Journal of Zhejiang University-SCIENCE B (Biomedicine & Biotechnology) 2025 26(5):409-420
www.jzus.zju.edu.cn; www.springer.com/journal/11585
E-mail: jzus_b@zju.edu.cn

Review

https://doi.org/10.1631/jzus.B2400387

Recent advances in antibody optimization based on deep learning methods

Ruofan JIN[1], Ruhong ZHOU[1,2]✉, Dong ZHANG[1]✉

[1]*Institute of Quantitative Biology, College of Life Sciences, Zhejiang University, Hangzhou 310058, China*
[2]*Department of Hepatobiliary and Pancreatic Surgery, The First Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou 310058, China*

Abstract: Antibodies currently comprise the predominant treatment modality for a variety of diseases; therefore, optimizing their properties rapidly and efficiently is an indispensable step in antibody-based drug development. Inspired by the great success of artificial intelligence-based algorithms, especially deep learning-based methods in the field of biology, various computational methods have been introduced into antibody optimization to reduce costs and increase the success rate of lead candidate generation and optimization. Herein, we briefly review recent progress in deep learning-based antibody optimization, focusing on the available datasets and algorithm input data types that are crucial for constructing appropriate deep learning models. Furthermore, we discuss the current challenges and potential solutions for the future development of general-purpose deep learning algorithms in antibody optimization.

Key words: Deep learning; Antibody optimization; Available dataset; Input data type

1 Introduction

Antibodies, also recognized as immunoglobulins (Igs), represent a class of specialized immune proteins synthesized by B lymphocytes in the adaptive immune system. As an important means for the human body to fight against foreign antigens in humoral immunity, antibodies demonstrate high specificity in recognizing and neutralizing a diverse array of foreign pathogens (Zurawski and McLendon, 2020; Young et al., 2022), including bacteria (Zheng et al., 2020), viruses (Sun et al., 2022), fungi (Boniche et al., 2020; Doron et al., 2021), parasites (Thirumalai et al., 2019), and other extraneous substances (Akter et al., 2019). Antibodies adopt a "Y"-shaped architecture, comprised of two paired heavy chains and light chains connected by disulfide bonds, with both heavy and light chains containing constant regions and variable regions (VRs). The high "variability" of these VRs arises from the recombination reaction of V(D)J (V, variable; D, diversity; J, joining) genes, acting as the basis of antibody specificity. This recombination process engenders a vast repertoire of VR combinations, empowering antibodies to exhibit specificity for countless antigens sourced from diverse origins. There are three loops within each VR, called complementarity-determining regions (CDRs), which are pivotal for binding specificity. In humoral immunity, the initiation of processes such as somatic hypermutation (SHM) and class switching, induced by factors like T follicular helper cells, is indispensable for the maturation of B cells. These processes promote the gradual development of high-affinity antibodies, known as affinity maturation. Notably, a multitude of amino acid point mutations that augment antibody affinity occur within CDRs, which is attributed to their characteristics of being hypervariable regions.

Over the past three decades, monoclonal antibodies (mAbs) have become the predominant treatment modality for a variety of diseases, including various mAbs for cancer treatment, due to their excellent binding affinity and specificity. Köhler and Milstein

✉ Ruhong ZHOU, rhzhou@zju.edu.cn
Dong ZHANG, zhangd_iqb@zju.edu.cn
Ruhong ZHOU, https://orcid.org/0000-0001-8624-5591
Dong ZHANG, https://orcid.org/0000-0001-7297-5083
Ruofan JIN, https://orcid.org/0009-0001-3017-2423

Received July 24, 2024; Revision accepted Nov. 9, 2024; Crosschecked May 27, 2025

(1975) introduced the hybridoma technique, which made it possible to obtain pure mAbs in large quantities, greatly enhancing their potential for basic research and clinical applications. Since then, major technological advances have facilitated the discovery and development of more effective mAb therapies. As of July 2024, the United States Food and Drug Administration (US FDA) had approved 142 mAbs, with efficacy ranging from mere blockade to the activation and enhancement of natural immune responses (Fig. 1) (Mullard, 2021; Baldo, 2022; Kaplon et al., 2023; Crescioli et al., 2024). Compared with conventional molecular-based medicine, therapeutic antibodies offer distinctive advantages, including superior specificity, prolonged serum half-life, enhanced affinity, and strong immune effector function, thus occupying a central position in drug research and development. It is predicted that the global market for cancer mAbs will soar to approximately $12 billion by the culmination of 2026 (do Pazo et al., 2021).

Despite the tremendous commercial success of antibodies, their discovery and optimization for targeted applications remain a time-consuming and costly venture. In fact, though effective individual mAb therapies can cost up to $100 000 per year (Hernandez et al., 2018). As a result, the need for advanced diagnostics, therapeutics, convenient research tools, and an overall quest for a healthier future has led to the development of computational methods, such as machine learning (ML)-based approaches, to accelerate the rapid, inexpensive, on-demand generation of fit-for-purpose antibodies (Akbar et al., 2022). These in silico methods attempt to leverage advances in computational processing power to reduce costs and increase the success rate of lead candidate generation and optimization. Recently, deep learning (DL)-based methods have achieved great success in the field of biology, both in terms of tackling well-established problems (e.g., protein three-dimensional structure prediction by Alphafold (Abramson et al., 2024)) and creating entirely new fields (e.g., novel antibody design by RFdiffusion (Bennett et al., 2024)). This has driven the introduction of DL-based approaches into antibody engineering in the last few years, yielding significant progress in areas such as the language-based modeling of antibody repertoires and the DL-based generation of novel sequences (Graves et al., 2020; Wilman et al., 2022). Here, we briefly review the current research progress in leveraging DL-based methodologies for antibody drug design and optimization, then discuss the current limitations, and propose possible future research directions.

2 Deep learning-based antibody optimization

Artificial intelligence (AI) is a form of intelligent behavior achieved through computer programs, which can simulate human intelligence by autonomous learning, reasoning, judging, and decision-making on

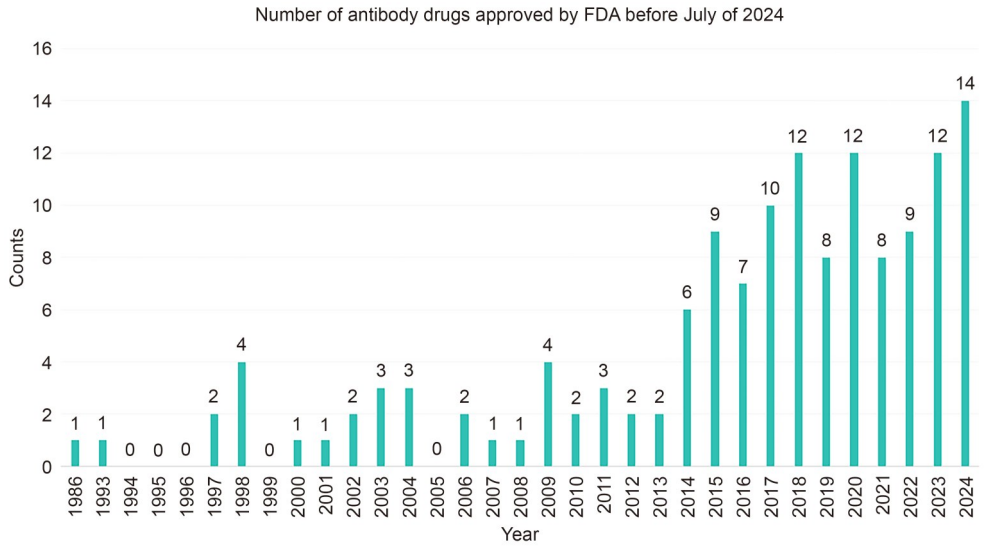

Fig. 1 Number of antibody drugs approved by the United States Food and Drug Administration (US FDA) as of July 2024. Data comes from official information released by the US FDA.

the basis of large amounts of data. Since the emergence of AlphaGo (Silver et al., 2016) in 2016, AI has triggered a global industrial revolution, brought about research innovations in various fields, and become a crucial factor in technological revolution and social development. AI algorithms accelerate the innovative development of biopharmaceuticals, reshape traditional clinical drug screening and design methods, and offer new approaches for exploring biophysical mechanisms and engineering drug molecules. For instance, ML- and DL-based methods have been widely utilized in drug discovery and optimization processes, including peptide synthesis, virtual screening, toxicity prediction, drug monitoring and release, pharmacophore modeling, quantitative structure–activity relationship, drug repurposing, polypharmacology, and bioactivity (Gupta et al., 2021).

In addition to drug design and optimization, AI methods have provided a reliable solution for predicting protein–protein interactions (PPIs), a crucial achievement in understanding important biological processes at the molecular level (Guo and Yamaguchi, 2022). Accurate PPI predictions enable us to explore cellular mechanisms, signal transduction pathways, and disease-related protein networks, offering valuable insights into therapeutic interventions such as antibody therapies. By leveraging large-scale protein datasets such as STRING (Szklarczyk et al., 2023) and the Protein Data Bank (PDB) (Berman et al., 2000), DL models have achieved high accuracy in predicting interaction interfaces and binding affinities, outperforming traditional computational approaches (Chen et al., 2013; Soleymani et al., 2022). For example, algorithms such as graph neural networks (Wu et al., 2021; Lee, 2023; Mastropietro et al., 2023; Réau et al., 2023) and convolutional neural networks (O'Shea and Nash, 2015; Guo and Yamaguchi, 2022; Soleymani et al., 2023) can capture the complex relationships between proteins encoded in sequence and structural features, thereby improving prediction accuracy. Moreover, the interpretability of DL models provides a reliable means for validating the physicochemical properties associated with PPIs, greatly helping researchers to explore and discover physicochemical principles using DL algorithms (Chen et al., 2013; Wang et al., 2023). These advances have not only accelerated research in structural biology but also provided new avenues to address key bioengineering challenges, such as antibody optimization and functional enzyme design (Listov et al., 2024; Notin et al., 2024).

In this review, we focus on using AI algorithms, especially DL-based methods, to assist in the design and optimization of antibodies. These emerging computational methods aim to predict and optimize antibody properties faster and more accurately, reduce costs and risks, and provide a deeper understanding of the underlying mechanisms, thereby accelerating the development and clinical application of related downstream antibody-based technologies (Fig. 2). On the one hand, DL-based methods are always data-intensive, as they rely on a curated dataset composed of a large amount of well-labeled data to effectively train and

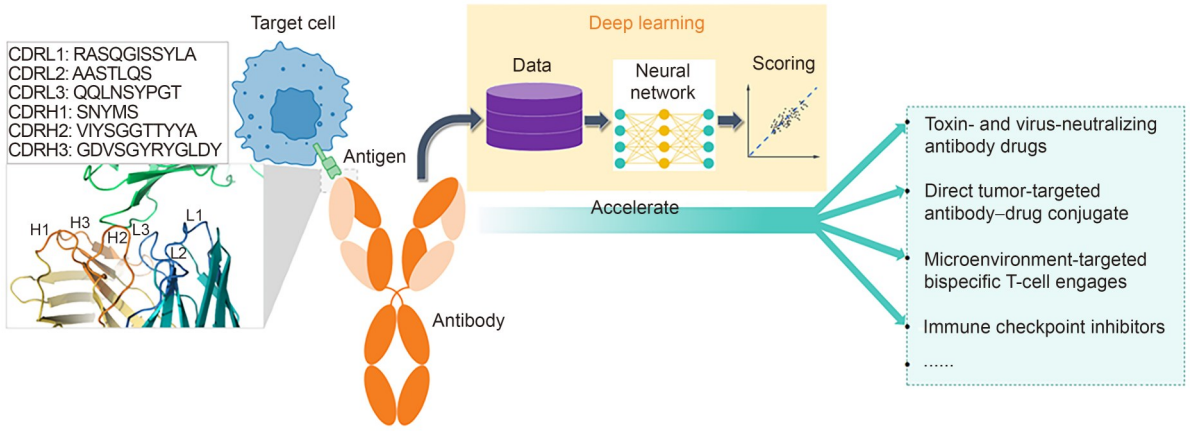

Fig. 2 Illustrative scheme showing that deep learning-based methods can accelerate the development of downstream antibody-based technologies, including the optimized ones for neutralizing toxins and viruses, antibody–drug conjugate (ADC)-targeting tumors, bispecific T-cell engagers (BiTEs) for microenvironment targeting, and engineered immune checkpoint inhibitors. When building deep learning methods, a well-curated dataset is indispensable for training and optimizing the models, whereas sequence information alone or combined with structural details can be used as model inputs.

optimize models. For instance, the high-accurate three-dimensional structure prediction of proteins by Alphafold2 (Jumper et al., 2021) largely benefits from the accumulation of around 140 000 experimental structures deposited in PDB. On the other hand, while sequence information is always the primary input of DL-based methods, the featurization of structural details as a complementary input is sometimes expected to improve model performance, such as binding affinity prediction (Meli et al., 2022). That is, DL-based methods can firstly use different inputs (sole sequence or combined sequence and structural information) and then combine varied model architectures to achieve their desired goals. Therefore, we introduce the recent progress in DL-based antibody optimization from two perspectives: available datasets and input data types. We hope that this review can guide experts to easily select appropriate datasets to train and optimize their own DL-based models, to choose the proper input data type according to their desired goals, and ultimately to estimate the upper/lower limits of the constructed model.

2.1 Available datasets

Firstly, we overviewed the main public datasets that are currently available for constructing DL-based antibody optimization models. These datasets contain PPI information related to antibodies or antigen–antibody complexes, as well as affinity change information such as dissociation constants and free energy changes. Therefore, they are essential for DL models aimed at optimizing antibodies. Several notable public datasets are outlined below, and a summary is displayed in Table 1.

Toseland et al. (2005) introduced the AntiJen database, a curated dataset of B and T cell antigens with experimental annotations, links to published experimental articles, and PDB entries. AntiJen has focused on continuous quantitative data on peptide binding with the transporter associated with antigen processing (TAP) protein complex and the major histocompatibility complex (MHC). Douguet et al. (2006) presented the DOCKGROUND database, which serves as a benchmark dataset for protein–protein complexes whose decoys were generated by the docking algorithm GRAMM-X (Tovchigrechko and Vakser, 2006). This database includes 61 real complexes and 100 generated negative data. Ansari et al. (2010) launched AntigenDB, a validated antigen database containing the structural, sequence, and binding data of verified antigens. Dunbar et al. (2014) introduced the Structural Antibody Database (SAbDab), an automated, regularly updated structural antibody database that includes annotations of affinity data, CDR classifications, and other antibody-specific information.

Since 2016, the quality and quantity of emerging datasets have been significantly improved. Sirin et al. (2016) presented the Antibody-Bind (AB-Bind) dataset, a collection of 1101 mutations in 32 different antibody-antigen structures, including the experimentally determined binding free energy change associated with each mutation, as well as the experimental conditions under which each mutation was tested. Vita et al. (2019) launched the Immune Epitope Database (IEDB), which contains sequence, experimental, and structural data information for 1 619 619 linear and discontinuous epitopes from 4505 antigens (as of July 3, 2024). Although the structure of each native protein complex is provided therein, it is up to the database users to simulate any changes to the native structure caused by the mutations. Jankauskaitė et al. (2019) presented the Structural Kinetic and Energetic Database of Mutant Protein Interactions (SKEMPI) database, which collects a set of structurally resolved protein mutations and their corresponding binding free energy changes. As one of the latest datasets, the SKEMPI database has been employed for model training and testing in several recent studies aimed at antibody optimization (Jankauskaitė et al., 2019; Soleymani et al., 2022).

In addition, there are some highly popular protein databases, such as PDB and PDBbind, which are also very useful for DL-based antibody optimization, although they are not specifically designed for this purpose. PDB was proposed by Berman et al. (2000) and is currently the largest and most famous protein database. Many curated datasets used in DL-based methods are collected from PDB. Wang et al. (2004) conducted a comprehensive screening of the complete protein database PDBbind and successfully identified 5671 protein–ligand complexes from a pool of 19 621 experimental structures (as of January 2004).

Overall, the aforementioned public datasets provide a rich source of sequence and structure information (as input) as well as binding data (as labels) for training and evaluating DL-based methods, forming the indispensable resources for developing DL models in antibody optimization. Furthermore, we hope that

Table 1 Summary of available datasets for deep learning-based antibody optimization

Database name	Reference	Number of entries	Website	Brief description
AntiJen	Toseland et al., 2005	24 000 (as of July 3, 2024)	https://www.ddg-pharmfac.net/antijen/AntiJen/antijenhomepage.htm	A curated dataset of B and T cell antigens with experimental annotations, links to articles, and Protein Data Bank (PDB) entries, which focuses on continuous quantitative peptide-binding data with transporter associated with antigen processing (TAP) and major histocompatibility complex (MHC)
DOCKGROUND	Douguet et al., 2006	102 678 (as of July 3, 2024)	https://dockground.compbio.ku.edu	A database offering extensive experimental data and computational models to aid in understanding protein–protein interactions (PPIs), including comprehensive datasets, stringent structural quality assessments, and intuitive interface for easy access and retrieval
AntigenDB	Ansari et al., 2010	504 (as of July 3, 2024)	http://crdd.osdd.net/raghava/antigendb	A validated antigen database containing structural, sequence, and binding data of verified antigens
SAbDab	Dunbar et al., 2014	8616 (as of July 3, 2024)	https://opig.stats.ox.ac.uk/webapps/sabdab-sabpred/sabdab	An automated, regularly updated structural antibody database including affinity data annotations and complementarity-determining region (CDR) classification
AB-Bind	Sirin et al., 2016	1101 (as of July 3, 2024)	https://github.com/sarahsirin/AB-Bind-Database	A collection of 1101 mutations in 32 antibody-antigen structures, with associated changes in binding free energy and experimental conditions for each mutation
Immune Epitope Database	Vita et al., 2019	1 619 619 (as of July 3, 2024)	https://www.iedb.org	A database containing sequences, experimental, and structural data of 1 619 619 linear and discontinuous epitopes from 4505 antigens
SKEMPI	Jankauskaitė et al., 2019	7085 (as of July 3, 2024)	https://life.bsc.es/pid/skempi2	A compilation of protein mutations with resolved structures, including changes in binding free energy
PDB	Berman et al., 2000	1 290 613 (as of July 3, 2024)	https://www.rcsb.org	Established as the largest and most renowned protein database, it serves as a primary source for numerous curated datasets utilized extensively in deep learning research.
PDBbind	Wang et al., 2004	27 408 (as of July 3, 2024)	https://www.pdbbind-plus.org.cn	This database was created on the basis of comprehensive screening of the entire protein database up to 2024, identifying over 22 000 protein–ligand complexes, with 19 443 containing experimental structures. It hosts 2852 protein–protein complexes with experimentally determined binding affinity data that are critical for understanding the molecular mechanisms of protein interactions, providing valuable resources for the study of PPIs. Openly accessible to researchers worldwide through the PDBbind database.

continuously updated and emerging datasets that cover more sequence space and have well-defined labels will further improve the performance of DL-based methods in antibody optimization.

2.2 Input data types

According to the input data type, the current DL-based antibody optimization algorithms can be roughly categorized into two classes: those based solely on sequence and those based on both sequence and structural data. Since different input data types may result in varied model performances, we introduce these two types of algorithms separately below (Tables 2 and 3, respectively).

2.2.1 DL-based antibody optimization methods using only sequence data

Protein sequence data are relatively abundant and convenient to obtain. By leveraging DL methods, it is possible to quickly generate a large number of virtual

Table 2 Summary of deep learning-based antibody optimization methods using only sequence data

Reference	Data source	Pros	Cons
Mason et al., 2021	Training set: approximately 5×10^4 antibody sequences generated through deep mutational scanning (DMS) along with their antigen-binding data, including both binding and non-binding variants Test set: approximately 3×10^6 antibody sequence variants predicted by the neural network model from a combinatorial mutagenesis library of around 1×10^8 sequences, with subsequent experimental validation of their antigen specificity Link: https://github.com/dahjan/DMS_opt	1. Utilizes deep learning (DL) to predict antigen specificity from diverse sequence space; 2. Identifies globally optimized antibody variants using clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated nuclease 9 (Cas9)-mediated homology-directed repair (HDR)	Requires a well-designed mutagenesis library for effective training
Kang et al., 2022	Mutated antibodies from AB-Bind Link: https://github.com/sarahsirin/AB-Bind-Database	1. Utilizes the Hag-Net neural network for antibody optimization; 2. Focuses on complementarity-determining regions (CDRs) for affinity contributions	Limited to certain regions of the antibody sequence; specificity beyond these regions not explored extensively
Makowski et al., 2022	Training set: approximately 1×10^6 antibody sequences generated by mutating the clinical-stage antibody emibetuzumab at eight positions within the heavy-chain CDRs. These sequences were sorted using yeast surface display and deep sequencing to collect binding data, forming the training set for model building Test set: a subset of the 1×10^6 sequences from the training set was used to evaluate the performance of the machine learning models in predicting antibody affinity and specificity Link: https://github.com/Tessier-Lab-UMich/Emi_Pareto_Opt_ML	1. Uses DL to collectively optimize various antibody properties; 2. Demonstrates a trade-off between affinity and specificity in antibody optimization	The trade-off between affinity and specificity limits the simultaneous enhancement of both attributes

mutant sequences, conserving significant biological experimental resources and reducing time costs. These methods have notably improved the efficiency of antibody optimization. Here, we focus on representative studies in antibody optimization using DL methods.

Mason et al. (2021) harnessed DL to inquire and predict antigen specificity from a vast and diverse sequence space. By building a well-designed experimental site-directed mutagenesis library using clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated nuclease 9 (Cas9)-mediated homology-directed repair (HDR) and using the resulting sequence data as input, the authors developed DL models to identify globally optimized antibody variants of the therapeutic antibody trastuzumab. These DL models were trained on 5×10^4 antibody sequence mutations and eventually screened out thousands of effective mutation sites as potential antibody optimization strategies. The results validated the feasibility and reliability of DL algorithms for antibody optimization.

Subsequently, Kang et al. (2022) introduced an antibody optimization algorithm on the basis of the Hag-Net neural network. They trained and validated their neural network on AB-Bind, the dataset from multiple antigen–antibody interaction pairs, and found that the neural network mainly focused on the CDRs of the antibody sequence, particularly the regions responsible for interactions with antigens, when learning the contributions of antibody sequences to binding affinity. This computational perspective corroborated the critical binding patterns in antigen–antibody interactions.

Makowski et al. (2022) employed DL methods to collectively optimize various properties of the antibody emibetuzumab. The authors experimentally mutated sites within the CDRs of the antibody, ranked the sequence library for high-affinity, low-affinity, and non-specific binding, and conducted deep sequencing on the enriched library. They discovered that ML models trained on datasets with binary labels could predict continuous indicators that are closely related to

Table 3 Overview of deep learning-based antibody optimization methods using both sequence and structural data

Model name	Reference	Data source	Pros	Cons
mCSM-AB	Pires and Ascher, 2016	Mutated antibodies from AB-Bind Link: https://github.com/sarahsirin/AB-Bind-Database	1. Utilizes graph neural networks for feature extraction from pre- and post-mutation antigen-antibody structural data; 2. Broad applicability to diverse antibody optimization scenarios; 3. Accessible as a comprehensive web-based tool	Requires a degree of familiarity with structural data analysis
Ens-Grad	Liu et al., 2020	Training set: approximately 572 647 unique complementarity-determining region heavy chain 3 (CDR-H3) sequences obtained from the first round of phage display panning experiments against ranibizumab, bevacizumab, etanercept, and trastuzumab, with subsequent rounds generating 297 290 and 171 568 unique sequences. These were used to train the machine learning models to predict antibody enrichment and binding affinity Test set: a subset of sequences including 558 400, 265 563, and 96 912 unique CDR-H3 sequences from a replicate experiment with ranibizumab, along with additional sequences from experiments under stringent washing conditions, used to evaluate the model's performance in predicting antibody affinity and specificity Link: https://github.com/gifford-lab/antibody-2019/tree/master	Designs CDRs of human immunoglobulin G (IgG) antibodies, resulting in higher affinity compared with phage display-derived antibodies	Specific to optimizing CDRs, potential limitations in broader antibody optimization scenarios
TopNetTree	Wang et al., 2020	AB-Bind, SKEMPI	1. Integrates protein crystal structure data for antibody optimization; 2. Introduces a novel approach for the digital extraction of antibody structural information, enhancing predictive powers	Requires specialized knowledge in structural biology for effective utilization
GeoPPI	Liu et al., 2021	Training set: comprised of approximately 7085 protein–protein interaction (PPI) mutation data points from the SKEMPI v2.0 database, including experimentally measured ΔG values and corresponding structural data from the Protein Data Bank (PDB) Test set: comprised of a subset of PPI mutations not included in the training process, used to evaluate the predictive accuracy of the GeoPPI model, with independent experimental ΔG measurements for validation Link: https://github.com/Liuxg16/GeoPPI	1. Utilizes self-supervised learning to derive geometric representations from protein structures for mutation impact modeling; 2. Combines deep learning with gradient boosting trees to predict binding affinity changes, demonstrating superior performance on benchmark datasets	Requires a pre-training step for geometric representation generation

To be continued

Table 3 (continued)

Model name	Reference	Data source	Pros	Cons
An attention-based geometric neural network	Shan et al., 2022	Training set: primarily comprised of experimentally determined structures of antibody-antigen complexes. These structures were sourced from publicly available databases, such as the PDB. The dataset was augmented with synthetic mutations in the CDRs to generate a diverse set of training examples. The model trained on this data was designed to predict changes in binding affinity caused by these mutations Test set: real-world examples, including various severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) variants with mutations in the spike protein that could potentially affect antibody binding. These variants were not part of the training data, ensuring that the model's performance was evaluated on its ability to generalize to unseen variants. Additionally, experimental validations were performed on selected antibodies optimized by the model to assess their neutralizing activity against these variants Link: https://github.com/HeliXonProtein/binding-ddg-predictor	1. Employs geometric deep learning algorithm to enhance antibody affinity, enabling broader neutralization activity against antibodies variations; 2. Valuable for therapeutic antibody drug development	Specific focus on affinity enhancement may overlook other critical attributes

antibody affinity and non-specific binding. The results indicated a strong trade-off between binding affinity and specificity, as increasing affinity along the Pareto frontier requires a gradual reduction in specificity, suggesting that DL methods can perform knowledge transfer in learning antibody affinity and specificity.

2.2.2 DL-based antibody optimization methods using both sequence and structural data

While sequence data are vastly predominant over structural data, many scientists believe that the inclusion of structural data can significantly enhance the accuracy of DL methods for antibody optimization. Therefore, many methods currently employ structural data (in addition to sequence information) to predict antibody optimization pathways (Table 3).

Pires and Ascher (2016) introduced mCSM-AB, which utilizes graph neural networks to learn and extract features from pre- and post-mutation antigen–antibody structural information. This tool can be broadly applied to various antibody optimization scenarios and is available as a comprehensive web-based tool.

Liu et al. (2020) presented Ens-Grad, a method for designing the CDRs of human immunoglobulin G (IgG) antibodies, which yields antibodies with higher affinity than candidate antibodies derived from phage display experiments.

The graph-based DL method is pivotal in the realm of DL-based antibody optimization with structural data. With the deepening understanding of antibody protein structure information, researchers are gradually integrating algorithm design with geometric representation.

Wang et al. (2020) integrated protein crystal structure information and proposed the TopNetTree model to predict changes in binding affinity caused by mutations in protein–protein complexes. The model innovatively extracts structural information topologically and introduces a practical approach for the digital extraction of antibody structural information.

Liu et al. (2021) developed GeoPPI, a DL framework designed to model the effects of mutations on protein–protein binding affinity using geometric representations. GeoPPI utilizes self-supervised learning to generate structural embeddings from protein complexes, which are then combined with gradient boosting trees to predict changes in binding affinity due to mutations.

Shan et al. (2022) introduced a geometric DL algorithm that can effectively enhance antibody affinity, thereby enabling broader and more effective neutralizing activity against antibody variations, ultimately facilitating the development of therapeutic antibody drugs.

3 Discussion and challenges

DL methods offer significant advantages for antibody optimization by efficiently processing large and complex datasets of antibody sequences and structures. They can rapidly generate virtual mutants, significantly reducing experimental costs and time demand. In addition, DL models can accurately identify and predict antibody specificity and affinity, improving the accuracy of antibody drug design (Chen and Wei, 2022). By integrating sequence and structural data, these models can provide a more comprehensive simulation of antibody–antigen interactions, thereby improving binding affinity predictions and offering a reliable option for the development of therapeutic antibodies.

Although significant advances have been made in DL-based antibody optimization in the past few years, various problems and challenges still need to be overcome to achieve the ultimate goal of rapid and efficient antibody drug design. Accordingly, we attempt to discuss the existing challenges and possible solutions in this field from the perspectives of datasets and algorithms (input data types) below.

3.1 Challenges in public datasets

At present, both the quantity and quality of available antibody-related public databases remain inadequate. Compared with extensive repositories such as PDB and UniProt (Bairoch and Apweiler, 1997; Apweiler et al., 2004), which offer abundant sequence and structural information for general protein-related ML tasks, the curated datasets that could provide extensive antigen-antibody data and high-quality affinity labels, both suitable for DL methods, are still limited in size and quality. Additionally, despite the significant contributions in generating high-throughput datasets through relevant biological experiments, many researchers have not yet made their data publicly accessible.

Although some datasets are dedicated to antigen–antibody interactions, most of them encompass a variety of different PPI data, which require meticulous filtering and analysis. For instance, AB-Bind and SAbDab stand out as databases that exclusively record antigen–antibody interactions and affinity changes due to missense mutations, addressing antibody optimization tasks. However, certain instances of non-B-cell-secreted antibody data still exist in the AB-Bind, which hinders the acquisition of the requisite pure antibody data for DL methods and necessitates essential manual filtering. Additionally, discrepancies in the length of antibody sequences (full-length antibodies or only their CDRs) across these databases further complicate the pre-processing steps of DL-based antibody optimization tasks.

3.2 Current algorithmic landscape

The current realm of DL algorithms built on sequence and structural data offers diverse approaches for antibody optimization. However, the general-purpose algorithms applicable to various antibody optimization prediction tasks are still subject to obvious limitations. That is, they almost rely heavily on structural information, which is usually expensive and low-throughput. While challenges such as the limited availability of well-labeled structural data persist, ongoing efforts are making significant strides to address these issues. Recent developments of AI-driven structure prediction tools like AlphaFold-Multimer (Evans et al., 2021) and AlphaFold3 (Abramson et al., 2024) are rapidly generating accurate antigen–antibody complex structures, reducing reliance on experimental methods, and have already demonstrated promising results in T-cell receptor and antibody structure prediction, as well as in addressing antigen–antibody interaction prediction problems (Zhao et al., 2023). We expect that the accurate predictions of the antigen–antibody complex by AI approaches could significantly accelerate the acquisition of structural information and promote DL-based antibody optimization in the future.

Currently, some efforts have been made to eliminate the reliance on structural data. Notably, the DL algorithm introduced in a recent study (Shan et al., 2022) utilizes site-specific information at the antigen–antibody interaction interface but does not depend on structural data entirely for antibody screening and optimization. Recently, a study employed a multi-head attention mechanism to perform in silico antibody optimization based solely on antigen–antibody sequence

information (Jin et al., 2024). This study introduced the innovative attention-based antibody sequence (AttABseq) model, which efficiently predicts affinity changes caused by amino acid point mutations and outperforms existing methods across multiple metrics.

Sequence-based algorithms may address multiple antibody optimization tasks. These algorithms are pivotal not only for leveraging the vast amount of available sequence data, which greatly exceeds current structural data, but also for controlling the costs associated with redundant model training. By utilizing the potential interpretability of DL methods, sequence-based approaches can facilitate the exploration of biophysical binding patterns across diverse antigen and antibody categories. Recently, current text-based language models exhibit significant advantages in predicting the properties of biomacromolecules and molecular generation (Hou et al., 2023; Lubiana et al., 2023; Lam et al., 2024), highlighting their potential in accelerating and enhancing research in this domain. Building on this progress, protein language model-based DL methods are poised to unlock further performance improvements in antibody optimization approaches, offering a promising avenue for advancement in this field.

Data availability statement

All relevant data are included in the main text.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 12104396), the National Key R&D Program of China (Nos. 2021YFF1200404 and 2021YFA1201200), the National Independent Innovation Demonstration Zone Shanghai Zhangjiang Major Projects (No. ZJZX2020014), the Starry Night Science Fund at Shanghai Institute for Advanced Study of Zhejiang University (No. SN-ZJU-SIAS-003), and the Shanghai Artificial Intelligence Lab (No. P22KN00272), China.

Author contributions

Ruofan JIN conducted the review and research for the study, and wrote and edited the manuscript. Ruhong ZHOU provided guidance for the study and manuscript preparation. Dong ZHANG contributed to the study's conceptualization, offered guidance for the research and manuscript preparation, and participated in writing and revising the manuscript. All authors have read and approved the final manuscript.

Compliance with ethics guidelines

Dong ZHANG is a Young Scientist Committee Member for *Journal of Zhejiang University-SCIENCE B (Biomedicine & Biotechnology)* and was not involved in the editorial review or the decision to publish this article. Ruofan JIN, Ruhong ZHOU, and Dong ZHANG declare that they have no conflicts of interest.

This review does not contain any studies with human or animal subjects performed by any of the authors.

References

Abramson J, Adler J, Dunger J, et al., 2024. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493-500. https://doi.org/10.1038/s41586-024-07487-w

Akbar R, Bashour H, Rawat P, et al., 2022. Progress and challenges for the machine learning-based design of fit-for-purpose monoclonal antibodies. *mAbs*, 14(1):2008790. https://doi.org/10.1080/19420862.2021.2008790

Akter J, Khoury DS, Aogo R, et al., 2019. *Plasmodium*-specific antibodies block in vivo parasite growth without clearing infected red blood cells. *PLoS Pathog*, 15(2):e1007599. https://doi.org/10.1371/journal.ppat.1007599

Ansari HR, Flower DR, Raghava GPS, 2010. AntigenDB: an immunoinformatics database of pathogen antigens. *Nucleic Acids Res*, 38(suppl_1):D847-D853. https://doi.org/10.1093/nar/gkp830

Apweiler R, Bairoch A, Wu CH, et al., 2004. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res*, 32(suppl_1): D115-D119. https://doi.org/10.1093/nar/gkh131

Bairoch A, Apweiler R, 1997. The SWISS-PROT protein sequence data bank and its supplement TrEMBL. *Nucleic Acids Res*, 25(1):31-36. https://doi.org/10.1093/nar/25.1.31

Baldo BA, 2022. Immune- and non-immune-mediated adverse effects of monoclonal antibody therapy: a survey of 110 approved antibodies. *Antibodies*, 11(1):17. https://doi.org/10.3390/antib11010017

Bennett NR, Watson JL, Ragotte RJ, et al., 2024. Atomically accurate de novo design of single-domain antibodies. bioRxiv, preprint. https://doi.org/10.1101/2024.03.14.585103

Berman HM, Westbrook J, Feng ZK, et al., 2000. The protein data bank. *Nucleic Acids Res*, 28(1):235-242. https://doi.org/10.1093/nar/28.1.235

Boniche C, Rossi SA, Kischkel B, et al., 2020. Immunotherapy against systemic fungal infections based on monoclonal antibodies. *J Fungi*, 6(1):31. https://doi.org/10.3390/jof6010031

Chen JH, Wei GW, 2022. Mathematical artificial intelligence design of mutation-proof COVID-19 monoclonal antibodies. *Commun Inf Syst*, 22(3):339-361. https://doi.org/10.4310/CIS.2022.v22.n3.a3

Chen JM, Sawyer N, Regan L, 2013. Protein–protein interactions: general trends in the relationship between binding affinity and interfacial buried surface area. *Protein Sci*, 22(4):510-515. https://doi.org/10.1002/pro.2230

Crescioli S, Kaplon H, Chenoweth A, et al., 2024. Antibodies to watch in 2024. *mAbs*, 16(1):2297450.

https://doi.org/10.1080/19420862.2023.2297450

do Pazo C, Nawaz K, Webster RM, 2021. The oncology market for antibody–drug conjugates. *Nat Rev Drug Discov*, 20(8):583-584. https://doi.org/10.1038/d41573-021-00054-2

Doron I, Mesko M, Li XV, et al., 2021. Mycobiota-induced IgA antibodies regulate fungal commensalism in the gut and are dysregulated in Crohn's disease. *Nat Microbiol*, 6(12):1493-1504. https://doi.org/10.1038/s41564-021-00983-z

Douguet D, Chen HC, Tovchigrechko A, et al., 2006. Dockground resource for studying protein–protein interfaces. *Bioinformatics*, 22(21):2612-2618. https://doi.org/10.1093/bioinformatics/btl447

Dunbar J, Krawczyk K, Leem J, et al., 2014. SAbDab: the structural antibody database. *Nucl Acids Res*, 42(D1): D1140-D1146. https://doi.org/10.1093/nar/gkt1043

Evans R, O'Neill M, Pritzel A, et al., 2021. Protein complex prediction with AlphaFold-Multimer. bioRxiv, preprint. https://doi.org/10.1101/2021.10.04.463034

Graves J, Byerly J, Priego E, et al., 2020. A review of deep learning methods for antibodies. *Antibodies*, 9(2):12. https://doi.org/10.3390/antib9020012

Guo ZL, Yamaguchi R, 2022. Machine learning methods for protein-protein binding affinity prediction in protein design. *Front Bioinform*, 2:1065703. https://doi.org/10.3389/fbinf.2022.1065703

Gupta R, Srivastava D, Sahu M, et al., 2021. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Mol Divers*, 25(3):1315-1360. https://doi.org/10.1007/s11030-021-10217-3

Hernandez I, Bott SW, Patel AS, et al., 2018. Pricing of monoclonal antibody therapies: higher if used for cancer? *Am J Manag Care*, 24(2):109-112.

Hou WP, Shang XY, Ji ZC, 2023. Benchmarking large language models for genomic knowledge with GeneTuring. bioRxiv, preprint. https://doi.org/10.1101/2023.03.11.532238

Jankauskaitė J, Jiménez-García B, Dapkūnas J, et al., 2019. SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics*, 35(3):462-469. https://doi.org/10.1093/bioinformatics/bty635

Jin RF, Ye Q, Wang JK, et al., 2024. AttABseq: an attention-based deep learning prediction method for antigen–antibody binding affinity changes based on protein sequences. *Brief Bioinform*, 25(4):bbae304. https://doi.org/10.1093/bib/bbae304

Jumper J, Evans R, Pritzel A, et al., 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873): 583-589. https://doi.org/10.1038/s41586-021-03819-2

Kang Y, Leng DW, Guo JJ, et al., 2022. Sequence-based deep learning antibody design for in silico antibody affinity maturation. arXiv:2103.03724. https://doi.org/10.48550/arXiv.2103.03724

Kaplon H, Crescioli S, Chenoweth A, et al., 2023. Antibodies to watch in 2023. *mAbs*, 15(1):2153410. https://doi.org/10.1080/19420862.2022.2153410

Köhler G, Milstein C, 1975. Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature*, 256(5517):495-497. https://doi.org/10.1038/256495a0

Lam HYI, Ong XE, Mutwil M, 2024. Large language models in plant biology. *Trends Plant Sci*, 29(10):1145-1155. https://doi.org/10.1016/j.tplants.2024.04.013

Lee M, 2023. Recent advances in deep learning for protein-protein interaction analysis: a comprehensive review. *Molecules*, 28(13):5169. https://doi.org/10.3390/molecules28135169

Listov D, Goverde CA, Correia BE, et al., 2024. Opportunities and challenges in design and optimization of protein function. *Nat Rev Mol Cell Biol*, 25(8):639-653. https://doi.org/10.1038/s41580-024-00718-y

Liu G, Zeng HY, Mueller J, et al., 2020. Antibody complementarity determining region design using high-capacity machine learning. *Bioinformatics*, 36(7):2126-2133. https://doi.org/10.1093/bioinformatics/btz895

Liu XG, Luo YN, Li PY, et al., 2021. Deep geometric representations for modeling effects of mutations on protein-protein binding affinity. *PLoS Comput Biol*, 17(8):e1009284. https://doi.org/10.1371/journal.pcbi.1009284

Lubiana T, Lopes R, Medeiros P, et al., 2023. Ten quick tips for harnessing the power of ChatGPT in computational biology. *PLoS Comput Biol*, 19(8):e1011319. https://doi.org/10.1371/journal.pcbi.1011319

Makowski EK, Kinnunen PC, Huang J, et al., 2022. Co-optimization of therapeutic antibody affinity and specificity using machine learning models that generalize to novel mutational space. *Nat Commun*, 13:3788. https://doi.org/10.1038/s41467-022-31457-3

Mason DM, Friedensohn S, Weber CR, et al., 2021. Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nat Biomed Eng*, 5(6):600-612. https://doi.org/10.1038/s41551-021-00699-9

Mastropietro A, Pasculli G, Bajorath J, 2023. Learning characteristics of graph neural networks predicting protein–ligand affinities. *Nat Mach Intell*, 5(12):1427-1436. https://doi.org/10.1038/s42256-023-00756-9

Meli R, Morris GM, Biggin PC, 2022. Scoring functions for protein-ligand binding affinity prediction using structure-based deep learning: a review. *Front Bioinform*, 2:885983. https://doi.org/10.3389/fbinf.2022.885983

Mullard A, 2021. FDA approves 100th monoclonal antibody product. *Nat Rev Drug Discov*, 20(7):491-495. https://doi.org/10.1038/d41573-021-00079-7

Notin P, Rollins N, Gal Y, et al., 2024. Machine learning for functional protein design. *Nat Biotechnol*, 42(2):216-228. https://doi.org/10.1038/s41587-024-02127-0

O'Shea K, Nash R, 2015. An introduction to convolutional neural networks. arXiv:1511.08458. https://doi.org/10.48550/arXiv.1511.08458

Pires DEV, Ascher DB, 2016. mCSM-AB: a web server for predicting antibody–antigen affinity changes upon mutation

with graph-based signatures. *Nucleic Acids Res*, 44(W1): W469-W473. https://doi.org/10.1093/nar/gkw458

Réau M, Renaud N, Xue LC, et al., 2023. DeepRank-GNN: a graph neural network framework to learn patterns in protein–protein interfaces. *Bioinformatics*, 39(1):btac759. https://doi.org/10.1093/bioinformatics/btac759

Shan SS, Luo ST, Yang ZQ, et al., 2022. Deep learning guided optimization of human antibody against SARS-CoV-2 variants with broad neutralization. *Proc Natl Acad Sci USA*, 119(11):e2122954119. https://doi.org/10.1073/pnas.2122954119

Silver D, Huang A, Maddison CJ, et al., 2016. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484-489. https://doi.org/10.1038/nature16961

Sirin S, Apgar JR, Bennett EM, et al., 2016. AB-bind: antibody binding mutational database for computational affinity predictions. *Protein Sci*, 25(2):393-409. https://doi.org/10.1002/pro.2829

Soleymani F, Paquet E, Viktor H, et al., 2022. Protein–protein interaction prediction with deep learning: a comprehensive review. *Comput Struct Biotechnol J*, 20:5316-5341. https://doi.org/10.1016/j.csbj.2022.08.070

Soleymani F, Paquet E, Viktor HL, et al., 2023. ProtInteract: a deep learning framework for predicting protein–protein interactions. *Comput Struct Biotechnol J*, 21:1324-1348. https://doi.org/10.1016/j.csbj.2023.01.028

Sun XY, Yi CY, Zhu YF, et al., 2022. Neutralization mechanism of a human antibody with pan-coronavirus reactivity including SARS-CoV-2. *Nat Microbiol*, 7(7):1063-1074. https://doi.org/10.1038/s41564-022-01155-3

Szklarczyk D, Kirsch R, Koutrouli M, et al., 2023. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res*, 51(D1):D638-D646. https://doi.org/10.1093/nar/gkac1000

Thirumalai D, Visaga Ambi S, Vieira-Pires RS, et al., 2019. Chicken egg yolk antibody (IgY) as diagnostics and therapeutics in parasitic infections–a review. *Int J Biol Macromol*, 136:755-763. https://doi.org/10.1016/j.ijbiomac.2019.06.118

Toseland CP, Clayton DJ, McSparron H, et al., 2005. AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data. *Immunome Res*, 1:4. https://doi.org/10.1186/1745-7580-1-4

Tovchigrechko A, Vakser IA, 2006. GRAMM-X public web server for protein–protein docking. *Nucleic Acids Res*, 34(suppl_2):W310-W314. https://doi.org/10.1093/nar/gkl206

Vita R, Mahajan S, Overton JA, et al., 2019. The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res*, 47(D1):D339-D343. https://doi.org/10.1093/nar/gky1006

Wang GY, Liu XH, Wang K, et al., 2023. Deep-learning-enabled protein–protein interaction analysis for prediction of SARS-CoV-2 infectivity and variant evolution. *Nat Med*, 29(8):2007-2018. https://doi.org/10.1038/s41591-023-02483-5

Wang ML, Cang ZX, Wei GW, 2020. A topology-based network tree for the prediction of protein–protein binding affinity changes following mutation. *Nat Mach Intell*, 2(2): 116-123. https://doi.org/10.1038/s42256-020-0149-6

Wang RX, Fang XL, Lu YP, et al., 2004. The PDBbind database: collection of binding affinities for protein–ligand complexes with known three-dimensional structures. *J Med Chem*, 47(12):2977-2980. https://doi.org/10.1021/jm030580l

Wilman W, Wróbel S, Bielska W, et al., 2022. Machine-designed biotherapeutics: opportunities, feasibility and advantages of deep learning in computational antibody discovery. *Brief Bioinform*, 23(4):bbac267. https://doi.org/10.1093/bib/bbac267

Wu ZH, Pan SR, Chen FW, et al., 2021. A comprehensive survey on graph neural networks. *IEEE Trans Neural Netw Learn Syst*, 32(1):4-24. https://doi.org/10.1109/TNNLS.2020.2978386

Young C, Lau AWY, Burnett DL, 2022. B cells in the balance: offsetting self-reactivity avoidance with protection against foreign. *Front Immunol*, 13:951385. https://doi.org/10.3389/fimmu.2022.951385

Zhao Y, He B, Xu F, et al., 2023. DeepAIR: a deep learning framework for effective integration of sequence and 3D structure to enable adaptive immune receptor analysis. *Sci Adv*, 9(32):eabo5128. https://doi.org/10.1126/sciadv.abo5128

Zheng W, Zhao WJ, Wu M, et al., 2020. Microbiota-targeted maternal antibodies protect neonates from enteric infection. *Nature*, 577(7791):543-548. https://doi.org/10.1038/s41586-019-1898-4

Zurawski DV, McLendon MK, 2020. Monoclonal antibodies as an antibacterial approach against bacterial pathogens. *Antibiotics*, 9(4):155. https://doi.org/10.3390/antibiotics9040155