Trends in Pharmacological Sciences

CellPress

Review

Machine Learning for Biologics: Opportunities for Protein Engineering, Developability, and Formulation

Harini Narayanan,[1] Fabian Dingfelder,[1,2] Alessandro Butté,[3] Nikolai Lorenzen,[2] Michael Sokolov,[3] and Paolo Arosio[1,*]

Successful biologics must satisfy multiple properties including activity and particular physicochemical features that are globally defined as developability. These multiple properties must be simultaneously optimized in a very broad design space of protein sequences and buffer compositions. In this context, artificial intelligence (AI), and especially machine learning (ML), have great potential to accelerate and improve the optimization of protein properties, increasing their activity and safety as well as decreasing their development time and manufacturing costs. We highlight the emerging applications of ML in biologics discovery and development, focusing on protein engineering, early biophysical screening, and formulation. We discuss the power of ML in extracting information from complex datasets and in reducing the necessary experimental effort to simultaneously achieve multiple quality targets. We finally anticipate possible future interventions of AI in several steps of the biological landscape.

Highlights

Biologics are an important class of therapeutics due to their high specificity, efficacy, and safety.

However, biomolecule discovery and optimal formulation development are time-and resource-intensive.

The search space is highly complex and multidimensional because multiple physicochemical properties must be optimized.

AI is emerging as a predictive and generative tool to aid in protein engineering for therapeutic applications. AI can also be employed to model multiple biophysical and chemical degradation properties.

The Need for AI in Biologics

Biologics include a plethora of different molecular formats such as enzymes, hormones, peptides, cytokines, fusion proteins, monoclonal antibodies (mAbs) and next-generation antibody formats such as antibody–drug conjugates (ADCs), bispecific antibodies, single-chain variable fragments (scFvs), vaccine components, and gene therapy vectors. Biologics are an important class of therapeutics[i], and eight of the top 10 best-selling drugs in 2018 were biologics [1,2]. Key benefits of biologics are their high specificity and affinity, longer-acting pharmacokinetics, and lower toxicity and side effects compared to small molecules [3–5]. This high potential should, however, be considered together with the very expensive development and manufacturing processes that precede drug commercialization. Figure 1 summarizes the complex procedure from discovery, through process development and clinical trials, to the manufacture of a mAb. The associated costs per successful drug are in the order of 2 billion USD[ii] and the timescale is ~10–15 years [6], which must be compared to the 25 years patent runtime of the molecule. There is economic pressure to move fast because of impending patent expiry and often fierce competition, as well as social pressure to make new biologics accessible to patients worldwide, and new technologies are therefore necessary to generate high quantities of in-specification products to meet market demand, at accessible costs, and in shorter timescales. These needs have become even more obvious with the onset of the COVID-19 pandemic in early 2020 [7].

Drug development is a highly complex multi-objective optimization problem in which a candidate molecule must satisfy multiple criteria, including activity against a biological target, suitable biophysical and pharmacokinetic properties, and safety [8]. Compared to small molecules, the space of protein sequence and solution conditions that must be screened is much wider. The

[1]Department of Chemistry and Applied Biosciences, Institute for Chemical and Bioengineering, Swiss Federal Institute of Technology, Zurich 8093, Switzerland
[2]Department of Biophysics and Injectable Formulation 2, Global Research Technologies, Novo Nordisk A/S, Måløv 2760, Denmark
[3]DataHow AG, Zurich 8093, Switzerland

*Correspondence:
paolo.arosio@chem.ethz.ch (P. Arosio).

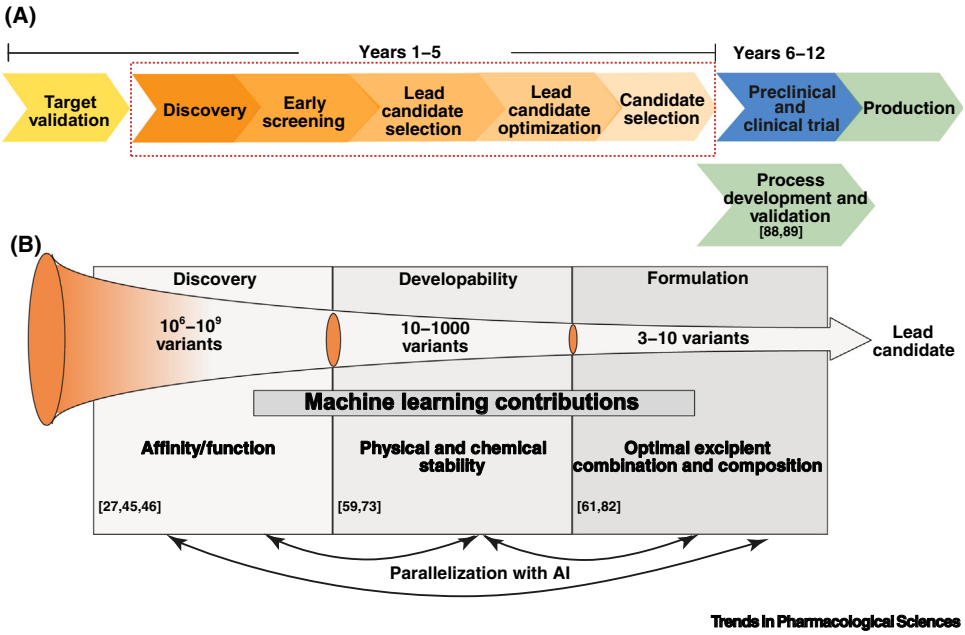

Figure 1. The Potential Contributions of Artificial Intelligence (AI) in the Biologics Development Cycle. (A) Key steps involved in the biologics development cycle from target identification to marketed product. (B) Outline of the areas discussed in this review with key examples of machine learning applications. 'Parallelization with AI' indicates opportunities of AI to simultaneously optimize multiple properties for successful drug development. (See [27,45,46,59,61,73,82,88,89].)

number of possible protein sequences is in the order of 20^k, where *k* is the number of amino acids to be changed in the sequence, each of which could be formulated in thousands (or even more) of possible buffer compositions. Moreover, complex molecules such as proteins must demonstrate appropriate chemical and physical properties. A combinatorial study of all these factors results in an extremely large space that is well outside the cognitive ability of humans, especially under restricted timelines and laboratory capacities. The functional sequence space (with a single desired functionality) is a very narrow subset of this wide space and is estimated to be as low as 1 in 10^{77} to a maximum of 1 in 10^{11} sequences [9]. **Artificial intelligence** (AI; see Glossary) has the potential to traverse this space efficiently while simultaneously optimizing different parameters. Machine learning (ML) is a subfield of AI that deals with algorithms that can learn from samples or instances that are often multidimensional and contain complex patterns, noise, and redundancies. Advances in computer hardware (e.g., graphics processing units, GPUs; faster central processing units, CPUs), cloud computing, and new software algorithms, coupled with the exponentially growing availability of data [10], have allowed conventional ML and advanced ML algorithms (**deep learning**, DL) to be applied in several domains. Some algorithms that are widely used in different applications are summarized in Box 1. ML applications in the field of computer vision and robotics typically deal with image, voice, and text data. Inspired by the success of ML in these areas, ML and particularly DL were readily applicable to medical images, signal data (speech), and electronic health records (text) for diagnostics. Recently, given the development of several high-throughput assays, ML is spreading to other domains of biomedicine such as molecular target prediction and functional genomic element prediction in fundamental biology and (small-molecule) drug discovery for disease treatment [11].

In this review we focus on the potential of AI in the design and development of biologics, with a particular focus on protein engineering, the selection of variants during early-stage screening, and formulation (Figure 1B). We start with a brief description of a typical ML workflow. We then

Glossary

Artificial intelligence (AI): a domain focusing on simulating human intelligence in a machine, resulting in smart machines.
Cell-line selection: it is a process in which the gene for the protein of interest is transfected into host cells leading to a heterogeneous cell pool. Cells are sorted into single-cell cultures, and the cell line that produces highest quality and quantity in sequentially scaled-up culture is selected for the master cell bank which is used for production during clinical trials and later for commercial use.
Classification: supervised learning tasks where the target is categorical, for instance, 'yes' or 'no'.
Deep learning (DL): a subclass of machine learning (ML) that uses sophisticated multilevel deep neural networks to train on unlabeled or labeled data.
***De novo* design:** designing completely new polypeptide sequences that can fold into a stable 3D structure and show desired functionality (existing or new).
Directed evolution: a protein engineering method that uses multiple rounds of mutagenesis and selection to improve existing functions.
Epitope: the part of an antigen that interacts with the antibody.
Feature engineering: the technique of obtaining meaningful information from the raw inputs while preparing a representation of a dataset that is compatible with ML algorithms. This can be based on domain knowledge or on black-box methods (such as DL).
Higher-level features: features built on top of existing features. For instance, during object identification in images, pixels are grouped to identify lines and edges (features or low-level features) and operations are performed to extract shapes from these features (higher-level features).
Inverse design: unlike the direct approach, that takes the input and predicts the output, inverse design determines the input that will lead to the output of interest.
Microfluidics: the science of controlling and manipulating fluids at a micrometer scale; this is governed by physical principles that differ from those operating at the macroscale. Microfluidic devices contain channel networks, require small sample volumes, and offer the potential to perform multiple experiments in parallel.

Box 1. Description of ML and DL Methods

Among the several unsupervised learning methods, clustering algorithms and principal component analysis (PCA) are commonly used to segregate different patterns in the data and reduce the dimensions of the data, respectively. The simplest supervised ML models are linear or logistic **regressions** that perform a linear combination of the input features to predict the output. Decision trees (DTs) lead to output estimation by consecutively splitting the dataset based on best split decisions taken on the input features. DTs are typically used in ensemble methods such as random forests (RFs) or gradient boosted trees. Ensemble techniques combine predictions from multiple estimators (or models) by taking an average or major-voting in regression and classification tasks, respectively. Kernel-based methods such as support vector machines (SVMs) impose non-linear transformation by using kernel functions to implicitly project the input features into a high-dimensional space without explicitly calculating the coordinates in the new space. Gaussian processes (GPs) additionally combine kernel methods with Bayesian learning to provide predictions as a probability distribution. Artificial neural networks (ANNs) consist of a group of interconnected nodes that take the input features and, after several non-linear transformations, connect them to the output(s). Deep learning (DL) is a development of ANNs that uses sophisticated architecture to perform automatic feature detection from highly structured data such as images or text. A prerequisite for DL methods is a large set of training data. Whereas ANNs use one or two hidden layers, DL uses neural network architectures with a large number of hidden layers. The simplest architecture is a fully connected feedforward neural network (FFNN), which is a directed analogy of traditional ANNs in which all the nodes in one layer are connected to every node in the subsequent layer. A convolutional neural network (CNN) is a DL architecture in which some of the hidden layers are only locally connected to the subsequent hidden layer, generating simple local features and hierarchically combining them into complex features. This architecture is well suited for image recognition or for cases where local features must be captured within a space containing many generic features. The third architecture is the recurrent neural network (RNN) in which the connections between the nodes form a directed graph along a sequence; this is used for data such as time-series and text. Variational autoencoders (VAEs) are unsupervised learning neural net architectures that encode the inputs in lower-dimensional latent space. Finally, generative adversarial networks (GANs) are a combination of two networks (any of the architectures mentioned above) in which one generates synthetic data and the other learns to differentiate between real and synthetic data.

Multitask learning: an ML paradigm in which the aim is to leverage information contained in multiple tasks to assist generalization in all tasks and also to facilitate efficient learning for related task with fewer datapoints.
Paratope: the part of an antibody that recognizes and binds to the antigen.
Rational design: protein engineering using *a priori* knowledge about protein residues, domains, and scaffolds to target specific interactions or functions. For instance, fusing well-characterized protein domains to create a single multidomain protein with distinct functions.
Reinforcement learning: an ML approach that interacts with its environment by producing actions and learning the relationship between possible actions and the outcomes.
Regression: supervised learning tasks where the target is continuous.
Soft sensors: mathematical models (or software) that use measurements from other physical sensors to estimate the values of variables that are difficult to measure.
Transfer learning: an ML paradigm in which knowledge obtained in a particular task is used in a related task by repurposing the model learned in one task as the starting point for the other.
Unstructured raw data: raw data that do not possess a fixed-length vector representation that is classically required as input to ML algorithms.

highlight the emerging applications and potential of ML for biologics development, reviewing the current works in protein engineering, biomolecule optimization, and formulation development.

ML Workflow

ML models learn input to output relationships (supervised learning), or the patterns that are present in a given input (unsupervised learning), from sample instances that are defined as training data. In biologics discovery, common inputs are represented by the protein sequence or structure and/or by environmental conditions (such as the formulation composition). The output is typically a functional product property such as activity, stability, or a specific physicochemical feature.

ML models can be used for different applications, as illustrated in Figure 2A. Models referred to as discriminative or predictive models are used to predict the properties of new input [route (i) in Figure 2A]. In a second type of applications, known as **inverse design**, models are not only used to predict properties but also to design new optimal (bio)molecules or experiments. For instance, predictive models can be coupled with optimization algorithms [route (ii)] to design optimal inputs to meet desired properties. In addition, a recent advance of DL models, called generative models, aim to create new inputs from the learned parts of the input - property space that can be used for inverse design, thus providing access to unexplored regions of the input space [route (iii)].

The first step in applying these models is to generate the training instances that determine the boundaries of the model [12] (Figure 2B). Information can be obtained from public databases that collect experiments previously conducted or by performing an initial set of experiments. These experiments may be performed either randomly or be guided by a design of experiments (DoE) strategy that maximizes the coverage of the design space for a defined experimental capacity [12].

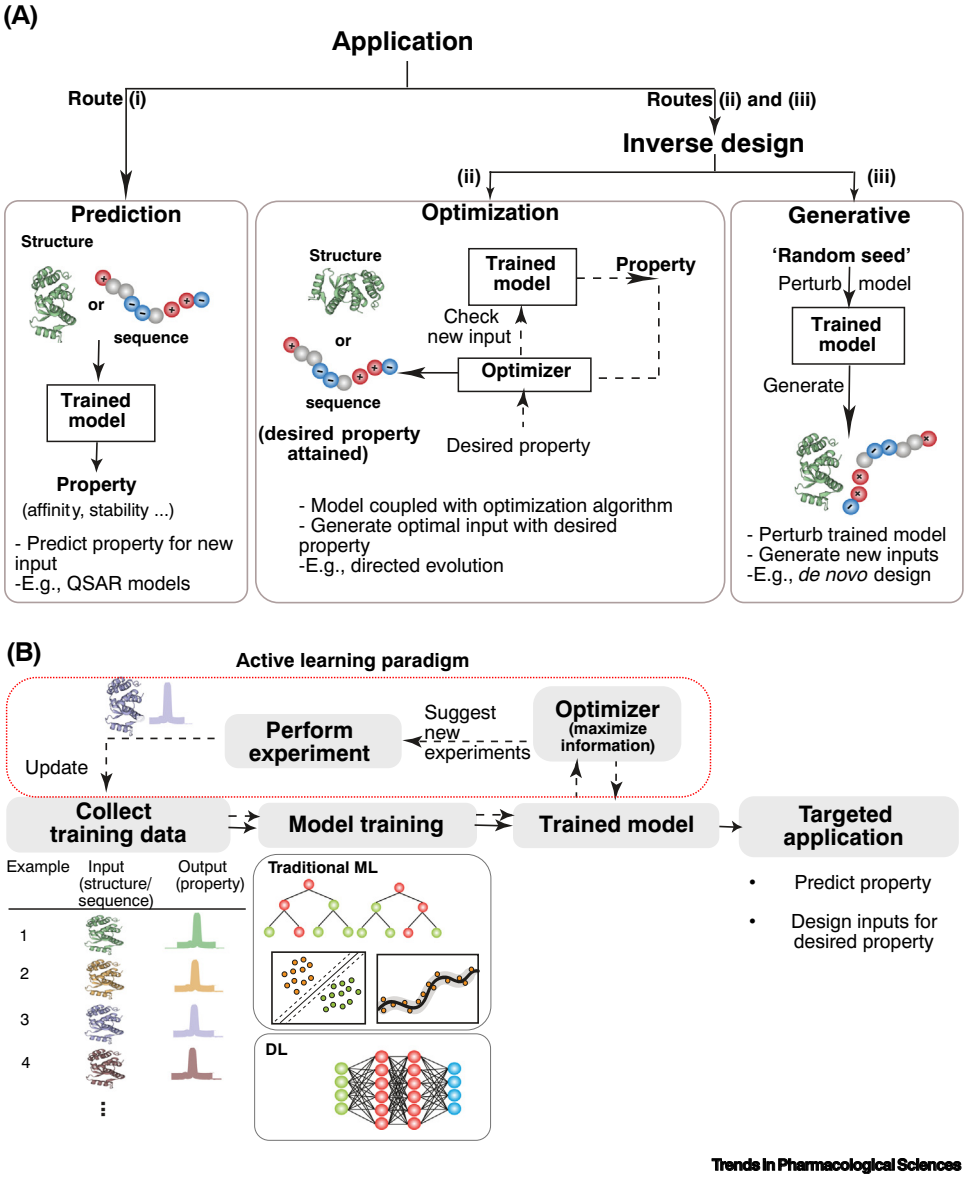

Figure 2. Schematic Illustration of Machine Learning (ML) Workflow. (A) Major types of ML applications in biologics development, namely prediction and design (optimization vs generative). (B) Schematic illustration of the different steps involved in developing an ML model with passive learning and additional steps involved in active learning (box with broken red lines). Abbreviations: DL, deep learning; QSAR, quantitative structure–activity relationship.

During passive learning, the model is built with fixed training data. By contrast, during active learning (Figure 2B) training datasets and thus the models are updated and used iteratively to design new experiments to ensure the efficient learning of input–output relationships in the entire design space. In comparison to classical DoE, active learning methods present the ability to learn the relevant input–output information more resource-efficiently.

Once a training dataset is obtained, the next steps involve the choice and training of a suitable type of ML algorithm (Figure 2B). This choice is mainly based on the application as well as on the amount and type of available data. Algorithms such as support vector machines (SVMs)

and random forests (RFs) have performed well across different application domains for small and moderate datasets, whereas DL methods require large amounts of data. In addition, classical ML requires a procedure called **feature engineering** that involves extraction of input features from **unstructured raw data** such as text (protein sequence), graphs (molecular graphs of entities or complexes), and images. By contrast, DL models eliminate the need for feature engineering by using different architectures that can translate unstructured data into a compact representation by progressively extracting **higher-level features** from raw data.

ML in Biologics Discovery

Applications of ML in biologics can take inspiration from successful stories in the field of small-molecule discovery and development (Box 2) in which ML tools have been the focus of study for the past decade. Several excellent reviews (e.g., [8,10,11,13]) have presented the algorithms, achievements, and challenges in AI-assisted discovery of small-molecule therapeutics.

In contrast to small molecules, the use of ML tools for biotherapeutics is still in its early stages. This can be attributed to a series of factors, including the complexity of biomolecules and the sparse, heterogeneous, and smaller datasets that are available or in the data warehouses of pharma companies.

However, biomanufacturing has already witnessed an infiltration of ML techniques for over a decade (Box 3), although with a significant delay compared to other process industries. Bioprocess data typically involve measurements of process variables that could be used directly

Box 2. ML in Small-Molecule Discovery and Design

The discovery of promising small-molecule candidates can be achieved through high-throughput screening of the commercially available molecules or through the design of novel molecules (*de novo* design).

ML has been employed for screening libraries of chemical compounds *in silico* (virtual screening, VS) and rank them based on their binding affinity to the biological target. In an approach known as quantitative structure–activity relationship (QSAR) modeling, classification or regression models are trained to predict active/inactive ligands or binding-affinity values of the ligand, respectively. These models are then used to predict the behavior of all the commercially available chemical compounds and rank them accordingly [route (i) in Figure 2A in main text] [90]. Such *a priori in silico* screening has the potential to significantly accelerate drug discovery in a resource- and cost-efficient manner.

VS can be applied in two ways: (i) based solely on the chemical molecule (ligand-based virtual screening, LBVS), or (ii) based on the target–ligand complex (structure-based virtual screening, SBVS). Taking advantages of large standardized databases for ligand activity data (e.g., PubChem, ZINC, ChEMBL, and ChemSpider) and binding data for protein/ligand complexes (e.g., DUD, PubBind, CSAR, and MUV) [90], different conventional ML techniques such as naïve Bayes classifier (NBC), k-nearest neighbor (kNN), SVMs, linear analysis, and RFs have been used for LBVS (reviewed in [91,92] and compared by Korkmaz *et al.* in [93]). However, with advances in DL, QSAR models based on **multitask learning** have been shown to have improved performance compared to classical ML tools [11,13,94]. In the context of SBVS, ML methods have been used as scoring function in docking programs [90,95,96] or for binding-affinity prediction based on drug–target complexes [90,97,98].

In addition to screening, generative DL models have been applied in *de novo* drug design [99–102].

In addition to activity, the physicochemical properties of molecules must be optimized because these are crucial for drug safety and efficacy [11,103]. In conventional approaches, the most promising candidates identified either by VS or *de novo* design are further screened for properties such as absorption, distribution, metabolism, excretion (ADME), and toxicity. Many classical ML algorithms such as kNN, SVM, partial least square (PLS), RF, and DL methods have been used to predict the ADME-toxicity properties [94,104–108].

Among these property-prediction algorithms, applications that couple generative DL models with **reinforcement learning** have recently demonstrated the ability of ML to identify optimized molecules that simultaneously fulfill multiple properties [103,109–112].

Box 3. ML in Biopharmaceutical Process Development and Manufacturing

Once a final candidate molecule is selected, the process development for manufacturing of the product is initiated. The biopharmaceutical preparation is a heterogeneous mixture of proteins with diverse structures, further compounded by various post-translational modifications and degradation pathways. Both the intrinsic properties of the protein and extrinsic factors such as media composition and process conditions affect post-translation modifications and hence product quality. As a result, designing a process suitable for producing a high-quality product is essential, as emphasized by the Quality by Design (QbD) initiative [113]. Initiatives by health authorities who require better quantitative understanding of the process and product interrelationship, combined with the increasing exploitation of smart digital solutions (Industry 4.0 or Pharma 4.0), are instrumental for the application of ML to bioprocess development and manufacturing [88]. The operating conditions (or the so-called process design space) for production are iteratively optimized during process development. This begins with **cell-line selection** to obtain a clone capable of producing a high quantity and consistent quality of the product. ML has been employed for image analysis in fluorescence-assisted cell sorting (FACS) to segregate the heterogeneous pool of cells into single-cell cultures [114,115]. The subsequent process development involves taking many different choices, from media selection to process conditions (e.g., pH, temperature, dissolved oxygen) and optimization of the control regime (structure, intervals, etc.) [116]. In addition, the process conditions must be optimized for all the other unit operations in the biomanufacturing process. ML is envisaged to support process optimization *in silico* as so-called digital twins [117,118] to minimize the experimental effort for process development and validation. In such a regulated environment, the models considered must also be validated before being utilized for manufacturing control, and this requires closer interaction of digital solution providers, users, and health authorities.

The use of classical DoEs for advanced ML-based techniques has been demonstrated for the development of optimal processes that maximize production while ensuring quality [119,120]. This type of optimization is termed open loop control and has been conventionally used in industry. However, real-time monitoring and control of the process has recently gained attention wherein critical quality attributes (CQAs) and product quantity are followed in real-time, and control decisions are continuously taken based on updated data [118,121]. As an alternative to complex analytical measurements of the CQAs, **soft sensors** based on ML methods such as PLS, GP, and SVM are used to predict CQAs using data available online or atline. After selecting an optimal process design, large-scale experiments are performed to demonstrate consistent product quality and a rigorous control scheme. Finally, the validated process is used for production. ML applications in different areas of process development and manufacturing have been recently reviewed [88,89].

as inputs to ML algorithms. Thus, classical ML tools could be applied relatively efficiently even if limited product- or project-specific data are available, and without being limited by feature-engineering challenges.

By contrast, modeling activities in biomolecule discovery require efficient feature representation from complex data types, such as protein sequences or structures. This was made possible only recently with the development of DL architectures that are capable of automatically extracting representative features from unstructured inputs such as texts, images, and graphs. Efficient software implementation of these techniques, together with improved computational (such as faster CPUs and GPUs) and laboratory support from novel high-throughput experimental assays (e.g., next-generation sequencing), is now driving the recent application of ML in biologics discovery.

The discovery of proteins with desired functions and properties typically occurs through three techniques of protein engineering or design [14]: **rational design**, **directed evolution** (DE), or ***de novo* design**. As demonstrated in Figure 2A, the application of ML models through route (i) can support rational design, whereas routes (ii) and (iii) highlighted under 'inverse design' assist DE and *de novo* design, respectively. ML can learn input–output relationships even when the underlying physical principles are not completely understood. This property makes ML very attractive in incompletely characterized fields such as protein science. Compared to other approaches such as molecular dynamics, the low computational time and resource requirements of ML provide an additional advantage to significantly accelerate the discovery process.

Successful application of ML-assisted DE [12] using Gaussian processes (GPs) has been demonstrated for non-therapeutic applications such as the engineering of GFP fluorescence [15], the

localization of membrane proteins [16], and protein thermostability [9]. Using ML to model sequence–function relationship aids systematic exploration of the protein landscape. In addition to DE, generative DL models have opened opportunities for *de novo* design of sequences with desired functionalities such as affinity, solubility, and stability [17–20]. There are also reports of the use of different DL architectures, such as variational autoencoder (VAE) [21] and recurrent neural network (RNN) [22,23], to learn a representative embedding of protein sequence that could be coupled to other property prediction tasks or used for *de novo* design [24,25]. Although most protein design work has involved non-therapeutic applications, the concepts and approaches are easily translatable to therapeutic proteins and peptides. For instance, ML has been used to design [26] and predict the activity and physicochemical properties of antimicrobial peptides [27–29], as well as the cell-penetrating ability of peptides for drug delivery applications.

In recent years, some applications have also emerged for the design of mAbs. With increasing availability of sequence (e.g., Abysis) and structure information (e.g., SAbDab, PDBbind) [30], ML can be used independently or in combination with conventional discovery techniques, which include animal immunization, human patients, antibody library screening based on display technology, rational design, and molecular dynamics (MD) simulations [31–34].

Initial ML applications emerged for the identification of **epitopes** and **paratopes**, which are crucial for the rational design of antibodies. In addition, epitope identification is also essential for diagnostics and vaccine development. On the one hand, models based on SVM [35] and DL [36–38] have been employed to predict linear epitope regions in antigen sequences. On the other, RF [39], SVM [40], and DL [41,42] models trained on complementarity-determining regions (CDRs) in the hypervariable loop region can be used to predict the probability of each residue being a paratope. In addition, Deac *et al.* [42] and Akbar *et al.* [43] presented DL-based models for epitope-specific paratope identification and for predicting antibody/antigen binding, respectively.

More recently, DL has been applied to facilitate virtual screening of sequence libraries [44]. Further, Liu *et al.* used DL to optimize the CDR H3 region, and highlighted the possibility of transferring knowledge across different antigen campaigns by using models trained in other campaigns as filters for binding specificity [45]. However, the synthetic screening libraries used for antibody discovery depend on random mutation of residues based on positional frequency analysis, which by itself suffers from limitations such as producing non-human-like antibody sequences and focusing only on the CDR to generate the library [46]. Amimeur *et al.* [46] addressed this challenge by learning a generative DL model (generative adversarial networks, GANs) from human antibodies to then generate a more human-like screening library than conventional synthetic libraries. In addition, the authors demonstrated the use of **transfer learning** to further tune the generative models and produce sequences with favorable properties such as length, immunogenicity, isoelectric point, and surface patches. In this approach generative models that are learned for one germline may be transferred to other germlines. The key advantages are the increase in the accuracy of the models with a smaller amount of training data, and the more efficient generation of screening libraries.

In summary, the development of DL methods has been key in increasing applications of ML in protein-to-property prediction, especially in protein engineering tasks. With the recent success of DL models in protein 3D structure modeling [47,48], the application of AI and the quality of ML models for function modeling and protein design is likely to increase further.

ML for Developability and Formulation

In addition to biological activity, the translation of a candidate biotherapeutic into a successful drug requires a series of other properties that guarantee the stability and safety of the drug during production and formulation. These properties, globally defined as 'developability', are complex and depend on the sequence of the molecule as well as on the buffer composition [4,49]. Moreover, stability must be guaranteed against a variety of stresses, including thermal, chemical, mechanical, and interface stresses. Therefore, developability is a multi-objective optimization problem wherein ML combined with high-throughput experimental techniques represents a very promising screening solution to improve efficiency and to identify optimal properties during the very early stages of the process. This is important to avoid instability problems during later steps of manufacturing or formulation development that can have severe economic implications.

The few hundred promising candidates from the discovery phase are often expressed in mammalian cells and purified in small amounts (100 µg to 1 mg) [50]. At this early stage of development it is already possible to derive biophysical properties relating to conformational or colloidal stability. For example, assays probing reversible self-association such as affinity-capture self-association (AC-SINS) are feasible at high throughput using low amounts of moderate-purity antibodies [51,52]. Furthermore, the thermal stability can be determined with high-throughput by applying for instance nano-differential scanning fluorimetry [53]. Assays probing for non-specificity, such as the non-specific ELISA assay, are also often used to screen for developability [54]. Finally, during lead candidate selection, both chemical and physical stability are assessed in detail. Typical labilities regarding chemical stability include Met and Trp oxidation, Asn deamidation, and Asp isomerization [55]. Deamidation and isomerization, in particular for residues in the CDRs, might lead to decreasing antigen binding affinity. Oxidized Trp residues in CDRs could be related to reduced thermal stability and increased aggregation propensity [55]. Biophysical characterization is performed to ensure low viscosity, low aggregation propensity, and high solubility of the final lead candidate. Ensuring moderate viscosity and high solubility is essential because mAbs are generally formulated at high concentrations (up to 150 mg/ml) to enable subcutaneous delivery and patient self-administration.

Properties that are material- and time-consuming to measure are often indirectly predicted from biophysical parameters that are more easily accessible. However, no single parameter is predictive for the entire developability potential of a biotherapeutic. Therefore, a combination of different methods is essential to be able to flag variants with unfavorable biophysical properties. For instance, Jain *et al.* investigated 137 antibodies that are either approved or in clinical trials with 12 different assays to probe their suitability as therapeutic entities [56]. This effort- and time-consuming combination of multiple experimental techniques has great potential to be supported and accelerated by ML methods.

Hedbitch *et al.* used the dataset published by Jain *et al.* to develop ML models for all the 12 biophysical properties using only the sequence-based descriptors as input [57]. Gentiluomo *et al.* used artificial neural networks (ANNs) to predict the biophysical properties [e.g., melting temperature (T_m), diffusion interaction parameter (K_D), and aggregation onset temperature (T_{agg})] of mAbs [58] using amino acid compositions of the protein sequence and buffer conditions such as pH and salt concentration. The same authors applied ANNs to predict the long-term stability of protein drugs based on accelerated stability studies and biophysical parameters (such as T_m and T_{agg}) on a dataset of 14 proteins formulated in 24 different conditions [59]. This prediction of long-term stability of biotherapeutics against aggregation is a key aspect because aggregates could potentially be immunogenic and must be reported to regulatory agencies [60], and this is currently a major challenge in the field.

In addition to ANNs, SVMs were also applied to predict thermal and pH stability from sequence data. The model could reasonably predict pH_{50} values, whereas prediction of thermal stability was more challenging [61].

Moreover, some ML algorithms are part of *in silico* predictors of particular liabilities that could allow virtual screening of biologics for their developability potential. Many of these tools are based on MD simulations or heuristics (e.g., [62–65] and references therein). However, some of the developed predictors are also based on ML algorithms, for example tools to predict changes in the thermal stability of proteins upon mutation [66–68], which could also be adopted for therapeutic applications. Models based on **gradient-boosted machines** [69], feedforward neural network (FFNNs) with GANs [70], and convolutional neural networks (CNNs) [71] are the current state-of-the-art for protein solubility prediction, whereas SVMs were the previous benchmark. Furthermore, ML tools are also available to predict the effect of individual amino acid substitutions on solubility, for instance based on RF [72], and have been already specifically applied to therapeutics. For example, Obrezanova *et al.* proposed a **classification** tree ensemble method for predicting the intrinsic aggregation propensity of mAbs using sequence-derived physicochemical properties [73]. In terms of chemical stability, Asn deamidation probability and (or) rate prediction methods based on RFs, SVMs, ANNs, naïve Bayes classifier (NBC), and k-nearest neighbor (kNN) have been presented in literature using protein sequence- and structure-derived descriptors as inputs [74,75]. Similarly, RF, SVM, and ANN for Met oxidation site and risk [76,77] prediction based on sequence and structure descriptors of mAbs have also been demonstrated. RF-based predictions of substitute metrics such as solvent accessible surface area (SASA) of Met residues are available [78]. However, one common disadvantage of such *in silico* tools is that they use only protein sequences or structure-based information as input and usually do not consider the impact of formulation conditions.

Indeed, the physical and chemical stability of the molecule is not exclusively an intrinsic property of the sequence, and is instead dictated by the combination of sequence and buffer composition. If not properly designed, excipients might have a detrimental effect [79]. Formulation design is therefore another crucial step in the development of biopharmaceuticals. A typical formulation for biologics involves multiple excipients such as buffering agents, sugars, salts, amino acids, and surfactants. The selection of the optimal composition thus presents a highly multidimensional non-linear optimization problem that can drastically benefit from ML applications [80], in particular from ML-assisted experimental design. For example, Johnson *et al.* applied ANNs to predict second virial coefficients B_{22} of unknown formulation conditions [81]. The model was trained on a small set of formulations and the B_{22} values were derived from self-interaction chromatography (SIC) of lysozyme as a model protein. The approach could easily be transferred to therapeutic proteins.

Furthermore, image analysis using DL methods can be coupled with image-based experimental techniques to analyze aggregates. This has great potential in the analysis of different types of aggregates generated under various stresses such as freeze-thawing or agitation [82]. Randolph and coworkers applied a CNN to differentiate images of aggregates based on the stress responsible for generating the particle. This is a very promising approach because it could identify the source of aggregation during processing, leading to strategies to prevent the formation of aggregates.

Concluding Remarks and Future Perspectives

Applications of ML are beginning to emerge in the field of protein engineering such as biocatalysts and biomaterials discovery, and we have reviewed recent progress in the discovery and

Outstanding Questions

Do the available databases contain data suitable (quantity and quality) for the efficient application of ML?

How should relevant datasets be developed?

What are the advances in terms of experimental techniques that will be necessary to support the application of AI to drug development?

How well do ML methods perform for applications where it is challenging to generate large amounts of experimental data (e.g., aggregation upon long-term storage of mAbs)?

How to translate AI beyond research in real-time industrial applications?

What changes must be made in the current approaches to integrate AI efficiently?

developability of biologics. It is important to note that, in these fields, the potential of ML is not limited to the extraction of information from a set of data that are often multidimensional, containing complex patterns, noise, and redundancies. ML also plays an important role in actively reducing the number of experiments that are necessary to achieve a given target by avoiding redundancy and suggesting the most suitable experiments to derive relevant information for decision support. This role of ML is particularly important in the field of protein engineering where experiments are time- and effort-consuming and the design space is extremely large. In addition, it is expected that ML-based workflows will become even more important for complex biologics such as extracellular vesicles [83], RNA-based therapies [84], and gene and cell therapy [85]. Consistent application of ML to datasets with similar structures will assist in making decisions more quickly, avoiding unnecessary experiments, and creating knowledge libraries that will generally improve scientific understanding of the problems.

Currently, ML tools mainly involve passive learning tasks in which the training dataset is fixed. A crucial aspect of these passive learning tasks is the amount of good-quality data that has sufficient variation and uniform coverage of the investigation space (see Outstanding Questions). Publicly available databases often contain collections of experiments performed with a different goal, and therefore they probe only limited regions of space, resulting in data bias [8].

To alleviate this problem, it is crucial to generate dedicated datasets that are suitable for modeling. With this aim, training samples can be developed with statistical DoE and experimental design strategies that maximize design space coverage (and therefore information content). In

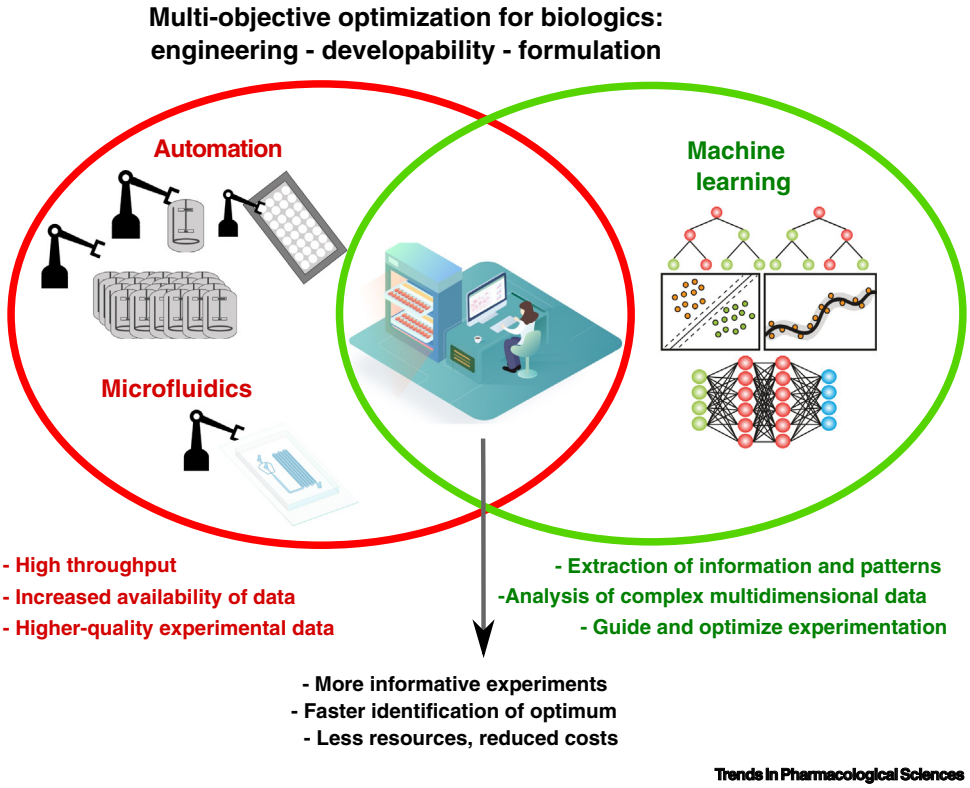

Figure 3. Schematic Representation of the Interplay between Experimental High-Throughput Data and Machine-Learning Algorithms To Enable Efficient Biologics Development.

this context, active learning, a subfield of ML that can efficiently couple data analysis with data generation (Figure 2B), can be particularly useful.

Furthermore, active learning can benefit from recent progress in high-throughput experimental assays that can generate large amounts of data in a relatively automated way, including for instance new sequencing technology. However, the bottleneck is to identify the properties of these sequences that are relevant in the context of a specific application, and this requires significant further experimentation [25]. It is expected that in the near future the potential of **microfluidics** [86,87] and of experimental automation with robotic platforms, possibly in synergy, could assist in performing this operation in a cost- and time-efficient manner.

Moreover, the emerging semi-supervised learning and transfer learning methods could further address the issue of limited data availability. Semi-supervised learning is a ML task in which only a small fraction of the input data are mapped to the properties, and a large fraction of data lack a mapped target. This is the classical scenario in protein modeling, where next-generation sequencing has led to an explosion of plausible protein sequences (unlabeled samples) whereas the properties or functions are quantified only for a handful of them (labeled samples). The unlabeled samples allow the algorithm to learn a better representation of the input, which in turn improves the accuracy of the model in predicting the target.

Transfer learning, by contrast, repurposes the model learned for one task for a second related task. For instance, common features of proteins could be learned from a generalized database, and refinement for a specific application could be performed with limited amounts of data. This has had success for medical image analysis where DL networks learned on natural images are then transfer-learned with medical images [11].

Overall, we envisage that advances in experimental techniques and ML methods have the potential to reduce the experimental burden and therefore the cost of biotherapeutics development (summarized schematically in Figure 3). To expand and accelerate this potential, it is important that ML methods are uniformly applied and combined with consistent data storage.

In light of the general movement to AI, (bio)pharmaceutical companies have created data warehouses and are digitally tracking the different stages of molecule and process development, mainly for regulatory purposes. Only a small fraction of the data generated are currently considered, and the large amount of data remaining unexplored represents a resource with huge potential to accelerate, automate, and optimize many decisions. Most importantly, AI technology can not only be incorporated into the current procedural framework but can also enable drastic changes in the biotherapeutics development process itself. Biologics development would benefit by moving from the traditional approach of serially optimizing properties to a parallel approach in which optimal molecules satisfying different properties (multi-objective optimization) can be designed (Figure 1B). This change in the molecule discovery and development protocol can only be facilitated by the use of AI to allow more effective management of complexity for similar resource utilization. However, a revolutionary breakthrough in AI for biopharma can only be achieved in combination with broad availability of automated experimental and analytical equipment, and this might be limited because not all relevant assays are available in a high-throughput format. Currently, challenges in data generation are still dominant compared to data analysis.

Although the focus of the review has been the emerging application of AI in biologics discovery and development (Figure 1B), the intervention of AI in the biologics landscape is not limited to

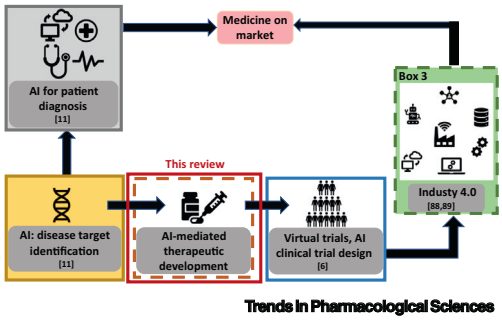

Figure 4. Futuristic Vision of Artificial Intelligence (AI) in a Health, Diagnostics, and Disease Treatment Cycle. The color coding of the boxes is in accordance with the biologics development cycle in Figure 1A. The broken boxes indicate areas covered in this review. Shaded boxes indicate areas where the application of AI has been demonstrated and incremental work is ongoing. Whereas the empty boxes indicate fields where application of AI has emerged only in the recent years. (See [6,11,88,89]).

these aspects. As highlighted in Box 3, significant application of AI has been established in biomanufacturing towards the goals of Industry 4.0. In addition, AI applications have been established in other areas of biomedicine such as fundamental disease biology (target identification), disease diagnosis, and patient categorization, as well as in clinical trial design [6]. Looking at the bigger picture, a futuristic vision could be AI-driven therapeutic treatment, as summarized in Figure 4. In the initial stages of discovery AI could identify new disease targets and suggest optimal therapeutic solutions. AI can then identify the best synthesis protocol to produce the therapeutic in the smart factories of the digital era. This approach will be essential for personalized medicine such as gene and cell therapies in which a customized drug, and therefore a specific production process and formulation recipe, must be optimized for each patient. Finally, AI will assist medical practitioners in patient diagnosis and drug prescription leading to an AI-supported therapeutic–treatment cycle.

Acknowledgments

We gratefully acknowledge Professor Massimo Morbidelli for scientific discussions and critically reading of the manuscript. F.D. acknowledges Novo Nordisk R&D STAR Programme for funding.

Resources

[i] https://www.fda.gov/about-fda/center-biologics-evaluation-and-research-cber/what-are-biologics-questions-and-answers

[ii] https://www2.deloitte.com/uk/en/pages/life-sciences-and-healthcare/articles/measuring-return-from-pharmaceutical-innovation.html

References

1. Kaplon, H. *et al.* (2020) Antibodies to watch in 2020. *MAbs* 12, 1703531
2. Lu, R.-M. *et al.* (2020) Development of therapeutic antibodies for the treatment of diseases. *J. Biomed. Sci.* 27, 1
3. Carter, P.J. and Lazar, G.A. (2018) Next generation antibody drugs: pursuit of the 'high-hanging fruit'. *Nat. Rev. Drug Discov.* 17, 197–223
4. Jarasch, A. *et al.* (2015) Developability assessment during the selection of novel therapeutic antibodies. *J. Pharm. Sci.* 104, 1885–1898
5. Kesik-Brodacka, M. (2018) Progress in biopharmaceutical development. *Biotechnol. Appl. Biochem.* 65, 306–322
6. Harrer, S. *et al.* (2019) Artificial intelligence for clinical trial design. *Trends Pharmacol. Sci.* 40, 577–591
7. Kelley, B. (2020) Developing therapeutic monoclonal antibodies at pandemic pace. *Nat. Biotechnol.* 38, 540–545
8. Schneider, P. *et al.* (2020) Rethinking drug design in the artificial intelligence era. *Nat. Rev. Drug Discov.* 19, 353–364
9. Romero, P.A. *et al.* (2013) Navigating the protein fitness landscape with Gaussian processes. *Proc. Natl. Acad. Sci. U. S. A.* 110, E193–E201
10. Vamathevan, J. *et al.* (2019) Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* 18, 463–477
11. Ching, T. *et al.* (2018) Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* 15, 20170387
12. Yang, K.K. *et al.* (2019) Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* 16, 687–694
13. Chan, H.C.S. *et al.* (2019) Advancing drug discovery via artificial intelligence. *Trends Pharmacol. Sci.* 40, 592–604
14. Bojar, D. and Fussenegger, M. (2020) The role of protein engineering in biomedical applications of mammalian synthetic biology. *Small* 16, e1903093
15. Saito, Y. *et al.* (2018) Machine-learning-guided mutagenesis for directed evolution of fluorescent proteins. *ACS Synth. Biol.* 7, 2014–2022
16. Bedbrook, C.N. *et al.* (2017) Machine learning to design integral membrane channelrhodopsins for efficient eukaryotic expression and plasma membrane localization. *PLoS Comput. Biol.* 13, e1005786

17. Greener, J.G. *et al.* (2018) Design of metalloproteins and novel protein folds using variational autoencoders. *Sci. Rep.* 8, 16189
18. Riesselman, A. *et al.* (2019) Accelerating protein design using autoregressive generative models. *BioRxiv* Published online September 5, 2019. http://doi.org/10.1101/757252
19. Ingraham, J. *et al.* (2019) Generative models for graph-based protein design. In *Proceedings of the Internation Conference Advances in Neural Information Processing Systems* (Vol. 32), pp. 15820–15831, Curran Associates, Inc.
20. Karimi, M. *et al.* (2020) De novo protein design for novel folds using guided conditional Wasserstein generative adversarial networks (gcWGAN). *J. Chem. Inf. Model* 60, 5667–5681
21. Sinai, S. *et al.* (2018) Variational auto-encoding of protein sequences. *ArXiv* Published online January 3, 2018. http://arxiv.org/abs/1712.03346
22. Bepler, T. and Berger, B. (2019) *Learning protein sequence embeddings using information from structure*. 7th Int. Conf. Learn. Represent. ICLR 2019
23. Yang, K.K. *et al.* (2018) Learned protein embeddings for machine learning. *Bioinformatics* 34, 2642–2648
24. Alley, E.C. *et al.* (2019) Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* 16, 1315–1322
25. Rao, R. *et al.* (2019) Evaluating protein transfer learning with TAPE. *Adv. Neural. Inf. Process Syst.* 32, 9689–9701
26. Müller, A.T. *et al.* (2018) Recurrent neural network model for constructive peptide design. *J. Chem. Inf. Model.* 58, 472–479
27. Khosravian, M. *et al.* (2013) Predicting antibacterial peptides by the concept of Chou's pseudo-amino acid composition and machine learning methods. *Protein Pept. Lett.* 20, 180–186
28. Sanders, W.S. *et al.* (2011) Prediction of cell penetrating peptides by support vector machines. *PLoS Comput. Biol.* 7, e1002101
29. Fjell, C.D. *et al.* (2009) Identification of novel antibacterial peptides by chemoinformatics and machine learning. *J. Med. Chem.* 52, 2006–2015
30. Norman, R.A. *et al.* (2020) Computational approaches to therapeutic antibody design: established methods and emerging trends. *Brief. Bioinform.* 21, 1549–1567
31. Traggiai, E. *et al.* (2004) An efficient method to make human monoclonal antibodies from memory B cells: potent neutralization of SARS coronavirus. *Nat. Med.* 10, 871–875
32. Reddy, S.T. *et al.* (2010) Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. *Nat. Biotechnol.* 28, 965–969
33. Jones, P.T. *et al.* (1986) Replacing the complementarity-determining regions in a human antibody with those from a mouse. *Nature* 321, 522–525
34. Sormanni, P. *et al.* (2018) Third generation antibody discovery methods: in silico rational design. *Chem. Soc. Rev.* 47, 9137–9157
35. Singh, H. *et al.* (2013) Improved method for linear B-cell epitope prediction using antigen's primary sequence. *PLoS One* 8, e62216
36. Cheng, B. *et al.* (2018) Prediction of continuous B-cell epitopes using long short term memory networks. *ACM Int. Conf. Proceeding Ser.* 55–59
37. Liu, L.Y. *et al.* (2019) Prediction of linear B-cell epitopes based on PCA and RNN network. In *Proceedings of the 2019 IEEE 7th International Conference on Bioinformatics and Computational Biology (ICBCB 2019)*, pp. 39–43, IEEE
38. Sun, P. *et al.* (2020) B-cell epitope prediction method based on deep ensemble architecture and sequences. In *Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 94–97
39. Olimpieri, P.P. *et al.* (2013) Prediction of site-specific interactions in antibody-antigen complexes: the proABC method and server. *Bioinformatics* 29, 2285–2291
40. Daberdaku, S. and Ferrari, C. (2019) Antibody interface prediction with 3D Zernike descriptors and SVM. *Bioinformatics* 35, 1870–1876
41. Liberis, E. *et al.* (2018) Parapred: antibody paratope prediction using convolutional and recurrent neural networks. *Bioinformatics* 34, 2944–2950
42. Deac, A. *et al.* (2019) Attentive cross-modal paratope prediction. *J. Comput. Biol.* 26, 536–545
43. Akbar, R. *et al.* (2019) A compact vocabulary of paratope–epitope interactions enables predictability of antibody–antigen binding. *BioRxiv* Published online September 9, 2019. https://doi.org/10.1101/759498
44. Mason, D.M. *et al.* (2019) Deep learning enables therapeutic antibody optimization in mammalian cells by deciphering high-dimensional protein sequence space. *BioRxiv* Published online May 30, 2019. https://doi.org/10.1101/617860
45. Liu, G. *et al.* (2020) Antibody complementarity determining region design using high-capacity machine learning. *Bioinformatics* 36, 2126–2133
46. Amimeur, T. *et al.* (2020) Designing feature-controlled humanoid antibody discovery libraries using generative adversarial networks. *BioRxiv* Published online April 13, 2019. https://doi.org/10.1101/2020.04.12.024844
47. Senior, A.W. *et al.* (2019) Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). *Proteins Struct. Funct. Bioinforma.* 87, 1141–1148
48. AlQuraishi, M. (2019) End-to-end differentiable learning of protein structure. *Cell Syst.* 8, 292–301
49. Zurdo, J. (2013) Developability assessment as an early de-risking tool for biopharmaceutical development. *Pharm. Bioprocess.* 1, 29–50
50. Bailly, M. *et al.* (2020) Predicting antibody developability profiles through early stage discovery screening. *MAbs* 12, 1743053
51. Liu, Y. *et al.* (2014) High-throughput screening for developability during early-stage antibody discovery using self-interaction nanoparticle spectroscopy. *MAbs* 6, 483–492
52. Wu, J. *et al.* (2015) Discovery of highly soluble antibodies prior to purification using affinity-capture self-interaction nanoparticle spectroscopy. *Protein Eng. Des. Sel.* 28, 403–414
53. Wen, J. *et al.* (2020) Nano differential scanning fluorimetry for comparability studies of therapeutic proteins. *Anal. Biochem.* 593, 113581
54. Avery, L.B. *et al.* (2018) Establishing in vitro in vivo correlations to screen monoclonal antibodies for physicochemical properties related to favorable human pharmacokinetics. *MAbs* 10, 244–255
55. Xu, Y. *et al.* (2019) Structure, heterogeneity and developability assessment of therapeutic antibodies. *MAbs* 11, 239–264
56. Jain, T. *et al.* (2017) Biophysical properties of the clinical-stage antibody landscape. *Proc. Natl. Acad. Sci. U. S. A.* 114, 944–949
57. Hebditch, M. and Warwicker, J. (2019) Charge and hydrophobicity are key features in sequence-trained machine learning models for predicting the biophysical properties of clinical-stage antibodies. *PeerJ* 7, e8199
58. Gentiluomo, L. *et al.* (2019) Application of interpretable artificial neural networks to early monoclonal antibodies development. *Eur. J. Pharm. Biopharm.* 141, 81–89
59. Gentiluomo, L. *et al.* (2020) Application of machine learning to predict monomer retention of therapeutic proteins after long term storage. *Int. J. Pharm.* 577, 119039
60. Roberts, C.J. (2014) Therapeutic protein aggregation: mechanisms, design, and control. *Trends Biotechnol.* 32, 372–380
61. King, A.C. *et al.* (2011) High-throughput measurement, correlation analysis, and machine-learning predictions for pH and thermal stabilities of Pfizer-generated antibodies. *Protein Sci.* 20, 1546–1557
62. Sankar, K. *et al.* (2018) AggScore: prediction of aggregation-prone regions in proteins based on the distribution of surface patches. *Proteins Struct. Funct. Bioinforma.* 86, 1147–1156
63. Lauer, T.M. *et al.* (2012) Developability Index: a rapid in silico tool for the screening of antibody aggregation propensity. *J. Pharm. Sci.* 101, 102–115
64. Sharma, V.K. *et al.* (2014) In silico selection of therapeutic antibodies for development: viscosity, clearance, and chemical stability. *Proc. Natl. Acad. Sci. U. S. A.* 111, 18601–18606
65. Sormanni, P. *et al.* (2015) The CamSol method of rational design of protein mutants with enhanced solubility. *J. Mol. Biol.* 427, 478–490

66. Fang, J. (2020) A critical review of five machine learning-based algorithms for predicting protein stability changes upon mutation. *Brief. Bioinform.* 21, 1285–1292
67. Cao, H. *et al.* (2019) DeepDDG: predicting the stability change of protein point mutations using neural networks. *J. Chem. Inf. Model.* 59, 1508–1514
68. Jokinen, E. *et al.* (2018) MGPfusion: predicting protein stability changes with Gaussian process kernel learning and data fusion. *Bioinformatics* 34, i274–i283
69. Rawi, R. *et al.* (2018) PaRSnIP: sequence-based protein solubility prediction using gradient boosting machine. *Bioinformatics* 34, 1092–1098
70. Khurana, S. *et al.* (2018) DeepSol: a deep learning framework for sequence-based protein solubility prediction. *Bioinformatics* 34, 2605–2613
71. Han, X. *et al.* (2019) ProGAN: protein solubility generative adversarial nets for data augmentation in DNN framework. *Comput. Chem. Eng.* 131, 106533
72. Yang, Y. *et al.* (2016) PON-Sol: prediction of effects of amino acid substitutions on protein solubility. *Bioinformatics* 32, 2032–2034
73. Obrezanova, O. *et al.* (2015) Aggregation risk prediction for antibodies and its application to biotherapeutic development. *MAbs* 7, 352–363
74. Delmar, J.A. *et al.* (2019) Machine learning enables accurate prediction of asparagine deamidation probability and rate. *Mol. Ther. Methods Clin. Dev.* 15, 264–274
75. Jia, L. and Sun, Y. (2017) Protein asparagine deamidation prediction based on structures with machine learning methods. *PLoS One* 12, e0181347
76. Aledo, J.C. *et al.* (2017) A machine learning approach for predicting methionine oxidation sites. *BMC Bioinforma.* 18, 430
77. Sankar, K. *et al.* (2018) Prediction of methionine oxidation risk in monoclonal antibodies using a machine learning method. *MAbs* 10, 1281–1290
78. Yang, R. *et al.* (2017) Rapid assessment of oxidation via middle-down LCMS correlates with methionine side-chain solvent-accessible surface area for 121 clinical stage monoclonal antibodies. *MAbs* 9, 646–653
79. Falconer, R.J. (2019) Advances in liquid formulations of parenteral therapeutic proteins. *Biotechnol. Adv.* 37, 107412
80. Yang, Y. *et al.* (2019) Deep learning for in vitro prediction of pharmaceutical formulations. *Acta Pharm. Sin. B* 9, 177–185
81. Johnson, D.H. *et al.* (2009) High-throughput self-interaction chromatography: Applications in protein formulation prediction. *Pharm. Res.* 26, 296–305
82. Calderon, C.P. *et al.* (2018) Deep convolutional neural network analysis of flow imaging microscopy data to classify subvisible particles in protein formulations. *J. Pharm. Sci.* 107, 999–1008
83. Paganini, C. *et al.* (2019) Scalable production and isolation of extracellular vesicles: available sources and lessons from current industrial bioprocesses. *Biotechnol. J.* 14, e1800528
84. Kis, Z. *et al.* (2019) Emerging technologies for low-cost, rapid vaccine manufacture. *Biotechnol. J.* 14, e1800376
85. Elverum, K. and Whitman, M. (2020) Delivering cellular and gene therapies to patients: solutions for realizing the potential of the next generation of medicine. *Gene Ther.* 27, 537–544
86. Kopp, M.R.G. and Arosio, P. (2018) Microfluidic approaches for the characterization of therapeutic proteins. *J. Pharm. Sci.* 107, 1228–1236
87. Kopp, M.R.G. *et al.* (2018) Microfluidic diffusion analysis of the size distribution and microrheological properties of antibody solutions at high concentrations. *Ind. Eng. Chem. Res.* 57, 7112–7120
88. Narayanan, H. *et al.* (2020) Bioprocessing in the digital age: the role of process models. *Biotechnol. J.* 15, e1900172
89. Steinwandter, V. *et al.* (2019) Data science tools and applications on the way to Pharma 4.0. *Drug Discov. Today* 24, 1795–1805
90. Ain, Q.U. *et al.* (2015) Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 5, 405–424
91. Lavecchia, A. (2015) Machine-learning approaches in drug discovery: methods and applications. *Drug Discov. Today* 20, 318–331
92. Lo, Y.C. *et al.* (2018) Machine learning in chemoinformatics and drug discovery. *Drug Discov. Today* 23, 1538–1546
93. Korkmaz, S. *et al.* (2015) MLViS: A web tool for machine learning-based virtual screening in early-phase of drug discovery and development. *PLoS One* 10, e0124600
94. Dixon, S.L. *et al.* (2016) AutoQSAR: an automated machine learning tool for best-practice quantitative structure-activity relationship modeling. *Future Med. Chem.* 8, 1825–1839
95. Pereira, J.C. *et al.* (2016) Boosting docking-based virtual screening with deep learning. *J. Chem. Inf. Model.* 56, 2495–2506
96. Ericksen, S.S. *et al.* (2017) Machine learning consensus scoring improves performance across targets in structure-based virtual screening. *J. Chem. Inf. Model.* 57, 1579–1590
97. Tian, K. *et al.* (2016) Boosting compound–protein interaction prediction by deep learning. *Methods* 110, 64–72
98. Gonczarek, A. *et al.* (2018) Interaction prediction in structure-based virtual screening using deep learning. *Comput. Biol. Med.* 100, 253–258
99. Sanchez-lengeling, B. (2018) Inverse molecular design using machine learning: generative models for matter engineering. *Science* 361, 360–365
100. Segler, M.H.S. *et al.* (2018) Models generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* 4, 120–131
101. Merk, D. *et al.* (2018) De novo design of bioactive small molecules by artificial intelligence. *Mol. Inform.* 37, 1700153
102. Yang, X. *et al.* (2017) ChemTS: an efficient python library for de novo molecular generation. *Sci. Technol. Adv. Mater.* 18, 972–976
103. Maziarka, Ł. *et al.* (2020) Mol-CycleGAN: a generative model for molecular optimization. *J. Cheminform.* 12, 2
104. Tao, L. *et al.* (2015) Recent progresses in the exploration of machine learning methods as in-silico ADME prediction tools. *Adv. Drug Deliv. Rev.* 86, 83–100
105. Maltarollo, V.G. *et al.* (2015) Applying machine learning techniques for ADME-Tox prediction: a review. *Expert Opin. Drug Metab. Toxicol.* 11, 259–271
106. Bhhatarai, B. *et al.* (2019) Opportunities and challenges using artificial intelligence in ADME/Tox. *Nat. Mater.* 18, 418–422
107. Korotcov, A. *et al.* (2017) Comparison of deep learning with multiple machine learning methods and metrics using diverse drug discovery data sets. *Mol. Pharm.* 14, 4462–4475
108. Mayr, A. *et al.* (2018) Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chem. Sci.* 9, 5441–5451
109. Olivecrona, M. *et al.* (2017) Molecular de-novo design through deep reinforcement learning. *J. Cheminform.* 9, 48
110. Popova, M. *et al.* (2018) Deep reinforcement learning for de novo drug design. *Sci. Adv.* 7, eaap7885
111. Zhou, Z. (2019) Optimization of molecules via deep reinforcement learning. *Sci. Rep.* 9, 10752
112. Fu, T. *et al.* (2020) CORE: automatic molecule optimization using copy & refine strategy. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34), pp. 638–645
113. Papathanasiou, M.M. and Kontoravdi, C. (2020) Engineering challenges in therapeutic protein product and process design. *Curr. Opin. Chem. Eng.* 27, 81–88
114. Gu, Y. *et al.* (2019) Machine learning based real-time image-guided cell sorting and classification. *Cytom. Part A* 95, 499–509
115. Nitta, N. *et al.* (2018) Intelligent image-activated cell sorting. *Cell* 175, 266–276
116. Sokolov, M. *et al.* (2017) Robust factor selection in early cell culture process development for the production of a biosimilar monoclonal antibody. *Biotechnol. Prog.* 33, 181–191
117. Zobel-Roos, S. *et al.* (2019) Accelerating biologics manufacturing by modeling or: is approval under the QbD and PAT approaches demanded by authorities acceptable without a digital-twin? *Processes* 7, 94

118. Guerra, A. *et al.* (2019) Toward biotherapeutic product real-time quality monitoring. *Crit. Rev. Biotechnol.* 39, 289–305
119. Sokolov, M. *et al.* (2018) Sequential multivariate cell culture modeling at multiple scales supports systematic shaping of a monoclonal antibody toward a quality target. *Biotechnol. J.* 13, e1700461
120. Brühlmann, D. *et al.* (2017) Parallel experimental design and multivariate analysis provides efficient screening of cell culture media supplements to improve biosimilar product quality. *Biotechnol. Bioeng.* 114, 1448–1458
121. Bayrak, E.S. *et al.* (2018) Product attribute forecast: adaptive model selection using real-time machine learning. *IFAC-PapersOnLine* 51, 121–125