ADCNet: a unified framework for predicting the activity of antibody-drug conjugates

Liye Chen‡, Biaoshun Li‡, Yihao Chen‡, Mujie Lin, Shipeng Zhang, Chenxin Li, Yu Pang, Ling Wang*

Guangdong Provincial Key Laboratory of Fermentation and Enzyme Engineering, Joint International Research Laboratory of Synthetic Biology and Medicine, Guangdong Provincial Engineering and Technology Research Center of Biopharmaceuticals, School of Biology and Biological Engineering, South China University of Technology, No. 382 Waihuan East Road, Higher Education Mega Center, Guangzhou 510006, China

*Corresponding author: Ling Wang, School of Biology and Biological Engineering, South China University of Technology, No. 382 Waihuan East Road, Higher Education Mega Center, Guangzhou 510006, China. E-mail: lingwang@scut.edu.cn

‡Liye Chen, Biaoshun Li, and Yihao Chen authors contributed equally.

Abstract

Antibody-drug conjugates (ADCs) have revolutionized the field of cancer treatment in the era of precision medicine due to their ability to precisely target cancer cells and release highly effective drugs. Nevertheless, the rational design and discovery of ADCs remain challenging because the relationship between their quintuple structures and activities is difficult to explore and understand. To address this issue, we first introduce a unified deep learning framework called ADCNet to explore such relationship and help design potential ADCs. The ADCNet highly integrates the protein representation learning language model ESM-2 and small-molecule representation learning language model functional group-based bidirectional encoder representations from transformers to achieve activity prediction through learning meaningful features from antigen and antibody protein sequences of ADC, SMILES strings of linker and payload, and drug-antibody ratio (DAR) value. Based on a carefully designed and manually tailored ADC data set, extensive evaluation results reveal that ADCNet performs best on the test set compared to baseline machine learning models across all evaluation metrics. For example, it achieves an average prediction accuracy of 87.12%, a balanced accuracy of 0.8689, and an area under receiver operating characteristic curve of 0.9293 on the test set. In addition, cross-validation, ablation experiments, and external independent testing results further prove the stability, advancement, and robustness of the ADCNet architecture. For the convenience of the community, we develop the first online platform (https://ADCNet.idruglab.cn) for the prediction of ADCs activity based on the optimal ADCNet model, and the source code is publicly available at https://github.com/idrugLab/ADCNet.

Keywords: antibody-drug conjugates; pretrained model; deep learning; molecular representation; online platform

Introduction

With the in-depth study of molecular biology, target-oriented drug discovery has become mainstream, with the aim of discovering more selective and promising clinical candidates [1]. Among them, the traditional small-molecule inhibitor-targeted therapeutic strategy has achieved remarkable results [2] whereas there are still some fatal shortcomings, such as off-target effects [3], limited therapeutic windows [4], and drug resistance concerns [5, 6]. These drawbacks underscore the need for innovative solutions in the field of targeted therapies. Remarkably, the specific expression of antigens has dramatically facilitated the development of monoclonal antibodies [7], partially overcoming the limitations of traditional small-molecule inhibitors [8]. Despite the notable advances made by monoclonal antibodies in cancer treatment, their utility against highly heterogeneous cancer cell populations remains elusive [4].

In response to such a challenge, antibody-drug conjugate (ADC) has emerged as an innovative drug design concept [9], which ingeniously combines the highly targeted nature of monoclonal antibody with the cytotoxicity of a toxin by means of a suitable linker [10]. Such precise anchoring and direct cytotoxicity dramatically increase the therapeutic effect on cancer while reducing the damage to normal cells. Typically, the antibody portion of an ADC provides a unique targeting for cells carrying the desired antigen. When the ADC binds to the surface of the intended cell, the cell initiates endocytosis and transports the ADC to the lysosome. Subsequently, the cytotoxic portion of the ADC is released into the cell to exert its cytotoxic effects [11]. Tracing the evolution of ADCs [12], vindesine-anti-carcinoembryonic antigen (CEA) conjugate was first applied in human clinical trials as early as 1983 [13]. As the field continues to advance, the design of ADCs has been progressively refined over several generations. First-generation ADCs, such as gemtuzumab ozogamicin, typically contain a humanized antibody, a single toxin structure, and an acid-labile linker [14], where the humanized antibody and toxin evolved from murine-derived antibodies and conventional chemotherapeutic agents, respectively [15]. Their conjugation patterns are commonly based on random conjugation of lysine and cysteine, producing mixtures with uneven drug-antibody ratio (DAR), which results in poorly defined therapeutic index criteria [16]. Obviously, first-generation ADCs are prone to unmanageable off-target toxicity and narrow therapeutic windows due to their characteristics. In contrast, second-generation ADCs, represented by brentuximab vedotin and trastuzumab emtansine, have introduced more optimized cytotoxicity payloads like tubulin polymerization inhibitors [17],

as well as steadier linkers [18], guaranteeing both clinical efficacy and stability to a large extent. In this context, the remarkable breakthrough of third-generation ADCs is precisely the application of site-specific coupling technology [19] and the adoption of hydrophilic linkers [20], which not only makes the DAR controllable, but also greatly enhances the safety and anticancer activity [18]. Nevertheless, opportunities and challenges normally accompany each other in this field. Constraints such as the choice of antibodies, toxic payloads, and the design of linkers have contributed substantially to the difficulties in the rational design of ADCs and the continued expansion of therapeutic index [21–23].

In the rational design of ADCs, it is particularly important to explore the relationship between the conformations and properties of ADCs through computational methods such as homology modeling, molecular docking, and molecular dynamics simulations, providing new perspectives for further accelerating the research process of ADCs and understanding the mechanism of action (MOA) [24]. For example, Gupta *et al.* proposed the use of a combination of molecular docking and molecular dynamics simulations to design a non-covalently bound multivalent and affinity-guided antibody empowerment technology (MAGNET) linker platform that largely preserves the ability of ADCs to bind to antigens and overcomes the limitations of existing synthetic strategies [25]. Molecular docking is also used to simulate the binding of a drug-linker to an antibody to determine the optimal site for conjugation, and the antibody could then be engineered to introduce cysteine residues at the immobilization site for more purposeful site-specific coupling [26]. Computational methods have also been used to investigate the MOAs of existing ADCs. Srivastava *et al.* explored the MOAs of nine Food and Drug Administration (FDA)-approved ADCs for cancer therapy by integrating multiple computational approaches; including ADMET risk prediction, potential target identification, molecular docking, molecular dynamics simulations, as well as non-covalent interactions in ADCs [27]. Ultimately, they highlighted the high toxicity of the payloads in these ADCs and identified specific binding sites on the target antigens. In addition, a traditional machine learning (ML) method has been used to mine human gene expression data to optimize the rational selection of breast cancer ADC targets [28]. Although current computational methods provide valuable assistance in designing ADC drugs and/or elucidating their mechanisms, they are often tailored to specific ADCs, restricting their applicability in general. Recently, deep learning (DL) methods have been widely and successfully applied in the fields of prediction and exploration of structure–activity/property relationships [29–33], molecular property prediction [34–41], drug-related interaction prediction [42–45] as well as ADMET prediction [46].

Currently, to the best of our knowledge, the application of DL methods in the field of ADCs has not been reported due to the lack of standardized and unified data in the ADC field. Encouragingly, with the accumulation of ADC research data, Zhu *et al.* recently presented the first comprehensive online database called antibody-drug conjugate database (ADCDB) to store ADCs and associated data [47]. To date, the ADCDB contains 6514 ADCs, 1157 antibodies, 493 linkers, 446 cytotoxic payloads, as well as their corresponding pharmacology assay data. Obviously, the database provides a platform for developing computational tools for the rational design of ADCs.

In the present study, we propose a unified and generalized deep neural network model known as ADCNet for activity prediction of ADCs. As shown in Fig. 1, it is capable of efficiently predicting the anticancer cellular activity of targeted ADCs based on protein sequence representations of antigens and antibodies, SMILES representations of linkers and payloads, together with DAR values that are highly correlated with ADC activities. To comprehensively learn and understand the interrelationships between the different components, ADCNet inputs different parts of a given antigen-ADC complex into the embedding layer separately to obtain distributed representations, which are then embedded into separate neural network modules for feature processing. The feature-spliced vectors are fed to a multilayer perceptron (MLP) with two fully connected layers (FCL) for processing to generate the final prediction. We used ten commonly used evaluation metrics for model evaluation, including average accuracy (ACC), area under the receiver operating characteristics curve (AUC), and balanced accuracy (BA), and ADCNet performed well on the test set. We further validated the generalization ability of the ADCNet model using 21 newly reported ADCs (involving 7 antigens, 7 antibodies, 20 linkers, and 6 payloads). Finally, we provide a free online platform to support the rational design and activity prediction of novel ADCs.

Methods
ADCs dataset preparation and labeling
The dataset used to construct the ADCNet model was uniformly sourced from ADCDB [47]. The original data were processed with the following steps: (i) ADCs with complete structural information on each item were retained, ensuring that the four important structural information of each ADC were available, including the heavy and light chain sequences of the antibody, the antigen sequence, and the isomeric SMILES of the linker and payload, while ADCs with incomplete information on the above were directly discarded; (ii) removing duplicates, cleaning and standardizing the raw data; (iii) uniformly converting units of bioactivity testing data including IC_{50}, EC_{50}, and GI_{50} (e.g. μM, pM, etc.) at cellular level into nM; (iv) if an ADC has multiple biological activity data, removing invalid values such as ">" and "<" and keeping and selecting the minimum value as the final value. Subsequently, we labeled the data as follows (i) if an ADC is in research status of marketed, clinical phases I, II, and III, we all labeled as "positive"; (ii) if an ADC is in the investigational stage, it was labeled as "positive" according to its IC_{50}, EC_{50}, and GI_{50} <= 100 nM, and vice versa (labeled as "negative"). Finally, the obtained ADC dataset was randomly divided into three subsets: 80% for training, 10% for validation, and 10% for testing.

ADCNet architecture and training strategies
Feature representations of antigen and antibody in ADC

In the current study, the antibodies in ADCs, as well as the antigens they act on, were uniformly classified as proteins based on their structural characteristics. The protein language model ESM-2 (freely available at https://github.com/facebookresearch/esm), developed by the Facebook team, is an unsupervised and Transformer-based pre-trained model that enables the prediction of protein structure from protein amino acid sequences and has achieved state-of-the-art (SOTA) performance across multiple downstream task experiments [48]. The ESM-2 model was used to efficiently extract features of antibodies and their corresponding antigens in this study. Specifically, the model processed the input sequences through its multi-layer neural network architecture, that is, the features of amino acids were abstracted and processed layer by layer. Finally, ESM-2 extracted features of the corresponding protein sequences through transfer learning and
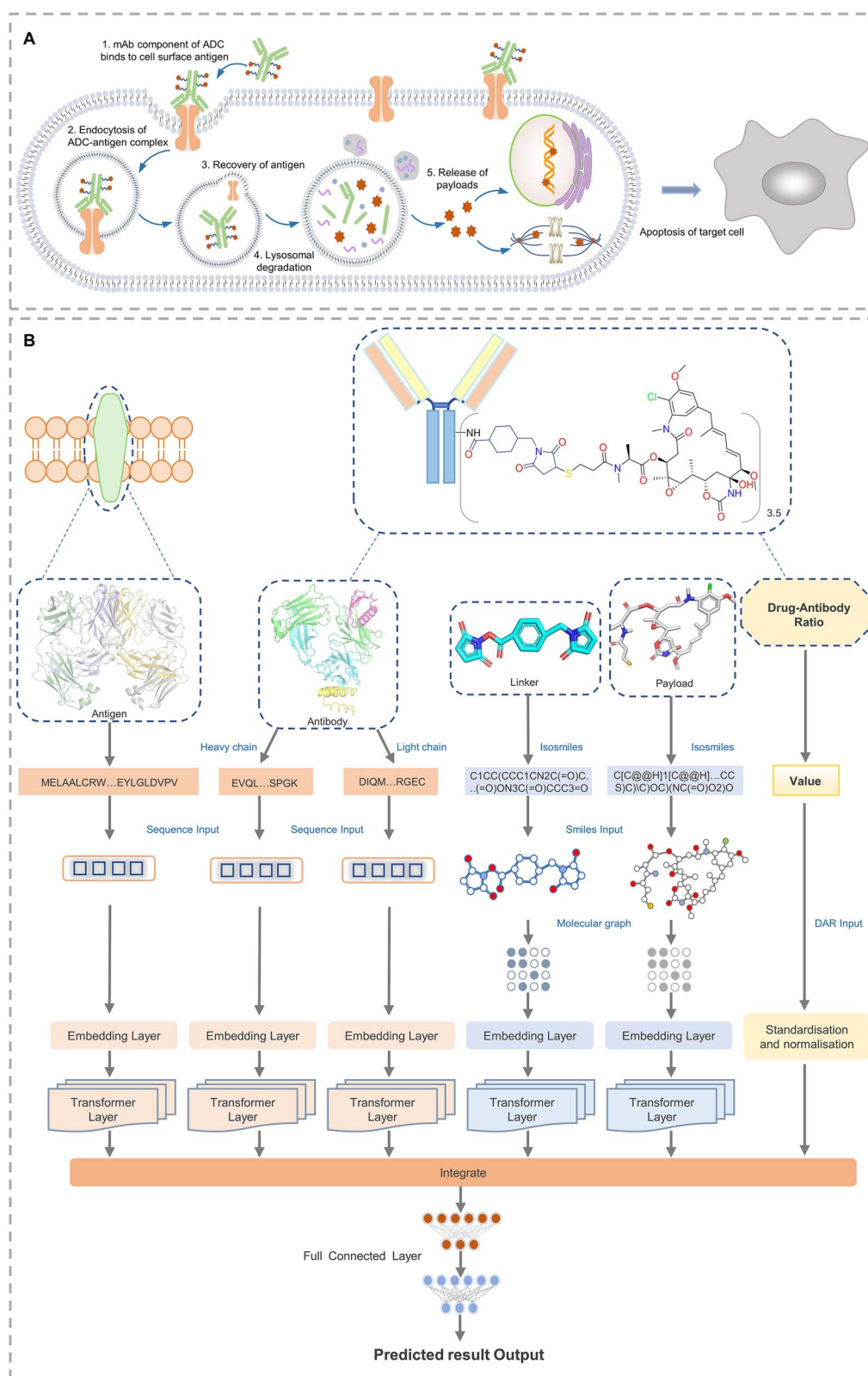
Figure 1. A diagram illustrating the ADCNet framework. (A) The MOA of ADCs. (B) The network architecture of ADCNet model. The protein-related features are extracted by the ESM-2 model, while the small molecule-related features are extracted by the FG-BERT model.

generated high-quality 1280-dimensional feature vectors, which can be effectively applied to the relevant representation of ADCs.

Feature representations of payload and linker in ADC

Recently, we introduced a generalized and self-supervised pre-training DL framework called functional group-based bidirectional encoder representations from transformers (FG-BERT) to learn meaningful representations of molecules from ~1.45 million unlabeled drug-like molecules [36]. The model showcases SOTA predictive capabilities across 44 benchmark datasets covering biological activities, physicochemical properties, physiology, and toxicity attributes. In this study, the payload and linker belonging to the ADC were rationally divided into small-molecule items based on their structures. Accordingly, we adopted the FG-BERT DL model (freely available at https://github.com/idrugLab/FG-BERT) to deeply comprehend and predict the chemical properties and biological activities of payloads and linkers in ADCs. FG-BERT is remarkable for its intensive analysis of intramolecular functional groups, which is critical for the accurate assessment of the efficacy and stability of ADCs. Initially, the input SMILES string of each payload and linker in ADC is converted into a molecular graph to capture molecular structural features as well as information about the connections between atoms. In this molecular graph, atoms are represented as nodes, while chemical bonds are represented as edges. Subsequently, the FG-BERT model was fine-tuned to learn ADC-related representations, and useful molecular representations were finally extracted through transfer learning. After this process, the small molecule structures of linkers and payloads are mapped into 256-dimensional eigenvectors, providing important molecular-level information for ADC studies.

Feature representation of DAR

DAR essentially represents the average number of drug molecules attached to a single antibody. Previous studies have pointed to the profound impact of this metric on multiple aspects of ADC efficacy, safety, and stability [49, 50]. Therefore, we specifically added DAR as a key input variable to the model. To increase the accuracy and effectiveness of data processing, we used the StandardScaler tool to standardize the extracted DAR data columns, ensuring that each variable in the dataset has a uniform zero mean and unit variance, thus eliminating the possible bias of different magnitudes of data. In addition, to further enhance the usability and consistency of the data, the standardized data was normalized by the normalize function in the TensorFlow framework. While this step ensures that the data is normalized to a paradigm of one on the specified axes, the model is able to capture the characteristics of the data more accurately when analyzing and interpreting this data, ensuring the accuracy and reliability of the findings.

ADCNet framework

As shown in Fig. 1B, ADCNet is a multi-channel feature fusion DL model that integrates the representations of antigens, antibodies, linkers, payloads, and DAR values. Remarkably, both ESM-2 and FG-BERT introduced in ADCNet are based on the BERT large language model, which are typically characterized by the inclusion of embeddings, Transformer encoders, and FCL. BERT pre-trains all layers for deep bi-directional representations by simultaneously considering left and right contexts. BERT has strong transfer learning capabilities and can be fine-tuned for a variety of downstream tasks by adding only one output layer [51]. The principles and specific implementations of the FG-BERT [36] and ESM-2 [48]

models have been specifically shown in previous studies. In brief, FG-BERT is pre-trained by masking-recovery of functional groups using a corpus of ~1.45 million unlabeled small molecules, while ESM-2 is pre-trained by masking-recovery of amino acids using a corpus of ~65 million unlabeled unique protein sequences. The two pre-trained models are then utilized to predict downstream tasks through transfer learning. In ADCNet, we use ESM-2 for feature extraction of large molecule antibodies and antigens. With this operation, we get the feature vectors t_1 (heavy chain), t_2 (light chain) for antibodies, and t_3 for antigens. Meanwhile, FG-BERT is used for feature extraction of small molecule linkers and payloads, and we then get the feature vector x_1 for linker and feature vector x_2 for payload. DAR is standardized and normalized to the $[-1, 1]$ interval to get t_4. We then splice these feature vectors using the concat function, and the formula is as follows:

$$x = concat(x_1, x_2, t_1, t_2, t_3, t_4) \quad (1)$$

The final feature fusion is performed by an MLP for prediction of ADC activity, and the formula is defined as follows:

$$\hat{y} = Wx + b \quad (2)$$

Where $\hat{y}$, W, x, b represent the output vector, weight matrix, input vector, and bias vector, respectively. In addition, LeakyReLU is used as an activation function in the ADCNet model. The detailed output dimensions of each layer of ADCNet are provided in Supplementary Table S1.

Model training protocol and hyperparameter optimization

The entire ADCNet model is implemented using Python and TensorFlow [52] with a unique training strategy. During model training, the dataset is divided into training, validation, and test sets, which are split using three different random seeds with a ratio of 8:1:1. Once all data is loaded, the architecture of shared models and the separate output layers is defined. To avoid model overfitting during training, ADCNet was trained using an early stopping strategy, where the number of training epochs was set to 200 and the tolerance value was set to 30. Concretely, after each round of training, the AUC value of the validation set is evaluated and recorded. If the AUC value of the validation set fails to improve within 30 consecutive epochs, the training loop is immediately interrupted, and the test set is entered for final testing. After each training iteration, we check whether the current validation set AUC value exceeds the previous best AUC value. If this value is exceeded, the current model weights are saved as the new best model. If it is not exceeded, the stopping monitor will increase by one. When the value of the tolerance detector reaches 30, the training process ends and the test set is evaluated using the saved best model weights. If the tolerance detector still does not reach 30 throughout 200 epochs, the training also ends, and the performance evaluation of the test set is finally performed. This process aims to ensure that the model does not fit while learning the data in order to obtain the model with the best generalization performance on the test set. During the training process of ADCNet, a Bayesian optimization technique based on Hyperopt and Python packages was introduced to improve and tune the hyperparameters of the model. The optimal hyperparameters were determined by performing 30 searches in the defined parameter space (Supplementary Table S2) to represent the best configuration of the model in terms of final performance results.

To comprehensively and impartially evaluate the predictive performance of ADCNet, we constructed four classical basic ML models for comparison, including logistic regression (LR) [53], random forest (RF) [54, 55], support vector machine (SVM) [56], and extreme gradient boosting (XGBoost) [57]. The characteristics of both linker and payload in the ADC are represented by two molecular fingerprints, namely the 166-bit MACCS keys [58] and the 1024-bit Morgan fingerprint [59]. The RF-, SVM-, XGBoost-, and LR-based models were built on the basis of the fingerprints of linker and payload, protein features of antigen and antibody, and DAR value. The scikit-learn library in Python (freely available at https://github.com/scikit-learn/scikit-learn, version 0.24.1) was used to construct the RF, SVM, and LR models, while the XGBoost Python library (freely available at https://github.com/dmlc/xgboost, version 1.3.3) was utilized to build the XGBoost model. All these traditional ML and as the ADCNet DL models, were trained on Graphics Processing Unit (GPU) [NVIDIA Corporation GV100GL (Tesla V100 PCIe 32 GB)] and Central Processing Unit (CPU) (Intel(R) Xeon(R) Silver 4216 CPU @ 2.10 GHz).

Evaluation of model performance

The performance of ADCNet and traditional baseline ML models is scored by a variety of metrics, including positive predictive value (PPV), negative predictive value (NPV), specificity (SP/TNR), sensitivity (SE/TPR/Recall), ACC, F1 score, BA, Matthew's correlation coefficient (MCC), area under the precision-recall curve (PR-AUC), and AUC. These metrics are defined as follows:

$$PPV = \frac{TP}{TP + FP} \tag{3}$$

$$NPV = \frac{TN}{TN + FN} \tag{4}$$

$$SE = \frac{TP}{TP + FN} \tag{5}$$

$$SP = \frac{TN}{TP + FP} \tag{6}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$F1 = \frac{2 \times TP}{2 \times TP + FN + FP} \tag{8}$$

$$BA = \frac{TPR + TNR}{2} = \frac{SE + SP}{2} \tag{9}$$

$$MCC = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FN) \times (TN + FP)}} \tag{10}$$

Where TP, TN, FP, FN, TNR, and TPR represent true positive, true negative, false positive, false negative, true negative rate, and true positive rate, respectively. Furthermore, AUC is the area under the ROC curve plotted by calculating TPR versus FPR at different discrimination thresholds, which is used to measure the overall performance of the model.

Web server implementation

We have developed an online platform that facilitates users to predict ADC activity for free. The platform consists of three main parts: front-end, back-end, and cloud server. User input is received using Streamlit framework and passed to the back-end TensorFlow model for processing, and the cloud server runs on CentOS and uses Nginx to handle user requests. The platform is hosted on Tencent Cloud to ensure its high speed and stable performance.

Results and discussion
Analysis and visualization of ADC modeling dataset

As described in the Methods section, we have defined for the first time a standard for collecting and processing ADC modeling datasets based on the fact that the efficacy of a given ADC is maintained by the complexities of how antibody, linker, and payload components interact with tumors and their microenvironment [10, 60], all of which have significant clinical implications. As shown in Fig. 2A, a total of 435 ADCs were obtained based on the mentioned standard. Figure 2B shows the research phase and corresponding proportions of these ADCs.

When constructing predictive models, the structural diversity and wide chemical space of the molecules in the dataset are crucial for improving the accuracy and robustness of the models [61]. Hence, based on the characteristics of the ADCNet architecture, we performed a statistical analysis of the antibodies, linkers, payloads, DAR, and antigens involved in the ADCs. As shown in Fig. 2C, among these 435 ADCs, the number of unique antibodies, linkers, payloads, DAR values, and antigens are 154, 82, 71, 75, and 66, respectively, indicating the overall diversity of ADC modeling dataset used in this study. In addition, MEGA11, a key tool in molecular evolutionary genetics, provides robust capabilities for sequence alignment and phylogenetic analysis, which was used to explore the genetic diversity and evolutionary patterns of antigens in the ADC modeling dataset. Figure 2D demonstrates that 66 antigens can be divided into six different categories based on sequence alignment results, intuitively illustrating the structural and functional diversity of these antigens in the construction of the ADCNet model. The top five antigens are ERBB2 (106), KIT (48), FGFR2 (36), TNFRSF1A (33), and EGFR (22). Meanwhile, humanized IgG1-kappa is most widely used to design ADCs (Fig. 2E), followed by Chimeric IgG1-kappa, humanized IgG1, and humanized IgG1-nd. Supplementary Fig. S1 displays the distribution of isoelectric point (pI) values and molecular weight (MW) of antigens and antibodies, further indicating the diversity of antibodies and antigens.

We further performed structural diversity and chemical space analysis on 82 linkers and 71 payloads. Bemis-Murcko scaffold analysis [62] can efficiently extract the basic scaffold of a compound, ignoring non-cyclic parts and functional groups, and is therefore widely used to judge the scaffold diversity and novelty of different compounds. As shown in Fig. 2F, we found that the scaffolds for linkers and payloads in the ADC dataset are between 58.54% and 67.61% through Bemis-Murcko scaffold analysis, demonstrating that the linkers and payloads are structurally more diverse. In addition, the chemical space of the linkers and payloads in the ADC dataset can be characterized by using MW and AlogP. Figure 2G and 2H show that both linkers and payloads are distributed over a wide range of MW (linkers: 120.195–1702.11, payloads: 233.097–1580.593) and AlogP (linkers: −6.504–6.908, payloads: −4.675–7.406), indicating that the linkers and payloads in the ADC dataset have a broad chemical space. Collectively, the above analysis results, together with 75 different DAR values (Fig. 2C), demonstrated that the five input components in the ADCNet framework are also diverse.

Among these ADCs, 281 ADCs were labeled as positives and 154 ADCs were labeled as negatives. Finally, the whole ADC modeling dataset was randomly split into the training (80%), validation (10%), and test (10%) sets. The training set is used to build a predictive model, the validation set is used to optimize and select hyperparameters, and the test set is utilized to test the final predicted performance of the model.
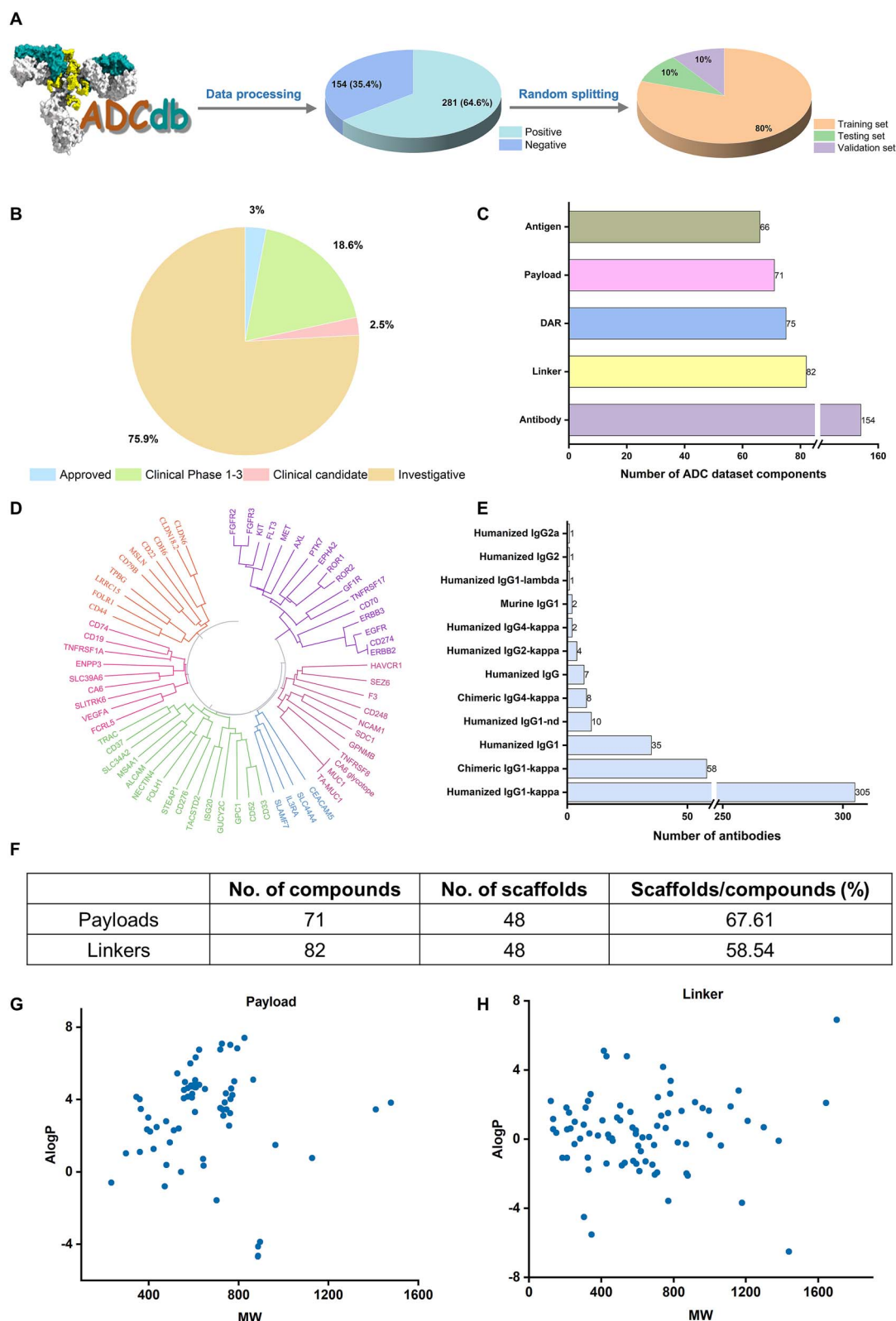
Figure 2. The overall analysis results of the ADC modeling dataset. (A), data acquisition and processing procedure. (B), research distribution of ADCs in the modeling dataset. (C), the number of components involved in antigens, antibodies, linkers, payloads, and DAR values in the ADC dataset. (D), developmental tree analysis of antigens in the ADC dataset. (E), the number of each antibody class in the ADC dataset. (F), scaffold diversity analysis of payloads and linkers in the ADC dataset. Chemical space analysis of payload (G) and linker (H) in the ADC dataset. The developmental tree was generated by MEGA11 software (https://www.megasoftware.net/). The scaffold of each payload or linker was calculated by RDKit software (https://www.rdkit.org/). The chemical space was defined using MW (X-axis) and AlogP (Y-axis). MW and AlogP were computed by RDKit software (https://www.rdkit.org/).

Performance of the ADCNet model

We utilized the carefully designed ADC dataset to evaluate the predictive ability of the ADCNet model. Detailed performance results of the ADCNet and eight classic ML models are provided in Table 1. Notably, all models are constructed based on the same dataset, data split method, and initial seeds for fair comparison. As shown in Table 1, the ADCNet achieves the best performance on the test set due to its highest values on all commonly used evaluation metrics. For one thing, the ADCNet shows the best overall accuracy compared to the baseline models, with the highest ACC, AUC, and PR-AUC values of 0.8712, 0.9293, and 0.9518, respectively. In addition, the ADCNet achieves the maximum values of SE (0.8713), SP (0.8665), PPV (0.9241), and NPV (0.7778), indicating that the prediction accuracy of the model for both ADC positives/actives and negatives/inactives is not only optimal, but also balanced, compared with the traditional ML models.

For another, the results of data collection and analysis have shown that the current ADC dataset is unbalanced, due to its inclusion of 281 actives and 154 inactives. Therefore, MCC, F1, and BA indicators were employed to further evaluate the ADCNet model. Compared to eight traditional ML-based baseline models, the ADCNet model performed the best, with the highest values of MCC (0.7196), F1 (0.8968), and BA (0.8689).

In fact, there are two main differences between the ADCNet model and the ML-based baseline models. First, the baseline models use the Morgan fingerprint or MACCS fingerprint to encode the linkers and payloads in the ADCs, while the ADCNet model uses the FG-BERT large language model to learn the representation of the linkers and payloads. Previous study has shown the superiority of FG-BERT in molecular learning representation [36], which may be one of the reasons for the explosive performance of the ADCNet model. Second, in view of the complex input features from the five components (antigen, antibody, linker, payload, and DAR), the traditional ML methods can only accept all the input features and then directly build the predictor and output the result, which obviously lacks the process of meaningful feature processing. In contrast, ADCNet as a DL model can learn meaningful features from the complex input features of these five components through self-learning and feedback mechanisms, which ultimately leads to the best performance in the ADCs activity prediction task. Incidentally, Supplementary Fig. S2 shows the iterative change in loss during the training of the ADCNet model and the changes in AUC curves for the training and validation sets.

Multi-level model validation
Cross-validation

It is common practice to employ a cross-validation approach in the model training and evaluation process to ensure the robustness and general applicability of the model. Accordingly, we adopted a 5-fold cross-validation approach to further evaluate the predictive performance of the ADCNet model in this study, which is particularly well-suited for smaller datasets because it can make more efficient use of limited data resources. As shown in Fig. 3A, the average BA, AUC, and ACC of 5-fold cross-validation are 0.8477 ± 0.0078, 0.9281 ± 0.0076, and 0.8521 ± 0.015, respectively, which were consistent with the results obtained by the conventionally trained ADCNet model. In addition, other evaluation indicators also show a similar trend. All of these results demonstrate the robustness of the ADCNet model.

Table 1. Performance evaluation results of ADCNet and traditional ML-based models on the test set

Model	SE	SP	MCC	ACC	AUC	F1	BA	PRAUC	PPV	NPV
LR-MACCS	0.8296 ± 0.0641	0.5318 ± 0.061	0.3791 ± 0.0303	0.7348 ± 0.0347	0.6807 ± 0.006	0.8068 ± 0.0387	0.6807 ± 0.006	0.8647 ± 0.0229	0.7862 ± 0.0239	0.6127 ± 0.0489
LR-Morgan	0.8527 ± 0.0442	0.6486 ± 0.0364	0.5105 ± 0.0352	0.7879 ± 0.0262	0.7506 ± 0.0111	0.8433 ± 0.0281	0.7506 ± 0.0111	0.8928 ± 0.016	0.8344 ± 0.0136	0.6859 ± 0.0422
RF-MACCS	0.8073 ± 0.0497	0.7028 ± 0.0728	0.4979 ± 0.1128	0.7727 ± 0.0601	0.7551 ± 0.0604	0.8259 ± 0.0548	0.7551 ± 0.0604	0.8908 ± 0.0422	0.8455 ± 0.0611	0.6405 ± 0.0453
RF-Morgan	0.8196 ± 0.042	0.7237 ± 0.0417	0.5315 ± 0.0581	0.7879 ± 0.0347	0.7716 ± 0.0304	0.838 ± 0.0343	0.7716 ± 0.0304	0.8995 ± 0.027	0.8581 ± 0.0397	0.6621 ± 0.0465
SVM-MACCS	0.7863 ± 0.0199	0.6218 ± 0.1251	0.4013 ± 0.1291	0.7348 ± 0.0473	0.7040 ± 0.0675	0.8004 ± 0.0293	0.7040 ± 0.0675	0.8728 ± 0.0231	0.8154 ± 0.0427	0.5794 ± 0.0836
SVM-Morgan	0.7859 ± 0.0527	0.6742 ± 0.0273	0.4492 ± 0.0817	0.7500 ± 0.0455	0.7301 ± 0.0391	0.8081 ± 0.0391	0.7301 ± 0.0391	0.8810 ± 0.0248	0.8321 ± 0.0299	0.6064 ± 0.0667
XGB-MACCS	0.8649 ± 0.0071	0.7237 ± 0.0417	0.5869 ± 0.0376	0.8182 ± 0.0227	0.7943 ± 0.0241	0.8647 ± 0.0216	0.7943 ± 0.0241	0.9103 ± 0.0216	0.8648 ± 0.0362	0.7206 ± 0.011
XGB-Morgan	0.8526 ± 0.0748	0.7445 ± 0.0279	0.5955 ± 0.0771	0.8182 ± 0.0455	0.7985 ± 0.0284	0.8616 ± 0.0444	0.7985 ± 0.0284	0.9118 ± 0.024	0.8726 ± 0.0253	0.7222 ± 0.0962
ADCNet	**0.8713 ± 0.0660**	**0.8665 ± 0.0387**	**0.7196 ± 0.0387**	**0.8712 ± 0.0572**	**0.9293 ± 0.0464**	**0.8968 ± 0.0569**	**0.8689 ± 0.0522**	**0.9518 ± 0.0366**	**0.9241 ± 0.0470**	**0.7778 ± 0.0481**

The best performing model for each metric is highlighted in bold.
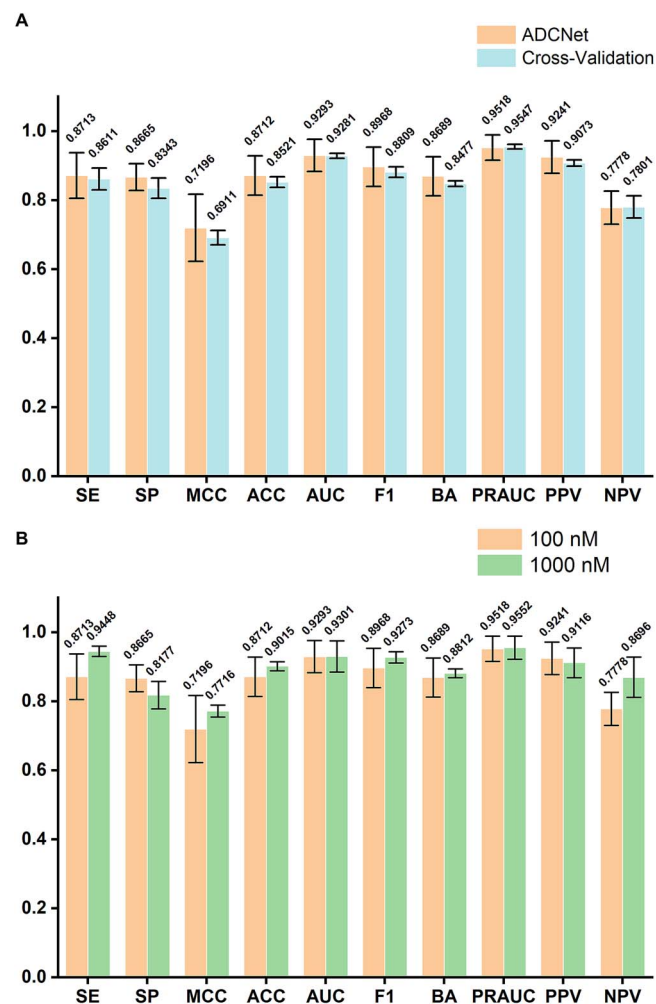
A

B

Figure 3. Results of ADCNet model validation comparison results of the original ADCNet model with the 5- fold cross-validation model (A) and the model based on a threshold of 1000 nM (B). The average metrics of three independent experiments on the test set were used to evaluate all models.

Activity threshold-based model performance analysis

In the present study, the ADCNet model was originally trained using a default active/inactive cutoff value of 100 nM for IC_{50}, EC_{50}, and GI_{50}. In order to further reduce the influence of artificial selection on the experimental results and to enhance the accuracy and reliability of the study, we introduced another active/inactive cut-off criterion (1000 nM). In other words, an ADC is considered active only if its IC_{50}, EC_{50}, or GI_{50} satisfies the cut-off condition, or if it is in clinical studies and approved. We then retrained the ADCNet model based on this new truncation criterion. As shown in Fig. 3B, the retrained model based on the new criterion achieves the average BA, AUC, PRAUC, and ACC values of 0.8812, 0.9301, 0.9552, and 0.9015, respectively, which were extremely close to the values obtained by using the original criterion. Other evaluation metrics also exhibit a similar trend. These findings confirm that our model exhibits good stability in response to different truncation criteria. Notably, for the sake of simplicity and the fact that ADC acts as a magic bullet due to its highly active nature, only the initial set of labels was selected to train the final ADCNet model.

Model architecture validation based on ablation experiments

As shown in Fig. 1B, ADCNet is a multipath framework that combines information about the antibodies, linkers, payloads, DAR values, and antigens involved in the ADCs. To further investigate the effectiveness and necessity of these five well-designed modules in the ADCNet architecture, we performed a series of ablation studies to decouple ADCNet networks. Concretely, we designed five variants of ADCNet as follows:

(i) ADCNet *without antigen information* (w/o antigen): removing the antigen module in Fig. 1B.
(ii) ADCNet *without antibody information* (w/o antibody): removing the antibody module in Fig. 1B.
(iii) ADCNet *without linker information* (w/o linker): removing the linker module in Fig. 1B.
(iv) ADCNet *without payload information* (w/o payload): removing the payload module in Fig. 1B.
(v) ADCNet *without DAR information* (w/o DAR): removing the DAR module in Fig. 1B.

All five variants of ADCNet were evaluated on the ADCNet dataset, and the experimental setup is consistent with the ADCNet for a fair comparison. Two key indicators, MCC and BA, were utilized to adjudicate the comparison of the final ablation experimental results due to the imbalance in the ADC dataset. As shown in Fig. 4A and 4B, the ADCNet model performed best according to the highest average MCC (0.7196) and BA (0.8689) values compared to the other five variants, demonstrating that the five well-designed modules are necessary in the ADCNet architecture and can complement each other to improve the final performance of the ADCNet model. For example, antigens are not intrinsic components of ADC molecules in principle, but the lack of antigen information module will cause the MCC and BA values of the variant model (w/o antigen) to decrease to 0.5155 and 0.738, with an overall decrease of 28.36% and 15.07%, respectively, compared to the complete ADCNet model. One possible explanation is that antigens play an important role in recognition of ADCs and guiding them into tumor cells, thereby supporting their anti-tumor activity. A similar trend was also observed on other variant models (Fig. 4A and 4B), verifying that the entire ADCNet architecture strongly dominates the simpler variants.

In addition, we used ACC and AUC metrics to judge the overall superiority of the ADCNet framework in ablation experiments. As shown in Fig. 4C, even if we do not consider the imbalance of ADC data, the ADCNet model also achieves the best ACC (0.8712) compared to the five variant models. Meanwhile, its AUC (0.9293) is also very close to the best AUC (0.9304) from the variant model (w/o DAR), with little difference (Fig. 4D). Collectively, these results further confirm the overall superiority and rationality of the ADCNet architecture proposed in this study.

The ADCNet framework uses two pre-trained models to generate embeddings as input vectors for the prediction subnetworks. To evaluate its feature learning capability, we employed t-distributed stochastic neighbor embedding (t-SNE), an excellent dimensionality reduction method, to visualize a quintuple consisting of all embeddings of antibody, antigen, linker, payload, and DAR value. We applied the entire collected data as input for dimensionality reduction, and used the ADCNet model to reduce the embedding space of the ADCs in the before and after training scenarios. As shown in Fig. 5B, the distribution of positive (green) and negative (orange) quintuples in the dataset shows that it is easy to separate positive and negative ADC samples by
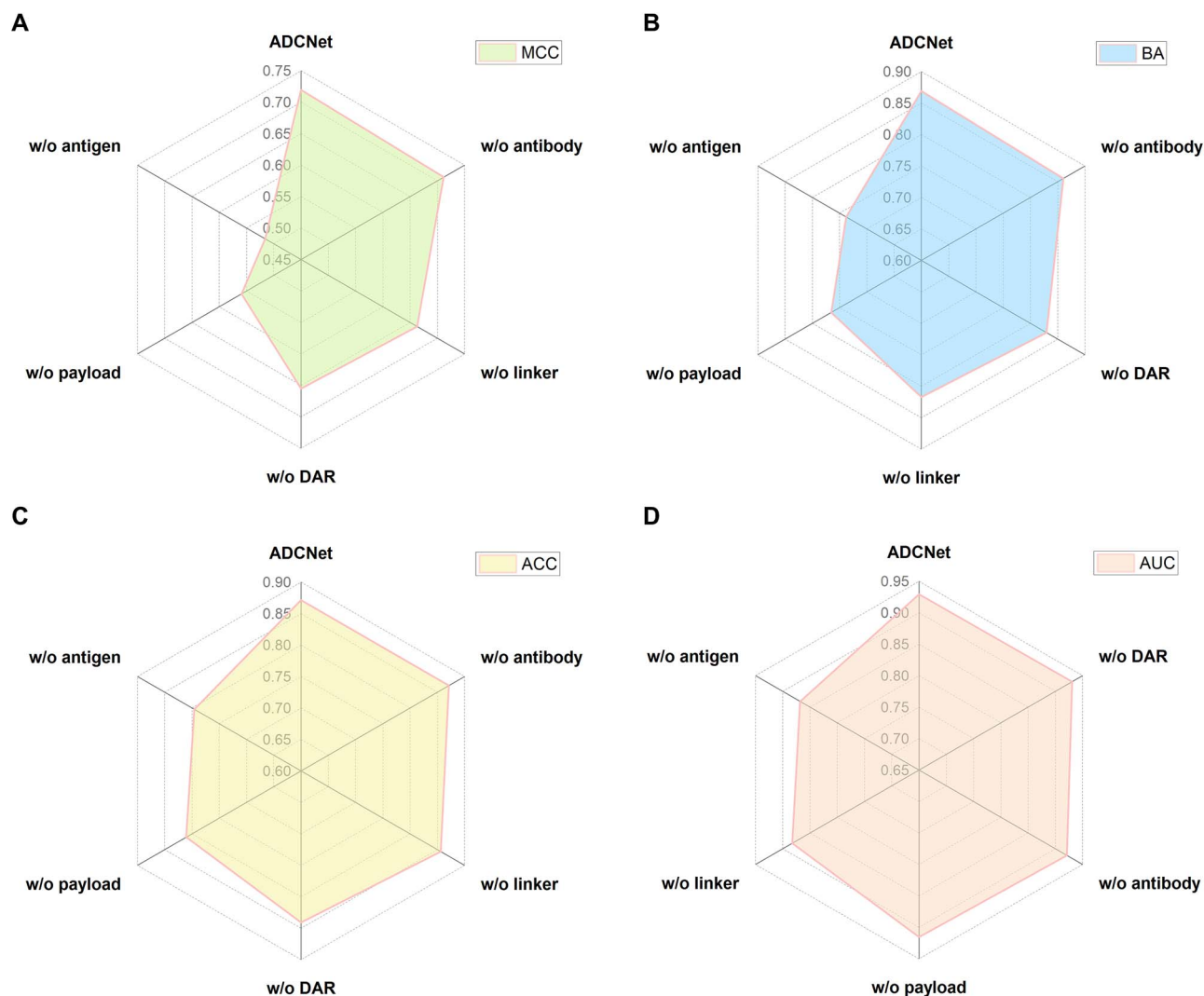
Figure 4. Results of ablation experiments. (A), (B), (C), and (D) represent the average MCC, BA, ACC, and AUC values of the test set, respectively.

the ADCNet model, compared to the corresponding quintuple distribution without model training. Taking ADC1 (negative) and ADC2 (positive) as an example, although they have different activities and are structurally different (Fig. 5A), they overlapped in an untrained space. However, it was easy to separate and successfully classify in the space after the ADCNet training. Meanwhile, a pair of ADC molecules (ADC3 and ADC4, Fig. 5C) with opposite activities were randomly distributed before training, but both were successfully categorized after training. The more clearly the quintet of positive and negative roles can be separated, the better the categorization will be. Collectively, the embeddings of both categories are more prominent in terms of data distribution after training compared to before, which further verifies the excellent performance of our model. To verify whether the results of the ADCNet model were due to coincidental associations, we conducted a Y-scrambling test by shuffling the labels in both the test set and the training set. Supplementary Fig. S3 shows that the ADCNet model outperformed the Y-scrambled model in terms of ACC, AUC, and PR-AUC, further confirming the robustness and reliability of the model's results.

Independent testing based on new external dataset
The above results have confirmed the robustness and stability of the ADCNet model. Here, we continue to test the generalization

ability of the ADCNet model because it is of great value for the application of the model. To this end, we collected an additional dataset of 21 ADCs that satisfy the defined criteria from the latest publications [25], [63–66] and a patent (WO2023237050). Details of these 21 ADCs are provided in Supplementary Table S3.

As shown in Table 2, there are 7 antigens, 7 antibodies, 20 linkers, 6 payloads (Supplementary Fig. S4), and 18 DAR values in these ADCs. Subsequently, sequence alignment-based similarity score was used to assess the similarity of the antibody and antigen in each newly collected ADC to the corresponding items in the original ADC dataset. Meanwhile, ECFP-4 (Extended Connectivity Fingerprint 4) fingerprint-based molecular similarity calculation was conducted to evaluate the novelty of linker and payload in these 21 ADCs. Detailed results are listed in Supplementary Table S3. Table S3 shows that these new ADCs are derived from the redesign of specific items, including the substitution, modification, and/or recombination of antibodies, antigens, and payloads. However, our analysis also shows that the ADCs in the external dataset exhibit significant novelty in the overall structure, especially in terms of the structure of the linkers. In addition, to further assess the structural novelty and diversity of these 21 ADCs, the reconciled mean similarity as an evaluation metric was adopted. This similarity metric is derived from an overall judgment of the similarity of antigens, antibodies,
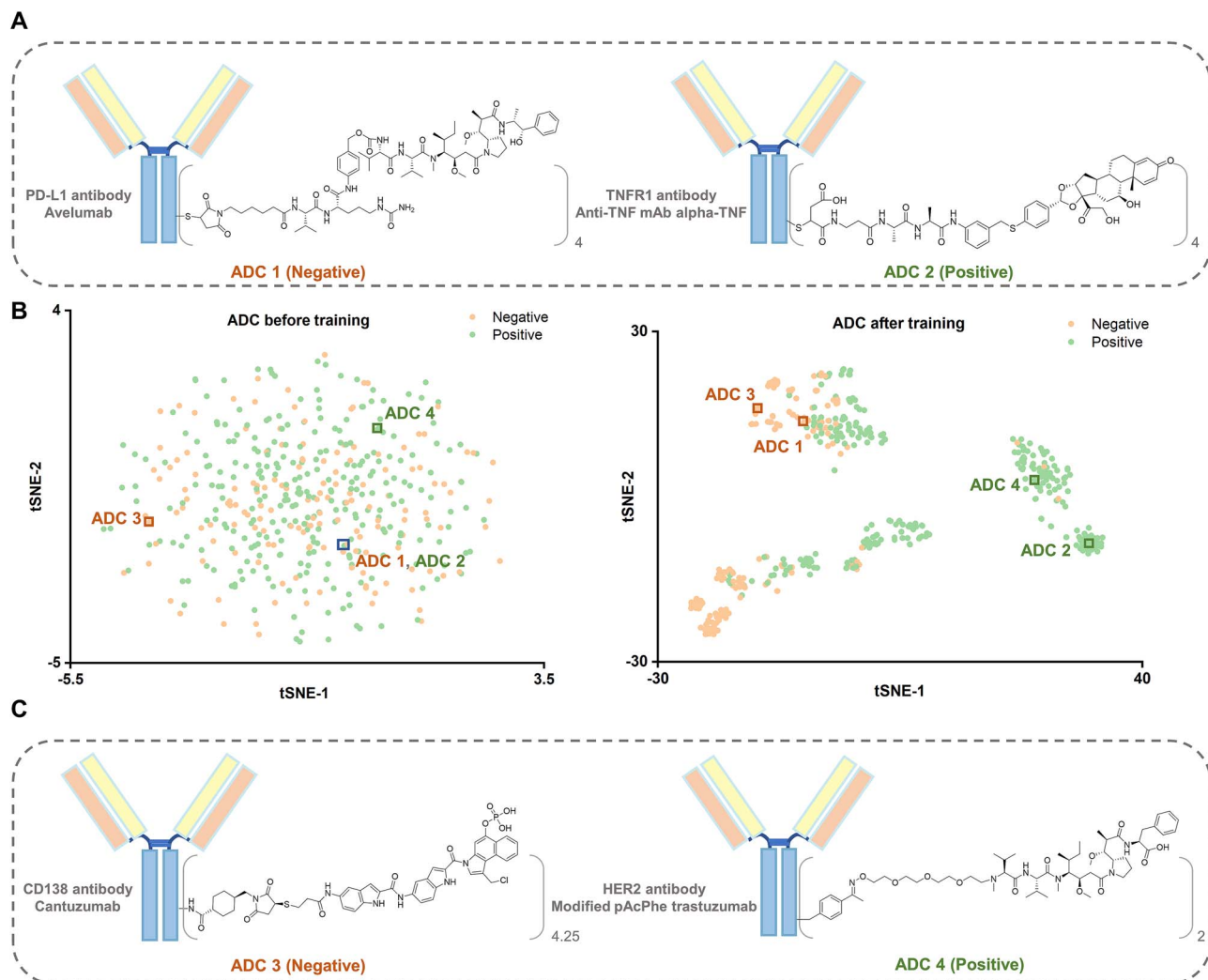
Figure 5. Visualization of the ADC dataset using t-SNE method. (A) Represents ADC1 (PD-L1-ADC-2, negative) and ADC2 (anti-TNF ADC 115, positive) with overlap in an untrained space. (B), spatial distribution of the ADC dataset before and after training. (C) Represent ADC3 (HuC242-SMCC-DC4, negative) and ADC4 (ARX-788, positive) with randomly distribution in an untrained space.

linkers, and payloads, compared to ADCs in the original dataset. As shown in Table 2, the reconciled mean similarity scores range from 0.4047 to 0.9662, which further reflects the moderate to high dissimilarity between these newly collected ADCs and ADCs in the modeling dataset. Therefore, this external dataset is particularly suitable for testing the generalization ability of the ADCNet model.

Detailed predictive results of the external data set are summarized in Table 2. It is clear that the ADCNet model can accurately forecast the anticancer capacities of 19 out of 21 ADCs, resulting in a prediction accuracy of 90.48%. Accordingly, the independent external testing results not only demonstrate the strong generalization ability of ADCNet in predicting the activity of different types of ADCs, but also indicate that our model can be used in a real-world scenario for ADC drug discovery.
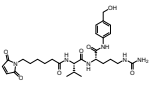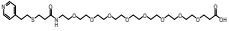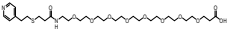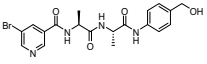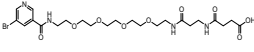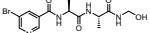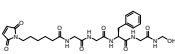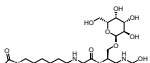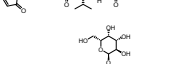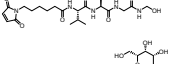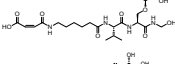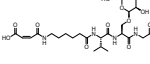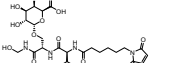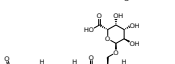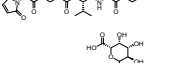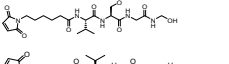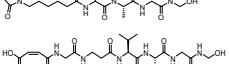
Web server construction and usage

To facilitate design and discovery of novel ADC drugs, we developed an online platform called DeepADC based on the optimal ADCNet model for experts and non-experts in the field. DeepADC is freely available at https://ADCNet.idruglab.cn. Users can achieve ADC activity prediction by entering the sequences of antigen and antibody, SMILES of linker and payload, and DAR

value for a given ADC (Fig. 6). It is worth noting that due to the complexity of ADC structures and mechanisms of action, we recommend that users input antigens and antibodies that are naturally occ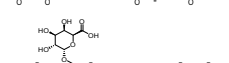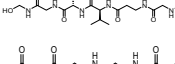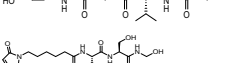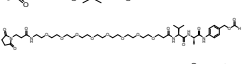urring and already paired; otherwise, it will be beyond the model's predictive scope. First, to enhance convenience, users can directly select antigen and antibody names from our database, and the corresponding sequences will be automatically filled in (Fig. 6A). If users choose to manually input sequences, they should rely on practical experience to verify the authenticity of the information provided. Second, users type or paste the smiles of linker and payload of the ADC in the "Linker Iso-SMILES" and "Payload Iso-SMILES" input boxes, respectively, or use the online Molecular Sketcher to draw their molecular structures individually (Fig. 6B). Finally, users enter a DAR value (Fig. 6C) and submit the calculation task. It is worth noting that such DAR value can be determined according to the actual situation, or can refer to the DAR range of the existing antibodies in "Help" function button to determine. Alternatively, users can upload a standard CSV file to perform the activity prediction of multiple ADCs (Fig. 6D). After completing the calculation, users can view the forecast results online (Fig. 6E) or download the forecast results in CSV format.

For example, sacituzumab-DM1, a novel ADC constructed by bottom-up aqueous nickel-catalyzed cross-coupling [64] was

Table 2. Predictive results of the ADCNet model on an independent external dataset

No.	Antibody[a]	Antigen[b]	Payload[c]	Linker	DAR	Similarity score[d]	Bioactivity value	Label	Predictive score
1	H1D8	CD44v5	MMAE		3.7	0.7867	$IC_{50} = 8.24\ \mu g/mL$	1	0.9720
2	Cetuximab	EGFR	Gemcitabine		6	0.4047	$IC_{50} = 10$ nM	1	0.9956
3	Trastuzumab	HER2	Gemcitabine		6	0.4051	$IC_{50} = 150$ nM	0	0.9907
4	Sacituzumab	TROP2	Auristatin E		6.1	0.7240	$IC_{50} = 0.03\text{–}0.09$ nM	1	0.9839
5	Sacituzumab	TROP2	DM1		6.2	0.6541	$IC_{50} = 0.05\text{–}0.1$ nM	1	0.9991
6	Sacituzumab	TROP2	DXd		7.3	0.5757	$IC_{50} = 0.4\text{–}0.7$ nM	1	0.9817
7	Ifinatamab	CD276	DXd		8.12	0.9662	$EC_{50} = 0.022\text{–}3.764$ nM	1	0.9889
8	Ifinatamab	CD276	DXd		8.76	0.8152	$EC_{50} = 0.008\text{–}3.291$ nM	1	0.9326
9	Ifinatamab	CD276	DXd		7.76	0.8019	$EC_{50} = 0.018\text{–}2.291$ nM	1	0.9648
10	Ifinatamab	CD276	DXd		7.98	0.7733	$EC_{50} = 0.038\text{–}1.245$ nM	1	0.9946
11	Ifinatamab	CD276	DXd		8.12	0.7616	$EC_{50} = 0.013\text{–}0.53$ nM	1	0.9950
12	Ifinatamab	CD276	DXd		8.06	0.8107	$EC_{50} = 0.007\text{–}0.894$ nM	1	0.9847
13	Ifinatamab	CD276	DXd		8.02	0.7271	$EC_{50} = 0.015\text{–}1.042$ nM	1	0.9946
14	Ifinatamab	CD276	DXd		8.18	0.7976	$EC_{50} = 0.024\text{–}1.647$ nM	1	0.9910
15	Ifinatamab	CD276	DXd		8.52	0.9125	$EC_{50} = 0.025\text{–}1.073$ nM	1	0.9855
16	Ifinatamab	CD276	DXd		8.08	0.8544	$EC_{50} = 0.025\text{–}1.234$ nM	1	0.9958
17	Ifinatamab	CD276	DXd		7.76	0.7273	$EC_{50} = 0.029\text{–}1.094$ nM	1	0.9958
18	Ifinatamab	CD276	DXd		7.78	0.8074	$EC_{50} = 0.01\text{–}0.861$ nM	1	0.9958
19	Ifinatamab	CD276	DXd		8.56	0.8914	$EC_{50} = 0.033\text{–}0.511$ nM	1	0.9616
20	hLL2-C220	CD22	SG3249		1.52	0.8196	$IC_{50} = 0.02\text{–}51.7$ nM	1	0.9998
21	FZ-A038	DLL3	DXd		8	0.6023	$IC_{50} = 0.15\text{–}4.29$ nM	1	0.0196

[a]The sequences of antibody are in Supplementary Table S3. [b]The sequences of antigen are in Supplementary Table S3. [c]The detailed structures of payloads are in Supplementary Fig. S4. [d]Similarity score refers to the harmonic mean similarity score, which is used to calculate the average level of similarity between points in a set of data points by taking the reciprocal of the arithmetic mean of the inverse of the similarity score.

considered as an active ADC due to its high prediction score of 0.9991. Previous studies have demonstrated that sacituzumab-DM1 has strong inhibitory activity against JIMT-1, HCC78, BxPC-3, and MDA-MB-468 cancer cell lines with an IC_{50} value of 0.05–0.1 nM, which further confirms the accuracy and availability of DeepADC.

Conclusion

In this study, we constructed a novel DL model called ADCNet, which was able to predict ADC activity. ADCNet creatively incorporates both ESM-2 and FG-BERT Large Language Model (LLM) to learn the characteristics of the five components involved in ADC efficacy. Based on our first carefully-designed ADC benchmark

Figure 6. The operation of DeepADC. (A), (B), and (C) represent the input cases of antibody, antigen, linker, payload, and DAR value, respectively. (D) Represents the input case for multiple ADCs. (E) Represents the online display or download of the predictive results.

data set, ADCNet achieved the optimal performance on the test set across all evaluation metrics. The 5-fold cross-validation and threshold reclassification experiments successfully confirmed the stability of the model and reduced the influence of human factors, respectively. It is worth mentioning that the ADCNet model achieved the best performance in a series of well-designed ablation experiments, which fully justified the rationality and necessity of the model architecture. Subsequently, external independent testing results also strongly demonstrated the potential of the ADCNet model to handle unknown samples. Finally, to better support the research community in the field, an easy-to-use online platform called DeepADC based on the optimal ADCNet model was constructed to support the discovery and design of novel ADC drugs. Biologically, the integration of ESM-2 and FG-BERT enables ADCNet to capture complex molecular interactions that drive ADC efficacy. By learning from a vast array of biological data, the model provides deeper insights into how changes in antibody, linker, and payload structures can influence therapeutic outcomes.

It is worth noting that we have developed the first DL-based prediction model for ADC drugs, and there is still room for improvement in the current work. For example, as the number of other ADC drugs gradually increases, we will build more accurate predictive models and update them on the DeepADC online platform. We will be able to integrate a variety of biomarkers and drug properties to develop new ADC predictive models in the future to improve the accuracy and understanding of models for complex biological systems. In addition, we will also use generative artificial intelligence to generate and optimize ADC linkers to discover better ADC drugs. In summary, this study opens the door to the rational design of ADC drugs.

Key Points

- We propose a unified and generalized deep neural network model known as ADCNet for activity prediction of antibody-drug conjugates (ADCs).
- ADCNet creatively incorporates both ESM-2 and FG-BERT LLM to learn the characteristics of the five components involved in ADCs efficacy.
- Based on our first carefully-designed ADC benchmark data set, ADCNet achieved the optimal performance on the test set across all evaluation metrics.
- The superior performance of ADCNet was consistently demonstrated through 5-fold cross-validation, threshold reclassification experiments, well-designed ablation studies, and external independent testing.
- For the convenience of the community, we develop the first online platform for the prediction of ADCs activity based on the optimal ADCNet model.

Acknowledgments

Author contributions

L.C. collected, annotated, and processed the ADCs data, helped train the ADCNet model, and wrote the manuscript; B.L. developed and trained the network model; Y.C. constructed the web server; M.L., S.Z., C.L., and Y.P. helped for data collection, annotation and processing; L.W. designed the entire project, revised the manuscript and solved the problems encountered.

Supplementary data

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

Conflict of interest: The authors declare no competing interests.

Funding

Data availability

The ADCs data used in the present study are freely available in ADCdb (http://adcdb.idrblab.net/). The remaining data or questions regarding this study are available to the corresponding author upon request (Ling Wang: lingwang@scut.edu.cn). Source data are provided with this paper.

Code availability

The source code of ADCNet and associated data preparation scripts are available at GitHub (https://github.com/idruglab/ADCNet). The optimized ADCNet model is also provided.

Correspondence and requests for materials should be addressed to Ling Wang.

References

1. Belfield GP, Delaney SJ. The impact of molecular biology on drug discovery. *Biochem Soc Trans* 2006;**34**:313–6. https://doi.org/10.1042/BST0340313
2. Zhong L, Li Y, Xiong L. *et al.* Small molecules in targeted cancer therapy: advances, challenges, and future perspectives. *Signal Transduct Target Ther* 2021;**6**:201. https://doi.org/10.1038/s41392-021-00572-w
3. Lin A, Giuliano CJ, Palladino A. *et al.* Off-target toxicity is a common mechanism of action of cancer drugs undergoing clinical trials. *Sci Transl Med* 2019;**11**:eaaw8412. https://doi.org/10.1126/scitranslmed.aaw8412
4. Chari RV. Targeted cancer therapy: conferring specificity to cytotoxic drugs. *Acc Chem Res* 2008;**41**:98–107. https://doi.org/10.1021/ar700108g
5. Schram AM, Chang MT, Jonsson P. *et al.* Fusions in solid tumours: diagnostic strategies, targeted therapy, and acquired resistance. *Nat Rev Clin Oncol* 2017;**14**:735–48. https://doi.org/10.1038/nrclinonc.2017.127
6. Catalano A, Iacopetta D, Ceramella J. *et al.* Multidrug resistance (MDR): a widespread phenomenon in pharmacological therapies. *Molecules* 2022;**27**:616. https://doi.org/10.3390/molecules27030616
7. Walts AE, Said JW. Specific tumor markers in diagnostic cytology. Immunoperoxidase studies of carcinoembryonic antigen,

lysozyme and other tissue antigens in effusions, washes and aspirates. *Acta Cytol* 1983;**27**:408–16.

8. Imai K, Takaoka A. Comparing antibody and small-molecule therapies for cancer. *Nat Rev Cancer* 2006;**6**:714–27. https://doi.org/10.1038/nrc1913

9. Latif ZA, Lozzio BB, Wust CJ. et al. Evaluation of drug-antibody conjugates in the treatment of human myelosarcomas transplanted in nude mice. *Cancer-Am Cancer Soc* 1980;**45**:1326–33.

10. Khongorzul P, Ling CJ, Khan FU. et al. Antibody-drug conjugates: a comprehensive review. *Mol Cancer Res* 2020;**18**:3–19. https://doi.org/10.1158/1541-7786.MCR-19-0582

11. Kovtun YV, Goldmacher VS. Cell killing by antibody-drug conjugates. *Cancer Lett* 2007;**255**:232–40. https://doi.org/10.1016/j.canlet.2007.04.010

12. Fu Z, Li S, Han S. et al. Antibody drug conjugate: the "biological missile" for targeted cancer therapy. *Signal Transduct Target Ther* 2022;**7**:93. https://doi.org/10.1038/s41392-022-00947-7

13. Ford CH, Newman CE, Johnson JR. et al. Localisation and toxicity study of a vindesine-anti-CEA conjugate in patients with advanced cancer. *Br J Cancer* 1983;**47**:35–42. https://doi.org/10.1038/bjc.1983.4

14. McGavin JK, Spencer CM. Gemtuzumab ozogamicin. *Drugs* 2001;**61**:1317–22. https://doi.org/10.2165/00003495-200161090-00007

15. Dosio F, Brusa P, Cattel L. Immunotoxins and anticancer drug conjugate assemblies: the role of the linkage between components. *Toxins (Basel)* 2011;**3**:848–83. https://doi.org/10.3390/toxins3070848

16. Lucas AT, Price L, Schorzman A. et al. Factors affecting the pharmacology of antibody-drug conjugates. *Antibodies (Basel)* 2018;**7**:10. https://doi.org/10.3390/antib7010010

17. Wang Z, Li H, Gou L. et al. Antibody-drug conjugates: recent advances in payloads. *Acta Pharm Sin B* 2023;**13**:4025–59. https://doi.org/10.1016/j.apsb.2023.06.015

18. Tsuchikama K, An Z. Antibody-drug conjugates: recent advances in conjugation and linker chemistries. *Protein Cell* 2018;**9**:33–46. https://doi.org/10.1007/s13238-016-0323-0

19. Strop P, Delaria K, Foletti D. et al. Site-specific conjugation improves therapeutic index of antibody drug conjugates with high drug loading. *Nat Biotechnol* 2015;**33**:694–6. https://doi.org/10.1038/nbt.3274

20. Shao T, Chen T, Chen Y. et al. Construction of paclitaxel-based antibody-drug conjugates with a PEGylated linker to achieve superior therapeutic index. *Signal Transduct Target Ther* 2020;**5**:132. https://doi.org/10.1038/s41392-020-00247-y

21. Beck A, Goetsch L, Dumontet C. et al. Strategies and challenges for the next generation of antibody-drug conjugates. *Nat Rev Drug Discov* 2017;**16**:315–37. https://doi.org/10.1038/nrd.2016.268

22. McCombs JR, Owen SC. Antibody drug conjugates: design and selection of linker, payload and conjugation chemistry. *AAPS J* 2015;**17**:339–51. https://doi.org/10.1208/s12248-014-9710-8

23. Nejadmoghaddam MR, Minai-Tehrani A, Ghahremanzadeh R. et al. Antibody-drug conjugates: possibilities and challenges. *Avicenna J Med Biotechnol* 2019;**11**:3–23.

24. Melo R, Lemos A, Preto AJ. et al. Computational approaches in antibody-drug conjugate optimization for targeted cancer therapy. *Curr Top Med Chem* 2018;**18**:1091–109. https://doi.org/10.2174/1568026618666180731165222

25. Gupta N, Ansari A, Dhoke GV. et al. Computationally designed antibody-drug conjugates self-assembled via affinity ligands. *Nat Biomed Eng* 2019;**3**:917–29. https://doi.org/10.1038/s41551-019-0470-8

26. Coumans R, Ariaans GJA, Spijker HJ. et al. A platform for the generation of site-specific antibody-drug conjugates that allows for selective reduction of engineered cysteines. *Bioconjug Chem* 2020;**31**:2136–46. https://doi.org/10.1021/acs.bioconjchem.0c00337

27. Srivastava R. Computational studies on antibody drug conjugates (ADCs) for precision oncology. *ChemistrySelect* 2022;**7**:e202202259. https://doi.org/10.1002/slct.202202259

28. Fauteux F, Hill JJ, Jaramillo ML. et al. Computational selection of antibody-drug conjugate targets for breast cancer. *Oncotarget* 2016;**7**:2555–71. https://doi.org/10.18632/oncotarget.6679

29. Ai D, Wu J, Cai H. et al. A multi-task FP-GNN framework enables accurate prediction of selective PARP inhibitors. *Front Pharmacol* 2022;**13**:971369. https://doi.org/10.3389/fphar.2022.971369

30. Ai D, Cai H, Wei J. et al. DEEPCYPs: a deep learning platform for enhanced cytochrome P450 activity prediction. *Front Pharmacol* 2023;**14**:1099093. https://doi.org/10.3389/fphar.2023.1099093

31. Wu J, Xiao Y, Lin M. et al. DeepCancerMap: a versatile deep learning platform for target- and cell-based anticancer drug discovery. *Eur J Med Chem* 2023;**255**:115401. https://doi.org/10.1016/j.ejmech.2023.115401

32. Li F, Hu Q, Zhang X. et al. DeepPROTACs is a deep learning-based targeted degradation predictor for PROTACs. *Nat Commun* 2022;**13**:7133. https://doi.org/10.1038/s41467-022-34807-3

33. Tao S, Chen Y, Wu J. et al. VDDB: a comprehensive resource and machine learning tool for antiviral drug discovery. *MedComm – Future Medicine* 2023;**2**:e32. https://doi.org/10.1002/mef2.32

34. Wu Z, Jiang D, Wang J. et al. Knowledge-based BERT: a method to extract molecular features like computational chemists. *Brief Bioinform* 2022;**23**:bbac131. https://doi.org/10.1093/bib/bbac131

35. Cai H, Zhang H, Zhao D. et al. FP-GNN: a versatile deep learning architecture for enhanced molecular property prediction. *Brief Bioinform* 2022;**23**:bbac408. https://doi.org/10.1093/bib/bbac408

36. Li B, Lin M, Chen T. et al. FG-BERT: a generalized and self-supervised functional group-based molecular representation learning framework for properties prediction. *Brief Bioinform* 2023;**24**:bbad398. https://doi.org/10.1093/bib/bbad398

37. Zhu W, Zhang Y, Zhao D. et al. HiGNN: a hierarchical informative graph neural network for molecular property prediction equipped with feature-wise attention. *J Chem Inf Model* 2023;**63**:43–55. https://doi.org/10.1021/acs.jcim.2c01099

38. Zeng X, Xiang H, Yu L. et al. Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework. *Nature Machine Intelligence* 2022;**4**:1004–16. https://doi.org/10.1038/s42256-022-00557-6

39. Shen WX, Zeng X, Zhu F. et al. Out-of-the-box deep learning prediction of pharmaceutical properties by broadly learned knowledge-based molecular representations. *Nature Machine Intelligence* 2021;**3**:334–43. https://doi.org/10.1038/s42256-021-00301-6

40. Xiong Z, Wang D, Liu X. et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *J Med Chem* 2020;**63**:8749–60. https://doi.org/10.1021/acs.jmedchem.9b00959

41. Wu Z, Jiang D, Hsieh CY. et al. Hyperbolic relational graph convolution networks plus: a simple but highly efficient QSAR-modeling method. *Brief Bioinform* 2021;**22**:bbab112. https://doi.org/10.1093/bib/bbab112

42. Zhao Y, Yin J, Zhang L. et al. Drug-drug interaction prediction: databases, web servers and computational models. *Brief Bioinform* 2023;**25**:bbad445. https://doi.org/10.1093/bib/bbad445

43. Wang CC, Zhao Y, Chen X. Drug-pathway association prediction: from experimental results to computational models. *Brief Bioinform* 2021;**22**:bbaa061. https://doi.org/10.1093/bib/bbaa061

44. Chen X, Yan CC, Zhang X. *et al.* Drug-target interaction prediction: databases, web servers and computational models. *Brief Bioinform* 2016;**17**:696–712. https://doi.org/10.1093/bib/bbv066

45. Chen X, Guan NN, Sun YZ. *et al.* MicroRNA-small molecule association identification: from experimental results to computational models. *Brief Bioinform* 2020;**21**:47–61. https://doi.org/10.1093/bib/bby098

46. Fralish Z, Chen A, Skaluba P. *et al.* DeepDelta: predicting ADMET improvements of molecular derivatives with deep learning. *J Chem* 2023;**15**:101. https://doi.org/10.1186/s13321-023-00769-x

47. Shen L, Sun X, Chen Z. *et al.* ADCdb: the database of antibody-drug conjugates. *Nucleic Acids Res* 2023;**52**:D1097–109. https://doi.org/10.1093/nar/gkad831

48. Lin Z, Akin H, Rao R. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023;**379**:1123–30. https://doi.org/10.1126/science.ade2574

49. Tarantino P, Ricciuti B, Pradhan SM. *et al.* Optimizing the safety of antibody-drug conjugates for patients with solid tumours. *Nat Rev Clin Oncol* 2023;**20**:558–76. https://doi.org/10.1038/s41571-023-00783-w

50. Tang Y, Tang F, Yang Y. *et al.* Real-time analysis on drug-antibody ratio of antibody-drug conjugates for synthesis, process optimization, and quality control. *Sci Rep* 2017;**7**:7763. https://doi.org/10.1038/s41598-017-08151-2

51. Devlin J, Ming-Wei C, Lee K. *et al. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.* arXiv preprint arXiv 2018:1810.04805.

52. Abadi M. *et al.* TensorFlow: large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv 2016:1603.04467.

53. Davis LJ, Offord KP. Logistic regression. *J Pers Assess* 1997;**68**:497–507. https://doi.org/10.1207/s15327752jpa6803_3

54. Breiman L. Random forests. *Mach Learn* 2001;**45**:5–32. https://doi.org/10.1023/A:1010933404324

55. Guo Q, Zhang H, Deng Y. *et al.* Ligand- and structural-based discovery of potential small molecules that target the colchicine site of tubulin for cancer treatment. *Eur J Med Chem* 2020;**196**:112328. https://doi.org/10.1016/j.ejmech.2020.112328

56. Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995;**20**:273–97. https://doi.org/10.1007/BF00994018

57. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. arXiv preprint arXiv 2016:1603.02754.

58. Durant JL, Leland BA, Henry DR. *et al.* Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Comput Sci* 2002;**42**:1273–80. https://doi.org/10.1021/ci010132r

59. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model* 2010;**50**:742–54. https://doi.org/10.1021/ci100050t

60. Drago JZ, Modi S, Chandarlapaty S. Unlocking the potential of antibody-drug conjugates for cancer therapy. *Nat Rev Clin Oncol* 2021;**18**:327–44. https://doi.org/10.1038/s41571-021-00470-8

61. Ross J, Belgodere B, Chenthamarakshan V. *et al.* Large-scale chemical language representations capture molecular structure and properties. *NATURE MACHINE INTELLIGENCE* 2022;**4**:1256–64. https://doi.org/10.1038/s42256-022-00580-7

62. Bemis GW, Murcko MA. The properties of known drugs. 1. Molecular frameworks. *J Med Chem* 1996;**39**:2887–93. https://doi.org/10.1021/jm9602928

63. Bei Y, He J, Dong X. *et al.* Targeting CD44 variant 5 with an antibody-drug conjugate is an effective therapeutic strategy for intrahepatic cholangiocarcinoma. *Cancer Res* 2023;**83**:2405–20. https://doi.org/10.1158/0008-5472.CAN-23-0510

64. Bacauanu V, Merz ZN, Hua ZL. *et al.* Nickel-catalyzed antibody bioconjugation. *J Am Chem Soc* 2023;**145**:25842–9. https://doi.org/10.1021/jacs.3c10185

65. Zammarchi F, Havenith KE, Sachini N. *et al.* ADCT-602, a novel PBD dimer-containing antibody-drug conjugate for treating CD22-positive hematologic malignancies. *Mol Cancer Ther* 2024;**23**:520–31. https://doi.org/10.1158/1535-7163.MCT-23-0506

66. Guo Q, Gao B, Song R. *et al.* FZ-AD005, a novel DLL3-targeted antibody-drug conjugate with topoisomerase I inhibitor, shows potent antitumor activity in preclinical models. *Mol Cancer Ther* 2024;**23**:1367–77. https://doi.org/10.1158/1535-7163.MCT-23-0701