


Review

Computational and artificial intelligence-based methods for antibody development

Jisun Kim,^{1,4} Matthew McFee,^{2,4} Qiao Fang,^{2,4} Osama Abidin,² and Philip M. Kim ^{1,2,3,*}

Due to their high target specificity and binding affinity, therapeutic antibodies are currently the largest class of biotherapeutics. The traditional largely empirical antibody development process is, while mature and robust, cumbersome and has significant limitations. Substantial recent advances in computational and artificial intelligence (AI) technologies are now starting to overcome many of these limitations and are increasingly integrated into development pipelines. Here, we provide an overview of AI methods relevant for antibody development, including databases, computational predictors of antibody properties and structure, and computational antibody design methods with an emphasis on machine learning (ML) models, and the design of complementarity-determining region (CDR) loops, antibody structural components critical for binding.

The need for computational/AI-based methods in antibody development

Therapeutic antibodies are highly successful biotherapeutics and account for four of the ten top therapeutics by sales in 2021 [1]. Moreover, antibody-based biotherapeutics, such as antibody–drug conjugates, and bispecific antibodies, are also promising therapeutic modalities. Antibody discovery and development has traditionally been driven by experimental approaches [2], for instance, using **directed evolution** (see [Glossary](#)) processes through phage or yeast display techniques, or through immunization of animals. However, these approaches are time-consuming and laborious and have several limitations, including difficulties in specifying the antibody-binding side (**epitope**), as well as issues with obtaining antibodies that can be manufactured at scale ([Box 1](#)).

Although several strategies to optimize the experimental workflow have been reported [3,4], significant challenges remain. In recent years, computational/AI-based methodologies for antibody development are becoming more important for many parts of this workflow. This is analogous to small-molecule drug discovery, where computational methods have already made significant inroads [5,6]; prediction of drug–target interactions in particular is largely driven by the significant improvement in the performance of computational methods. Furthermore, new biotechnology companies with their own computational methods for antibody discovery are emerging. In this review, we cover recent advances made in computational and AI-based methods relevant for antibody development. In particular, we provide an overview of databases that have been amassed to allow for data-driven development of antibodies, predictors that have been developed that cover sequence, structure, and functional properties of antibodies, and computational models that leverage these databases and predictors to improve on experimental antibody development. We highlight the strengths and limitations of these approaches, as well as the steps needed to allow for practical application of these approaches to therapeutic antibody development.

Highlights

Recent advances in computational/artificial intelligence (AI)-based methodologies for antibody engineering and discovery hold great promise for accelerating and improving the development of therapeutic antibodies.

Databases hold large repertoires of antibody sequences, but only limited structural data; data on biophysical properties is also available.

A large suite of predictors of different biophysical and other properties of antibodies have been developed.

Deep learning approaches are improving the performance of structure prediction of antibodies, including CDRs, while *de novo* design remains a challenging problem.

Protein language models are showing very promising results for the improvement of antibody activity and properties.

¹Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto M5S 3E1, Canada

²Department of Molecular Genetics, University of Toronto, Toronto M5S 1A8, Canada

³Department of Computer Science, University of Toronto, Toronto M5S 2E4, Canada

⁴These authors contributed equally to this work.

*Correspondence: pi@kimlab.org (P.M. Kim).



Box 1. Limitations of traditional antibody engineering methods

Antibody binders to many targets can be obtained from existing well-established experimental approaches, such as immunization or directed evolution, but there still remain substantial limitations. First of all, it remains quite difficult to target specific epitopes on the antigen. Most screening campaigns only yield binders to one given epitope, which may not be the one needed to achieve the desired biological or therapeutic effect. Second, only a relatively small subset of antibodies has the biophysical and other properties needed to become therapeutics (e.g., solubility in high concentrations, manufacturability at large scale); these properties are referred to as 'developability' of an antibody. Most initial hits coming out of screening campaigns tend to have poor developability, leading to challenges later in their development. Moreover, once a candidate antibody binder has been obtained, a structure of the antibody/antigen complex is highly useful for further engineering, but requires either X-ray crystallography or NMR, which is typically time-consuming and low throughput.

Databases for antibody sequences and structures

For antibody development and engineering, antigen targeting ability and functional properties, including antigen binding **affinity**, target specificity, biological efficacy through epitope analysis, and **developability** properties are considered (Figure 1A). These abilities and properties are determined by antibody sequences and structures. Thus, information about antibody sequences, structures, and their associated properties has been informative to the design of novel antibodies. A number of databases of antibody sequences, structures, and their properties have been released, which enable the development of later computational methods (Table 1). Such databases are crucial to provide training data for **deep learning (DL)** models.

Sequence databases of antibody repertoires

In the human immune system, antibodies are produced by B cells and its repertoire is estimated at approximately 10^{13} unique sequences [7]. Antibodies are composed of two types of protein chains, known as the heavy chain (HC) and light chain (Figure 1B). Each of the chains is encoded by multiple gene segments (V, D, and J segments), which are spliced together using a V(D)J recombination process [7]. Through this process, a diverse range of antibody sequences can be produced. Large snapshots of this repertoire can now be obtained using **next-generation sequencing (NGS)** approaches. Efforts have been made to create standardized, publicly available repositories for these sequencing data [8,9]. These databases have provided researchers with easy access to a vast number of sequences and created opportunities for large-scale data mining. The observed antibody space (OAS) collates variable fragment (Fv) sequences and contains nearly 2 billion sequences spanning 68 different studies [8]. Several DL models have been trained on the OAS database as a means of discerning and generating humanized antibody sequences (details are described in the 'Computational design of antibodies' section).

Furthermore, some novel tools leverage these datasets for comprehensive analysis of antibody sequences [10,11] (Table 1). For example, AbDiver is a tool that employs the publicly curated B cell receptor NGS datasets and compares designed sequences to natural repertoires [10]. This tool facilitates the navigation of the vast antibody mutation space for the purpose of rational therapeutic antibody design and engineering. Moreover, the developer identified that it could find suitable profiles for 742 therapeutic antibodies. Another tool, the repertoire sequencing dataset analysis platform with an integrated antibody database (RAPID), includes more than 300 million clones extracted from human HC repertoire sequencing [11]. It additionally incorporates a large antibody database, including 521 therapeutic antibodies and 88 059 antibodies targeting specific antigens or arising in patients with particular diseases. Such curated antibody-antigen information could eventually lead to a purely sequence-based solution to the antibody-antigen problem (i.e., predicting which antigen any given antibody binds), but likely much larger datasets will be needed for this.

Glossary

Affinity: the interaction strength at which the antibody/drug is attracted to the target.

Artificial intelligence (AI): a theory and development of computer systems that can simulate human intelligence processes and perform complicated tasks. One of the branches of AI is machine learning (ML), which focuses on the use of data and algorithms to mimic the way that humans learn.

B and T cell epitopes: B cell epitopes are either the conformational or linear part of the protein antigen that antibodies bind to. T cell epitopes are usually protein antigen-derived peptides presented by MHC molecules on antigen-presenting cells and recognized by T cell receptors.

Complementarity-determining regions (CDRs): flexible loop regions of diversity in immunoglobulin variable domains that are mainly responsible for antigen binding. Both the VH and VL domains contain three CDR loops.

Deep learning (DL): a subclass of ML with a focus on mimicking human brain by training artificial deep neural networks to train on data.

Developability: the drug-like properties of therapeutic antibody candidates, including immunogenicity, solubility, stability, viscosity, charge profiles, post-translational modifications, pharmacokinetic and pharmacodynamic profiles, hydrophobicity, and manufacturability.

Directed evolution: a representative method used in protein/antibody engineering through iterative rounds of library screening. For this, phage display technology and yeast display technology are commonly used.

Epitope: the region of an antigen that binds to an antibody.

Humanization: humanization is required to produce antibodies with reduced immunogenicity risks. Antibody therapeutics are often produced from mice or other non-human organisms and therefore are likely to elicit an immunogenic response in humans. In order to reduce the immunogenicity, the variable region of a non-human antibody can be fused to human antibody constant region to generate a chimeric antibody.

Immunogenicity: an ability of therapeutics to trigger an undesirable immune response against the therapeutics.

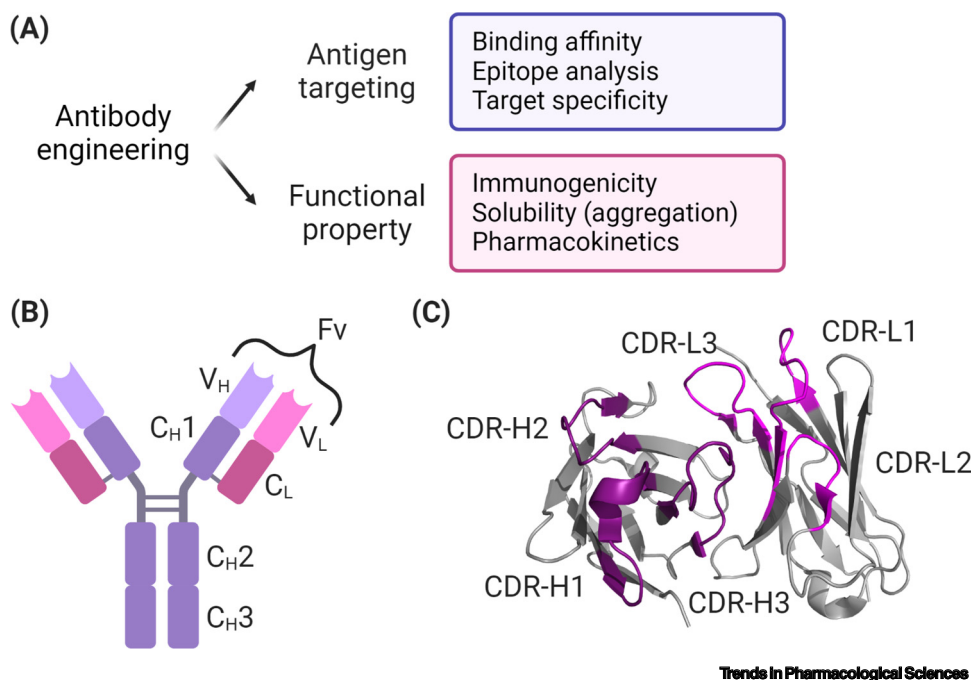


Figure 1. Antibody engineering for better efficacy and functional properties. (A) The antibody hits/candidates are usually further developed using protein engineering to enhance not only antigen binding affinity, target specificity, and biological efficacy through epitope analysis, but also developability properties, including immunogenicity, solubility, and pharmacokinetics, to ensure later manufacturability. (B) Schematic of antibody (immunoglobulin G). The heavy (H) chain of the antibody is depicted in purple, while the light (L) chain is shown in magenta. Both chains show labels C for constant region and V for variable region. The variable fragment (Fv) region is composed of the interaction between the variable domains of the heavy and light chains. (C) Representative arrangement of the complementarity-determining regions (CDRs) of the Fv region. The CDRs are composed of three respective loops of the variable chains of the heavy and light chains. The six CDRs can interact with antigens to mediate the specificity and potency of an antibody. CDR loops are highlighted and labeled (Protein Data Bank: 1N8Z)

Molecular dynamics (MD)

simulations: a computer simulation technique for analyzing the physical movements of each individual atoms or molecules. This technique permits the prediction of time evolution of an interacting particular system involving the generation of atomic trajectories of a system under different conditions, such as temperature, pH, and pressure.

Neural networks: mathematical models that consist of a network of connected nodes which compute a linear combination of inputs and then pass the result through a nonlinear activation function.

Next-generation sequencing (NGS): a massively parallel or deep sequencing technology that provides ultra-high throughput, large-scale, and high speed.

Paratope: the region of an antibody that recognizes and binds to the antigen.

Scoring functions: physical/mathematical models that attempt to describe the potential energy of a protein system.

Antibody structure databases

Antibody Fv domains consist of **CDRs**, highly variable sequences and frameworks of conserved sequences (Figure 1C). The six CDR loops, three on each of the variable domain of the heavy chain (VH) and the light chain (VL), are involved in antigen binding. Therefore, the 3D structure of the antibody determines how it interacts with an antigen and governs its binding properties. Researchers can use the information obtained from antibody structures to increase binding affinity, or to develop methods to predict the epitope and **paratope** [12–15]. Structure-based methods remain the most promising route for antibody design, hence structure databases are important to train and benchmark such models.

Antibody structures in the Protein Data Bank (PDB) have been extracted and compiled into various datasets [16]. Fv regions from the antibody have been updated in the Antibody Structure Database (AbDb) using information from the Summary of Antibody Crystal Structures (SACS) dataset [12]. Similarly, according to specific criteria, antibody structures from the PDB are listed in the Structural Antibody Database (SAbDab) [13], a database containing 12 367 Fv region structures as of June 2022, and abYsis [14]. Additionally, the Therapeutic Structural Antibody Database (Thera-SAbDab) is a therapeutic structural antibody database including antibody- and nanobody-related biotherapeutics [15]. As of July 2022, this database is tracking 748 unique therapeutics, including monoclonal antibody and bispecific therapeutics. It also provides additional metadata

Table 1. Current computational databases and tools for analysis and prediction of antibody structure, sequence, and properties

Name	Description	Refs
OAS	A sequence database containing 1.5 billion paired Fv and unpaired sequences from 80 studies	[8]
PIRD	A sequence database containing annotated T cell receptor and B cell receptor repertoire sequencing data	[9]
AbDiver	A bioinformatics tool that employs the B cell receptor sequencing data and compared query sequences to natural repertoires	[10]
RAPID	A bioinformatics tool that allows users to process and analyze human HC repertoire sequencing data and a large antibody database	[11]
AbDb	A structure database containing Fv region structures extracted from the RCSB PDB and SACS dataset	[12,16]
SAbDab	A bioinformatics resource containing all the publicly available antibody structures	[13]
abYsis	A bioinformatics tool including an integrated database of antibody sequence and structure data	[14]
Thera-SAbDab	A structure database including all antibody- and nanobody-related therapeutics recognized by the World Health Organization (WHO)	[15]
SKEMPI	A structural database containing more than 3000 binding free energy changes upon mutation from published literature	[17]
AB-Bind	A database including a diverse set of antibody binding data with experimentally determined binding free energy	[18]
Cov-AbDab	A focused repository for antibodies targeting SARS-CoV-2, SARS-CoV-1, and MERS-CoV	[19]
sdAb-DB	A focused repository for single-domain antibodies and related classes of proteins	[20]
TAP	A comprehensive platform including tools for evaluating five properties of antibody	[23]
Camsol	A predictor of protein solubility and potential sites to be modified for improvement through sequence- and structure-based algorithm	[25]
SOLart	A predictor of protein solubility and aggregation through structure-based algorithm	[26]
A3D	A predictor of aggregation-prone regions through sequence- and structure-based algorithms	[28]
IEDB-AR	A comprehensive platform including tools for prediction and analysis of immune epitopes	[31]
Hu-mAb	A classifier that can distinguish between human and non-human antibody Fv sequences	[32]
BioPhi	A comprehensive platform for antibody design, humanization, and humanness evaluation	[33]

for each therapeutic entry, such as clinical trial status, target antigen specificity, and companies involved in development.

Additional dedicated datasets summarize curated information about properties and structure of antibodies. These databases include the structural database of kinetics and energetics of mutant protein interactions (SKEMPI) [17], and the antibody binding mutational database (AB-Bind) [18]. Furthermore, there are more focused repositories of antibodies, such as CoV-AbDab [19] for anti-coronavirus antibodies, and sdAb-DB [20] for single-domain antibodies and related classes of proteins. Moreover, Antibodypedia provides validation information on commercially available antibodies, another useful resource for dataset curation [21]. These individual repositories could lead to development of specific DL models by providing training datasets, or for analysis of already designed antibody sequences. While these datasets remain relatively small on their own, these repositories can be used in conjunction with other datasets through techniques such as transfer learning (a process in which information learned from one task is applied to a related task, i.e., refining a model trained on data for one task with additional data from another).

Predictors of antibody properties

Apart from binding specificity and affinity, developability is crucial for the development of novel antibody therapeutics. Developability properties influence the likelihood that an antibody candidate can be advanced to clinical use, given proper efficacy. Several main properties comprise developability: intrinsic **immunogenicity**, aggregation/insolubility, viscosity, and half-life span. In order to reduce developability issues for industrial application, the properties of therapeutic antibodies need to be evaluated during early-stage development. Recently, computational tools or methods based on either classic statistics or ML have been developed and can be used to rapidly predict the developability of antibody candidates.

In general, the developability of an antibody can be mostly predicted by physicochemical properties of the amino acid sequence, such as hydrophobicity, electrostatic charge, and the interaction in their topology pattern [22]. For instance, therapeutic antibody profiling (TAP) models a set of post-Phase 1 clinical stage antibody therapeutics as distributions of five metrics thought to be linked to poor developability: the total length of CDRs, the extent of surface hydrophobicity, positive- and negative-charge in the CDRs, and asymmetry in the net heavy- and light-chain surface charges. TAP also provides guideline cut-offs for each metric and those cut-off values are collectively used for filtering out antibody candidates with poor developability [23]. In addition, there are some ML models that predict overall developability using antibody sequences or structures [14,24]. For example, Chen *et al.* [24] built an ML pipeline to predict antibody developability using a sequence dataset of 2400 antibodies from the SAbDab database.

While the aforementioned tools predict the overall developability, there are other tools that estimate one specific property for therapeutic antibody candidates.

Aggregation

Hydrophobicity can relate to aggregation propensity, solubility, viscosity, self-interaction, and protein stability, making it useful to predict potential downstream risks. Aggregation of antibody therapeutics can lead to precipitation and shortened storage period of drugs before administration, while aggregation *in vivo* can increase the immunogenicity of the drug [22].

To predict the solubility and aggregation propensity of protein, several prediction tools have been developed, such as the Camsol [25], and SOLart [26], with SOLart currently holding state-of-the-art performance. Furthermore, there are some ML models that predict aggregation propensity using antibody sequences or structures [27,28]. A structure-based aggregation prediction tool, AGGRESCAN 3D (A3D) [28], allows for the design of antibodies displaying significantly reduced aggregation propensity. The Trout group proposed an ML framework obtained from various models, including linear regression, support vector regression, and nearest neighbors' regression to predict antibodies' aggregation rate, which followed the form:

$$\text{Aggregation rate} = -0.34 \times \text{SCM positive}_{Fv} + 0.29 \times \text{SASA}_{CDRH2H3} + 0.84. \quad (1)$$

The group identified the spatial charge map (SCM) positive of the Fv region, which describes the sum of all the positive charge patches in the variable region, and surface accessible surface area (SASA) of the CDR-H2 and H3 as the two most important **molecular dynamics (MD) simulation** features in their model [29]. Such methods can be used to guide the engineering of

antibody candidates and avoid ‘developability traps’ (e.g., antibodies with very poor aggregation or viscosity properties).

Immunogenicity

Antibody therapeutics are often derived from mice or other non-human organisms, which are therefore potential epitopes recognized by B and T cells (**B and T cell epitopes**). Immunogenic response occurs when the immune system recognizes the B and T cell epitopes of antibody therapeutics, thereby eliciting the production of anti-drug antibodies. Immunogenicity negatively affects therapeutic efficacy and impacts safety through the induction of adverse drug reactions [30]. Therefore, predicting immunogenicity is a critical part of evaluating the clinical safety and efficacy of antibody therapeutics. As such, methods to predict immunogenic sequences and humanize them have been developed.

Classical techniques and AI-based computational tools to predict epitopes from proteins/antibodies have been developed [31–33]. These tools usually examine the primary sequences of antibodies to identify B and T cell epitopes. Immune epitope database analysis resource (IEDB-AR) [31] is a comprehensive website that provides several computational tools focused on the prediction and analysis of B and T cell epitopes.

Moreover, the Deane group used nearly 2 billion antibody sequences, available in OAS, across organisms where all human sequences were labeled as positive class, while others were labeled as negative class, and trained a random forest classifier model to produce a ‘human-ness score’. The random forest classifier consists of 200 decision trees and each decision tree individually separates the input sequences into classes (human or other species) using their features. The output score not only distinguishes human V genes from non-human sequences in the variable region but also indicates the level of immunogenicity risks [32]. Furthermore, the score was used to construct a **humanization** tool called HumAb, which suggests mutations to reduce immunogenicity risk [32]. Likewise, Merck® and the Bitton group trained a Transformer encoder-based language model called Sapiens to humanize sequences. Transformer architectures are natural language processing (NLP) models that employ the concept of attention. Here, each input token embedding (e.g., an amino acid) will receive updates from other tokens in the input, weighted by learned importance, the aforementioned attention. Please consult <http://nlp.seas.harvard.edu/annotated-transformer> for a fine-detailed introduction to Transformers and attention. They randomly masked or mutated amino acids in the natural human Fv sequences from OAS database and required the Sapiens to predict the original amino acids based on the remaining sequences. The Transformer used attention mechanisms to put different weights in different amino acids in the input sequence, revealing the different dependencies between residues and whole sequences. They provide an open-source platform, named BioPhi [33], incorporating the Transformer, and an evaluation method to aid antibody humanization.

Pharmacokinetic clearance

Pharmacokinetic (PK) clearance determines the half-life of an antibody and therefore determines how long an antibody drug can remain in the body after administration. Isoelectric point (pI), viscosity, immunogenicity, non-specificity, aggregation, and stability are factors related to clinical PK clearance rate [34]. To predict clearance rate through comprehensive evaluation of these factors, Grinshpun *et al.* [35] collected 64 mAbs that have been approved or are under Phase 2 and 3 clinical trials, and their published clearance data, and trained a random forest classifier to distinguish antibodies with different clearance levels. They found the pI is the most essential

feature to discriminate between fast and slow clearing antibodies. Also, Labute *et al.* [36] reported structure-based charge calculations to predict PK clearance rate.

Computational design of antibodies

Antibody structure modeling based on ML

Antibody structure information is critical to developing an understanding of the characteristic of designed antibodies, such as specificity and affinity, and so experimental prediction of the structures of designs is crucial for model development. For five of the six CDRs (H1, H2, L1, L2, and L3), structural diversity is limited and the loops typically follow canonical conformations; thus, these loops have significantly more structural constraints, allowing for easier prediction. Conversely, CDR-H3 loops have high conformational diversity, even in cases where sequence similarity is high [37], and so CDR-H3 modeling is a significantly less constrained, more difficult problem.

Many experimental antibody design techniques yield sequence information without associated structures. When performing large screens, independently determining all antibody loop structures and their interface with the antigen is impractical. Antibody structures have conventionally been predicted using either physics-based modeling, such as MD simulations [38], or homology-based modeling [39], or a combined method, such as MODELLER [40].

In contrast to physics-based modeling, computationally efficient DL models have also been developed for antibody structure modeling (Table 2). For example, ABodyBuilder [41], an automated antibody homology modeling pipeline, follows four steps: template selection, orientation prediction, CDR loop modeling, and side chain prediction. ABlooper [42] uses a graph **neural network**, which can directly work on 3D coordinate data from structure files, to predict the positions of all backbone atoms (Ca, N, C, and Cb) for the six CDR loops. In contrast, DeepAb [43], from the Gray group, predicts invariant features which are then used for reconstructing structures with Rosetta [44]. In the end, DeepAb achieved a prediction accuracy for H3 loop at root-mean-square deviation (RMSD) of 2.33 Å while ABlooper and ABodyBuilder achieved RMSD of 2.49 Å and 3.25 Å, respectively [41]. Due to DeepAb depending on Rosetta, its prediction is slow (10 min for one structure). The Gray group recently established a protein-language-based model called IgFold, which can predict structures faster (one structure under 1 min). Unlike the

Table 2. Antibody structure modeling methods

Name	Workflow	Refs
ABodyBuilder	1. Annotate and number antibody sequence 2. Select framework template with high homology 3. Model VH-VL orientation 4. Model CDR loop templates by <i>ab initio</i> modeling 5. Model side chains using SCWRL4 software	[41]
ABlooper	1. Input geometry for each CDR loop with its residues 2. Introduce it into graph neural network, named E(n)-Equivariant Graph Neural Networks [E(n)-EGNN], to predict the positions of all backbone atoms 3. Predict CDR loop structures	[42]
DeepAb	1. Adopt a recurrent neural network encoder-decoder model, where a bidirectional long short-term memory can learn features from sequences, which are evolutionary features and structural features. 2. Predict invariant features which are then used for reconstructing structures with Rosetta	[43]
IgFold	1. Convert antibody sequences into contextual embeddings using AntiBERTy 2. Predict atomic coordinates using a series of Transformer layers 3. Refine the predicted structures using Rosetta	[45]

ABlooper and DeepAb, the IgFold incorporated template structures, leading to more accurate predictions, especially on nanobody structures [45].

Although these results present a significant advancement of antibody structure modeling (largely driven by the adoption of DL techniques), more work is needed and the reliable modeling of an antibody/antigen complex structure remains an unsolved problem. CDR-H3 loop modeling remains challenging and improved architectures that better leverage domain-specific knowledge may prove to increase model prediction accuracy (i.e., adapting existing model architectures to incorporate biological knowledge about CDR loops). However, as previously mentioned, this is challenging as biologists have found CDR-H3 loops with very similar sequences adopting totally different structural conformations. AlphaFold2 [46] has proven to be revolutionary in predicting protein structure from sequence. It functions by taking in an input multiple sequence alignment (MSA) which provides important structural information via evolutionary history, as well as template PDB structures which provide further structural constraints to the model. AlphaFold2 extensively uses the attention mechanism that dominates NLP, as well as geometrically inspired attention variants such as triangle self-attention. Invariant point attention (IPA), also introduced in AlphaFold2, allows 3D atom coordinates in a global frame to be generated. Currently it struggles to predict the structures of orphan proteins, and MSA generation proves to be a major computational bottleneck. In the context of antibody design, because AlphaFold2 depends on MSA of homologous sequences, it is not applicable to CDR-H3 modeling due to limited antibody structures in available databases and the overwhelming diversity of CDR-H3 loops. Current work is attempting to generate structure from single sequence inputs alone to address the challenges of dependence on MSAs [47–50]. EquiFold [50] from Genentech® does well on antibody structure. AlphaFold2-Multimer [51] could also potentially be used to guide antibody design, by generating the complexed structure of input antibody and antigen sequences, with good binders presumably having very confident structural predictions. However, the performance of AlphaFold2-Multimer is still not optimal. In addition, Rosetta antibody package-based methods such as DeepAb are limited by the accuracy of their built-in **scoring functions**, mathematical models of the energy of protein systems. Finally, since binding to antigen could lead to conformational changes, it is critical to extend current models to include antigen information when predicting Fv or CDR-H3 loop structures.

DL has proven useful in aiding antibody screening. For example, DLAB [52] localizes the interface of an antibody and antigen and uses a 3D convolution-based neural network to generate binding scores. Limitations of these models include the necessary discretization (segmenting the protein structure into 3D grid of user-defined fineness, where each grid element may contain atom density information about different types, such as carbon and oxygen); since the correct fineness is not known *a priori*, there is the potential for significant information loss. Furthermore, dataset limitations include being dependent on a specific docking algorithm for data generation or having clonally related sequences in the dataset.

Structure-based DL models for design and analysis

To overcome the limitations of classic antibody design techniques (Box 2), recent efforts in applying ML to the design of antibodies, specifically CDR-H3 loops, typically fall into two categories (Table 3, Key table). First, there are models that attempt to design realistic backbones for CDR-H3s by generating 3D coordinates. For example, the IG-VAE model [53], which builds on previous work from the Huang group [54], generates complete backbones including all 3D coordinates by implementing a variational autoencoder trained to accurately reconstruct 3D coordinates, torsional angles, and distance maps from a learned latent space. Using random initializations in this space and gradient descent, the model is able to produce full 3D backbone structures of

Box 2. Classical antibody design techniques

Prior to the recent development of ML-based techniques for antibody design, more classical techniques involved the use of classical sampling and scoring methods. Two antibody design suites, each using different non-ML techniques, will be presented as examples of traditional structure-guided antibody design. RosettaAntibodyDesign (RABD) [89] relies on two algorithmic loops referred to the outer and inner loops of design. In the outer loop design, a CDR set to be designed has a structure sampled from a database of known structures and this loop is grafted in place of the existing loop. Multiple inner cycles are then conducted where amino acids are substituted at different positions in the antibody loops and an inner cycle is accepted or rejected based on the Metropolis criteria and an estimate in change of binding energy. Upon completion of the inner cycles, the final design is then compared energetically to the loop generated by the last cycle and again accepted or rejected according to the Metropolis criterion. Optimal method for antibody variable region engineering (OptMAVE) [90] involves a three-step workflow. First, using structural information of a database of 750 antibody–antigen complexes, potential positions of the antigen with respect to the antibody are selected. Secondly, the antibody is divided into six components, modular antibody parts (MAPs), each having their own reference database of structures. Each MAP in the antibody has its interaction energies with the antigen measured, the best combination being selected by linear programming before further refinement [91]. Nimrod *et al.* [92] proposed a strategy of re-epitoping [i.e., using similarity to epitopes of known antibody–antigen complexes for their antibody design strategy. Most of these methods require the generation of libraries of many candidates (~millions)], adopting highly parallel screening strategies that have been developed previously [93], reflecting relatively low positive predictive value of the scoring functions to predict candidate binding. However, recently, a strategy of grafting paratope fragments based on epitope similarity has been used to obtain nanomolar binders without parallel screening [94].

However, there are many limitations to these techniques, as they are dependent on databases of known structure, are computationally intensive (such as involving significant sampling), as well as being dependent on classical scoring functions, such as the Rosetta all-atom scoring function, which often have severe limitations [95]. With the explosion of DL research in the early 2010s, considerable effort has been used to shift towards applying DL models for the application of antibody design.

immunoglobulins with specified constraints. Shan *et al.* [55] developed severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) antibodies by redesigning CDR regions utilizing a model involving learning graph embeddings of protein structures to predict the change in $\Delta\Delta G$ upon introducing specific mutations in the loops. The drawback to backbone only generation is that they rely on external tools from Rosetta to design sequences that would fold into the specified structure. Furthermore, these structure-based models do not allow for conditioning on specific epitopes, which is also not desirable.

Sequence-based DL models design and analysis

In the second category, there are DL models that attempt to learn overall features of antibodies from their sequence alone (Table 3). At their core, such models learn interdependencies of antibody sequence; the rise of autoregressive or generative models then allowed for models that can generate ‘antibody-like’ sequences. In an early method, Shin *et al.* [56] applied causal dilated convolutions to learn long-range relationships in sequences in an autoregressive model (i.e., a model that predicts each amino acid one at a time, using the previous predictions as the next model input). Their trained model was used to generate libraries of nanobodies, fully designing the CDR1, CDR2, and CDR3 regions. Since then, there have been great advances in NLP due to developments of Transformer-based models such as BERT [57]. Such models were then applied to protein sequence: the BERT architecture was directly trained on the Uniref100 and BFD-100 datasets to create the ProtBERT model [58] for protein sequence prediction. Similarly, in ESM-1b [59], 250 million protein sequences were used to train a language model which learned embeddings that encoded important biological properties, which could be easily recovered using trained linear layers. Both ProtBERT and ESM-1b were highly successful at capturing many or most properties of protein sequences. The AbLang model [60] showed strong performance in predicting missing amino acids in the OAS database where 40% of the sequences are missing 15 or more amino acids. Akbar *et al.* [61] have shown promising performance in predicting binding affinity to 3D antigen structures using sequence information alone. The AntiBERTy model [62], an antibody specific BERT-based model, was trained on 558 million antibody sequences, with embeddings that cluster into directed evolution trajectories and the ability to detect paratope

Key table

Table 3. Deep learning models for antibody design

Model	Input type	Description	Refs
AbLang	Sequence	Predicts missing amino acids in antibody sequences, embeddings outperform other state-of-the-art models	[60]
Akbar <i>et al.</i>	Sequence	Generative model trained on sequences that can be used to design epitope-specific antibodies	[61]
Anand <i>et al.</i>	Sequence + structure	Diffusion-based model utilizing sequence and structure information	[75]
AntiBERTy	Sequence	Model used to embed antibody sequences into a low dimensional space revealing affinity maturation trajectories as well as detection and analysis of paratope residues	[62]
ESM-1b	Sequence	Transformer-based model that learns useful protein embeddings similar to BERT	[59]
Hie <i>et al.</i>	Sequence	Language-based model used to guide affinity maturation process for improved antibody binding	[64]
Ig-VAE	Structure	Generated novel antibody backbone structures with a variational autoencoder, design can be constrained with specified structural elements	[53]
IgLM	Sequence	Language-based model that redesigns specified regions of antibodies using an infilling technique	[63]
Jin <i>et al.</i>	Sequence + structure	Generates antibody CDR-H3 loops iteratively, predicting both sequence and all-atom structures	[69]
Mason <i>et al.</i>	Sequence	Model predicts binding allowing for screening of libraries designed with combinatorial mutagenesis	[65]
ProtBERT	Sequence	BERT architecture trained on large number of protein sequences which produces highly informative embeddings for multiple tasks	[58]
ProtDiff	Sequence + structure	Diffusion-based model allowing for the generation of protein backbones	[74,75]
Protein hallucinations	Sequence	Predicts protein backbone structure from sequence with design done using Monte Carlo sampling to maximize different from background probability distributions	[72]
Shan <i>et al.</i>	Structure	Predicts the ddG impact of mutations at any given point to guide antibody design	[55]
Shin <i>et al.</i>	Sequence	Dilated convolution based network applied to nanobody design	[56]

binding residues. Shuai *et al.* [63] used the Immunoglobulin Language Model (IgLM), to great success in generating immunoglobulin sequences. Hie *et al.* [64] showed that NLP models can be used to perform affinity maturation on even clinical grade antibodies, which was a surprising result as the affinity maturation could be performed without any explicit modeling of the antigen; in a sense they optimize the antibody by itself.

By contrast, another type of sequence-only approach does makes use of antigen-specific data: in a type of methodology that is starting to see adoption in industry, Mason *et al.* [65] presented a method to train a convolutional neural network on the results of a directed evolution experiment to generate improved libraries. A number of other related approaches that use ML models to learn sequence models from directed evolution experiments have been developed [66–68]. It is likely that future models would incorporate both features from language models and subsequently train on some sort of experimental data to obtain realistic antibodies that are optimized for antigen binding. However, at the moment, purely sequence-based models are either antigen agnostic or need outside experimental data and would benefit from incorporating structural information during the generation process.

DL models that utilize structure and sequence information as the future for ML-based antibody design

In addition to the classical work discussed earlier, more AI-based methods are starting to be developed. For instance, Jin *et al.* [69] developed a model that generates both the sequence and structure of a CDR-H3 loop in an autoregressive manner, similar to work conducted by the Huang group [70] on general protein sequence design. The next amino acid prediction is conditioned on a coarse representation of the antibody framework and the current amino acids in the

generated loop, including structure. The model generates probabilistic predictions for the next amino acid in the loop and can generate corresponding 3D atom coordinates for the next residue generated in the loop. They showed promising results in developing antibodies that can potentially interfere with SARS-CoV-1 and SARS-CoV-2. Although exciting, this model does not condition on a specific epitope, leaving room for further model improvement. A simpler approach was taken by Kang *et al.* [71] where the antibody and antigen amino acids are represented as nodes in a graph and interface structure is captured via edges between interface residue nodes. Although somewhat successful, this model showed a significant level of over-fitting to the training data. The future of CDR-H3 loop design will be in models that can fully generate sequence and structure while properly condition on target epitopes.

Incorporating AI protein design methods could improve antibody development

Finally, it is also important to look towards the developments in general protein design with DL to adapt successful techniques to the task of antibody design. There has also been recent success in applying the hallucination technique to general protein design [72]. In this methodology, random protein sequences are selected and then ‘folded’ into 3D structures (via a backbone alpha-carbon distance map prediction) using the existing ML model [73], providing a starting point, which then has mutations introduced via a Monte Carlo process in such a way that the protein becomes more realistic when compared with the folds of real proteins. Furthermore, diffusion probabilistic models have been applied with great success to the generation of protein backbone design [74,75]. In this framework, representations of proteins have noise applied until the protein representation is fully Gaussian noise. A DL model is then trained to undo this noising process, allowing for Gaussian noise to be sampled (a simple task) and converted into realistic protein structure. Very recently, RFDiffusion [76] and Chroma [77], both score-based generative models, have been able to generate all-atom coordinates for novel proteins, with Chroma creating proteins with several thousand amino acids. It is clear that these methodologies are applicable to the design of CDR regions and future work would involve implementing them.

Concluding remarks and future perspectives

We summarized current computational methods for therapeutic antibody development, applicable to the prediction of structures, epitopes, and addressing developability issues (Figure 2A,B). These approaches could not only improve developability, but lower costs, reduce labor, and support broader access to other biotherapeutics. For example, AlphaFold2 reduces weeks of work to hours or minutes. One illustration of the value of such computational tools has been drug discovery during the coronavirus disease 2019 (COVID-19) pandemic. Among research for COVID-19 drug discovery, many cases have been either exclusively computational or computer-aided experimental studies [78]. The improvement of computational approaches could be helpful to accelerate the development of novel drugs to resolve future pandemics.

An ideal computational antibody design platform would allow for the joint generation of antibody sequence and structure while conditioning on the target epitope. There has been progress towards this goal and there exist approaches that address subsets of the components of this problem [53–56,63–65,68–70]. Further progress will likely come through the continued development of structure-based DL approaches for protein design (see Outstanding questions). Improved solutions to related, more modular, problems in protein design continue to be developed. The problem of predicting a protein sequence that folds into a particular structure (inverse protein folding) has, for example, seen continuous improvements; state-of-the-art approaches integrate contemporary architectures that capture protein invariances while reasoning about protein residues in local and global coordinate frames [79–81]. This problem directly relates to the design of a CDR sequence given the backbone structure. Similarly, there has been progress

Outstanding questions

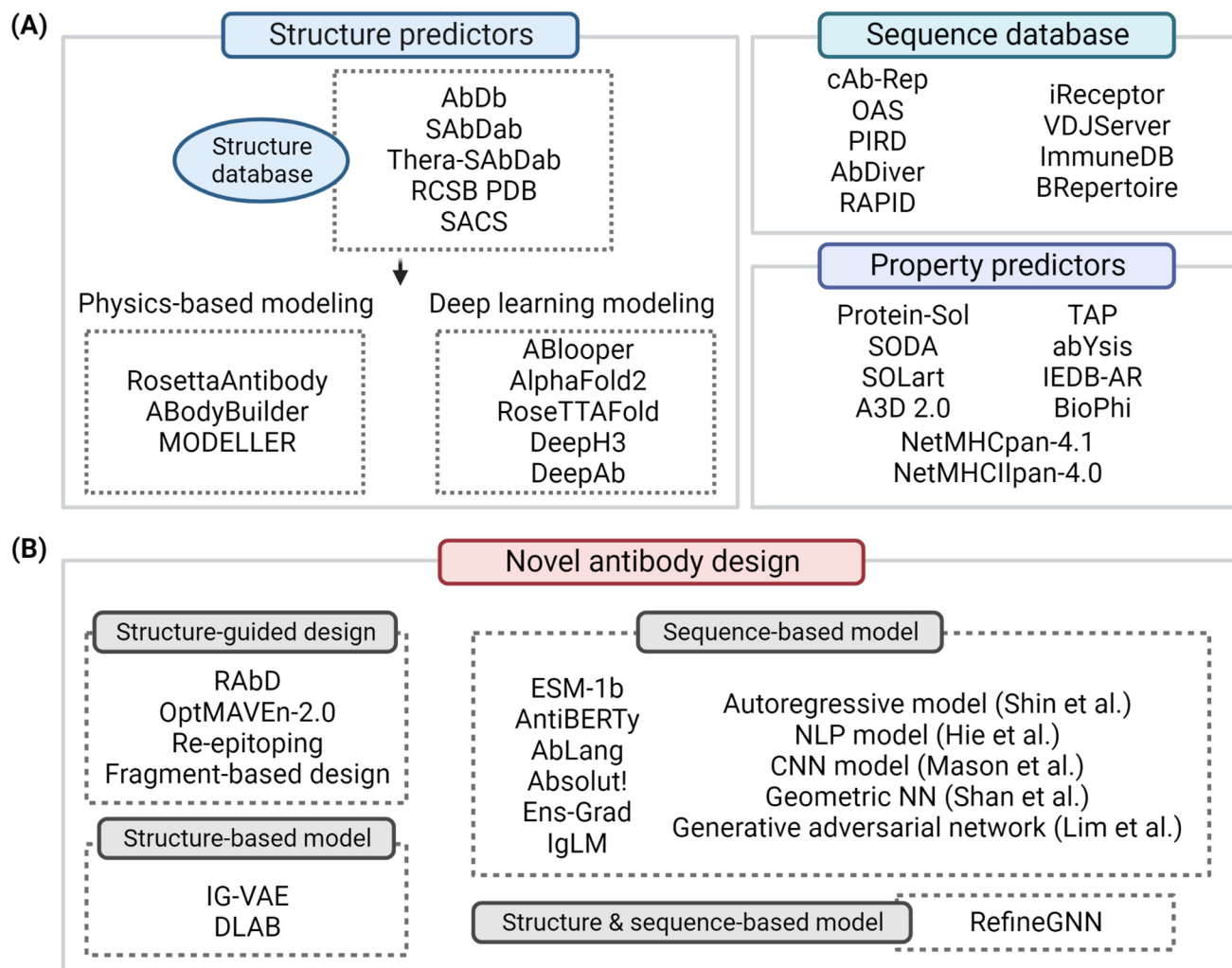
How can novel deep learning methods be translated beyond research to practical applications in the biopharma industry?

Will recent advances in deep learning-based approaches in general protein design result in commensurate improvements to therapeutic antibody design?

Can the landmark success of AlphaFold2 in protein structure prediction be leveraged for antibody modeling and design?

Can existing deep learning methods addressing subsets of the antibody design problem be consolidated to an effective antibody engineering platform?

What should be improved in current approaches to efficiently integrate deep learning techniques?



Trends in Pharmacological Sciences

Figure 2. Recently developed computational methods for antibody design. (A) Overview of current computational tool for analysis and prediction of antibody structure, sequence, and properties. The antibody structures are predicted through physics-based or deep learning modeling methods based on structure databases. Sequence databases provide the next-generation sequencing (NGS) data of antibody repertoire or repository of published antibody sequences and related information. Finally, several antibody property predictors have been released. (B) Novel antibody design via deep learning method based on the database/tools. These methods include sequence-based models, structure-based models, and structure and sequence-based models. Abbreviations: AbDb, Antibody Structure Database; CNN, convolutional neural network; IgLM, Immunoglobulin Language Model; NLP, natural language processing; NN, neural network; OAS, observed antibody space; PDB, Protein Data Bank; RAbD, RosettaAntibodyDesign; SAbDab, Structural Antibody Database; SACS, Summary of Antibody Crystal Structures; TAP, therapeutic antibody profiling; Thera-SAbDab, Therapeutic Structural Antibody Database. See [55,56,64–66].

in the use of DL for the complementary problem of unconditional, and constrained, protein backbone generation. These approaches have notably benefited from the development of modern generative frameworks, namely diffusion, and are directly applicable to loop generation in an antibody framework [74,75]. Very recently, score-based generative models (a modern formulation of diffusion models), have been shown to perform well on generating novel protein folds [82] and it seems likely that they could be also applied towards antibody generation.

DL-based protein design methods have additionally benefited from the paradigm-shifting work of AlphaFold2 [46] in protein structure prediction. Using AlphaFold2 for data augmentation, where

AlphaFold2 predictions are used as additional training data, has contributed to improved inverse protein folding [79]. The competing RoseTTAFold approach has allowed for *de novo* design through network hallucination by training a similar model specially for inpainting portions of proteins given context [72,83]. This type of guided search, based on Markov chain Monte Carlo methods or gradient updates, is often used for conditional generation [84] and can be used to guide antibody design in structure-based generative frameworks. More recently, the IPA module introduced by AlphaFold2 was adapted for use in a generative model for protein structure and sequence design [75]. IPA is just one of the many architectural innovations introduced by AlphaFold2. Future work that directly incorporates these ideas into generative contexts has potential to further advance the performance of DL-based protein engineering.

In terms of a useful future direction, it may also be important to include codon optimization to finalize designs. Codon optimization can not only result in an improvement in protein expression yield, but also antibody assembly, affinity, and biological activity [85,86]. Novimmune® optimized the assembly of bispecific antibodies, through codon optimization [87]. Furthermore, Rosenberg *et al.* [88] showed the mathematical calculation of the protein structure coordinates depends on the codon usage. Based on this approach, the DL method could be developed.

It remains to be seen if the various advances underlying protein structure and sequence-based design can be consolidated into an end-to-end differentiable pipeline for epitope-conditioned antibody generation without degradation in the performance of the individual components. This would contrast the conventional approach of integrating unconstrained generation with a binding predictor and extensive searches of sequence and structure space. Even with guided search strategies, it is likely that directly generating antibodies in the context of the epitope will allow for more efficient identification of binders. This is especially important when enforcing additional developability constraints that are necessary to mediate effective therapeutic development. Given the rapidly changing landscape of DL in structural and protein biology, it is likely that the coming years will see large progress towards this goal of a comprehensive antibody development platform.

Acknowledgments

P.M.K. acknowledges funding from the Canadian Institute of Health Research (PJT-159750 and PJT-153279).

Declaration of interests

P.M.K. is a cofounder and consultant to multiple companies, including Oracle Therapeutics, TBG Therapeutics, and Zymedi. O.A. and M.M. consult for Oracle Therapeutics.

References

1. Urquhart, L. (2022) Top companies and drugs by sales in 2021. *Nat. Rev. Drug Discov.* 21, 251
2. Kandari, D. and Bhatnagar, R. (2021) Antibody engineering and its therapeutic applications. *Int. Rev. Immunol.* Published online August 6, 2021. <https://doi.org/10.1080/08830185.2021.1960986>
3. Goydel, R.S. *et al.* (2020) Affinity maturation, humanization, and co-crystallization of a rabbit anti-human ROR2 monoclonal antibody for therapeutic applications. *J. Biol. Chem.* 295, 5995–6006
4. Perveen, R. *et al.* (2021) A rapid novel strategy for screening of antibody phage libraries for production, purification, and functional characterization of amber stop codons containing single-chain antibody fragments. *Biotechnol. Prog.* 37, e3136
5. Wan, F. *et al.* (2019) DeepCPI: a deep learning-based framework for large-scale in silico drug screening. *Genomics Proteomics Bioinformatics* 17, 478–495
6. Peng, J. *et al.* (2020) A learning-based method for drug-target interaction prediction based on feature representation learning and deep neural network. *BMC Bioinformatics* 21, 394
7. Rees, A.R. (2020) Understanding the human antibody repertoire. *mAbs* 12, 1729683
8. Olsen, T.H. *et al.* (2022) Observed antibody space: a diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Sci.* 31, 141–146
9. Zhang, W. *et al.* (2020) PIRD: pan immune repertoire database. *Bioinformatics* 36, 897–903
10. Młokosiewicz, J. *et al.* (2022) AbDiver—a tool to explore the natural antibody landscape to aid therapeutic design. *Bioinformatics* 38, 2628–2630
11. Zhang, Y. *et al.* (2021) RAPID: a rep-seq dataset analysis platform with an integrated antibody database. *Front. Immunol.* 12, 717496
12. Ferdous, S. and Martin, A.C.R. (2018) AbDb: antibody structure database—a database of PDB-derived antibody structures. *Database (Oxford)* 2018, bay040
13. Dunbar, J. *et al.* (2014) SAbDab: the structural antibody database. *Nucleic Acids Res.* 42, D1140–D1146

14. Swindells, M.B. *et al.* (2017) abYsis: integrated antibody sequence and structure-management, analysis, and prediction. *J. Mol. Biol.* 429, 356–364
15. Raybould, M.I.J. *et al.* (2020) Thera-SAbDab: the therapeutic structural antibody database. *Nucleic Acids Res.* 48, D383–D388
16. Allcorn, L.C. and Martin, A.C. (2002) SACS—self-maintaining database of antibody crystal structure information. *Bioinformatics* 18, 175–181
17. Jankauskaite, J. *et al.* (2019) SKEMPI 2.0: an updated benchmark of changes in protein-protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics* 35, 462–469
18. Sirin, S. *et al.* (2016) AB-bind: antibody binding mutational database for computational affinity predictions. *Protein Sci.* 25, 393–409
19. Raybould, M.I.J. *et al.* (2021) CoV-AbDab: the coronavirus antibody database. *Bioinformatics* 37, 734–735
20. Wilton, E.E. *et al.* (2018) sdAb-DB: the single domain antibody database. *ACS Synth. Biol.* 7, 2480–2484
21. Kiemer, V. (2008) Antibodypedia. *Nat. Methods* 5, 860
22. Bailly, M. *et al.* (2020) Predicting antibody developability profiles through early stage discovery screening. *mAbs* 12, 1743053
23. Raybould, M.I.J. and Deane, C.M. (2022) The therapeutic antibody profiler for computational developability assessment. *Methods Mol. Biol.* 2313, 115–125
24. Chen, X. *et al.* (2020) Predicting antibody developability from sequence using machine learning. *bioRxiv* Published online June 20, 2020. <https://doi.org/10.1101/2020.06.18.159798>
25. Sormanni, P. *et al.* (2015) The CamSol method of rational design of protein mutants with enhanced solubility. *J. Mol. Biol.* 427, 478–490
26. Hou, Q. *et al.* (2020) SOLart: a structure-based method to predict protein solubility and aggregation. *Bioinformatics* 36, 1445–1452
27. Lai, P.K. *et al.* (2022) Machine learning prediction of antibody aggregation and viscosity for high concentration formulation development of protein therapeutics. *mAbs* 14, 2026208
28. Pujols, J. *et al.* (2022) A3D 2.0 update for the prediction and optimization of protein solubility. *Methods Mol. Biol.* 2406, 65–84
29. Lai, P.K. *et al.* (2021) Machine learning feature selection for predicting high concentration therapeutic antibody aggregation. *J. Pharm. Sci.* 110, 1583–1591
30. Vaisman-Mentesh, A. *et al.* (2019) Molecular landscape of anti-drug antibodies reveals the mechanism of the immune response following treatment with TNF α antagonists. *Front. Immunol.* 10, 2921
31. Dhanda, S.K. *et al.* (2019) IEDB-AR: immune epitope database-analysis resource in 2019. *Nucleic Acids Res.* 47, W502–W506
32. Marks, C. *et al.* (2021) Humanization of antibodies using a machine learning approach on large-scale repertoire data. *Bioinformatics* 37, 4041–4047
33. Prihoda, D. *et al.* (2022) BioPhi: a platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning. *mAbs* 14, 2020203
34. Ovacki, M. and Lin, K. (2018) Tutorial on monoclonal antibody pharmacokinetics and its considerations in early development. *Clin. Transl. Sci.* 11, 540–552
35. Grinshpun, B. *et al.* (2021) Identifying biophysical assays and in silico properties that enrich for slow clearance in clinical-stage therapeutic antibodies. *mAbs* 13, 1932230
36. Thorsteinson, N. *et al.* (2021) Structure-based charge calculations for predicting isoelectric point, viscosity, clearance, and profiling antibody therapeutics. *mAbs* 13, 1981805
37. North, B. *et al.* (2011) A new clustering of antibody CDR loop conformations. *J. Mol. Biol.* 406, 228–256
38. Shirai, H. *et al.* (2014) High-resolution modeling of antibody structures by a combination of bioinformatics, expert knowledge, and molecular simulations. *Proteins* 82, 1624–1635
39. Sircar, A. *et al.* (2009) RosettaAntibody: antibody variable region homology modeling server. *Nucleic Acids Res.* 37, W474–W479
40. Webb, B. and Sali, A. (2016) Comparative protein structure modeling using MODELLER. *Curr. Protoc. Protein Sci.* 86, 2
41. Leem, J. *et al.* (2016) ABodyBuilder: automated antibody structure prediction with data-driven accuracy estimation. *mAbs* 8, 1259–1268
42. Abanades, B. *et al.* (2022) ABlooper: fast accurate antibody CDR loop structure prediction with accuracy estimation. *Bioinformatics* 38, 1877–1880
43. Ruffolo, J.A. *et al.* (2022) Antibody structure prediction using interpretable deep learning. *Patterns (N Y)* 3, 100406
44. Leman, J.K. *et al.* (2020) Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat. Methods* 17, 665–680
45. Ruffolo, J.A. *et al.* (2022) Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *bioRxiv* Published online April 21, 2022. <https://doi.org/10.1101/2022.04.20.488972>
46. Jumper, J. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589
47. Chowdhury, R. *et al.* (2022) Single-sequence protein structure prediction using a language model and deep learning. *Nat. Biotechnol.* 40, 1617–1623
48. Lin, Z. *et al.* (2022) Evolutionary-scale prediction of atomic level protein structure with a language model. *bioRxiv* Published online December 21, 2022. <https://doi.org/10.1101/2022.07.20.500902>
49. Wu, R. *et al.* (2022) High-resolution de novo structure prediction from primary sequence. *bioRxiv* Published online July 22, 2022. <https://doi.org/10.1101/2022.07.21.500999>
50. Lee, J.H. *et al.* (2022) EquiFold: protein structure prediction with a novel coarse-grained structure representation. *bioRxiv* Published online January 02, 2023. <https://doi.org/10.1101/2022.10.07.511322>
51. Evans, R. *et al.* (2022) Protein complex prediction with AlphaFold-multimer. *bioRxiv* Published online March 10, 2022. <https://doi.org/10.1101/2021.10.04.463034>
52. Schneider, C. *et al.* (2021) DLAB-deep learning methods for structure-based virtual screening of antibodies. *Bioinformatics* 38, 377–383
53. Eguchi, R.R. *et al.* (2022) Ig-VAE: generative modeling of protein structure by direct 3D coordinate generation. *PLoS Comput. Biol.* 18, e1010271
54. Anand, N. and Huang, P. (2018) Generative modeling for protein structures. *Adv. Neural Inf. Process. Syst.* 31, 54062141
55. Shan, S. *et al.* (2022) Deep learning guided optimization of human antibody against SARS-CoV-2 variants with broad neutralization. *Proc. Natl. Acad. Sci. U. S. A.* 119, e2122954119
56. Shin, J.E. *et al.* (2021) Protein design and variant prediction using autoregressive generative models. *Nat. Commun.* 12, 2403
57. Devlin, J. *et al.* (2018) Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv* Published online October 11, 2018. <https://doi.org/10.48550/arXiv.1810.04805>
58. Elnaggar, A. *et al.* (2020) ProtTrans: towards cracking the language of life's code through self-supervised deep learning and high performance computing. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 7112–7127
59. Rives, A. *et al.* (2021) Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U. S. A.* 118, e2016239118
60. Olsen, T.H. *et al.* (2022) AbLang: an antibody language model for completing antibody sequences. *Bioinformatics Adv.* 2, vbac046
61. Akbar, R. *et al.* (2022) In silico proof of principle of machine learning-based antibody design at unconstrained scale. *mAbs* 14, 2031482
62. Ruffolo, J.A. *et al.* (2021) Deciphering antibody affinity maturation with language models and weakly supervised learning. *arXiv* Published online December 14, 2021. <https://doi.org/10.48550/arXiv.2112.07782>
63. Shuai, R.W. *et al.* (2021) Generative language modeling for antibody design. *bioRxiv* Published online December 20, 2022. <https://doi.org/10.1101/2021.12.13.472419>
64. Hie, B.L. *et al.* (2022) Efficient evolution of human antibodies from general protein language models and sequence information alone. *bioRxiv* Published online April 11, 2022. <https://doi.org/10.1101/2022.04.10.487811>
65. Mason, D.M. *et al.* (2021) Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nat. Biomed. Eng.* 5, 600–612
66. Lim, Y.W. *et al.* (2022) Predicting antibody binders and generating synthetic antibodies using deep learning. *mAbs* 14, 2069075
67. Saka, K. *et al.* (2021) Antibody design using LSTM based deep generative model from phage display library for affinity maturation. *Sci. Rep.* 11, 5852
68. Liu, G. *et al.* (2020) Antibody complementarity determining region design using high-capacity machine learning. *Bioinformatics* 36, 2126–2133

69. Jin, W. *et al.* (2021) Iterative refinement graph neural network for antibody sequence-structure co-design. *arXiv* Published online October 9, 2021. <https://doi.org/10.48550/arXiv.2110.04624>
70. Anand, N. *et al.* (2022) Protein sequence design with a learned potential. *Nat. Commun.* 13, 746
71. Kang, Y. *et al.* (2021) Sequence-based deep learning antibody design for in silico antibody affinity maturation. *arXiv* Published online February 21, 2021. <https://doi.org/10.48550/arXiv.2103.03724>
72. Anishchenko, I. *et al.* (2021) De novo protein design by deep network hallucination. *Nature* 600, 547–552
73. Yang, J. *et al.* (2020) Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci. U. S. A.* 117, 1496–1503
74. Trippe, B.L. *et al.* (2022) Diffusion probabilistic modeling of protein backbones in 3D for the motif-scaffolding problem. *arXiv* Published online June 8, 2022. <https://doi.org/10.48550/arXiv.2206.04119>
75. Anand, N. and Achim, T. (2022) Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. *arXiv* Published online May 26, 2022. <https://doi.org/10.48550/arXiv.2205.15019>
76. Watson, J.L. *et al.* (2022) Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. *bioRxiv* Published online December 14, 2022. <https://doi.org/10.1101/2022.12.09.519842>
77. Ingraham, J. *et al.* (2022) Illuminating protein space with a programmable generative model. *bioRxiv* Published online December 2, 2022. <https://doi.org/10.1101/2022.12.01.518682>
78. Muratov, E.N. *et al.* (2021) A critical overview of computational approaches employed for COVID-19 drug discovery. *Chem. Soc. Rev.* 50, 9121–9151
79. Hsu, C. *et al.* (2022) Learning inverse folding from millions of predicted structures. *bioRxiv* Published online September 06, 2022. <https://doi.org/10.1101/2022.04.10.487779>
80. McPartlon, M. *et al.* (2022) A deep SE(3)-equivariant model for learning inverse protein folding. *bioRxiv* Published online April 16, 2022. <https://doi.org/10.1101/2022.04.15.488492>
81. Strokach, A. *et al.* (2021) Computational generation of proteins with predetermined three-dimensional shapes using ProteinSolver. *STAR Protoc.* 2, 100505
82. Lee, J.S. and Kim, P.M. (2022) ProteinSGM: score-based generative modeling for de novo protein design. *bioRxiv* Published online July 13, 2022. <https://doi.org/10.1101/2022.07.13.499967>
83. Wang, J. *et al.* (2021) Deep learning methods for designing proteins scaffolding functional sites. *bioRxiv* Published online November 15, 2021. <https://doi.org/10.1101/2021.11.10.468128>
84. Castro, E. *et al.* (2022) ReLSO: a transformer-based model for latent space optimization and generation of proteins. *arXiv* Published online January 24, 2022. <https://doi.org/10.48550/arXiv.2201.09948>
85. Zhao, F. *et al.* (2017) Codon usage regulates protein structure and function by affecting translation elongation speed in *Drosophila* cells. *Nucleic Acids Res.* 45, 8484–8492
86. Liu, Y. (2020) A code within the genetic code: codon usage regulates co-translational protein folding. *Cell Commun. Signal.* 18, 145
87. Magistrelli, G. *et al.* (2017) Optimizing assembly and production of native bispecific antibodies by codon de-optimization. *MAbs* 9, 231–239
88. Rosenberg, A.A. *et al.* (2022) Codon-specific Ramachandran plots show amino acid backbone conformation depends on identity of the translated codon. *Nat. Commun.* 13, 2815
89. Adolf-Bryfogle, J. *et al.* (2018) RosettaAntibodyDesign (RABD): a general framework for computational antibody design. *PLoS Comput. Biol.* 14, e1006112
90. Chowdhury, R. *et al.* (2018) OptMAVEN-2.0: de novo design of variable antibody regions against targeted antigen epitopes. *Antibodies (Basel)* 7, 23
91. Saraf, M.C. *et al.* (2006) IPRO: an iterative computational protein library redesign and optimization procedure. *Biophys. J.* 90, 4167–4180
92. Nimrod, G. *et al.* (2018) Computational design of epitope-specific functional antibodies. *Cell Rep.* 25, 2121–2131
93. Sun, M.G. *et al.* (2016) Protein engineering by highly parallel screening of computationally designed variants. *Sci. Adv.* 2, e1600692
94. Aguilar Rangel, M. *et al.* (2022) Fragment-based computational design of antibodies targeting structured epitopes. *Sci. Adv.* 8, eabp9540
95. Guedes, I.A. *et al.* (2018) Empirical scoring functions for structure-based virtual screening: applications, critical aspects, and challenges. *Front. Pharmacol.* 9, 1089