Computational and Structural Biotechnology Journal 23 (2024) 2648–2660

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Mini-review

A comprehensive overview of recent advances in generative models for antibodies

Fanxu Meng [a,1], Na Zhou [b,c,1], Guangchun Hu [d], Ruotong Liu [b,c], Yuanyuan Zhang [a], Ming Jing [e,f,*], Qingzhen Hou [b,c,**]

[a] College of Chemical Engineering, Qingdao University of Science and Technology, Qingdao 266042, China
[b] Department of Biostatistics, School of Public Health, Cheeloo College of Medicine, Shandong University, Jinan 250100, China
[c] National Institute of Health Data Science of China, Shandong University, Jinan 250100, China
[d] School of Information Science and Engineering, University of Jinan, Jinan 250022, China
[e] Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250353, China
[f] Shandong Provincial Key Laboratory of Computer Networks, Shandong Fundamental Research Center for Computer Science, Jinan 250000, China

ARTICLE INFO

Keywords:
Antibody generative model
Therapeutic antibody
Deep learning
Protein sequence
Protein structure

ABSTRACT

Therapeutic antibodies are an important class of biopharmaceuticals. With the rapid development of deep learning methods and the increasing amount of antibody data, antibody generative models have made great progress recently. They aim to solve the antibody space searching problems and are widely incorporated into the antibody development process. Therefore, a comprehensive introduction to the development methods in this field is imperative. Here, we collected 34 representative antibody generative models published recently and all generative models can be divided into three categories: sequence-generating models, structure-generating models, and hybrid models, based on their principles and algorithms. We further studied their performance

Abbreviations: **AAR,**, Amino acid recovery; **AbGAN-LMG**, Language Model Guided Antibody Generative Adversarial Network; **ACE,**, Activity-specific cell-enrichment; **AC-SINS**, Affinity-capture self-interaction nanoparticle spectroscopy; **ADCC,**, Antibody-dependent cellular cytotoxicity; **ADCP,**, Antibody-dependent cellular phagocytosis; **ADP**, Antibody-dependent phagocytosis; **AI,**, Artificial Intelligence; **ALM**, Antibody language model; **APMixer**, Aligned Protein Mixer; **APR**, Average precision scores with recall; **AUC**, Area Under the Curve; **BALM**, Bio-Inspired Antibody Language Model; **BART**, Bidirectional and Auto-Regressive Transformers; **BCR**, B-cell receptor; **BERT**, Bidirectional Encoder Representations from Transformers; **BLAST**, Basic Local Alignment Search Tool; **BLI**, Biolayer Interferometry; **BSL2**, Biosafety Level 2; **CDC**, Complement-dependent cytotoxicity; **CDR**, Complementarity-determining region; **CDRH3**, The third CDR of heavy chain; **CDRL2,**, The second CDR of light chain; **CNN**, Convolutional neural network; **COVID-19,**, Coronavirus disease 2019; **CoV-AbDab**, The coronavirus antibody database; **EAGLE**, Epitope-specific Antibody Generation using Language model Embeddings; **EGNN**, Equivariant Graph Neural Network; **ELISA**, The enzyme-linked immunosorbent assay; **Fab,**, Fragment antigen-binding region; **FACS**, Fluorescence-activated cell sorting; **Fc,**, Fragment crystallizable; **FLAb**, Fitness Landscape for Antibodies; **Fv**, Variable fragment; **GAN,**, Generative adversarial network; **GGNN**, Geometric graph neural network; **HC**, Heavy chain; **HIC**, Hydrophobic interaction chromatography; **HIC-HPLC,**, Hydrophobic Interaction Chromatography-High Performance Liquid Chromatography; **HSRN,**, Hierarchical Structure Refinement Network; **hsVNA**, homogeneous surrogate Virus Neutralization Assay; **HTP**, Hierarchical training paradigm; **IEDB,**, Immune Epitope Database; **IgLM**, Immunoglobulin Language Model; **LC**, Light chain; **LSTM**, Long Short-Term Memory network; **mAb**, monoclonal Antibody; **MF-CNN**, Multi-fusion convolutional neural network; **ML**, Machine learning; **MLSA**, Multi-task learning model with Loop Specific Attention; **MNA,**, Microneutralization assay; **mRNA,**, Messenger RNA; **MSA**, Multiple Sequence Alignment; **Nb**, Nanobody; **NGS**, Next-generation sequencing; **NLL**, Negative Log-Likelihood; **NOS**, diffusioN Optimized Sampling; **OAS**, Observed Antibody Space; **OCD**, Orientational coordinate distance; **PALM**, Pre-training antibody language model; **PARA**, Pre-training with A Rational Approach; **PDB**, Protein Data Bank; **PLM**, Protein language model; **PNA**, Pseudotyped virus neutralization assay; **PRNT**, Plaque reduction neutralization test; **PR-AUC**, Precision-recall area under the curve; **RAbD,**, RosettaAntibodyDesign; **RF**, Random Forest; **RFdiffusion**, RoseTTAFold diffusion; **RMSD**, Root mean square deviation; **RNA,**, Ribonucleic acid; **ROC-AUC**, Area under the receiver operating characteristic; **RP-HPLC**, Reversed Phase High Performance Liquid Chromatography; **SAbDab**, Structural Antibody Database; **SACS**, Summary of Antibody Crystal Structures; **scFv**, Single-chain variable fragment; **scRMSD,**, Self-consistent RMSD; **SdAb-DB**, Single Domain Antibody Database; **SEC**, Size exclusion chromatography; **SPR**, Surface plasmon resonance; **TAP**, Therapeutic Antibody Profiler; **VH**, Heavy chain variable domain; **VL,**, Light chain variable domain.

* Corresponding author at: Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250353, China.
** Corresponding author at: Department of Biostatistics, School of Public Health, Cheeloo College of Medicine, Shandong University, Jinan 250100, China.
E-mail addresses: jingming@sdu.edu.cn (M. Jing), houqingzhen@sdu.edu.cn (Q. Hou).
[1] Equal contribution

https://doi.org/10.1016/j.csbj.2024.06.016
Received 31 March 2024; Received in revised form 15 June 2024; Accepted 18 June 2024
Available online 20 June 2024

and contributions to antibody sequence prediction, structure optimization, and affinity enhancement. Our manuscript will provide a comprehensive overview of the status of antibody generative models and also offer guidance for selecting different approaches.

1. Introduction

Antibodies are crucial immunoglobulins in the immune response and play an important role in identifying and neutralizing pathogens. In addition, antibodies also induce a variety of other effector functions, such as ADCP [1], ADCC [2] and CDC [3].

There are 177 antibody therapies approved according to The Antibody Society until 28 May 2024 [4]. One of the major challenges in therapeutic antibody discovery is to find diverse new sequences that provide high-quality candidates for antibody therapy. The antibody sequence space has been estimated to contain up to 10^18 molecules [5]. One traditional approach is animal immunization, which is time-consuming and requires molecular engineering for reasonable exploitability. In vitro display techniques (e.g., phage display) are an effective method for antibody discovery, with libraries created in the form of phage displays up to 10^11 unique molecules [6,7]. However, the searching space for antibody CDRs is so large that it is nearly impossible to experimentally test all possible sequences. Therefore, antibody development requires more efficient methods to thoroughly explore more possibilities.

Compared to experimental methods, the generative models can capture the distribution characteristics of the input data during the training process and infer new sequences based on this distribution, thereby quickly generating new candidate antibody sequences [8]. Furthermore, deep generative models can detect high-order amino acid interactions within antibodies as well as between antibodies and antigens, thereby enabling more efficient antibody design [9,10].

Recently, various deep generative models have been developed for antibody design, focusing on generating amino acid sequences of CDRs. For example, AntiBARTy utilizes its latent space guided by antibody-specific language models to aid antibody design, generating novel antibodies with improved properties [11]. IgLM creates synthetic libraries by redesigning variable-length antibody sequences [12]. EAGLE integrates sequential latent space diffusion and antigenic epitope amino acids into antibody sequence sampling to design epitope-specific antibodies [13].

In addition to generating antibody sequences, it is also important to design antibody structures with binding specificity. The modeling of CDRH3 is a key step in antibody modeling, because CDRH3 has significantly high variability in length, sequence and structure [14]. IG-VAE generates a complete antibody skeleton with a variant autoencoder that accurately reconstructs three-dimensional coordinates, torsion angles, and distance maps from the learned latent space [15]. Recently, IgFold developed specifically to predict antibody structures, combines pre-trained language models with graph networks to predict the atomic coordinates of antibody structures directly [16]. In addition, the diffusion model RFdiffusion based on RoseTTAFold structure prediction network can generate a variety of protein structures [17]. Chroma successfully applied Bayesian inference in protein design by the diffusion process, neural network architecture and low-temperature sampling algorithm [18], and conditionally guide the generation process based on the required properties and functions. AlphaFold 3 shows significant improvement in antibody-protein interaction prediction [19]. According to Abramson et al., AlphaFold-based methods are able to model the chemical and physical properties of molecular interactions without relying on MSA. HelixFold-Multimer is also skilled at predicting antigen-antibody and peptide-protein interfaces [20]. These methods are also suitable for studying the structure and sequence of CDRs.

Considering the rapid development of antibody generative models, it is now urgent to summarize the latest generative models. Aiming for this, we collected 34 antibody generative models released up to May 2024, and provided a comprehensive description of the principles, algorithms, and performance of each model. This paper outlines the current state of antibody generative models and provides new insights for the development and application of novel antibody generative approaches.

2. Overview of antibody generative models

A typical antibody molecule is Y-shaped, with two identical LC and two identical HC. Antibodies have two functional domains, namely Fab and Fc. Each antibody has two Fab domains and one Fc domain [21]. CDRs in Fab are responsible for antigen binding, and CDRH3 is crucial for antibody-antigen interaction. The framework regions within Fab constitute approximately 85 % of the variable region, displaying relatively conserved amino acid sequences [22]. They play a relatively minor role in antibody binding specificity, but serve as scaffolds for CDRs to maintain the overall structural stability [23]. In contrast, the sequence of Fc is relatively stable, which determines the effector functions of antibodies, such as activating immune cells, mediating cytotoxicity, or promoting immune response [24].

Comparing with predicting antibody properties by classification and regression models, the main goal of generative models is to learn the distribution from existing antibody sequence and structure data, and then use this distribution to generate new antibody sequences or optimize existing antibodies, rather than classifying the data or predicting continuous numerical outputs. Recent advances in antibody generative models include transformers [25], graph neural networks [26], diffusion models [27], etc. Together, these innovations provide more efficient methods to search sequence and structure spaces.

Fig. 1 shows the schematic diagram of the antibody generative models. The first step in building antibody generative models is to organize the dataset from a large repository and/or published papers for model training. Then, various methods are implemented to encode the antibody sequences or structures as the input of the generative models. After that, different algorithms are used to train and optimize the generative models to generate antibody sequences and structures. According to the type of output, we divide the generative models into three categories: sequence-generating models, structure-generating models, and hybrid models.

2.1. Databases

The first key aspect of developing a generative model is to select a large amount of high-quality antibody data. Here, we introduce a widely used database in antibody generative models and briefly discuss the impact of possible redundancy on sequence and structure spaces.

OAS specializes in providing cleaned, annotated antibody library data, and it currently normalizes more than 1 billion sequence data from 80 studies [28], covering different immune states, biological species and individuals [29]. Access OAS at https://opig.stats.ox.ac.uk/webapps/oas/.

SAbDab is dedicated to collecting, organizing and displaying all publicly available antibody structures and annotating them in a consistent manner [30]. As of May 16, 2024, SAbDab contains 8438 antibody structures and 1609 nanobody structures. The database can be accessed in https://opig.stats.ox.ac.uk/webapps/sabdab-sabpred/sabdab. AbDb summarizes the antibody structures in the PDB and collects information on the Fv region from these structures using SACS [31,32]. AbDb is able to handle a variety of antibody structures and multi-chain

antigen situations. Up to May 2024, AbDb contains 5025 whole antibodies, 251 light chain antibodies, and 700 heavy chain antibodies. AbDb can be accessed at http://www.abybank.org/abdb/. CoV-AbDab aims to integrate molecular information, metadata of antibodies and nanobodies that bind to betacoronaviruses such as SARS-CoV-2, SARS-CoV-1, and MERS-CoV [33]. Until February 2024, CoV-AbDab includes 12,916 entries, and it is available at https://opig.stats.ox.ac.uk/webapps/covabdab.

SdAb-DB is the first library providing single domain antibodies and related proteins. It aggregates data from protein databases, published scientific articles, and user submissions, providing researchers with a publicly available, manually curated single domain antibody data platform [34]. As of May 2024, SdAb-DB contains 1446 sequences, SdAb-DB can be accessed through https://sdab-db.ca/.

The antibody sequences or structures in the database might be highly similar or repeated. Redundant data affects the performance and query efficiency of the database. In addition, this redundancy can cause the learning model to overfit to repeated sequences or structural patterns, thereby reducing the generalization ability of the models and hindering their performance in generating new antibodies.

In sequence spaces, clustering algorithms (such as CD-HIT [35], MMseqs2 [36]) can be used to aggregate similar sequences into representative sequences to reduce redundancy. In addition, sequence alignment tools (such as Clustal Omega [37]) can be used to align sequences in the database and then filter based on similarity and relatedness to remove highly similar sequences. For example, Madsen et al. used the CD-HIT algorithm to cluster individual VH and VL sequences separately at a 95 % sequence identity cutoff, which is a common strategy for dealing with redundant antibody sequences [38].

In the structural spaces, structural alignment tools (such as MM-align [39], DALI [40]) are often used in combination to ensure that redundancy is reduced, thereby improving data quality and increasing model training efficiency. For example, to test the efficiency of ABodyBuilder [41], Abysis [42] and SAbDab [30] were based on a 99 % Fv sequence identity cutoff.

2.2. Antibody generative models

We summarized key information of 34 antibody generative models in Table 1. In general, there are three categories based on the output: sequence-generating models, structure-generating models, and hybrid models. The most representative generative models within each category are then described.

2.2.1. Sequence-generating Models

Antibody sequence generative models aim to predict amino acid sequences of antibodies. As shown in Table 1, over half of the generative models fall into the category of sequence-generating models. These models learn from known antibody sequences to understand and capture patterns in antibody sequences, including the relative positions and conservation of amino acid sequences. On the other hand, they use algorithms such as GAN to generate new antibody sequences, providing strong support for vaccine development, disease treatment, etc. [43,44].

2.2.1.1. Learning features of antibody sequences. Protein sequences can be viewed as a form of language. PLMs can capture key features and structural patterns in antibody sequences by learning from a large number of natural antibody sequences [45,46]. For example, they identify highly conserved amino acid residues and specific sequence motifs in CDRs, thus facilitating subsequent tasks related to sequence information, including tuning antibody affinity or specificity by changing specific residues. They accelerate the antibody design process, reduce reliance on natural antibodies, and provide a faster and more efficient method for learning sequence features.

Protein-specific pre-trained models have been used to extract latent features from protein sequences. For example, ProteinBERT uses the Transformer/BERT architecture and combines local (character level) and global (whole sequence level) representations to support multi-task processing. The performance of ProteinBERT is limited by the quality and scale of pre-training data. If the pre-training data is not rich or representative enough, it may affect the generalization ability of the model [47].

AbLang was trained using a large collection of antibody sequences from the OAS database. It trains two AbLang models for heavy chain and light chain sequences respectively. The main purpose of AbLang is to solve the problem of missing residues in antibody sequences, which is of great significance in antibody drug discovery [48]. For deletions in the first 15 positions, AbLang recovered about 96 % of the light chains and 98 % of the heavy chains accurately. It was able to process 100 sequences in 6.5 s using four Intel Core i7–10700 cores.

AntiBERTy is trained on OAS non-redundant sequences, totaling 558 M, to parse antibody sequences and help understand the affinity maturation process of antibodies [49]. AntiBERTa aims to provide contextual representations of BCR sequences. The AntiBERTa model was pre-trained on human antibody sequences consisting of unpaired heavy and light chains, and fine-tuned on sequence and structure data from SAbDab [50]. AntiBERTa performs well in predicting antigen binding sites with a precision of 0.738 and a recall of 0.722 in the test set.

PARA provides a latent representation of antibodies for relevant field studies, such as antibody structure prediction, affinity prediction, and novel antibody design [51]. In the CDR3 region reconstruction task, when the mask ratio is 70 %, PARA70-L can accurately restore approximately 21 % of the residues. BERT2DAb differs from previous

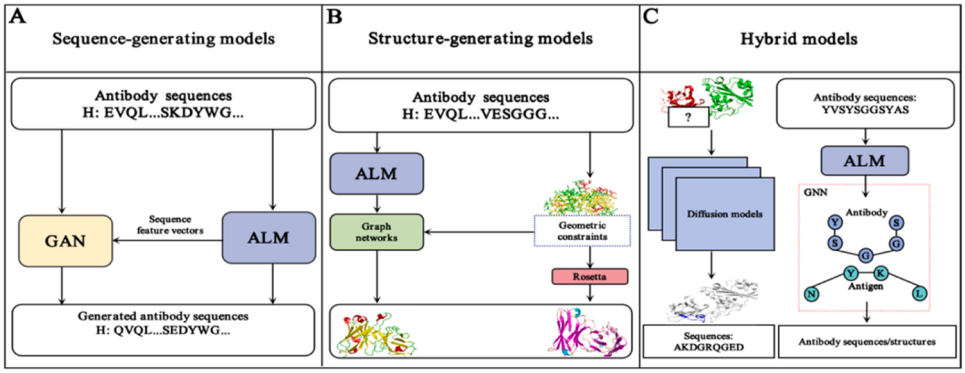

Fig. 1. Schematic diagrams of generative models for antibodies. (A) The sequence-generating models capture sequence features through ALM and generate new antibody sequences by neural network using these features. (B) The structure-generating models generate initial geometric constraints and predict structure using tools like Rosetta, or generate 3D structures of antibodies by ALM and Graph networks. (C) The hybrid models generate both antibody structures and sequences by diffusion models (adapted from Peng et al., 2023) and language models (adapted from Gao et al., 2023).

Table 1
Summary of generative models for antibodies.

Categories	Methods	Classifications	Input	Training data	Availability	References
Sequences	AB-Gen	Pre-trained transformer, reinforcement learning	Sequences	OAS	https://doi.org/10.5281/zenodo.7657016	[10]
	Venderley et al.	AntiBARTy, conditional diffusion model	Sequences	OAS, UniProtKB	/	[11]
	EAGLE	Diffusion model	Structures	SAbDab, OAS	/	[13]
	He et al.	PALM, Roformer	Sequences	OAS, Cov-AbDab, 14 H, 14 L, BioMap	https://github.com/TencentAILabHealthcare/PALM	[53]
	p-IgGen	Decoder language model	Sequences	OAS	/	[56]
	AbGAN-LMG	GAN	Sequences	CoV-AbDab	https://github.com/Zhaowenbin98/AbGAN-LMG	[57]
	Li et al.	Bayesian optimization, language model	Sequences	Pfam, OAS, Ab-14-H, Ab-14-L	https://github.com/AIforGreatGood/biotransfer	[62]
	AbMPNN	ProteinMPNN	Structures	SAbDab, OAS	/	[85]
	ReprogBert	Model reprogramming	Sequences	/	https://github.com/IBM/ReprogBERT	[136]
	IgBert, IgT5	Language Model	Sequences	OAS	https://huggingface.co/Exscientia	[143]
	IgDesign	IgMPNN, ESM2-3B	Structures	SAbDab	/	[145]
	Gruver et al.	NOS, LaMBO-2	Sequences	/	https://github.com/ngruver/NOS	[146]
	GLMAb-BO	PLM, batch bayesian optimization	Sequences	OAS	/	[147]
	Shanker et al.	Inverse folding protein language model	Structures	/	/	[148]
	AntiFold	Reverse folding model	Structures	SAbDab, OAS	https://opig.stats.ox.ac.uk/data/downloads/AntiFold	[149]
	pAbT5	Encoder-decoder	Sequences	OAS	/	[150]
	Hie et al.	Language model	Sequences	UniRef 90, UniRef 50	https://github.com/brianhie/efficient-evolution	[151]
	Frey et al.	Langevin Markov chain Monte Carlo	Sequences	OAS	/	[152]
Structures	IgFold	Language model, graph network	Structures	OAS, SAbDab	https://github.com/Graylab/IgFold	[16]
	Shanehsazzadeh et al.	MaskedDesign	Sequences, backbones	SAbDab	/	[60]
	BALMFold	Bio-inspired antibody language model, folding block	Sequences	OAS, SAbDab	https://github.com/BEAM-Labs/BALM	[69]
	AbMAP	Refine original PLM embedding	Sequences	SAbDab, LIBRA-seq	https://github.com/rs239/ablm	[71]
	MLSA	Multi-task learning with attention mechanism	Structures	SAbDab	/	[72]
	LoopGen	Score-based generative model	Structures, epitope	PDB 90, SAbDab	https://github.com/mgreenig/loopgen	[153]
Hybrid	AlphaPanda	transformer, 3DCNN, diffusion model	Structures	/	https://github.com/YueHuLab/AlphaPanda	[25]
	ABGNN	Language model, GNN	Sequences	OAS	https://github.com/KyGao/ABGNN	[26]
	AbDesign	AbDesign, AbDock	Sequences, backbones, epitope	PDB	https://github.com/pengzhangzhi/ab_opt	[27]
	AbDiffuser	Equivariant diffusion model	Structures	OAS, HER2 binder dataset (Mason et al., 2021)	/	[74]
	Villegas-Morcillo et al.	Diffusion probabilistic model	Structures	SAbDab	https://github.com/amelvim/antibody-diffusion-properties	[76]
	AntibodySGM	Score-based diffusion model	Structures	AbDb, SAbDab	/	[77]
	IgDiff	Diffusion Model	Structures	OAS	/	[83]
	ADesigner	Cross-Gate MLP, one-shot decoding	Structures	SAbDab	/	[154]
	PF-ABGen	Poisson flow generative model	Structures	SAbDab	/	[155]
	AbODE	Graphing partial differential equations	Structures	SAbDab	/	[156]

pre-trained models in that it not only considers the embedded representation of amino acids or k-mers in antibody sequences, but also explicitly considers the role of secondary structural features. This design enables BERT2DAb to better capture the features in antibody sequences, thereby achieving optimal performance in downstream tasks such as antibody screening and optimization [52]. On the Tra_dataset, the AUC of the BERT2DAb-based classifier is 0.90, and the average accuracy is 91.0 %; on the balanced Tra_dataset, the F1 score of the BERT2DAb-based classifier is 86.26 %, the recall rate is 87.41 %, and the accuracy rate is 85.46 %. The recall rate is nearly 1.5 % higher than AbLang model.

2.2.1.2. Generating antibody sequences. Understanding the characteristics of antibody sequences can provide valuable information for designing or predicting new antibodies with reasonable binding affinity and specificity for target antigens.

PALM proposed by He et al. is an antibody generation method based on a large-scale pre-trained language model, using the ESM2-based Antigen model as an encoder and Antibody Roformer as a decoder [53]. Roformer is a new position embedding method that adds explicit relative position dependencies to the self-attention mechanism to improve the performance of the Transformer architecture. This method combines absolute and relative position information into the Transformer model, thereby improving the performance of the model when processing long texts [54]. PALM can learn the characteristics and patterns of antibody sequences and generate antibodies with specific antigen binding specificity by performing large-scale pre-training. During the pre-training phase, the prediction accuracy of heavy chain Roformer

was 92.74 %, and the prediction accuracy of light chain Roformer was 94.14 %.

A2binder can be used to evaluate the binding affinity of PALM-generated antibodies. It uses a large-scale pre-trained model for sequence feature extraction, then uses a multi-scale feature fusion neural network for feature integration, and finally uses MF-CNN for affinity prediction. A2binder considers the features of antigen and antibody sequences, fuses them together for affinity prediction, and can make accurate and generalizable predictions even in the absence of known antigens. Due to insufficient paired data, the researchers had to train the models for antigens and antibodies separately instead of directly performing encoder-decoder pre-training. On the CoV-AbDab dataset, the ROC-AUC and PR-AUC of A2binder are 0.930 and 3.01, respectively. On the 14 H and 14 L datasets [55], the Spearman rank correlation coefficients of A2binder are 0.553 and 0.688, respectively.

AntiBARTy is based on BART and is used to guide the de novo design of IgG in antibody discovery. AntiBARTy is trained on large-scale antibody sequence data and is able to capture key features in the antibody sequence spaces [11]. In the samples used to assess protein solubility, more than 99.9 % of the synthesized sequences were classified as heavy chains and were unique.

IgLM regards antibody design as an autoregressive sequence generation task based on natural language text filling, which is used to redesign sequence-length agnostic sequences. The model was trained on approximately 558 M antibody heavy and light chain variant sequences, and conditioned on chain type and species of origin [12]. IgLM achieves an average infilling perplexity of 1.53 on the test set while maintaining random selection of spans. p-IgGen is a decoder-only language model designed to generate paired heavy and light chain sequences that is as diverse as natural sequences at a sampling temperature of 1.25 [56].

AbGAN-LMG is used to generate high-quality antibody libraries. Different from traditional GAN, it uses language models as input and combines the representation ability of language models to improve the effect of generating high-quality antibodies [57]. GANs inherently have a certain degree of randomness, which allows them to explore a wider space in the process of sequence generation [58]. However, the randomness of GAN often leads to a low probability of obtaining the desired sequences, which poses a challenge to enhancing GAN's ability to generate high-quality candidate antibodies. AbGAN-LMG improves the efficiency and quality of generating antibody libraries by integrating information from language models. The current model is only designed for antibody heavy chains, without considering the pairing information of light and heavy chains, which may limit its applicability in practical applications.

Recent studies have demonstrated innovative approaches to antibody discovery and optimization. Arras et al. used LSTM networks for sequence prediction based on NGS data. LSTM models are able to learn and capture complex patterns and dependencies in sequences. Most VHH SEEDbodies exhibited good biophysical properties after protein A purification, showing more than 90 % protein purity in SEC analysis [59]. Shanehsazzadeh et al. reported a process for therapeutic antibody discovery using a generative AI model, with an accuracy of approximately 60 % and a recall of $>$ 95 % for correctly classifying binding antibodies in the ACE assay. A total of 440,354 antibody variants targeting CDRH3 were designed and screened, and 421 designs were confirmed to bind to HER2, of which 71 designs had an affinity of less than 10 nM [60].

Bachas et al. describe a method for improving the binding affinity of antibodies to target antigens using deep contextual language models and high-throughput affinity data. The model can quantitatively predict the binding affinity of unseen antibody sequence variants, enabling virtual screening and expansion of accessible sequence spaces. This provides a more efficient and accurate method for selecting drug candidates. However, the performance of the model is limited by the training data. If the data quality is not high or the dataset is not comprehensive enough, the predictive ability of the model may be affected [61].

Li et al. introduced an end-to-end method based on Bayesian optimization and language model for designing large-scale high-affinity scFvs libraries. The model reduces time and cost by optimizing the antibody design process through ML methods. The pre-trained Spearman correlation coefficient of the heavy chain model was 0.51, and that of the light chain was 0.69. The Spearman correlation coefficient of the heavy chain ensemble model for the designed Ab-14-H variants is 0.69, while the Spearman correlation coefficient of the Gaussian process model for the designed Ab-14-H variants is − 0.42 [62].

2.2.2. *Structure-generating models*

The binding of antibodies to antigens is mainly mediated by CDRs. CDRs are constituted by short loop regions that typically lack standard helical or β-sheet structures and exhibit structural diversity. The spatial structure and geometric constraints of CDRs are crucial to ensure the binding properties of antibodies to antigens.

2.2.2.1. Predicting geometric constraints. By predicting the distances and orientation angles between residues, accurate constraints can be provided for antibody structure modeling, thereby improving the accuracy of CDRH3 loop predictions [63].

DeepAb focuses on predicting the Fv structure of antibodies from sequences. It introduces an interpretable attention mechanism that enables the network to focus on physically important residue pairs (e.g., proximal aromatic and key hydrogen bond interactions), which is a significant improvement over previous "black box" deep learning methods. It consists of two main stages, using a deep residual convolutional network to predict distances and orientation angles, and using Rosetta to implement structures [64]. The average OCD of the structures predicted by DeepAb was less than 4, indicating that the predicted structures were within one standard deviation. The average RMSD of the CDRH3 loop structures was 2.33 Å (± 1.32 Å) in the RosettaAntibody benchmark and 2.52 Å (± 1.50 Å) in the Therapeutic Antibody benchmark.

2.2.2.2. Generating atomic coordinates. Predicting the atomic coordinates of antibody structures can elucidate the mechanism of interaction between antibodies and antigens. By predicting the three-dimensional structure of antibodies, their functional regions can be revealed, providing key information for antibody design and optimization.

ABlooper leverages EGNN, a neural network that can handle rotation and translation invariance in 3D space. This allows ABlooper to learn directly from atomic coordinates, significantly improving the accuracy of its structure predictions [65]. ABlooper predicted 5 structures for each loop. The RMSD cutoff between predictions was set to 1.5 Å, meaning that if the average RMSD between the five predicted structures exceeded 1.5 Å, the predictions were considered inconsistent and therefore excluded. This process reduced the average RMSD value for the CDRH3 loop from the original 2.49 Å to 2.05 Å. In the automated antibody modeling workflow ABodyBuilder, the average backbone RMSD for CDRL2 was 0.5 Å and for CDRH3 was 1.9 Å. For most CDRs, ABlooper achieved lower average RMSDs than ABodyBuilder [66], and for all CDRs except CDRH3, ABlooper's loop prediction accuracy was comparable to AlphaFold2 [67] and DeepAb.

NanoNet is a deep learning end-to-end model designed to accurately predict the 3D coordinates of Nb structures. Compared to traditional antibody modeling methods, NanoNet directly predicts the 3D coordinates of the entire Nb backbone and Cβ atoms, rather than subdividing the modeling into framework and CDR regions. However, NanoNet currently ignores side chain representation, which may lead to less accurate structure predictions in some cases [68]. For the mAb test set, the average RMSD values for CDR1 and CDR2 were 0.88 Å and 0.95 Å, respectively, with an average RMSD of 2.38 Å for CDR3. For the

Nb test set, the average RMSD values for CDR1 and CDR2 were 2.65 Å and 1.73 Å, respectively, with an average RMSD of 3.16 Å for CDR3.

When combined with techniques such as graph networks, language model embeddings can serve as a powerful starting point for fast and accurate prediction of antibody structures. BALMFold is an end-to-end method derived from the BALM model for fast prediction of complete antibody structures from a single sequence. It bypasses traditional methods that rely on MSA, thereby improving the speed and accuracy of predictions [69]. When evaluating antigen binding ability, BALMFold achieved an APR of 92.9. The average prediction time of BALMFold was 5 s, which was faster than AlphaFold2 [67] and AlphaFold2-Multimer [70].

AbMAP is a transfer learning framework designed to fine-tune protein language models for the hypervariable regions of antibodies, thereby enabling accurate prediction of antibody structures and large-scale analysis of the human antibody repertoire [71]. IgFold uses the embedding of the pre-trained language model AntiBERTy on a dataset of 558 million natural antibody sequences, combined with a graph network to directly predict the main chain atomic coordinates of the antibody structure [16]. IgFold extracts the final hidden layer states and attention matrices from all layers of AntiBERTy. It uses residual embeddings (nodes) and residual attention features (edges) as input, and updates nodes and edges using graph transformation layers and triangular multiplication operations. In addition, IgFold provides error estimates for predictions, providing important information for evaluating the reliability of the prediction structure. For paired antibodies, IgFold takes an average of 23 s to predict the complete atomic structure. For nanobodies, it takes an average of 15 s to predict the complete atomic structure. For the CDRH3 loop, IgFold's average RMSD is between 2.01 Å and 2.34 Å.

2.2.2.3. Joint prediction of three CDR loops. Compared to CDRH1 and CDRH2 loops, the length variation range of CDRH3 loop is wide and the structure is diverse, so accurately predicting CDRH3 loop remains a challenge. Existing strategies usually try to predict each CDR structure individually rather than jointly. Therefore, multi-task learning can effectively explore the commonalities between CDRH123 loops, thereby improving model performance. The MLSA proposed by Giovanoudi et al. first utilized the similarities and differences in structure and function of CDR loops, and simultaneously learned the three loops through a multi-task learning strategy. In addition, the researchers proposed a loop specific attention mechanism to control the impact of each CDR loop on MLSA model training [72]. For the 49 targets in the RosettaAntibody benchmark set, this method achieved a resolution of 2.5 Å or higher, and the MLSA model improved by 19 % compared to the best baseline model RefineGNN [73]. In PyMOL, the RMSD of the CDRH3 loop prediction structure for PDB: 1NLB is 0.7459 Å.

The MLSA model architecture includes a shared layer and a loop specific attention mechanism. In the shared layer, the model encodes the sequence and three-dimensional coordinates of the antibody into a graph, where nodes represent amino acid residues and edges represent interactions between residues. The shared layer uses this information for prediction. In the loop specific attention mechanism, the model divides the node features into three subsets, corresponding to each CDR loop, and then selectively focuses on features related to the current loop. In this way, the model can balance the influence of each loop in joint learning, thereby improving the accuracy of CDR structure prediction. MLSA can be extended to perform multi-task learning not only on the CDR loops of the heavy chain, but also on the CDR loops of the light chain.

2.2.3. Hybrid generative models

Recent advances in generative methods have made it possible to co-design antibody sequences and structures, and these models can simulate antibody sequences and structures at different scales, from atomic-level details to overall structures. Here we focus on the application of language models and diffusion models in the co-design of antibody structures and sequences.

In order to maximize the use of antibody sequence data, Gao et al. designed a new sequence pre-training method, called ABGNN, to pre-train antibody language models. This method enhances the representation ability of antibodies and addresses the limitations of structural data. Compared with traditional autoregressive methods, ABGNN uses a one-shot approach to simultaneously generate all amino acids of CDR, which avoids the error accumulation problem in autoregressive methods and improves decoding efficiency [26]. The ABGNN framework consists of a pre-trained antibody language model AbBERT and a hierarchical graph neural network: Hseq for sequence encoding and Hstr for structure prediction. The RMSD of ABGNN in CDRH1/CDRH2 is less than 1.0 Å. In CDRH3 design, the AAR of ABGNN reaches 39.63 % and the RMSD is reduced to 1.56 Å.

Luo et al. introduced DiffAb, the first deep generative model based on diffusion probability model and equivariant neural network, which can model antibody sequence and structure at atomic resolution. It requires the structure of antibody framework and antigen as input to design CDR. This may limit its applicability in situations where the structure of the complex is unavailable [9]. Peng et al. introduced a diffusion model-based antibody optimization method that includes two key components: AbDesign for antibody design and AbDock for screening. AbDesign designs the optimal antibody sequences and structures; AbDock generates docking poses of the antibodies binding site with a specific antigen to screen potential antibody candidates [27]. In terms of structural design, AbDesign achieved an RMSD of 2.56 Å, which was more than 1 Å better than DiffAb [9] and 0.13 Å better than RefineGNN [73]. For the antibody optimization process, among the 19 designed 1G5.3 variants, MUT2, MUT5, MUT7, and MUT12 showed higher binding activity, especially MUT7, which had additional specificity for DENV-4 and JEV. In terms of virtual screening, AbDock achieved the highest DockQ score of 0.44.

AbDiffuser uses an equivariant and physics-informed diffusion model to deal with changes in sequence length, and gradually converts noise into a complete 3D structure by adding side chains and optimizing atomic positions [74]. AbDiffuser introduces a new equivariant neural network called APMixer. This innovative approach implicitly models the relationships between residues and significantly reduces memory usage. The RMSD of the structure created by AbDiffuser and the test set structure is 0.4962 Å.

The AlphaPanda algorithm combines the transformer model, 3DCNN model, and diffusion model. The transformer model captures global information, the 3DCNN model captures local structural features of antibody antigen complexes, and the diffusion model generates antibody sequences and structures [25]. Among the 500 antibodies generated by AlphaPanda, only two antibody structures had particularly large RMSD values. Compared with DiffAb, AlphaPanda's RMSD value was significantly lower, about 0.5 Å.

Bennett et al. used a fine-tuned RFdiffusion network to design nanobodies. This new method does not require the specification of protein structures in advance and can design binding proteins with shape complementarity based on user-specified epitopes. The researchers successfully designed nanobodies that bind to influenza hemagglutinin with an RMSD of 1.45 Å. However, the designed antibodies have a low binding affinity, indicating that this method needs further optimization and improvement [75].

Chroma can directly sample novel protein structures and sequences without relying on a library of known proteins. The model can be tuned during generation to meet user-specified property and functional requirements, such as symmetries and shape, or even natural language cues [18]. Villegas-Morcillo et al. used a diffusion probability model when designing CDRs related to antigen structures, while considering key properties such as solubility and folding stability. The flexibility of this property-conditional diffusion model provides a promising avenue

for computational antibody design in therapeutic applications. However, wet lab experiments are still needed to confirm its enhanced effects in terms of solubility and binding energy values [76].

Antibody-SGM employs a score-based generative diffusion model that enables iterative refinement of sequences and structures [77]. Most of the data generated by AntibodySGM have a sequence homology of more than 65 % with the real data. In addition, the Rosetta score of most of the generated structures is less than − 250 REU, indicating that the generated structures are reasonable and conflict-free.

In the field of nanoantibody engineering and antibody discovery, it also covers the application of methods such as directed mutagenesis technology, deep learning models, efficient screening methods and de novo design.

Bai et al. performed site-directed mutagenesis to enhance the affinity and stability of nanobodies for specific antigens. In addition, they designed bifunctional molecules by combining two nanobodies to simultaneously target two clinically relevant antigens, TNF-α and IL-23 [78]. HTP combines GGNN with PLM to fully exploit the evolutionary information in protein sequences and complex binding structures through multi-level training tasks [79]. However, the structural data of antibody-antigen complexes are still limited, and the number of structures in the SAbDab and RAbD [80] databases is significantly lower than the number required to achieve major breakthroughs in deep learning.

The deep screening method developed by Porebski et al. can screen about 10^8 antibody-antigen interactions in 3 days. The current implementation of this method mainly relies on the Illumina HiSeq platform. In addition, they also proposed BERT-DS, a natural language model based on the BERT architecture, which is pre-trained on the OAS database and then uses antigen-specific data directly obtained from deep screening experiments to generate new high-affinity antibody sequences. By constructing BERT-DS, they verified that language models pre-trained on large antibody sequence datasets can provide implicit representations of functional human antibody sequences [81].

Aguilar Rangel et al. described a fragment-based antibody design method that can design antibody CDR loops for pre-selected antigenic epitopes and graft them onto antibody frameworks. The effectiveness of this method depends on whether there are suitable CDR-like fragments in the database. The median coverage of the antigen surface using this design strategy is 78 %, and there are usually up to 19 design CDR candidates for each antigen surface region [82].

IgDiff is based on the existing SE(3) diffusion framework FrameDiff and is fine-tuned for antibody variable regions [83,84]. The training objective is to generate paired heavy and light chain backbone structures. The structures generated by IgDiff need to be predicted by an inverse folding model (e.g., AbMPNN) [85]. This means that the accuracy of the generated sequences is limited by the performance of the inverse folding model. If the inverse folding model performs poorly, the overall design performance will be affected. The scRMSD of the best predicted sequences for all generated antibodies is less than 2 Å, and 88 % of the generated antibodies have scRMSD less than 2 Å in all CDR loops. The average RMSD of the CDRH3 loop of the antibodies generated by IgDiff and the closest match in the training dataset is 1.39 Å ± 0.56 Å.

2.2.4. Evaluation Metrics

As shown in Table 2, the metrics for evaluating the generated antibodies and their example tools are described.

2.2.4.1. Sequences evaluation metrics.

- Diversity: The diversity of antibody sequences refers to the degree of variation at the sequence level. This diversity underlies the functional diversity of antibodies, enabling the immune system to recognize and neutralize a variety of antigens. Sequence alignment is used to identify similar regions between sequences. Sequence alignment can be roughly divided into pairwise sequence alignment and multiple sequence alignment. The former compares two sequences at a time, while the latter compares multiple sequences simultaneously. There are many popular sequence alignment tools, such as BLAST [86] for pairwise alignment and Clustal Omega [87] for multiple sequence alignment.

Table 2
Summary of common evaluation metrics.

Categories	Metrics	Methods
Sequences	Diversity	BLAST/Alignment tools
	Immunogenicity	BioPhi/OASis/IEDB/IEDB-AR/AbNatiV/Hu-mAb
	Solubility	Protein–Sol/Aggrescan4D/Solubis/CamSol/CamSol 3.0/SOLart
	Perplexity	Comparing language model prediction probabilities with actual data distribution
	Sequence recovery	Calculating the number of amino acid matches between the predicted and actual sequences
Structures	Visualization	PyMOL
	Homology	RMSD
	Dynamic properties（stability，interaction, etc.）	GROMACS/NAMD/LAMMPS
	Secondary structure	JPred/PSIPRED/PredictProtein/JPred4
	Binding ability	CSM-AB
	Stability	RosettaDDGPrediction/iStable 2.0/FoldX
Experiments	Binding affinity	SPR
	Detection and quantification	Virus Neutralization Assay
	Specificity	BLI/ELISA/FACS
	Detection of aggregates	AC-SINS
	Heterogeneity	icIEF/HIC-HPLC/RP-HPLC
Others	Developability	TAP
	Epitope prediction	EpitopeVec/EPMLR/SEMA/SeRenDIP-CE/HSRN
	Benchmark	FLAb

- Immunogenicity: Naturalness is generally considered to be related to immunogenicity risk. The naturalness metric is calculated by comparing the given antibody sequences with known antibody repertoires from humans. For example, The BioPhi platform can be used for humanized evaluation. Among them, OASis is calculated based on the sequences in the OAS database. The higher the OASis percentile, the greater the similarity of the antibody sequences to known human antibodies, indicating a higher degree of naturalness [88]. IEDB collects and organizes data related to immune epitopes. It is updated every two weeks to ensure consistency with the latest scientific literature [89]. IEDB-AR is a companion website of IEDB, which provides computational tools focused on predicting and analyzing B cell and T cell epitopes [90]. AbNatiV is able to evaluate the naturalness of Fv sequences from any source. It provides an interpretable score to predict immunogenicity and a residue-level naturalness distribution map [91]. Hu-mAb aims to optimize the humanization process of mAbs by suggesting mutations. Hu-mAb uses an RF classifier that is trained on large antibody sequence databases (such as OAS) and can accurately distinguish between human V genes and non-human variable region sequences [92].

- Solubility: High solubility is crucial for antibody therapeutics and research applications, as insoluble or aggregated antibodies can lead to reduced efficacy. Several computational methods have been developed to predict protein solubility. SOLart [93] and Protein–Sol [94] use machine learning techniques to predict solubility. Other methods use physicochemical descriptors of amino acid sequences to infer solubility, including Aggrescan4D [95] and Solubis [96]. An

extended version of CamSol, CamSol 3.0, has been able to accurately predict the solubility of proteins of various sizes (including nanobodies, full-length antibodies, and intrinsically disordered proteins) at different pH values [97].

- Perplexity/Sequence recovery: Perplexity is a metric used to evaluate the predictive performance of a language model, while natural sequence recovery refers to how many amino acids in the sequences generated by the model match the original sequences [12]. Perplexity is used to measure the difference between the predicted probabilities of a language model and the actual data distribution. It is the exponentiated average of the Negative Log-Likelihood (NLL). Generally, a lower perplexity indicates a better fit between the model and the data. When calculating sequence recovery, the sequences generated by the models can be compared with the true sequences and the edit distance between them can be calculated. A higher native sequence recovery rate indicates that the model is more capable of retaining the original sequence information.

2.2.4.2. Structures evaluation metrics.

- Visualization: Through visualization, we can better study the conformation of the antibody and understand how it binds to the antigen, such as by presenting the three-dimensional structure through PyMOL [98].

- Homology: Homology generally refers to the similarity between different antibody structures, which can be compared with the generated antibody structures known in SAbDab. Through structural alignment algorithms, such as TM-align, the three-dimensional coordinates of the two antibody structures are aligned, and the structural similarity score or RMSD value between them is calculated [99]. If the structures of two antibodies are very similar, their RMSD values will be low.

- Dynamic properties: Dynamic property studies can help evaluate the stability of antibodies under different temperature, pH and solvent conditions. In addition, they can help identify which domains in the antibody are critical to its function, thereby focusing on these areas during the design and modification process. Molecular dynamics is implemented in many software packages, including GROMACS [100], NAMD [101], and LAMMPS [102].

- Secondary structure: Secondary structure prediction provides key constraints for homology modeling and tertiary structure prediction methods, which helps to identify functional domains. Widely used methods include JPred [103], PSIPRED [104], and PredictProtein [105]. JPred4 is the latest version of the JPred protein secondary structure prediction server [106].

- Binding ability: CSM-AB is the first scoring function developed specifically for antibody-antigen docking and affinity prediction by modeling the interaction interface as graph-based signatures [107]. Guest et al. introduced an extended benchmark set for testing antibody-antigen docking and affinity prediction. This benchmark set contains 67 non-redundant antibody-antigen docking cases and 51 affinity cases [108].

- Stability: There are also algorithms and software for predicting stability, such as RosettaDDGPrediction [109], iStable 2.0 [110] and FoldX [111]. RosettaDDGPrediction uses the Rosetta modeling suite to predict changes in folding/unfolding free energies (ΔΔGs) caused by amino acid substitutions in protein monomers, as well as to calculate changes in binding free energies in protein complexes. iStable 2.0 integrates 11 sequence-based and structure-based protein stability prediction tools. It uses machine learning methods to integrate the results of these tools, constructing classification and regression models for structure and sequence inputs. FoldX is used to predict protein stability mutation sites to help enhance protein stability. This target-oriented mutation prediction helps reduce experimental workload and costs.

2.2.4.3. Experimental evaluation metrics.

- Binding affinity: SPR is used to evaluate the interaction between antibodies and targets. In SPR experiments, the target is usually fixed on the surface of the chip, and the antibody solution flows through the surface. The binding of antibodies to targets leads to changes in SPR signals, which can be monitored and provide detailed kinetic parameters such as binding constants and dissociation constants [112].

- Detection and quantification: Virus neutralization assays are primary tools for developing vaccines and therapeutic strategies. In this experiment, the virus is mixed with a certain concentration of antibodies and then co-cultured with host cells. If the antibodies have neutralizing capabilities, they will prevent the virus from invading host cells. The neutralizing effect of antibodies can be assessed by measuring the extent of host cell infection. Additionally, Bewley et al. proposed three methods for measuring SARS-CoV-2 neutralizing antibodies: PRNT, MNA, and PNA. MNA and PNA improve upon traditional PRNT by increasing experimental efficiency and throughput, with MNA using 96-well plates and immunostaining to increase automation and precision. PNA uses pseudotyped viruses and can be performed in BSL2 laboratories, which is safer compared to the BSL3 facilities required for PRNT and MNA [113]. Also, Neu-SATiN is an improved version of hsVNA used to measure antibody neutralizing activity directly from the sera of recovered patients or vaccinated patients, and the hands-on time of Neu-SATiN is less than 30 min. Neu-SATiN in its current form cannot measure other antiviral activities such as ADP or ADCC, and it is primarily suitable for vaccine evaluation based on the SARS-CoV-2 spike protein [114].

- Specificity: Specificity ensures that the antibody can accurately recognize and bind to the target antigen. BLI uses light interference technology to detect molecular interactions [115]. When the antibody binds to the target antigen, the interference pattern of light is changed, and the binding event is detected. ELISA is based on the specific binding of antibodies to antigens and enzymatic reactions. ELISA typically immobilizes antibodies on solid-phase carriers, then adds target antigens and binds to them through enzyme labeled secondary antibodies. By adding substrates for colorimetric reactions and measuring enzyme activity, the concentration of antigens or antibodies in the sample can be quantified through standard curves [116]. FACS uses fluorescently labeled antibodies to bind to targets on the cell surface, and then sorts these cells using flow cytometry [117].

- Detection of aggregates: In vivo, aggregates can be recognized by the immune system as foreign entities, leading to immunogenicity and adverse reactions. Additionally, aggregates may obscure the active sites of antibodies, thereby impairing their antigen-binding capability. AC-SINS utilizes gold nanoparticles to concentrate antibodies and detect their self-interactions. It is suitable for large-scale antibody screening and can screen thousands of antibodies in one day. Compared with traditional HIC, this method has higher throughput and lower sample requirements [118,119].

- Heterogeneity: During the production process, antibodies may undergo different forms of changes or modifications. For example, glycosylation affects antibody function and stability, antibodies may exist in different structural isomers, and may be affected by oxidation and degradation during storage. icIEF is an electrophoresis technique that can be used to separate and quantify variants with different isoelectric points, and can detect antibody isomers and glycosylation heterogeneity [120]. HIC-HPLC separates protein variants by utilizing hydrophobic interactions between proteins [121]. This method is very effective for detecting hydrophobic heterogeneity and isomers of antibodies. RP-HPLC is a commonly used high-performance liquid chromatography technique that can be used to separate and quantify different components in proteins and detect oxidation and degradation products of antibodies [122].

2.2.4.4. Others.

- Developability: Developability assessments must be performed early so that potential problems can be identified during the discovery phase. There are also methods that assess multiple aspects of developability, such as Raybould et al. updated the TAP (including the total length of the CDRs, the extent and magnitude of surface hydrophobicity, positive charge, negative charge, and asymmetry in the net HC and LC surface charges) tool and found a proportion of low-risk drug candidates in lambda antibodies [123,124]. Rosace et al. introduced a fully automated computational strategy to simultaneously optimize solubility and conformational stability [125]. The pipeline works by eliminating aggregation hotspots that cause poor solubility and proposes mutations that are expected to enhance conformational stability and solubility. This approach uses phylogenetic information to reduce false positive predictions and prevent modifications of functionally relevant sites. The process then relies on the CamSol method to predict solubility changes after mutation [126] and uses the FoldX energy function to predict associated stability changes [111].

- Epitope prediction: An antigen epitope is a specific region recognized and bound by an antibody. By predicting antigen epitopes, we can determine whether the generated antibody can effectively recognize and bind to the target antigen. This is crucial for developing highly specific and effective antibodies. There are currently several epitope prediction methods, such as: linear epitope prediction (EpitopeVec [127], EPMLR [128]); conformational epitope prediction (SEMA [129], SeRenDIP-CE [130], Sun et al. [131],); and conformational epitope prediction using antibody-antigen docking (HSRN [132]).

- Benchmark: The lack of synthetic datasets to compare the performance of different models is another obstacle in this field. Most antibody generative models use generated data and training databases for comparative analysis to evaluate model performance, or use the generated molecules for downstream experimental analysis. In addition, these models in the field use different training datasets, which complicates the determination of which model is best suited for a given application. There are currently some experimental datasets that can be used to evaluate model performance. For example, Marks et al. obtained an immunogenicity set of 217 therapeutic agents from clinical papers [92]. FLAb is the largest benchmark for therapeutic antibody design to date [133]. FLAb includes six properties of therapeutic antibodies: (1) expression, (2) thermal stability, (3) immunogenicity, (4) aggregation, (5) multireactivity, and (6) binding affinity. By evaluating the performance of existing pre-trained deep learning models (such as IgLM [12], AntiBERTy [49], ProtGPT2 [134], ProGen2 [135]) and the physics-based model Rosetta, it was found that no model performed well on datasets with all properties or similar properties. Guest et al. described an extended benchmark dataset for evaluating antibody antigen docking and affinity prediction. However, this still relies on the existence of experimental structures. For antibodies or antigens without experimental structures, the authors suggest using modeling techniques, which may introduce prediction errors [108].

3. Discussion

The complexity and high costs associated with the antibody design process are significant challenges. Generative AI is leading a transformation in this field. Antibody generative models can efficiently predict and generate highly specific antibody sequences and structures, thereby increasing the success rate of antibody design. The overall trend shows that antibody generative models are becoming more and more complex, and the application of pre-trained models in the field of antibody generation is receiving attention. The researchers used a multi-model integration approach to combine different models. For example, the antibody binding affinity optimization method proposed by Peng et al. includes two key models: AbDesign and AbDock, which are used to design and screen CDRs respectively [27]. In addition, pre-trained models are fine-tuned on specific tasks, which can accelerate model convergence and improve performance. For example, ReprogBert uses a pretrained English language model (BERT) to design CDR sequences for antibodies. ReprogBert performs well in data-scarce environments and requires only a small number of training parameters for efficient training [136]. Also, AbMAP fine-tunes PLM through transfer learning to accurately predict antibody structures [71].

At the same time, the emergence of graph neural networks is driving the antibody generation task from sequence-based methods to structure-based methods, ultimately achieving the fusion generation of sequence and structure. For example, RefineGNN solves the design of sequence and three-dimensional structure as a graph generation problem, which can model the conditional dependency between CDR and its context at the sequence and structure levels [73]. HTP combines geometric neural networks and large-scale protein language models to effectively mine the evolutionary information encoded in rich protein sequences and complex binding structures [79].

It is noteworthy that most antibody generative models ignore antigen information. This can lead to inaccurate or lack of specificity in the predicted antibody structure, as the interaction between the antibody and the antigen is crucial for binding specificity. For example, ignoring antigen information may limit model performance when optimizing antibody affinity against the novel COVID-19 virus variant XBB.1.5 [137]. Antigen information (e.g., antigen structure, antigen epitope, and biological properties of the antigen) should be used to improve the accuracy of antibody generative models.

Structure-based datasets provide static features of antibody-antigen binding, and epitope prediction lays the foundation for identifying binding sites. Initially identified antibodies are often weak binders, but their binding affinity to antigens can be enhanced through computational affinity maturation. Recent methods have predicted the effects of mutations on binding affinity, such as Kurumida et al. [138], AB-Bind [139], and mCSM-AB [140]. Some studies (e.g., DiffAb [9], Peng et al. [27]) have also discussed methods for co-designing antibody sequences and structures using antigen information, but they do not explicitly consider conditional epitopes.

Binding to more conserved epitopes may reduce the chance of viral escape and prolong the effectiveness of the antibody [113]. However, there is general consensus that antibodies can recognize almost any accessible surface region of an antigen and that the amino acid composition of epitope residues is often indistinguishable from other surface residues. The generation of conditionally specific antibodies against epitopes remains to be seen.

In addition, most generative models focus on the design of antibody heavy chains, while the pairing of antibody light and heavy chains is often ignored. Since VH and VL are encoded by different mRNA

transcripts, determining the natural pairing of VH-VL remains a challenge. The rules of VH-VL pairing have not been fully explored, and chain pairing is achieved by simple random pairing of VH and VL chains. Usually, generating heavy or light chains alone is not enough to reflect their heterodimeric properties and may lead to the loss of their co-evolutionary information. Therefore, current generative models may not accurately simulate the overall structure of antibodies. DeKosky et al. described a low-cost single-cell emulsion technology for sequencing the VH-VL recombinant components of antibodies [141]. The technology can process millions of cells in a short period of time with up to 97 % pairing accuracy. BALDR was used to accurately reconstruct paired heavy and light chain immunoglobulin gene sequences from Illumina single-cell RNA sequencing data [142].

From another perspective, the VH domain is the smallest active antibody fragment that can function as a single domain. These VH domains are ideal building blocks for a variety of antibody-based biologics and may help target epitopes that are difficult to access with traditional antibodies. Currently, there are several models that can independently generate new antibody VH or VL chains, such as IgLM [12]. However, none of these models can generate paired heavy and light chain sequences. It is known that some methods for generating antibody paired sequences have recently emerged, such as p-IgGen [56], IgBert, IgT5 [143], and IgDiff [83].

At a sampling temperature of 1.25, the sequences generated by p-IgGen are as diverse as natural sequences. IgBert and IgT5 perform well in processing paired and unpaired variable region sequences. They use more than 2 billion unpaired sequences and 2 million paired sequences to train these language models, which is much larger than the amount of training data in the past, providing the model with richer learning materials. The scRMSD of the best predicted sequences of all antibodies generated by IgDiff is less than 2 Å, and the average RMSD of the CDRH3 loop is 1.39 Å ± 0.56 Å.

Finally, developability considerations are critical for therapeutic advancement. The AntiBARTy Diffusion and Villegas-Morcillo et al. studies also considered developability factors such as solubility [11,76]. Harmalkar et al. used ML approaches to predict and enhance the thermal stability of scFvs. They described two ML approaches: one using a pre-trained language model and the other using a trained CNN. The study utilized temperature-specific data (TS50 measurements) extracted from multiple scFv sequence libraries. By training on this data, the model was able to extract features related to thermal stability and identify mutations that may increase thermal stability [144]. Future efforts should focus on finding a balance between increasing the binding affinity of antibodies for specific antigens and minimizing development-related risks.

4. Conclusion

In this paper, we have compiled 34 representative antibody generative models, providing an overview of their algorithms and applications. Additionally, we have summarized the limitations of these methods and potential strategies for improvement. Our study will serve as a guide for selecting diverse antibody generative approaches and offer valuable insights for the development of novel methods.

CRediT authorship contribution statement

Fanxu Meng: Writing – original draft, Conceptualization, Formal analysis. **Na Zhou:** Writing – original draft, Validation. **Qingzhen Hou:** Writing – review & editing, Supervision, Funding acquisition. **Guangchun Hu:** Investigation. **Ruotong Liu:** Visualization. **Yuanyuan Zhang:** Validation. **Ming Jing:** Writing – review & editing, Validation.

Declaration of Competing Interest

The authors declare no conflicts of interest.

Acknowledgements

QH was supported by Shandong Province Key R&D Program (2021SFGC0504) and the Young Scholars Program of Shandong University (21320082064101).

References

[1] Tay MZ, Wiehe K, Pollara J. Antibody-dependent cellular phagocytosis in antiviral immune responses. Front Immunol 2019;10:332.
[2] Iannello A, Ahmad A. Role of antibody-dependent cell-mediated cytotoxicity in the efficacy of therapeutic anti-cancer monoclonal antibodies. Cancer Metastas–Rev 2005;24:487–99.
[3] Quast I, Keller CW, Maurer MA, Giddens JP, Tackenberg B, et al. Sialylation of IgG Fc domain impairs complement-dependent cytotoxicity. J Clin Investig 2015;125(11):4160–70.
[4] The Antibody Society. (2024, May 28). Antibody therapeutics product data. Antibodysociety.org. https://www.antibodysociety.org/antibody-therapeutics-product-data/.
[5] Wilman W, Wróbel S, Bielska W, Deszynski P, Dudzic P, et al. Machine-designed biotherapeutics: opportunities, feasibility and advantages of deep learning in computational antibody discovery. Brief Bioinforma 2022;23(4). bbac267.
[6] Jafari N, Abediankenari S. Phage Particles as Vaccine Delivery Vehicles: Concepts, Applications and Prospects. Asian Pac J Cancer Prev 2015;16(18):8019–29.
[7] Almagro JC, Pedraza-Escalona M, Arrieta HI, Pérez-Tapia SM. Phage Display Libraries for Antibody Therapeutic Discovery and Development. Antibodies 2019;8(3):44.
[8] Mardikoraem M, Wang Z, Pascual N, Woldring D. Generative models for protein sequence modeling: recent advances and future directions. Brief Bioinforma 2023;24(6). bbad358.
[9] Luo S, Su Y, Peng X, Wang S, Peng J, et al. Antigen-Specific Antibody Design and Optimization with Diffusion-Based Generative Models for Protein Structures. Adv Neural Inf Process Syst 2022;35:9754–67.
[10] Xu X, Xu T, Zhou J, Liao X, Zhang R, et al. AB-Gen: Antibody Library Design with Generative Pre-trained Transformer and Deep Reinforcement Learning. Genom, Proteom Bioinforma 2023;21:1043–53.
[11] Venderley J. AntiBARTy Diffusion for Property Guided Antibody Design. arXiv preprint arXiv:2309.13129, 2023.
[12] Shuai RW, Ruffolo JA, Gray JJ. IgLM: Infilling language modeling for antibody sequence design. Cell Syst 2023;14(11):979–89.
[13] Cohen T, Schneidman-Duhovny D. Epitope-specific antibody design using diffusion models on the latent space of ESM embeddings. NeurIPS 2023 Gener AI Biol (GenBio) Workshop 2023.
[14] Vajda S, Porter KA, Kozakov D. Progress toward improved understanding of antibody maturation. Curr Opin Struct Biol 2021;67:226–31.
[15] Eguchi RR, Choe CA, Huang PS. Ig-VAE: Generative modeling of protein structure by direct 3D coordinate generation. PLoS Comput Biol 2022;18(6):e1010271.
[16] Ruffolo JA, Chu LS, Mahajan SP, Gray JJ. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. Nat Commun 2023;14(1):2389.
[17] Watson JL, Juergens D, Bennett NR, Trippe BL, Yim J, et al. De novo design of protein structure and function with RFdiffusion. Nature 2023;620(7976):1089–100.
[18] Ingraham JB, Baranov M, Costello Z, Barber KW, Wang W, et al. Illuminating protein space with a programmable generative model. Nature 2023;623(7989):1070–8.
[19] Abramson J, Adler J, Dunger J, Evans R, Green T, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 2024:1–3.
[20] Fang X., Gao J., Hu J., Liu L., Xue Y., et al. HelixFold-Multimer: Elevating Protein Complex Structure Prediction to New Heights. arXiv preprint arXiv:2404.10260, 2024.
[21] Khatib SE, Salla M. The mosaic puzzle of the therapeutic monoclonal antibodies and antibody fragments - A modular transition from full-length immunoglobulins to antibody mimetics. Leuk Res Rep 2022;18:100335.
[22] Mak TW, Saunders ME. B cell receptor structure and effector function. Immune Response 2006:93–120.
[23] Chiu ML, Goulet DR, Teplyakov A, Gilliland GL. Antibody Structure and Function: The Basis for Engineering Therapeutics. Antibodies 2019;8(4):55.
[24] Lobner E, Traxlmayr MW, Obinger C, Hasenhindl C. Engineered IgG1-Fc–one fragment to bind them all. Immunol Rev 2016;270(1):113–31.
[25] Hu Y., Tao F., Lan W., Zhang J. Combining transformer and 3DCNN models to achieve co-design of structures and sequences of antibodies in a diffusional manner. bioRxiv, 2024: 2024.04. 25.587828.
[26] Gao K, Wu L, Zhu J, Peng T, Xia Y, et al. Pre-training Antibody Language Models for Antigen-Specific Computational Antibody Design. Proc 29th ACM SIGKDD Conf Knowl Discov Data Min 2023:506–17.
[27] Peng Z., Han C., Wang X., Li D., Yuan F. Generative Diffusion Models for Antibody Design, Docking, and Optimization. bioRxiv, 2023: 2023.09. 25.559190.
[28] Olsen TH, Boyles F, Deane CM. Observed antibody space: a diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. Protein Sci 2022;31(1):141–6.

[29] Kovaltsuk A, Leem J, Kelm S, Snowden J, Deane CM, et al. Observed antibody space: a resource for data mining next-generation sequencing of antibody repertoires. J Immunol 2018;201(8):2502–9.
[30] Dunbar J, Krawczyk K, Leem J, Baker T, Fuchs A, et al. SAbDab: the structural antibody database. Nucleic Acids Res 2014;42(D1):D1140–6.
[31] Ferdous S., Martin A.C.R. AbDb: antibody structure database-a database of PDB-derived antibody structures. Database 2018;2018:bay040.
[32] Allcorn LC, Martin ACR. SACS–self-maintaining database of antibody crystal structure information. Bioinformatics 2002;18(1):175–81.
[33] Raybould MIJ, Kovaltsuk A, Marks C, Deane CM. CoV-AbDab: the coronavirus antibody database. Bioinformatics 2021;37(5):734–5.
[34] Wilton EE, Opyr MP, Kailasam S, Kothe RF, Wieden HJ, et al. sdAb-DB: the single domain antibody database. ACS Synth Biol 2018;7(11):2480–4.
[35] Chen Q., Wan Y., Lei Y., Zobel J., Verspoor K., et al. Evaluation of CD-HIT for constructing non-redundant databases. 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 2016:703–706.
[36] Mirdita M, Steinegger M, Söding J. MMseqs2 desktop and local web server app for fast, interactive sequence searches. Bioinformatics 2019;35(16):2856–8.
[37] Davies AM, Beavil RL, Barbolov M, Sandhar BS, Gould HJ, et al. Crystal structures of the human IgD Fab reveal insights into CH1 domain diversity. Mol Immunol 2023;159:28–37.
[38] Madsen AV, Mejias-Gomez O, Pedersen LE, Morth JP, Kristensen P, et al. Structural trends in antibody-antigen binding interfaces: a computational analysis of 1833 experimentally determined 3D structures. Comput Struct Biotechnol J 2023;23:199–211.
[39] Mukherjee S, Zhang Y. MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. Nucleic Acids Res 2009;37(11). e83-e83.
[40] Holm L, Laakso LM. Dali server update. Nucleic Acids Res 2016;44(W1):W351–5.
[41] Leem J, Dunbar J, Georges G, Shi J, Deane CM. ABodyBuilder: Automated antibody structure prediction with data–driven accuracy estimation. MAbs 2016;8(7):1259–68.
[42] Swindells MB, Porter CT, Couch M, Hurst J, Abhinandan KR, et al. abYsis: integrated antibody sequence and structure—management, analysis, and prediction. J Mol Biol 2017;429(3):356–64.
[43] Bauer J, Rajagopal N, Gupta P, Gupta P, Nixon AE, et al. How can we discover developable antibody-based biotherapeutics? Front Mol Biosci 2023;10:1221626.
[44] Akbar R, Bashour H, Rawat P, Robert PA, Smorodina E, et al. Progress and challenges for the machine learning-based design of fit-for-purpose monoclonal antibodies. MAbs 2022;14(1):2008790.
[45] Diaz DJ, Kulikova AV, Ellington AD, Wilke CO. Using machine learning to predict the effects and consequences of mutations in proteins. Curr Opin Struct Biol 2023;78:102518.
[46] Horne J, Shukla D. Recent advances in machine learning variant effect prediction tools for protein engineering. Ind Eng Chem Res 2022;61(19):6235–45.
[47] Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M. ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics 2022;38(8):2102–10.
[48] Olsen TH, Moal IH, Deane CM. AbLang: an antibody language model for completing antibody sequences. Bioinforma Adv 2022;2(1). vbac046.
[49] Ruffolo J.A., Gray J.J., Sulam J. Deciphering antibody affinity maturation with language models and weakly supervised learning. arXiv preprint arXiv:2112.07782, 2021.
[50] Leem J, Mitchell LS, Farmery JHR, Barton J, Galson JD. Deciphering the language of antibodies using self-supervised learning. Patterns 2022;3(7):100513.
[51] Gao X., Cao C., Lai L. Pre-training with A rational approach for antibody. bioRxiv, 2023: 2023.01. 19.524683.
[52] Luo X, Tong F, Zhao W, Zheng X, Li J, et al. BERT2DAb: a pre-trained model for antibody representation based on amino acid sequences and 2D-structure. MAbs 2023;15(1):2285904.
[53] He H, He B, Guan L, Zhao Y, Chen G, et al. De novo generation of antibody CDRH3 with a pre-trained generative large language model. bioRxiv 2023;2023.10. 17.562827.
[54] Su J., Lu Y., Pan S., Murtadha A., Wen B., et al. Roformer: Enhanced transformer with rotary position embedding. arXiv preprint arXiv:2104.09864, 2023.
[55] Engelhart E, Emerson R, Shing L, Lennartz C, Guion D, et al. A dataset comprised of binding interactions for 104,972 antibodies against a SARS-CoV-2 peptide. Sci Data 2022;9(1):653.
[56] Turnbull O.M., Oglic D., Deane C. p-IgGen: A Paired Antibody Generative Language Model. ICLR 2024 Workshop on Generative and Experimental Perspectives for Biomolecular Design.
[57] Zhao W, Luo X, Tong F, Zheng X, Li J, et al. Improving antibody optimization ability of generative adversarial network through large language model. Comput Struct Biotechnol J 2023;21:5839–50.
[58] Amimeur T., Shaver J.M., Ketchem R.R., Taylor J.A., Clark R.H., et al. Designing Feature-Controlled Humanoid Antibody Discovery Libraries Using Generative Adversarial Networks. bioRxiv, 2020: 2020.04. 12.024844.
[59] Arras P, Yoo HB, Pekar L, Clarke T, Friedrich L, et al. AI/ML combined with next-generation sequencing of VHH immune repertoires enables the rapid identification of de novo humanized and sequence-optimized single domain antibodies: a prospective case study. Front Mol Biosci 2023;10:1249247.
[60] Shanehsazzadeh A., McPartlon M., Kasun G., Steiger A.K., Sutton J.M., et al. Unlocking de novo antibody design with generative artificial intelligence. bioRxiv, 2023: 2023.01. 08.523187.
[61] Bachas S., Rakocevic G., Spencer D., Sastry A.V., Haile R., et al. Antibody optimization enabled by artificial intelligence predictions of binding affinity and naturalness. bioRxiv, 2022: 2022.08. 16.504181.
[62] Li L, Gupta E, Spaeth J, Shing L, Jaimes R, et al. Machine learning optimization of candidate antibody yields highly diverse sub-nanomolar affinity antibody libraries. Nat Commun 2023;14(1):3454.
[63] Weitzner BD, Gray JJ. Accurate Structure Prediction of CDR H3 Loops Enabled by a Novel Structure-Based C-Terminal Constraint. J Immunol 2017;198(1):505–15.
[64] Ruffolo JA, Sulam J, Gray JJ. Antibody structure prediction using interpretable deep learning. Patterns 2022;3(2):100406.
[65] Abanades B, Georges G, Bujotzek A, Deane CM. ABlooper: fast accurate antibody CDR loop structure prediction with accuracy estimation. Bioinformatics 2022;38(7):1877–80.
[66] Leem J, Deane CM. High-throughput antibody structure modeling and design using abodybuilder. Comput Methods Protein Evol 2019:367–80.
[67] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596(7873):583–9.
[68] Cohen T, Halfon M, Schneidman-Duhovny D. NanoNet: Rapid and accurate end-to-end nanobody modeling by deep learning. Front Immunol 2022;13:958584.
[69] Jing H., Gao Z., Xu S., Shen T., Peng Z., et al. Accurate Prediction of Antibody Function and Structure Using Bio-Inspired Antibody Language Model. bioRxiv, 2023: 2023.08. 30.555473.
[70] Evans R., O'Neill M., Pritzel A., Antropova N., Senior A., et al. Protein complex prediction with AlphaFold-Multimer. biorxiv, 2021: 2021.10. 04.463034.
[71] Singh R., Im C., Sorenson T., Qiu Y., Wendt M., et al. Learning the Language of Antibody Hypervariability. bioRxiv, 2023: 2023.04. 26.538476.
[72] Giovanoudi E., Rafailidis D. Multi-Task Learning with Loop Specific Attention for CDR Structure Prediction. arXiv preprint arXiv:2306.13045, 2023.
[73] Jin W., Wohlwend J., Barzilay R., Jaakkola T.S. Iterative refinement graph neural network for antibody sequence-structure co-design. arXiv preprint arXiv:2110.04624, 2021.
[74] Martinkus K., Ludwiczak J., Cho K., Liang W.C., Lafrance-Vanasse J., et al. AbDiffuser: Full-Atom Generation of in vitro Functioning Antibodies. arXiv preprint arXiv:2308.05027, 2023.
[75] Bennett N.R., Watson J.L., Ragotte R.J., Borst A.J., See D.L., et al. Atomically accurate de novo design of single-domain antibodies. bioRxiv, 2024: 2024.03. 14.585103.
[76] Villegas-MorcilloA., Weber J.M., Reinders MJT. Guiding diffusion models for antibody sequence and structure co-design with developability properties. bioRxiv, 2023: 2023.11. 22.568230.
[77] Xie X, Lee JS, Kim D, Jo J, Kim J, et al. Antibody-SGM: Antigen-Specific Joint Design of Antibody Sequence and Structure using Diffusion Models. 2023 ICML Workshop Comput Biol 2023.
[78] Bai Z, Wang J, Li J, Yuan H, Wang P, et al. Design of nanobody-based bispecific constructs by in silico affinity maturation and umbrella sampling simulations. Comput Struct Biotechnol J 2023;21:601–13.
[79] Wu F., Li S.Z. A Hierarchical Training Paradigm for Antibody Structure-sequence Co-design. arXiv preprint arXiv:2311.16126, 2023.
[80] Adolf-Bryfogle J, Kalyuzhniy O, Kubitz M, Weitzner BD, Hu X, et al. RosettaAntibodyDesign (RAbD): A general framework for computational antibody design. PLoS Comput Biol 2018;14(4):e1006112.
[81] Porebski BT, Balmforth M, Browne G, Riley A, Jamali K, et al. Rapid discovery of high-affinity antibodies via massively parallel sequencing, ribosome display and affinity screening. Nat Biomed Eng 2024;8(3):214–32.
[82] Aguilar Rangel M, Bedwell A, Costanzi E, Taylor RJ, Russo R, et al. Fragment-based computational design of antibodies targeting structured epitopes. Sci Adv 2022;8(45). eabp9540.
[83] Cutting D., Dreyer F.A., Errington D., Schneider C., Deane C.M., et al. De novo antibody design with SE (3) diffusion. arXiv preprint arXiv:2405.07622, 2024.
[84] Yim J., Trippe B.L., De Bortoli V., Mathieu E., Doucet A., et al. Se (3) diffusion model with application to protein backbone generation. arXiv preprint arXiv:2302.02277, 2023.
[85] Dreyer F.A., Cutting D., Schneider C., Kenlay H., Deane C.M., et al. Inverse folding for antibody sequence design using deep learning. arXiv preprint arXiv:2310.19513, 2023.
[86] Ye J, McGinnis S, Madden TL. BLAST: improvements for better sequence analysis. Nucleic Acids Res 2006;34(suppl_2):W6–9.
[87] Sievers F, Higgins DG. Clustal Omega for making accurate alignments of many protein sequences. Protein Sci 2018;27(1):135–45.
[88] Prihoda D, Maamary J, Waight A, Juan V, Fayadat-Dilman L, et al. BioPhi: A platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning. MAbs 2022;14(1):2020203.
[89] Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, et al. The immune epitope database (IEDB): 2018 update. Nucleic Acids Res 2019;47(D1):D339–43.
[90] Dhanda SK, Mahajan S, Paul S, Yan Z, Kim H, et al. IEDB-AR: immune epitope database—analysis resource in 2019. Nucleic Acids Res 2019;47(W1):W502–6.
[91] Ramon A, Ali M, Atkinson M, Saturnino A, Didi K, et al. Assessing antibody and nanobody nativeness for hit selection and humanization with AbNatiV. Nat Mach Intell 2024:1–18.
[92] Marks C, Hummer AM, Chin M, Deane CM. Humanization of antibodies using a machine learning approach on large-scale repertoire data. Bioinformatics 2021;37(22):4041–7.
[93] Hou Q, Kwasigroch JM, Rooman M, Pucci F. SOLart: a structure-based method to predict protein solubility and aggregation. Bioinformatics 2020;36(5):1445–52.

[94] Hebditch M, Carballo-Amador MA, Charonis S, Curtis R, Warwicker J. Protein–Sol: a web tool for predicting protein solubility from sequence. Bioinformatics 2017;33(19):3098–100.
[95] Bárcenas O, Kuriata A, Zalewski M, Iglesias V, Pintado-Grima C, et al. Aggrescan4D: structure-informed analysis of pH-dependent protein aggregation. Nucleic Acids Res 2024:gkae382.
[96] Van Durme J, De Baets G, Van Der Kant R, Ramakers M, Ganesan A, et al. Solubis: a webserver to reduce protein aggregation through mutation. Protein Eng, Des Sel 2016;29(8):285–9.
[97] Oeller M, Kang R, Bell R, Ausserwöger H, Sormanni P, et al. Sequence-based prediction of pH-dependent protein solubility using CamSol. Brief Bioinforma 2023;24(2). bbad004.
[98] Rosignoli S, Paiardini A. Boosting the full potential of PyMOL with structural biology plugins. Biomolecules 2022;12(12):1764.
[99] Cankara F., Tuncbag N., Gursoy A., Keskin O. Comparative Analysis of Structural Alignment Algorithms for Protein-Protein Interfaces in Template-Based Docking Studies. bioRxiv, 2024: 2024.04. 03.587755.
[100] Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX 2015;1:19–25.
[101] Phillips JC, Hardy DJ, Maia JDC, Stone JE, Ribeiro JV, et al. Scalable molecular dynamics on CPU and GPU architectures with NAMD. J Chem Phys 2020;153(4).
[102] Thompson AP, Aktulga HM, Berger R, Bolintineanu DS, Brown WM, et al. LAMMPS-a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. Comput Phys Commun 2022;271:108171.
[103] Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ. JPred: a consensus secondary structure prediction server. Bioinformatics 1998;14(10):892–3.
[104] Buchan DWA, Jones DT. The PSIPRED protein analysis workbench: 20 years on. Nucleic Acids Res 2019;47(W1). W402-W407.
[105] Bernhofer M, Dallago C, Karl T, Satagopam V, Heinzinger M, et al. PredictProtein-predicting protein structure and function for 29 years. Nucleic Acids Res 2021;49 (W1). W535-W540.
[106] Drozdetskiy A, Cole C, Procter J, Barton GJ. JPred4: a protein secondary structure prediction server. Nucleic Acids Res 2015;43(W1). W389-W394.
[107] Myung Y, Pires DEV, Ascher DB. CSM-AB: graph-based antibody-antigen binding affinity prediction and docking scoring function. Bioinformatics 2022;38(4): 1141–3.
[108] Guest JD, Vreven T, Zhou J, Moal I, Jeliazkov JR, et al. An expanded benchmark for antibody-antigen docking and affinity prediction reveals insights into antibody recognition determinants. Structure 2021;29(6):606–21. e5.
[109] Peccati F, Alunno-Rufini S, Jiménez-Osés G. Accurate prediction of enzyme thermostabilization with Rosetta using AlphaFold ensembles. J Chem Inf Model 2023;63(3):898–909.
[110] Chen CW, Lin MH, Liao CC, Chang HP, Chu YW. iStable 2.0: Predicting protein thermal stability changes by integrating various characteristic modules. Comput Struct Biotechnol J 2020;18:622–30.
[111] Buß O, Rudat J, Ochsenreither K. FoldX as protein engineering tool: better than random based approaches? Comput Struct Biotechnol J 2018;16:25–33.
[112] Davidoff SN, Ditto NT, Brooks AE, Josh Eckman J, Brooks BD. Surface plasmon resonance for therapeutic antibody characterization. Label-Free Biosens Methods Drug Discov 2015:35–76.
[113] Bewley KR, Coombes NS, Gagnon L, McInroy L, Baker N, et al. Quantification of SARS-CoV-2 neutralizing antibody by wild-type plaque reduction neutralization, microneutralization and pseudotyped virus neutralization assays. Nat Protoc 2021;16(6):3114–40.
[114] Kim SJ, Yao Z, Marsh MC, Eckert DM, Kay MS, et al. Homogeneous surrogate virus neutralization assay to rapidly assess neutralization activity of anti-SARS-CoV-2 antibodies. Nat Commun 2022;13(1):3716.
[115] Kamat V, Rafique A. Designing binding kinetic assay on the bio-layer interferometry (BLI) biosensor to characterize antibody-antigen interactions. Anal Biochem 2017;536:16–31.
[116] Aydin S. A short history, principles, and types of ELISA, and our laboratory experience with peptide/protein analyses using ELISA. Peptides 2015;72:4–15.
[117] Fei C, Nie L, Zhang J, Chen J. Potential applications of fluorescence-activated cell sorting (facs) and droplet-based microfluidics in promoting the discovery of specific antibodies for characterizations of fish immune cells. Front Immunol 2021;12:771231.
[118] Estep P, Caffry I, Yu Y, Sun T, Cao Y, et al. An alternative assay to hydrophobic interaction chromatography for high-throughput characterization of monoclonal antibodies. MAbs 2015;7(3):553–61.
[119] Liu Y, Caffry I, Wu J, Geng SB, Jain T, et al. High-throughput screening for developability during early-stage antibody discovery using self-interaction nanoparticle spectroscopy. MAbs 2014;6(2):483–92.
[120] Schlecht J, Moritz B, Kiessig S, Neusüß C. Characterization of therapeutic mAb charge heterogeneity by iCIEF coupled to mass spectrometry (iCIEF–MS). Electrophoresis 2023;44(5-6):540–8.
[121] Haverick M, Mengisen S, Shameem M, Ambrogelly A. Separation of mAbs molecular variants by analytical hydrophobic interaction chromatography HPLC: overview and applications. MAbs 2014;6(4):852–8.
[122] Kirthi A, Shanmugam R, Prathyusha MS, Basha DJ. A review on bioanalytical method development and validation by RP-HPLC. J Glob Trends Pharm Sci 2014; 5(4):2265–71.
[123] Raybould MIJ, Deane CM. The Therapeutic Antibody Profiler for Computational Developability Assessment. Methods Mol Biol 2022;2313:115–25.
[124] Raybould MIJ, Turnbull OM, Suter A, Guloglu B, Deane CM. Contextualising the developability risk of antibodies with lambda light chains using enhanced therapeutic antibody profiling. Commun Biol 2024;7(1):62.
[125] Rosace A, Bennett A, Oeller M, Mortensen MM, Sakhnini L, et al. Automated optimisation of solubility and conformational stability of antibodies and proteins. Nat Commun 2023;14(1):1937.
[126] Sormanni P, Vendruscolo M. Protein solubility predictions using the camsol method in the study of protein homeostasis. Cold Spring Harb Perspect Biol 2019; 11(12):a033845.
[127] Bahai A, Asgari E, Mofrad MRK, Kloetgen A, McHardy AC. EpitopeVec: linear epitope prediction using deep protein sequence embeddings. Bioinformatics 2021;37(23):4517–25.
[128] Lian Y, Ge M, Pan XM. EPMLR: sequence-based linear B-cell epitope prediction method using multiple linear regression. BMC Bioinforma 2014;15(1):414.
[129] Shashkova TI, Umerenkov D, Salnikov M, Strashnov PV, Konstantinova AV, et al. SEMA: Antigen B-cell conformational epitope prediction using deep transfer learning. Front Immunol 2022;13:960985.
[130] Hou Q, Stringer B, Waury K, Capel H, Haydarlou R, et al. SeRenDIP-CE: sequence-based interface prediction for conformational epitopes. Bioinformatics 2021;37 (20):3421–7.
[131] Sun P, Ju H, Zhang B, Gu Y, Liu B, et al. Conformational B-cell epitope prediction method based on antigen preprocessing and mimotopes analysis. BioMed Res Int 2015;2015:257030.
[132] Jin W, Barzilay R, Jaakkola T. Antibody-antigen docking and design via hierarchical structure refinement. Int Conf Mach Learn PMLR 2022:10217–27.
[133] Chungyoun M., Ruffolo J., Gray J. FLAb: Benchmarking deep learning methods for antibody fitness prediction. bioRxiv, 2024: 2024.01. 13.575504.
[134] Ferruz N, Schmidt S, Höcker B. ProtGPT2 is a deep unsupervised language model for protein design. Nat Commun 2022;13(1):4348.
[135] Nijkamp E, Ruffolo JA, Weinstein EN, Naik N, Madani A. Progen2: exploring the boundaries of protein language models. Cell Syst 2023;14(11):968–78. e3.
[136] Melnyk I, Chenthamarakshan V, Chen PY, Das P, Dhurandhar A, et al. Reprogramming Pretrained Language Models for Antibody Sequence Infilling. Int Conf Mach Learn 2023.
[137] Mannar D, Saville JW, Poloni C, Zhu X, Bezeruk A, et al. Altered receptor binding, antibody evasion and retention of T cell recognition by the SARS-CoV-2 XBB. 1.5 spike protein. Nat Commun 2024;15(1):1854.
[138] Kurumida Y, Saito Y, Kameda T. Predicting antibody affinity changes upon mutations by combining multiple predictors. Sci Rep 2020;10(1):19533.
[139] Sirin S, Apgar JR, Bennett EM, Keating AE. AB-Bind: Antibody binding mutational database for computational affinity predictions. Protein Sci 2016;25(2):393–409.
[140] Pires DEV, Ascher DB. mCSM-AB: a web server for predicting antibody-antigen affinity changes upon mutation with graph-based signatures. Nucleic Acids Res 2016;44(W1):W469–73.
[141] DeKosky BJ, Kojima T, Rodin A, Charab W, Ippolito GC, et al. In-depth determination and analysis of the human paired heavy-and light-chain antibody repertoire. Nat Med 2015;21(1):86–91.
[142] Upadhyay AA, Kauffman RC, Wolabaugh AN, Cho A, Patel NB, et al. BALDR: a computational pipeline for paired heavy and light chain immunoglobulin reconstruction in single-cell RNA-seq data. Genome Med 2018;10:1–18.
[143] Kenlay H., Dreyer F.A., Kovaltsuk A., Miketa D., Pires D., et al. Large scale paired antibody language models. arXiv preprint arXiv:2403.17889, 2024.
[144] Harmalkar A, Rao R, Richard Xie Y, Honer J, Deisting W, et al. Toward generalizable prediction of antibody thermostability using machine learning on sequence and structure features. MAbs 2023;15(1):2163584.
[145] Shanehsazzadeh A., Alverio J., Kasun G., Levine S., Khan J.A., et al. In vitro validated antibody design against multiple therapeutic antigens using generative inverse folding. bioRxiv, 2023: 2023.12. 08.570889.
[146] Gruver N., Stanton S., Frey N.C., Rudner TGJ, Hotzel I., et al. Protein Design with Guided Discrete Diffusion. arXiv preprint arXiv:2305.20009, 2023.
[147] Wang Y., Wang B., Shi T., Fu J., Zhou Y., et al. Sample-efficient Antibody Design through Protein Language Model for Risk-aware Batch Bayesian Optimization. bioRxiv, 2023: 2023.11. 06.565922.
[148] Shanker V.R., Bruun TUJ, Hie B.L., Kim P.S. Inverse folding of protein complexes with a structure-informed language model enables unsupervised antibody evolution. bioRxiv, 2023: 2023.12. 19.572475.
[149] Høie M., Hummer A., Olsen T., Nielsen M., Deane C. AntiFold: Improved antibody structure design using inverse folding. NeurIPS 2023 Generative AI and Biology (GenBio) Workshop, 2023.
[150] Chu SKS, Wei K.Y. Generative Antibody Design for Complementary Chain Pairing Sequences through Encoder-Decoder Language Model. arXiv preprint arXiv: 2301.02748, 2023.
[151] Hie BL, Shanker VR, Xu D, Bruun TUJ, Weidenbacher PA, et al. Efficient evolution of human antibodies from general protein language models. Nat Biotechnol 2024; 42(2):275–83.
[152] Frey N.C., Berenberg D., Kleinhenz J., Hotzel I., Lafrance-Vanasse J., et al. Learning protein family manifolds with smoothed energy-based models. ICLR 2023 Workshop on Physics for Machine Learning, 2023.
[153] Boom J.D., Greenig M., Sormanni P., Liò P. Score-Based Generative Models for Designing Binding Peptide Backbones. arXiv preprint arXiv:2310.07051, 2023.

[154] Tan C., Gao Z., Wu L., Xia J., Zheng J., et al. Protein Complex Invariant E60bedding with Cross-Gate MLP is A One-Shot Antibody Designer. arXiv preprint arXiv:2305.09480, 2023.
[155] Huang C., Liu Z., Bai S., Zhang L., Xu C., et al. PF-ABGen: A Reliable and Efficient Antibody Generator via Poisson Flow. ICLR 2023-Machine Learning for Drug Discovery workshop, 2023.
[156] Verma Y., Heinonen M., Garg V. AbODE: Ab Initio Antibody Design using Conjoined ODEs. arXiv preprint arXiv:2306.01005, 2023.