafold.ebi.ac.uk/files/AF-O43826-F1-model_v4.pdb, https://alphafold.ebi.ac.uk/files/AF-P42224-F1-model_v4.pdb, and https://files.rcsb.org/download/1MBA.cif. Source data are provided with this paper on figshare: https://doi.org/10.6084/m9.figshare.30372982.

Code availability

The source code used to develop the model, perform the analyses, and generate the results in this study is publicly available under the Apache-2.0 license at: https://github.com/Yigang-Chen/SCOPE-DTI. The specific version of the code associated with this publication has been archived in Zenodo: https://doi.org/10.5281/zenodo.17217391[65]. Benchmark implementations of other methods used in this study are available at: MolTrans: https://github.com/kexinhuang12345/MolTrans, TransformerCPI: https://github.com/lifanchen-simm/transformerCPI, BertDTI: https://github.com/hskang0906/DTI-Prediction, DrugBAN: https://github.com/peizhenbai/DrugBAN.

References

1. Luo, Y. et al. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nat. Commun.* **8**, 573 (2017).
2. Pushpakom, S. et al. Drug repurposing: progress, challenges and recommendations. *Nat. Rev. Drug Discov.* **18**, 41–58 (2019).
3. Mizutani, S., Pauwels, E., Stoven, V., Goto, S. & Yamanishi, Y. Relating drug-protein interaction network with drug side effects. *Bioinformatics* **28**, i522–i528 (2012).
4. Smit, I. A. et al. Systematic analysis of protein targets associated with adverse events of drugs from clinical trials and postmarketing reports. *Chem. Res. Toxicol.* **34**, 365–384 (2021).
5. Friman, T. Mass spectrometry-based Cellular Thermal Shift Assay (CETSA®) for target deconvolution in phenotypic drug discovery. *Bioorg. Med. Chem.* **28**, 115174 (2020).
6. Huang, Y. et al. A robust drug-target interaction prediction framework with capsule network and transfer learning. *Int. J. Mol. Sci.* **24**, 14061 (2023).
7. Koutsoukas, A. et al. From in silico target prediction to multi-target drug design: current databases, methods and applications. *J. Proteom.* **74**, 2554–2574 (2011).
8. Chen, W., Liu, X., Zhang, S. & Chen, S. Artificial intelligence for drug discovery: resources, methods, and applications. *Mol. Ther. Nucleic Acids* **31**, 691–702 (2023).
9. Bagherian, M. et al. Machine learning approaches and databases for prediction of drug-target interaction: a survey paper. *Brief. Bioinforma.* **22**, 247–269 (2021).
10. Van Westen, G. J., Wegner, J. K., IJzerman, A. P., Van Vlijmen, H. W. & Bender, A. Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. *MedChemComm* **2**, 16–30 (2011).
11. Öztürk, H., Özgür, A. & Ozkirimli, E. DeepDTA: deep drug-target binding affinity prediction. *Bioinformatics* **34**, i821–i829 (2018).
12. Lee, I., Keum, J. & Nam, H. DeepConv-DTI: prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS Comput. Biol.* **15**, e1007129 (2019).
13. Karimi, M., Wu, D., Wang, Z. & Shen, Y. DeepAffinity: Interpretable deep learning of compound-protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics* **35**, 3329–3338 (2019).
14. Kang, H. et al. Fine-tuning of BERT model to accurately predict drug-target interactions. *Pharmaceutics* **14**, https://doi.org/10.3390/pharmaceutics14081710 (2022).
15. Bai, P., Miljković, F., John, B. & Lu, H. Interpretable bilinear attention network with domain adaptation improves drug–target prediction. *Nat. Mach. Intell.* **5**, 126–136 (2023).
16. Tsubaki, M., Tomii, K. & Sese, J. Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics* **35**, 309–318 (2019).
17. Luo, Y., Liu, Y. & Peng, J. Calibrated geometric deep learning improves kinase–drug binding predictions. *Nat. Mach. Intell.* **5**, 1390–1401 (2023).
18. Zheng, S., Li, Y., Chen, S., Xu, J. & Yang, Y. Predicting drug–protein interaction using quasi-visual question answering system. *Nat. Mach. Intell.* **2**, 134–140 (2020).
19. Huang, K., Xiao, C., Glass, L. M. & Sun, J. MolTrans: Molecular Interaction Transformer for drug-target interaction prediction. *Bioinformatics* **37**, 830–836 (2021).
20. Gilson, M. K. et al. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* **44**, D1045–1053 (2016).
21. Bai, P. et al. Hierarchical clustering split for low-bias evaluation of drug-target interaction prediction. In: *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 641–644, https://doi.org/10.1109/BIBM52615.2021.9669515 (IEEE, 2021).
22. Liu, H., Sun, J., Guan, J., Zheng, J. & Zhou, S. Improving compound-protein interaction prediction by building up highly credible negative samples. *Bioinformatics* **31**, i221–229 (2015).
23. Tang, J. et al. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *J. Chem. Inf. Model.* **54**, 735–743 (2014).
24. Chatterjee, A. et al. Improving the generalizability of protein-ligand binding predictions with AI-Bind. *Nat. Commun.* **14**, 1989 (2023).
25. Chen, L. et al. TransformerCPI: Improving compound-protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics* **36**, 4406–4414 (2020).
26. UniProt Consortium. UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2023).
27. Harding, S. D. et al. The IUPHAR/BPS Guide to PHARMACOLOGY in 2024. *Nucleic Acids Res.* **52**, D1438–D1449 (2024).
28. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
29. Kim, S. et al. PubChem 2023 update. *Nucleic Acids Res.* **51**, D1373–D1380 (2023).
30. Zdrazil, B. et al. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res.* **52**, D1180–D1192 (2024).
31. Tosco, P., Stiefl, N. & Landrum, G. Bringing the MMFF force field to the RDKit: Implementation and validation. *J. Cheminformat.* **6**, 37 (2014).
32. Zhang, C., Song, D., Huang, C., Swami, A. & Chawla, N. V. Heterogeneous Graph Neural Network. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 793–803, https://doi.org/10.1145/3292500.3330961 (ACM, Anchorage, AK, USA, 2019).
33. Wu, L. et al. MAPE-PPI: Towards effective and efficient protein-protein interaction prediction via microenvironment-aware protein

embedding, https://doi.org/10.48550/arXiv.2402.14391 (2024).
34. Jing, B., Eismann, S., Suriana, P., Townshend, R. J. L. & Dror, R. Learning from protein structure with geometric vector perceptrons, https://doi.org/10.48550/arXiv.2009.01411 (2021).
35. Li, M. et al. DGL-LifeSci: an open-source toolkit for deep learning on graphs in life science. *ACS Omega* **6**, 27233–27238 (2021).
36. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
37. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986).
38. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. https://doi.org/10.48550/arXiv.1802.03426 (2018).
39. Ankerst, M., Breunig, M. M., Kriegel, H.-P. & Sander, J. Optics: Ordering points to identify the clustering structure. *ACM Sigmod. Rec.* **28**, 49–60 (1999).
40. Atanasov, A. G., Zotchev, S. B., Dirsch, V. M. & Supuran, C. T. Natural products in drug discovery: Advances and opportunities. *Nat. Rev. Drug Discov.* **20**, 200–216 (2021).
41. Tang, Y. et al. Panax notoginseng alleviates oxidative stress through miRNA regulations based on systems biology approach. *Chin. Med.* **18**, 74 (2023).
42. Tam, D. N. H. et al. Ginsenoside Rh1: a systematic review of its pharmacological properties. *Planta Med.* **84**, 139–152 (2018).
43. Lyu, X. et al. Ginsenoside Rh1 inhibits colorectal cancer cell migration and invasion in vitro and tumor growth in vivo. *Oncol. Lett.* **18**, 4160–4166 (2019).
44. Wang, C. et al. Celastrol as an emerging anticancer agent: current status, challenges and therapeutic strategies. *Biomed. Pharmacother.* **163**, 114882 (2023).
45. Savage, S. R. et al. Pan-cancer proteogenomics expands the landscape of therapeutic targets. *Cell* **187**, 4389–4407.e15 (2024).
46. Uhlén, M. et al. Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
47. Hu, M. et al. Celastrol-induced Nur77 interaction with TRAF2 alleviates inflammation by promoting mitochondrial ubiquitination and autophagy. *Mol. Cell* **66**, 141–153.e6 (2017).
48. Menichetti, G., Barabási, A.-L. & Loscalzo, J. Chemical complexity of food and implications for therapeutics. *N. Engl. J. Med.* **392**, 1836–1845 (2025).
49. Coates, P. M. et al. The evolution of science and regulation of dietary supplements: past, present, and future. *J. Nutr.* **154**, 2335–2345 (2024).
50. Zhang, T. et al. Red rice extract as a biological UV filter and its photoprotective enhancement effects. *J. Dermatol. Sci. Cosmet. Technol.* 100102, https://doi.org/10.1016/j.jdsct.2025.100102 (2025).
51. Terwilliger, T. C. et al. AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. *Nat. Methods* **21**, 110–116 (2024).
52. Knox, C. et al. DrugBank 6.0: The DrugBank Knowledgebase for 2024. *Nucleic Acids Res.* **52**, D1265–D1275 (2024).
53. Avram, S. et al. DrugCentral 2023 extends human clinical data and integrates veterinary drugs. *Nucleic Acids Res.* **51**, D1276–D1287 (2023).
54. Zhou, Y. et al. TTD: Therapeutic Target Database describing target druggability information. *Nucleic Acids Res.* **52**, D1465–D1477 (2024).
55. Kelleher, K. J. et al. Pharos 2023: an integrated resource for the understudied human proteome. *Nucleic Acids Res.* **51**, D1405–D1416 (2023).
56. Gallo, K. et al. PROMISCUOUS 2.0: a resource for drug-repositioning. *Nucleic Acids Res.* **49**, D1373–D1380 (2021).
57. Leskovec, J. & Sosič, R. SNAP: a general-purpose network analysis and graph-mining library. *ACM Trans. Intell. Syst. Technol.* **8**, 1:1–1:20 (2016).
58. Davis, M. I. et al. Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* **29**, 1046–1051 (2011).
59. Townshend, R. J. L. et al. *ATOM3D: Tasks On Molecules in Three Dimensions* https://doi.org/10.48550/arXiv.2012.04035 (2022).
60. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library, https://doi.org/10.48550/arXiv.1912.01703 (2019).
61. Fey, M. & Lenssen, J. E. Fast graph representation learning with PyTorch geometric, https://doi.org/10.48550/arXiv.1903.02428 (2019).
62. Wang, M. et al. Deep graph library: a graph-centric, highly-performant package for graph neural networks, https://doi.org/10.48550/arXiv.1909.01315 (2020).
63. Pedregosa, F. et al. Scikit-learn: machine learning in Python, https://doi.org/10.48550/arXiv.1201.0490 (2018).
64. Harris, C. R. et al. Array programming with NumPy, https://doi.org/10.48550/arXiv.2006.10256 (2020).
65. Chen, Y. et al. Semi-inductive dataset construction and framework optimization for practical drug target interaction prediction with ScopeDTI. SCOPE-DTI, https://doi.org/10.5281/zenodo.17217391 (2025).

Acknowledgements

This work was financially supported by Shenzhen Science and Technology Program (JCYJ20220530143615035, H.-D.H); the Warshel Institute for Computational Biology funding from Shenzhen City and Longgang District (LGKCSDPT2025001, A.W.); Shenzhen-Hong Kong Cooperation Zone for Technology and Innovation (HZQB-KCZYB-2020056, P2-2022-HDH-001-A, H.-D.H); Guangdong Young Scholar Development Fund of Shenzhen Ganghong Group Co., Ltd. (2021E0005, 2022E0035, H.-D.H); Phase III Government Matching Fund of Shenzhen Ganghong Group Co., Ltd. (2023E0012, H.-D.H); Guangdong S&T Program (2024A0505050001 Y.-C.-D.L, 2024A0505050002 H.-Y.H); 2023 The Second Affiliated Hospital of the Chinese University of Hong Kong, Shenzhen Joint Fund Project (HUUF-MS-202308 H.-Y.H, HUUF-MS-202309 Y.-C.-D.L); Better Way Group - Chinese University of Hong Kong (Shenzhen) Warshel Joint Laboratory for skin health and active molecule innovation (2024E0087, H.-D.H); Shenzhen Science and Technology Innovation Program (JCYJ20250604141235046 H.-Y.H, JCYJ20250604141041017 Y.-C.-D.L). We would like to express our sincere gratitude to the Vincent & Lily Woo Foundation for their generous support of the Vincent & Lily Woo Fellowship in Memory of Dr Albert Wong (To Y.C.). This fellowship, endowed by the Vincent & Lily Woo Foundation, is provided through MCMIA Foundation Limited, and we are deeply grateful for their contribution to our research. We would also like to thank Mr. Cheng Xiang for his valuable assistance in the wet lab experiments, which greatly contributed to this work.

Author contributions

Y.C., X.J., Z.Z. (Ziyue Zhang), Z.Z. (Zihao Zhu), H.-Y.H. and H.-D.H. conceived the study. Data collection was conducted by Y.C. and Z.Z. (Ziyue Zhang), X.L., J.F., and Y.H. Y.C., Z.Z. (Ziyue Zhang), K.C. and A.W. designed the methodology. Computational analysis was performed by Y.C. and Z.Z. (Ziyue Zhang), Y.Z. (Yuming Zhou), K.C., C.S., X.L., S.C., and Y.H., with interpretability validation contributed by Y.C., and Z.Z. (Ziyue Zhang). Wetlab experiments were executed by X.J., Y.L., and Z.Z. (Zihao Zhu), Y.Z. (Yangyi Zhang), K.W., Y.-C.-D.L. The original manuscript was drafted by Y.C., X.J., and Z.Z. (Ziyue Zhang), C.S., H.-Y.H., S.-C.Y., T.Z., Y.-C.-D.L., and H.-D.H. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-66311-9.

Correspondence and requests for materials should be addressed to Hsien-Da Huang.

Peer review information *Nature Communications* thanks Anna Rita Bizzarri and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.