Best practices for machine learning in antibody discovery and development

Leonard Wossnig [1,2,*], Norbert Furtmann [3], Andrew Buchanan [4], Sandeep Kumar [5], Victor Greiff [6]

[1] LabGenius Ltd, The Biscuit Factory, 100 Drummond Road, London SE16 4DG, UK
[2] Department of Computer Science, University College London, 66–72 Gower St, London WC1E 6EA, UK
[3] R&D Large Molecules Research Platform, Sanofi Deutschland GmbH, Industriepark Höchst, Frankfurt Am Main, Germany
[4] Biologics Engineering, R&D, AstraZeneca, Cambridge CB2 0AA, UK
[5] Computational Protein Design and Modeling Group, Computational Science, Moderna Therapeutics, 200 Technology Square, Cambridge, MA 02139, USA
[6] Department of Immunology and Oslo University Hospital, University of Oslo, Oslo, Norway

In the past 40 years, therapeutic antibody discovery and development have advanced considerably, with machine learning (ML) offering a promising way to speed up the process by reducing costs and the number of experiments required. Recent progress in ML-guided antibody design and development (D&D) has been hindered by the diversity of data sets and evaluation methods, which makes it difficult to conduct comparisons and assess utility. Establishing standards and guidelines will be crucial for the wider adoption of ML and the advancement of the field. This perspective critically reviews current practices, highlights common pitfalls and proposes method development and evaluation guidelines for various ML-based techniques in therapeutic antibody D&D. Addressing challenges across the ML process, best practices are recommended for each stage to enhance reproducibility and progress.

Leonard Wossnig is the chief technical officer at LabGenius, where he leads the development of LabGenius's data-driven platform capabilities and the application of computational methods to the company's antibody drug discovery programs. He is an honorary research fellow in computer science at University College London and a fellow of the Royal Society of Biology. Prior to LabGenius, he was vice president of AI at Odyssey Therapeutics, leading a team of machine learning experts to develop a computational platform for generative drug design in the areas of cancer and inflammation. In 2018, he co-founded Rahko, a machine learning for drug discovery company that he led from inception to acquisition in 2021. He holds a PhD from University College London, where he was a Google PhD fellow, and is a scientific advisor to multiple start-up companies, such as Shift Biosciences.

Victor Greiff has been an associate professor for systems immunology at the University of Oslo since 2018. His group develops machine learning, computational and experimental tools for immune-repertoire-based *in silico* immunodiagnostic and immunotherapeutic discovery and design. He is the president of the Norwegian Society for Immunology and the chair of the AIRR Community. He received his PhD in systems immunology from Humboldt University (Germany, 2012) and performed his postdoctoral training at ETH Zürich (Switzerland, 2013–17).

Keywords: machine learning; data curation; drug discovery; data standardisation; FAIR data; model performance; metrics; model evaluation; antibodies; protein language models

* *Corresponding author.* Wossnig, L. (*leonard.wossnig@labgeni.us*)

1359-6446/© 2024 The Author(s). Published by Elsevier Ltd.
https://doi.org/10.1016/j.drudis.2024.104025

Introduction

The development of antibody-based drugs has revolutionised the field of medicine, providing effective treatments for a wide range of diseases.[p1],[p2] However, although the drug-discovery process has proved effective, it is complex, expensive, time-consuming and has room for improvement.[p3],[p4],[p5],[p6],[p7] Recent advances in machine learning (ML) have the potential to accelerate and optimise this process by enabling the identification of better biotherapeutic drug candidates more rapidly, and thereby reducing the cost and timelines for antibody drug discovery.[p8],[p9],[p10]

The rapidly evolving landscape of ML-guided therapeutic antibody design and development (D&D) holds immense potential for the biopharmaceutical industry. However, although initial success stories have emerged, the 'real world' impact of ML-guided therapeutic antibody D&D has thus far been relatively minimal.[p11] To unlock its full potential and make a significant impact on commercial drug discovery and development, standardised guidelines and best practices for applications of ML must be established at every step of biologic drug discovery and development. These guidelines and best practices must cover processes such as the *in silico* design of antibody candidate drugs; the computational identification or design of high-affinity, specific-function-relevant epitopes; the accurate prediction of biophysical attributes; screening for formulations; and the development of digital twins of manufacturing processes.[p12],[p13]

Improving the probability of success in clinical trials generally has the highest positive impact on the costs and timelines for an individual program.[p14] Nevertheless, a reduction of time and costs in the preclinical stages can still enable significant overall savings owing to the high volume of early-stage programs.[p15] Therefore, in recent years there has been a surge in the application of ML-based models at each stage of the drug discovery and development cycle. However, little to no attention is being paid to benchmarking the general applicability of these models, as well as benchmarking practices.

This review article aims to address the need for establishing best practices for the use of ML in the biopharmaceutical industry. We critically examine current methodologies, identifying common pitfalls and providing recommendations for ML-based approaches to therapeutic antibody D&D, akin to other reviews for chemistry or broader ML research.[p16],[p17],[p18],[p19] Unlike other reviews,[p20],[p21],[p22],[p23],[p24] we focus on data aspects (see Table 1 and Box 1) and model validation (see Fig. 1 for an overview).

BOX 1 IIn the context of enabling ML, important aspects of data generation include:

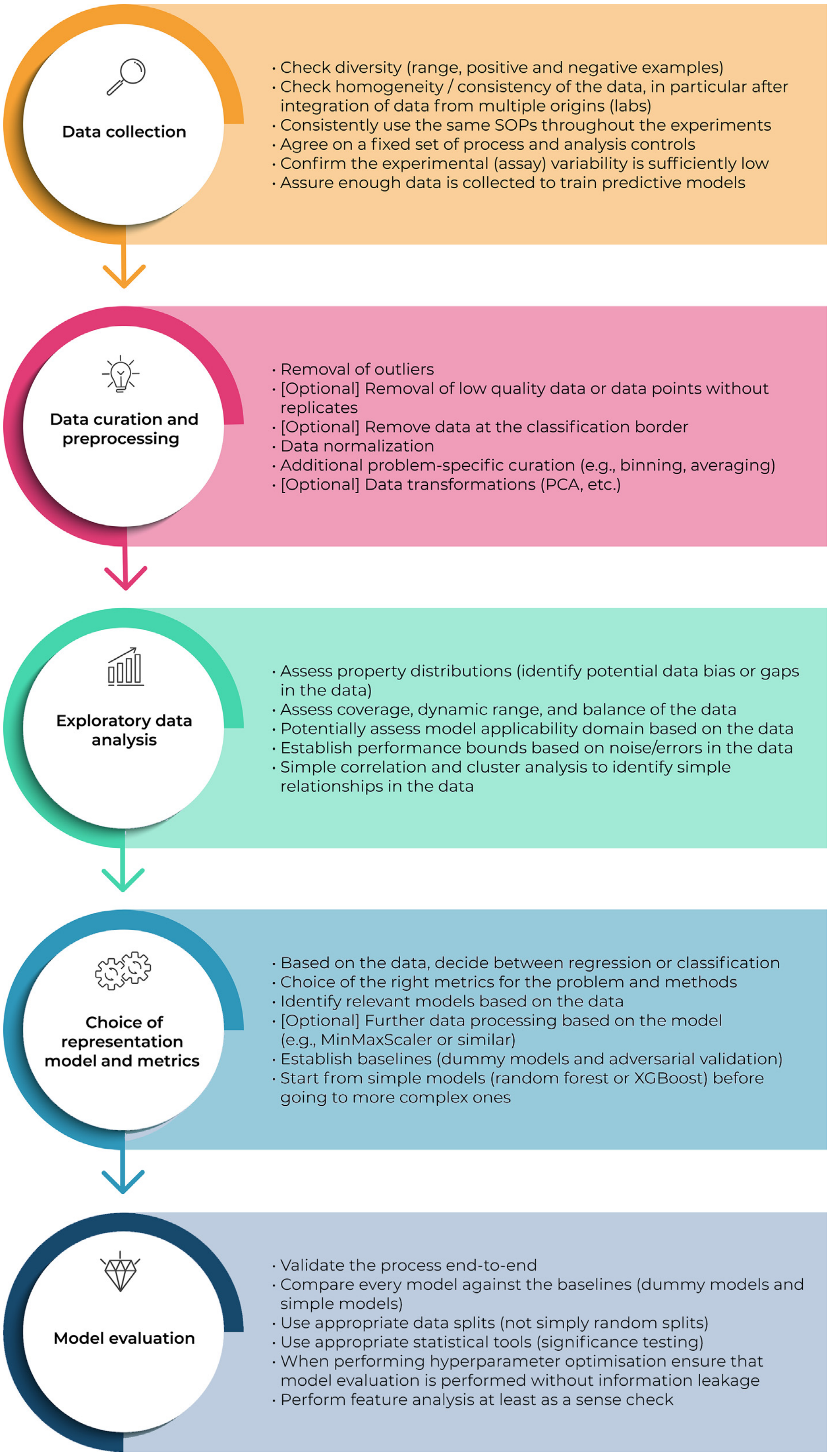
- Agreed upon standard protocols (including plate layouts) for all assays if data are generated internally.
- Capturing data with minimal manual handling and storing them in a findable, accessible, interoperable and reusable (FAIR) way.[p192] If pre-existing data, such as public data, are used, check the origin (lab), assay type of the data set, assay variability, readout and units of measurement, and confirm that these match. Apply stringent filter criteria to remove data that are not suitable for ML (e.g., data with too high variability or outliers). Check whether enough data remain to train a model.[p87],[p193],[p194] A general guideline is that there should be more data points than model parameters.
- To enable team-wide tracking of data processing, establish data and processing lineage and versioning.
- Employ technical repeats on the plate controls for assay quality controls (robust Z-prime and variance of repeats) and track assay behaviour (e.g., drift). The robustness and reproducibility of the assay protocol need to be validated beforehand. This should include process controls (e.g., for expression).
- We advise deploying clear acceptance criteria for data quality before use in modelling.
- For data from multiple classes (e.g., two classes in the case of a quality-control method: fail/pass), ensure that data points are available for each class[p195]
- For regression, check that there are sufficient real-value data points. If there are too many data points with a cut-off (e.g., pIC$_{50}$ > 100 IM), there might not be sufficient data to train a regression model, and only a classifier might be possible.
- If real-valued data (not multiple classes) are used, the data should include a wide range (e.g., pEC$_{50}$ values of 7 to 11).
- Track relevant confounding variables through metadata.

TABLE 1

Illustration of the data generation from a typical assay cascade[a]

Discovery platform	Platform output	No. of variants	Technical repeats	Readout	Z′
Primary (1–2 point) screen	Function in human assay Sequence diversity and liabilities	1,000–30,000	1	Rank order	>0.4
Confirmation (3–5 point) screen	Function in species-relevant assays	100 s	1–3	Approx. IC$_{50}$, ±SE	0.5–0.7
Profiling *in vitro* (12 point) assay	Affinity	10 s	3	K_D	>0.9
	Endogenous cell-based function		>3	IC$_{50}$, ±SE	>0.8
	Developability		3	Risks ranked	–
Profiling *in vivo*	Efficacy	1–5	–	Pharmacodynamic (PD) end point	–
Early lead	Optimised or final candidate selection for First time in human (FTiH)	1–3	–	–	–

[a] The details of throughput, repeats, data quality assessed by Z′ and end points will be dependent upon aspects of the platform and biology of the target.



Data collection

- Check diversity (range, positive and negative examples)
- Check homogeneity / consistency of the data, in particular after integration of data from multiple origins (labs)
- Consistently use the same SOPs throughout the experiments
- Agree on a fixed set of process and analysis controls
- Confirm the experimental (assay) variability is sufficiently low
- Assure enough data is collected to train predictive models

Data curation and preprocessing

- Removal of outliers
- [Optional] Removal of low quality data or data points without replicates
- [Optional] Remove data at the classification border
- Data normalization
- Additional problem-specific curation (e.g., binning, averaging)
- [Optional] Data transformations (PCA, etc.)

Exploratory data analysis

- Assess property distributions (identify potential data bias or gaps in the data)
- Assess coverage, dynamic range, and balance of the data
- Potentially assess model applicability domain based on the data
- Establish performance bounds based on noise/errors in the data
- Simple correlation and cluster analysis to identify simple relationships in the data

Choice of representation model and metrics

- Based on the data, decide between regression or classification
- Choice of the right metrics for the problem and methods
- Identify relevant models based on the data
- [Optional] Further data processing based on the model (e.g., MinMaxScaler or similar)
- Establish baselines (dummy models and adversarial validation)
- Start from simple models (random forest or XGBoost) before going to more complex ones

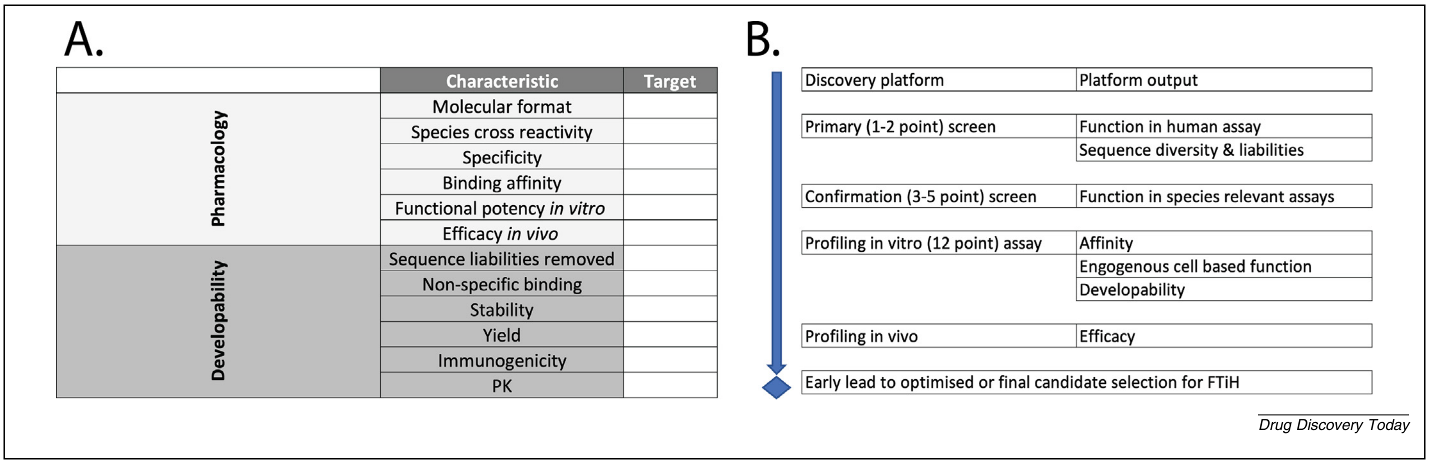
Model evaluation

- Validate the process end-to-end
- Compare every model against the baselines (dummy models and simple models)
- Use appropriate data splits (not simply random splits)
- Use appropriate statistical tools (significance testing)
- When performing hyperparameter optimisation ensure that model evaluation is performed without information leakage
- Perform feature analysis at least as a sense check

Drug Discovery Today

FIG. 1

Overview of the entire machine learning (ML) process for antibody R&D, from data collection to model evaluation. Abbreviations: PCA, principal component analysis; SOP, standard of practice.



A.

	Characteristic	Target
Pharmacology	Molecular format	
	Species cross reactivity	
	Specificity	
	Binding affinity	
	Functional potency *in vitro*	
	Efficacy *in vivo*	
Developability	Sequence liabilities removed	
	Non-specific binding	
	Stability	
	Yield	
	Immunogenicity	
	PK	

B.

Discovery platform	Platform output
Primary (1-2 point) screen	Function in human assay Sequence diversity & liabilities
Confirmation (3-5 point) screen	Function in species relevant assays
Profiling in vitro (12 point) assay	Affinity Engogenous cell based function Developability
Profiling in vivo	Efficacy
Early lead to optimised or final candidate selection for FTiH	

Drug Discovery Today

FIG. 2

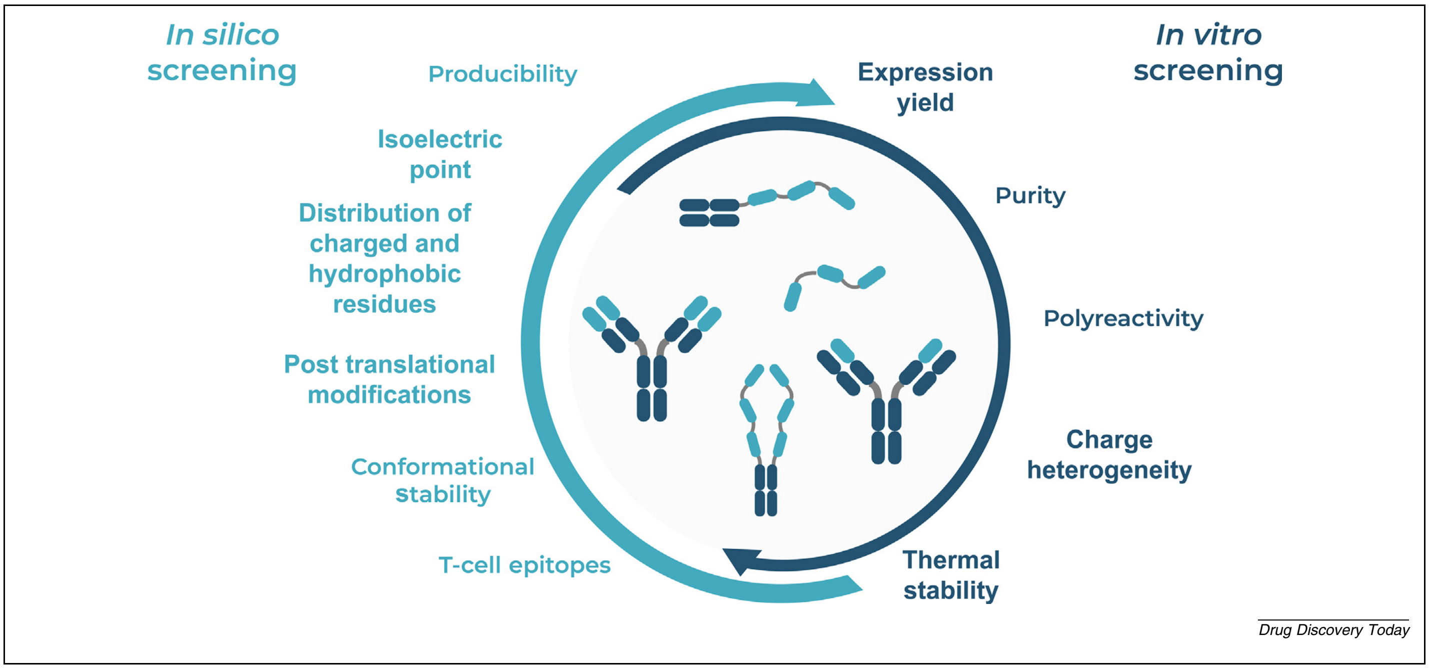
Exemplary candidate drug target profile and assay cascade. **(a)** An illustrative CDTP for any biologic modality, covering key attributes of both pharmacology and developability. **(b)** Exemplary assay cascade for lead selection during D&D. Successful completion may be achieved following one or a number of iterations of the cascade dependent upon aspects of the platform and biology of the target.

Define an appropriate experimental strategy for antibody D&D

Prior to discussing the intricacies of the ML process in antibody drug discovery and design, any antibody discovery program needs to have a clear experimental strategy aligned with the candidate drug target profile (CDTP) (see Table 1).[p25] The CDTP should include defined targets for both the desired pharmacology and the appropriate developability package (Fig. 2a). This specification sheet will inform the assay cascade for lead generation, optimisation and final candidate selection (Fig. 2b).

Building confidence in both the target and the candidate molecule's pharmacology and developability are key milestones before larger investments in drug product and clinical development are made. To maximise the probability of success, it is crucial to establish assays that relate to species cross-reactivity, desired mechanism of action and aspects of developability (Fig. 3). Focusing on function rather than affinity or other simplistic properties enables the identification of rarer functional biologics, especially in the context of complex targets, agonists and multi-specific or multivalent biologics.[p26],[p27],[p28] For example, affinity optimisation can have a limited correlation with the desired complex function (Fig. 4).

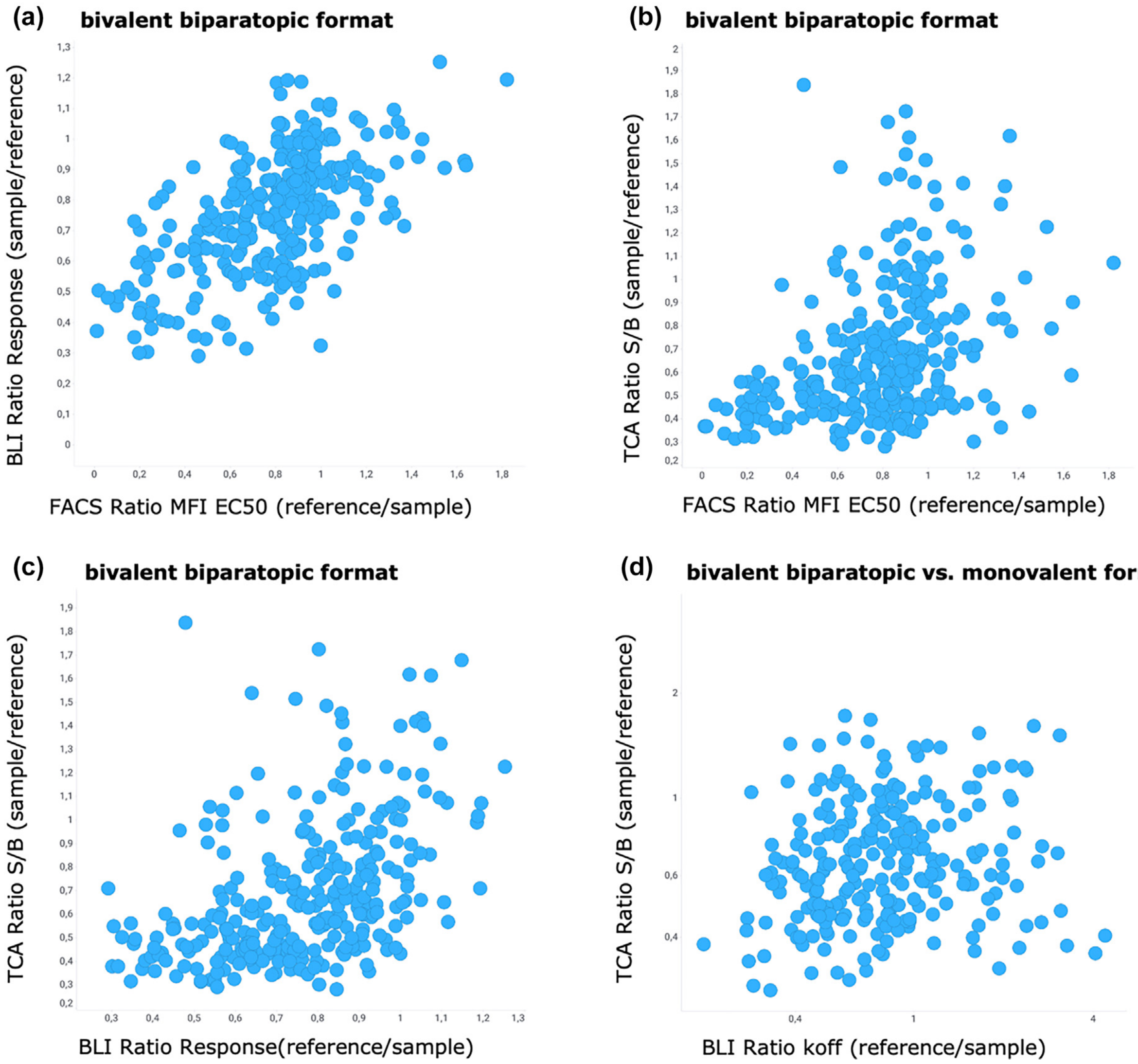
Designing a project-specific assay cascade entails balancing assay feasibility, throughput, robustness, data quality and translational relevance. The integration of wet-lab automation with the goal of generating structured, consistent, machine-readable data will enhance data-generation efficiency and accuracy, and is hence essential for the successful application of ML.[p29],[p30]



Drug Discovery Today

FIG. 3

Key properties that are typically co-optimised to achieve specified antibody developability attributes of the CDTP.



Drug Discovery Today

FIG. 4

This figure illustrates correlations between binding and activity behaviours, assessed using different technologies for VHH domains in both mono- and bivalent (biparatopic) formats against a specific therapeutic target that requires agonistic activity. VHH-target binding was evaluated through biolayer interferometry (BLI), cell binding through fluorescence-activated cell sorting (FACS) and activity within a T-cell activation assay. In this context, agonistic activity is achievable only by combining two binding domains against the same target within a biparatopic format. **(a)** Binding against the immobilised target measured via BLI (response) correlates with cell binding assessed via FACS (EC_{50}) for compounds in the biparatopic format. **(b,c)** Binding to the immobilised target (BLI response) and binding to cells (FACS EC_{50}) poorly correlate with activity in the T-cell activation assay. **(d)** Binding (BLI off-rates) of monovalent building blocks does not correlate at all with the activation behaviour of the same variants in the biparatopic format in the T-cell activation assay.

Components of a (good) ML process

All applied ML requires process validation. Process validation, unlike model validation, is crucial because we need to validate that the entire process, from data collection and processing to the actual model predictions, is applicable to the given problem. For an overview, see Fig. 5.

An ML process consists of the following steps.

(i) Data collection

Obtain diverse, high-quality, and relevant data from relevant sources, including the scientific literature, public and commercial databases, and proprietary wet-lab experiments. Particular care needs to be taken when mixing data sources owing to high variability in execution and data-collection standards, particularly when dealing with human-labelled training data.[(p31)]

(ii) Data curation, preprocessing and standardisation

Clean, organise and transform the data to ensure consistency and reduce noise. It is helpful to adopt similar practices to the ones that have been established for quantitative structure–activity relationship (QSAR) models[(p32),(p33)]: that is, to ensure the data is collected using standardised experimental protocols and/or molecules with common formats.

(iii) Exploratory data analysis

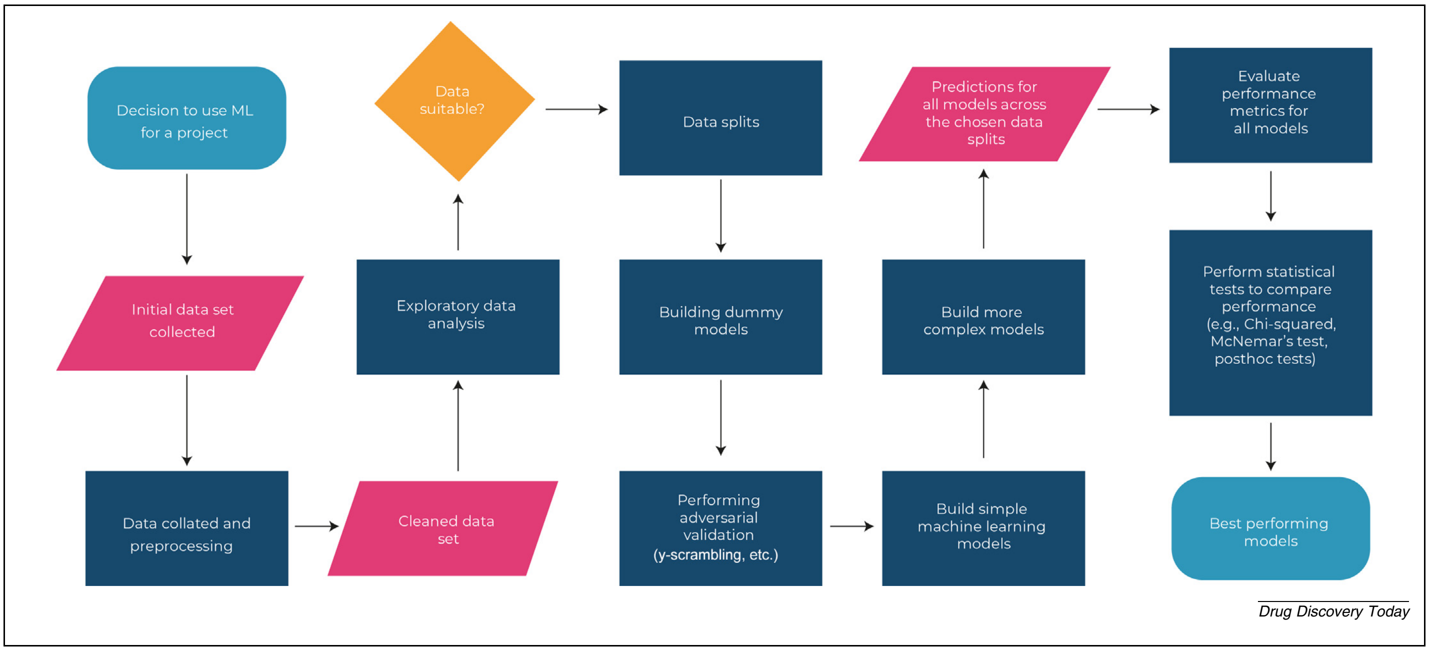

Decision to use ML for a project
Data suitable?
Data splits
Predictions for all models across the chosen data splits
Evaluate performance metrics for all models
Initial data set collected
Exploratory data analysis
Building dummy models
Build more complex models
Perform statistical tests to compare performance (e.g., Chi-squared, McNemar's test, posthoc tests)
Data collated and preprocessing
Cleaned data set
Performing adversarial validation (y-scrambling, etc.)
Build simple machine learning models
Best performing models
Drug Discovery Today

FIG. 5

Systematic process overview of the different ML evaluation steps. These steps are generic and apply to most ML processes. Light blue, final inputs and outputs; yellow, decision point; dark blue, process steps; pink, results of a processing/modelling step.

Examine the data to understand their characteristics, imbalances and distributions before building complex models. The collaboration of data scientists with experimental scientists is essential here for a deeper understanding of the data.[p34]

(iv) Choosing a model performance metric

Select appropriate metrics that align with the nature of the data and the application.[p18],[p35],[p36],[p37] Note that metrics might be different in specific applications: for example, the metrics used in protein structure prediction[p38],[p39],[p40] might be different from those used in the prediction of biophysical measures associated with expression, purification, conformational stability and the colloidal interactions of proteins.

(v) Model components and model choice

Determine the most suitable model on the basis of the complexity of the problem, the available data and other constraints.[p18] When mimicking biological processes such as the generation of unwanted immune responses against therapeutic antibodies, it might be essential to develop multiple models, each representing a specific step in the process, and then tie them together to obtain an improved understanding of the process itself.

(vi) Evaluation

Assess model performance using techniques like cross-validation, different data splits and comparison to baseline methods to understand realistic model performance. Based on our experience, consensus predictions or other ways of using multiple models will often perform better than individual models developed using specific ML methods.

(vii) Putting the model into production

Ensure scalability, computational efficiency, compatibility and interpretability, and monitor the data and model performance over time in a production environment. This might require the repeated validation and deployment of models as new data arrive, and hence the building of strong ML operations pipelines.

Data collection

In this section, we delve into the essential aspects and considerations to be aware of when gathering high-quality data for ML in a drug discovery setting. The goal of this step is to ensure that the collected data is accurate, relevant and suitable for training ML models to make reliable predictions.

Predictive validity of assays

The predictive validity of an assay refers to its ability to predict a desired outcome accurately.[p41] In this case, we refer to it as the likelihood of an assay outcome translating into a more complex experiment, which could be a more complicated assay or even a clinical trial. It is crucial to validate that the end point that is used correlates with the desired outcome whenever proxy assays are used (e.g., the end point/prediction needs to be correlated with go/no-go criteria). For example, a biochemical antagonism assay should first be validated to translate into a more complex cell-based readout. This is essential to ensure that the optimisation process guarantees useful candidates.

Determine the correct set-up of the assay

As previously discussed, deciding the number of repeats and type of assay that is used will have a large impact on the data quality and quantity. The correct set-up should be chosen and then maintained throughout the whole process, and the set-up chosen might vary depending on the developmental stage of the program. We further recommend establishing, maintaining and adhering to versioned business rules for all data (pre)processing.

These should encompass normalisation, transformation and scaling processes for the data.

Confirming data accuracy

Ensuring data accuracy involves proper assay calibration and reproducibility to confirm that measurements reflect the desired behaviour. Maintaining consistent conditions is crucial in ML, because later changes in the assay can introduce inconsistencies and confuse the model, resulting in poor performance. The use of control molecules and the conduction of regular quality assessments of the assays is advisable. Furthermore, the collection of metadata can help to improve the quality and understanding of the collected data sets.(p42),(p43) In particular, regular visual inspection through, for example, box- or scatterplots of the data allows the detection of errors, which can be subsequently incorporated as automated quality controls.

Choosing the correct measurement metric and process

Selecting a measurement metric that is appropriate for the specific ML problem [e.g., area under the curve (AUC), half maximal effective concentration (EC_{50}) or maximum activation/inhibition value] and process (including data) directly affects ML performance and the suitability of combining data from different sources.(p31) Consistency across data processing steps and protocol standardisation is key for optimal model training. Input from data-science colleagues must be sought to inform these decisions, but ultimately the decisions must be made by experimental experts. For example, when AUC values from different sources are used, it is essential to confirm that baseline subtraction, hook-effect removal or curve-fitting processes have been consistently applied to the data in order to ensure comparability.

Minimise biological variability

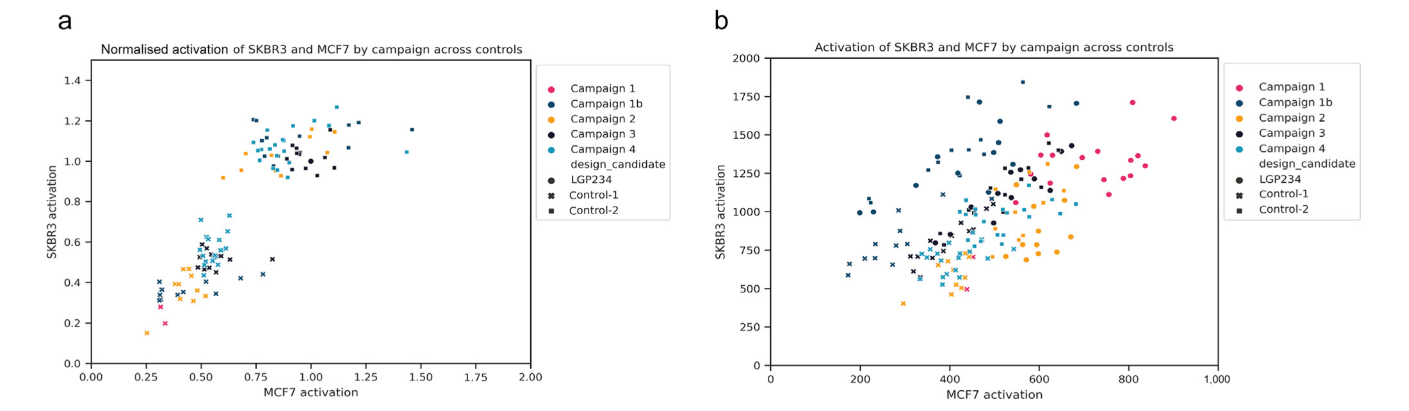
Biological data often exhibit inherent variability and noise, so it is important to rely on biological and technical repeats. Understanding the variance between repeated measurements helps to set a baseline for the best-case model performance, because a model cannot realistically outperform assay accuracy.(p44),(p45) This can be done by using the experimentally measured error distributions to simulate repeated measurements for a larger amount of data and then evaluating different correlation metrics on the basis of the resulting simulations, as done by Brown *et al.*(p46) The impact of experimental errors is usually more significant for data sets with a limited dynamic range and is less problematic for larger ranges. For example, when the error is nearly half of the data set's dynamic range, achieving a meaningful correlation becomes nearly impossible. This implies that for early projects (e.g., hit finding) experimental errors usually pose a smaller problem than in late-stage lead optimisation or when we want to increase the selectivity in each campaign by a small amount. Notably, the variability will vary between different assays. Variabilities in cell-based assays tend to be around 20–30%, whereas biophysical measurements usually exhibit much lower variability: for example, melting temperature can be accurate to 0.1° when done with calorimetry. Additionally, controls can be used to discard readouts that exhibit excessive variability.

Use data normalisation based on controls

Data normalisation using controls is essential when the data come from the same laboratory, as it helps to calibrate data, particularly for cellular assays. For example, normalising AUC measurements using a combination of plate controls and average control AUC ensures consistency (Fig. 6) or the normalisation of expression data across different plates, batches and days. However, this can be challenging when using public data sources or data from different service providers, which might lack the appropriate controls, harmonisation of experimental conditions and protocol standardisation. The use of controls to normalise data sets to a common reference is essential for improving the consistency of the data.(p47)

Other challenges with drug-discovery data

Drug discovery data often have set cut-offs, such as maximum concentration limits in concentration–response curves. They are derived from the specific targets of the program (e.g., the required target potency) and will cover a reasonable range around these targets. For example, concentration–response curves have a number of dilution steps and will hence cover a



Drug Discovery Today

FIG. 6

Data from different controls evaluated in cell-based assays from different campaigns before **(a)** and after **(b)** normalisation can dramatically change the picture. Here we clearly see two clusters (Control-1 and Control-2) emerging after normalisation.

few orders of magnitude of potencies. Although the values are usually set to the most relevant range for the program, the properties of the molecules might not always fall within these ranges, leading to values beyond the limits. For example, for an assay limit of 100 μM on the upper side, a compound might receive a cut-off value of >100 μM.

Such cut-offs can restrict the ability to train regression models, requiring the use of categorical models (classification) with lower-resolution predictions (e.g., >5 μM or <5 μM potency). Therefore, it is crucial to understand the type of data to collect to maximise the model's ability to make relevant predictions, especially when relying on public data sources.(p48),(p49),(p50)

A second challenge with drug-discovery data is the harmonisation of experimental data across different functional units of an organisation. For example, the data on protein expression, purification and biophysical characterization performed at preformulation stages might not be easily translatable to similar experiments performed at the CMC (chemistry, manufacturing and controls) development and formulation stages.

A third challenge with the experimental data collected in a discovery setting is time-lapse. Typically, it is harder to use data from older projects than from current ones, even if the same target is being worked on again: this is because there are likely to have been changes over time in the methods for used for data collection (e.g., the instruments) and data storage.

Detecting and dealing with data drift

Being aware of changes in experimental set-ups, such as reagent changes, equipment accuracy variations or data processing alterations, is vital, because these can have significant impacts on the ML models trained. Consistently monitoring and addressing data drift, for example via the deployment of an appropriate set of controls, helps to maintain model performance and ensures that any changes in experimental conditions do not go unnoticed. In addition, periodic updates of the model using the greater amounts of available experimental data are also recommended to minimise the impact of data drift and to maximise performance. We generally recommend the visualisation of process data alongside molecular profiles throughout the program.

Data curation and preprocessing

Antibody engineering by means of ML requires stringent data handling and preprocessing, which are the key to attaining reliable, meaningful outputs. Outliers, noisy data or data that might confuse a model [e.g., using a logarithmic half maximal inhibitory concentration (pIC_{50}) of 5 in classification to distinguish more clearly between active and inactive molecules] should be carefully addressed. Although when using AI it is generally advised to use as much data as possible, the removal of some data can be beneficial for the predictive performance of an ML model.

Data curation and preprocessing have a major impact on the behaviour and performance of an ML model. The cheminformatics community has over the years established clear guidelines on how to process and filter data to make them suitable for ML purposes.(p32),(p33),(p51),(p52),(p53),(p54)

Although there have been some attempts to provide similar guidelines for bioinformatics and ML for biology,(p55) most publications are limited to best practices for computational modelling, and do not offer advice for data preparation and quality.(p17),(p22),(p56),(p57),(p58),(p59),(p60),(p61) The AIRR (Adaptive Immune Receptor Repertoire) community has established guidelines for preprocessing antibody and T-cell receptor data that enable standardised data input and output, which enables the communication of AIRR-compliant bioinformatics tools.(p62),(p63),(p64),(p65),(p66)

In the following, we highlight key steps that should be taken in order to obtain good predictive performance and to ensure that this performance is representative of the actual application of a program.

Data integration

Many large company databases and public data sources are not ideal for training ML models owing to the lack of appropriate controls for normalisation, along with varying or changing assay protocols over time. Generally, better-quality data result in better model performance. Nonetheless, training models on combinations of different data sources (for example, a mix of public and in-house data, or different sources of public data) can be beneficial under certain circumstances.(p67),(p68),(p69) On the basis of a large-scale analysis of small-molecule (ChEMBL) data, Kramer *et al.*(p67),(p68),(p69) concluded that augmenting mixed public IC_{50} data with public K_i data does not deteriorate the quality of the mixed IC_{50} data if the K_i is corrected with an offset. However, in our experience, this is not the case for biologics, and consistent internal data typically lead to the best outcomes; this is because the production process has a much bigger impact on the final molecule, and variations in it have a non-proportional effect on the results. In order to decide how much data you need, a common rule of thumb is at least ten times as many data instances (data points) as there are data features. However, this depends strongly on the selected features, the quality of the data and the complexity of the problem.(p70)

Data cleaning

Model accuracy can be enhanced by removing noisy data points or outliers. For example, when training a classification model, excluding data points with $pIC_{50} = 5$ (commonly considered the threshold for inactive compounds) can improve model performance. This step might involve techniques such as outlier detection, data imputation and standardisation of units and formats. In many databases, active compounds are overrepresented owing to reporting biases. For example, Parks *et al.*(p71) reported that the average pIC_{50} value of the whole distribution of data in ChEMBL25 is 6.57 for small molecules. This is likely to be different from prospective unbiased library screens, where it is common for 95% of the compounds to have $pIC_{50} < 5$, a large train-test distribution shift. For biologics, these ratios will vary, but being aware of such changes in the distribution is important in order to allow adequate processing (e.g., resampling) of the data to create a more representative data distribution for the actual application.

Binning of data

For classification, it might be necessary to bin data into active and inactive groups. The choice of group ranges can significantly affect model performance. Initial surveys of the data, as well as feedback from the experimentalists, can help when setting appropriate data bins. For example, immunoglobulin G (IgG) antibody purification data is often obtained using size exclusion chromatography (SEC). The experimental result consists of a chromatogram showing relative abundance (peaks) for the extents of high molecular mass species, monomer content and low molecular mass species in terms of the percentages of the areas under the peaks. Percent monomer is often used as an indicator of quality of the antibodies when these measurements are performed over a large set of them. The range of percent monomers in these samples might vary from 0 to 100%. A typical quartile-based binning of the data without any inputs from the experts in the field might sort the samples into bins of poor quality (percent monomer below 25%), good quality (25–75%) and high quality (>75%). However, domain experts would often consider IgG antibodies with a percent monomer below 90% to be of poor quality, instead recognising a value of 90–95% as good quality and a value of > 95% as high quality.(p72),(p73)

Averaging data over technical and/or biological repeats

Data are typically averaged over technical and biological repeats, helping to reduce noise and error. If the data distribution is narrow, averaging is advisable. However, if the data points are disparate, outliers are better discarded, unless a clear rationale for their preservation is presented. Based on our experience, averaging repeated measurements and using the mean to train ML models is a best practice. However, individual repeats might be more suitable for certain models, especially when data errors or noise levels are used for calibrating model uncertainties, as is commonly the case for Bayesian methods.(p74),(p75)

Feature extraction and selection

Converting raw data into a set of features or descriptors is crucial for effectively training ML models in drug discovery. In the antibody or protein engineering space, features might include protein sequence or sequence-derived structural features, designs (e.g., a combination of combinatorial antibody parts), 3D structure, learned representations or more conventional molecular descriptors (amino acid composition, dipeptide composition, tripeptide composition or pseudo amino acid composition). The physicochemical properties of amino acids,(p76),(p77),(p78) such as hydrophobicity, charge, size and polarity, can also be used to compute various descriptors, including autocorrelation and Moran and Geary coefficients (e.g., see(p79)). Finally, the incorporation of evolutionary information, for example through multiple sequence alignment,(p80) has been particularly helpful for computational structure prediction.

For protein sequence-based features, values can be directly extracted from the sequences, such as amino acid type, evolutionary information [e.g., profile representations in the form of position-specific scoring matrixes (PSSMs) from PSI-BLAST] or features predicted by other tools, such as secondary structure and solvent accessibility. Representation models such as transformers can be used to learn features from large volumes of unlabelled data, aiming to represent the innate structure of the data. Several pretrained representation models are available for proteins (e.g., CPCProt, DeepGraphGO, ESM-1b, ProtTrans/ProtBert, rawMSA, SeqVec, GearNet, UniRep, AntiBerty and(p81)), and we recommend carefully evaluating these, depending on the specific task at hand.(p82),(p83),(p84),(p85),(p86),(p87),(p88),(p89),(p90),(p91),(p92),(p93),(p94)

Feature selection is important when dealing with protein sequence-based features because many features can be redundant. Feature selection offers several advantages, including a decrease in the overall number of tuneable parameters in the algorithm, reducing the likelihood of overfitting. A reduced number of input features can also increase the algorithm's speed, which is crucial for large-scale applications. Most importantly, a concise list of relevant features aids in understanding the essential characteristics of the problem at hand. Feature selection can be categorised into three types: (i) 'wrappers' use the ML algorithm as a black box to select features based on their performance; (ii) 'filters' select feature subsets without considering the ML algorithm; and (iii) 'embedded' techniques are part of the ML algorithm training procedure.(p95) Choosing the correct representation for the task at hand is crucial and is typically more important than the choice of the model used. It is recommended to evaluate a range of representations in combination with simpler models to identify the most suitable approach for the specific problem. In drug discovery, simple models have repeatedly been shown to outperform more complex ones,(p96),(p97),(p98) and should hence be used at least as a baseline before moving on to more complex ones such as deep learning.(p99),(p100)

Feature scaling

Real-world data sets often contain features that vary in degrees of magnitude, range and/or units. In order for ML models to interpret these features on the same scale, we hence need to perform a step called feature scaling that involves standardising the range of feature values to ensure that no single feature dominates the model. Common techniques include normalisation (scaling features to a range of 0 to 1 or −1 to 1) and standardisation (scaling features to have zero mean and unit variance).(p101),(p102)

Data transformation

Apply transformations to the data to reduce dimensionality, enhance interpretability or improve model performance. Examples include principal component analysis (PCA), t-distributed stochastic neighbour embedding (t-SNE) and log transformation. Dimensionality reduction methods can be used to reduce the number of features and hence reduce training times. They can also be used to assess the importance of features, for sense checks, to confirm biological hypotheses or to act as regularisation.(p103),(p104)

Simulations

There is a lack of large-scale ground truth experimental data. This hinders the development and benchmarking of robust and interpretable ML approaches.(p105),(p106) To address this problem, there is a need to complement analyses on experimental data with simulated ground-truth data. The challenge is to generate simulated data, such that they incorporate key features observed in experimental repertoires that render ML problems challenging. Simulation frameworks for antibodies range from VDJ-recombination-like antibody generation(p107),(p108),(p109) to synthetic antibody–antigen structures.(p110) Together, these tools allow for large-scale, high-throughput and real-world relevant synthetic data generation. The here-cited simulation tools have been tested for nativeness *vis-à-vis* experimental data. The extension of experimental observations via simulations can also help when exploring deeper correlations among different attributes, such as the aggregation and immunogenicity of antibody-based therapeutics. For example, molecular dynamics trajectories could be used as inputs to ML models to enable better prediction of molecular properties such as binding, similar to small molecules.(p111),(p112) Publicly available data sets are required for the comparison of methods based on a single agreed-upon data sets.(p13) The OAS (Observed Antibody Space)(p113) and iReceptor databases(p114) represent starting points for the integration of novel data sets that are associated with function metadata. These data sets could be set for different ML tasks, ranging from antibody structure prediction and antibody–antigen docking to antibody developability prediction. These data sets would not only represent a data standard, but would also be necessary building blocks for public competitions.(p39),(p115),(p116),(p117),(p118),(p119) Competitions are integral to mapping both those areas where predictability is good and those where blank spots exist in our knowledge.

Exploratory data analysis

Exploratory data analysis (EDA) is an essential step in the ML process. It allows for a better understanding of the data and helps to inform subsequent modelling decisions. In this section, we discuss several aspects of EDA in the context of drug discovery.

Assessing property distributions

This step involves analysing the distribution of biophysical properties, such as activation (e.g., IC_{50} or K_i value range), potency, selectivity, positional amino acid frequency, antibody topology or other relevant features in the data set. This analysis can help to identify potential biases, outliers or trends that might affect the model's performance. It will also affect the model's prediction ability and enable determination if additional preprocessing or normalisation steps are required.

Coverage of the target space and dynamic range of the data

EDA must include an analysis of the coverage of the target space (the final outputs of inputs) and dynamic range of the data. Evaluate the data's coverage of the target space to ensure that the model can make accurate predictions across the entire range. Assess the dynamic range of the data, because small ranges can lead to poor model performance. For example, if the selectivity ranges are small, it is unlikely that the model will be able to make reliable predictions far outside these ranges (c.f. 'Model applicability domain' below). When dealing with very small dynamic ranges, the question of sufficient resolution of the corresponding assay might arise. If the underlying assay is not able to distinguish variants with diverse properties in a significant manner, computational methods built on top of such data will probably fail as well. The dynamic range of a data set can have a large impact on the apparent correlation between experimental and predicted activity, and the literature is full of examples of what seem to be impressive correlations on data sets that span an unrealistically high range. So, when testing a model, it is important to ensure that the evaluation range is representative of the application it is intended for. When data within this typical range are considered, these apparent correlations can decrease dramatically.(p120)

Evaluating data imbalance

This step involves assessing the balance between different classes or ranges of values in the data set, because imbalanced data might negatively affect the performance of ML models. Techniques such as re- or oversampling,(p121),(p122),(p123) undersampling,(p121),(p122),(p123) changing the decision threshold (for classification)(p124) or using weighted loss functions(p123) can help to address this issue and should be considered where appropriate.(p125),(p126),(p127),(p128)

Model applicability domain

It might be possible to evaluate the applicability domain of the model on the basis of the similarity of training to test/production data or uncertainty.(p129),(p130) For similarity, consider factors such as the similarity of the training set property range to the target property range, input sequence similarity or clustering representations. Keep in mind that sequence similarity does not always imply phenotypic similarity, because sequence-similar antibodies might bind to different antigens. The applicability domain should in particular be considered when deploying generative models,(p131) because these can easily exploit weaknesses of the scoring functions.(p132),(p133)

Correlation and cluster analysis

Perform correlation and cluster analysis to gain insights into the relationships between variables, to identify potential outliers or patterns, and to inform the choice of features and models. This information can be valuable for improving model performance and interpretability. Additional methods for outlier detection should also be considered (e.g.,(p134)).

Choosing the correct model loss function and performance metrics

Selecting a set of appropriate metrics is crucial for assessing the performance of ML models in drug discovery.(p135) This section discusses various metrics for regression and classification tasks and when to apply them. Rather than being comprehensive, we provide a flavour of the variety of methods and refer to literature for further reading.

It is important to note that metrics are different from loss functions. Loss functions measure the model performance during training and are used to optimise ML models by minimising the loss in order to derive the optimal performing model. The loss function hence usually needs to be differentiable (i.e., a gradient can be calculated) with respect to the model's parameters. Conversely, metrics are used to monitor and measure the performance of a model both during training and testing. Not all metrics are differentiable, in particular if they are only used for the final evaluation of the model and are not used for training. Although there is a substantial overlap between the two, here we focus on the model performance metric.(p136) For the final model evaluation, we typically look at multiple different metrics, whereas for the loss function, typically a single one is chosen, although additional terms can be added for regularisation.[p137] Both need to be chosen carefully alongside the optimiser.

Regression metrics

It is crucial to evaluate regression models using multiple metrics, including Pearson and Spearman correlation coefficients and Kendall's tau. Combinations of different metrics can provide insights into model performance because they highlight the strengths and weaknesses in different regimes. Examples for common metrics are covered in references.[p18],[p20],[p138] Correlation coefficients have an associated error, dependent on both the correlation coefficient itself and the number of data points that are used to obtain it. It is hence good practice to evaluate confidence intervals for the correlation coefficients, in particular when comparing these across different models. There is only a clear advantage if the confidence intervals of two correlations do not overlap. However, the existence of an overlap does not necessarily imply a lack of difference between the two models.[p139]

Classification metrics

For classification tasks, it is important to select the appropriate threshold for optimal performance. There is again a range of different metrics available for classification models. In general, we advise using several of them, including (balanced) accuracy, precision, recall, F1 score and receiver operating characteristic AUC (ROC-AUC). For all of these, it is important to carefully choose which metric to use in which circumstance. For example, ROC-AUC should be used for balanced data, because it measures the trade-off between true positive rate and false positive rate,[p138] but it should be avoided for imbalanced data.[p140] ROC-AUC can summarise the performance, with perfect classifiers having an AUC of 1 and a random one having an AUC of 0.5. It is another measure of model calibration because it assesses model performance across all possible decision boundaries and is directly related to the Mann–Whitney statistic.[p141] Other metrics include partial AUC, precision-recall AUC (PR-AUC), Matthews correlation coefficient (MCC) and Cohen's kappa.

For multiclass predictions, metrics such as MCC or extensions of the above scores can be used, but require additional care.[p142],[p143] Choosing the metric with care is crucial because these can otherwise be misleading, and Cohen's kappa, for example, should be avoided.[p144] In addition, when predicting probabilities (rather than classes), other metrics, such as the Kullback–Leibler divergence, need to be chosen.[p145]

Model components and model choice

Choosing the right ML model for drug-discovery projects is crucial for achieving the desired results. This section discusses various aspects of ML models, from data considerations to model types and best practices.

General data and program-specific data

Depending on the data-set size and problem scope, you can use pretrained or multiclass models or train new models from scratch. When you have a large, diverse data set that covers various aspects of the problem, pretrained models can be fine-tuned for your specific application. For smaller, domain-specific data sets or unique problems, training a new model might be necessary.

Model requirements

Consider the specific goals of the drug-discovery project and the characteristics of the data when selecting a model. This includes interpretability, computational resources for training and the model's ability to generalise to new data.

Data preprocessing

Data preprocessing techniques such as StandardScaler, MinMaxScaler and PowerTransformer can be crucial for achieving good results.[p146],[p147],[p148] StandardScaler removes the mean and scales features to unit variance, whereas MinMaxScaler scales features between a specified range. PowerTransformer applies a power transformation to make data more Gaussian-like. Standardising data is important for many ML estimators because they might perform poorly if the individual features do not resemble standard, normally distributed data.

Model types

Depending on the problem, you might want to use a different set of descriptors (e.g., structure, sequence or physicochemical descriptors). Structure-based models use 3D molecular structures, sequence-based models focus on the amino acid sequences and descriptor-based models use calculated features to represent molecules. You can use different descriptors with different models, and in many cases, different descriptor-model combinations will result in different performance. Automating the process so that you can test a variety of combinations is hence generally useful and a best practice.

Dummy models

Start with simple baselines to evaluate the performance of more complex models. The simplest models are dummy models, which for example can just make random predictions or majority class predictions. These are helpful to establish an absolute minimum baseline to beat.

Adversarial validation

Adversarial validation[p149],[p150] is a technique used to assess the degree of similarity between training and testing data sets in terms of feature distribution. A common use of this approach is to determine the similarity of training and test sets which can be used to indicate whether conventional validation techniques should work well or to invalidate wrong model hypotheses.[p151] A related method is y-scrambling,[p152] which can also be used to test the robustness of the model predictions and overfitting.

Conventional ML models (random forest, support vector machine, etc.)

Use these models as a starting point before exploring more complex (deep learning) models. These will usually be easy and quick to implement and will give you a sense if there is a signal in the data. They will also allow you to meaningfully assess any improvements of more complex models and are typically more robust.

Deep learning models

Consider using deep learning models when you have a large, complex data set and the problem requires advanced feature extraction. Certain types of architecture, such as equivariant neural networks, can be used even if little data are available. Protein language models also belong to this category and can be fine-tuned for specific tasks (c.f. previous section on encodings).

Physics-based models and simulations

These models can be used when you have structural information and require a more detailed understanding of the interaction mechanisms. A wide range of tools, such as Rosetta,[(p153),(p154)] ZDOCK[(p155)] and HDOCK,[(p156)] are available and should be chosen depending on the problem.[(p157),(p158)]

Best practices

Combining models into ensembles

Ensemble methods can improve model performance by combining the strengths of multiple models. This is particularly the case if you combine different types of models (not the same type with different initialisation), such as physics-based with conventional ones.

Common pitfalls and best practices

Always validate your models using appropriate metrics, avoid overfitting and tune hyperparameters carefully. It is also essential to use a correct data split and avoid data leakage.[(p159)] This should be carefully tested beforehand. Data leakage in particular has become a large problem in computational protein structure prediction[(p160)] and needs to be carefully considered, because sequence alone is not indicative of good data splits, and data leakage and protein homology also need to be taken into account (c.f. section 9).

In summary, choosing the right ML model for drug-discovery projects involves considering various factors such as data size, model requirements, preprocessing and best practices. Always start with simple models and move towards more complex models as necessary, keeping in mind the specific goals and characteristics of your drug-discovery project.

Evaluation

Model evaluation is a crucial aspect of drug-discovery projects that ensures the chosen models are effective and accurate.[(p135),(p151)] This section delves into various components of model evaluation, from validation to data splits and feature analysis.

Process validation versus model validation

In drug discovery, every model validation is a process validation rather than just a model validation.[(p30)] This is because the entire process, including data preprocessing, feature selection and model training, contributes to the overall performance of the model. Therefore, it is essential to validate and optimise the entire process end-to-end to ensure the best possible results.

Baselines

Establishing baselines, such as dummy or minimal (e.g., linear) models is crucial for evaluating the effectiveness of complex models. By comparing the performance of the proposed model with existing methods, it is possible to demonstrate the superiority of the new model and justify its use in the drug-discovery project.

Metrics

As discussed earlier, choosing the correct metrics for model evaluation is crucial. Ensure that the selected metrics are appropriate for the specific drug-discovery problem, the model and the task, and that they also adequately capture the desired aspects of model performance.

Significance testing

An important part of model evaluation is the comparison between different tested models and whether there is a significant difference in their performance. For comparison of the two models, it is recommended to use McNemar's test in cases where there are a limited amount of data and each algorithm can only be evaluated once[(p161)] or a resampling method with 10 × 10-fold cross-validation with a corrected paired Student's *t*-test.[(p162),(p163)] Comparing multiple models at once requires different tests, such as the Holm–Bonferroni method, a Wilcoxon signed rank test with adjustment for multiple testing or the Friedman two-way analysis of variance by ranks test (Friedman's test for short).[(p164),(p165),(p166)] Although it is generally recommended to use adjustments of the p-values for comparison of multiple models, there is no consensus as to which method should be used. We refer the reader to[(p167)] for a practical tutorial and note that looking at the overlap of error bars is insufficient.[(p168),(p169)]

Model performance versus program impact

Although academic and research groups are mainly interested in improving the model quality, in real-world drug-discovery scenarios, the real impact comes from the improved quality of decision making, leading to the generation of better drug candidates. The performance of a model should hence be evaluated in the context of its impact on the drug-discovery program. In many cases, a few percentage points of uplift in the model performance is irrelevant given that the noise in the data is usually high and the models simply overfit. Getting a small amount of extra, high-quality data would usually result in much better outcomes and should be considered as an alternative to working for weeks on new or better models.

Drug-discovery-specific metrics

There are additional metrics for ML-driven discovery processes. Although these are likely to be less familiar to an ML expert, they can be used to evaluate the model's ability to discover new drugs.

These include, for example, 'enrichment', which is defined as the fraction of active compounds (e.g., binders) in a selected subset of compounds compared to the fraction of active compounds in a randomly chosen subset.[(p170),(p171),(p172)] A related metric is the normalised discounted cumulative gain (NDCG), which indicates whether the top predicted results are enriched for truly high performers.[(p173)] In contrast to the Spearman correlation, NDCG weights the ranking of the top of the list higher.

Data splits

Defining data splits that better assess real model performance is crucial for accurate evaluation. Predictive models are not universally applicable, and they generally perform better when predicting the activity of molecules that are similar to those in the training set, and perform worse if the molecules are too dissimilar.(p174) In our experience, simple data splits rarely reflect the true prospective model performance in a drug-discovery program.(p175) For this reason, a set of different data splits with varying degrees of difficulty typically allows one to get a better sense of the true prospective model performance. For proteins, we recommend testing several of the below data-set splitting approaches. Various data splits based on sequence identity can be used for protein models,(p19) such as variation in sequence, mutation location, number of mutations, amino acids, physicochemical properties of the sequence and property ranges. It is also useful to utilise the concept of adversarial validation(p149) to establish whether training and test data are easy to distinguish and the splits are done well.

Consider using 80/20 or other fixed splits, cross-validation, time splits (e.g., campaign-based),(p176) simulated time splits(p177), or cluster-based splits. In general, a variety of different data splits with increasing difficulty should be chosen. Of particular importance when splitting the data is to check for any data leakage. This can happen if proteins with high sequence identity or homology are present in the data set, and it is hence essential to check the similarities among training, test and validation sets. A typical approach that is chosen by many researchers is to use a threshold of 25–30% for sequence identity of the training set proteins to any protein in the test set. This is enough to exclude many homologous pairs of proteins, but it is well-known(p38),(p178),(p179) that some homologous proteins can have almost no sequence similarity. Such challenges can also be addressed by using additional tools such as CD-Hit,(p180) BlastClust(p181) or TESE.(p182) Preprocessing the sequences into domains using tools such as PFAM(p183) can be an option as well.

However, for antibodies or related scaffolds, sequence identity thresholds of 25–30% need some further consideration, because the constant regions of antibodies are highly similar, and diversity is mainly limited to the Fv region and complementarity-determining region (CDR). Custom data splits (as discussed in the next section) that take the program or project-specific sequence diversity and design into account are often the methods of choice. Other types of data leakage are of course also possible: for example, the availability of a feature during training that is not available during testing or in the application. We refer the reader to the literature (e.g.,(p184)).

In antibody D&D, one typically engages one of the following data forms: (i) mutational variants; (ii) different 'wild types' (no mutational relationship; directed against same or different targets); or (iii) a mix of the above. For 'general' models such as for expression or stability, (iii) is the most common case, whereas (i) and (ii) are common in program-specific predictions such as function during lead identification (ii) or optimization of lead candidates (i).

For data representative for cases (i) and (iii), generic splits should be avoided, and other approaches should be sought depending on the application and objective. For example, training and test splits should consider meaningful distributions of mutational positions, amino acid type and biophysical properties (polar *versus* hydrophobic, etc.). For data of the form (i) and (iii), data leakage can become a problem, and random splits should be avoided because they are likely to result in overestimation of the performance and a lack of generalisation abilities. In general, we always advise understanding the data set (e.g., the origin and purpose of the sequence designs) at hand and forming a clear definition of similarity and the prediction goals in order to create data splits that are representative of real-world performance. One last point is related to the hyperparameters. No parameter should be selected on the basis of the test data, and this includes hyperparameters. A simple approach to guard against overestimation due to choice of the hyperparameters based on the test set is to introduce a third validation set and select hyperparameters based on it before testing it on the test set (ideally using cross-validation).

With regards to the impact of training data composition, there is emerging evidence that the choice of negative data can affect prediction accuracy and generalisation in both antibodies(p110),(p185),(p186) and T-cell receptor (TCR) specificity prediction.(p187),(p188),(p189),(p190) Therefore, care must be taken as to how negative and positive data sets are defined.

For each model and metric, there is a trade-off between false-positives and false-negatives. For each project and problem, there will be specific choices, and these need to be informed by the overall objectives. Clearly understanding the trade-offs will enable us to obtain optimal performance.

Feature analysis and interpretation of results

Assessing the importance of features helps in understanding their contribution to the model's performance and in refining the model. For example, one can use p-values to assess feature significance, or other methods such as GINI impurity for random forest models or Shapley additive explanations (SHAP) analysis.(p191)

In summary, model evaluation is a multifaceted process that involves validation, benchmarking, metric selection, performance assessment and feature analysis. By carefully considering each aspect, researchers can ensure that the chosen models are accurate and effective, ultimately leading to more successful drug-discovery projects.

Conclusions

This review establishes a set of best practices, primarily focused on robust data generation and capture, and model building. The focus on practical considerations ensures that ML applications not only accelerate the R&D process, but also contribute to the development of safer and more effective biotherapeutics. Overall, by adhering to best practices and robust validation approaches, the field can progress to produce higher-quality antibodies, thus offering better therapeutic options and meeting unmet medical needs. Future work should continue to emphasise the importance of robust end-to-end processes, from data generation and storage to model validation and deployment.

Funding

The Leona M. and Harry B. Helmsley Charitable Trust (#2019PG-T1D011, to V.G.), UiO World-Leading Research Community (to V.G.), UiO: LifeScience Convergence Environment Immunolingo (to V.G.), European Union (EU) Horizon 2020 iReceptorplus (#825821) (to V.G.), a Norwegian Cancer Society Grant (#215817, to V.G.) and Research Council of Norway projects

(#300740, #311341, #331890, to V.G.). This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 101007799 (Inno4Vac). This Joint Undertaking receives support from the EU's Horizon 2020 research and innovation programme and the European Federation of Pharmaceutical Industries and Associations (EFPIA) (to V.G.). This communication reflects the author's view, and neither the Innovative Medicines Initiative nor the EU, EFPIA or any associated partners are responsible for any use that might be made of the information contained therein.

Conflicts of interest

V.G. declares advisory board positions in aiNET GmbH, Enpicom B.V, Absci, Omniscope and Diagonal Therapeutics. V.G. is a consultant for Adaptyv Biosystems, Specifica Inc, Roche/Genentech, immunai, Proteinea, LabGenius and Fairjourney Biologics.

L.W. is an employee of LabGenius Ltd, a biotech company developing next-generation antibody therapeutics. S.K. is an employee at Moderna Therapeutics, a pharmaceutical and biotechnology company based in Cambridge, Massachusetts, that focuses on RNA therapeutics. N.F. is an employee of Sanofi S.A., a multinational pharmaceutical and health-care company. A.B. is an employee of AstraZeneca, a patient-focused pharmaceutical company.

CRediT authorship contribution statement

Leonard Wossnig: Writing – review & editing, Writing – original draft, Conceptualization. **Norbert Furtmann:** Writing – review & editing, Writing – original draft, Conceptualization. **Andrew Buchanan:** Writing – review & editing, Writing – original draft, Conceptualization. **Sandeep Kumar:** Writing – review & editing, Writing – original draft, Conceptualization. **Victor Greiff:** Writing – review & editing, Writing – original draft, Conceptualization.

Data availability

No data was used for the research described in the article.

Acknowledgements

We would like to thank Gino van Heeke and Wade Davis for their detailed comments and careful review of the manuscript, and Lucy Shaw for reviewing the manuscript and the creation of the figures for this article. We would also like to thank Marion Schneider and Eduard Kober for generating some of the figures.

References

1. Senior M. Fresh from the biotech pipeline: fewer approvals, but biologics gain share. *Nat Biotechnol*. 2023;41:174–182.
2. Wang Y, Yang S. Multispecific drugs: the fourth wave of biopharmaceutical innovation. *Signal Transduct Target Ther*. 2020;5:86.
3. Durán CO et al. Implementation of digital health technology in clinical trials: the 6R framework. *Nat Med*. 2023;29:2693–2697.
4. Paul SM et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Discov*. 2010;9:203–214.
5. Schlander M et al. How much does it cost to research and develop a new drug? A systematic review and assessment. *PharmacoEconomics*. 2021;39:1243–1269.
6. Wouters OJ et al. Estimated research and development investment needed to bring a new medicine to market, 2009–2018. *J Am Med Assoc*. 2020;323:844–853.
7. Morgan S et al. The cost of drug development: a systematic review. *Health Policy*. 2011;100:4–17.
8. Kelley B. Developing therapeutic monoclonal antibodies at pandemic pace. *Nat Biotechnol*. 2020;38:540–545.
9. Akbar R et al. Progress and challenges for the machine learning-based design of fit-for-purpose monoclonal antibodies. *mAbs*. 2022;14, 2008790.
10. Narayanan H et al. Machine learning for biologics: opportunities for protein engineering, developability, and formulation. *Trends Pharmacol Sci*. 2021;42:151–165.
11. Glatt S et al. First-in-human randomized study of bimekizumab, a humanized monoclonal antibody and selective dual inhibitor of IL-17A and IL-17F, in mild psoriasis. *Br J Clin Pharmacol*. 2017;83:991–1001.
12. Bauer J et al. How can we discover developable antibody-based biotherapeutics? *Front Mol Biosci*. 2023;10, 1221626.
13. Mock M et al. AI can help to speed up drug discovery—but only if we give it the right data. *Nature*. 2023;621:467–470.
14. Bender A, Cortés-Ciriano I. Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 1: ways to make an impact, and why we are not there yet. *Drug Discov Today*. 2021;26:511–524.
15. Fernández-Quintero ML et al. Assessing developability early in the discovery process for novel biologics. *mAbs*. 2023;15, 2171248.
16. Bender A et al. Evaluation guidelines for machine learning tools in the chemical sciences. *Nat Rev Chem*. 2022;6:428–442.
17. Lee BD et al. Ten quick tips for deep learning in biology. *PLoS Comput Biol*. 2022;18, e1009803.
18. Lones MA. How to avoid machine learning pitfalls: a guide for academic researchers. *arXiv*. 2021. https://doi.org/10.48550/arxiv.2108.02497.
19. Walsh I et al. Correct machine learning on protein sequences: a peer-reviewing perspective. *Brief Bioinform*. 2015;17:831–840.
20. Greener JG et al. A guide to machine learning for biologists. *Nat Rev Mol Cell Biol*. 2022;23:40–55.
21. Sapoval N et al. Current progress and open challenges for applying deep learning across the biosciences. *Nat Commun*. 2022;13:1728.
22. Johnston KE et al. Machine learning for protein engineering. *arXiv*. 2023. https://doi.org/10.48550/arxiv.2305.16634.
23. Xu Y et al. Deep dive into machine learning models for protein engineering. *J Chem Inf Model*. 2020;60:2773–2790.
24. Kouba P et al. Machine learning-guided protein engineering. *ACS Catal*. 2023;13:13863–13895.
25. Bergström F, Lindmark B. Accelerated drug discovery by rapid candidate drug identification. *Drug Discov Today*. 2019;24:1237–1241.
26. Austin M et al. Structural and functional characterization of C0021158, a high-affinity monoclonal antibody that inhibits arginase 2 function via a novel non-competitive mechanism of action. *mAbs*. 2020;12, 1801230.
27. Rossant CJ et al. Phage display and hybridoma generation of antibodies to human CXCR2 yields antibodies with distinct mechanisms and epitopes. *mAbs*. 2014;6:1425–1438.
28. Furtmann N et al. An end-to-end automated platform process for high-throughput engineering of next-generation multi-specific antibody therapeutics. *mAbs*. 2021;13, 1955433.
29. Rodrigues T. The good, the bad, and the ugly in chemical and biological data for machine learning. *Drug Discov Today Technol*. 2019;32:3–8.
30. Bender A, Cortes-Ciriano I. Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 2: a discussion of chemical and biological data. *Drug Discov Today*. 2021;26:1040–1052.
31. Geiger RS et al. "Garbage in, garbage out" revisited: What do machine learning application papers report about human-labeled training data? *Quant Sci Stud*. 2021;2:795–827.
32. Fourches D et al. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J Chem Inf Model*. 2010;50:1189–1204.
33. Fourches D. Trust, but verify II: a practical guide to chemogenomics data curation. *J Chem Inf Model*. 2016;56:1243–1252.
34. Littmann M et al. Validity of machine learning in biology and medicine increased through collaborations across fields of expertise. *Nat Mach Intell*. 2020;2:18–24.

35. Jiao Y, Du P. Performance measures in evaluating machine learning based bioinformatics predictors for classifications. *Quant Biol*. 2016;4:320–330.
36. Erickson BJ, Kitamura F. Magician's corner: 9. Performance metrics for machine learning models. *Radiol Artif Intell*. 2021;3, e200126.
37. Vishwakarma G et al. Metrics for benchmarking and uncertainty quantification: quality, applicability, and best practices for machine learning in chemistry. *Trends Chem*. 2021;3:146–156.
38. Söding J, Remmert M. Protein sequence comparison and fold recognition: progress and good-practice benchmarking. *Curr Opin Struct Biol*. 2011;21:404–411.
39. Won J et al. Assessment of protein model structure accuracy estimation in CASP13: challenges in the era of deep learning. *Proteins*. 2019;87:1351–1360.
40. Bashour H et al. Biophysical cartography of the native and human-engineered antibody landscapes quantifies the plasticity of antibody developability. *bioRxiv*. 2023. https://doi.org/10.1101/2023.10.26.563958.
41. Scannell JW et al. Predictive validity in drug discovery: what it is, why it matters and how to improve it. *Nat Rev Drug Discov*. 2022;21:915–931.
42. Minot M, Reddy ST. Meta learning improves robustness and performance in machine learning-guided protein engineering. *bioRxiv*. 2023. https://doi.org/10.1101/2023.01.30.526201.
43. Pavlović M et al. Improving generalization of machine learning-identified biomarkers with causal modeling: an investigation into immune receptor diagnostics. *arXiv*. 2022. https://doi.org/10.48550/arXiv.2204.09291.
44. Kolmar SS, Grulke CM. The effect of noise on the predictive limit of QSAR models. *J Cheminform*. 2021;13:92.
45. Li G et al. Performance of regression models as a function of experiment noise. *Bioinform Biol Insights*. 2021;15, 11779322211020316.
46. Brown SP et al. Healthy skepticism: assessing realistic model performance. *Drug Discov Today*. 2009;14:420–427.
47. Campbell RM. Data standardization for results management. In: Markossian S, ed. *Assay Guidance Manual*. Eli Lilly & Company and the National Center for Advancing Translational Sciences; 2004.
48. Schisterman EF et al. The limitations due to exposure detection limits for regression models. *Am J Epidemiol*. 2006;163:374–383.
49. Lubin JH et al. Epidemiologic evaluation of measurement data in the presence of detection limits. *Environ Health Perspect*. 2004;112:1691–1696.
50. Anger LT et al. Generalized workflow for generating highly predictive in silico off-target activity models. *J Chem Inf Model*. 2014;54:2411–2422.
51. Tropsha A. Best practices for QSAR model development, validation, and exploitation. *Mol Inform*. 2010;29:476–488.
52. Young D et al. Are the chemical structures in your QSAR correct? *QSAR Comb Sci*. 2008;27:1337–1345.
53. OECD. Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models. *OECD Series on Testing and Assessment*. 2014. https://doi.org/10.1787/9789264085442-en.
54. Muratov EN et al. QSAR without borders. *Chem Soc Rev*. 2020;49:3525–3564.
55. Apiletti D et al. Data cleaning and semantic improvement in biological databases. *J Integr Bioinform*. 2006;3:219–229.
56. Chicco D. Ten quick tips for machine learning in computational biology. *BioData Min*. 2017;10:35.
57. Walsh I et al. DOME: recommendations for supervised machine learning validation in biology. *Nat Methods*. 2021;18:1122–1127.
58. Jones DT. Setting the standards for machine learning in biology. *Nat Rev Mol Cell Biol*. 2019;20:659–660.
59. Xu C, Jackson SA. Machine learning and complex biological data. *Genome Biol*. 2019;20:76.
60. Shugay M et al. Towards error-free profiling of immune repertoires. *Nat Methods*. 2014;11:653–655.
61. Pavlović M et al. The immuneML ecosystem for machine learning analysis of adaptive immune receptor repertoires. *Nat Mach Intell*. 2021;3:936–944.
62. Breden F et al. Reproducibility and reuse of adaptive immune receptor repertoire data. *Front Immunol*. 2017;8:1418.
63. Christley S et al. The ADC API: a web API for the programmatic query of the AIRR data commons. *Front Big Data*. 2020;3:22.
64. Community TA et al. Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data. *Nat Immunol*. 2017;18:1274–1278.
65. Heiden JAV et al. AIRR Community standardized representations for annotated immune repertoires. *Front Immunol*. 2018;9:2206.
66. Mhanna V et al. Adaptive immune receptor repertoire analysis. *Nat Rev Methods Prim*. 2024;4:6.
67. Kramer C et al. The experimental uncertainty of heterogeneous public K_i data. *J Med Chem*. 2012;55:5165–5173.
68. Kramer C et al. A comprehensive company database analysis of biological assay variability. *Drug Discov Today*. 2016;21:1213–1221.
69. Kalliokoski T et al. Comparability of mixed IC_{50} data – a statistical analysis. *PLoS One*. 2013;8:e61007.
70. Aldeghi M et al. Roughness of molecular property landscapes and its impact on modellability. *J Chem Inf Model*. 2022;62:4660–4671.
71. Parks C et al. An analysis of proteochemometric and conformal prediction machine learning protein-ligand binding affinity models. *Front Mol Biosci*. 2020;7:93.
72. Jain T et al. Biophysical properties of the clinical-stage antibody landscape. *Proc Natl Acad Sci USA*. 2017;114:944–949.
73. Jain T et al. Identifying developability risks for clinical progression of antibodies using high-throughput *in vitro* and *in silico* approaches. *mAbs*. 2023;15:2200540.
74. Bellamy H et al. Batched Bayesian optimization for drug design in noisy environments. *J Chem Inf Model*. 2022;62:3970–3981.
75. Wang D et al. A statistical framework for assessing pharmacological responses and biomarkers using uncertainty estimates. *eLife*. 2020;9:e60352.
76. Kawashima S et al. AAindex: amino acid index database. *Nucleic Acids Res*. 1999;27:368–369.
77. Georgiev AG. Interpretable numerical descriptors of amino acid space. *J Comput Biol*. 2009;16:703–723.
78. Wittmann BJ et al. Informed training set design enables efficient machine learning-assisted directed protein evolution. *Cell Syst*. 2021;12:1026–1045.e7.
79. Chen W et al. PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. *Bioinformatics*. 2015;31:119–120.
80. Zhang Y et al. A survey on the algorithm and development of multiple sequence alignment. *Brief Bioinform*. 2022;23:bbac069.
81. Leem J et al. Deciphering the language of antibodies using self-supervised learning. *Patterns*. 2022;3, 100513.
82. Fenoy E et al. Transfer learning in proteins: evaluating novel protein learned representations for bioinformatics tasks. *Brief Bioinform*. 2022;23:bbac232.
83. Alley EC et al. Unified rational protein engineering with sequence-based deep representation learning. *Nat Methods*. 2019;16:1315–1322.
84. Brandes N et al. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics*. 2022;38:2102–2110.
85. Rives A et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci USA*. 2021;118, e2016239118.
86. Wu Z et al. Protein sequence design with deep generative models. *Curr Opin Chem Biol*. 2021;65:18–27.
87. Li L et al. Machine learning optimization of candidate antibody yields highly diverse sub-nanomolar affinity antibody libraries. *Nat Commun*. 2023;14:3454.
88. Choi Y. Artificial intelligence for antibody reading comprehension: AntiBERTa. *Patterns*. 2022;3, 100535.
89. Dounas A et al. Learning immune receptor representations with protein language models. *arXiv*. 2024. https://doi.org/10.48550/arxiv.2402.03823.
90. You R et al. DeepGraphGO: graph neural network for large-scale, multispecies protein function prediction. *Bioinformatics*. 2021;37:i262–i271.
91. Heinzinger M et al. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinform*. 2019;20:723.
92. Lu AX et al. Self-supervised contrastive learning of protein representations by mutual information maximization. *bioRxiv*. 2020. https://doi.org/10.1101/2020.09.04.283929.
93. Mirabello C, Wallner B. rawMSA: end-to-end deep learning using raw multiple sequence alignments. *PLoS One*. 2019;14:e0220182.
94. Ruffolo JA et al. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Nat Commun*. 2023;14:2389.
95. Dash M, Liu H. Feature selection for classification. *Intell Data Anal*. 1997;1:131–156.
96. van Tilborg D et al. Exposing the limitations of molecular machine learning with activity cliffs. *J Chem Inf Model*. 2022;62:5938–5951.
97. Janela T, Bajorath J. Rationalizing general limitations in assessing and comparing methods for compound potency prediction. *Sci Rep*. 2023;13:17816.
98. Hsu C et al. Learning protein fitness models from evolutionary and assay-labeled data. *Nat Biotechnol*. 2022;40:1114–1122.
99. Raybould MIJ et al. Five computational developability guidelines for therapeutic antibody profiling. *Proc Natl Acad Sci USA*. 2019;116:4025–4030.
100. Ahmed L et al. Intrinsic physicochemical profile of marketed antibody-based biotherapeutics. *Proc Natl Acad Sci USA*. 2021;118, e2020577118.
101. Ozsahin DU et al*Impact of Feature Scaling on Machine Learning Models for the Diagnosis of Diabetes*. IEEE; 2022:87–94.

102. Wan X. Influence of feature scaling on convergence of gradient iterative algorithm. *J Phys Conf Ser*. 2019;1213, 032021.
103. Jia W et al. Feature dimensionality reduction: a review. *Complex Intell Syst*. 2022;8:2663–2693.
104. Velliangiri S et al. A review of dimensionality reduction techniques for efficient computation. *Proc Comput Sci*. 2019;165:104–111.
105. Sandve GK, Greiff V. Access to ground truth at unconstrained size makes simulated data as indispensable as experimental data for bioinformatics methods development and benchmarking. *Bioinformatics*. 2022;38:4994–4996.
106. Chen V et al. Best practices for interpretable machine learning in computational biology. *bioRxiv*. 2022. https://doi.org/10.1101/2022.10.28.513978.
107. Marcou Q et al. High-throughput immune repertoire analysis with IGoR. *Nat Commun*. 2018;9:561.
108. Weber CR et al. immuneSIM: tunable multi-feature simulation of B- and T-cell receptor repertoires for immunoinformatics benchmarking. *Bioinformatics*. 2020;36:3594–3596.
109. Chernigovskaya M et al. Simulation of adaptive immune receptors and repertoires with complex immune information to guide the development and benchmarking of AIRR machine learning. *bioRxiv*. 2023. https://doi.org/10.1101/2023.10.20.562936.
110. Robert PA et al. Unconstrained generation of synthetic antibody-antigen structures to guide machine learning methodology for real-world antibody specificity prediction. *bioRxiv*. 2022. https://doi.org/10.1101/2021.07.06.451258.
111. Jamal S et al. Machine learning from molecular dynamics trajectories to predict caspase-8 inhibitors against Alzheimer's disease. *Front Pharmacol*. 2019;10:780.
112. Min Y et al. From static to dynamic structures: improving binding affinity prediction with a graph-based deep learning model. *arXiv*. 2022. https://doi.org/10.48550/arxiv.2208.10230.
113. Olsen TH et al. Observed Antibody Space: a diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Sci*. 2022;31:141–146.
114. Corrie BD et al. iReceptor: a platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. *Immunol Rev*. 2018;284:24–41.
115. Janin J. Welcome to CAPRI: a critical assessment of PRedicted interactions. *Proteins*. 2002;47:257.
116. Janin J. Assessing predictions of protein–protein interaction: the CAPRI experiment. *Protein Sci*. 2005;14:278–283.
117. Kryshtafovych A et al. Critical assessment of methods of protein structure prediction (CASP)—Round XIV. *Proteins*. 2021;89:1607–1617.
118. Armer C et al. The Protein Engineering Tournament: an open science benchmark for protein modeling and design. *arXiv*. 2023. https://doi.org/10.48550/arXiv.2309.09955.
119. Meysman P et al. Benchmarking solutions to the T-cell receptor epitope prediction problem: IMMREP22 workshop report. *ImmunoInformatics*. 2023;9, 100024.
120. Walters WP. What are our models really telling us? A practical tutorial on avoiding common mistakes when building predictive models. In: Bajorath J, ed. *Chemoinformatics for Drug Discovery*. John Wiley & Sons; 2013.
121. Estabrooks A et al. A multiple resampling method for learning from imbalanced data sets. *Comput Intell*. 2004;20:18–36.
122. Cao H et al. Integrated oversampling for imbalanced time series classification. *IEEE Trans Knowl Data Eng*. 2013;25:2809–2822.
123. Anand A et al. An approach for classification of highly imbalanced data using weighting and undersampling. *Amino Acids*. 2010;39:1385–1391.
124. Esposito C et al. GHOST: adjusting the decision threshold to handle imbalanced data in machine learning. *J Chem Inf Model*. 2021;61:2623–2640.
125. Haixiang G et al. Learning from class-imbalanced data: review of methods and applications. *Expert Syst Appl*. 2017;73:220–239.
126. Kaur H et al. A systematic review on imbalanced data challenges in machine learning. *ACM Comput Surv*. 2019;52:1–36.
127. Kumar P et al. Classification of imbalanced data: review of methods and applications. *IOP Conf Ser Mater Sci Eng*. 2021;1099, 012077.
128. García V. Exploring the performance of resampling strategies for the class imbalance problem. In: García-Pedrajas N, ed. *Trends in Applied Intelligent Systems. IEA/AIE 2010 (Lecture Notes in Computer Science, Vol. 6096)*. Springer; 2010:541–549.
129. Sheridan RP. The relative importance of domain applicability metrics for estimating prediction errors in QSAR varies with training set diversity. *J Chem Inf Model*. 2015;55:1098–1107.
130. Sugita S, Ohue M. Drug-target affinity prediction using applicability domain based on data density. *ChemRxiv*. 2021. https://doi.org/10.26434/chemrxiv-2021-hp2p9-v2.
131. Langevin M et al. Impact of applicability domains to generative artificial intelligence. *ACS Omega*. 2023;8:23148–23167.
132. Renz P et al. On failure modes in molecule generation and optimization. *Drug Discov Today Technol*. 2019;32:55–63.
133. Langevin M et al. Explaining and avoiding failure modes in goal-directed generation of small molecules. *J Cheminform*. 2022;14:20.
134. Motulsky HJ, Brown RE. Detecting outliers when fitting data with nonlinear regression – a new method based on robust nonlinear regression and the false discovery rate. *BMC Bioinform*. 2006;7:123.
135. Robinson MC et al. Validating the validation: reanalyzing a large-scale comparison of deep learning and machine learning models for bioactivity prediction. *J Comput Aided Mol Des*. 2020;34:717–730.
136. López OAM et al. Overfitting, model tuning, and evaluation of prediction performance*Multivariate Statistical Machine Learning Methods for Genomic Prediction*. Springer; 2022:109–139.
137. Hastie T et al. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer; 2009.
138. Ozenne B et al. The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *J Clin Epidemiol*. 2015;68:855–859.
139. Schenker N, Gentleman JF. On judging the significance of differences by examining the overlap between confidence intervals. *Am Stat*. 2001;55:182–186.
140. Davis J, Goadrich M*The Relationship between Precision-Recall and ROC Curves*. Association for Computing Machinery; 2006:233–240.
141. Xu W et al. Estimating the area under a receiver operating characteristic (ROC) curve: parametric and nonparametric ways. *Signal Process*. 2013;93:3111–3123.
142. Grandini M et al. Metrics for multi-class classification: an overview. *arXiv*. 2020. https://doi.org/10.48550/arxiv.2008.05756.
143. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. *Inf Process Manag*. 2009;45:427–437.
144. Delgado R, Tibau XA. Why Cohen's Kappa should be avoided as performance measure in classification. *PLoS One*. 2019;14:e0222916.
145. Bishop CM, Nasrabadi NM. *Pattern Recognition and Machine Learning*. Springer; 2006.
146. Raju VNG et al*Study the Influence of Normalization/transformation Process on the Accuracy of Supervised Classification*. IEEE; 2020:729–735.
147. de Amorim LBV et al. The choice of scaling technique matters for classification performance. *Appl Soft Comput*. 2023;133, 109924.
148. Patro S, Sahu KK. Normalization: a preprocessing stage. *arXiv*. 2015. https://doi.org/10.48550/arXiv.1503.06462.
149. Chuang KV, Keiser MJ. Adversarial controls for scientific machine learning. *ACS Chem Biol*. 2018;13:2819–2821.
150. Rücker C et al. y-Randomization and its variants in QSPR/QSAR. *J Chem Inf Model*. 2007;47:2345–2357.
151. Tropsha A et al. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb Sci*. 2003;22:69–77.
152. Lipiński PFJ, Szurmak P. SCRAMBLE'N'GAMBLE: a tool for fast and facile generation of random data for statistical evaluation of QSAR models. *Chem Pap*. 2017;71:2217–2232.
153. Lyskov S, Gray JJ. The RosettaDock server for local protein–protein docking. *Nucleic Acids Res*. 2008;36:W233–W238.
154. Weitzner BD et al. Modeling and docking of antibody structures with Rosetta. *Nat Protoc*. 2017;12:401–416.
155. Pierce BG et al. ZDOCK server: interactive docking prediction of protein–protein complexes and symmetric multimers. *Bioinformatics*. 2014;30:1771–1773.
156. Yan Y et al. The HDOCK server for integrated protein–protein docking. *Nat Protoc*. 2020;15:1829–1852.
157. Desta IT et al. Performance and its limits in rigid body protein-protein docking. *Structure*. 2020;28:1071–1081.e3.
158. Fan W et al. Online bioinformatics teaching practice: comparison of popular docking programs using SARS-CoV-2 spike RBD–ACE2 complex as a benchmark. *Biochem Mol Biol Educ*. 2021;49:833–840.
159. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in ML-based science. *arXiv*. 2022. https://doi.org/10.48550/arxiv.2207.07048.
160. Bernett J et al. Cracking the black box of deep sequence-based protein-protein interaction prediction. *bioRxiv*. 2023. https://doi.org/10.1101/2023.01.18.524543.

161. Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput*. 1998;10:1895–1923.
162. Nadeau C, Bengio Y. Inference for the generalization error. *Mach Learn*. 2003;52:239–281.
163. Bouckaert RR, Frank E. Evaluating the replicability of significance tests for comparing learning algorithms. In: Dai H, ed. *Advances in Knowledge Discovery and Data Mining. PAKDD 2004 (Lecture Notes in Computer Science, Vol. 3056)*. Springer; 2004:3–12.
164. Berrar D. Using p-values for the comparison of classifiers: pitfalls and alternatives. *Data Min Knowl Discov*. 2022;36:1102–1139.
165. Benavoli A et al. Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis. *J Mach Learn Res*. 2017;18:2653–2688.
166. Dem̌sar J. Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res*. 2006;7:1–30.
167. Walters WP. Comparing classification models—a practical tutorial. *J Comput Aided Mol Des*. 2022;36:381–389.
168. Nicholls A. Confidence limits, error bars and method comparison in molecular modeling. Part 1: the calculation of confidence intervals. *J Comput Aided Mol Des*. 2014;28:887–918.
169. Nicholls A. Confidence limits, error bars and method comparison in molecular modeling. Part 2: comparing methods. *J Comput Aided Mol Des*. 2016;30:103–126.
170. Bender A, Glen RC. A discussion of measures of enrichment in virtual screening: comparing the information content of descriptors with increasing levels of sophistication. *J Chem Inf Model*. 2005;45:1369–1375.
171. Lopes JCD et al. The power metric: a new statistically robust enrichment-type metric for virtual screening applications with early recovery capability. *J Cheminform*. 2017;9:7.
172. Huang N et al. Benchmarking sets for molecular docking. *J Med Chem*. 2006;49:6789–6801.
173. Järvelin K, Kekäläinen J. Cumulated gain-based evaluation of IR techniques. *ACM Trans Inf Syst*. 2002;20:422–446.
174. Sheridan RP et al. Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *J Chem Inf Comput Sci*. 2004;44:1912–1928.
175. Kearnes S. Pursuing a prospective perspective. *Trends Chem*. 2021;3:77–79.
176. Sheridan RP. Time-split cross-validation as a method for estimating the goodness of prospective prediction. *J Chem Inf Model*. 2013;53:783–790.
177. Landrum GA et al. SIMPD: an algorithm for generating simulated time splits for validating machine learning approaches. *ChemRxiv*. 2023. https://doi.org/10.26434/chemrxiv-2023-x9pjf.
178. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. *EMBO J*. 1986;5:823–826.
179. Li Y, Yang J. Structural and sequence similarity makes a significant impact on machine-learning-based scoring functions for protein–ligand interactions. *J Chem Inf Model*. 2017;57:1007–1012.
180. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. 2006;22:1658–1659.
181. Altschul SF et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997;25:3389–3402.
182. Sirocco F, Tosatto SCE. TESE: generating specific protein structure test set ensembles. *Bioinformatics*. 2008;24:2632–2633.
183. Finn RD et al. Pfam: the protein families database. *Nucleic Acids Res*. 2014;42:D222–D230.
184. Nayak SK, Ojha AC. Data leakage detection and prevention: review and research directions. In: Swain D, ed. *Machine Learning and Information Processing (Advances in Intelligent Systems and Computing)*. Springer; 2020:203–212.
185. Krützfeldt LM et al. The impact of different negative training data on regulatory sequence predictions. *PLoS One*. 2020;15:e0237412.
186. Schneider C et al. DLAB—Deep learning methods for structure-based virtual screening of antibodies. *Bioinformatics*. 2021;38:btab660.
187. Dens C et al. The pitfalls of negative data bias for the T-cell epitope specificity challenge. *bioRxiv*. 2023. https://doi.org/10.1101/2023.04.06.535863.
188. Gao Y et al. Reply to: The pitfalls of negative data bias for the T-cell epitope specificity challenge. *bioRxiv*. 2023. https://doi.org/10.1101/2023.04.07.535967.
189. Montemurro A et al. NetTCR-2.1: Lessons and guidance on how to develop models for TCR specificity predictions. *Front Immunol*. 2022;13:1055151.
190. Grazioli F et al. On TCR binding predictors failing to generalize to unseen peptides. *Front Immunol*. 2022;13, 1014256.
191. Lundberg S, Lee SI. A unified approach to interpreting model predictions. *arXiv*. 2017. https://doi.org/10.48550/arxiv.1705.07874.
192. Wilkinson MD et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016;3, 160018.
193. Yang A et al. Deploying synthetic coevolution and machine learning to engineer protein-protein interactions. *Science*. 2023;381, eadh1720.
194. Mason DM et al. Deep learning enables therapeutic antibody optimization in mammalian cells by deciphering high-dimensional protein sequence space. *bioRxiv*. 2019. https://doi.org/10.1101/617860.
195. Maloney MP et al. Negative data in data sets for machine learning training. *Org Lett*. 2023;25:2945–2947.

KEYNOTE (GREEN)