Validation guidelines for drug-target prediction methods

Tanoli, Ziaurrehman; Schulman, Aron; Aittokallio, Tero

2025-01-02
Taylor and Francis Ltd.
http://hdl.handle.net/10138/588728
Tanoli, Z, Schulman, A & Aittokallio, T 2025, 'Validation guidelines for drug-target prediction methods', Expert opinion on drug discovery, vol. 20, no. 1, pp. 31-45. https://doi.org/10.1080/17460441.2024.2430955

Taylor & Francis
Taylor & Francis Group

Expert Opinion on Drug Discovery

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/iedc20

Validation guidelines for drug-target prediction methods

Ziaurrehman Tanoli, Aron Schulman & Tero Aittokallio

To cite this article: Ziaurrehman Tanoli, Aron Schulman & Tero Aittokallio (2025) Validation guidelines for drug-target prediction methods, Expert Opinion on Drug Discovery, 20:1, 31-45, DOI: 10.1080/17460441.2024.2430955

To link to this article: https://doi.org/10.1080/17460441.2024.2430955

EXPERT OPINION ON DRUG DISCOVERY
2025, VOL. 20, NO. 1, 31–45
https://doi.org/10.1080/17460441.2024.2430955

Taylor & Francis
Taylor & Francis Group

REVIEW

Validation guidelines for drug-target prediction methods

Ziaurrehman Tanoli[a,b], Aron Schulman[a] and Tero Aittokallio[a,b,c,d]

[a]Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Helsinki, Finland; [b]iCAN Digital Precision Cancer Medicine Flagship, University of Helsinki and Helsinki University Hospital, Helsinki, Finland; [c]Institute for Cancer Research, Department of Cancer Genetics, Oslo University Hospital, Oslo, Norway; [d]Oslo Centre for Biostatistics and Epidemiology (OCBE), Faculty of Medicine, University of Oslo, Oslo, Norway

ABSTRACT

Introduction: Mapping the interactions between pharmaceutical compounds and their molecular targets is a fundamental aspect of drug discovery and repurposing. Drug-target interactions are important for elucidating mechanisms of action and optimizing drug efficacy and safety profiles. Several computational methods have been developed to systematically predict drug-target interactions. However, computational and experimental validation of the drug-target predictions greatly vary across the studies.

Areas covered: Through a PubMed query, a corpus comprising 3,286 articles on drug-target interaction prediction published within the past decade was covered. Natural language processing was used for automated abstract classification to study the evolution of computational methods, validation strategies and performance assessment metrics in the 3,286 articles. Additionally, a manual analysis of 259 studies that performed experimental validation of computational predictions revealed prevalent experimental protocols.

Expert opinion: Starting from 2014, there has been a noticeable increase in articles focusing on drug-target interaction prediction. Docking and regression stands out as the most commonly used techniques among computational methods, and cross-validation is frequently employed as the computational validation strategy. Testing the predictions using multiple, orthogonal validation strategies is recommended and should be reported for the specific target prediction applications. Experimental validation remains relatively rare and should be performed more routinely to evaluate biological relevance of predictions.

ARTICLE HISTORY
Received 30 April 2024
Accepted 14 November 2024

KEYWORDS
Drug-target interaction prediction; computational validation; experimental validation; target activity mapping; drug repurposing

1. Introduction

Identification of protein targets of investigational compounds and approved drugs is a fundamental task in drug discovery and repurposing efforts, respectively. Drug-target interactions guide the development of more effective and safer therapeutics with reduced risks, costs and time [1]. By mapping the molecular targets implicated in disease pathways, one can design drugs that specifically modulate these targets to enhance therapeutic efficacy. Similarly, systematic mapping of potent off-targets also helps to avoid anti-targets, i.e. proteins or pathways which may lead to drug side effects. This targeted approach not only accelerates drug development but also improves the likelihood of successful clinical outcomes, such as treatment efficacy and tolerability. Furthermore, repurposing of existing drugs for new therapeutic indications relies heavily on the identification of new disease-targets. Such off-target repurposing enables the exploration of novel treatment avenues at reduced costs and timeframes [2]. Over the past two decades, there has been a notable surge in both experimental mapping of compound-target activities [3–6], and in the development of numerous databases housing drug-target interactions for various drug and target classes [7,8]. This significant advancement in the field underscores the increasing emphasis on cataloging and disseminating comprehensive information pertaining to drug-target interactions.

The human proteome presents a vast array of potential therapeutic targets. Despite the rapid expansion in our knowledge of potent drug-target interactions, the current drugs still cover only a small subset of the potentially druggable target space [9]. Furthermore, many rare diseases still lack targeted treatments, and protein families such as ion channels and nuclear receptors are under-represented in the drug-target activity landscape. Comprehensive testing of all approved drugs against the entire proteome is not yet experimentally feasible using multi-dose affinity or target engagement assays. Furthermore, the increasing number of investigational or probe compounds, and the use of more disease-relevant experimental assays and model systems makes the experimental mapping studies costly and time-consuming. Consequently, there is a pressing need for computational prediction algorithms capable of identifying novel, potent targets of small-molecules. Ideally, these algorithms should be able to predict the binding

CONTACT Ziaurrehman Tanoli ziaurrehman.tanoli@helsinki.fi; Tero Aittokallio tero.aittokallio@helsinki.fi Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Helsinki, Finland

Supplemental data for this article can be accessed online at https://doi.org/10.1080/17460441.2024.2430955

Article highlights

- Bioactivity or binding affinity prediction with regression and docking methods were the most common computational methods for drug-target interaction prediction during the last decade. Deep learning algorithms and AlphaFold-based docking have also become popular in recent years.
- Cross-validation and benchmarking against state-of-the-art methods remain more popular computational validation strategies than independent testing. Correlation coefficients are widely used in regression-based studies and accuracy is the most popular metric in the activity classification.
- Six different computational cross-validation scenarios are commonly used to assess the generalization accuracy of drug-target predictions. However, some of the scenarios may lead to over-optimistic results due to potential data leakage, and are therefore not recommended.
- Various assay types are utilized for experimental validation of drug-target predictions, including biochemical, *in vitro* and *in vivo* approaches. Notably, we found that most of the computational methods for predicting drug-target interactions remain experimentally unvalidated.
- Predictive models could replace initial target activity screening and lead to an improved accuracy when predicting binding affinity profiles. This would reduce experimental costs, provided that large-enough target activity training data are available for the drug classes of interest.
- Natural language processing document classification enables a fast automated survey of a large number of research articles with a relatively high classification accuracy. We have made the scripts available on GitHub (https://github.com/AronSchulman/DTI-abstract-classifier).

affinity of drugs across a large set of potential targets spanning several protein families. Such algorithms would significantly broaden the targetable space, particularly for approved drugs, which would enhance their therapeutic potential and expand the scope of drug discovery and repurposing efforts.

Over the past decade, thousands of drug-target interaction prediction methods have been developed and extensively described. Numerous reviews cover the computational methodologies utilized in predicting drug-target affinities or activity classes, reflecting the substantial advancements made in this field. For example, Ding et al. [10] reviewed similarity-based machine learning (ML) techniques for predicting drug – target interactions. Bagherian et al. [11] published an overview of ML-based approaches and databases utilized in prediction of drug-target interactions, providing comparisons of their respective merits and drawbacks. Koutsoukas et al. [12] conducted a survey of prominent databases suitable for ligand-based target prediction and delineated methods applicable for target prediction based on both ligand bioactivity knowledge and protein structural information. Ezzat et al. [13] explored various drug-target interaction prediction methods and elucidated the types of input data employed for such predictions. Abbasi et al. [14] conducted a comprehensive overview of deep learning (DL) methods for drug-target prediction. Paananen et al. [15] compared several drug target discovery platforms leveraging omics knowledge to identify novel drug-targets. Lastly, Sachdev et al. [16] reviewed feature-based methodologies and discussed metrics utilized in diverse drug-target interaction prediction approaches.

To our knowledge, none of the previously published reviews or surveys have specifically focused on evaluating the validation strategies employed in drug-target prediction studies. Therefore, the primary aim of this survey was to evaluate the diverse computational and experimental validation strategies employed in drug-target interaction prediction studies. To fill this gap, we devised a PubMed search query and retrieved 3,286 drug-target interaction prediction articles published in the last decade. We further implemented scripts to automatically extract pertinent metadata, including journal information, abstracts, titles, and publication years for all the articles. Subsequently, we developed NLP-based document classifiers to automatically categorize the 3,286 articles into 16 distinct categories. These categories were related to the model type (e.g. machine learning or docking-based), the validation methodology employed (such as cross-validation or experimental validation), and the evaluation metrics utilized in the studies. These automated methods enabled us to implement a fast literature analysis and study the evolution of computational methodologies and validation strategies over the years. Additionally, we carried out a manual analysis of 259 studies that performed experimental analyses to survey various experimental protocols used in these studies to validate the drug-target predictions from the computational models.

2. Literature survey and NLP-based article classification

2.1. PubMed search query

To retrieve articles on drug-target interaction prediction published within the last decade, we adapted the following PubMed query:

((‘ligand receptor’ OR ‘ligand protein’ OR ‘compound target’ OR ‘target affinity’ OR ‘drug target’ OR ‘binding affinity’ OR ‘Ligand-receptor binding’ OR ‘target potency’ OR ‘drug-target interactions’) AND (‘prediction’ OR ‘algorithm’) NOT (review [Publication Type]) NOT (news[Publication Type]) NOT (newspaper article[Publication Type]) NOT (systematic review [Publication Type]) NOT (editorial[Publication Type]) AND (y_10[Filter]))

We tailored our search query based on the methodology outlined by Cichońska et al. [17] in their PubMed literature scan section. In addition to the original search terms, we included three more specific terms: ‘ligand-receptor binding,’ ‘target potency,’ and ‘drug-target interactions.’ We aimed to maximize the overlap between the articles retrieved from PubMed query and those citing three well-known drug-target interaction studies: Yamanishi et al. [4], Pahikkala et al. [18], and Tang et al [19]. Each of the three studies has received over 400 citations to date, and our PubMed search query retrieved drug-target articles having > 42% overlap with the articles citing these three studies based on PubMed citations (Supplementary Figure S1).

2.2. Data extraction using PubMed API

The PubMed query yielded PubMed IDs of the 3,286 articles published in the last decade (2014–2023). To obtain additional metadata necessary for our literature analysis, such as journal names, abstracts, titles, and publication years, we developed

scripts utilizing PubMed's API. This facilitated the systematic extraction of comprehensive information essential for two main objectives: (1) to analyze the evolution of drug-target interaction prediction methods over the years, and (2) to train NLP-based document classifiers, as elaborated in the subsequent sections. The scripts utilized in this study are openly available, facilitating their reuse for various applications by the community (https://github.com/AronSchulman/DTI-abstract-classifier).

Upon programmatic extraction of journal information for the 3,286 articles, it became apparent that 'Journal of Biomolecular Structure and Dynamics' (4.38% of the articles), 'Journal of Chemical Information and Modeling' (3.83%), and 'Briefings in Bioinformatics' (2.89%) emerged as the three most prominent journals in this PubMed query (Figure 1). However, there were also other, more computationally oriented journals, that had published several drug-target interaction prediction studies, such as Bioinformatics and BMC Bioinformatics. Most of these journals do not require experimental validations of computational model predictions.

2.3. NLP-based PubMed document classifiers

We developed 16 NLP-based document binary classifiers to predict 16 distinct sentiments associated with each of the 3,286 articles (yes or no classification). Six of these sentiments pertain to drug-target interaction prediction methodologies: machine learning, deep learning, attention-based, docking-based, classification and regression. Four sentiments focus on validation methods: experimental validation, cross-validation, independent testing, and state-of-the-art comparison. The remaining six sentiments involve evaluation metrics used: correlation (Spearman or Pearson), root mean square error (RMSE), AUROC, accuracy, F1 measure, and any other relevant metric.

For each of the 16 document classifiers, we utilized a pre-trained BioMed-RoBERTa [20] model obtained from Hugging Face [21] as the base model. BioMed-RoBERTa has been pre-trained on 2.68 million full-text scientific articles, amounting to 7.55 billion training tokens. The model achieves state-of-the-art results on natural language processing tasks within the biomedical domain, surpassing more general-purpose language models [20]. Thus, this base model was an obvious choice for our use case.

The classifiers were fine-tuned for 10 epochs using manually labeled training data. In total, 16 separate training datasets were collected manually from the 3,286 articles retrieved by the PubMed query. We concatenated the title and abstract of each article to form a document as the input, since obtaining full text for all the articles proved challenging. The model output is a binary label indicating the presence (1) or absence (0) of a particular sentiment. Training datasets were balanced to ensure an equal proportion of positive and negative documents, with training data sizes ranging from 32 to 180 documents, depending on the complexity of the 16 classification tasks (Figure 2(a)).

We optimized the model hyperparameters, including the learning rate, weight decay, and batch size, using random search over 20 iterations. The best hyperparameters are detailed in Supplementary Table S1. To evaluate the model performance, we employed a stratified 5-fold cross-validation and obtained reasonable accuracies (>0.86) and F1 scores (>0.89) for each of the 16 classification tasks (Figure 2(b)). The varying difficulty of the classification tasks was reflected in the model accuracies, with some performing very well on relatively small training data (e.g. 'Docking'), and others requiring more training data for comparatively weaker performance (e.g. 'Machine Learning'). Notably, there was no correlation either between the training data set size and accuracy (Pearson correlation −0.04), or the data size and F1 score (Pearson correlation −0.11) across the different categories (Supplementary Figure S2).

We further validated 50 predicted articles through a manual review of the classification results. These 50 articles were randomly selected from the list of predictions and were not part of the training data used for the model development. 6 articles were excluded from the analysis due to either lack of relevance to drug targets or unavailability of full-text access. The remaining 44 articles were manually validated. The model demonstrated a relatively accurate classification of abstracts into 16 categories, related to modeling and validation methodologies, achieving an accuracy above 0.79 and an F1 score exceeding 0.47 (Supplementary Figure S3). However, categories related to

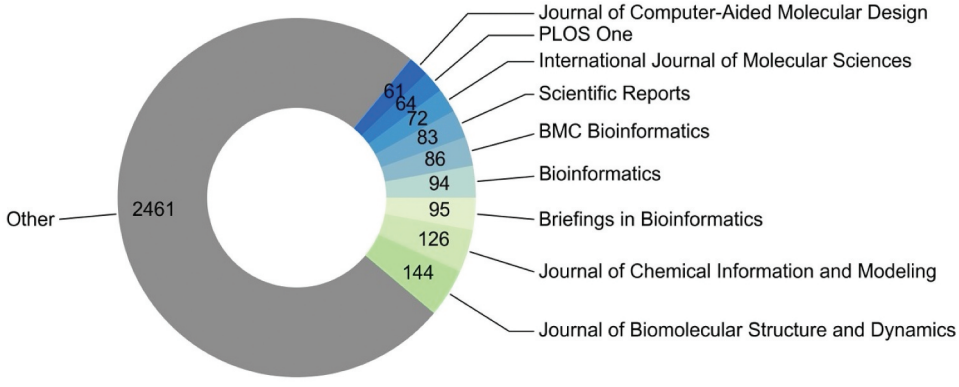

Figure 1. Distribution of journals of the 3,286 publications in our survey. The publications were retrieved from PubMed with the search query related to drug-target interaction prediction studies (see Section 2.1). The nine most prominent journals are highlighted, with the journal of biomolecular structure and dynamics being the most frequently observed journal.

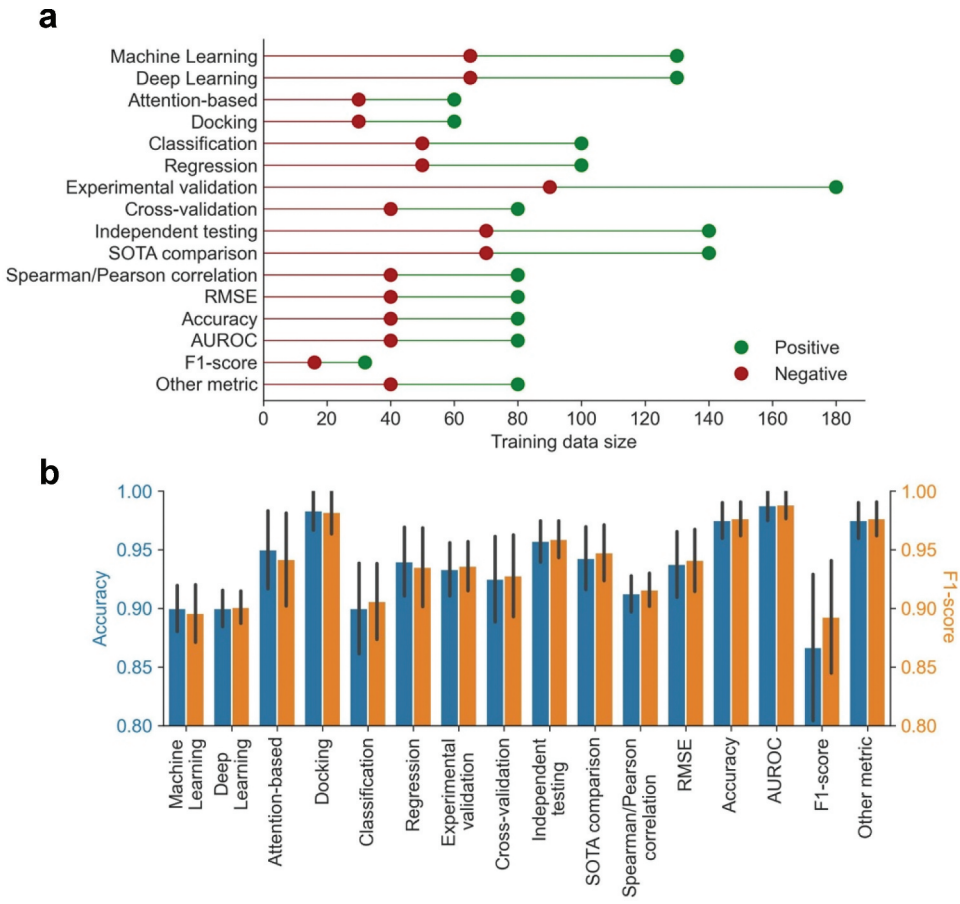

Figure 2. Nlp-based PubMed document classifier. (a) The composition of training data for each document classifier, balanced to contain equal proportions of positive and negative documents as shown in green and red, respectively. (b) The document classifiers were trained and tested using a stratified five-fold CV for each document category, i.e. we trained and evaluated five binary classifiers for each category. The blue and orange bars show the average accuracies and F1 scores of the five classifiers, respectively, and the black vertical lines indicate the standard errors of the mean (SEM). Note: the y-axis starts from 0.80 to show the differences between the 16 classification tasks; the theoretical minimum value of both the accuracy and F1 score is zero.

performance metrics proved more challenging for automated classification, likely due to inconsistent references to metrics in the abstracts. Independent test results can be found in

3. Computational validation of drug-target interaction predictions

Figure 3 shows an overall workflow for drug-target prediction modeling and their validation. The first phase represents different types of compound and target representations used as modeling input, available in-house or from public data resources, while the second phase outlines prediction model types, such as machine learning and docking methods. The third phase shows the prediction labels, including bioactivity values or docking scores for the top hits. The final phase covers computational and experimental validation assays for evaluating prediction accuracy and relevance. In the below, we focus first on the main prediction method types and their computational validation, while Section 4 describes the main experimental validation assays.

3.1. Classification and regression methods for drug-target prediction

In drug discovery, predicting how strongly drugs interact with their protein targets can be computationally approached in two main ways: classification or regression. Classification models require training data from both positive (active) and negative (inactive) classes, which can be obtained from databases like ChEMBL [22] and Drug Target Commons (DTC) [23,24]. Information on inactive targets of compounds (i.e. protein targets against which the compound is non-potent based on biochemical or cell-based assays) is typically available in the 'Activities' table of these databases, often labeled as 'inactive' in the 'Activity_comment' column. This information is often based on the annotations made by the original screening researchers, not the database maintainers. It is important to note that missing measurements among compound-target pairs do not necessarily mean that the compound is inactive against that protein. Instead, missing pairs simply indicate the absence of data from biochemical or cell-based assays for that specific compound-target pair in the database. This could be due to the pair not having been experimentally studied so far, or because the measurement not yet been incorporated into the database.

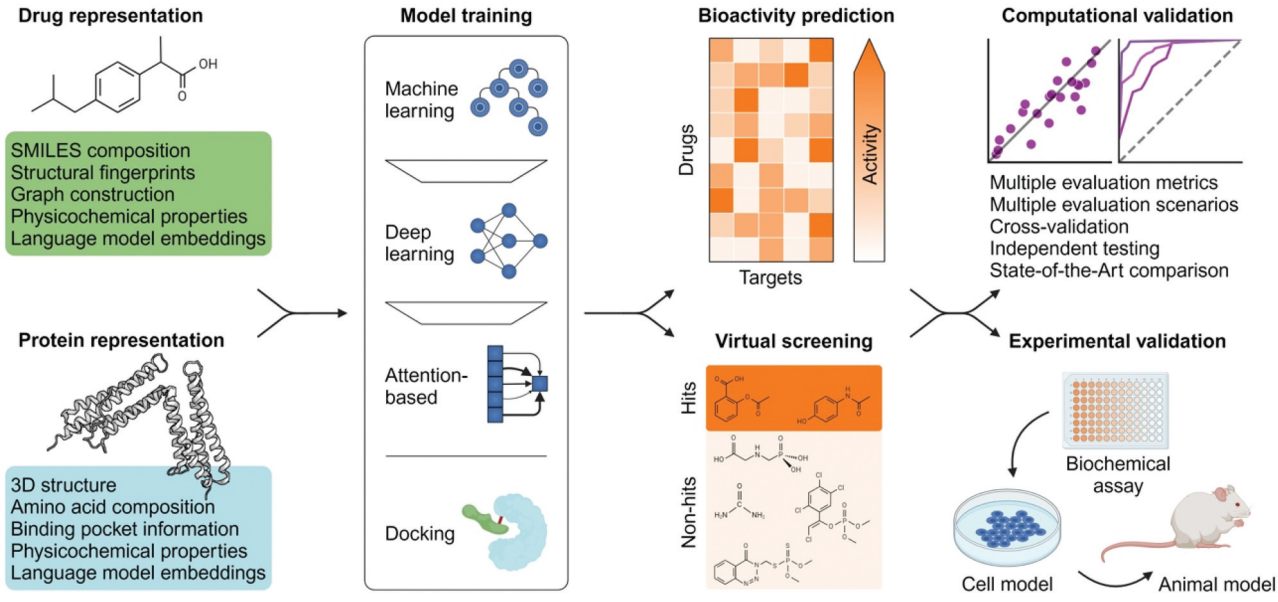

Figure 3. A schematic workflow for drug-target interaction prediction and validation. Relevant biological and structural features of drug-target pairs available in-house or in public bioactivity data resources are converted into numerical representations as model inputs for computational prediction approaches. The model predictions should be evaluated with multiple computational and experimental validation schemes, depending on the target class and application.

Regression models, on the other hand, use the whole continuum of bioactivity values without the need to distinguish the positive and negative interaction classes. This makes the training data collection somewhat easier, since no predefined activity cut-offs are needed. However, bioactivity databases often report various types of dose-response readouts, such as IC_{50}, K_d and K_i, which can complicate the preparation of the training datasets. To address this, databases like ChEMBL have unified these affinities into a single score, such as pChEMBL value, which represents the negative logarithm of molar concentration across seven affinity types, namely IC_{50}, EC_{50}, XC_{50}, AC_{50}, K_i, K_d and potency. It is important to note that pChEMBL value does not incorporate information on the specific protein target families or assay types. To overcome this limitation, Drug Target Profiler [25] introduced a scoring system that maps bioactivity affinity values into continuous scores ranging from zero to one. This system accounts for variations in target families and assay formats, assigning different weights to specific target superfamilies or assay types. By doing so, it enhances the development of accurate regression models across a wide range of target superfamilies.

3.2. Computational evaluation strategies for drug-target predictions

Generally, there are three types of computational validation approaches: cross-validation (CV), state-of-the-art (SOTA) comparison, and independent testing (i.e. hold-out validation). In CV, the dataset is split into *k* subsets or folds. The model is then trained on *k*-1 folds and tested on the remaining fold. This process is repeated *k* times, with each fold being used as the test set once. The performance metrics are finally averaged over the *k* runs. The SOTA comparison identifies current state-of-the-art method(s), against which the proposed model is compared on one or more benchmark datasets. Finally, independent testing involves evaluating the performance of the model predictions on a separate dataset that was not used during the model development or training. For different protein target superfamilies, we recommend applying each of the validation approach independently to each superfamily to ensure robust and accurate evaluation across the various drug target classes.

Target superfamilies often encompass a wide range of structurally diverse proteins with distinct binding pockets and functional mechanisms, hence making a model trained in one superfamily most likely inaccurate at predicting compound-target interactions in another. Computational methods need to account for this variability to accurately predict ligand-binding affinities, target-selectivity profiles, and functional effects. However, it requires additional curation efforts to subdivide the existing benchmark datasets in terms of the superfamilies. Lack of experimentally tested compound-target interactions across some superfamilies (e.g. ion channels) may further deteriorate the performance of the models for the smaller superfamilies. We therefore suggest that experimental scientists should generate more drug-target activity data on less-studied superfamilies, including ion channels, nuclear receptors, and epigenetic regulators.

Computational validation methods may share similarities between the classification and regression problems, whereas the evaluation metrics employed differ significantly. In drug-target classification models, binary metrics such as accuracy, F1 score, and Area Under the Receiver Operating Characteristic Curve (AUROC) are commonly utilized. Conversely, regression-based models are typically assessed using Spearman or Pearson correlation coefficients, concordance index and root mean squared error (RMSE). Additionally, there exist also other evaluation metrics tailored to specific modeling objectives [26]. Employing a variety of evaluation metrics enables a comprehensive assessment of the performance and

robustness of drug-target prediction models. Such multi-metric approach aids in gaining insights into the model's accuracy across various aspects of classification and regression tasks.

3.3. Computational evaluation of virtual screening methods

Evaluation approaches can also share similarities with virtual screening (VS) studies. VS is a related computational process of ranking molecules in large databases based on their predicted binding affinity toward a specific target. Similar to ML-based drug-target interaction prediction, VS aims to prioritize the most potential drug candidates that should proceed to experimental testing, thus saving time and costs. The accuracy of a VS algorithm can be evaluated using existing benchmark datasets. However, some of these benchmarks, including DUD-E benchmark [27], contain biases that may reward memorization, rather than true learning and generalization [28,29]. A common source of bias is the extensive use of decoys, i.e. ligands assumed to be inactive that are designed to have similar physical properties to true actives, but differing topologically [28,30]. DUDE-Z benchmark improves on the limitations of DUD-E by revising its method of generating the decoys [31]. Additionally, the DUDE-Z introduced Extrema and Goldilocks benchmarks, which address charge imbalances and account for the broad features of larger docking libraries, respectively [31]. Another recommended, bias-aware benchmark is LIT-PCBA [32], a carefully-curated subset of the PubChem BioAssay database [33]. Notably, there are no decoys present in the LIT-PCBA. The unbiased nature of LIT-PCBA renders it a considerably more challenging evaluation dataset compared to many other VS benchmarks [32].

Evaluation metrics for VS algorithms should properly address the 'early recognition problem,' where active ligands must be ranked at the very top of the hit list to make sure they are included in the final selection. Common evaluation metrics, such as AUROC, do not consider the early recognition problem properly, and are thus not ideal for VS method evaluation, especially when ultra-large libraries are screened [34]. The Boltzmann-Enhanced Discrimination of ROC (BEDROC) [35] improves the standard AUROC with an exponential weighting to accommodate early recognition. Analogous to AUROC, the worst and best values of BEDROC are 0 and 1, respectively. Enrichment factor (EF) is another commonly-used VS evaluation metric that measures how many active ligands are within the predefined top fraction of the hit list, compared to what is expected by a random chance. EF remains a widely used metric in VS, even though it is dependent on the arbitrarily chosen threshold and may be potentially misleading [31].

3.4. Evolution of drug-target prediction methods during 2014–2023

After fine-tuning and optimizing the 16 document classifiers, we utilized them to automatically predict sentiments across the 3,286 drug-target prediction studies to study their methodological aspects (Supplementary Figure S4). In terms of prediction methods, regression and docking were the most common sentiments, appearing in 2116 (64.4%) and 1736 (52.8%) studies, respectively. Cross-validation and SOTA validation strategies were implemented in 758 (23.1%) and 659 (20.1%) studies, respectively, which makes them more popular compared to independent testing, observed only in 295 (9.0%) studies. Spearman/Pearson correlations and RMSE are the two most widely used metrics in regression-based studies found in 1,599 (48.7%) and 1150 (35.0%) studies. Similarly, accuracy was the most popular metric in the activity classification studies, adapted in 1,593 (48.5%) of drug-target classification studies (Supplementary Figure S4).

To study the evolution of computational methods over the years, we visualized the trends of drug-target prediction studies using each of the 16 sentiments during the last decade (Figure 4). To make more robust document classifications, we used voting of the 5 model ensemble classifiers, where the predicted label was assigned according to the consensus of at least 3 out of the five models in the 5-fold CV. When using a stricter prediction scheme (all five models must agree on the outcome), we observed a reduced number of positive labels across sentiments, especially in CV, Spearman/Pearson correlation, RMSE, accuracy, and F1 score (see Supplementary Figure S5). It is likely that the stricter scheme introduces more false negatives; however, the results provide a good starting point for the manual analysis of the articles, due to the smaller number of positives ($n = 259$). This can be alleviated in future by increasing training data set sizes.

Compared to drug-target activity classification, regression-based methods are designed to predict the continuous binding affinities between drug-target pairs. Throughout the years, research articles focusing on regression-based approaches have consistently outnumbered those studied focused on activity classification for drug-target interactions (Figure 4). Especially, starting from 2021, there has been a notable surge in the proportion of articles employing regression-based techniques. This increase may be attributed to the significant expansion of the available drug-target binding affinity datasets for various drug and target classes, such as those curated in ChEMBL and other databases. For example, ChEMBL 25th release contained 15.5 million bioactivities [36], while its latest release (ChEMBL 33rd) encompasses approximately 20 million bioactivities [22].

As expected, studies using DL and attention-based methods began to emerge prominently only after 2020 (Figure 4). Notably, the number of DL-based studies even surpassed traditional ML drug-target interaction prediction methods from 2022 onwards, as DL is generally considered to offer higher accuracy compared to traditional methods when the amount of high-quality training dataset is large enough for DL model estimation [37]. Interestingly, our document classifier revealed that the ratio of purely computational prediction studies to those incorporating experimental validation has remained relatively constant over the years (approximately 25%; see Supplementary Figure 6). Even if a notable rise in the number of drug-target prediction studies has taken place during this period, there has not been a similar increase in the number of studies

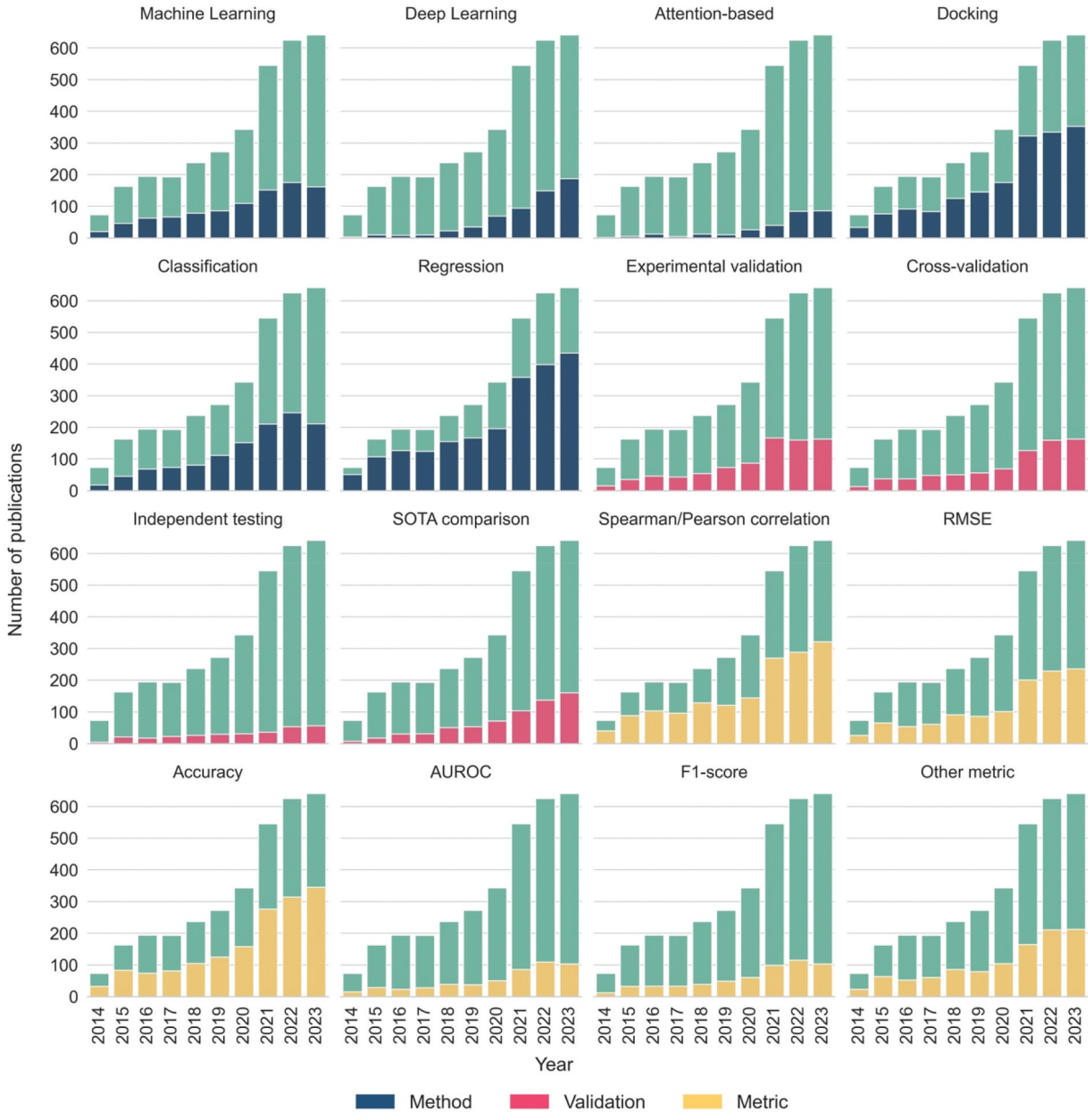

Figure 4. The evolution of computational drug-target prediction methods and evaluation practices over the past decade. The publications are obtained from PubMed with the search query (Section 2.1). The green bars show the total number of articles published each year, and the colored parts of the stacked bars indicate the proportion of studies that include the given method class (blue), validation strategy (red), or evaluation metric (yellow). These proportions were predicted with our suite of transformer-based ensemble document classifiers.

performing experimental validation, which should be done more routinely.

In the beginning of the decade (year 2014), articles focusing on docking methods (45.2%) were much more numerous than those based on the supervised ML approaches for drug-target interaction prediction (27.4%). Docking studies further experienced a sudden surge observed between 2021 and 2023 (Figure 4). This surge in docking is most likely attributed to the first release of AlphaFold [38] in 2021, which may have prompted a shift in the research focus toward molecular docking. We expect many more AlphaFold-based docking studies published in the coming years, as new and extended versions of AlphaFold are being published and predictions become available in databases [39,40]. However, it is also important to keep on evaluating the prediction accuracy and confidence of AlphaFold and other DL-based protein structure prediction methods [41], when used as part of drug-target interaction prediction pipelines [42].

3.5. Data splitting scenarios for validating compound-target predictions

In supervised machine learning tasks, the model evaluation often involves *k*-fold cross-validation or independent test set. In drug-target prediction context, these validation strategies assume that the pairs to be predicted are randomly distributed within the known (experimentally measured) interaction matrix. However, in paired input scenarios, such as predicting

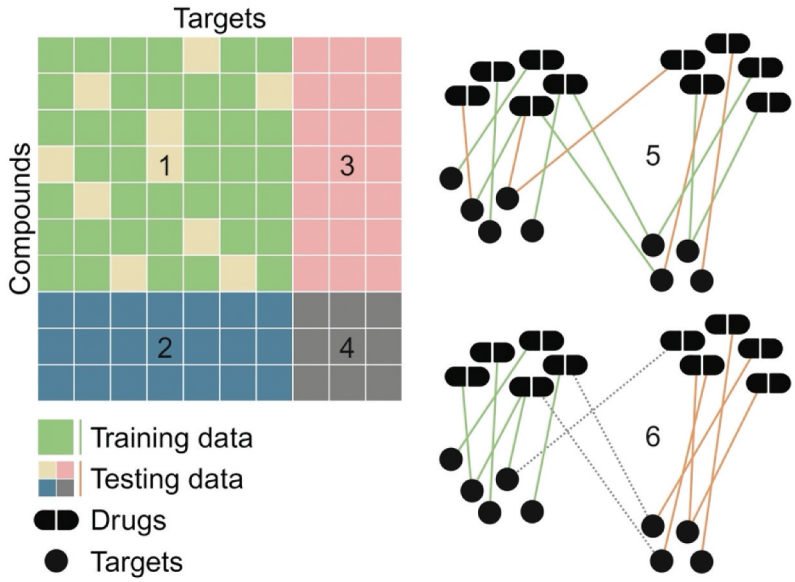

Figure 5. Common data splitting scenarios for validating compound-target prediction methods. 1: bioactivity imputation, 2: new compounds, 3: new targets, 4: new compounds and new targets, 5: random split, and 6: fully dissimilar split. A test drug-target pair qualifies for scenario 1 if the pairwise interaction involves a compound and target that are both present in the training data; for scenario 2 if the compound is absent and target is present in the training data; for scenario 3 if the compound is present and target is absent in the training set; for scenario 4 if both the compound and target are absent in the training data. Scenario 5 can include interactions between any compounds and target pairs. For scenario 6, all the test compounds and targets must be structurally dissimilar enough to those found in the training data. Structural similarity is represented by compound and target clusters in the figure. The dashed lines signify interactions that must be discarded from both training and testing, as otherwise they would violate the dissimilarity assumption.

drug-target interactions, it is crucial to consider different prediction scenarios, where the training and test data share either common or different sets of drugs or targets. This ensures a more accurate assessment of the model's performance in these specific contexts. Six scenarios are commonly employed to validate predictions of drug-target affinities (Figure 5):

(1) **Bioactivity imputation**: testing currently missing interactions among compounds and targets that are all present in the training set.
(2) **New compounds**: testing interactions between unseen compounds (absent in the training set) and the targets present in the training data.
(3) **New targets**: testing interactions between unseen targets (absent in the training set) with the compounds present in the training data.
(4) **New compounds and new targets**: testing interactions between unseen compounds and targets, neither of which were present in the training data.

These four scenarios were introduced by Pahikkala et al. [18], and they are typically applied to independent testing data separate from the dataset used for model training (either using CV or hold-out set). Scenarios 1 and 3 are particularly relevant for drug repurposing efforts, where the goal is to predict new potent targets for approved drugs that may extend their usage for other diseases where the new targets play a role. On the other hand, scenarios 2 and 4 are more suitable for target-based drug discovery applications, where the primary aim is to identify new targeted inhibitors for specific disease-related targets. The last two scenarios, recommended by Guvenilir et al. [43], involve other types of splitting the original dataset into training and test sets:

(5) **Random split**: divides the dataset randomly into training and test folds. However, this often leads to over-optimistic results, due to potential data leakage in terms of drug and target similarity. Highly similar compounds between training and test sets may lead to overfitting, where the model is unable to generalize to novel chemical domains. Thus, random split is not recommended as a general validation strategy.
(6) **Dissimilar split**: neither similar compounds nor targets are shared between the training and test folds. Similarity can be determined based on pairwise comparison between compound structures (SMILES) and target sequences (fasta).

3.6. Chemical properties affect drug-target prediction models and evaluation

Considering the chemical structural similarities is crucial for drug-target prediction and evaluation. Structural analysis helps computational models recognize patterns in molecular structures that may correlate with biological activity, often resulting in improved test set accuracies, especially when the training dataset contains very similar compounds. However, highly similar compounds between training and test sets may lead to overfitting, and that the model is unable to generalize to novel chemical domains. Therefore, the best ML models can make use of also a structurally diverse set of compounds to provide accurate

test compound activity predictions [17]. Structural similarity can be computed using various fingerprint algorithms, such as Molecular Access System (MACCS) [44] and featmorgan fingerprints [45]. The RDKit package [46] in Python or R computes these and other structural fingerprints. Structural similarity analysis is particularly useful in hit identification and lead optimization by leveraging the relationship between chemical structure and biological function [47].

When developing and evaluating drug-target interaction prediction models, it is also essential to consider physicochemical and geometric properties such as molecular protonation state and chirality [48]. The protonation state of a molecule, which is determined by the pH of the environment, affects its charge distribution, geometry, and overall stability. This state can significantly influence the binding affinity of a drug to its target by affecting hydrogen bonding, electrostatic interactions, and the molecule's conformation [49]. Since many drug targets are enzymes operating within specific pH ranges, the protonation state is crucial for ensuring a good fit within an enzyme's active site. Chirality refers to the chemical property where a molecule and its mirror image are not superimposable. These mirror images, or enantiomers, are labeled as R (rectus) and S (sinister). Enantiomers can exhibit different biological activities: one might be beneficial while the other could be inactive or harmful. Therefore, identifying and utilizing the correct chiral form of a drug is vital for its efficacy and safety [50].

Many molecular docking software, such as Schrödinger [51], can predict and optimize protonation states and chiral configurations. In Schrödinger's Maestro interface, the Epik module predicts and optimizes protonation states. Users can import the molecule into Maestro, open the Epik tool, set the desired pH range, and run the prediction. Epik then analyzes the local environment and pKa values to generate and rank possible protonation states, allowing the selection of the most relevant state for further analysis. For chirality, Schrödinger's ConfGen module within Maestro can be used. After importing the molecule, ConfGen specifies chiral centers and generates possible stereoisomers, enabling the evaluation of different chiral configurations. This aids in identifying the most biologically active form of the molecule. Together, these tools provide a comprehensive analysis of protonation states and chiral configurations, thereby enhancing the precision of drug-target prediction models.

In summary, taking into account structural similarity, chirality, and protonation states is vital for accurate and generalizable drug-target prediction models. Therefore, these properties should be considered when splitting the data for model training and validation, in addition to considering the drug and target identifiers in the paired input prediction problems. Ensuring the use of the correct chiral forms in both training and testing enhances the model's predictive accuracy, and aligning protonation states between training and test sets improves the reliability of predicting drug-target interactions. Integrating these molecular properties facilitates the development of robust and precise prediction models applicable to real-world scenarios.

3.7. Large language models in drug discovery

Large language models (LLMs), including transformers, present a cost-effective and efficient solution for various drug discovery applications, and several LLMs have been recently proposed and used in this domain. For example, recent studies [52,53] have utilized LLMs such as BERT [54], SciBERT [55], BioBERT [56], BioMed-RoBERTa [57] and BlueBERT [58] to identify drug-target interactions from experimental literature. Additionally, another recent study successfully applied BERT to predict protein and peptide interactions, demonstrating its versatility in biological drug discovery [59]. LLMs have also been used in translating between drug molecules and indications [60]. Furthermore, GPT-based architectures have been employed in other drug discovery tasks, including *de novo* drug design [61] and biological question answering [62]. These examples underscore the broad applicability of LLMs in advancing research and innovation in drug discovery and related areas. However, similar to the other computational methods, predictions from LLMs need to be carefully validated using both computational and experimental approaches.

4. Experimental validation of drug-target interaction predictions

To further validate that the predicted drug-target interactions are either biologically or pharmaceutically meaningful, in a specific disease or cellular context, the predictions should be validated in wet-lab or animal experiments. Ideally, multiple orthogonal assays and disease models should be employed. Even though experimental validations require dedicated infrastructure, additional resources and sometimes lengthy wet-lab or animal experiments, which may not be available at all sites, at least targeted experimental investigations in cell models should be done for selected drugs-target interactions, for instance, as a collaboration or service. This should be also required by the scientific journals. Target engagement or bioactivity validation is the only way to really confirm that the computational predictions have the potential to progress to the next phases in the drug discovery or development process. Ideally, the drug-target binding experiments should be designed after the predictions, hence serving as fully blinded validation data [63]. This is in contrast with the computational validation strategies, using e.g. CV or hold-out sets of experimental data. In these strategies, the validation data are already available during the model training [64], which may lead to intentional or unintentional information leakage, and hence to over-optimistic prediction accuracies. However, if the wet-lab experiments are designed after making the predictions, the experimental validation data could not be seen by the model or the modeler. This blinded approach helps to avoid any potential information leakage, provided that the modeler (person who trains the model and makes the predictions) and the experimentalist (who makes the experimental validations) do not share the outcome information in the process. Ideally, both positive predictions (drug-target activities) and negative predictions (inactive predictions [52]) should be tested and randomized in the experimental setting, that is, the experimentalist is blinded to the predicted outcome.

Many of the drug-target interaction studies classified by the NLP classifier as 'experimental validation' involved studies that did not validate drug-protein interactions *per se*, rather they validated predicted drug activities in a specific cellular or disease context. In the list of 259 articles from the 'stricter prediction scheme' (see Section 3.4), there were also studies where the drug-target interactions did not come from a specific docking, statistical or ML model. Instead, the drug-target interaction predictions were made based on a biological hypothesis or general exploration of previously collected biological or cellular data, followed by pathway or network analyses to come up with a handful of target activity hypotheses. In the following sections, we focus on studies that use experimental investigations to validate a specific computational model, whether classification or regression-based, which has the potential to systematically predict either compound-target affinities or their activity classes across a wide variety of compound classes and/or protein families. This excludes targeted prediction models that can be applied to a particular drug or target only, e.g. standard structure-activity relationship (QSAR) models [65]. However, the recommendations below should be useful also for target-specific QSAR and other model classes in their experimental validation. The specific examples below were selected such that they could be applied to various drug classes and protein families. We note that the selected studies are not meant to be the only recommended studies but serve as examples of studies that have performed relatively rigorous experimental validations of the computational model predictions.

4.1. Biochemical large-scale validation studies

Biochemical target activity experiments can be performed as a large-scale validation of hundreds of compound-target interactions in dose-response assays. If done at a single fixed dose only, these assays can be easily scaled-up to thousands or even tens of thousands of compounds tested across hundreds of proteins in high-throughput screening (HTS). However, dose-response biochemical assays are recommended because the activity window of compounds may vary drastically depending on the drug and target class. Consequently, single-dose assays at a selected concentration (e.g. 1000 nM) may not be easily comparable between the tested compound-protein pairs from various drug or target classes.

Similarity Ensemble Approach (SEA) is a quantitative prediction method that groups a wide range of proteins based on topological chemical similarity among their ligands annotated into known drug targets [66]. Statistical significance of the resulting similarity scores, expressed as a minimum spanning tree, can then be used to map the protein target sets together. This relatively simple statistical model has revealed links among unexpected targets, some of which were experimentally validated using biochemical binding assays (and later with cell-based functional assays, see Section 4.2) [66]. The SEA approach was also applied to systematically predict new targets for 3,665 FDA-approved drugs. Overall, 30 novel off-targets were tested experimentally, covering many protein families, using radioligand competition binding assays, and 23 of these (77%) yielded inhibition constants (K_i) less than 15 μM, five of which were relatively potent (K_i <100 nM) [67].

VirtualKinomeProfiler extends the SEA strategy and captures distinct representations of chemical similarity space using an ensemble support vector machine (eSVM) algorithm for systematic kinase activity classification [68]. The web-tool enables the profiling of compounds across 248 kinases simultaneously. The experimental validations were performed using both single-dose and dose-response biochemical activity assays. The eSVM prediction model led to a 1.5-fold increase in precision and 2.8-fold decrease in false discovery rate, when compared with single-dose biochemical screening. A similar validation strategy was used in a comparative benchmarking of predictive algorithms for kinase inhibitor target activity prediction across kinase families. The validations used dose-response dissociation constant (K_d) assays and found that an ensemble of the top-performing ML models led to an accuracy exceeding that of single-dose activity screening [17].

Recently, many DL models have been developed for target activity prediction. Currently, protein kinases have arguably the largest amount of high-quality and quantitative binding affinity data available for DL model training and computational validation. For instance, the transfer learning model ConPLex predicts drug-target binding on many types of target classes by using a pre-trained protein language model [69] . It makes predictions of binding based on the distance between learned representations, enabling predictions at the scale of massive compound libraries and the full human proteome. Experimental testing of 19 kinase-drug predictions validated 12 interactions (63%), including four with subnanomolar affinity, and a strongly binding EPHB1 inhibitor (K_d = 1.3 nM).

For antiviral applications, Han et al. developed a webserver and DL model, D3AI-Spike, based on convolutional neural network (CNN), transformer and attention mechanisms to quickly predict binding affinity between SARS-CoV-2 spike receptor binding domain (RBD) mutations and human angiotensin-converting enzyme 2 (hACE2), the receptor of SARS-CoV-2 [70]. D3AI-Spike was validated by first predicting and then experimentally determining the binding affinity of a variant under monitoring IHU (B.1.640.2), which harbors many mutations and deletions in the spike protein. A noncompetitive ELISA assay was used to measure the affinity constant (K_{aff}) of spike protein RBD, demonstrating a strong predictive power of the D3AI-Spike platform.

While biochemical assays can provide the first level of validation that a compound has the potential to modulate a given target protein, activity from such assays does not yet confirm that the compound-target interaction would lead either to desired cellular or disease phenotypes. Furthermore, biochemical testing of the drug-target interactions across a large panel of protein targets, both those predicted to be active and inactive, is important for the selectivity analysis, i.e. determining how target-specific the compound activity is against a particular target protein (e.g. disease target of interest) [71]. Negative predictions of inactive interactions should also be tested experimentally to investigate the negative prediction accuracy of the computational model [63].

4.2. In vitro validation studies in cell models

Cell-based phenotypic assays provide a second-level confirmation that the predicted compound-target interactions lead to desired cellular phenotypes, such as selective killing of cancer cells or virus-infected cells, in the disease-relevant cellular context. In addition to anticancer and antiviral applications, cell-based validation can be used to guide phenotypic screening when identifying bioactive, antibacterial agents. Based on a large-scale screen performed against Mycobacterium tuberculosis (Mtb), Mugumbate et al. constructed an ensemble of SEA, a naive Bayesian classifier and docking algorithm to predict potential targets for an anti-tuberculosis phenotypic effect [72]. They confirmed two compounds in *M. bovis* resistant strains as potent inhibitors of dihydrofolate reductase (DHFR), an essential Mtb gene that is clinically validated as a drug target.

Ariey-Bonnet et al. used a data-driven predictive tool, MolTarPred [73], to investigate the mechanism of action (MoA) of mebendazole (MBZ), an antihelminthic drug, repurposed in the treatment of brain tumors [74]. MolTarPred returns the most similar target-annotated molecules to the user-supplied query molecule, here MBZ, enabling the identification of putative targets based on chemical similarity. Validation experiments in glioblastoma cells using thermal shift and NanoBRET target engagement assays showed that MBZ binds to MAPK14/p38α and inhibits its kinase activity *in vitro* and *in cellular* in a dose-dependent manner. Furthermore, gene silencing by RNA interference confirmed that MAPK14 plays a key role in the cytotoxic activity of MBZ against glioblastoma (GBM) cells, representing a promising druggable target in GBM.

Another structure-based model, DRUIDom, presents a statistical approach to identify new interactions between drug compounds and their biological targets by utilizing the modular structure of proteins [75]. The method enables a large-scale mapping of small molecule compounds and their clustering based on molecular similarities. The DRUIDom predictions were tested on selected proteins that play critical roles in the progression of numerous types of cancer, and the cell-based experiments indicated that the predicted inhibitors are effective even on drug-resistant cancer cells. In particular, the authors showed that compounds predicted to target LIM-kinase proteins, earlier implicated in regulating cell motility and cell cycle progression, significantly block the cancer cell migration by inhibiting LIMK phosphorylation and the downstream protein cofilin.

DEEPScreen uses CNNs to learn features from the compounds' 2D representations, instead of using structural fingerprints [76]. DEEPScreen predicted JAK proteins as new targets of cladribine, an anti-neoplastic drug approved for specific forms of lymphoma and leukemia. The model predictions were experimentally validated in hepatocellular carcinoma (HCC) cell lines through phosphorylation of STAT3, a downstream effector of JAK proteins. The experimental data suggests that cladribine acts on JAK/STAT3 signaling and induces apoptosis in HCC cells. Another DL model, OverfitDTI, uses deep neural networks (DNNs) to learn predictive features of both the drugs' chemical space and the targets' molecular space [77]. Two predicted compounds were experimentally shown to inhibit TEK kinase activity and block endothelial cell tube formation.

While cell-based assays can confirm predicted target modulation or target engagement in the selected cellular context, it is important to choose the cell models carefully so that they present well the desired disease context to make the validations translationally meaningful. Furthermore, it is important to include also multiple control cell models, in which the computationally predicted activities are not expected to be observed, e.g. with either low/high expression of the target protein or wild-type version in cancer applications, to assess how context- and target-specific the predictions are. For instance, inhibiting the same target both in mutant and wild type cell lines indicates that the compound may lead to toxic side effects on nonmalignant cells.

4.3. In vivo validation studies in animal models

Animal studies are used to study the drug treatment responses *in vivo*, as well as the MoA of the drugs using molecular profiling of biospecimens from the treated animals. Such experiments can naturally be done only for a few selected drugs, and the profiling of the samples after the treatment provides only indirect evidence for target validation. For instance, Guo et al. investigated the MoA of a traditional Chinese medicine, Danhong injection (DHI), used in the treatment of acute myocardial infarction (AMI). The authors combined molecular docking with compound-target and disease-target protein-protein interaction network analyses, and validated the predictions using protein analyses in cardiac tissue specimens of the treated Sprague-Dawley rats [78]. They demonstrated calcium signaling pathway as a potential mechanism of DHI.

Zhao et al. developed a DL-based drug-target interaction prediction method, DLDTI, based on network representation learning and CNNs. Low-dimensional feature vectors were used to train DLDTI to achieve an optimal mapping space and to infer new drug-target interactions by ranking them according to their proximity to the optimal mapping space [79]. Experimental validation of the predicted targets of tetramethylpyrazine (TMPZ) on atherosclerosis progression was carried out in hamsters. More specifically, protein analysis of platelets was carried out by Western blotting, and the atherosclerotic plaque cell composition was studied with immunohistochemistry analysis of the aortic root of the treated animals. The authors concluded that TMPZ could attenuate atherosclerosis by inhibiting signal transductions in platelets.

While animal models can assess the compound activity *in vivo*, it is important to note that for many diseases and drugs there are not yet representative animal models that would be predictive of the eventual human *in vivo* effects and clinical outcomes. The evaluation of a potential toxicity of the treatment and the toxicity-related off-target effects is also not routinely studied in most animal experiments, even though that should be the case. It should be noted that testing the drugs and their MoA in multiple concentrations is incompatible with the '4 R' principles of animal ethics, due to addition of extra animal groups per each dose. Furthermore, predicting the clinically applicable dose based on the animal treatment data is not straightforward, and dose-escalating studies in humans are still needed to evaluate the tolerable doses and potential drug adverse effects.

5. Conclusion

In this study, we conducted a survey of drug-target interaction prediction studies and developed NLP-based document classifiers to automatically examine the methodology of 3,286 articles published in the last decade for drug-target interaction prediction. The use of the automated document classification facilitated a fast survey of a large corpus of articles relatively accurately (Figure 2, Supplementary Figure S3). We have made the analysis scripts available on GitHub (https://github.com/AronSchulman/DTI-abstract-classifier), so that others can employ and expand upon similar approaches in future studies. We further manually analyzed 259 out of the 3,286 studies that additionally performed experimental validation, since the automated analysis was not capable of distinguishing whether the experiments were done before or after the computational predictions.

Based on the literature survey, we found that most of the computational algorithms for predicting drug-target interactions remain experimentally unvalidated. Even though rigorous computational validations, such as CV and hold-out sets, help to avoid reporting over-optimistic prediction accuracies when performed correctly, experimental validation remains critical. Experimental investigations can demonstrate that the predicted target activities are either biologically or pharmaceutically meaningful in a specific disease or cellular context, and therefore have also the potential to progress to the next phases of the drug discovery or development process. Therefore, both computational and experimental validations are required to fully evaluate the generalization and translational power of a prediction model, respectively.

This review provides a comprehensive overview of various types of approaches to identify drug-target interactions, with different validation strategies and evaluation metrics employed across diverse prediction approaches. We also described advantages and limitations of various data splitting strategies for validating compound-target predictions (Section 3.5), and highlighted the impact of chemical properties on the performance and evaluation of drug-target prediction models (Section 3.6). By providing insights into these critical aspects, the review underscores the importance of selecting appropriate methodologies and validation techniques to enhance the accuracy and reliability of drug-target interaction predictions. Overall, this analysis serves as a resource for guiding future research and optimizing model performance in drug discovery applications.

6. Expert opinion

Our analysis revealed that most computational methods are trained and tested on relatively small kinase target benchmark datasets, such as Davis [6] and KIBA [19], which contain only a few hundred kinase targets and kinase inhibitors. Therefore, we recommend that future developers should train their prediction methods on larger bioactivity datasets, available in databases such as ChEMBL [22], Drug Target Commons [24], and BindingDB [80], which contain continuous target activity data across multiple protein families and drug classes. These data can be used to investigate the correct applicability domain and potential cross-activity of the compounds across protein families. A comprehensive and rigorous validation approach significantly enhances model's reliability for drug discovery or repurposing applications.

Furthermore, we recommend that benchmark datasets, including Davis [6], Metz [5], KIBA [19], and Yamanishi [4], should be either completely reserved for testing, as hold-out validation data, or that the exact CV folds used for training and testing will be clearly reported, as demonstrated in the DTITR method [81]. Otherwise, it may become challenging to compare the prediction accuracies in future studies, due to variations in data preprocessing and splitting across studies, potentially biasing continuous model performance evaluations. To enable continuous benchmarking of existing and emerging methods, one should also make the training and testing data available in public repositories, such as GitHub or Zenodo, in addition to the codes of the prediction algorithm for others to freely use and extend.

Additionally, we recommend the incorporation of uncertainty estimation approaches in drug-target prediction research, such as confidence scoring by conformal prediction [82]. Conformal prediction is a lightweight statistical method that generates prediction sets (for classification) or intervals (for regression) instead of point predictions. The prediction sets or intervals are then used for quantifying the confidence of individual predictions. A conformal classifier can inform the end-user about a specific prediction whether it is too uncertain about the outcome. In regression, it is difficult to predict narrow, more actionable activity intervals with high confidence. The relationship between model confidence levels and the resulting prediction interval widths will facilitate rigorous model comparisons [83].

Experimental validation is critical for confirming that the compound-target interaction leads to desired cellular phenotypes in the predicted cell-context. However, control experiments, including both control cell lines (e.g. nonmalignant or wild-type cells) as well as negative predictions (i.e. non-active drug-target interactions) should be included to test the cell context-selectivity and target-specificity of the effects, respectively. Single-dose biochemical activity assays are often used as the first HTS in kinome-wide experimental studies, and only those drug-kinase interactions showing single-dose activity are further tested in dose-response assays [6]. However, ML prediction models could replace the single-dose screening, with an improved accuracy when predicting affinity levels [17,68]. This will lead to reduced experimental costs, provided large-enough target activity training data are available for the drug and target classes of interest.

Funding

T.A. was supported by Academy of Finland [grants 326238, 340141, 344698, and 345803], Norwegian Health Authority South-East [grants 2020026 and 2023105], the Cancer Society of Finland, the Norwegian Cancer Society, and the Sigrid Jusélius Foundation. ZT was supported by the Academy of Finland [grant 351507].

Declaration of interest

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict

with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

Reviewers disclosure

Peer reviewers on this manuscript have no relevant financial or other relationships to disclose.

References

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. Schuhmacher A, Hinder M, Und Stein AVS, et al. Analysis of pharma R&D productivity – a new perspective needed. Drug Discovery Today. 2023;28(10):103726. doi: 10.1016/j.drudis.2023.103726
2. Pushpakom S, Iorio F, Eyers PA, et al. Drug repurposing: progress, challenges and recommendations. Nat Rev Drug Discov. 2019;18(1):41–58. doi: 10.1038/nrd.2018.168

•• **This review presents several approaches and challenges for systematic drug repurposing.**

3. Reinecke M, Brear P, Vornholz L, et al. Chemical proteomics reveals the target landscape of 1,000 kinase inhibitors. Nat Chem Biol. 2024;2023(5):577–585. doi: 10.1038/s41589-023-01459-3
4. Yamanishi Y, Kotera M, Kanehisa M, et al. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. Bioinformatics. 2010;26(12):i246–i254. doi: 10.1093/bioinformatics/btq176

• **Provides binary benchmark datasets for four superfamilies of protein targets.**

5. Metz JT, Johnson EF, Soni NB, et al. Navigating the kinome. Nat Chem Biol. 2011;7(4):200–202. doi: 10.1038/nchembio.530
6. Davis MI, Hunt JP, Herrgard S, et al. Comprehensive analysis of kinase inhibitor selectivity. Nat Biotechnol. 2011;29(11):1046–1051. doi: 10.1038/nbt.1990

•• **One of the first large-scale screens of kinase inhibitor target activities.**

7. Tanoli Z, Seemab U, Scherer A, et al. Exploration of databases and methods supporting drug repurposing: a comprehensive survey. Brief Bioinform. 2020;22(2):1656–1678. doi: 10.1093/bib/bbaa003

•• **Review presents pros and cons of 102 public resources supporting drug repurposing.**

8. Carpenter KA, Altman RB. Databases of ligand-binding pockets and protein-ligand interactions. Comput Struct Biotechnol J. 2024;23:1320–1338. doi: 10.1016/j.csbj.2024.03.015
9. Santos R, Ursu O, Gaulton A, et al. A comprehensive map of molecular drug targets. Nat Rev Drug Discov. 2017;16(1):19–34. doi: 10.1038/nrd.2016.230
10. Ding H, Takigawa I, Mamitsuka H, et al. Similarity-based machine learning methods for predicting drug–target interactions: a brief review. Brief Bioinform. 2014;15(5):734–747. doi: 10.1093/bib/bbt056
11. Bagherian M, Sabeti E, Wang K, et al. Machine learning approaches and databases for prediction of drug–target interaction: a survey paper. Brief Bioinform. 2021;22(1):247–269. doi: 10.1093/bib/bbz157
12. Koutsoukas A, Simms B, Kirchmair J, et al. From in silico target prediction to multi-target drug design: current databases, methods and applications. J Proteomics. 2011;74(12):2554–2574. doi: 10.1016/j.jprot.2011.05.011
13. Ezzat A, Wu M, Li X-L, et al. Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey. Brief Bioinform. 2019;20(4):1337–1357. doi: 10.1093/bib/bby002
14. Abbasi K, Razzaghi P, Poso A, et al. Deep learning in drug target interaction prediction: current and future perspectives. Curr Med Chem. 2021;28(11):2100–2113. doi: 10.2174/0929867327666200907141016
15. Paananen J, Fortino V. An omics perspective on drug target discovery platforms. Brief Bioinform. 2020;21(6):1937–1953. doi: 10.1093/bib/bbz122
16. Sachdev K, Gupta MK. A comprehensive review of feature based methods for drug target interaction prediction. J Biomed Inform. 2019;93:103159. doi: 10.1016/j.jbi.2019.103159
17. Cichońska A, Ravikumar B, Allaway RJ, et al. Crowdsourced mapping of unexplored target space of kinase inhibitors. Nat Commun. 2021;12(1):3307. doi: 10.1038/s41467-021-23165-1

•• **Crowdsourced community-wide benchmarking of predictive algorithms for kinase inhibitor potencies across multiple kinase families.**

18. Pahikkala T, Airola A, Pietilä S, et al. Toward more realistic drug–target interaction predictions. Brief Bioinform. 2015;16(2):325–337. doi: 10.1093/bib/bbu010

•• **The first study that describes the various challenges and gives guidelines for computational evaluation of drug-target interaction predictions.**

19. Tang J, Szwajda A, Shakyawar S, et al. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. J Chem Inf Model. 2014;54(3):735–743. doi: 10.1021/ci400709d
20. Gururangan S, Marasović A, Swayamdipta S, et al. Don't stop pretraining: adapt language models to domains and tasks. arXiv preprint arXiv:2004.10964. 2020.
21. Wolf T, Debut L, Sanh V, et al. Huggingface's transformers: state-of-the-art natural language processing. *ArXiv*. arXiv preprint arXiv:1910.03771. 2019.
22. Zdrazil B, Felix E, Hunter F, et al. The ChEMBL database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Res. 2023;52(D1):D1180–D1192. doi: 10.1093/nar/gkad1004
23. Tang J, Tanoli Z-R, Ravikumar B, et al. Drug target commons: a community effort to build a consensus knowledge base for drug-target interactions. Cell Chem Biol. 2018;25(2):224–229. doi: 10.1016/J.CHEMBIOL.2017.11.009
24. Tanoli Z, Alam Z, Vähä-Koskela M, et al. Drug target commons 2.0: a community platform for systematic analysis of drug–target interaction profiles. Database. 2018;2018:1–13. doi: 10.1093/database/bay083
25. Tanoli Z, Alam Z, Ianevski A, et al. Interactive visual analysis of drug–target interaction networks using drug target profiler, with applications to precision medicine and drug repurposing. Brief Bioinform. 2020;21(1):211–220. doi: 10.1093/bib/bby119
26. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128–138. doi: 10.1097/EDE.0b013e3181c30fb2
27. Mysinger MM, Carchia M, Irwin JJ, et al. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. J Med Chem. 2012;55(14):6582–6594. doi: 10.1021/jm300687e
28. Chen L, Cruz A, Ramsey S, et al. Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. PLOS ONE. 2019;14(8):e0220113. doi: 10.1371/journal.pone.0220113
29. Wallach I, Heifets A. Most ligand-based classification benchmarks reward memorization rather than generalization. J Chem Inf Model. 2018;58(5):916–932. doi: 10.1021/acs.jcim.7b00403
30. Chaput L, Martinez-Sanz J, Saettel N, et al. Benchmark of four popular virtual screening programs: construction of the active/decoy dataset remains a major determinant of measured performance. J Cheminform. 2016;8(1):1–17. doi: 10.1186/s13321-016-0112-z
31. Stein RM, Yang Y, Balius TE, et al. Property-unmatched decoys in docking benchmarks. J Chem Inf Model. 2021;61(2):699–714. doi: 10.1021/acs.jcim.0c00598
32. Tran-Nguyen V-K, Jacquemard C, Rognan D. LIT-PCBA: an unbiased data set for machine learning and virtual screening. J Chem Inf Model. 2020;60(9):4263–4273. doi: 10.1021/acs.jcim.0c00155

33. Wang Y, Bryant SH, Cheng T, et al. PubChem BioAssay: 2017 update. Nucleic Acids Res. 2016;45(D1):D955–D963. doi: 10.1093/nar/gkw1118
34. Lyu J, Irwin JJ, Shoichet BK. Modeling the expansion of virtual screening libraries. Nat Chem Biol. 2023;19(6):712–718. doi: 10.1038/s41589-022-01234-w
35. Truchon J-F, Bayly CI. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. J Chem Inf Model. 2007;47(2):488–508. doi: 10.1021/ci600426e
36. Mendez D, Gaulton A, Bento AP, et al. ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res. 2019;47(D1):D930–D940. doi: 10.1093/nar/gky1075
37. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–444. doi: 10.1038/nature14539
38. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–589. doi: 10.1038/s41586-021-03819-2
•• **The first version of AlphaFold protein structure prediction algorithm.**
39. Varadi M, Bertoni D, Magana P, et al. AlphaFold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. Nucleic Acids Res. 2024;52(D1):D368–D375. doi: 10.1093/nar/gkad1011
40. Abramson J, Adler J, Dunger J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024;2024(8016):493–500. doi: 10.1038/s41586-024-07487-w
41. Terwilliger TC, Liebschner D, Croll TI, et al. AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. Nat Methods. 2024;21(1):110–116. doi: 10.1038/s41592-023-02087-4
42. Voitsitskyi T, Stratiichuk R, Koleiev I, et al. 3DProtDTA: a deep learning model for drug-target affinity prediction based on residue-level protein graphs. RSC Adv. 2023;13(15):10261–10272. doi: 10.1039/D3RA00281K
43. Atas Guvenilir H, Doğan T. How to approach machine learning-based prediction of drug/compound–target interactions. J Cheminform. 2023;15(1):16. doi: 10.1186/s13321-023-00689-w
44. Durant JL, Leland BA, Henry DR, et al. Reoptimization of MDL keys for use in drug discovery. J Chem Inf Comput Sci. 2002;42(6):1273–1280. doi: 10.1021/ci010132r
45. Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model. 2010;50(5):742–754. doi: 10.1021/ci100050t
46. Bento AP, Hersey A, Félix E, et al. An open source chemical structure curation pipeline using RDKit. J Cheminform. 2020;12(1):1–16. doi: 10.1186/s13321-020-00456-1
47. Yi J, Lee S, Lim S, et al. Exploring chemical space for lead identification by propagating on chemical similarity network. Comput Struct Biotechnol J. 2023;21:4187–4195. doi: 10.1016/j.csbj.2023.08.016
48. Rudrapal M, Egbuna C. Computer aided drug design (CADD): from ligand-based methods to structure-based approaches. Elsevier; 2022. https://shop.elsevier.com/books/computer-aided-drug-design-cadd-from-ligand-based-methods-to-structure-based-approaches/rudrapal/978-0-323-90608-1.
49. Lasham J, Djurabekova A, Zickermann V, et al. Role of protonation states in the stability of molecular dynamics simulations of high-resolution membrane protein structures. J Phys Chem B. 2024;128(10):2304–2316. doi: 10.1021/acs.jpcb.3c07421
50. Ceramella J, Iacopetta D, Franchini A, et al. A look at the importance of chirality in drug activity: some significative examples. Appl Sci. 2022;12(21):10909. doi: 10.3390/app122110909
51. Friesner RA, Banks JL, Murphy RB, et al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J Med Chem. 2004;47(7):1739–1749. doi: 10.1021/jm0306430
52. Aldahdooh J, Vähä-Koskela M, Tang J, et al. Using BERT to identify drug-target interactions from whole PubMed. BMC Bioinformatics. 2022;23(1):1–13. doi: 10.1186/s12859-022-04768-x
53. Aldahdooh J, Tanoli Z, Tang J, et al. Mining drug–target interactions from biomedical literature using chemical and gene descriptions-based ensemble transformer model. Bioinf Adv. 2024;4(1):vbae106. doi: 10.1093/bioadv/vbae106
54. Devlin J, Chang M-W, Lee K, et al. Bert: pre-training of deep bidirectional transformers for language understanding. arXiv Prepr arXiv: 1810.04805. 2018.
55. Beltagy I, Lo K, Cohan A. Scibert: a pretrained language model for scientific text. arXiv Prepr arXiv: 1903.10676. 2019.
56. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234–1240. doi: 10.1093/bioinformatics/btz682
57. Liu Y, Ott M, Goyal N, et al. Roberta: a robustly optimized BERT pretraining approach. arXiv Prepr arXiv: 1907.11692. 2019.
58. Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on Ten benchmarking datasets. arXiv Prepr arXiv: 1906.05474. 2019.
59. Gurvich R, Markel G, Tanoli Z, et al. Peptriever: a Bi-encoder approach for large-scale protein–peptide binding search. Bioinformatics. 2024;40(5):btae303. doi: 10.1093/bioinformatics/btae303
60. Oniani D, Hilsman J, Zang C, et al. Emerging opportunities of using large language models for translation between drug molecules and indications. Sci Rep. 2024;14(1):10738. doi: 10.1038/s41598-024-61124-0
61. Zhavoronkov A, Ivanenkov YA, Aliper A, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat Biotechnol. 2019;37(9):1038–1040. doi: 10.1038/s41587-019-0224-x
62. Wang C, Long Q, Meng X, et al. BioRAG: a RAG-LLM framework for biological question reasoning. arXiv Prepr arXiv: 2408.01107. 2024.
63. Cichonska A, Ravikumar B, Parri E, et al. Computational-experimental approach to drug-target interaction mapping: a case study on kinase inhibitors. PLoS Comput Biol. 2017;13(8):e1005678. doi: 10.1371/journal.pcbi.1005678
64. Halder AK, Cordeiro MNDS. Development of multi-target chemometric models for the inhibition of class I PI3K enzyme isoforms: a case study using QSAR-Co tool. Int J Mol Sci. 2019;20(17):4191. doi: 10.3390/ijms20174191
65. Tropsha A. Best practices for QSAR model development, validation, and exploitation. Mol Inform. 2010;29(6-7):476–488. doi: 10.1002/minf.201000061
66. Keiser MJ, Roth BL, Armbruster BN, et al. Relating protein pharmacology by ligand chemistry. Nat Biotechnol. 2007;25(2):197–206. doi: 10.1038/nbt1284
• **Introduction of the similarity ensemble approach (SEA) to rapidly search large compound databases and to build cross-target similarity maps.**
67. Keiser MJ, Setola V, Irwin JJ, et al. Predicting new molecular targets for known drugs. Nature. 2009;462(7270):175–181. doi: 10.1038/nature08506
68. Ravikumar B, Timonen S, Alam Z, et al. Chemogenomic analysis of the druggable kinome and its application to repositioning and lead identification studies. Cell Chem Biol. 2019;26(11):1608–1622. doi: 10.1016/j.chembiol.2019.08.007
69. Singh R, Sledzieski S, Bryson B, et al. Contrastive learning in protein language space predicts interactions between drugs and protein targets. Proc Natl Acad Sci. 2023;120(24):e2220778120. doi: 10.1073/pnas.2220778120
70. Han J, Liu T, Zhang X, et al. D3AI-Spike: a deep learning platform for predicting binding affinity between SARS-CoV-2 spike receptor binding domain with multiple amino acid mutations and human angiotensin-converting enzyme 2. Comput Biol Med. 2022;151:106212. doi: 10.1016/j.compbiomed.2022.106212
71. Wang T, Pulkkinen OI, Aittokallio T. Target-specific compound selectivity for multi-target drug discovery and repurposing. Front Pharmacol. 2022;13:1003480. doi: 10.3389/fphar.2022.1003480
72. Mugumbate G, Abrahams KA, Cox JAG, et al. Mycobacterial dihydrofolate reductase inhibitors identified using chemogenomic methods and in vitro validation. PLOS ONE. 2015;10(3):e0121492. doi: 10.1371/journal.pone.0121492

73. Peón A, Li H, Ghislat G, et al. MolTarPred: a web tool for comprehensive target prediction with reliability estimation. Chem Biol Drug Des. 2019;94(1):1390–1401. doi: 10.1111/cbdd.13516
74. Ariey-Bonnet J, Carrasco K, Le Grand M, et al. In silico molecular target prediction unveils mebendazole as a potent MAPK14 inhibitor. Mol Oncol. 2020;14(12):3083–3099. doi: 10.1002/1878-0261.12810
75. Doğan T, Akhan Güzelcan E, Baumann M, et al. Protein domain-based prediction of drug/Compound–target interactions and experimental validation on LIM kinases. PLoS Comput Biol. 2021;17(11):e1009171. doi: 10.1371/journal.pcbi.1009171
76. Rifaioglu AS, Nalbat E, Atalay V, et al. Deepscreen: high performance drug–target interaction prediction with convolutional neural networks using 2-D structural compound representations. Chem Sci. 2020;11(9):2531–2557. doi: 10.1039/C9SC03414E
77. Xiaolin X, Xiaozhi L, Guoping H, et al. Overfit deep neural network for predicting drug-target interactions. iScience. 2023;26(9):107646. doi: 10.1016/j.isci.2023.107646
78. Guo S, Tan Y, Huang Z, et al. Revealing calcium signaling pathway as novel mechanism of danhong injection for treating acute myocardial infarction by systems pharmacology and experiment validation. Front Pharmacol. 2022;13:839936. doi: 10.3389/fphar.2022.839936
79. Zhao Y, Zheng K, Guan B, et al. DLDTI: a learning-based framework for drug-target interaction identification using neural networks and network representation. J Transl Med. 2020;18(1):1–15. doi: 10.1186/s12967-020-02602-7
80. Gilson MK, Liu T, Baitaluk M, et al. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res. 2016;44(D1):D1045–D1053. doi: 10.1093/nar/gkv1072
81. Monteiro NRC, Oliveira JL, Arrais JP. DTITR: end-to-end drug–target binding affinity prediction with transformers. Comput Biol Med. 2022;147:105772. doi: 10.1016/j.compbiomed.2022.105772
82. Alvarsson J, McShane SA, Norinder U, et al. Predicting with confidence: using conformal prediction in drug discovery. J Pharm Sci. 2021;110(1):42–49. doi: 10.1016/j.xphs.2020.09.055
83. Oršolić D, Šmuc T, Martelli PL. Dynamic applicability domain (DAD): compound–target binding affinity estimates with local conformal prediction. Bioinformatics. 2023;39(8):btad465. doi: 10.1093/bioinformatics/btad465